Methods for improving species distribution models in data-poor areas: example of sub-Antarctic benthic species on the Kerguelen Plateau.

Charlène Guillaumot, Alexis Martin, Marc Eléaume, Thomas Saucède

Methodological clues for improving species distribution models in data-poor areas: example of sub-Antarctic benthic species on the Kerguelen Plateau

Charlène Guillaumot[*1], *Alexis Martin*[2], *Marc Eléaume*[3], *Thomas Saucède*[4]

[1] Université Libre de Bruxelles, Laboratoire de Biologie Marine. Avenue F.D. Roosevelt, 50. CP 160/15 1150 Bruxelles, Belgique

[2] Muséum national d'Histoire naturelle, Département Milieux et Peuplements Aquatiques, UMR BOREA 7208, 57 rue Cuvier, F-75231 Paris Cedex 05, France

[3]Muséum national d'Histoire naturelle, Département Systématique et Évolution, UMR ISYEB 7205, 57 rue Cuvier, F-75231 Paris Cedex 05, France

[4]UMR 6282 Biogéosciences, Univ. Bourgogne Franche-Comté, CNRS, 6 bd Gabriel F-21000 Dijon, France

Corresponding author: *Charlène Guillaumot* (charleneguillaumot21@gmail.com)

ABSTRACT

Species distribution models (SDM) are essential tools for conservation biologists to evaluate the combined effects of environmental changes and direct human activities on natural habitats and develop relevant conservation plans. However, modeling species distribution in vast and remote regions is often challenging due to poor and heterogeneous datasets and questions the relevance of modeling procedures. In the past few years, there have been many methodological developments in SDM procedures using virtual species and broad datasets but few solutions have been proposed to deal with poor and heterogeneous datasets. In the present work, we address this methodological challenge by studying the performance of different modeling procedures based on four real species, presence-only data compiled from various oceanographic surveys on the Kerguelen Plateau (Southern Ocean). We followed a practical protocol to test for the reliability and the performance of models and to correct for the incompleteness of data, and for spatial and temporal sampling biases. Our results show that producing reliable species distribution models is feasible as long as the number and quality of available data allow testing and correcting for these biases. However, SDM could be

corrected for spatial and temporal heterogeneities in one species only, showing the need to consider all these potential biases when modeling the distribution of species. Finally, we show that model reliability and performance also depend on the interaction between the incompleteness of data and species niches, the distribution of narrow niche species being less sensitive to data gaps than wider niche species.

Keywords:

Species distribution modeling, model performance, historical datasets, Kerguelen Plateau, presence-only data

INTRODUCTION

Today, species distribution models (SDM) constitute essential tools for conservation biologists to understand species distribution patterns and their underpinning drivers (see Guillera-Arroita et al. 2015 for a review), assess the combined effects of environmental changes and direct human pressures (i.e. economic activities including tourism) on natural habitats (Gutt et al. 2012), define conservation priorities (Vierod et al. 2014, Greathead et al. 2014), and develop relevant management plans (Reiss et al. 2014, Koubbi et al. 2016). SDM allow to interpolate the known distribution of single species, assemblages or communities (Ferrier and Guisan 2006) to little-accessed or under-sampled areas (Reiss et al. 2011, Robinson et al. 2011) and help improve our knowledge of the distribution of rare species (McCune 2016).

In regions subject to fast environmental changes and significant anthropogenic activities, SDM can provide useful tools for conservation purposes (Guisan et al. 2013, Reiss et al. 2014). However, modeling species distribution over vast and remote areas is challenging

55 and questions the relevance of the method compared to more traditional and qualitative
56 approaches (Koubbi et al. 2016). In such regions, our knowledge of species distribution
57 usually relies on historical and heterogeneous presence-only datasets, which concentrate
58 many gaps and can induce methodological biases altering the level of SDM performance
59 (Loiselle et al. 2008, Costa et al. 2010, Newbold 2010). The use of historical data in SDM has
60 been widely discussed (Reutter et al. 2003, Hortal et al. 2007, 2008), for instance with regards
61 to the spatial and temporal heterogeneities induced by the practice of different sampling
62 strategies. Limitations to SDM performance are mainly due to uncertainties in data location
63 and detection (Costa et al. 2010, Naimi et al. 2014, Tessarolo et al. 2014), to over-estimations
64 of habitat suitability in intensively sampled areas (Guillera-Arroita et al. 2015), and to
65 artefacts in niche descriptions (Hortal et al. 2008). The lack of available data from remote
66 areas also constitutes a limitation to SDM, which are restricted to presence-only data, and are
67 regarded as less reliable and less efficient than presence-absence and abundance-based
68 models (Brotons et al. 2004). Over the past few years, many methodological developments in
69 SDM procedures have been produced to correct for such biases (Dormann 2007, Phillips et al.
70 2009, Barbet-Massin et al. 2012) but no single procedure emerged (Qiao et al. 2015) and few
71 practical solutions have been proposed to deal with poor and heterogeneous datasets.
72 Our knowledge of Southern Ocean species distribution remains patchy (Koubbi et al.
73 2016). Therefore, the growing interest of marine biologists and biogeographers for the region
74 has led to the conception of collaborative projects compiling past and present marine
75 biodiversity data in information networks like the SCAR-Marine Biodiversity Information
76 Network (SCAR-MarBIN) (Griffiths et al. 2011), the Biogeographic Atlas of the Southern
77 Ocean (De Broyer et al. 2014) and other open access databases (Danis et al. 2013, Gutt et al.
78 2013, Van de Putte et al. 2014). However, running species distribution models in the region
79 still requires a significant effort of data compilation (Guillaumot et al. 2016) to complement

80 the existing open access data sources, and check for data quality. In addition, modeling
81 Southern Ocean species distribution poses auxiliary problems due to the paucity of data and
82 model performances that can vary with ecological niche width (Qiao et al. 2015). Recent
83 works have developed methodologies to adapt SDM to rare species and poorly-sampled areas
84 but none was tested for the Southern Ocean (Pokharel et al. 2016, Phillips et al. 2017).
85
86 In this work, we analysed the reliability of modeling procedures with regards to the
87 heterogeneous nature of data available and the gaps in our knowledge of species distribution.
88 We compiled echinoid presence-only data collected from several ancient and recent
89 oceanographic campaigns carried out on the Kerguelen Plateau (sub-Antarctic region) for one
90 and a half century. The distribution of four echinoid species with contrasting ecological
91 niches was modeled, and the reliability and the performance of modeling procedures were
92 tested. We propose methodological clues to correct for spatial and temporal biases and assess
93 the sensitivity of modeling procedures to species ecological niche width. This is the first
94 methodological approach to correct for potential biases in SDM in the Southern Ocean. Our
95 objective is to offer useful perspectives for future modeling works along with a practical and
96 transferable protocol to test for the reliability and performance of modeling procedures.
97

98 # MATERIAL AND METHODS

99 **Biological data**

100 Species occurrence data were taken from Guillaumot et al. (2016) and Pierrat et al. (2012).
101 The dataset includes presence-only data of echinoid species collected during scientific cruises
102 carried out on the Kerguelen Plateau (63°/81°E; -46°/-56°S) since 1872 (Fig. 1). Scientific
103 objectives, dates, sampling efforts, gears, and surveyed areas have differed between cruises,
104 leading to spatial and temporal heterogeneities (Guillaumot et al. 2016). In the dataset, four

105 echinoid species with contrasting ecological preferences and a high number of presence-only
106 records were selected. Species include two sediment feeders of the family Schizasteridae, one
107 shallow water species, *Abatus cordatus*, and a deeper one, *Brisaster antarcticus,* one
108 carnivorous/detritivorous and eurybathic species of the family Cidaridae, *Ctenocidaris nutrix*,
109 and one omnivorous and eurybathic species of Echinidae, *Sterechinus diadema* (David et al.
110 2005) (Fig. 1). *A. cordatus* is a coastal species that is endemic to the Kerguelen Plateau, *B.*
111 *antarcticus* is known in the Kerguelen and Crozet archipelagoes and has broader
112 environmental preferences than *A. cordatus* (Fig. 1). *C. nutrix* and *S. diadema* are widespread
113 in the Southern Ocean and have contrasting environmental preferences (Fig. 1).
114
115 **Environmental descriptors**
116 Environmental descriptors were taken from Guillaumot et al. (2016). The dataset covers the
117 geographic extent of the Kerguelen Plateau (63°/81°E and -46°/-56°S) and compiles
118 environmental data for six decades included in [1955-2012]. Environmental data are available
119 at a grid cell resolution of 10km precision. Environmental layers include no data pixels,
120 particularly in seafloor related descriptors. Data were not interpolated to avoid the potential
121 biases due to interpolation procedures.
122 Collinearity between descriptors can alter modeling performances (Phillips et al. 2006)
123 because collinear data may (1) inflate standard errors, (2) induce the violation of residual
124 independency during model validation and (3) generate noise that can be interpreted as a link
125 between descriptors (Dormann et al. 2013). To reduce the collinearity effect, we computed the
126 Variance Inflation Factor (VIF) and the Spearman correlation coefficient (rs) between all
127 available descriptors from Guillaumot et al. (2016). VIF analysis was performed using a
128 stepwise procedure, using the *vifstep* function proposed in the 'usdm' R package (Naimi et al.
129 2014). Descriptor pairs with high VIF and rs values were omitted based on the commonly

used thresholds of VIF < 5 and rs < 0.85 (Pierrat et al. 2012, Dormann et al. 2013, Duque-Lazo et al. 2016). Environmental descriptors finally selected to model species distribution are displayed in Table 1.

Environmental changes were tested between 1955 and 2012. The comparison of pixel values between periods was generated using a Wilcoxon signed-rank test with the Bonferroni correction.

Analytical procedures

The flow chart of Figure 2 details the analytical procedure used in the present work.

Model selection

Due to the growing interest of ecologists for species distribution modeling, a large range of modeling techniques is now available (Reiss et al. 2011, Guillera-Arroita et al. 2015, Qiao et al. 2015). Running the most appropriate model involves selecting the best modeling technique for the data under analysis and also involves considering the scientific objectives to be addressed (Reiss et al. 2011, Qiao et al. 2015).

Here we compared several modeling techniques using the 'biomod2' R3.3.0 library (Thuiller et al. 2016) and we tested the performance of these approaches with regards to the chronological addition of new data and to the transferability of models between areas. Several models were generated with an increasing number of occurrence data (Fig. S1). The best modeling techniques were then compared with each other using a non-random cross-validation procedure (Fig. S2, Wengen and Olden 2012) in order to determine the approach

154 with the best accuracy in transferability performances (Randin et al. 2006, Wengen and Olden
155 2012).
156 Results show high performance and stability values for Random Forest (RF) and Boosted
157 Regression Trees (BRT) in our case study (Appendix 1). However, BRT performed better in
158 transferability in comparison with RF (Heikkinen et al. 2012), and previous works showed
159 that RF does not deal correctly with missing values and patchy datasets (Breiman 2001,
160 Barbet-Massin et al. 2012, Qiao et al. 2015, see Table S1 for a review). Therefore, BRT was
161 chosen in the present work to generate the analyses.
162 BRT calibration was realised using the 'gbm' R package (Ridgeway 2015, Elith et al. 2008).
163 The three main parameters (learning rate lr, tree complexity tc and bag fraction bg) were
164 selected using the method developed by Elith et al. (2008) to figure out the combination of
165 values minimizing the predicted deviance of the models (Elith and Leathwick 2014). The
166 parameters were finally set at respectively lr= 0.0001, tc=2 and bf= 0.75.
167 Following Barbet-Massin et al. (2012), we sampled the same number of background data as
168 the number of presence data available for computing BRT models. Considering the low
169 number of presence data available, 100 model replicates (i.e. background sampling) were
170 generated for each analysis. Finally, to correct for data aggregation in space, presence
171 duplicates were removed when present in a same 10km resolution pixel.
172
173 Model performance was assessed by measuring AUC score values (Area Under the Curve of
174 the Receiver Operating Curve) of each model replicate using the 'dismo' R library (Hijmans
175 et al. 2016). AUC expresses the relationship between model sensitivity and the commission
176 error (1-specificity). The sensitivity corresponds to the number of « presence » pixels
177 correctly predicted as present and the specificity the number of « absence » pixels correctly
178 predicted as absent (Fielding and Bell 1997). The use of the AUC to evaluate SDM

179 performance has been repeatedly discussed (Lobo et al. 2007, Peterson et al. 2008) but the
180 AUC remains the most appropriate metric for presence-background models as values stay
181 stable with low-prevalence datasets and are not sensitive to threshold effects (Hand 2009,
182 Proosdij et al. 2016). Following the recommendation of Jimenez-Valverde et al. (2012), we
183 used the AUC to estimate the robustness of models but not for direct comparisons between
184 models that were generated for different species, on different studied areas and with different
185 training samples.
186
187 **Correcting for sampling bias**
188 Data collected during the various scientific cruises led over the Kerguelen Plateau for the last
189 145 years present conspicuous spatial heterogeneities. The resulting bias can generate an
190 unequal number of records in the different sectors of the study area and heterogeneous
191 patterns in record distribution. Such heterogeneities might increase the risk of over-estimating
192 the contribution of environmental conditions to the models in the most sampled areas (Araújo
193 and Guisan 2006).
194
195 The effect of spatial heterogeneities on the quality of distribution models was tested using a
196 null model approach. A first null model, null model #1, was generated by sampling presence
197 data at random within the total set of sites that were visited during the different campaigns,
198 whether echinoid specimens were collected at these sites or not (Fig. 3). Because absence data
199 are not available, this approach allows us to assess the weight of sampling bias in the models.
200 If a sampling bias is significant, null model #1 is expected to produce distribution maps with
201 higher suitability values in the most sampled areas (Merckx et al. 2011).
202

A second null model, null model #2, was built by simulating presence data sampled at random over the entire studied area. Null model #2 is expected to produce distribution maps of equal suitability over the entire study area. If sampling is spatially biased, we expect that null model #1 deviates from null model #2 (Raes and ter Steege 2007).

The two null models were generated for the four selected species. The number of presence-only data used in the models was contained between the number of data collected until the MD04 campaign and until the PROTEKER campaign, between 1974 and 2015, which corresponds to periods of high sampling effort (Fig. 1). In each null model, 100 replicates were produced. Time-averaged environmental descriptors [1955-2012] were used for the analysis.

To correct for sampling bias when null models #1 and #2 significantly differ between each other, we used the methodology proposed by Phillips et al. (2009), which has been shown to improve modeling performances (Phillips et al. 2009, Aguirre-Gutiérrez et al. 2013). A grid layer was built using a kernel density estimation (KDE) to represent sampling spatial bias. The layer was calculated from the map of visited sites. The estimated proportion of presence-only data present in each pixel was determined using the 'kde2d' function of 'MASS' R package (Venables and Ripley 2002). Background data were sampled according to the weighting scheme of the KDE layer, to reduce discrepancies between presence-only records and background data (Phillips et al. 2009, Barbet-Massin et al. 2012). In order to test for the efficiency of model correction based on the KDE, the Pearson r correlation was computed between pixel values of the KDE layer (a proxy for the sampling effort) and the predicted probabilities of models, before and after the KDE correction.

228 Spatial heterogeneities in data collection can also generate spatial autocorrelation (SAC)
229 between presence records, which can violate model calibration assumptions, and affect model
230 accuracy with wrong parameter estimations (Segurado et al. 2006, Dormann 2007, Crase et al.
231 2012). Several approaches have been developed to take into account SAC in SDM (Crase et
232 al. 2012 for a review). They consist in including an additional term in models (the auto-
233 covariate), which represents the influence of neighboring records on modeling predictions.
234 The significance of SAC was tested using the Moran I autocorrelation index computed on
235 model residuals (Luoto et al. 2005, Crase et al. 2012) for both original and corrected models.
236 Models were built using time-averaged environmental descriptors [1955-2012].
237
238 **Testing for the effect of the chronological addition of new records on model**
239 **performance**
240 Our dataset compiles presence-only data collected during various scientific cruises and with
241 distinct sampling protocols, which may alter the performance of the models (Fig. 1). To test
242 for model reliability, we separately analysed (1) the influence of the chronological addition of
243 presence records, (2) the influence of data number alone and (3) the influence of sampling
244 patterns (distribution of data in space). The analyses were performed for *A. cordatus*, *C.*
245 *nutrix* and *S. diadema*. Not enough data were available for *B. antarcticus*. We used time-
246 averaged environmental descriptors [1955-2012] to generate the models.
247
248 To test for the potential effect of the chronological addition of new data on model
249 performance, we followed the protocol proposed by Aguiar et al. (2015). The dataset was split
250 into distinct subsets corresponding to main periods of sampling effort (1975: including
251 Marion Dufresne campaigns, 1993: including ANARE campaigns, 2010: including POKER II
252 campaign, 2015: including PROTEKER campaigns). New presence data were progressively

added to the models, following the chronological collection of new records. The influence of the chronological addition of data was assessed by measuring the correlation between models using the Schoener's D statistic. The Schoener's D is a correlation metric adapted to the study of niche similarities (Warren et al. 2008, Rödder and Engler 2011). It evaluates the similarity of pixel values between two distribution grids. A D value of 0 means that the two maps are perfectly different, a D value of 1 means that maps are perfectly similar. Values were computed using the *niche.overlap* function of the 'ENMeval' R package (Muscarella et al. 2014).

The significance of correlations was tested following a null model protocol, using 100 replicates, pairwise-compared using the Schoener's D statistic (Raes and ter Steege 2007, Warren et al. 2008, Ficetola et al. 2009).

The distinct effect of data addition and sampling patterns were tested separately. To test for the effect of data addition alone, models were built by sampling an increasing number of presence data at random in the total area for *A. cordatus* [n=54, 76, 95], *C. nutrix* [n=46, 54, 106, 114] and *S. diadema* [n=54, 66, 98]. These thresholds correspond to the number of presence-only data used in the chronological addition analysis.

Finally, to test for the effect of sampling patterns, different models were produced by sampling presence data at random either within a subset of real data collected along transects (MD03 campaign) or within a subset of real data collected at random (POKER, PROTEKER campaigns). All models were compared between each other.

Testing for the effect of temporal variations on model performance

277 To test for the effect of environmental shifts on models, different distribution models were
278 generated using distinct environmental descriptors for four periods ([1955-1964]; [1965-
279 1974]; [1975-1994]; [2005-2012]) and the complete set of presence data available.
280 Similarities between models were measured using the Schoener's D statistic.
281

282 # RESULTS

283 **Environmental shifts**

284 Mean sea surface temperature and amplitude, mean seafloor temperature and amplitude, mean
285 sea surface salinity and amplitude were all tested significantly different between all the
286 studied decades ($p < 0.001$). Seafloor temperature amplitude only was not proved
287 significantly different between decades [2005-2012] and [1955-1964]. These results indicate
288 that significant environmental shifts occurred during the studied time period and may induce
289 important variations in models as the dataset extends over 145 years.
290

291 **Spatial bias**

292 Null model #1 predicts higher suitability values in areas with most intense sampling effort
293 corresponding to the northern part of the Kerguelen Plateau and the vicinity of the Kerguelen
294 archipelago (Fig. 3A). In contrast, null model #2 predicts equally medium suitability values
295 over the entire Kerguelen Plateau because presence data were sampled at random in the area
296 (Fig. 3B). The difference between null models #1 and #2 was tested significant for the four
297 species (Fig. 3) showing that sampling bias has a significant impact on model outputs, which
298 will over estimate environment suitability in areas with the highest number of sampling sites
299 if no correction is applied.
300 Correlation between visited areas and predicted probability distribution decreases in models
301 built with the KDE-correction compared to non-corrected models (Table 2), showing that the

correction is efficient to reduce the influence of sampling bias on modeling performances. However, the correction was proved less efficient in models of the coastal and narrow niche species *A. cordatus* for which correlation values after the KDE correction remain high (r=0.44) (Table 2).

Spatial autocorrelation (SAC) was tested significant for non-corrected models (Moran index, I_{min}=0.05, I_{max}=0.16) but values were not significant in corrected models (I_{min}=0.04, I_{max}=0.06), except for *A. cordatus* (Table S2, Figure S3). This shows that the KDE procedure also corrected for SAC for three of the four studied species.

Chronological addition of new records

The different models built with a chronological addition of new data show high AUC values for *C. nutrix* and *A. cordatus* (0.814±0.018 < $AUC_{C.nutrix}$ < 0.883±0.024 and 0.908±0.023 < $AUC_{A.cordatus}$ < 0.909±0.018 respectively) demonstrating the relevance of all models (Fig. 4, Fig. S4). For these two species, Schoener's D correlation values are high ($\overline{D}_{A.cordatus}$= 0.978±0.023, $\overline{D}_{C.nutrix}$=0.968±0.020) and were tested significant, showing that models are similar to each other.

In contrast, models generated for *S. diadema* significantly differ between each other with lower Schoener's D statistics ($\overline{D}_{S.diadema}$ = 0.932±0.036) (Fig. S5). Therefore, the chronological addition of new data has contrasting impacts on model outputs according to the studied species, which may be explained by a various sensitivity of models to data addition and to sampling patterns.

Data addition and sampling patterns

Comparison of models produced with an increasing number of data presents high and significant Schoener's D values (*minimum*=0.979 ± 0.031 for *S. diadema*, *maximum*=0.985 ±

0.020 for *C. nutrix*), showing that model outputs do not vary significantly with increasing data in our case study (Table 3).

To test for the influence of sampling patterns, models built using subsets with contrasting distribution patterns (radial versus random patterns) were compared. Schoener's D statistics measured between these two types of models present low values. This suggests a significant influence of sampling patterns on model outputs (Table 3).

Environmental change and model performance

The different models generated with contrasting environmental descriptors are highly similar as shown by high Schoener's D and low standard deviation values ($\overline{D}$= 0.981±0.005). This proves that environmental shifts have no significant impact on model outputs. In addition, the respective contributions of environmental descriptors to models do not vary significantly between periods for the four species. However, *A. cordatus* seems to be less impacted by environmental shifts in comparison with other species (Fig. 5).

Finally, the contribution of time-averaged environmental descriptors over the total studied period [1955-2012] tends to differ from contributions computed for each decade separately (Fig. 5).

Final species distribution models

Sampling bias analyses and model corrections show that reliable distribution models can be built for *C. nutrix* only. It is the only dataset in which spatial and temporal heterogeneities do not impact prediction performances significantly. A final, reliable model was produced for *C. nutrix* over the Kerguelen Plateau (Fig. 6).

DISCUSSION

Data scarcity and heterogeneity

First research surveys of the Kerguelen Plateau date back to the oceanographic campaign of the HMS Challenger in 1872. One and a half century later, our knowledge of benthic species distribution on the Kerguelen Plateau has significantly increased but remains patchy (Koubbi et al. 2016). As in most parts of the Southern Ocean, modeling species distribution on the Kerguelen Plateau faces significant limitations due to data gaps and heterogeneities (Guillaumot et al. 2016). Such limitations seriously question the relevance of modeling procedures, which are required by environmental managers for conservation purposes (Féral et al. 2016, Koubbi et al. 2016). In the present work, we follow a step by step protocol to assess, quantify, and correct the potential effects of data scarcity and heterogeneity on models, a critical issue when considering the growing interest for modeling approaches in Antarctic and Sub-Antarctic regions (Gutt et al. 2012). Our results demonstrate that such approaches can prove feasible and reliable in certain case studies, when data quality and sampling bias can be tested and corrected.

Coping with spatial and temporal bias in presence-only datasets

Spatial bias and spatial autocorrelation (SAC)

Building SDM for remote and little-accessed regions often implies the use of spatially biased datasets conditioned by sampling caveats. Because parts of these regions that are the most easily accessed aggregate most of the available presence data, more weight is given to the most sampled sites and model performance is reduced (Phillips et al. 2009). In the present work, a significant difference was measured between the two null models (that were generated by selecting presence data either from visited stations only or at random over the

total investigated area), highlighting the strong heterogeneity of sampling effort with more data collected in the northern part of the Kerguelen Plateau and in coastal, shallow areas.

The significant spatial autocorrelation (SAC) values that were computed from model residuals also reveal the impact of sampling bias. The significance of SAC on uncorrected model residuals can be partly explained by the relative accumulation and high density of presence data in shallow areas of the Kerguelen Plateau where species presence probability is over-predicted. One could argue that SAC analysis does not apply to SDM as species presence proximities must be considered in the environmental niche space, not in the geography. However, in the present study, the difference between null models constitutes an operational evidence of the impact of sample clumping on model outputs, which is also revealed by significant SAC values.

To correct for sampling bias, we used a background-based correction method (Phillips et al. 2009) that was already used in former studies based on presence-only and limited datasets (Mateo et al. 2010, Pokharel et al. 2016, Phillips et al. 2017). These methods allow to reduce the effect of sample spatial bias on modeling performance by weighting background records according to sampling patterns. In the present study, the correction was proved efficient to correct both for the influence of the uneven sampling effort on predicted distributions (Table 2) and for SAC on all SDM except for models of *A. cordatus*. *A. cordatus* is a coastal, shallow marine species that was mainly sampled in the northern part of the Kerguelen Plateau. Species presence records are strongly conditioned by the location of most important sampling efforts. This is in line with previous studies that already highlighted the difficulties of modeling the distribution of narrow niche species with low prevalence distribution (*i.e*, corresponding to the proportion of the area where presence records are located) (Barbet Massin et al. 2012, Qiao et al. 2015). In small presence-only datasets, the methodologies used to correct for spatial bias are not as efficient for narrow niche species as

401 for broader niche species. Reducing the extent of distribution modeling of narrow niche
402 species to the boundaries of their environmental limits could prove a good alternative.
403
404 *Influence of record addition*
405 The chronological addition of new data has a limited impact on certain model outputs as
406 demonstrated by high similarities between the chronological models generated for *A. cordatus*
407 and *C. nutrix*. In contrast, chronological models of *S. diadema* significantly differ between
408 each other. The detailed analysis of data increment proved that the increasing number of
409 presences has no impact on modeling performance, which is not in line with previous works
410 (Stockwell and Peterson 2002, Wisz et al. 2008). However, these results can be altered by our
411 incomplete knowledge of species full distribution due to the limited number of data available
412 and to sampling bias (Hernandez et al. 2006, Bean et al. 2012). In models of *S. diadema*,
413 differences between the chronological models are due to contrasted spatial patterns between
414 datasets (transect versus random patterns).
415
416 *Historical data and environmental change*
417 Significant environmental shifts were measured for the descriptors analysed between 1955
418 and 2012 over the Kerguelen Plateau (*i.e.* mean sea surface temperature and amplitude, mean
419 surface salinity and amplitude). However, for all species, distribution models built for each
420 decade are highly similar between each other. These results confirm that temporal
421 heterogeneities in datasets do not necessarily impact the robustness of models, because
422 species preferences for their environment may be wider than the magnitude of changes in
423 time. Working with both present and historical data to improve the completeness of
424 occurrence records proved reliable when assuming that species niche and distribution have
425 not significantly changed during the studied time period.

Between the five decades, the respective contributions of temperature and salinity to the models did not vary over the range of within-decade variation for *B. antarcticus*, *C. nutrix* and *S. diadema*. Variations between decades are more marked in models produced for *A. cordatus*. This near-shore species is found in shallow waters of Kerguelen and Heard islands, where environmental descriptors include many no data pixels (Guillaumot et al. 2016). Consequently, the varying contributions of temperature and salinity to the models of *A. cordatus* between decades cannot be attributed with certainty to the effect of environmental change but to modeling limitations.

Sea surface temperature and salinity amplitudes have significant contributions to the models, contributing more than averaged parameters (i.e. *A. cordatus* and *B. antarcticus*, Fig. 5). This is in line with the results of Bradie and Leung (2016) who tested for the contribution of several environmental descriptors on a wide panel of taxa. They showed the importance of including seasonal means and extremes in models to further depict species distribution, considering their stronger relationships with species niche width and ecological traits (*i.e.* growth and survival, see Franklin 2009).

Using time-averaged descriptors over the entire period [1955-2012] could have been considered the best approach to produce representative models, independent from short-term environmental variations. Unexpectedly, our results show that for all species, contributions of time-averaged descriptors to the models are much more different than all differences between decadal descriptors (Fig. 5). This shows that using time-averaged descriptors for long time periods does not necessarily improve model reliability in comparison with using descriptors averaged for a shorter time period. This also shows the importance of the descriptor selection

in modeling procedures, a critical issue for improving model performance as already stressed in previous studies (Bradie and Leung 2016). This is particularly relevant for certain regions of the Southern Ocean like the Western Antarctic Peninsula that has experienced among the most significant environmental changes in the world ocean during the last decades (Turner 2015).

Influence of species niche width in modeling performances

Among the four studied species, *A. cordatus* has the narrowest ecological niche and most restricted distribution in the vicinity of coastal areas of the Kerguelen and Heard archipelagoes. Such limited geographic and environmental distributions compared to the total extent of the studied area implies that similar environmental conditions prevail in geographically close occurrence sites. This induces a strong SAC pattern that explains the difficulties to correct for spatial bias in comparison to other species models. Moreover, the limited environmental variability between coastal sampling sites of the different oceanographic surveys can also explain the absence of data addition effect on modeling performances for *A. cordatus*.

In contrast, *C. nutrix* and *S. diadema* have wider ecological niches than *A. cordatus* (Fig.1). For these two species, record data are more widely distributed and show contrasting sampling patterns (*i.e.* transect-like versus random patterns) that have been shown to influence modeling performance in *S. diadema* only (Table 3). This can be explained by the higher number of presence records available for *C. nutrix* (n=114 and n=98 for *C. nutrix* and *S. diadema* respectively) that allowed a more complete survey of *C. nutrix* distribution. Finally, *C. nutrix* dataset only presents a quality and number of occurrence records that fulfill all methodological requirements to produce reliable distribution models.

Considering species niche width to cope with spatial and temporal bias in SDM is important, as already shown by Tessarolo et al. (2014) who studied the influence of survey designs on the performance of distribution models for endemic species with narrow ecological niches. They concluded that survey designs have a low impact on models in comparison with the effect of niche width, data number, and type of modeling technique used. However, they did not generate any analysis of species with broad ecological niche as a comparison. Our results are also in line with other modeling studies in which distribution models of species with broad niche were the least stable (Reiss et al. 2011, Qiao et al. 2015, Guo et al. 2015, Ranc et al. 2016).

Conclusion

The use of SDMs has gained in importance during the last decades. They can provide complementary information for environmental managers. Modeling results can help interpolate species distribution, identify the potential drivers of species distribution and predict the potential effects of environmental changes on habitat suitability. However, modeling species distribution over vast and remote marine areas like the Southern Ocean using poor and heterogeneous datasets remains challenging and improvement of biological and environmental datasets is still required.

In the present study, we showed that reliable species distribution models can be produced in such areas as long as the number and quality of data allow testing and correcting for the effects of biases. Using historical data requires proper environmental descriptors for modeling the effect of environmental changes on species distribution. Using time-averaged predictors over long time periods can generate unrealistic models.

501 Model selection is also crucial at this stage and the statistical performance of models is
502 not the only criteria to be considered. Modeling procedures must be chosen with regards to
503 the scientific issues that are addressed. Two procedures (BRT and RF) performed best in our
504 case study, but one of them (BRT) was proved to be more relevant because it better dealt with
505 transferability and data patchiness.
506 Modeling species distribution in data-poor areas poses the practical problem of the
507 minimum number of presence-only data required to run reliable models, although this is not
508 the only and most critical issue. Beforehand, the number of occurrence records must be high
509 enough for testing model robustness and reliability. In regions with limited access, sampling
510 effort may be heterogeneous, which influences model performance. We showed that sampling
511 bias can be corrected, but the efficiency of correction depends on species niche width, narrow
512 niche species models being more troublesome to correct. In our study, *A. cordatus* is a species
513 limited to coastal shallow areas, which implies a strong correlation between species
514 occurrence and sampling patterns. Restricting the model to a more reduced area could allow
515 to correct for spatial bias and improve modeling performances.
516 There is also a crucial need for improving the quality of datasets (Kennicutt et al.
517 2014) and running more accurate models to better tackle conservation issues (Guisan et al.
518 2013, Rodríguez et al. 2007). For the time being, producing uncertainty maps can be an
519 alternative (Rocchini et al. 2011, Tessarolo et al. 2014) and can provide additional
520 information to environmental managers and stakeholders (Addison et al. 2013, Guisan et al.
521 2013).
522 Model reliability and performance also rely on the interaction between dataset
523 completeness and species intrinsic ecological properties. Hence, we showed that the type and
524 width of ecological niches were important to consider, distribution of narrow niche species
525 being easier to model and less sensitive to incomplete datasets (Guo et al. 2015, Ranc et al.

2016). However, narrow niches usually imply that species are distributed over small areas for which distribution models will be highly sensitive to extrapolations.

Our protocol shows that reliable SDMs can be produced when enough data are available and dataset bias can be tested and corrected. In the present study, one SDM only (*C. nutrix*) could be corrected for spatial and temporal heterogeneities to generate reliable distribution predictions. However, our results stress the need to consider methodological issues when modeling species distribution based on poor and spatially biased datasets. They should contribute to bring new insights and enhance modeling performance in future studies.

Acknowledgements

This work is a contribution to the IPEV program PROTEKER funded by the French polar institute (IPEV program n°1044) and contribution no. 21 to the vERSO project (www.versoproject.be) funded by the Belgian Science Policy Office (BELSPO, contract n°BR/132/A1/vERSO).

References

Addison PF, Rumpff L, Bau SS, Carey JM and others (2013) Practical solutions for making models indispensable in conservation decision-making. Divers. Distrib. 19: 490–502.

Aguiar LMDS, Rosa ROL, Jones G and Machado RB (2015) Effect of chronological addition of records to species distribution maps: The case of *Tonatia saurophila maresi* (Chiroptera, Phyllostomidae) in South America. Austral Ecol. 40: 836–844.

549 Aguirre-Gutiérrez J, Carvalheiro LG, Polce C, van Loon EE and others (2013) Fit-for-
550 purpose: species distribution model performance depends on evaluation criteria–Dutch
551 hoverflies as a case study. PloS one 8: e63708.
552
553 Araújo MB and Guisan A (2006) Five (or so) challenges for species distribution modelling. J.
554 Biogeogr. 33: 1677–1688.
555
556 Barbet-Massin M, Jiguet F, Albert CH and Thuiller W (2012) Selecting pseudo-absences for
557 species distribution models: how, where and how many? Method. Ecol. Evol. 3: 327–338.
558
559 Bradie J and Leung B (2016) A quantitative synthesis of the importance of variables used in
560 MaxEnt species distribution models. J. Biogeogr. 44: 1344-1361.
561
562 Breiman L (2001) Random forests. Machine learning. 45: 5–32.
563
564 Brotons L, Thuiller W, Araújo MB and Hirzel AH (2004). Presence-absence versus presence-
565 only modelling methods for predicting bird habitat suitability. Ecography. *27*(4): 437-448.
566
567 Costa GC, Nogueira C, Machado RB and Colli GR (2010) Sampling bias and the use of
568 ecological niche modeling in conservation planning: a field evaluation in a biodiversity
569 hotspot. Biodiv. Conserv. 19: 883–899.
570
571 Crase B, Liedloff AC and Wintle BA (2012) A new method for dealing with residual spatial
572 autocorrelation in species distribution models. Ecography. 35: 879–888.
573

574 Danis B, Van de Putte A, Renaudier S and Griffiths H (2013) Connecting biodiversity data
575 during the IPY: the path towards e-polar science. Adapt. Evol. Mar. Env. Springer Berlin
576 Heidelberg. 2: 21–32.
577
578 David B, Choné T, Mooi R, De Ridder C (2005) Antarctic Echinoidea – Synopses of the
579 Antarctic benthos. 10: 273.
580
581 De Broyer C, Koubbi P, Griffiths H, Grant A and others (2014) Biogeographic Atlas of the
582 Southern Ocean. Scientific Committee on Antarctic Research, Cambridge. 498pp.
583
584 Dormann CF (2007) Effects of incorporating spatial autocorrelation into the analysis of
585 species distribution data. Glob. Ecol. Biogeogr. 16:129–138.
586
587 Dormann CF, Elith J, Bacher S, Buchmann C and others (2013) Collinearity: a review of
588 methods to deal with it and a simulation study evaluating their performance. Ecography. 36:
589 27–46.
590
591 Douglass LL, Turner J, Grantham HS, Kaiser S, Constable A and others (2014) A hierarchical
592 classification of benthic biodiversity and assessment of protected areas in the Southern Ocean.
593 PloS one. 9: e100551.
594
595 Duque-Lazo J, Van Gils HAMJ, Groen TA and Navarro-Cerrillo RM (2016) Transferability
596 of species distribution models: The case of *Phytophthora cinnamomi* in Southwest Spain and
597 Southwest Australia. Ecol. Model. 320: 62–70.
598

599 Elith J, Leathwick JR and Hastie T (2008) A working guide to boosted regression trees. J.
600 Anim. Ecol. 77: 802–813.
601
602 Elith J and Leathwick JR (2009) Species distribution models: ecological explanation and
603 prediction across space and time. Annu. Rev. Ecol. Evol. Syst. 40: 677–697.
604
605 Elith J and Leathwick J (2014). Boosted Regression Trees for ecological modeling. At
606 http://www.lcis.com.tw/paper_store/paper_store/brt(5)-2015115131033846.pdf , accessed
607 02/2016.
608
609 Féral JP et al. (2016) PROTEKER: implementation of a submarine observatory at the
610 Kerguelen Islands (Southern Ocean) – Underwater Technol. 3: 3–10.
611
612 Ferrier S and Guisan A (2006) Spatial modelling of biodiversity at the community level. J.
613 Appl. Ecol. 43: 393–404.
614
615 Ficetola GF, Thuiller W and Padoa-Schioppa E (2009) From introduction to the establishment
616 of alien species: bioclimatic differences between presence and reproduction localities in the
617 slider turtle. Divers. Distrib. 15: 108–116.
618
619 Fielding AH and Bell JF (1997) A review of methods for the assessment of prediction errors
620 in conservation presence/absence models. Environ. Conserv. 24: 38–49.
621
622 Franklin J (2009) Mapping species distributions: spatial inference and prediction. University
623 Press, Cambridge: 320pp.

624

625 Greathead C, González-Irusta JM, Clarke J, Boulcott P and others (2014) Environmental
626 requirements for three sea pen species: relevance to distribution and conservation. ICES J.
627 Mar. Sci. 72: 576–586.

628

629 Griffiths HJ, Danis B and Clarke A (2011) Quantifying Antarctic marine biodiversity: The
630 SCAR-MarBIN data portal. Deep Sea Res. Pt II. 58: 18–29.

631

632 Guillaumot C, Martin A, Fabri-Ruiz S, Eléaume M and Saucède T (2016) Echinoids of the
633 Kerguelen Plateau: occurrence data and environmental setting for past, present, and future
634 species distribution modelling. ZooKeys. 630: 1–17.

635

636 Guillera-Arroita G, Lahoz-Monfort JJ, Elith J, Gordon A and others (2015) Is my species
637 distribution model fit for purpose? Matching data and models to applications. Glob. Ecol.
638 Biogeogr. 24: 276–292.

639

640 Guisan A, Tingley R, Baumgartner JB, Naujokaitis-Lewis I and others (2013) Predicting
641 species distributions for conservation decisions. Ecol. Lett. 16: 1424–1435.

642

643 Guo C, Lek S, Ye S, Li W, Liu J and Li Z (2015) Uncertainty in ensemble modelling of large-
644 scale species distribution: Effects from species characteristics and model techniques. Ecol.
645 Model. 306: 67–75.

646

647 Gutt J, Zurell D, Bracegridle T, Cheung W and others (2012) Correlative and dynamic species
648 distribution modelling for ecological predictions in the Antarctic: a cross-disciplinary
649 concept. Polar Res. 31:11091.
650
651 Gutt J, Barnes D, Lockhart SJ and van de Putte A. (2013) Antarctic macrobenthic
652 communities: A compilation of circumpolar information. Nat. Conserv. 4: 1–13.
653
654 Hand D.J. (2009). Measuring classifier performance: a coherent alternative to the area under
655 the ROC curve. Machine learning. *77*(1): 103-123.
656
657 Heikkinen RK, Marmion M and Luoto M (2012) Does the interpolation accuracy of species
658 distribution models come at the expense of transferability? Ecography. 35: 276–288.
659
660 Hernandez PA, Graham CH, Master LL and Albert DL (2006). The effect of sample size and
661 species characteristics on performance of different species distribution modeling methods.
662 Ecography. 29(5): 773–785.
663
664 Hijmans RJ, Phillips S, Leathwick J and Elith J (2016) dismo: Species Distribution Modeling.
665 R package version 1.1-1. https://CRAN.R-project.org/package=dismo.
666
667 Hortal J, Lobo JM and Jimenez-Valverde A (2007) Limitations of biodiversity databases: case
668 study on seed-plant diversity in Tenerife, Canary Islands. Conserv. Biol. 21: 853–863.
669

670 Hortal J, Jiménez-Valverde A, Gómez JF, Lobo JM and Baselga A (2008) Historical bias in
671 biodiversity inventories affects the observed environmental niche of the species. Oikos. 117:
672 847–858.
673
674 Jackson AL, Inger R, Parnell AC and Bearhop S (2011) Comparing isotopic niche widths
675 among and within communities: SIBER–Stable Isotope Bayesian Ellipses in R. J. Ann. Ecol.
676 80: 595–602.
677
678 Jiménez-Valverde A (2012) Insights into the area under the receiver operating characteristic
679 curve (AUC) as a discrimination measure in species distribution modelling. Glob. Ecol.
680 Biogeogr. 21: 498–507.
681
682 Kennicutt MC, Chown SL, Cassano JJ, Liggett D and others (2014) Six priorities for
683 Antarctic science. Nature. 512: 23–25.
684
685 Koubbi P, Mignard C, Causse R, Da Silva O and others (2016) Ecoregionalisation of the
686 Kerguelen and Crozet islands oceanic zone Part I: Introduction and Kerguelen oceanic zone.
687 Workshop Paris. 06/6-9/2016.
688
689 Lobo JM, Jiménez-Valverde A and Real R (2007). AUC: a misleading measure of the
690 performance of predictive distribution models. Glob. Ecol. Biogeog. 17(2): 145-151.
691
692 Loiselle BA, Jørgensen PM, Consiglio T, Jiménez I and others (2008) Predicting species
693 distributions from herbarium collections: does climate bias in collection sampling influence
694 model outcomes? J. Biogeogr. 35: 105–116.

695

696 Luoto M, Pöyry J, Heikkinen RK and Saarinen K (2005) Uncertainty of bioclimate envelope
697 models based on the geographical distribution of species. Glob. Ecol. Biog. 14: 575–584.

698

699 Mateo RG, Croat TB, Felicisimo AM and Munoz J. (2010). Profile or group discriminative
700 techniques? Generating reliable species distribution models using pseudo-absences and target-
701 group absences from natural history collections – Divers. Distrib. 16:84–94.

702

703 McCoy FW (1991) Southern Ocean sediments: circum-Antarctic to 30°S. Marine Geological
704 and Geophysical Atlas of the circum-Antarctic to 30°S. (ed. by D.E. Hayes) – Antarctic
705 Research Series.

706

707 McCune JL (2016) Species distribution models predict rare species occurrences despite
708 significant effects of landscape context – J. Appl. Ecol. 53: 1871–1879.

709

710 Merckx B, Steyaert M, Vanreusel A, Vincx M and Vanaverbeke J (2011) Null models reveal
711 preferential sampling, spatial autocorrelation and overfitting in habitat suitability modelling.
712 Ecol. Model. 222: 588–597.

713

714 Muscarella R, Galante PJ, Soley-Guardia M, Boria RA and others (2014) ENMeval: An R
715 package for conducting spatially independent evaluations and estimating optimal model
716 complexity for ecological niche models. Method. Ecol. Evol. 5: 1198–1205.

717

718 Naimi B, Hamm NA, Groen TA, Skidmore AK and Toxopeus AG (2014) Where is positional
719 uncertainty a problem for species distribution modelling? Ecography. 37: 191–203.

720

721 Newbold T (2010) Applications and limitations of museum data for conservation and
722 ecology, with particular attention to species distribution models. Prog. Phys. Geogr. 34: 3–22.
723
724 Paradis E, Claude J and Strimmer K (2004) APE: analyses of phylogenetics and evolution in
725 R language. Bioinformatics. 20: 289–290.
726
727 Peterson AT, Papeş M and Soberón J (2008). Rethinking receiver operating characteristic
728 analysis applications in ecological niche modeling. Ecol. Model. 213(1): 63–72.
729
730 Phillips SJ, Anderson RP and Schapire RE (2006) Maximum entropy modeling of species
731 geographic distributions. Ecol. Model. 190: 231259.
732
733 Phillips SJ, Dudík M, Elith J, Graham CH and others (2009) Sample selection bias and
734 presence-only distribution models: implications for background and pseudo-absence data.
735 Ecol. Appl. 19: 181–197.
736
737 Phillips ND, Reid N, Thys T, Harrod C, Payne N and others (2017) Applying species
738 distribution modelling to a data poor, pelagic fish complex: the ocean sunfishes. J. Biogeogr.
739
740 Pierrat B, Saucède T, Laffont R, De Ridder C and others (2012) Large-scale distribution
741 analysis of Antarctic echinoids using ecological niche modelling. Mar. Ecol. Prog. Ser. 463:
742 215–230.harel KP, Ludwig T and Storch I (2016). Predicting potential distribution of poorly
743 known species with small database: the case of four horned antelope *Tetracerus quadricornis*
744 on the Indian subcontinent. Ecol. Evol. 6: 2297-2307.
745

746 Proosdij A.S., Sosef M.S., Wieringa J.J. and Raes N. (2016). Minimum required number of
747 specimen records to develop accurate species distribution models. Ecography. *39*(6): 542–
748 552.
749
750 Qiao H, Soberón J and Peterson AT (2015) No silver bullets in correlative ecological niche
751 modelling: insights from testing among many potential algorithms for niche estimation.
752 Method. Ecol. Evol. 6: 1126–1136.
753
754 Raes N and ter Steege H (2007) A null-model for significance testing of presence-only
755 species distribution models. Ecography. 30: 727–736.
756
757 Ranc N, Santini L, Rondinini C, Boitani L and others (2016) Performance tradeoffs in target-
758 group bias correction for species distribution models. Ecography. 39: 1–12.
759
760 Randin CF, Dirnböck T, Dullinger S, Zimmermann NE and others (2006) Are niche-based
761 species distribution models transferable in space? J. Biogeogr. 33: 1689–1703.
762
763 Reiss H, Cunze S, König K, Neumann H and Kröncke I (2011) Species distribution modelling
764 of marine benthos: a North Sea case study. Mar. Ecol. Prog. Ser. 442: 71–86.
765
766 Reiss H, Birchenough S, Borja A, Buhl-Mortensen L and others (2014) Benthos distribution
767 modelling and its relevance for marine ecosystem management. ICES J. Mar. Sci. 72: 297–
768 315.
769

770 Reutter BA, Helfer V, Hirzel AH and Vogel P (2003) Modelling habitat suitability using
771 museum collections: an example with three sympatric *Apodemus* species from the Alps. J.
772 Biogeogr. 30: 581–590.
773
774 Ridgeway G (2015) gbm: Generalized Boosted Regression Models. R package version 2.1.1.
775 https://CRAN.R-project.org/package=gbm.
776
777 Robinson LM, Elith J, Hobday AJ, Pearson RG and others (2011) Pushing the limits in
778 marine species distribution modelling: lessons from the land present challenges and
779 opportunities. Glob. Ecol. Biogeogr. 20: 789–802.
780
781 Rocchini D, Hortal J, Lengyel S, Lobo JM and others (2011) Accounting for uncertainty
782 when mapping species distributions: the need for maps of ignorance. Prog. Phys. Geogr. 35:
783 211–226.
784
785 Rödder D and Engler JO (2011) Quantitative metrics of overlaps in Grinnellian niches:
786 advances and possible drawbacks. Glob. Ecol. Biogeogr. 20: 915–927.
787
788 Rodríguez JP, Brotons L, Bustamante J and Seoane J (2007) The application of predictive
789 modelling of species distribution to biodiversity conservation. Divers. Distrib. 13: 243–251.
790
791 Segurado PA, Araújo MB and Kunin WE (2006) Consequences of spatial autocorrelation for
792 niche-based models. J. Appl. Ecol. 43: 433–444.
793

794 Smith WH and Sandwell DT (1997) Global sea floor topography from satellite altimetry and
795 ship depth soundings. Science. 277: 1956–1962.
796
797 Stockwell DR and Peterson AT (2002) Effects of sample size on accuracy of species
798 distribution models. Ecol. Model. 148: 1–13.
799
800 Tessarolo G, Rangel TF, Araújo MB and Hortal J (2014) Uncertainty associated with survey
801 design in Species Distribution Models. Divers. Distrib. 20: 1258–1269.
802
803 Thuiller W, Georges D, Engler R, Breiner F and others (2016) biomod2: Ensemble Platform
804 for Species Distribution Modeling. R package version 3.3-7. https://CRAN.R-
805 project.org/package=biomod2.
806
807 Van de Putte A. et al. (2014) Chapter 2.1. Data and Mapping. In De Broyer C., Koubbi P.,
808 Griffiths H.J., Raymond B., Udekem d'Acoz C. d', et al. (eds.). Biogeographic Atlas of the
809 Southern Ocean. Scientific Committee on Antarctic Research, Cambridge: 14–15.
810
811 Venables WN and Ripley BD (2002) Modern Applied Statistics with S. Fourth Edition.
812 Springer, New York. ISBN 0-387-95457-0.
813
814 Vierod AD, Guinotte JM and Davies AJ (2014) Predicting the distribution of vulnerable
815 marine ecosystems in the deep sea using presence-background models. Deep Sea Res. II. 99:
816 6–18.
817

Warren DL, Glor RE and Turelli M (2008) Environmental niche equivalency versus conservatism: quantitative approaches to niche evolution. Evolution. 62: 2868–2883.

Wenger SJ and Olden JD (2012) Assessing transferability of ecological models: an underappreciated aspect of statistical validation. Method. Ecol. Evol. 3: 260–267.

Wisz MS, Hijmans RJ, Li J, Peterson AT, Graham CH and Guisan A (2008) Effects of sample size on the performance of species distribution models. Divers. Distrib. 14: 763–773.

Zucchetta M, Venier C, Taji MA, Mangin A and Pastres R (2016) Modelling the spatial distribution of the seagrass *Posidonia oceanica* along the North African coast: Implications for the assessment of Good Environmental Status. Ecol. Indic. 61: 1011–1023.

Figures legend list

Figure 1: (A) Map showing occurrence data of the four studied echinoid species over the Kerguelen Plateau: *Brisaster antarcticus* (orange diamond), *Ctenocidaris nutrix* (grey circle), *Sterechinus diadema* (green triangle) and *Abatus cordatus* (purple square). (B) Evolution of sampling effort (in presence-only records) through time with the position of main scientific cruises during which the four studied species were collected on the Kerguelen Plateau. (C) Species presence data plotted according to depth (z axis), seafloor salinity (y axis) and seafloor temperature (x axis) on the extent of the Kerguelen Plateau with projection of standardized distribution ellipsoids (see Jackson et al. 2011 for details) on bivariate plots. (D) Species depth range over the Kerguelen Plateau based on occurrence data (solid line: median, box: upper and lower quartiles, whiskers: 75% ± 1.5 interquartile range, dots: outliers).

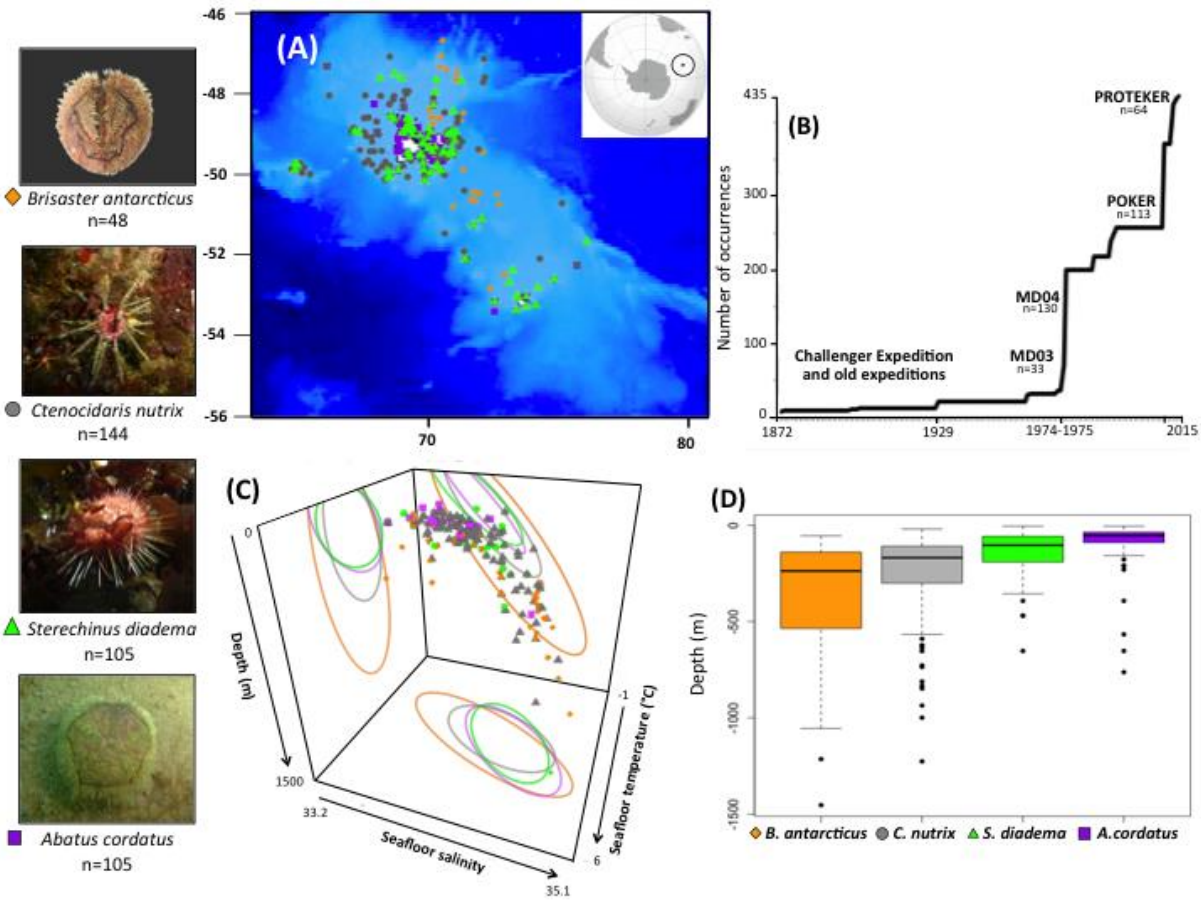
Figure 2. Tests and procedures carried out in the present work. Arrows indicate the stepwise procedure with statistical validation conducting either to the following step or correction/stepback requirements.

Figure 3: (A) Null model #1 and (B) Null model #2 for the different species under study. Mean AUC value and standard deviation are given for the 100 replicates. Comparisons between models compiled with Pearson r correlation value and associated probabilities.

Figure 4: First row: Distribution models of *Ctenocidaris nutrix* species with increasing number of presence data to build the model, for four periods. Averaged maps of 100 model replicates. Second row: (A) Difference in probability distribution between n=54 and n=46, (B) between n=106 and n=54, (C) between n=114 and n=106.

Figure 5: Mean contributions of environmental descriptors to the models with standard deviation (error bars) for the four time periods and species under study. sst= sea surface temperature, sst amp= sea surface temperature amplitude, sssalinity= sea surface salinity, sst amp= sea surface salinity amplitude, chl a= chlorophyll-a (see Guillaumot et al. 2016 for details).

Figure 6: Species distribution model generated for *Ctenocidaris nutrix* using all presence-only data available and present environmental descriptors (2005-2012). AUC=0.813±0.02



884

885 **Figure 1:** (A) Map showing occurrence data of the four studied echinoid species on the
886 Kerguelen Plateau: *Brisaster antarcticus* (orange diamond), *Ctenocidaris nutrix* (grey circle),
887 *Sterechinus diadema* (green triangle) and *Abatus cordatus* (purple square). (B) Evolution of
888 sampling effort (in presence-only records) through time with the position of main scientific
889 cruises during which the four studied species were collected on the Kerguelen Plateau. (C)
890 Species presence data plotted according to depth (z axis), seafloor salinity (y axis) and
891 seafloor temperature (x axis) on the extent of the Kerguelen Plateau with projection of
892 standardized distribution ellipsoids (see Jackson et al. 2011 for details) on bivariate plots. (D)
893 Species depth range over the Kerguelen Plateau based on occurrence data (solid line: median,
894 box: upper and lower quartiles, whiskers: 75% ± 1.5 interquartile range, dots: outliers).

Table 1: List of environmental descriptors selected for SDM. Asterisks (*) indicate that environmental layers are available for the following time periods: [1955-2012], [1955-1964], [1965-1974], [1975-1994], [2005-2012]. Minimum and maximum values are indicated for [1955-2012]. Spatial resolution of layers: 10km resolution grid-cell pixels.

Environmental descriptors	Unit	Description	Min value	Max value	Source	Spatial extent
Depth	Meters	Bathymetric grid around the Kerguelen Plateau	-4977.0000	-1.0000	This study. Derived from the Biogeographic Atlas of the Southern Ocean (De Broyer et al. 2014)	46_56°S/63_81°E
Sea surface mean temperature*	°Celsius degrees	Mean sea surface temperature	3.0566	7.6223	World Ocean Circulation Experiment 2013	46_56°S/63_81°E
Sea surface temperature amplitude*	°Celsius degrees	Amplitude between mean summer and mean winter sea surface temperature	-3.3036	-1.4108	World Ocean Circulation Experiment 2013	46_56°S/63_81°E
Seafloor mean temperature*	°Celsius degrees	Mean seafloor temperature	-0.2978	4.6422	This study. Derived from World Ocean Circulation Experiment 2013 sea surface temperature layers	46_56°S/63_81°E
Seafloor temperature amplitude*	°Celsius degrees	Amplitude between mean summer and mean winter seafloor temperature	-2.5757	0.8867	This study. Derived from World Ocean Circulation Experiment 2013 sea surface temperature layers	46_56°S/63_81°E
Sea surface mean salinity*	PSS	Mean sea surface salinity	33.6849	33.8251	World Ocean Circulation Experiment 2013	46_56°S/63_81°E
Sea surface salinity amplitude*	PSS	Amplitude between mean summer and mean winter sea surface salinity	-0.0859	0.3165	World Ocean Circulation Experiment 2013	46_56°S/63_81°E
Seafloor salinity amplitude*	PSS	Amplitude between mean summer and mean winter seafloor salinity	- 169	0.0937	This study. Derived from World Ocean Circulation Experiment 2013 sea surface salinity layers	46_56°S/63_81°E
Mean surface chlorophyll a	mg/m^3	Surface chlorophyll a concentration. Summer mean over 2002-2009	0.1358	2.7324	MODIS AQUA (NASA) 2010	46_56°S/63_81°E
Sediments	Categorical	Sediment features	14 categories		McCoy (1991), updated by Griffiths 2014 (unpublished)	46_56°S/63_81°E
Geomorphology	Categorical	Geomorphologic features	27 categories		ATLAS ETOPO2 2014 (Douglass et al. 2014)	46_56°S/63_81°E
Slope	Degrees	Bathymetric slope	$4.8229.\ 10^{-5}$	0.1547	Biogeographic Atlas of the Southern Ocean (De Broyer et al. 2014)	46_56°S/63_81°E
Mean seafloor oxygen concentration	mL/L	Mean seafloor oxygen concentration over 1955-2012	4.0080	7.6223	This study. Derived from World Ocean Circulation Experiment 2013 sea surface oxygen concentration layers	46_56°S/63_81°E

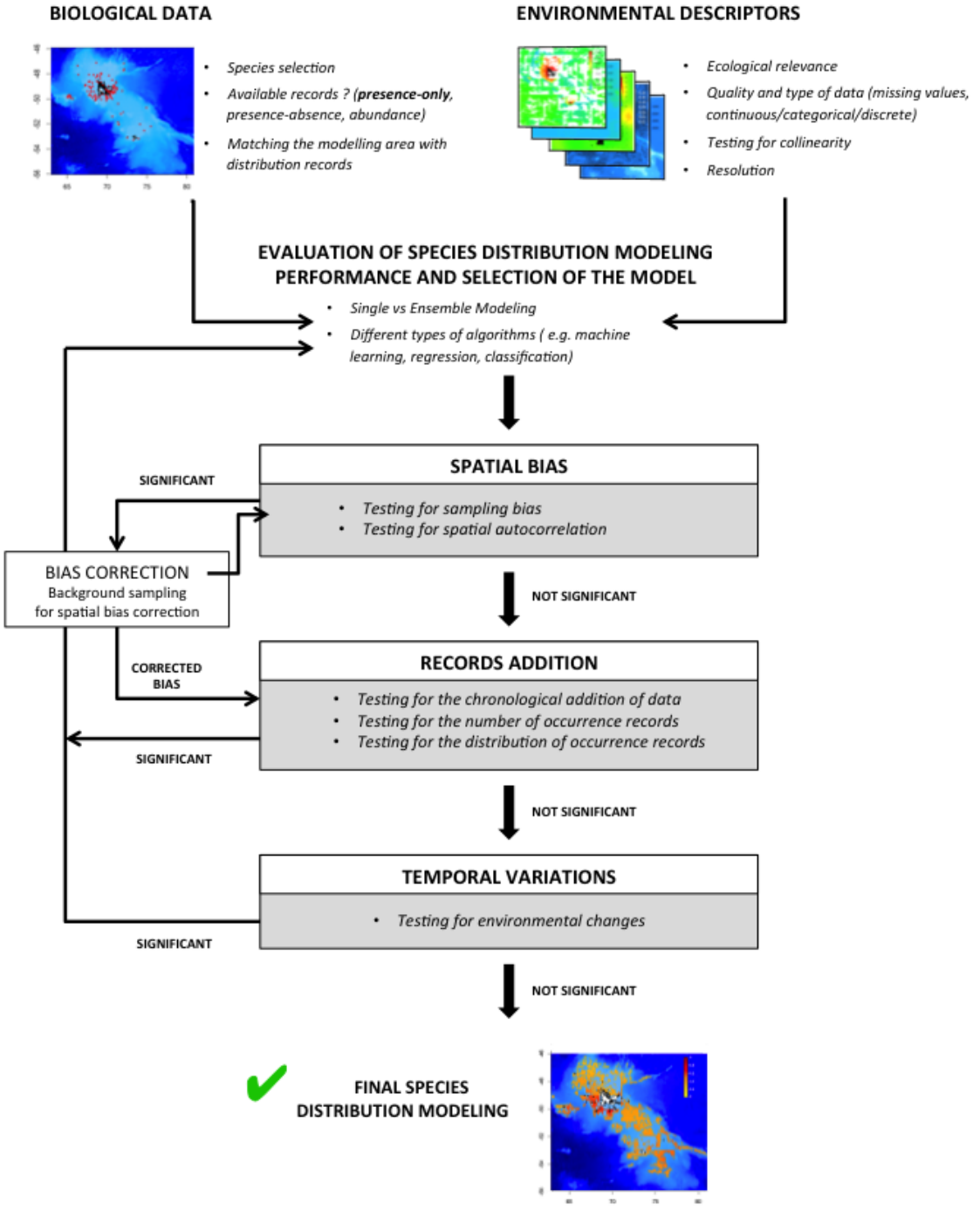


Figure 2. Tests and procedures carried out in the present work. Arrows indicate the stepwise procedure with statistical validation conducting either to the following step or correction/stepback requirements.

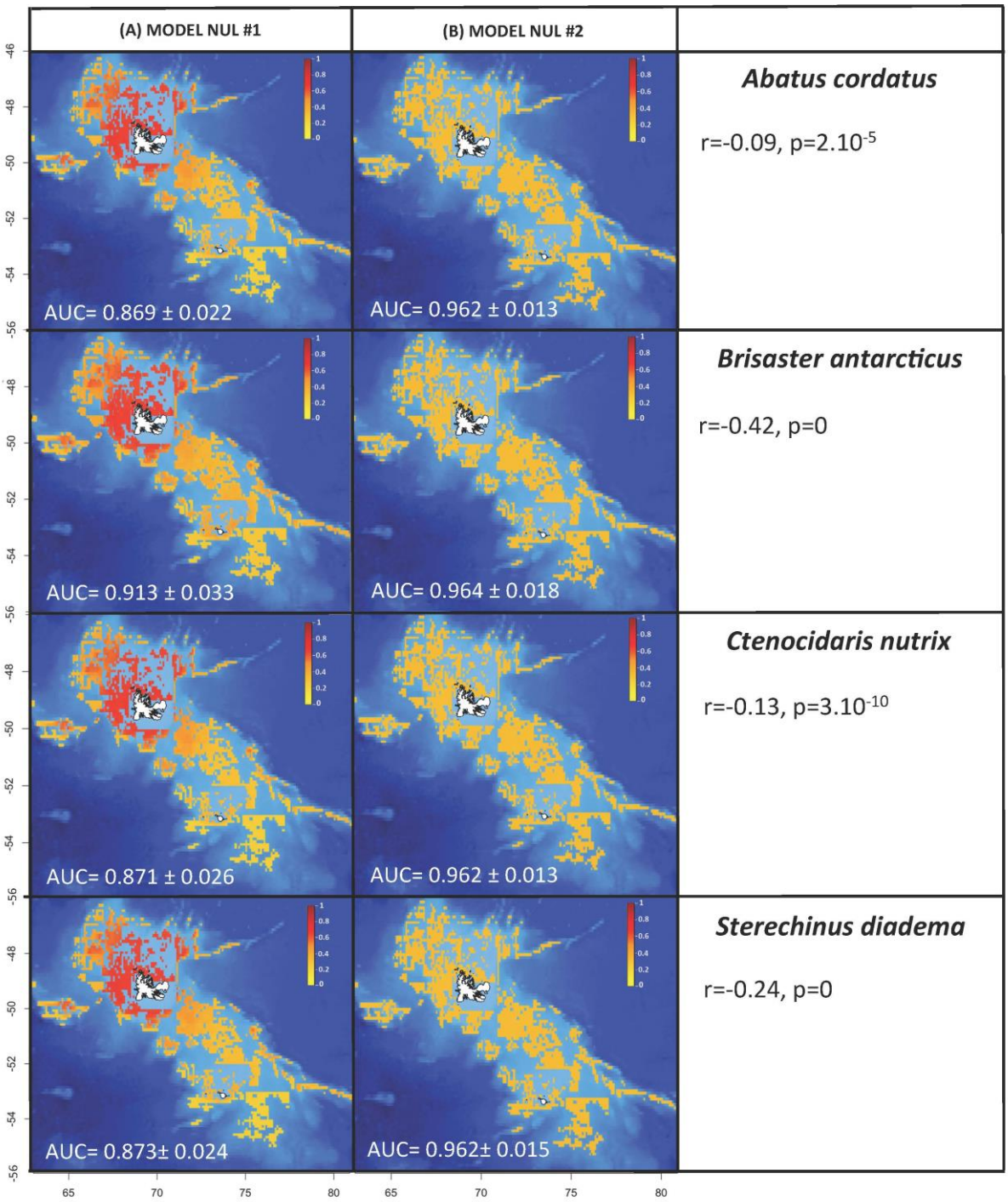


Figure 3: (A) Null model #1 and (B) Null model #2 for the different species under study. Mean AUC value and standard deviation are given for the 100 replicates. Comparisons between models compiled with Pearson r correlation value and associated probabilities.

Table 2: r Pearson correlation of pixel values between the KDE layer and the predicted probability of each species model. Statistic probabilities are all < 0.05.

	Before KDE correction	**After KDE correction**
Abatus cordatus	**0.72**	**0.44**
Brisaster antarcticus	**0.60**	**-0.17**
Ctenocidaris nutrix	**0.80**	**0.11**
Sterechinus diadema	**0.61**	**0.20**

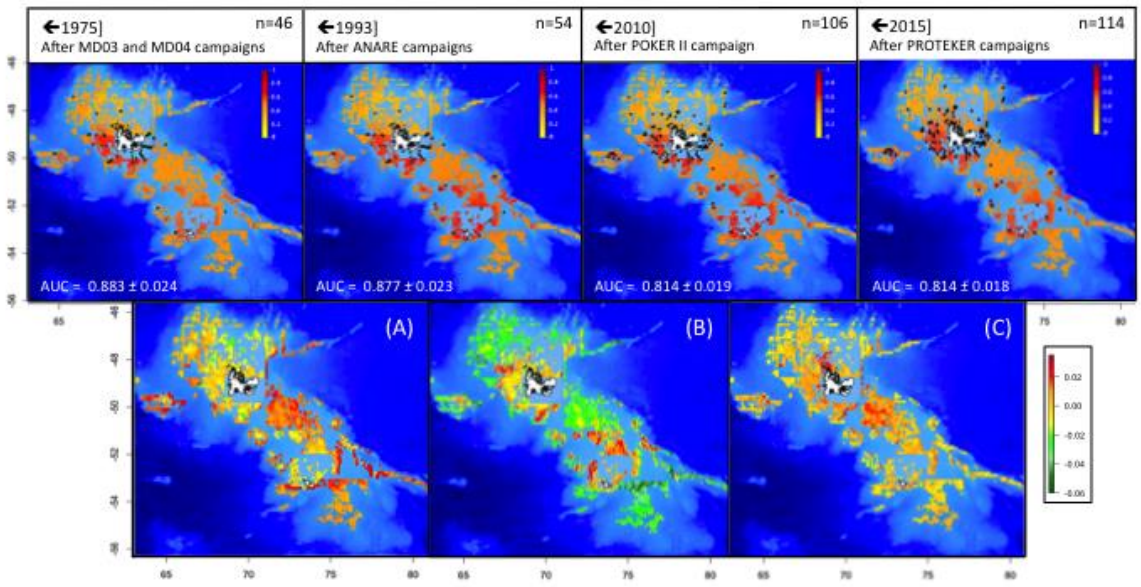


Figure 4: First row: Distribution models of *Ctenocidaris nutrix* species with increasing number of presence data to build the model, for four periods. Averaged maps of 100 model replicates. Second row: (A) Difference in probability distribution between n=54 and n=46, (B) between n=106 and n=54, (C) between n=114 and n=106.

Table 3: Influence of data addition and sampling patterns on models for *Abatus cordatus, Ctenocidaris nutrix* and *Sterechinus diadema.* Column 1: mean Schoener's D and associated *p*-value computed between models (100 replicates) produced respectively with {n=54, 76, 95}, {n=46, 54, 106, 114} and {n=54, 66, 98} occurrences randomly sampled from the total dataset. Column 2: mean Schoener's D and associated p-value computed between models (100 replicates) produced with subsets contrasting in data distribution patterns (transect versus random sampling).

Species	occurrence number		sampling pattern	
	Mean D_{obs}	Mean p-value	D_{obs}	p-value
Abatus cordatus	0.981 ± 0.025	<0.05	-	
Ctenocidaris nutrix	0.985 ± 0.020	<0.05	0.941 ± 0.030	0.147
Sterechinus diadema	0.979 ± 0.031	<0.05	0.842 ± 0.040	0.941

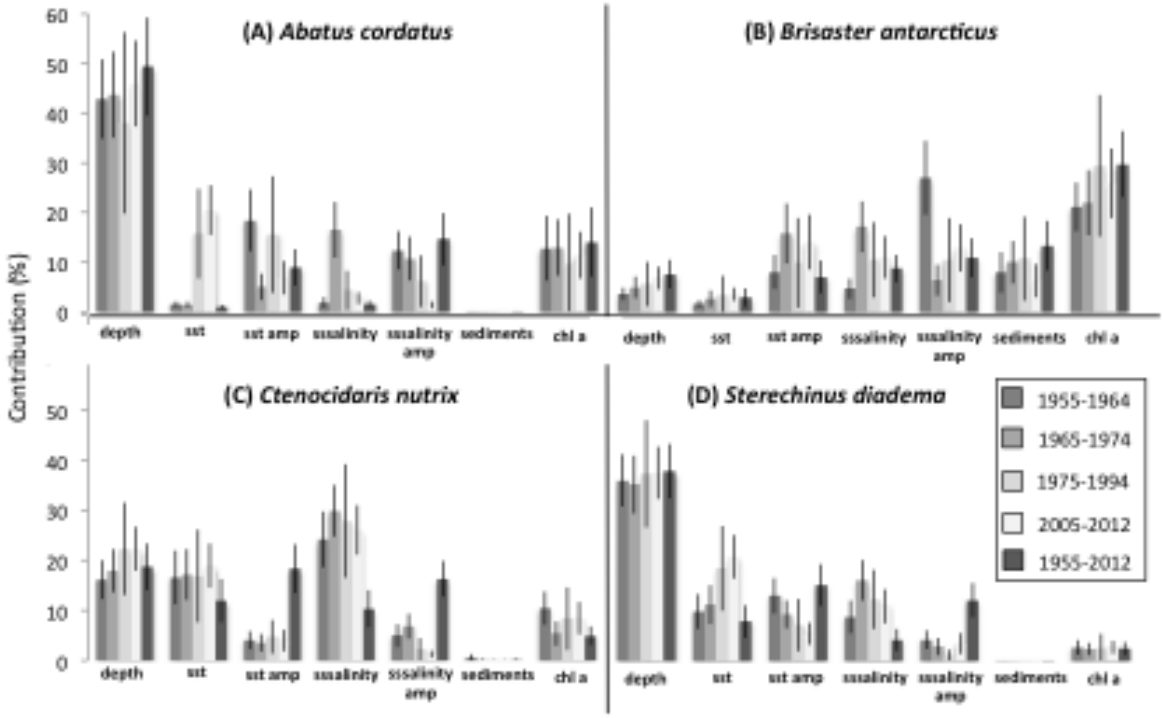


Figure 5: Mean contributions of environmental descriptors to the models with standard deviation (error bars) for the four time periods and species under study. sst= sea surface temperature, sst amp= sea surface temperature amplitude, sssalinity= sea surface salinity, sst amp= sea surface salinity amplitude, chl a= chlorophyll-a (see Guillaumot et al. 2016 for details).

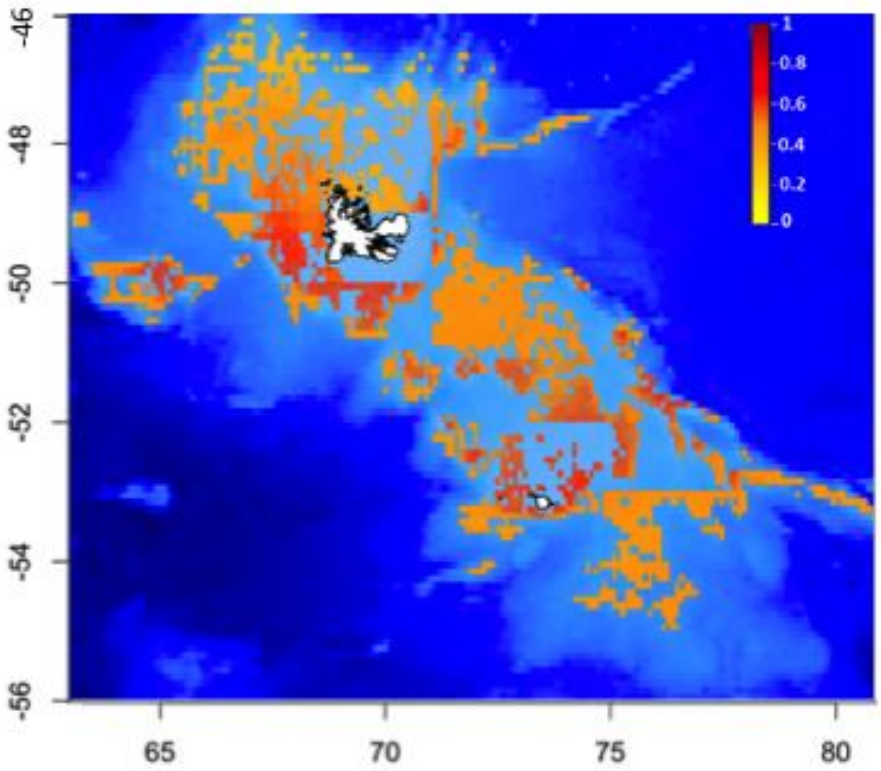


Figure 6: Species distribution model generated for *Ctenocidaris nutrix* using all presence-only data available and present environmental descriptors (2005-2012). AUC=0.813±0.02