CONSULTANCY ON VME MAPPING

Boris Leroy, Berta Ramiro Sánchez and Alexis Martin

FEBRUARY 1, 2022

Laboratory of biology of aquatic organisms and ecosystems
MUSEUM NATURAL D'HISTOIRE NATURELLE
43 rue Cuvier, CP 2675231 Paris Cedex 05, France

Funded by the European Union

1. Introduction

Vulnerable Marine Ecosystems (VMEs) are ecosystems at potential risk from the effects of fishing activities or other anthropogenic disturbances, as determined by the vulnerability of their components (e.g., species, communities, or habitats) (FAO, 2009). The United Nations require states and regional fisheries management organizations/agreements to implement measures to prevent significant adverse impacts on VMEs, which can include closing areas to fishing (UNGA, 2007). Several species and taxonomic groups have been identified as indicators of VMEs in particular ocean regions to assist management authorities in their protection. VME taxa possess traits that make them particularly vulnerable to disturbance (slow growth, late maturity, longevity, and fragility), and include species that can form structurally complex frameworks such as coral reefs, coral gardens, and sponge grounds- which provide three-dimensional structures that enhance local biodiversity and act as areas of functional significance (e.g., spawning, breeding, and nursery grounds for fish stocks and particular life-history stages, and habitat for rare, threatened or endangered species of the habitat) (FAO, 2009).

In the Southern Indian Ocean, the Southern Indian Ocean Fisheries Agreement (SIOFA) has recently taken a series of steps towards the protection of VMEs. SIOFA's management efforts have identified five Benthic Protected Areas (BPAs) where bottom-fishing trawling is not permitted and have adopted a list of VME indicator taxa following UNGA resolutions. For SIOFA to meet its management obligations (UNGA, 2007), it requires scientific advice informed by maps of where VMEs are known to occur, or likely to occur, in the Agreement Area. The aim of this consultancy is to map the biogeographical regions of indicator taxa of VME, and VME habitat, in the Southern Indian Ocean to improve knowledge of VMEs within SIOFA's management area and help inform additional management measures to mitigate risks to VMEs. We aim to achieve this through the Terms of Reference (ToR) and applied deliverables established for the consultancy (Table 1).

The SIOFA area encompasses a large area with a rich biodiversity, especially in terms of VME indicator taxa, with an observed species richness of 1921 species and an estimated total richness of 2906 species (updated from Ramiro-Sánchez et al., 2021). Occurrence records of VME indicator taxa are typically scarce and collected without standardised methodologies. In ToR1 and ToR2, we assess their accuracy, quality, and comprehensiveness to properly evaluate the uncertainties of any map that will be derived from them. In ToR2, we further engage in habitat suitability modelling for predicting distribution patterns of VME indicator taxa, as such models are considered useful for marine ecosystem management (Ross and Howell, 2013; Reiss et al., 2014). In ToR3, we develop several bioregionalization approaches that provide maps of the extent of VME-based bioregions.

Because interpreting at once the distribution of hundreds of species is complex and makes it difficult to derive management recommendations, biogeographers often engage in biogeographical classifications, also known as bioregionalization (Ebach and Parenti, 2015). Bioregionalizations partition the geographical space into biological and physical units based on the distribution of multiple species, communities, ecosystems, or other biological characteristics– these units are called biogeographical regions or bioregions. Bioregions are therefore a simplification (a model) of the true biogeographical distribution of biodiversity,

which is generally based on taxa that share similar distributions because of similar ecological and physical preferences and of a shared history (Lomolino et al., 2016; Leroy et al., 2019; Woolley et al., 2020). These bioregional classifications are an indispensable component of the planning and implementation process of protected areas (CBD, 2010).

There are many predictive approaches to map bioregions, which can be generally classified into three categories (Woolley et al. 2020). The first one consists in grouping biological features into bioregions first, and then spatially predict these bioregions ("group first, then predict"). The second one consists in spatially predicting all biological features first, and then group them with a clustering approach ("predict first, then group"). The last one consists in grouping and predicting bioregions in a single modelling approach ("analyze simultaneously"). Each method has its set of advantages and disadvantages that make them complementary in the information they provide; although, ultimately, the kind of data available will often limit the choice of method (Woolley et al., 2020).

In this progress report, we synthesise results for Terms of Reference 1 to 3 of this consultancy (Table 1) and update preliminary observations presented at the PAEWG Annual meeting in March 2021. We note that improvements of the models will continue until the end of the consultancy (ToR4) and thus, results shown here should be interpreted with caution.

Table 1. Terms of reference and corresponding deliverables for the consultancy.

Terms of Reference	Deliverables
ToR1: Compile, clean and verify occurrence and environmental data	Consolidated VME occurrence dataset. Consolidated environmental variables dataset. Maps of VME taxa occurrence.
ToR2: Evaluate the quality of occurrence data and determine spatial scale and taxonomic resolution.	Indicators of data quality. Maps of optimal resolutions of analysis. Maps of predicted VME taxa occurrences.
ToR3: Develop and compare multiple bioregionalization schemes and approaches.	Maps of bioregionalization based on occurrence data. Maps of bioregionalization based on taxa distribution models. Maps of bioregionalization based on generalised dissimilarity models or other appropriate modelling techniques.
ToR4: Reporting	Scientific reports to PAEWG and SC annual meetings, 30 days prior to meeting commencement date. Final report to PAEWG and SC.

2. Methods

2.1 Data

Biological data

Based on the list of VME indicators adopted by SIOFA[1], we downloaded occurrence records from the public databases the Ocean Biodiversity Information System[2] (OBIS), the Global Biodiversity Information Facility[3] (GBIF), NOAA's Deep-sea Corals Data Portal[4], and Smithsonian Natural History Museum[5]. We also obtained occurrence records from SIOFA's observer programme and research campaigns led by the Muséum National d'Histoire Naturelle. We obtained records for the whole Southern Indian Ocean to account for ecological continuity and because of the scant availability of data within SIOFA's management area.

We applied verification procedures for taxonomic consistency, error detection, as well as evaluation of records in the environmental space. Specifically, we first checked species names against the most updated authority, the World Register of Marine Species (WoRMS 2021), for synonyms and fossil records. Secondly, we applied automatic error and outlier detection using the function clean-coordinates from the R package CoordinateCleaner version 2.0-18 (Zizka et al., 2019). We tested for equal coordinates, coordinates over land using the Natural Earth data ocean shapefile version 4.1.0 (www.naturalearthdata.com, accessed November 2020), and zero coordinates. For duplicates of GBIF and OBIS, we used the catalogue number and geographical coordinates to filter out potential duplicates.

The broad taxonomic resolution of the VME taxa list rendered the download of shallow water species, particularly of zooxanthellate corals, whose environmental requirements are different from deep-sea taxa. Although the distribution of many deep-sea corals (i.e., azooxanthellate) expands to shallower depths, the deep sea is typically defined as waters below 200 m. Therefore, in order to retain only deep-water species, we applied several filters to the dataset to exclude zooxanthellate corals, which generally do not occur below 50m water depth (Cairns, 2007). First, we individually assessed the depth distribution of each species and applied the following rule: a species was kept if 90% of its records were below 200m water depth. With such rule we adopt the general definition of deep sea as waters below 200 m, but also incorporate the ecology of species and avoid biased predictions led by strict depth cut-offs. To implement the rule and assess the depth range of records, we relied on their original recorded depth as indicated in the sample record. In the cases where this information was not available, we used the General Bathymetric Chart of the Oceans (GEBCO, see next paragraph) to assign depths. Second, we further filtered species known to be tropical as we worked through with peer-reviewed deep-sea taxa lists (Cairns, 2021; Kocsis et al., 2018). The working dataset comprised 1991 species (Table 2).

[1] http://apsoi.org/meetings/sc4
[2] https://obis.org/
[3] https://www.gbif.org/
[4] https://deepseacoraldata.noaa.gov/
[5] https://collections.nmnh.si.edu/

Table 2. Number of species, genera and family used for analyses. Depending on how many occurrences each taxa has (column « Cut-off threshold » in the table), the total number of records for each taxonomic level will vary. Cut-off thresholds were chosen as: "All": all species in the dataset; "5": a taxon has at least 5 occurrences; "10": a taxon has at least 10 occurrences; "30": a taxon has at least 30 occurrences. These thresholds were used for tuning the model parameters of species.

Cut-off occurrence threshold	Species	Genus	Family
All	1991	755	267
≥ 5	734	443	213
≥ 10	466	288	171
≥30	155	109	93

Environmental data

Environmental data (Table 3) included a global bathymetry model (GEBCO 2021 Grid; GEBCO Compilation Group, 2021) downloaded at a resolution of 0.004°. We downloaded additional environmental variables from the Bio-Oracle v2.0 and v2.1 datasets (Assis et al., 2018; Tyberghein et al., 2021) that represent conditions between 2000 and 2014 to use as predictors to model distributions. The variables included sea bottom temperature, dissolved O_2 concentration, salinity, SiO_4^{4-}, NO_3^{2-} and PO_4^{3-} concentrations at the seafloor. For all variables, we downloaded annual mean, minimum and maximum values. We also included current velocity at the seafloor and primary productivity at sea surface. Particulate organic flux at seafloor was downloaded from the Global Marine Environment Datasets[6] (GMED). In addition, we included the layers aragonite and calcite saturation horizon downloaded from OCEANSODA[7] at a resolution of 1°.

In order to calculate spatial autocorrelation and evaluate models with equal area grid cells, we used the Albers Equal Area projection centred at latitude 30°S and longitude 80°E to reproject the environmental variables.

[6] http://gmed.auckland.ac.nz/
[7] https://esa-oceansoda.org/outputs/

Table 3. Environmental variables used as predictors of taxa distributions.

Data source	Description
Depth	General Bathymetric Chart of the Oceans – GEBCO2021 – 0.004 degrees. Raster data.
Temperature (min, max, mean, range)	Bio-Oracle, 0.083 degrees. Raster data.
Salinity (min, max, mean, range)	Bio-Oracle, 0.083 degrees. Raster data.
Dissolved Oxygen (min, max, mean, range)	Bio-Oracle, 0.083 degrees. Raster data.
Silicone (min, max, mean, range)	Bio-Oracle, 0.083 degrees. Raster data.
Phosphate (min, max, mean, range)	Bio-Oracle, 0.083 degrees. Raster data.
Nitrate (min, max, mean, range)	Bio-Oracle, 0.083 degrees. Raster data.
Primary productivity at surface (min, max, mean)	Bio-Oracle, 0.083 degrees. Raster data.
Currents velocity (min, max, mean, range)	Bio-Oracle, 0.083 degrees. Raster data.
Omega Aragonite	OCEANSODA, 1 degree. Raster data.
Omega Calcite	OCEANSODA, 1 degree. Raster data.
Particulate Organic Carbon	GMED, 0.083 degrees. Raster data.

2.2 Indicators of data quality

To investigate the accuracy of the consolidated occurrence dataset, completeness, and uncertainty levels, we produced several maps of distribution of VME indicator taxa and indicators of data quality (Soberón et al., 2007; Leroy et al., 2017).

Specifically, we first mapped the point-locality distributions of VME indicator taxa and the gridded observed richness. Second, in each grid cell, we estimated the theoretical total species richness with three non-parametric incidence-based richness estimators: ICE (Gotelli & Colwell, 2011), Chao2 and Jackknife (Chiu et al., 2014). These methods project the estimated total number of species based on the observed number of species in the sample plus a correction based on the number of rare species (typically, singletons or doubletons), to account for the unobserved fraction of rare species. We chose non-parametric richness estimators because they have been proven to perform better than other richness estimators (Walther & Moore, 2005). Third, we produced maps of estimated sampling completeness by dividing the observed species richness in each cell by the estimated total richness. Because species richness estimators can be biased when the sampling intensity and observed richness are low, resulting in overestimations of data completeness (Leroy, 2012), we defined a richness threshold of 30 below which the completeness was set to zero. We produced these different maps at the 1° x 1° spatial resolution. We chose this resolution as a trade-off between data quantity and usefulness of the bioregion predictions. Exploration of the environmental gradients of records at different resolutions assisted in the selection.

2.3 Bioregionalization schemes

2.3.1 Bioregions based on observed occurrences

This first method consists in grouping observed distributions of VME indicator taxa into bioregions.

We divided the Southern Indian Ocean into a 1° latitude-longitude grid. We created a species/grid cell contingency table of presence-absence and transformed it into a bipartite occurrence network (Vilhena and Antonelli, 2015). In essence, a network is a representation of how species are distributed and connected between sites (here, grid cells). A network comprises nodes that can be connected to each other by links (or edges). Here, nodes represent grid cells and, separately, the species that occur in them, while edges link grid cell and species nodes indicating whether a species is present in a grid cell. The network is bipartite because each type of node, grid cell and species, cannot connect to another node of the same type; that is, species nodes can only connect to grid cell nodes and grid cell nodes can only connect to species nodes.

We created networks using the R package "biogeonetworks" (Leroy, 2021). We visualised networks with the software Gephi 0.9.2 (Bastian et al., 2009) and used the ForceAtlas2 algorithm (Jacomy et al., 2014) to spatialise them. The ForceAtlas2 algorithm groups nodes that are interconnected (i.e., grid cells and species belonging to the same biogeographical region) and separates groups of nodes that are not interconnected (different biogeographical regions). The graphical representation of the network facilitates the analysis of the data. For example, we can detect the erroneous inclusion of shallow water species.

In order to detect biogeographical regions, we applied a community-detection algorithm to the network. Community-detection algorithms aim at grouping together nodes that belonged to the same biogeographical region. As a community detection algorithm, we chose "Map Equation" (Rosvall and Bergstrom, 2008), as it has been recommended in previous biogeographical works (Bloomfield et al., 2017; Edler et al., 2017; Rojas et al., 2017; Vilhena and Antonelli, 2015) and it features hierarchical clustering. We ran Map Equation with 1000 trials to find the optimal clustering and we ran it with hierarchical clustering to test whether larger regions showed a nested hierarchy of subregions.

2.3.2 Predicted bioregions

2.3.2.1 Group first, then predict

We aimed to model the distribution of biogeographical regions identified in step 2.3.1. We proceeded as follows: (1) we collated environmental predictors for VME indicator taxa; (2) we modelled the relationship between bioregions and predictors using an ensemble modelling approach based on multiple algorithms, in order to obtain predictive suitability maps for each biogeographical region, (3) we stacked these predictive suitability maps in order to obtain in each pixel the predicted suitability of biogeographical regions of VME indicator taxa.

Our models comprised all species under the taxonomic categories listed in SIOFA's VME taxa list for the Southern Indian Ocean (Longitude 20°E-147°E, Latitude 13°N-65°S) at a 1° x 1°

spatial resolution. We chose this resolution after preliminary observations of the occurrence records: there were few records to work at finer resolutions (i.e., < 1° latitude-longitude), risking high variability due to sampling bias; on the other hand, coarser resolutions (i.e., 2° latitude-longitude) were not representative of the environmental conditions driving distributions. Thus, 1° x 1° spatial resolution appeared sufficient to identify broad geographic patterns. We selected temperature, salinity, dissolved oxygen, nitrate, silicate and phosphate concentrations, speed of currents, particulate organic carbon (POC) flux, aragonite and calcite saturation state horizon at the bottom, primary productivity at sea surface, and bathymetry. These environmental variables have been regarded as important in determining large-scale distribution patterns of deep-sea habitat forming species (Davies and Guinotte, 2011; Morato et al., 2020; Yesson et al., 2012; Yesson et al., 2016). We also included tectonic plates to account for dispersal restrictions.

We fitted predictive models for the three biogeographical regions identified at level 1 of the hierarchy, for which we had at least 30 available occurrences (Table 4). Biogeographical region occurrences and environmental variables were projected to Albers Equal Area, centered at 80°E longitude and 52°S latitude. We fitted SDMs in biomod2 v.3.4.13 (Thuiller et al., 2020) using seven algorithms: general linear model, general additive model, artificial neural network, multivariate adaptive regression splines, generalised boosted models, factorial discriminant analysis, and random forest. Only the random forest was tuned with 1000 pruned trees and equal sample size of presence and absence classes. We used absence data where, for each biogeographical region, an absence corresponds to the presence of another biogeographical region.

We evaluated model performance with Jaccard indices, true skill statistic (TSS) and area under the curve (AUC) (Leroy et al., 2018). We ran a 4-fold block cross-validation procedure using a block size of 441 000m² (corresponding to the minimum autocorrelation range in our variables) (Roberts et al., 2017; Valavi et al., 2019) to improve model evaluation, where data were randomly split in calibration (75%) and evaluation subsets (25%). We discarded models with a poor performance (Jaccard index < 0.3 with cross-validation) and produced an ensemble model based on the median of predictions. In the final post-processing step, we stacked all three-ensemble models and derived one map where each pixel showed the highest probability of all ensembles.

Table 4. Number of occurrences and selected variables used to predict the biogeographical regions of level 1 of the hierarchy.

Cluster	Number of occurrences	Selected variables
1	392	Mean Dissolved oxygen Maximum depth
2	71	Mean Dissolved oxygen
3	59	Maximum depth Mean Dissolved oxygen

Table 5. Number of occurrences and selected variables used to predict the biogeographical regions of level 2 of the hierarchy.

Cluster	Number of occurrences	Selected variables
1.1	87	Mean Dissolved oxygen Maximum depth
1.2	21	Mean Dissolved oxygen Maximum depth Minimum salinity
1.3	25	Minimum Primary Productivity Mean Dissolved oxygen Mean Temperature
1.5	25	Salinity range
1.7	11	Mean Dissolved oxygen Minimum salinity
2.1	23	Mean Dissolved oxygen
2.4	11	Mean Dissolved oxygen
3.1	16	Maximum depth Mean Dissolved oxygen

For the eight sub biogeographical regions obtained at level 2 of the hierarchy, we also fitted predictive models following the same procedure described above. However, because these biogeographical regions had less than 30 occurrences (Table 5), it was not possible to apply block cross-validation for the model evaluation. Instead, all occurrences were used in the models and no evaluation of model performance was carried out.

For the modelling of biogeographical regions of both hierarchical levels, the selection of variables followed a semi-automatic process using permutation-based variable importance that we describe as follows (Bellard et al., 2016; Leroy et al., 2014). First, we assessed the correlation between variables and groups of intercorrelated variables were retained. We calibrated models for each group of intercorrelated variables and the most important variable from each group was retained. If several variables had the same importance, we kept arbitrarily the first variable. Then, we calibrated one model with the non-correlated variables and the most important variables selected from the group of intercorrelated variables. From this model, we again calculated the importance of the variables and we only kept variables that reached 10% of importance for at least 50 % of the models (median value). Finally, we calculated the correlation between these variables to identify if there was a negative correlation; these variables were individually assessed and discarded. The final set of variables selected to predict each biogeographical region is presented in Table 4 and Table 5, correspondingly.

2.3.2.2 Predict first, then group

We aimed to model the current distribution of biogeographical regions of VME indicator taxa in the Southern Indian Ocean in a « predict first, then group » approach. We proceeded as follows: (1) we collated environmental predictors for VME indicator taxa; (2) we modelled the relationship between taxa and predictors with random forest models, in order to obtain predictive suitability maps for each taxon; (3) we converted suitability into binary suitable/unsuitable maps using a threshold optimised on the Jaccard index; (4) we detected bioregions on these predictive maps using the same clustering approach as in 2.3.1. Our models comprised all taxa under the taxonomic categories listed in SIOFA's VME taxa list in the Southern Indian Ocean (Longitude 20°E-147°E, Latitude 13°N-65°S). We predicted maps at the species, genus, and family levels.

We worked at a 0.083° x 0.083° resolution. This resolution is the native resolution of the environmental variables we selected, which are the same as for Region Distribution Modelling (Methods 2.2.1).

All taxa (species, genus, family) with more than five records (total: 1,390) were modelled with down-sampled presence-only random forests, using a set of 50,000 randomly sampled background points. We chose down-sampled random forests because these techniques are one of the best performing techniques for presence-only distribution models, especially in situation with low occurrence numbers (i.e., between 5 and 30 – see Valavi et al. 2021a, b). In addition, random forests are relatively fast single models. Because most VME indicator taxa have only a limited number of occurrences (Table 2), we adapted our modelling process depending on the number of occurrences, with a decreasingly complex procedure.

For taxa with more than 30 records, we applied a variable selection procedure based on variable importance permutations, in order to select the most relevant variables. For each taxon, we filtered important variables with the following criteria: (1) median importance across all permutations above 0.05; (2) in case of pairwise negative correlations below -0.4 in importance among variables, the least important variable was excluded; (3) in order to limit overfitting, we kept a maximum of one variable per 10 occurrence points, excluding the least important variables every time (e.g., for a species with 50 occurrence points, a maximum of 5 variables could be kept). Once the important variables were selected for each taxon, we calibrated and evaluated models with a 3-fold block cross-validation procedure using a block size of 697 246m² (corresponding to the minimum autocorrelation range in our variables) (Roberts et al., 2017; Valavi et al., 2019). Both presence and background points were used for the block cross-validation. To account for variations among random forest runs, we repeated each model three times, in each cross-validation fold, resulting in a total of nine models per taxon. We evaluated each model with two metrics: the Boyce index (R package enmSdm, Smith, 2020), and the gain in Area under Precision Recall (R package prg, Kull and Flach, 2022), two metrics recommended for presence-only distribution models (Leroy et al., 2018, Valavi et al., 2021b). To produce the final suitability map, we averaged all projections from the nine models, excluding models with a poor predictive ability, i.e., models with a Boyce index below 0.3.

For taxa with records ranging from 5 to 30, we chose to not apply a statistical variable selection procedure, since such a procedure presents the risk of representing sampling biases

rather than true effect of predictors. Instead, for each taxon, we chose the most frequently selected predictors in the same taxonomic order (on the basis of the selection procedure for taxa with more than 30 records). For those taxa which were in an order where there was no other taxon upon which the variable selection procedure had been run, we selected the most frequently selected predictors across the entire pool of taxa. We modelled these species with no cross-validation procedure because of the low number of records. Instead, we evaluated models on the calibration dataset, only to remove the most spurious models, using the same metrics and cut-off as above. Such a procedure implies that we cannot have a proper evaluation of the predictive performance models – hence, these models are to be interpreted with caution. We repeated each random forest three times. To produce the final suitability map, we averaged all projections from the three models, excluding models with a Boyce index below 0.3.

Once we had predicted the distributions of suitability for all taxa, we applied the bioregionalization procedure at the family, genus, and species levels. To be able to apply the bioregionalization procedure, we had to convert the predicted suitability maps into binary suitable/unsuitable maps. To do that, for each taxon, we look for the optimal suitability cut-off by maximising the Jaccard metric on the full set of presences and an equal number set of randomly selected background points, with 10 repetitions in total. We did that to limit the effect of sample prevalence bias because of the large number of background points (Leroy et al., 2018). Note that because we do not have information on the true absence of species, we cannot confirm with confidence that areas deemed as "unsuitable" are truly unsuitable. Rather, unsuitability is here defined as "either unsuitable or never sampled in similar environmental conditions". Conversely, we have confidence on areas deemed as "suitable" by our models, because we have samples for these taxa in similar environmental conditions.

On the basis of these binary suitability maps, we built two biogeographical networks. A first one, based on all taxa, including those for which models could not be reliably estimated, and a second one, based only on the taxa for which models could be reliably evaluated (i.e., taxa with more than 30 occurrences). We applied upon these networks the Map Equation community-detection algorithm in order to detect predictive bioregions.

3. Results

3.1 Indicators of data quantity and quality

Most of the available records were obtained from the publicly available repositories GBIF and OBIS, where we could find information for most taxonomic resolutions. Records collated from NOAA and Smithsonian also stored information up to species level, whereas records coming from the SIOFA Secretariat presented a coarser taxonomic rank (Order) of limited use for biodiversity assessments, although useful for model validation.

Records were mainly distributed in waters within national jurisdictions where survey effort is typically concentrated, whereas large areas of scant information dominated the centre of the SIOFA range (Figure 1**Error! Reference source not found.**). The majority of the records

occurred above 1,000 m water depth, although the distribution spanned all depths with some typically deep-sea taxa found in deeper areas (Figure 2).

The completeness index, the ratio between observed richness and estimated richness, indicated that 72% of the species richness of the area is known (Table 6). This index varied for each of the phyla, with sponges (Porifera) amongst the worst sampled and corals (Cnidaria) amongst the best sampled. However, looking in detail at the spatial distribution of the completeness index it suggests that the SIOFA area is critically under-sampled (Figure 3). Only few sites have a high completeness, which coincide with areas of high species richness and sampling intensity (Figure 3).

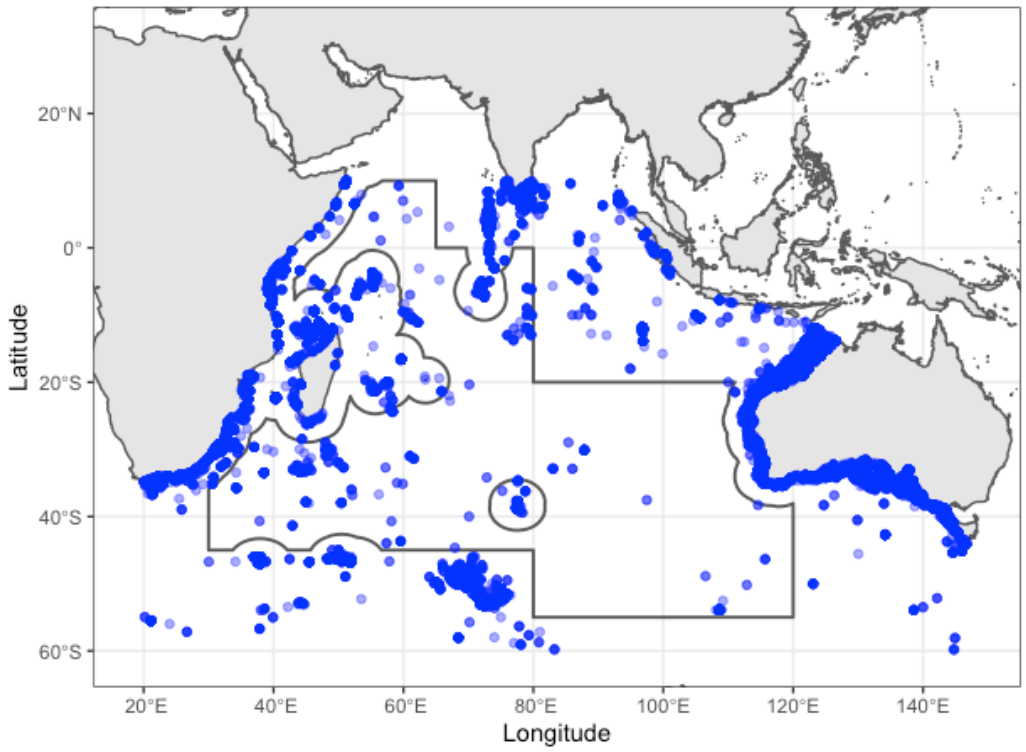

Figure 1. Distribution of taxa records over the considered area of study. Shades of blue indicate the number of occurrences, where darker shades reflect higher concentration of records.

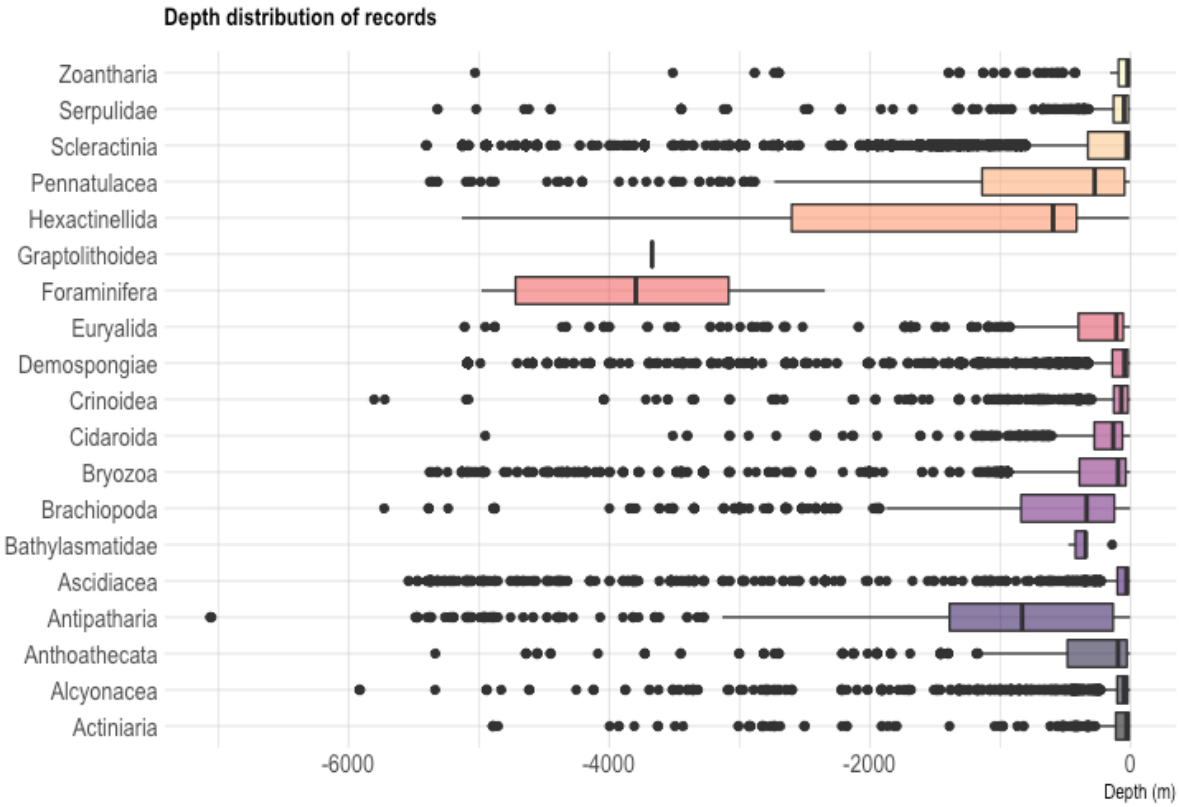

Figure 2. Depth distribution of records by VME taxa.

Table 6. Completeness and species richness of each phylum of the database. Completeness is the ratio between observed and estimated species richness (Soberón et al., 2007). The completeness is based on three estimators (Chao2, ICE and Jack1) averaged across all sites.

	Species richness	Average estimated richness	Completeness index
Database	1306	1812	0.72
Cnidaria	599	753	0.80
Echinodermata	86	118	0.73
Chordata	110	138	0.80
Porifera	264	727	0.36
Bryozoa	177	219	0.81
Foraminifera	13	19	0.68
Brachiopoda	33	41	0.80
Annelida	22	29	0.76

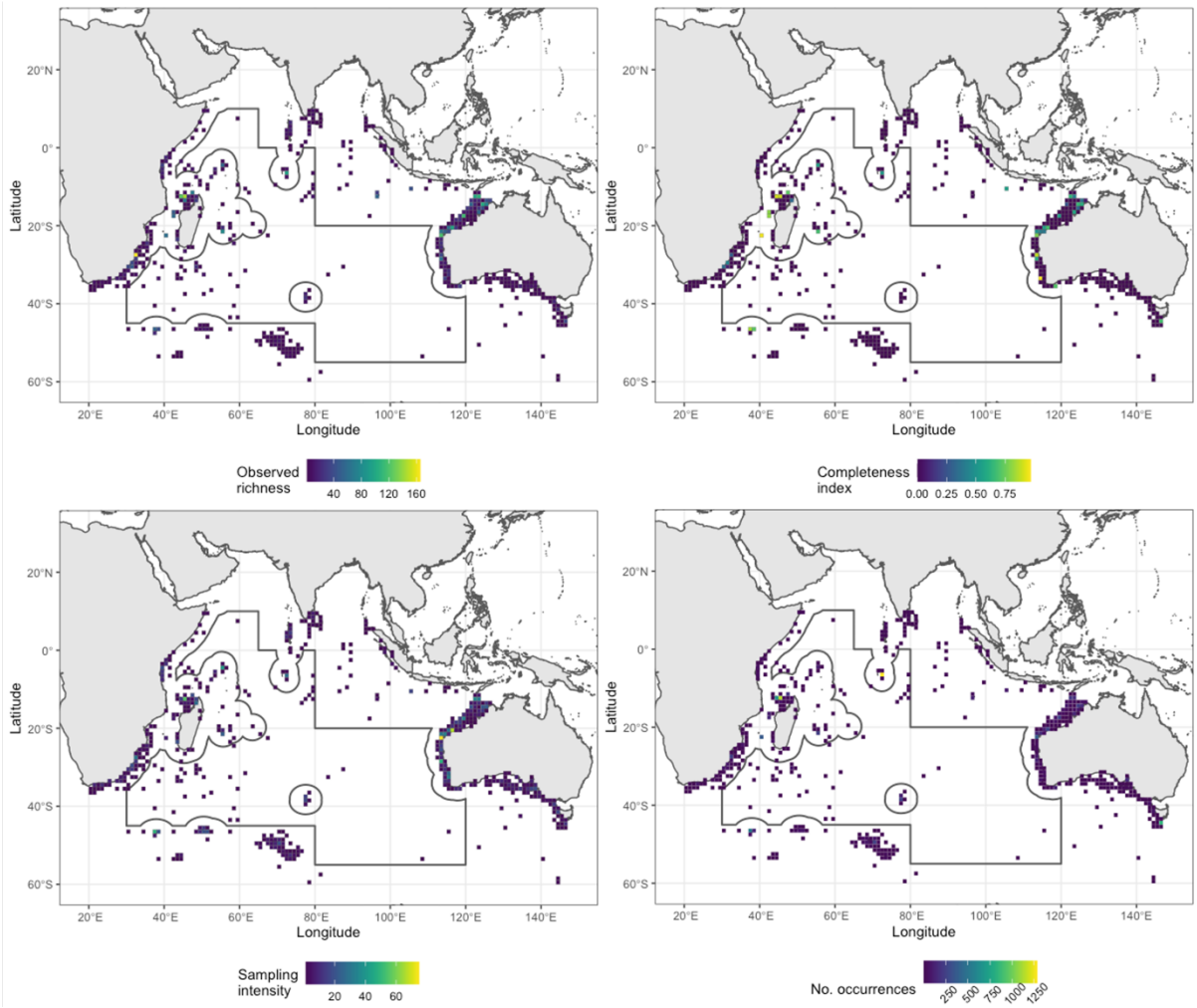

Figure 3. Indicators of data quantity and quality. From top to right: observed richness; completeness index; sampling intensity, and number of occurrences.

3.2 Bioregionalization schemes

3.1.1 Observed bioregions

We detected three biogeographical regions in the Southern Indian Ocean at the first hierarchical level (Figure 4A). Their distribution reflects the availability of the input data, which were distributed within jurisdictional waters mainly, although many records were not included in the analysis given that these were not at species level, rather at coarser taxonomic levels (e.g., order)[8]. Biogeographical region 1 is the largest encompassing a broad latitudinal and longitudinal span. This biogeographical region seems to represent the continental shelf and slope. Biogeographical region 2 represents the Southern Ocean. Biogeographical region 3 is the predominant group observed in SIOFA's management area. This group is sparsely distributed around Walter's Shoal, the Southwest Indian Ridge, east of the Chagos Lacadive Plateau, and north of Ninety-east Ridge.

At the second level of hierarchy, we observed a finer distribution within the larger biogeographical regions with distinct geographic differences (Figure 4B). Biogeographical region 1.1 was the most extended, mainly present at latitudes above 20°S. Biogeographical region 1.2 was characteristic of the continental slope of northwest Australia. Biogeographical region 1.3 was well defined around the east of South Africa. Biogeographical region 1.5 had a distinct distribution along the southeast of Australia, covering the continental shelf between latitudes 30°S – 50°S. At those latitudes on the continental slope, biogeographical region 1.7 was present. The Southern Ocean biogeographical region was divided into two subregions, where subregion 2.1 had a localised distribution north of the Kerguelen Plateau and subregion 2.4 had a disperse distribution north and south of the Sub Antarctic Front at 50°S latitude. Finally, region 3.1 was distributed across some abyssal areas.

[8] For example, all records sent by the SIOFA were excluded from all our bioregionalization analyses, because they were at a taxonomic resolution too coarse to be useful.

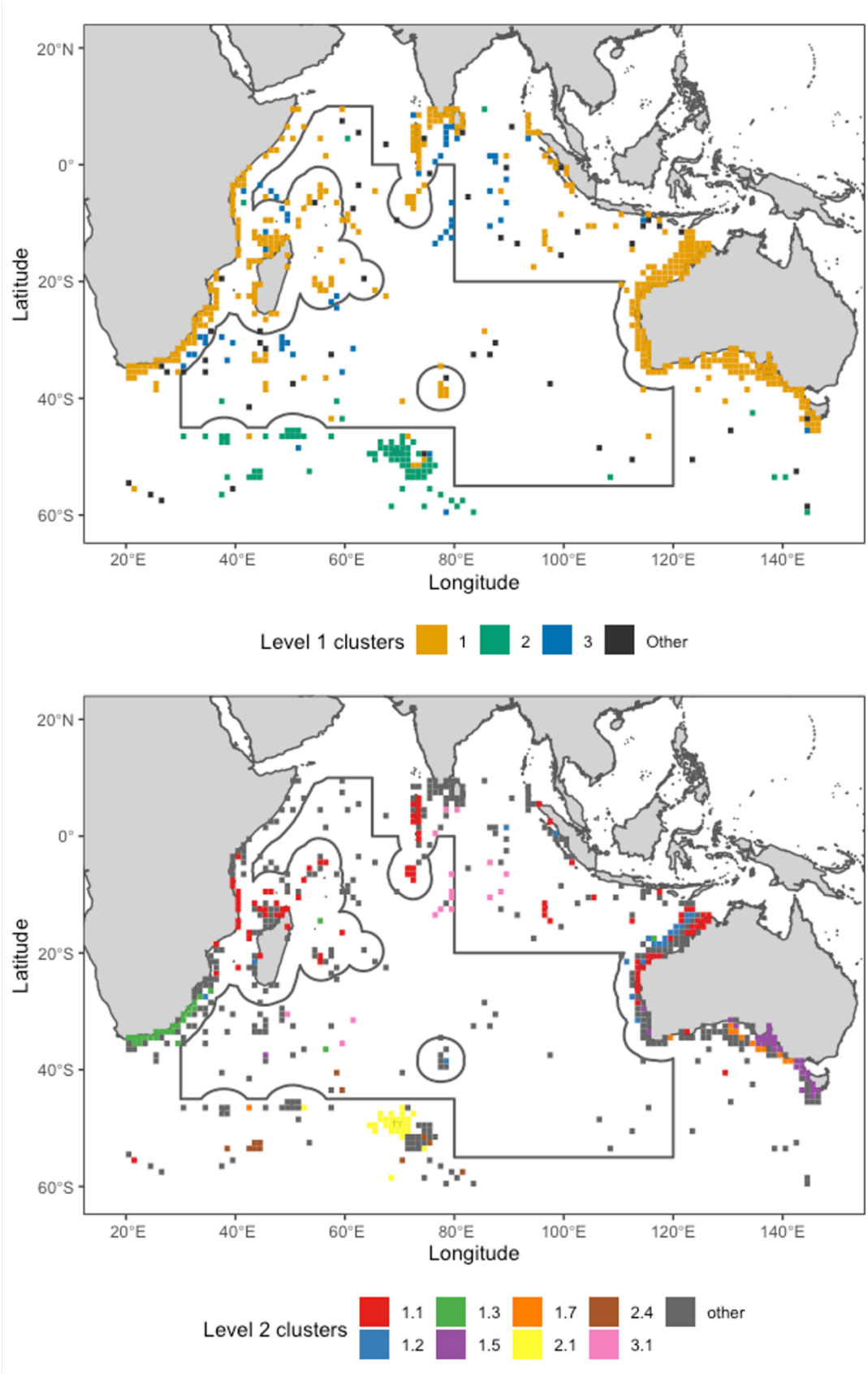

Figure 4. Biogeographical regions (clusters) of VME indicator taxa in the Southern Indian Ocean: (A) Three biogeographical regions of similar species composition detected at the first hierarchical level. (B) Finer sub-biogeographical regions detected at the second hierarchical level; black indicates no assigned subregion.

The distribution of species between grid cells is illustrated in the biogeographical network for hierarchical level 1 (Figure 5) and level 2 (Figure 6). At level 1 of the hierarchy, the biogeographical regions detected have very distinct faunas, which is observed in the network as the groups of nodes are well separated. In addition, the number of links between biogeographical regions is limited, that is, not many species are shared between regions. Specifically, we can note that the bioregion specific to the SIOFA area is very distinct from the other regions, with a limited number of links. This observation suggests that the SIOFA area has a unique composition of VME indicator taxa, not found outside this region. These observations are well reflected in the endemism patterns observed for each bioregion (Table 7).

At level 2 of the hierarchy, the biogeographical regions detected have also distinct faunas, although it is not as strong as at level 1. Because this level reflects finer patterns of level 1, the number of links is similar but the distribution of these within and among regions become well-defined. The sub biogeographical region predominant within SIOFA (biogeographical region 3.1) remains still quite unique in terms of number of links and nodes. The endemism patterns observed for each bioregion are presented in Table 8.

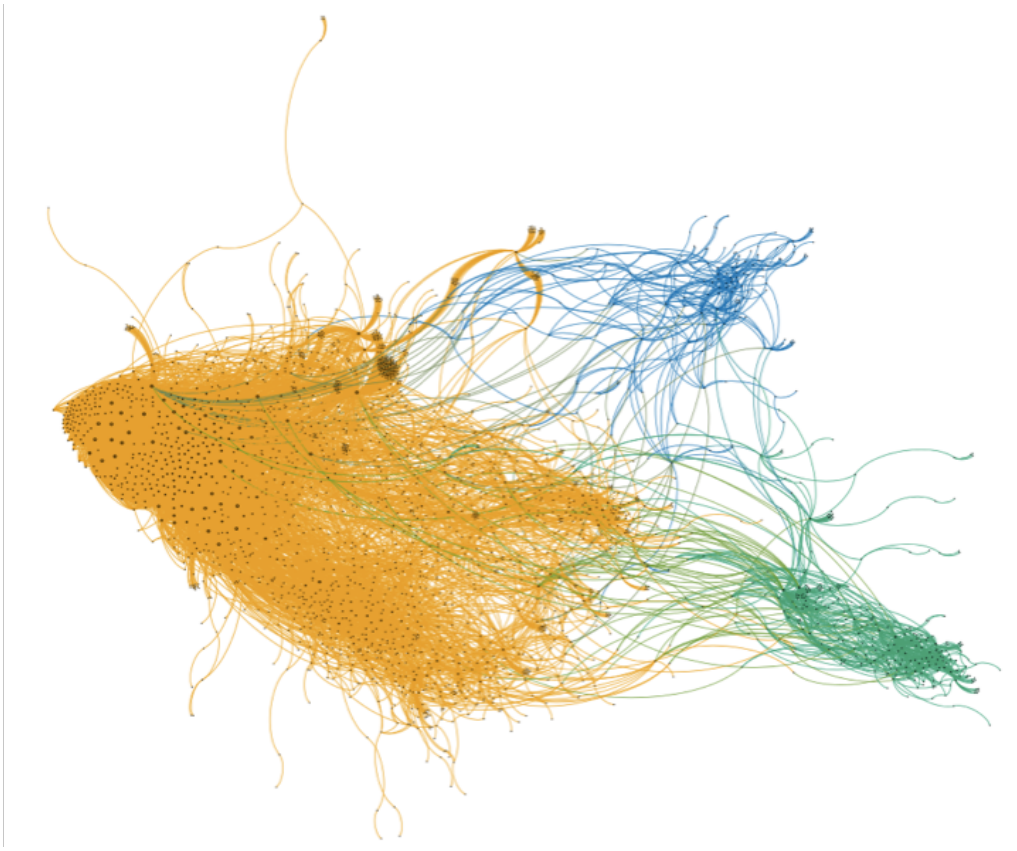

Figure 5. Biogeographical network of VME indicator taxa in the Southern Indian Ocean at the first level of hierarchy. Each biogeographical region is represented in different colours. The predominant biogeographical region found within SIOFA's management area is coloured in blue.

Figure 6. Biogeographical network of VME indicator taxa in the Southern Indian Ocean at the second level of hierarchy. Each sub biogeographical region is represented in different colours. The predominant biogeographical region found within SIOFA's management area is coloured in green. Grey indicates no assigned biogeographical region.

Table 7. Number of species and endemicity of each biogeographical region at level 1.

Level 1 Biogeographical region	Total richness	Assigned richness	Endemic richness	% Endemic species	% Indian Ocean species
1	1667	1620	1586	95.14	80.11
2	234	204	180	76.92	11.24
3	121	108	79	65.29	5.81

Table 8. Number of species and endemicity of each biogeographical region at level 2.

Level 2 Biogeographical region	Total richness	Assigned richness	Endemic richness	% Endemic species	% Indian Ocean species
1.1	754	414	219	29.05	18.81
1.3	223	133	88	39.46	5.56
1.2	189	86	26	13.76	4.71
2.1	96	60	41	42.71	2.39
1.5	96	41	12	12.50	2.39
1.7	65	20	6	9.23	1.62
3.1	22	12	4	18.18	0.55
2.4	8	3	1	12.50	0.20

3.1.2 Bioregions based on the "Group first, then predict" approach

We were able to successfully model the distribution of the three identified bioregions with our ensemble modelling approach. The map in Figure 7 shows the predicted distribution of bioregions. We added on this map the degree of uncertainty (index of confidence from 0 to 1000) that we have in our predictions, such that we are less confident of predictions in darker areas. These areas are uncertain because none of the samples we gathered were located in similar conditions – hence, model predictions in these areas are extrapolations.

The final predictive map of bioregions showed three bioregions encompassing the SIOFA area. The biogeographical regions 1 and 2 prevail from 20°N to 10°S latitude. Below 40°S latitude, the Southern Ocean biogeographical dominates the area. Areas of uncertainty in the models' predictions coincide with areas of heterogeneity in biogeographical composition and with areas where there was no initial data. Notably, the area west of St Paul and Amsterdam islands and the middle of SIOFA's management area are poorly predicted.

Although no formal evaluation of the model performance was carried out for predictions at the second level of the hierarchy, we modelled the distribution of the eight subregions (Figure 8). As previously, we incorporate confidence levels in the predictions for consistency. Note that, because of the low number of points, we cannot reliably evaluate these predictions.

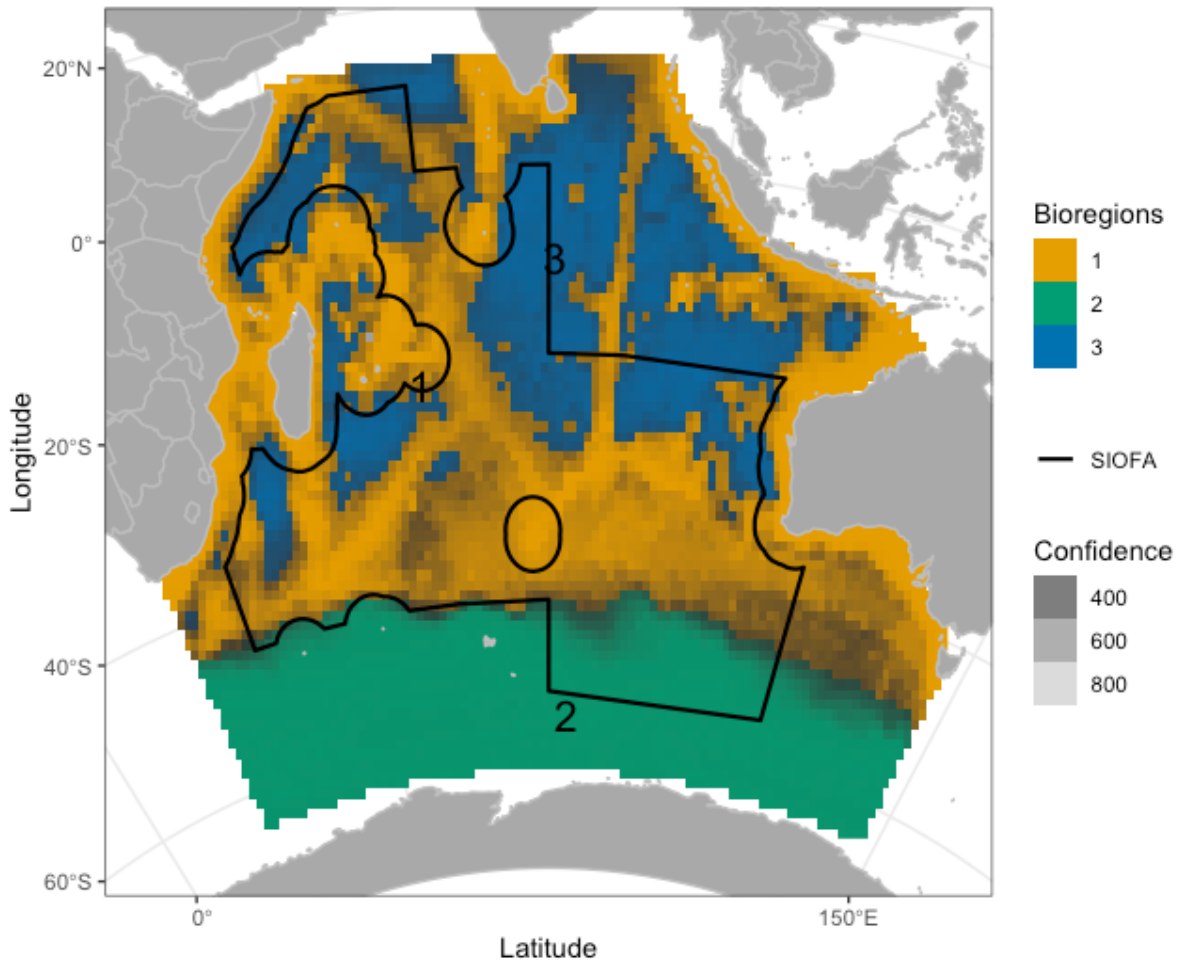

Figure 7. Predicted biogeographical regions of VME indicator taxa in the Southern Indian Ocean at the first level of the hierarchy. Areas with low confidence in the prediction are shown in darker shades of grey.

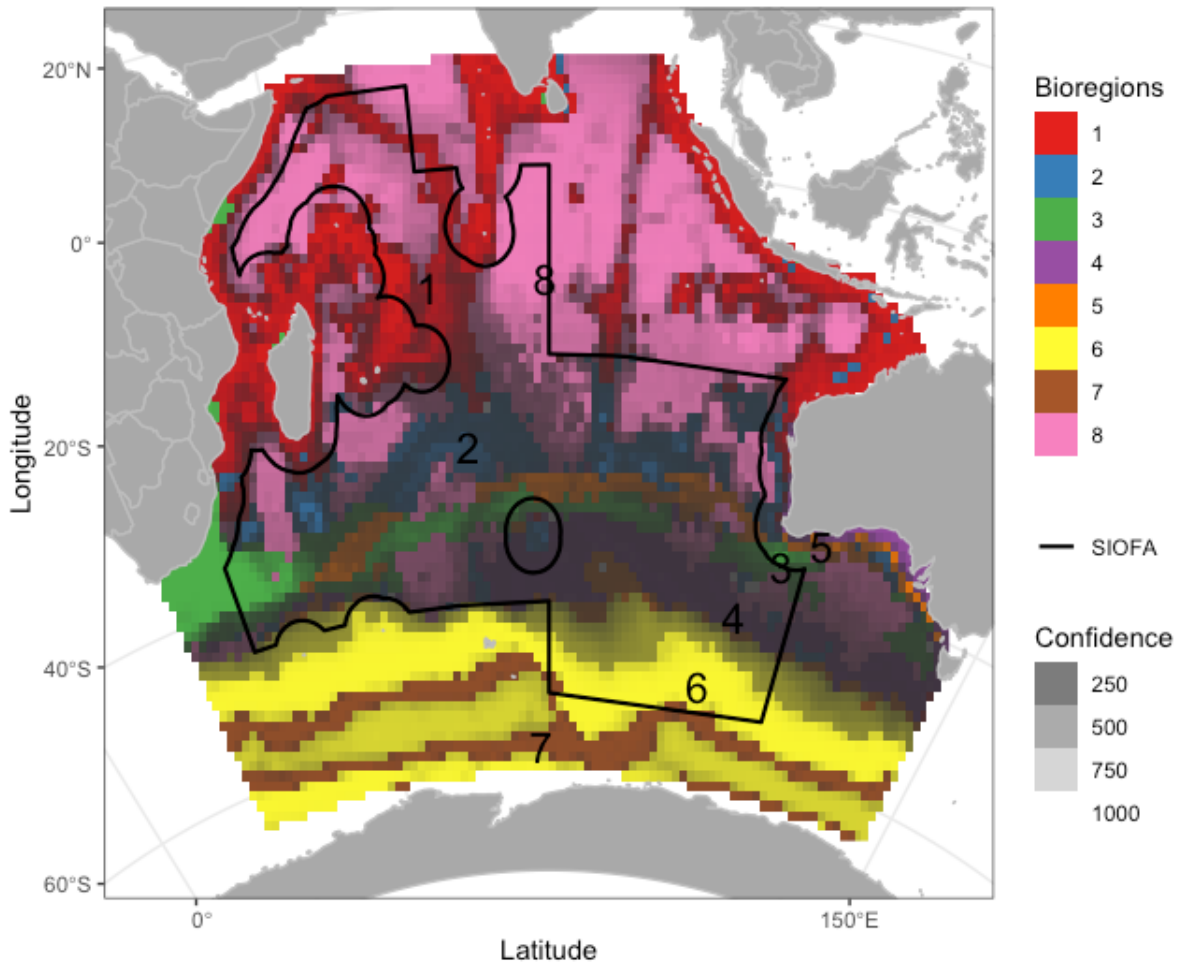

Figure 8. Predicted biogeographical regions of VME indicator taxa in the Southern Indian Ocean at the second level of the hierarchy. Areas with low confidence in the prediction are shown in darker shades of grey. Note that, because of the low number of points, we cannot reliably evaluate these predictions.

3.1.3 Bioregions based on the "Predict first, then group" approach

We produced bioregionalizations based on family, genus, and species level for all taxa and for taxa with more than 30 occurrences (Figure 9-14). Bioregionalizations based on this approach have a much finer resolution compared to the previous method, because they are based on the fine-scale predictive distribution maps for each taxon. This increase in resolution is an improvement. However, it is important to note beforehand that these predictive models are based on a limited set of variables, and do not account for species dispersal abilities. Hence, these maps are likely to connect together areas with similar environmental conditions, despite the fact that they may not share the same taxa in reality. Consequently, the maps

presented here are complementary to the maps provided by the previous predictive approach.

Family-based bioregionalization for all taxa produced 10 clusters (Figure 9). Cluster 1 was the largest, covering the entire latitudinal range of the area of study. It followed the continental shelf, slope, and shallower depths of topographically complex areas (i.e., ridges). Cluster 2 covered some of the abyssal plains up to 50°S latitude. Cluster 3 was present in the depth zone just below cluster 1 across the entire Indian Ocean. Cluster 5 represented another abyssal region. Cluster 7 was predominant south of 50°S latitude. Cluster 8 was abyssal, refining areas in cluster 2, while cluster 9, also abyssal shared the plains with cluster 5.

Family-based bioregionalization for taxa with more than 30 occurrences produced 7 clusters (Figure 10). As opposed to the bioregionalization with all taxa, details in the depth zones were lost and predictions became more homogeneous. For example, clusters 2 and 3 in Figure 9 became part of cluster 1 in Figure 12. Cluster 7 in Figure 9 also disappears in Figure 10. The abyssal areas were less predicted when using a threshold of occurrences.

Genus-based bioregionalization for all taxa produced 12 clusters (Figure 11). At this level, there was a larger spatial coverage than at family-based clusters. Ocean basins (clusters 1 and 5) were well defined as well as clusters representing the bathyal zone (clusters 2, 3). At this level, we found more detail in latitudes below 50°S with four clusters (cluster 7, 8, 9, 12).

Genus-based bioregionalization for taxa with more than 30 occurrences produced 8 clusters (Figure 12). Very deep areas were not predicted and, in general, predictions were narrower in their depth zones.

Both species-based bioregionalizations produced 8 clusters (Figure 13 and 14). The spatial predictions are very similar to those at the genus-level, although more conservative when including only species with more than 30 occurrences. With such threshold, shallower and upper bathyal depths are predicted, reflecting the more intense sampling typical of these depth zones.

Overall, three main observations can be derived from these maps. First, the spatial coverage of the prediction increases when all taxa are included regardless of the number of occurrences, because it reflects a wider diversity of the input data. However, it is not possible to reliably evaluate models for taxa with less than 30 occurrences. Conversely, bioregionalizations with an occurrence threshold are more conservative, partially representing the diversity of the area. We can notice that the deepest areas of the Indian Ocean are less covered when taxa with less than 30 occurrences are excluded. By investigating bioregion composition and indicator taxa it is possible to understand if these resulting bioregions can be used as proxies for habitats. Second, with increasing resolution in taxonomic level, predictions are reduced spatially, as expected, as there is less information at each level. Third, these maps suggest that the SIOFA has a greater diversity of bioregions than observed in Figure 7-8, which is insightful for the ultimate objective of the identification of key areas for conservation.

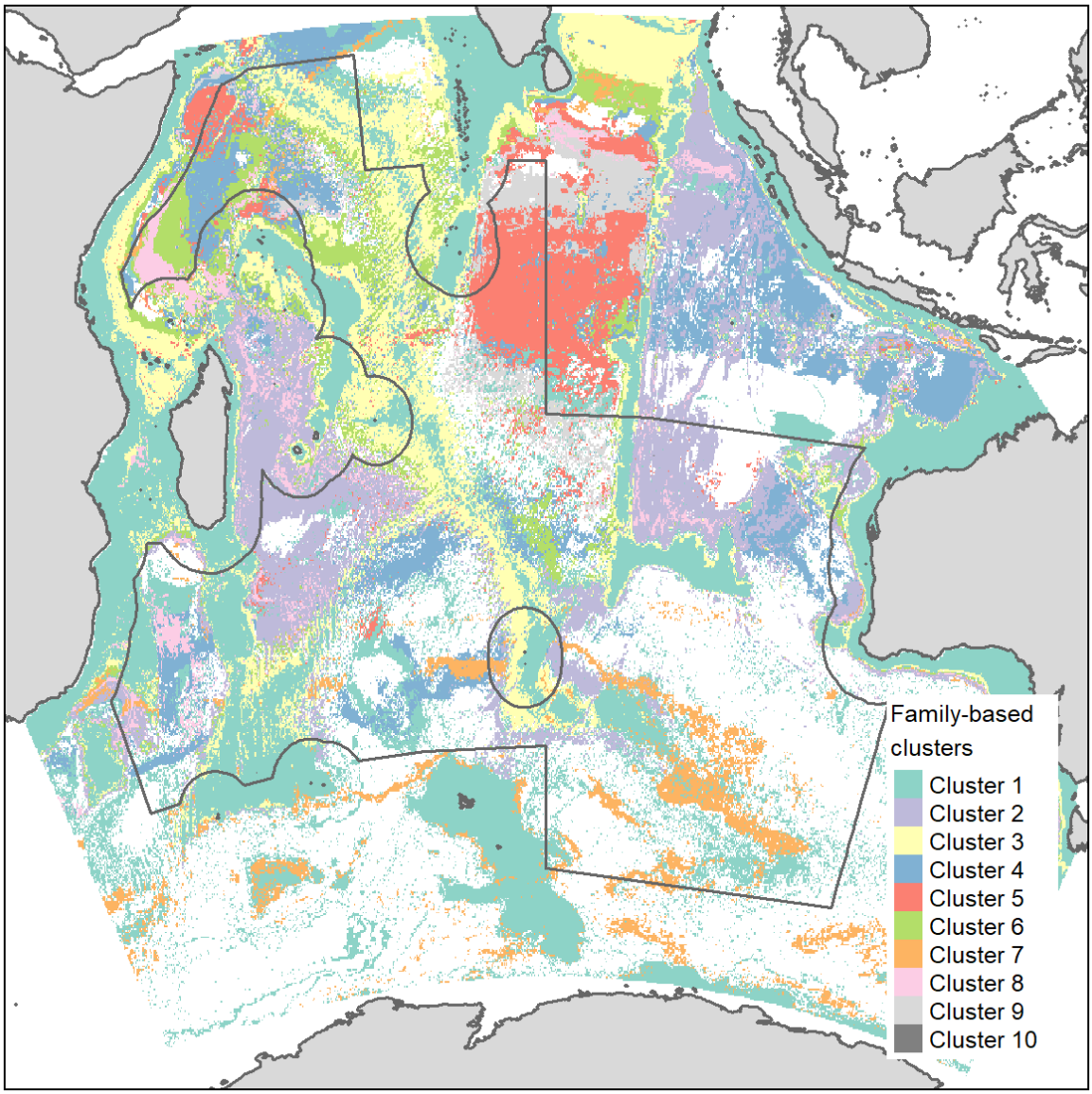

Figure 9. Family-level bioregionalization of VME indicator taxa in the Southern Indian Ocean based on all species. There are 10 bioregions depicted in different colours. White areas indicate no prediction.

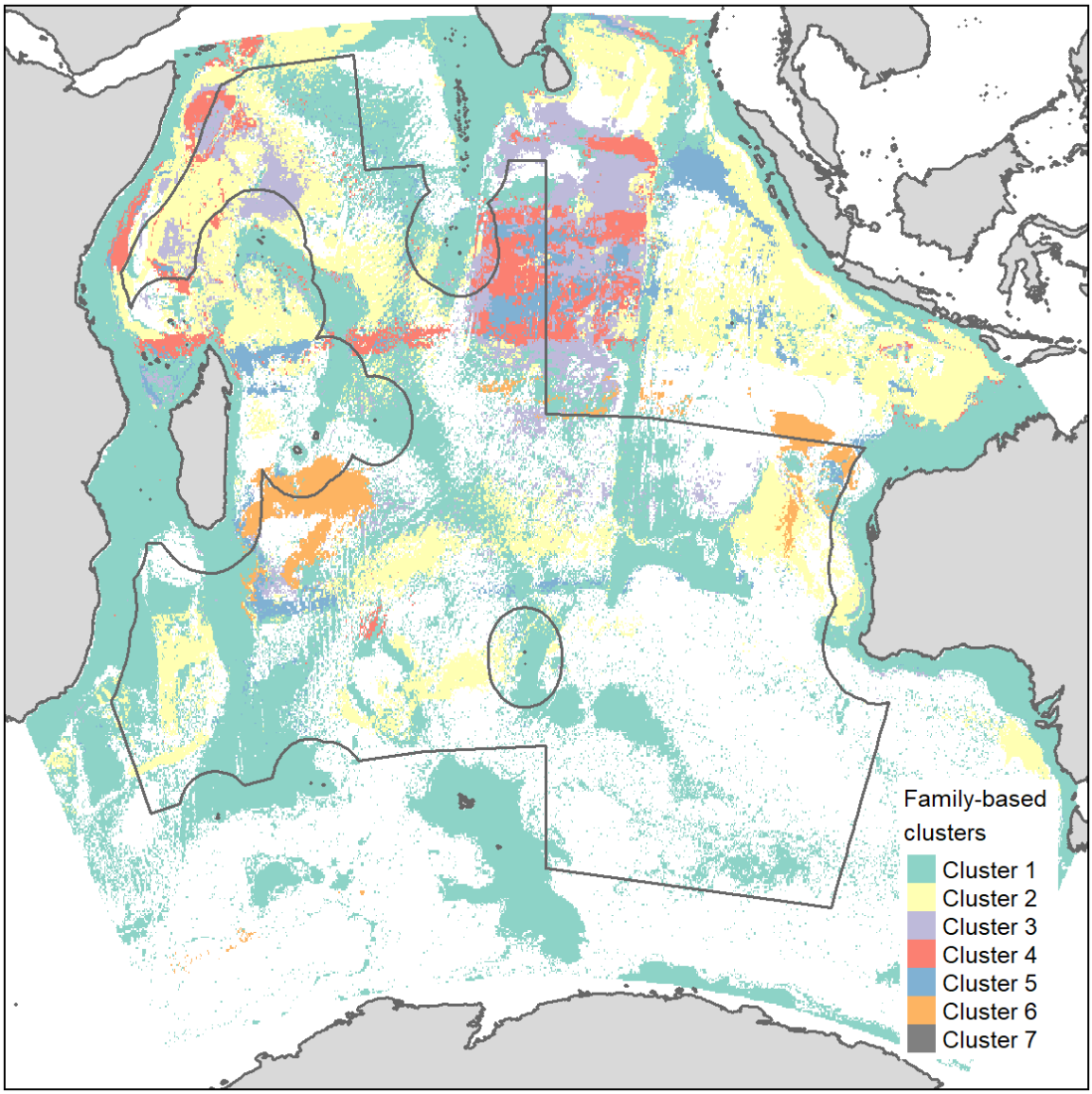

Figure 10. Family-level bioregionalization of VME indicator taxa in the Southern Indian Ocean based on taxa with more than 30 occurrences. There are 7 bioregions depicted in different colours. White areas indicate no prediction.

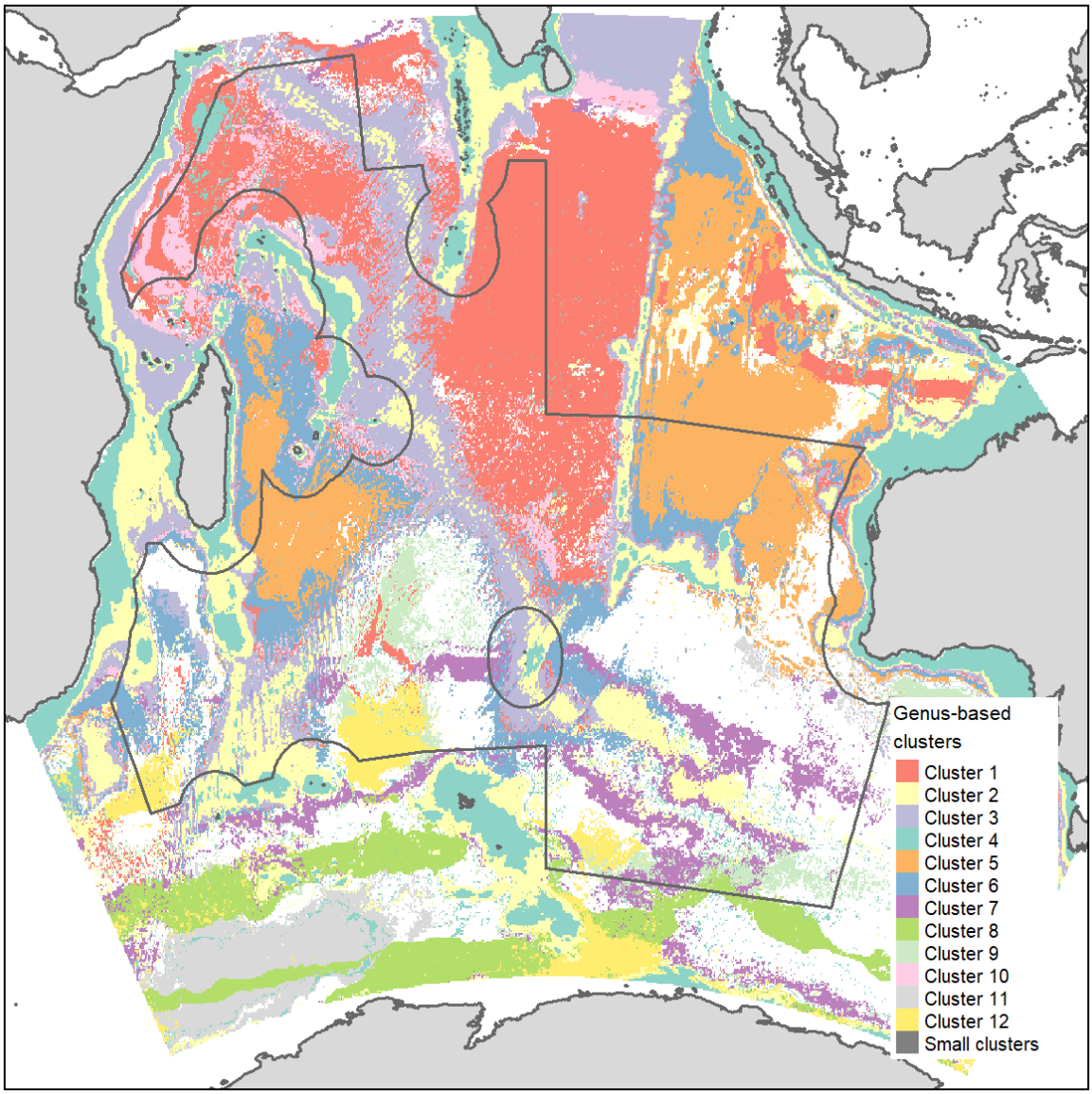

Figure 11. Genus-level bioregionalization of VME indicator taxa in the Southern Indian Ocean based on all species. There are 12 bioregions depicted in different colours. White areas indicate no prediction.

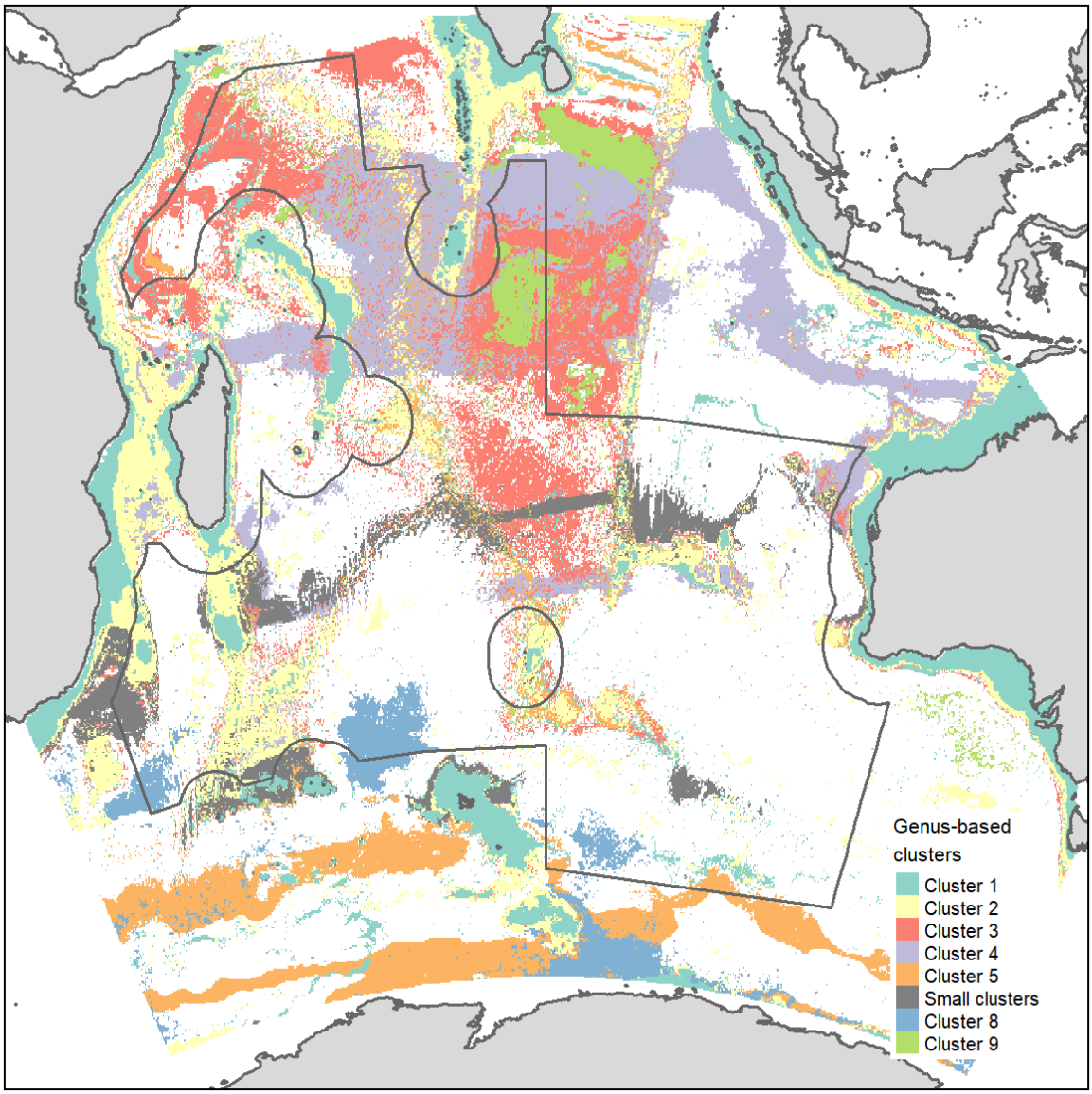

Figure 12. Genus-level bioregionalization of VME indicator taxa in the Southern Indian Ocean based on taxa with more than 30 occurrences. There are 7 bioregions depicted in different colours. White areas indicate no prediction.

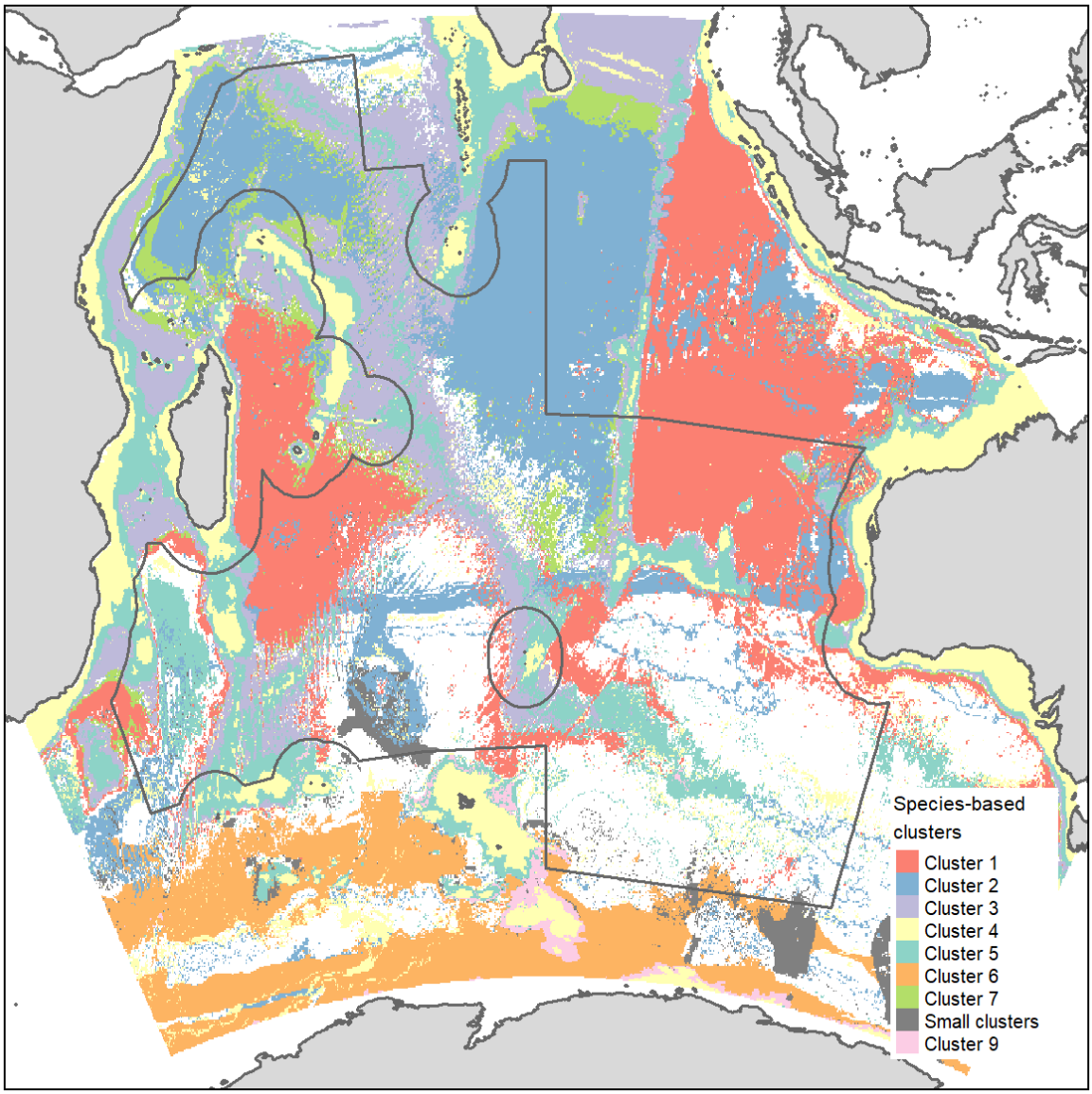

Figure 13. Species-level bioregionalization of VME indicator taxa in the Southern Indian Ocean based on all species. There are 8 bioregions depicted in different colours. White areas indicate no prediction.

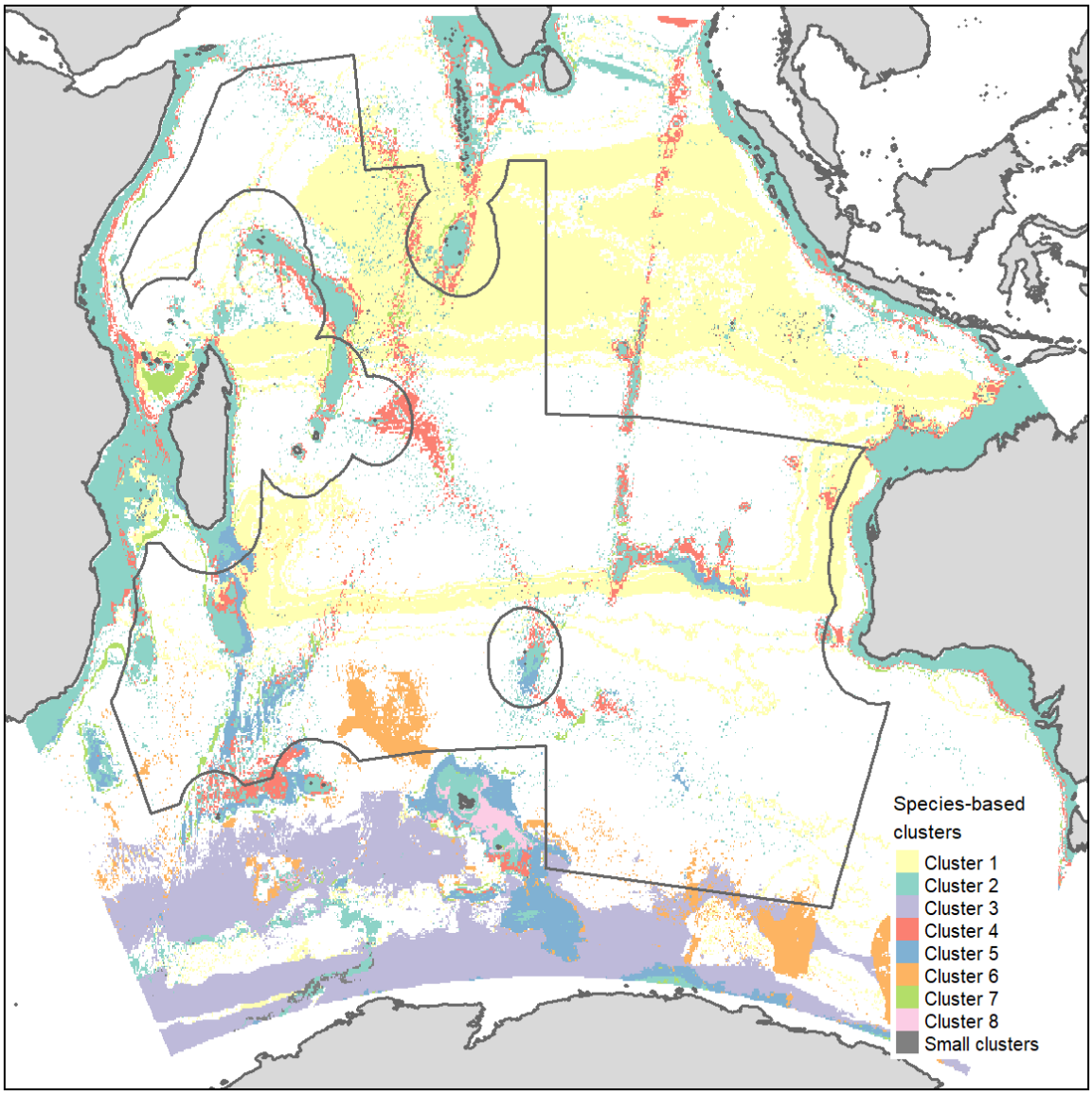

Figure 14. Species-level bioregionalization of VME indicator taxa in the Southern Indian Ocean based on species with more than 30 occurrences. There are 8 bioregions depicted in different colours. White areas indicate no prediction.

4. Discussion

Biodiversity knowledge of VME indicator taxa in the Southern Indian Ocean is both limited and spatially aggregated. The taxonomic resolution of biological data is at generic levels for areas of interest, specifically in SIOFA's management area, which reduces the dataset for comprehensive biogeographical analyses. Indeed, the majority of resolved data at species level is available from territorial waters. Despite this data-poor situation, we were able to detect three main and eight nested biogeographical regions based on observed occurrence records. These results have two main implications. First, SIOFA's management area hosts a unique biogeographical region, which suggests a decisive responsibility for SIOFA to preserve deep-sea biodiversity. Second, data shortfalls make it impossible to identify nested biogeographical regions based on observed samples and map them at a resolution relevant for conservation decisions. Thus, conservation decisions will rely on the predictive modelling approaches presented here to deal with the data-poor situation of the Southern Indian Ocean. Together, the resulting predictive maps and associated uncertainty of the models' predictions are complementary in the information they convey. Therefore, despite the poor quantity and quality of underlying data, these different approaches will be instrumental to inform and guide the future conservation planning procedure of this consultancy.

Due to data constraints, we acknowledge that the resolution of the analyses of our first predictive bioregionalization approach is too coarse (1° latitude-longitude). However, the resultant bioregions are supported by and refine existing global biogeographic classifications. The marine realms (based on endemicity; Costello et al., 2017) resemble the distribution of our observed biogeographical regions, although the representation of the central Indian realm is significantly overrepresented according to our predictions. On the other hand, the Global Open Oceans and Deep Seabed (GOODS; Watling et al., 2013) classification depicts the Indian Ocean as one homogeneous lower bathyal (800 – 3,500 m) province and one homogenous abyssal (> 3,500 m) province. This classification, which is based on oceanographic proxies and refined by expert knowledge, does not take into account potential dispersal restrictions due to geographical distance and, according to authors; it should especially be refined in the bathyal depth zone. Although, our first level bioregionalization supports the GOODS classification, the nested biogeographical regions at the second level suggest greater spatial divergences. Thus, our first modelling approach is an improvement over these classifications. Nevertheless, because of its resolution, these first maps may be of limited use for applied management decisions at a scale relevant for fisheries, that is, at seamount scale.

The taxa distribution models improved the first modelling approach two-fold. First, we were able to model the individual environmental requirements of each taxon. Second, because models are individually undertaken for each taxon, the original resolution of the environmental variables remains unchanged. Consequently, the spatial resolution of the output maps is at a finer level increasing their value for management applications. Here, previous works also support our results. The bioregions obtained at the family level resemble the preliminary predicted distributions of statistical bioregions for the SIOFA area based on OBIS data and The ABNJ Deep Seas Project[9], which was based on records at suborder taxonomic level. In addition, our bioregions obtained at species taxonomic level refine both schemes and enlarge our biogeographical knowledge of deep-sea biodiversity in the Indian

Ocean. However, it is worth noting that these maps reflect bioregions based on the modelled distributions of taxa and, consequently, do not take into account aspects such as dispersal limitation (e.g., note the ocean-wide distribution of Cluster 2 on the species-based map). Thus, some species and bioregions are over predicted in areas where geographic distance or local conditions impose a barrier to dispersal. Hence, observed bioregions and expert-opinion are important to inform modelled maps.

Altogether, these two first approaches are complementary in their relative strengths and weaknesses: the first map is based on observed distributions, and thus accounts for species dispersal limitations, but is at a coarse resolution, whereas the second map is based on predicted distributions, and thus allows a very fine resolution, but does not account for dispersal limitations. Consequently, in the future developments we will tentatively investigate novel methods to couple both maps, in order to produce a final map at a fine resolution all-the-while accounting for major limits to species dispersal in the SIOFA area.

As a third modelling approach, we aim to develop Regions of Common Profile (RCPs, Foster et al., 2013) statistical bioregions, the state-of-the-art modelling technique for bioregionalization (Hill et al., 2020). We hope to refine the bioregionalization predictions of the region distribution models and the taxa distribution models with the outputs of this "analyse simultaneously" modelling approach that considers simultaneously biological and environmental interactions for species distributions. Currently, we are undertaking analyses in collaboration with Dr Piers Dunstan and Dr Skipton Woolley, who are part of team who developed these models. These models are computationally expensive and require days to run. We will present the outputs of these models in March 2022.

To date, work is in progress to refine the modelling approaches presented here. The model outputs are preliminary and should not be taken as the final product until the end of the consultancy. This report provides the opportunity to follow the work in progress and keep up to date with the methodology used in the analyses.

References

Assis, J., Tyberghein, L., Bosch, S., Verbruggen, H., Serrão, E.A. and De Clerck, O., 2018. Bio-ORACLE v2. 0: Extending marine data layers for bioclimatic modelling. *Global Ecology and Biogeography*, *27*(3), pp.277-284.

Bastian, M., Heymann, S. and Jacomy, M., 2009, March. Gephi: an open-source software for exploring and manipulating networks. In *Third international AAAI conference on weblogs and social media*.

Bellard, C., Leroy, B., Thuiller, W., Rysman, J.-F. J.-F., and Courchamp, F., 2016. Major drivers of invasion risks throughout the world. Ecosphere, 7(3), e01241. https://doi.org/10.1002/ecs2.1241

Bloomfield, N.J., Knerr, N. and Encinas-Viso, F., 2018. A comparison of network and clustering methods to detect biogeographical regions. *Ecography*, *41*(1), pp.1-10.

Cairns, S.D., 2007. Deep-water corals: an overview with special reference to diversity and distribution of deep-water scleractinian corals. *Bulletin of marine Science*, *81*(3), pp.311-322.

Cairns, S.D., 2021. Cold-water corals. Online Appendix. Phylogenetic list of the valid Recent azooxanthellate scleractinian species, with their Junior synonyms and depth ranges. Available at www.lophelia.org/coldwatercoralsbook

Chiu, C. H., Wang, Y. T., Walther, B. A., and Chao, A. 2014. An improved nonparametric lower bound of species richness via a modified good–turing frequency formula. *Biometrics*, *70*(3), 671-682.

Costello, M.J., Tsai, P., Wong, P.S., Cheung, A.K.L., Basher, Z. and Chaudhary, C., 2017. Marine biogeographic realms and species endemicity. *Nature communications*, *8*(1), pp.1-10.

Davies, A.J. and Guinotte, J.M., 2011. Global habitat suitability for framework-forming cold-water corals. *PloS one*, *6*(4), p.e18483.

Ebach, M.C. and Parenti, L.R., 2015. The dichotomy of the modern bioregionalization revival. *Journal of Biogeography*, *42*(10), pp.1801-1808.

Edler, D., Guedes, T., Zizka, A., Rosvall, M., and Antonelli, A. 2017. Infomap bioregions: interactive mapping of biogeographical regions from species distributions. *Systematic biology*, *66*(2), 197-204.

FAO, 2009. *International Guidelines for the Management of Deep-Sea Fisheries in the High-Seas.* Food and Agriculture Organisation of the United Nations, Rome.

Foster, S.D., Givens, G.H., Dornan, G.J., Dunstan, P.K. and Darnell, R., 2013. Modelling biological regions from multi-species and environmental data. *Environmetrics*, *24*(7), pp.489-499.

GEBCO Compilation Group, 2021. GEBCO 2021 Grid (doi:10.5285/c6612cbe-50b3-0cff-e053-6c86abc09f8f)

Gotelli, N. J. and Colwell, R. K. 2011. Estimating species richness. *Biological diversity: frontiers in measurement and assessment*, *12*(39-54), 35.

Jacomy, M., Venturini, T., Heymann, S. and Bastian, M., 2014. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PloS one*, *9*(6), p.e98679.

Kocsis, Á.T., Reddin, C.J. and Kiessling, W., 2018. The stability of coastal benthic biogeography over the last 10 million years. *Global Ecology and Biogeography*, *27*(9), pp.1106-1120.

Kull, M. and Flach, P. 2022. prg: Creates the Precision-Recall-Gain curve and calculates the area under the curve. R package version 0.5.1.

Leroy, B., 2021. biogeonetworks: Biogeographical Network Manipulation And Analysis. R package version 0.1.2.

Leroy, B., Bellard, C., Dubos, N., Colliot, A., Vasseur, M., Courtial, C., Bakkenes, M., Canard, A., and Ysnel, F., 2014. Forecasted climate and land use changes, and protected areas: the contrasting case of spiders. *Diversity and Distributions*, *20*, 686–697.

Leroy, B., Delsol, R., Hugueny, B., Meynard, C. N., Barhoumi, C., Barbet-Massin, M., and Bellard, C., 2018. Without quality presence-absence data, discrimination metrics such as TSS can be misleading measures of model performance. *Journal of Biogeography*, *45*(9), 1994–2002. https://doi.org/10.1111/jbi.13402

Leroy, B., Dias, M. S., Giraud, E., Hugueny, B., Jézéquel, C., Leprieur, F., Oberdorff, T., and Tedesco, P. A., 2019. Global biogeographical regions of freshwater fish species. *Journal of Biogeography*, *July*, 2407–2419.

Leroy, B., Gallon, R., Feunteun, E., Robuchon, M., and Ysnel, F. 2017. Cross-taxon congruence in the rarity of subtidal rocky marine assemblages: No taxonomic shortcut for conservation monitoring. *Ecological Indicators*, *77*, 239-249.

Lomolino, M. V, Riddle, B. R., and Whittaker, R. J., 2016. *Biogeography 5th edition* (5th ed.). Sinauer Associates, Oxford University Press.

Morato, T., González-Irusta, J.M., Dominguez-Carrió, C., Wei, C.L., Davies, A., Sweetman, A.K., Taranto, G.H., Beazley, L., García-Alegre, A., Grehan, A. and Laffargue, P., 2020. Climate-induced changes in the suitable habitat of cold-water corals and commercially

important deep-sea fishes in the North Atlantic. *Global change biology*, *26*(4), pp.2181-2202.

Ramiro-Sánchez, B., Martin, A. and Leroy, B. 2021. SIOFA Vulnerable Marine Ecosystem Mapping. *3rd Meeting of the Protected Areas and Ecosystems Working Group (PAEWG3), 1-4 March 2021*. Available at http://www.apsoi.org/meetings/paewg3

Reiss, H., Birchenough, S., Borja, A., Buhl-Mortensen, L., Craeymeersch, J., Dannheim, J., ... and Degraer, S. 2015. Benthos distribution modelling and its relevance for marine ecosystem management. *ICES Journal of Marine Science*, *72*(2), 297-315.

Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J.J., Schröder, B., Thuiller, W. and Warton, D.I., 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*(8), pp.913-929.

Rojas, A., Patarroyo, P., Mao, L., Bengtson, P. and Kowalewski, M., 2017. Global biogeography of Albian ammonoids: a network-based approach. *Geology*, *45*(7), pp.659-662.

Ross, R. E. and Howell, K. L. 2013. Use of predictive habitat modelling to assess the distribution and extent of the current protection of 'listed' deep-sea habitats. *Diversity and Distributions*, *19*(4), 433-445.

Rosvall, M. and Bergstrom, C.T., 2008. Maps of random walks on complex networks reveal community structure. *Proceedings of the national academy of sciences*, *105*(4), pp.1118-1123.

Soberón, J., Jiménez, R., Golubov, J., and Koleff, P. 2007. Assessing completeness of biodiversity databases at different spatial scales. *Ecography*, *30*(1), 152-160.

Smith, A. B. 2020. enmSdm: tools for modeling niches and distributions of species. R package version 0.5.3.0.

Thuiller, W., Georges, D., Engler, R., Breiner, F., Georges, M.D. and Thuiller, C.W., 2016. Package 'biomod2'. *Species distribution modeling within an ensemble forecasting framework*.

Tyberghein, L., Verbruggen, H., Pauly, K., Troupin, C., Mineur, F. and De Clerck, O., 2012. Bio-ORACLE: a global environmental dataset for marine species distribution modelling. *Global ecology and biogeography*, *21*(2), pp.272-281.

UNGA, 2007. UNGA Resolution 61/105. Sustainable Fisheries, Including Through the 1995 Agreement for the Implementation of the Provisions of the United Nations Convention on the Law of the Sea of 10 December 1982 Relating to the Conservation and Management of Straddling Fish Stocks and Highly Migratory Fish Stocks, and Related Instruments. United Nations General Assembly (UNGA) A/RES/61/105 (2007).

Valavi, R., Elith, J., Lahoz-Monfort, J.J. and Guillera-Arroita, G., 2018. blockCV: An r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *bioRxiv*, p.357798.

Valavi, R., Shafizadeh-Moghadam, H., Matkan, A., Shakiba, A., Mirbagheri, B. and Kia, S.H., 2019. Modelling climate change effects on Zagros forests in Iran using individual and ensemble forecasting approaches. *Theoretical and Applied Climatology*, *137*(1), pp.1015-1025.

Valavi, R., Guillera-Arroita, G., Lahoz-Monfort, J. J., and Elith, J., 2021a. Predictive performance of presence-only species distribution models: a benchmark study with reproducible code. *Ecological Monographs*, *0*(0), 1–27. https://doi.org/10.1002/ecm.1486

Valavi, R., Elith, J., Lahoz-Monfort, J. J., and Guillera-Arroita, G., 2021b. Modelling species presence-only data with random forests. *Ecography*, 1–12. https://doi.org/10.1111/ecog.05615

Vilhena, D.A. and Antonelli, A., 2015. A network approach for identifying and delimiting biogeographical regions. *Nature communications*, *6*(1), pp.1-9.

Walther, B. A., and Moore, J. L. 2005. The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. *Ecography*, *28*(6), 815-829.

Watling, L., Guinotte, J., Clark, M.R. and Smith, C.R., 2013. A proposed biogeography of the deep ocean floor. *Progress in Oceanography*, *111*, pp.91-112.

Woolley, S.N., Foster, S.D., Bax, N.J., Currie, J.C., Dunn, D.C., Hansen, C., Hill, N., O'Hara, T.D., Ovaskainen, O., Sayre, R. and Vanhatalo, J.P., 2020. Bioregions in marine environments: combining biological and environmental data for management and scientific understanding. *BioScience*, *70*(1), pp.48-59.

Yesson, C., Bedford, F., Rogers, A.D. and Taylor, M.L., 2017. The global distribution of deep-water Antipatharia habitat. *Deep Sea Research Part II: Topical Studies in Oceanography*, *145*, pp.79-86.

Yesson, C., Taylor, M.L., Tittensor, D.P., Davies, A.J., Guinotte, J., Baco, A., Black, J., Hall-Spencer, J.M. and Rogers, A.D., 2012. Global habitat suitability of cold-water octocorals. *Journal of Biogeography*, *39*(7), pp.1278-1292.

Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Duarte Ritter, C., Edler, D., ... and Antonelli, A. 2019. CoordinateCleaner: Standardized cleaning of occurrence records from biological collection databases. *Methods in Ecology and Evolution*, *10*(5), 744-751.