

**Bioregions in marine environments: Combining Biological and Environmental Data for Management and Scientific Understanding**

Journal:	<i>BioScience</i>
Manuscript ID	BIOS-19-0020.R2
Manuscript Type:	Overview Article
Date Submitted by the Author:	n/a
Complete List of Authors:	<p>Woolley, Skipton; CSIRO, Oceans and Atmosphere          Foster, Scott; CSIRO, Data61          Bax, Nicolas; University of Tasmania, IMAS          Currie, Jock; Nelson Mandela University, Institute for Coastal and Marine Research          Dunn, Daniel; Duke University, Marine Geospatial Ecology Lab          Hansen, Cecilie; Institute of Marine Research, Ecosystem Modelling          Hill, Nicole; University of Tasmania Institute for Marine and Antarctic Studies, biodiversity modelling          O'Hara, Timothy; Museums Victoria, Sciences Department          Ovaskainen, Otso; Helsingin Yliopisto, Organismal and Evolutionary Biology Research Programme          Sayre, Roger; US Geological Survey, Land Change Science Program          Vanhatalo, Jarno; Helsingin Yliopisto, Department of Mathematics and Statistics          Dunstan, Piers; CSIRO, Oceans and Atmosphere</p>
Key words:	biogeography, community ecology, statistics, marine biology
Abstract:	<p>Bioregions are important tools for understanding and managing natural resources. Bioregions should describe where relatively homogenous assemblages of species, enabling managers to better regulate activities that might affect these assemblages. Many existing bioregionalisation approaches, which rely on expert derived, delphic comparisons or environmental surrogates do not explicitly include observed biological data in such analyses. We highlight that for bioregionalisations to be useful and reliable for systems scientists and managers, bioregionalisations need to be based on biological data, include an easily understood assessment of uncertainty, preferably in a spatial format matching the bioregions, and be scientifically transparent and reproducible. Statistical models provide a scientifically robust, transparent and interpretable approach for ensuring that bioregions are formed based on observed biological and physical data. Using statistically-derived bioregions provides a repeatable framework for the spatial representation of biodiversity at multiple spatial scales. This results in better informed management decisions and biodiversity conservation outcomes.</p>

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60



SCHOLARONE™  
Manuscripts

# Bioregions in marine environments: Combining Biological and Environmental Data for Management and Scientific Understanding

Skipton N.C Woolley<sup>1</sup>, Scott D Foster<sup>2</sup>, Nicholas J Bax<sup>1,3</sup>, Jock C Currie<sup>4</sup>, Daniel C. Dunn<sup>5</sup>, Cecilie Hansen<sup>6</sup>, Nicole Hill<sup>3</sup>, Timothy D O'Hara<sup>7</sup>, Otso Ovaskainen<sup>8,9</sup>, Roger Sayre<sup>10</sup>, Jarno P Vanhatalo<sup>8,11</sup> & Piers K Dunstan<sup>1</sup>

<sup>1</sup>*Oceans and Atmospheres, CSIRO, Hobart, AUS.*

<sup>2</sup>*Data61, CSIRO, Hobart, AUS.*

<sup>3</sup>*Institute for Marine and Antarctic Studies, University of Tasmania, Hobart, AUS*

<sup>4</sup>*Nelson Mandela University & South African National Biodiversity Institute, Cape Town, South Africa*

<sup>5</sup>*Marine Geospatial Ecology Lab, Duke University, Durham, NC, USA*

<sup>6</sup>*Institute of Marine Research, Bergen, Norway*

<sup>7</sup>*Museums Victoria, GPO Box 666, Melbourne, VIC 3001, Australia*

<sup>8</sup>*Organismal and Evolutionary Biology Research Programme, P.O. Box 65, 00014 University of Helsinki, Finland.*

<sup>9</sup>*Centre for Biodiversity Dynamics, Department of Biology, Norwegian University of Science and Technology, N-7491 Trondheim, Norway.*

<sup>10</sup>*Land Change Science Program, U.S. Geological Survey, Reston, VA, USA*

<sup>11</sup>*Department of Mathematics and Statistics, University of Helsinki, Finland.*

## Abstract

Bioregions are important tools for understanding and managing natural resources. Bioregions should describe where relatively homogenous assemblages of species, enabling managers to better regulate activities that might affect these assemblages. Many existing bioregionalisation approaches, which rely on expert derived, delphic comparisons or environmental surrogates do not explicitly include observed biological data in such analyses. We highlight that for bioregionalisations to be useful and reliable for systems scientists and managers, bioregionalisations need to be based on biological data, include an easily understood assessment of uncertainty, preferably in a spatial format matching the bioregions, and be scientifically transparent and reproducible. Statistical models provide a scientifically robust, transparent and interpretable approach for ensuring that bioregions are formed based on observed biological and physical data. Using statistically-derived bioregions provides a repeatable framework for the spatial representation of biodiversity at multiple spatial scales. This results in better informed management decisions and biodiversity conservation outcomes.

## 37 **Introduction**

38 The distribution of species is a function of many controlling influences operating at a diversity of  
39 scales, including environmental heterogeneity and stability in space and time (Rohde 2007), genetic  
40 and evolutionary history (Webb et al. 2002), intra- and inter-specific species interactions such as  
41 predation, competition and facilitation (Polechová & Barton 2005), dispersal dynamics (Ronce  
42 2007), and environmental disturbances (Sheil 2016). Irrespective of history, the present patterns can  
43 be organised spatially creating a biogeography (Last et al. 2010; Ebach & Parenti 2015). With  
44 knowledge of where all species exist, scientists would be in a better position to understand why  
45 species are distributed as they are, a fundamental line of biogeographic inquiry. Moreover,  
46 managers would be in a better position to manage species and their assemblages for a variety of  
47 applications, including the conservation and sustainable use of biodiversity and ecosystems, and  
48 their associated goods and services.

49  
50 The quantity and quality of observations of data required to precisely understand where all species  
51 are located is impractical to achieve. This is particularly so in ecosystems that are vast and  
52 inaccessible, such as our focus: the marine ecosystem. All individual species cannot be surveyed,  
53 and, even for well-studied species, complete knowledge on their distributions remains highly  
54 uncertain. From a scientific perspective, knowledge of the distribution of species is still  
55 fundamentally lacking, despite long-term, ongoing efforts to compile observational datasets for a  
56 broad range of taxonomic groups (e.g. Ocean Biogeographic Information System; Grassle 2000).  
57 Consequently, to better understand how species are placed within their environment, tools like  
58 species distribution models are used to describe their distributions (Guisan & Zimmermann 2000).  
59 When considering a few species, the use of individual species distribution models is a logical  
60 approach to describe their distributions. However, with many species in a geographical region,  
61 researchers may want to move beyond individual species' distributions and better understand the  
62 distributions of species assemblages, communities, ecosystems and bioregions (Fig 1a; Ferrier &  
63 Guisan 2006; Warton et al. 2015). To do this, scientists often engage in biogeographic  
64 classification, otherwise known as bioregionalisation, ecoregionalisation, zoogeographic  
65 classification, and ecological mapping (Ebach & Parenti 2015). Here we consider that  
66 bioregionalisations are a biological and physical partitioning of geographic space based on the  
67 spatial distribution of multiple species, communities, ecosystems, or other biological characteristics.  
68 This description shares many concepts with approaches such as vegetation classification, ecosystem  
69 characterisation, ecoregions and fisheries regions (Begg et al. 1999).

1  
2 71 Bioregions are a simplification (a model) of the true distribution of multiple species that share a  
3  
4 72 similar ecological and abiotic preference and sometimes an evolutionary history. Bioregion maps  
5  
6 73 may be useful for managing multiple different human activities in a region because they simplify  
7  
8 74 complex information into a form that humans are inherently good at understanding (May 1976).  
9  
10 75 Bioregions should define the key physical and biological attributes of a region, provide a simplified  
11  
12 76 understanding of the ecosystems and can be an effective way to compare geographic differences in  
13  
14 77 species composition from local to global scales (Fig. 1b). Bioregions can help contextualise spatial  
15  
16 78 management in a framework that is transparent for decision-making. Decision makers often need to  
17  
18 79 know the spatial extent of biological diversity in order to assess a management action in the context  
19  
20 80 of the biological component of interest (Fig. 1c). Bioregional classifications provide the  
21  
22 81 fundamental building blocks to inform the most appropriate management tools for any geographic  
23  
24 82 area (CBD 2010). Managers might want to assess the impact of human activity (Leaper et al. 2012),  
25  
26 83 gauge the representativeness of a protected area system (Brunckhorst & Bridgewater 1995), or  
27  
28 84 establish representative monitoring programs (Hutchings et al. 2009), within and across bioregions.  
29  
30 85

31  
32 86 Herein, we advocate for the continued development of statistical bioregions to increase scientific  
33  
34 87 understanding of the distribution of biodiversity and to support resource management. We identify  
35  
36 88 the desired characteristics of bioregions, emphasising the importance of appropriate statistical  
37  
38 89 methods in their derivation. We provide a case study in the marine environment to demonstrate one  
39  
40 90 example of how a statistical bioregionalisation can be conducted. As implementation of management  
41  
42 91 based on biogeographic classification continues to be developed, there is a need for rigorous,  
43  
44 92 transparent and well-accepted statistical biogeographic characterisations to deliver improved  
45  
46 93 management tools to support sustainable use and conservation at local, national and global levels.  
47  
48 94

## 49 95 **The current state of marine bioregionalisation approaches**

50 96 Bioregionalisations often rely heavily on physical, spatial or biological surrogates to describe the  
51  
52 97 distribution of more complex assemblages, communities, ecosystems or bioregions. This approach  
53  
54 98 has been implemented with both expert knowledge (UNESCO 2009) and statistical modelling  
55  
56 99 (Reygondeau et al. 2017; Sayre et al. 2017). These approaches are useful across broad geographic  
57  
58 100 regions where equivalent biological data are lacking or too sparse to inform reliable biological  
59  
60 101 models (Beier & Albuquerque 2015), if their uncertainty is recognised and presented. Global  
102  
103 102 marine bioregional maps have been produced, which ecologically partition the planet based on only  
104  
105 103 abiotic characteristics (environmental drivers) rather than in combination with biotic distributions

1  
2 104 (UNESCO 2009; Longhurst 2010; Sayre et al. 2017). These maps attempt to depict ecological  
3  
4 105 zonations based on environmental variation and are often labelled ‘ecoregions’. Despite the term  
5  
6 106 ‘ecoregions’, we stress that these are not ecological, because there is no explicit link to biological  
7  
8 107 data. They do, however, provide a useful partitioning of environment, which might be better coined  
9  
10 108 as ‘enviro-regions’. An example of this is the Global Open Oceans and Deep Seabed (GOODS;  
11  
12 109 UNESCO 2009) biogeographic classification that assumes that ocean basins delineate species.  
13  
14 110 However, at least for some taxa this assumption, which seems quite plausible at face-value, is not  
15  
16 111 supported by more recent research (O’Hara et al. 2011).

16 112  
17  
18 113 Recent work by Sayre *et al.*, (2017) provides an example of ‘enviro-regions’ – regions based on  
19  
20 114 physical data. This physical regionalisation is both broad-scale (global) and relatively fine-  
21  
22 115 resolution ( $1/4^\circ$  across the globe). Compared to bioregions, ‘enviro-regions’ are relatively easy to  
23  
24 116 create as physical data are usually more accessible and comprehensive. These environmental  
25  
26 117 regionalisations can correlate with variation in biotic distributions and are assumed to be  
27  
28 118 representative of biotic distributions. However, some studies have demonstrated less partitioning of  
29  
30 119 geographic space with the inclusion of biological data (Woolley et al. 2013), suggesting that species  
31  
32 120 might persist at broader ranges of environmental variation than the variation generated from  
33  
34 121 physical data alone. Other studies have shown the danger of over-prediction when using physical  
35  
36 122 data alone, without a better understanding of the biology of the species or community, important  
37  
38 123 physical environmental variables may be lacking from the analysis (Anderson et al. 2016). It  
39  
40 124 follows that addition of biological and ecological information can improve delineation of ecological  
41  
42 125 patterns, through improved accuracy, granularity and reduced bias (Warton et al. 2015).

41 126  
42  
43 127 Biological information is often incorporated into bioregionalisations as expert-derived products  
44  
45 128 (Spalding et al. 2007; UNESCO 2009; Longhurst 2010) and in such form rarely includes estimates  
46  
47 129 of uncertainty (Robinson et al. 2017). In such cases, bioregion boundaries are outlined by experts  
48  
49 130 (humans), often as part of a committee, and likely influenced by anthropogenic requirements (e.g.  
50  
51 131 fisheries regions and stocks; Begg et al. 1999; Longhurst 2010). Good examples of this are fisheries  
52  
53 132 areas that reflect governance boundaries and consequently might not accurately describe the  
54  
55 133 distributions of species or ecosystems of interest (Department of the Environment and Heritage,  
56  
57 134 2006). Expert-derived bioregionalisations are typically quite coarse resolution and often broad in  
58  
59 135 spatial extent (Ekman 1953). While expert information is often easily communicated and is  
60  
136 applicable in situations where data are inadequate, it may not be objective or reproducible.

137

1  
2 138 An alternative approach (which could be based on physical or expert-derived data), is to use  
3  
4 139 *statistical* models that estimate the distribution and content of bioregions based on biological and  
5  
6 140 physical data. Increasingly improved global datasets that incorporate remotely sensed information  
7  
8 141 (including both satellites and autonomous platforms to provide information on the surface and sub-  
9  
10 142 surface physical properties) are reducing the need to rely on physical surrogates or expert-derived  
11  
12 143 processes when generating bioregions. Much of the difficulty in producing a bioregion stems from  
13  
14 144 relating (dense) interpolated physical data layers to (sparser) biological data.

15 145

## 17 146 **What data can inform statistical bioregions?**

19 147 The generation of bioregions requires data. The three main types of data for bioregionalisation are  
20  
21 148 biological data, physical data and expert-derived knowledge (which itself is usually a mental model  
22  
23 149 based on the first two data types). The volume and variety of biological and physical datasets are  
24  
25 150 increasing, and in many areas, we have reached a key point in time where bioregionalisation can  
26  
27 151 now evolve towards data-driven analyses and products based on observed data and expressed with  
28  
29 152 accompanying measures of uncertainty.

30 153

31 154 Physical data have many desirable features which make them good datasets for informing  
32  
33 155 bioregionalisation; notably good spatial coverage. ‘Physical data’ is a term we use here to represent  
34  
35 156 many types of environmental, abiotic, geomorphic or spatial data used to inform the classification  
36  
37 157 of biological data and represent the physical world. The main sources of synoptic physical data  
38  
39 158 include remotely sensed data, model outputs and interpolations of *in situ* physical data. At broad  
40  
41 159 spatial scales, most remotely sensed data for marine habitats comes from satellites, and *in situ*  
42  
43 160 physical data which has been collected at discrete locations in time, which is modelled and then  
44  
45 161 predicted across space based on these observations. These datasets can come with inherent biases  
46  
47 162 which are often overlooked in broad-scale modelling (Foster et al. 2012). Despite this, physical data  
48  
49 163 can be used to inform the distribution of biological data. Like approaches such as species  
50  
51 164 distribution models, we can generate bioregions based on model relationships between the physical  
52  
53 165 and biological data (Foster et al. 2013).

52 166

53 167 Biological data comes in many forms, such as genes, traits, populations, species and higher  
54  
55 168 taxonomic units. For our purposes, we will focus on biological data that can be readily incorporated  
56  
57 169 in statistical models to build bioregions or components of bioregions. This largely constrains us to  
58  
59 170 the use of observational data about species (and/or Operational Taxonomic Units) in time and  
60  
171 171 space. These observations can be grouped into two broad categories; data from scientific surveys



1  
2 172 and ad-hoc datasets (Graham et al. 2004). Scientific survey data tends to be more systematic and are  
3  
4 173 usually more suitable for scientific endeavours. They include information on the amount of  
5  
6 174 biological material (presence-absence, abundance or biomass) at relatively fine taxonomic  
7  
8 175 resolutions and can include additional biological data like genetic information and/or trait  
9  
10 176 information. The short-coming of survey data in marine environments is that they usually focus on  
11  
12 177 relatively small geographical regions, however there are exceptions to this rule (Edgar & Stuart-  
13  
14 178 Smith 2014). Ad-hoc datasets come from a variety of sources, including museum records and  
15  
16 179 citizen science programs. Generally, they are collected without a rigorous scientific survey design  
17  
18 180 (Warton & Shepherd 2010). Often the location where species were observed is recorded, but  
19  
20 181 corresponding information on absences, survey effort and observation methods are generally  
21  
22 182 lacking. These data are widely referred to as 'presence-only' data and are included in biodiversity  
23  
24 183 databases such as Ocean Biogeographic Information System (OBIS; Grassele 2000; ). Presence-only  
25  
26 184 data obtained from biogeographic databases are widely used for modelling broad scale biodiversity  
27  
28 185 patterns. This is because they have the greatest spatial coverage at regional and global scales,  
29  
30 186 however the lack of an appropriate sampling design, and the frequent lack of recorded absences,  
31  
32 187 means that they should be treated with care in statistical biogeographic models, or indeed any  
33  
34 188 inference (Beck et al. 2014).

35  
36 189 Expert opinion has had a prominent role in the development of bioregions (Ekman 1953). This is  
37  
38 190 because a major limiting factor to developing many broad-scale bioregionalisations has been the  
39  
40 191 lack of biological (and to a lesser extent, physical) data. Therefore, past bioregionalisation efforts  
41  
42 192 have heavily relied on expert elicitation from taxonomists, marine ecologists, biogeographers and  
43  
44 193 stake-holders to delineate important biogeographic regions based on the current status of literature  
45  
46 194 and local knowledge. Expert opinion is still likely to play an important role in bioregional analyses,  
47  
48 195 as it contains implicit information on a region and how species might be distributed within it, as  
49  
50 196 well as an understanding of the biases associated with different data types or surveys. One major  
51  
52 197 issue with expert-based bioregionalisations is reproducibility and assessing the uncertainty in  
53  
54 198 predictions, as expert knowledge is a synthesis of mental, rather than statistical, models. However,  
55  
56 199 there are promising methods that can explicitly include expert knowledge as prior information into  
57  
58 200 statistical models, which we discuss below.

## 201 202 **Developing statistical bioregions**

203 A bioregion can be defined as a geographic region with some relatively constant biological  
204 characteristics, while the biology across different bioregions are relatively different (Brunckhorst &  
205 Bridgewater 1995). This definition is intuitively appealing in its logic, but it is not specific enough



1  
2 206 to guide formal data analysis. To formalise it, we need to define characteristics of a bioregion and  
3  
4 207 specify how they should reflect in biological data. A formal definition of bioregions enables their  
5  
6 208 description in the context of their spatial domain and their relationships to physical data, which can  
7  
8 209 be used as explanatory variables to inform a model. Under all appropriate statistical approaches, we  
9  
10 210 suggest a useful characteristic of a bioregion is an area in which the community composition (the  
11  
12 211 set of species attributes, such as their abundances) is approximately constant. Different bioregions  
13  
14 212 are characterised by different community composition and their respective relationships to the  
15  
16 213 physical data (Fig. 1). A similar formal definition was introduced by Ter Braak et al. (2003) and  
17  
18 214 Foster et al. (2013) using presence and absence data, and expanded to count and biomass data of  
19  
20 215 each species to include a constant abundance within each bioregion (Foster et al. 2017). However,  
21  
22 216 such a definition requires careful implementation when the data arise from samples that have  
23  
24 217 different areal or temporal units of measurement. In such cases, the scale of the data is different and  
25  
26 218 must be adjusted for during the analysis – generally using an offset in the model (e.g. Foster et al.  
27  
28 219 2017). However, using quantities such as probability of occurrence needs to be interpreted with  
29  
30 220 information on how the data were collected to effectively describe the probability with reference to  
31  
32 221 the sampling unit (Warton & Shepherd 2010). Like most classification approaches we assume that  
33  
34 222 once bioregions are defined, the species composition remains constant per bioregion.

35  
36 223  
37  
38 224 Currently, there are many statistical approaches available to classify biological data in to  
39  
40 225 bioregions. However, the choice of which approach to take will often be dictated by the type of data  
41  
42 226 available and the inferences the researchers wish to make. We suggest a useful delineation of  
43  
44 227 possible approaches into the following four categories (like those suggested by Ferrier & Guisan  
45  
46 228 2006)

- 47 229 1. **Predict First, then Group:** A two-step procedure that involves predicting the value of each  
48 230 species at a grid of locations and then clustering those predictions. The environmental  
49 231 conditions are incorporated in the first step through species distribution models (Guisan &  
50 232 Zimmermann 2000), which output species prediction maps. The set of individual species  
51 233 maps are used as inputs into a spatial clustering analysis in a second step. There are multiple  
52 234 model choices available for each step of this analysis. For the prediction step, any kind of  
53 235 species distribution modelling procedure appropriate for the input data could be used  
54 236 (Guisan & Zimmermann 2000). For the clustering step, the analytical method should ideally  
55 237 have methods that will help inform the number of bioregions/clusters (e.g. k-means  
56 238 clustering or model-based clustering; Fraley & Raftery 2002).

- 239 2. **Jointly Predict, then Group:** This is an extension of the previous method, where recent  
240 developments in joint species distribution modelling (Thorson et al. 2016; Ovaskainen et al.  
241 2017; Vanhatalo et al. 2018) enable the joint estimation of multiple species and their  
242 interspecific correlations (Thorson et al. 2016; Ovaskainen et al. 2017; Vanhatalo et al.  
243 2018). Predictions from the multispecies JSDM are passed to an appropriate clustering  
244 method to group species into regions. This remains a two-step procedure for delineating  
245 bioregions and does not explicitly aim for spatially contiguous regions (Ovaskainen et al.  
246 2017).
- 247 3. **Group First, and then Predict:** Another two-step approach involves first clustering  
248 biological data alone, and then predicting the clusters into unsampled locations using a  
249 variant of a ‘species’ distribution model (Miller & Franklin 2002; Ohmann & Gregory 2002;  
250 Vogiatzakis & Griffiths 2006). These are similar steps to the previous methods but are  
251 performed in the reverse order. Like before, there are multiple choices for appropriate  
252 methods to be used in each step.
- 253 4. **Analyse Simultaneously:** Perform both clustering and spatial predictions within a single  
254 model that defines the assumptions/requirements of a bioregion, propagates uncertainty  
255 throughout the process and appropriately handles the multivariate spatial data (Ter Braak et  
256 al. 2003; Foster et al. 2013, 2017; Valle et al. 2014).

257 Each method has its positive and negative attributes, but some are inappropriate for certain  
258 situations. The choice among them will depend on the kind of results required and the kind of data  
259 available. As we describe above, there are two main sources of biological data, those that come  
260 from scientific surveys and those collected in an ad-hoc manner. Currently, many of these methods  
261 have been described and build on scientific survey datasets, where the collection of biological data  
262 is relatively consistent between observations. This means that for many of these approaches  
263 systematic sampling is required to generate robust bioregional outputs. Currently, the ‘Predict First,  
264 then Group’ approach is one of the few approaches which can assemble bioregions based on ad-hoc  
265 data. This approach allows for the development and prediction of species-specific presence-only  
266 models (Warton & Shepherd 2010). These single species predictions can then be classified into  
267 bioregions based on species which have similar ad-hoc collection sightings. At a broad geographical  
268 extent, the use of the ‘Predict First, then Group’ approach is useful with reference to presence-only  
269 datasets and methods account for issues associated with variability in occurrence records (Warton &  
270 Shepherd 2010). Broad scale bioregionalisations have been achieved using multiple presence-only  
271 species distribution models, which are then clustered to provide insight into major biogeographical  
272 configuration (O’Hara et al. 2011; El-Gabbas & Dormann 2018). These approaches still suffer from

1  
2 273 a range of issues, especially related to observational biases in occurrence record data. Correcting for  
3  
4 274 observational and taxonomic biases in broad scale occurrence data is an active area of statistical  
5  
6 275 research (Renner et al. 2015).  
7

8 276  
9  
10 277 Some commonly used ‘Predict First, then Group’ approaches are based on biological distances (e.g.  
11  
12 278 Bray-Curtis dissimilarity). These are fed into regression type models to predict biological  
13  
14 279 dissimilarities in unobserved regions, based on site-pair differences in the physical data (Ferrier &  
15  
16 280 Guisan 2006), and subsequently classified into similar ecological regions or clusters. This approach  
17  
18 281 fails to capture several key statistical principles making it inappropriate as a statistical method for  
19  
20 282 bioregional classification: Firstly, they do not model the observed data, but rather an algorithmic  
21  
22 283 abstraction of it (dissimilarities), which means that concepts like mean-variance relationships are  
23  
24 284 often violated (Warton et al. 2012). Secondly, the model likelihoods are often inappropriately  
25  
26 285 specified as it is based on models for single observations, not pairs of observations; so derived  
27  
28 286 metrics from the likelihood such as information criteria and deviance are unreliable (Warton et al.  
29  
30 287 2015). Thirdly, they typically ignore uncertainty or are unable to compute it directly (Woolley et al.  
31  
32 288 2016).  
33

34 290 Recent development of joint species distribution models has seen their application in specific  
35  
36 291 ecological and biogeographic contexts. Joint species distribution models (JSDMs) are a powerful  
37  
38 292 extension to the ‘Predict First, then Group’ approach, because they jointly estimate the covariance  
39  
40 293 between species, which improves prediction and provides insight into how species are related and  
41  
42 294 structured (Hui et al. 2013; Thorson et al. 2016). They require subsequent clustering on the species  
43  
44 295 predictions, which if done with appropriate clustering methods, should produce reliable results.  
45  
46 296 While these powerful approaches are at the cutting edge of ecological statistics, they currently fail  
47  
48 297 to propagate the uncertainty from the species level predictions through to the bioregional  
49  
50 298 classification step (Ovaskainen et al. 2017). As a result, their predicted bioregional classifications  
51  
52 299 lacks an estimate of uncertainty in the bioregional predictions, however this information can be  
53  
54 300 obtained at the species level (Warton et al. 2015).  
55

56 301  
57 302 The ‘Group First, then Predict’ method suffers from many of the criticisms, and benefits from  
58  
59 303 similar strengths, to those of the ‘Predict First, then Group’ method. A positive, compared to  
60  
304 ‘Predict First, then Group’, is that the number of prediction models is greatly reduced. This enables  
305 the analyst to really focus on fitting good models and diagnosing them well (Miller & Franklin

1  
2 306 2002; Vogiatzakis & Griffiths 2006). Unlike the ‘Predict First, then Group’ method, grouping first  
3  
4 307 currently restricts completely the use of ad-hoc data as methods to cluster only presences are  
5  
6 308 undeveloped. This severely limits the breadth of applications it is available for. Lastly, both the  
7  
8 309 ‘Group First, then Predict’ and the ‘Predict First, then Group’ methods typically fail to propagate  
9  
10 310 uncertainty from the data through to the final bioregional classification.

11 311  
12  
13 312 Examples of the ‘Analyse Simultaneously’ bioregional methods have recently emerged (Dunstan et  
14  
15 313 al. 2011; Foster et al. 2013, 2017). These approaches build upon the concepts of modelling physical  
16  
17 314 and biological data together, but do the prediction and clustering within a single model. These  
18  
19 315 approaches model observed data directly and transfer the variance of the data all the way through to  
20  
21 316 final bioregional prediction (Woolley et al. 2013; Hill et al. 2017).

22  
23 317  
24  
25 318 We argue for the purposes of bioregionalisation using a model which is designed specifically for  
26  
27 319 estimating bioregions should be used. The ‘Analyse Simultaneously’ approaches can account for  
28  
29 320 inter-dependencies between biological and physical data when estimating bioregional classifications  
30  
31 321 (Foster et al. 2013, 2017). Researchers might achieve what they consider ecologically informative  
32  
33 322 regionalisation using any of these four approaches, but must be aware of the information lost at  
34  
35 323 each modelling step in a bioregional analysis (Hill et al. In Prep).

36 324  
37

## 38 325 **Case study**

39  
40 326 To illustrate how one might implement a statistical bioregionalisation, we present a  
41  
42 327 bioregionalisation of fish on the North-West shelf of Australia. The analysis was performed using  
43  
44 328 an extension of the Regions of Common Profiles (RCP) model (Foster et al. 2013) that allows for  
45  
46 329 spatial coherency (Vanhatalo et al. In Review) and is an example of an ‘Analyse Simultaneously’  
47  
48 330 method. There are several important decisions which need to be considered when developing these  
49  
50 331 approaches. Firstly, from a biological perspective we need to consider the number of species to  
51  
52 332 include in the model and what are the minimum number of observations a species requires to be  
53  
54 333 included. As a rule of thumb, multiple species models such as JSMD and mixture models can  
55  
56 334 handle rarer species compared to single species models (Hui et al. 2013; Norberg et al. 2019). In  
57  
58 335 this case study, the entire dataset consisted of 854 demersal trawls taken in depths of 20 to 450 m  
59  
60 336 from October 1986 to August 1997. Each trawl sampled approximately the same amount of seabed,  
337  
338 337 so no adjustment is necessary for varying sample effort. We based the bioregionalisation on 253  
338  
339 338 teleost and chondrichthyan species, from a total of 579 species. We chose this subset as of species as

1  
2 339 they were observed in a least 15 or more trawls. As a rule of thumb, multiple species models such  
3  
4 340 as JSDM and mixture models can handle rarer species compared to single species models (Hui et al.  
5  
6 341 2013; Norberg et al. 2019). Species observed in fewer trawls could have been included in the  
7  
8 342 analyses if the distribution of rare (and potentially threatened) species was a management priority.  
9  
10 343 However, the inclusion of these species would likely add additional noise making it harder to  
11  
12 344 extract relevant information.

13 345  
14  
15 346 As per the biological data, the choice of physical data used as covariates in the modelling will have  
16  
17 347 important ramifications for the bioregionalisation produced. Ideally, the covariates used in the  
18  
19 348 model should best describe the environmental and abiotic factors which characterise each species  
20  
21 349 distribution. For our case study, we chose intra-annual standard deviation (SD) of nitrate, intra-  
22  
23 350 annual SD of dissolved oxygen, annual mean of salinity, intra-annual SD of silicate and intra-annual  
24  
25 351 SD of sea surface temperature as physical data to define bioregions (see Foster et al., 2013 for  
26  
27 352 details). Intra-annual variation can be important to ecological systems as it measures the range of  
28  
29 353 environmental conditions that a single location may encounter. In this example, we did not include  
30  
31 354 information on geomorphic data like soft and hard substrate. These types of variables are likely to  
32  
33 355 be important for describing marine species distributions and will help inform species distributions  
34  
35 356 and assemblages which respond strongly to physical features, rather than environmental gradients  
36  
37 357 (like azonal ecosystems in terrestrial environments; Olson et al. 2001). It is quite plausible that  
38  
39 358 different bioregions can exist in the same covariate space. In these instances, this would likely be an  
40  
41 359 effect of missing covariates, which could be added to an analysis (if available) to help differentiate  
42  
43 360 bioregions. Different physical data will tend to operate on different spatial and temporal scales  
44  
45 361 which could have important implications for bioregionalisation and the variation of assemblages  
46  
47 362 (Austin 2002).

48 363  
49  
50 364 The number of groups chosen during the bioregionalisation process can drastically change the  
51  
52 365 bioregional outcomes (Miller 1996). In the RCP approach we estimated the number of groups from  
53  
54 366 the data based on the model likelihood. Using a single step, sites are grouped based on the species  
55  
56 367 composition and their relationship to the physical data. The model likelihood was then used to  
57  
58 368 inform the number of groups based on Bayesian Information Criterion (BIC; Burnham & Anderson  
59  
60 369 2004). Our model identified four bioregions. Choosing the number of groups which best represents  
370  
371 the available data appears to be one of the key advantages of the 'Analyse Simultaneously' method.  
Other approaches can generate similar groupings, but they must be done in a two-step process,



1  
2 372 which potentially divorces the link between the biological data and number of groups (Hill et al. In  
3  
4 373 Prep).

5  
6 374  
7  
8 375 In this analysis we kept outputs simple for illustration purposes, but note that this analysis can  
9  
10 376 provide more complex outputs. We present a discrete (or hard) classification by assigning site labels  
11  
12 377 based on the most probable bioregion at that site, even though the probability of each site belonging  
13  
14 378 to each bioregion (RCP group) is estimated. The discrete clustered bioregions are given in Fig. 2b,  
15  
16 379 which suggests that there is a coastal region, an inner continental shelf bioregion, a patchy mid-  
17  
18 380 shelf bioregion, and an outer shelf and slope bioregion. Greater information can be gained by  
19  
20 381 examining the probabilities of each bioregion being present across the same study region (Figs. 2c-  
21  
22 382 f). There appears to be quite high probability for Bioregion 1 throughout the entire shallow and  
23  
24 383 medium-water environment and this overlaps substantially with Bioregions 3 and 4. Conversely, the  
25  
26 384 deep-water bioregion (Bioregion 2) appears to have a sharp boundary where the continental margin  
27  
28 385 descends more steeply. Uncertainty maps are available for the probabilistic prediction, as illustrated  
29  
30 386 for Bioregion 3 (See Fig. 3a-c). There are many spatial locations where the predicted presence of  
31  
32 387 this bioregion has low certainty. This is evidenced by the interval estimates (95% confidence  
33  
34 388 interval), Fig. 3a and Fig. 3c, covering the probabilities between (almost) zero and (almost) one.  
35  
36 389 However, there are locations where the probability is certain. These include locations where the  
37  
38 390 bioregion is not (e.g. in the deeper and shallowest water) and locations where there is high  
39  
40 391 confidence in the prediction.

41 392  
42 393 To understand the predicted biological content of each bioregion, we can inspect its species profile.  
43  
44 394 An example, again for Bioregion 3, is given in Fig. 3d: this Bioregion is represented by a small  
45  
46 395 number of species that are very likely present (probability of observation  $> 0.5$ ); a moderate number  
47  
48 396 of species that are moderately likely to be found; and many species that are unlikely to be present.  
49  
50 397 Summing these probabilities gives an indication of species richness in the bioregion. In this case,  
51  
52 398 we would expect to encounter approximately 37 species each time a trawl is performed at a site  
53  
54 399 estimated to have high probability (probability of observation  $> 0.5$ ) of belonging to this bioregion.  
55  
56 400 The species profiles also enable contrasts between bioregions based on their biological content, the  
57  
58 401 profile is the prevalence of every species in each bioregion (we have depicted the profile as a line in  
59  
60 402 Fig. 3, but the identities and their prevalence can be compared across regions; e.g. Hill et al. 2017):  
61  
62 403 if two bioregions share a similar species profile, then they are less different than two bioregions that  
63  
64 404 do not.

1  
2 405  
3  
4 406 Statistical bioregionalisations offer a robust means for identifying, framing and predicting the  
5  
6 407 distribution of biodiversity patterns. The example above shows how quantifying the distributions of  
7  
8 408 multiple species can be distilled into bioregional predictions. These predictions and their associated  
9  
10 409 uncertainties can be assessed against management actions or industrial activities within a bioregion  
11 410 (Fig. 1).  
12

## 13 411

### 16 412 **How do statistical bioregions help improve management decisions?**

17  
18 413 The choice to undertake a bioregionalisation process is often driven by the desire to understand how  
19  
20 414 multiple species are assembled and how best to manage them (Fig. 1a). It is important that  
21  
22 415 information obtained from a bioregionalisation analysis be directly applicable to the current  
23  
24 416 management or scientific question at hand *and* it should be presented with an appropriate level of  
25  
26 417 detail so that it can be understood by those who use it. To us, key considerations in this context  
27 418 include:

- 28  
29  
30 419 (i) *Identify which species are likely to be found in each bioregion (noting that some species*  
31 420 *may be found in multiple bioregions)*. Understanding the membership of species to each  
32  
33 421 bioregion is a critical step for management because it gives managers the capacity to  
34  
35 422 identify which species will be affected by activities (protective or threatening) in a region.  
36  
37 423 We can see from figure one (Fig. 1b), that many of the species are present across multiple  
38 424 bioregions, but their relative intensity is specific to each region. It is this combination of  
39  
40 425 predicted abundances (or prevalence) for a set of species which represents the community  
41  
42 426 composition present within that bioregion relative to the variation in the physical data (e.g.  
43  
44 427 environmental gradients). For all bioregional approaches, except ‘enviro-regionalisation’,  
45 428 the species composition of groups can be identified. For all the two-step approaches species  
46  
47 429 in groups can be identified by summarising the observed species’ data at classified survey  
48  
49 430 sites, while the ‘analyse simultaneously’ approach we can estimate the species membership  
50  
51 431 from parameters in the models and the associated uncertainty that each species belongs to a  
52 432 bioregion. Reporting estimates of species density (or prevalence) and the uncertainties  
53  
54 433 associated with those estimates, will further help managers avoid or protect critical areas  
55  
56 434 within a bioregion(s) where key species or assemblages need to be managed (Fig. 1b).  
57 435

- 58  
59 436 (ii) *Identify which physical data characterise each bioregion*. All approaches should enable the  
60 437 characterisation of physical data used to describe each bioregion (except for expert derived).



1  
2 438 Like describing species membership, the two-step approach can summarise the observed  
3  
4 439 environmental data at classified survey site. While the one-step approach can report these  
5  
6 440 characteristics via model parameters (Hill et al. 2017).  
7  
8 441

9 442 (iii) *Identifying the number of bioregions is an important part of any bioregionalisation process.*

10 443 Choosing the number of bioregions is often driven by the requirements of managers or is  
11  
12 444 chosen to reflect governance boundaries (Department of the Environment and Heritage,  
13  
14 445 2006). Ideally, the number of bioregions should be informed by the data, the ‘Analyse  
15  
16 446 Simultaneously’ approach is currently the only approach which can estimate the number of  
17  
18 447 groups with reference to the original data. Having said this, all approaches should perform  
19  
20 448 similarly if the number of bioregions is known (Hill et al. In Prep). When the number of  
21  
22 449 bioregions is unknown, additional information like phylogeny might help inform this step  
23  
24 450 (Ebach & Parenti 2015).  
25  
26 451

27 452 (iv) *Bioregional classification should be undertaken using a transparent analytical process, so*

28 453 that it is clear to an interested onlooker what was done, why certain decisions were made  
29  
30 454 and what assumptions (ecological and statistical) these decisions reflect. This is a clear  
31  
32 455 advantage of statistical bioregionalisation over expert derived or delphic approaches. Under  
33  
34 456 all statistical bioregional approaches the steps from data, through to analysis, outputs and  
35  
36 457 interpretation can be clearly reported and reproduced based on the data and methods used.  
37  
38 458

39 459 (v) *Bioregional classification should be updatable with the availability of new information, so*

40 460 that the bioregions can be updated in a coherent and consistent manner when additional data  
41  
42 461 become available. This is clearly an advantage of all statistical bioregionalisation  
43  
44 462 approaches where outputs are derived based on modelled data and clearly reported steps and  
45  
46 463 assumptions in a way that expert-derived products are not.  
47  
48 464

49 465 (vi) *Understanding how uncertainty informs confidence in the location of bioregions, along*

50 466 with the confidence in the description of physical and biological characteristics within each  
51  
52 467 bioregion (Brown 1998; Fiorentino et al. 2018). Assessments of uncertainty and variance are  
53  
54 468 already standard in many management actions, for example fisheries ecosystem-based  
55  
56 469 management (Koen-Alonso et al. 2019) and are likely to become more important in  
57  
58 470 bioregionalisation decision-making, where economic, social and biodiversity values are  
59  
60 471 often traded to meet competing objectives. Uncertainty helps modellers, ecologists,  
472  
managers understand how reliable a bioregional classification might be, what its limitations

1  
2 473 are (e.g. for widely ranging species) and should be based on the propagation of variance  
3  
4 474 from the data through to the estimated model. It is important to recognise that many  
5  
6 475 management processes (e.g. design of a representative marine reserve network, or an  
7  
8 476 environmental offset program), require stability from bioregional analyses, especially over  
9  
10 477 the time period that policies are being developed. Defining a coherent and consistent process  
11  
12 478 to update bioregionalisations is an important aspect of their value to government.

## 13 479 14 480 **Future directions for statistical bioregions**

16 481 Throughout this paper we have advocated the use of data informed statistical bioregionalisations.  
17  
18 482 We acknowledge that a lack of quantitative data has been a limiting factor in many cases, leading to  
19  
20 483 the use of physical data or expert knowledge to characterise bioregions (Reygondeau et al. 2017).  
21  
22 484 The availability of biological occurrence data is escalating rapidly with improved data sharing and  
23  
24 485 collection technologies. However, broad-scale biological datasets frequently contain significant  
25  
26 486 biases and error, such as spatial bias in where occurrences were recorded (near major human  
27  
28 487 populations) and an over representation of rare species (taxonomists are inherently interested in rare  
29  
30 488 species) and gear or observer selectivity (Graham et al. 2004). These present challenges which need  
31  
32 489 to be addressed when developing broad scale bioregional models. In our case study, we used  
33  
34 490 biological data collected in a systematic and consistent way, meaning that abundances (presence  
35  
36 491 and absence) of species was explicitly recorded. If these data had been ad-hoc collections, we  
37  
38 492 currently would not have been able to use RCP model approach to describe bioregions, as the model  
39  
40 493 has not been correctly formulated to handle presence-only occurrence data where absences are not  
41  
42 494 systematically recorded. With a lack of recorded absences, an appropriate model would be a spatial  
43  
44 495 point process (Cressie 1993). Under a situation where we only had presence-only data from ad-hoc  
45  
46 496 surveys, we might choose to use a ‘Predict First, then Group’ approach where we generate many  
47  
48 497 point process species distribution models independently and then undertake a clustering of the  
49  
50 498 species Poisson point process predictions across the seafloor to generate bioregions (e.g. O’Hara et  
51  
52 499 al. 2011). Future work that can extend presence-only species distribution models (Warton &  
53  
54 500 Shepherd 2010) to multiple species might provide a promising solution. However, as is the case  
55  
56 501 with single-species presence-only approaches, the underlying biases will need to be clearly  
57  
58 502 documented to make users aware of their potential effects in the outputs. Another promising field of  
59  
60 503 statistical model development consists of augmented approaches that combine *ad-hoc* and scientific  
61  
62 504 survey data to generate more robust predictions without having to remove potentially biased *ad-hoc*  
63  
64 505 occurrence records (Fithian et al. 2015; Renner et al. 2015). Such augmented models could possibly  
65  
66 506 be expanded to accommodate multiple species to inform statistical bioregionalisation.

1  
2 507  
3  
4 508 Throughout this manuscript we have largely focused on species as the biological data for  
5  
6 509 bioregionalisation. But it is plausible that these methods could be extended to encapsulate other  
7  
8 510 sources and types of biological data. For example, a similar approach to the RCP model we  
9  
10 511 described in the case study, has been used to understand population genetic structure in stocks of  
11  
12 512 commercially valuable fisheries (Grewe et al. 2015). Similar models could be extended to multiple  
13  
14 513 species, to understand where genetic populations for numerous species are differentiating. For  
15  
16 514 example, the development of eDNA sampling protocols appears to be a promising area where  
17  
18 515 bioregionalisation could be undertaken on Operational Taxonomic Units, to describe the  
19  
20 516 biogeography of important groups such as bacteria and phytoplankton (Rees et al. 2014). Genetic  
21  
22 517 data can also be used to understand the evolutionary processes that shape the distributions of extant  
23  
24 518 species (Webb et al. 2002; Ebach & Parenti 2015) and is an important source of historical  
25  
26 519 information we have largely ignored. The development of new multiples species models (JSDMs)  
27  
28 520 which can explicitly include information on species traits and phylogeny is likely to facilitate new  
29  
30 521 bioregionalisations which incorporate the role of historical processes with reference to observed  
31  
32 522 species in a joint model (Ives & Helmus 2011; Ovaskainen et al. 2017).

33  
34 523  
35 524 Until comprehensive and broad-scale biological datasets preclude its utility, expert knowledge can  
36  
37 525 continue to play a powerful role in bioregional analyses. We acknowledge the importance of expert  
38  
39 526 knowledge as an information source in many cases, but suggest that it be included as informative  
40  
41 527 prior information in a Bayesian framework where possible (Gelman et al. 2013). An effective way  
42  
43 528 to do so might be to elicit information from experts with the aid of probability training (Hosack et  
44  
45 529 al. 2017). Under such an approach, the development of priors based on expert opinion inform  
46  
47 530 bioregional outputs in low data situations, but as greater volumes of biological data become  
48  
49 531 available, bioregion predictions will increasingly shift towards data driven outcomes.

50 532

## 51 533 **Conclusion**

52  
53 534 Statistical bioregions can be used to frame existing ecosystem-based approaches and provide novel  
54  
55 535 insight into how biodiversity is structured. The development of statistical bioregions, and the  
56  
57 536 methodological developments which underpin them, can help build upon existing bioregional  
58  
59 537 classifications by reducing the ambiguity in what a bioregion is, which we have formally defined,  
60 538 along with how this definition can be matched to data. These bioregions add consistency and

1  
2 539 reproducibility of classifications over approaches like expert elicitation, whilst making direct  
3  
4 540 assessment of bioregions and the information contained within them derived directly from the data  
5  
6 541 used to generate them. Assessing uncertainty in the quality of the estimated model can improve the  
7  
8 542 decision making based on these bioregionalisations. The implementation of statistical bioregions  
9  
10 543 provides a robust path forward for many scientific problems, but to do so, will require that  
11  
12 544 taxonomists, ecologist, biogeographers, statisticians and stakeholders to work on a common set of  
13  
14 545 problems and integrate their skills into a coherent set of meaningful bioregional products that serve  
15  
16 546 their unique purposes (e.g. scientific inquiry, management, or spatial planning).

## 17 547 **Acknowledgments**

18  
19  
20 548 The ‘Statistical Bioregional Workshop’ that facilitated this work was funded by the Global Ocean  
21  
22 549 Biodiversity Initiative (GOBI). GOBI is supported by the International Climate Initiative (IKI).  
23  
24 550 The German Federal Ministry for the Environment, Nature Conservation and Nuclear Safety  
25  
26 551 (BMU) supports this initiative on the basis of a decision adopted by the German Bundestag.  
27  
28 552 S.N.C.W was supported by GOBI. C.H was supported by the EAF-Nansen scientific program.  
29  
30 553 N.J.B was supported by the Marine Biodiversity Hub through funding from the Australian  
31  
32 554 Government's National Environmental Science Program. O.O acknowledges funding by the  
33  
34 555 Academy of Finland (grants 273253 and 284601), the Research Council of Norway (SFF-III grant  
35  
36 556 223257), and by the Jane and Aatos Erkko Foundation (grant to Research Centre for Ecological  
37  
38 557 Change).

## 39 558 40 559 **References**

- 41  
42 560 Anderson OF, Guinotte JM, Rowden AA, Clark MR, Mormede S, Davies AJ, Bowden DA. 2016. Field  
43 561 validation of habitat suitability models for vulnerable marine ecosystems in the South Pacific Ocean:  
44 562 implications for the use of broad-scale models in fisheries management. *Ocean & Coastal Management*  
45 563 **120**:110–126.
- 46  
47 564 Austin MP. 2002. Spatial prediction of species distribution: an interface between ecological theory and  
48 565 statistical modelling. *Ecological Modelling* **157**:101–118.
- 49  
50 566 Beck J, Böller M, Erhardt A, Schwanghart W. 2014. Spatial bias in the GBIF database and its effect on  
51 567 modeling species' geographic distributions. *Ecological Informatics* **19**:10–15.
- 52  
53  
54 568 Begg GA, Friedland KD, Pearce JB. 1999. Stock identification and its role in stock assessment and fisheries  
55 569 management: an overview. *Fisheries Research* **43**:1–8.
- 56  
57 570 Beier P, Albuquerque FS. 2015. Environmental diversity as a surrogate for species representation.  
58 571 *Conservation Biology* **29**:1401–1410.
- 59  
60 572 Brown DG. 1998. Mapping historical forest types in Baraga County Michigan, USA as fuzzy sets. *Plant*  
573 573 *Ecology* **134**:97–111.

- 1  
2 574 Brunckhorst D, Bridgewater P. 1995. Marine bioregional planning: a strategic framework for identifying  
3 575 marine reserve networks, and planning sustainable use and management. Pages 105–16 Proceedings of the  
4 576 Symposium on Marine Protected Areas and Sustainable Fisheries conducted at the Second International  
5 577 Conference on Science and the Management of Protected Areas.  
6  
7 578 Burnham KP, Anderson RP. 2004. Multimodel Inference: Understanding AIC and BIC in Model Selection.  
8 579 *Sociological Methods & Research* **33**:261–304.  
9  
10 580 CBD. 2010. The strategic plan for biodiversity 2011-2020 and the aichi biodiversity targets. Document  
11 581 UNEP/CBD/COP/DEC/X/2. Secretariat of the Convention on Biological Diversity, Nagoya, Japan.  
12  
13 582 Cressie N. 1993. *Statistics for spatial data*. John Wiley & Sons.  
14  
15 583 Department of the Environment and Heritage, 2006. (n.d.). A guide to The Integrated Marine and Coastal  
16 584 Regionalisation of Australia - version 4.0 June 2006 (IMCRA v4.0). Available from  
17 585 <http://www.environment.gov.au/> (accessed February 6, 2018).  
18  
19 586 Dunstan PK, Foster SD, Darnell R. 2011. Model based grouping of species across environmental gradients.  
20 587 *Ecological Modelling* **222**:955–963.  
21  
22 588 Ebach MC, Parenti LR. 2015. The dichotomy of the modern bioregionalization revival. *Journal of*  
23 589 *Biogeography* **42**:1801–1808.  
24  
25 590 Edgar GJ, Stuart-Smith RD. 2014. Systematic global assessment of reef fish communities by the Reef Life  
26 591 Survey program. *Scientific Data* **1**:140007.  
27  
28 592 Ekman S. 1953. *Zoogeography of the seas*. London: Sidgwick & Jackson **953**:415.  
29  
30 593 El-Gabbas A, Dormann CF. 2018. Improved species-occurrence predictions in data-poor regions: using large-  
31 594 scale data and bias correction with down-weighted Poisson regression and Maxent. *Ecography* **41**:1161–  
32 595 1172.  
33  
34 596 Ferrier S, Guisan A. 2006. Spatial modelling of biodiversity at the community level. *Journal of Applied*  
35 597 *Ecology* **43**:393–404.  
36  
37 598 Fiorentino D, Lecours V, Brey T. 2018. On the art of classification in spatial ecology: fuzziness a way to map  
38 599 uncertainty. *Frontiers in Ecology and Evolution* **6**:231.  
39  
40 600 Fithian W, Elith J, Hastie T, Keith DA. 2015. Bias correction in species distribution models: pooling survey  
41 601 and collection data for multiple species. *Methods in Ecology and Evolution* **6**:424–438.  
42  
43 602 Foster SD, Givens GH, Dornan GJ, Dunstan PK, Darnell R. 2013. Modelling biological regions from multi-  
44 603 species and environmental data. *Environmetrics* **24**:489–499.  
45  
46 604 Foster SD, Hill NA, Lyons M. 2017. Ecological grouping of survey sites when sampling artefacts are present.  
47 605 *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **66**:1031–1047.  
48  
49 606 Foster SD, Shimadzu H, Darnell R. 2012. Uncertainty in spatially predicted covariates: Is it ignorable? *Journal*  
50 607 *of the Royal Statistical Society. Series C: Applied Statistics* **61**:637–652.  
51  
52 608 Fraley C, Raftery AE. 2002. Model-based clustering, discriminant analysis, and density estimation. *J Am Stat*  
53 609 *Assoc* **97**:611–631.  
54  
55 610 Gelman A, Carlin J, Stern H, Dunson D, Ventari A, Rubin D. 2013. *Bayesian Data Analysis*. Chapman &  
56 611 Hall/CRC Boca Raton, FL, USA.

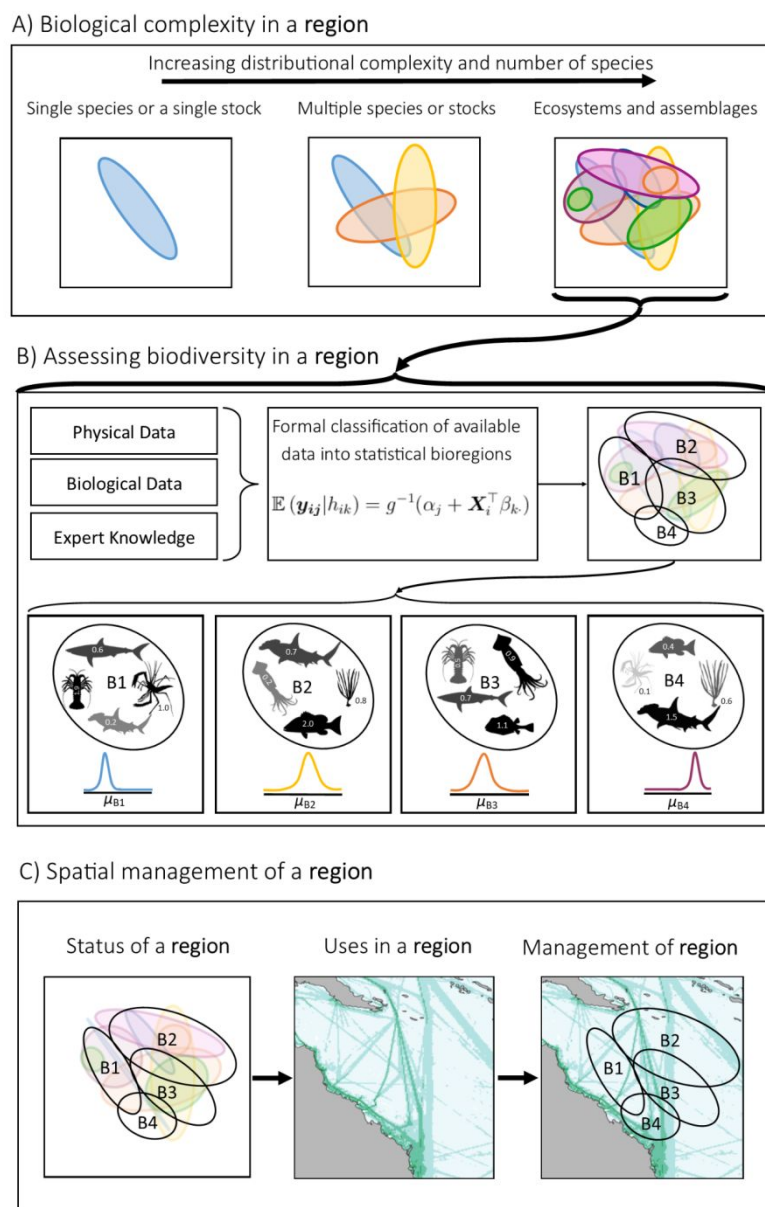


- 1  
2 612 Graham C, Ferrier S, Huettman F, Moritz C, Peterson A. 2004. New developments in museum-based  
3 613 informatics and applications in biodiversity analysis. *Trends in Ecology & Evolution* **19**:497–503.  
4
- 5 614 Grassle JF. 2000. The Ocean Biogeographic Information System (OBIS): an on-line, worldwide atlas for  
6 615 accessing, modeling and mapping marine biological data in a multidimensional geographic context.  
7 616 *Oceanography* **13**:5–7.  
8
- 9 617 Grewe P, Feutry P, Hill P, Gunasekera R, Schaefer K, Itano D, Fuller D, Foster S, Davies C. 2015. Evidence of  
10 618 discrete yellowfin tuna (*Thunnus albacares*) populations demands rethink of management for this globally  
11 619 important resource. *Scientific reports* **5**:16916.  
12
- 13 620 Guisan A, Zimmermann NE. 2000. Predictive habitat distribution models in ecology. *Ecological Modelling*  
14 621 **135**:147–186.  
15
- 16 622 Hill NA, Foster SD, Woolley S, Dunstan PK, Mckinlay J, Ovaskainen O, Johnson CR. In Prep. Determining  
17 623 marine bioregions: a review by example.  
18
- 19 624 Hill NA, Foster SD, Duhamel G, Welsford D, Koubbi P, Johnson CR. 2017. Model-based mapping of  
20 625 assemblages for ecology and conservation management: A case study of demersal fish on the Kerguelen  
21 626 Plateau. *Diversity and Distributions* **23**:1216–1230.  
22
- 23 627 Hosack GR, Hayes KR, Barry SC. 2017. Prior elicitation for Bayesian generalised linear models with  
24 628 application to risk control option assessment. *Reliability Engineering & System Safety* **167**:351–361.  
25
- 26 629 Hui FK, Warton DI, Foster SD, Dunstan PK. 2013. To mix or not to mix: comparing the predictive  
27 630 performance of mixture models vs. separate species distribution models. *Ecology* **94**:1913–1919.  
28
- 29 631 Hutchings L, Roberts MR, Verheye HM. 2009. Marine environmental monitoring programmes in South  
30 632 Africa: a review. *South African Journal of Marine Science* **105**:94–102.  
31
- 32 633 Ives AR, Helmus MR. 2011. Generalized linear mixed models for phylogenetic analyses of community  
33 634 structure. *Ecological Monographs* **81**:511–525.  
34
- 35 635 Koen-Alonso M, Pepin P, Fogarty MJ, Kenny A, Kenchington E. 2019. The Northwest Atlantic Fisheries  
36 636 Organization Roadmap for the development and implementation of an Ecosystem Approach to Fisheries:  
37 637 structure, state of development, and challenges. *Marine Policy* **100**:342–352.  
38
- 39 638 Last PR, Lyne VD, Williams A, Davies CR, Butler AJ, Yearsley GK. 2010. A hierarchical framework for  
40 639 classifying seabed biodiversity with application to planning and managing Australia's marine biological  
41 640 resources. *Biological Conservation* **143**:1675–1686.  
42
- 43 641 Leaper R, Dunstan PK, Foster SD, Barrett NJ, Edgar GJ. 2012. Comparing large-scale bioregions and fine-  
44 642 scale community-level biodiversity predictions from subtidal rocky reefs across south-eastern Australia.  
45 643 *Journal of applied ecology* **49**:851–860.  
46
- 47 644 Longhurst AR. 2010. *Ecological geography of the sea*. Elsevier.  
48
- 49 645 May RM. 1976. Simple mathematical models with very complicated dynamics. *Nature* **261**:459.  
50
- 51 646 Miller J, Franklin J. 2002. Modeling the distribution of four vegetation alliances using generalized linear  
52 647 models and classification trees with spatial dependence. *Ecological Modelling* **157**:227–247.  
53
- 54 648 Miller KR, others. 1996. *Balancing the scales: guidelines for increasing biodiversity's chances through*  
55 649 *bioregional management*. World Resources Institute.  
56  
57  
58  
59  
60

- 1  
2 650 Norberg A et al. 2019. A comprehensive evaluation of predictive performance of 33 species distribution  
3 651 models at species and community levels. *Ecological Monographs*:e01370.  
4
- 5 652 O'Hara TD, Rowden AA, Bax NJ. 2011. A Southern Hemisphere bathyal fauna is distributed in latitudinal  
6 653 bands. *Current Biology* **21**:226–230.  
7
- 8 654 Ohmann, JL, Gregory, MJ. 2002. Predictive mapping of forest composition and structure with direct  
9 655 gradient analysis and nearest-neighbor imputation in coastal Oregon, USA. *Canadian Journal of Forest*  
10 656 *Research*, **32**:725-741.  
11
- 12  
13 657 Olson DM, Dinerstein E, Wikramanayake ED, Burgess ND, Powell GV, Underwood EC, D'amico JA, Itoua I,  
14 658 Strand HE, Morrison JC. 2001. Terrestrial Ecoregions of the World: A New Map of Life on Earth: A new  
15 659 global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity. *BioScience*  
16 660 **51**:933–938.  
17
- 18  
19 661 Ovaskainen O, Tikhonov G, Norberg A, Guillaume Blanchet F, Duan L, Dunson D, Roslin T, Abrego N. 2017.  
20 662 How to make more out of community data? A conceptual framework and its implementation as models and  
21 663 software. *Ecology Letters* **20**:561–576.  
22
- 23 664 Polechová J, Barton NH. 2005. Speciation through competition: a critical review. *Evolution* **59**:1194–1210.  
24
- 25 665 Rees HC, Maddison BC, Middleditch DJ, Patmore JR, Gough KC. 2014. The detection of aquatic animal  
26 666 species using environmental DNA—a review of eDNA as a survey tool in ecology. *Journal of Applied Ecology*  
27 667 **51**:1450–1459.  
28
- 29  
30 668 Renner IW, Elith J, Baddeley A, Fithian W, Hastie T, Phillips SJ, Popovic G, Warton DI. 2015. Point process  
31 669 models for presence-only analysis. *Methods in Ecology and Evolution* **6**:366–379.  
32
- 33 670 Reygondeau G, Guieu C, Benedetti F, Irisson J-O, Ayata S-D, Gasparini S, Koubbi P. 2017. Biogeochemical  
34 671 regions of the Mediterranean Sea: an objective multidimensional and multivariate environmental  
35 672 approach. *Progress in oceanography* **151**:138–148.  
36
- 37 673 Robinson NM, Nelson WA, Costello MJ, Sutherland JE, Lundquist CJ. 2017. A Systematic Review of Marine-  
38 674 Based Species Distribution Models (SDMs) with Recommendations for Best Practice. *Frontiers in Marine*  
39 675 *Science* **4**. Available from <https://www.frontiersin.org/articles/10.3389/fmars.2017.00421/full> (accessed  
40 676 February 5, 2018).  
41
- 42  
43 677 Rohde K. 2007. Latitudinal gradients in species diversity : the search for the primary cause. *Oikos* **65**:514–  
44 678 527.  
45
- 46 679 Ronce O. 2007. How does it feel to be like a rolling stone? Ten questions about dispersal evolution. *Annu.*  
47 680 *Rev. Ecol. Evol. Syst.* **38**:231–253.  
48
- 49 681 Sayre RG, Wright DJ, Breyer SP, Butler KA, Van Graafeiland K, Costello MJ, Harris PT, Goodin KL, Guinotte  
50 682 JM, Basher Z. 2017a. A three-dimensional mapping of the ocean based on environmental data.  
51 683 *Oceanography* **30**:90–103.  
52
- 53 684 Sheil D. 2016. Disturbance and distributions: avoiding exclusion in a warming world. *Ecology and Society* **21**.  
54
- 55  
56 685 Spalding MD et al. 2007. Marine Ecoregions of the World: A Bioregionalization of Coastal and Shelf Areas.  
57 686 *BioScience* **57**:573.  
58
- 59 687 Ter Braak CJ, Hoijsink H, Akkermans W, Verdonschot PF. 2003. Bayesian model-based cluster analysis for  
60 688 predicting macrofaunal communities. *Ecological Modelling* **160**:235–248.



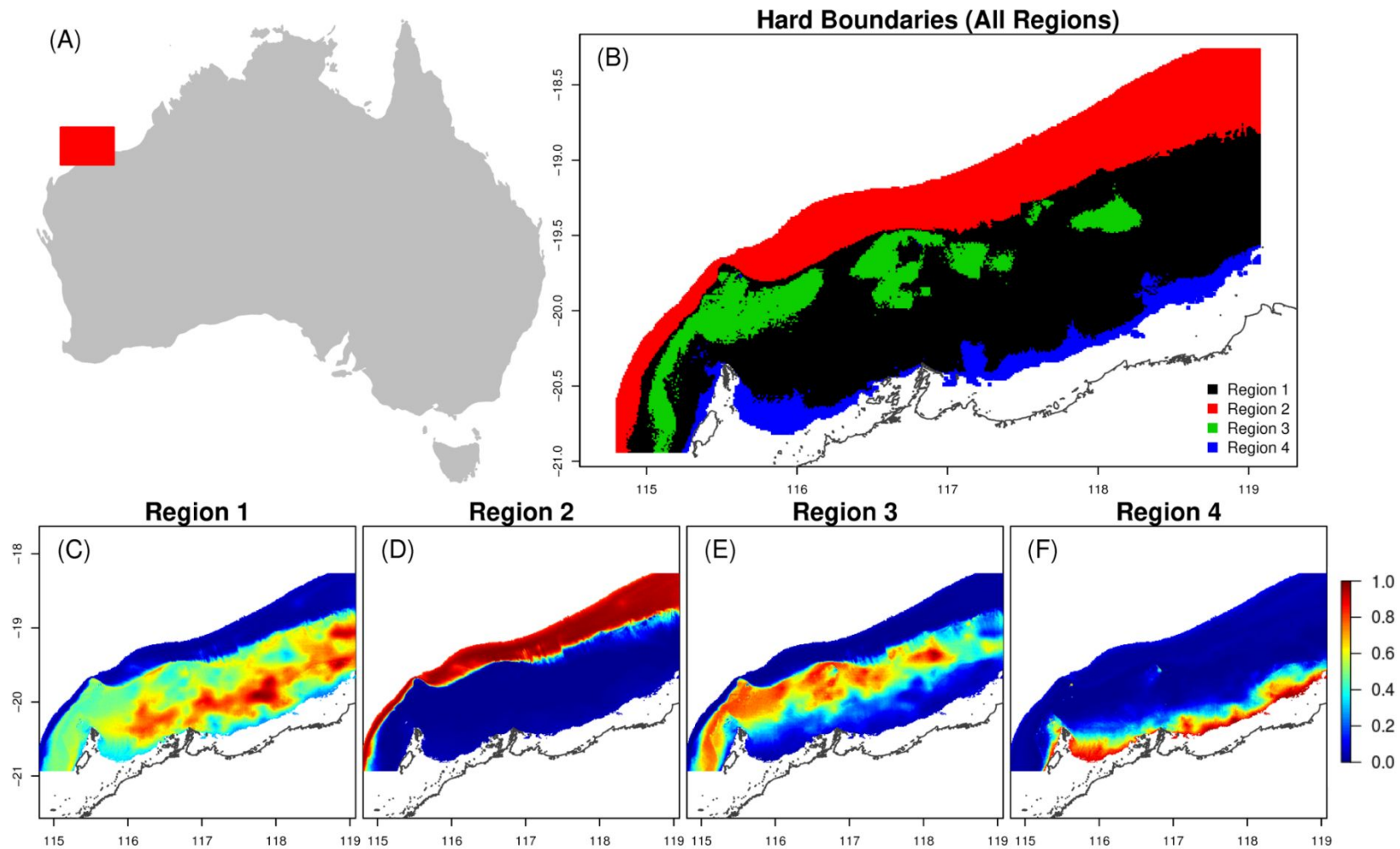
- 1  
2 689 Thorson JT, Ianelli JN, Larsen EA, Ries L, Scheuerell MD, Szuwalski C, Zipkin EF. 2016. Joint dynamic species  
3 690 distribution models: a tool for community ordination and spatio-temporal monitoring. *Global Ecology and*  
4 691 *Biogeography* **25**:1144–1158.
- 5  
6 692 UNESCO. 2009. Global Open Oceans and Deep Seabed (GOODS) - biogeographic classification. *IOC Technical*  
7 693 *Series* **84**:84.
- 8  
9 694 Valle D, Baiser B, Woodall CW, Chazdon R. 2014. Decomposing biodiversity data using the Latent Dirichlet  
10 695 Allocation model, a probabilistic multivariate statistical method. *Ecology letters* **17**:1591–1601.
- 11  
12  
13 696 Vanhatalo J, Foster SD, Hosack GR. In Review. Spatiotemporal Clustering Using Gaussian Processes in a  
14 697 Mixture Model.
- 15  
16 698 Vanhatalo J, Hartmann M, Veneranta L, others. 2018. Additive Multivariate Gaussian Processes for Joint  
17 699 Species Distribution Modeling with Heterogeneous Data. *Bayesian Analysis*.
- 18  
19 700 Vogiatzakis IN, Griffiths GH. 2006. A GIS-based empirical model for vegetation prediction in Lefka Ori, Crete.  
20 701 *Plant ecology* **184**:311–323.
- 21  
22 702 Warton DI, Foster SD, De’ath G, Stoklosa J, Dunstan PK. 2015. Model-based thinking for community  
23 703 ecology. *Plant Ecology* **216**:669–682.
- 24  
25  
26 704 Warton DI, Shepherd LC. 2010. Poisson point process models solve the “pseudo-absence problem” for  
27 705 presence-only data in ecology. *Annals of Applied Statistics* **4**:1383–1402.
- 28  
29 706 Warton DI, Wright ST, Wang Y. 2012. Distance-based multivariate analyses confound location and  
30 707 dispersion effects. *Methods in Ecology and Evolution* **3**:89–101.
- 31  
32 708 Webb CO, Ackerly DD, McPeck M a., Donoghue MJ. 2002. Phylogenies and Community Ecology. *Annual*  
33 709 *Review of Ecology and Systematics* **33**:475–505.
- 34  
35 710 Woolley SNC, Foster SD, Dunstan PK, O’Hara TD, Wintle BA. 2016. Characterising Uncertainty in Generalised  
36 711 Dissimilarity Modelling. *Methods in Ecology and Evolution*.
- 37  
38  
39 712 Woolley SNC, McCallum AW, Wilson R, O’Hara TD, Dunstan PK. 2013. Fathom out: Biogeographical  
40 713 subdivision across the Western Australian continental margin - a multispecies modelling approach. *Diversity*  
41 714 *and Distributions* **19**:1506–1517.
- 42  
43 715
- 44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60



716

42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62

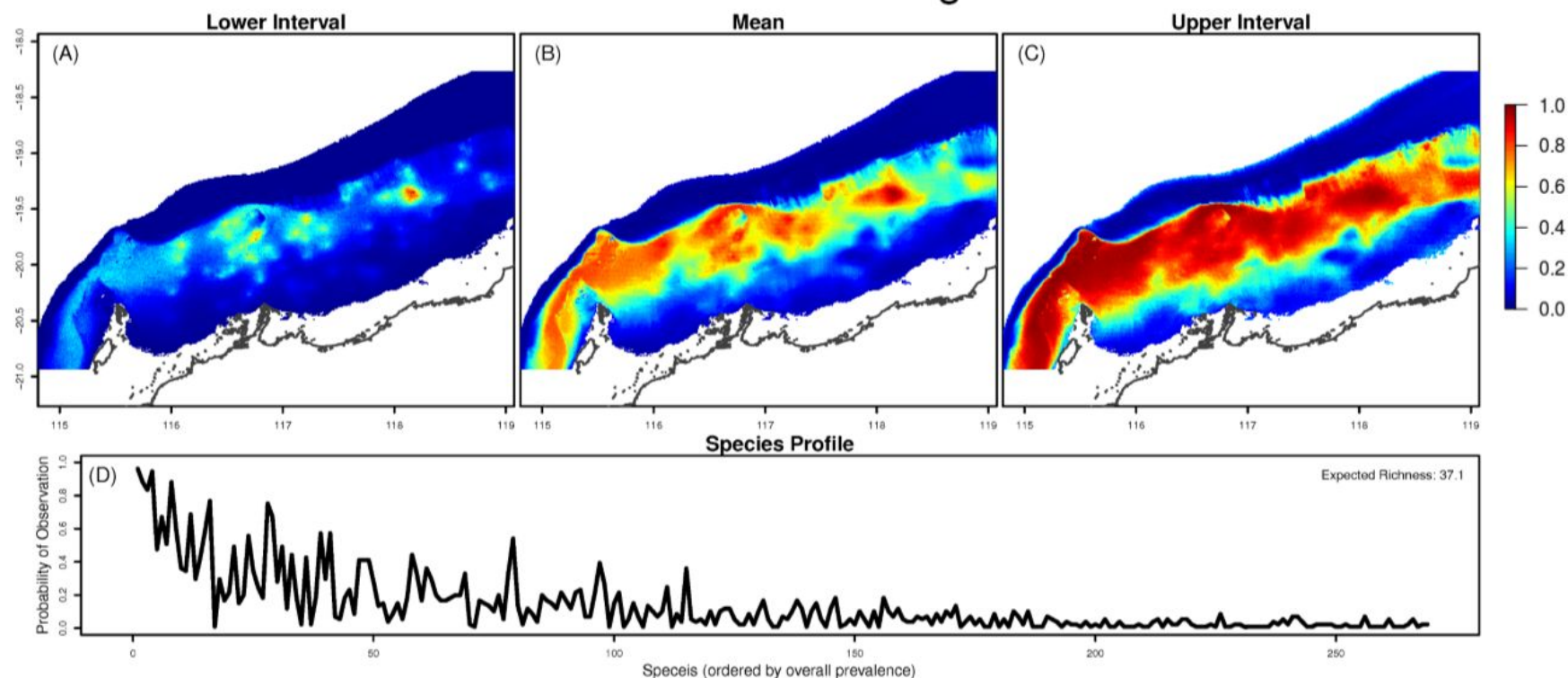
**Figure 1. Why do we need bioregions, how can we describe them from data and how can bioregions be used to manage biodiversity in a region?** A) For the management of a single species (or stock) it is often clear how we might model its distribution. But, as the number of species increases it becomes more challenging to model and interpret hundreds to thousands of species. B) Statistical bioregionalisations offer a solution, as they help contextualise and simplify complex ecosystems or species assemblages into units that are understandable and describe the physical and biological characteristics present in each bioregion. Knowledge on the distribution of biological and physical data can be formally incorporated into statistical models and then be used to distil bioregional level predictions or species assemblages. The colour and the numeric value of the species depicted in each bioregion represents the predicted intensity of each species in that region. We can see that some species occur across multiple regions, but their intensities are specific to each bioregion. Although all species will have a predicted intensity within each bioregion, for plotting purposes we have excluded species from the figure where their predicted intensity is effectively zero. C) Once bioregions have been quantified, the distributions of the bioregions and the species therein can be assessed with reference to uses in that region. Bioregions can then help inform decisions about human activities in a region with reference to their impacts on species assemblages within bioregions.



**Figure 2. Bioregionalisation of the North-West Shelf area of Australia.** (A) Shows where the region is. (B) Shows a set of 4 discrete (hard-clustered) bioregions. (C)-(F) Shows the estimated probability of observing each bioregion in each location. Note that blue colour corresponds to low probability (zero) and red to high (one). Results obtained after applying the regions of common profiles (RCP; Foster et al. 2013), as implemented in Vanhatalo et al (in review).

740

### More Information for Region 3



741 **Figure 3. Further details for Bioregion 3.** (A) Lower interval estimate of probability of each site belonging to Bioregion 3. (B) Mean estimate  
 742 of probability. (C) Upper interval of probability. (D) The profile of species within Bioregion 3 -- species have been ordered according to their  
 743 overall prevalence (across all bioregions), each species' identity is preserved and can be used to understand the composition of each bioregion  
 744 (e.g. Hill et al. 2017). Results obtained by applying the methods in Foster et al., (2013) as implemented in Vanhatalo et al (in review).