



HAL
open science

Extrapolation in species distribution modelling. application to Southern Ocean marine species.

Charlène Guillaumot, Camille Moreau, Bruno Danis, Thomas Saucède

► To cite this version:

Charlène Guillaumot, Camille Moreau, Bruno Danis, Thomas Saucède. Extrapolation in species distribution modelling. application to Southern Ocean marine species.. Progress in Oceanography, 2020, 188, pp.102438. 10.1016/j.pocean.2020.102438 . hal-02985372

HAL Id: hal-02985372

<https://hal.science/hal-02985372v1>

Submitted on 17 Oct 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

1 **Extrapolation in species distribution modelling.** 2 **Application to Southern Ocean marine species**

3
4 **Authors:** Guillaumot Charlène^{1,2}, Moreau Camille^{1,2}, Danis Bruno¹, Saucède Thomas²

5 1 Laboratoire de Biologie Marine, Université Libre de Bruxelles, Avenue F.D.Roosevelt, 50. CP 160/15. 1050
6 BRUXELLES, BELGIUM

7 2 UMR 6282 Biogéosciences, Univ. Bourgogne Franche-Comté, CNRS, EPHE, 6 bd Gabriel F-21000 Dijon,
8 France

9
10 **Keywords:** Multivariate Environmental Similarity Surface (MESS), marine species, Antarctic,
11 modelling relevance, conservation issues

12 13 **ABSTRACT**

14 Species distribution modelling (SDM) has been increasingly applied to Southern Ocean case
15 studies over the past decades, to map the distribution of species and highlight environmental
16 settings driving species distribution. Predictive models have been commonly used for conservation
17 purposes and supporting the delineation of marine protected areas, but model predictions are
18 rarely associated with extrapolation uncertainty maps.

19 In this study, we used the Multivariate Environmental Similarity Surface (MESS) index to quantify
20 model uncertainty associated to extrapolation. Considering the reference dataset of environmental
21 conditions for which species presence-only records are modelled, extrapolation corresponds to the
22 part of the projection area for which one environmental value at least falls outside of the reference
23 dataset.

24 Six abundant and common sea star species of marine benthic communities of the Southern Ocean
25 were used as case studies. Results show that up to 78% of the projection area is extrapolation, i.e.
26 beyond conditions used for model calibration. Restricting the projection space by the known
27 species ecological requirements (e.g. maximal depth, upper temperature tolerance) and increasing
28 the size of presence datasets were proved efficient to reduce the proportion of extrapolation areas.
29 We estimate that multiplying sampling effort by 2 or 3 fold should help reduce the proportion of
30 extrapolation areas down to 10% in the six studied species.

31 Considering the unexpectedly high levels of extrapolation uncertainty measured in SDM
32 predictions, we strongly recommend that studies report information related to the level of
33 extrapolation. Waiting for improved datasets, adapting modelling methods and providing such
34 uncertain information in distribution modelling studies are a necessity to accurately interpret
35 model outputs and their reliability.

36 Introduction

37 Among the broad array of analytical tools developed for marine ecology studies over the last two
38 decades, Species Distribution Modelling (SDM) has been increasingly used ([Peterson 2001](#), [Elith
39 et al. 2006](#), [Austin 2007](#), [Gobeyn et al. 2019](#)) and applied to Southern Ocean pelagic ([Pinkerton et
40 al. 2010](#), [Freer et al. 2019](#)), benthic organisms ([Loots et al. 2007](#), [Pierrat et al. 2012](#), [Basher and
41 Costello 2016](#), [Xavier et al. 2016](#), [Gallego et al. 2017](#), [Guillaumot et al. 2018a, 2018b](#), [Fabri-Ruiz
42 et al. 2019](#), [Jerosch et al. 2019](#)) and even marine mammals ([Nachtsheim et al. 2017](#)). SDM
43 represents a complementary approach to individual-based modelling and eco-physiological
44 experiments, quickly and synthetically identifying environmental correlates of species distribution
45 ([Brotons et al. 2012](#), [Feng and Papes 2017](#), [Feng et al. 2020](#)). SDM is also used to define species
46 distribution spatial range ([Nori et al. 2011](#), [Walsh and Hudiburg 2018](#)) and can be used as decision
47 criteria for conservation purposes ([Guisan et al. 2013](#), [Marshall et al. 2014](#)). For instance, it is
48 currently used in proposals developed by national committees of the CCAMLR (Commission for
49 the Conservation of Antarctic Marine Living Resources) to support the definition and delineation of
50 marine protected areas ([Ballard et al. 2012](#), [CCAMLR report WG-FSA-15/64](#), [Arthur et al. 2018](#)).

51
52 Applying SDM to Southern Ocean case studies is particularly challenging due to major constraints
53 and biases that may reduce modelling performance. As for many oceanographic studies, access to
54 environmental data with high temporal and spatial resolutions is difficult ([Davies et al. 2008](#),
55 [Robinson et al. 2011](#)). Antarctic coastal areas, in particular, are rarely accessed and documented
56 due to logistical constraints, access being for example impossible during the austral winter due to
57 sea ice cover ([De Broyer et al. 2014](#)). The availability of species absence records is also a limiting
58 factor to modelling performances and model calibrations ([Brotons et al. 2004](#), [Wisz and Guisan
59 2009](#)). Models are usually based on a limited number of presence-only records and limited number
60 of sampling sites, which are both spatially aggregated in the vicinity of scientific stations, where
61 access is frequent and datasets from different seasons, have been compiled over decades and
62 even beyond ([De Broyer et al. 2014](#), [Guillaumot et al. 2018a](#), [Fabri-Ruiz et al. 2019](#), [Guillaumot et
63 al. 2019](#)).

64
65 When generating a SDM, the model is fit to data with a given range of value for each
66 environmental descriptor (i.e. the calibration range). When transferring model predictions, a portion
67 of the environment may cover additional conditions that are outside this calibration range: these
68 are non-analog conditions and the model extrapolates ([Randin et al. 2006](#), [Williams and Jackson
69 2007](#), [Williams et al. 2007](#), [Fitzpatrick and Hargrove 2009](#), [Owens et al. 2013](#), [Yates et al. 2018](#)).
70 Considering the limited number of species presence-only records occupied by each marine benthic
71 species, and the poor quality and precision of environmental descriptors available for modelling

72 Southern Ocean species distributions (Guillaumot et al. 2018a, Fabri-Ruiz et al. 2019), a large
73 proportion of cells might be expected to be extrapolations beyond the calibration range of the
74 model.

75
76 The Multivariate Environmental Similarity Surface (MESS) approach analyses spatial extrapolation
77 by extracting environmental values covered by presence-only records and estimates areas where
78 environmental conditions are outside the range of conditions contained in the calibration area (Elith
79 et al. 2010). The method considers that extrapolation occurs when at least one environmental
80 descriptor value is outside the range of the environment envelop for model calibration (more details
81 given in Appendix 4).

82 The MESS approach was initially used to determine the environmental barriers to the invasion of
83 the cane toad in Australia, when facing new environments and under future conditions (Elith et al.
84 2010). Implemented in MaxEnt (Elith et al. 2011), MESS was subsequently used by several
85 authors for defining the climatic limits to the colonisation of new environments by non-native
86 species, such as the American bullfrog in Argentina (Nori et al. 2011), for studying contrasts
87 between native and potential ecological niches like in the study of the spotted knapweed
88 (*Centaurea stoebe*) (Broennimann et al. 2014), or for defining the limits to model transferability and
89 predicting the distribution of trees under future environmental conditions (Walsh and Hudiburg
90 2018).

91 More recently, the MESS approach was used to define model uncertainties related to extrapolation
92 (Escobar et al. 2015, Li et al. 2015, Cardador et al. 2016, Luizza et al. 2016, Iannella et al. 2017,
93 Milanesi et al. 2017, Silva et al. 2019) and extrapolation areas where environmental conditions are
94 non-analog to conditions of model calibration (Fitzpatrick and Hargrove 2009, Anderson 2013).
95 Associating uncertainty information to model predictions has been acknowledged as a necessity
96 for reliable interpretations of model predictions (Grimm and Berger 2016, Yates et al. 2018). It is
97 also a requirement for specifying the level of risk associated with predictions and evaluating
98 whether uncertainty can be mitigated to improve model outcomes (Guisan et al. 2013).

99
100 This study addresses the importance of extrapolation and associated uncertainties in SDMs
101 generated at broad spatial scale for Southern Ocean species: an analysis that is seldom performed
102 although important to characterise model reliability. Using the case study of six abundant and
103 common sea star species in marine benthic communities, objectives of this work are to evaluate
104 the importance of extrapolation proportions in wide projection areas, and to provide some
105 methodological clues to mitigate the effects of extrapolation and improve model accuracy.

106

107

108 **Methods**

109 **Studied species and environmental descriptors**

110 The distribution of six sea star species (Asteroidea : Echinodermata) was studied (Table 1). The
111 six species, *Acodontaster hodgsoni* (Bell, 1908), *Bathybiaster loripes* (Sladen, 1889), *Glabraster*
112 *antarctica* (Smith, 1876), *Labidiaster annulatus* Sladen, 1889, *Odontaster validus* Koehler, 1906
113 and *Psilaster charcoti* (Koehler, 1906) are abundant and common in benthic communities in the
114 Southern Ocean. The biology, ecology and distribution of these species have been extensively
115 studied and are relatively well documented ([McClintock et al. 2008](#), [Mah and Blake 2012](#),
116 [Lawrence 2013](#)). Presence-only records were compiled from a recently updated database,
117 thoroughly scrutinised with the World Register of Marine Species ([WoRMS Editorial Board 2016](#)),
118 to delete potential discrepancies, update taxonomy and correct for georeferencing errors ([Moreau](#)
119 [et al. 2018](#)).

120 Models were generated for the different species using 298 to 851 presence-only records, and
121 projected at different depth ranges (Table 1). The distributions of these presence-only records are
122 contrasting between species (Appendix 1), with *A. hodgsoni*, *B. loripes* and *G. antarctica* having an
123 Antarctic and sub-Antarctic distribution, with an important number of data available for *B. loripes*
124 and *G. antarctica* but less data for *A. hodgsoni* (respectively 591, 851 and 298 presence-only
125 records). *L. annulatus* has a distribution mainly gathered in the sub-Antarctic region with few data
126 available (375 presence-only records). *O. validus* and *P. charcoti* are mainly present on the coasts
127 of the Antarctic shelf.

128

129

130

131

132

133

134

135

136

137

138

139





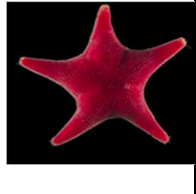

140

141

142

143

144 **Table 1.** Sea star species investigated in the present study. The number of presence-only records
 145 available was summed up after removal of duplicates from each grid cell pixel. Image sources:
 146 [Brueggeman 1998](#), BIOMAR ULB database (P. Pernet), proteker.net, B121 expedition (Q.
 147 Jossart).

	<i>Acodontaster hodgsoni</i> (Bell, 1908)	<i>Bathybiaster loripes</i> (Sladen, 1889)	<i>Glabraster antarctica</i> (Smith, 1876)	<i>Labidiaster annulatus</i> Sladen, 1889	<i>Odontaster validus</i> Koehler, 1906	<i>Psilaster charcoti</i> (Koehler, 1906)
						
Presence-only records number	298	591	851	375	337	353
Model maximum depth	1500 m	4000 m	4000 m	1500 m	1500 m	4000 m

148
 149 Environmental descriptors were selected from the dataset provided at
 150 https://data.aad.gov.au/metadata/records/environmental_layers. These are oceanography raster
 151 layers that mostly describe the physical and geochemical environment south of 45°S with a 0.1°
 152 grid-cell resolution (approximately 11km wide in latitude). Among the 58 environmental descriptors
 153 provided, only those that fulfilled the analysis performed by [Guillaumot et al. \(2020\)](#) were selected:
 154 ‘distance’ layers and ‘extreme’ layers were not selected because the interpretation of their
 155 respective contributions to niche models is complex or weak and collinear descriptors were also
 156 discarded for a Variance Inflation Factor (VIF) > 10 ([Naimi et al. 2014](#)). A set of 14 to 16 species-
 157 specific layers that characterise temperature, salinity, food availability and habitat characteristics
 158 were therefore used for model calibration (Table S2).

159
 160 **Models calibration**

161 Species Distribution Models (SDMs) were generated using the Boosted Regression Trees (BRT), a
 162 machine-learning approach that was already calibrated for Southern Ocean case studies
 163 ([Guillaumot et al. 2018](#), [Guillaumot et al. 2019](#)) and was proved efficient to provide accurate
 164 models with good transferability performance, that is good ability to project model in space and
 165 time ([Elith et al. 2008](#), [Reiss et al. 2011](#), [Heikkinen et al. 2012](#), [Guillaumot et al. 2019](#)). In order to
 166 minimize the effect of presence-only records aggregation on model predictions, background data
 167 were randomly sampled in the environment following the probabilities defined by a Kernel Density
 168 Estimation (KDE) (see [Phillips et al. 2009](#) for general principles, [Guillaumot et al. 2018a](#), [2018b](#)

169 and [Fabri-Ruiz et al. 2019](#) for applications). The number of background records was selected
170 equal to the number of presence-only records ([Barbet-Massin et al. 2012](#)). The KDE was
171 established based on the aggregation of benthos sampling effort provided in the Biogeographic
172 Atlas of the Southern Ocean ([De Broyer et al. 2014](#), map available in supplementary material of
173 [Guillaumot et al. \(2019\)](#)). One hundred SDMs were generated and averaged for each species, with
174 background data randomly sampled following the KDE for each replicate.

175 SDMs were calibrated and reliability tested using a spatial cross-validation procedure. For each
176 species, several procedures were compared following [Guillaumot et al. \(2019\)](#). The studied area
177 was randomly subdivided into 2 to 6 areas of similar surfaces (longitude-split spatial folds), with
178 presence and background data selected from one to three areas for model training and from the
179 remaining areas for model testing. The “6-fold CLOCK” cross-validation approach was selected for
180 *B. loripes*, *G. antarctica*, *L. annulatus* and *O. validus* and the “2-fold CLOCK” procedure was
181 selected for *A. hodgsoni* and *P. charcoti*, according to the best percentage of test data correctly
182 classified (Appendix 3).

183 The Maximum sensitivity plus specificity threshold (MaxSSS), considered the most appropriate
184 threshold for presence-only SDM ([Liu et al. 2013](#)) was used to binarize models into suitable
185 (>MaxSSS value) and unsuitable areas (<MaxSSS value). This threshold was used to measure the
186 proportion of test data correctly classified. Modelling performances were also assessed using the
187 three following metrics: Area Under the Receiver Operating Curve (AUC, [Fielding and Bell 1997](#)),
188 the Point Biserial Correlation between predicted and observed values (COR, [Elith et al. 2006](#)) and
189 the True Skill Statistics (TSS, [Allouche et al. 2006](#)).

190
191 Two analyses were performed: in Analysis #0 ('no-depth limited'), SDMs were projected on the
192 entire Southern Ocean surface (south of 45°S) and in Analysis #1 ('depth limited'), SDM
193 projections and background samplings were restricted to areas limited by a maximum depth
194 threshold defined for each species based on the available species presence-only records (Table
195 1).

196 197 **MESS calculation**

198 The MESS was measured using the *dismo* R package ([Hijmans et al. 2017](#)) and following the
199 guidelines provided in [Elith et al. \(2010\)](#). Pixels for which at least one environmental descriptor has
200 a value that is outside the range of environmental values defined by presence-only records
201 (calibration range) were considered to be extrapolation (i.e when MESS estimate gets negative
202 values, Appendix 4). The proportion of extrapolation areas (i.e. the proportion of cells defined as
203 extrapolations over the total projection area) was calculated and compared between species. On
204 SDM projection maps, extrapolated pixels were displayed in black.

205 Environmental parameters responsible for extrapolation were estimated by modifying the code
206 provided in [Elith et al. \(2010\)](#). Detailed R scripts are available at
207 <https://github.com/charleneguillaumot/THESIS>. Methodological details are provided in Appendix 4.

208
209

210 **Influence of the number and distribution of presence-only records on extrapolation**

211 The proportion of extrapolation areas may vary with presence-only sampling effort. In order to
212 study the influence of the number and distribution of these presence-only records on the proportion
213 of extrapolation areas, two analyses were performed. First, several SDMs were generated with
214 different numbers of presence-only records, following the chronological addition of new presence-
215 only records through time, from 1980 to 2016. Second, SDMs generated with 10% to 100% (10%
216 increments, so 10 subsets) of the entire presence-only dataset were compared. In this analysis, in
217 contrast to the previous one, presence-only records are randomly sampled among the datasets
218 available.

219 In these two analyses, SDMs were projected on the environmental space limited by the maximum
220 depth defined for each species (Table 1), 100 model replicates were generated and averaged in
221 each case and spatial autocorrelation (SAC) was estimated to assess the influence of presence-
222 only records aggregation on modelling performances. The significance of SAC was tested using
223 the Moran I index computed on model residuals ([Luoto et al. 2005](#), [Crase et al. 2012](#)).

224
225 The relationship between the number of presence-only records used in SDM and the relative
226 proportion of extrapolation areas was characterised using linear regressions. This allowed, for
227 each model, estimation of the minimum number of presence-only records required to obtain a
228 'reasonable' proportion of extrapolation area arbitrarily set to a 10% threshold.

229

230 **Results**

231 **Extrapolation and the extent of projection areas**

232 All generated SDMs are accurate and performant, with high AUC ($AUC > 0.91$), TSS ($TSS > 0.559$)
233 and COR ($COR > 0.68$) values, low standard deviations and good percentages of correctly
234 classified presence-only test data (77 to 90 %) (Table 2). Descriptors that contribute the most to
235 SDMs are depth (22 to 34%), minimum POC (6 to 21%), POC standard deviation (8 to 20%), mean
236 ice cover depth (7 to 17%) and mixed layer depth (3 to 10%). Contrasts between species are in the
237 respective percentage of contribution of these descriptors. Descriptors that drive the most species
238 distribution are similar between species (Appendix 5).

239

240 Models projected on the entire Southern Ocean (Analysis #0, 'no-depth limited') extrapolate on an
241 area covering between 15 to 78% of the entire projection area, and 19 to 45% of the area initially
242 predicted as suitable to the species distribution (Table 2, Fig. 1). Extrapolation areas cover more
243 than 50% of the projection area for *A. hodgsoni* (78.6%), *P. charcoti* (67.8%), *L. annulatus* (64.8%)
244 and *O. validus* (51.9%) and more than 30% of suitable areas (Table 2). For these four species,
245 depth is responsible for 25 to 68% of extrapolation (Appendix 5). Geomorphology, mean ice cover
246 and POC standard deviation are layers also contributing to 2 to 7% for extrapolation (Appendix 5).
247 These descriptors that highly contribute to MESS also contribute to the model, and there are no
248 descriptors for which the contribution to MESS is important whereas the contribution to the model
249 is not substantial (Appendix 5).

250
251 In models projected on areas restrained in depth (Analysis #1, 'depth limited'), the percentage of
252 extrapolation area sharply decreases from 59 to 18% according to the species (Table 2). However,
253 model performances also decrease, with AUC values going down to 0.885, TSS values to 0.419
254 and COR values to 0.475. The percentage of correctly classified test data is much lower and more
255 variable for the shallowest species *A. hodgsoni* (from $90 \pm 6.26\%$ to $45.5 \pm 8.1\%$), *L. annulatus*
256 ($77.7 \pm 15.2\%$ to $57.98 \pm 20\%$) and *O. validus* (from $85.4 \pm 9.6\%$ to $57.68 \pm 21\%$). For all species,
257 predicted suitable areas increase two-fold.

258 Overall, descriptor contributions to the model remain unchanged between the two analyses, except
259 for depth contribution that decreases to around 10% on average for all the species. In contrast, in
260 Analysis #1, depth contribution to the MESS is very low (0.64 to 5.8%), except for *P. charcoti*
261 (16.3%). Mean ice cover is the layer that contributes the most to extrapolation, extrapolation areas
262 mainly corresponding to Weddel and Amundsen seas.

263
264
265
266
267
268
269
270
271
272
273
274
275
276

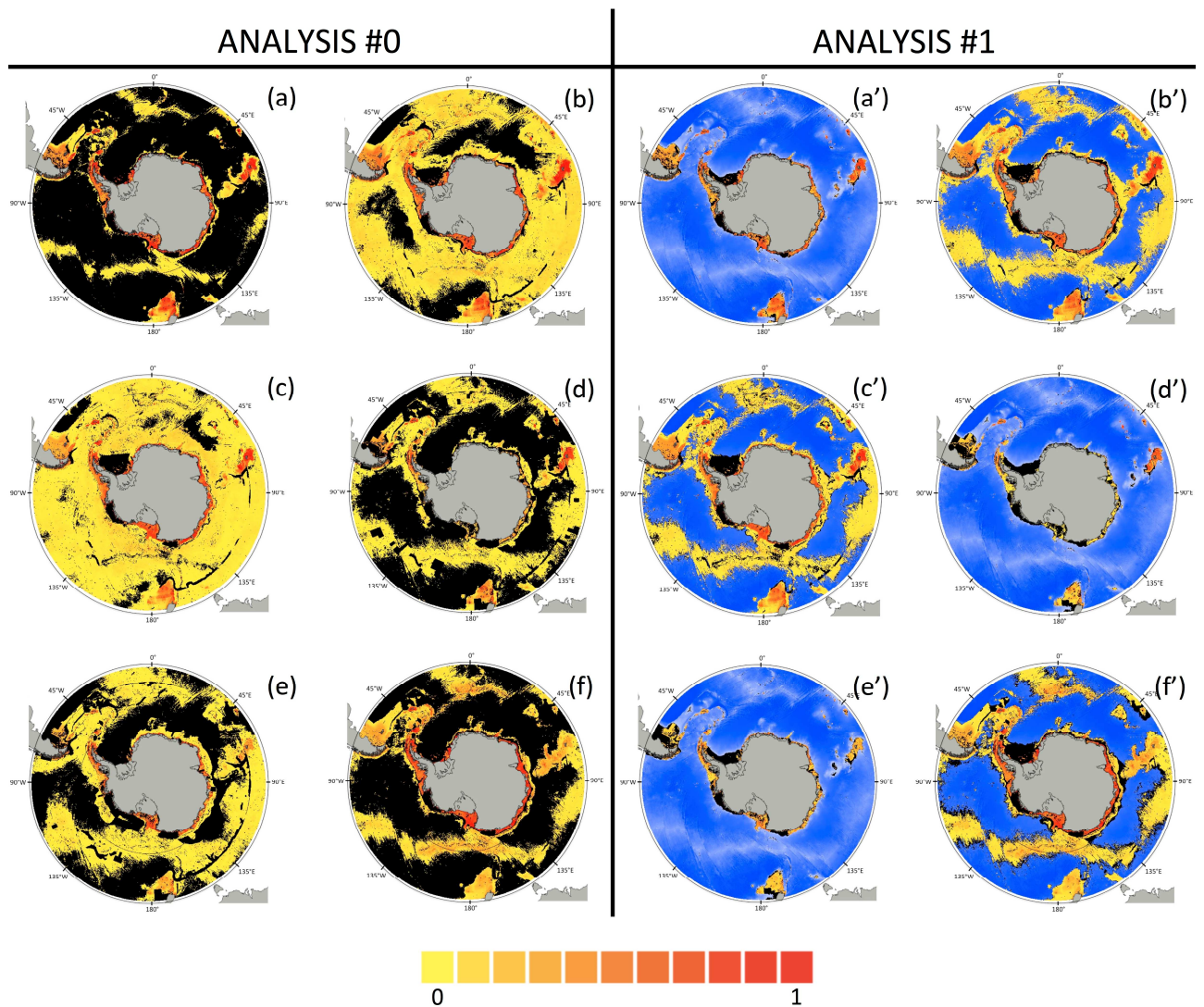
277 **Table 2.** Modelling performances for each species. Average and standard deviation values of the
 278 100 model replicates. Pres. NB: number of presences-only records available for modelling
 279 (duplicates excluded); AUC: Area Under the Curve; TSS: True Skills Statistic; COR: Biserial
 280 Correlation.
 281

Analysis #0, no-depth limited								
Species	Pres. NB	AUC	TSS	COR	Correctly classified test data (%)	Suitable area (% total area)	Extrapolation area (% total area)	Extrapolation area (% total area)
<i>Acodontaster hodgsoni</i>	298	0.925 ± 0.02	0.579 ± 0.04	0.735 ± 0.06	90 ± 6.26	8.86	78.6	35.3 ± 4.1
<i>Bathybiaster loripes</i>	591	0.910 ± 0.02	0.559 ± 0.07	0.68 ± 0.09	80.6 ± 10.9	8.55	29.1	21.9 ± 4.4
<i>Glabraster antarctica</i>	851	0.929 ± 0.01	0.58 ± 0.05	0.719 ± 0.07	85.45 ± 6.34	7.95	15.73	19.9 ± 3.9
<i>Labidiaster annulatus</i>	375	0.95 ± 0.03	0.598 ± 0.07	0.730 ± 0.14	77.7 ± 15.2	3.33	64.83	42.1 ± 10.5
<i>Odontaster validus</i>	337	0.953 ± 0.01	0.605 ± 0.05	0.746 ± 0.09	85.4 ± 9.6	6.89	51.9	45.2 ± 5.65
<i>Psilaster charcoti</i>	353	0.911 ± 0.02	0.58 ± 0.03	0.723 ± 0.04	87.7 ± 4.8	8.90	67.9	32.5 ± 4.71

282

Analysis #1, depth limited								
Species	Pres. NB	AUC	TSS	COR	Correctly classified test data (%)	Suitable area (% total area)	Extrapolation area (% total area)	Extrapolation area (% total area)
<i>Acodontaster hodgsoni</i>	298	0.823 ± 0.05	0.419 ± 0.1	0.475 ± 0.14	45.5 ± 18.1	17.49	40.6	27.5 ± 8.5
<i>Bathybiaster loripes</i>	591	0.887 ± 0.03	0.513 ± 0.08	0.607 ± 0.12	78.4 ± 11	15.75	18.2	20.8 ± 4.8
<i>Glabraster antarctica</i>	851	0.915 ± 0.01	0.537 ± 0.08	0.654 ± 0.1	81.8 ± 7.7	14.08	23.9	18.64 ± 3.5
<i>Labidiaster annulatus</i>	375	0.918 ± 0.03	0.482 ± 0.16	0.563 ± 0.25	57.98 ± 20	8.88	59.5	38.7 ± 14.6
<i>Odontaster validus</i>	337	0.908 ± 0.03	0.504 ± 0.13	0.586 ± 0.17	57.68 ± 21	11.64	51.5	38.3 ± 6.97
<i>Psilaster charcoti</i>	353	0.885 ± 0.02	0.546 ± 0.04	0.665 ± 0.06	83 ± 6.6	15.40	35.78	33.2 ± 5.1

283



284
 285 **Figure 1.** Maps of extrapolation areas covering SDM predictions, generated with all presence-only
 286 records available for the studied species. Left panel: projection area not limited in depth (Analysis
 287 #0), right panel: projection area limited to -1,500 m and -4,000 m depth (Analysis #1), according to
 288 the species (*A. hodgsoni*, *L. annulatus*, *O. validus* until 1,500 m; *B. loripes*, *G. antarctica*, *P.*
 289 *charcoti* until 4,000 m; Table 1). (a) *Acodontaster hodgsoni*, (b) *Bathyiaster loripes*, (c) *Glabiraster*
 290 *antarctica*, (d) *Labidiaster annulatus*, (e) *Odontaster validus*, (f) *Psilaster charcoti*. Extrapolation
 291 areas displayed in black; pixels colored by the yellow-red color palette provide SDM distribution
 292 probabilities (comprised between 0 and 1); bathymetric chart in shades of blue.

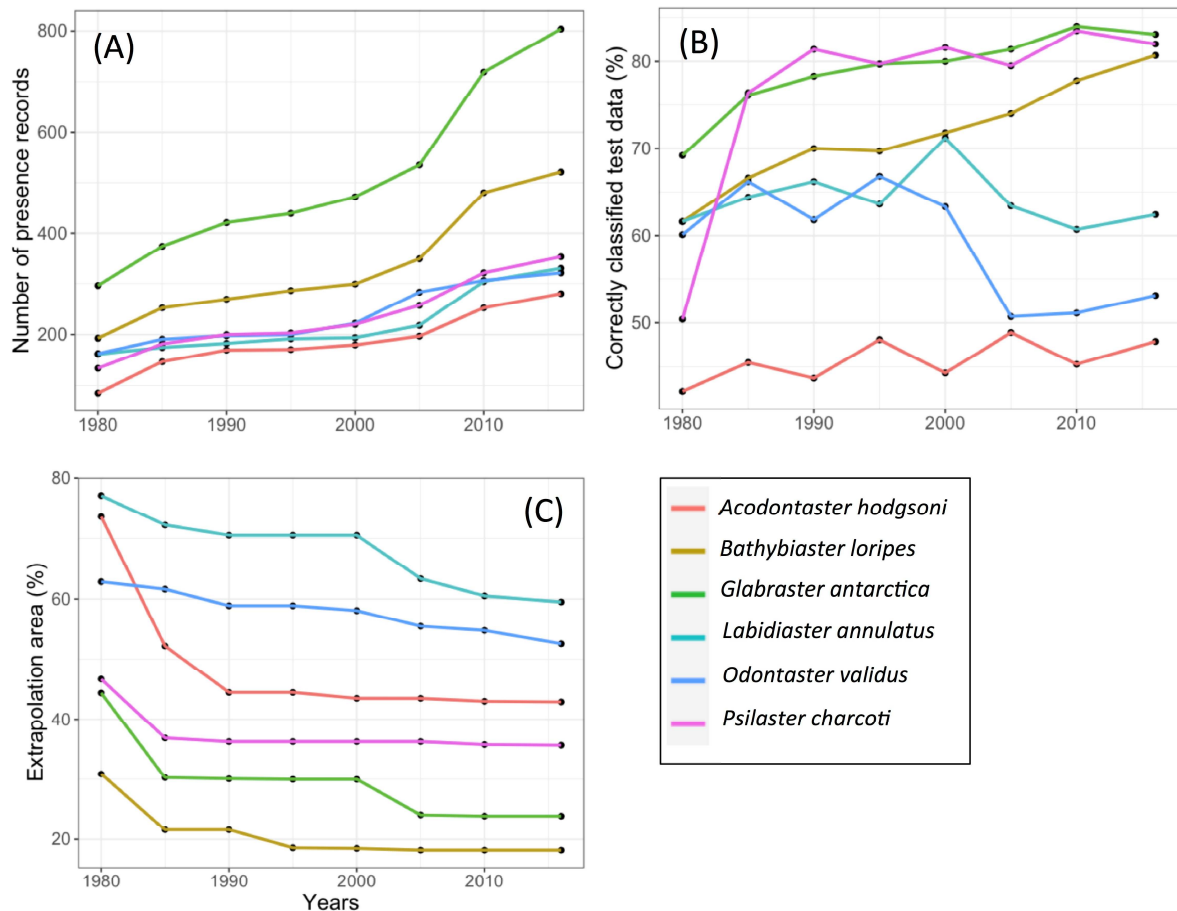
293
 294 **Extrapolation and the number of presence-only records**
 295 Model performance and size of extrapolation area were compared between models run with
 296 different numbers of presence-only records, following the chronological addition of new samples
 297 (from 1980 to 2016). From 1980 to 2016, the number of presence-only records collected during
 298 oceanographic campaigns has increased from 1.9 to 3.3 times according to the species (1.9 times
 299 for *O. validus*, 3.3 times for *A. hodgsoni*)(Fig. 2a). Spatial autocorrelation between presence-only

300 records varies between species, with the highest Moran's I scores obtained for *L. annulatus*, *O.*
 301 *validus* and *A. hodgsoni*. The highest Moran's I values were mainly calculated for the oldest
 302 presence-only subsets (1980), strenghtening the fact that the addition of new presence-only
 303 records with additional campaigns reduces spatial autocorrelation (Table S7).

304
 305 Model performance increases (higher AUC scores) with the addition of new presence-only records,
 306 for all species except for models of *A. hodgsoni* and *B. loripes* for which AUC values are stable
 307 (Table S6). Similarly, the percentage of correctly classified test data presents important standard
 308 deviation values and improves with the addition of new presence-only records, except for *O.*
 309 *validus* (10% decrease) (Fig. 2).

310
 311 For all species, the addition of new data reduces the percentage of extrapolation over the total
 312 projection area (-30.7% for *A. hodgsoni*, -12.7% for *B. loripes*, -20.5% for *G. antarctica*, -17.6% for
 313 *L. annulatus*, -10.2% for *O. validus* and -11% for *P. charcoti*, i.e. differences between the two
 314 extrapolation % values) and over the species suitable area as well (Fig. 2, Table S6).

315



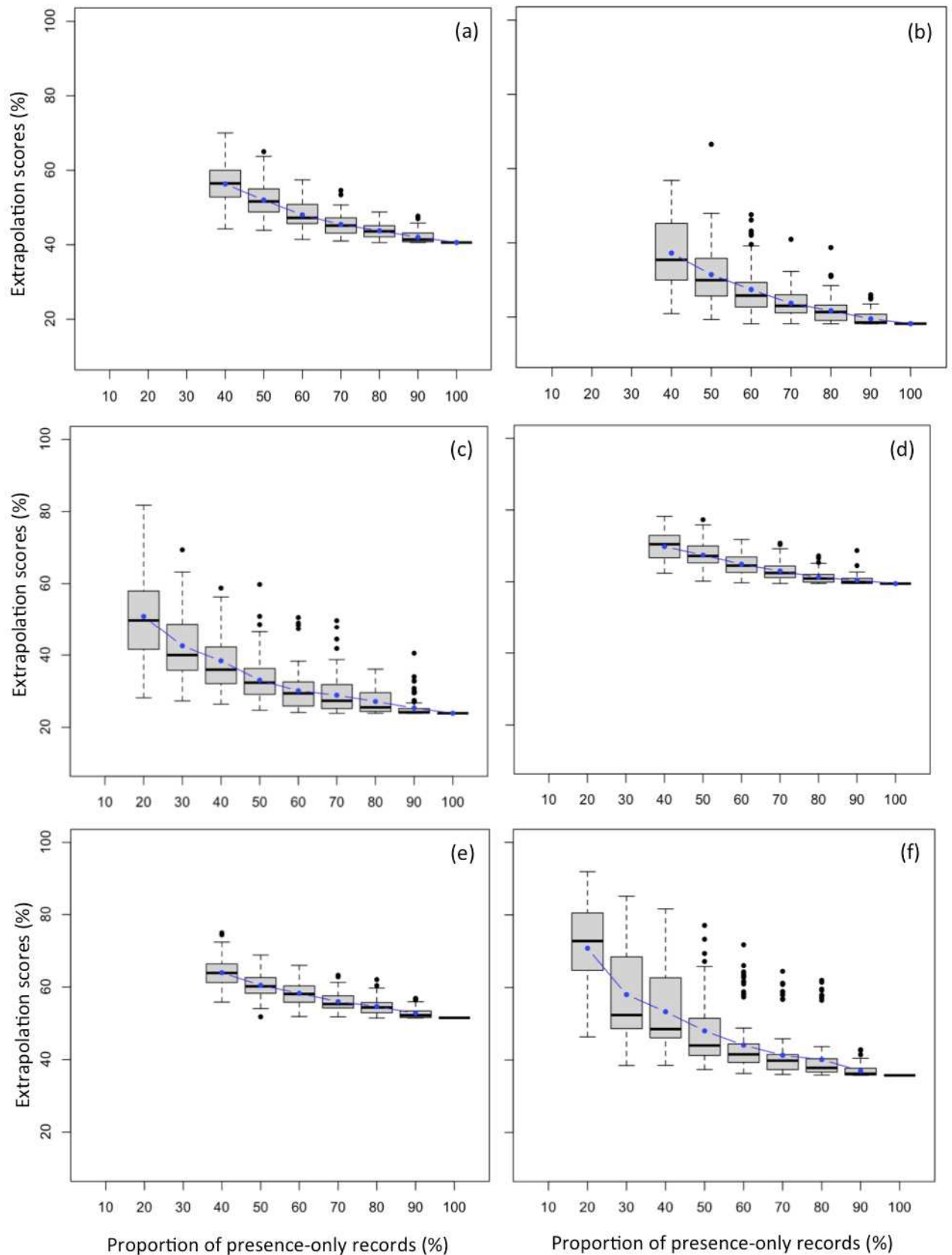
316
 317 **Figure 2.** Evolution of model performances with the increase of data (chronological addition of
 318 presence-only records, by 5-year periods, from 1980 to 2016). (A) Number of presence-only

319 records available to generate the model; (B) Mean correctly classified test data (%) (standard
320 deviation values available in Table S7); (C) Proportion of grid-cell pixels of the projection area that
321 are extrapolations (%). The maximal number of presence-only records present in Table 2 may not
322 be reached here because some collection dates remain unknown.

323
324 The decrease of extrapolation with the addition of presence-only records was tested by running, for
325 each species a series of models with different subsets of presence-only records randomly sampled
326 from the total dataset. One hundred model replicates were progressively run with 10 to 100% of
327 the total dataset and proportions of extrapolation areas were computed accordingly (Fig. 3, Table
328 S7). Results confirm that the addition of presence-only records strongly reduces proportions of
329 extrapolation areas. Proportions of extrapolation areas also vary between species models as a
330 function of depth. Low proportions of extrapolation areas are obtained in models run for deep
331 species and large datasets (e.g. 8.2% for 591 records in *B. loripes* and 23.9% for 851 records in *G.*
332 *antarctica*). In contrast, models run for shallower species show higher proportions of extrapolation
333 areas (40.6% for 298 records in *A. hodgsoni*, 51.5% for 375 records in *L. annulatus* and 35.8% for
334 337 records in *O. validus*). For these last species, spatial autocorrelation values are also higher
335 compared to other species (Table S7).

336

337



338
 339 **Figure 3.** Boxplot diagrams representing the decrease of proportions of extrapolation areas (in %
 340 of the total projection area) with addition of presence-only records used to generate model
 341 replicates (in % of data available, see Table 1 and Table S7), for: (a) *Acodontaster hodgsoni*, (b)

342 *Bathybiaster loripes*, (c) *Glabraster antarctica*, (d) *Labidiaster annulatus*, (e) *Odontaster validus*, (f)
 343 *Psilaster charcoti*. For each box, mean values (blue dots) and outliers (black dots) are shown for
 344 the 100 model replicates. Some boxes are missing for low percentages of presence-only records
 345 (10 to 30%, corresponding to close or less than 100 presence-only records) that do not allow
 346 models to be generated.

347
 348 A linear regression model was fit to the relationship between the number of presence-only records
 349 and proportions of extrapolation areas. For all species, regression coefficients are all negative and
 350 tested significant showing that proportions of extrapolation areas decrease with the addition of new
 351 records (Table 3). The intersection point between regression models and the (arbitrary) 10%
 352 extrapolation threshold was used to provide an estimate of the minimum number of records
 353 required for each species model to have an "adequate" proportion of extrapolation areas of 10%.
 354 This minimum number of presence-only records is reached for none of the studied species, and
 355 according to species, the number of presence-only records available should be increased at least
 356 by 1.6 to 3.3 times (Table 3).

357
 358 **Table 3.** Equations of simple linear regressions between the number of presence-only records X
 359 and the average proportion of extrapolation areas Y (Table 2, significance levels: * $p < 0.1$, **
 360 $p < 0.05$). The estimate of the number of presence-only records necessary to have a minimum
 361 "adequate" arbitrary proportion of extrapolation areas of 10% is given in the last column

Species	Equation	R ²	Estimated Pres.NB. (with multiplier of actual Pres.NB. available)
<i>Acodontaster hodgsoni</i>	$Y = -0.1358X + 73.616^{**}$	0.60	468 (x 1.6)
<i>Bathybiaster loripes</i>	$Y = -0.0249X + 28.974^*$	0.42	762 (x 1.3)
<i>Glabraster antarctica</i>	$Y = -0.0304X + 44.991^{**}$	0.61	1151 (x 1.4)
<i>Labidiaster annulatus</i>	$Y = -0.0913X + 88.078^{**}$	0.85	855 (x 2.3)
<i>Odontaster validus</i>	$Y = -0.0561X + 71.112^{**}$	0.93	1089 (x 3.2)
<i>Psilaster charcoti</i>	$Y = -0.0301X + 44.613^*$	0.37	1150 (x 3.3)

362

363

364

365 Discussion

366 Modelling performances and extrapolation

367 SDMs were generated for Southern Ocean sea star species, with contrasting distributions and
368 different numbers of presence-only records available (Table 1, Appendix 1). Overall, species
369 presence-only records are spatially concentrated in the most accessible and visited areas of the
370 Southern Ocean. Most of the sea star samples were collected close to the coasts of the Western
371 Antarctic Peninsula, the Ross Sea and sub-Antarctic Islands such as the Kerguelen Islands.
372 Consequently, high spatial autocorrelation values were computed, for *L. annulatus* and *O. validus*
373 in particular (Table S6).

374
375 Overall, models all show good performances (Table 2), the spatial cross-validation procedure
376 ensuring a relevant evaluation of modelling performances when using spatially aggregated data
377 ([Muscarella et al. 2014](#), [Dhingra et al. 2016](#), [Guillaumot et al. 2019](#)). However, models show high
378 proportions of extrapolation areas, with extrapolation covering up to 78% of the projection area in
379 *A. hodgsoni* model (Table 2). This means that even if models are evaluated as accurate, model
380 extrapolation area can concern up to three quarters of the projection area! Assessing the
381 proportion of the projection area for which models extrapolate is therefore necessary as a
382 complementary statistic to adapt modelling methods and improve model predictions. Masking
383 projections by extrapolation uncertainties is also important to perform accurate interpretations.

384
385 Extrapolation uncertainty maps have already been associated to SDM projections once in the
386 context of the Southern Ocean, by [Torres et al. \(2015\)](#) in their study of the grey petrel *Procellaria*
387 *cinerea*, performed at the scale of the Southern Ocean. More recently, the MESS approach has
388 been introduced in the methodological paper of [Guillaumot et al. \(2019\)](#), showing an extrapolation
389 area covering 64% of the projection area for the distribution model of the sea star *O. validus*, the
390 most studied benthic invertebrate of the Southern Ocean. However, uncertainties associated to
391 extrapolation were not provided in most model projections performed for Southern Ocean species
392 studies. For instance, modelled distributions performed for the sea urchins *Sterechinus neumayeri*
393 and *Sterechinus diadema* ([Pierrat et al. 2012](#)) were generated using a relative low number of
394 presence-only records (241 and 332, respectively). Based on results of the present study,
395 extrapolation could be expected to cover up to 60% of modelled distribution areas for these last
396 two species. Further Southern Ocean species distribution models were generated with sometimes
397 less than 100 presence-only records (see [Guillaumot et al. 2018b](#) and [Fabri-Ruiz et al. 2019](#) for
398 instance), suggesting that extrapolation could cover up to 70% of projection areas as visible in
399 models of *A. hodgsoni* and *P. charcoti* performed in our study with few records (Fig. 2, Table S6,
400 Table S7).

401
402 In addition to model uncertainties associated to extrapolation, other biases can alter the
403 performance of SDMs generated at broad spatial scales including the spatial and temporal

404 aggregation of data (Hortal et al. 2008, Tassarolo et al. 2014, 2017), the selection and quality of
405 environmental descriptors (Davies et al. 2008, Synes and Osborne 2011), the choice of modelling
406 algorithms and the definition of model settings (Hartley et al. 2006, Marmion et al. 2009). Providing
407 such uncertainty information, highlighted with some model statistics is very much encouraged
408 here, as they are essential to model interpretation (Beale and Lennon 2012, Guisan et al. 2013,
409 Yates et al. 2018).

410

411 **How can we reduce model extrapolation? Enriching SDMs with knowledge of species** 412 **ecology**

413 One objective of this work was to provide some methods to mitigate the effect of extrapolation on
414 model uncertainties. Our results show clear contrasts between models generated for “deep” and
415 “shallow” species, with lower proportions of extrapolation areas computed for deep species models
416 (29.1 and 15.73% respectively for *B. loripes* and *G. antarctica*). The model generated for *P.*
417 *charcoti* departs from this general scheme, with extrapolation reaching 67.9% of the projection
418 area. This is due to the strong spatial aggregation of records and the small presence-only record
419 dataset available in deeper habitats. Depth is indeed responsible for 58.1% of the extrapolation for
420 *P. charcoti* (Appendix 5). Indeed, the erroneous characterization of species occupied space, due to
421 an incomplete sampling, has been identified as a significant source of bias in SDM predictions
422 (Hortal et al. 2007, 2008, Rocchini et al. 2011, Sánchez-Fernández et al. 2011, Titeux et al. 2017,
423 El-Gabbas and Dormann 2018).

424

425 Limiting model projection areas to biogeographically, or ecologically “realistic” depth ranges can
426 help reduce extrapolation as exemplified in the present study, for models of *A. hodgsoni* and *P.*
427 *charcoti*, for which extrapolation was reduced from 78.6 to 40.6% and 67.9 to 35.8% respectively
428 (Table 2). Restraining model projection areas based on species ecological or physiological
429 tolerance thresholds is a common approach in ecological modelling using experimental data or
430 field observations (Kearney and Porter 2009, Hare et al. 2012, De Villiers et al. 2013). Knowledge
431 of species ecology and physiology can also be useful to delineate transferability areas (Feng and
432 Papes 2017) and improve distribution models, as recently shown for Southern Ocean species
433 (Guillaumot et al. 2018a, Guillaumot et al. 2019). Feng et al. (2020) developed a new modelling
434 algorithm, called *Plateau*, which uses experimental data to define upper temperature conditions in
435 distribution models. For temperature and salinity, physiological experiments and field observations
436 can be used in models to determine species tolerance thresholds. This requires knowledge about
437 the species ecology and physiology and the input from specialists, all conditions that remain
438 difficult to meet, regarding deep sea species of the Southern Ocean (Gage 2004, Gutt et al. 2010,
439 De Broyer and Danis 2011). Moreover, several studies suggested that some Southern Ocean
440 species might have found refuges in deep sea habitats in the past, during glacial maxima, which

441 makes species depth range difficult to precise when deep and shallow populations have not been
442 differentiated into distinct taxonomic units yet ([Rogers 2007](#), [Arango et al. 2011](#), [Havermans et al.](#)
443 [2011](#), [Near et al. 2012](#)).

444
445 **How can we reduce extrapolation? Improving sampling effort**
446 Increased sampling effort over enlarged areas allows the production of larger datasets from which
447 many records can be used to generate reliable models with reduced extrapolation areas. In this
448 study, proportions of extrapolation areas proportionally decreased when increased numbers of
449 presence-only records were used to generate models. The occurrence datasets were significantly
450 augmented between 1980 and 2016, with a number of presence-only records multiplied by 1.9 to
451 3.3 times according to the studied species, which allowed reduction of model extrapolation from
452 10.2 to 30.7% according to the species (Fig. 2, Table S6). However, results suggest that about
453 twice the number of presence-only records actually available would be necessary to reduce
454 extrapolation down to a “satisfactory” threshold of 10% of the projection area (Table 3).

455
456 Generating reliable and stable models using a sufficient number of presence-only records is
457 essential. In this study, some models could not be run when the number of presence-only records
458 was too low (approaching 150 presence-only records or less) compared to the broad extent of the
459 projection area and the spatial aggregation of these data (Table S7). Considering that the spatial
460 cross-validation procedure splits the initial dataset into training and test data, and that at each step,
461 75% of these training data are randomly sampled by BRT to iterately create a model tree (and
462 generate stochasticity in the procedure), the final number of presence-only records available to
463 describe the presence data - environment relationship becomes too low (around 37.5% of the
464 initial number of presence-only records).

465 The lowest number of presence-only records required to build a reliable model is species-
466 dependent as not all presence-only records are equally informative, due to species-specific
467 relationships between records and the environment. When models are generated using BRT,
468 records that bring no new environmental information to the model are dropped because they are
469 not informative enough to improve the construction of BRT trees. Pruning non-informative data
470 also reduces the total number of presence-only records available to generate a model ([Elith et al.](#)
471 [2008](#)). This is strongly related to prevalence that is, the ratio between the number of presence-only
472 records and the size of the projection area ([Jiménez-Valverde et al. 2009](#), [Santika 2011](#), [Barbet-](#)
473 [Massin et al. 2012](#)). In order to accurately describe a vast projection area and be able to create a
474 model, it is necessary to gather a substantial amount of information about the geographic
475 environmental conditions and about species known distribution. If a limited number of records is
476 available and these data are aggregated in space (i.e. weakly informative), the first trees produced
477 by BRT will contain most of the model deviance, but as no new information is provided, the model

478 will quickly overfit because redundant information is provided by close presence-only records.
479 Eventually, this will make the model collapse.

480 Increasing the number of presence-only records is proved an efficient alternative to generate more
481 relevant models (Stockwell and Peterson 2002, Feeley and Silman 2011, van Proosdij et al. 2016),
482 but the spatial distribution of these records is of importance as well (Yates et al. 2018). A uniform
483 distribution of records over the entire projection area reduces spatial autocorrelation and optimizes
484 the sampling and representativeness of environmental conditions under which species can thrive.
485 In this study, the spatial aggregation of species records was particularly high for two species, *O.*
486 *validus* and *L. annulatus*. It was estimated that the number of supplementary presence-only
487 records necessary to reach a proportion of extrapolation areas of 10% should be twice as high as it
488 is for other species (Table 3). Additional data are necessary to improve the establishment of the
489 relationship between species distribution and the environment because species records are less
490 informative when aggregated than when they are evenly distributed.

491
492 The Southern Ocean covers contrasting environmental conditions, biogeographic regions and
493 ecoregions (Pierrat 2011, Fabri-Ruiz et al. 2020). Ideally, both species presence and absence
494 should be recorded in each ecoregion for an accurate description of the occupied space (Torres et
495 al. 2015). Because such a sampling effort is usually not achievable, nor realistic, alternatives would
496 consist of (1) a relevant adjustment of projection areas, with for instance the combination of
497 several SDM projections using different grid sizes according to what is available. Generating SDM
498 projections for large areas and combining results with projections zoomed in on areas where more
499 environmental detail is available would provide more relevant and realistic modelled species
500 distributions (Seo et al. 2009, Anderson and Raza 2010). (2) In order to compensate for the lack of
501 presence-record availability, the 'ensembles of small models' approach is another alternative. This
502 method fits a set of bivariate models (i.e. generated with two environmental descriptors only),
503 within a hierarchic multi-scale framework (i.e. zooming in and out in space from local to regional
504 predictions), and finally averages this ensemble of models with a weighted ensemble approach,
505 which subsequently provides more accurate and robust model predictions (Lomba et al. 2010,
506 Breiner et al. 2015, Habibzadeh and Ludwig 2019).

507

508 **Some limitations to the MESS approach**

509 The MESS approach can reveal parts of projection areas where models extrapolate. Extrapolation
510 however can be over-estimated. Indeed, extrapolation is considered as soon as the value of a
511 single environmental descriptor falls outside the range of the known species environmental
512 requirements. But, some extreme values would not limit but can promote species presence: this is
513 the case for descriptors relating to food resource availability (e.g. chlorophyll a, POC
514 concentrations...), for which a high pixel value exceeding the range of values recorded based on

515 species presences will be still considered as extrapolation, although more food usually means
516 suitable conditions for species distribution.

517 Some fine-tuning of the MESS approach would imply to identify, for each pixel, which descriptor is
518 responsible for extrapolation and filter the conditions for which the model should really extrapolate.
519 Such an approach was developed by [Owens et al. \(2013\)](#), who used the MOP method (Mobility
520 Oriented Parity). Based on multivariate analyses, they determined if pixels contain a combination
521 of environmental conditions that should induce extrapolation. In contrast to the MESS approach,
522 the MOP method can directly differentiate proportions of extrapolation areas according to the
523 combination of descriptors responsible for extrapolation. Another complex alternative is the ExDet
524 tool, developed by [Mesgaran et al. \(2014\)](#), which also accounts for multivariate extrapolation
525 possibilities, i.e. extrapolation linked to novel combinations between covariates.

526 In this study, the MESS approach was favored as a more strict and conservative method to
527 highlight the importance of extrapolation, the effect of data quantity and quality, and the relevance
528 of the proposed corrections. The MESS is also simpler to apply and well suited to exploratory
529 studies.

530

531 **Conclusions**

532 This study shows that when modelling species distribution on broad scale areas, such as the
533 Southern Ocean, important proportions of predicted distribution probabilities (suitable or not) are
534 model extrapolations. This extrapolation uncertainty relies on the completeness of species
535 sampling, and the definition of its occupied space to calibrate the model. Extrapolation occurs in
536 areas where habitat suitability is unknown as no information on species presence or absence is
537 provided.

538
539 Reducing extrapolation is possible by combining SDM with ecological and physiological knowledge
540 of species requirements (e.g. depth range, temperature tolerance thresholds). Increased sampling
541 effort over enlarged areas also allows the production of more reliable models with reduced
542 extrapolation areas and our study shows that doubling the number of presence-only records
543 available to generate the model would help reduce the extrapolation area down to 10% of the
544 projected area.

545 While more data samples remain unavailable, some methods are increasingly developed to
546 improve model performances, by adjusting the extent of the projection area or by generating and
547 aggregating several small ensemble models.

548

549 Finally, present results call for a widespread use of extrapolation maps and uncertainties
550 associated to model predictions in model outputs, along with information about the quantity of
551 presence-only records available, the quality and resolution of environmental descriptors and the
552 state of our knowledge of species ecology. These are all essential information needed to support
553 model interpretations, as also stated in recent publications that review best practices in ecological
554 modelling ([Araújo et al. 2019](#), [Zurell et al. 2020](#)).

555

556 **Acknowledgements**

557 This work was supported by a “Fonds pour la formation à la Recherche dans l’Industrie et
558 l’Agriculture” (FRIA) and “Bourse fondation de la mer” grants to C. Guillaumot.

559 This is contribution no. 46 to the vERSO project (www.versoproject.be), funded by the Belgian
560 Science Policy Office (BELSPO, contract n°BR/132/A1/vERSO). Research was also financed by
561 the “Refugia and Ecosystem Tolerance in the Southern Ocean” project (RECTO;
562 BR/154/A1/RECTO) funded by the Belgian Science Policy Office (BELSPO), this study being
563 contribution number 23.

564

565

566

567

568

569

570

571

572

573

574

575

576

577

578

579 **References**

580 [Allouche](#), O., Tsoar, A. & Kadmon, R. (2006). Assessing the accuracy of species distribution
581 models: prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, 43(6),
582 1223-1232.

583
584 [Anderson](#), R.P. & Raza, A. (2010). The effect of the extent of the study region on GIS models of
585 species geographic distributions and estimates of niche evolution: preliminary tests with montane
586 rodents (genus *Nephelomys*) in Venezuela. *Journal of Biogeography*, 37(7), 1378-1393.

587
588 [Anderson](#), R.P. (2013). A framework for using niche models to estimate impacts of climate change
589 on species distributions. *Annals of the New York Academy of Sciences*, 1297(1), 8-28.

590
591 [Arango](#), C.P., Soler-Membrives, A. & Miller, K.J. (2011). Genetic differentiation in the circum—
592 Antarctic sea spider *Nymphon australe* (Pycnogonida; Nymphonidae). *Deep Sea Research Part II:*
593 *Topical Studies in Oceanography*, 58(1-2), 212-219.

594
595 [Araújo](#), M. B., Anderson, R. P., Barbosa, A. M., Beale, C. M., Dormann, C. F., Early, R., ... &
596 O'Hara, R. B. (2019). Standards for distribution models in biodiversity assessments. *Science*
597 *Advances*, 5(1), eaat4858.

598
599 [Arthur](#), B., Hindell, M., Bester, M., De Bruyn, P.N., Goebel, M.E., Trathan, P. & Lea, M.A. (2018).
600 Managing for change: Using vertebrate at sea habitat use to direct management efforts. *Ecological*
601 *Indicators*, 91, 338-349.

602
603 [Austin](#), M. (2007). Species distribution models and ecological theory: a critical assessment and
604 some possible new approaches. *Ecological modelling*, 200(1-2), 1-19.

605
606 [Austin](#), M.P. & Van Niel, K.P. (2011). Improving species distribution models for climate change
607 studies: variable selection and scale. *Journal of Biogeography*, 38(1), 1-8.

608
609 [Ballard](#), G., Jongsomjit, D., Veloz, S.D. & Ainley, D.G. (2012). Coexistence of mesopredators in an
610 intact polar ocean ecosystem: the basis for defining a Ross Sea marine protected area. *Biological*
611 *Conservation*, 156, 72-82.

612
613 [Barbet-Massin](#), M., Jiguet, F., Albert, C.H. & Thuiller, W. (2012). Selecting pseudo-absences for
614 species distribution models: how, where and how many?. *Methods in ecology and evolution*, 3(2),
615 327-338.

616
617 [Basher](#), Z. & Costello, M.J. (2016). The past, present and future distribution of a deep-sea shrimp
618 in the Southern Ocean. *PeerJ*, 4, e1713.
619
620 [Beale](#), C.M., & Lennon, J.J. (2012). Incorporating uncertainty in predictive species distribution
621 modelling. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1586), 247-
622 258.
623
624 [Breiner](#), F.T., Guisan, A., Bergamini, A. & Nobis, M.P. (2015). Overcoming limitations of modelling
625 rare species by using ensembles of small models. *Methods in Ecology and Evolution*, 6(10), 1210-
626 1218.
627
628 [Broennimann](#), O., Mráz, P., Petitpierre, B., Guisan, A. & Müller-Schärer, H. (2014). Contrasting
629 spatio-temporal climatic niche dynamics during the eastern and western invasions of spotted
630 knapweed in North America. *Journal of Biogeography*, 41(6), 1126-1136.
631
632 [Brotons](#), L., Thuiller, W., Araújo, M.B. & Hirzel, A.H. (2004). Presence-absence versus presence-
633 only modelling methods for predicting bird habitat suitability. *Ecography*, 27(4), 437-448.
634
635 [Brotons](#), L., De Cáceres, M., Fall, A. & Fortin, M.J. (2012). Modeling bird species distribution
636 change in fire prone Mediterranean landscapes: incorporating species dispersal and landscape
637 dynamics. *Ecography*, 35(5), 458-467.
638
639 [Brueggeman](#), P. (1998). Underwater Field Guide to Ross Island & McMurdo Sound, Antarctica.
640 The National Science Foundation's Office of Polar Programs sponsored Norbert Wu.—Univ.
641 California, San Diego.
642
643 [Cardador](#), L., Carrete, M., Gallardo, B. & Tella, J.L. (2016). Combining trade data and niche
644 modelling improves predictions of the origin and distribution of non-native European populations of
645 a globally invasive species. *Journal of Biogeography*, 43(5), 967-978.
646
647 [CCAMLR report](#) WG-FSA-15/64, access at <https://www.ccamlr.org/fr/wg-fsa-15/64>. August 2020.
648
649 [Crase](#), B., Liedloff, A.C. & Wintle, B.A. (2012). A new method for dealing with residual spatial
650 autocorrelation in species distribution models. *Ecography*, 35(10), 879-888.
651

652 [Davies](#), A. J., Wisshak, M., Orr, J.C. & Roberts, J.M. (2008). Predicting suitable habitat for the
653 cold-water coral *Lophelia pertusa* (Scleractinia). *Deep Sea Research Part I: Oceanographic*
654 *Research Papers*, 55(8), 1048-1062.

655

656 [De Broyer](#), C. & Danis, B. (2011). How many species in the Southern Ocean? Towards a dynamic
657 inventory of the Antarctic marine species. *Deep sea research Part II: Topical studies in*
658 *oceanography*, 58(1-2), 5-17.

659

660 [De Broyer](#), C., Koubbi, P., Griffiths, H.J., Raymond, B., d'Udekem d'Acoz, C., Van de Putte, A.P.,
661 ... Ropert-Coudert, Y. (2014). Biogeographic atlas of the Southern Ocean (p. 498). C. De Broyer,
662 & P. Koubbi (Eds.). Cambridge: Scientific Committee on Antarctic Research.

663

664 [De Villiers](#), M., Hattingh, V. & Kriticos, D.J. (2013). Combining field phenological observations with
665 distribution data to model the potential distribution of the fruit fly *Ceratitis rosa* Karsch (Diptera:
666 Tephritidae). *Bulletin of Entomological Research*, 103(1), 60-73.

667

668 [Dhingra](#), M.S., Artois, J., Robinson, T.P., Linard, C., Chaiban, C., Xenarios, I. ... & Von
669 Dobschuetz, S. (2016). Global mapping of highly pathogenic avian influenza H5N1 and H5Nx
670 clade 2.3. 4.4 viruses with spatial cross-validation. *Elife*, 5, e19571.

671

672 [El-Gabbas](#), A., & Dormann, C.F. (2018). Wrong, but useful: regional species distribution models
673 may not be improved by range-wide data under biased sampling. *Ecology and evolution*, 8(4),
674 2196-2206.

675

676 [Elith](#), J., Graham, H., Anderson, C.P., Dudík, R., Ferrier, M., Guisan, A. ... & Li, J. (2006). Novel
677 methods improve prediction of species' distributions from occurrence data. *Ecography*, 29(2), 129-
678 151.

679

680 [Elith](#), J., Leathwick, J.R. & Hastie, T. (2008). A working guide to boosted regression trees. *Journal*
681 *of Animal Ecology*, 77(4), 802-813.

682

683 [Elith](#), J., Kearney, M. & Phillips, S. (2010). The art of modelling range-shifting species. *Methods in*
684 *Ecology and Evolution*, 1(4), 330-342.

685

686 [Elith](#), J., Phillips, S.J., Hastie, T., Dudík, M., Chee, Y.E., Yates, C.J., 2011. A statistical explanation
687 of MaxEnt for ecologists. *Diversity and Distributions*, 17, 43–57.

688

689 [Escobar](#), L.E., Ryan, S.J., Stewart-Ibarra, A.M., Finkelstein, J.L., King, C.A., Qiao, H. & Polhemus,
690 M.E. (2015). A global map of suitability for coastal *Vibrio cholerae* under current and future climate
691 conditions. *Acta tropica*, 149, 202-211.
692

693 [Fabri-Ruiz](#), S., Danis, B., David, B. & Saucède, T. (2019). Can we generate robust species
694 distribution models at the scale of the Southern Ocean?. *Diversity and distributions*, 25(1), 21-37.
695

696 [Fabri-Ruiz](#), S., Danis, B., Navarro, N., Koubbi, P., Laffont, R. & Saucède, T. (2020). Benthic
697 ecoregionalization based on echinoid fauna of the Southern Ocean supports current proposals of
698 Antarctic Marine Protected Areas under IPCC scenarios of climate change. *Global Change*
699 *Biology*.
700

701 [Feeley](#), K.J. & Silman, M.R. (2011). Keep collecting: accurate species distribution modelling
702 requires more collections than previously thought. *Diversity and distributions*, 17(6), 1132-1140.
703

704 [Feng](#), X. & Papeş, M. (2017). Can incomplete knowledge of species' physiology facilitate
705 ecological niche modelling? A case study with virtual species. *Diversity and Distributions*, 23(10),
706 1157-1168.
707

708 [Feng](#), X., Liang, Y., Gallardo, B. & Papeş, M. (2020). Physiology in ecological niche modeling:
709 using zebra mussel's upper thermal tolerance to refine model predictions through Bayesian
710 analysis. *Ecography*, 43(2), 270-282.
711

712 [Fielding](#), A.H. & Bell, J.F. (1997). A review of methods for the assessment of prediction errors in
713 conservation presence/absence models. *Environmental Conservation*, 24(1), 38-49.
714

715 [Fitzpatrick](#), M.C. & Hargrove, W.W. (2009). The projection of species distribution models and the
716 problem of non-analog climate. *Biodiversity and Conservation*, 18(8), 2255.
717

718 [Freer](#), J.J., Tarling, G.A., Collins, M.A., Partridge, J.C. & Genner, M.J. (2019). Predicting future
719 distributions of lanternfish, a significant ecological resource within the Southern Ocean. *Diversity*
720 *and Distributions*, 25(8), 1259-1272.
721

722 [Gage](#), J.D. (2004). Diversity in deep-sea benthic macrofauna: the importance of local ecology, the
723 larger scale, history and the Antarctic. *Deep Sea Research Part II: Topical Studies in*
724 *Oceanography*, 51(14-16), 1689-1708.
725

726 [Gallego](#), R., Dennis, T.E., Basher, Z., Lavery, S. & Sewell, M.A. (2017). On the need to consider
727 multiphasic sensitivity of marine organisms to climate change: A case study of the Antarctic acorn
728 barnacle. *Journal of Biogeography*, 44, 2165–2175.

729

730 [Gobeyn](#), S., Mouton, A.M., Cord, A.F., Kaim, A., Volk, M. & Goethals, P.L. (2019). Evolutionary
731 algorithms for species distribution modelling: A review in the context of machine learning.
732 *Ecological modelling*, 392, 179-195.

733

734 [Grimm](#), V. & Berger, U. (2016). Robustness analysis: Deconstructing computational models for
735 ecological theory and applications. *Ecological Modelling*, 326, 162-167.

736

737 [Guillaumot](#), C., Martin, A., Eléaume, M., & Saucède, T. (2018a). Methods for improving species
738 distribution models in data-poor areas: example of sub-Antarctic benthic species on the Kerguelen
739 Plateau. *Marine Ecology Progress Series*, 594, 149-164.

740

741 [Guillaumot](#), C., Fabri-Ruiz, S., Martin, A., Eléaume, M., Danis, B., Féral, J.P. & Saucède, T.
742 (2018b). Benthic species of the Kerguelen Plateau show contrasting distribution shifts in response
743 to environmental changes. *Ecology and evolution*, 8(12), 6210-6225.

744

745 [Guillaumot](#), C., Danis B. & Saucède, T. (2020). Selecting environmental descriptors is critical to
746 modelling the distribution of Antarctic benthic species. *Polar Biology*. 1-19.

747

748 [Guillaumot](#), C., Artois, J., Saucède, T., Demoustier, L., Moreau, C., Eléaume, M. ... & Danis, B.
749 (2019). Broad-scale species distribution models applied to data-poor areas. *Progress in*
750 *oceanography*, 175, 198-207.

751

752 [Guisan](#), A., Tingley, R., Baumgartner, J.B., Naujokaitis-Lewis, I., Sutcliffe, P.R., Tulloch, A.I. ... &
753 Martin, T.G. (2013). Predicting species distributions for conservation decisions. *Ecology letters*,
754 16(12), 1424-1435.

755

756 [Gutt](#), J., Hosie, G. & Stoddart, M., (2010). Marine Life in the Antarctic. In: McIntyre A.D. (Ed.). Life
757 in the World's Oceans: Diversity, Distribution and Abundance. Wiley-Blackwell, Oxford, pp. 203–
758 220.

759

760 [Habibzadeh](#), N. & Ludwig, T. (2019). Ensemble of small models for estimating potential abundance
761 of Caucasian grouse (*Lyrurus mlokosiewiczzi*) in Iran. *Ornis Fennica*, 96(2), 77-89.

762

763 Hare, J. A., Wuenschel, M.J. & Kimball, M.E. (2012). Projecting range limits with coupled thermal
764 tolerance-climate change models: an example based on gray snapper (*Lutjanus griseus*) along the
765 US east coast. *PLoS One*, 7(12), e52294.

766

767 Hartley, S., Harris, R., & Lester, P.J. (2006). Quantifying uncertainty in the potential distribution of
768 an invasive species: climate and the Argentine ant. *Ecology letters*, 9(9), 1068-1079.

769

770 Havermans, C., Nagy, Z.T., Sonet, G., De Broyer, C. & Martin, P. (2011). DNA barcoding reveals
771 new insights into the diversity of Antarctic species of Orchomene sensu lato (Crustacea:
772 Amphipoda: Lysianassoidea). *Deep Sea Research Part II: Topical Studies in Oceanography*, 58(1-
773 2), 230-241.

774

775 Heikkinen, R.K., Marmion, M. & Luoto, M. (2012). Does the interpolation accuracy of species
776 distribution models come at the expense of transferability?. *Ecography*, 35(3), 276-288.

777

778 Hijmans, R.J., Phillips, S., Leathwick, J., Elith, J. & Hijmans, M.R. (2017). Package 'dismo'. *Circles*,
779 9(1). <https://CRAN.R-project.org/package=dismo>

780

781 Hortal, J., Lobo, J.M. & Jiménez-Valverde, A. (2007). Limitations of biodiversity databases: case
782 study on seed-plant diversity in Tenerife, Canary Islands. *Conservation Biology*, 21(3), 853-863.

783

784 Hortal, J., Jiménez-Valverde, A., Gómez, J.F., Lobo, J.M. & Baselga, A. (2008). Historical bias in
785 biodiversity inventories affects the observed environmental niche of the species. *Oikos*, 117(6),
786 847-858.

787

788 Iannella, M., Cerasoli, F. & Biondi, M. (2017). Unraveling climate influences on the distribution of
789 the parapatric newts *Lissotriton vulgaris meridionalis* and *L. italicus*. *Frontiers in Zoology*, 14(1),
790 55.

791

792 Jerosch, K., Scharf, F.K., Deregibus, D., Campana, G.L., Zacher, K., Pehlke, H. ... & Abele, D.
793 (2019). Ensemble modelling of Antarctic macroalgal habitats exposed to glacial melt in a polar
794 fjord. *Frontiers in Ecology and Evolution*, 7, 207.

795

796 Jiménez-Valverde, A., Lobo, J.M. & Hortal, J. (2009). The effect of prevalence and its interaction
797 with sample size on the reliability of species distribution models. *Community Ecology*, 10(2), 196-
798 205.

799

800 [Kearney](#), M. & Porter, W. (2009). Mechanistic niche modelling: combining physiological and spatial
801 data to predict species' ranges. *Ecology letters*, 12(4), 334-350.
802

803 [Lawrence](#), J.M. (Ed.). (2013). *Starfish: biology and ecology of the Asteroidea*. JHU Press.
804

805 [Li](#), G., Du, S. & Guo, K. (2015). Correction: Evaluation of Limiting Climatic Factors and Simulation
806 of a Climatically Suitable Habitat for Chinese Sea Buckthorn. *PloS one*, 10(8), e0136001.
807

808 [Liu](#), C., White, M. & Newell, G. (2013). Selecting thresholds for the prediction of species
809 occurrence with presence-only data. *Journal of Biogeography*, 40(4), 778-789.
810

811 [Lomba](#), A., Pellissier, L., Randin, C., Vicente, J., Moreira, F., Honrado, J. & Guisan, A. (2010).
812 Overcoming the rare species modelling paradox: a novel hierarchical framework applied to an
813 Iberian endemic plant. *Biological conservation*, 143(11), 2647-2657.
814

815 [Loots](#), C., Koubbi, P. & Duhamel, G. (2007). Habitat modelling of *Electrona antarctica*
816 (Myctophidae, Pisces) in Kerguelen by generalized additive models and geographic information
817 systems. *Polar Biology*, 30, 951-959.
818

819 [Luizza](#), M.W., Wakie, T., Evangelista, P.H. & Jarnevich, C.S. (2016). Integrating local pastoral
820 knowledge, participatory mapping, and species distribution modeling for risk assessment of
821 invasive rubber vine (*Cryptostegia grandiflora*) in Ethiopia's Afar region. *Ecology and Society*,
822 21(1), 1-22.
823

824 [Luoto](#), M., Pöyry, J., Heikkinen, R.K. & Saarinen, K. (2005). Uncertainty of bioclimate envelope
825 models based on the geographical distribution of species. *Global Ecology and Biogeography*,
826 14(6), 575-584.
827

828 [Mah](#), C.L. & Blake, D.B. (2012). Global diversity and phylogeny of the Asteroidea (Echinodermata).
829 *PloS one*, 7(4), e35644.
830

831 [Marmion](#), M., Luoto, M., Heikkinen, R.K. & Thuiller, W. (2009). The performance of state-of-the-art
832 modelling techniques depends on geographical distribution of species. *Ecological Modelling*,
833 220(24), 3512-3520.
834

835 [Marshall](#), C.E., Glegg, G.A. & Howell, K.L. (2014). Species distribution modelling to support marine
836 conservation planning: the next steps. *Marine Policy*, 45, 330-332.

837
838 [McClintock](#), J.B., Angus, R.A., Ho, C.P., Amsler, C.D. & Baker, B.J. (2008). Intraspecific agonistic
839 arm-fencing behavior in the Antarctic keystone sea star *Odontaster validus* influences prey
840 acquisition. *Marine Ecology Progress Series*, 371, 297-300.
841
842 [Mesgaran](#), M. B., Cousens, R. D., & Webber, B. L. (2014). Here be dragons: a tool for quantifying
843 novelty due to covariate range and correlation change when projecting species distribution models.
844 *Diversity and Distributions*, 20(10), 1147-1159.
845
846 [Milanesi](#), P., Herrando, S., Pla, M., Villero, D. & Keller, V. (2017). Towards continental bird
847 distribution models: environmental variables for the second European breeding bird atlas and
848 identification of priorities for further surveys. *Vogelwelt*, 137, 53-60.
849
850 [Moreau](#), C., Mah, C., Agüera, A., Améziane, N., Barnes, D., Crokaert, G. ... & Jażdżewska, A.
851 (2018). Antarctic and sub-Antarctic Asteroidea database. *ZooKeys*, 747, 141-156.
852
853 [Muscarella](#), R., Galante, P.J., Soley-Guardia, M., Boria, R.A., Kass, J.M., Uriarte, M. & Anderson,
854 R.P. (2014). ENM eval: An R package for conducting spatially independent evaluations and
855 estimating optimal model complexity for Maxent ecological niche models. *Methods in Ecology and*
856 *Evolution*, 5(11), 1198-1205.
857
858 [Nachtsheim](#), D.A., Jerosch, K., Hagen, W., Plötz, J. & Bornemann, H. (2017). Habitat modelling of
859 crabeater seals (*Lobodon carcinophaga*) in the Weddell Sea using the multivariate approach
860 Maxent. *Polar Biology*, 40(5), 961-976.
861
862 [Naimi](#) B., Hamm N.A., Groen T.A., Skidmore A.K. & Toxopeus A.G. (2014). Where is positional
863 uncertainty a problem for species distribution modelling? *Ecography*, 37(2), 191-203.
864
865 [Near](#), T.J., Dornburg, A., Kuhn, K.L., Eastman, J.T., Pennington, J.N., Patarnello, T. ... & Jones,
866 C.D. (2012). Ancient climate change, antifreeze, and the evolutionary diversification of Antarctic
867 fishes. *Proceedings of the National Academy of Sciences*, 109(9), 3434-3439.
868
869 [Nori](#), J., Akmentins, M.S., Ghirardi, R., Frutos, N. & Leynaud, G.C. (2011). American bullfrog
870 invasion in Argentina: where should we take urgent measures?. *Biodiversity and Conservation*,
871 20(5), 1125-1132.
872

873 [Owens](#), H.L., Campbell, L.P., Dornak, L.L., Saupe, E.E., Barve, N., Soberón, J. ... & Peterson, A.T.
874 (2013). Constraints on interpretation of ecological niche models by limited environmental ranges
875 on calibration areas. *Ecological Modelling*, 263, 10-18.

876

877 [Peterson](#), A.T. (2001). Predicting species' geographic distributions based on ecological niche
878 modeling. *The Condor*, 103(3), 599-605.

879

880 [Phillips](#), S.J., Dudík, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J. & Ferrier, S. (2009).
881 Sample selection bias and presence-only distribution models: implications for background and
882 pseudo-absence data. *Ecological Applications*, 19(1), 181-197.

883

884 [Pierrat](#), B. (2011). *Macroécologie des échinides de l'océan Austral: Distribution, Biogéographie et*
885 *Modélisation* (Doctoral dissertation).

886

887 [Pierrat](#), B., Saucède, T., Laffont, R., De Ridder, C., Festeau, A., & David, B. (2012). Large-scale
888 distribution analysis of Antarctic echinoids using ecological niche modelling. *Marine Ecology*
889 *Progress Series*, 463, 215-230.

890

891 [Pinkerton](#), M.H., Smith, A.N., Raymond, B., Hosie, G.W., Sharp, B., Leathwick, J.R. & Bradford-
892 Grieve, J.M. (2010). Spatial and seasonal distribution of adult *Oithona similis* in the Southern
893 Ocean: predictions using boosted regression trees. *Deep Sea Research Part I: Oceanographic*
894 *Research Papers*, 57(4), 469-485.

895

896 [Randin](#), C.F., Dirnböck, T., Dullinger, S., Zimmermann, N.E., Zappa, M. & Guisan, A. (2006). Are
897 niche-based species distribution models transferable in space?. *Journal of biogeography*, 33(10),
898 1689-1703.

899

900 [Reiss](#), H., Cunze, S., König, K., Neumann, H. & Kröncke, I. (2011). Species distribution modelling
901 of marine benthos: a North Sea case study. *Marine Ecology Progress Series*, 442, 71-86.

902

903 [Rocchini](#), D., Hortal, J., Lengyel, S., Lobo, J.M., Jimenez-Valverde, A., Ricotta, C. ... & Chiarucci,
904 A. (2011). Accounting for uncertainty when mapping species distributions: the need for maps of
905 ignorance. *Progress in Physical Geography*, 35(2), 211-226.

906

907 [Robinson](#), L.M., Elith, J., Hobday, A.J., Pearson, R.G., Kendall, B.E., Possingham, H.P. &
908 Richardson, A.J. (2011). Pushing the limits in marine species distribution modelling: lessons from
909 the land present challenges and opportunities. *Global Ecology and Biogeography*, 20(6), 789-802.

910
911 [Rogers](#), A.D. (2007). Evolution and biodiversity of Antarctic organisms: a molecular perspective.
912 *Philosophical transactions of the royal society B: Biological sciences*, 362(1488), 2191-2214.
913
914 [Sánchez-Fernández](#), D., Lobo, J.M. & Hernández-Manrique, O.L. (2011). Species distribution
915 models that do not incorporate global data misrepresent potential distributions: a case study using
916 Iberian diving beetles. *Diversity and Distributions*, 17(1), 163-171.
917
918 [Santika](#), T. (2011). Assessing the effect of prevalence on the predictive performance of species
919 distribution models using simulated data. *Global Ecology and Biogeography*, 20(1), 181-192.
920
921 [Seo](#), C., Thorne, J.H., Hannah, L. & Thuiller, W. (2009). Scale effects in species distribution
922 models: implications for conservation planning under climate change. *Biology letters*, 5(1), 39-43.
923
924 [Silva](#), B.P.C., Silva, M.L.N., Avalos, F.A.P., de Menezes, M.D. & Curi, N. (2019). Digital soil
925 mapping including additional point sampling in Posses ecosystem services pilot watershed,
926 southeastern Brazil. *Scientific Reports*, 9(1), 1-12.
927
928 [Stockwell](#), D.R., & Peterson, A.T. (2002). Effects of sample size on accuracy of species distribution
929 models. *Ecological modelling*, 148(1), 1-13.
930
931 [Synes](#), N.W. & Osborne, P.E. (2011). Choice of predictor variables as a source of uncertainty in
932 continental-scale species distribution modelling under climate change. *Global Ecology and*
933 *Biogeography*, 20(6), 904-914.
934
935 [Tessarolo](#), G., Rangel, T.F., Araújo, M.B. & Hortal, J. (2014). Uncertainty associated with survey
936 design in Species Distribution Models. *Diversity and Distributions*, 20(11), 1258-1269.
937
938 [Tessarolo](#), G., Ladle, R., Rangel, T. & Hortal, J. (2017). Temporal degradation of data limits
939 biodiversity research. *Ecology and evolution*, 7(17), 6863-6870.
940
941 [Titeux](#), N., Maes, D., Van Daele, T., Onkelinx, T., Heikkinen, R.K., Romo, H. ... & Schweiger, O.
942 (2017). The need for large-scale distribution data to estimate regional changes in species richness
943 under future climate change. *Diversity and Distributions*, 23(12), 1393-1407.
944

945 [Torres](#), L.G., Sutton, P.J., Thompson, D.R., Delord, K., Weimerskirch, H., Sagar, P.M. ... & Phillips,
946 R.A. (2015). Poor transferability of species distribution models for a pelagic predator, the grey
947 petrel, indicates contrasting habitat preferences across ocean basins. *PLoS One*, 10(3), e0120014.
948

949 [van Proosdij](#), A.S., Sosef, M.S., Wieringa, J.J. & Raes, N. (2016). Minimum required number of
950 specimen records to develop accurate species distribution models. *Ecography*, 39(6), 542-552.
951

952 [Walsh](#), E. & Hudiburg, T.W. (2018). A Framework for Forest Landscape and Habitat Suitability
953 Model Integration to Evaluate Forest Ecosystem Response to Climate Change. *AGUFM, 2018*,
954 GC11G-0989.
955

956 [Williams](#), J.W. & Jackson, S.T. (2007). Novel climates, no-analog communities, and ecological
957 surprises. *Frontiers in Ecology and the Environment*, 5(9), 475-482.
958

959 [Williams](#), J.W., Jackson, S.T. & Kutzbach, J.E. (2007). Projected distributions of novel and
960 disappearing climates by 2100 AD. *Proceedings of the National Academy of Sciences*, 104(14),
961 5738-5742.
962

963 [Wisz](#), M.S. & Guisan, A. (2009). Do pseudo-absence selection strategies influence species
964 distribution models and their predictions? An information-theoretic approach based on simulated
965 data. *BMC Ecology*, 9(1), 8.
966

967 [WoRMS](#) Editorial Board (2016) World Register of Marine Species. <http://www.marinespecies.org>
968 [Accessed: 2016-05-23]
969

970 [Xavier](#), J.C., Raymond, B., Jones, D.C. & Griffiths, H. (2016). Biogeography of Cephalopods in the
971 Southern Ocean using habitat suitability prediction models. *Ecosystems*, 19, 220–247.
972

973 [Yates](#), K. L., Bouchet, P. J., Caley, M. J., Mengersen, K., Randin, C. F., Parnell, S., ... & Dormann,
974 C. F. (2018). Outstanding challenges in the transferability of ecological models. *Trends in ecology*
975 *& evolution*, 33(10), 790-802.
976

977 [Zurell](#), D., Zimmermann, N. E., Gross, H., Baltensweiler, A., Sattler, T., & Wüest, R. O. (2020).
978 Testing species assemblage predictions from stacked and joint species distribution models.
979 *Journal of Biogeography*, 47(1), 101-113.
980