SIOFA BIOREGIONALISATION AND VME PROJECT

BERTA RAMIRO SÁNCHEZ, BORIS LEROY

Final Report

NOVEMBER 2020 - MAY 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

Table of Contents

Summary

The aim of this consultancy is to map where indicator taxa of VMEs are known to occur, and are likely to occur, in the Southern Indian Ocean Fisheries Agreement Area, to eventually delineate biogeographical regions for indicator taxa. We aimed to achieve this through the Terms of Reference (ToR) and applied deliverables established for the consultancy. In this final report we first summarise the results obtained in the mapping of VME indicator taxa, and the associated indicators of data quality. Then, we present maps of biogeographical regions for VME indicator taxa using three complementary predictive modelling approaches and discuss the resulting biogeographical regions in detail. In the first approach, "group first, then predict", we applied the Map Equation community detection algorithm to a network of sites to detect observed bioregions. We then modelled the relationship between these bioregions and environmental predictors using an ensemble modelling approach based on multiple algorithms. This allowed to obtain predictive suitability maps for each biogeographical region. In the second approach, "predict first, then group", we first modelled the relationship between taxa and environmental predictors with random forest models to obtain predictive suitability maps for each taxon. Then, we delineated bioregions on these predictive maps using the Map Equation algorithm. In an "analyse simultaneously" approach, we obtained preliminary bioregions and their predictions in a single step. For the "group first, then predict" approach, we detected three biogeographical regions at a first hierarchical level, which broadly represented the upper and lower bathyal, the abyssal, and the Southern Ocean. At a second hierarchical level we detected eight nested biogeographical regions that displayed distinct geographical and bathymetric differences across the region. Bioregions at the first level are a reflection of the overlapping species ranges, while bioregions at the second level reflect limits in larval dispersal likely driven by the circulation of the water masses in the area. In the "predict first, then group" approach, biogeographical regions varied from six to nine groups depending on the taxonomic level. These bioregions can be interpreted as reflecting the diversity of habitats in the area. In the third approach, we obtained seven bioregions that resemble the bioregions from the first approach. This model is however preliminary, and interpretation must be exerted with caution. The SIOFA area encompassed the entire range of bioregions from the predictive approaches. These maps suggest that the SIOFA has a great diversity of bioregions, which is insightful for the ultimate objective of the identification of key areas for conservation. We discuss the interpretation of both approaches in relation to their future use in management applications.

1. Introduction

The United Nations require states and regional fisheries management organisations and agreements to implement measures to prevent significant adverse impacts on vulnerable marine ecosystems (VMEs) (UNGA, 2007). 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). Following UNGA resolutions, the Southern Indian Ocean Fisheries Agreement (SIOFA) has taken a series of steps towards the protection of VMEs. SIOFA have adopted a list of VME indicator taxa and have identified five Benthic Protected Areas (BPAs) where bottom-fishing trawling is not permitted. For SIOFA to progress meeting 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 where indicator taxa of VMEs are known to occur, and are likely to occur, in the Southern Indian Ocean Fisheries Agreement Area, to eventually delineate biogeographical regions for these indicator taxa. We aimed to achieve this through the Terms of Reference (ToR) and applied deliverables established for the consultancy (Table 1).

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

Terms of Reference	Deliverables	Deliverable deadline
ToR1: Compile, clean and verify occurrence and environmental data	1.1 Consolidated VME occurrence dataset. 1.2 Consolidated environmental variables dataset. 1.3 Maps of VME taxa occurrence.	
ToR2: Evaluate the quality of occurrence data and determine spatial scale and taxonomic resolution.	2.1 Indicators of data quality. 2.2 Maps of optimal resolutions of analysis. 2.3 Maps of predicted VME taxa occurrences.	February 2021
ToR3: Develop and compare multiple bioregionalisation schemes and approaches.	3.1 Maps of bioregionalisation based on occurrence data. 3.2 Maps of bioregionalisation based on taxa distribution models. 3.3 Maps of bioregionalisation based on generalised dissimilarity models or other appropriate modelling techniques.	February 2022
ToR4: Scientific reporting	4.1 Scientific reports to PAEWG and SC annual meetings, 30 days prior to meeting commencement date. 4.2Final report to PAEWG and SC.	May 2022

In ToR1 and ToR2, we assess the accuracy, quality, and comprehensiveness of occurrence records of VME indicator taxa in the Southern Indian Ocean to properly evaluate the uncertainties of any map that will be derived from them. In ToR2, we further engage in habitat suitability modelling to predict 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 bioregionalisation approaches that provide maps of the extent of VME-based bioregions. Biogeographical classifications, or bioregionalisations (Ebach & Parenti, 2015), permit an understanding of the distribution of biodiversity over large spatial scales in a simplification (a model) of the true biogeographical distribution of their components. Bioregionalisations 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. The resulting bioregions are generally based on taxa that share similar distributions because of similar ecological and physical preferences and of a shared history (Leroy et al., 2019; Lomolino et al., 2017; Woolley et al., 2020). These classifications are an indispensable component of the planning and implementation process of protected areas (Rice et al., 2011).

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 individually 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 ("analyse simultaneously"). Each method has its set of advantages and shortcomings that make them complementary in the information they provide. Ultimately, however, the kind of data available will often limit the choice of method (Woolley et al., 2020).

For ToR3 we provide bioregionalisation maps using the three predictive approaches. Note, however, that we modified some of the methods to follow state-of-the-art recommendations. Specifically, a "group first, then predict" modelling approach known as Generalised Dissimilarity Models was initially planned; however, it has been recently criticised for their inadequacy in identifying relevant predictors of turnover in species composition between regions (Woolley et al., 2017). Consequently, to provide the most accurate results, we decided to change the methods in the consultancy (Generalised Dissimilarity Models or other appropriate techniques; Deliverable 3.3) to Regions of Common Profile (Foster et al., 2013), which corresponds to the approach "analyse simultaneously". Likewise, we changed the modelling techniques initially proposed to model each indicator taxa in the "predict first, then group" approach of the VME mapping consultancy. Indeed, the modelling of indicator taxa requires to apply presence-only modelling techniques in a data-poor situation, and these techniques have very recently been significantly improved (Valavi, Elith, et al., 2021; Valavi, Guillera-Arroita, et al., 2021). Therefore, we implemented these new methods in order to improve the predictions of individual taxa, which in turn has significant positive impacts on the prediction of bioregions in the "predict first, then group" approach.

In this final report, we fully report on and discuss results (ToR4; Table 1). Note, however, that because we modified some of the methods to follow state-of-the-art recommendations, results of the third modelling approach (Regions of Common Profile; Foster et al., 2013) will be revaluated in the next consultancy, although we provide preliminary results. The results from the present consultancy may serve as the building blocks for a conservation planning procedure.

2. Methodology

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.

Because the VME taxa list is based on a very broad taxonomic resolution, when we downloaded data, many species considered to be shallow water were automatically downloaded, 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 50 m water depth (Cairns, 2007). First, we individually assessed the depth distribution of each

[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/

species and applied the following rule: a species was kept if 90% of its records were below 200 m 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 zooxanthellate corals 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).

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., 2012) 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 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 an overestimation of data completeness (Leroy, 2012), we defined a richness threshold of 30 below which the completeness was set to zero.

2.3 Bioregionalisation schemes

2.3.1 Bioregionalisation procedure

To identify bioregions in both the "group first, then predict" and "predict first, then group" approaches, we used a procedure based on biogeographical networks. To delineate bioregions on the basis of an occurrence dataset, we created a taxon/grid cell contingency table of taxon occurrence and transformed it into a bipartite occurrence network (Vilhena & Antonelli, 2015). In essence, a network is a representation of how taxa 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 taxa that occur in them, while edges link grid cell and taxa nodes indicating whether a taxon is present in a grid cell. The network is bipartite because each type of node, grid cell and taxon, cannot connect to another node of the same type; that is, connections can only occur between taxon and grid cell 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 taxa 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 taxa.

In order to detect biogeographical regions, we applied a community-detection algorithm to the network. Community-detection algorithms aim at grouping together nodes that belong to the same biogeographical region. As a community detection algorithm, we chose "Map Equation" (Rosvall & Bergstrom, 2008), as it has been recommended in previous biogeographical works (Bloomfield et al., 2018; Edler et al., 2017; Rojas et al., 2017; Vilhena & 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.

We described the observed bioregions at the first and second hierarchical levels based on their species composition and environmental characteristics. Indicator values were obtained as a function of species affinity ($A_i = R_i/Z$ where R_i is the range of species i in region Z; the higher the value, the more widespread the species is in the region) and fidelity ($B_i = R_i/D_i$ where D_i is the total distribution range of species i; the higher the value, the more exclusive the distribution range of the species is located in its associated region). Bioregions were described as a function of the following environmental variables: depth (m); Shannon index for depth; salinity (PPS); silicate concentration (mol m^{-3}); speed of currents (m^{-1}); dissolved molecular oxygen (mol m^{-3}); temperature (°C); surface primary productivity (g m^{-3} day^{-1}) and its range (i.e., difference between maximum and minimum annual mean values); particulate organic carbon (POC) flux (mol m^{-3}); omega aragonite; terrain ruggedness index (TRI); TRI Shannon index; percentage of shelf depth zone (0 – 200 m) covered; percentage of upper bathyal depth zone (200 – 800 m) covered; percentage of lower bathyal depth zone (800 – 3,500 m) covered; percentage of abyssal depth zone (3,500 – 6,500 m) covered. Differences

between bioregions for each environmental variable were tested using Tukey's HSD *post hoc* test (p-value < 0.05).

2.3.2 Predicted bioregions

2.3.2.1 Group first, then predict

In this first approach, we delineated biogeographical regions at the species level using the network approach, on the basis of the observed occurrence datasets. We included all taxa 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.

Once we identified the bioregions, we aimed to map their distribution into the entire study area with predictive models. We summarise the procedure as follows: (1) we modelled the relationship between VME bioregions and predictors (Table 3) using an ensemble modelling approach based on seven algorithms, in order to obtain predictive suitability maps for each bioregion, (2) we stacked these predictive suitability maps in order to identify in each pixel the bioregion most likely to occur.

To model the distributions of bioregions, 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 & Guinotte, 2011; Morato et al., 2020; Yesson et al., 2012, 2017). We also explored the inclusion of tectonic plates to account for dispersal limitations.

We fitted predictive models for the biogeographical regions for which we had at least 30 grid cells (hereafter called bioregion occurrences) (Table 4). Bioregion occurrences and environmental variables were projected to Albers Equal Area. We fitted SDMs in biomod2 v.3.4.13 (Thuiller et al., 2020) using seven algorithms: general linear models, general additive models, artificial neural networks, multivariate adaptive regression splines, generalised boosted models, factorial discriminant analysis, and random forests. We used absence data for each biogeographical region where an absence corresponds to the presence of another biogeographical region.

We selected variables for each region with a semi-automatic process using permutation-based variable importance (Bellard et al., 2016; Leroy et al., 2014) that we describe as follows. 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. The relationship between the environmental layers and the predicted probability of presence was analysed using response curves and produced using the evaluation strip method described by Elith et al. (2005). To derive the response curve to a specific variable, we explored the outputs of models when this variable varies at 1000 points across its range whilst all other variables are held constant at their mean value.

We evaluated model performance with Jaccard indices, true skill statistic (TSS) and area under the curve (AUC) (Leroy et al., 2018). We applied a block cross-validation procedure to limit over-optimistic performance evaluations due to autocorrelation between calibration and evaluation data. 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), where data were randomly split in calibration (75%) and evaluation subsets (25%). We could not apply block cross-validation for the sub-bioregions obtained at level 2 of the bioregionalisation hierarchy because they had less than 30 occurrences (Table 4). For these sub-bioregions, all occurrences were used in the models and no evaluation of model performance was carried out – hence predictions for these bioregions should be considered preliminary until further data is collected.

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 for each bioregion. In the final post-processing step, we stacked all predictive maps of bioregions and derived one map where each pixel showed the bioregion with the highest probability. Upon this map we superimposed an estimation of the uncertainty in our predictions, calculated as $1 - P_{max}$ where P_{max} is the maximum probability of occurrence across all regions in a given pixel. This index of uncertainty was scaled between 0 and 1 (0: low confidence; 1: high confidence in the models).

Table 4. Number of occurrences and selected variables used to predict the level 1 and level 2 biogeographical regions of the "group first, then predict" approach.

Cluster	Number of occurrences	Selected variables	Model evaluation
1	392	Mean Dissolved oxygen Maximum depth	Yes
2	71	Mean Dissolved oxygen	Yes
3	59	Maximum depth Mean Dissolved oxygen	Yes
1.1	87	Mean Dissolved oxygen Maximum depth	Yes
1.2	21	Mean Dissolved oxygen Maximum depth Minimum salinity	No
1.3	25	Minimum Primary Productivity Mean Dissolved oxygen Mean Temperature	No
1.5	25	Salinity range	No
1.7	11	Mean Dissolved oxygen Minimum salinity	No
2.1	23	Mean Dissolved oxygen	No
2.4	11	Mean Dissolved oxygen	No
3.1	16	Maximum depth Mean Dissolved oxygen	No

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 modelled the relationship between the occurrence of each taxon and environmental 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 network bioregionalisation procedure described 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, which is the native resolution of the environmental variables (Table 3).

All taxa (species, genus, family) with more than five records (total: 1,390; Table 2) 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 situations with low occurrence numbers (i.e., between 5 and 30 – see Valavi, Elith, et al., 2021; Valavi, Guillera-Arroita, et al., 2021). In addition, random forests are

relatively fast single models, which is an important aspect here given the large number of taxa to model. 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 prg, Smith, 2020) and the gain in Area under Precision Recall (R package prg, Kull & Flach, 2022), two metrics recommended for presence-only distribution models (Leroy et al., 2018; Valavi, Guillera-Arroita, et al., 2021). 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 until further data are collected. 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 bioregionalisation procedure at the family, genus, and species levels. To be able to apply the bioregionalisation 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.

2.3.3 Regions of Common Profile (analyse simultaneously)

As a third approach to bioregionalisation (i.e., group and predict simultaneously), we used an extension of the Regions of Common Profile (RCPs; Foster et al., 2013) approach. The RCP approach is a multivariate adaptation of a finite mixture-of-experts model (Jacobs et al., 1991), jointly considering multi-species and environmental data. It allows mapping of bioregions within which species share the same probability of being sampled at any site within that bioregion and predict these groups at unsampled sites through the relationship with the environmental data (Foster et al., 2013, 2017) in a single analysis. This is an advantage compared to the previous two-step classifications “first group, then predict” and the “predict first, then group” approaches since those do not allow to propagate uncertainty throughout the modelling process (Warton et al., 2015). Moreover, this approach allows to incorporate sampling artefacts in the analysis that otherwise may have the effect of conflating patterns in sampling artefacts with ecological patterns, leading to incorrect inferences (Foster et al., 2017).

The framework described above considers scientifically collected data for which we have information of where species are present and where species were not observed (absences, e.g., Warton et al., 2013). However, at large scales like the one considered here, data are a compilation of records from scientific surveys and ad-hoc (museum) collections; the latter being typically kept in natural history collections and in on-line repositories. These ad-hoc data are missing in information of where species were not observed (absence) and, as such, it is not possible to truly estimate the probability of occupancy for these data (Guillera-Arroita et al., 2015). Instead, these ad hoc data require special consideration, which has not been formulated into Foster et al.'s (2013) approach (Woolley et al., 2020).

A promising solution is to extend Foster et al.'s (2013) method to presence-only cases by implementing a Poisson point process and accounting for sampling imbalance in the models (Woolley et al., 2020). Here we therefore extend the RCP model to be a spatial point process (Cressie, 1993; Warton & Shepherd, 2010), where values (i.e., presence points) will follow a Poisson distribution. The resultant model generates a probability of sighting a species within a bioregion (a density based on the presence points) rather than a probability of occurrence of a species within a bioregion. To account for the source of variation in the data collection,

we estimated sampling bias for each grid cell using the target-group approach, as it has been shown to perform the best at emulating sampling effort (Barber et al., 2022). Basically, we used all records of VMEs indicators in our dataset (at any taxonomic level) and used a 2D kernel density estimation to convert single points into a continuous probability surface (see Barber et al. (2022) for methods and code). By including it in the model, it is possible to estimate the distribution of bioregions with the confounding of sampling bias much reduced (Foster et al., 2017).

When fitting a RCP model, the number of groups (i.e., RCPs) needs to be specified. The optimal number of RCPs is rarely known *a priori*, as we are trying to determine here, but can be inferred by fitting models with varying numbers of groups and using the Bayesian Information Criteria (BIC) to choose the "best" number of groups (Foster et al., 2013; Hill et al., 2017). Here we ran models with 2 to 12 RCPs with a meaningful number of environmental variables. Specifically, we described the probability of sighting a species within a bioregion as a function of linear and quadratic polynomials of the explanatory variables for each of the 2-12 RCPs, with 50 random starts. Random starts are necessary to avoid getting stuck in a local likelihood maximum (Foster et al., 2013). Specifically, we fitted a preliminary model with linear and quadratic polynomials of depth, dissolved oxygen, temperature, and salinity at seafloor, for 153 species that had more than 30 occurrences in the dataset. We restricted the model to depths shallower than 3,500 m because of the large number of grid cells for prediction. The resolution of the environmental variables was 0.083 degrees. We projected the variables to equal area as with the other two bioregionalisation approaches.

Model uncertainty was estimated via 20 Bayesian bootstrap replications, providing 95% confidence intervals (CI). Finally, we predicted the preliminary optimal RCPs with the selected number of groups. Further details of RCP models can be found in Refs Foster et al. (2013) and Foster et al. (2017).

We used the "ecomix" package in R (https://github.com/skiptoniam) to implement the RCP models.

All analyses were conducted in the R computing environment (R Development Core Team, 2021) in 4.1.2.

3. Results

3.1 Maps of VME indicator taxa and 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 to some extent.

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). 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).

Deliverables 1.1 (occurrence dataset), **1.2** (environmental layers) and **1.3** (observed taxa distribution maps) are provided to SIOFA alongside this report in electronic format.

Deliverable 2.1 (Indicators of data quality): The completeness index of our gathered distributional database, the ratio between observed richness and estimated richness, indicated that 72% of the species richness of the area is known (Table 5) – therefore 28% remains to be recorded or described. 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 in coastal areas (Figure 3).

Table 5. 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. ***Deliverable 2.1.***

	Species richness	Average estimated richness	Completeness index
Database	1,306	1,812	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

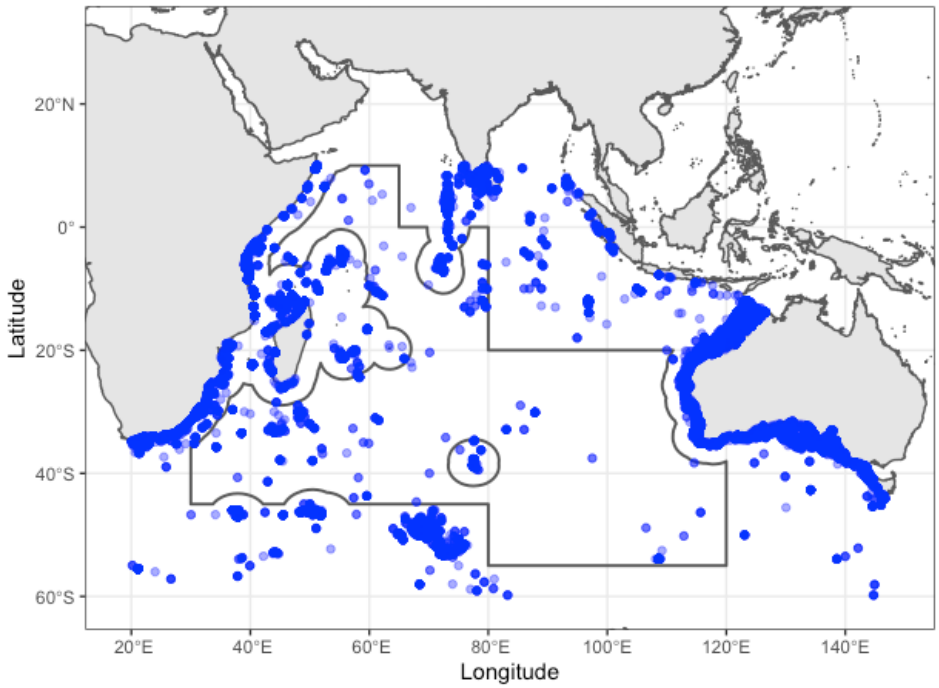

Figure 1. Distribution of taxa records over the considered area of study. Shades of blue indicate the number of occurrences, where darker shades reflect higher concentration of records. The black line represents the area managed by SIOFA.

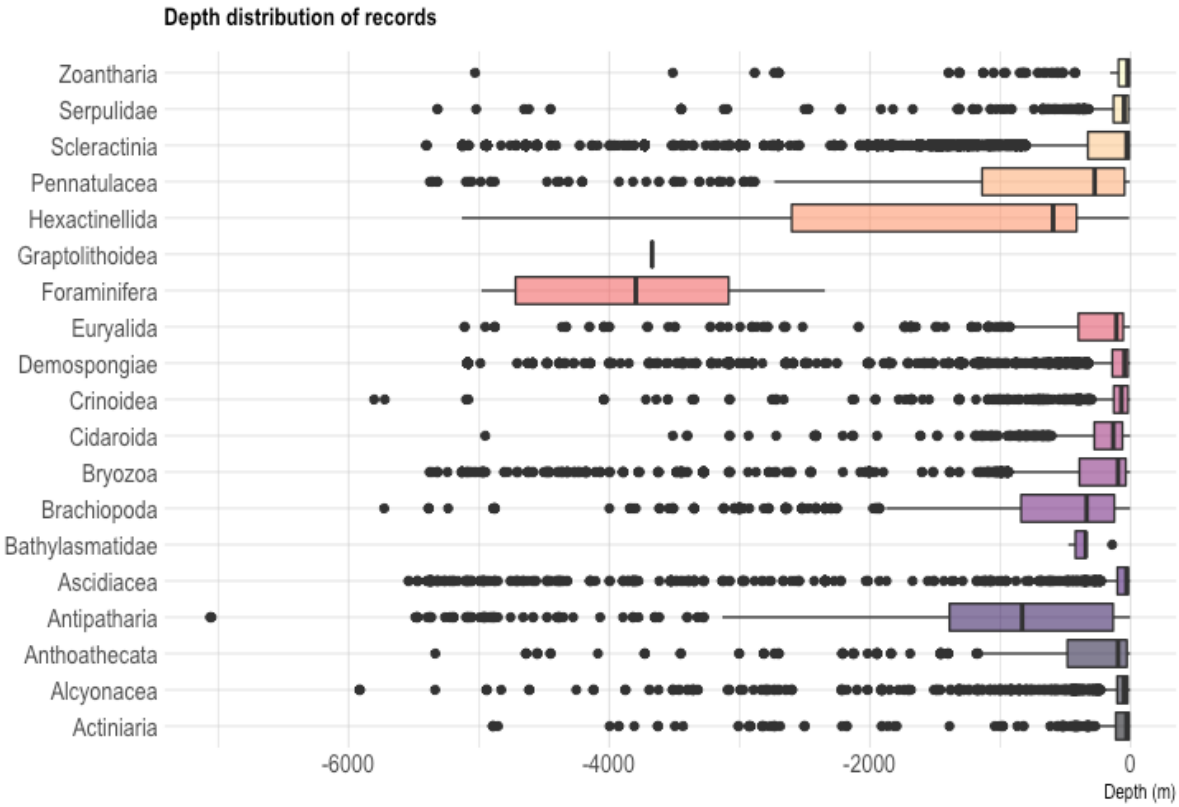

Figure 2. Depth distribution of records by VME taxa. Within each boxplot the line indicates the 50th percentile (median), the box encompasses 50% of the data, from the 25th to the 75th percentile and the dashed vertical lines extend to 1.5 times the interquartile range, with circles indicating "outliers".

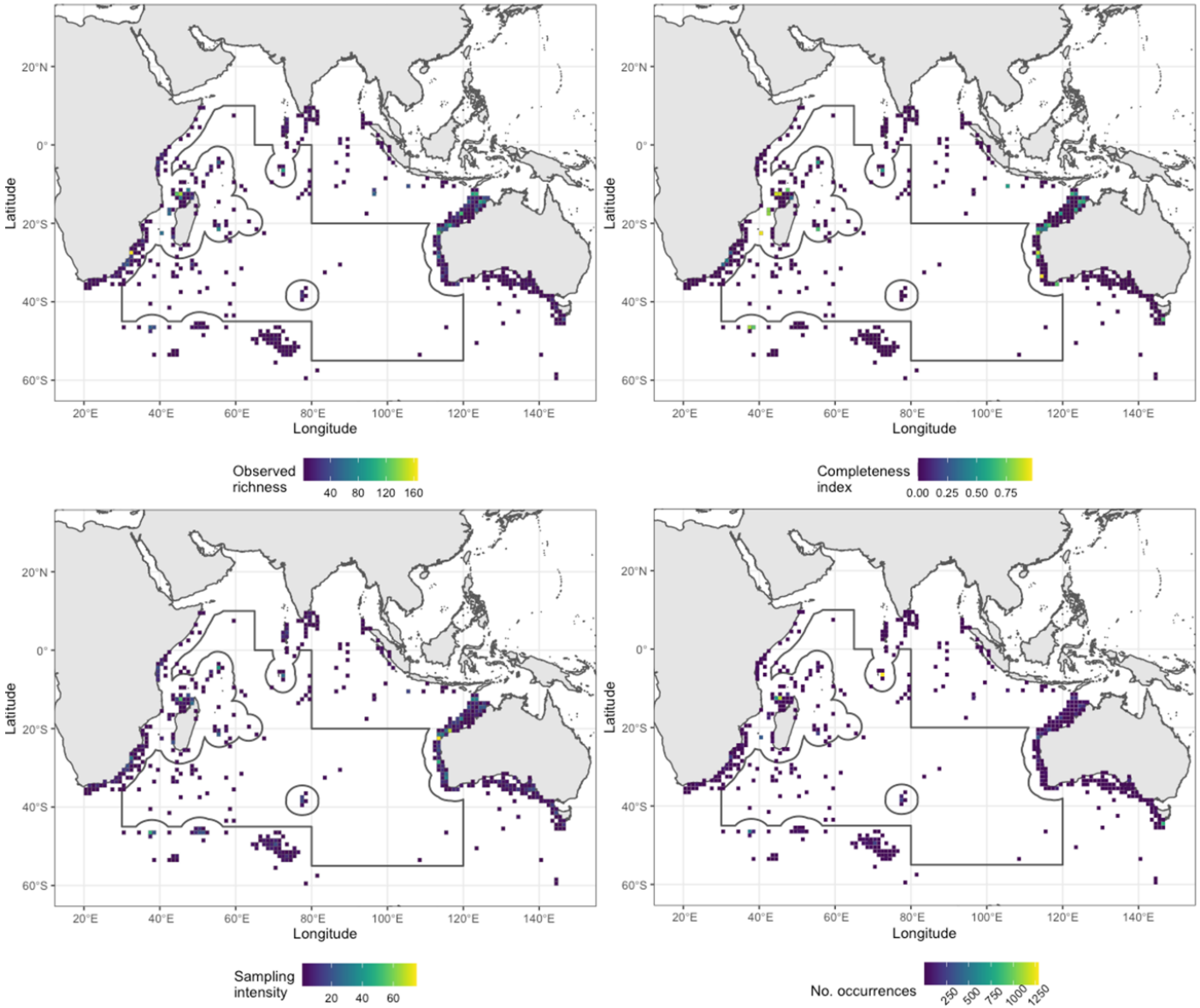

Figure 3. Indicators of data quantity and quality. From top to right: observed richness; completeness index; sampling intensity, and number of occurrences. ***Deliverable 2.1***

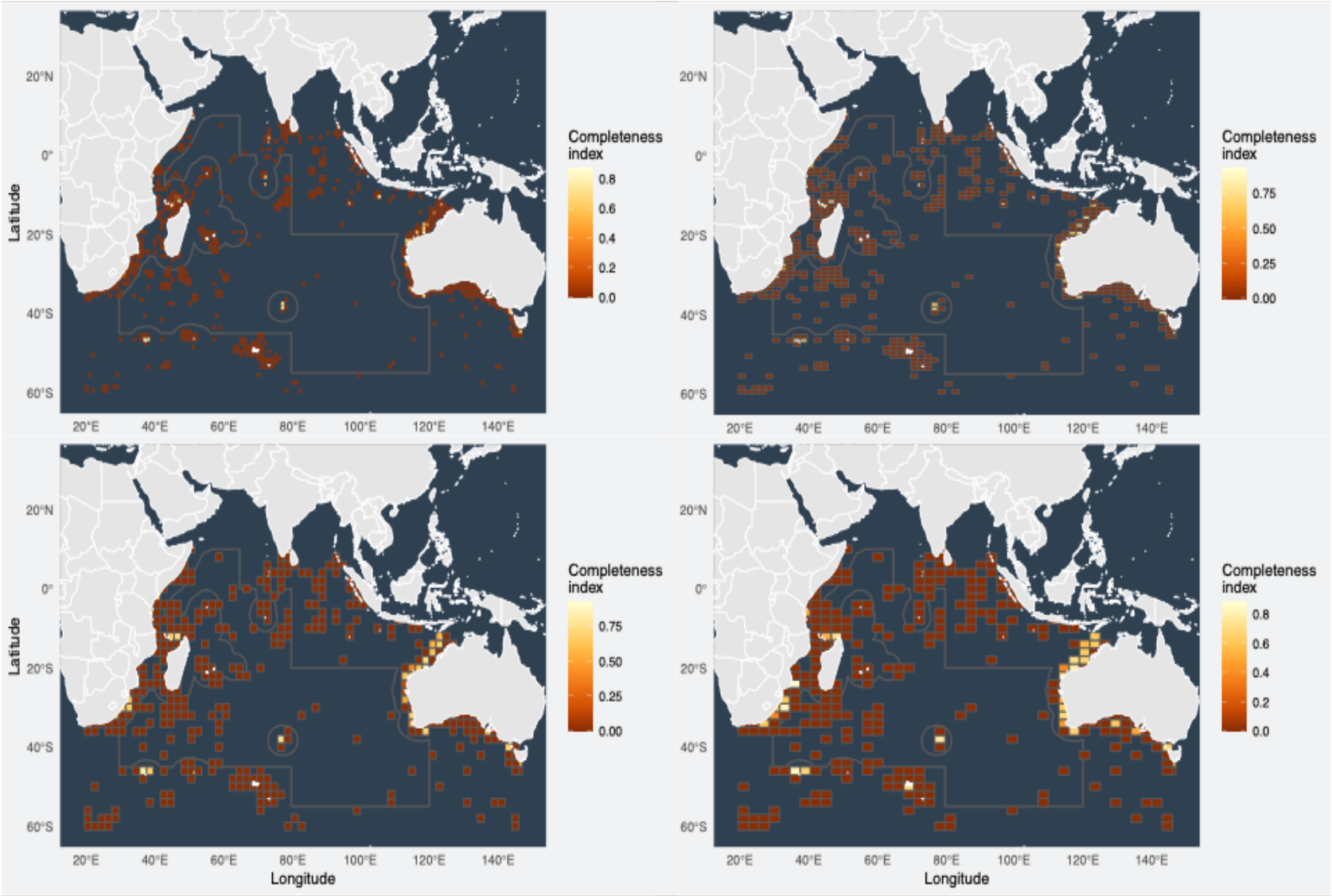

Figure 4. Indicators of data quantity and quality. From top to right: observed richness; completeness index; sampling intensity, and number of occurrences. ***Deliverable 2.2***

Deliverable 2.2 (Maps of optimal resolutions of analysis): We produced maps of completeness at varying spatial resolutions, specifically at 1° x 1°, 2° x 1°, 2° x 2°, and 3° x 2° latitude-longitude (Figure 4). We chose 1°x1° latitude-longitude as the optimal resolution for delineating assemblages. Occurrence data need to be pooled by grid cell to identify co-occurrences. This resolution was as a trade-off between data quantity and usefulness of the bioregion predictions. Whilst coarser resolutions allow to cover more space of the study area, the coarser the resolution the less useful an environmental predictor becomes as it lacks in detail. In contrast, the optimal resolution for modelling individual taxa is finer as it can be at the same resolution as the environmental variables, with the assumption that occurrence records have a precision higher than this resolution.

Deliverable 2.3 (Maps of predicted VME taxa occurrences): We were able to successfully model 1,390 taxa with more than five occurrences (734 species, 443 genera, and 213 families), and 357 taxa with more than 30 occurrences (155 species, 109 genera, and 93 families) (see Table 2). As explained in section 2.3.2.2, the number of occurrences per taxon will limit the number of variables that are used for modelling its distribution. Generally, it is accepted to include one environmental predictor per ten occurrences. Given the low number of taxa with more than 30 occurrences, we were forced to reduce the number of predictor variables. As a result, the models provide a coarse approximation of the niche of the taxa. The consequence of such limitation is that for those low-occurrence taxa, the predicted maps are likely to be an overprediction of actual suitable areas. In Figure 5, we provide an example of the predicted distribution models at family, genus, and species level for taxa with high ($\geq$ 30) and low (< 10) number of occurrences.

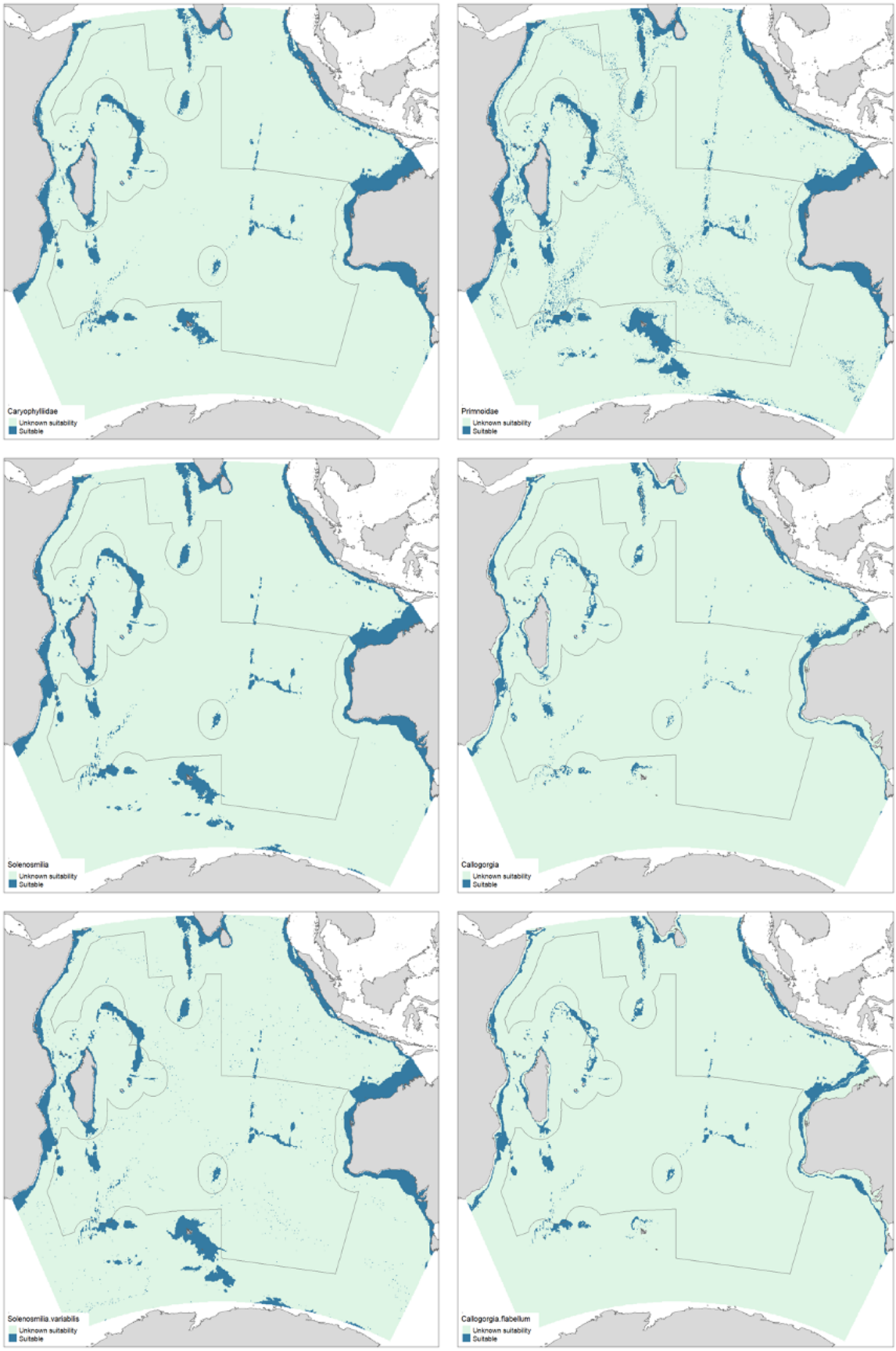

Figure 5. TDMs at family, genus, and species level for taxa with more than 30 occurrences (left column) and taxa with less than ten occurrences (right column). With fewer number of taxon occurrences, fewer number of environmental predictors are included in the model.

3.2 Bioregionalisation schemes

3.2.1 Group first, then predict **(Deliverable 3.1)**

3.2.1.1 Observed bioregions

We detected three biogeographical regions in the Southern Indian Ocean at the first hierarchical level (Figure 6A). 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]. The three biogeographical regions were geographically cohesive, they had distinct composition, and occurred in different depth zones and environmental conditions.

Cluster 1 was the largest of the three bioregions spanning all latitudes and longitudes but restricted to inshore areas (Figure 6A). The median depth range spanned from shallow waters to 2,800 m, as observed in the bathymetry and the percentage coverage of depth zones, where shelf and lower bathyal were the most representative environments (Figure 7). Terrain ruggedness, a proxy for elevated areas, showed a wide range of values corresponding to the large spanning depths of this cluster. This cluster had the highest values of salinity, temperature, and primary productivity due to its proximity to land. It had the widest variability in currents velocity with the highest values. The Map Equation algorithm assigned 1,620 out of the total 1,991 species to this cluster – in other words, these 1,620 species have the majority of their distribution located in cluster 1.

Cluster 2 was a high latitude cluster, present below 40°S below the Agulhas Retroflection Current (ARC) and the Subtropical Front (STF) (Figure 6A). The Kerguelen Plateau had the largest coverage of sites belonging to this cluster, with sites west present around Conrad Rise and along Del Caño Rise. In the eastern side, there were some sites along the Southeast Indian Ridge. This cluster had a deeper median depth of ~ 2,000 m, further observed in the percentage cover of lower bathyal depth zones, although it had some representation of upper bathyal and abyssal areas too (Figure 7). Salinity values were lower than the other main clusters. Silicate values were larger than for Cluster 1. Temperatures and surface PP were low. This Cluster had the highest values of dissolved oxygen. The Map Equation algorithm assigned 204 out of the 1,991 species to cluster 2.

Cluster 3 was the deepest of all clusters (Figure 6A), with fewer deep-sea habitats, supported by less variability in TRI (Figure 7). Indeed, the majority of sites covered the lower bathyal and abyssal areas with effectively no representatives in waters above 800 m (Figure 7). Salinity values were higher than for Cluster 2 and lower than Cluster 1. Cluster 3 had the highest silicate values of all clusters. Dissolved oxygen values spanned a wide range, although generally below 225 mol m^{-3}. Temperatures were similar to Cluster 2, with slightly higher speed of currents.
The Map Equation algorithm assigned 108 species in total to cluster 3.

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

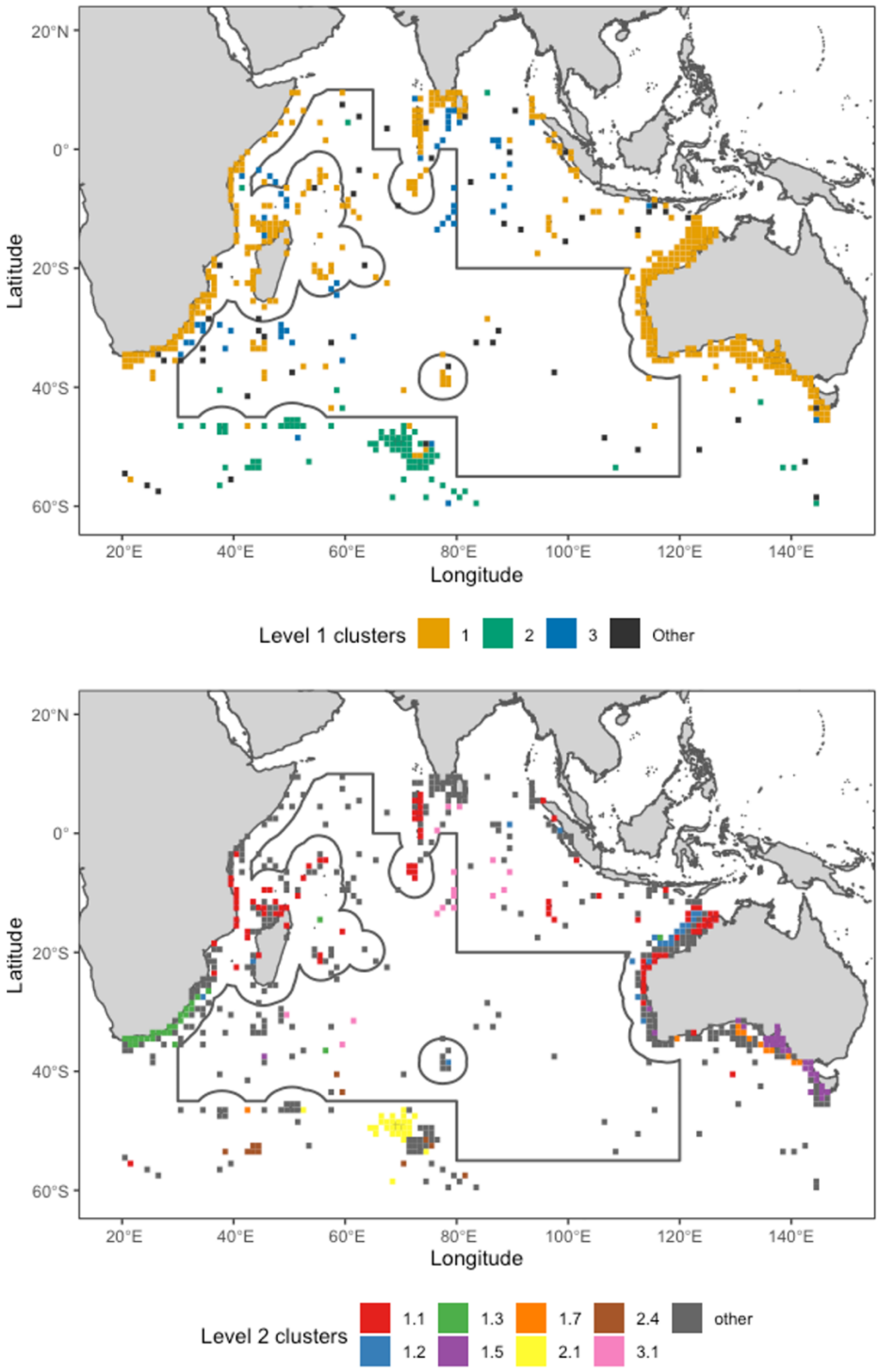

Figure 6. Biogeographical regions (clusters) of VME indicator taxa in the Southern Indian Ocean: (A) Three biogeographical regions of similar species composition detected at the first hierarchical level. (B) Finer sub-biogeographical regions detected at the second hierarchical level; black indicates no assigned subregion.

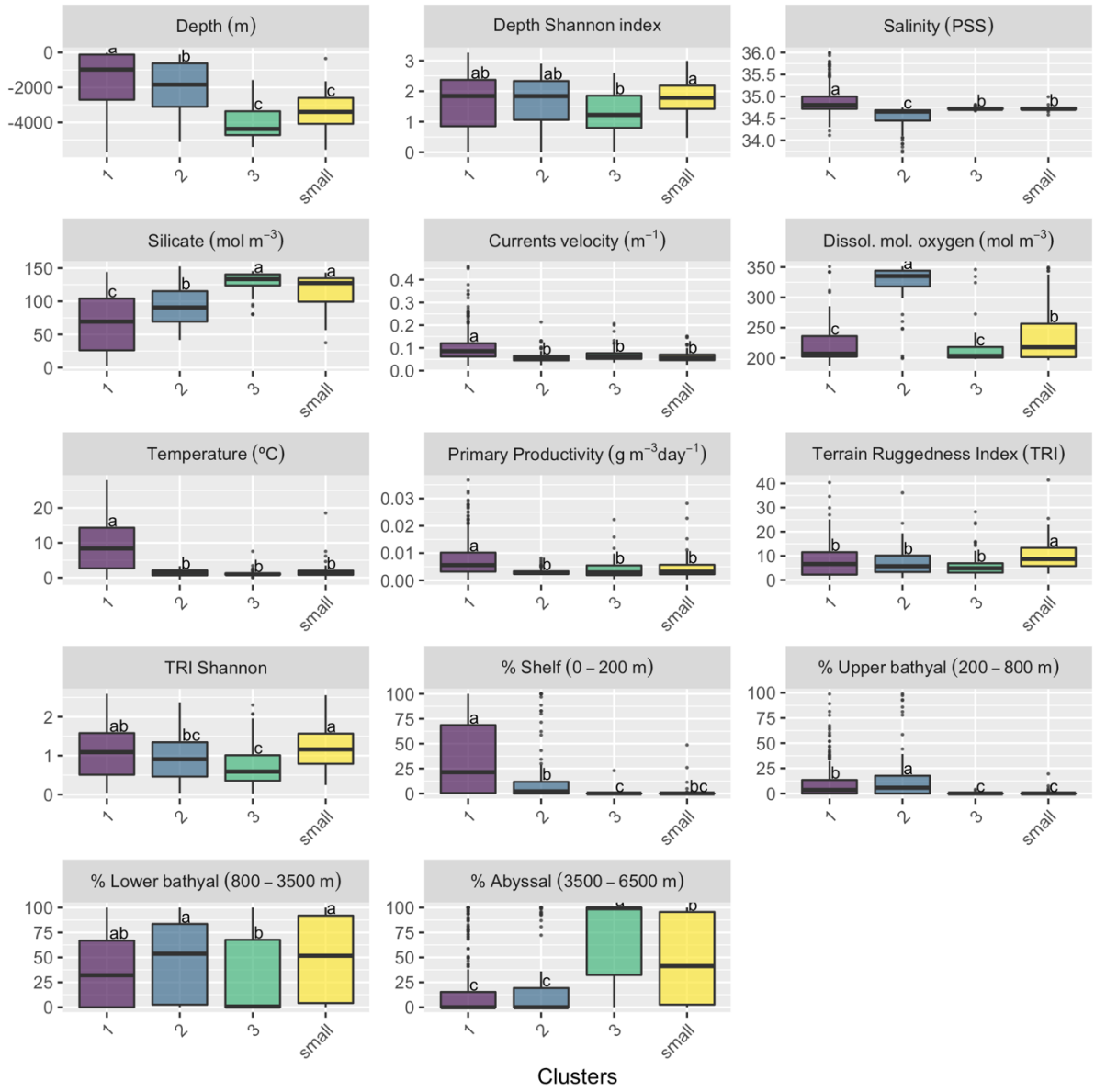

Figure 7. Environmental variables distribution for each of the three main biogeographical observed regions detected at the first hierarchical level. Small clusters (labelled 'small' in the graph) are also shown for comparison. Variables from top left to bottom right: depth (m); depth Shannon index; salinity (PPS); silicate concentration (mol m^{-3}); speed of currents (m^{-1}); dissolved molecular oxygen (mol m^{-3}); temperature (°C); surface primary productivity (g m^{-3} day^{-1}); terrain ruggedness index (TRI); TRI Shannon index; percentage of shelf depth zone (0 – 200 m) covered; percentage of upper bathyal depth zone (200 – 800 m) covered; percentage of lower bathyal depth zone (800 – 3,500 m) covered; percentage of abyssal depth zone (3,500 – 6,500 m) covered. Differences between subclusters per environmental variable are shown through pairwise comparisons (post-hoc Tukey test) with letters on the boxplots: mean values with at least one common letter are not significantly different.

The distribution of species between grid cells is illustrated in the biogeographical network (Figure 8). 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. These observations are well reflected in the high degree of endemism observed for each bioregion (65 to 95%, Table 6).

Figure 8. Biogeographical network of observed bioregions at the first hierarchical level showing the interaction between the three regions detected. Satellite nodes (in grey) are nodes not assigned to any bioregion by Map Equation.

Table 6. Number of species and endemicity of each biogeographical region at level 1 and level 2 of the hierarchy.

Level 1 Biogeographical region	Total richness	Characteristic richness	Endemic richness	% Endemic species	% Indian Ocean species
1	1,667	1,620	1,586	95.14	80.11
2	234	204	180	76.92	11.24
3	121	108	79	65.29	5.81
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

At the second level of hierarchy, we detected eight subregions (Figure 6B) within the larger biogeographical regions. The eight subregions presented a finer distribution with distinct geographic and bathymetric differences (Figure 9), and species composition (Table 6). Cluster 1 subdivided into five subregions: biogeographical region 1.1, 1.2, 1.3, 1.5 and 1.7; Cluster 2 into two subregions: 2.1 and 2.4; Cluster 3 into one subregion: 3.1 (Figure 6B).

Cluster 1.1(Figure 6B) had a median depth of 1,000 m, with spanning depths from 200 to 2,700 m (Figure 9). Within these depths, the majority of observations in this cluster covered the lower bathyal, followed by shelf observations and few upper bathyal sites. Thus, the variability of deep-sea habitats was relatively high, as observed by the depth Shannon index and the TRI (Figure 9). Temperature values were similar to the mean temperature of that of Cluster 1.
The Map Equation algorithm assigned 447 species to this cluster (Table 6), of which the most indicative species belonged to the phylum Cnidaria (Table A1, Appendix A).

Cluster 1.2 was mainly present in the northwest and western Australia, although there were some sparse sites occurring in Ninety East Ridge, west of Madagascar, south of Mozambique Channel, Saint Paul and Amsterdam (Figure 6B). Cluster 1.2 had 75% of its observations in waters above 1,500m, with a median depth of 800 m and a maximum depth of 3,000 m (Figure 9). The decomposition of percentage cover of different depth zones showed it as the predominant subcluster at upper bathyal depths. Currents speeds were slow. Together with Cluster 1.1 and Cluster 3.1, Cluster 1.2 has the lowest levels of dissolved oxygen. Surface PP over the areas of the observations was low.
The Map Equation algorithm assigned 93 species to this cluster (Table 6). Indicator species corresponded to taxa within the families Flabellidae, Actiniiidae, Schizopathidae,

Dendrobrachiidae, Leiopathidae, Pennatulidae, Micrabaciidae, Caryophylliidae, Oculiniadae and Fungiacyathidae (Table A1, Appendix A).

Cluster 1.3 (Figure 6B) had a median depth of 400-500 m, ranging from 200 to 2,700 m (Figure 9). This was reflected in the decomposition by depth zones, where most of records covered the shelf and lower bathyal area, and there were few observations in the upper bathyal (Figure 9). Silicate values were among the lowest values (around 37 mol m^{-3}). This Cluster showed the strongest velocity of currents. The PP over the area was the highest of all subclusters.
The Map Equation algorithm assigned 142 species to this cluster (Table 6). In general, species from this cluster showed very high fidelity (Table A1, Appendix A).

Cluster 1.5 represented southeast Australia and west of Tasmania (Figure 6B). This cluster had a shallow median depth of 200 m, and only some outliers were present at deeper depths (Figure 9). The shallow distribution of the cluster was reflected in the percentage coverage of the shelf depth zone, which contributed to 100% of the observations (Figure 9). Moreover, TRI were also the lowest of all subclusters. Salinity values had the largest median value of all clusters, while silicate was the lowest. Currents velocity values were in the group of faster areas. Cluster 1.5 had the highest median temperature of all subregions and the third largest surface PP.
The Map Equation algorithm assigned 45 species to this cluster (Table 6), all of them with low indicator values, where affinity and fidelity for sites was not strong (Table A1, Appendix A).

Cluster 1.7 was present mainly along south Australia (Figure 6B), deeper than cluster 1.5. There was one site on Del Caño Rise too. Cluster 1.7 had a median depth of 400 m, with a maximum of 1,800 m (Figure 9). Most of the observations covered the shelf, the lower bathyal and the upper bathyal, resulting in a high Shannon Index and the highest TRI values. Currents velocity values belonged to the faster group. Oxygen values were similar to Cluster 1.5, while the difference in depth with this subregion was reflected in a much lower median temperature of 8°C. Surface PP was the second largest among all subregions (Figure 9).
The Map Equation algorithm assigned 25 species to this cluster (Table 6). The most indicative species belonged to the families Petrosiidae, Aulocalycidae, Theonellidae, Niphatidae, Pachastrellidae, Caryophyllidae, Tethyidae, and Flabellidae (Table A1, Appendix A).

Cluster 2.1 was present in the northern area of the Kerguelen Plateau, just north of the Polar Front (Figure 6B). Cluster 2.1 had a median depth of 600 m, with depths spanning from 200 to 1,500 m (Figure 9). The observations covered roughly equally the shelf, the upper and the lower bathyal, with a varied diversity of habitats and TRI values (Figure 9). This subregion presented the lowest median salinity, the lowest speed for currents and temperature together with cluster 2.4. On the other hand, dissolved oxygen values were the highest. Surface PP was higher than for cluster 2.4.
The Map Equation algorithm assigned 62 species to this cluster (Table 6), of which the most indicative species were sea urchins (order Cidaroida), sponges (order Suberitida, Polymastiida, Poecilosclerida, Tetractinellida, Bubarida), one primnoid (order Alcyonacea),

one crinoid (order Comatulida), and two brachiopods (order Rhynchonellida and Terebratula) (Table A1, Appendix A).

Cluster 2.4 was distributed around Conrad Rise (Figure 6B). Cluster 2.4 was a deep cluster with a median depth of 2,900 m, where 50% of the observations fell between 2,100 and 4,000 m (Figure 9). Surprisingly, the depth Shannon index was higher than for subcluster 2.1. Salinity range values for Cluster 2.4 were very narrow from 34.6-34.7 PSS. This cluster had the second largest median value of silicate at ~ 120 mol m^{-3} and the highest dissolved oxygen value at 340 mol m^{-3}. On the contrary, cluster 2.4 held the lowest temperature at almost 0°C. Currents velocity were the lowest together with cluster 2.1. Surface PP was the lowest together with subcluster 3.1.
The Map Equation algorithm assigned three species to this cluster (Table 6): two brittle stars (family Gorgonocephalidae, order Euryalida) and one brachiopod (order Terebratula) (Table A1, Appendix A).

Cluster 3.1 observations were located in the Mid-Indian Basin, northern Ninetyeast Ridge, south of Sri Lanka and northeast of the Chagos Laccadive Plateau (Figure 6B). In the southern hemisphere, some observations were at the far east extent of the Southwest Indian Ridge and on the east of the Madagascar Plateau (Figure 6B). Cluster 3.1 was the deepest of all subregions with a median depth of 5,000 m (Figure 9). Indeed, all the observations were in abyssal depths (Figure 9). The diversity of deep-sea habitats (depth Shannon index) was therefore low as it was the TRI index too. The median salinity value was almost constant at 34.7 PSS. Cluster 3.1 had the largest silicate mean concentration at 140 mol m^{-3}, low temperature at almost 0°C and very low surface PP. Currents speed were low but slighter faster than Cluster 2.4.
The Map Equation algorithm assigned 16 species to this cluster (Table 6), for which the indicator values were not particularly high. Species belonged to families Cladorhizidae, Hyalonematidae, Agneziidae, Styelidae, Umbellulidae, Schizopathidae and Cerelasmidae (Table A1, Appendix A).

At level 2 of the hierarchy, the biogeographical regions detected had also distinct faunas as seen in the network (Figure 10), although patterns were not as strong as at level 1. Because this level reflected finer patterns of level 1, the number of links was similar but the distribution of these within and among regions became well defined. The endemism patterns observed for each bioregion are presented in Table 6.

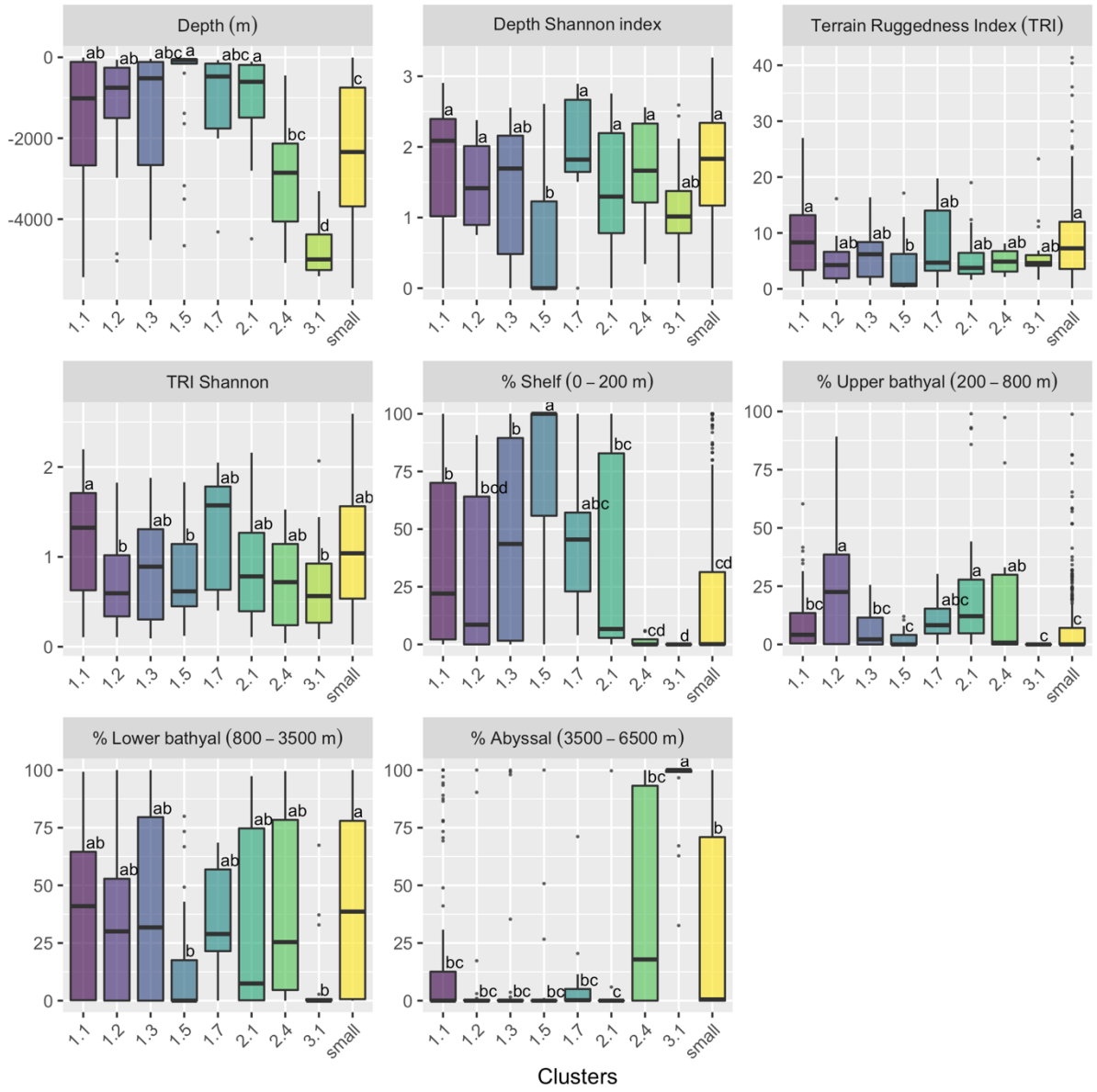

Figure 9. Environmental variables distribution for each of the eight nested biogeographical regions detected at the second hierarchical level. Small clusters (labelled 'small' in the graph) are also shown for comparison. Variables from top left to bottom right: depth (m); depth Shannon index; terrain ruggedness index (TRI); TRI Shannon index; percentage of shelf depth zone (0 – 200 m) covered; percentage of upper bathyal depth zone (200 – 800 m) covered; percentage of lower bathyal depth zone (800 – 3,500 m) covered; percentage of abyssal depth zone (3,500 – 6,500 m) covered. Differences between subclusters per environmental variable are shown through pairwise comparisons (post-hoc Tukey test) with letters on the boxplots: mean values with at least one common letter are not significantly different.

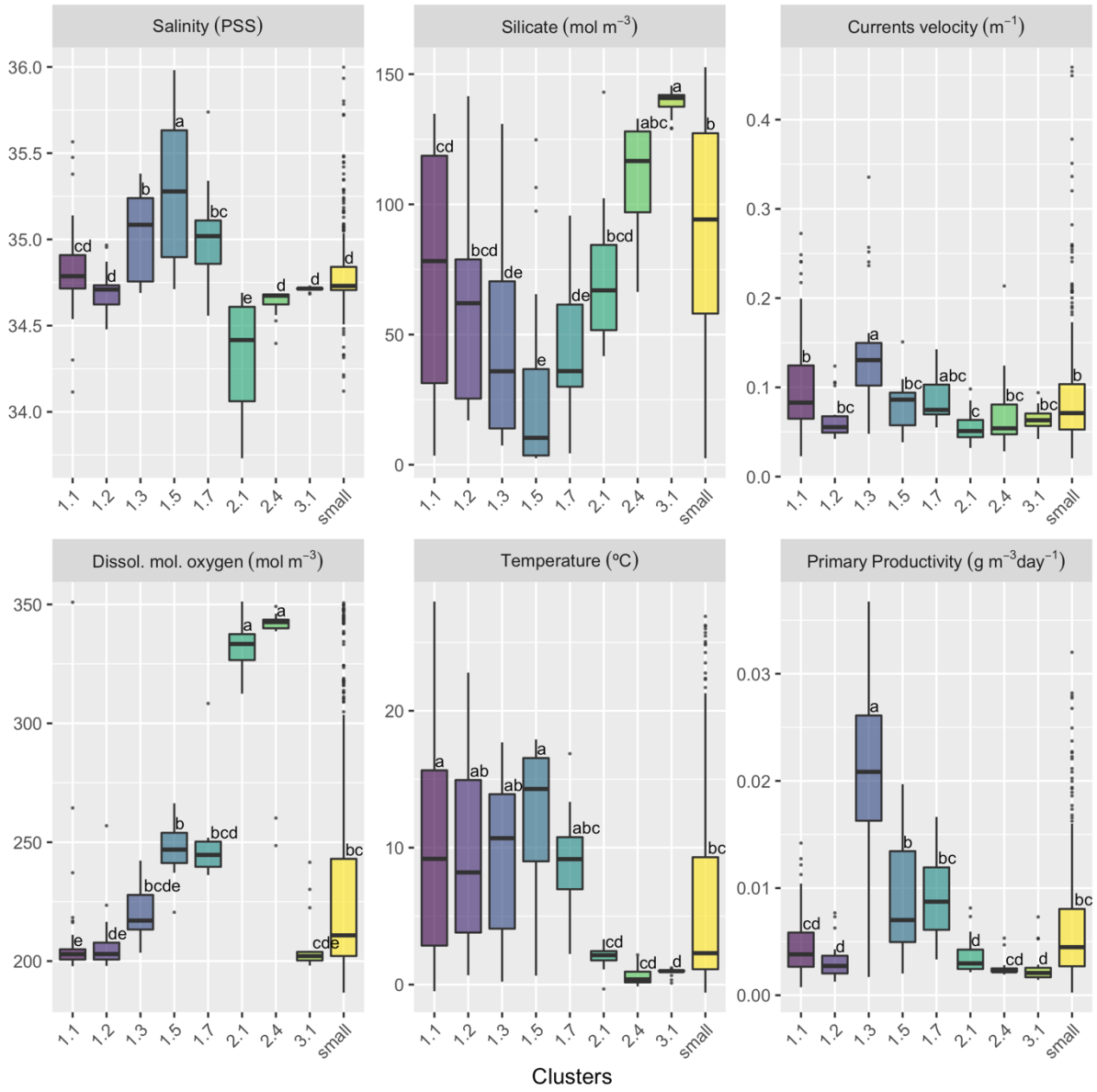

Figure 9. (Cont). Environmental variables distribution for each of the eight biogeographical regions detected at the second hierarchical level. Small clusters (labelled 'small' in the graph) are also shown for comparison. Variables from top left to bottom right: salinity (PPS); silicate concentration (mol m^{-3}); speed of currents (m-1); dissolved molecular oxygen (mol m^{-3}); temperature (°C); surface primary productivity (g m^{-3} day^{-1}). Differences between subclusters per environmental variable are shown through pairwise comparisons (post-hoc Tukey test) with letters on the boxplots: mean values with at least one common letter are not significantly different.

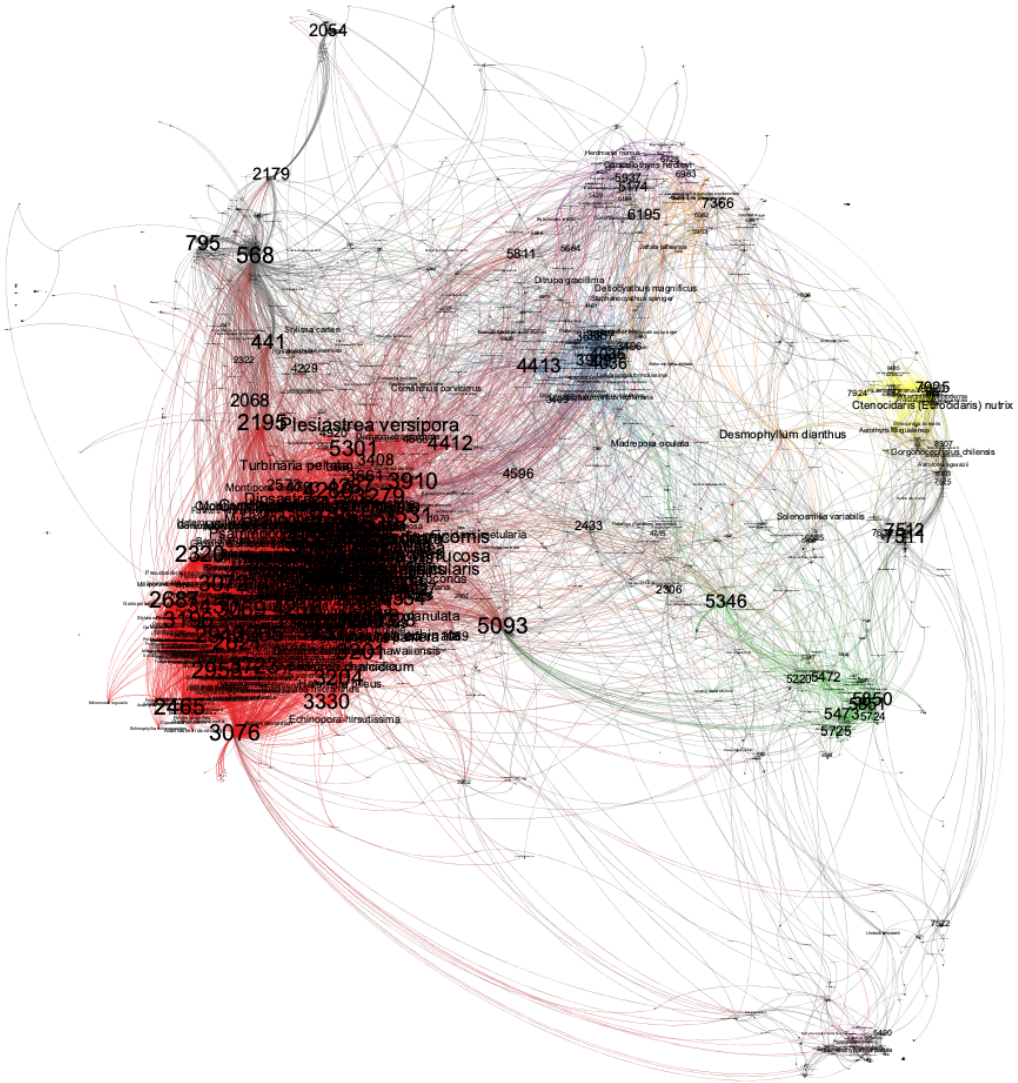

Figure 10. Biogeographical network of observed bioregions at the second hierarchical level showing the interplay of sites and species between the eight subregions detected.

3.2.1.2 Predictive modelling of the observed bioregions ("Group first, then predict" approach)

We were able to successfully model the distribution of the three bioregions (TSS and ROC > 0.7; Jaccard index > 0.6) identified at the first hierarchical level with our ensemble modelling approach (evaluation metrics in Appendix B, Figures B1-B2). Variable importance for models generating Bioregion 1 included dissolved molecular oxygen, temperature, and maximum depth (Table 7). For Bioregion 2, the variable most important was dissolved oxygen, while for Bioregion 3 was maximum depth and to a much lesser degree, dissolved molecular oxygen (Table 7). Bioregion 1 is predicted in areas where dissolved oxygen values vary from low to high and at intermediate to shallow depths (Figure 11), while bioregion 3 is predicted in areas with very high levels of dissolved oxygen (Figure 13). Bioregion 2 is predicted in very deep areas, where low values of dissolved oxygen dominate, and where surface PP remains relatively constant (Figure 12). The variable importance for all predictors considered are shown in Appendix B, Figures B5-B7.

The map in Figure 14 shows the predicted distribution of bioregions at the first hierarchical level. We added on the map the degree of uncertainty (index of confidence from 0 to 1, see methods) that we had 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 environmental conditions – thus, model predictions in these areas are extrapolations. The final predictive map of bioregions showed that the three bioregions encompassed the SIOFA area. The biogeographical regions 1 and 3 prevail from 20°N to 10°S latitude. Below 40°S latitude, the Southern Ocean bioregion 2 dominates the area. Areas of uncertainty in predictions coincide with areas of heterogeneity in biogeographical composition and with areas where there was no initial data. Notably, areas of low confidence coincide with shallower depths in the Arabian Basin, northern part of the Mid-Indian Basin, the Crozet Basin, the South Australian Basin, and the Southeast Indian Ridge.

We modelled the distribution of the eight subregions detected at the second hierarchical level, although no formal evaluation of the model performance was carried out for predictions because of the low number of occurrence points (Appendix B, Figures B2-B3). Table 7 shows the final environmental variables used for the prediction of the subregions, while response curves are shown in Figures B16-B23. Cluster 1.1 was predicted in areas with extreme values of dissolved oxygen at upper bathyal depths (Figure B16). Cluster 1.2 was predicted to occur in areas with lower levels of dissolved oxygen, intermediate depths, and at a restricted range of salinity values (Figure B17). Cluster 1.3 was predicted to occur in areas with high surface PP, across a large thermal range and low dissolved oxygen values (Figure B18). Cluster 1.5 was predicted in areas with high salinity (Figure B19). Cluster 1.7 showed a bimodal response in its predicted distribution, predicted to occur in areas with high and low salinity, but not at intermediate values (Figure B20). Cluster 2.1 and Cluster 2.4 were predicted in areas of high dissolved oxygen, although the distribution of Cluster 2.4 was less influenced by this predictor (Figure B21-B22). Cluster 1.3 was predicted to occur in very deep areas and low values of dissolved oxygen (Figure B23). Variable importance plots for each of the subregions are in Appendix B (Figures B8-B15).

The map in Figure 15 shows the predicted distribution of bioregions at the second hierarchical level. As before, we incorporated confidence levels in the predictions for consistency. Cluster 1.1 was predicted in latitudes north of 23°S on the west and 30°S on the east, across the coasts and the Ninety East Ridge, the Chagos Laccadive Plateau, the Carlsberg Ridge, the Mascarene Plateau, the northern part of the Madagascar Plateau, Mozambique Channel (Figure 15). While the northwest of Australia shows an extensive distribution, on the west the Cluster 1.1 prediction is restricted to the areas closest to land. Cluster 1.2 is present in the northwest of Australia, deeper depth zones on the west from 21° to 35°S, extending to the south of Ninety East Ridge, Broken Ridge, Mid-Indian Ridge, Madagascar Plateau, and the Mozambique Plateau. Cluster 1.3 was predicted over the Agulhas Plateau, in the southwestern part of the Mozambique Channel, a small area on the Madagascar Plateau, on the southernmost coast of Somalia, and across the entire 40°S latitude reaching Australia. It is further predicted in the southeast Australia. Cluster 1.5 is predicted only in the south of Australia, while Cluster 1.7 is predicted in the deeper south Australia extending at approximately 38°S up to the south of the Madagascar Plateau. Cluster 2.1 is predicted south of 45°S in two sections divided by Cluster 2.4 at approximately 50°S, both with circumpolar distribution.

Table 7. Number of occurrences and selected variables used to predict the biogeographical regions of level 1 and level 2 of the hierarchy.

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

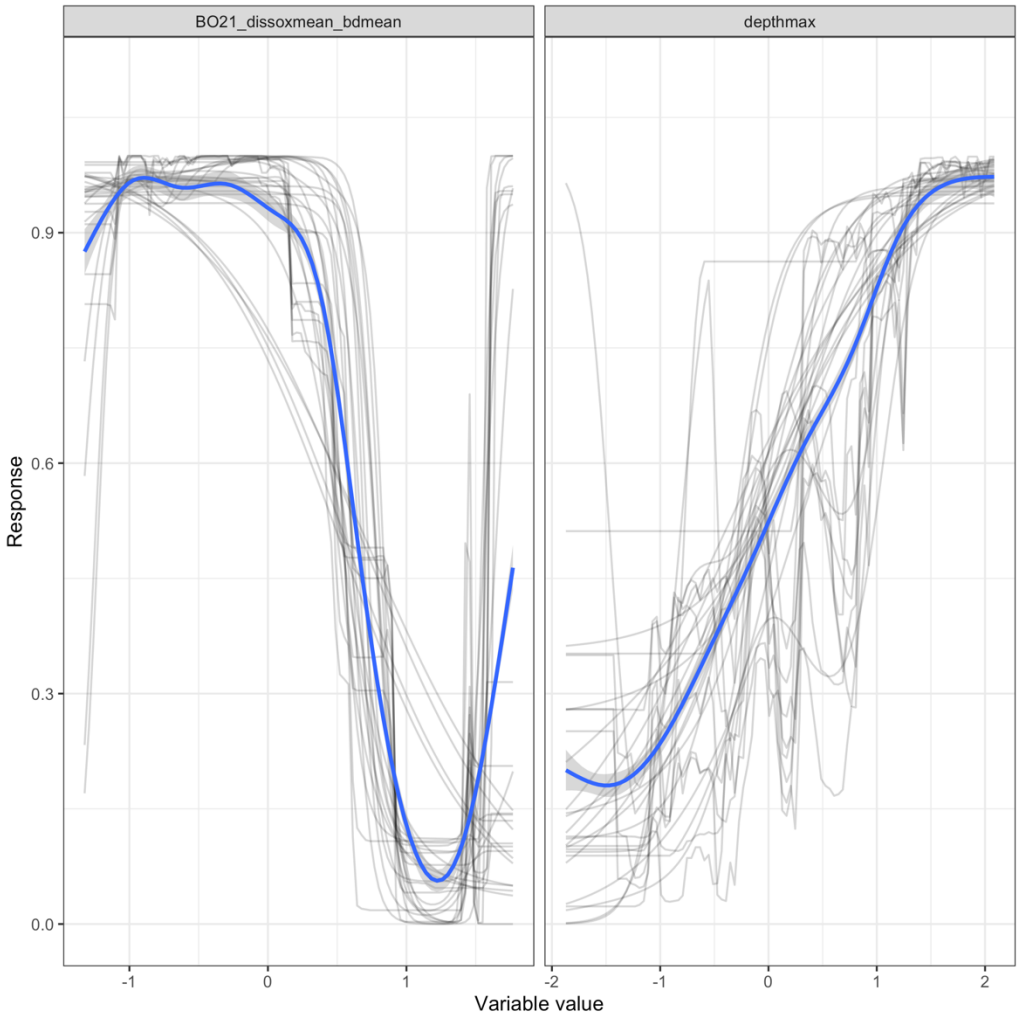

Figure 11. Response curves of the probability of presence of Cluster 1 for each predictor variable (bottom dissolved oxygen and maximum depth) based on the seven algorithms used for the prediction. The response curve is the dependence of the probability of the presence on one predictor variable after averaging out the effects of the other predictor variables in the model.

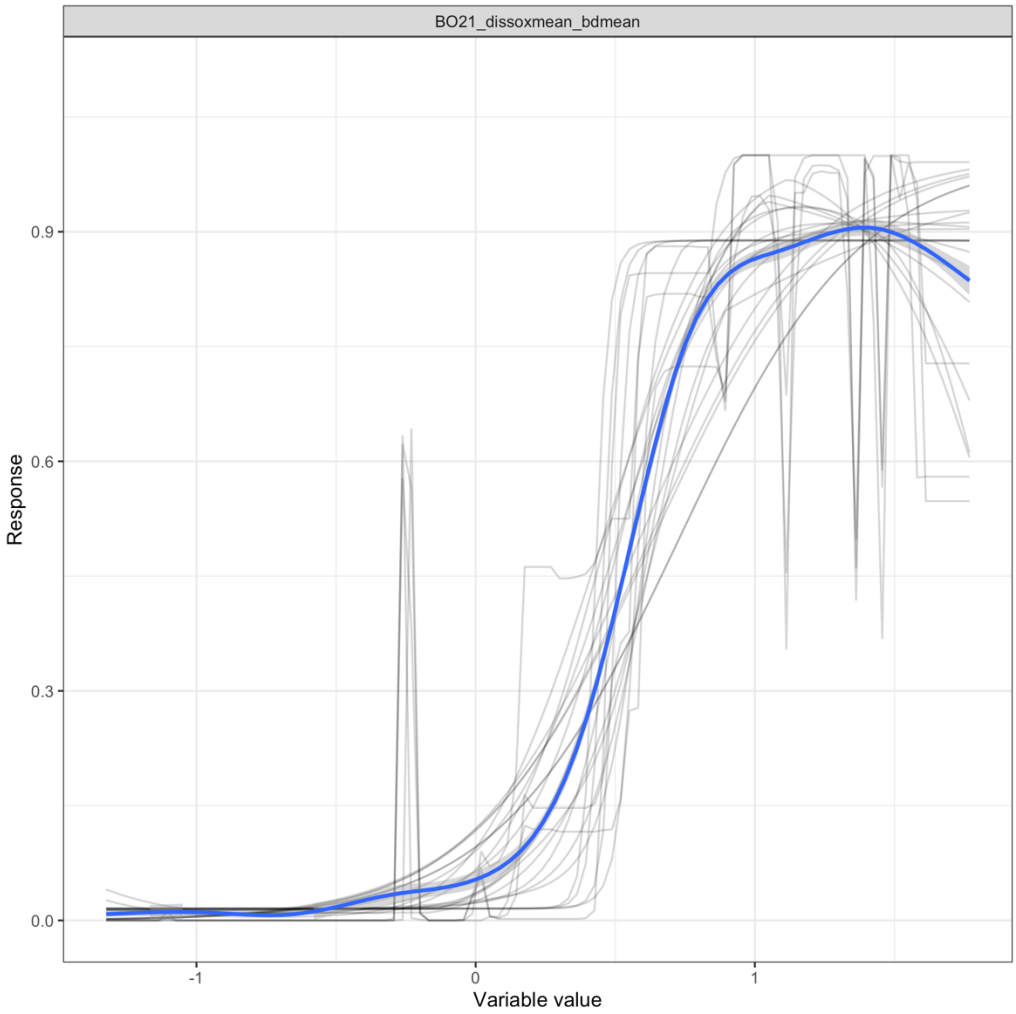

Figure 12. Response curves of the probability of presence of Cluster 2 for each predictor variable (bottom dissolved oxygen) based on the seven algorithms used for the prediction. The response curve is the dependence of the probability of the presence on one predictor variable after averaging out the effects of the other predictor variables in the model.

.

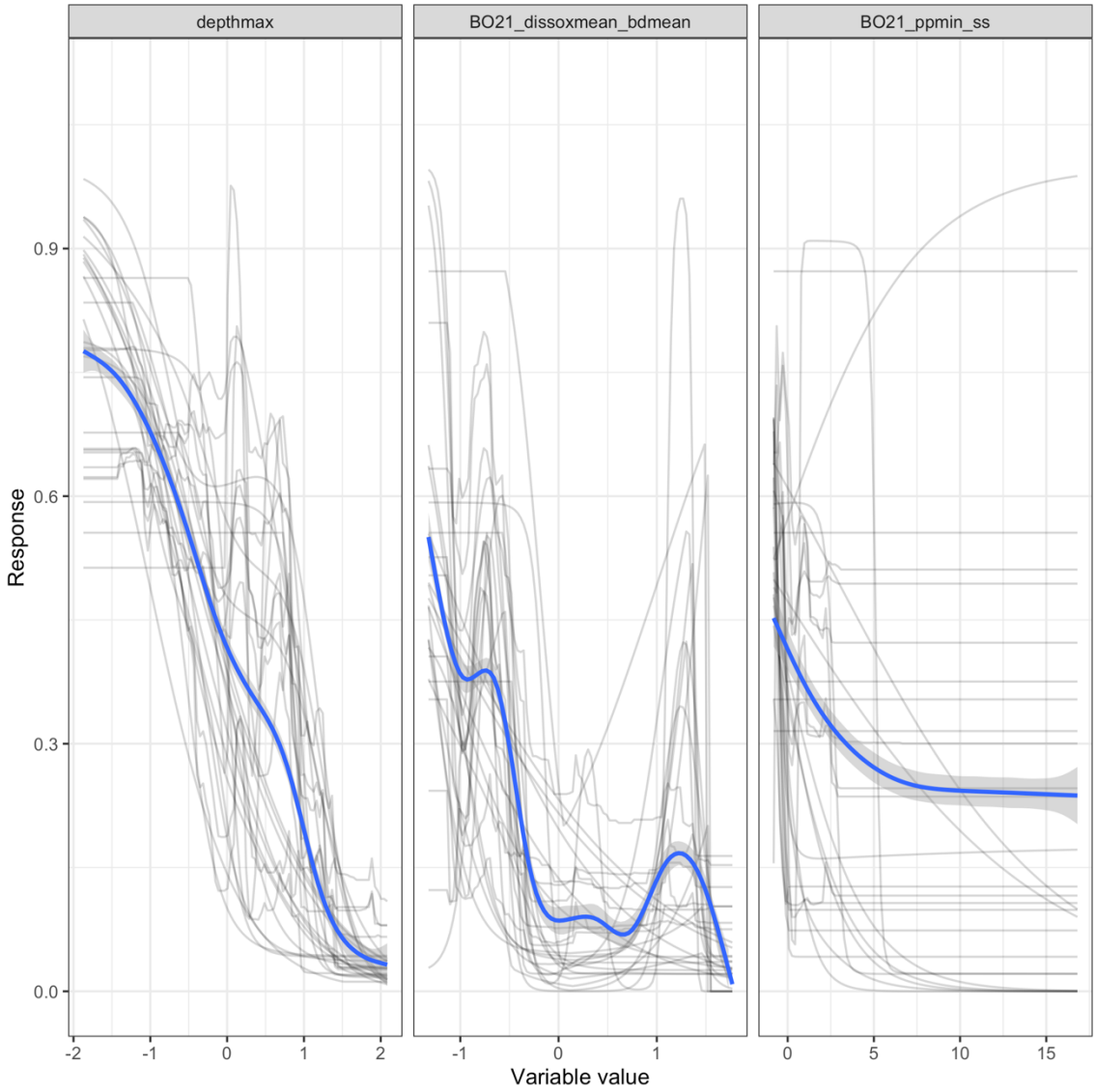

Figure 13. Response curves of the probability of presence of Cluster 3 for each predictor variable (bottom dissolved oxygen, maximum bottom depth, and minimum primary productivity at surface) based on the seven algorithms used for the prediction. The response curve is the dependence of the probability of the presence on one predictor variable after averaging out the effects of the other predictor variables in the model.

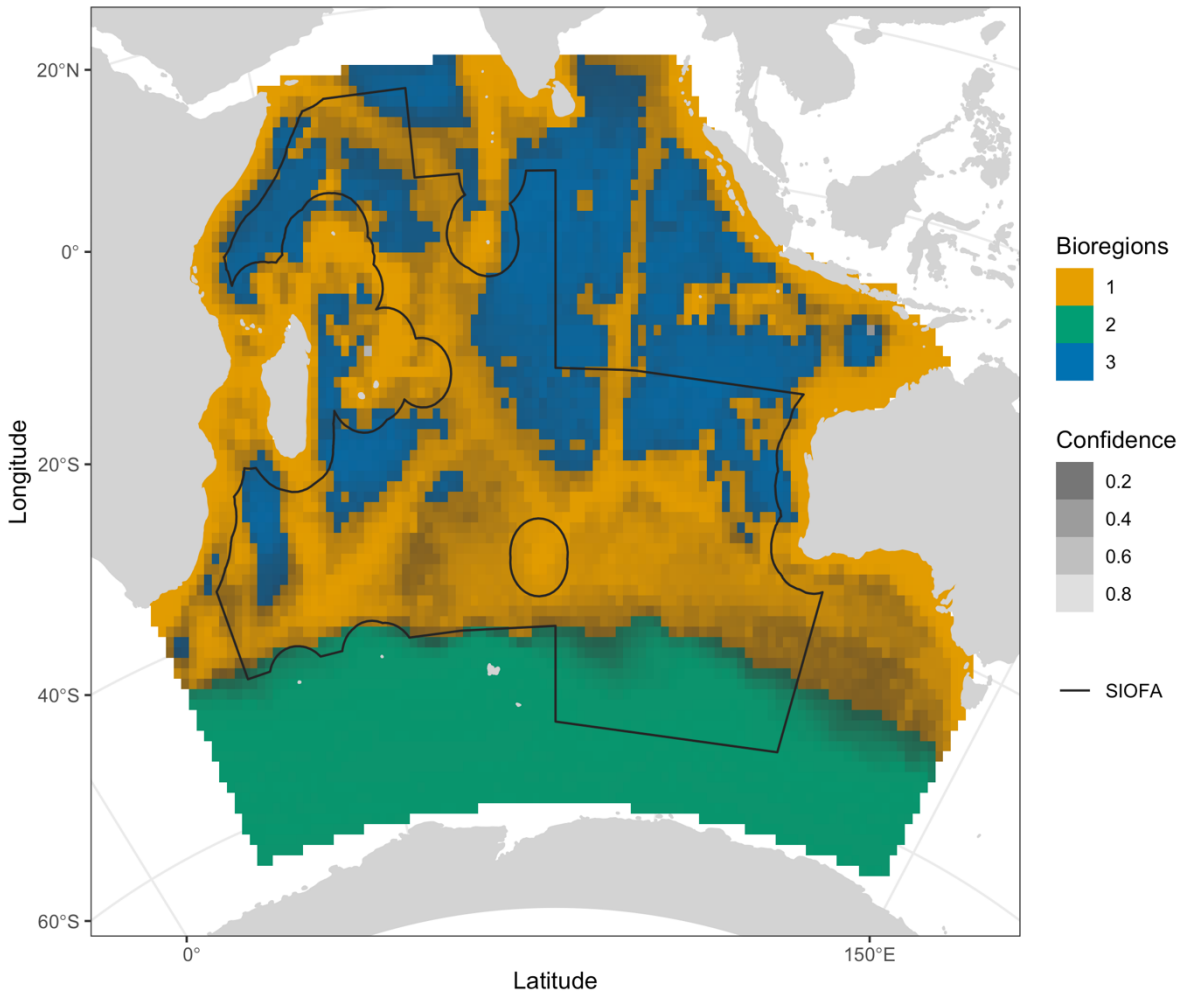

Figure 14. Predicted biogeographical regions of VME indicator taxa in the Southern Indian Ocean at the first level of the hierarchy. Areas with low confidence in the prediction are shown in darker shades of grey.

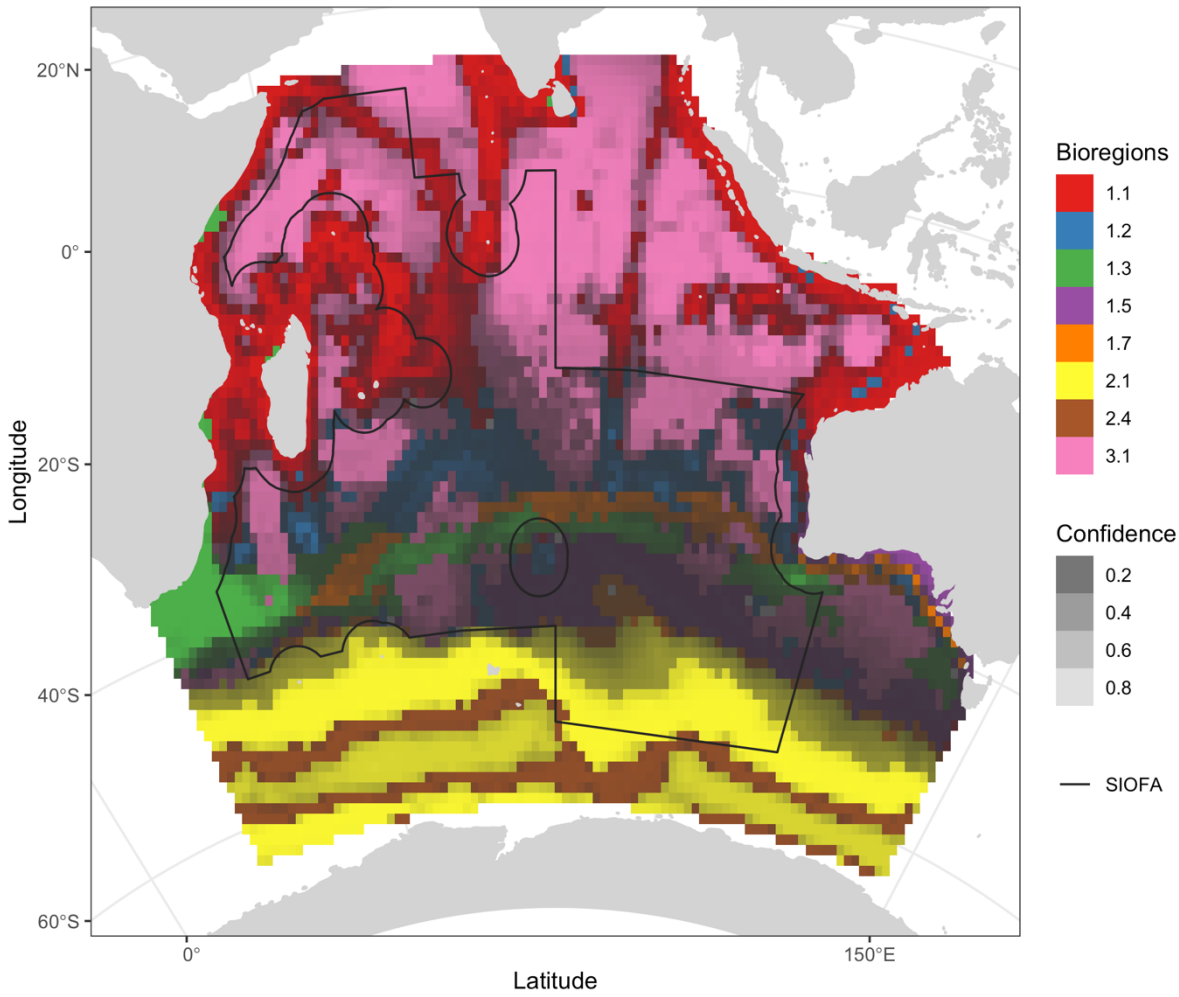

Figure 15. Predicted biogeographical regions of VME indicator taxa in the Southern Indian Ocean at the second level of the hierarchy. Areas with low confidence in the prediction are shown in darker shades of grey. Note that, because of the low number of points, we cannot reliably evaluate these predictions.

3.2.2 Bioregions based on the "Predict first, then group" approach **(Deliverable 3.2)**

We produced bioregionalisations based on family, genus, and species level for the complete dataset, and for a subset composed only of taxa with more than 30 occurrences. Because we cannot reliably evaluate models for taxa with less than 30 occurrences, we only analyse here results for bioregionalisation maps for taxa with more than 30 occurrences. Nevertheless, we provide in Appendix D bioregionalisation maps and networks for the complete dataset.

Family level bioregionalisation

Family level bioregionalisation predicted 6 bioregions (Figure 16). In general, the clusters were highly asymmetric, with a dominant one composed of the majority of taxa, and smaller ones composed by only one to two taxa (Table 8). Environmental conditions were relatively similar among clusters, although specific differences were identified for some variables (Figure 17). Further, although individual clusters (except cluster 1) were generally exclusively explained by 1 to 2 taxa, these taxa generally had distributions expanding beyond the clusters. We describe each cluster in detail below in terms of distribution (Figure 16), environmental (Figure 17) and taxa composition (Figure 18; Table C1 in Appendix C).

Cluster 1 followed the continental shelf and slope, and the main ridges of the Indian Ocean (Figure 16). Median depth was ~ 4,000 m, with distribution of values skewed towards shallower depths (< 2,000 m) (Figure 17). This cluster includes areas with strong speed of currents, and a broad range of salinities given the large extent of the cluster. Dissolved molecular oxygen median values were ~250 mol m^{-3}, with values positively skewed towards higher levels. Temperature, surface primary productivity, POC export flux, and range primary productivity vary largely across the cluster (Figure 17).
The biogeographical network is dominated by Cluster 1 and the large number of families shared among the locations covered by this cluster (Figure 18). Some of the families had a large number of connections, showing as distinct nodes in the network. and consequently, having a higher indicator value than the others (Table C1, Appendix C). Nevertheless, indicator values were not high (< 0.5) and were driven by fidelity (i.e., species are site-specific). Some of these families were Chlidonophoridae (Terebratulida), Umbellulidae (Pennatulacea), Platidiidae (Terebratulida), Euplectillidae (Lyssacinosida), Cellariidae (Cheilostomatida), Hyalonematidae (Amphidiscosida), Molgulidae (Stolidobranchia) and Primnoidae (Alcyonacea).

Cluster 2 was present north of 40°S in the shallower depths of the Mozambique Basin, the Crozet Basin, Madagascar Basin, east of the South Australian Basin, Perth and Wharton Basin, Mid-Indian Basin, North Australian Basin, Cocos Basin, Mascarene Basin, Somali Basin, Arabian Basin (Figure 16). The distribution of Cluster 2 in shallower areas of the main basins was observed in the median depth for this cluster which was approximately 4,500 m (Figure 17). Silicate values were similar to Cluster 1 and currents velocity to Clusters 3-6. Omega aragonite values were slightly lower than for other clusters. Salinity, temperature, POC export flux values were in the same range as the other clusters. Dissolved oxygen values had lower variability than Cluster 1 (Figure 17).
Cluster 2 was characterised uniquely by the family Agneziidae (Ascidiacea:Phlebobranchia), a good indicator of Cluster 2 because it occured in all sites belonging to this group (i.e., A =

1.00), although its predicted distribution also expands beyond Cluster 2 (i.e., F = 0.56) (Table C1, Appendix C). In addition, although cluster 2 was characterised by Agneziidae, some other families were also predicted to occur in sites of cluster 2, as shown by the links in Figure 18.

Cluster 3 was restricted to latitudes north of 15°S, in the Mid-Indian Basin and the Somali and Arabian Basin (Figure 16). Cluster 3 was slightly shallower than other clusters, together with Cluster 1 and Cluster 6. Silicate values are slightly lower than for Cluster 1 and 2, while salinity and surface primary productivity were more restricted than for these clusters (Figure 17). Dissolved oxygen presented low variability, with a median value of 235 mol m^{-3} (Figure 17).
Cluster 3 was characterised by the family Schizopathidae (order Antipatharia), a good indicator because it occured in all sites belonging to this group (i.e., A = 1.00), although its range was predicted to expand beyond Cluster 3 (i.e., F = 0.37) (Table C1, Appendix C). Sites belonging to Cluster 3 shared species with other clusters as observed in the network (Figure 18).

Cluster 4 had a similar geographical distribution to Cluster 3 (Figure 16), although it was slightly deeper. Environmental values for all variables were similar to Cluster 3, except for dissolved oxygen which were lower on average (Figure 17).
Cluster 4 was characterised by the family Styelidae (Stolidobranchia). The family was present in all sites of the cluster (A =1.00), although its distribution was predicted to be much larger than cluster 4 alone (F = 0.26) (Table C1, Appendix C). The network suggested more similarities with Cluster 1 than to other clusters (Figure 18).

Cluster 5 was geographically restricted, interweaved with Cluster 4 in the Mid Indian Basin, the Mascarene Basin and the Madagascar Basin (Figure 16). Cluster 5 was slightly deeper than Cluster 4, and therefore silicate values were slightly higher too (Figure 17). The remaining of the environmental variables were similar to other clusters, except for dissolved oxygen values which were the lowest among all clusters (Figure 17).
Cluster 5 was characterised by the family Pyuridae (Stolidobranchia). It occurred in all sites of the cluster (A =1.00), but also had a distribution predicted to be much larger than just this cluster (F = 0.31) (Table C1, Appendix C). Cluster 5 did not share many connections with other clusters (Figure 18).

Cluster 6 was latitudinally restricted mainly between 10°S – 30°S in the northwest of Australia in the Wharton Basin, and east of Madagascar in the Madagascar Basin (Figure 16). Cluster 6 had a median depth above 4,000 m (Figure 17). Silicate values were similar to other clusters, although always higher than 50 mol m^{-3}. Currents speed were the fastest on average, while omega aragonite was always lower than 1.4. Salinity values were strictly lower than 34.7 PSS. Dissolved oxygen was the highest on average although it varied largely across sites. Temperature was always lower than 4°C. Primary productivity, its range, and POC export flux varied the least among clusters (Figure 17).
Cluster 6 was characterised by the family Candidae (Order Cheilostomatida), present in all sites of this cluster with a restricted distribution (Table C1, Appendix C). The environmental similarity with other Clusters is observed through the position in the network (Figure 18).

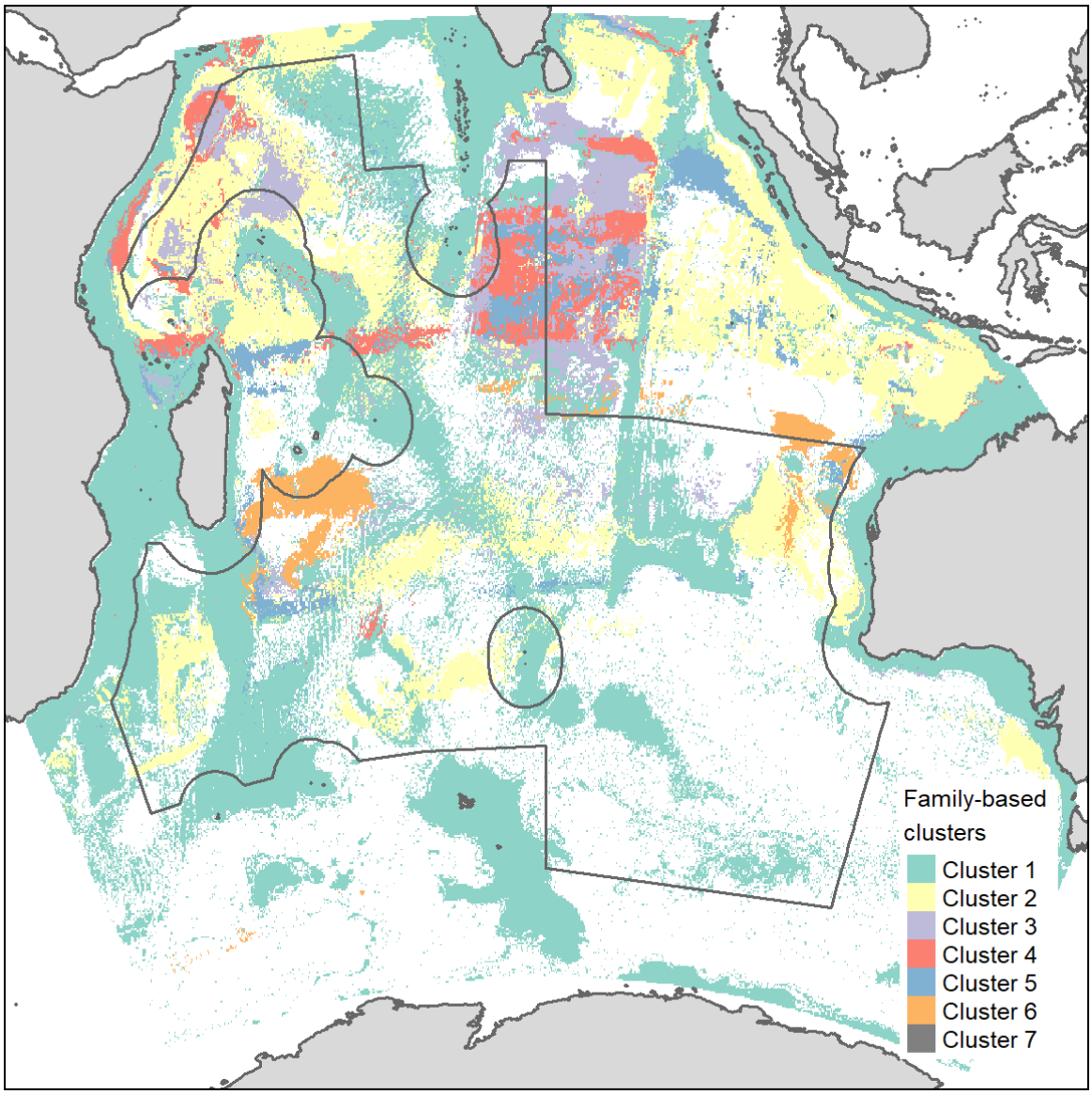

Figure 16. Family-level bioregionalisation of VME indicator taxa in the Southern Indian Ocean based on taxa with more than 30 occurrences. There are 7 bioregions depicted in different colours. Cluster 7 represents a combination of very small clusters. White areas indicate no prediction.

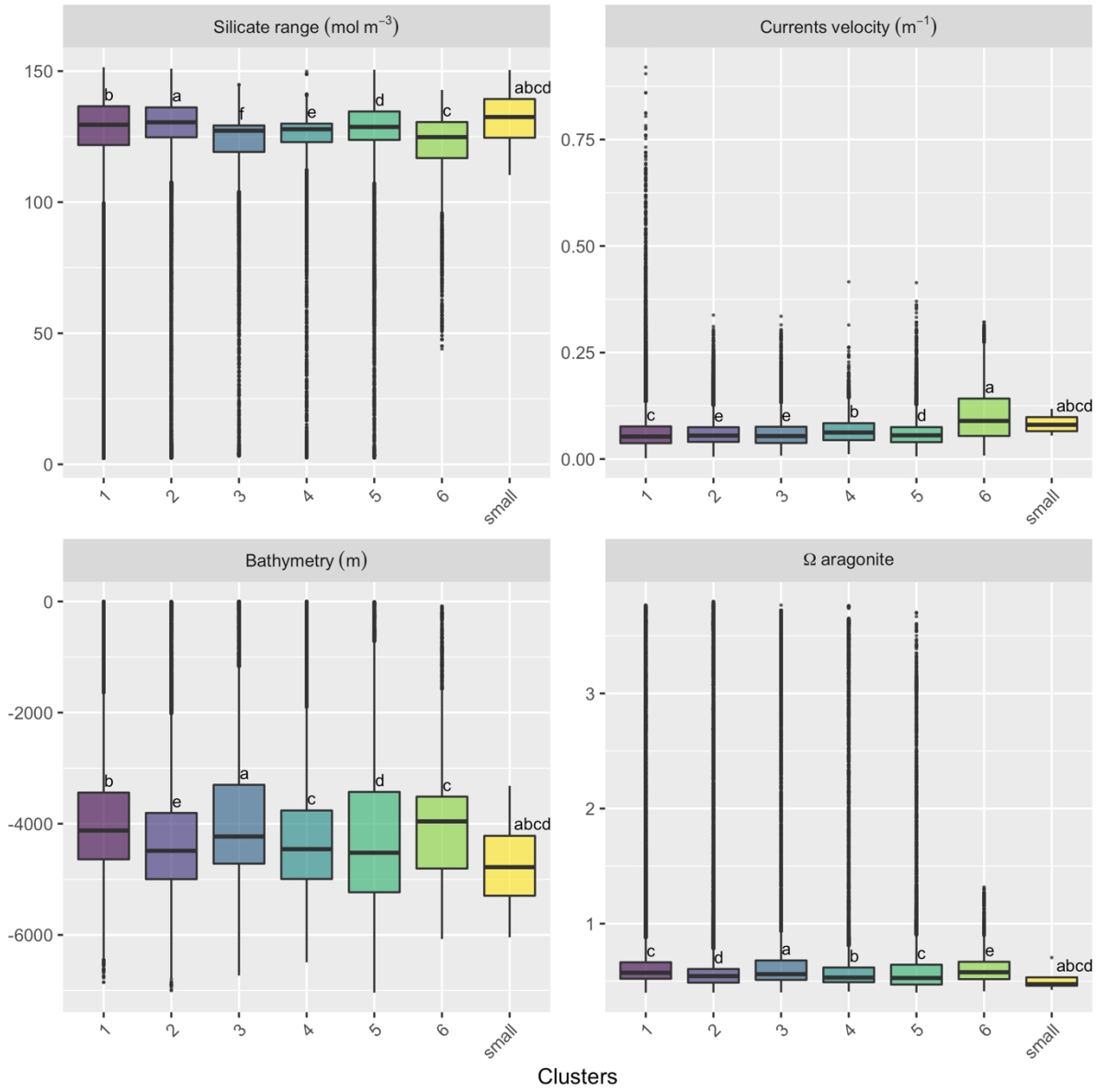

Figure 17. Environmental variables distribution of the clusters detected at family level. Small clusters (labelled 'small' in the graph) are also shown for comparison. Variables from top left to bottom right: silicate concentration (mol m^{-3}); speed of currents (m^{-1}); bathymetry (m); omega aragonite. Differences between subclusters per environmental variable are shown through pairwise comparisons (post-hoc Tukey test) with letters on the boxplots: mean values with at least one common letter are not significantly different.

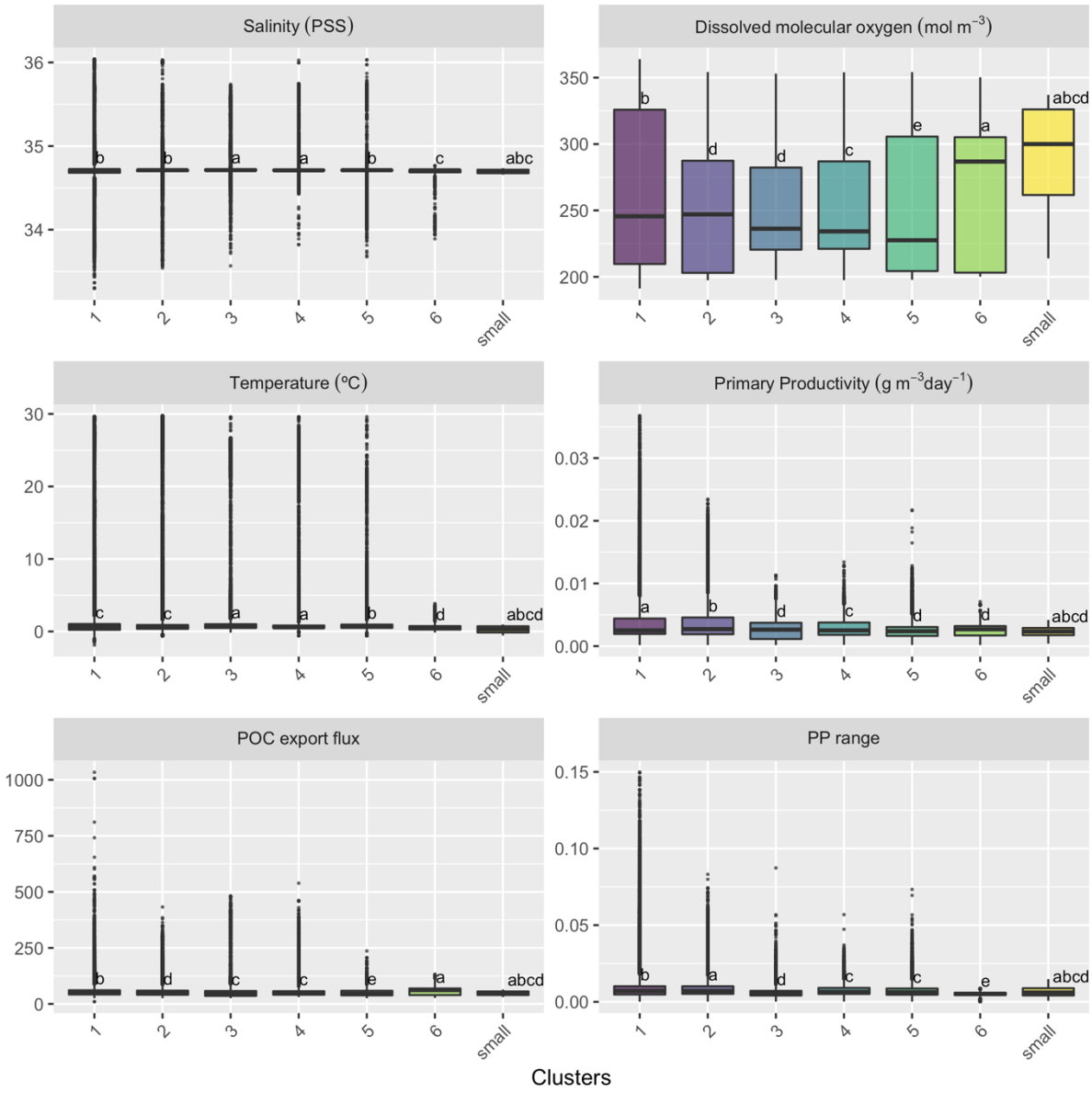

Figure 17. (Cont.). *Environmental variables distribution of the clusters detected at family level. Small clusters (labelled 'small' in the graph) are also shown for comparison. Variables from top left to bottom right: salinity (PPS); dissolved molecular oxygen (mol m^{-3}); temperature (°C); surface primary productivity (PP) (g m^{-3} day^{-1}); particulate organic carbon (POC) flux (mol m^{-3}); primary productivity (PP range.)*

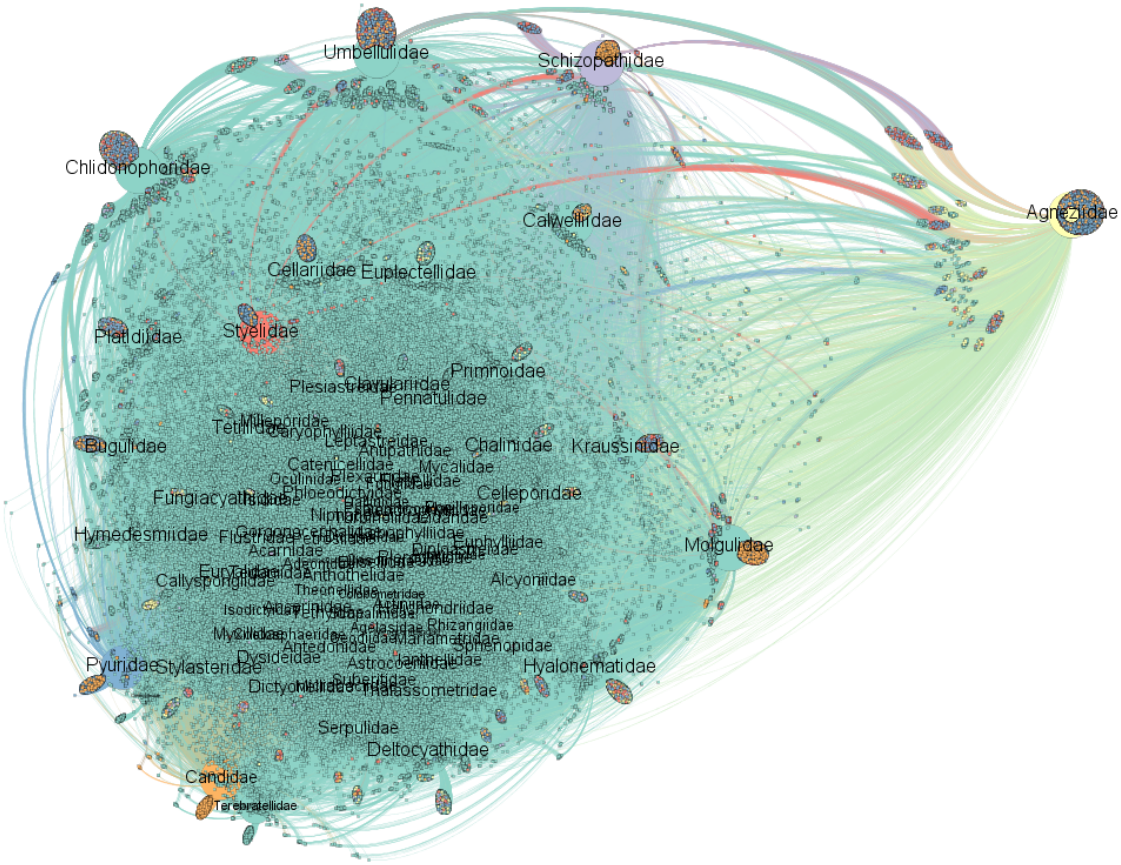

Figure 18. Biogeographical network at family level. Filter of families with at least 30 occurrences.

Table 8. Family taxa composition of the family-based bioregionalisation clusters.

cluster	Total Richness	Assigned richness	Endemic richness	% Endemic families	% Endemic families southern Indian Ocean
1	92	87	21	0.23	0.95
2	53	1	0	0	0.02
3	57	1	0	0	0.02
4	58	1	0	0	0.02
5	47	1	0	0	0.02
6	41	1	0	0	0.02
7	6	0	0	0	0

Genus level bioregionalisation

Genus level bioregionalisation predicted 7 bioregions (Figure 19). In general, the clusters were highly asymmetric, with a dominant one composed of the majority of taxa, and smaller ones compose by only one to two taxa (Table 9). Environmental conditions were relatively similar among clusters, although specific differences were identified on some variables. Further, although individual clusters (except cluster 1) were generally exclusively explained by 1 to 2 taxa, these taxa generally had distributions expanding beyond the clusters. We describe each cluster in detail below in terms of distribution (Figure 19), environmental (Figure 20) and taxa composition (Figure 21; Table C2 in Appendix C).

Cluster 1 extended across the continental shelf, slope, plateaus, and shallower offshore areas across the entire Indian Ocean (Figure 19), covering a wide range of depths (Figure 20). Given the coastal extent of the cluster salinity, temperature and surface PP varied largely (Figure 20). Dissolved oxygen values were on average 224 mol m^{-3}, although the majority of values were higher. Currents velocity and aragonite saturation horizon were similar to Cluster 2 (Figure 20).

Cluster 1 was comprised by 97 out of 108 genera (Table 9). Indicator species for the cluster showed high fidelity (F > 0.8) to the cluster with a wide distribution, but these indicator genera were found in other clusters too (Table C2, Appendix C). Some of the indicator genera were *Tedania* (Poecilosclerida: Tedaniidae), *Solenosmilia* (Scleractinia: Caryophylliidae), *Myxilla* (Poecilosclerida:Myxilliidae), *Stephanocyathus* (Scleractinia: Caryophylliidae), *Ascidia* (Phlebobranchia:Ascidiidae), *Lobactis* (Scleractinia:Fungiidae), *Spirobranchus* (Sabellida:Serpulidae) (Table C2, Appendix C). Cluster 1 was the largest of all genus-based clusters and the amount genera shared across sites was evident in the network through the tight connections between nodes of this cluster (Figure 21). At the same time, the cluster was highly distinct from other clusters (Figure 21).

Cluster 2 followed the same geographic distribution as Cluster 1 (Figure 19). Salinity and temperature values ranged broadly depending on the site (Figure 20). Surface PP over predicted areas was generally lower than Cluster 1, although POC flux reached a maximum of 500 mol m^{-3} in some sites (Figure 20).
Cluster 2 comprised three genera (Table 9), including *Platidia* (Terebratulida:Platidiidae), *Stylaster* (Anthoathecata:Stylasteridae) and *Umbellula* (Pennatulacea: Umbellulidae) (Table C2, Appendix C). These three genera could generally be found in most of the sites of Cluster 2 (A= 0.45-0.75) and in other clusters (F= 0.43-0.48), as observed in the connections with other sites in the network (Figure 21).

Cluster 3 was predominantly distributed in the Mid Indian Ocean Basin, the Somali and Arabian Basin, with some presences east of the Madagascar Plateau, Broken Ridge and east of South Africa (Figure 19). Cluster 3 had the lowest median depth together with Cluster 9 (Figure 20). Salinity values were generally higher than 34 PSS, reaching a maximum of 35.8 PSS. Dissolved oxygen was on average ~230 mol m^{-3}, although the distribution of values was skewed towards higher values. Temperature spanned largely (Figure 20).
Cluster 3 was characterised by the deep-sea genera *Bathypathes* (Antipatharia:Schizopathidae), which was present in all sites of the cluster (A=1.00), but also in sites outside of the cluster (F=0.45) (Table C2, Appendix C). Cluster 3 appeared as a distinct cluster in the network, with connections to other clusters (Figure 21).

Cluster 4 was also geographically distributed north of 35°S, in the north of Saint Paul and Amsterdam, northeast and southeast of Madagascar, deeper flanks of the Carlsberg Ridge, Chagos-Laccadive Plateau and the Cocos Basin (Figure 19). Cluster 4 appeared contiguous to Cluster 2 in deeper environments. Silicate values were on average ~125 mol m^{-3} (Figure 20). Salinity ranged widel and dissolved oxygen had a median 240 mol m^{-3} (Figure 20). Temperature values were the same as Cluster 8 (Figure 20).
Cluster 4 was characterised by the indicator genus *Styela* (Stolidobranchiae:Styelidae), which was present in all sites of the cluster (A=1.00), but also in other sites outside the cluster (F= 0.54) (Table C2, Appendix C). It appeared as a distinct cluster in the network with multiple connections (Figure 21).

Cluster 5 was a subantarctic cluster, found west and east of the Kerguelen Plateau and as far north as Del Caño Rise at 45°S (Figure 19). Salinity values were fresher than other clusters, dissolved oxygen were on average higher than 260 mol m^{-3} and temperature values were generally low (Figure 20).
Cluster 5 was characterised by two genera: *Aerothyris* (Terebratulida: Terebratellidae) and Cornucopina (Cheilostomatida: Bugulidae) (Table C2, Appendix C). Both genera were not distributed in all sites; *Cornucopina* occurred in most sites of the cluster (A=0.85), whereas *Aerothyris* was restricted to a number of sites (A=0.22). Both genera had distributions predicted to extend beyond cluster 5 (F = 0.41-0.47) (Table C2, Appendix C). These two genera appeared as distinct nodes in the network, sharing connections with Cluster 1 and Cluster 2, mainly (Figure 21).

Cluster 8 was restricted to latitudes south of 35°S, in the Agulhas Basin, south of the Crozet Basin, and east and south of the Kerguelen Plateau (Figure 19). Cluster 8 showed

environmental similarities to Cluster 5, however dissolved oxygen was lower on average and temperature and surface PP more variable than Cluster 5 (Figure 20).
Cluster 8 was characterised by the genus *Astrotoma* (Euryalida: Gorgonocephalidae), a good indicator as it was found in all sites of the cluster (A=1.00), and its distribution was largely predicted to be within the cluster (F=0.76) (Table C2, Appendix C). It appeared as a distinct cluster in the network with connections notably with Cluster 9 (Figure 21).

Finally, Cluster 9 was confined to the Mid Ocean Basin in the northern Indian Ocean between 10°N – 15°S (Figure 19). Unexpectedly, currents speed was the fastest among the clusters for its deep bathymetric range (Figure 20). Salinity values were restricted to values between 34.1 and 35.5 PSS and silicate below 125 mol m^{-3} (Figure 20).
Cluster 9 was characterised by the genus *Gorgonocephalus* (Euryalida, Gorgonocephalidae), present in all sites of the cluster (A= 1.00), although its distribution was predicted to expand beyond the cluster (F= 0.62) (Table C2, Appendix C). Connections with sites from Cluster 8 were seen in the network, as well as with other clusters (Figure 21).

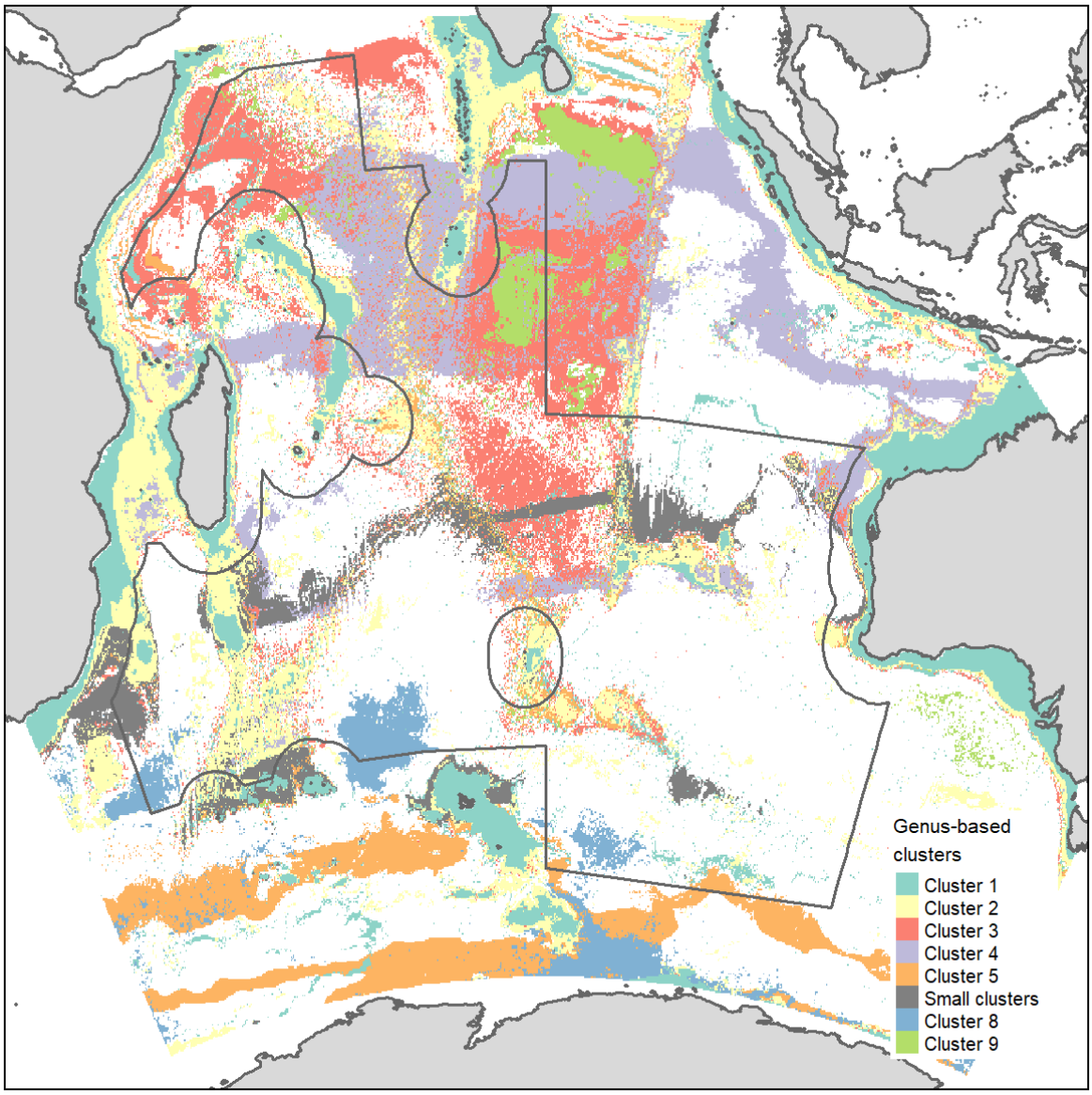

Figure 19. Genus-level bioregionalisation of VME indicator taxa in the Southern Indian Ocean based on taxa with more than 30 occurrences. There are 7 bioregions depicted in different colours. White areas indicate no prediction.

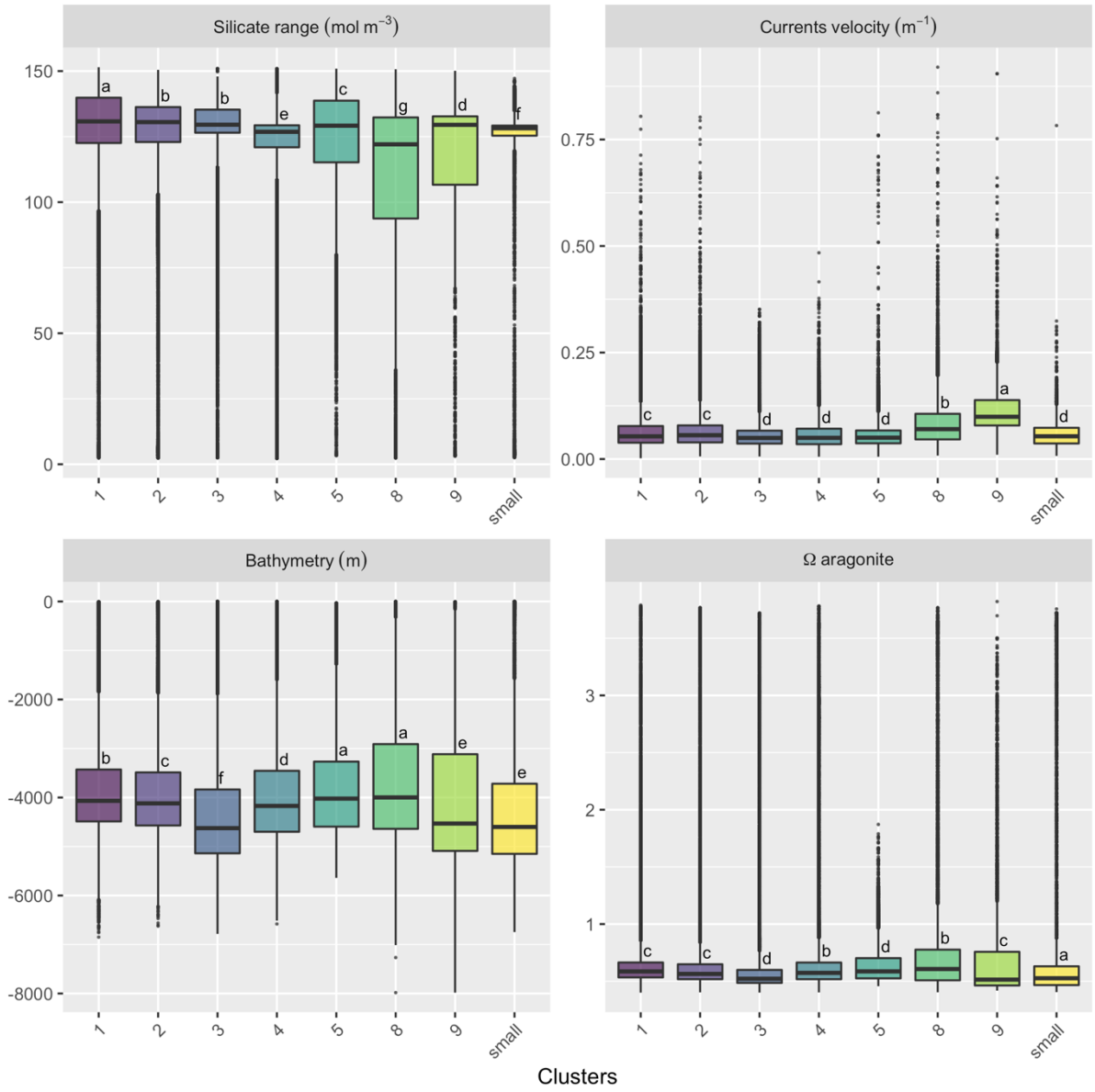

Figure 20. Environmental variables distribution of the clusters detected at genus level. Small clusters (labelled 'small' in the graph) are also shown for comparison. Variables from top left to bottom right: silicate concentration (mol m-3); speed of currents (m-1); bathymetry (m); omega aragonite. Differences between subclusters per environmental variable are shown through pairwise comparisons (post-hoc Tukey test) with letters on the boxplots: mean values with at least one common letter are not significantly different.

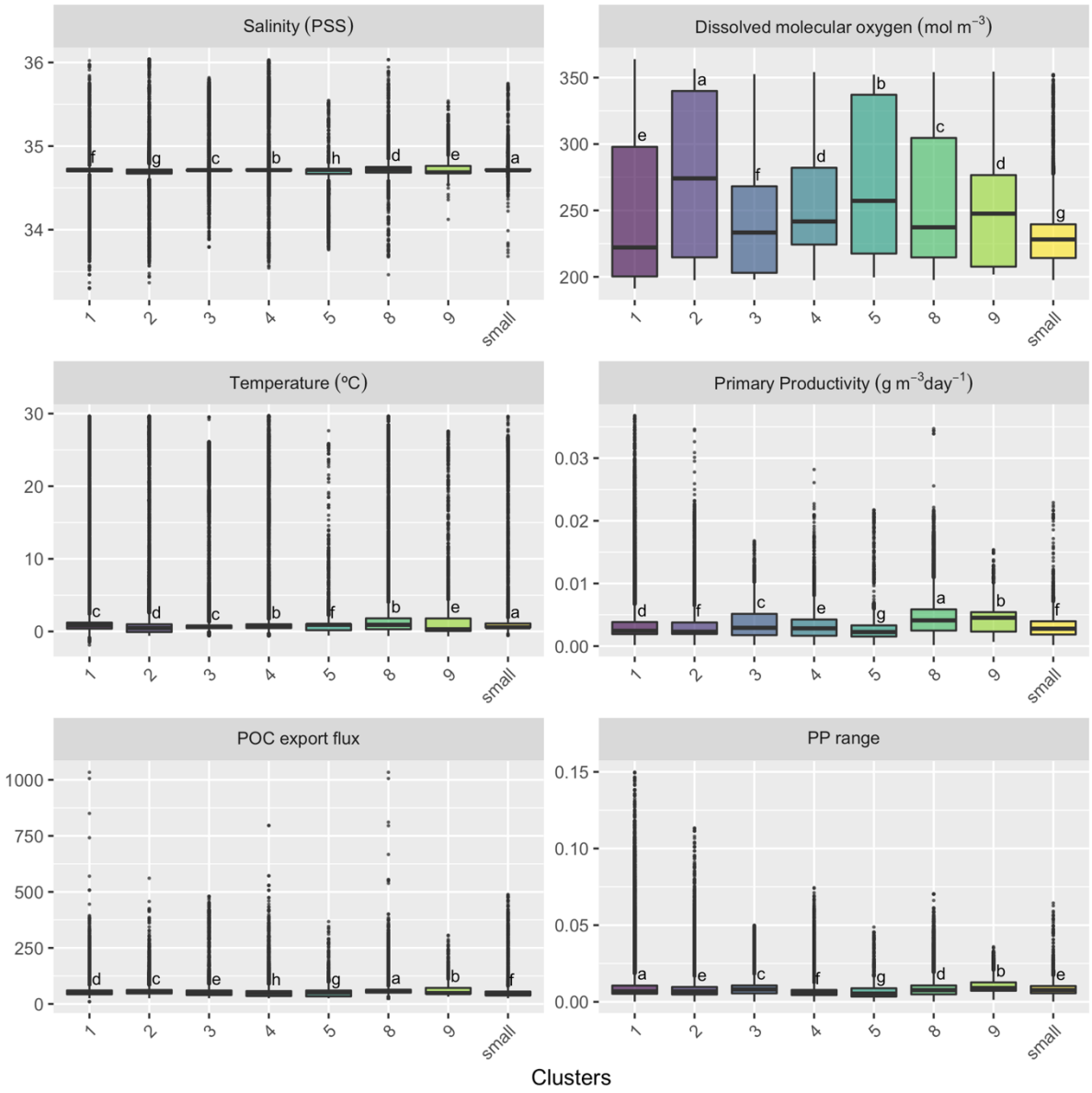

Figure 20. (Cont.). Environmental variables distribution of the clusters detected at genus level. Small clusters (labelled 'small' in the graph) are also shown for comparison. Variables from top left to bottom right: salinity (PPS); dissolved molecular oxygen (mol m^{-3}); temperature (°C); surface primary productivity (PP) (g m-3 day-1); particulate organic carbon (POC) flux (mol m^{-3}); primary productivity (PP range.)

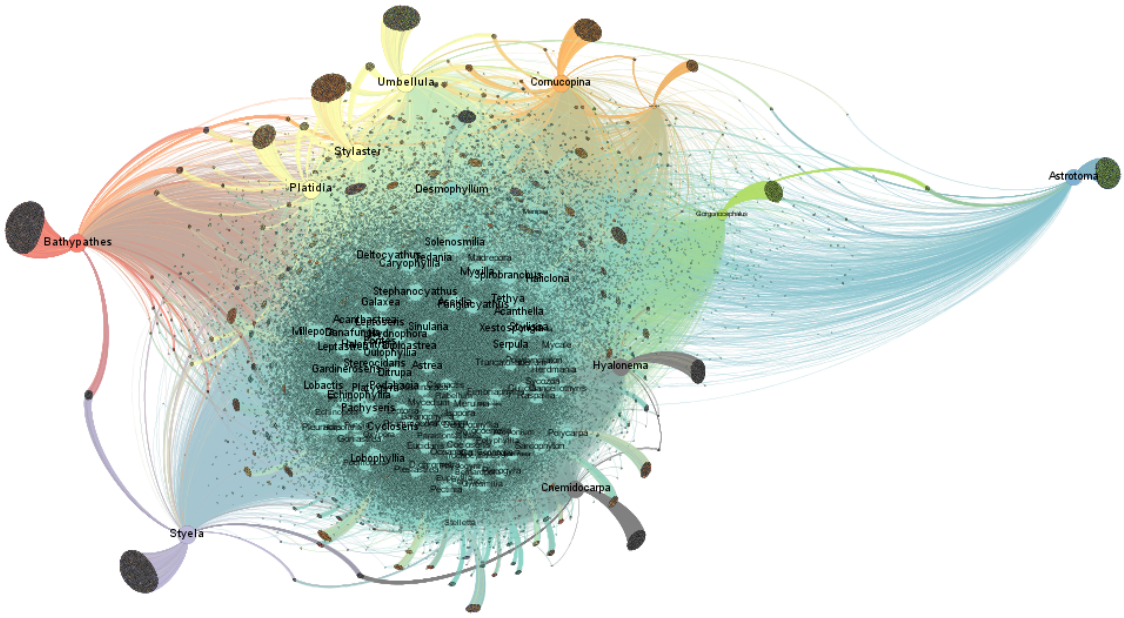

Figure 21. Biogeographical network at genus level. Filter of genera with at least 30 occurrences.

Table 9. Genus taxa composition of the genus-based bioregionalisation clusters.

Cluster	Total richness	Assigned richness	Endemic richness	% Endemic genera	% Endemic genera Southern Indian Ocean
1	108	97	30	0.28	0.90
2	73	3	0	0.00	0.04
3	53	1	0	0.00	0.02
4	23	1	0	0.00	0.04
5	55	2	0	0.00	0.04
8	31	1	0	0.00	0.03
9	26	1	0	0.00	0.04

Species level bioregionalisation

Species level bioregionalisation suggested 8 bioregions (Figure 22). In general, clusters are highly asymmetric, with a dominant one composed of the majority of taxa, and smaller ones compose by only one to two taxa (Table 10). Environmental conditions are relatively similar among clusters, although specific differences can be identified on some variables. Further, although individual clusters (except cluster 1) are generally exclusively explained by 1 to 2 taxa, these taxa generally have distributions expanding beyond the clusters. We describe each cluster in detail below in terms of distribution (Figure 22), environmental (Figure 23) and taxa composition (Figure 24; Table C3 in Appendix C).

Cluster 1 was distributed mainly from 5°N to 30°S (Figure 22). From 5°N to 15°S it was homogenously distributed, while south of 15°S it was present at eastern and western coasts (Figure 22). Median depth was at 4,000 m, although it spanned all depths (Figure 23). Currents velocity were among the lowest values with a maximum around 0.375 m^{-1}. Silicate values were around 135 mol m^{-3}, with most sites with lower values. Salinity, temperature and dissolved oxygen varied according to areas (Figure 23).
Cluster 1 was characterised by 145 out of 152 species (Table 10). Indicator values were driven mostly by fidelity, but indicator species were present in other clusters too (Table C3, Appendix C). Cluster 1 was the largest of clusters with many connections among sites of the same cluster, as observed in the network (Figure 24).

Cluster 2 was distributed along the coasts, plateaus and shallower depths of the ridges (Figure 22). Dissolved oxygen was generally low in the cluster with a median value of 225 mol m^{-3} (Figure 23). This cluster had the lowest aragonite (Figure 23).
Cluster 2 was characterised by *Bathypathes patula* (Antipatharia: Schizopathidae). This species was found in all sites of the cluster (A = 1.00), and largely predicted to occur in the

cluster (Table C3, Appendix C). It appeared in the network as a distinct cluster with connections notably to Cluster 7 (Figure 24).

Cluster 3 had a subantarctic distribution, found at latitudes south of 45°S across all longitudes, with a split distribution at 50°S east of the Kerguelen Plateau (Figure 22). Cluster 3 had a median bathymetry of 3,800 m, the lowest median silicate values, and currents velocity among the fastest (Figure 23). Temperature, surface PP and POC flux export were slightly higher than other clusters (Figure 23).
Cluster 3 was characterised by *Astrotoma agassizii* (Euryalida: Gorgonocephalidae), found in all sites of the cluster (A=1.00), rarely predicted to occur beyond the cluster (F=0.82) (Table C3, Appendix C). Cluster 3 was clearly separated from other clusters with only a few connections, suggesting a specific niche, as it seemed the case for Cluster 2 (Figure 24).

Cluster 4 was distributed contiguously to Cluster 1 in deeper environments, following the same distribution across the entire Indian Ocean (Figure 22). Bathymetry had a median of 4,000 m, although it was usually shallower (Figure 23). Silicate values were lower than 130 mol m^{-3} and currents velocity were quite restricted. Salinity values were the most varied and dissolved oxygen had the second highest value at 270 mol m^{-3}. Surface PP and POC export flux were similar to those of Cluster 5 and 7 (Figure 23).
Cluster 4 was characterised by *Platidia anomioides* (Terebratulida:Platidiidae), although this species was also largely distributed outside the cluster (F = 0.30) (Table C3, Appendix C). The network suggested similarities to Cluster 2, with some environmental overlap (Figure 24).

Cluster 5 was limited to latitudes south of 27°S, in the Madagascar Plateau, Southwest Indian Ridge and Del Caño Rise, Agulhas Plateau, Kerguelen Plateau, and Broken Ridge (Figure 22). Bathymetry on average was shallower than 4,000 m, and it had relatively higher speed of currents and aragonite saturation than other clusters (Figure 23). Cluster 5 had the second lowest silicate values at ~ 115 mol m^{-3}, a salinity restricted between 34 and 35.5 PSS (Figure 23). Surface PP over areas of Cluster 5 was reduced (Figure 23).
Cluster 5 was characterised by *Aerothyris kerguelensis*. Although it was a good indicator of the cluster (A=1.00), it was also predicted to frequently occur outside of it (F=0.46) (Table C3, Appendix C). The cluster was very close to Cluster 2 in the network, suggesting some depth overlap and environmental conditions (Figure 24).

Cluster 6 was present south of 35°S in the Crozet Basin, Australian-Antarctic Basin (Figure 22). It was the deepest cluster, although most of the records were in shallower areas (Figure 23). Silicate range was narrow and currents speed were the fastest (Figure 23). Salinity was fresh with small variability. Cluster 6 showed the highest POC export flux and second highest values of surface PP (Figure 23).
Cluster 6 was characterised by *Gorgonocephalus chilensis* (Euryalida: Gorgonocephalidae), occurring in all sites of the cluster (A = 1.00). It also occurred in sites of other clusters (F=0.70) (Table C3, Appendix C). In the network, Cluster 6 appeared at the intersection between Cluster 1 and Cluster 3 (Figure 24).

Cluster 7 was predicted mainly in the Mozambique Channel, east of the Southwestern Indian Ridge, Madagascar Plateau and Agulhas Current (Figure 22). Bathymetry was skewed towards

shallower depths and dissolved oxygen exhibited the highest median value (Figure 23). Temperature varied with a gap between 20° and 25°C (Figure 23).
Cluster 7 was characterised by *Deltocyathus magnificus* (Scleractinia:Deltocyathidae), which occurred in all sites of the cluster, but with a distribution largely predicted to expand beyond the cluster (F=0.20) (Table C3, Appendix C). In the network, this cluster was embedded among the larger Cluster 1 suggesting depth and environmental overlap (Figure 24).

Finally, Cluster 8 was exclusively present in the Kerguelen Plateau (Figure 22). Bathymetry ranged from 2,500 to 5,000 m (Figure 23). Currents velocity and aragonite saturation exhibited distinct values (Figure 23).
Cluster 8 was characterised by *Antarctotetilla leptoderma* (Tetractinellida: Tetillidae), occurring in all sites of the clusters, but also predicted to occur outside of it (F=0.53) (Table C3, Appendix C).

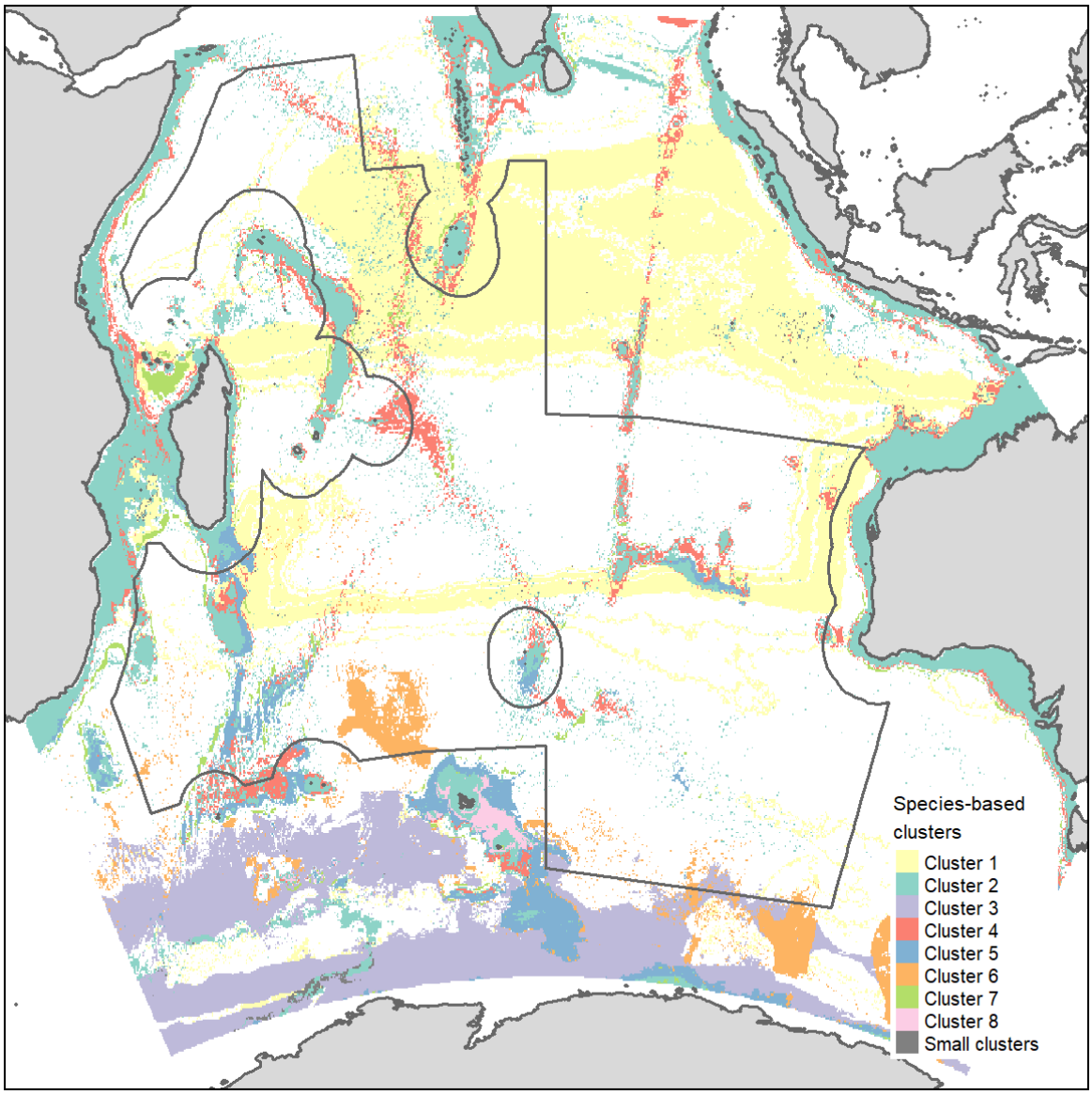

Figure 22. Species-level bioregionalisation of VME indicator taxa in the Southern Indian Ocean based on species with more than 30 occurrences. There are 8 bioregions depicted in different colours. White areas indicate no prediction.

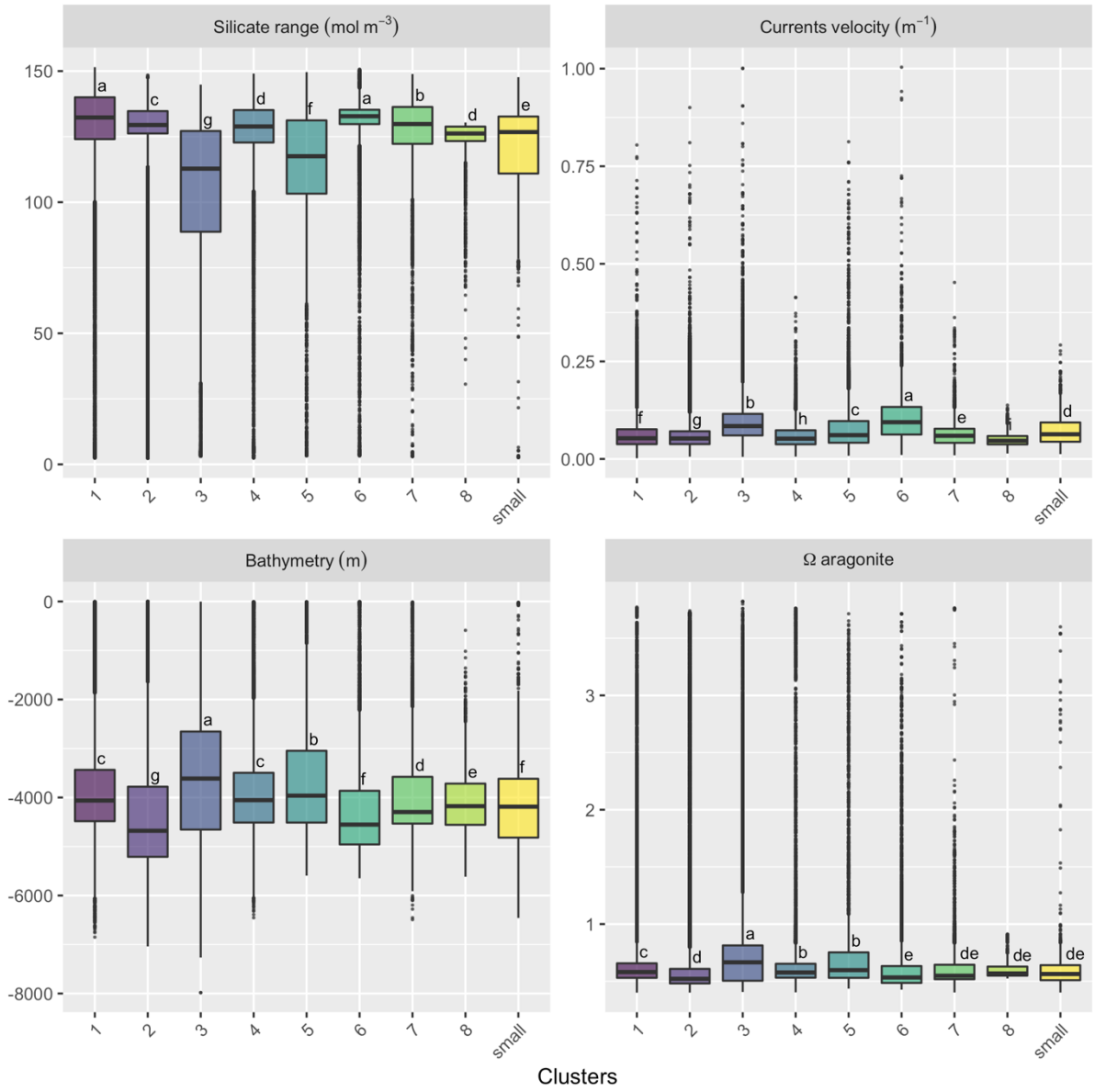

Figure 23. Environmental variables distribution of the clusters detected at species level. Small clusters (labelled 'small' in the graph) are also shown for comparison. Variables from top left to bottom right: silicate concentration (mol m-3); speed of currents (m-1); bathymetry (m); omega aragonite. Differences between subclusters per environmental variable are shown through pairwise comparisons (post-hoc Tukey test) with letters on the boxplots: mean values with at least one common letter are not significantly different.

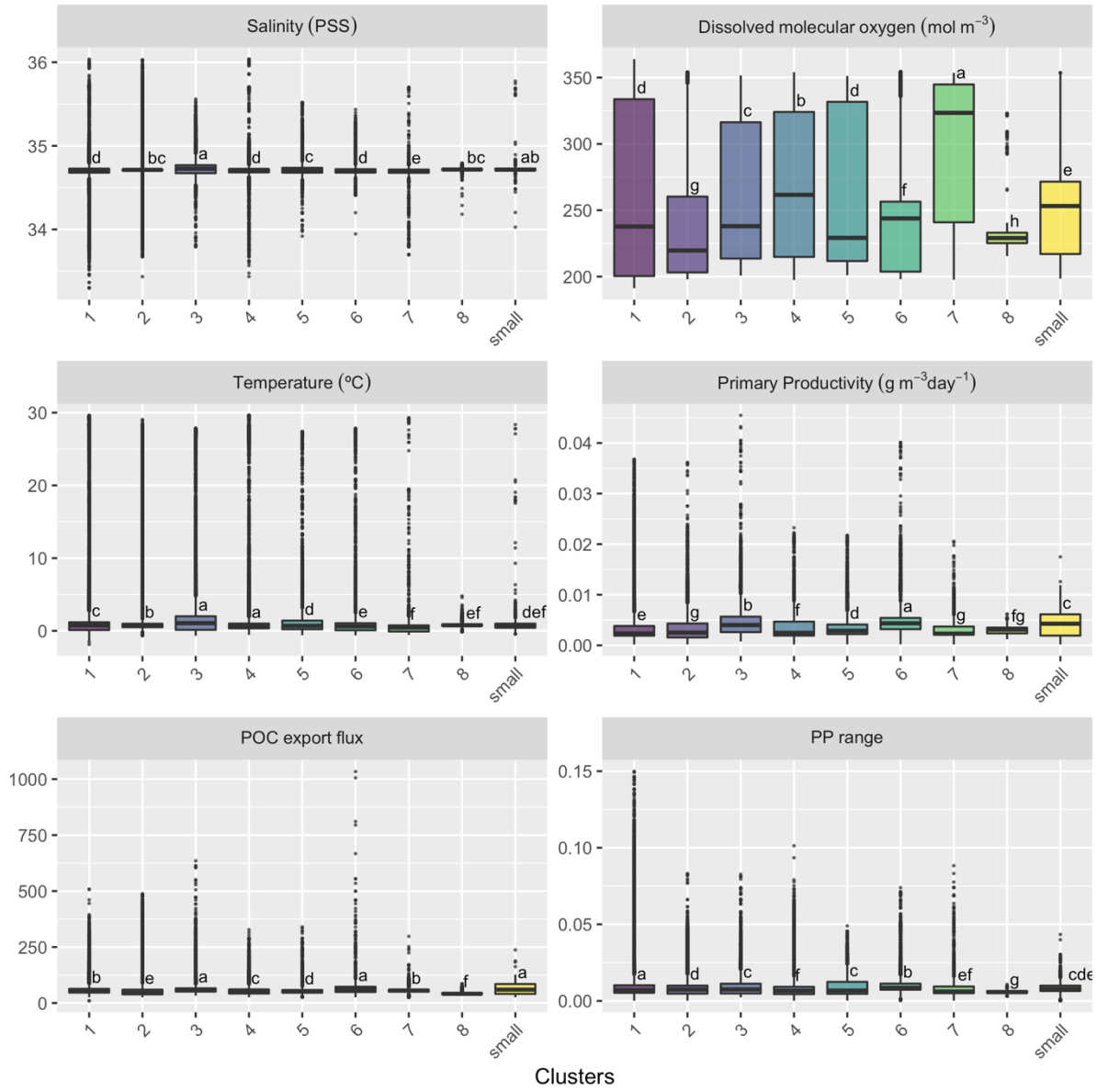

Figure 23. (Cont.). Environmental variables distribution of the clusters detected at species level. Small clusters (labelled 'small' in the graph) are also shown for comparison. Variables from top left to bottom right: salinity (PPS); dissolved molecular oxygen (mol m^{-3}); temperature (°C); surface primary productivity (PP) (g m-3 day-1); particulate organic carbon (POC) flux (mol m^{-3}); primary productivity (PP range).

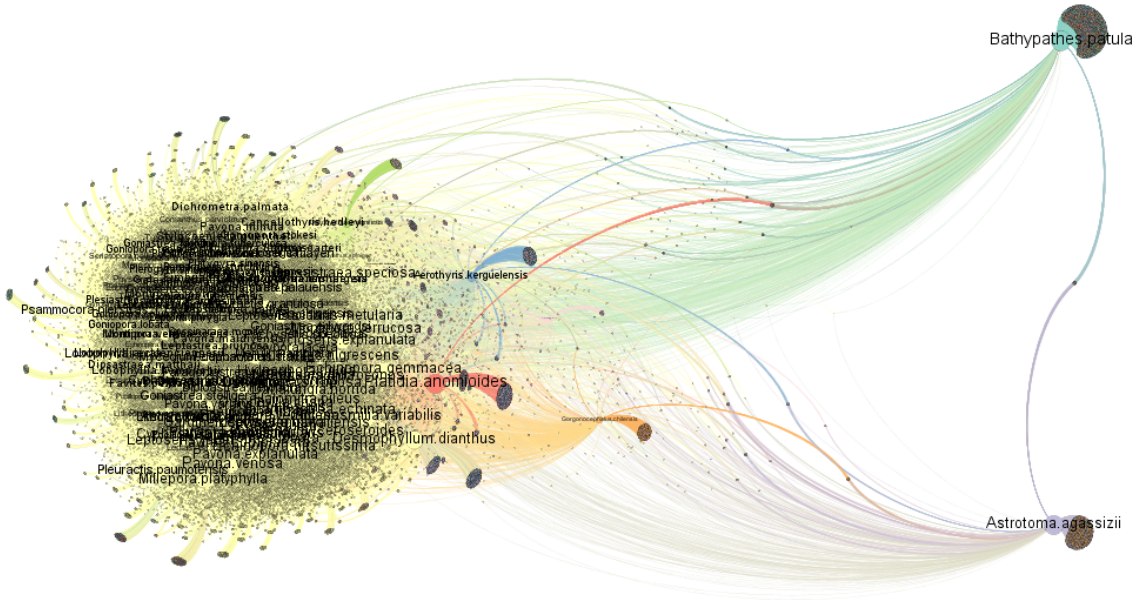

Figure 24. Biogeographical network at species level. Filter of species with at least 30 occurrences.

Table 10. Species taxa composition of the species-based bioregionalisation clusters.

Cluster	Total Richness	Assigned richness	Endemic richness	% Endemic species	% Endemic species Southern Indian Ocean
1	152	145	75	0.49	0.95
2	32	1	0	0.00	0.03
3	10	1	0	0.00	0.10
4	34	1	0	0.00	0.03
5	51	1	0	0.00	0.02
6	13	1	0	0.00	0.08
7	51	1	0	0.00	0.02
8	35	1	0	0.00	0.03

3.2.3 Bioregions based on the "analyse simultaneously" approach **(Deliverable 3.3)**

We provide preliminary results of the third bioregionalisation approach, where clustering and prediction of bioregions are performed in a single analysis. Out of the 600 models fitted (50 runs per each number of RCPs), 540 were successfully modelled. Models fitted with 10, 11, and 12 number of RCPs were considered misfits as the BIC spanned a large range of negative and positive values, and the log-likelihoods were greater than zero (Figure 25).

Initially, the lowest BIC (excluding RCPs > 10) suggested 8 RCPs as the best number of groups in the data (Figure 25). Inspection of this model suggested a misfit as there was only one group predicted and the remaining seven groups were zero groups. This was likely a result of a spike in the log-likelihood, as it can be observed in Figure 25B as a solitary very negative value. Although model selection in mixture-of-experts models (i.e., RCPs) is performed using BIC values, BIC is not entirely reliable because it sometimes can suggest a RCP for a single site (i.e., grid cell), as it was the case here. These models need to be inspected to understand whether they are a misfit. Consequently, we focused on the next lowest BIC, which corresponded to 9 RCPs. We selected this model (Log-likelihood = -40,181) and performed 20 Bayesian bootstrap samples to calculate uncertainty in the predictions resulting from the model. Spatially predicting this model showed seven bioregions and two groups with zero predictions, corresponding to RCP6 and RCP9 (Figure 26).

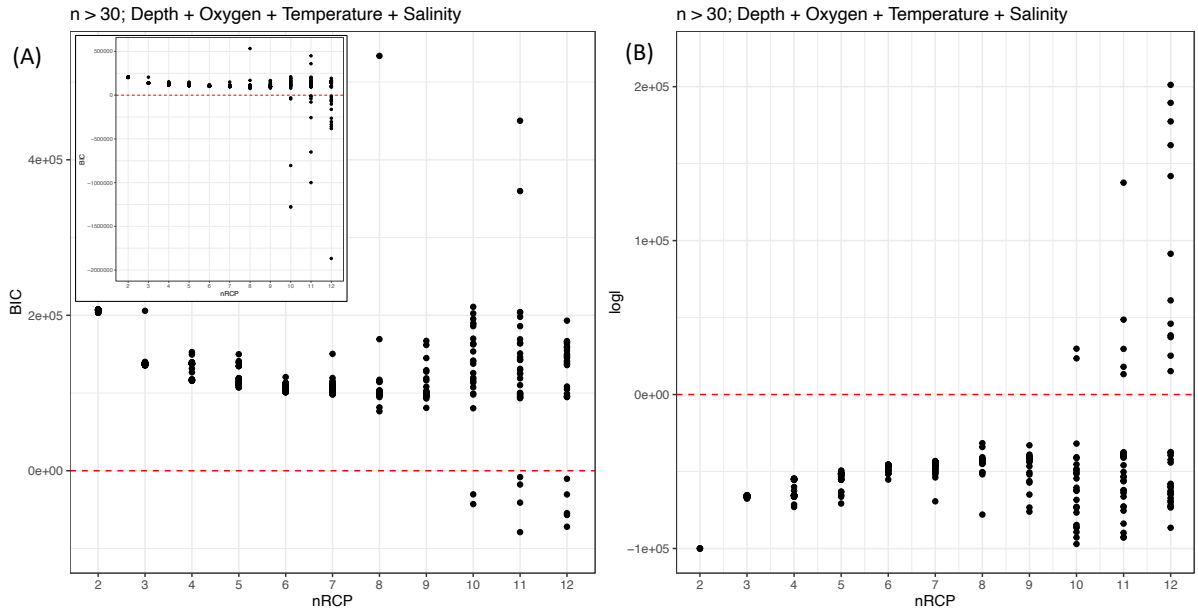

Figure 25. Selection of "best" number of RCPs based on (A) the Bayesian Information Criteria (BIC) and (B) maximum log likelihood. We truncated the y-axis in (A) to display better the BIC values and the characteristic exponential decay curve they should follow. We provide as an inset the full spectrum of models, where RCPs ≥ 10 are misfits given their large spanning positive and negative BIC values. Similarly, models for RCPs ≥ 10 exhibit log-likelihoods greater than zero also indicating a misfit of the models (B).

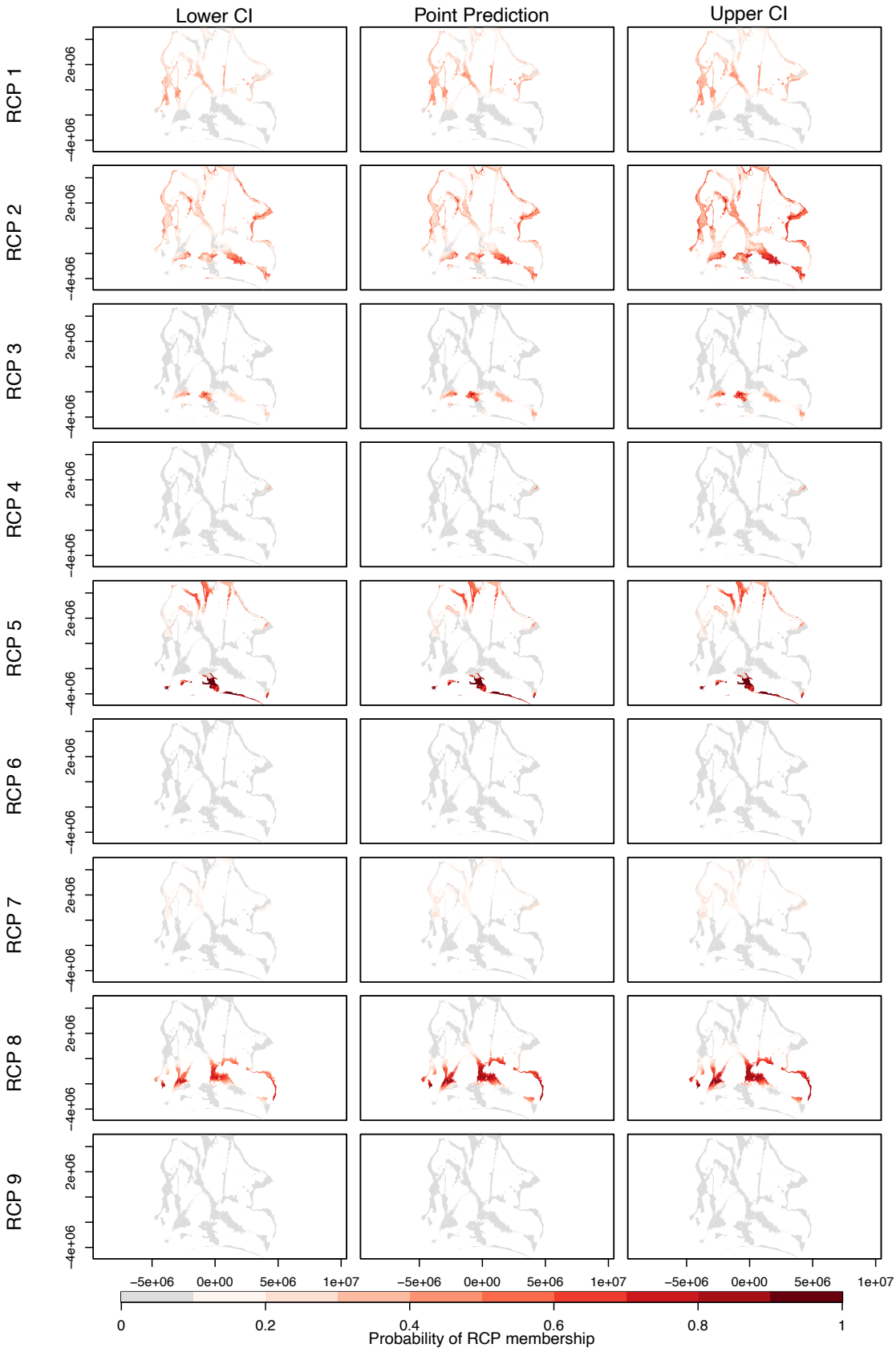

Figure 26. Distribution of seven predicted RCPs across the study extent (20°N-65°S, 20°-147°E). RCP6 and RCP9 represent zero group observations. White areas are not included in the model; grey areas represent the maximum depth for prediction (depths ≤ 3,500 m). Colours represents the probability distribution of the seven PPM-RCPs. Lower and upper confidence intervals (95%) are also provided.

RCP1 was predicted at latitudes north of 35°S with probabilities lower than 0.5. The area with highest membership probability was a small coastal region in the African coast from 15°-20°S. Other predicted areas with lower probabilities were the Mozambique Plateau, Madagascar Plateau up to 30°S, the shallowest depths of the Mid-Indian Ridge, Ninety-east Ridge from approximately 10°S up to the start of Broken Ridge, and Australia from 15°- 30°S. Depths spanned from the continental shelf in the Mozambique Channel up to 3,000 m, but always deeper than 1,500 m.

RCP2 showed some contrasting latitudinal patterns. It was predicted across the entire study extent, although basically absent from a narrow latitudinal band from 35°-45°S. These patterns are probably a result of contrasting environmental conditions or a negative interaction of the chosen preliminary variables for the model at the predicted depths. Specifically, RCP2 was predicted in the shallow depths of the Mascarene Plateau, the continental shelf of northeast Africa, the shallow depths of the Chagos Archipelago and the Chagos-Laccadive Plateau, the Indonesian coast, the continental mid-slope of Australia, south of the Southwest Indian Ocean Ridge, the northern edge of the Kerguelen Plateau, and the Southeast Indian Ridge. While at latitudes north of 35°S the predictions were at depths < 1,500 m, at latitudes south of 35°S predictions were at depths > 2,000 m.

RCP3 was predicted at latitudes from 45°-50°S approximately, in the Conrad Rise, especially around the Crozet Islands, and in the north of the Kerguelen Plateau. Depths spanned up to 1,500 m.

RCP4 had a geographic prediction relatively restricted to the northwest mid continental slope of Australia from 10°-20°S at approximately 1,500 m. It was also predicted along the Java coast and in a very small area of the Saya de Malha Bank.

RCP5 was predicted in the northern Indian Ocean, at latitudes north of 10°S. The highest membership probability was found in the Carlsberg Ridge, the deepest areas of the Chagos Laccadive Plateau, and northern Ninety East Ridge. Depths spanned from 1,700 to 2,500 m approximately.

RCP7 had a very low probability of membership from 0.2 to 0.3. On the western Indian Ocean RCP7 was predicted up to 25°S in areas deeper than 3,000 m. On the eastern side it was predicted north of 15°S, where it encompassed shallow depths in the northwest Australian shelf.

RCP8 was strongly predicted at latitudes 30°-50°S in the Agulhas Plateau, Southwest Indian Ocean Ridge, Madagascar Plateau (Walter's Shoal), Southeast Indian Ridge, south of Broken Ridge, and along the south coast of Australia. RCP8 encompassed depths from 800 to 2,000 m mainly, with some predictions in depths up to 2,500 m.

Finally, RCP6 and RCP9 were zero groups. These "zero" predictions resulted from the large number of "absences" (i.e., grid cells with zero observations) in the model emerging as "absence predictions". These need to be removed from the set of predicted RCPs.

4. Discussion

The biodiversity knowledge of VME indicator taxa in the Southern Indian Ocean was found to be both limited and spatially aggregated. The taxonomic resolution of biological data was at generic levels for areas of interest, specifically in SIOFA's management area, which reduced the dataset for comprehensive biogeographical analyses. Indeed, the majority of resolved data at species level was available from territorial waters. The data shortfalls make it impossible to identify comprehensive biogeographical regions based on observed data and map them at a resolution relevant for conservation decisions. As such, data shortfalls will also render high uncertainty in the individual mapping of the identified 1,500 taxa, even with state-of-the-art predictive techniques. Nevertheless, given the deficient sampling effort in areas beyond national jurisdiction, predictive approaches will be instrumental to inform and guide future conservation planning procedures in the Southern Indian Ocean.

Our work provides quantitative and fine-scale mapping of VME indicator taxa bioregions in the Southern Indian Ocean. Defining and mapping the extent of biological assemblages, or bioregions, is essential for the development of conservation strategies. These bioregions can provide a deeper understanding of the species composition and their environmental drivers as well as reveal broader biogeographic patterns. Despite the data-poor situation (in quantity and quality), we identified three main and eight nested bioregions across the Southern Indian Ocean with a "group first, then predict" approach; six, seven and eight bioregions at family, genus, and species level, respectively, with a "predict first, then group" approach; and preliminary seven bioregions with a "analyse simultaneously" approach. Together, the resulting predictive maps and associated uncertainty of the models' predictions are complementary in the information they convey. In the following sections we discuss biogeographic information and insights that the three predictive approaches undertaken in the consultancy give with respect to SIOFA's management area.

"Group first, then predict" bioregions

We identified two biogeographical levels through the "group first, then predict" approach that explain the distribution of VME indicator biodiversity through two angles. First, they illustrate differences in habitat complexity in the deep sea and species physiological adaptations with depth. Second, they show that there is spatial nestedness in the structure of VME biodiversity within the SIOFA area, with smaller partitions of biodiversity nested within larger ones. These nested patterns can be explained throughout drivers of biodiversity in the deep sea, water masses and topography, that influence larval dispersal. Although these maps are accurate in representing biogeographic patterns of VME indicator assemblages, we acknowledge they may be of limited use for applied management decisions at a scale relevant for fisheries given their coarse resolution (1° latitude-longitude grid).

First level bioregions

The three biogeographical regions observed in the Southern Indian Ocean depict the distribution of fauna across two distinct environments, the bathyal and abyssal depth zones, and a specific Southern Ocean bioregion. The bathyal zone (200 – 3,500 m) reveals a complex

topography represented by the continental slope, continental rise, and canyons near continents, and seamounts and mid-ocean ridges further offshore. These varying slopes and features enhance the heterogeneity and availability of habitats in the deep sea (Ramirez-Llodra et al., 2010). In contrast, the abyssal seafloor is generally uniform with level muddy bottoms and low-speed currents (Carney, 2005a; Mortensen et al., 2009). Thus, changes in topography coupled with physiological adaptations to tolerate increasing hydrostatic pressure at depth result in distinct faunas from bathyal and abyssal depths, as captured in our analysis.

Existing global biogeographical classifications for the Southern Indian Ocean reflect these depth zones. The marine realms (based on species endemicity; Costello et al., 2017) resemble the distribution of our observed biogeographical regions at the first hierarchical level, although the extent of the central Indian realm in Costello et al.'s (2017) 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. The GOODS 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. Here, our first level bioregions support the GOODS classification and further improves it with the proposal of eight nested biogeographical regions (discussed in the following section).

Second level bioregions

There were eight nested biogeographical regions with distinctive geographic and bathymetric characteristics reflecting spatially restricted water masses and topography of the Southern Indian Ocean. Changes in the physical, chemical, and biological properties of water masses commonly drive trends in the distribution of biodiversity (e.g., Buhl-Mortensen et al., 2015; Carney, 2005; Levin et al., 2001; Watling et al., 2013). Adaptation to environmental properties of specific water mass properties can denote the influence of niche specialisation in species distributions. In addition, major topographic features are known to influence beta diversity patterns along depth gradients (McClain & Rex, 2015). Depending on taxa and depth of topographic features, these may act as potential barriers or as stepping stones to dispersal by restricting the flow of currents and/or facilitating habitat for dispersing larvae across vast distances (McClain & Hardy, 2010). Here we will discuss the detected eight nested bioregions in relation to water masses and topographic features of the Southern Indian Ocean.

The large, predicted extent of Cluster 1.1 suggests strong connectivity of the surface and upper ocean layers throughout the cluster's latitudinal range, sustained by the complex interplay of currents in the northern Indian Ocean. The strong connectivity seems to be explained by the flow of the Indian Throughout Flow (ITF). The ITF is a unidirectional westward extension of the Pacific Warm Pool centered at 12°S, marking the division between the tropics (0°-23.5°) and subtropics (20°-40°) (Talley et al., 2011). At latitudes north of 12°S, there is seasonal reversal of currents by monsoon winds that produce switching from up to downwelling, modifications of concentration of oxygen, nutrients, and phytoplankton species composition (Hood et al., 2017). The depth influence, salinity, and temperature of the

currents at play change with the monsoon. For example, the Somalia and East African Coastal Current (EACC) has its maximum at 500 m during the Southwest Moon (summer monsoon) and 150 m during the northeast monsoon (winter monsoon) (Hood et al., 2017). The ITF integrates with the South Equatorial Current (SEC), which creates a boundary at 10°-15°S. The SEC reaches Madagascar and it splits into a northern (North East Madagascar Current; NEMC) and southern (South East Madagascar Current; SEMC) component. The northern flow (NEMC) further splits into a portion that flows into the Mozambique Channel, which will eventually join the Agulhas Current (AC), and another portion that continues north joining the EACC (Talley et al., 2011). This flow path appears to be well reflected by the predicted subregion 1.1.

Based on Decapoda, Ophiuroidea and Polychaeta species, Woolley et al. (2013) found two main regions along the northwest shelf of Australia, one mid-slope assemblage inhabiting depths between 150 – 850 m, from latitudes 11° S –22° S, and an outer-shelf assemblage at latitudes 15° S – 27° S and depths from 50 to 150 m. The latter is suggested to coincide with a region from Kalbarri (27° S) to the Broome (15° S) that may extend to the Timor Sea (11° S), supporting the prediction of Cluster 1.1 along the entire shelf. The former might coincide with the mid-slope assemblage observed in our predictions as Cluster 1.2. Although our prediction shows no contiguity, the observed bioregions showed a band of Cluster 1.2. Cluster 1.2 continues south to latitude 35° S as a mid-slope assemblage, while Cluster 1.1 continues also to the same latitude as an outer-shelf assemblage (Figure 4B). At the same latitudes, Woolley et al. (2013) found two different outer-shelf (from 50-250 m at latitudes 30° to 35.6° S, that probably extended too along 22° to 30° S deeper at depths from 250-450 m) and mid-slope (450-850 m at 23°S, depths from 350-1200 m at 36°S) assemblages. Differences with our predictions might be due to taxa-specific biogeographies or the lack of defined subregions at those latitudes in our analysis (Figure 1).

The eastward distribution of Cluster 1.2 following mid-ocean ridges seems to be driven by the presence of some pixels in these ridges. The meridional distribution can indeed be explained for some taxa. O'Hara et al. (2011) found ophiuroid species known from southwestern Australia, Tasmania, and the Southwest Pacific at similar depths around the Amsterdam and St. Paul islands. O'Hara et al. (2011) suggested that fauna extend into latitudinal bands, where the change between tropical (shelf/bathyal) and temperate (shelf/bathyal faunas) in the Indian Ocean would be around 30° S. Our predictions for Cluster 1.2 seem to follow the flow of the Subtropical Gyre (Talley et al., 2011), supporting these widespread meridional ranges (and therefore dispersal) within their latitudinal bands and depth bands. Further, the Leewin Current explains some of the pixels in the observed map (Figure 4B) that were not predicted in the final map (Figure 14). The Leewin Current follows the continental break within 100 km from the coast from 22° to 35° S. Then turns east where it continues along the shelf break as the Leewin Current Extension and then the South Australian Current, bringing warm, saline waters as a result of the inflow of the Subtropical Gyre. Finally, it veers between the south of Australia and Tasmania as a boundary current known as Zeehan Current, where we observed some Cluster 1.2 pixels.

The area covered by Cluster 1.3 is influenced by the Agulhas Current, which reaches the seafloor at 33° S. The Agulhas Current is a poleward current that makes favourable upwelling and that carries saline and oxygenated North Atlantic Deep Water (NADW) waters (Talley et

al., 2011), all of which was observed in the description of our observed clusters (Figure 7). The prediction of some sites of Cluster 1.3 in the Mozambique Channel and in the Somalian coast could be explained by the currents feeding and originating the Agulhas Current (explained above) and the northward flow of the NADW at intermediate depths (1,500-2,000 m) (Charles et al., 2020; van Aken et al., 2004). Besides, the eastward prediction of Cluster 1.3 up to Cape Naturalist in Australia, was predicted poorly. However, it is known that the Agulhas Retroflection Current (ARC) reaches longitudes of 50° E with unreduced flow (Pollard & Read, 2017), and it extends eastwards as a narrow band, which follows the Subtropical Front (South Indian Current) also called the Agulhas Return Current (Talley et al., 2011). Yet, this could be an overprediction of the cluster.

Cluster 1.5 was a relatively shallow, geographically restricted cluster in the south of Australia. The predicted extent of Cluster 1.5 across Australia's southern coast is supported by findings from Tanner et al. (2018) who, using historical museum benthic records, showed similarity across different assemblages in the southern Australia. In our observed bioregions, we found some presences of the cluster in western Australia. These were shelf assemblages (as shown in Figure 4B), which corresponds well with results by Tanner et al. (2018) and O'Hara (2008) for the area. These shallow assemblages are connected through the South Australian Current, which is a continuation of the Leewin Current, bringing warm saline waters into the region (Talley et al., 2011).

In contrast, Cluster 1.7 was restricted to bathyal depths (Figure 4B), a bioregion observed for ophiuroid data too (O'Hara et al., 2011). Connectivity at these depths might be explained by the Flinders Current, supporting our predictions (Figure 14). The Flinders Current is an eastward/equatorward current flowing at depths from 300 - 800 m with its maximum at 400 m. The Flinders Current flows from Tasmania up to Cape Leewin, where it turns right following Australia as the Leewin Undercurrent and continuing eastward and northwest. This current interacts with the Leewin Current at the shelf break and slope (Woo et al., 2007). The Flinders Current has sources of Subantarctic Mode Water. The prediction of Cluster 1.7 on the western side of the Indian Ocean coincides with another area of Mode Water, the Indian Ocean Subtropical Mode Water. Thus, prediction of Cluster 1.7 seems to be driven by environmental conditions rather than observed bioregions, as no pixel was observed on the western area.

The prediction of cluster 2.1 and cluster 2.4 are driven by the presence of two topographic features altering the eastward flow circulation of Antarctic water masses and the position of the major Southern Ocean fronts. On one hand, the Kerguelen Plateau obstructs the movement of Antarctic Circumpolar Current (ACC), affecting the regional circulation and creating complex patterns. In particular, the Polar Front is found on the southern flank of the Kerguelen Island shoal at 50° 35'S rounding the islands from the south and east (Park & Vivier, 2011), which could be observed in our predictions (see Figure 14). The northern limit of Cluster 2.1 is demarcated by the Subantarctic Front (SAF) sitting at 45° S tightly merged with the Subtropical Front (STF). This latitude had very uncertain predictions, suggesting a transition zone and lack of sufficient sample data.

On the other hand, Cluster 2.4 was observed over Conrad Rise, another feature strongly affecting the eastward flow of Antarctic water masses at all depths fragmenting the Antarctic Polar Front along its northern and southern extremities (Durgadoo et al., 2008). The southern

latitudinal limit prediction of Cluster 2.4 coincides with the Southern ACC Front (SACCF), which in turn coincides with the Fawn Trough Current in the Kerguelen Plateau (Park et al., 2009). The SAF and SACCF present pronounced properties in salinity and density, which have a clear imprint in our predictions.

Finally, Cluster 3.1 represents the main ocean basins of the Indian Ocean. All the basins are connected, allowing the circulation of the deep and bottom waters of the Indian Ocean. At those depths, Circumpolar Deep Water (CDW) enters the Indian Ocean from the south. In the west, NADW enters from the South Atlantic. Indian Deep Water (IDW) is formed by diffusion and upwelling, it is low in oxygen and high in nutrients (see silicate, Figure 7), reflecting its high age as it advects back to Pacific. Therefore, at any depth in the deep Indian Ocean, we find both CDW and IDW, where differences among both are regional (Talley et al., 2011). These differences are not observed in our predictions, and it explains the homogeneity of Cluster 3.1 across the whole Indian Ocean.

"Predict first, then group" bioregions

This second approach, based on taxa distribution models, complements the first modelling approach in two ways. 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, which in turn adds value for management applications. However, it is important to note 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-bioregionalisation map, Figure 20). Thus, some species and bioregions are overpredicted 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.

One of the main findings of this approach was that, for all taxonomic levels, we found a large first cluster composed of the majority of taxa, and multiple additional clusters generally driven by one or a few taxa. This consistent pattern across taxonomic levels can be explained by several factors. First, upper bathyal (and upper lower bathyal) and shelf faunas share similar complex environments. This cluster parallels the distribution of the inshore cluster detected at the first hierarchical level in the "group first, then predict" approach. As mentioned earlier, at bathyal depths there is greater topographic complexity that translates into greater heterogeneity and availability of habitats. In addition, our models do not account for dispersal limitations or biotic interactions, which, in turn, will result in potential distributions largely overpredicting the true distributions of taxa, as opposed to the geographical and bathymetric nestedness observed in the "group first, then predict" approach. These three characteristics, greater habitat complexity of the bathyal zone and omission of both dispersal limitations and biotic interactions in the models lead to a predicted continuum of overlapping distributions which makes it impossible to detect distinct clusters for the upper bathyal to shelf areas. Second, the severe lack of distributional data for most taxa has important impacts on the predictive models for each taxon. In a nutshell, because of

the limited number of occurrences available for each taxon, we had to limit the number of predictors in our taxa distribution models in order to reduce the risks of overfitting. However, this limitation in the number of predictors implies that our models are coarse approximations of the niches of taxa, which, in turn, is likely to result in overpredictions of their distributions. Last, sampling biases towards coastal areas in the study area can lead to identifying inadequate environmental predictors during the variable selection process. In the next paragraphs we elaborate further on these three hypotheses.

The broad latitudinal and longitudinal extent of our TDM-based bioregions could be explained by broad and overlapping tolerances of species to environmental conditions (Woolley et al., 2013). O'Hara et al. (2011) found that ophiuroid taxa from the upper continental slope was more similar to that of the continental shelf, and in other occasions, lower-bathyal taxa patterns aligned more with those of the abyssal (O'Hara et al., 2020). A gradual depth boundary combined with the fact that the bioregionalisation was based on an "arbitrary" binary transformation of the species "presence" would lend support to the extensive distribution of Cluster 1 across the three taxonomic predictions. In addition, Cluster 1 could represent all bathyal habitats that without further species information and finer environmental predictors (such as substrate type) cannot be disentangled into finer biogeographic units. Nevertheless, that the taxa composition of Cluster 1 was composed of practically the totality of families, genera, or species, respectively, whereas the remainder of the bioregions were characterised by one type of taxa, raises the question whether a stricter binary transformation should be implemented. However, the selection of binary thresholds always entails these difficulties.

The bioregions connected areas with similar environmental conditions over large distances despite the fact that they may not share the same taxa (in contrast, see the observed bioregions). This is because our models do not account for species dispersal abilities and biotic interactions, which are two factors that shape a species range. Such a limitation is due to the underlying hypotheses of correlative niche models. To account for both aspects, additional modelling approaches would be necessary, but these are unlikely to be realistic for the target taxa for two reasons. First, implementing biotic interactions in distribution models is still unresolved in the literature, and, when successfully implemented, is generally based on *a priori* knowledge on the interactions which does not exist for the target VME taxa. Second, implementing dispersal models would require additional knowledge on the dispersal mechanisms for the target taxa, which, again, requires a level of information that currently does not exist for VME taxa. Hence, these bioregions cannot be interpreted with relation to water masses in the Indian Ocean. Possible future avenues to circumvent the impossibility of accounting directly for dispersal limitations and biotic interactions would be to include *a priori* information based on the results of the observed bioregions. For example, it could be possible to investigate in the future different bioregionalisations of predictive taxa distribution models based on sets of species retained in the generation of the observed bioregions.

Finally, an important limitation of this approach inherent to modelling taxa with low number occurrences is that it can result in a less precise estimation of taxa niches due to the number of environmental variables used to calibrate the model. Thus, a less precise estimation of the niche may result in overprediction of taxa distribution ranges, which, in turn, results in overlapping distribution ranges creating large bioregions in space. The relatively low number

of occurrences per taxa allowed the inclusion of three variables at best. The resulting predicted distributions are likely overpredicted given that the environmental niche is not being well characterised. In particular, this was observed in the species and genus maps as large areas predicted that were not predicted at family level (Figure 14, 17, 20). Thus, with more occurrences more explanatory variables can be incorporated in the calibration of the model without overfitting the data. Currently, we estimate that the genus-level bioregionalisation might be the best compromise between a fine taxonomic resolution and a reduced (albeit still very present) overprediction of ranges due to the lack of data. However, this observation is speculative, and, because of the above discussion, we deem that all bioregionalisation maps in this "predict first, then group" approach are subject to a high degree of uncertainty. We urge the need for more data in the SIOFA area.

"Analyse simultaneously" bioregions

The PPM-RCP approach allowed to group and predict, in a single step, bioregions based on 155 VME indicator species, suggesting seven groups at depths shallower than 3,500 m. These results are preliminary and require careful interpretation. Here we provided progress on the development of this approach for SIOFA and discuss results in relation to oceanographic patterns in the Indian Ocean. Mixture-of-experts models are difficult to implement, although advances with user-friendly packages are rapidly developing to facilitate their use. In addition, these methods are computationally intensive. These results will be subject to reevaluation during the next consultancy.

The preliminary seven RCPs have similarities to the "group first, then predict" bioregions, although there are differences rising from the selected environmental variables, which will require adjustment for a final RCP model. RCP1 resembles the distribution of bioregion 1.2, although on the Western Indian Ocean it combines with bioregion 1.1 in the Madagascar Plateau and the Mozambique Plateau. RCP1 seems to follow the westward flow of the South Equatorial Current (SEC) and the eastward flow of the East Madagascar Current (EMC) when it reaches the southern tip of Madagascar. This branch then feeds into the Leeuwin Current at the Australian coast (Talley et al., 2011), which explains the prediction on western Australia.

RCP2 displayed some contrasting ocean wide predictions. North of 30°S it strongly resembled bioregion 1.1, where upper bathyal areas were the dominant depth zone. South of 30°S, it was strongly predicted along a latitudinal belt at 45°S. This prediction followed closely the Subantarctic Front (SAF) flow around Del Caño Rise and the Crozet Plateau at 2,000 m water depth (Pollard et al., 2007).

RCP3 was a geographically restricted bioregion predicted in the northern part of the Kerguelen Plateau and the Crozet Plateau. RCP3, which mirrors the distribution of bioregion 2.1, is driven by the position of the major Southern Ocean fronts. As previously explained, the Polar Front is found on the southern flank of the Kerguelen Island shoal at 50° 35'S rounding the islands from the south and east (Park & Vivier, 2011), limiting RCP3 to the SAF in the north (also for the Crozet Plateau) and the Polar Front in its southernmost prediction.

RCP4 was restricted to the northwest shelf of Australia. The Australian coast had the highest sampling effort (Figure 3), which probably allows for a greater refinement of biogeographical regions. This RCP had a similar distribution to a bioregion found by Woolley et al. (2013) based on Decapoda, Ophiuroidea and Polychaeta. Woolley et al. (2013) described this region to extend from 116°-124°E, down to 20°S, at depths between 150-850 m. Comparing it with environmental drivers, Woolley et al. (2013) describe the probability of occurrence of this bioregion in an envelope of low oxygen.

RCP5 exhibited contrasting predicting patterns, with a wide latitudinal gap between predicted areas. The northernmost prediction (20°N to 10°S) was distributed in the deepest areas around RCP2, and it was also similar to some parts of the predicted distribution of bioregion 1.1. This latitudinal belt coincides with an area of warmest sea surface temperature in the Indian Ocean (Talley et al., 2011). Here, confined to north of the South Equatorial Current (SEC) front at 10°S, there is a high salinity layer. The deeper part of this high salinity layer is referred to as Red Sea Overflow Water, which has its core at 2,000 m at 95°E. Further at 10°S, the upper ocean water mass the Indonesian Throughflow Water marks the limit between the tropics and subtropics. At intermediate depths, it becomes the Indonesian Intermediate Water, distinct by a salinity minimum in the north-south direction, and also vertically (Talley et al., 2011). On the other hand, the southernmost prediction of RCP5 in the south of the Kerguelen Plateau and Conrad Rise, coincides with the northernmost limit of the Polar Front. It is likely that the prediction picked on the salinity minima despite contrasting latitudes.

RCP7 was faintly predicted with a similar distribution to the northern predicted latitudes of RCP5 and RCP1. In some areas, like in the northwest shelf of Australia, the prediction was centered on the continental shelf, while in the Chagos-Laccadive Rdige was predicted at the deepest depths.

Finally, RCP8 had a strong prediction in the latitudinal band from 30°S to 45°S. RCP8 prediction coincided with the predicted distributions for bioregions 1.2, 1.3, 1.5, and 1.7 in the “group first, then predict” approach. For the latter, it coincided with an area of low model confidence. While the southern limit of RCP8 coincided with the SAF, the longitudinal spread of the RCP could be explained by the complex interplay of the Agulhas Retroflection Current (ARC), the Subtropical Front (STF) and north of the SAF at those latitudes. On one hand, it is known that the Agulhas Retroflection Current (ARC) reaches longitudes of 50° E with unreduced flow (Pollard & Read, 2017), and it extends eastwards as a narrow band, which follows the Subtropical Front (South Indian Current) also called the Agulhas Return Current (Talley et al., 2011). On the other hand, the SAF sits at 45°S marking the southern limit distribution of RCP8. The depth range of RCP8 spans all the available range and, at least on the Southwest Indian Ocean Ridge (SWIR), there might be connectivity of water masses at depth as suggested by the similarities in the coral species found at Coral Seamount, Middle of What seamount, Melville seamount and Atlantis Bank (Pratt et al., 2019). Pratt et al. (2019) further noted within-species variation for some coral species sampled at those locations as a possible response to environmental conditions such as temperature and/or food, supporting the notion of the SWIR as a transitional zone. RCP8 indeed showed less partition of the environment than the “group first, then predict” predictions for the SWIR, which is in line with faunal groups defined by transitions rather than abrupt geographical breaks (Woolley et al., 2020). It is however worth noting that video footage from these surveys suggested the

greatest diversity of habitats and species richness for corals and sponges at Coral Seamount (Frinault, 2017; Rogers & Taylor, 2011), which is located just south of the STF at 45°S.

The importance of developing one-stage approaches such as RCPs is threefold. First, they offer repeatability because the mathematical model defines formally bioregions and the relationship with the environment (Hill et al., 2020; Warton et al., 2015). Second, currently, they are the only approach able to quantify uncertainty in the final bioregionalisation map given the probabilistic nature of the models (i.e., a site has some chance of belonging to more than one bioregion). Furthermore, 95% confidence intervals of the predictions can be estimated allowing to have a more conservative or a more optimistic scenario of the predictions. These appropriate measures of uncertainty are essential for spatial management applications to understand the risks associated with applying or not a certain measure in a location (Hill et al., 2020). Third, the simultaneous grouping and prediction for taxa has the potential of integrating dispersal limitation in the prediction. As noted, the RCP bioregions had similarities with the "group first, then predict" approach rather than with the "predict first, then group" approach. Additionally, this approach provides less partitioning of the environment by taking into account species tolerance to environmental conditions. For instance, this was observed in the prediction of RCP8 in the Southwest Indian Ocean Ridge, which represented one bioregion as opposed to the partitioning of the "group first, then predict" approach at the same location. Nevertheless, it is worth reminding that any prediction will require model validation and further sampling given the limited number of samples in areas beyond national jurisdiction of the study area.

5. Evaluation of the applicability of the different bioregionalisation schemes to the conservation of SIOFA VMEs

In this work, despite the scarcity in data, we managed to identify, predict, and map a number of biogeographical regions with several predictive approaches. In total, we provided here three types of classifications of the SIOFA area into key regions for VME indicator taxa. Our first classification was based on the "group first, then predict" approach; the second classification was based on the "predict first, then group" approach"; and the third was based on the "analyse simultaneously" approach. The ultimate purpose of these different classifications will be to support management decisions on the implementation of benthic protected areas. As we have discussed extensively, the results differ between the different approaches, and especially between bioregions of the "group first, then predict" approach and the bioregions of the "predict first, then group" approach.

The "group first, then predict" approach suggests several bioregions and subregions whose distribution could be mainly explained by depth and water masses in the Indian Ocean. Because these bioregions are based on the known patterns of species co-occurrence in the area, they account for the known limits of species dispersal. The resulting bioregions can be interpreted as known bioregions to occur in the area based on observed samples and, as such, they provide an approximation of the expected number of bioregions to occur in the Southern

Indian Ocean. Consequently, we have confidence that the bioregions and sub-bioregions delineated in this first approach reflect actual patterns of geographical distinction in the distribution of VME indicator taxa. In addition, the fact that the preliminary RCPs resembled the distribution of bioregions from this group, also lends support to our results., although it will be necessary to complete the modelling of the RCPs to compare objectively results from these two approaches. Therefore, we recommend that conservation efforts in the SIOFA area should represent all the delineated bioregions and sub-bioregions in the "group first, then predict" approach. Nonetheless, there were two main limits to these results. First, the resolution at which we delineated bioregions and predicted them spatially is coarse (1° latitude-longitude). Second, there were large areas of uncertainty in our predictions, i.e., areas where none of our models were able to predict the suitability for any of the bioregions. Consequently, the inclusion of the final RCPs will inform future steps on how best to capitalise on the results and circumvent these limits to inform conservation plans for SIOFA VME taxa.

In our opinion, the "group first, then predict" results can be refined with results from the "predict first, then group" approach. Indeed, although the latter approach does not account for actual distribution patterns of VME indicator taxa resulting from dispersal limitations and biotic interactions, it provides a classification of the environment into distinct groups at a fine resolution. As such, this second set of methods can be interpreted as an illustration of the diversity of habitats for VME taxa, based on the available environmental in the study area. In that sense, it is complementary to the "group first, then predict" approach, and future steps will need to explore how to combine these different classifications to best inform SIOFA conservation plans. Finally, we insist on the very limited availability of data in the SIOFA area, which implies that interpretation must be exerted with caution, and we urge for the necessity of acquiring more data on VME indicator taxa in the SIOFA area.

References

Assis, J., Tyberghein, L., Bosch, S., Verbruggen, H., Serrão, E. A., & de Clerck, O. (2018). Bio-ORACLE v2.0: Extending marine data layers for bioclimatic modelling. *Global Ecology and Biogeography*, *27*(3), 277–284. https://doi.org/10.1111/GEB.12693

Barber, R. A., Ball, S. G., Morris, R. K. A., & Gilbert, F. (2022). Target-group backgrounds prove effective at correcting sampling bias in Maxent models. *Diversity and Distributions*, *28*(1), 128–141. https://doi.org/10.1111/DDI.13442

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi : An Open Source Software for Exploring and Manipulating Networks Visualization and Exploration of Large Graphs. In *Association for the Advancement of Artificial Intelligence*. Association for the Advancement of Artificial Intelligence. www.aaai.org

Bellard, C., Leroy, B., Thuiller, W., Rysman, J. F., & 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., & Encinas-Viso, F. (2018). A comparison of network and clustering methods to detect biogeographical regions. *Ecography*, *41*(1), 1–10. https://doi.org/10.1111/ECOG.02596

Buhl-Mortensen, L., Olafsdottir, S. H., Buhl-Mortensen, P., Burgos, J. M., & Ragnarsson, S. A. (2015). Distribution of nine cold-water coral species (Scleractinia and Gorgonacea) in the cold temperate North Atlantic: effects of bathymetry and hydrography. *Hydrobiologia*, *759*(1), 39–61. https://doi.org/10.1007/S10750-014-2116-X/FIGURES/9

Cairns, S. D. (n.d.). On line appendix: Phylogenetic list of the 711 valid Recent azooxanthellate scleractinian species with their junior synonyms and depth ranges. In *Cold-Water Corals: The Biology and Geology of Deep-Sea Coral Habitats.* (p. 28). Cambridge University Press. Retrieved April 4, 2022, from https://www.marinespecies.org/introduced/aphia.php?p=sourcedetails&id=142534

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), 311–322.

Carney, R. S. (2005). Zonation of Deep Biota on Continental Margins. In *Oceanography and Marine Biology* (1st ed., pp. 221–288). CRC Press. https://doi.org/10.1201/9781420037449-8

Charles, C., Pelleter, E., Révillon, S., Nonnotte, P., Jorry, S. J., & Kluska, J. M. (2020). Intermediate and deep ocean current circulation in the Mozambique Channel: New insights from ferromanganese crust Nd isotopes. *Marine Geology*, *430*, 106356. https://doi.org/10.1016/J.MARGEO.2020.106356

Chiu, C. H., Wang, Y. T., Walther, B. A., & Chao, A. (2014). An improved nonparametric lower bound of species richness via a modified good–turing frequency formula. *Biometrics*, *70*(3), 671–682. https://doi.org/10.1111/BIOM.12200

Costello, M. J., Tsai, P., Wong, P. S., Cheung, A. K. L., Basher, Z., & Chaudhary, C. (2017). Marine biogeographic realms and species endemicity. *Nature Communications*, *8*(1), 1–10. https://doi.org/10.1038/s41467-017-01121-2

Cressie, N. (1993). *Statistics for spatial data*. John Wiley & Sons.

Davies, A. J., & Guinotte, J. M. (2011). Global Habitat Suitability for Framework-Forming Cold-Water Corals. *PLOS ONE*, *6*(4), e18483. https://doi.org/10.1371/JOURNAL.PONE.0018483

Durgadoo, J. v., Lutjeharms, J. R. E., Biastoch, A., & Ansorge, I. J. (2008). The Conrad Rise as an obstruction to the Antarctic Circumpolar Current. *Geophysical Research Letters*, *35*(20), 20606. https://doi.org/10.1029/2008GL035382

Ebach, M. C., & Parenti, L. R. (2015). The dichotomy of the modern bioregionalization revival. *Journal of Biogeography*, *42*(10), 1801–1808. https://doi.org/10.1111/JBI.12558

Edler, D., Guedes, T., Zizka, A., Rosvall, M., & Antonelli, A. (2017). Infomap Bioregions: Interactive Mapping of Biogeographical Regions from Species Distributions. *Systematic Biology*, *66*(2), 197–204. https://doi.org/10.1093/SYSBIO/SYW087

Elith, J., Ferrier, S., Huettmann, F., & Leathwick, J. (2005). The evaluation strip: A new and robust method for plotting predicted responses from species distribution models. *Ecological Modelling*, *186*(3), 280–289. https://doi.org/10.1016/J.ECOLMODEL.2004.12.007

Foster, S. D., Givens, G. H., Dornan, G. J., Dunstan, P. K., & Darnell, R. (2013). Modelling biological regions from multi-species and environmental data. *Environmetrics*, *24*(7), 489–499. https://doi.org/10.1002/ENV.2245

Foster, S. D., Hill, N. A., & Lyons, M. (2017). Ecological grouping of survey sites when sampling artefacts are present. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *66*(5), 1031–1047. https://doi.org/10.1111/RSSC.12211

Frinault, B. A. v. (2017). *The Deep-Sea: Protecting Areas Beyond National Jurisdiction; Particular Focus on Communities of Southwest Indian Ocean Ridge Seamounts* [MSc]. University of Oxford.

Gotelli, N. J., & Colwell, R. K. (2011). Estimating Species Richness. In *Biological Diversity Frontiers in Measurement and Assessment* (pp. 39–54). Oxford University Press. https://www.scirp.org/(S(vtj3fa45qm1ean45vvffcz55))/reference/ReferencesPapers.aspx?ReferenceID=1358018

Guillera-Arroita, G., Lahoz-Monfort, J. J., Elith, J., Gordon, A., Kujala, H., Lentini, P. E., Mccarthy, M. A., Tingley, R., & Wintle, B. A. (2015). Is my species distribution model fit for purpose? Matching data and models to applications. *Global Ecology and Biogeography*, *24*(3), 276–292. https://doi.org/10.1111/GEB.12268/SUPPINFO

Hill, N. A., Foster, S. D., Duhamel, G., Welsford, D., Koubbi, P., & Johnson, C. R. (2017). Model-based mapping of assemblages for ecology and conservation management: A case study of demersal fish on the Kerguelen Plateau. *Diversity and Distributions*, *23*(10), 1216–1230. https://doi.org/10.1111/DDI.12613

Hill, N. A., Woolley, S. N. C., Foster, S. D., Dunstan, P. K., McKinlay, J., Ovaskainen, O., & Johnson, C. (2020). Determining marine bioregions: A comparison of quantitative approaches. *Methods in Ecology and Evolution*, *11*(10), 1258–1272. https://doi.org/10.1111/2041-210X.13447

Hood, R. R., Beckley, L. E., & Wiggert, J. D. (2017). Biogeochemical and ecological impacts of boundary currents in the Indian Ocean. *Progress in Oceanography*, *156*, 290–325. https://doi.org/10.1016/J.POCEAN.2017.04.011

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive Mixtures of Local Experts. *Neural Computation*, *3*(1), 79–87. https://doi.org/10.1162/NECO.1991.3.1.79

Jacomy, M., Venturini, T., Heymann, S., & Bastian, M. (2014). ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLOS ONE*, *9*(6), e98679. https://doi.org/10.1371/JOURNAL.PONE.0098679

Kocsis, Á. T., Reddin, C. J., & Kiessling, W. (2018). The stability of coastal benthic biogeography over the last 10 million years. *Global Ecology and Biogeography*, *27*(9), 1106–1120. https://doi.org/10.1111/GEB.12771

Kull, M., & Flach, P. (2022). *prg: Software to create Precision-Recall-Gain curves and calculate area under the curve* (0.5.1). https://github.com/meeliskull/prg

Leroy, B. (2012). *Utilisation des bases de données biodiversité pour la conservation des taxons d'invertébrés: indices de rareté des assemblages d'espèces et modèles de prédiction de répartition d'espèces* [Doctoral dissertation]. Muséum national d'histoire naturelle.

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., & Ysnel, F. (2014). Forecasted climate and land use changes, and protected areas: the contrasting case of spiders. *Diversity and Distributions*, *20*(6), 686–697. https://doi.org/10.1111/DDI.12191

Leroy, B., Delsol, R., Hugueny, B., Meynard, C. N., Barhoumi, C., Barbet-Massin, M., & 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., & Tedesco, P. A. (2019). Global biogeographical regions of freshwater fish species. *Journal of Biogeography*, *46*(11), 2407–2419. https://doi.org/10.1111/JBI.13674

Levin, L. A., Etter, R. J., Rex, M. A., Gooday, A. J., Smith, C. R., Pineda, J., Stuart, C. T., Hessler, R. R., & Pawson, D. (2001). Environmental Influences on Regional Deep-Sea Species Diversity1. *Annual Review of Ecology and Systematics*, *32*, 51–93. https://doi.org/10.1146/ANNUREV.ECOLSYS.32.081501.114002

Lomolino, M. v., Riddle, B. R., & Whittaker, R. J. (2017). *Biogeography* (Fifth). Oxford University Press.

McClain, C. R., & Rex, M. A. (2015). Toward a Conceptual Understanding of β-Diversity in the Deep-Sea Benthos. *Annual Reviews of Ecology, Evolution, and Systematics*, *46*, 623–642. https://doi.org/10.1146/ANNUREV-ECOLSYS-120213-091640

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., Laffargue, P., Murillo, F. J., Sacau, M., Vaz, S., Kenchington, E., Arnaud-Haond, S., Callery, O., Chimienti, G., Cordes, E., … Carreiro-Silva, M. (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), 2181–2202. https://doi.org/10.1111/GCB.14996

Mortensen, P. B., Buhl-Mortensen, L., Dolan, M., Dannheim, J., & Kröger, K. (2009). Megafaunal diversity associated with marine landscapes of northern Norway: a preliminary assessment. *Norwegian Journal of Geology*, *89*, 163–171.

O'Hara, T. D. (2008). *Bioregionalisation of Australian waters using brittle stars (Echinodermata: Ophiuroidea), a major group of marine benthic invertebrates*. http://www.ag.gov.au/cca

O'Hara, T. D., Rowden, A. A., & Bax, N. J. (2011). A Southern Hemisphere Bathyal Fauna Is Distributed in Latitudinal Bands. *Current Biology*, *21*(3), 226–230. https://doi.org/10.1016/J.CUB.2011.01.002

O'Hara, T. D., Williams, A., Woolley, S. N. C., Nau, A. W., & Bax, N. J. (2020). Deep-sea temperate-tropical faunal transition across uniform environmental gradients. *Deep Sea Research Part I: Oceanographic Research Papers*, *161*, 103283. https://doi.org/10.1016/J.DSR.2020.103283

Park, Y.-H., & Vivier, F. (2011). Circulation and hydrography over the Kerguelen Plateau. In *The Kerguelen Plateau: marine ecosystems and fisheries* (Vol. 35, pp. 43–55). https://sfi-cybium.fr/en/node/2924

Park, Y.-H., Vivier, F., Roquet, F., Kestenare, E., Park, Y.-H., Vivier, F., Roquet, F., & Kestenare, E. (2009). Direct observations of the ACC transport across the Kerguelen Plateau. *Geophysical Research Letters*, *36*(18). https://doi.org/10.1029/2009GL039617

Pollard, R. T., & Read, J. F. (2017). Circulation, stratification and seamounts in the Southwest Indian Ocean. *Deep Sea Research Part II: Topical Studies in Oceanography*, *136*, 36–43. https://doi.org/10.1016/J.DSR2.2015.02.018

Pollard, R. T., Venables, H. J., Read, J. F., & Allen, J. T. (2007). Large-scale circulation around the Crozet Plateau controls an annual phytoplankton bloom in the Crozet Basin. *Deep Sea Research Part II: Topical Studies in Oceanography*, *54*(18–20), 1915–1929. https://doi.org/10.1016/J.DSR2.2007.06.012

Pratt, N., Chen, T., Li, T., Wilson, D. J., van de Flierdt, T., Little, S. H., Taylor, M. L., Robinson, L. F., Rogers, A. D., & Santodomingo, N. (2019). Temporal distribution and diversity of cold-water corals in the southwest Indian Ocean over the past 25,000 years. *Deep Sea Research Part I: Oceanographic Research Papers*, *149*, 103049. https://doi.org/10.1016/J.DSR.2019.05.009

Ramirez-Llodra, E., Brandt, A., Danovaro, R., de Mol, B., Escobar, E., German, C. R., Levin, L. A., Martinez Arbizu, P., Menot, L., Buhl-Mortensen, P., Narayanaswamy, B. E., Smith, C. R., Tittensor, D. P., Tyler, P. A., Vanreusel, A., & Vecchione, M. (2010). Deep, diverse and definitely different: Unique attributes of the world's largest ecosystem. *Biogeosciences*, *7*(9), 2851–2899. https://doi.org/10.5194/BG-7-2851-2010

Rice, J., Gjerde, K. M., Ardron, J., Arico, S., Cresswell, I., Escobar, E., Grant, S., & Vierros, M. (2011). Policy relevance of biogeographic classification for conservation and management of marine biodiversity beyond national jurisdiction, and the GOODS biogeographic classification. *Ocean & Coastal Management*, *54*(2), 110–122. https://doi.org/10.1016/J.OCECOAMAN.2010.10.010

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., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*(8), 913–929. https://doi.org/10.1111/ECOG.02881

Rogers, A. D., & Taylor, M. L. (2011). *Benthic Biodiversity of Seamounts in the Southwest Indian Ocean: Cruise Report - R/V James Cook 066 Southwest Indian Ocean Seamounts Expedition*.

Rojas, A., Patarroyo, P., Mao, L., Bengtson, P., & Kowalewski, M. (2017). Global biogeography of Albian ammonoids: A network-based approach. *Geology*, *45*(7), 659–662. https://doi.org/10.1130/G38944.1

Rosvall, M., & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(4), 1118–1123. https://doi.org/10.1073/PNAS.0706851105

Smith, A. B. (2020). *enmSdm: Faster, better, smarter ecological niche modeling* (5.3.0). https://github.com/adamlilith/enmSdm

Talley, L. D., Pickard, G. L., Emery, W. J., & Swift, J. H. (2011). Indian Ocean. In *Descriptive Physical Oceanography: An Introduction* (6th ed., pp. 1–555). Elsevier Ltd. https://doi.org/10.1016/C2009-0-24322-4

Tanner, J. E., Althaus, F., Sorokin, S. J., & Williams, A. (2018). Benthic biogeographic patterns in the southern Australian deep sea: Do historical museum records accord with recent systematic, but spatially limited, survey data? *Ecology and Evolution*, *8*(23), 11423–11433. https://doi.org/10.1002/ECE3.4565

Thuiller, W., Georges, D., Gueguen, M., Engler, R., & Breiner, F. (2020). *"biomod2": Ensemble Platform for Species Distribution Modeling* (3.5.0). https://r-forge.r-project.org/forum/forum.php?eforum_id=995&group_id=302

Tyberghein, L., Verbruggen, H., Pauly, K., Troupin, C., Mineur, F., & de Clerck, O. (2012). Bio-ORACLE: a global environmental dataset for marine species distribution modelling. *Global Ecology and Biogeography*, *21*(2), 272–281. https://doi.org/10.1111/J.1466-8238.2011.00656.X

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, UNGA A/RES/61/105 21 (2007).

Valavi, R., Elith, J., Lahoz-Monfort, J. J., & Guillera-Arroita, G. (2019). blockCV: An r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *Methods in Ecology and Evolution*, *10*(2), 225–232. https://doi.org/10.1111/2041-210X.13107

Valavi, R., Elith, J., Lahoz-Monfort, J. J., & Guillera-Arroita, G. (2021). Modelling species presence-only data with random forests. *Ecography*, *44*(12), 1731–1742. https://doi.org/10.1111/ECOG.05615

Valavi, R., Guillera-Arroita, G., Lahoz-Monfort, J. J., & Elith, J. (2021). Predictive performance of presence-only species distribution models: a benchmark study with reproducible code. *Ecological Monographs*, *92*(1), e01486. https://doi.org/10.1002/ECM.1486

van Aken, H. M., Ridderinkhof, H., & de Ruijter, W. P. M. (2004). North Atlantic deep water in the south-western Indian Ocean. *Deep Sea Research Part I: Oceanographic Research Papers*, *51*(6), 755–776. https://doi.org/10.1016/J.DSR.2004.01.008

Vilhena, D. A., & Antonelli, A. (2015). A network approach for identifying and delimiting biogeographical regions. *Nature Communications*, *6*(1), 1–9. https://doi.org/10.1038/ncomms7848

Walther, B. A., & 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. https://doi.org/10.1111/J.2005.0906-7590.04112.X

Warton, D. I., Foster, S. D., De'ath, G., Stoklosa, J., & Dunstan, P. K. (2015). Model-based thinking for community ecology. *Plant Ecology*, *216*(5), 669–682. https://doi.org/10.1007/S11258-014-0366-3/FIGURES/3

Warton, D. I., Renner, I. W., & Ramp, D. (2013). Model-Based Control of Observer Bias for the Analysis of Presence-Only Data in Ecology. *PLOS ONE*, *8*(11), e79168. https://doi.org/10.1371/JOURNAL.PONE.0079168

Warton, D. I., & Shepherd, L. C. (2010). Poisson point process models solve the "pseudo-absence problem" for presence-only data in ecology. *Https://Doi.Org/10.1214/10-AOAS331*, *4*(3), 1383–1402. https://doi.org/10.1214/10-AOAS331

Watling, L., Guinotte, J., Clark, M. R., & Smith, C. R. (2013). A proposed biogeography of the deep ocean floor. *Progress in Oceanography*, *111*, 91–112. https://doi.org/10.1016/J.POCEAN.2012.11.003

Woo, M., Pattiaratchi, C., & Feng, M. (2007). *Hydrography and water masses in the south-eastern Indian Ocean*. https://doi.org/10.13140/2.1.3119.2001

Woolley, S. N. C., Foster, S. D., Bax, N. J., Currie, J. C., Dunn, D. C., Hansen, C., Hill, N. A., O'Hara, T. D., Ovaskainen, O., Sayre, R., Vanhatalo, J. P., & Dunstan, P. K. (2020). Bioregions in Marine Environments: Combining Biological and Environmental Data for Management and Scientific Understanding. *BioScience*, *70*(1), 48–59. https://doi.org/10.1093/BIOSCI/BIZ133

Woolley, S. N. C., Foster, S. D., O'Hara, T. D., Wintle, B. A., & Dunstan, P. K. (2017). Characterising uncertainty in generalised dissimilarity models. *Methods in Ecology and Evolution*, *8*(8), 985–995. https://doi.org/10.1111/2041-210X.12710

Woolley, S. N. C., Mccallum, A. W., Wilson, R., O'Hara, T. D., & Dunstan, P. K. (2013). Fathom out: biogeographical subdivision across the Western Australian continental margin – a multispecies modelling approach. *Diversity and Distributions*, *19*(12), 1506–1517. https://doi.org/10.1111/DDI.12119

WoRMS. (2021). *WoRMS - World Register of Marine Species*. https://www.marinespecies.org/

Yesson, C., Bedford, F., Rogers, A. D., & Taylor, M. L. (2017). The global distribution of deep-water Antipatharia habitat. *Deep Sea Research Part II: Topical Studies in Oceanography*, *145*, 79–86. https://doi.org/10.1016/J.DSR2.2015.12.004

Yesson, C., Taylor, M. L., Tittensor, D. P., Davies, A. J., Guinotte, J., Baco, A., Black, J., Hall-Spencer, J. M., & Rogers, A. D. (2012). Global habitat suitability of cold-water octocorals. *Journal of Biogeography*, *39*(7), 1278–1292. https://doi.org/10.1111/J.1365-2699.2011.02681.X

Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Duarte Ritter, C., Daniel Edler, |, Farooq, H., Herdean, A., Ariza, M., Ruud Scharn, |, Svantesson, S., Wengström, N., Zizka, V., & Antonelli, | Alexandre. (2019). CoordinateCleaner: Standardized cleaning of occurrence records from biological collection databases. *Methods Ecol Evol*, 1–8. https://doi.org/10.1111/2041-210X.13152

Appendix A

Table A1. Indicator species for each cluster. Ai: affinity; Fi: Fidelity.

species	cluster	Ai	Fi	IndVal	order	family
Acanthastrea simplex	1.1	1	1	1	Scleractinia	Lobophylliidae
Aldersladum sodwanum	1.1	1	1	1	Alcyonacea	Alcyoniidae
Cladiella kashmani	1.1	1	1	1	Alcyonacea	Alcyoniidae
Lobophytum latilobatum	1.1	1	1	1	Alcyonacea	Alcyoniidae
Lobophytum venustum	1.1	1	1	1	Alcyonacea	Alcyoniidae
Sarcophyton flexuosum	1.1	1	1	1	Alcyonacea	Alcyoniidae
Sarcophyton infundibuliforme	1.1	1	1	1	Alcyonacea	Alcyoniidae
Scyphopodium ingolfi	1.1	1	1	1	Alcyonacea	Clavulariidae
Sinularia erecta	1.1	1	1	1	Alcyonacea	Alcyoniidae
Sinularia gyrosa	1.1	1	1	1	Alcyonacea	Alcyoniidae
Sinularia notanda	1.1	1	1	1	Alcyonacea	Alcyoniidae
Sinularia schleyeri	1.1	1	1	1	Alcyonacea	Alcyoniidae
Sinularia triangula	1.1	1	1	1	Alcyonacea	Alcyoniidae
Sinularia variabilis	1.1	1	1	1	Alcyonacea	Alcyoniidae
Xenia garciae	1.1	1	1	1	Alcyonacea	Xeniidae
Xenia kuekenthali	1.1	1	1	1	Alcyonacea	Xeniidae
Lobophyllia hemprichii	1.1	0.55	0.98	0.54	Scleractinia	Lobophylliidae
Galaxea fascicularis	1.1	0.57	0.93	0.53	Scleractinia	Euphylliidae
Favites pentagona	1.1	0.56	0.93	0.52	Scleractinia	Merulinidae
Trachycladus spinispirulifer	1.1	1	0.52	0.52	Trachycladida	Trachycladidae
Bugulella gracilis	1.1	1	0.51	0.51	Cheilostomatida	Bugulidae
Fungia fungites	1.1	0.57	0.9	0.51	Scleractinia	Fungiidae
Palythoa natalensis	1.1	1	0.51	0.51	Zoantharia	Sphenopidae
Xenia crassa	1.1	1	0.51	0.51	Alcyonacea	Xeniidae
Pavona varians	1.1	0.56	0.91	0.51	Scleractinia	Agariciidae
Porites lutea	1.1	0.55	0.93	0.51	Scleractinia	Poritidae
Costaticella carotica	1.1	1	0.5	0.5	Cheilostomatida	Catenicellidae
Platygyra daedalea	1.1	0.54	0.93	0.5	Scleractinia	Merulinidae
Astrea curta	1.1	0.51	0.98	0.5	Scleractinia	Merulinidae
Flabellum magnificum	1.2	0.63	0.81	0.51	Scleractinia	Flabellidae
Stylobates loisetteae	1.2	0.54	0.92	0.49	Actiniaria	Actiniidae
Flabellum (Flabellum) magnificum	1.2	0.58	0.76	0.44	Scleractinia	Flabellidae

Flabellum (Ulocyathus) hoffmeisteri	1.2	0.49	0.65	0.32	Scleractinia	Flabellidae
Parantipathes helicosticha	1.2	0.49	0.58	0.29	Antipatharia	Schizopathidae
Dendrobrachia paucispina	1.2	0.37	0.75	0.28	Alcyonacea	Dendrobrachiidae
Leiopathes acanthophora	1.2	0.25	1	0.25	Antipatharia	Leiopathidae
Pennatula indica	1.2	0.25	1	0.25	Pennatulacea	Pennatulidae
Rhombopsammia niphada	1.2	0.34	0.7	0.24	Scleractinia	Micrabaciidae
Caryophyllia (Caryophyllia) grandis	1.2	0.39	0.62	0.24	Scleractinia	Caryophylliidae
Stephanocyathus (Acinocyathus) explanans	1.2	0.44	0.53	0.23	Scleractinia	Caryophylliidae
Flabellum (Flabellum) lamellulosum	1.2	0.39	0.58	0.23	Scleractinia	Flabellidae
Stephanocyathus explanans	1.2	0.49	0.46	0.22	Scleractinia	Caryophylliidae
Madrepora oculata	1.2	0.53	0.41	0.22	Scleractinia	Oculinidae
Fungiacyathus (Fungiacyathus) stephanus	1.2	0.33	0.59	0.2	Scleractinia	Fungiacyathidae
Caryophyllia grandis	1.2	0.39	0.5	0.2	Scleractinia	Caryophylliidae
Monniotus ramosus	1.3	0.51	1	0.51	Aplousobranchia	Protopolyclinidae
Sphenotrochus (Sphenotrochus) imbricaticostatus	1.3	0.74	0.5	0.38	Scleractinia	Turbinoliidae
Discoporella umbellata	1.3	0.27	1	0.27	Cheilostomatida	Cupuladriidae
Kraussina rubra	1.3	0.26	1	0.26	Terebratulida	Kraussinidae
Sphenotrochus aurantiacus	1.3	0.27	0.88	0.24	Scleractinia	Turbinoliidae
Scleranthelia thomsoni	1.3	0.23	1	0.23	Alcyonacea	Clavulariidae
Pourtalopsammia togata	1.3	0.23	0.85	0.2	Scleractinia	Dendrophylliidae
Sinularia muralis	1.5	0.5	0.51	0.26	Alcyonacea	Alcyoniidae
Iconometra intermedia	1.5	0.5	0.49	0.24	Comatulida	Colobometridae
Cancellothyris hedleyi	1.5	0.48	0.5	0.24	Terebratulida	Cancellothyrididae
Neopetrosia tuberosa	1.7	1	0.67	0.67	Haplosclerida	Petrosiidae
Aulocalyx serialis	1.7	0.5	1	0.5	Lyssacinosida	Aulocalycidae
Discodermia tuberosa	1.7	0.5	1	0.5	Tetractinellida	Theonellidae
Hemigellius calyx	1.7	0.5	1	0.5	Haplosclerida	Niphatidae
Pachastrella ovisternata	1.7	0.5	1	0.5	Tetractinellida	Pachastrellidae

Caryophyllia planilamellata	1.7	0.67	0.54	0.36	Scleractinia	Caryophylliidae
Tethya irisae	1.7	0.34	1	0.34	Tethyida	Tethyidae
Caryophyllia (Caryophyllia) planilamellata	1.7	0.59	0.54	0.32	Scleractinia	Caryophylliidae
Flabellum (Ulocyathus) tuthilli	1.7	0.34	0.68	0.23	Scleractinia	Flabellidae
Rhynchocidaris triplopora	2.1	1	0.75	0.75	Cidaroida	Ctenocidaridae
Ctenocidaris (Eurocidaris) nutrix	2.1	0.85	0.71	0.6	Cidaroida	Ctenocidaridae
Plicatellopsis antarctica	2.1	0.52	0.88	0.46	Suberitida	Suberitidae
Pemphixina pyxidata	2.1	0.44	0.93	0.41	Rhynchonellida	Hemithirididae
Thouarella (Epithouarella) chilensis	2.1	0.32	1	0.32	Alcyonacea	Primnoidae
Tentorium papillatum	2.1	0.3	1	0.3	Polymastiida	Polymastiidae
Aerothyris kerguelensis	2.1	0.48	0.61	0.3	Terebratulida	Terebratellidae
Myxilla (Myxilla) basimucronata	2.1	0.29	1	0.29	Poecilosclerida	Myxillidae
Myxilla basimucronata	2.1	0.29	1	0.29	Poecilosclerida	Myxillidae
Antarctotetilla leptoderma	2.1	0.45	0.65	0.29	Tetractinellida	Tetillidae
Promachocrinus kerguelensis	2.1	0.41	0.66	0.27	Comatulida	Antedonidae
Myxilla (Ectyomyxilla) kerguelensis	2.1	0.26	1	0.26	Poecilosclerida	Myxillidae
Megaciella pilosa	2.1	0.3	0.88	0.26	Poecilosclerida	Acarnidae
Hymedesmia (Hymedesmia) mariondufresni	2.1	0.26	1	0.26	Poecilosclerida	Hymedesmiidae
Tedania (Tedaniopsis) charcoti	2.1	0.27	0.89	0.24	Poecilosclerida	Tedaniidae
Tedania charcoti	2.1	0.27	0.89	0.24	Poecilosclerida	Tedaniidae
Bubaris vermiculata	2.1	0.26	0.82	0.21	Bubarida	Bubaridae
Tetilla coronida	2.1	0.23	0.87	0.2	Tetractinellida	Tetillidae
Astrotoma agassizii	2.4	0.74	0.47	0.35	Euryalida	Gorgonocephalidae
Gorgonocephalus chilensis	2.4	0.76	0.34	0.26	Euryalida	Gorgonocephalidae
Magellania joubini	2.4	0.1	1	0.1	Terebratulida	Terebratellidae
Chondrocladia (Chondrocladia) clavata	3.1	1	1	1	Poecilosclerida	Cladorhizidae
Hyalonema (Cyliconema) madagascarense	3.1	0.6	1	0.6	Amphidiscosida	Hyalonematidae

Hyalonema (Cyliconemaoida) choaniferum	3.1	0.6	1	0.6	Amphidiscosida	Hyalonematidae
Proagnesia depressa	3.1	0.63	0.76	0.48	Phlebobranchia	Agneziidae
Styela calva	3.1	0.45	0.87	0.39	Stolidobranchia	Styelidae
Styela ordinaria	3.1	0.31	1	0.31	Stolidobranchia	Styelidae
Umbellula thomsoni	3.1	0.37	0.79	0.29	Pennatulacea	Umbellulidae
Abyssopathes lyra	3.1	0.39	0.56	0.22	Antipatharia	Schizopathidae
Bathypathes patula	3.1	0.51	0.41	0.21	Antipatharia	Schizopathidae
Cerelasma massa	3.1	0.6	0.33	0.2	Foraminifera	Cerelasmidae

Appendix B

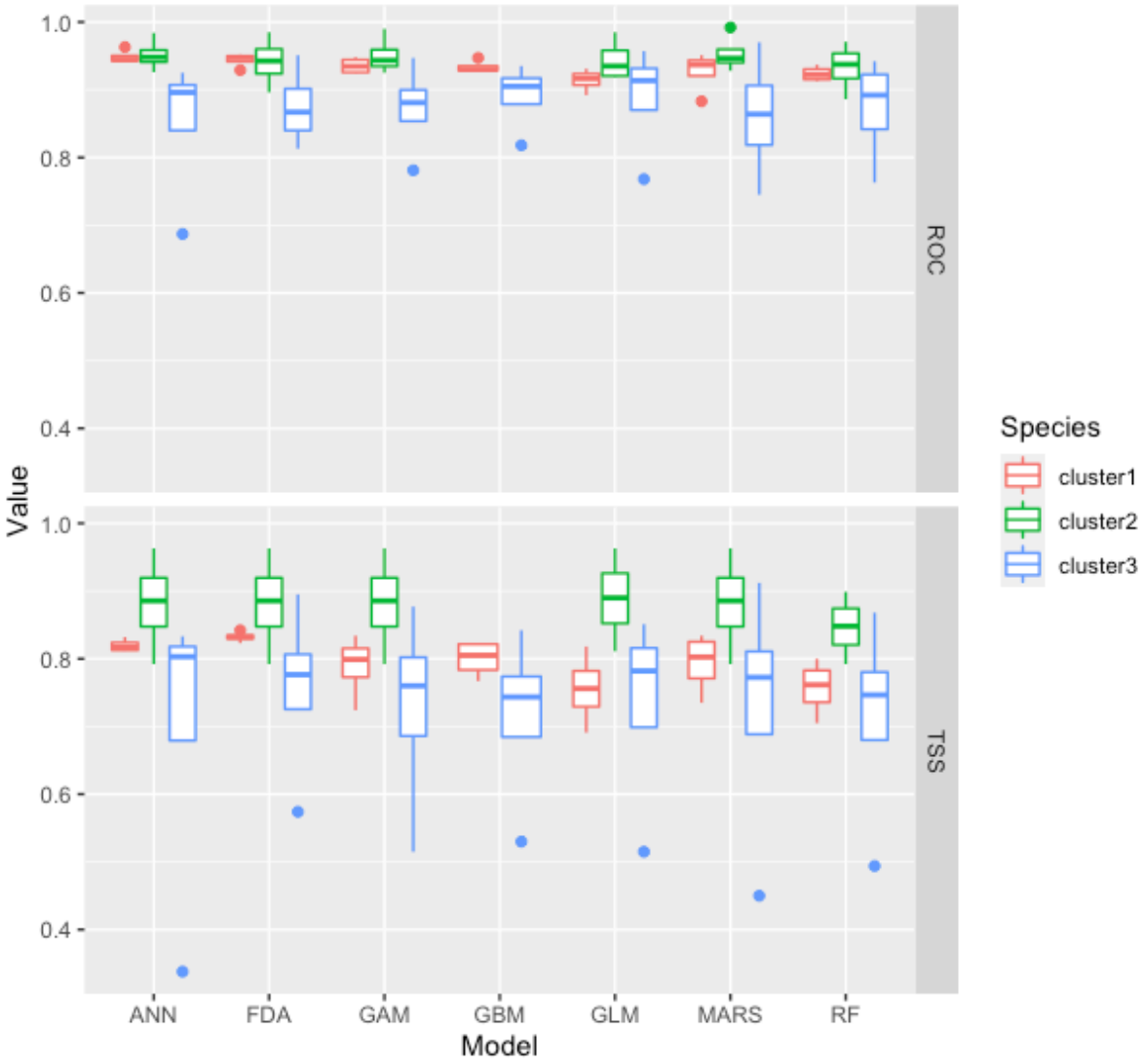

Figure B1. Model evaluation metrics for the three biogeographical regions at the first hierarchical level. ROC and TSS values for each of the seven models included for the ensemble model.

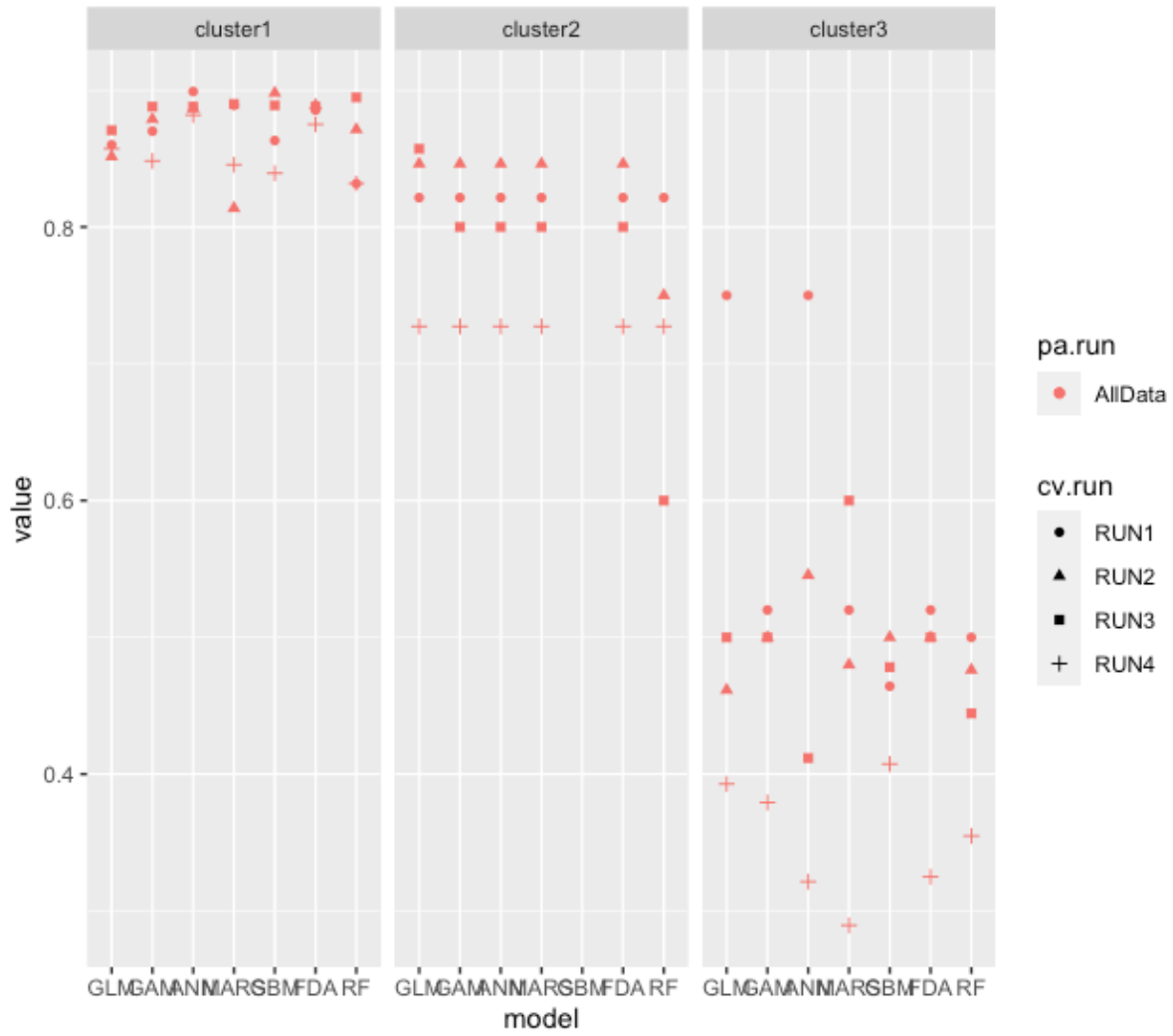

Figure B2. Model evaluation metrics for the three biogeographical regions at the first hierarchical level. Jaccard index for each of the seven models included for the ensemble model.

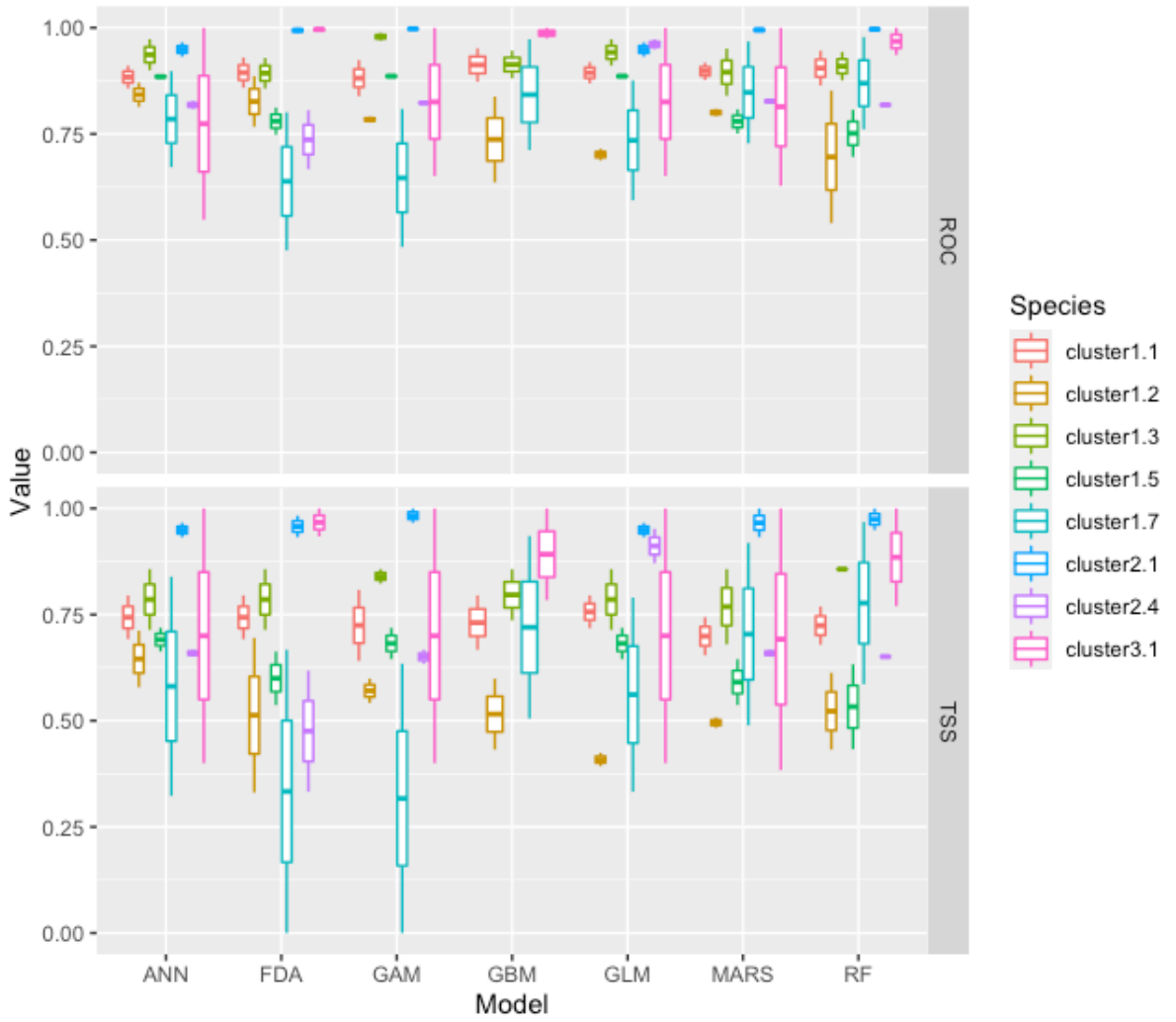

Figure B3. Model evaluation metrics for the eight biogeographical regions at the second hierarchical level. ROC and TSS values for each of the seven models included for the ensemble model. Evaluations are not reliable due to the low number of occurrences per subregions (< 30 occurrences).

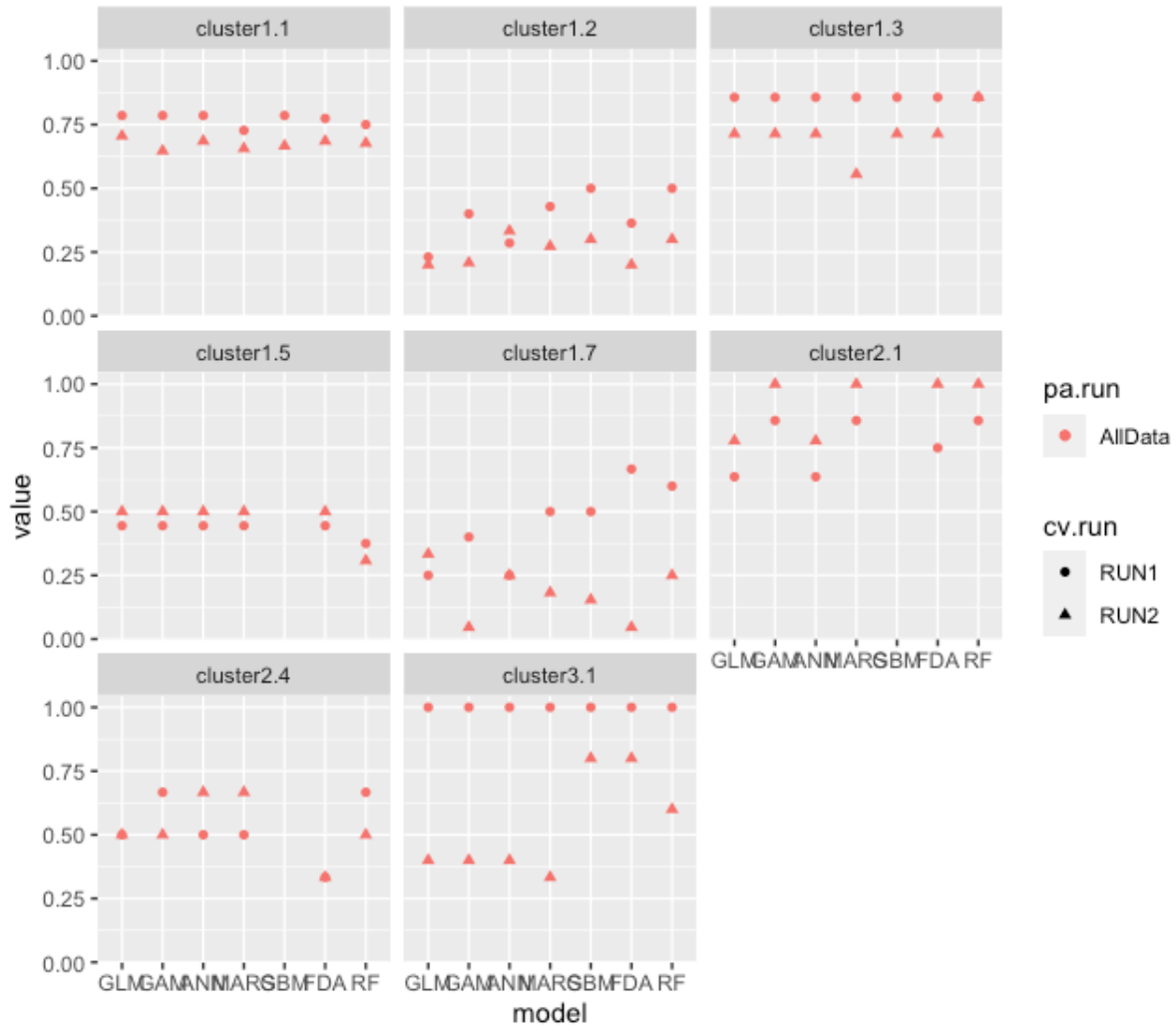

Figure B4. Model evaluation metrics for the three biogeographical regions at the first hierarchical level. Jaccard index for each of the seven models included for the ensemble model. Evaluations are not reliable due to the low number of occurrences per subregions (< 30 occurrences).

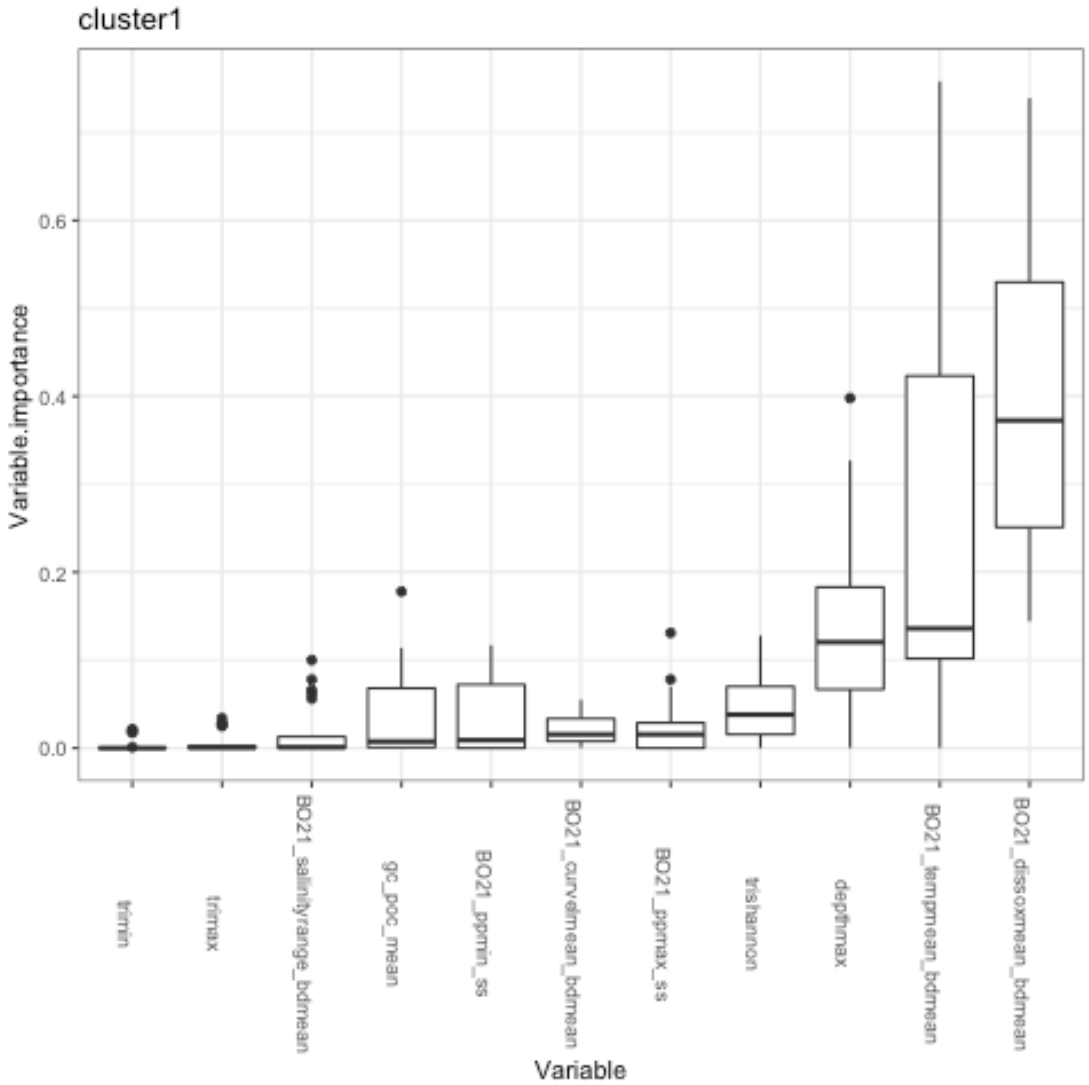

Figure B5. Variable importance plot for Cluster 1. Variables are considered important if they are so for at least 10% of the models, represented by the median values of the boxplot.

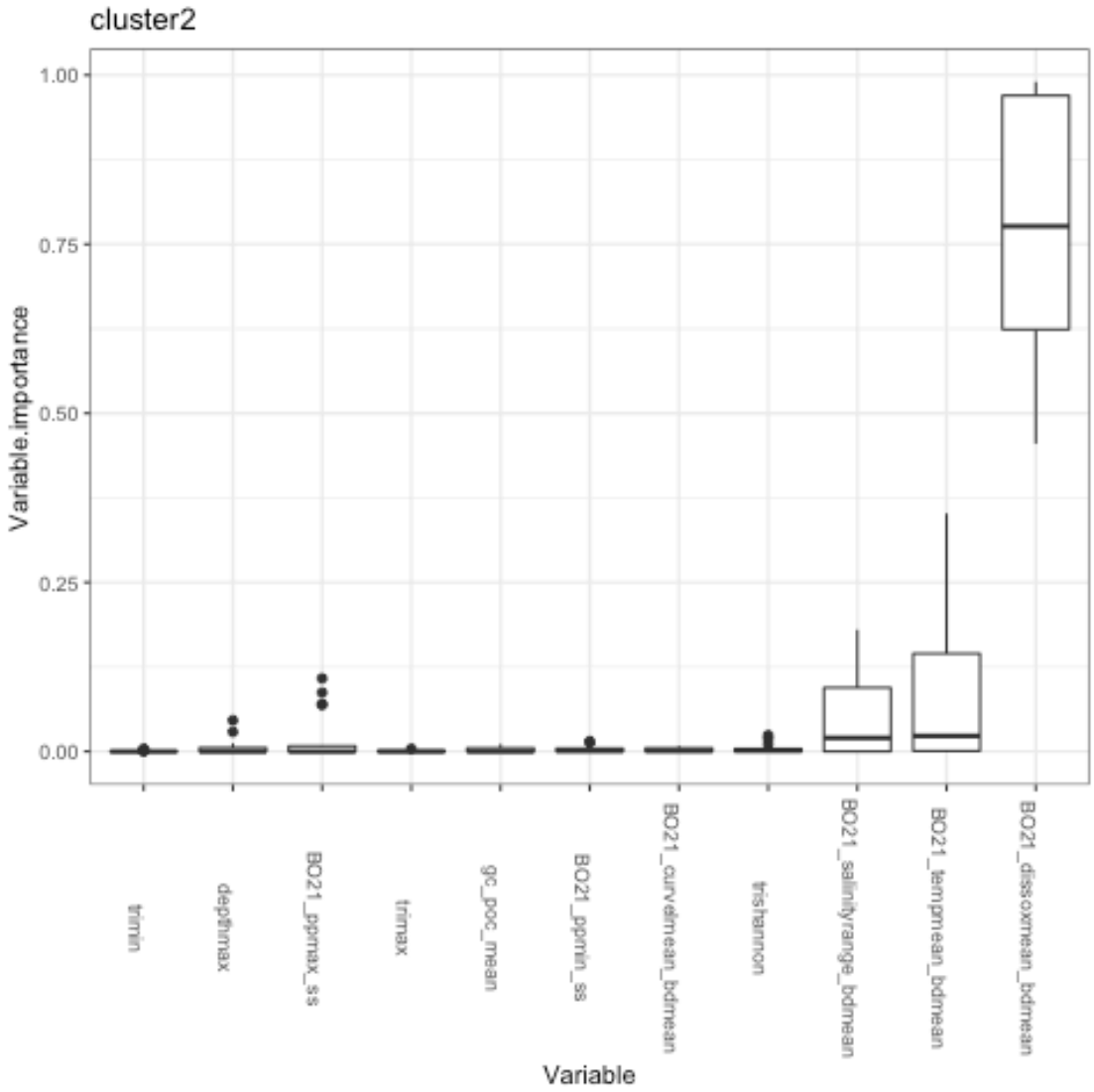

Figure B6. Variable importance plot for Cluster 2. Variables are considered important if they are so for at least 10% of the models, represented by the median values of the boxplot.

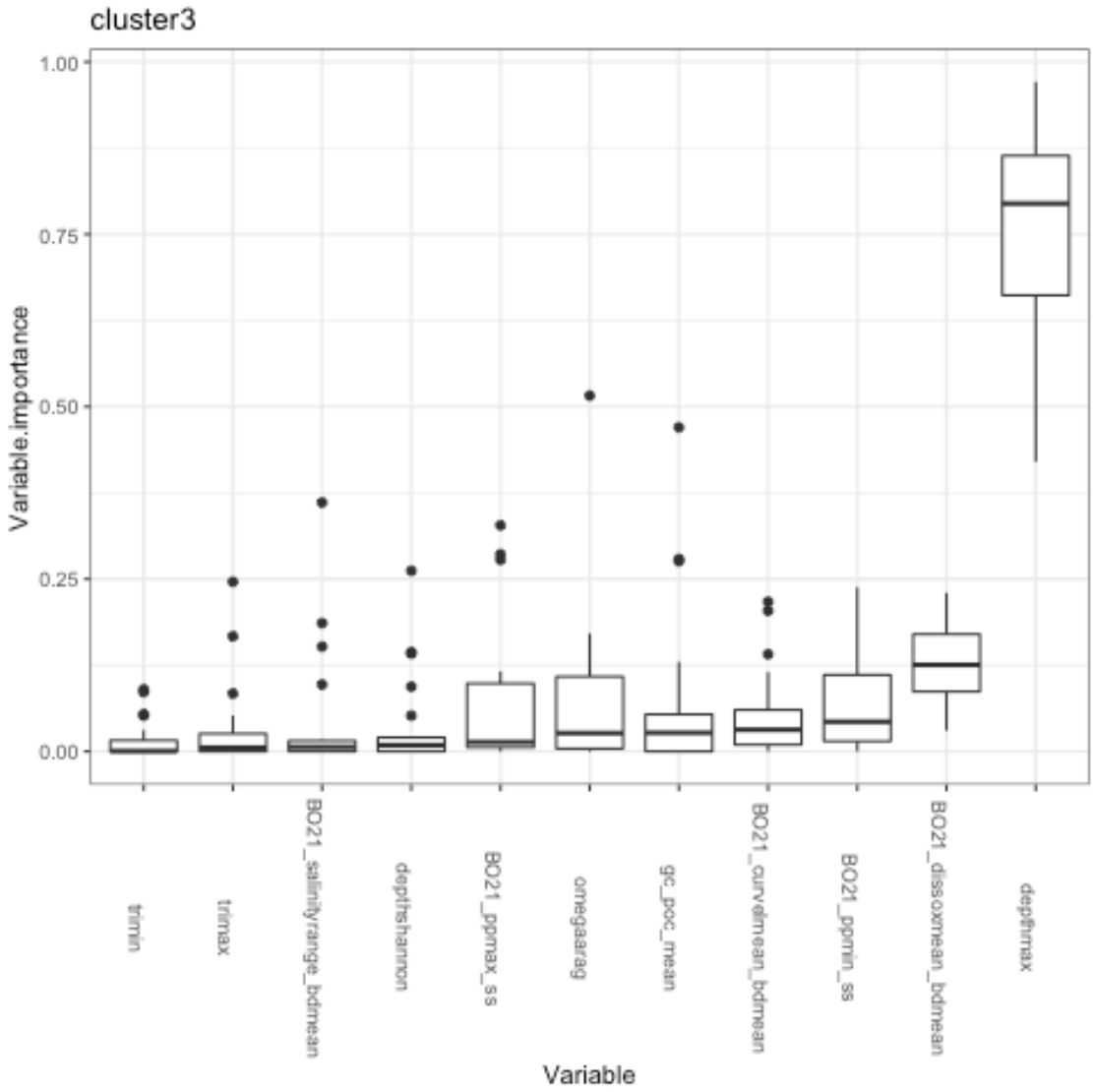

Figure B7. Variable importance plot for Cluster 3. Variables are considered important if they are so for at least 10% of the models, represented by the median values of the boxplot.

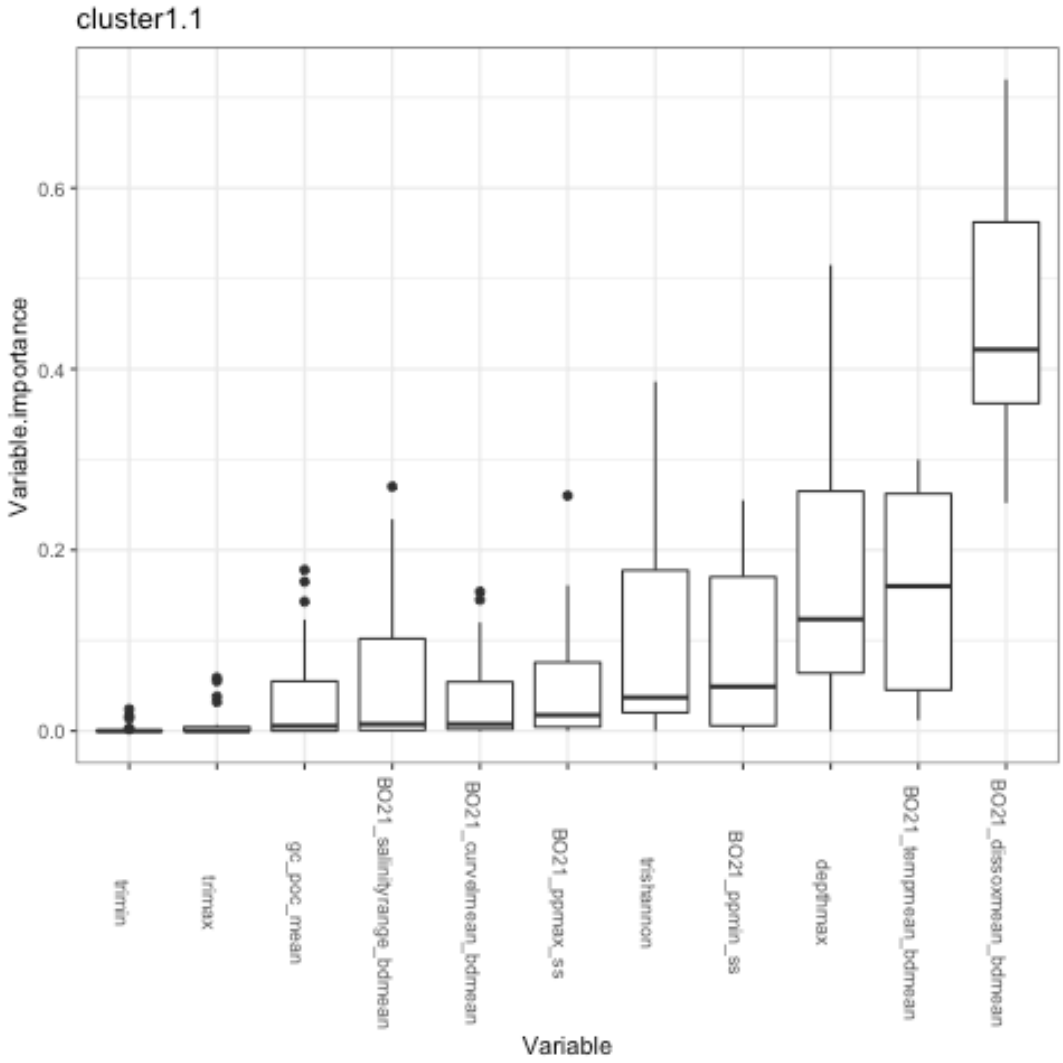

Figure B8. Variable importance plot for Cluster 1.1. Variables are considered important if they are so for at least 10% of the models, represented by the median values of the boxplot.

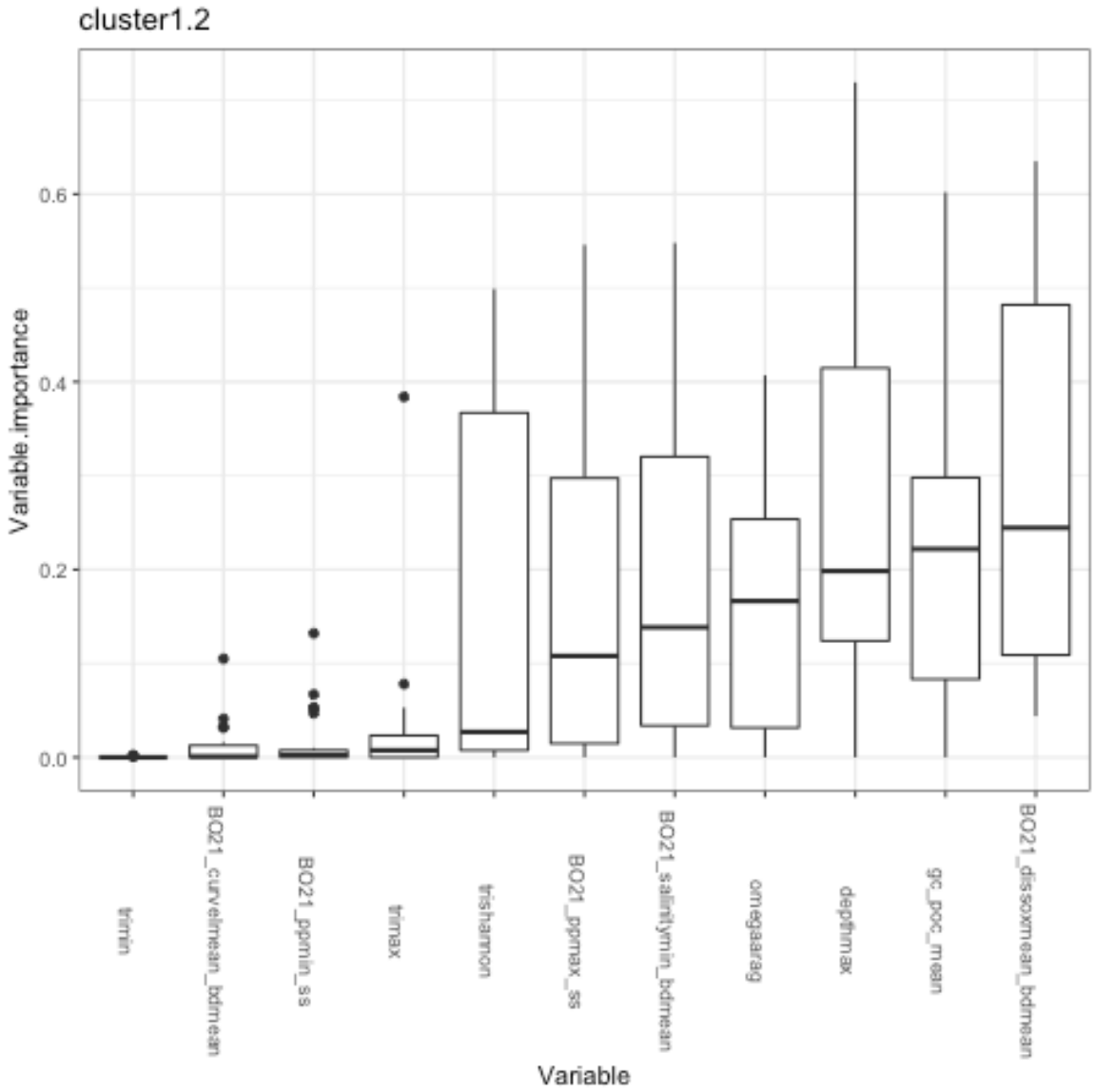

Figure B9. Variable importance plot for Cluster 1.2. Variables are considered important if they are so for at least 10% of the models, represented by the median values of the boxplot.

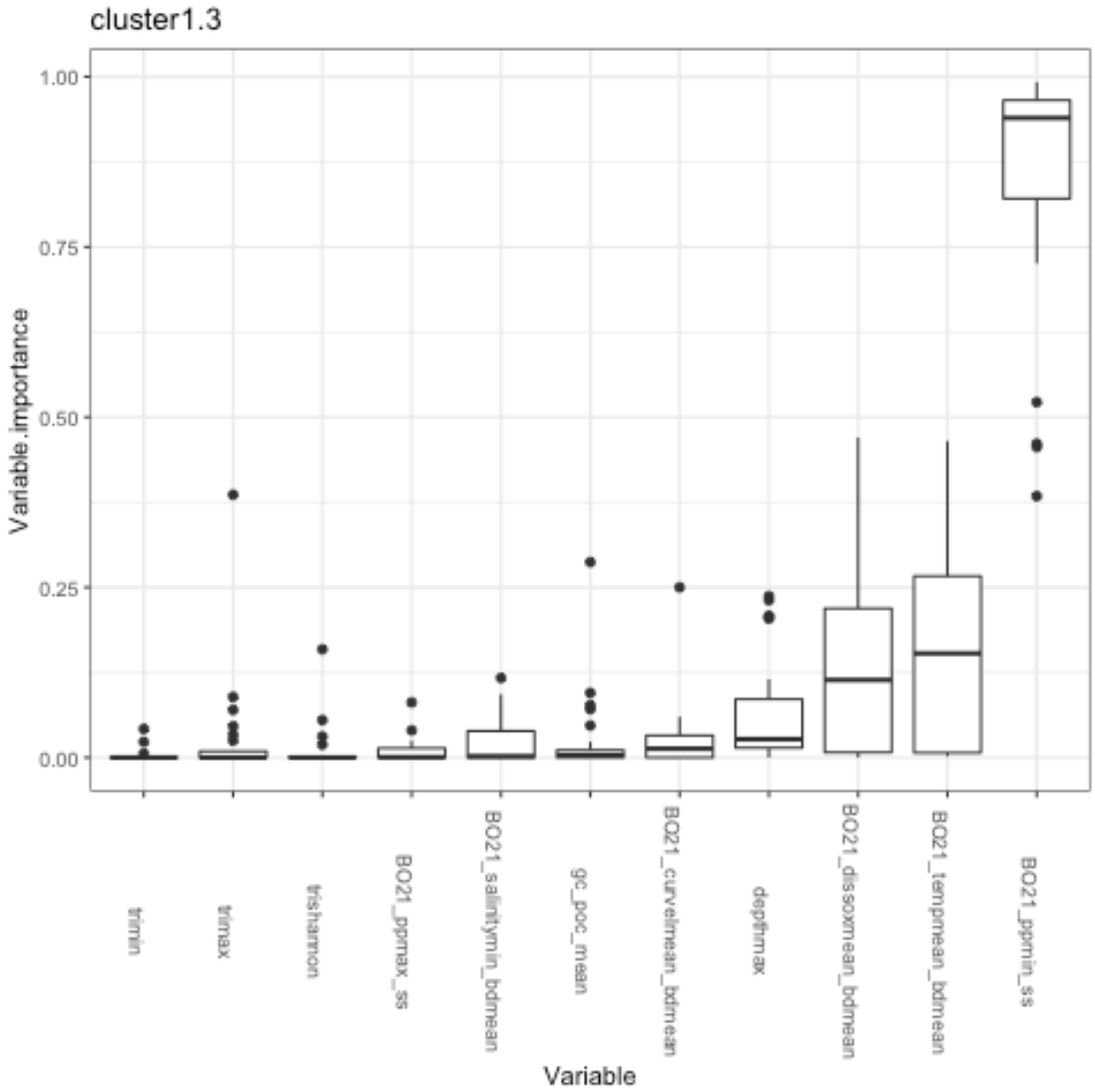

Figure B10. Variable importance plot for Cluster 1.3. Variables are considered important if they are so for at least 10% of the models, represented by the median values of the boxplot.

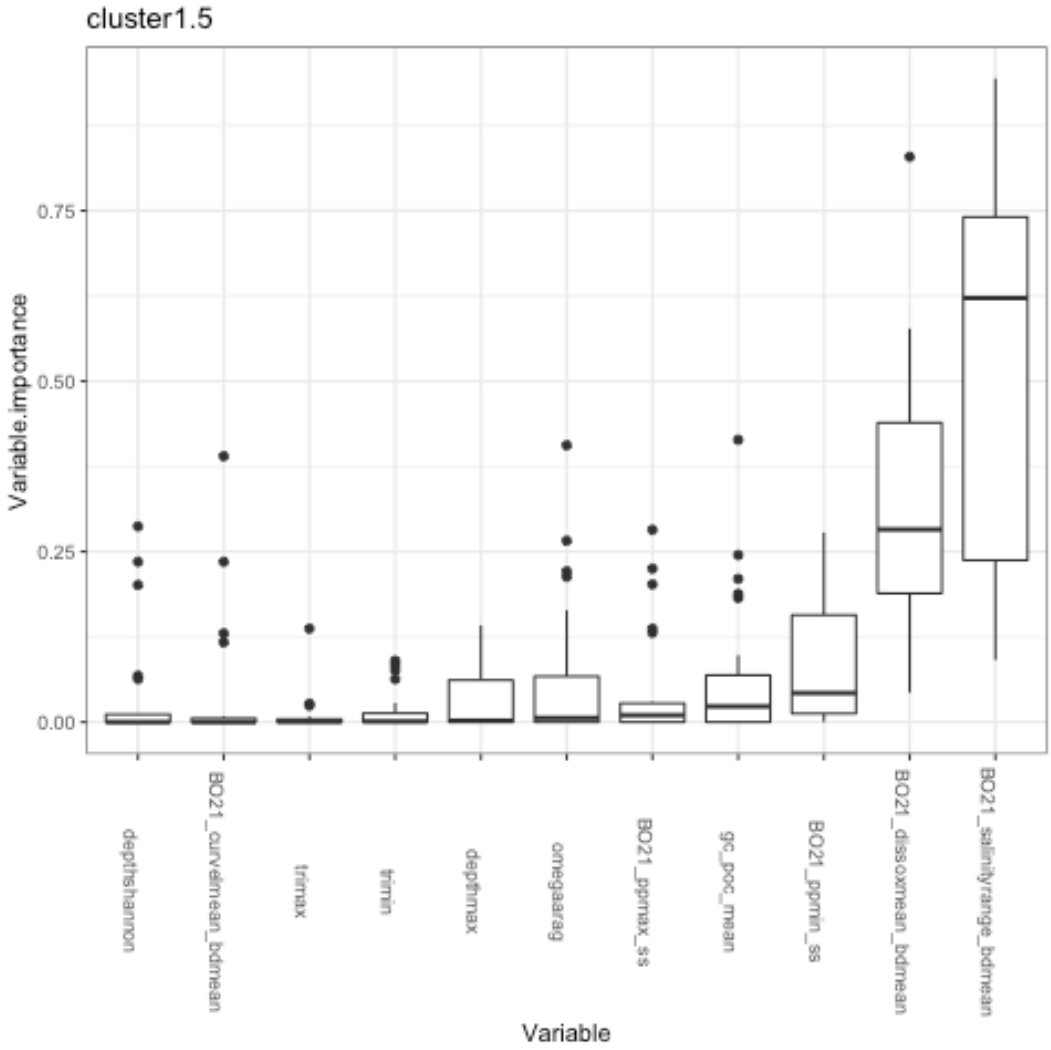

Figure B11. Variable importance plot for Cluster 1.5. Variables are considered important if they are so for at least 10% of the models, represented by the median values of the boxplot.

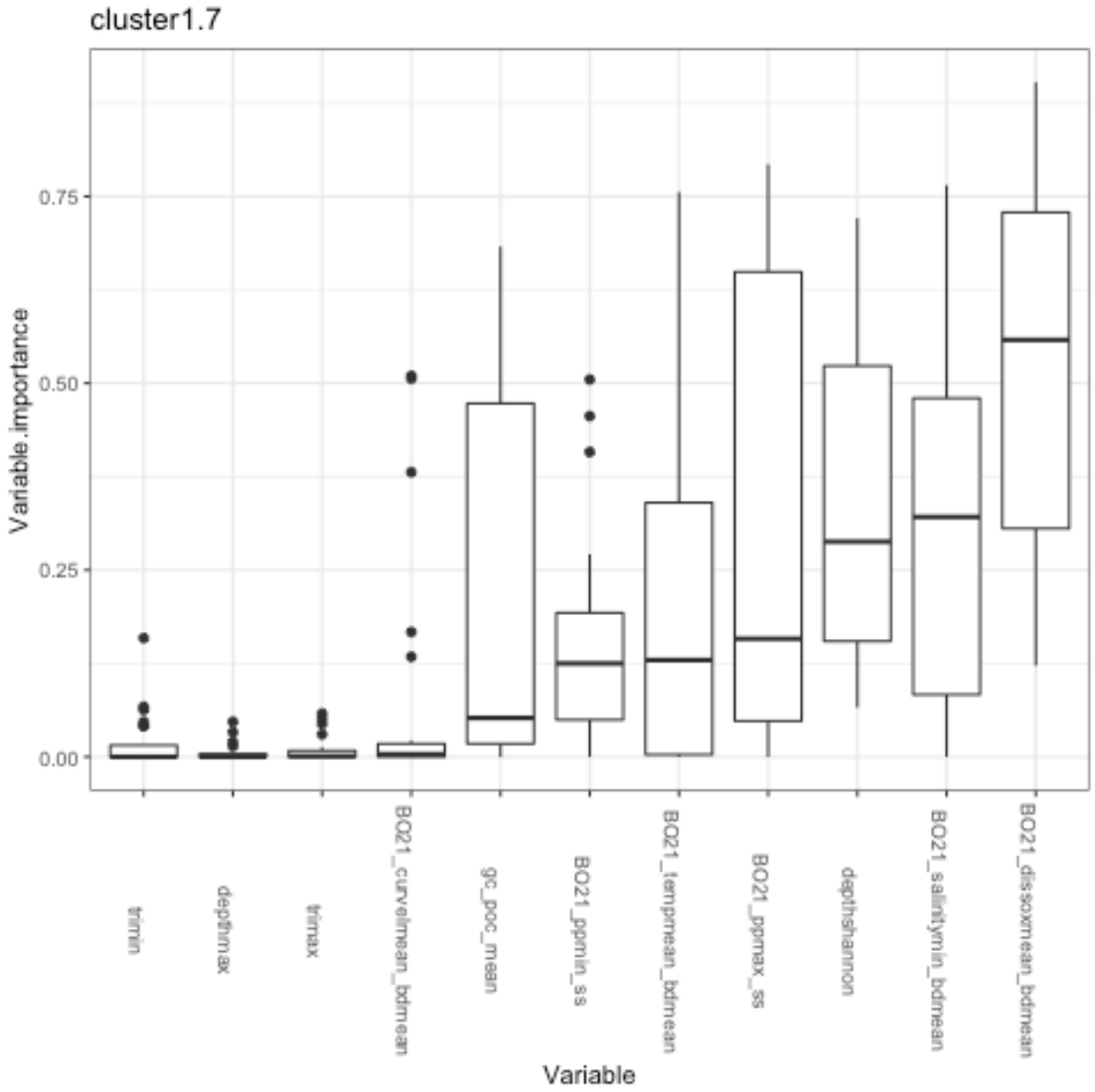

Figure B12. Variable importance plot for Cluster 1.7. Variables are considered important if they are so for at least 10% of the models, represented by the median values of the boxplot.

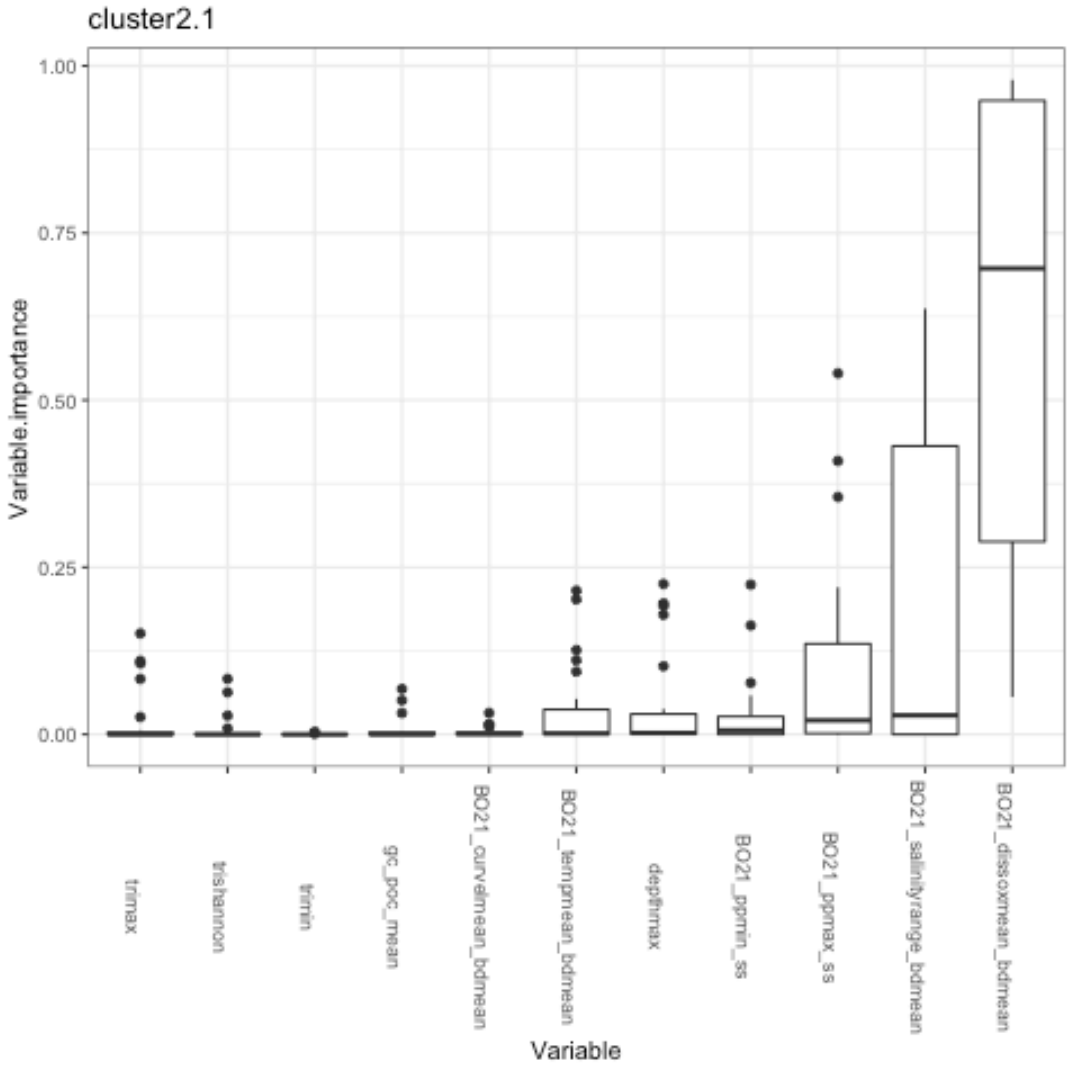

Figure B13. Variable importance plot for Cluster 2.1. Variables are considered important if they are so for at least 10% of the models, represented by the median values of the boxplot.

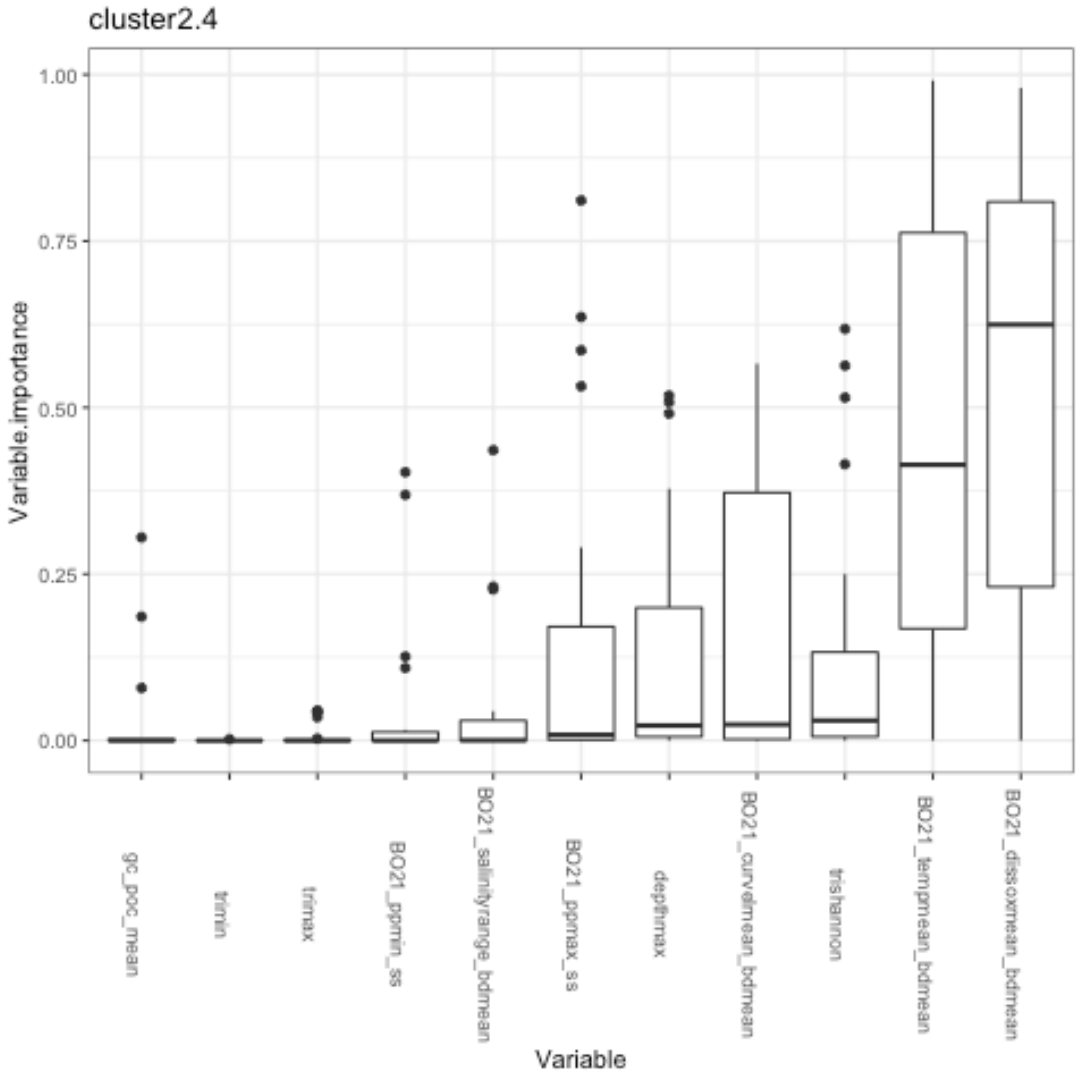

Figure B14. Variable importance plot for Cluster 2.4. Variables are considered important if they are so for at least 10% of the models, represented by the median values of the boxplot.

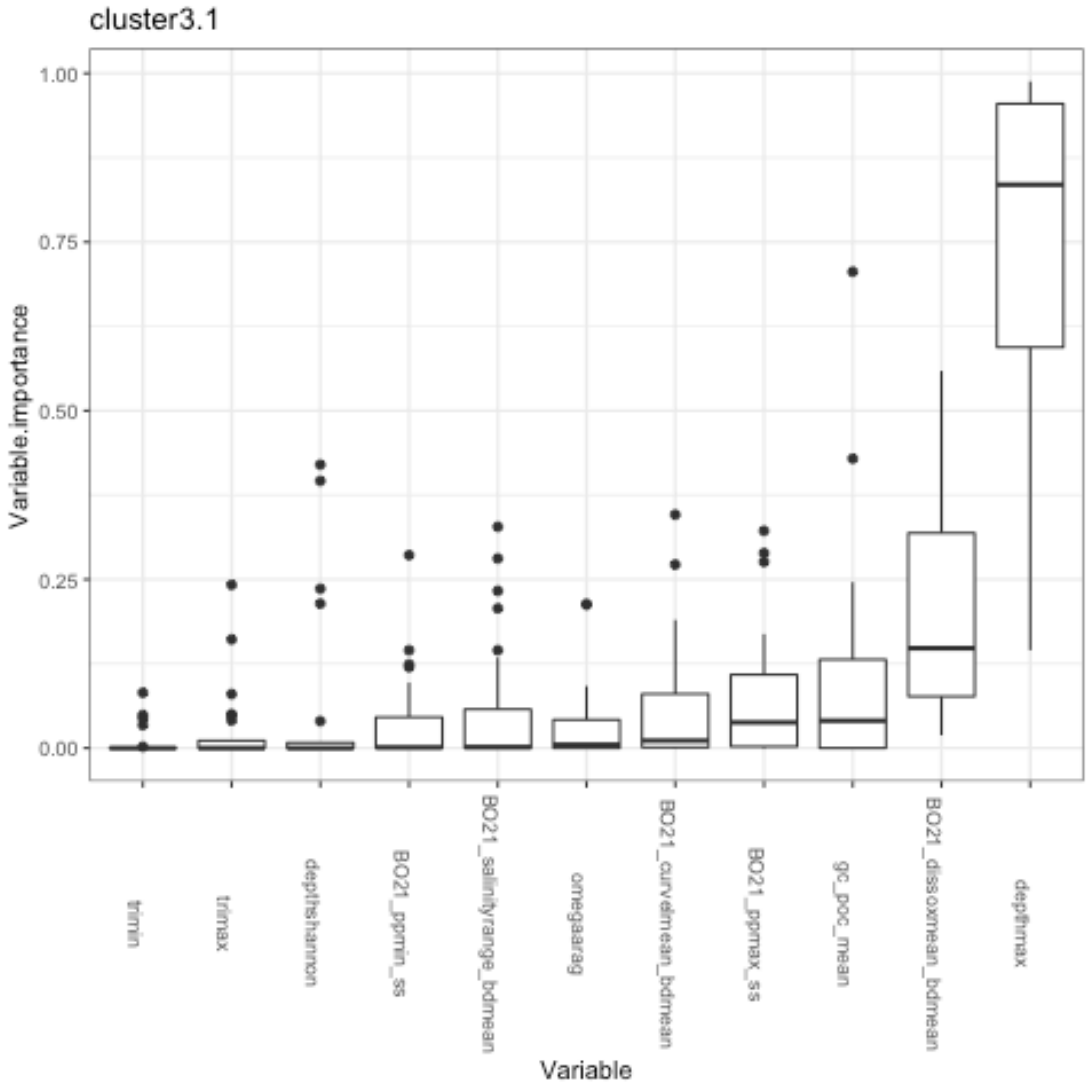

Figure B15. Variable importance plot for Cluster 3.1. Variables are considered important if they are so for at least 10% of the models, represented by the median values of the boxplot.

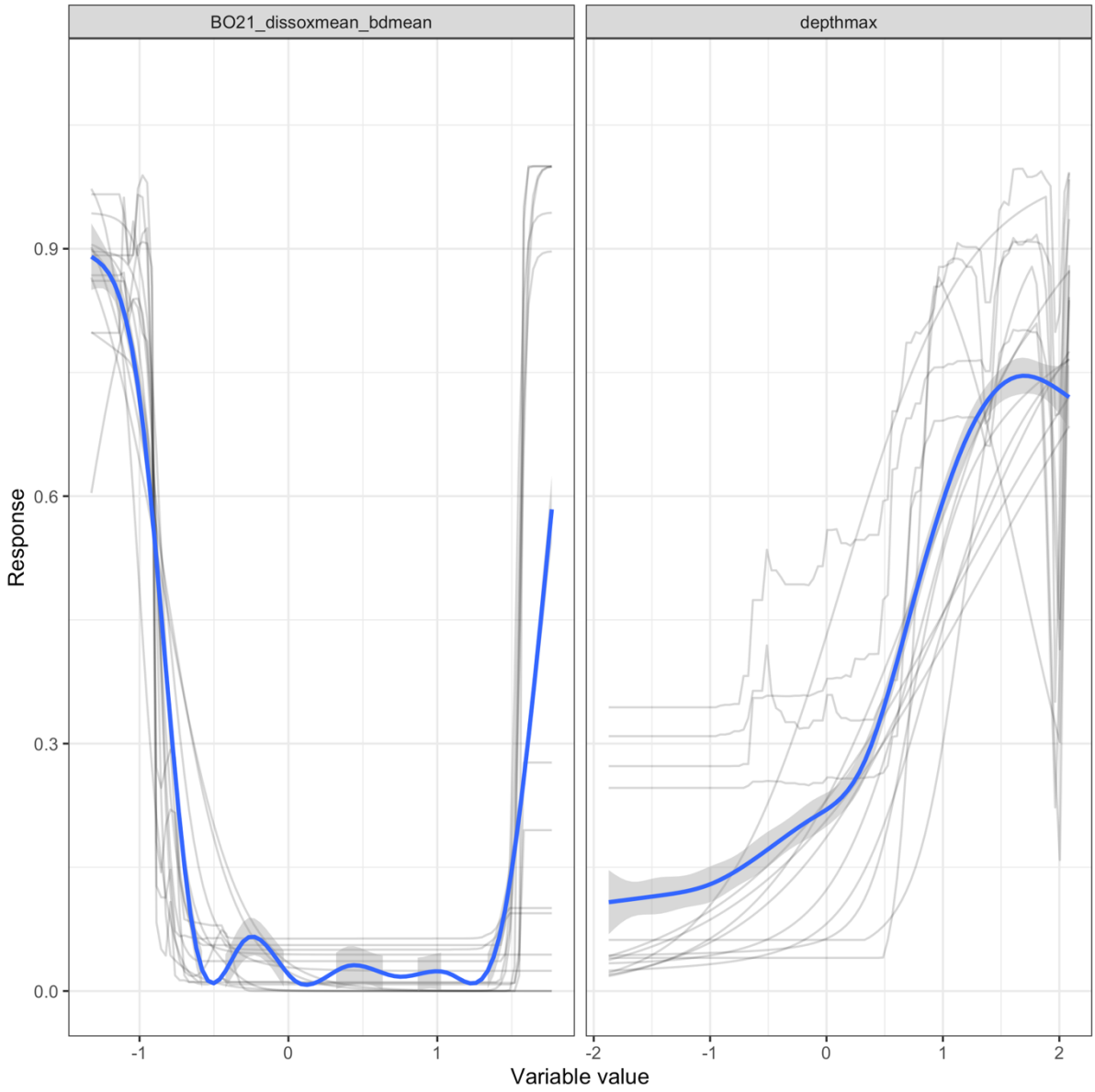

Figure B16. Response plot of predicted Cluster 1.1 based on mean bottom dissolved oxygen and maximum bottom depth.

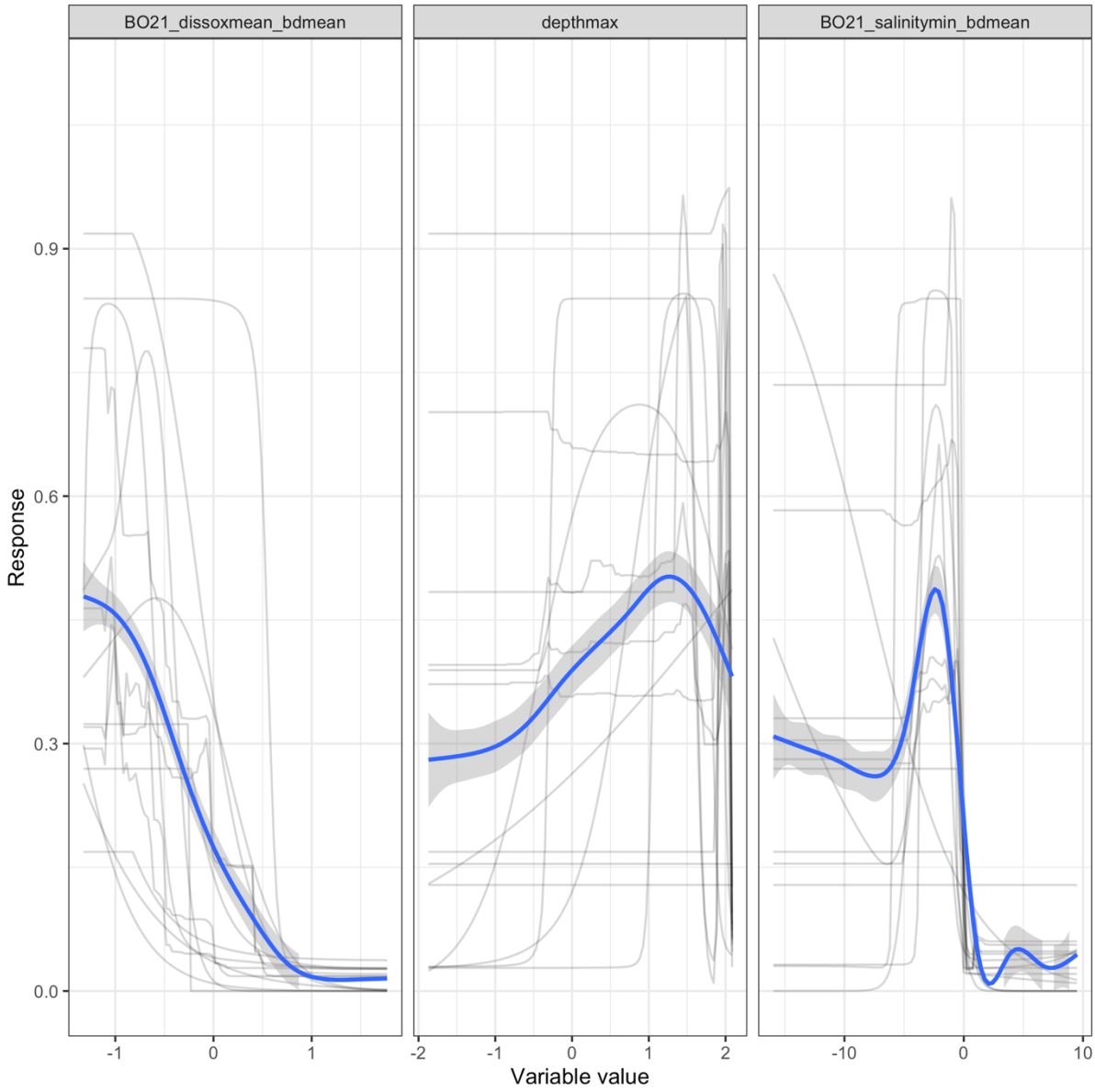

Figure B17. Response plot of predicted Cluster 1.2 based on mean bottom dissolved oxygen, maximum bottom depth and minimum bottom salinity.

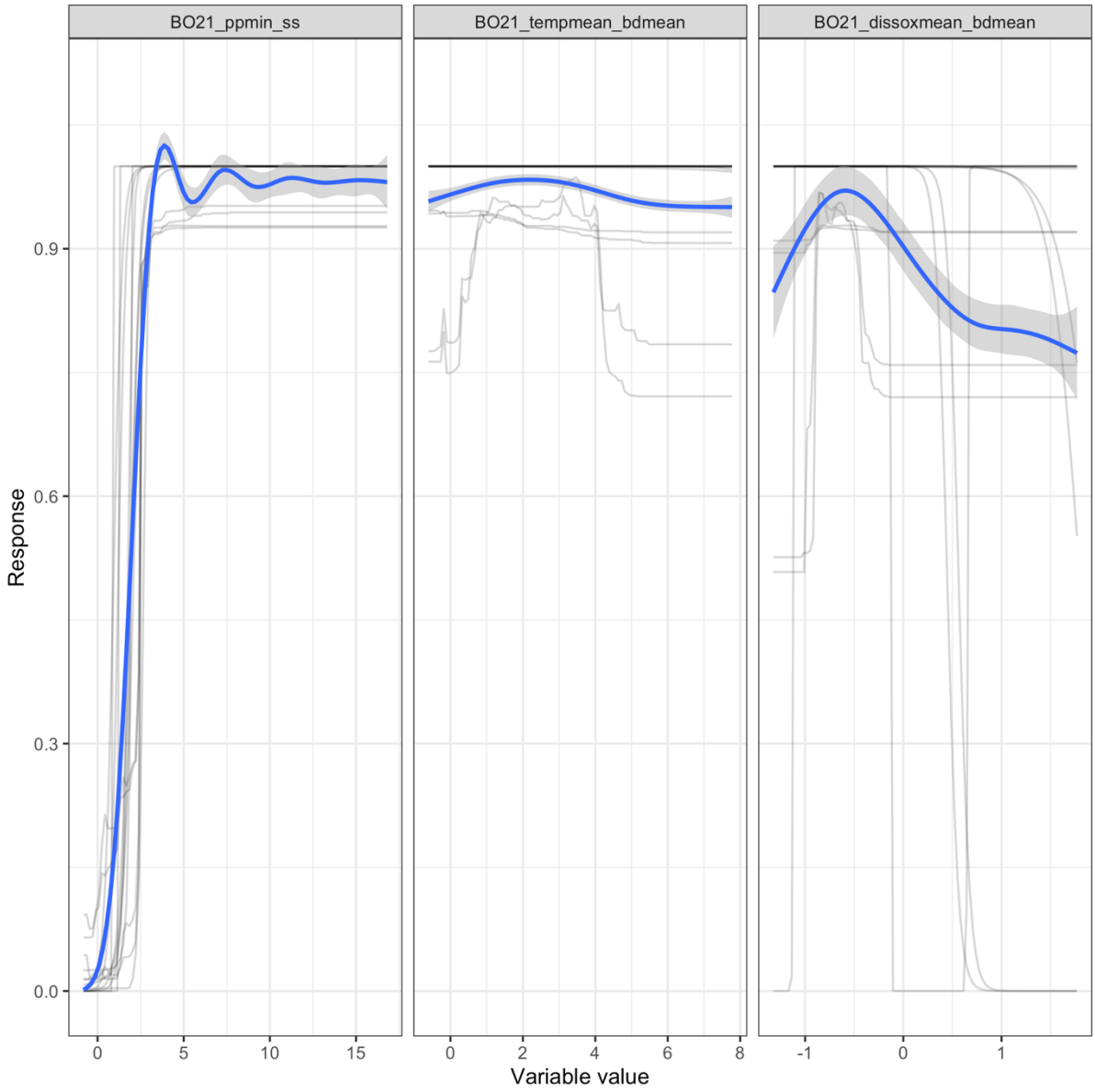

Figure B18. Response plot of predicted Cluster 1.3 based on minimum primary productivity at surface, mean bottom dissolved oxygen and temperature.

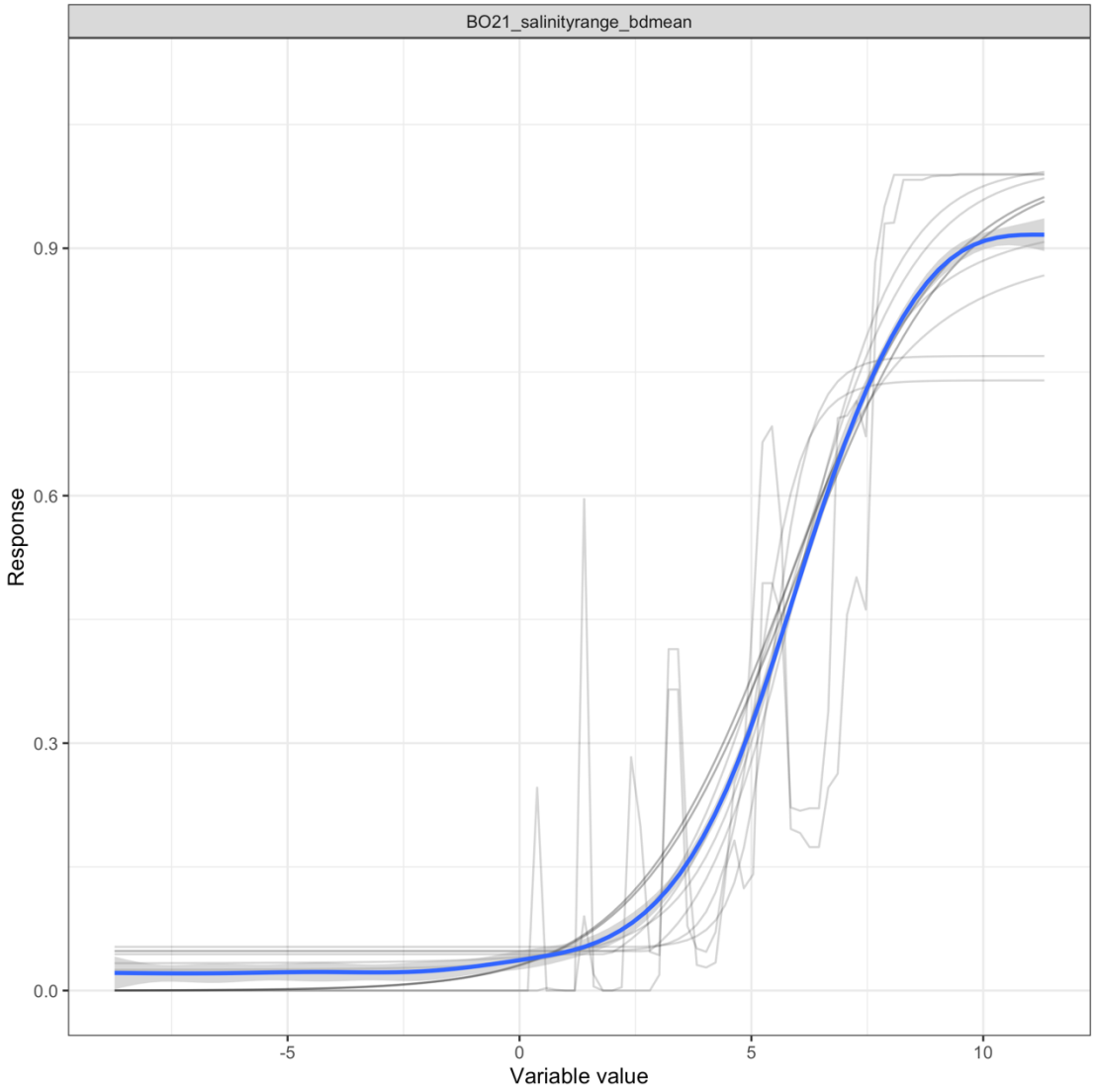

Figure B19. Response plot of predicted Cluster 1.5 based on salinity range.

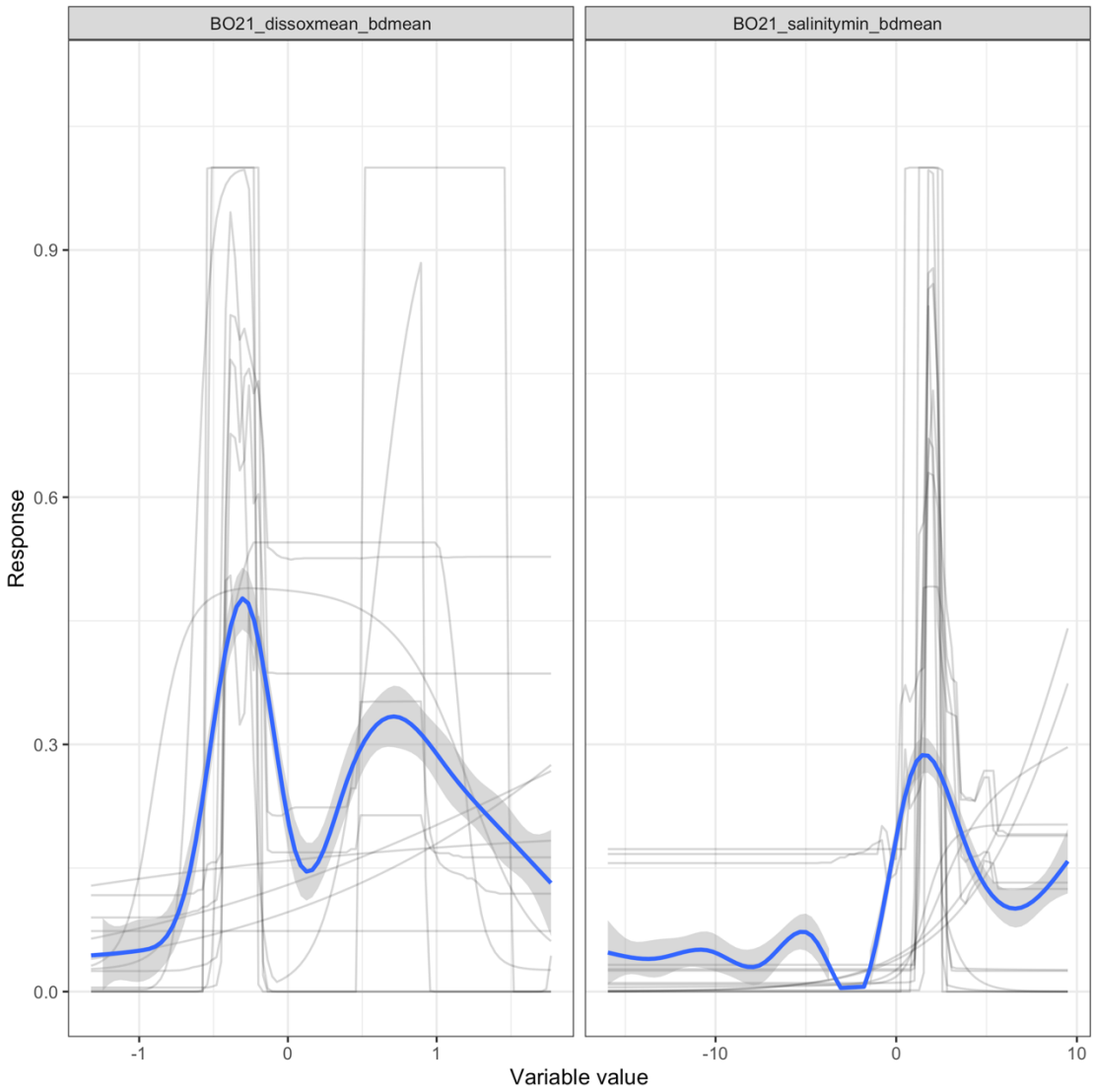

Figure B20. Response plot of predicted Cluster 1.7 based on mean bottom dissolved oxygen and minimum bottom salinity.

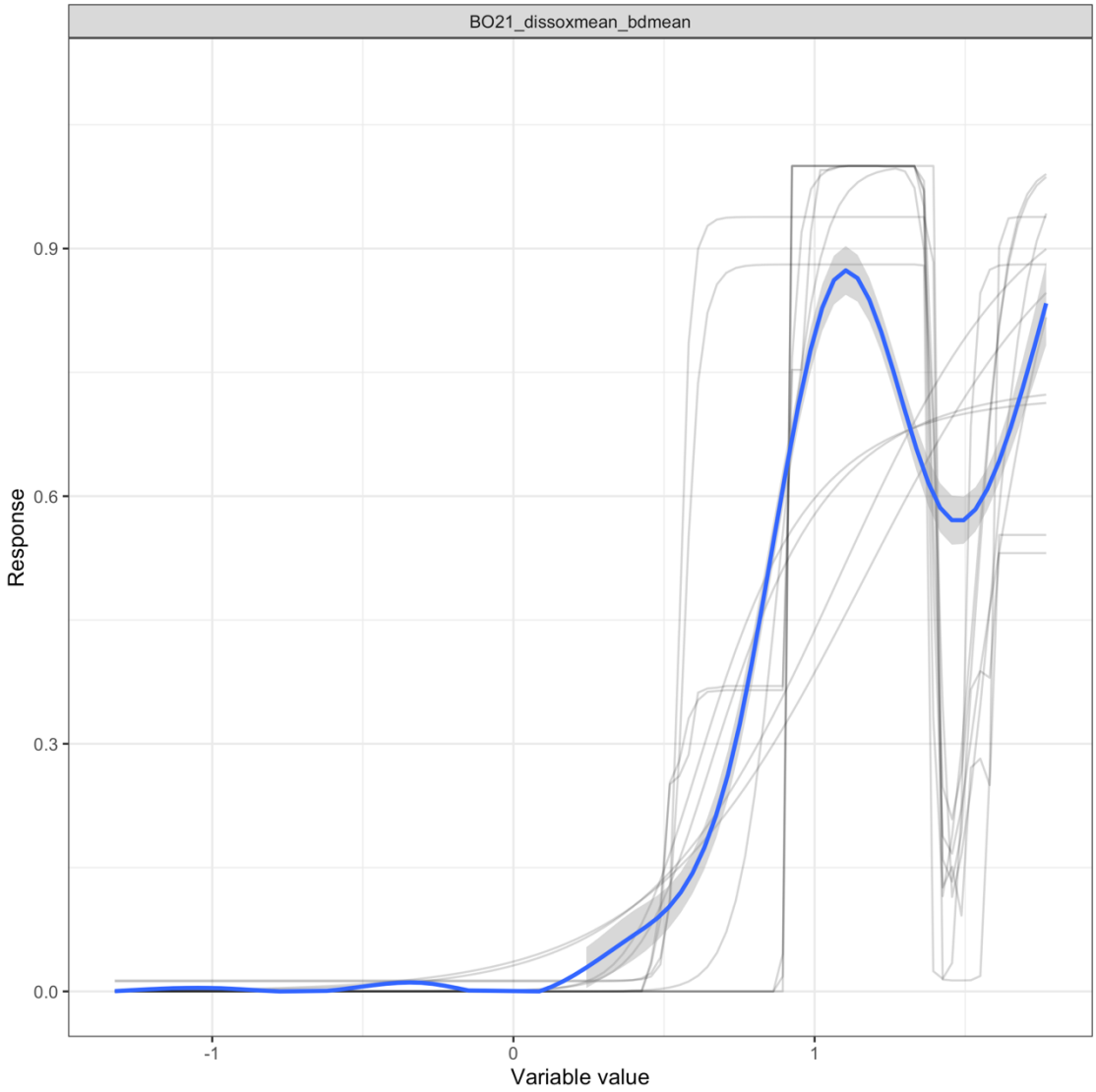

Figure B21. Response plot of predicted Cluster 2.1 based on mean bottom dissolved oxygen.

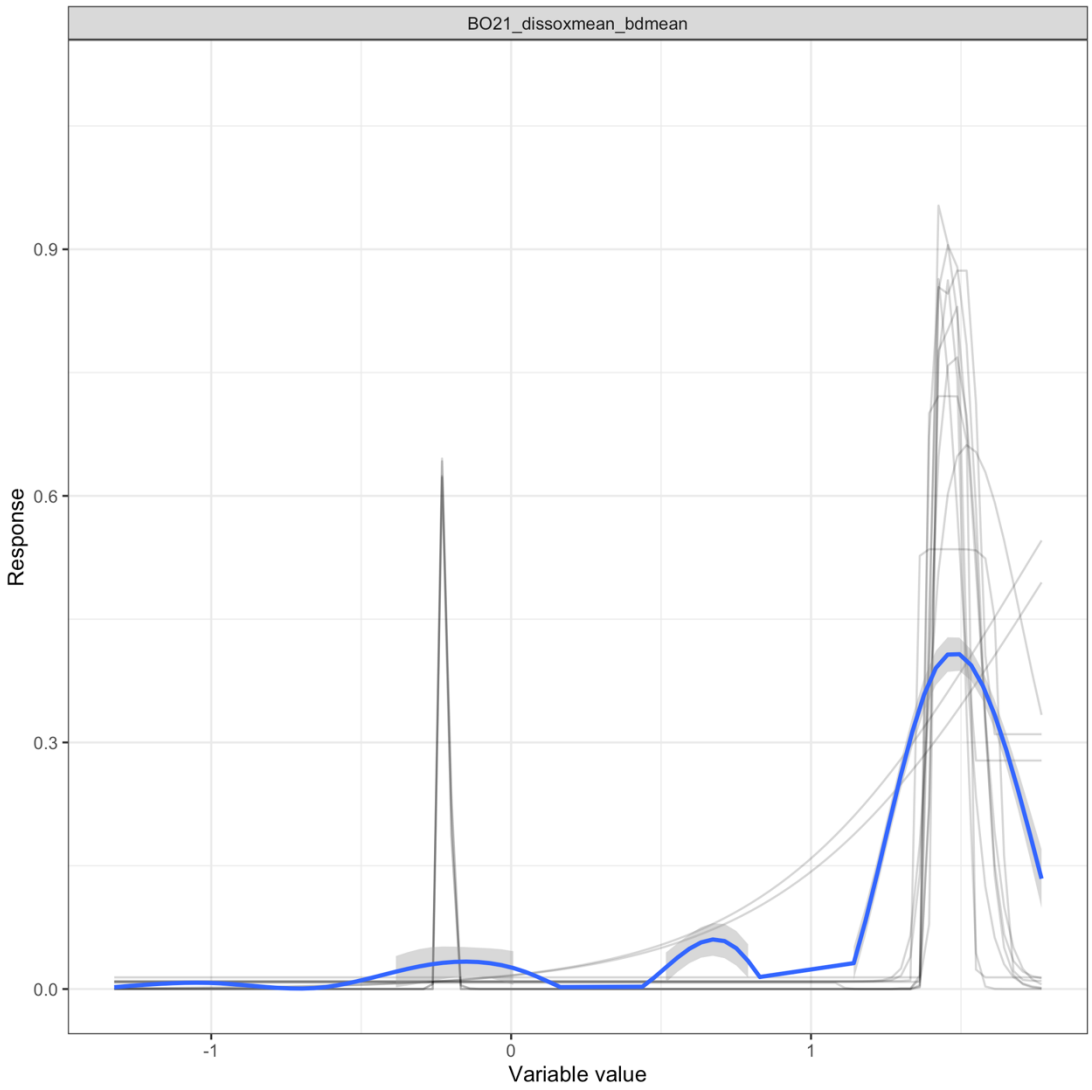

Figure B22. Response plot of predicted Cluster 2.4 based on mean bottom dissolved oxygen.

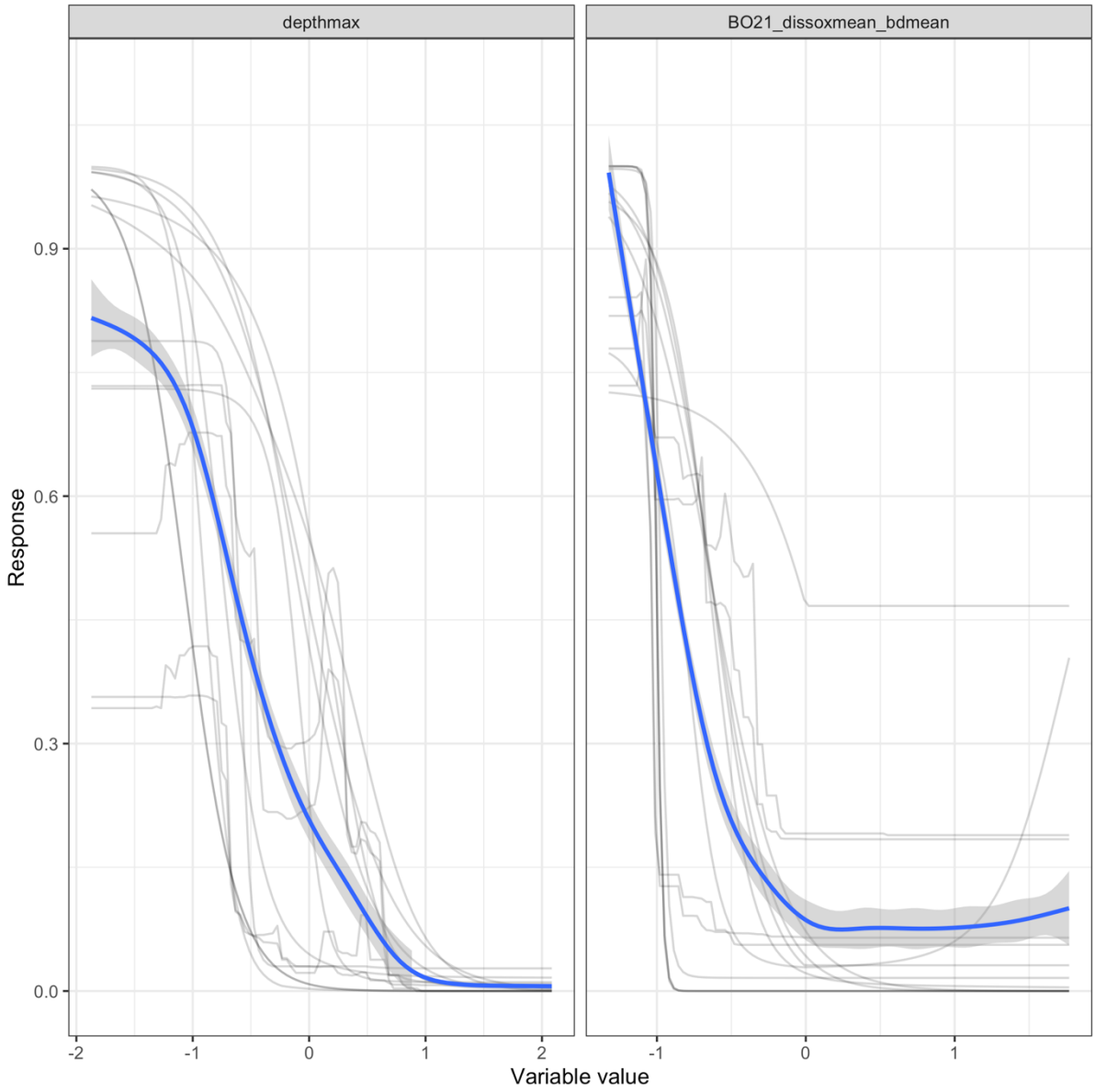

Figure B23. Response plot of predicted Cluster 3.1 based on mean bottom dissolved oxygen and maximum bottom depth.

Appendix C

Table C1. Indicator values for Family bioregionalisation.

family	*cluster*	*A*	*F*	*IndVal*	*order*
Chlidonophoridae	1	0.62	0.92	0.57	Terebratulida
Umbellulidae	1	0.57	0.81	0.46	Pennatulacea
Platidiidae	1	0.41	0.98	0.40	Terebratulida
Euplectellidae	1	0.37	1.00	0.37	Lyssacinosida
Cellariidae	1	0.38	0.92	0.36	Cheilostomatida
Hyalonematidae	1	0.33	0.97	0.32	Amphidiscosida
Molgulidae	1	0.37	0.83	0.31	Stolidobranchia
Primnoidae	1	0.31	1.00	0.30	Alcyonacea
Bugulidae	1	0.32	0.94	0.30	Cheilostomatida
Hymedesmiidae	1	0.28	1.00	0.28	Poecilosclerida
Deltocyathidae	1	0.27	0.99	0.27	Scleractinia
Pennatulidae	1	0.27	1.00	0.27	Pennatulacea
Celleporidae	1	0.26	0.99	0.26	Cheilostomatida
Tetillidae	1	0.25	1.00	0.25	Tetractinellida
Agneziidae	2	1	0.57	0.57	Phlebobranchia
Schizopathidae	3	1	0.38	0.38	Antipatharia
Styelidae	4	1	0.27	0.27	Stolidobranchia
Pyuridae	5	1	0.32	0.32	Stolidobranchia
Candidae	6	1	0.26	0.26	Cheilostomatida

Table C2. Indicator values for Clusters based on genus bioregionalisation.

genus	*cluster*	*A*	*F*	*IndVal*	*order*	*family*
Tedania	1	0.72	0.95	0.68	Poecilosclerida	Tedaniidae
Solenosmilia	1	0.74	0.93	0.68	Scleractinia	Caryophylliidae
Myxilla	1	0.70	0.97	0.68	Poecilosclerida	Myxillidae
Stephanocyathus	1	0.67	0.99	0.67	Scleractinia	Caryophylliidae
Ascidia	1	0.65	1.00	0.65	Phlebobranchia	Ascidiidae
Lobactis	1	0.66	0.98	0.64	Scleractinia	Fungiidae
Spirobranchus	1	0.66	0.96	0.64	Sabellida	Serpulidae
Leptastrea	1	0.63	1.00	0.63	Scleractinia	Leptastreidae
Acanthastrea	1	0.63	0.99	0.63	Scleractinia	Lobophylliidae
Halomitra	1	0.63	1.00	0.63	Scleractinia	Fungiidae
Danafungia	1	0.63	1.00	0.63	Scleractinia	Fungiidae
Tethya	1	0.63	1.00	0.63	Tethyida	Tethyidae
Gardineroseris	1	0.62	1.00	0.62	Scleractinia	Agariciidae
Leptoseris	1	0.62	0.99	0.62	Scleractinia	Agariciidae
Sinularia	1	0.62	1.00	0.62	Alcyonacea	Alcyoniidae
Acanthella	1	0.62	0.99	0.61	Bubarida	Dictyonellidae
Porites	1	0.61	1.00	0.61	Scleractinia	Poritidae
Echinophyllia	1	0.61	1.00	0.61	Scleractinia	Lobophylliidae
Galaxea	1	0.61	0.99	0.61	Scleractinia	Euphylliidae
Oulophyllia	1	0.61	1.00	0.61	Scleractinia	Merulinidae
Stereocidaris	1	0.60	1.00	0.60	Cidaroida	Cidaridae
Umbellula	2	0.66	0.49	0.32	Pennatulacea	Umbellulidae
Stylaster	2	0.75	0.43	0.32	Anthoathecata	Stylasteridae
Platidia	2	0.45	0.43	0.20	Terebratulida	Platidiidae
Bathypathes	3	1	0.45	0.45	Antipatharia	Schizopathidae
Styela	4	1	0.54	0.54	Stolidobranchia	Styelidae
Cornucopina	5	0.85	0.41	0.35	Cheilostomatida	Bugulidae
Aerothyris	5	0.22	0.47	0.11	Terebratulida	Terebratellidae
Cnemidocarpa	6	1.00	0.34	0.34	Stolidobranchia	Styelidae
Hyalonema	7	1.00	0.29	0.29	Amphidiscosida	Hyalonematidae
Astrotoma	8	1	0.76	0.76	Euryalida	Gorgonocephalidae
Gorgonocephalus	9	1	0.63	0.63	Euryalida	Gorgonocephalidae

Table C3. Species level bioregionalisation indicator values.

species	*cluster*	*A*	*F*	*IndVal*	*order*	*family*
Pavona venosa	1	0.67	1.00	0.67	Scleractinia	Agariciidae
Leptoseris mycetoseroides	1	0.64	0.99	0.64	Scleractinia	Agariciidae
Echinopora hirsutissima	1	0.64	1.00	0.63	Scleractinia	Merulinidae
Favites complanata	1	0.62	1.00	0.62	Scleractinia	Merulinidae
Acanthastrea echinata	1	0.62	1.00	0.62	Scleractinia	Lobophylliidae
Pavona explanulata	1	0.61	1.00	0.61	Scleractinia	Agariciidae
Echinopora gemmacea	1	0.70	0.87	0.61	Scleractinia	Merulinidae
Porites lobata	1	0.61	1.00	0.61	Scleractinia	Poritidae
Leptoseris hawaiiensis	1	0.60	1.00	0.60	Scleractinia	Agariciidae
Galaxea fascicularis	1	0.60	1.00	0.60	Scleractinia	Euphylliidae
Porites lutea	1	0.60	1.00	0.60	Scleractinia	Poritidae
Danafungia horrida	1	0.60	1.00	0.60	Scleractinia	Fungiidae
Pocillopora verrucosa	1	0.60	1.00	0.60	Scleractinia	Pocilloporidae
Porites rus	1	0.60	1.00	0.60	Scleractinia	Poritidae
Bathypathes patula	2	1	0.85	0.85	Antipatharia	Schizopathidae
Astrotoma agassizii	3	1	0.82	0.82	Euryalida	Gorgonocephalidae
Platidia anomioides	4	1	0.30	0.30	Terebratulida	Platidiidae
Aerothyris kerguelensis	5	1	0.46	0.46	Terebratulida	Terebratellidae
Gorgonocephalus chilensis	6	1	0.69	0.69	Euryalida	Gorgonocephalidae
Deltocyathus magnificus	7	1	0.21	0.21	Scleractinia	Deltocyathidae
Antarctotetilla leptoderma	8	1	0.53	0.53	Tetractinellida	Tetillidae

Appendix D

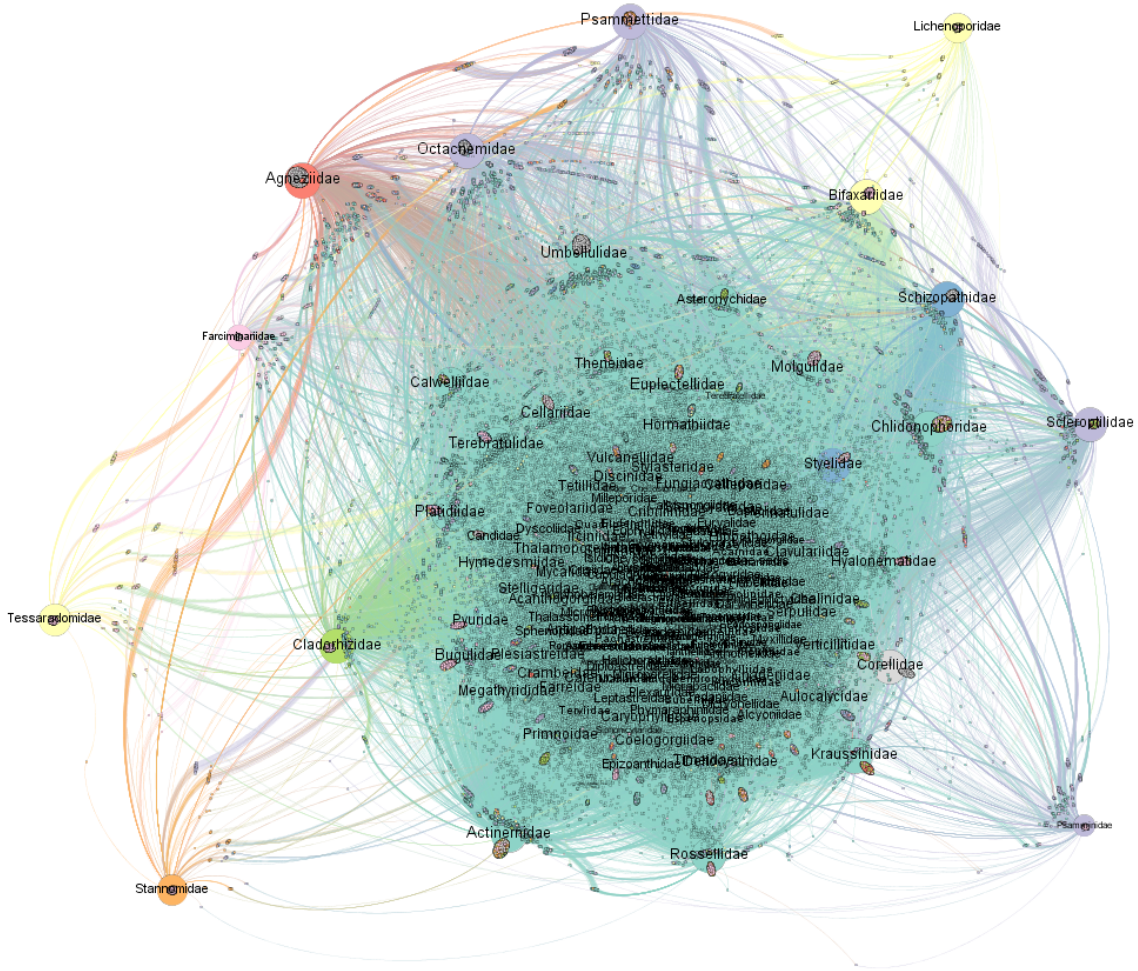

Figure D1. Biogeographical network at family level with no occurrence threshold.

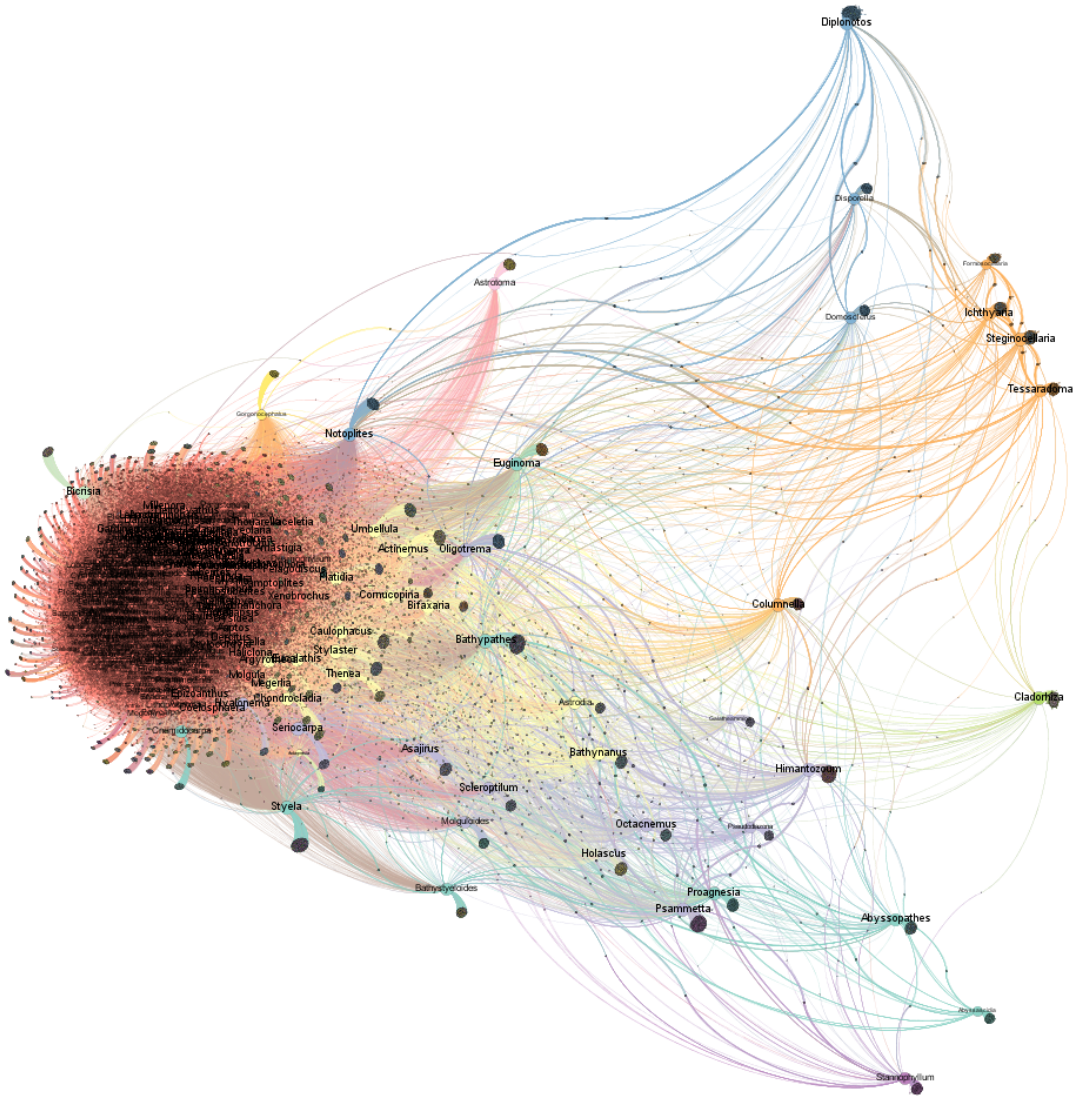

Figure D2. Biogeographical network at genus level. No threshold of occurrences.

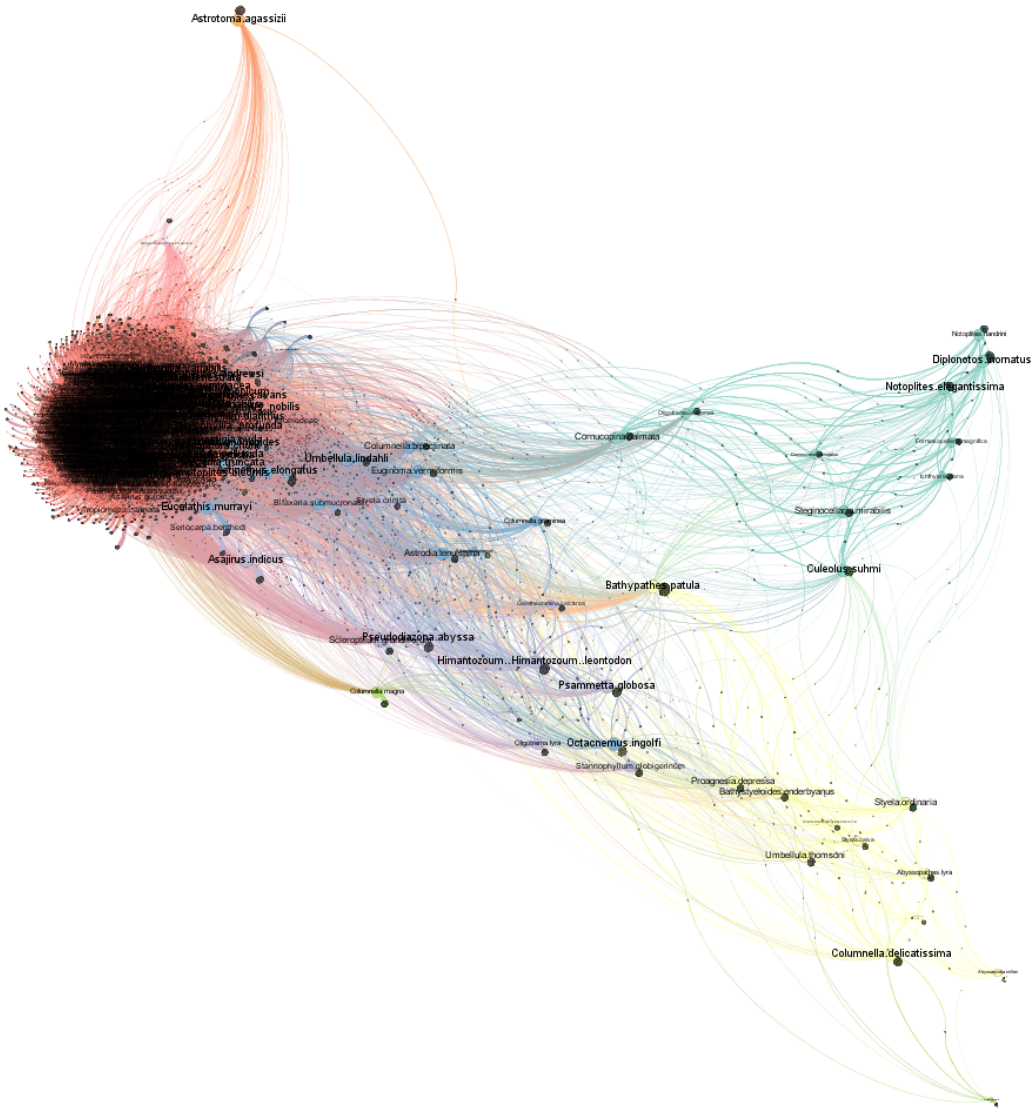

Figure D3. Biogeographical network at species level. No threshold of occurrences.

Table D1. Number of families and endemicity of each biogeographical region for the family-based bioregionalisation with no threshold of occurrences.

Cluster	Total richness	Characteristic richness	Endemic richness	% Endemic families	% Endemic families Southern Indian Ocean
1	210	196	76	0.36	0.93
2	103	4	0	0.00	0.04
3	69	3	0	0.00	0.04
4	99	2	0	0.00	0.02
5	77	1	0	0.00	0.01
6	76	1	0	0.00	0.01
7	50	1	0	0.00	0.02
8	71	1	0	0.00	0.01
9	33	1	0	0.00	0.03

Table D2. Number of genera and endemicity of each biogeographical region for the genus-based bioregionalisation with no threshold of occurrences.

Cluster	Total richness	Characteristic richness	Endemic richness	% Endemic species	% Endemic species of the Southern Indian Ocean
1	427	378	88	0.21	0.89
2	221	17	0	0.00	0.08
3	159	11	0	0.00	0.07
4	130	8	0	0.00	0.06
5	81	5	0	0.00	0.06
6	58	4	0	0.00	0.07
7	54	1	0	0.00	0.02
8	72	1	0	0.00	0.01
9	24	1	0	0.00	0.04
10	46	1	0	0.00	0.02
11	57	1	0	0.00	0.02
12	57	1	0	0.00	0.02

Table D3. Number of species and endemicity of each biogeographical region for the species-based bioregionalisation with no threshold of occurrences.

Cluster	Total richness	Characteristic richness	Endemic richness	% Endemic species	% Endemic species Southern Indian Ocean
1	697	647	404	0.58	0.93
2	224	16	0	0.00	0.07
3	163	11	0	0.00	0.07
4	129	11	1	0.01	0.09
5	96	10	0	0	0.10
6	41	1	0	0	0.02
7	60	2	0	0	0.03
8	92	1	0	0	0.01
9	43	1	0	0	0.02