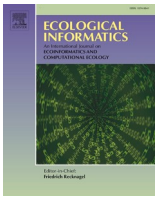
Ecological Informatics 86 (2025) 103058

Contents lists available at ScienceDirect

Ecological Informatics

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



Leveraging citizen science to classify and track benthic habitat states: An unsupervised UMAP-HDBSCAN pipeline applied to the global reef life survey dataset

Clément Violet [a,*], Aurélien Boyé [a], Stanislas Dubois [a], Graham J. Edgar [b], Elizabeth S. Oh [b], Rick D. Stuart-Smith [b], Martin P. Marzloff [a]

[a] Ifremer, Centre de Bretagne, DYNECO LEBCO, Plouzané, France
[b] Institute for Marine and Antarctic Studies (IMAS), University of Tasmania, Hobart, Tasmania, Australia

ARTICLE INFO

Keywords:
Benthic habitat
Citizen science
Clustering
Habitat states
HDBSCAN
UMAP
SHAP

ABSTRACT

Benthic biogenic habitats are crucial for coastal marine ecosystems, supporting food and shelter for a large range of marine species, but they are increasingly threatened by increasing anthropogenic impacts. While large-scale monitoring data are increasingly available, tools to describe benthic habitat changes in standardised and yet finely resolved manner are still needed. The aim of this study was to define reef benthic habitat states and explore their spatial and temporal variability on a global scale using an innovative clustering pipeline. For this purpose, we used substrate cover data collected along 6554 transects worldwide by citizen scientists contributing to the *Reef Life Survey* program. We applied an innovative clustering pipeline that combines three algorithms — *Uniform Manifold Approximation and Projection* (*UMAP*) for dimension reduction; *Hierarchical Density-Based Spatial Clustering of Applications with Noise* (*HDBSCAN*) — to identify benthic habitat states and Shapley values to interpret the clusters identified. This unsupervised pipeline identified 17 distinct clusters worldwide, representing typical temperate and tropical benthic habitats such as large canopy forming algae and branching corals, respectively, as well as transitional states between different habitat states. Temporal site-specific analyses further demonstrated the pipeline's effectiveness in capturing fine-scale habitat dynamics. By providing a standardised, scalable approach, this work enables consistent tracking of benthic habitat changes across spatial and temporal scales worldwide. This study also showcases the potential of integrating the *UMAP-HDBSCAN* pipeline with Shapley values for clustering noisy ecological data from citizen science initiatives.

1. Introduction

Benthic habitats host diverse species and communities (Sunday et al., 2017). They contribute to the functioning of marine coastal ecosystems and the services they provide to humanity (Barbier et al., 2011). These services include shoreline protection (Barbier, 2017), carbon sequestration (Fourqurean et al., 2012), and commercial fisheries (Barbier, 2017). As modifiers to abiotic substrates, foundation species, such as kelp, seagrass, and coral, engineer biogenic habitats that contribute to specific functions of coastal ecosystems (Elith and Leathwick, 2009). For instance, the three-dimensional structure of coral reefs can shelter fish assemblages from predators (Hixon and Beets, 1993); seaweed or mussel beds can buffer environmental conditions (Jurgens et al., 2022; Whitaker et al., 2023); and kelp forests are both habitat and food sources for various fish and invertebrate species (Edgar et al., 2004). Thus, changes in coastal benthic habitats have direct cascading consequences on marine ecosystem structure, functioning, and services.

As hotspots of human activity, coastal ecosystems can be adversely affected by multiple anthropogenic stressors (Halpern et al., 2019), including global climate change (Bowler et al., 2020; Burrows et al., 2014). The impact of these multiple stressors on benthic communities and ecosystems is frequently mediated by the response of biogenic habitats such as kelp, seagrass or coral (Harley et al., 2006; Rocha et al., 2015a). For example, in the vicinity of urban areas, eutrophication can induce replacement of kelp forests by turf algae (Filbee-Dexter and Wernberg, 2018; Pessarrodona et al., 2021); marine heatwaves can lead

* Corresponding author.
E-mail address: clement.violet@ifremer.fr (C. Violet).

https://doi.org/10.1016/j.ecoinf.2025.103058
Received 15 April 2024; Received in revised form 20 December 2024; Accepted 27 January 2025
Available online 31 January 2025

to coral bleaching, and, with intensification in magnitude and frequency, long-term decline in tropical coral reefs (Bellwood et al., 2004); and overfishing of herbivorous coral reef fishes can lead to macroalgae overgrowth (Hughes et al., 2007). Therefore, habitat changes are among the greatest symptoms of anthropogenic impacts on shallow marine systems, with major consequences for marine biodiversity (Rocha et al., 2015b).

As anthropogenic stressors in marine ecosystems increase and diversify (Halpern et al., 2019), habitat changes will likely become more frequent in the world's seas (Conversi et al., 2015). Detecting and anticipating future habitat changes in benthic ecosystems requires a thorough understanding of the current state and distribution of benthic habitats and characterisation of underlying drivers across multiple scales. Currently, detailed knowledge of habitat distribution is mostly local, i.e., at scales ranging from study sites (10 m - 100 m) or bay (100 m - 10 km) up to regions (10 km - 100 km) (e.g., Robert et al. (2015); Wicaksono et al. (2019)). At larger scales, habitat distribution maps are primarily based on physical, geomorphological, and biogeochemical ocean properties (e.g., Brown et al. (2011); Lecours et al. (2015); Sonnewald et al. (2020)). At such scales, habitat maps either disregard biogenic habitats or focus on a small number of specific habitat-formers (Assis et al., 2020; McKenzie et al., 2020), and thus rarely provide information on community composition. Large-scale seafloor habitat maps of either abiotic or biogenic features also tend to integrate data over large timescales (e.g. decades).

Knowledge of benthic habitat changes thus remains highly regional (see, e.g., Cattano et al. (2020)). In that context, several global studies have collated heterogeneous regional monitoring data to document changes in emblematic habitat-formers, such as seagrass spp. (Dunic et al., 2021; Waycott et al., 2009), kelp beds (Filbee-Dexter and Wernberg, 2018; Krumhansl et al., 2016), or coral reefs (Eddy et al., 2021). However, these independent studies on specific habitat forms are not sufficient to gain a comprehensive understanding of how the seafloor habitat mosaic has changed through time at a global scale in the face of anthropogenic pressures. Our understanding of current changes in the mosaic of habitats on the seafloor is impeded by the lack of a large-scale, standardised, data-driven definition, and maps describing benthic habitat and their potential states.

In this study, we aim to develop a data-driven pipeline that distinguishes different iconic benthic habitats observed spatially, and apply this to characterise stepwise changes in habitat ecological states through time. Because scientific monitoring programs are often expensive (Edgar et al., 2016) and restricted in their spatial and/or temporal cover (Rhodes et al., 2015), participatory science programmes have emerged as a valuable means to increase monitoring programme cover and resolution. In this study, we leverage the benefits of a citizen science program to characterise benthic habitat states at the global scale.

Reef Life Survey (*RLS*) relies on standardised diver-based assessment along 50-m-long transects to estimate fish and invertebrate species abundance, as well as image-based percentage cover of coastal benthic habitats (Edgar and Stuart-Smith, 2014). Estimates of habitat percentage cover have already proven useful for defining habitat states at a regional scale through the use of unsupervised machine learning techniques (Cresswell et al., 2017; Pelletier et al., 2020). However, the methods proposed in these previous studies come with a number of limitations when upscaled at a global level. In particular, the occurrence and abundance of habitat-forming species are expected to show nonlinear responses to environmental changes (Oksanen and Minchin, 2002), especially across large environmental gradients.

Consequently, we applied a new workflow that combines two algorithms selected for their effectiveness in both handling heterogeneous, noisy, high-dimensional data, such as the *RLS* dataset, and capturing non-linear relationships as commonly described in ecological datasets. Specifically, we used (1) *Uniform Manifold Approximation and Projection* (*UMAP*), a dimension reduction technique that preserves complex nonlinear structures and patterns (McInnes et al., 2020), and (2) *Hierarchical Density-Based Spatial Clustering of Applications with Noise algorithm* (*HDBSCAN*) that can identify clusters of varying shapes and sizes while filtering outlier noise (Campello et al., 2013; McInnes et al., 2017).

These two algorithms have been successfully applied across a wide range of research fields. For example, *UMAP* has been used to analyse and visualise genomic datasets from single-cell analysis (see for instance Becht et al. (2019), Funnell et al. (2022) or Packer et al. (2019)), while *HDBSCAN* has been employed for unsupervised classification of stellar objects (Logan and Fotopoulou, 2020) and clustering molecular dynamics data (Melvin et al., 2018). In combination, *UMAP* and *HDBSCAN* have proven effective in clustering solar wind data (Bloch et al., 2020) and also for clustering data from scanning transmission electron microscopes (Blanco-Portals et al., 2022). Within the field of ecology, *UMAP* has been utilised for community visualisation and as a preprocessing step prior to clustering (Milošević et al., 2022), and the combination of *UMAP* and *HDBSCAN* has been applied to identify areas of similar food web structures (Ohlsson and Eklöf, 2020). Our study presents a novel application of these algorithms to analyse benthic organisms' cover data, with the goal of classifying coastal marine habitats and potentially distinguishing between different habitat states, ultimately providing a template for tracking seafloor changes on a global scale.

In this study, our primary aim is to identify habitat states on a global scale using *UMAP* and *HDBSCAN*. Secondly, we aim to validate these groupings by thoroughly characterising the composition of each cluster and examining their biogeographical distribution. Lastly, we demonstrate the utility of our classification by investigating the temporal variability in the composition of habitat states at selected sites.

2. Materials and methods

2.1. Data

2.1.1. Reef Life Survey photoquadrat dataset

The *RLS* http://www.reeflifesurvey.com/ is a hybrid citizen science/professional researcher program that monitors reef communities around the world using scuba-diving visual census. Survey methodology, including protocols, diver training, data quality assurance and data management, is detailed by Edgar and Stuart-Smith (2014), Edgar et al. (2020) and Cooper and Oh (2023). Following a dive-based visual census of mobile megafauna along a 50-m transect, trained citizen scientists collect digital pictures of the seabed (hereafter referred to as "photoquadrats"). A total of 20 photoquadrats, which each approximately covers 0.3 m $\times$ 0.3 m, are collected every 2.5 m along the transect (Edgar et al., 2020). Photoquadrats are then visually processed using the *Squidle*+ https://squidle.org/ software and point counts annotation. A minimum of 100 point counts per transect (i.e. $\geq$ 5 point counts per photoquadrat) are thus used to estimate percentage covers for various benthic substrate categories defined based on the *CATAMI* benthic imagery classification scheme (Althaus and Hill, 2015; Edgar et al., 2020). All analyses presented in this paper rely on these mean percentage cover estimates at the 50-m transect level. To reduce biases related to high-level classification into detailed *CATAMI* groups for specific applications, we grouped the original benthic habitat categories into 24 broader categories that can overall contribute to seascape features along *RLS* transects (see Table 1 and Appendix A for descriptions and rationale). This aggregation was informed by the expertise of *RLS* specialists to better capture the range of dominant coastal substratum types present along *RLS* transects at a global scale. The majority of aggregations aligned with a 'parent node' on the *CATAMI* scheme, and were deemed necessary where data from the child nodes were patchy. With the exception of habitat-forming corals, sessile invertebrates were grouped to provide consistency between regions, and because image resolution sometimes prohibits consistent classification of these taxa, for instance ascidians and sponges can be difficult to tell apart.

Table 1
Description of the 24 categories used in this study to capture the overall diversity of habitat types sampled by *Reef Life Surveys* around the world. The 50 original *RLS* categories were grouped into these 24 categories that represent ecologically consistent groups associated with different levels of structural complexity.

Habitat Categories	Description
	Erect algae
Large canopy forming algae	*Large overstorey algae forming a canopy, including kelps or large fucoids*
Bushy Fucoid like	*Robust erect leaf-shaped brown algae*
Other Brown algae	*Thick or thin-sheet like erect brown algae*
Red algae	*Foliose erect red algae*
Green algae	*Thin-sheet like, thick, or ribbon-like, erect green algae*
	Erect calcareous algae
Geniculate coralline algae	*Red erect calcified segmented algae*
Green calcified algae	*Small calcified green algae*
	Encrusting algae
Crustose coralline algae	*Red algae forming a small calcified crust over hard substrate*
Encrusting algae	*Non-coralline algae forming a leathery crust over substrate*
	Mat-forming Algae
Filamentous algae	*Filamentous algae, epiphyte or rock-attached*
Turf algae	*Fine and mat-forming filamentous algae growing on hard substrate*
	Plant
Seagrass	*Vertical ribbon-like marine plant*
	Sessile invertebrates
Encrusting corals	*Stony corals forming a crust over hard substrate*
Branching coral	*Branching coral forming large colonies*
Foliose/Plate corals	*Stony corals forming tabular or foliaceous colonies*
Massive corals	*Stony corals characterised by large, ball- or boulder-shaped colonies with a compact structure*
Large-polyp stony corals	*Large lobed stony coral, usually free-living*
Soft corals and gorgonians	*Soft coral or gorgonian in the sub-class Octocorallia*
Calcareous hydrocorals and octocorals	*Branching or foliaceous coral-like*
Other sessile invertebrates	*Habitat-forming sessile invertebrates (e.g. sponges, ascidians, bryozoans or molluscs) excluding corals*
	Seabed Materials
Dead coral	*Dead attached coral skeleton*
Bare rocky substrate	*Bare rock*
Unconsolidated substrate	*Gravel, shell, coral rubble*
Sand	*Sand and fine sediments*

We extracted the *RLS* photoquadrat dataset on 24 January 2023. From the original 8154 transects, we removed partially scored transects. For transects annotated multiple times on various research projects, mean percentage cover estimates were considered. After fully curating the dataset, the photoquadrat dataset consisted of 6554 transects at 2249 sites worldwide. All subsequent analyses were performed at the transect level to consider local-scale variation in the state of benthic habitats.

2.2. Dimension reduction and clustering pipeline

To account for the nonlinear, high-dimensional, and complex nature of the ecological data, we combined a graph theoretical dimension reduction technique and a density-based classification technique, as previously applied to identify ecoprovinces based on biogeochemical ocean data at a global scale (Sonnewald et al., 2020). Among alternative methods available, we applied the *UMAP* algorithm (McInnes et al., 2020) and the *HDBSCAN* algorithm (Campello et al., 2013; McInnes et al., 2017) for dimension reduction and clustering, respectively.

2.3. Dimension reduction - UMAP

The *UMAP* algorithm is a nonlinear reduction technique (McInnes et al., 2020). Unlike more traditional methods applied in ecology, such as *Principal Component Analysis*, *UMAP* preserves both the local structure (the distance between neighbouring points) and the global structure (the distances between the most different points) of the raw dataset (McInnes et al., 2020). These two key properties are useful in reducing the dimension of complex genomic (Dorrity et al., 2020) or ecological (Milošević et al., 2022) data before clustering.

2.3.1. Principles of UMAP and details of its main hyperparameters

UMAP reduces the dimensionality of a dataset by first creating a high-dimensional graph that connects each data point to its k-nearest neighbours. Then, *UMAP* produces a low-dimensional representation of this high-dimensional graph that reflects the original dataset (McInnes et al., 2020). *UMAP* requires a distance matrix to construct the initial k-nearest-neighbour graph. Here, we applied the Chord transformation to standardise percentage cover data as relative cover per transect before computing Euclidean distances between transects (Legendre and Gallagher, 2001).

In addition to the choice of a suitable distance metric, two *UMAP* hyperparameters can influence the dimension reduction. The first is the number of neighbours (*n_neighbors*) that *UMAP* must consider when creating its k-nearest neighbour graph. Low values of *n_neighbors* allow the embedding to preserve the local structure of the original distance matrix while larger ones better preserve the global structure (McInnes et al., 2020). The second parameter, *min_dist*, controls the packing density at which *UMAP* is allowed to clump similar points in the reduced-dimensional space. A high value of *min_dist* will tend to preserve the overall topological structure of the data, whereas a low value allows *UMAP* to clump closely similar points on the embedding. The value of *n_neighbors* has been tuned in this study (see following section and Appendix B for details), while the value of *min_dist* has been set to 0.0, since this value allows a denser representation of the low-dimensional dataset, an important transformation before using a density-based classification algorithm (Vermeulen et al., 2021).

2.4. Clustering - HDBSCAN

After embedding our data into a two-dimensional space, we clustered the generated projections of the data with the machine learning algorithm named "*Hierarchical Density-Based Spatial Clustering of Applications with Noise*" (*HDBSCAN*).

HDBSCAN shares both similarities and differences with traditional clustering algorithms used in ecology. Like *Ward's hierarchical clustering*, *HDBSCAN* does not require the user to define the number of clusters in advance and provides a *hierarchical* clustering solution. However, unlike

K-means and *Ward's* clustering algorithms, *HDBSCAN* allows for clusters with varying shapes and densities. Additionally, *HDBSCAN* offers both hard clustering (where each sample is assigned to a single cluster) and soft clustering (where samples are assigned probabilities of belonging to different clusters). Beyond identifying clusters of different shapes and densities from a dendrogram, *HDBSCAN* has several advantages for ecological applications. Firstly, it can exclude noisy observations, leaving them unclustered, a particularly useful property when dealing with citizen science dataset where some outliers are possible. Secondly, *HDBSCAN* can highlight most representative members of each cluster, enhancing both classification and interpretation (Campello et al., 2013; McInnes et al., 2017).

2.4.1. Principles of HDBSCAN and details of its main hyperparameters

The *HDBSCAN* clustering algorithm firstly computes the core distance for the k-nearest neighbours for all points in the dataset. Then, it computes the extended minimum spanning tree from a weighted graph, where the edges are weighted by the distance between two points while taking into account the density of points around them. Then, *HDBSCAN* builds a hierarchy from the extended minimum spanning tree by cutting it at different levels of density. If the cut results in the creation of clusters smaller than the minimal number of observations set by the user *min_cluster_size*, all points members of these clusters are declared as noise by the algorithm. The algorithm stops when it declares all points as noise and produces a tree-like structure where each node corresponds to a cluster varying in shape and density (Campello et al., 2013; McInnes et al., 2017). In this study, we tuned only one parameter – the *minimum_cluster_size*, controlling for the minimal number of observations required to form a cluster, and otherwise used the default parameters.

2.5. Hyperparameter tuning and evaluation of the clustering output

For this pipeline, we search for the best combination of hyperparameters for both *UMAP* (*n_neighbors*) and *HDBSCAN* (*minimum_cluster_size*) using a complete grid search. We exhaustively explored the results sensitivity to the two hyperparameters from 10 to 500, resulting in 241081 models evaluated. The best combination was found by optimising both the quality of the embedding and the clustering, using two criteria. The *UMAP* embedding was evaluated with the trustworthiness metric (Venna and Kaski, 2001), ranging from 0 to 1 (the higher the index, the more the local structure of the original data is preserved). The quality of the clustering was evaluated using the *Density-Based Clustering Validation* (*DBCV*) as it is one of the few evaluation metrics capable of handling both noise in the data and the non-convex shape of clusters (see Moulavi et al. (2014) for a detailed discussion on these challenges and comparisons with other clustering evaluation indices). *DBCV* measures both compactness within and separations between clusters (Moulavi et al., 2014) and ranges between −1 and 1, where higher scores indicate better clustering quality (Moulavi et al., 2014). The sensitivity of the pipeline to the choice of hyperparameters is described in Appendix B.

A previous analysis by Cresswell et al. (2017) on a national subset of this dataset yielded nine groups of habitats. We expected to find at least that many groups at the global scale, and thus constrained our search of the best hyperparameter combinations to the solution yielding at least the same number of clusters as Cresswell et al. (2017). Among these solutions, we selected the best combination of hyperparameters (*n_neighbors* = 400; *min_cluster_size* = 74) in terms of both their trustworthiness and *DBCV* scores, while maximizing the number of clusters for a finer granularity. We finally compared outputs from the best *UMAP-HDBSCAN* solution to Agglomerative clustering used in Cresswell et al. (2017) at 212 common transects and also compared the pipeline results with *K-means* as well as *Ward's hierarchical clustering* (Appendix E).

2.6. Interpretation of the clusters

To interpret individual clusters identified with *UMAP-HDBSCAN*, we calculated the mean percentage cover of each habitat in each cluster. Then we used the *SHAP* framework to further explore how potential nonlinear interactions between variables can determine clustering outcomes (Lundberg and Lee, 2017). Because of the computational cost of applying *SHAP* to our complete pipeline, we used a classification tree (Breiman et al., 1984) to approximate the clustering pipeline (i.e., to predict label cluster membership based on the raw percentage cover variables) before applying the *SHAP* framework. Classification trees were trained with the default scikit-learn (version 1.2.0) with the following hyperparameter (i.e. *criterion = 'gini', splitter = 'best', max_depth = None, min_samples_split = 2, min_samples_leaf = 1, min_weight_fraction_leaf = 0.0, max_features = None, random_state = None, max_leaf_nodes = None, min_impurity_decrease = 0.0, class_weight = None, ccp_alpha = 0.0, monotonic_cst = None*). We trained the classification trees and estimated their ability to mimic our clustering pipeline using a stratified train-test split to ensure that the relative frequency of each cluster label is preserved in the train and test fold. The training and test sets contain 80 % and 20 % of the data, respectively. Then, we used a minimal cost-complexity pruning algorithm to avoid overfitting of our classification tree (Breiman et al., 1984). We estimated classification error rates using the F1-score (Van Rijsbergen, 1979). The classification error rates were satisfactory, with F1-scores of 0.99 and 0.94 on the training and test sets, respectively. Based on the *SHAP* values that estimate the influence of each variable in the definition of the cluster, we examined potential interactions between the two most characteristic variables for each cluster by performing a piecewise-linear interpolation of the *SHAP* values. Finally, we completed the interpretation by extracting the photoquadrats for these transects considered by HDBSCAN as the most representative members of their cluster.

2.6.1. Spatio-temporal distribution of benthic habitat states

We first explored the latitudinal distribution of each cluster. We also summarised their occurrence within each of the *Marine Ecoregions of the*

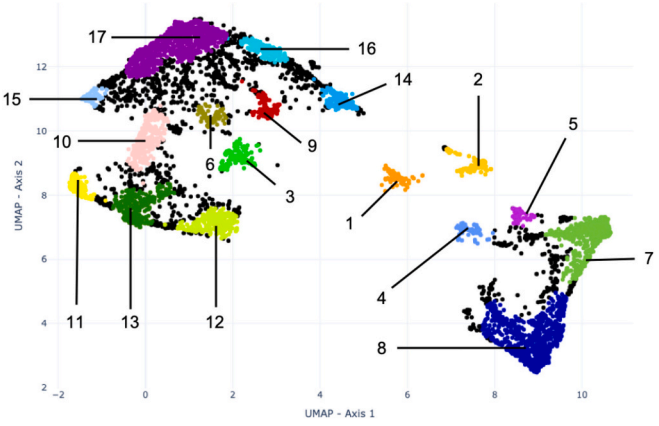


Fig. 1. Two-dimensional *UMAP* embedding of the benthic cover data of the 6554 *RLS* transects. Each point corresponds to an *RLS* transect, colored according to membership for the selected *UMAP-HDBSCAN* pipeline. Black dots represent points classified as noise (n = 1464). The 17 clusters can be interpreted as follows (see Fig. 2 and 1–17 in Appendix B): 1. Foliose brown algae (n = 148) 2. Filamentous algae (n = 208) 3. Other Sessile invertebrates (n = 185) 4. Foliose red algae (n = 123) 5. Seagrass (n = 83) 6. Soft coral and gorgonians (n = 98) 7. Bushy fucoids (n = 577) 8. Large Canopy forming algae (n = 894) 9. Unconsolidated substrate (n = 151) 10. Crustose coralline and turf algae (n = 286) 11. Green calcified algae (n = 166) 12. Bare substrates (n = 329) 13. Crustose coralline algae (n = 409) 14. Sand (n = 220) 15. Branching coral (n = 110) 16. Turf and sand (n = 207) 17. Turf algae (n = 897). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

World (*MEOW*; Spalding et al. (2007)) sampled by the *RLS*. In addition to examining dominant clusters per ecoregion, we also computed the proportion of transects classified as noise, as well as the Gini-Simpson diversity index. We chose this diversity index because it focuses on changes in dominance patterns, is more indicative of changes in landscapes, and is more robust to low sampling issues than other diversity indices (Lande et al., 2000).

Finally, we investigated temporal trends at the site level by comparing the proportion of transects classified into the different habitat states at five different temperate Australian locations previously reported by Stuart-Smith et al. (2022). We were interested in determining the extent to which changes in the proportion of transects classified in the different habitat states at a location could be an indicator of ecological changes.

3. Results

Based on exploration of hyperparameter space, both the trustworthiness score of 0.98 ± 0.002 (mean $\pm$ sd) and *DBCV* score of 0.46 ± 0.08 (mean $\pm$ sd) for solutions containing at least 9 groups suggest that these solutions produce reliable clustering of the *RLS* photoquadrat dataset (Fig. 1 in Appendix B). Among the top 100 solutions, the optimal number of clusters varied between 9 and 184 (number of clusters 22.81 ± 18.37; mean $\pm$ sd), while the mean number of points classified as noise was $2,207.33 \pm 364.95$ (mean $\pm$ sd). Hereafter, we focus on the single solution yielding the highest resolution (i.e. the greatest number of clusters; see Fig. 1 for a description of the habitat states uncovered), and the smallest number of transects classified as noise (1464 transects). This solution has a trustworthiness score of 0.98 for *UMAP* and a *DBCV* score of 0.60 for *HDBSCAN*. The number of clusters identified by this set of hyperparameters is 17 (Fig. 1).

The 17 clusters identified can be summarised as four broad habitat groups (Fig. 2; Fig. 3; see Fig. 2-19 in Appendix B for their distribution on the globe and 1-17 in Appendix C for their interpretation with *SHAP* framework): (1) temperate habitats, (2) subtropical and tropical habitats, (3) broadly distributed habitats and (4) opportunistic habitats (i.e., habitats characterised by the presence of filamentous algal species or turf, generally with strong anthropogenic influences).

Transects within temperate regions can be classified into five major clusters with contrasting dominance of sessile invertebrates, foliose red algae, seagrass, bushy foliose algae and canopy-forming algae, as follows: cluster 3 is dominated by at least 30 % and on average 42 % of sessile invertebrates. Cluster 4 is dominated by at least 40 % cover of foliose red algae. Cluster 5 is dominated by at least 30 % and on average 40 % seagrass. Cluster 7 is dominated by at least 20 % cover, an average of 56 % fucoid bushy algae, and the absence of canopy-forming algae. Cluster 8 is characterised by a cover of at least 20 % and an average of 55 % of canopy-forming algae with fucoid bushy algae absent.

Three clusters correspond to tropical and subtropical habitat types. Cluster 6 which is characterised by at least 30 % and on average 37 % of soft corals and gorgonians. Cluster 11 is composed of 20 % cover and an average of 35 % green calcified algae. Finally, cluster 15 is composed of at least 35 % and on average 55 % branching coral. This is the only group of corals identified in the dataset, which is unusual given four categories of corals that form colonies were originally defined.

Five clusters correspond to broadly distributed habitats that can occur across both temperate and tropical latitudes. Cluster 1 is dominated by at least 30 % and on average 46 % brown foliose algae. Cluster 9 is dominated by the presence of at least 30 % and on average 41 % unconsolidated substrate. Cluster 12 has at least 30 % and on average 42 % bare substrate. Cluster 13 is characterised by 40 % and on average 51 % of crustose coralline algae with an absence of turf algae. Cluster 14 has at least 30 % and an average of 53 % sand without turf algae.

Finally, four groups correspond to opportunistic habitats. Cluster 2 is dominated by at least 30 % cover and an average of 39 % filamentous algae. Clusters 10, 11 and 17 are all dominated by turf algae. Cluster 10 is composed of at least 30 % and on average 39 % turf algae, at least 20 % and on average 28 % crustose coralline algae. Cluster 16 is characterised by the presence of at least 30 % and on average 48 % turf algae, and a minimum cover of 20 % and on average 26 % sand. Cluster 17 is composed of at least 40 % and on average 60 % turf algae with crustose coralline algae absent.

These 17 clusters are consistent with (i) previous identifications of nine major clusters along the Australian coastline (Cresswell et al. (2017) based on a subset of the dataset used here; see Fig. 1 and 2 in Appendix E); as well as (ii) clustering results obtained using commonly-used clustering algorithms (i.e. *Ward's* clustering and *K-means*; Fig. 3 and 4 in Appendix E). Overall, *UMAP-HDBSCAN* yields more homogeneous clusters than alternative algorithms (constrained to the same number of clusters) as assigning a fraction (22 % of observations) to the noise category, rather than forcing these into the closest group. The 17 clusters identified offer a finer insights into reef habitat states relative to Cresswell et al. (2017)'s nine broad categories (for instance, the "*barren*" habitat identified by Cresswell et al. (2017) gets subdivided in our classification according to dominance of crustose coralline algae, other sessile invertebrates or bare substrate).

The clusters identified by the *UMAP-HDBSCAN* pipeline show a marked latitudinal gradient (Fig. 3). *Red algae*, *filamentous algae*, *fucoids*, *large canopy-forming algae* and *seagrass* are essentially distributed overall in the temperate zones across latitudes higher than 25° (Fig. 3). In contrast, four habitat states, namely *soft corals and gorgonians*, *green calcified algae*, *sand and turf*, and *branching coral*, essentially occur in tropical latitudes (lower than 25°; Fig. 3). However, some groups are relatively ubiquitous across all surveyed latitudes, such as those associated with transects classified as *bare substrate* and *unconsolidated substrate*, *brown algae*, *crustose coralline algae* with and without *turf algae*, and *turf algae* (Fig. 3). It should also be noted that the transects considered noisy are also evenly distributed across all latitudes (Fig. 3).

The spatial distribution of transects sampled by *RLS* volunteers is particularly concentrated in Australia (Fig. 4). However, other areas such as the Caribbean, the Canary Islands and French Polynesia have also been extensively surveyed with more than 50 transects (Fig. 4. a). Globally, three habitat types dominate in terms of occurrences in all ecoregions surveyed, namely *bare substrate* ($n = 20$), *turf algae* ($n = 17$), and *large canopy-forming algae* ($n = 11$). These three habitat types dominate in 37 % of the ecoregions sampled by the *RLS* (Fig. 4 b). Two habitat types identified by the *UMAP-HDBSCAN* pipeline, *seagrass* and *red algae*, are not dominant on the surveys of reef habitats in any of the world's ecoregions. The patterns of dominance of the different clusters also vary along the latitudinal gradient (Fig. 4 b), in line with the latitudinal distribution of each cluster (Fig. 3). These latitudinal variations of dominance are visible both at a global scale, but also along the coasts of certain regions. For instance, a decrease in prevalence of sites in the *large canopy-forming algae* cluster accompanies an increase in sites in the *turf algae* cluster along the coastline from southern to northern Australia (Fig. 4).

The proportion of noisy transects is highly heterogeneous across the globe (Fig. 4 c). *Noisy* transects represent 23 % of all transects analysed but are present in some areas more than in others. For example, in the Southern California Bight (Western USA), Bight of Sofala/Swamp Coast (Eastern Africa), the Seychelles, and in Three Kings-North Cape (Northern New Zealand), at least 60 % of transects are classified as noisy (Fig. 4 c). Although these four ecoregions share a low number of transects sampled (Fig. 4 a), no significant correlation was found between the proportion of transects classified as noisy and the number of transects done in each ecoregion ($\tau_{Kendall} = 0.05, p = 0.54$; Fig. 1 in Appendix D). Furthermore, 12 of the 83 ecoregions sampled by the *RLS* did not have transects that were classified as noisy (Fig. 4 c). Overall, a large proportion of temperate transects identified as noise by the *UMAP-HDBSCAN* pipeline corresponds to transects classified as "*barren*" or "*kelp*" by Cresswell et al. (2017). In tropical areas, transects identified as *noisy* are spread more or less evenly in the five different clusters they

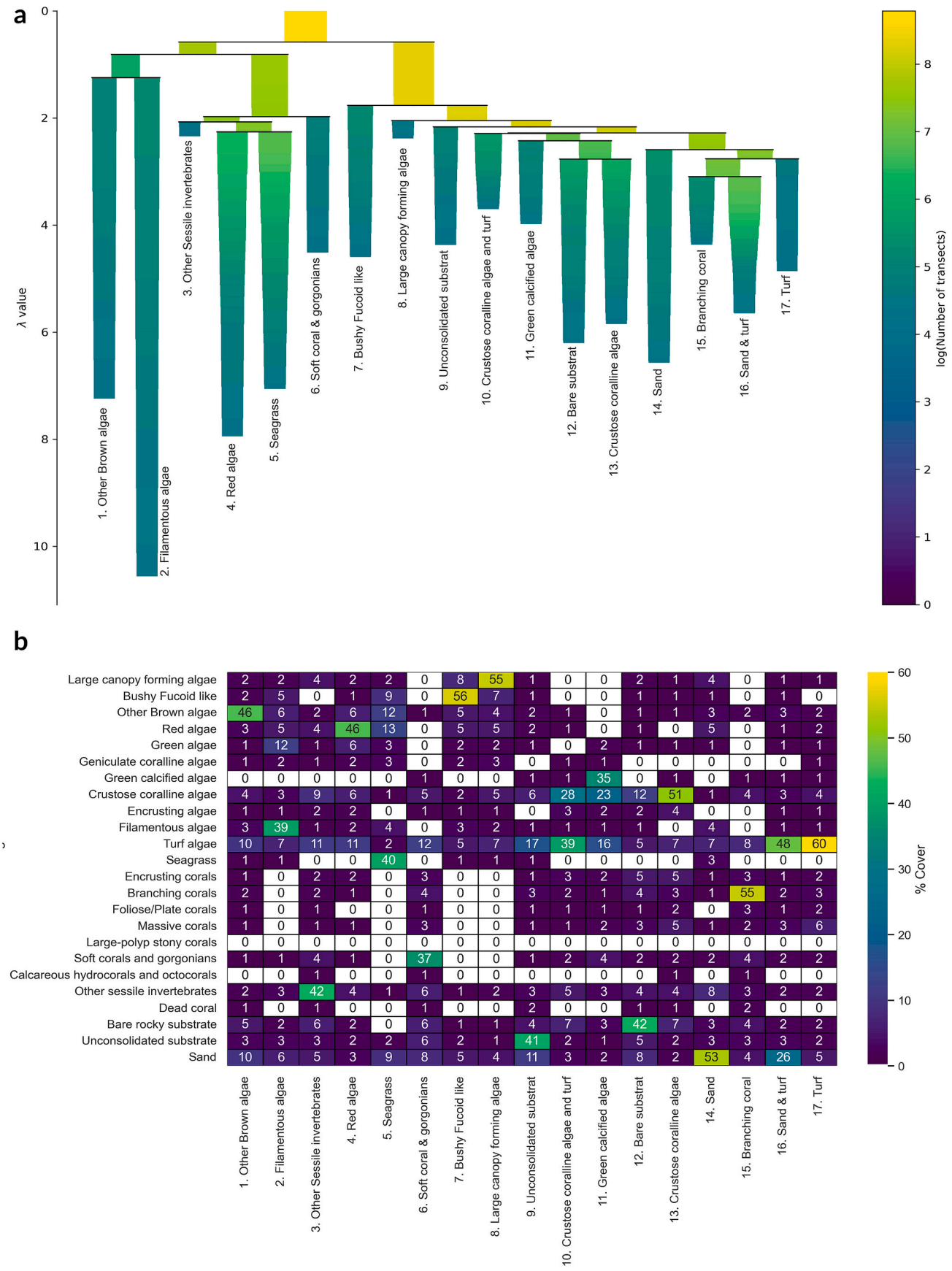


Fig. 2. a. *HDBSCAN* condensed clustering tree of the *UMAP* 2D embedding b. Heatmap of the mean substrate cover (rounded to the nearest integer) for each cluster identified by the *UMAP-HDBSCAN* pipeline.

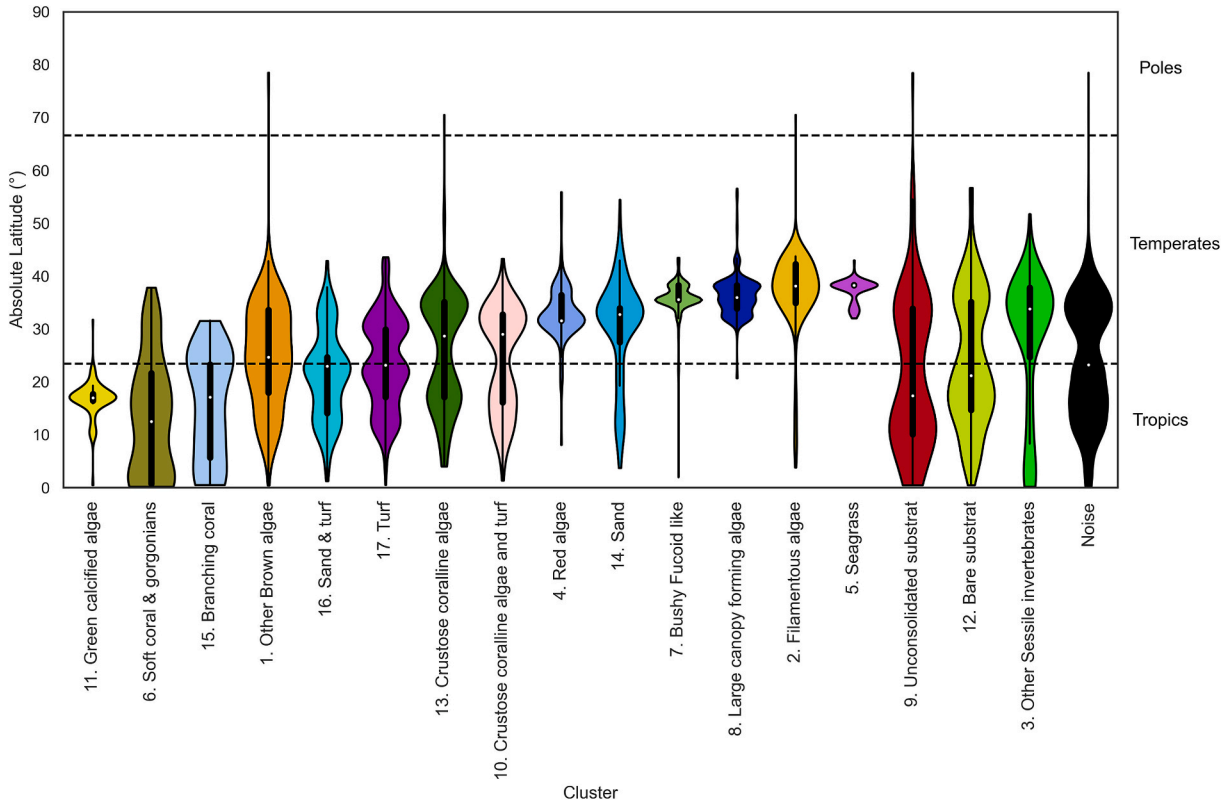


Fig. 3. Violin plot of the absolute latitudinal distribution of the different hard cluster solutions.

defined (Fig. 2 in Appendix E).

Areas with the highest diversity of habitat types, based on both the number of clusters occurring and their relative proportions in the ecoregions, are concentrated in Eastern and Western Australia, as well as in the Caribbean and Tuamotus (Fig. 4). The areas with the lowest Gini-Simpson values are the Southern California Bight (Western USA) and the Bight of Sofala/Swamp (Eastern Africa) coast with a Gini index of 0 (Fig. 4 d). However, it should be noted that there is a weak correlation between the Gini-Simpson index and the number of transects carried out in the ecoregion ($\tau_{Kendall} = 0.29, p < 0.001$; Fig. 2 in Appendix D).

At the *Reef Life Survey* monitoring location level, temporal changes in the occurrence of the different clusters can provide useful indicators of ecological changes (Fig. 5). For instance, at a given monitoring location, changes in yearly proportions of transects classified as *large canopy forming algae* tend to match with annual mean percentage cover of *large canopy forming algae* estimated across transects (Fig. 5 and Fig. 3 in Appendix D). Moreover, at certain locations where the cover of *large canopy forming algae* decreased, changes in the dominance patterns of habitat states offer insights about ongoing ecological changes. At Beware Reef (Fig. 5 b), *large canopy forming algae* disappeared in favour of *other sessile invertebrates*, whereas at Port Phillip Heads (Fig. 5 d), a decrease in the proportion of transects classified as *large canopy forming algae* between 2013 and 2017 was counterbalanced by an increase in the proportion of transects classified as *bushy fucoids* or *filamentous algae*. This area also experienced moderate interannual variability in the proportion of transects classified as noise. However, a long-term decrease in the proportion of transects classified as noise was observed at other sites (e.g., Batemans or Beware Reef; Fig. 5 a and b), where a turnover through time in the dominating habitat state occurred.

4. Discussion

The *UMAP-HDBSCAN* clustering pipeline identified 17 distinct clusters within all *RLS* transects surveyed globally in a range of coastal temperate and tropical regions. Within these groups, we found different biogenic habitats whose distribution patterns match with current biogeographic knowledge of benthic ecosystems: for example, *bushy fucoid algae*, and *large canopy-forming algae* predominantly occur in temperate waters (Assis et al., 2020; Jayathilake and Costello, 2020), while *soft corals and gorgonians*, and *branching coral*, are more frequent in tropical waters (Jones et al., 2019; Wirabuana et al., 2019). Our analysis also highlights habitat types that occur throughout the world, including (1) different granulometric facies such as *sand*, *unconsolidated substrate*, and *bare substrate*, as well as (2) different habitat types dominated by low-profile algae, such as *crustose coralline algae* or *turf algae*. The latter are known to occur globally and can dominate benthic substrates under a wide range of conditions (Connell et al., 2014; Liu et al., 2018). Thus, this classification distinguished between different ecological states of these habitats (hereafter referred to as "habitat state"), including known alternative succession stages, or different degradation states of these habitats (Fig. 6).

Our results align with those of Cresswell et al. (2017), who focused on a subset of data centered on the Australian region. However, our findings offer additional insights into habitat states on both regional and global scales. Specifically, our analysis, which identified 17 clusters, includes all nine clusters defined by Cresswell et al. (2017), with a similar distribution of transects across categories (e.g., *large canopy-forming algae*, *turf algae*, *filamentous algae*, and *branching coral*). These match closely with Cresswell et al. (2017)'s classifications of "*Canopy algae*", "*Turf*", "*Epiphytic filamentous algae–caulerpa*", and "*Coral*" (see Figs. 1 & 2 in Appendix E).

Furthermore, our results reveal a more refined resolution of habitat states compared to Cresswell et al. (2017). For example, our separate clusters for *crustose coralline algae* and *bare substrate* are combined into a single "*Barren*" cluster by Cresswell et al. (2017). Similarly, our classification distinguishes between *red algae* and *other brown algae clusters*, refining the broader "*Foliose algae*" group from Cresswell et al. (2017). Thus, while our large-scale spatial approach generally confirms the

habitat types defined by Cresswell et al. (2017), it also provides a more nuanced distinction among habitats, identifying transitional states such as *crustose coralline algae and turf*.

Our comparative analysis with Cresswell et al. (2017) also sheds light on the nature of transects categorised as noise in our *UMAP-HDBSCAN* pipeline. In temperate regions, most noisy transects correspond to "*Barren*" or "*Kelp*" in Cresswell et al. (2017), suggesting they may represent heterogeneous or transitional temperate habitats where kelp forests are either declining or patchy. In tropical regions, noisy transects span various groups defined by Cresswell et al. (2017), highlighting the advantage of our *UMAP-HDBSCAN* pipeline, which provides a more finely resolved classification of habitat states, effectively handling outliers or noisy observations.

Unexpectedly, our analysis captured only a single coral reef habitat state, while Cresswell et al. (2017) described several coral related habitat states. Potential reasons for this include the significant variability in coral cover across reefs, as noted by De'ath et al. (2012), which complicates the characterisation of coral reef states based on cover data alone. Furthermore, the morphological diversity of these reefs is extensive, including variations in surface areas (Zawada et al., 2019). Such high variability in their percentage cover, might explain why *branching coral* is the single group of corals identified, since some species of *Acropora* spp. are able to establish colonies with expansive surface areas. Conversely, other coral reef habitat states probably failed to be detected due to their rarity in the dataset and are classified by our pipeline as noise, while categorised by Cresswell et al. (2017) in one of the four categories characterised by the presence of corals (see Fig. 2 in Appendix E). Subsequent analysis could help interpret noisy transects, and further refine habitat state categories.

This pipeline also offers a valuable framework for describing habitat transitions and assessing their ecological impacts. Our results reveal a habitat state gradient transitioning from dominance by *crustose coralline algae*, through *crustose coralline algae and turf*, to a *turf* algae-dominated state. This progression has major ecological implications, including by modifying the carbonate production within reefs (Cornwall et al., 2023). Likewise, clusters such as *branching coral*, *turf and sand*, and *turf* provide a standardised way to characterise the increasing transitions between *coral* and *turf* dominated habitats, driven by anthropogenic pressures (Jouffray et al., 2015). Fig. 5 b illustrates examples along southeastern Australia's coastline, where dense macroalgal canopies, typically

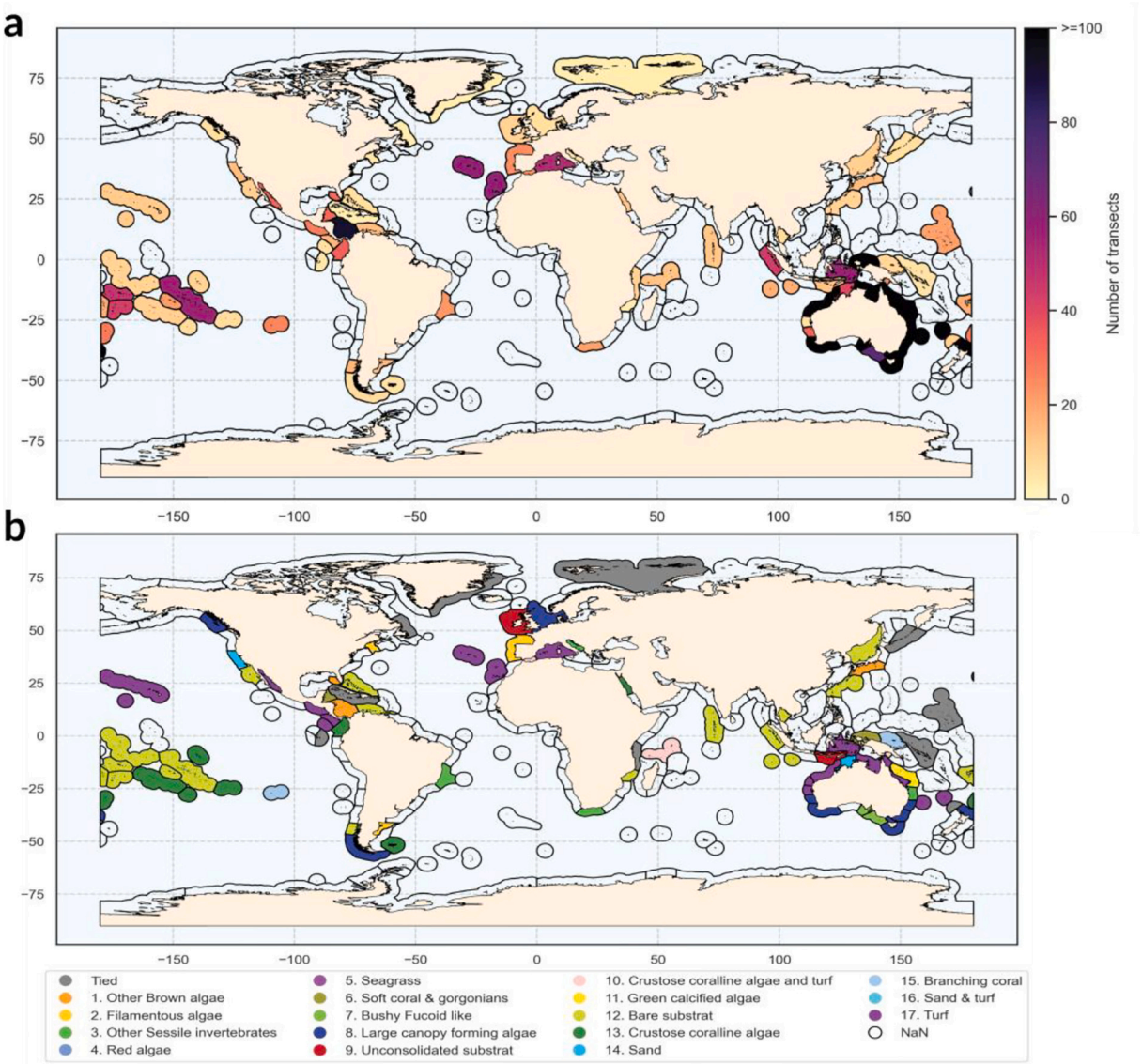


Fig. 4. a. Spatial distribution of reef surveys from the *Reef Life Survey* database used for analyses. b. Map of dominant clusters in each *MEOW* ecoregion. Dominant clusters were determined as the greatest count of transect labels in each ecoregion. c. Spatial distribution of the proportion of transects classified as noise in each ecoregion. d. Gini-Simpson diversity index calculated by the occurrence of clusters in each ecoregion of the world.

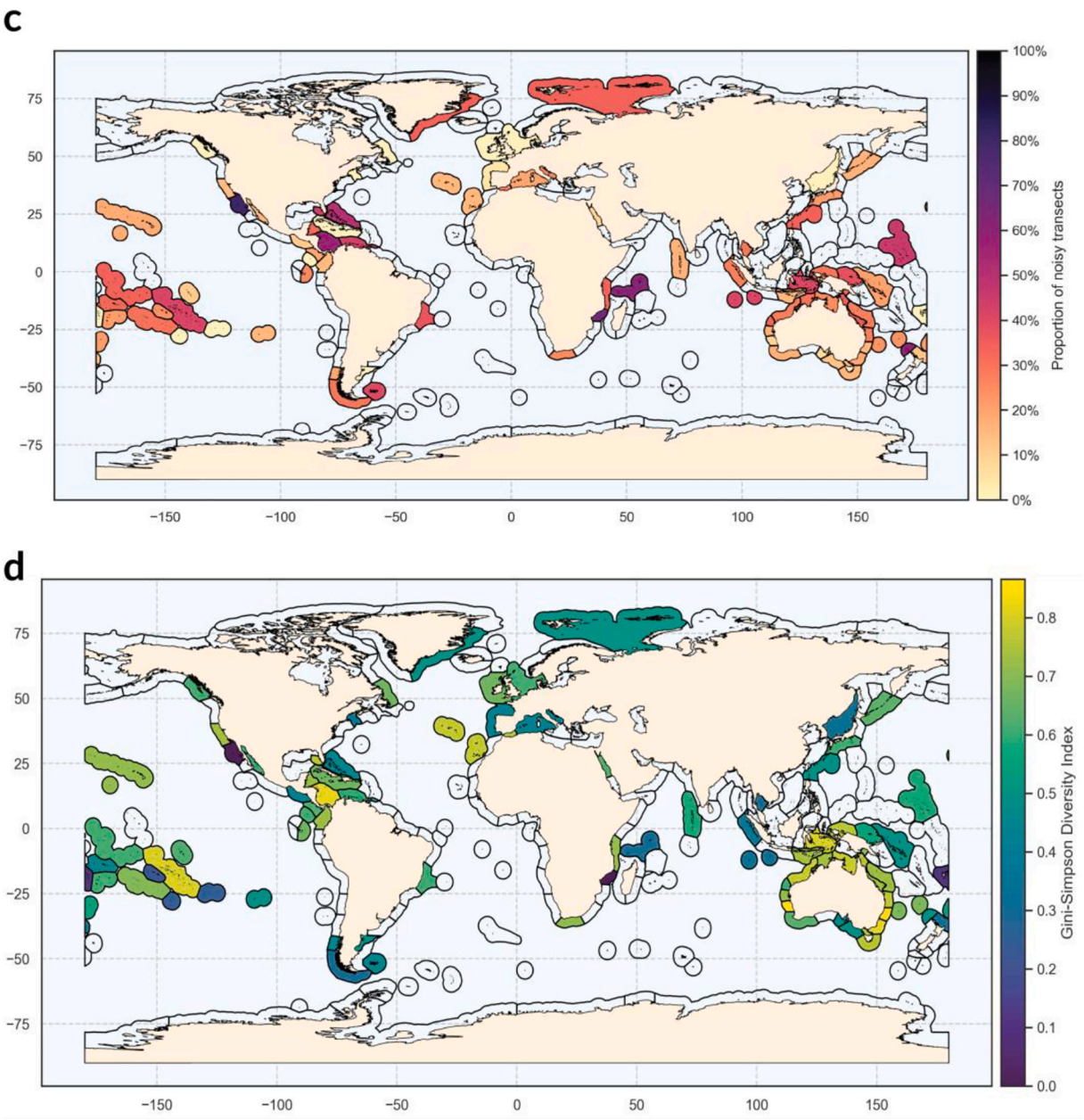


Fig. 4. (*continued*).

dominated by the kelp *Ecklonia radiata* (*large canopy-forming algae*), can shift to barren substrates (*bare substrate* or *crustose coralline algae*) following intense grazing by the long-spined sea urchin *Centrostephanus rodgersii* (Ling, 2008). Consequently, our approach enables classification of reef cover data collected globally with the *RLS* protocol into an ecologically robust framework for exploring common reef habitat transitions under anthropogenic influence (Donovan et al., 2018). The strength of this data-driven classification lies in its ability to encapsulate common benthic habitats (e.g., seagrass meadows, coral reefs, kelp forests) shared across major seafloor habitat classification systems (e.g., European Nature Information System; Bajjouk et al. (2015)).

At a global scale, our global classification of *RLS* data highlights hotspots of diversity in terms of benthic habitats and habitat states. Four ecoregions in particular, the Eastern (Manning-Hawkesbury ecoregion) and Western Australia (Houtman ecoregion), the Caribbean, and the Tuamotus Archipelago, showed a high diversity of habitat types (considering both richness and evenness). In the transition zones between temperate and tropical waters, such as the Manning-Hawkesbury or the Houtman ecoregions in Eastern and Western Australia, respectively, the high diversity of benthic habitat types we observe potentially results from a high diversity of foundation species. Indeed, high biodiversity is typical of transitional environmental conditions where species with temperate and tropical environmental niches overlap (Ferro and Morrone, 2014). This phenomenon is well known for multiple taxa such as birds (Altamirano et al., 2020), plants (Lemessa et al., 2023) or reef fishes (Pinheiro et al., 2018), and also apparently applies to biogenic habitats such as coralline red algae (Sissini et al., 2022). Such subtropical or warm temperate zones are also recognized as regions where both mobile fauna (Vergés et al., 2014) or sessile habitat-forming species assemblages (Marzloff et al., 2018) are likely to undergo tropicalisation, with poleward climate-driven range shifts of warmer water species at the expense of temperate species. Our finely resolved classification could be modelled against environmental predictors in future work to understand and predict the state of benthic habitats under

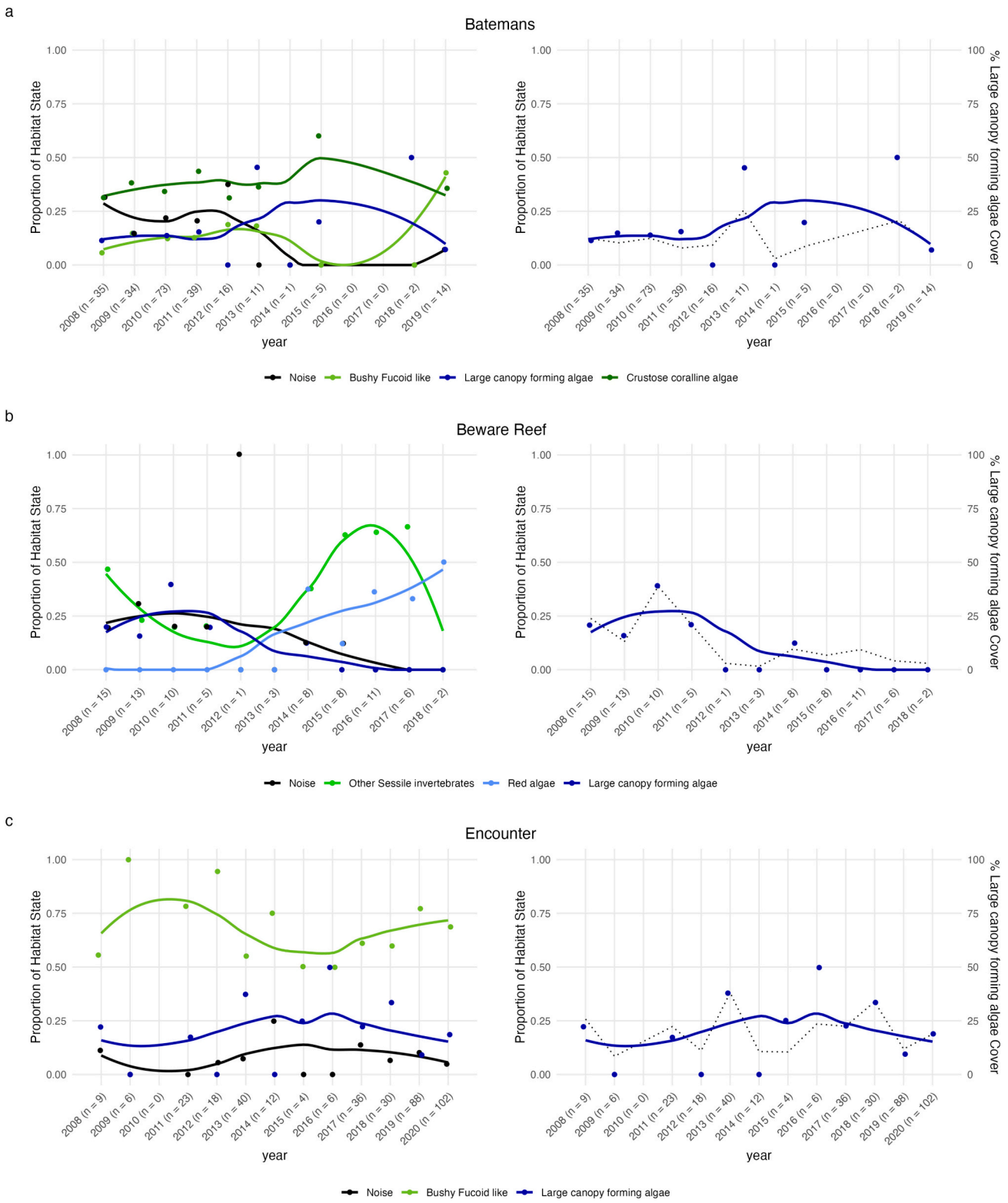


Fig. 5. Temporal evolution of the proportion of transects classified into different habitat states at different sites: a. Batemans, b. Beware Reef, c. Encounter, d. Port Phillip Heads, e. Port Stephens. Habitat states are colour-coded as follows: light blue for large canopy-forming algae; pink for crustose coralline and turf; deep green for crustose coralline algae; deep blue for sand; yellow for filamentous algae; light green for bushy fucoid; neon green for other sessile invertebrates; light blue for red algae; purple for turf. Noise appears in black). The dots represent the proportion of transects in each category in each year, and the trend lines are *LOESS* regression models weighted by the number of transects per year. For visual clarity, the left column only shows the subset of groups, which proportion most varied over time (see Fig. 4 in Appendix D). The right-hand column focuses solely on the yearly proportion of transects classified as Canopy forming algae (blue line) compared with the annual mean percentage cover of Canopy forming algae estimated across all transects (black dotted line). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

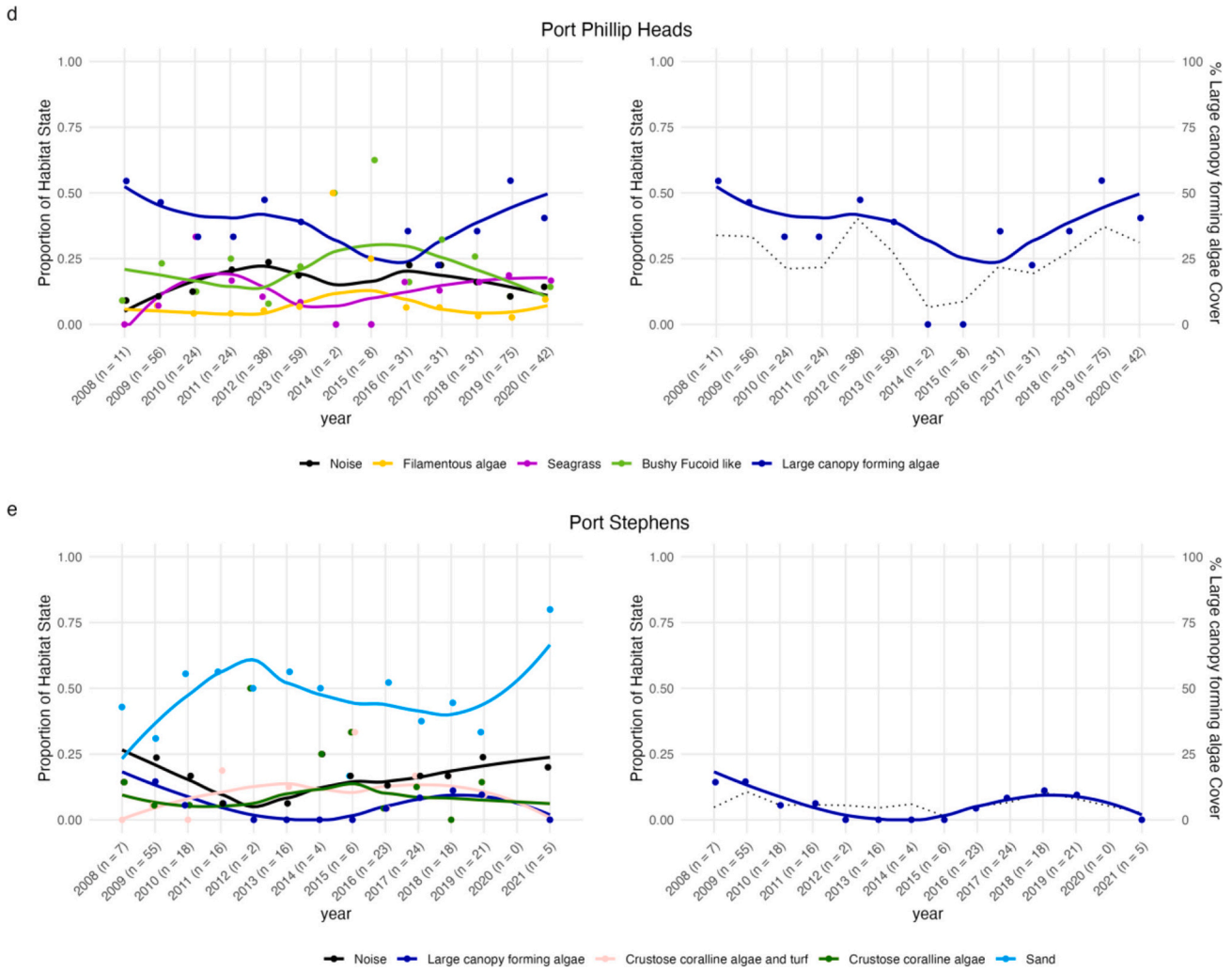


Fig. 5. (*continued*).

current and future conditions (see, e.g., Belanger et al. (2012)).

Beyond exploring spatial patterns of benthic biodiversity, our classification of the *RLS* dataset offers a new perspective to explore temporal changes in benthic habitat states at the site level. In fact, the proportion of transects classified in the different groups seems to be an interesting indicator for researchers and managers, providing a metric for expressing the dominant ecological state, less affected by the inherent variability at the scale of individual transects. This metric may be particularly useful for identifying regional-scale changes in the structure of ecosystems. One of the temperate reef monitoring locations investigated by (Stuart-Smith et al., 2022), Beware Reef, underwent a major ecological change following an overgrazing event in 2013 by *Centrostephanus rodgersii* sea urchins (Barrett et al., 2014). Our metric allows us to observe that the regime change towards a urchin barren state is not complete. In fact, this species of sea urchin is omnivorous, capable of attacking sessile invertebrates (Byrne and Andrew, 2013), but showing a clear preference for species belonging to the large canopy forming algae group (Hill et al., 2003). Our metric indicates that many transects at this site are classified as red algae, or other sessile invertebrates, showing that the proliferation of these urchins appears to be constrained, in agreement with in situ observations (*State of the marine and coastal environment report,* 2021).

Overall, changes in benthic habitat whether on an ecoregional or site level may reflect a variety of processes. These processes include ecological factors, such as temporal variability in cover of habitat-forming species or response to climate-driven environmental changes (i.e., marine heatwaves Wernberg et al. (2016)), tropicalisation of tropical-temperate transition zones (Horta e Costa et al., 2014). Additionally, changes in benthic habitat can also reflect gradients in human stressors (i.e. nutrients and organic pollution runoffs, impacts from coastal human populations; Halpern et al. (2019)). Moreover, some methodological factors can also explain some of the variability observed in this study due to irregular transect location or sampling effort through time (e.g., Stuble et al. (2021)). Identifying the processes driving the observed habitat transitions could better characterise the impact of anthropogenic activities on benthic habitats (see, for example, Donovan et al. (2018) for a similar approach at a finer spatial scale). Our classification could thus provide an interesting template to further explore changes in benthic habitats across the world with expanded monitoring efforts (Edgar et al., 2023).

Nevertheless, not all expected transitions between habitats or alternative ecological states resolve among different clusters. Some transitory states may be classified as noise if they are insufficiently observed to constitute a cluster of their own. Understanding the drivers behind the transects classified as noise can reveal valuable information about the factors influencing habitat variability, and the ecological processes driving shifts between different states. This includes deciphering the reasons for a noise classification, such as variations in environmental conditions, biotic interactions, or anthropogenic disturbances. By investigating these aspects, researchers should gain crucial insight into

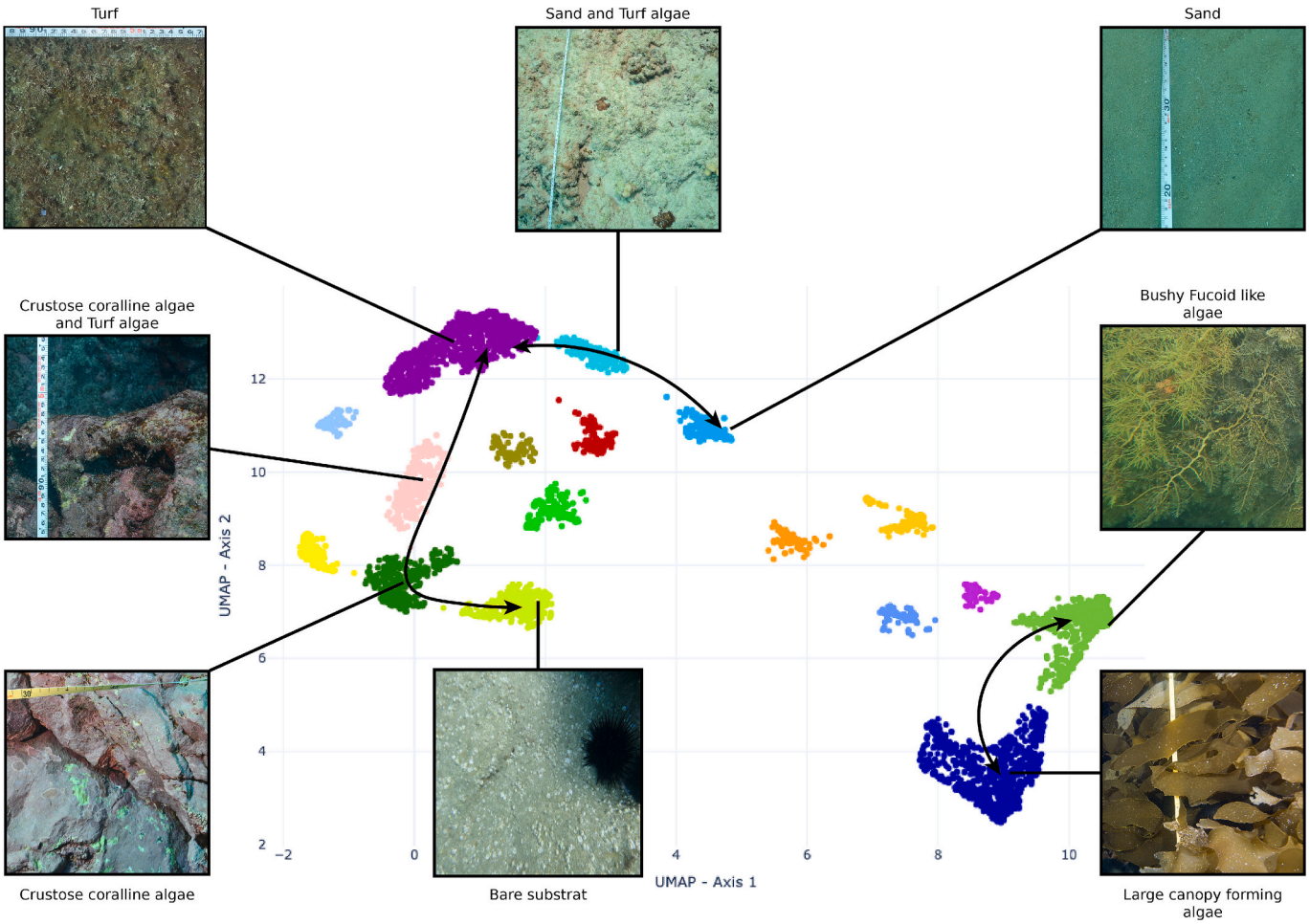


Fig. 6. Two-dimensional *UMAP* embedding of the 5090 clustered *RLS* transects, with clusters identified using *HDBSCAN* and colored accordingly. A subset of clusters was selected for which pictures are shown, as they highlight potential transitions previously identified in Jouffray et al. (2015) and Cornwall et al. (2023). These pictures represent transects identified as representative of their respective clusters by *HDBSCAN*. Arrows denote the direction of these potential transitions.

the dynamics and transitions that occur between habitat states and alternative ecological states.

5. Conclusion

In conclusion, this study highlights the value of *UMAP-HDBSCAN* for ecological clustering. Due to its hierarchical structure, this pipeline aligns with established classification standards and facilitates an initial data-driven description of global patterns in habitat states. Furthermore, the pipeline's ability to handle non-linear data and accommodate for noise underscores its adaptability to various ecological contexts and data sources. Thus, this clustering pipeline could help revisit the definition of groups in community ecology and for instance more finely distinguish nuances within functional groups from trait data.

Here, the *UMAP-HDBSCAN* clustering pipeline helped leverage a global dataset coming from citizen science to identify fine-scale habitat patterns within coastal temperate and tropical benthic ecosystems. Compared with a previous study using a national subset of the data used here, we highlight a more nuanced distinction between similar habitats previously considered homogeneous. We moreover identify habitat groups associated with different ecological states, which makes it possible to monitor ecosystem health across broad spatial and temporal scales. For instance, at the scale of monitoring locations, temporal changes in the proportion of transects classified according to the different categories can provide a relevant indicator of ecological dynamics. Thus, applying this habitat state classification on a fine spatial scale can effectively help assess ongoing trends and monitor outcomes of management interventions. At a regional scale, similar metrics related to relative proportion of habitat types can also help track consequences of global changes on coastal ecosystems and explore the influence of local and global drivers on benthic habitat states. This classification therefore provides a standardised template for tracking benthic habitat change across space and time at a global scale.

CRediT authorship contribution statement

Clément Violet: Writing – review & editing, Writing – original draft, Visualization, Methodology, Formal analysis, Conceptualization. **Aurélien Boyé:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Conceptualization. **Stanislas Dubois:** Writing – review & editing, Supervision, Conceptualization. **Graham J. Edgar:** Writing – review & editing, Investigation, Data curation. **Elizabeth S. Oh:** Writing – review & editing, Investigation, Data curation. **Rick D. Stuart-Smith:** Writing – review & editing, Investigation, Data curation. **Martin P. Marzloff:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Conceptualization.

Acknowledgments

We are grateful to all volunteers and staff involved in the Reef Life Survey monitoring programme. The authors would also like to acknowledge the Pôle de Calcul et de Données Marines (PCDM) for

providing DATARMOR storage and computational resources for this study. https://pcdm.ifremer.fr. C.V.'s research was supported by funding from Ifremer and an ISBlue mobility grant. This work was conducted as part of the TRIDENT ANR project (ANR early career grant ANR-21-CE02-0006 granted to MPM). RLS data management is supported by Australia's Integrated Marine Observing System (IMOS)—IMOS is enabled by the National Collaborative Research Infrastructure Strategy (NCRIS). We thank the two anonymous reviewers whose comments/suggestions helped improve and clarify this manuscript.

Appendix A. Supplemantary data

Data availability

The data and the code used in this study are available in ZENODO archive with the following DOI: https://doi.org/10.5281/zenodo.10974718.

References

Altamirano, T.A., de Zwaan, D.R., Ibarra, J., Wilson, S., Martin, K., 2020. Treeline ecotones shape the distribution of avian species richness and functional diversity in south temperate mountains. Sci. Rep. 10, 18428.
Althaus, N.A.F., Hill, Franziska, 2015. A standardised vocabulary for identifying benthic biota and substrata from underwater imagery: the CATAMI classification scheme. PLoS One 10, 1–18.
Assis, J., Fragkopoulou, E., Frade, D., Neiva, J., Oliveira, A., Abecasis, D., et al., 2020. A fine-tuned global distribution dataset of marine forests. Sci. Data 7, 119.
Bajjouk, T., Guillaumont, B., Michez, N., Thouin, B., Croguennec, C., Populus, J., et al., 2015. Classification EUNIS, système d'information européen sur la nature : Traduction française des habitats benthiques des régions atlantique et méditerranée. In: Habitats subtidaux & complexes d'habitats (Report). France, 2.
Barbier, E.B., 2017. Marine ecosystem services. Curr. Biol. 27, R507–R510.
Barbier, E.B., Hacker, S.D., Kennedy, C., Koch, E.W., Stier, A.C., Silliman, B.R., 2011. The value of estuarine and coastal ecosystem services. Ecol. Monogr. 81, 169–193.
Barrett, N., Bates, A., Beger, M., Stuart-Smith, R., Syms, C., Holbrook, N., et al., 2014. Adaptive Management of Temperate Reefs to Minimise Effects of Climate Change: Developing New Effective Approaches for Ecological Monitoring and Predictive Modelling. University of Tasmania, Institute for Marine. Antarctic Studies.
Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I.W.H., Ng, L.G., et al., 2019. Dimensionality reduction for visualizing single-cell data using UMAP. Nat. Biotechnol. 37, 38–44.
Belanger, C.L., Jablonski, D., Roy, K., Berke, S.K., Krug, A.Z., Valentine, J.W., 2012. Global environmental predictors of benthic marine biogeographic structure. Proc. Natl. Acad. Sci. 109, 14046–14051.
Bellwood, D.R., Hughes, T.P., Folke, C., Nyström, M., 2004. Confronting the coral reef crisis. Nature 429, 827–833.
Blanco-Portals, J., Peiró, F., Estradé, S., 2022. Strategies for EELS data analysis. Introducing UMAP and HDBSCAN for dimensionality reduction and clustering. Microsc. Microanal. 28, 109–122.
Bloch, T., Watt, C., Owens, M., McInnes, L., Macneil, A.R., 2020. Data-driven classification of coronal hole and streamer belt solar wind. Sol. Phys. 295, 41.
Bowler, D.E., Bjorkman, A.D., Dornelas, M., Myers-Smith, I.H., Navarro, L.M., Niamir, A., et al., 2020. Mapping human pressures on biodiversity across the planet uncovers anthropogenic threat complexes. People Nat. 2, 380–394.
Breiman, L., Friedman, J., Stone, C.J., A., O.R, 1984. Classification and Regression Trees. Hall/CRC, Chapman.
Brown, C.J., Smith, S.J., Lawton, P., Anderson, J.T., 2011. Benthic habitat mapping: A review of progress towards improved understanding of the spatial ecology of the seafloor using acoustic techniques. Estuar. Coast. Shelf Sci. 92, 502–520.
Burrows, M.T., Schoeman, D.S., Richardson, A.J., Molinos, J.G., Hoffmann, A., Buckley, L.B., et al., 2014. Geographical limits to species-range shifts are suggested by climate velocity. Nature 507, 492–495.
Byrne, M., Andrew, N., 2013. Chapter 17 - Centrostephanus rodgersii. In: Sea Urchins: Biology and Ecology, Developments in Aquaculture and Fisheries Science. Elsevier, Lawrence, JM, pp. 243–256.
Campello, R.J.G.B., Moulavi, D., Sander, J., 2013. Density-based clustering based on hierarchical density estimates. In: Pei, J., Tseng, V.S., Cao, L., Motoda, H., Xu, G. (Eds.), Advances in Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 160–172.
Cattano, C., Agostini, S., Harvey, B.P., Wada, S., Quattrocchi, F., Turco, G., et al., 2020. Changes in fish communities due to benthic habitat shifts under ocean acidification conditions. Sci. Total Environ. 725, 138501.
Connell, S., Foster, M., Airoldi, L., 2014. What are algal turfs? Towards a better description of turfs. Mar. Ecol. Prog. Ser. 495, 299–307.
Conversi, A., Dakos, V., Gårdmark, A., Ling, S., Folke, C., Mumby, P.J., et al., 2015. A holistic view of marine regime shifts. Philosoph. Trans. Royal Soc. B Biol. Sci. 370, 20130279.
Cooper, A., Oh, E., 2023. NRMN database QA/QC protocols. Version 1.4.. Reef Life Survey.
Cornwall, C.E., Carlot, J., Branson, O., Courtney, T.A., Harvey, B.P., Perry, C.T., et al., 2023. Crustose coralline algae can contribute more than corals to coral reef carbonate production. Commun. Earth Environ. 4, 105.
Cresswell, A.K., Edgar, G.J., Stuart-Smith, R.D., Thomson, R.J., Barrett, N.S., Johnson, C. R., 2017. Translating local benthic community structure to national biogenic reef habitat types. Glob. Ecol. Biogeogr. 26, 1112–1125.
De'ath, G., Fabricius, K.E., Sweatman, H., Puotinen, M., 2012. The 27–year decline of coral cover on the great barrier reef and its causes. Proc. Natl. Acad. Sci. 109, 17995–17999.
Donovan, M.K., Friedlander, A.M., Lecky, J., Jouffray, J.-B., Williams, G.J., Wedding, L. M., et al., 2018. Combining fish and benthic communities into multiple regimes reveals complex reef dynamics. Sci. Rep. 8, 16943.
Dorrity, M.W., Saunders, L.M., Queitsch, C., Fields, S., Trapnell, C., 2020. Dimensionality reduction by UMAP to visualize physical and genetic interactions. Nat. Commun. 11, 1537.
Dunic, J.C., Brown, C.J., Connolly, R.M., Turschwell, M.P., Côté, I.M., 2021. Long-term declines and recovery of meadow area across the world's seagrass bioregions. Glob. Chang. Biol. 27, 4096–4109.
Eddy, T.D., Lam, V.W.Y., Reygondeau, G., Cisneros-Montemayor, A.M., Greer, K., Palomares, M.L.D., et al., 2021. Global decline in capacity of coral reefs to provide ecosystem services. One Earth 4, 1278–1285.
Edgar, G.J., Stuart-Smith, R.D., 2014. Systematic global assessment of reef fish communities by the reef life survey program. Sci. Data 1.
Edgar, G.J., Barrett, N.S., Morton, A.J., 2004. Biases associated with the use of underwater visual census techniques to quantify the density and size-structure of fish populations. J. Exp. Mar. Biol. Ecol. 308, 269–290.
Edgar, G.J., Bates, A.E., Bird, T.J., Jones, A.H., Kininmonth, S., Stuart-Smith, R.D., et al., 2016. New approaches to marine conservation through the scaling up of ecological data. Annu. Rev. Mar. Sci. 8, 435–461.
Edgar, G.J., Cooper, A., Baker, S.C., Barker, W., Barrett, N.S., Becerro, M.A., et al., 2020. Establishing the ecological basis for conservation of shallow marine life using reef life survey. Biol. Conserv. 252, 108855.
Edgar, G.J., Stuart-Smith, R.D., Heather, F.J., Barrett, N.S., Turak, E., Sweatman, H., et al., 2023. Continent-wide declines in shallow reef life over a decade of ocean warming. Nature 615, 858–865.
Elith, J., Leathwick, J.R., 2009. Species distribution models: ecological explanation and prediction across space and time. Annu. Rev. Ecol. Evol. Syst. 40, 677–697.
Ferro, I., Morrone, J.J., 2014. Biogeographical transition zones: A search for conceptual synthesis. Biol. J. Linn. Soc. 113, 1–12.
Filbee-Dexter, K., Wernberg, T., 2018. Rise of turfs: a new battlefront for globally declining kelp forests. BioScience 68, 64–76.
Fourqurean, J.W., Duarte, C.M., Kennedy, H., Marbà, N., Holmer, M., Mateo, M.A., et al., 2012. Seagrass ecosystems as a globally significant carbon stock. Nat. Geosci. 5, 505–509.
Funnell, T., O'Flanagan, C.H., Williams, M.J., McPherson, A., McKinney, S., Kabeer, F., et al., 2022. Single-cell genomic variation induced by mutational processes in cancer. Nature 612, 106–115.
Halpern, B.S., Frazier, M., Afflerbach, J., Lowndes, J.S., Micheli, F., O'Hara, C., et al., 2019. Recent pace of change in human impact on the world's ocean. Sci. Rep. 9, 11609.
Harley, C.D.G., Randall Hughes, A., Hultgren, K.M., Miner, B.G., Sorte, C.J.B., Thornber, C.S., et al., 2006. The impacts of climate change in coastal marine systems. Ecol. Lett. 9, 228–241.
Hill, N.A., Blount, C., Poore, A.G.B., Worthington, D., Steinberg, P.D., 2003. Grazing effects of the sea urchin *Centrostephanus rodgersii* in two contrasting rocky reef habitats: effects of urchin density and its implications for the fishery. Mar. Freshw. Res. 54, 691–700.
Hixon, M.A., Beets, J.P., 1993. Predation, prey refuges, and the structure of coral-reef fish assemblages. Ecol. Monogr. 63, 77–101.
Horta e Costa, B., Assis, J., Franco, G., Erzini, K., Henriques, M., J., G.E, et al., 2014. Tropicalization of fish assemblages in temperate biogeographic transition zones. Mar. Ecol. Prog. Ser. 504, 241–252.
Hughes, T.P., Bellwood, D.R., Folke, C.S., McCook, L.J., Pandolfi, J.M., 2007. No-take areas, herbivory and coral reef resilience. Trends Ecol. Evol. 22, 1–3.
Jayathilake, D.R.M., Costello, M.J., 2020. A modelled global distribution of the kelp biome. Biol. Conserv. 252, 108815.
Jones, L.A., Mannion, P.D., Farnsworth, A., Valdes, P.J., Kelland, S.-J., Allison, P.A., 2019. Coupling of palaeontological and neontological reef coral data improves forecasts of biodiversity responses under global climatic change. R. Soc. Open Sci. 6, 182111.

Jouffray, J.-B., Nyström, M., Norström, A.V., Williams, I.D., Wedding, L.M., Kittinger, J. N., et al., 2015. Identifying multiple coral reef regimes and their drivers across the hawaiian archipelago. Philos. Trans. R. Soc. B 370, 20130268.
Jurgens, L.J., Ashlock, L.W., Gaylord, B., 2022. Facilitation alters climate change risk on rocky shores. Ecology 103, e03596.
Krumhansl, K.A., Okamoto, D.K., Rassweiler, A., Novak, M., Bolton, J.J., Cavanaugh, K. C., et al., 2016. Global patterns of kelp forest change over the past half-century. Proc. Natl. Acad. Sci. 113, 13785–13790.
Lande, R., DeVries, P.J., Walla, T.R., 2000. When species accumulation curves intersect: implications for ranking diversity using small samples. Oikos 89, 601–605.
Lecours, V., Deviller, R., Schneider, D.C., Lucieer, V.L., Brown, C.J., Edinger, E.N., 2015. Spatial scale and geographic context in benthic habitat mapping: review and future directions. Mar. Ecol. Prog. Ser. 535, 259–284.
Legendre, P., Gallagher, E.D., 2001. Ecologically meaningful transformations for ordination of species data. Oecologia 129, 271–280.
Lemessa, D., Mewded, B., Alemu, S., 2023. Vegetation ecotones are rich in unique and endemic woody species and can be a focus of community-based conservation areas. Bot. Lett. 0, 1–11.
Ling, S.D., 2008. Range expansion of a habitat-modifying species leads to loss of taxonomic diversity: A new and impoverished reef state. Oecologia 156, 883–894.
Liu, L.-C., Lin, S.-M., Caragnano, A., Payri, C., 2018. Species diversity and molecular phylogeny of non-geniculate coralline algae (corallinophycidae, rhodophyta) from taoyuan algal reefs in northern Taiwan, including crustaphytum gen. Nov. and three new species. J. Appl. Phycol. 30, 3455–3469.
Logan, C.H.A., Fotopoulou, S., 2020. Unsupervised star, galaxy, QSO classification - application of HDBSCAN⋆. A&A 633, A154.
Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., et al. (Eds.), Advances in Neural Information Processing Systems, 30. Curran Associates, Inc., pp. 4765–4774
Marzloff, M.P., Oliver, E.C.J., Barrett, N.S., Holbrook, N.J., James, L., Wotherspoon, S.J., et al., 2018. Differential vulnerability to climate change yields novel deep-reef communities. Nat. Clim. Chang. 8, 873–878.
McInnes, L., Healy, J., Astels, S., 2017. Hdbscan: hierarchical density based clustering. J. Open Source Softw. 2.
McInnes, L., Healy, J., Melville, J., 2020. UMAP: uniform manifold approximation and projection for dimension reduction. arXiv. https://arxiv.org/abs/1802.03426.
McKenzie, L.J., Nordlund, L.M., Jones, B.L., Cullen-Unsworth, L.C., Roelfsema, C., Unsworth, R.K.F., 2020. The global distribution of seagrass meadows. Environ. Res. Lett. 15, 074041.
Melvin, R.L., Xiao, J., Godwin, R.C., Berenhaut, K.S., Salsbury Jr., F.R., 2018. Visualizing correlated motion with HDBSCAN clustering. Protein Sci. 27, 62–75.
Milošević, D., Medeiros, A.S., Stojković Piperac, M., Cvijanović, D., Soininen, J., Milosavljević, A., et al., 2022. The application of uniform manifold approximation and projection (UMAP) for unconstrained ordination and classification of biological indicators in aquatic ecology. Sci. Total Environ. 815, 152365.
Moulavi, D., Jaskowiak, P.A., Campello, R.J.G.B., Zimek, A., Sander, J., 2014. Density-based clustering validation. In: Proceedings of the 2014 SIAM International Conference on Data Mining (SDM), pp. 839–847.
Ohlsson, M., Eklöf, A., 2020. Spatial resolution and location impact group structure in a marine food web. Ecol. Lett. 23, 1451–1459.
Oksanen, J., Minchin, P.R., 2002. Continuum theory revisited: what shape are species responses along ecological gradients? Ecological Modelling 157 (2–3), 119–129.
Packer, J.S., Zhu, Q., Huynh, C., Sivaramakrishnan, P., Preston, E., Dueck, H., et al., 2019. A lineage-resolved molecular atlas of *C. elegans* embryogenesis at single-cell resolution. Science 365, eaax1971.
Pelletier, D., Selmaoui-Folcher, N., Bockel, T., Schohn, T., 2020. A regionally scalable habitat typology for assessing benthic habitats and fish communities: application to New Caledonia reefs and lagoons. Ecol. Evol. 10, 7021–7049.
Pessarrodona, A., Filbee-Dexter, K., Alcoverro, T., Boada, J., Feehan, C.J., Fredriksen, S., et al., 2021. Homogenization and miniaturization of habitat structure in temperate marine forests. Glob. Chang. Biol. 27, 5262–5275.
Pinheiro, H.T., Rocha, L.A., Macieira, R.M., Carvalho-Filho, A., Anderson, A.B., Bender, M.G., et al., 2018. South-Western Atlantic reef fishes: zoogeographical patterns and ecological drivers reveal a secondary biodiversity Centre in the Atlantic Ocean. Divers. Distrib. 24, 951–965.
Rhodes, C.J., Henrys, P., Siriwardena, G.M., Whittingham, M.J., Norton, L.R., 2015. The relative value of field survey and remote sensing for biodiversity assessment. Methods Ecol. Evol. 6, 772–781.
Robert, K., Jones, D.O.B., Tyler, P.A., Van Rooij, D., Huvenne, V.A.I., 2015. Finding the hotspots within a biodiversity hotspot: fine-scale biological predictions within a submarine canyon using high-resolution acoustic mapping techniques. Mar. Ecol. 36, 1256–1276.
Rocha, J.C., Peterson, G.D., Biggs, R., 2015a. Regime shifts in the anthropocene: drivers, risks, and resilience. PLoS One 10, 1–16.
Rocha, J., Yletyinen, J., Biggs, R., Blenckner, T., Peterson, G., 2015b. Marine regime shifts: drivers and impacts on ecosystems services. Philosoph. Trans. Royal Soc. B Biol. Sci. 370, 20130273.
Sissini, M.N., Koerich, G., de Barros-Barreto, M.B., Coutinho, L.M., Gomes, F.P., Oliveira, W., et al., 2022. Diversity, distribution, and environmental drivers of coralline red algae: the major reef builders in the southwestern Atlantic. Coral Reefs 41, 711–725.
Sonnewald, M., Dutkiewicz, S., Hill, C., Forget, G., 2020. Elucidating ecological complexity: unsupervised learning determines global marine eco-provinces. Sci. Adv. 6.
Spalding, M.D., Fox, H.E., Allen, G.R., Davidson, N., Ferdaña, Z.A., Finlayson, M., et al., 2007. Marine ecoregions of the world: a bioregionalization of coastal and shelf areas. BioScience 57, 573–583.
State of the marine and coastal environment report, 2021. Commissioner for Environmental Sustainability.
Stuart-Smith, R.D., Edgar, G.J., Clausius, E., Oh, E.S., Barrett, N.S., Emslie, M.J., et al., 2022. Tracking widespread climate-driven change on temperate and tropical reefs. Curr. Biol. 32, 4128–4138.e3.
Stuble, K.L., Bewick, S., Fisher, M., Forister, M.L., Harrison, S.P., Shapiro, A.M., et al., 2021. The promise and the perils of resurveying to understand global change impacts. Ecol. Monogr. 91, e01435.
Sunday, J.M., Fabricius, K.E., Kroeker, K.J., Anderson, K.M., Brown, N.E., Barry, J.P., et al., 2017. Ocean acidification can mediate biodiversity shifts by changing biogenic habitat. Nat. Clim. Chang. 7, 81–85.
Van Rijsbergen, C.J., 1979. Information Retrieval, 2nd edn. Butterworths.
Venna, J., Kaski, S., 2001. Neighborhood preservation in nonlinear projection methods: An experimental study. In: Dorffner, G., Bischof, H., Hornik, K. (Eds.), Artificial Neural Networks — ICANN 2001. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 485–491.
Vergés, A., Tomas, F., Cebrian, E., Ballesteros, E., Kizilkaya, Z., Dendrinos, P., et al., 2014. Tropical rabbitfish and the deforestation of a warming temperate sea. J. Ecol. 102, 1518–1527.
Vermeulen, M., Smith, K., Eremin, K., Rayner, G., Walton, M., 2021. Application of uniform manifold approximation and projection (UMAP) in spectral imaging of artworks. Spectrochim. Acta A Mol. Biomol. Spectrosc. 252, 119547.
Waycott, M., Duarte, C.M., Carruthers, T.J.B., Orth, R.J., Dennison, W.C., Olyarnik, S., et al., 2009. Accelerating loss of seagrasses across the globe threatens coastal ecosystems. Proc. Natl. Acad. Sci. 106, 12377–12381.
Wernberg, T., Bennett, S., Babcock, R.C., de Bettignies, T., Cure, K., Depczynski, M., et al., 2016. Climate-driven regime shift of a temperate marine ecosystem. Science 353, 169–172.
Whitaker, S.G., Ambrose, R.F., Anderson, L.M., Fales, R.J., Smith, J.R., Sutton, S., et al., 2023. Ecological restoration using intertidal foundation species: considerations and potential for rockweed restoration. Ecosphere 14, e4411.
Wicaksono, P., Aryaguna, P.A., Lazuardi, W., 2019. Benthic habitat mapping model and cross validation using machine-learning classification algorithms. Remote Sens. 11.
Wirabuana, A., Litaay, M., Moka, W., Priosambodo, D., 2019. Distribution of soft coral octocorallia (alcyonacea) in coastal waters of gonda, polewali-mandar, West Sulawesi. In: IOP Conference Series: Earth and Environmental Science, 279, p. 012053.
Zawada, K.J.A., Dornelas, M., Madin, J.S., 2019. Quantifying coral morphology. Coral Reefs 38, 1281–1292.