Determining marine bioregions: A comparison of quantitative approaches

Authors

Nicole Hill[1]*, Skipton N.C Woolley[2], Scott Foster[2], Piers K. Dunstan[2], John McKinlay[3], Otso Ovaskainen[4] and Craig Johnson[1]

[1] Institute for Marine and Antarctic Studies, University of Tasmania, Hobart, Tasmania, Australia

[2] Commonwealth Scientific and Industrial Research Organisation (CSIRO), Hobart, Tasmania, Australia

[3] Australian Antarctic Division, Kingston, Tasmania, Australia

[4] Organismal and Evolutionary Biology Research Programme, University of Helsinki, Helsinki, Finland

*Corresponding author email: Nicole.Hilll@utas.edu.au

Running Headline: Comparing approaches for quantifying marine bioregions

Abstract

1. Areas that contain ecologically distinct biological content, called bioregions, are a central component to spatial and ecosystem-based management. We review and describe a variety of commonly-used and newly-developed statistical approaches for quantitatively determining bioregions.
2. Statistical approaches to bioregionalisation can broadly be classified as two-stage approaches that either 'Group First, then Predict' or 'Predict First, then Group', or a newer class of one-stage approaches that simultaneously analyse biological data with reference to environmental data to generate bioregions. We demonstrate these approaches using a selection of methods applied to simulated data and real data on demersal fish. The methods are assessed against their ability to answer several common scientific or management questions.
3. The true number of simulated bioregions was only identified by both of the one-stage methods and one two-stage method. When the number of bioregions was known, many of the methods, but not all, could adequately infer the species, environmental, and spatial characteristics of bioregions. One-stage approaches however, do so directly via a single model without the need for separate post-hoc analyses and additionally provide an appropriate characterisation of uncertainty.
4. One-stage approaches provide a comprehensive and consistent method for objectively identifying and characterising bioregions using both biological and environmental data. Potential avenues of future development in one-stage methods include incorporating presence-only and multiple data types as well as considering functional aspects of bioregions.

Key words:

Bioregionalisation, ecoregionalisation, biogeography, ecological statistics, community ecology

1. Introduction

As human pressures on natural systems increase, understanding and predicting the distribution of biodiversity has become vital for managing marine habitats. One important task is to define coherent and ecologically meaningful spatial units that can aid in planning, evaluating and implementing spatial management options. These spatial units are particularly useful for a diverse range of applications including: designing monitoring efforts, managing human activities (especially in marine protected area designation) and informing the relative scales required for ecosystem based assessments (Koubbi *et al.* 2011; Baker & Hollowed 2014; Rose *et al.* 2016; Hill *et al.* 2017; Stephenson *et al.* 2018; Koen-Alonso *et al.* 2019). This task requires identifying where different groups of species, or distinct assemblages, are found and has been variously termed ecoregionalisation, biogeographic classification, ecological mapping, and bioregionalisation (Woolley *et al.* 2019). Here we use 'bioregionalisation' to describe the process of identifying individual 'bioregions' which are geographic regions that are relatively homogeneous and distinct in terms of their biological contents.

Bioregionalisation is not a new concept. Early marine bioregionalisations drew on data from limited biological collections and expert knowledge to draw spatial boundaries (Ekman 1953; Hedgpeth 1957) and many global or large-scale bioregionalisations still rely heavily on the input of expert knowledge in various forms (GOODS UNESCO (2009), MEOW Spalding *et al.* (2007)). Since the widespread availability of remotely-sensed data, many bioregionalisations have used statistical methods to classify environmental data into distinct groups (Raymond 2014; Roberson *et al.* 2017; Sayre *et al.* 2017). The assumption underlying this approach is that different environments are representative of distinct habitats and should contain different assemblages of species, thus reflecting biogeographic patterns. However, evidence supporting this assumption is equivocal (Rickbeil *et al.* 2013; Ware *et al.* 2018). Where a reasonable amount of biological data exists for a region of interest, an alternative, and arguably more representative, approach is to explicitly

incorporate it into a quantitative analysis. Quantitative, biologically-derived bioregions incorporate patchy biological data into statistical models that directly relate the distribution and abundance of multiple species to broader-coverage environmental data (Rubidge, Gale & Curtis 2016; Hill *et al.* 2017; Woolley *et al.* 2019). We refer readers to (Woolley *et al.* 2019) for a detailed discussion on the current state of marine bioregionalisation. Presently, quantitative bioregionalisations that explicitly incorporate biological data are most feasible at small to large regional scales. We also note that while we focus on marine systems, terrestrial bioregionalisation and vegetation classification have undergone analogous evolution (Köppen 1884; Lyons, Foster & Keith 2017) and most of the concepts and analytical approaches that we discuss are applicable to terrestrial systems.

Analytical approaches to bioregionalisation that incorporate both biological and environmental data can broadly be classified into two-stage or one-stage approaches (Woolley *et al.* 2019). Two-stage approaches are most common, in which either biological groups are first determined and then related to their environment ('Group First, then Predict') or species are related to their environment and then biological groups identified ('Predict First, then Group'). Ferrier and Guisan (2006) provide a definition of these approaches in a related setting. Within the 'Predict First, then Group' approach, methods that predict the turnover in community composition (beta diversity) rather than the species themselves are becoming increasingly popular for bioregionalisation (Ferrier *et al.* 2007; Leaper *et al.* 2011; Ellis, Smith & Pitcher 2012; Stephenson *et al.* 2018). The introduction of models that jointly predict multiple species distributions, but not bioregions *per se* (e.g. Warton *et al.* (2015a); Ovaskainen *et al.* (2017)) with reported superiority in predicting community-level patterns (Norberg *et al.* 2019) also advance two-stage methods. In a one-stage approach, biological groups and their relationship with the environment are defined in a single model (i.e. analysed simultaneously), and various implementations of this have recently become available (ter Braak *et al.* 2003; Dunstan, Foster & Darnell 2011; Foster *et al.* 2013). Noted advantages of one-stage approaches are the direct ecological interpretation of bioregions and appropriate characterisation of

uncertainty in the distribution of bioregions (Hill *et al.* 2017; Lyons *et al.* 2017; Fiorentino, Lecours & Brey 2018).

As many jurisdictions are moving rapidly toward implementation of marine spatial planning and ecosystem approaches to management that require bioregion information as a key input (e.g. Koen-Alonso *et al.* (2019)), it is timely to review recent methodological developments for bioregionalisation. We categorise a range of modelling approaches available for bioregionalisation into one of the three approaches listed above and apply a selection of methods to simulated data and a more complex, real dataset of occurrences of demersal fishes on the Kerguelen Plateau.

We demonstrate each of the approaches and focus our comparison on how the approaches answer five core questions that allow ecologists and managers to interpret and use bioregionalisations:

i) How many bioregions are there?
ii) What is the spatial distribution of each bioregion across our region of interest?
iii) What species characterise these bioregions?
iv) What are the environmental characteristics of each bioregion?
v) How certain are we about the distribution of bioregions and their composition?

We acknowledge that for particular applications other aspects, such as spatial scale and coherence, may also be relevant, but do not consider them in detail here. We then explore the advantages and disadvantages of the approaches from a statistical and ecological viewpoint. Finally, we discuss future research directions for statistical approaches to bioregionalisation.

2. Materials and Methods

2.1 Categorising quantitative approaches to bioregionalisation

Quantitative approaches are categorised as two-stage if they separate the two components of bioregionalisation (i.e. identifying biological groups and relating biology to environmental characteristics), and one-stage if they delineate *bioregions* based on a *simultaneous* use of biological

data and their relationship to environment. Two-stage approaches can be further divided according to which component occurs first, and by how the biological components are modelled. Here we give an overview of the approaches (Fig. 1) and their ability to address key questions (Table 1). We very briefly introduce the methods selected for comparison and refer readers to Appendix 1 for a comprehensive description of each method.

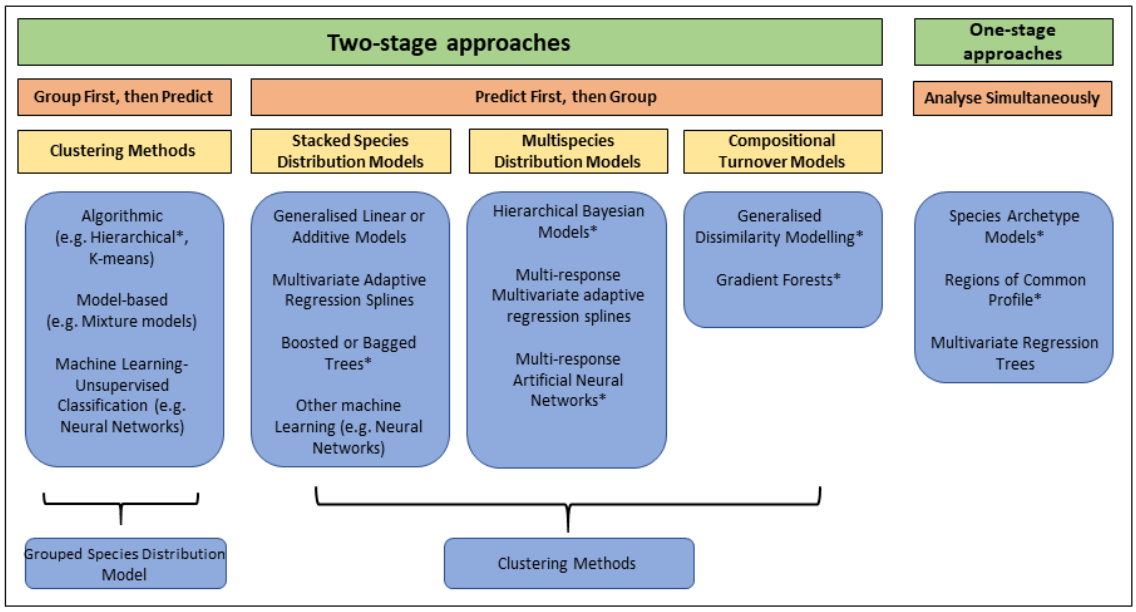


Fig. 1. Conceptual framework for bioregionalisation approaches and a selection of approaches that fall within each of these categories. Asterix (*) indicate methods compared in this paper and described in detail in Appendix 1.

2.1.1. Two-Stage Analyses: Group First, then Predict

In a 'Group First, then Predict' approach, biological data at sampled sites are first clustered to represent groups of relatively homogenous species composition, and these groups are secondarily related to environmental data. The first stage (clustering) addresses how many groups or bioregions can be defined. While there are many approaches to clustering data (Kaufman & Rousseeuw 1990), here we focus on hierarchical clustering because it is a popular approach used by ecologists (e.g. Rubidge *et al.* (2016); Bloomfield, Knerr and Encinas-Viso (2018)). Similarly, many metrics are available to determine the optimal number of clusters and we use the popular metric, average

131 silhouette width (Rousseeuw 1987). The second stage relates thegroups into which each site has
132 been clustered toenvironmental data to allow prediction of bioregions. This is typically done using a
133 single model for each group (e.g. Cooper *et al.* (2019)), or by using a multinomial technique (e.g.
134 Rubidge *et al.* (2016)). Here for the second stage of our analysis we used Random Forests (RF;
135 Breiman (2001), a method that produces an ensemble of classification trees, because it generally has
136 a high predictive power and is becoming increasingly popular for single species distribution
137 modelling. Characterising the bioregions produced by the 'Group First, then Predict' approaches
138 usually involves generating summary statistics from the clustered site data (Table 1). Note that
139 many methods for predicting bioregions can produce estimates of uncertainty, but these only
140 represent a portion of the variability in the analysis as they do not account for variability in
141 clustering. This includes the methods demonstrated in this work.

142 **2.1.2. Two-Stage Analyses: Predict First, then Group**

143 Under a 'Predict First, then Group' approach, the distribution of individual species or a
144 representation of community turnover is modelled and predicted across the region of interest, and
145 these predictions are subsequently clustered to represent bioregions. We divide this approach into: i)
146 stacked species distribution models, which model each species independently and then compile
147 ('stack') predictions to generate species composition for each prediction cell (Norberg *et al.* 2019); ii)
148 multi- species distribution models , that jointly model and predict the distribution of multiple species
149 at once (Ovaskainen *et al.* 2017); and iii) community turnover approaches, which depict how the
150 composition of communities change through space as a function of the environment (Ferrier *et al.*
151 2007; Ellis *et al.* 2012). Common to all 'Predict First, then Group' approaches, the number of groups
152 and their spatial distribution is determined in the second stage of the analysis by clustering the
153 predicted species composition, turnover of species composition or transformed environmental
154 space at cells in the region of interest. Like the 'Group First, then Predict' approaches, the species
155 and environmental characteristics of groups are usually determined by summarising classified site
156 data and uncertainty is only characterised for one of the stages (Table 1).

In the stacked species distribution approach, there are a multitude of methods for modelling single species distributions ranging from variations on linear and generalised linear models (GLM) to a vast array of machine learning approaches. We use Random Forests to model the distribution of each species individually because of the advantages noted above and to facilitate fair comparisons between the approaches.

In the multi-species distribution approach we use the recently-developed, Bayesian Joint Species Distribution modelling framework called Hierarchical Modelling of Species Communities (HMSC; (Ovaskainen *et al.* 2017)). This framework is built upon multi-response generalised linear models (GLMs) and has shown promise for a number of distribution modelling applications (Ovaskainen *et al.* 2017). Our implementation uses latent variables to account for spatially structured species' co-occurrences and enhance spatial prediction capacity. We also use a recently- developed multi-species implementation of the machine learning method Artificial Neural Networks (Mistnet, Harris (2015)) as neural networks are inherently able to model complex and non-linear relationships and interactions, a counter point to HMSC which is based on GLMs, and have been shown to have good predictive ability.

Of the compositional turnover (beta diversity) approaches, we used the popular Generalised Dissimilarity Modelling (GDM) and Gradient Forests (GF) methods. In GDM a pairwise biological dissimilarity metric (e.g. Jaccard) is modelled as the response variable and the corresponding site-wise differences in each of the environmental variables as the predictor variables in a regression spline GLM (Ferrier *et al.* 2007). Spatial predictions are made by transforming the environmental differences between pairs of prediction cells using the function identified by the GDM model and processing the outputs as described in section 3.1.1. Recently a bootstrapped version of GDM (bbGDM) has been developed to account for that fact that pairwise dissimilarities are not independent and violate the assumptions of GLMs (Woolley *et al.* 2017). Gradient Forests aggregate information from single-species Random Forests to build functions of how species composition

changes along environmental gradients (Ellis *et al.* 2012). Predictions are made by transforming the environmental covariates at all cells across the region of interest using these functions followed by clustering.

2.1.3. One-stage Analyses

In a One-Stage approach to bioregionalisation, biological *groups* and their relationship with environmental data are defined in a single model or analysed simultaneously. This means that groups (and their associated species composition) can be directly predicted across the region of interest with measures of uncertainty that encapsulate the entire analytical process. Also, the species composition and environmental characteristics of groups are derived directly from model parameters (Woolley *et al.* 2013; Leaper *et al.* 2014; Hill *et al.* 2017). As opposed to silhouette width or other discrimination metrics, the number of groups within one-stage approaches is currently chosen based on the model likelihood, using the Bayesian Information Criterion (BIC). Limited methods are available for one-stage approaches, which currently include Species Archetype Models (SAMs; Dunstan *et al.* (2011); Dunstan *et al.* (2013)), Regions of Common Profile models (RCPs; Foster *et al.* (2013) but also see ter Braak *et al.* (2003)) and Multivariate Regression Trees (MRTs; De'ath (2002) and Appendix 1). Here we focus on SAMs and RCPs that are both types of finite mixture models. This means that they can both handle data with non-constant mean-variance relationships (e.g. abundance data; Warton *et al.* (2015b)). The difference between SAMs and RCPs is that SAMs form *groups of species* based on the species' responses to environmental data (Dunstan *et al.* 2011), whereas RCPs *group sites* and model those sites grouping as a function of the environment data (Foster *et al.* 2013).

2.1.1. Comparison of methods using simulated and real data:

In this section, we run a selection of methods for bioregionalisation on a simulated and a real dataset. For the simulated data, we generated eight environmental variables across a hypothetical region of interest. We randomly assigned thirty species exclusively to one of three groups. These

groups responded to two of the eight environmental variables (temperature and oxygen, Table A3.1, Fig. A3.2) and we refer to the spatial distributions of these groups as the 'true' distributions. These data were designed to generate a distinct bioregional pattern with minimal spatial overlap between groups (see Figs. A3.3-5). Presence-absence data were randomly drawn from the probability of occurrence for each of the 30 species at 200 sites (a subset of all sites) and used as the biological data input for all methods. Further details for the simulation process are given in Appendix 3.

Our real dataset consisted of the presence-absence of demersal fish recorded in 524 trawls from random stratified surveys conducted during 2006, 2010 and 2013 on the Kerguelen Plateau in the Southern Indian Ocean. For illustration purposes, the 20 species that occurred in at least 10 trawls were retained for analyses. Eight environmental variables representing seafloor (e.g. depth) and sea surface (e.g. chlorophyll-a) conditions likely to affect the distribution of demersal fish were sourced at a 0.1 degree resolution. Details on the demersal fish and associated environmental data are in Appendix 4.

We compared nine modelling methods spread across the three broad modelling approaches discussed above. The approaches and the way that they answer our five key bioregionalisation questions are outlined in Table 1. As a final comparison we clustered the environmental data directly, representing bioregionalisations that do not incorporate biological data. Overall, we tried to ensure as much consistency as possible amongst the analysis steps for the different approaches to enable a fair comparison. Implementation details for each method and the derivation of comparison plots and statistics are in Appendix 2. R code to run the analyses for both the simulation and demersal fish data are provided in the supplementary material.

3. Results

3.1. Simulated Data

3.1.1. How many bioregions are there and what is their spatial distribution?

The type of bioregion outputs produced by the different methods can be described as either hard-class, where each site is assigned uniquely to a particular bioregion, or probabilistic, where each site has some chance of belonging to more than one bioregion. The three 'true' bioregions derived from the simulated data have a probabilistic and distinct spatial distribution (Fig. 2a). Two-stage approaches that used hierarchical clustering in the second stage produce hard classes (Fig. 2b), while the one-stage approaches have a probabilistic output (fig. 2c).

Most two-stage methods identified two bioregions as optimal (Fig. 2b,c). The exceptions were the Hierarchical Bayesian Model (HMSC_HC) and multi-response neural network (MNet_HC) 'Predict First, then Cluster' methods, where three and five bioregions respectively were selected as optimal.

Most methods that identified two bioregions discriminated bioregions 1 and 3 but did not distinguish bioregion 2. There are several options for presenting the outputs of naïve and bootstrapped GDM models (Ferrier *et al.* 2007). Here we cluster the predicted cell-wise dissimilarities directly (Fig. 2b, GDM_Dissim_HC) as well as the environmental space which has been transformed using the GDM model's spline functions (Fig. 2b, GDM_TransEnv_HC). The latter is most comparable to the Gradient Forest approach. In this instance, the overall pattern in the distribution of bioregions is similar using either technique. Clustering the environmental data directly, and without any biological information, results in 11 bioregions whose distribution looks like a 'patchwork quilt' and does not resemble the distribution of 'true' bioregion distributions (Fig. 2a). While the two-stage BioHC_RF method produces a probabilistic output that broadly distinguishes two groups, it has an increased degree of patchiness in predictions compared to the methods with

hard-class outputs (Fig. 2c). There are no estimates of uncertainty of the entire analysis process for BioHC_RF though – these probabilistic maps are only for the second stage.

The one-stage approaches produce probabilistic outputs of the entire analytical process and distinguish three bioregions that largely correspond to the 'true' distribution of bioregions. It should be noted however, that the model used to simulate bioregions most closely resembles the SAMs model. The RCP method predicts distinct groups that have a high probability of occurrence and do not overlap except at the boundaries of the bioregions. The SAM method produces bioregions with a lower probability of occurrence with less distinct boundaries between groups (Fig. 2c). This results from a fundamental difference in philosophy and implementation of the SAM and RCP methodologies; SAM models groups of *species* with a common response to the environmental data, while RCP models groups of *sites* with a common species composition and environmental profile. Therefore, often more than one SAM group is likely at a location. The RCP model assumes that there is a single assemblage type at each location, and the model is trying to find that type.

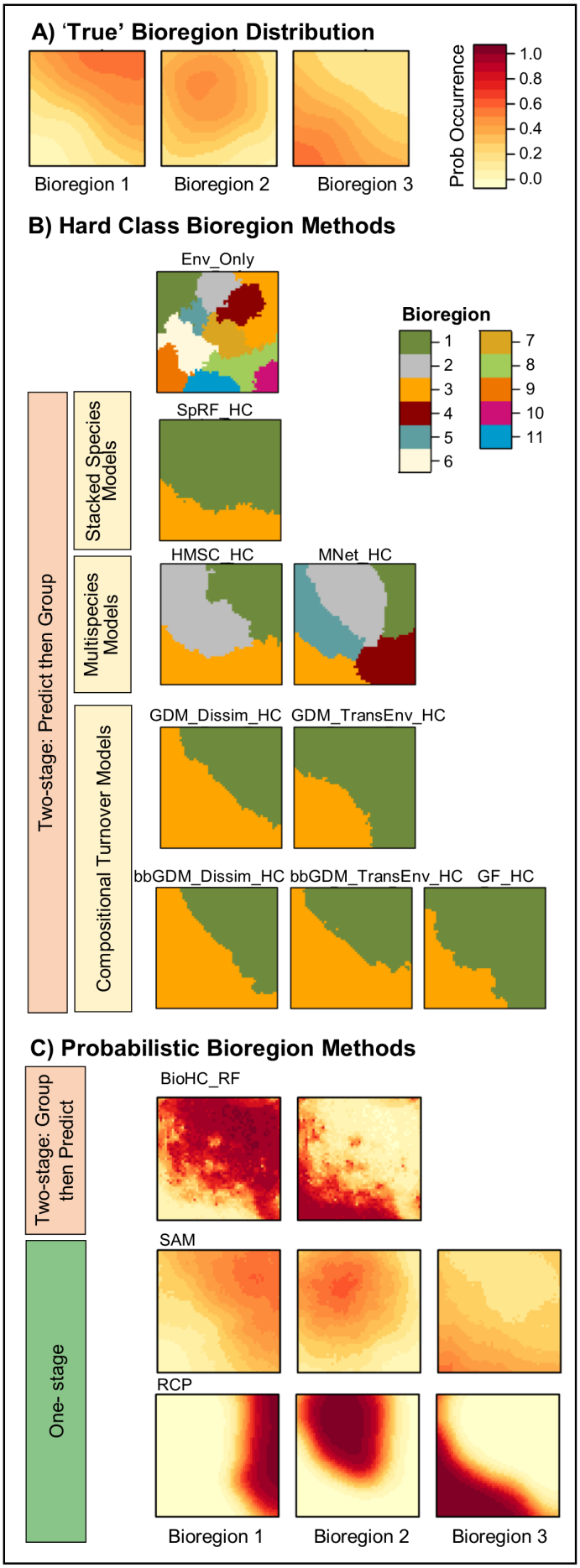
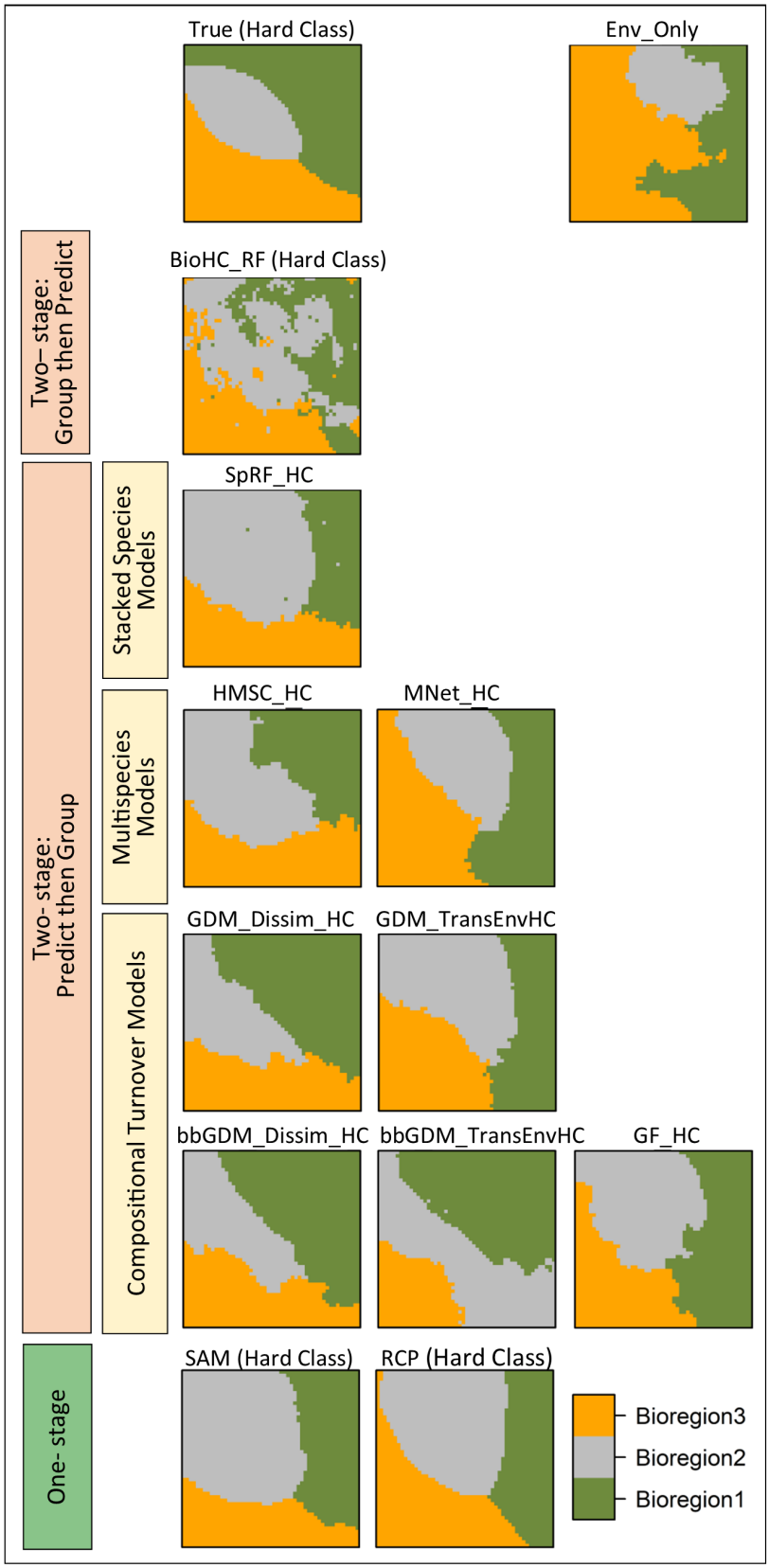

A) 'True' Bioregion Distribution
Prob Occurrence
1.0
0.8
0.6
0.4
0.2
0.0
Bioregion 1
Bioregion 2
Bioregion 3
B) Hard Class Bioregion Methods
Env_Only
Bioregion
1
2
3
4
5
6
7
8
9
10
11
Two-stage: Predict then Group
Stacked Species Models
SpRF_HC
Multispecies Models
HMSC_HC
MNet_HC
Compositional Turnover Models
GDM_Dissim_HC
GDM_TransEnv_HC
bbGDM_Dissim_HC
bbGDM_TransEnv_HC
GF_HC
C) Probabilistic Bioregion Methods
Two-stage: Group then Predict
BioHC_RF
One- stage
SAM
RCP
Bioregion 1
Bioregion 2
Bioregion 3

Fig. 2. The number and spatial distribution of bioregions selected as optimal for each method. A) The 'true' bioregion distribution from the simulation. Colour ramp corresponds to probability of occurrence. B) Hard-classes resulting from hierarchical classification in the 'Predict First, then Group' methods. Groups are colour-coded to reflect the best match to the 'true' bioregions. C) Probability of occurrence for one-stage and 'Group First, then Predict' methods. Only the one-stage methods (SAM and RCP) and the two-stage method HMSC_HC correctly identify the number of bioregions and their approximate distribution. Note that the BioHC_RF probabilities represent only the second-stage of the analysis. Acronyms match those in Table 1.

The number of bioregions chosen as optimal has a large influence on the results of the different approaches. For the remainder of the simulation results, we remove this influence and assume we know there are three bioregions (Fig. 3). Approaches that produce probabilistic outputs were converted to hard-class bioregions by assigning each cell its most probable bioregion. When the number of bioregions was fixed at three, the distribution of groups in many of the approaches bears strong resemblance to the simulated true number of bioregions. Nearly all methods overestimate the spatial extent of bioregion 2. The clustering of the environment alone (Env_Only), divides bioregions 1 and 3 into an E-W direction and displaces the distribution of bioregion 2 to the NE of its true region. The 'Group First, then Predict' method (BioHC_RF) again produces groups with a patchier distribution than other methods. Clustering the spline transformed environmental space from the bootstrapped GDM model (bbGDM_TransEnvHC) produces a different spatial pattern, although investigations (not shown here) using many different starts and numbers of bootstraps produced one of two contrasting patterns suggesting some instability in the model or influential sites even after many bootstraps.



286

287 **Fig. 3. Distribution of simulated bioregions for each method when the number of bioregions has been fixed**

288 **at three and where cells for methods with probabilistic outputs are assigned their most likely bioregion**

(denoted by 'Hard Class'). The distribution of bioregions for most methods are more similar to the 'true' bioregions when the number of bioregions are known (except Env_Only, and the bbGDM methods). Bioregions have been colour-coded to best match the 'true' bioregions. Method abbreviations match Table 1.

3.1.2. How certain are we about the distribution of these bioregions?

The only methods that generate appropriate measures of uncertainty for predicted distributions of the bioregions are the one-stage approaches. For SAMs uncertainty is low overall with few consistent patterns between the bioregions, while for RCPs the highest uncertainties lie in the transition between areas of high and low predicted RCP bioregion probability (Fig. 4).

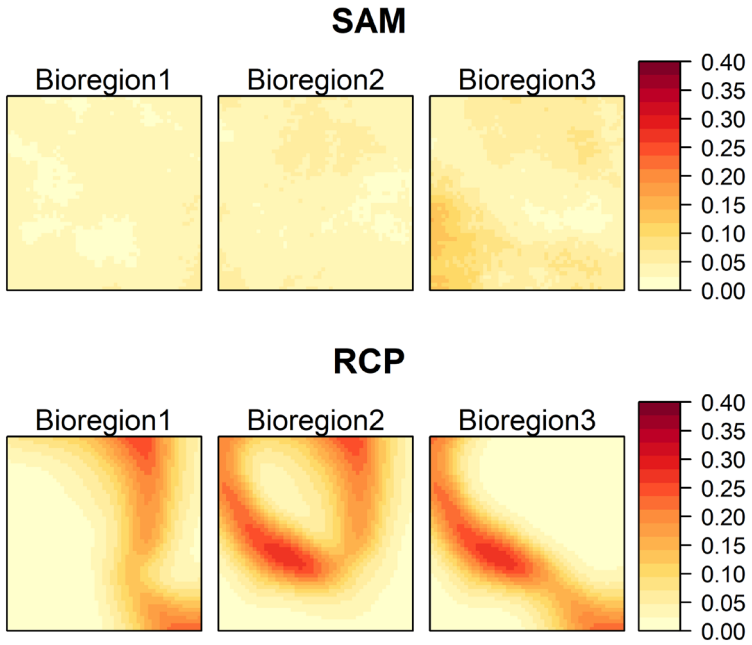


Fig. 4. Standard error of predicted probability of simulated bioregion occurrence for SAM and RCP, the only methods that appropriately characterise uncertainty for the entire bioregionalisation process. Uncertainty is low overall for SAM and for RCP is most uncertain along transitions between bioregions.

3.1.3. What is the species composition of each bioregion?

We derived the species composition of each bioregion for two-stage approaches by summarising the observed species' data at clustered survey sites. The one-stage models estimate species membership for each group directly and probabilistically. RCP models estimate parameters for the species composition of each RCP bioregion (its 'profile'), while SAMs estimate the relationships between species and the environment and thus do not directly provide species composition at site-based bioregions. For a fair comparison between approaches, we calculated average species responses for each of the three hard-class version of bioregions from the 'true' (simulated), one- and two-step approaches. In addition, for the RCP method we directly interpreted the model's estimated parameters (Fig. 5). Overall most methods recovered the 'true' tabulated distribution of species reasonably well, with some intra-method variability amongst the species composition (see Figs. 5 and A3.6 for all species). Surprisingly this included, the bbGDM_TransEnvHC that had a distinctly different spatial distribution of bioregions. The probability of species' occurrence estimates from the RCP model coefficients were slightly larger than the standard errors from the tabulated results.

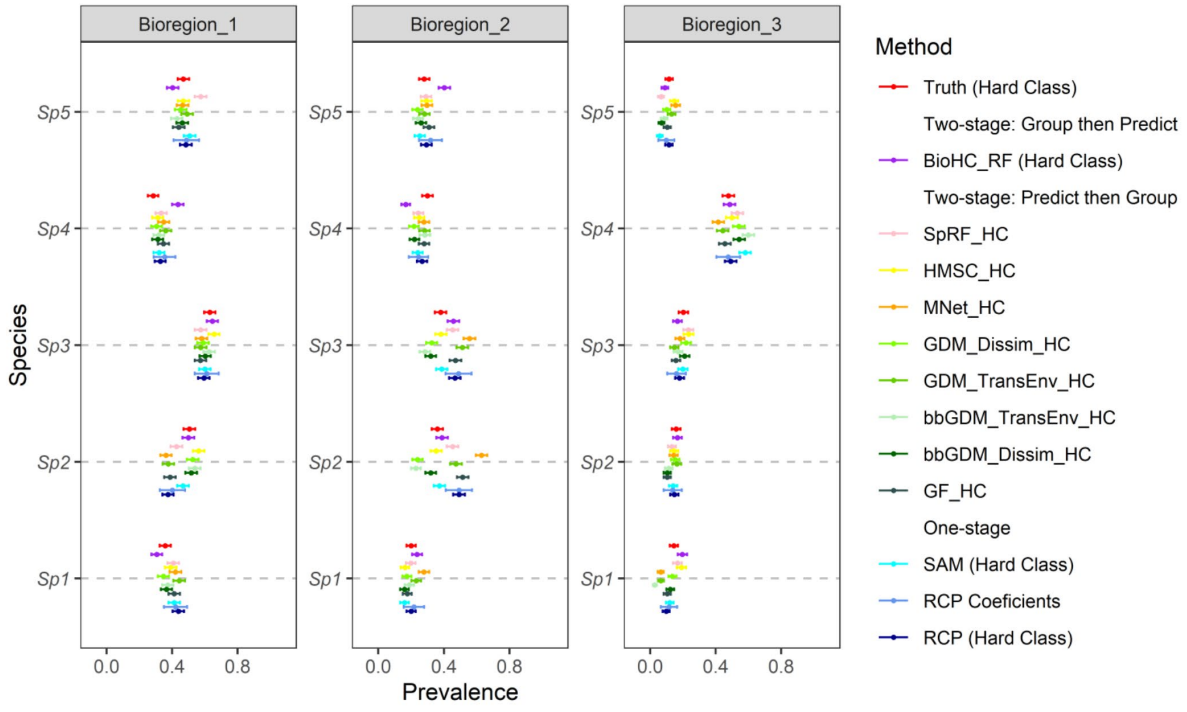


Fig. 5. Abridged species composition of each simulated bioregion, when the number of groups has been fixed to three. Most methods recovered the 'true' tabulated distribution of species reasonably well, with some intra-method variability amongst the species composition. Mean and standard error of the prevalence (or probability of occurrence for RCP Coeficients) for species 1- 5 (all 30 species are in Fig. A3.7). For comparative purposes, SAM, RCP and BioHC_RF results were also calculated using the hardclass conversion of the probability predictions (denoted "Hard Class")). Estimates directly from model parameters are also included for RCP (RCP Coefficients). Acronyms match Table 1.

3.1.4. What are the environmental characteristics of each bioregion?

Temperature and oxygen were the environmental variables that determined the simulated bioregions but all variables were considered in the models. Examination of the environmental characteristics for each bioregion *a posteriori* showed that each bioregion has a distinct combination of environmental values that are largely consistent with the truth (Fig. 6a, Fig. A3.7). The clustering of the environmental data only (Env_Only) is most different, followed by the bootstrapped GDM transformed environment (bbGDM_TransEnvHC, Fig. 6a, Fig. A3.7), reflecting differences in the spatial distribution of bioregions for these methods. For all two-stage approaches, the environmental characteristics of each bioregion were derived by summarising environmental covariates at clustered sites (Fig. 6a). In order to enable a fair comparison, we also calculated environmental characteristics for methods with probabilistic outputs ('true', BioHC_RF, SAM and RCP) by summarising the observed environmental conditions at sites assigned their most likely bioregion (denoted by "Hard Class").

The 'Group First, then Predict' and one-stage approaches provide additional information on the response of bioregions to each environmental variable in the form of partial response plots (see Figs. A3.8-11 and associated explanation). Responses of bioregions to simulated variables are largely in line with expectations from the simulation set up.

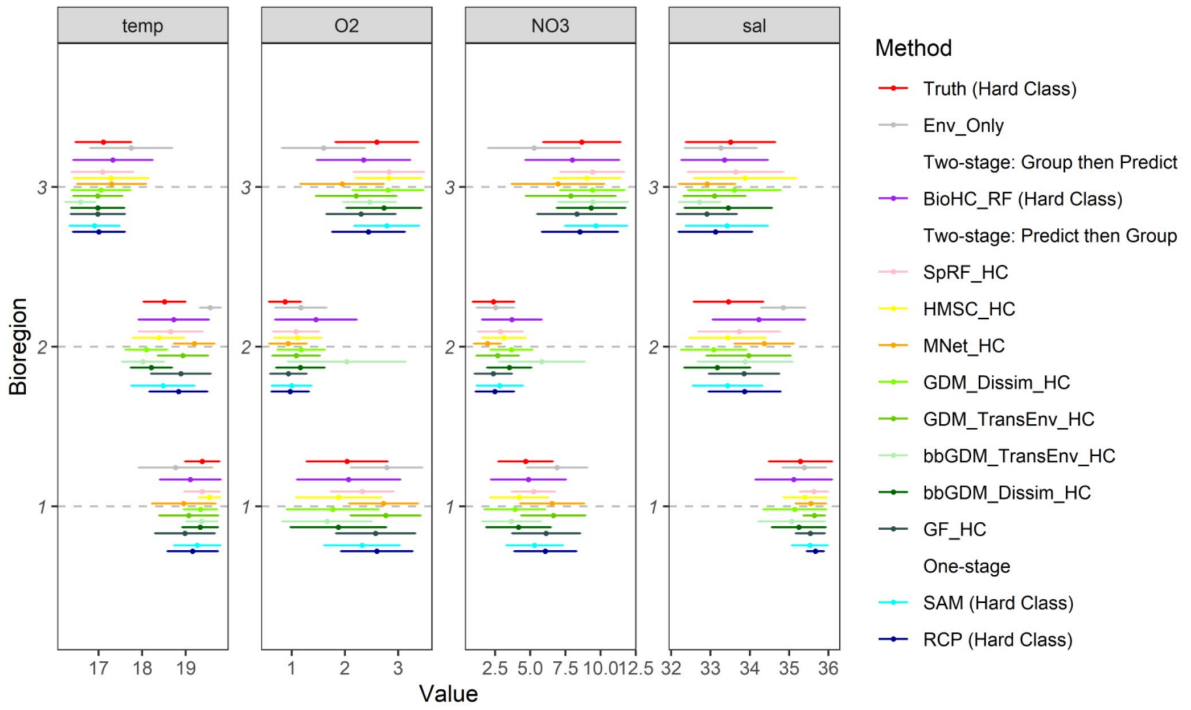


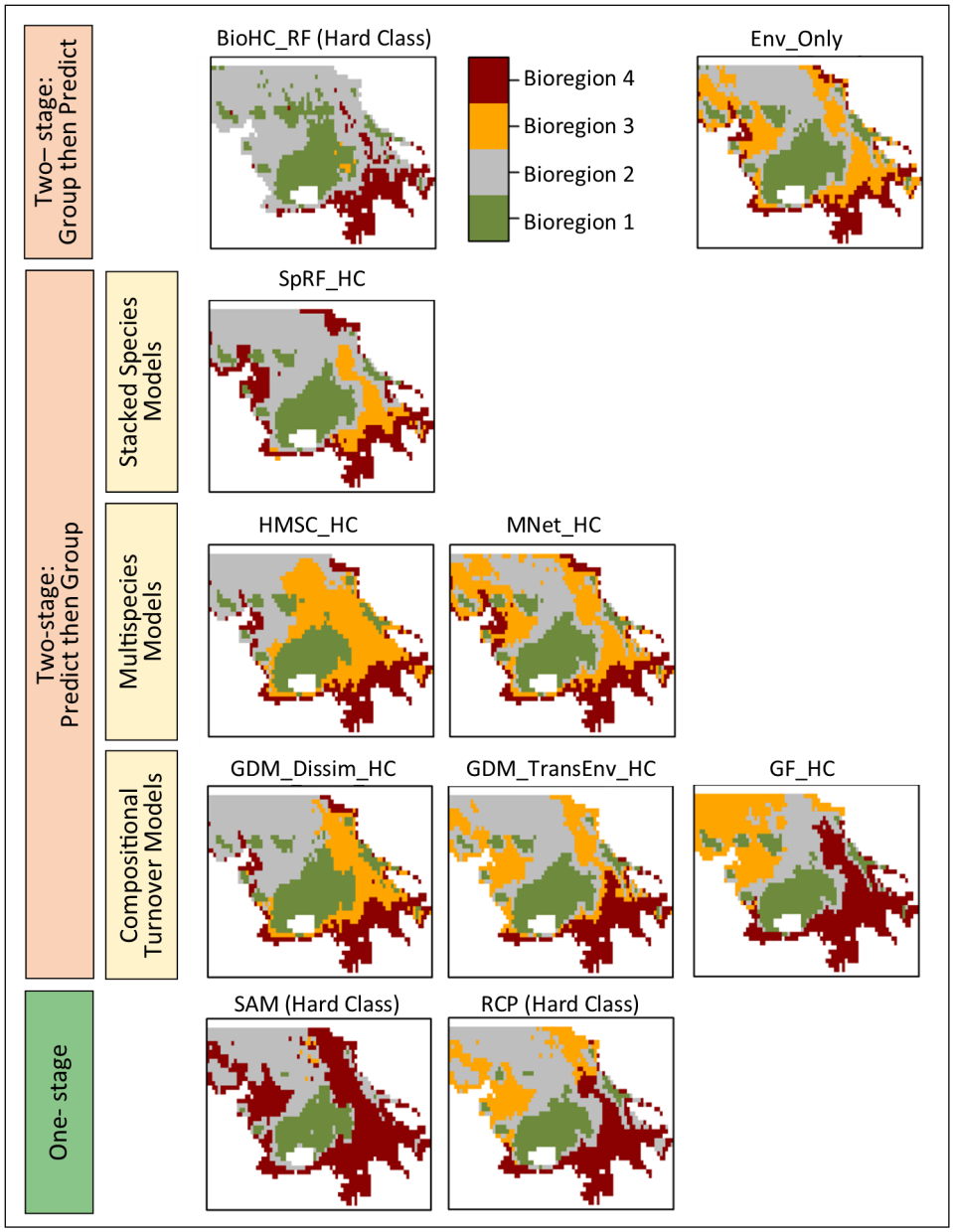
Fig. 6. A subset of the environmental characteristics of each simulated bioregion determined for each method when the number of groups is fixed at three. Each bioregion has a distinct combination of environmental values that are largely consistent with the truth with Env_Only and bbGDM_TransEnv_HC most different. Data are are the summarised from classified site data and representmean (+/- 1 SD) environmental conditions in each bioregion. For comparative purposes, SAM, RCP and BioHC_RF results were also calculated using the hard class conversion of the probability predictions (denoted "Hard Class").

3.2. Kerguelen Plateau demersal fish

Analysis of the Kerguelen Plateau demersal fish data yielded many similarities to the simulation results. (). Most of the two-stage methods discriminate two bioregions (exceptions were MNet_HC and both spline transformed GDM methods), while the one-stage methods SAM and RCP identify four and five bioregions respectively (Fig. A4.3). If we assume four bioregions for ease of comparison, then patterns in the spatial distribution of bioregions are more similar, including when clustering only the environmental data (Fig. 7, Fig. 8). Most methods consistently distinguished a shallow bioregion and a deep bioregion with varying boundaries for intermediate bioregions (Fig. 7). The

hard-class of SAM archetype probabilities in Fig. 7 results in three dominant bioregions, with the fourth spatially limited. This is because bioregion 3 has a low predicted probability of occurrence across the region which is more clearly seen in Fig. 8 and illustrates a key difference in the SAM compared to RCP methodology. RCP bioregions generally have a high and distinct probability of occurrence across the study region, whereas some of the SAM bioregions overlap and have moderate probability of occurrence (Fig. 8).

Patterns in the species associated with and environmental characteristics of each bioregion are complex. For many species and methods there is reasonable agreement in composition (Fig. A4.10) particularly for Bioregions 1 and 4. There was less agreement in the composition of Bioregions 2 and 3. Most methods distinguished depth bands for the different bioregions, while these appear to overlap more for SAM bioregions (Fig A4.6). Of all the methods the BioHC_RF most often stood apart from the others in its environmental characteristic. See Appendix 4 for additional results and their discussion.



369

370 **Fig. 7. Distribution of Kerguelen Plateau demersal fish bioregions for each method when the number of**
371 **bioregions has been fixed at four and where cells for methods with probabilistic outputs are assigned their**
372 **most likely bioregion (denoted by 'Hard Class').** The spatial distribution of bioregions are more similar
373 between methods when the number of bioregions are set to four and largely reflect depth-related patterns.
374 Bioregions have been colour-coded to best highlight similarities in the distribution of bioregions. Method
375 abbreviations match Table 1.

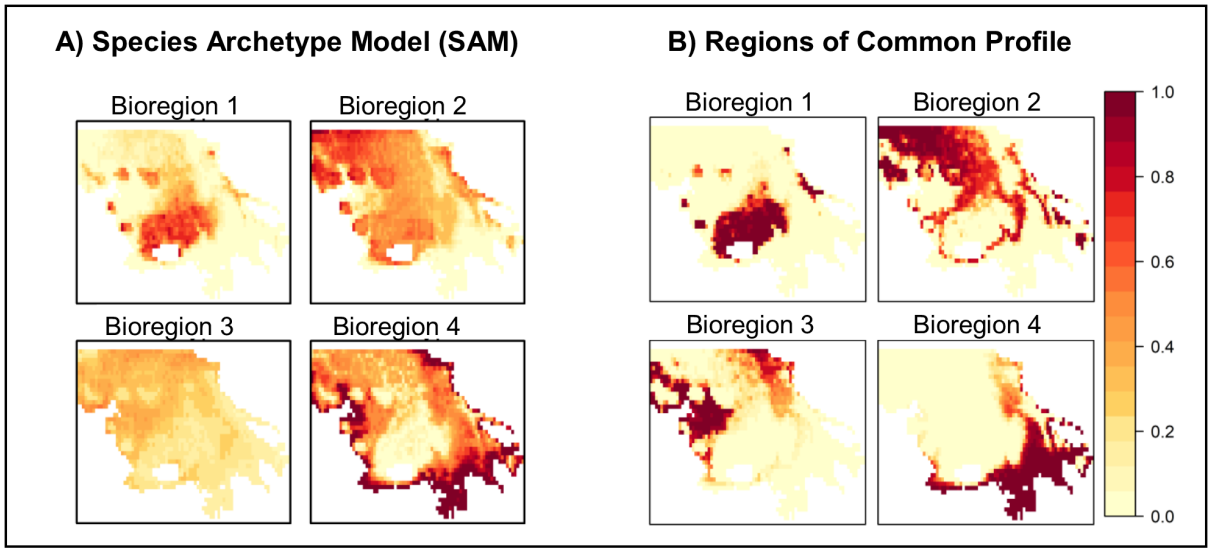


Fig. 8. Probabilistic predictions of the distribution of Kerguelen Plateau demersal fish bioregions from the one-stage methods, when the number of groups is set to four. A) Species Archetype Model (SAM) and B) Regions of Common Profile (RCP). Colour ramp indicated probability of bioregion presence. RCP bioregions generally have a high and distinct probability of occurrence across the study region, whereas some of the SAM bioregions overlap and have moderate probability of occurrence.

4. Discussion

We have categorised three broad approaches for quantitatively generating bioregions using both biological data and environmental data and compared some common and recently developed methods within each approach. We demonstrated that most methods could adequately delineate and characterise bioregions, but only if the number of bioregions was known. The exception is clustering only environmental data, which was unable to give any information on the expected species in each bioregion. In reality however, the optimal number of bioregions are not known *a priori* as they are not directly observed. In our simulation where we set the number of bioregions, only both of the one-stagemethods and one two-stage method correctly identified the true number of bioregions. We argue that in addition to correctly identifying the number of bioregions in the simulated data, one-stage approaches offer advantages over many of the other methods in terms of appropriately characterising uncertainty, direct interpretation, and transparency in what is actually

being modelled, and therefore represent a promising direction for quantitative bioregionalisation. We discuss some challenges and opportunities for bioregionalisation and one-stage methods into the future.

4.1. Advantages and Disadvantages of approaches

Most methods can provide answers to many of the questions ecologists and managers ask in order to identify, interpret and use bioregionalisations. In the two-stage approaches however, the clustering and prediction stages are decoupled from each other and often the original data which means that additional *post-hoc* analyses must be conducted to interpret the bioregions. For example, any method that uses a dissimilarity metric in either stage of the analyses loses the information about individual species needed to interpret the composition of bioregions. An advantage of one-stage approaches is that this information is recoverable directly via estimated model parameters, with variances that explicitly account for estimating bioregional groups (ter Braak *et al.* 2003; Foster *et al.* 2013). An important feature that sets one-stage approaches apart is that the mathematical model provides a formal definition of bioregions and their relationship with the environment. Thus, this explicitly provides transparency and repeatability (Warton *et al.* 2015b).

Currently only one-stage methods are able to appropriately quantify the uncertainty in the final bioregionalisation map.Many applications of two-stage approaches either do not consider uncertainty (Koubbi *et al.* 2011) or present 'final stage' uncertainty estimates that are incomplete or optimistic (Lasram *et al.* 2015; Rubidge *et al.* 2016) because of the difficulty in appropriately propagating uncertainty through both stages . Theoretically improvements to two-stage Bayesian models to allow uncertainty to propagate across stages are possible, but we are unaware of any current implementations in this context. In contrast, one-stage methods directly model and predict biological groups based on environmental data, explicitly quantifying the uncertainty in the predictions of the groupings themselves. Appropriate measures of uncertainty are important in many applications of bioregionalisation because they allow an assessment of risk associated with

applying (or not applying) spatial management to a location. These assessments are already standard in fisheries assessments (e.g. Koen-Alonso *et al.* (2019)) and are likely to become more important in bioregionalisation decision-making, where financial costs and biological costs are often traded to meet competing objectives.

There are several other trade-offs that may influence the method chosen to conduct bioregionalisation analyses (see Table .1). In terms of implementation, most methods have some model diagnostics, although only the diagnostics for the one-stage approaches are appropriate for the entire bioregionalisation process. Some methods (e.g. hierarchical clustering, random forests) are easy to implement within common software, while some of the newer methods, particularly the one-stage methods, require more investment. For example, currently optimising the parameters for the multi-response artificial neural networks and deriving the species' profiles for RCPs are not trivial. However, code is publicly available for these tasks and new packages are under development for facilitating the 'user-friendliness' of one-stage approaches. Similarly, some methods are relatively computationally intensive for moderate size datasets (Bayesian Bootstrapped GDM, Hierarchical Bayesian models, RCPs, artificial neural network). Finally, some methods, such as random forests and artificial neural networks, more naturally handle non-linear species responses to environmental variables and interactions, than methods based on GLMs which include the one-stage methods.

Focussing on the one-stage approaches used here, Species Archetype Models (SAM) and Regions of Common Profile (RCP), formulate and describe different types of groups that are useful for different applications. SAMs group species based on a similar environmental response, while RCPs group sites based on environments with similar species' composition. SAMs' species-centric approach fits with ecological theory about how species assemble and is well suited to answering questions surrounding species' responses to environmental factors now and into the future. RCP's site-based approach makes it particularly well suited to many explicitly spatial applications such as assessing the

444 comprehensiveness and representativeness of marine protected areas (Hill *et al.* 2017), developing
445 monitoring programs (Rose *et al.* 2016), and considering fisheries management under an ecosystem
446 based management approach (Baker & Hollowed 2014). In cases where groups of species have very
447 distinct responses to their environment and environmental gradients are strong, then the two
448 approaches are likely to coincide.

449 **4.2. Challenges and Future Directions**

450 A key challenge for bioregionalisation is determining the appropriate number of bioregions since
451 they represent the simplification of a complex system and are not directly observed. For all methods,
452 the number of groups identified as optimal makes the largest difference to the final
453 bioregionalisation, affecting the location of bioregions and the interpretation of the species and
454 environmental conditions they represent. Clearly this has ramifications when using bioregions for
455 spatial or ecosystem-based management. In our study, the optimal number of groups was often
456 underestimated by the two-stage methods which used hierarchical clustering and average silhouette
457 width, while better discriminated in the one-stage methods using information criteria (BIC, AIC). In
458 simulations, Hui (2017) also found information criteria superior to a range of common clustering
459 algorithms for discriminating groups. Explicitly considering the spatial nature of the data when
460 clustering may improve the discrimination of groups for two-stage methods (Liu *et al.* 2012) and
461 spatial clustering is also an area of active research (Alfó, Nieddu & Vicari 2009). Alternatives for
462 determining the number of groups may be more pragmatic than statistical, such as the feasibility of
463 managing bioregions or the desire to have bioregions at multiple scales. However, we emphasise
464 that statistical approaches are repeatable, and that information criteria and model-based
465 approaches appear a promising way forward to objectively identify the optimal number of
466 bioregions using the data themselves.

467 Two exciting areas of potential development for one-stage bioregionalisation models include
468 incorporating multiple data types and considering other aspects of biodiversity. Incorporating

multiple data types, especially presence-only data, will become ever more important as the spatial scale of management applications increases and online databases grow (Isaac *et al.* 2020). This could be achieved using Inhomogeneous Poisson Point Process Models (IPPM) which have demonstrated advantages for modelling presence-only data (Warton & Shepherd 2010; Renner *et al.* 2015), including multiple data types and attempting to account for sampling biases (Warton, Renner & Ramp 2013; Fithian *et al.* 2014). Similarly genetic, phylogenetic and functional aspects of diversity are becoming increasingly important considerations for conservation and ecosystem-based management (Guilhaumon *et al.* 2015). These metrics are tractable by replacing species with functions or traits as the unit for analyses. Alternatively, the concept used in the hierarchical Bayesian framework (HMSC) where species are modelled, but an explanation for their response is sought using functional or phylogenetic factors at a higher level (Ovaskainen *et al.* 2017), could be extended to one-stage methods.

While we have shown that most methods demonstrated here can potentially provide answers to questions commonly posed by ecologists and resource managers, it is our view that one-stage approaches offer the most comprehensive and consistent method for objectively identifying bioregions, and for describing their biological and environmental characteristics. As the need for and spatial scale of bioregionalisation increases, future developments of one-stage methods that can incorporate presence-only and multiple data types as well as considering functional aspects of bioregions will see the broader uptake and application of quantitative bioregionalisations that incorporate both biological and environmental data.

Acknowledgments

This work was completed as part of Australian Antarctic Science Grant 4124. We thank the companies, skipper and observers that collected the Kerguelen Plateau fish data, Tim Lamb for curating the data, Australian Fisheries Management Authority for permission to use it, Dirk Welsford for useful discussions and anonymous reviewers for constructive comments. The work by OO was by

the Academy of Finland (grant 309581 to OO) and by the Strategic Research Council of the Academy of Finland (grant 312650 to the BlueAdapt consortium).

Author Contributions

NH, SF, PD, CJ, JM conceived the study. NH and SW conducted analyses; all co-authors contributed to interpretation of results. NH wrote the majority of the paper with substantial contribution from SF and critical input from all co-authors.

Data Availability

Kerguelen Plateau fish and environmental data are archived at Australian Antarctic Data centre:

http://dx.doi.org/doi:10.26179/5f0528de8c1d2

http://dx.doi.org/doi:10.26179/5f055cd217aa8

Code to reproduce simulated data and run analyses on simulated and demersal fish data are archived in Zenodo: https://zenodo.org/record/3936354

References

Alfó, M., Nieddu, L. & Vicari, D. (2009) Finite Mixture Models for Mapping Spatially Dependent Disease Counts. *Biometrical Journal,* **51,** 84-97.

Baker, M.R. & Hollowed, A.B. (2014) Delineating ecological regions in marine systems: Integrating physical structure and community composition to inform spatial management in the eastern Bering Sea. *Deep Sea Research Part II: Topical Studies in Oceanography,* **109,** 215-240.

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

Breiman, L. (2001) Random forests. *Machine Learning***,** 15-32.

Cooper, K.M., Bolam, S.G., Downie, A.-L. & Barry, J. (2019) Biological-based habitat classification approaches promote cost-efficient monitoring: An example using seabed assemblages. *Journal of Applied Ecology,* **56,** 1085-1098.

De'ath, G. (2002) Multivariate Regression Trees: A New Technique for Modeling Species-Environment Relationships. *Ecology,* **83,** 1105-1117.

De'ath, G. (2007) Boosted regression trees for ecological modeling and prediction. *Ecology,* **88,** 243-251.

Dunstan, P., Foster, S., Hui, F.C. & Warton, D. (2013) Finite mixture of regression modeling for high-dimensional count and biomass data in ecology. *Journal of Agricultural, Biological, and Environmental Statistics,* **18,** 357-375.

Dunstan, P.K., Foster, S.D. & Darnell, R. (2011) Model based grouping of species across environmental gradients. *Ecological Modelling,* **222,** 955-963.

Ekman, S. (1953) *Zoogeography of the seas*. Sidgwick & Jackson, London.

Elith, J. & Graham, C.H. (2009) Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. *Ecography,* **32,** 66-77.

Elith, J. & Leathwick, J.R. (2009) Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology, Evolution, and Systematics,* **40,** 677-697.

Ellis, N., Smith, S.J. & Pitcher, C.R. (2012) Gradient forests: calculating importance gradients on physical predictors. *Ecology,* **93,** 156-168.

Ferrier, S. & Guisan, A. (2006) Spatial modelling of biodiversity at the community level. *Journal of Applied Ecology,* **43,** 393-404.
Ferrier, S., Manion, G., Elith, J. & Richardson, K. (2007) Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment. *Diversity and Distributions,* **13,** 252-264.
Fiorentino, D., Lecours, V. & Brey, T. (2018) On the Art of Classification in Spatial Ecology: Fuzziness as an Alternative for Mapping Uncertainty. *Frontiers in Ecology and Evolution,* **6**.
Fithian, W., Elith, J., Hastie, T. & Keith, D.A. (2014) Bias correction in species distribution models: pooling survey and collection data for multiple species. *Methods in Ecology and Evolution***,** n/a-n/a.
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,** 489-499.
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,** 1031-1047.
Guilhaumon, F., Albouy, C., Claudet, J., Velez, L., Ben Rais Lasram, F., Tomasini, J.-A., . . . Mouillot, D. (2015) Representing taxonomic, phylogenetic and functional diversity: new challenges for Mediterranean marine-protected areas. *Diversity and Distributions,* **21,** 175-187.
Harris, D.J. (2015) Generating realistic assemblages with a joint species distribution model. *Methods in Ecology and Evolution,* **6,** 465-473.
Hedgpeth, J.W. (1957) Classification of marine environments. *Treatise on marine ecology and paleoecology,,* **50,** 17-28.
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,** 1216-1230.
Hui, F.K.C. (2017) Model-based simultaneous clustering and ordination of multivariate abundance in ecology. *Computations Statistics and Data Analysis,* **105,** 1-10.
Isaac, N.J.B., Jarzyna, M.A., Keil, P., Dambly, L.I., Boersch-Supan, P.H., Browning, E., . . . O'Hara, R.B. (2020) Data Integration for Large-Scale Models of Species Distributions. *Trends in Ecology & Evolution,* **35,** 56-67.
Kaufman, L. & Rousseeuw, P.J. (1990) *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, Hoboken, NJ, USA.
Koen-Alonso, M., Pepin, P., Fogarty, M.J., Kenny, A. & Kenchington, E. (2019) The Northwest Atlantic Fisheries Organization Roadmap for the development and implementation of an Ecosystem Approach to Fisheries: structure, state of development, and challenges. . *Marine Policy,* **100,** 342-352.
Köppen, W. (1884) Die Wärmezonen der Erde, nach der Dauer der heissen, gemässigten und kalten Zeit und nach der Wirkung der Wärme auf die organische Welt betrachtet (The thermal zones of the Earth according to the duration of hot, moderate and cold periods and of the impact of heat on the organic world). Meteorol. Z., 1, . *Meteorologische Zeitschrift,* **1,** 215-226.
Koubbi, P., Moteki, M., Duhamel, G., Goarant, A., Hulley, P.-A., O'Driscoll, R., . . . Hosie, G. (2011) Ecoregionalization of myctophid fish in the Indian sector of the Southern Ocean: Results from generalized dissimilarity models. *Deep Sea Research Part II: Topical Studies in Oceanography,* **58,** 170-180.
Lasram, F.B.R., Hattab, T., Halouani, G., Romdhane, M.S. & Le Loc'h, F. (2015) Modeling of Beta Diversity in Tunisian Waters: Predictions Using Generalized Dissimilarity Modeling and Bioregionalisation Using Fuzzy Clustering. *PLoS ONE,* **10,** e0131728.
Leaper, R., Dunstan, P.K., Foster, S.D., Barrett, N.S. & Edgar, G.J. (2014) Do communities exist? Complex patterns of overlapping marine species distributions. *Ecology,* **95,** 2016-2025.

585 Leaper, R., Hill, N., Edgar, G.J., Ellis, N., Lawrence, E., Pitcher, C.R., . . . Thomson, R. (2011) Predictions
586 of beta diversity for reef macroalgae across southeastern Australia. *Ecosphere,* **2,** 1-18.
587 Liu, Q., Deng, M., Shi, Y. & Wang, J. (2012) A density-based spatial clustering algorithm considering
588 both spatial proximity and attribute similarity. *Computers & Geosciences,* **46,** 296-309.
589 Lyons, M.B., Foster, S.D. & Keith, D.A. (2017) Simultaneous vegetation classification and mapping at
590 large spatial scales. *Journal of Biogeography,* **44,** 2891-2902.
591 Norberg, A., Abrego, N., Blanchet, F.G., Adler, F.R., Anderson, B.J., Anttila, J., . . . Ovaskainen, O.
592 (2019) A comprehensive evaluation of predictive performance of 33 species distribution
593 models at species and community levels. *Ecological Monographs,* **0,** e01370.
594 Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume, B.F., Duan, L., Dunson, D., . . . Abrego, N. (2017)
595 How to make more out of community data? A conceptual framework and its implementation
596 as models and software. *Ecology Letters,* **20,** 561-576.
597 Pollock, L.J., Tingley, R., Morris, W.K., Golding, N., O'Hara, R.B., Parris, K.M., . . . McCarthy, M.A.
598 (2014) Understanding co-occurrence by modelling species simultaneously with a Joint
599 Species Distribution Model (JSDM). *Methods in Ecology and Evolution,* **5,** 397-406.
600 Raymond, B. (2014) Pelagic Regionalisation. *Biogeographic Atlas of the Southern Ocean* (eds C. De
601 Broyer, P. Koubbi, H.J. Griffiths, B. Raymond, C. d'Udekem d'Acoz, A.P. Van de Putte, B. Danis,
602 B. David, S. Grant, J. Gutt, C. Held, G. Hosie, F. Huettmann, A. Post & Y. Ropert-Coudert), pp.
603 418–421. Scientific Committee on Antarctic Research, Cambridge UK.
604 Renner, I.W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S., . . . Warton, D.I. (2015) Point
605 process models for presence-only analysis-a review. *Methods in Ecology and Evolution,* **6,**
606 366-379.
607 Rickbeil, G.J.M., Coops, N.C., Andrew, M.E., Bolton, D.K., Mahony, N. & Nelson, T.A. (2013) Assessing
608 conservation regionalization schemes: employing a beta diversity metric to test the
609 environmental surrogacy approach. *Diversity and Distributions***,** n/a-n/a.
610 Roberson, L.A., Lagabrielle, E., Lombard, A.T., Sink, K., Livingstone, T., Grantham, H. & Harris, J.M.
611 (2017) Pelagic bioregionalisation using open-access data for better planning of marine
612 protected area networks. *Ocean & Coastal Management,* **148,** 214-230.
613 Rose, P.M., Kennard, M.J., Sheldon, F., Moffatt, D.B. & Butler, G.L. (2016) A data-driven method for
614 selecting candidate reference sites for stream bioassessment programs using generalised
615 dissimilarity models. *Marine and Freshwater Research,* **67,** 440-454.
616 Rousseeuw, P.J. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster
617 analysis. *Journal of computational and applied mathematics,* **20,** 53-65.
618 Rubidge, M.E., Gale, K.S.P. & Curtis, J.M.R. (2016) Community ecological modelling as an alternative
619 to physiographic classifications for marine conservation planning. *Biodiversity and*
620 *Conservation,* **25,** 1899-1920.
621 Sayre, R.G., Wright, D.J., Breyer, S.P., Butler, K.A., Van Graafeiland, K., Costello, M.J., . . . Basher, Z.
622 (2017) A three-dimensional mapping of the ocean based on environmental data.
623 *Oceanography,,* **30,** 90–103.
624 Spalding, M.D., Fox, H.E., Allen, G.R., Davidson, N., Ferdaña, Z.A., Finlayson, M., . . . Robertson, J.
625 (2007) Marine Ecoregions of the World: A Bioregionalization of Coastal and Shelf Areas.
626 *Bioscience,* **57,** 573-583.
627 Stephenson, F., Leathwick, J.R., Geange, S.W., Bulmer, R.H., Hewitt, J.E., Anderson, O.F., . . .
628 Lundquist, C.J. (2018) Using Gradient Forests to summarize patterns in species turnover
629 across large spatial scales and inform conservation planning. *Diversity and Distributions,* **0**.
630 ter Braak, C., Hoijtink, H., Akkermans, W. & Verdonschot, P. (2003) Bayesian model-based cluster
631 analysis for predicting macrofaunal communities. *Ecological Modelling,* **160**.
632 Thorson, J.T., Ianelli, J.N., Larsen, E.A., Ries, L., Scheuerell, M.D., Szuwalski, C. & Zipkin, E.F. (2016)
633 Joint dynamic species distribution models: a tool for community ordination and spatio-
634 temporal monitoring. *Global Ecology and Biogeography,* **25,** 1144-1158.

UNESCO (2009) Global Open Oceans and Deep Seabed (GOODS) - Biogeographic Classification. *IOCTechnical Series*. UNESCO-IOC, Paris.

Ware, C., Williams, K.J., Harding, J., Hawkins, B., Harwood, T., Manion, G., . . . Ferrier, S. (2018) Improving biodiversity surrogates for conservation assessment: A test of methods and the value of targeted biological surveys. *Diversity and Distributions,* **0,** 1-14.

Warton, D.I., Blanchet, F.G., O'Hara, R.B., Ovaskainen, O., Taskinen, S., Walker, S.C. & Hui, F.K.C. (2015a) So many variables: Joint modeling in community ecology. *Trends in Ecology & Evolution,* **30,** 766-779.

Warton, D.I., Foster, S.D., De'ath, G., Stoklosa, J. & Dunstan, P.K. (2015b) Model-based thinking for community ecology. *Plant Ecology,* **216,** 669-682.

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,** e79168.

Warton, D.I. & Shepherd, L.C. (2010) Poisson Point Process models solve the 'pseudo-absence problem' for presence-only data in ecology. *The Annals of Applied Statistics,* **4,** 1383-1402.

Woolley, S.N.C., Foster, S.D., Bax, N.J., Currie, J.C., Dunn, D.C., Hansen, C., . . . Dunstan, P.K. (2019) Bioregions in Marine Environments: Combining Biological and Environmental Data for Management and Scientific Understanding. *Bioscience*.

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***,** n/a-n/a.

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,** 1506-1517.

Table 1. Summary of the methods selected to illustrate the broad modelling approaches, how they answer the five core bioregionalisation questions, as well as other key features of the selected methods. Boxes relating to bioregional questions are coloured to indicate: orange = not possible, yellow = possible; green = better approach because it does not require separate post-hoc analyses or considers the entire bioregionalisation process. Max. sil width= maximum silhouette width which was used in this study to determine the number of groups in two-stage approaches. Uncertainty is classed as 'No' unless it can be quantified throughout the entire analysis, not just part of it. BIC= Bayesian Information Criteria. * Only the one-stage methods have appropriate diagnostics to capture the bioregionalisation process. Other methods have diagnostics for a single stage of the analyses, listed in brackets. + Relative indication of computational requirements. Note that even simple hierarchical clustering methods may run into computational and memory issues with a large dataset.

		Approach					
		Two-stage					**One-stage**
		Environment Only	Group First, then Predict	Predict First, then Group			Analyse Simultaneously
				Stacked Distribution Models	Multispecies Distribution Model	Compositional Turnover Models	Mixture Models
Method/s Used	Method (abbreviation)	Hierarchical Clustering of environmental data only (Env_Only)	Hierarchical Clustering of biological data then Random Forest prediction (BioHC_RF)	Species distribution models using Random Forests then Hierarchical Clustering (SpRF_HC)	Hierarchical Modelling of Species Communities then Hierarchical Clustering (HMSC_HC) Neural Net then Hierarchical Clustering (MNet_HC)	Generalised Dissimilarity Model then Hierarchical Cluster Dissimilarities/ Spline transformed environment (GDM_Dissim_HC/ GDM_TransEnv_HC) Bayesian Bootstrap GDM then Hierarchical Cluster Dissimilarities/ Spline transformed environment (bbGDM_Dissim_HC/ bGDM_TransEnv_HC) Gradient Forest then Hierarchical Clustering (GF_HC)	Species Archetype Model (SAM) Regions of Common Profile (RCP)
Bioregional questions	Number of groups	Max. sil. Width	Max. sil. width	Max. sil. width	Max. sil. width	Max. sil. width	BIC
	Spatial distribution of groups	Plot clusters	Predict 1st stage clusters	Plot 2nd stage clusters	Plot 2nd stage clusters	Plot 2nd stage clusters	Direct prediction

	Appropriate uncertainty in bioregion maps	No	No	No	No	No	**Yes**
	Species Composition of groups	No	Summaries from clustered site data	Summaries from clustered site data	Summaries from clustered site data	Summaries from clustered site data	Directly from model parameters
	Environmental Characteristics of groups	Summaries from clustered site data	Summaries from clustered site data	Summaries from clustered site data	Summaries from clustered site data	Summaries from clustered site data	Directly from model parameters
Other Features	Model Diagnostics*	No	Confusion matrix (2nd stage)	Confusion matrix (1st stage)	Goodness of fit at species level (1st stage)	GDM and bbGDM methods: Dissimilarity residual plots (1st stage) GF_HC: Goodness of fit at species level (1st stage)	Residual plots
	Implementation	Easy	Easy	Easy	Moderate	GDM and bbGDM methods: Moderate GF: Easy	Moderate
	Computational Requirements+	Low	Low	Low	HMSC_HC: High MNet_HC: Moderate	GDM and bbGDM methods: High GF_HC: Low	SAM: Low-moderate RCP: Moderate- High
R package/s used		cluster	cluster, extendedForest	extendedForest, cluster	**HMSC_HC:** HMSC, cluster MNet_HC: mistnet, cluster, custom code	GDM and bbGDM methods: gdm, cluster, custom code GF_HC: gradientForest, cluster	SAM: ecomix RCP: RCPmod, custom code
Selected References		Kaufman and Rousseeuw (1990) Rousseeuw (1987)	Elith and Graham (2009) Elith and Leathwick (2009) De'ath (2007) Breiman (2001)		Pollock *et al.* (2014) Warton *et al.* (2015a) Harris (2015) Thorson *et al.* (2016) Ovaskainen *et al.* (2017)	Ferrier *et al.* (2007) Ellis *et al.* (2012) Woolley *et al.* (2017)	Dunstan *et al.* (2011); Dunstan *et al.* (2013) Foster *et al.* (2013); Foster, Hill and Lyons (2017)

666

Appendix S1:
Detailed description of approaches and selected methods

What data goes into bioregionalisation analyses?

Bioregional analyses at regional scales are generally underpinned by data collected during scientific surveys. Surveys typically collect data on the presence-absence, abundance, or biomass of multiple species sampled at sites across an area. Data from several surveys may be collated to gain better geographic coverage or sampling intensity. Environmental data usually consist of variables with synoptic coverage across the region of interest. In the marine realm, these can include data from satellites but may also be interpolated data or output from oceanographic models. Using synoptic environmental data allows us to 'fill the gaps' and map bioregions at sites where biological data has not been collected.

Two-stage approach

All two-stage approaches involve a clustering step where either the biological data or the predictions of single-species distribution models are clustered. Clustering data into groups is a long-standing analytical problem and numerous clustering techniques have been developed (Kaufman & Rousseeuw 1990). Here we use hierarchical clustering, a type of 'algorithmic' clustering techniques because it a popular method used by ecologists to group biological data (e.g. Chiba *et al.* (2001); Schiele, Darr and Zettler (2013). We note that there have been recent developments in model-based and machine -learning clustering approaches that are applicable to this task (e.g. Pledger and Arnold (2014); Hui (2017); Du (2010)).

'Algorithmic' clustering techniques, such as hierarchical clustering, are based on calculating pairwise dissimilarity between sites. They aim to simultaneously minimise the dissimilarity of sites within groups and maximise the dissimilarity of sites between groups (Kaufman & Rousseeuw 1990). Hierarchical clustering iteratively groups the data either divisively (where all sites are initially treated as one group) or agglomeratively (where individual sites are initially treated as groups) producing a tree-like structure. In this work we consider agglomerative hierarchical clustering based on Ward's linkage as a typical algorithmic method often employed in ecological studies.

Two-stage Analyses: Predict first, then Group

Stacked Species Distribution Models

In the stacked species distribution methods, each species is modelled independently and then predictions are compiled ('stacked') to generate species composition for each prediction cell (Norberg *et al.*). There are a multitude of approaches for modelling the distribution of individual species ranging from variations on linear and generalised linear models (GLM) to a vast array of machine learning approaches (described in Elith *et al.* (2006); Elith and Leathwick (2009); Franklin (2009). Here we briefly discuss tree-based machine learning methods because they are good at prediction, are becoming increasingly popular for single SDMs (De'ath & Fabricius 2000; De'ath 2007; Elith, Leathwick & Hastie 2008) and form the basis for more complex models discussed under other approaches. We also note that many of the methods applicable to modelling single species

distributions are applicable to modelling the relationship between groups and environmental factors in the second stage of the 'Group first, then Predict' approach.

Tree-based methods

Tree-based methods recursively partition species data into smaller and smaller groups basing the splits on environmental variables that reduce the error or variance within groups (De'ath & Fabricius 2000). Single trees are extremely sensitive to the input data and are highly variable, which reduces their predictive capacity. Ensemble trees that take random subsets of the data, build many trees and combine the predictions from all trees solve this issue. The best-known ensemble methods are Boosted Regression Trees (BRT; Elith *et al.* (2008) and Random Forests (RF; Breiman (2001). Here we focus on Random Forests as an example of an ensemble method that has been shown to have high predictive power (Lawler *et al.* 2006; Cutler *et al.* 2007) and is increasingly used in ecological applications (Knudby, LeDrew & Brenning 2010; Wei *et al.* 2011). Random Forests take a bootstrap sample of the data and of the environmental predictors to build independent trees (a forest). The final predicted value for the forest is the summary of all the predictions from all the trees in the forest. Interpreting which variables are important and how they relate to the distribution of species relies on aggregating information on the split points and their influence in reducing group error or variance for each environmental variable across all trees in the ensemble (Breiman 2001). Tree-based methods inherently model non-linearity and interactions in the data. We also use Random Forests in the second stage of the "Group first, then Predict' approach for comparability.

Multispecies Distribution Models

As opposed to stacked species distribution models, multispecies distribution models simultaneously model the distribution of multiple species within the single model to generate predictions of community composition. He we focus on a subset of recently-developed methods based on multivariate GLMs and a machine-learning technique.

Multivariate Models, which model more than one response at a time, provide a flexible means to simultaneously estimate the distribution of multiple species using environmental and other data and encompass a variety of specific models. Often also called Joint Species Distribution Models (Pollock *et al.* 2014; Warton *et al.* 2015; Ovaskainen *et al.* 2017), these models recognise that species' distributions are correlated due to factors such as biological interactions. JSDMs seek to model species' co-occurrence or other types of interaction through explicit joint correlation structures. This is often in the form of a multivariate response GLM with carefully designed random effects to account for interspecific interactions (Pollock *et al.* 2014; Warton *et al.* 2015; Ovaskainen *et al.* 2016b). As the number of required correlation terms between species grows quickly with the number of species, a clever solution is to use a small number of latent (unobserved) variables that map to the correlation structure to reduce dimensionality and simplify the problem (Hui 2016; Ovaskainen *et al.* 2016a; Thorson *et al.* 2016). Such a model effectively performs dimension reduction whilst simultaneously conditioning on the species' responses to the environment. Many variations of JSDMs exist and, depending on the data and context, can include temporal (Thorson *et al.* 2016) and spatial correlation structures (Latimer *et al.* 2009; Thorson *et al.* 2015; Ovaskainen *et al.* 2016b), experimental design considerations (Ovaskainen *et al.* 2016a), can accommodate the influence of functional traits (Sebastián-González *et al.* 2010; Abrego, Norberg & Ovaskainen 2017) or phylogeny (Ovaskainen *et al.* 2017) on the distribution of species. JSDMs can be more accurate at predicting individual species (Maguire *et al.* 2016; Norberg *et al.* 2019) and community-level properties such as species richness (Norberg *et al.* 2019) than stacked species distribution models (but see Caradima, Schuwirth and Reichert (2019)). Rarer species can more effectively be modelled than when using single species models as they 'borrow strength' from other species via correlation

structures (Hui *et al.* 2015; Norberg *et al.* 2019). The particular type of Joint Species Distribution Model we use here is a Bayesian Hierarchical model within the Hierarchical Modelling of Species Communities (HMSC) that uses latent variables to account for spatially-structured species' co-occurrences (Ovaskainen *et al.* 2016b; Tikhonov *et al.* 2020). In this instance the spatial structure aids in the prediction to unsampled locations within the region of interest.

Artificial Neural Networks (ANN) are a machine learning technique that is gaining popularity in species distribution modelling due to their generally high predictive performance and ability to accommodate interactions and non-linear responses (Olden 2003; Olden, Joy & Death 2006; McKenna, Carlson & Payne-Wynne 2013). Artificial Neural Networks consists of three types of layers, the environmental (input) layer, at least one hidden layer and the species (output) layer. All layers consist of neurons, akin to variables, or functions of them (Olden *et al.* 2006). Sometimes it is also advantageous to use random variables as inputs (Harris 2015), as these can improve prediction properties. The hidden layer takes transformations and combinations of the environmental covariates to form new variables – the hidden layer's neurons – with the number of neurons optimised using cross-validation. Weighting is applied to connections and determines the influence of the neurons in one layer on the neurons in the next layer. The ANN is trained to the data by iteratively adjusting the connection weights to find the set that minimises the error in the network as it is sequentially presented samples (Olden *et al.* 2006). The relative influence of environmental variables are quantified using the connection weights across the layers for each input neuron.

In the context of bioregionalisation, the outputs of multispecies ANNs and JSDMs are typically predictions of the probability of occurrence or abundance of each species in each spatial cell over the domain of interest. These predictions, although potentially more accurate than stacked single SDMs, still need to be clustered to define bioregions.

Compositional turnover models

Two methods have been developed in recent years to model the turnover of community composition (beta diversity), which can act as a proxy for defining bioregions; Generalised Dissimilarity Modelling (GDM; Ferrier *et al.* (2007) and Gradient Forests (GF; Ellis, Smith and Pitcher (2012).

GDM uses dissimilarity metrics as the basis for model building and inference. These pairwise dissimilarities are treated as the response variable, and the predictor variables are the corresponding site-wise differences in each of the environmental variables (Ferrier *et al.* 2007). This ecological dissimilarity is modelled using a regression spline within Generalised Linear Model (GLM) that enforces the constraint that sites that have greater environmental differences must be more ecologically dissimilar (Ferrier *et al.* 2007; Woolley *et al.* 2017). Spatial predictions are made by transforming the environmental differences between pairs of prediction cells using the function identified by the GDM model. Predictions are the dissimilarity of every cell to each other cell and are therefore difficult to visualise. Generally, the dimension of these dissimilarities is either reduced using multidimensional scaling (MDS) and the first three or four MDS axes plotted (Lasram *et al.* 2015) or the entire dissimilarity matrix is directly clustered to produce bioregions (Ferrier *et al.* 2007; Koubbi *et al.* 2011). Alternatively, the spline transformed environmental differences between pairs of prediction cells can be either ordinated or clustered (Leaper *et al.* 2011). We classify GDM as a two-stage method because a clustering step is necessary to derive bioregions.

One criticism of GDM is that the pairwise dissimilarities number m(m -1)/2, where m is the number of sites. GLMs assume stochastic independence, which cannot be obtained when modelling dissimilarities. This has important implications, chiefly that there is an overstatement of the amount

of information in the data that will under-estimate uncertainty and exaggerate the statistical significance of environmental factors that propagates through to the predictions of mean dissimilarities between sites (Woolley *et al.* 2017). A recent modification of GDM uses Bayesian bootstrap sampling to obtain better estimates of the significance of environmental variables for determining community turnover (Woolley *et al.* 2017).

Gradient Forests (GF) aggregate the information from single-species Random Forests (RF, discussed earlier) to generate a picture of which environmental predictors are important in determining distribution across all species and where compositional changes occur along each environmental gradient (Ellis *et al.* 2012). To do this GF aggregates the (cumulative) distribution of split values along each environmental variable, weighted by its importance in determining the split in the forest and the goodness of fit of the RF for each species. These cumulative split distributions are treated as functions that describe community turnover along environmental gradients. Predictions are made by transforming the environmental covariates at all cells across the region of interest using the cumulative importance functions (Pitcher *et al.* 2012). These 'biologically informed' environmental variables can then either be ordinated (Pitcher *et al.* 2012; Thomson *et al.* 2014) or clustered directly (Baker & Hollowed 2014; Stephenson *et al.* 2018) in a similar way to that performed in Leaper *et al.* (2011) using GDM, to produce groups that represent different assemblages or bioregions.

One-Stage Analyses

One-stage approaches delineate bioregions based on a simultaneous use of biological data and their relationship to environment. Limited methods are available for one-stage approaches, which currently include Species Archetype Models (SAMs; (Dunstan, Foster & Darnell 2011; Dunstan *et al.* 2013), Regions of Common Profiles models (RCPs; Foster *et al.* (2013) but also see ter Braak *et al.* (2003) and Multivariate Regression Trees (MRTs; De'ath (2002)).

In Species Archetype Models (SAMs), species are grouped based on their response to environmental gradients through application of a finite mixture of GLMs of species' data onto a set of environmental variables (Dunstan *et al.* 2011). The aim is to find subsets of species that can be described by a set of common environmental responses. Because these groups are unobserved, SAMs can also be classified as a latent factor model and has the property that rarer species can 'borrow strength' from more common species in terms of their environmental responses (Hui *et al.* 2013). The number of groups (species archetypes) supported by the data is determined using the Bayesian Information Criteria (BIC), but this measure is not infallible. BIC can also be used to select and quantify the relative importance of environmental variables in discriminating groups of species. Each species has a probability of belonging to each archetype via estimated model coefficients (Dunstan *et al.* 2011). Each archetype is defined by its response to the environment, which means that the environmental characteristic of each group are also defined by the model's coefficients. SAM has been used mainly in the marine environment to examine diversity patterns for conservation management and to examine ecological paradigms (Woolley *et al.* 2013; Leaper *et al.* 2014; Jansen *et al.* 2018).

Using the groups' responses to environmental variables, the probability of finding each archetype can be directly predicted into areas with synoptic coverage for environmental covariates but with limited biological sampling. It is important to emphasise that SAM groups *species* and not sites, therefore more than one archetype (i.e. group with common response to the environment) may be likely at any location. Importantly, the uncertainty in finding an archetype at a new location is quantifiable and appropriate (i.e. it captures the uncertainty in both the grouping of species and their response to the environment).

Regions of Common Profile (RCP; Foster (2013) models are another model-based, one-stage statistical approach to bioregionalisation. RCP defines environmental regions with a distinct species profile and essentially simultaneously uses biological and environmental data to cluster sites. Like SAMs, RCPs are based on multivariate GLMs. Technically, they are a mixture-of-experts model (Foster, Hill & Lyons 2017) where region is a latent factor whose probability of occurrence varies as a function of environment. The number of regions and influential environmental variables can be chosen using BIC (Hill *et al.* 2017). The expected prevalence or abundance of each species in each region (i.e. the species profile or composition) is defined directly by model coefficients. Because each region is defined by environmental variables, the environmental characteristics of each region are also defined by model coefficients and the probability of finding each region can be directly predicted at new sites. Similarly to SAM, RCP appropriately quantifies uncertainty in the probability of finding each RCP at each new site. In contrast to SAM, RCP groups *sites* based on their species composition and environment. RCPs are relatively new and are starting to see uptake in the marine and terrestrial realms to inform conservation management (Hill *et al.* 2017; Lyons, Foster & Keith 2017).

Multivariate regression trees (MRTs) are extensions of univariate regression trees, where each split in the tree is based on a division of the environmental predictor that minimises the sums of squares about the multivariate mean (De'ath 2002). The terminal nodes of MRTs indicate a relatively homogenous group of sites, characterised by the mean values of their associated species. Thus, MRTs also give information about the environment and species' composition of groups directly from a single model. Multivariate Regression Trees have been implemented primarily for abundance data using various standardisations (e.g. site standardisation) that are amenable to the sums of squares metrics. These can equate to different inter-site distances (e.g. Chi-squared distances) (De'ath 2002). Theoretically, it is possible to run MRTs directly on *any dissimilarity matrix* but interpreting the resulting tree and generating predictions are problematic and similar to the two-stage methods that explicitly model dissimilarity metrics. Similarly, to single univariate trees, multivariate trees can be sensitive to outliers. A Random Forest version of MRTs has been developed that builds ensembles of trees using bootstrapped samples (Segal & Xiao 2011). While the random sampling of data for each tree in the forest increases the robustness of individual species' predictions, it complicates interpreting which sites belong to which terminal node or group. The proximity matrix of each tree describes how sites are associated and when aggregated over re-sampled trees can be converted into a distance matrix can then be clustered to determine the number of groups (Miller *et al.* 2014). Because of this second classification step, we do not consider multivariate random forests as a truly one-stage method and because we are interested in presence-absence data (MRTs are available only for continuous data), we do not consider MRTs or multivariate random forests in this paper.

References

Abrego, N., Norberg, A. & Ovaskainen, O. (2017) Measuring and predicting the influence of traits on the assembly processes of wood-inhabiting fungi. *Journal of Ecology,* **105,** 1070-1081.

Baker, M.R. & Hollowed, A.B. (2014) Delineating ecological regions in marine systems: Integrating physical structure and community composition to inform spatial management in the eastern Bering Sea. *Deep Sea Research Part II: Topical Studies in Oceanography,* **109,** 215-240.

Breiman, L. (2001) Random forests. *Machine Learning***,** 15-32.

Caradima, B., Schuwirth, N. & Reichert, P. From individual to joint species distribution models: A comparison of model complexity and predictive performance. *Journal of Biogeography,* **46**.

Chiba, S., Ishimaru, T., Hosie, G.W. & Fukuchi, M. (2001) Spatio-temporal variability of zooplankton community structure off east Antarctica (90 to 160°E). *Marine Ecology Progress Series,* **216,** 95-108.

Cutler, D.R., Jr, T.C.E., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J. & Lawler., J.J. (2007) Random Forests for Classification in Ecology. *Ecology,* **88,** 2783-2792.

De'ath, G. (2002) Multivariate Regression Trees: A New Technique for Modeling Species-Environment Relationships. *Ecology,* **83,** 1105-1117.

De'ath, G. (2007) Boosted regression trees for ecological modeling and prediction. *Ecology,* **88,** 243-251.

De'ath, G. & Fabricius, K.E. (2000) Classification and Regression Trees: A Powerful Yet Simple Technique for Ecological Data Analysis. *Ecology,* **81,** 3178-3192.

Du, K.L. (2010) Clustering: A neural network approach. *Neural Networks,* **23,** 89-107.

Dunstan, P., Foster, S., Hui, F.C. & Warton, D. (2013) Finite mixture of regression modeling for high-dimensional count and biomass data in ecology. *Journal of Agricultural, Biological, and Environmental Statistics,* **18,** 357-375.

Dunstan, P.K., Foster, S.D. & Darnell, R. (2011) Model based grouping of species across environmental gradients. *Ecological Modelling,* **222,** 955-963.

Elith, J., Graham, C.H., Anderson, R.P., Dudík, M., Ferrier, S., Guisan, A., . . . Zimmerman, N.E. (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography,* **29,** 129-151.

Elith, J. & Leathwick, J.R. (2009) Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology, Evolution, and Systematics,* **40,** 677-697.

Elith, J., Leathwick, J.R. & Hastie, T. (2008) A working guide to boosted regression trees. *Journal of Animal Ecology,* **77,** 802-813.

Ellis, N., Smith, S.J. & Pitcher, C.R. (2012) Gradient forests: calculating importance gradients on physical predictors. *Ecology,* **93,** 156-168.

Ferrier, S., Manion, G., Elith, J. & Richardson, K. (2007) Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment. *Diversity and Distributions,* **13,** 252-264.

Foster, S.D. (2013) RCPmod: Regions of common profiles modelling with mixtures-of-experts. R package version

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,** 489-499.

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,** 1031-1047.

Franklin, J. (2009) *Mapping species distributions*. Cambridge University Press, New York.

Harris, D.J. (2015) Generating realistic assemblages with a joint species distribution model. *Methods in Ecology and Evolution,* **6,** 465-473.

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,** 1216-1230.

Hui, F.K.C. (2016) boral – Bayesian Ordination and Regression Analysis of Multivariate Abundance Data in r. *Methods in Ecology and Evolution,* **7,** 744-750.

Hui, F.K.C. (2017) Model-based simultaneous clustering and ordination of multivariate abundance in ecology. *Computations Statistics and Data Analysis,* **105,** 1-10.

Hui, F.K.C., Taskinen, S., Pledger, S., Foster, S.D. & Warton, D.I. (2015) Model-based approaches to unconstrained ordination. *Methods in Ecology and Evolution,* **6,** 399-411.

Hui, F.K.C., Warton, D.I., Foster, S.D. & Dunstan, P.K. (2013) To mix or not to mix: comparing the predictive performance of mixture models vs. separate species distribution models. *Ecology,* **94,** 1913-1919.

Jansen, J., Hill, N.A., Dunstan, P.K., Eléaume, M.P. & Johnson, C.R. (2018) Taxonomic Resolution, Functional Traits, and the Influence of Species Groupings on Mapping Antarctic Seafloor Biodiversity. *Frontiers in Ecology and Evolution,* **6**.

Kaufman, L. & Rousseeuw, P.J. (1990) *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, Hoboken, NJ, USA.

Knudby, A., LeDrew, E. & Brenning, A. (2010) Predictive mapping of reef fish species richness, diversity and biomass in Zanzibar using IKONOS imagery and machine-learning techniques. *Remote Sensing of Environment,* **114,** 1230-1241.

Koubbi, P., Moteki, M., Duhamel, G., Goarant, A., Hulley, P.-A., O'Driscoll, R., . . . Hosie, G. (2011) Ecoregionalization of myctophid fish in the Indian sector of the Southern Ocean: Results from generalized dissimilarity models. *Deep Sea Research Part II: Topical Studies in Oceanography,* **58,** 170-180.

Lasram, F.B.R., Hattab, T., Halouani, G., Romdhane, M.S. & Le Loc'h, F. (2015) Modeling of Beta Diversity in Tunisian Waters: Predictions Using Generalized Dissimilarity Modeling and Bioregionalisation Using Fuzzy Clustering. *PLoS ONE,* **10,** e0131728.

Latimer, A.M., Banerjee, S., Sang Jr, H., Mosher, E.S. & Silander Jr, J.A. (2009) Hierarchical models facilitate spatial analysis of large data sets: a case study on invasive plant species in the northeastern United States. *Ecology Letters,* **12,** 144-154.

Lawler, J.J., White, D., Neilson, R.P. & Blaustein, A.R. (2006) Predicting Climate-Induced Range Shifts: Model Differences and Model Reliability. *Global Change Biology,* **12,** 1568-1584.

Leaper, R., Dunstan, P.K., Foster, S.D., Barrett, N.S. & Edgar, G.J. (2014) Do communities exist? Complex patterns of overlapping marine species distributions. *Ecology,* **95,** 2016-2025.

Leaper, R., Hill, N., Edgar, G.J., Ellis, N., Lawrence, E., Pitcher, C.R., . . . Thomson, R. (2011) Predictions of beta diversity for reef macroalgae across southeastern Australia. *Ecosphere,* **2,** 1-18.

Lyons, M.B., Foster, S.D. & Keith, D.A. (2017) Simultaneous vegetation classification and mapping at large spatial scales. *Journal of Biogeography,* **44,** 2891-2902.

Maguire, K.C., Nieto-Lugilde, D., Blois, J.L., Fitzpatrick, M.C., Williams, J.W., Ferrier, S. & Lorenz, D.J. (2016) Controlled comparison of species- and community-level models across novel climates and communities. *Proceedings of the Royal Society B: Biological Sciences,* **283,** 20152817.

McKenna, J.E., Carlson, D.M. & Payne-Wynne, M.L. (2013) Predicting locations of rare aquatic species' habitat with a combination of species-specific and assemblage-based models. *Diversity and Distributions,* **19,** 503-517.

Miller, K., Huettmann, F., Norcross, B. & Lorenz, M. (2014) Multivariate random forest models of estuarine-associated fish and invertebrate communities. *Marine Ecology Progress Series,* **500,** 159-174.

Norberg, A., Abrego, N., Blanchet, F.G., Adler, F.R., Anderson, B.J., Anttila, J., . . . Ovaskainen, O. (2019) A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. *Ecological Monographs,* **89,** e01370.

Olden, J.D. (2003) A Species-Specific Approach to Modeling Biological Communities and Its Potential for Conservation. *Conservation Biology,* **17,** 854-863.

Olden, J.D., Joy, M.K. & Death, R.G. (2006) Rediscovering the species in community-wide predictive modeling. *Ecological Applications,* **16,** 1449-1460.

Ovaskainen, O., Abrego, N., Halme, P. & Dunson, D. (2016a) Using latent variable models to identify large networks of species-to-species associations at different spatial scales. *Methods in Ecology and Evolution,* **7,** 549-555.

Ovaskainen, O., Roy, D.B., Fox, R. & Anderson, B.J. (2016b) Uncovering hidden spatial structure in species communities with spatially explicit joint species distribution models. *Methods in Ecology and Evolution,* **7,** 428-436.

Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume, B.F., Duan, L., Dunson, D., . . . Abrego, N. (2017) How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecology Letters,* **20,** 561-576.

Pitcher, C.R., Lawton, P., Ellis, N., Smith, S.J., Incze, L.S., Wei, C.-L., . . . Snelgrove, P.V.R. (2012) Exploring the role of environmental variables in shaping patterns of seabed biodiversity composition in regional-scale ecosystems. *Journal of Applied Ecology,* **49,** 670-679.

Pledger, S. & Arnold, R. (2014) Multivariate methods using mixtures: Correspondence analysis, scaling and pattern-detection. *Computational Statistics & Data Analysis,* **71,** 241-261.

Pollock, L.J., Tingley, R., Morris, W.K., Golding, N., O'Hara, R.B., Parris, K.M., . . . McCarthy, M.A. (2014) Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). *Methods in Ecology and Evolution,* **5,** 397-406.

Schiele, K.S., Darr, A. & Zettler, M.L. (2013) Verifying a biotope classification using benthic communities – An analysis towards the implementation of the European Marine Strategy Framework Directive. *Marine Pollution Bulletin,* **78**.

Sebastián-González, E., Sánchez-Zapata, J.A., Botella, F. & Ovaskainen, O. (2010) Testing the heterospecific attraction hypothesis with time-series data on species co-occurrence. *Proceedings of the Royal Society B: Biological Sciences,* **277,** 2983-2990.

Segal, M. & Xiao, Y. (2011) Multivariate random forests. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* **1,** 80-87.

Stephenson, F., Leathwick, J.R., Geange, S.W., Bulmer, R.H., Hewitt, J.E., Anderson, O.F., . . . Lundquist, C.J. (2018) Using Gradient Forests to summarize patterns in species turnover across large spatial scales and inform conservation planning. *Diversity and Distributions,* **24,** 1641-1656.

ter Braak, C., Hoijtink, H., Akkermans, W. & Verdonschot, P. (2003) Bayesian model-based cluster analysis for predicting macrofaunal communities. *Ecological Modelling,* **160**.

Thomson, R.J., Hill, N.A., Leaper, R., Ellis, N., Pitcher, C.R., Barrett, N.S. & Edgar, G.J. (2014) Congruence in demersal fish, macroinvertebrate, and macroalgal community turnover on shallow temperate reefs. *Ecological Applications,* **24,** 287-299.

Thorson, J.T., Ianelli, J.N., Larsen, E.A., Ries, L., Scheuerell, M.D., Szuwalski, C. & Zipkin, E.F. (2016) Joint dynamic species distribution models: a tool for community ordination and spatio-temporal monitoring. *Global Ecology and Biogeography,* **25,** 1144-1158.

Thorson, J.T., Scheuerell, M.D., Shelton, A.O., See, K.E., Skaug, H.J. & Kristensen, K. (2015) Spatial factor analysis: a new tool for estimating joint species distributions and correlations in species range. *Methods in Ecology and Evolution,* **6,** 627-637.

Tikhonov, G., Opedal, Ø.H., Abrego, N., Lehikoinen, A., de Jonge, M.M.J., Oksanen, J. & Ovaskainen, O. (2020) Joint species distribution modelling with the R-package Hmsc. *Methods in Ecology and Evolution,* **11,** 442-447.

Warton, D.I., Blanchet, F.G., O'Hara, R.B., Ovaskainen, O., Taskinen, S., Walker, S.C. & Hui, F.K.C. (2015) So many variables: Joint modeling in community ecology. *Trends in Ecology & Evolution,* **30,** 766-779.

Wei, C.-L., Rowe, G.T., Escobar-Briones, E., Boetius, A., Soltwedel, T., Caley, M.J., . . . Narayanaswamy, B.E. (2011) Global Patterns and Predictions of Seafloor Biomass Using Random Forests. *PLOS ONE,* **5,** e15323.

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,** 985-995.

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,** 1506-1517.

Appendix S2:
Detailed description of implementation of selected methods

General

In order to keep the results as comparable as possible, hierarchical clustering was used in the clustering step of the two-stage methods. We used Ward's Distance as the linkage criteria (Kaufman & Rousseeuw 1990) and the maximum average silhouette width to determine the number of clusters supported by the data (i.e. where to cut the dendrogram; Rousseeuw (1987). All analyses were conducted in the R statistical environment (R Development Core Team 2015). The main packages used for each method are listed in Table 1.

All code is publicly available via Zenodo: https://zenodo.org/record/3936354

Within each of the folders 'Simulation' and 'KP_Fish' devoted to the analysis of the two datasets, code within '…_Run_Models.R' implements the analyses described below. Additional code files are provided that were used for interpreting and plotting model outputs for each dataset.

Environment Only

We performed a cluster analysis on the environmental data (Env_Only) available for the simulated and Kerguelen Plateau (KP) regions. This represents a common scenario where no biological data are available or incorporated into bioregionalisation analyses. Environmental data were scaled and centred, before being clustering based on Euclidean distances. The average (and SD) environmental characteristics of each group were determined by tabulating the environmental conditions of the hard clusters assigned to each cell in the simulated or KP region.

Group first, then Predict

Hierarchical cluster, Random Forest predict (BioHC_RF)

Hierarchical clustering was performed on species' presence-absence data converted to Jaccard dissimilarity, which represents the number of species shared between pairs of sites. The groups assigned to the sites by the hierarchical clustering were then related to environmental data using a classification Random Forest (RF) implemented in the R package 'extendedForest'. Default settings were used for the number of variables to try at each step (mtry= 2) and the number of trees to build (500). The correlation threshold was set to 0.65 to account for highly correlated predictor variables. Model fit was assessed by calculating a confusion matrix using the model's out of bag (OOB) samples. Variable importance was assessed using the mean decrease in accuracy and the mean decrease in node purity. The form of the relationship between environmental variables and the groups was assessed using partial plots. The species and environmental characteristics of each group were tabulated from the RF classification of the survey sites. Group predictions for the entire simulation and KP region were generated from the RF model using the 'predict' function.

Predict first, then Group

Stacked Single Species Distribution Models

Random Forest species' predictions, Hierarchical clustering (SpRF_HC)

Classification Random Forests were performed on presence-absence of each species separately using the same settings as above. Predictions for the probability of occurrence of each species across the simulated or survey region from each RF model were generated using the 'predict' function. The probability of finding each species in each cell was used as the input to the hierarchical clustering that was based on the Euclidean distance between the cells. The species and environmental characteristics of each group were tabulated from the classification of the survey sites.

Multispecies Distribution Models

Multiresponse Artificial Neural Networks prediction, Hierarchical clustering (MNet_HC)

Multi-response neural networks with latent variables that are able to capture unmeasured environmental and/or biological correlations, were implemented using source code for the R package 'mistnet' (Harris 2015) and optimisation code available on https://github.com/davharris/mistnet . We built a network with three layers; the input layer, 1 hidden layer and the output layer. We used 5-fold cross-validation and a modification of the github optimisation code to optimise the number of latent variables used as well as the number of nodes in hidden layer, while keeping other settings at their default. The final model for the simulation data had 2 latent input variables and 8 nodes in the hidden layer, while the final model for the KP fish data had 4 latent input variables and 12 nodes in the hidden layer. The species and environmental characteristics of each group were tabulated from the classification of the survey sites.

Hierarchical Bayesian Model prediction, Hierarchical clustering (HMSC_HC)

Hierarchical Bayesian Models were implemented using the Hierarchical Modelling of Species Communities (HMSC) framework (Ovaskainen *et al.* 2017) in the R package 'HMSC' (Blanchet 2013; Tikhonov *et al.* 2020). Environmental variables were centred and scaled (linear terms for simulation) or orthogonal quadratic polynomials created (KP fish analyses) before input into the HMSC models. A spatially explicit model was run where the latitude and longitude of survey sites were used to generate spatially-structured latent variables that had an exponential spatial covariance function (Ovaskainen *et al.* 2016). A spatially explicit model was run to improve predictive capacity across the study region. The model used a probit link and parameters were estimated using 10,000 MCMC iterations with a burn in of 1000 and a thinning rate of 10. The mixing of chains was assessed visually. Model fit was assessed by calculating Tjur's R for each species. The spatially explicit model was used to generate predictions of the occurrence of each species across the simulated or survey region. These predictions were used as the input into the hierarchical cluster analysis. The species and environmental characteristics of each group were tabulated from the classification of the survey sites.

Compositional Turnover Models

Generalised Dissimilarity Modelling, Hierarchical clustering (GDM_Dissim_HC, GDM_TransEnv_HC and bbGDM_Dissim_HC, bbGDM_TransEnv_HC)

A bootstrapped and naïve (i.e. non-bootstrapped) Generalised Dissimilarity Model were run using the presence-absence version of the Bray-Curtis metric. The default settings were used for the splines (i-spline, with 2 degrees of freedom and 1 knot) and geographic predictors were not used. For the bootstrapped model, 10,000 and 5,000 Bayesian bootstraps were used to estimate the distribution of simulation and fish model parameters respectively, using a bootstrapping wrapper

function to 'gdm' in the 'gdm' package *sensu* (Woolley *et al.* 2017). To capture the various ways that GDM outputs have been clustered, we generated two different hierarchical clustering results. These were; i) direct clustering of pairwise site dissimilarities (e.g. Koubbi *et al.* (2011)) which is the most intuitive step and ii) clustering the spline transformed environmental variables (e.g. Leaper *et al.* (2011); this is most similar to the approach used in gradient forests). Predictions of pairwise site dissimilarities were generated using a function analogous to 'gdm.predict' that could accommodate bootstrap estimates. Similarly, the spline transformed environment at each site differences was generated using a function analogous to 'gdm.transform' that could accommodate bootstrap estimates. Similar to the other two stage methods, the species and environmental characteristics of each group were tabulated from the classification of the survey sites. GDM models were implemented in the R package 'gdm' and bootstrapped versions of functions are available in our code on zenodo.

Gradient Forests prediction, Hierarchical clustering (GF_HC)

Gradient Forests were run on the presence-absence of species at sites. Forests used the default settings, with 500 trees. Predictions of compositional turnover were made for the entire simulation or survey region using the 'predict' function, which transforms the environmental data at prediction sites using the cumulative splits determined by the individual species' forests, weighted by their fit. These transformed environmental variables were used as the input to hierarchical clustering using Euclidean distance (Ellis, Smith & Pitcher 2012). The species and environmental characteristics of each group were tabulated from the classification of the survey sites

Analyse Simultaneously

Species Archetype Models (SAM)

Scaled and centred linear (simulation) or orthogonal quadratic polynomials (KP fish) of each of the predictor variables were generated and used as input into SAM models with a Bernoulli error distribution. Models with between 1 and 6 archetypes were examined. The final number of archetypes/groups was selected by considering the first group where the change in BIC from the previous group was positive, where the prior probability of any one group was greater than 1/number species (this removes archetypes that are unlikely to be useful) and where each archetype contained at least 1 species. The appropriateness of the model was checked using randomised quantile residuals (Dunn & Smyth 1996). Predictions of the probability of occurrence of each archetype were generated for the survey region using the environmental covariates used in the training model. The response of each group to the environmental variables was quantified using model co-efficients and partial plots where the value of all environmental variables, except the variable of interest, were held at their mean value. Although, the species belonging to each archetype can be identified directly from the model using the species' posterior probability of group membership (tau), we obtained *site-based* species' composition and environmental characteristic of each group for comparison with the other methods. To do this the SAM group predictions at survey sites were hard clustered and the observed prevalence of species and value of environmental variables tabulated. SAMs were implemented in the R package 'ecomix' available on: https://github.com/skiptoniam/ecomix

Regions of Common Profile (RCP)

Scaled and centred linear (simulation) or orthogonal quadratic polynomials (KP fish) of each of the predictor variables were generated and used as input into RCPs with a Bernoulli error distribution. For the fish model, survey was used as the sampling factor that affects the catchability of species. What this does is increase or decrease, for a particular level of the sampling factor, the expectation of each species by the same amount for all RCPs and so can account for seeing more or less of a particular species in a particular year. Five hundred model restarts were run to avoid local maxima. A forward selection procedure was used to select environmental variables and the number of RCPs

simultaneously (Hill *et al.* 2017). Starting from the null model for each step we considered the addition of each environmental variable (linear and quadratic term simultaneously) for between 1 and 8 RCPs. The best model for that step was the combination of environmental variables and number of RCPs that minimised BIC. The process was repeated until there was no improvement in BIC between selection steps. Model assumptions were checked by examining randomised quantile residuals. Five hundred Bayesian bootstraps were used to estimate uncertainty in model parameters. RCP predictions were generated using the 'predict' function using the environmental covariates used in the training model. For the fish data, which included survey as a sampling factor, predictions represent the values expected for the first survey. The species composition of RCP groups was calculated directly from the model co-efficients. The response of each group to the environmental variables was quantified using model co-efficients and partial plots where the value of all environmental variables, except the variable of interest, were held at their mean value. For comparison with the other methods, the RCP group predictions at survey sites were also hard clustered and the observed prevalence of species and value of environmental variables tabulated. The importance of environmental variables was assessed using the change in BIC between models in the forward selection procedure. Models were implemented in the R package 'RCPmod' (Foster 2013).

References

Blanchet, F.G. (2013) HMSC: Hierarchical Modelling of Species Community. CRAN, http://rpackages.ianhowson.com/rforge/HMSC/man/HMSC-package.html.

Dunn, P.K. & Smyth, G.K. (1996) Randomised quantile residuals. *Journal of Computations and Graphical Statisitics,* **5,** 236-244.

Ellis, N., Smith, S.J. & Pitcher, C.R. (2012) Gradient forests: calculating importance gradients on physical predictors. *Ecology,* **93,** 156-168.

Foster, S.D. (2013) RCPmod: Regions of common profiles modelling with mixtures-of-experts. R package version

Harris, D.J. (2015) Generating realistic assemblages with a joint species distribution model. *Methods in Ecology and Evolution,* **6,** 465-473.

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,** 1216-1230.

Kaufman, L. & Rousseeuw, P.J. (1990) *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, Hoboken, NJ, USA.

Koubbi, P., Moteki, M., Duhamel, G., Goarant, A., Hulley, P.-A., O'Driscoll, R., . . . Hosie, G. (2011) Ecoregionalization of myctophid fish in the Indian sector of the Southern Ocean: Results from generalized dissimilarity models. *Deep Sea Research Part II: Topical Studies in Oceanography,* **58,** 170-180.

Leaper, R., Hill, N., Edgar, G.J., Ellis, N., Lawrence, E., Pitcher, C.R., . . . Thomson, R. (2011) Predictions of beta diversity for reef macroalgae across southeastern Australia. *Ecosphere,* **2,** 1-18.

Ovaskainen, O., Roy, D.B., Fox, R. & Anderson, B.J. (2016) Uncovering hidden spatial structure in species communities with spatially explicit joint species distribution models. *Methods in Ecology and Evolution,* **7,** 428-436.

Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume, B.F., Duan, L., Dunson, D., . . . Abrego, N. (2017) How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecology Letters,* **20,** 561-576.

R Development Core Team (2015) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rousseeuw, P.J. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics,* **20,** 53-65.

Tikhonov, G., Opedal, Ø.H., Abrego, N., Lehikoinen, A., de Jonge, M.M.J., Oksanen, J. & Ovaskainen, O. (2020) Joint species distribution modelling with the R-package Hmsc. *Methods in Ecology and Evolution,* **11,** 442-447.

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,** 985-995.

Appendix S3: Simulation Study

Simulation Set up

All code for generating the simulated data and its subsequent analysis is hosted on Zenodo: https://zenodo.org/record/3936354

The simulation data consisted of a set of environmental variables and the probability of occurrence of 30 species across a hypothetical region.

Environmental Data

A set of eight environmental variables loosely modelled on patterns observed in real environmental data were generated for the simulation region. Two types of variables were simulated; variables that exhibit a gradient over the regions, and variables that exhibit a patchier distribution. This was achieved by manipulating the spatial dependency between cells in the simulation region, multiplying this dependency by cells values randomly chosen from a normal distribution and re-scaling to give values within a realistic range of the environmental variables that the simulated variables were intended to represent. The code for generating the simulated environmental variables can be found in the file: simulation_env_070518.r

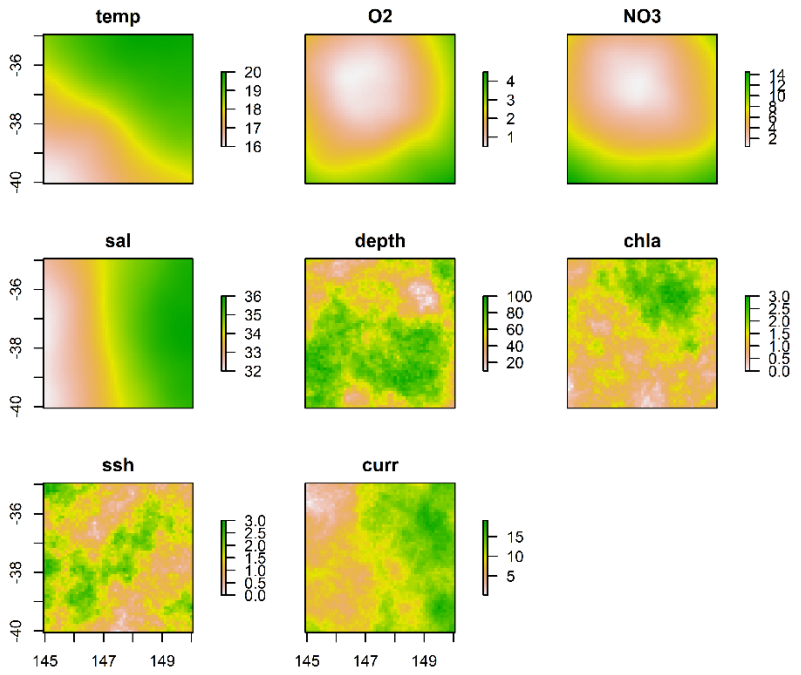


Table A3.1. Pearson's correlation between simulated environmental variables.

	temp	O2	NO3	sal	depth	chla	ssh	curr
temp	1							
O2	-0.26	1						
NO3	-0.57	**0.89**	1					
sal	**0.66**	0.34	-0.07	1				
depth	-0.44	0.19	0.20	-0.08	1			
chla	**0.67**	-0.39	-0.53	0.36	-0.44	1		
ssh	-0.21	-0.14	-0.04	-0.28	-0.09	0.01	1	
curr	0.46	0.40	0.11	0.83	-0.06	0.35	-0.26	1

Simulated species' and group distributions

The distribution of 30 species across the simulation region was generated using a multivariate normal mixture model using code adapted from (Woolley *et al.* 2017). The mixture model was parameterised to generate species with three groups of responses to the environmental variables. The groups were designed to represent bioregions with minimal spatial overlap. These three groupings were determined by temperature and O2 with the remaining six environmental variables having little or no influence. The response of each species belonging to each group to the above environmental variables (betas) was drawn randomly from a multivariate normal distribution with the mean betas for each group tabulated in Table A3.2 and a variance of 0.05. The prevalence (alphas) of simulated species was drawn from a beta distribution and informed by the prevalence of species observed in the Kerguelen Plateau fish dataset. The mixing parameter (that determines the number of species allocated to each group) was set to 0.3, 0.4, 0.3, to generate a roughly even allocation of species to groups. These parameters were used to generate a probability of occurrence for each species in each cell of the simulation region, and these simulated data were simplified to additionally provide a realisation of species' presence/absence. From the realisation of presence/absences, 200 samples were randomly chosen and form the 'sites' used for analysis and methods comparison.

The 'true' distribution of groups was calculated using the average alpha of all species and the mean beta values set in the simulation (Table A3.2) applied to the values of the environmental variables in each cell in the simulation region.

This method of generating the simulated distribution of species is most similar to the model underpinning Species Archetype Models (SAMs). The code for generating simulated species' and group distributions can be found: Sim_Setup/simulate_communities_final.R.

Table A3.2. Mean response (betas) of each group (bioregion) to environmental variables used in simulation study.

Group	temp	O2	NO3	sal	depth	chla	ssh	curr
1	0.75	0	0	0	0	0	0	0
2	0	-0.5	0	0	0	0	0	0
3	-0.5	0	0	0	0	0	0	0

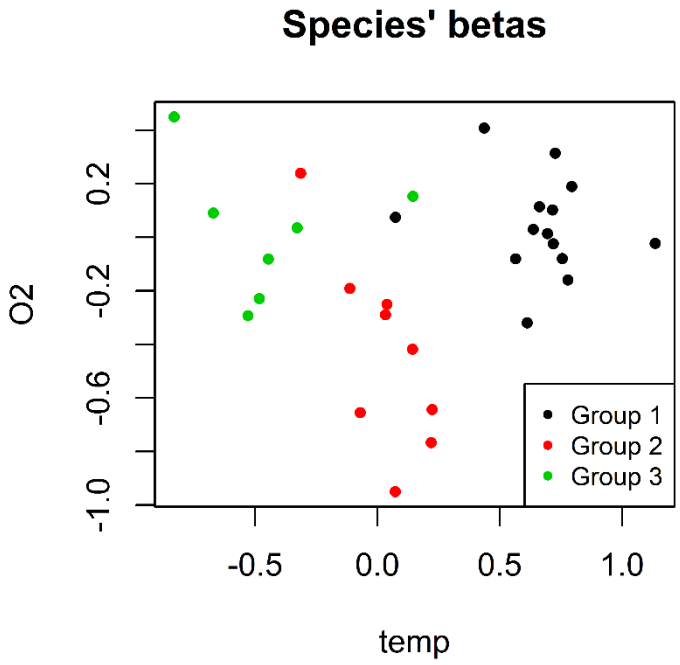


Fig. A3.2. Response (betas) of each species to environmental variables colour coded by group (representing a bioregion).

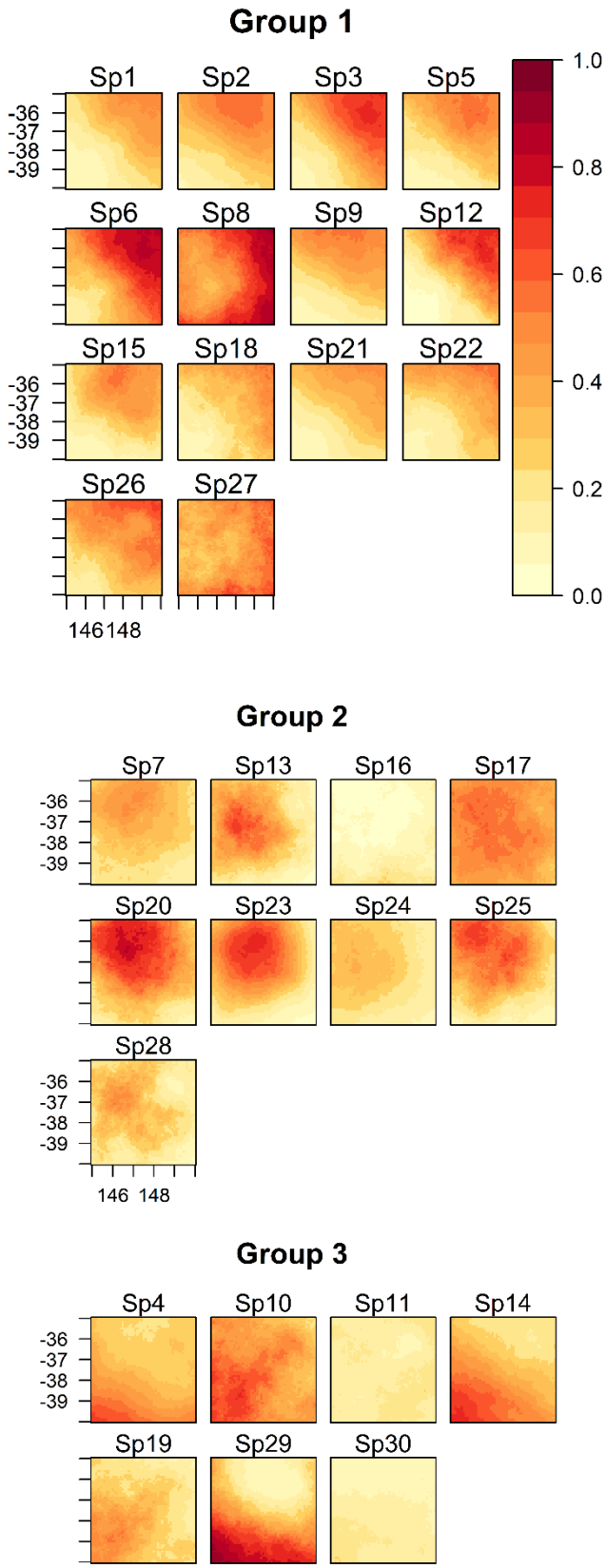


Fig. A3.3. Probability of occurrence of each simulated species across the simulation region, plotted according to their grouped response to the environmental variables.

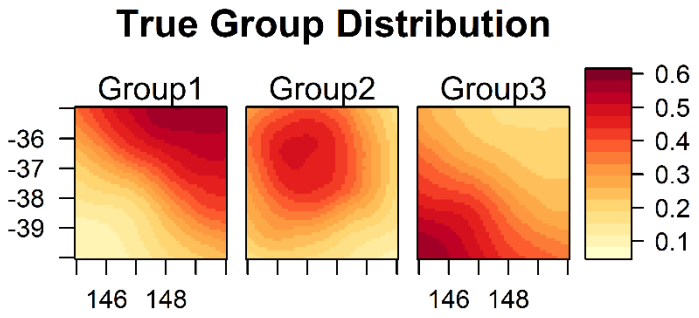


Fig. A3.4. 'True' distribution of species' groups in simulation used to represent bioregions.

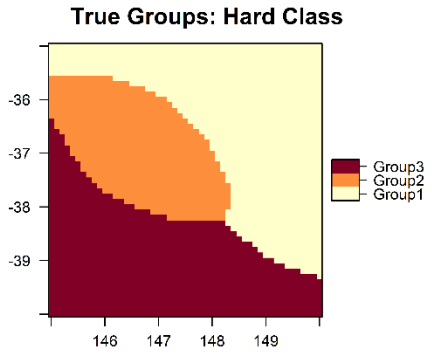


Fig. A3.5. Hard-class version of 'true' distribution of species' groups. Hard classes generated by assigning each cell it's most probable group.

Additional Simulation Results

Code to run the selected models can be found in 'Simulation_Run_Models.R' and additional code provided in the 'Simulation' folder was used to interpret and generate plots.

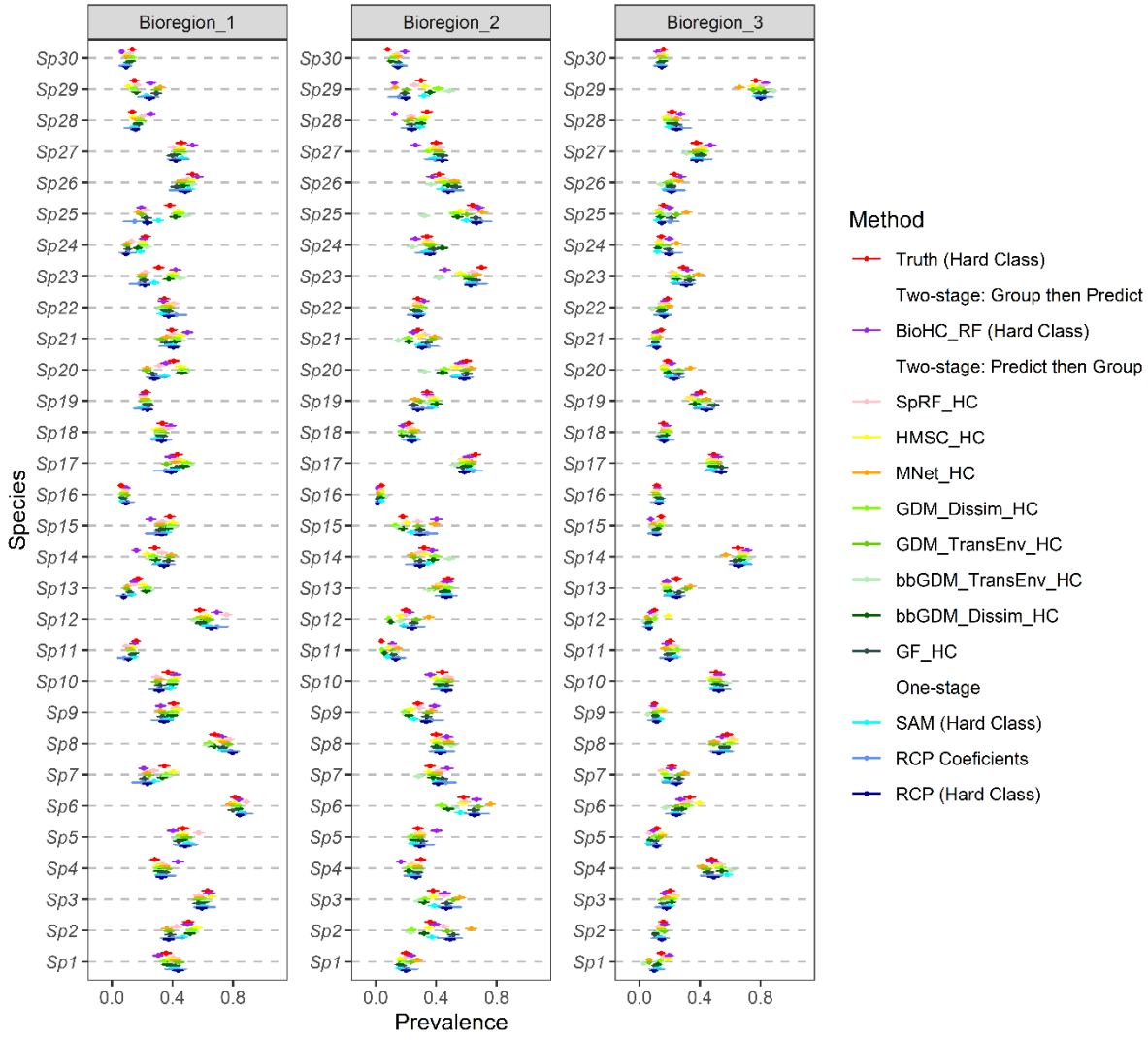


Fig. A3.6. Full species composition of each simulated nioregion, when the number of groups has been fixed to three. Mean and standard error of the prevalence (or probability of occurrence for RCPCoeficients) of each species in each group. Species composition for all two-stage methods was calculated by summarising the prevalence of species at clustered survey sites. To calculate equivalent, site-based measures for the 'true' distribution, SAM and RCP, the probabilistic predictions were converted to a hard class by assigning each site it's nost probable bioregion (denoted 'Hard Class'), and the prevalence of each species observed in each group calculated. For RCP, the expected probability of occurrence of each species in each group was also calculated directly from the model using model coefficients and bootstrap sampling (RCP Coeficients). Acronyms match those specified in Table 1 of the main paper.

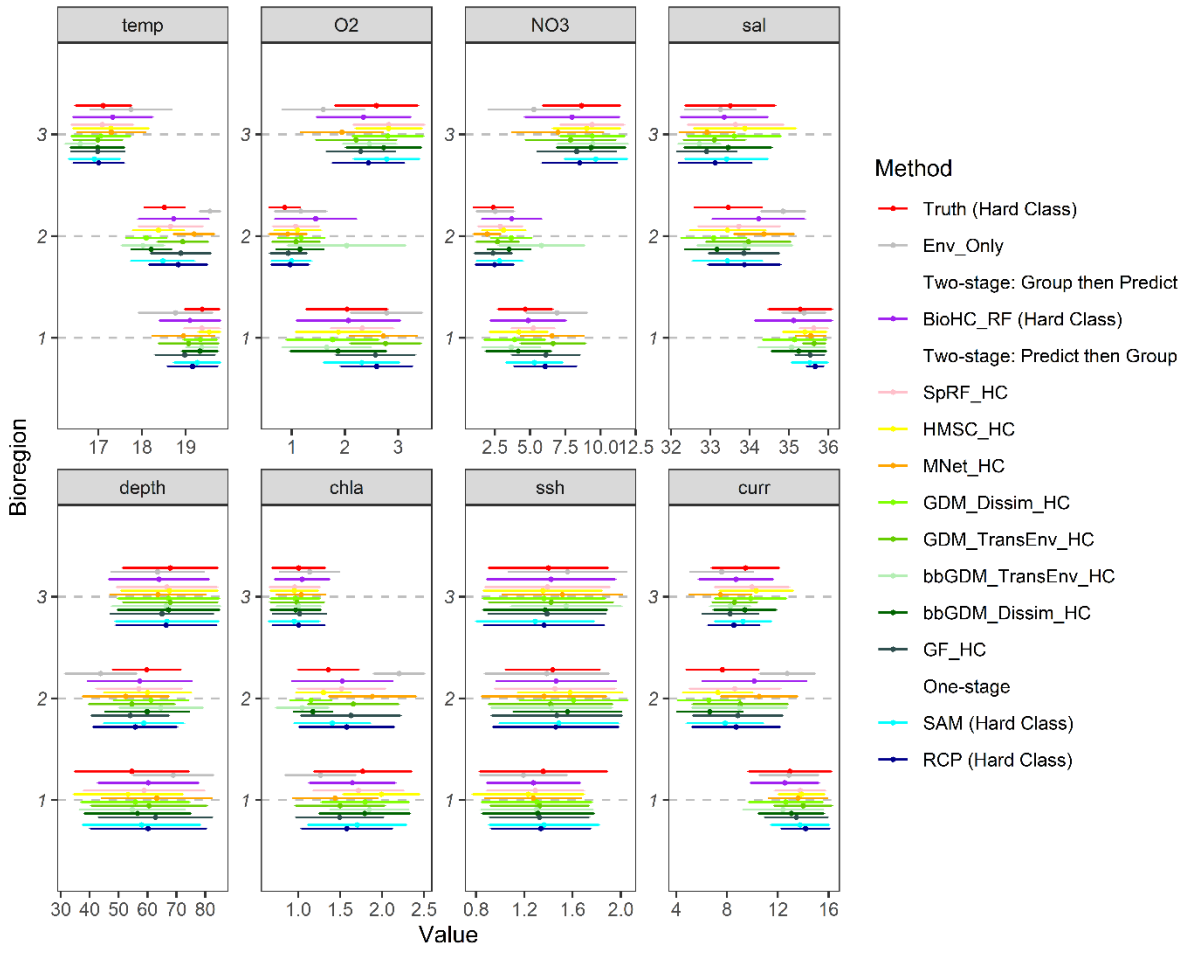


Fig. A3.7. Full environmental characteristics of each simulated bioregion determined for each method when the number of groups is fixed at three. For all two-stage methods, the average (and SD) environmental conditions for each group were calculated from the clustered survey sites. For comparative purposes, SAM, RCP and BioHC_RF results were also calculated using the hard class conversion of the probability predictions (denoted "Hard Class").Acronyms match those specified in Table 1 of the main paper.

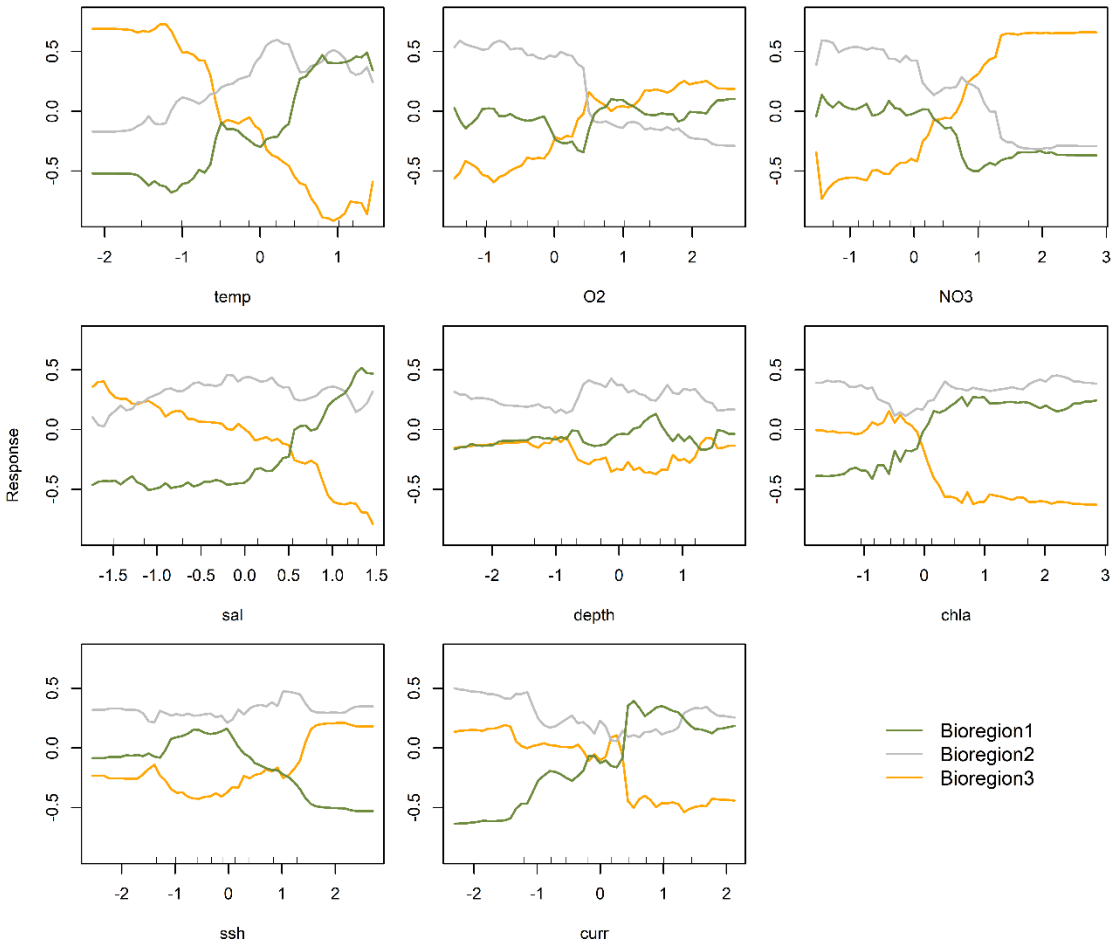


Fig. A3.8. Partial response plots for all environmental variables considered in the BioHC_RF method when the number of groups is fixed at three. The non-linear and non-smooth bioregion responses to environmental variables that are characteristic of tree-based methods.

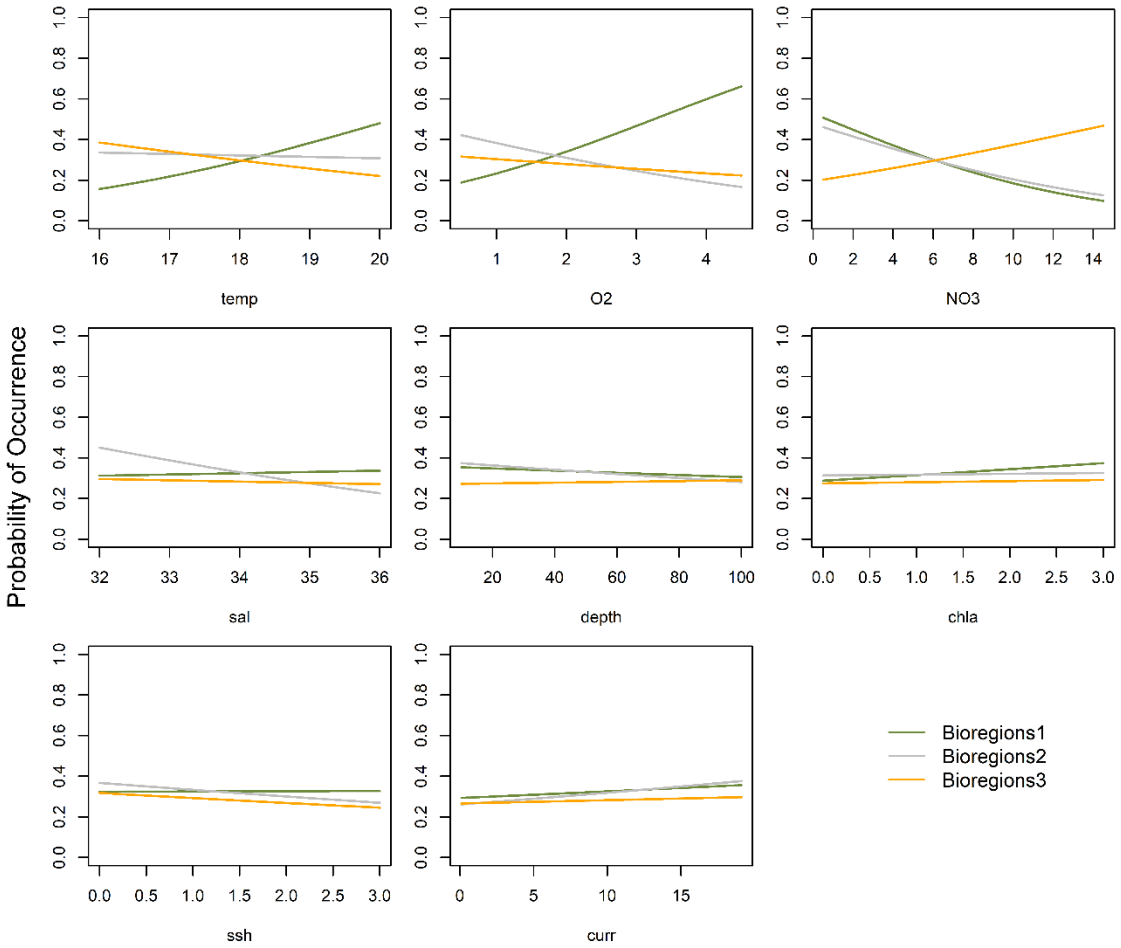


Fig. A3.9. Partial response plots for all environmental variables considered in the SAM method, when the number of groups is fixed at three. The SAM bioregions exhibit a mostly linear response and at any one site the sum of all groups present can exceed one.

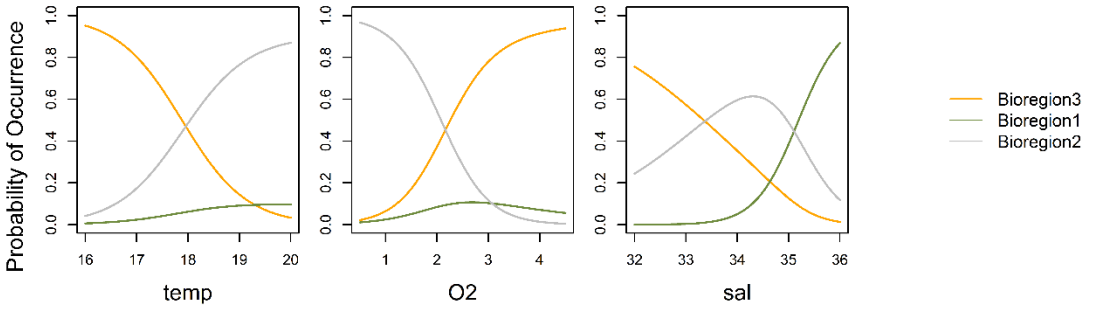


Fig. A3.10. Partial response plots for all environmental variables considered in the RCP method, when the number of groups is fixed at three. The RCP bioregions have a smooth non-linear partial response which highlights the fact the probability of the sum of all bioregions at any one site is constrained to one for RCPs, unlike in SAMs.

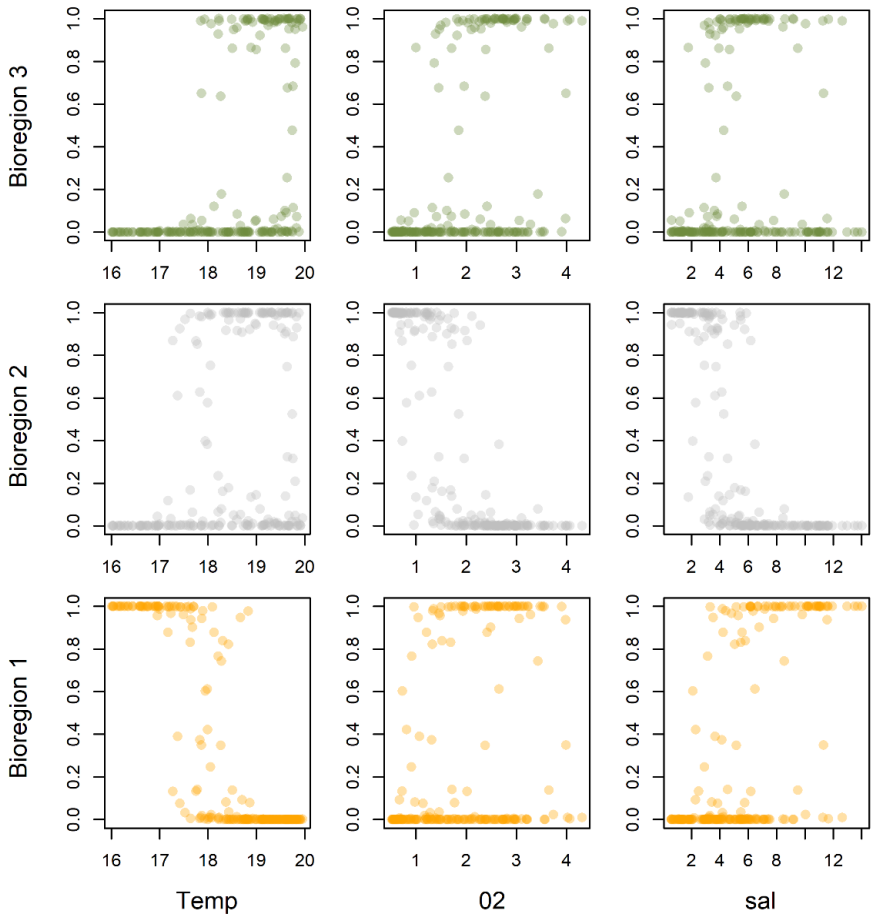


Fig. A3.11. RCP bioregional membership probabilities (pi; when the number of groups is fixed at three) for each modelled site plotted against corresponding environmental variables used to construct the RCP model. These values are directly from the RCP model and are most analogous to predictions that use only environmental data. Bioregion 1 is characterised by higher values of temperature; Bioregion 2 is characterised by lower values of O2 and salinity (which are moderately correlated); Bioregion 3 is characterised by low temperatures. This corresponds with Fig. A3.7.

References:

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,** 985-995.

Appendix S4:
Kerguelen Plateau Fish Analysis

Kerguelen Plateau fish and environmental data are archived at Australian Antarctic Data centre:

http://dx.doi.org/doi:10.26179/5f0528de8c1d2

http://dx.doi.org/doi:10.26179/5f055cd217aa8

All code for analysing the demersal fish data on Zenodo: https://zenodo.org/record/3936354

Biological data:

Biological data were obtained from Random Stratified Trawl Surveys conducted in the Australian EEZ surrounding Heard and MacDonald Islands. The surveys were conducted primarily for stock assessment purposes and stratification was based on depth and geomorphology (Duhamel & Hautecoeur 2009). Surveys were conducted on a commercial trawling vessel by a trained scientific observer. Data consists of 524 trawls between 100 and 1200 m depth evenly spread across four surveys in 2006, 2010 and 2013. Otter bottom trawls were towed for ~ 30 minutes (Duhamel & Hautecoeur 2009; Nowara, Lamb & Welsford 2014) and trawls on the shelf were conducted during the daytime to capture icefish which diurnally aggregate near the seafloor. All fish species caught in trawls were recorded. Species nomenclature was based on names published in appendix 5 of the Biogeographic Atlas of the Southern Ocean (Duhamel *et al.* 2014) and common names based on Gon and Heemstra (1990). Species that are primarily pelagic were removed from analyses and some species were aggregated (e.g. *Paraliparis spp., Macrourus spp.* and *Muraenolepis spp*.) Data are presence-absence, and the twenty species that occur in at least 10 trawls (2% sites) were retained for analyses. Our choice of cut-off value for species' occurrence is somewhat arbitrary. For our purposes here we do not feel that very rare species add much to the analysis and are generally poorly modelled by SDM methods (although multi-species methods will tend to do better than single species methods (Hui *et al.* 2013)). The choice of any cut-off value, and what they represent for conservation, is left to the practitioner's discretion.

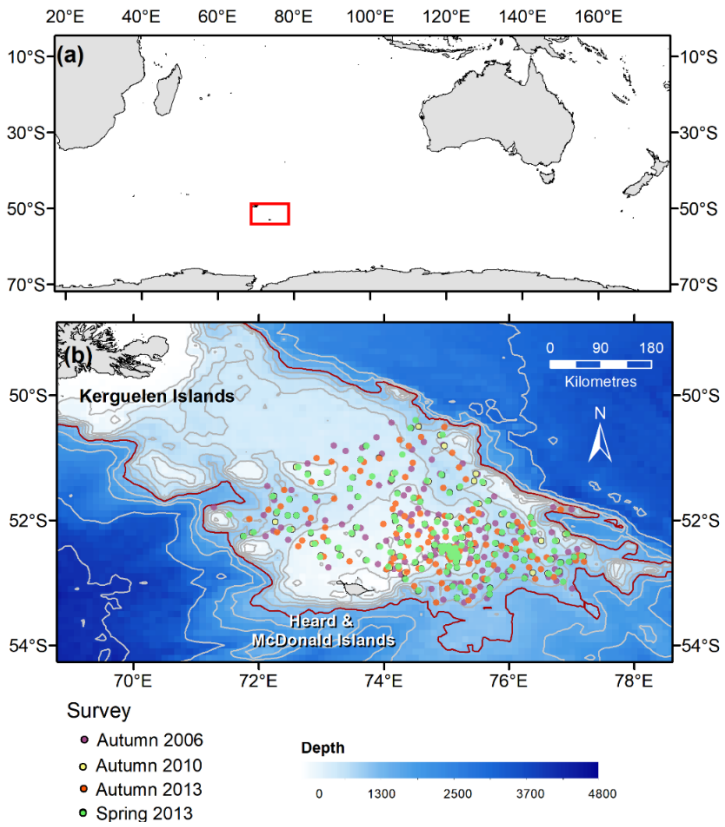


Fig. A4.1. Location of A) Kerguelen Plateau and B) survey sites. Sites are colour-coded by survey and the red contour line is the 1200 m the limit of the deepest trawls.

Environmental data:

Environmental climatological variables representing sea floor and sea surface conditions likely to affect the distribution of demersal fish were obtained from various sources (outlined in (Hill *et al.* 2017)) at a resolution of 0.1 degree. Variables were screened so that those retained for analyses were not highly correlated (< |0.7|). This process left the eight variables described in Table A4.1 and plotted in Fig. A4.2.

Table A4.1. Description and source of environmental variables used in analyses of Kerguelen Plateau demersal fish

	Variable	Description	Units	Source	Reference
Sea- floor	Depth	Seafloor depth	m	Trawl data/ Polar Environmental Data	Raymond (2012)
	Slope	Seafloor slope	degrees	Polar Environmental Data	Raymond (2012)
	Floor_temp	Average temperature near seafloor	° C	Polar Environmental Data	Raymond (2012)
	Current	Average current speed near seafloor	m/s^2	Polar Environmental Data	Raymond (2012)
	NO3_mean	Average nitrate concentration near seafloor	µmol/l	CSIRO Atlas of Regional seas (CARS)	Ridgway, Dunn and Wilkin (2002)
Sea- surface	Surface_temp	Average of daily surface temperature (1982- 2014)	° C	NOAA OI SST v2	Reynolds *et al.* (2007)
	Chla_SD	Standard deviation of yearly mean chl-a (1997-2010)	mg/m^3	L3 SeaWiFs data corrected for Southern Ocean	Johnson *et al.* (2013)
	ssha_SD	Standard deviation of sea surface height (indicates surface currents and fronts)	mm/km	Polar Environmental Data	Raymond (2012)

* Slope and current were log transformed prior to analyses.

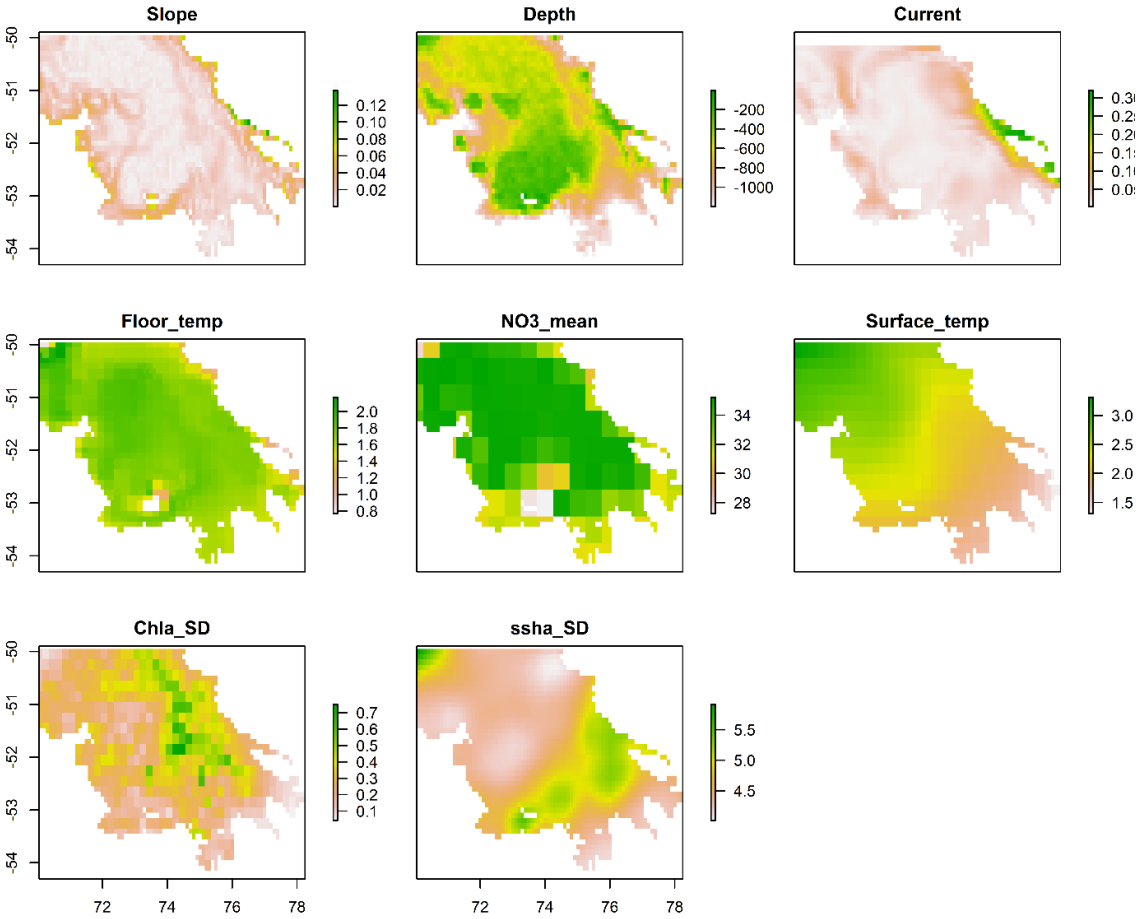


Fig. A4.2. Maps of environmental variables used in analyses of Kerguelen Plateau demersal fish. All maps were cropped at 1200 m corresponding to the depth limit of the trawls.

Additional Kerguelen Plateau Fish results

Analysis of a real dataset, demersal fish on the Kerguelen Plateau, yielded results with many similarities to the simulation results. The number of bioregions identified as optimal varied between the methods. Many of the two-stage methods only identified two bioregions (Fig. A4.3), the exceptions being MNet_HC (four bioregions), and the clustering of the transformed environmental spaces of the naïve and bootstrapped GDM (three bioregions). The one-stage methods, SAM and RCP identified four and five bioregions respectively. Nearly all methods distinguished a shallow - water bioregion (Bioregion 1) surrounding Heard and MacDonald Islands (HIMI) and on the banks. The spatial distribution of the Env_only and MNet bioregions were very similar and showed greater discrimination of depths as did the one-stage methods, SAM and RCP (Fig. A4.3).

If we assume there are in fact four groups from this point forward and compare hard-clustered predictions, patterns in distribution of groups were generally more similar amongst the methods (Fig. A4.4). Most methods consistently identified a shallow (Bioregion 1) and deep bioregion (Bioregion 4) with boundaries of the intermediate-depth bioregions more variable (Fig. A4.4). The one-stage methods are the only methods that capture uncertainty in the entire bioregionalisation process. For SAM, there are no clear patterns in uncertainty within bioregions, however, the highest uncertainty across all bioregions appears in the NW of the study region (Fig. A4.5). For the RCP bioregions, uncertainty was highest between the boundaries of bioregions 1 and 4, and in areas of moderate probability of occurrence in bioregions 2 and 3 (Fig A4.5)

Patterns between methods in the environmental conditions characterising the groups are difficult to discern. However, most methods distinguish depth bands for the different groups, while these appear to overlap more for SAM bioregions (Fig A4.6). Of all the methods the BioHC_RF most often stood apart from the others in its environmental profile (Fig A4.6). Methods that produce partial plots again highlight the depth niche of the different bioregions (Figs. A4.7-9), with some bioregions also differentiated by other environmental variables (e.g. surface temp for BioHC_RF and RCP (Fig. A4. 7 & 9) and chla for SAMs (Fig. A4. 8). RCP groups have a more curvi-linear pattern that is due to the fact that the probability of occurrence of all groups at a particular site is constrained to one. This means that as the probability of a site being one group increases the probability of it being other groups must decrease.

Patterns in the species associated with each bioregion are also complex. For many species and methods there is reasonable agreement in composition (Fig. A4.10) particularly for Bioregions 1 and 4. There was less agreement in the composition of Bioregions 2 and 3. No one particular methods consistently stood apart from the rest.

Finally we provide and interpretation of bioregions from the RCP method that combines the species composition (Fig A4.11) and environmental characteristics (Figs A4.9) derived directly from model parameters. RCP bioregions are mostly distinguished by depth, with bioregion 1 a shallow bioregion occurring in depths <300m and with a high prevalence of endemic and/or shelf species such as *G.acuta, C. gunnari* and *C. rhinoceratus*. Bioregion 2 is most likely found in depth around 300-400m and contains some similar species to bioregion 1 (e.g. *C. rhinoceratus*) with additional species becoming more prevalent (e.g. *L. squamiforms*). Bioregion 3 is deeper again, most prevalent around 600 m and on the NW of the Plateau corresponding with warmer surface temperatures and contains known deeper-water species (e.g. *Macrourus spp*.). The deepest bioregion, increasingly likely to be found at depth greater than 600m and cooler temperatures, is a species poor bioregion predicted to mostly contain *Macrourus spp*. and *D. eleginoides*. Some species such as the Patagonian Toothfish (*D. eleginoides*) are ubiquitous on the plateau and highly prevalent in all bioregions. This

interpretation is consistent with what is known of the biogeography and ecology of the region (See Hill *et al.* (2017) more detailed discussion).

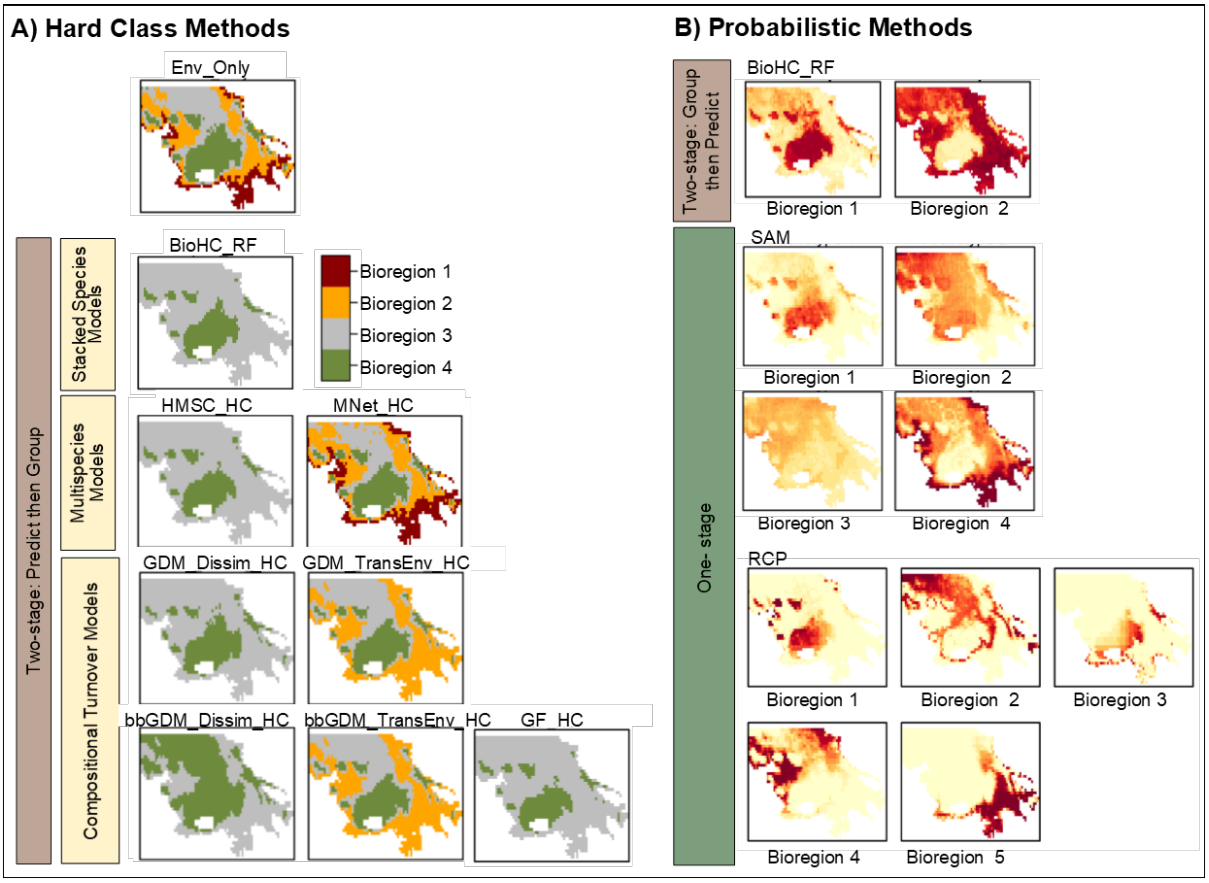


Fig. A4.3. Distribution of bioregions identified for demersal fish on the Kerguelen Plateau using the various modelling methods. A) Bioregions identified by clustering the environmental data only (Env_Only) and by most two-stage approaches that produce hard classes from the hierarchical clustering. B) Probability of occurrence of each bioregion for methods that produce probabilistic outputs. Note that the BioHC_RF probabilities represent only the second stage of the analysis. Method abbreviations match those in Table 1.

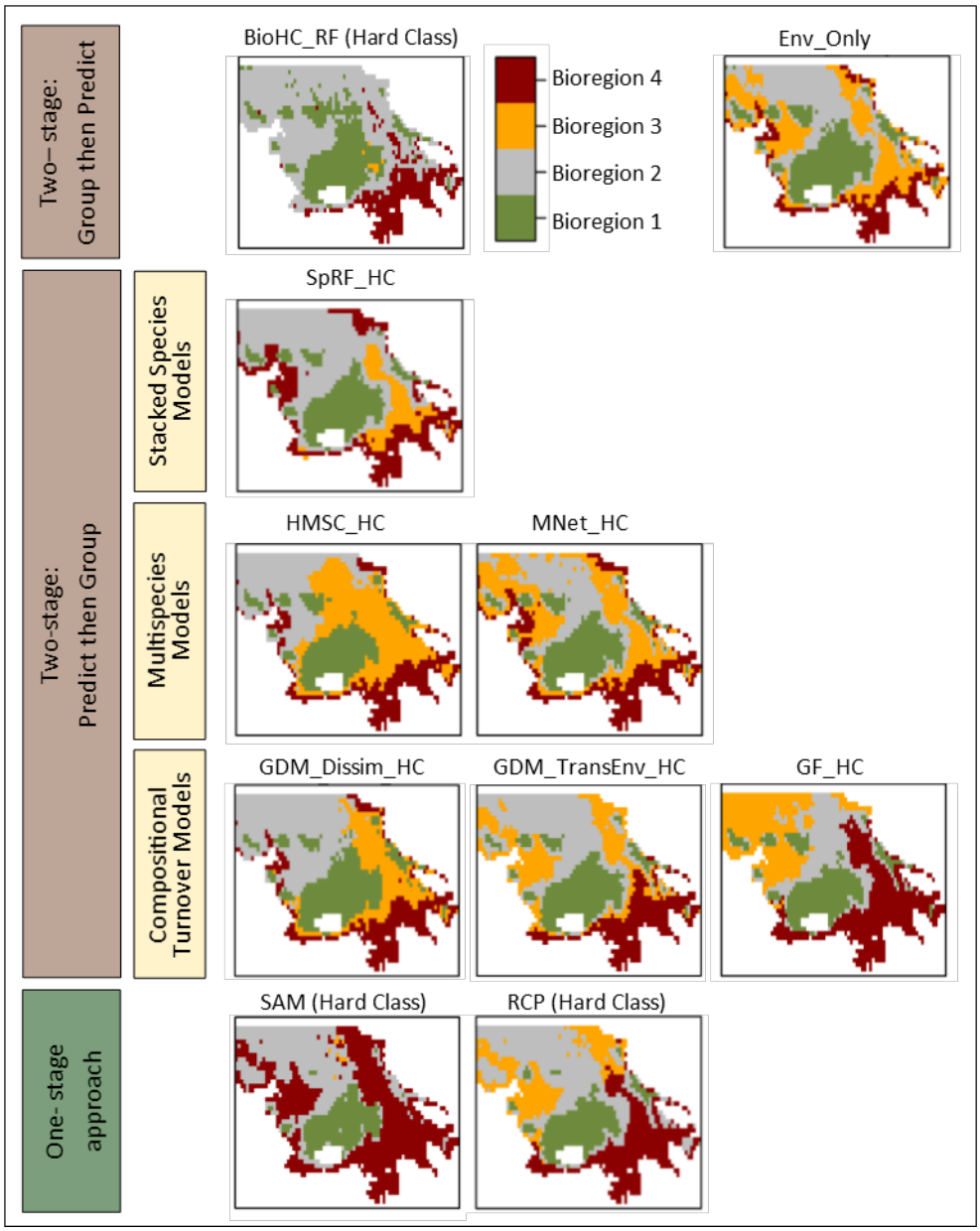


Fig. A4.4. Distribution of Kerguelen Plateau demersal fish bioregions for each method when the number of bioregions has been fixed at four and where cells for methods with probabilistic outputs are assigned their most likely bioregion (denoted by 'Hard Class'). The spatial distribution of bioregions are more similar between methods when the number of bioregions are set to four and largely reflect depth-related patterns. Bioregions have been colour-coded to best highlight similarities in the distribution of bioregions. Method abbreviations match Table 1.

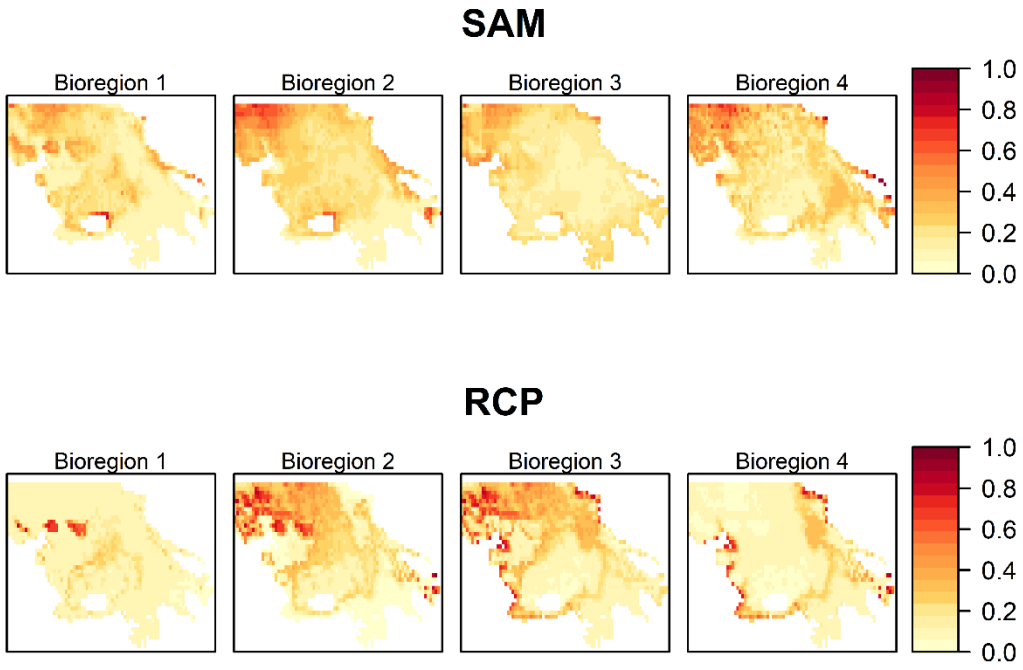


Fig. A4.5. Uncertainty associated with the one-stage approaches, Species Archetype Models (SAM) and Regions of Common Profile (RCP).

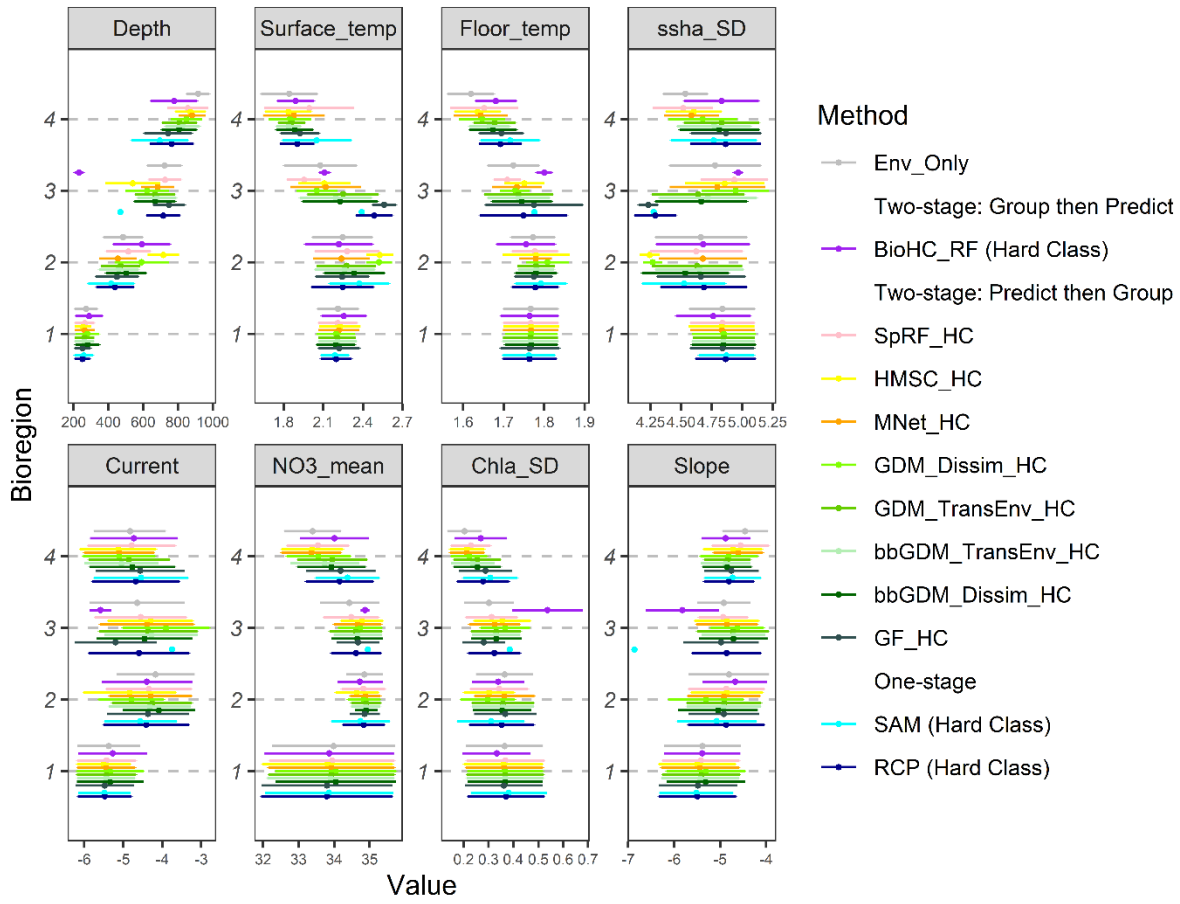


Fig. A4.6. Environmental characteristics of each bioregion determined for each method when the number of groups is fixed at four. For all two-stage methods, the average (and SD) environmental conditions for each group is summarised from the clustered survey sites. For comparison BioHC_RF, SAM and RCP values were calculated by first converting probabilistic predictions to hard classes then summarising the environmental conditions observed at each site belonging to each group. Method abbreviations match those in Table 1 of the main paper.

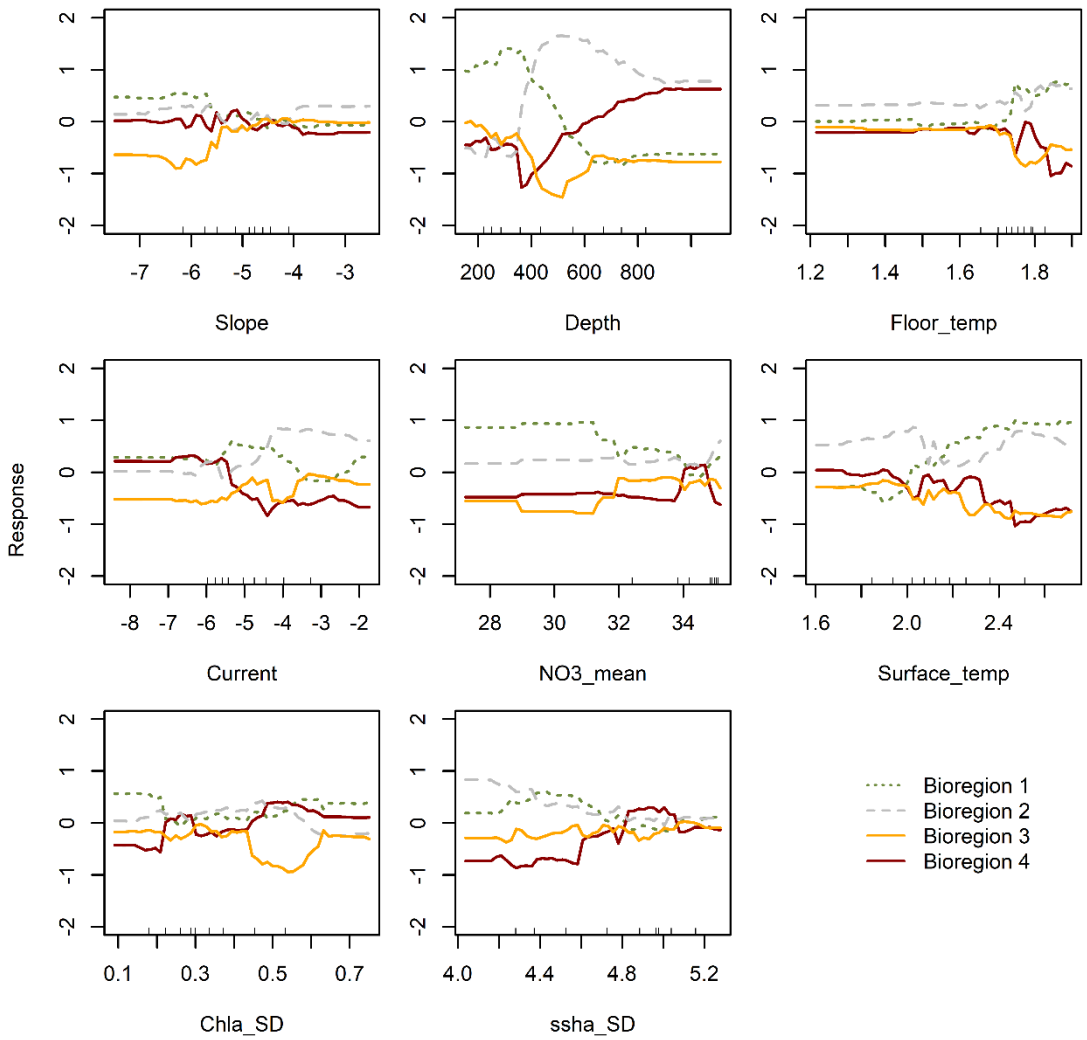


Fig. A4.7. Partial response plots for the BioHC_RF method when the number of groups is fixed at four. Response plots are calculated using random forests of the hierarchically clustered biological data.

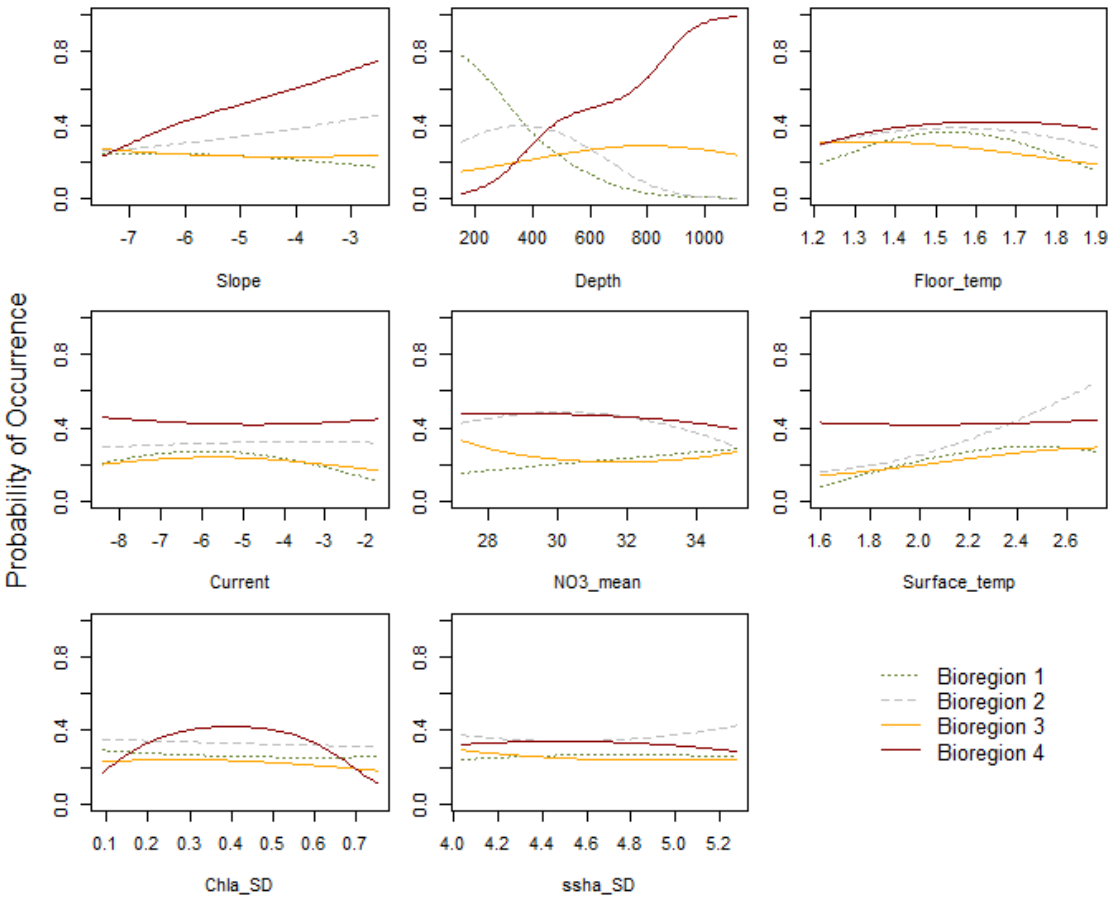


Fig. A4.8. Partial response plots for the one-stage method, SAM, when the number of groups is fixed at four. Response plots are calculated using the SAM model coefficients, which describe the relationship between groups and environmental variables, and holding all variables except the variable of interest at their mean values.

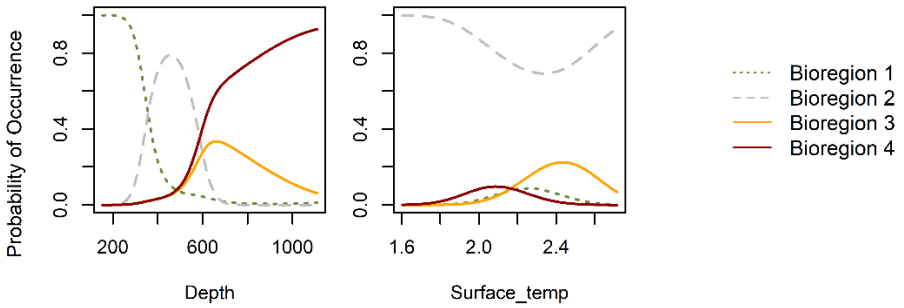


Fig. A4.9. Partial response plots for the one-stage method, RCP, when the number of groups is fixed at four. Response plots are calculated using the RCP model coefficients, which describe the relationship between groups (and species) and environmental variables, and holding all variables except the variable of interest at their mean values. Only two variables are included because these were selected in the final model.

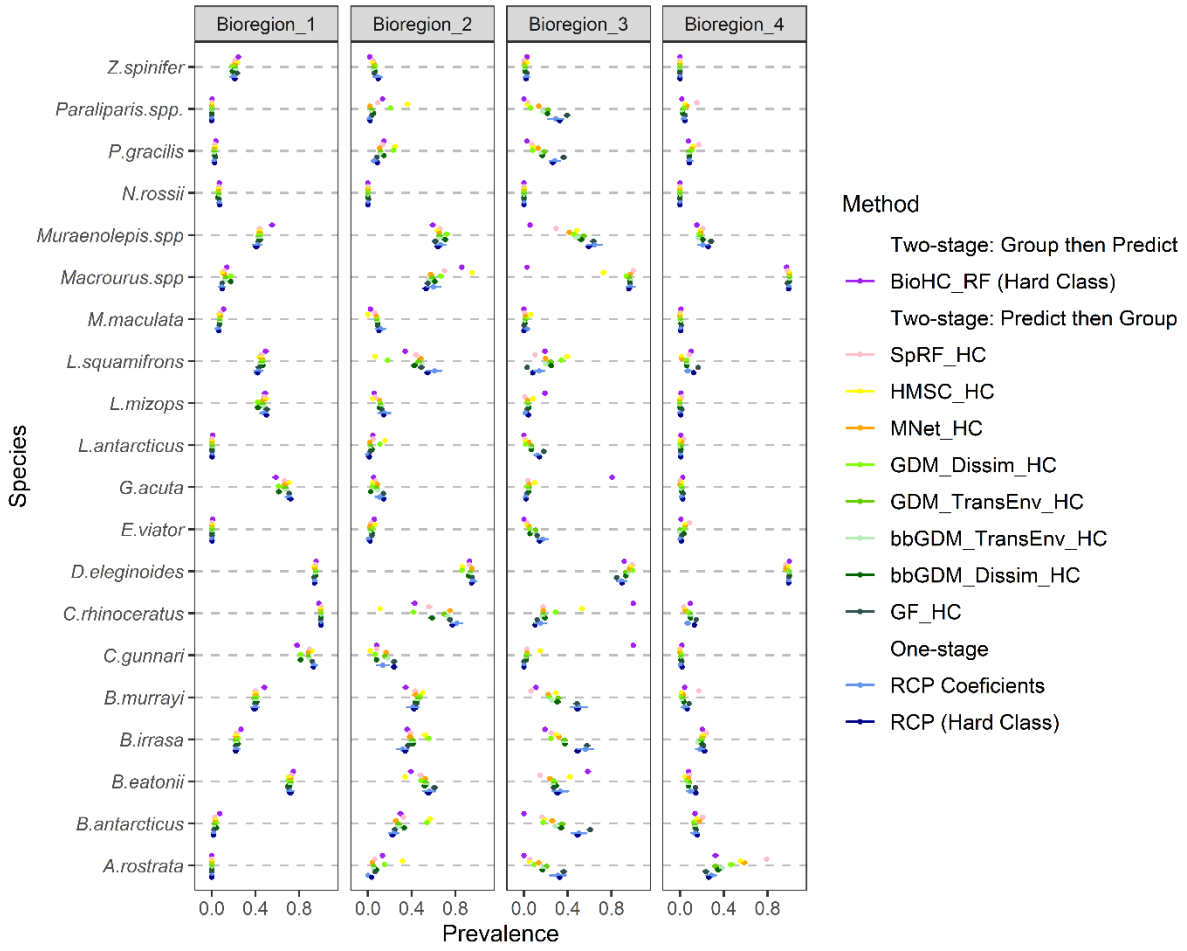


Fig. A4.10. Composition of species in each bioregion determined for each method when the number of groups is fixed at four. Mean and standard error of the prevalence (or probability of occurrence for RCP_Coefs) of each species in each group. Species composition for all two-stage methods was calculated by tabulating the prevalence of species at classified survey sites. To calculate equivalent, site-based measures for RCP (RCP_Hard), the probabilistic predictions were converted to hard classes and the prevalence of each species observed in each group tabulated. For RCP, the expected probability of occurrence of each species in each group was also calculated directly from the model using model coefficients and bootstrap sampling (RCP_Coeficients). Method abbreviations match those in Table 1 of the main paper.

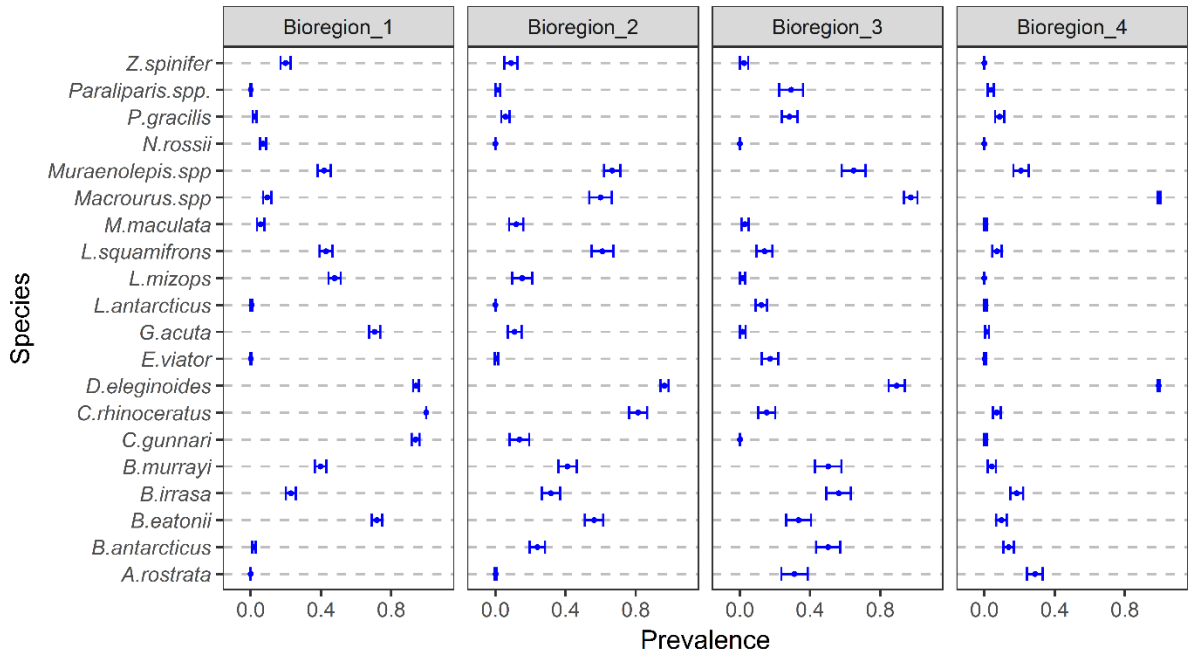


Fig. A4.11. Species profile for each RCP when the number of groups is fixed at four. Species' mean (and SE) prevalence in each RCP was determined using model co-efficients and Bayesian bootstrap sampling.

References:

Duhamel, G. & Hautecoeur, M. (2009) Biomass, abundance and distribution of fish in the Kerguelen Islands EEZ (CCAMLR statistical division 58.5.1). *CCAMLR Science,* **16,** 1-32.

Duhamel, G., Hulley, P.A., Causse, R., Koubbi, P., Vacchi, M., Pruvost, P., . . . Van de Putte, A.P. (2014) Biogeographic patterns of fish. *Biogeographic Atlas of the Southern Ocean* (eds C. De Broyer, P. Koubbi, H.J. Griffiths, B. Raymond, C. d'Udekem d'Acoz, A.P. Van de Putte, B. Danis, B. David, S. Grant, J. Gutt, C. Held, G. Hosie, F. Huettmann, A. Post & Y. Ropert-Coudert), pp. 328-362. Scientific Committee on Antarctic Research, Cambridge UK.

Gon, O. & Heemstra, P.C. (1990) *Fishes of the Southern Ocean*. J.L.B. Smith Institute of Ichthyology, Grahamstown, R.S.A.

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,** 1216-1230.

Hui, F.K.C., Warton, D.I., Foster, S.D. & Dunstan, P.K. (2013) To mix or not to mix: comparing the predictive performance of mixture models vs. separate species distribution models. *Ecology,* **94,** 1913-1919.

Johnson, R., Strutton, P.G., Wright, S.W., McMinn, A. & Meiners, K.M. (2013) Three improved satellite chlorophyll algorithms for the Southern Ocean. *Journal of Geophysical Research: Oceans,* **118,** 3694-3703.

Nowara, G.B., Lamb, T.D. & Welsford, D.C. (2014) The 2014 annual random stratified trawl survey in the waters of Heard Island (Division 58.5.2) to estimate the abundance of *Dissostichus eleginoides* and *Champsocephalus gunnari. CCAMLR Document*

Raymond, B. (2012) Polar environmental data layers. Australian Antarctic Data Centre, CAASM Metadata.

Reynolds, R.W., Smith, T.M., Liu, C., Chelton, D.B., Casey, K.S. & Schlax, M.G. (2007) Daily High-Resolution-Blended Analyses for Sea Surface Temperature. *Journal of Climate,* **20,** 5473-5496.

Ridgway, K.R., Dunn, J.R. & Wilkin, J.L. (2002) Ocean interpolation by four-dimensional least squares -Application to the waters around Australia. *Journal of Atmospheric and Ocean Technology,* **19,** 1357-1375.