Vol. 594: 149–164, 2018
https://doi.org/10.3354/meps12538

MARINE ECOLOGY PROGRESS SERIES
Mar Ecol Prog Ser

Published April 26

Methods 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]Laboratoire de Biologie Marine, Université Libre de Bruxelles, Avenue FD Roosevelt 50, CP 160/15, 1150 Bruxelles, Belgique
[2]Département adaptation du vivant, Muséum national d'Histoire naturelle, UMR BOREA 7208, 57 rue Cuvier, 75231 Paris Cedex 05, France
[3]Département Origine et Évolution, Muséum national d'Histoire naturelle, UMR ISYEB 7205, 57 rue Cuvier, 75231 Paris Cedex 05, France
[4]Biogéosciences, UMR 6282, Université Bourgogne Franche-Comté, CNRS, 6 bd Gabriel 21000 Dijon, France

ABSTRACT: Species distribution models (SDMs) are essential tools to aid conservation biologists in evaluating the combined effects of environmental change and human activities on natural habitats and for the development of relevant conservation plans. However, modeling species distributions over vast and remote regions is often challenging due to poor and heterogeneous data sets, and this raises questions regarding the relevance of the modeling procedures. In recent years, there have been many methodological developments in SDM procedures using virtual species and broad data sets, but few solutions have been proposed to deal with poor or heterogeneous data. In the present work, we address this methodological challenge by studying the performance of different modeling procedures based on 4 real species, using 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 performance of the models and to correct for limited and aggregated data, as well as accounting for spatial and temporal sampling biases. Our results show that producing reliable SDMs is feasible as long as the amount and quality of available data allow testing and correcting for these biases. However, we found that SDMs could be corrected for spatial and temporal heterogeneities in only 1 of the 4 species we examined, highlighting the need to consider all potential biases when modeling species distributions. Finally, we show that model reliability and performance also depend on the interaction between the incompleteness of the data and species niches, with the distribution of narrow-niche species being less sensitive to data gaps than species occupying wider niches.

KEY WORDS: Species distribution modeling · Model performance · Historical datasets · Kerguelen Plateau · Presence-only data

Resale or republication not permitted without written consent of the publisher

INTRODUCTION

Species distribution models (SDMs) are essential tools used by conservation biologists for understanding species distribution patterns and their drivers (see Guillera-Arroita et al. 2015 for a review), assessing the combined effects of environmental change and direct human pressure (i.e. economic activities including tourism) on natural habitats (Gutt et al. 2012), defining conservation priorities (Vierod et al. 2014, Greathead et al. 2015) and developing relevant management plans (Reiss et al. 2015, Koubbi et al. 2016). SDMs allow scientists to interpolate the known distribution of single species, assemblages or communities

*Corresponding author: charleneguillaumot21@gmail.com

© Inter-Research 2018 · www.int-res.com

(Ferrier & Guisan 2006) to little-accessed or undersampled 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 rapid environmental change and significant anthropogenic activities, SDMs can be useful tools in planning conservation measures (Guisan et al. 2013, Reiss et al. 2015). However, modeling species distributions over vast and remote areas is challenging and raises questions regarding the relevance of this method compared to more traditional and qualitative approaches (Koubbi et al. 2016). In such regions, our knowledge of species distributions is usually based on historical and heterogeneous presence-only data sets, which may include many gaps, and may induce methodological biases that affect the level of SDM performance (Loiselle et al. 2008, Costa et al. 2010, Newbold 2010). The use of historical data in SDMs has been widely discussed (Reutter et al. 2003, Hortal et al. 2007, 2008); for instance, regarding the spatial and temporal heterogeneities induced by the use of different sampling strategies. Limitations to SDM performance are mainly due to uncertainties in data location and detection (Costa et al. 2010, Naimi et al. 2014, Tessarolo et al. 2014), overestimations of habitat suitability in intensively sampled areas (Guillera-Arroita et al. 2015) and artefacts in niche descriptions (Hortal et al. 2008). The lack of available data from remote areas also constitutes a limitation to SDMs, which are restricted to presence-only data and are regarded as being less reliable and less efficient than presence–absence and abundance-based models (Brotons et al. 2004). Over the past few years, many methodological developments in SDM procedures have been produced to correct for such biases (Dormann 2007, Phillips et al. 2009, Barbet-Massin et al. 2012), but no single corrective procedure has emerged (Qiao et al. 2015) and few practical solutions have been proposed to deal with poor and heterogeneous data sets.

Our knowledge of species distribution in the Southern Ocean is still patchy (Koubbi et al. 2016). Therefore, the growing interest of marine biologists and biogeographers in the region has led to the conception of collaborative projects compiling past and present marine biodiversity data in information networks such as the SCAR-Marine Biodiversity Information Network (SCAR-MarBIN) (Griffiths et al. 2011), the Biogeographic Atlas of the Southern Ocean (De Broyer et al. 2014) and other open access databases (Danis et al. 2013, Gutt et al. 2013, Van de Putte et al. 2014). However, running SDMs in the region still requires a significant data compilation effort (Guillaumot et al. 2016) to complement the existing open access data sources and to check for data quality. In addition, modeling Southern Ocean species distributions poses auxiliary problems due to the paucity of data and model performances that can vary with ecological niche width (Qiao et al. 2015). Recent works have developed methodologies to adapt SDMs to rare species and poorly sampled areas, but none have been tested for the Southern Ocean (Pokharel et al. 2016, Phillips et al. 2017).

In this work, we analysed the reliability of modeling procedures with regards to the heterogeneous nature of data available and the gaps in our knowledge of species distributions. We compiled echinoid presence-only data collected from several ancient and recent oceanographic campaigns that have been carried out on the Kerguelen Plateau (sub-Antarctic region) over the past 145 yr. The distributions of 4 echinoid species with contrasting ecological niches were modeled and the reliability and performance of the modeling procedures were tested. We propose methodological procedures to correct for spatial and temporal biases and assess the sensitivity of modeling procedures to a species' ecological niche width. This is the first methodological approach to correct for potential biases in SDMs in the Southern Ocean. Our objective is to offer useful perspectives for future modeling, along with a practical and transferable protocol to test for the reliability and performance of modeling procedures.

MATERIALS AND METHODS

Biological data

Species occurrence data were taken from Guillaumot et al. (2016) and Pierrat et al. (2012). The data set includes presence-only data of echinoid species collected during 19 scientific cruises carried out on the Kerguelen Plateau (46 to 56° S, 63 to 81° E) since 1872 (Fig. 1). Fig. 1B illustrates the expeditions that mainly contributed to the dataset. The full list is available in Guillaumot et al. (2016). Scientific objectives, dates, sampling effort, gears and surveyed areas differed between cruises, leading to spatial and temporal heterogeneities (Guillaumot et al. 2016). From this data set, 4 echinoid species with contrasting ecological preferences and a high number of presence-only records were selected. Species included 2 sediment feeders of the family Schizasteridae (1 shallow water species, *Abatus cordatus*, and a deeper one, *Brisaster antarcticus*), 1 carnivorous/detritivorous and eury-

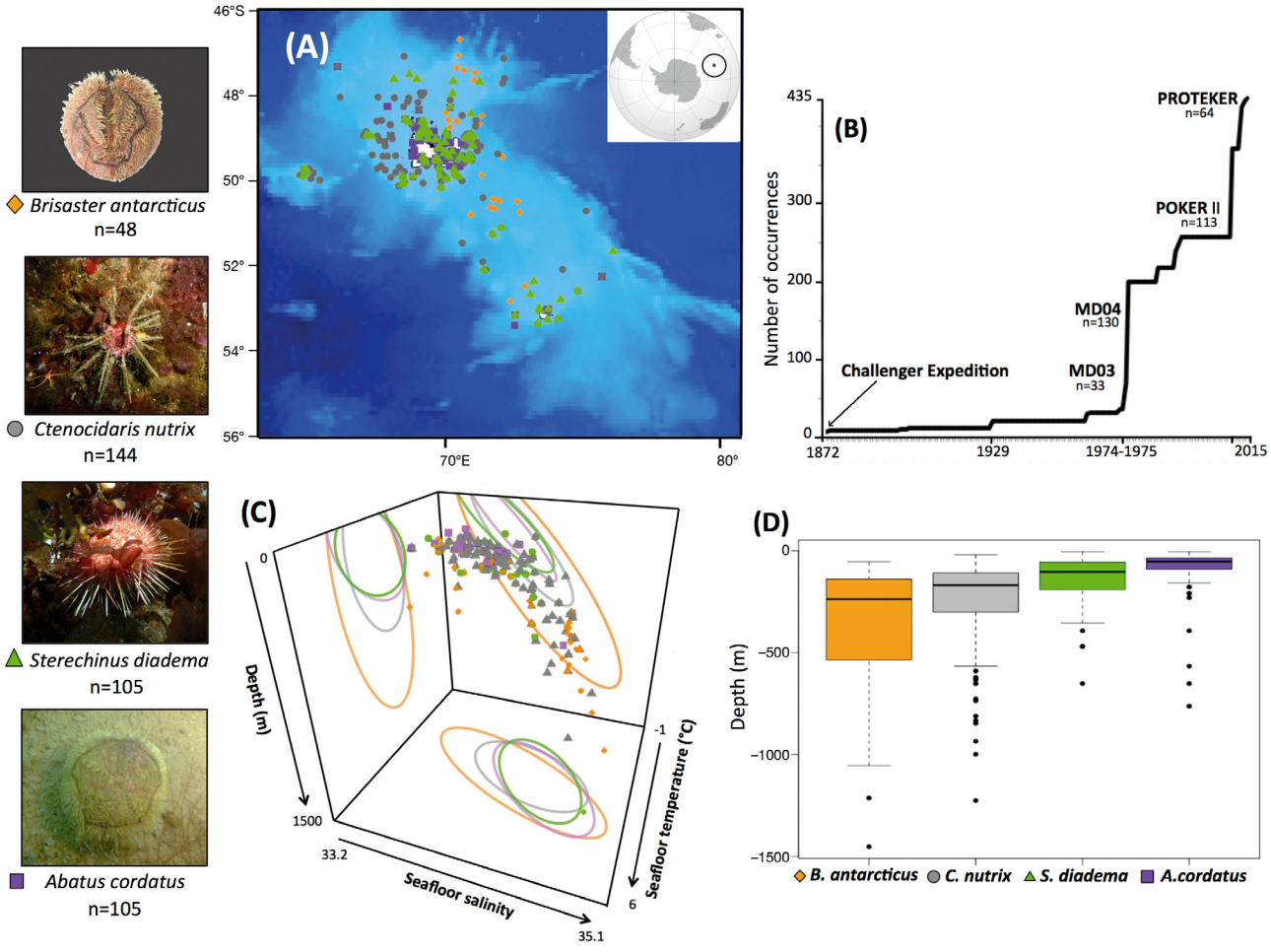

Fig. 1. (A) Occurrence data of the 4 studied echinoid species over the Kerguelen Plateau: *Brisaster antarcticus*, *Ctenocidaris nutrix*, *Sterechinus diadema*, *Abatus cordatus*. (B) Sampling effort (in presence-only records) through time by main scientific cruises during which the 4 studied species were collected on the Kerguelen Plateau. (C) Species presence data plotted according to depth , seafloor salinity and seafloor temperature on 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)

bathic species of the family Cidaridae, *Ctenocidaris nutrix* and 1 omnivorous and eurybathic species of Echinidae, *Sterechinus diadema* (David et al. 2005) (Fig. 1). *A. cordatus* is a coastal species endemic to the Kerguelen Plateau, *B. antarcticus* is known to occur in the Kerguelen and Crozet archipelagoes and has broader environmental preferences than *A. cordatus*, and *C. nutrix* and *S. diadema* are widespread in the Southern Ocean and have contrasting environmental preferences (Fig. 1).

Environmental descriptors

Environmental descriptors were taken from Guillaumot et al. (2016). The data set covers the geographic extent of the Kerguelen Plateau and comprises environmental data encompassing 6 decades (1955–2012). Environmental data are available at a grid cell resolution of 10 km. Environmental layers include no-data pixels, particularly in seafloor-related descriptors. Data were not interpolated to avoid potential biases due to interpolation procedures.

Collinearity between descriptors can alter modeling performances (Phillips et al. 2006) because collinear data may (1) inflate standard errors, (2) induce the violation of residual independency during model validation and (3) generate noise that can be interpreted as a link between descriptors (Dormann et al. 2013). To reduce the collinearity effect, we computed the variance inflation factor (VIF) and Spearman correlation coefficient (r_S) between all available descriptors from Guillaumot et al. (2016). VIF analysis was performed in a stepwise procedure using the 'vifstep'

function in the R package 'usdm' (Naimi et al. 2014). Descriptor pairs with high VIF and r_S values were omitted based on the commonly used thresholds of VIF < 5 and r_S < 0.85 (Pierrat et al. 2012, Dormann et al. 2013, Duque-Lazo et al. 2016). Environmental descriptors finally selected to model species distribution are given 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 procedure

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

Model selection

Due to the growing interest of ecologists in 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' library in R v.3.3.0 (Thuiller et al. 2016) and tested the performance of these approaches with regards to the chronological addition of new data and the transferability performance of models between areas. Several models were generated with an increasing number of occurrence data (see Fig. S1 in Supplement 1 at www.int-res.com/articles/suppl/m594p149_supp.pdf). The best modeling techniques were then compared with each other using a non-random cross-validation procedure (Fig. S2; Wenger & Olden 2012) in order to determine the approach with the best accuracy in transferability performances (Randin et al. 2006, Wenger & Olden 2012).

Results showed high performance and stability values for random forest (RF) and boosted regression trees (BRT) in our case study (see Supplement 1). However, BRT performed better in transferability than did RF (Heikkinen et al. 2012). Previous works have shown that RF does not deal correctly with missing values and patchy data sets (Breiman 2001, Barbet-Massin et al. 2012, Qiao et al. 2015; see Table S1 in Supplement 1 for a review). Therefore, BRT was chosen in the present work to generate the analyses.

BRT calibration was completed using the 'gbm' R package (Elith et al. 2008, Ridgeway 2015). The 3 main parameters (learning rate [lr], tree complexity [tc], bag fraction [bg]) were selected using the method developed by Elith et al. (2008) to determine the combination of values that would minimize the predicted deviance of the models (Elith & Leathwick 2014). The parameters were finally set at lr = 0.0001, tc = 2 and bf = 0.75.

Following Barbet-Massin et al. (2012), we sampled the same number of background data as the number of presence data available for computing BRT models. Considering the low number of presence data points available, 100 model replicates (i.e. background sampling) were generated for each analysis. Finally, to correct for data aggregation in space, presence duplicates were removed when present in the same 10 km resolution pixel.

Model performance was assessed by measuring the area under the receiver operating curve (AUC) of each model replicate using the 'dismo' R library (Hijmans et al. 2016). AUC expresses the relationship between model sensitivity and the commission error (1 − specificity), where sensitivity corresponds to the number of presence pixels correctly predicted as present, and specificity is the number of absence pixels correctly predicted as absent (Fielding & Bell 1997). The use of the AUC to evaluate SDM performance has been debated (Lobo et al. 2008, Peterson et al. 2008), but the AUC remains the most appropriate metric for presence-background models since values remain stable with low-prevalence data sets and are not sensitive to threshold effects (Hand 2009, van Proosdij et al. 2016). Following the recommendation of Jiménez-Valverde (2012), we used the AUC to estimate the robustness of the models but not for direct comparisons between models that were generated for different species, on different study areas or with different training samples.

Correcting for sampling bias

The data collected during the various scientific cruises over the Kerguelen Plateau over the last 145 yr present conspicuous spatial heterogeneities. The resulting biases can generate an unequal number of records in different sectors of the study area and heterogeneous patterns in record distribution. Such heterogeneities can increase the risk of overestimating the contribution of environmental conditions to the models in the most frequently sampled areas (Araújo & Guisan 2006).

Table 1. Environmental descriptors selected for species distribution models. *indicates that environmental layers were available for the following time periods: 1955–2012, 1955–1964, 1965–1974, 1975–1994 and 2005–2012. Minimum and maximum values are shown for the period 1955–2012. Spatial resolution of layers: 10 km grid-cell pixels. Spatial extent: 46–56° S, 63–81° E

Environmental descriptors	Units	Description	Min. value	Max. value	Source
Depth	m	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)
Sea surface mean temperature*	°C	Mean sea surface temperature	3.0566	7.6223	World Ocean Atlas (2013)
Sea surface temperature amplitude*	°C	Amplitude between mean summer and mean winter sea surface temperature	−3.3036	−1.4108	World Ocean Atlas (2013)
Seafloor mean temperature*	°C	Mean seafloor temperature	−0.2978	4.6422	This study. Derived from World Ocean Atlas (2013) sea surface temperature layers
Seafloor temperature amplitude*	°C	Amplitude between mean summer and mean winter seafloor temperature	−2.5757	0.8867	This study. Derived from World Ocean Atlas (2013) sea surface temperature layers
Sea surface mean salinity*	PSS	Mean sea surface salinity	33.6849	33.8251	World Ocean Atlas (2013)
Sea surface salinity amplitude*	PSS	Amplitude between mean summer and mean winter sea surface salinity	−0.0859	0.3165	World Ocean Atlas (2013)
Seafloor salinity amplitude*	PSS	Amplitude between mean summer and mean winter seafloor salinity	− 169	0.0937	This study. Derived from World Ocean Atlas (2013) sea surface salinity layers
Mean surface chl *a*	mg m^{-3}	Surface chlorophyll *a* concentration. Summer mean over 2002–2009	0.1358	2.7324	MODIS AQUA (NASA) 2010
Sediments	Categorical	Sediment features	14 categories		McCoy (1991), updated by H. J. Griffiths (unpubl. data)
Geomorphology	Categorical	Geomorphologic features	27 categories		ATLAS ETOPO2 2014 (Douglass et al. 2014)
Slope	Degrees	Bathymetric slope	4.8229×10^{-5}	0.1547	Biogeographic Atlas of the Southern Ocean (De Broyer et al. 2014)
Mean seafloor oxygen concentration	ml l^{-1}	Mean seafloor oxygen concentration over 1955–2012	4.0080	7.6223	This study. Derived from World Ocean Atlas (2013) sea surface oxygen concentration layers

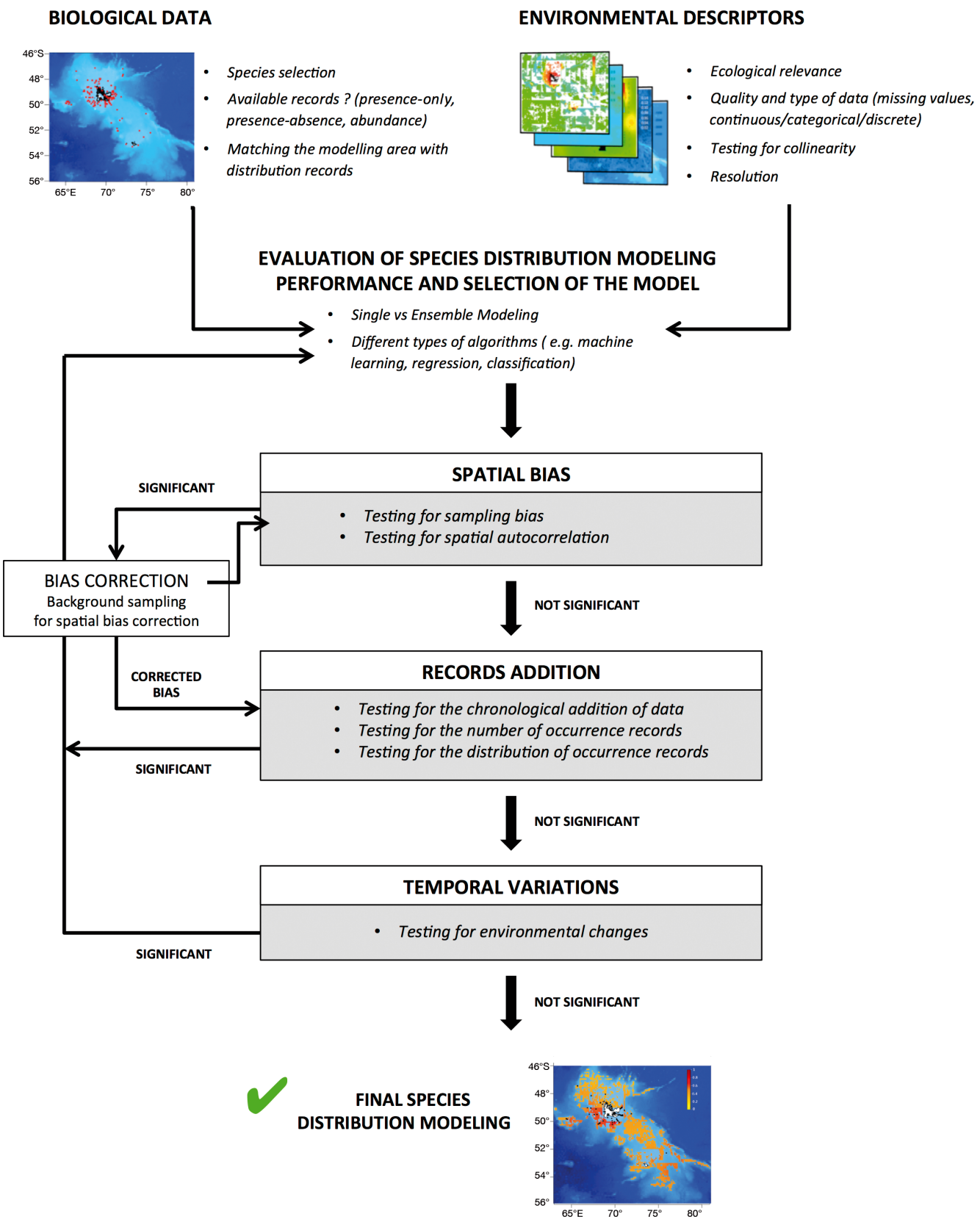

Fig. 2. Tests and procedures carried out in the present work. Arrows indicate the stepwise procedure with statistical validation leading either to the following step or correction/stepback requirements

The effect of spatial heterogeneities on the quality of distribution models was tested using a null model approach. The first null model (null model #1) was generated by sampling presence data at random within the total set of sites that were visited during the different campaigns, whether echinoid specimens were collected at these sites or not (see Fig. 3). Because absence data were not available, this approach allowed us to assess the weight of sampling bias in the models. If a sampling bias is significant, null model #1 is expected to produce distribution maps with higher suitability values in the most frequently sampled areas (Merckx et al. 2011).

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

The 2 null models were generated for the 4 selected species. The number of presence-only data used in the models was contained between the number of data points collected from the MD04 campaign until the PROTEKER campaign, between 1974 and 2015, which corresponds to periods of high sampling effort (Fig. 1B). 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 differed from each other, we used the methodology proposed by Phillips et al. (2009), which has been shown to improve modeling performance (Phillips et al. 2009, Aguirre-Gutiérrez et al. 2013). A grid layer was built using a kernel density estimation (KDE) to represent spatial sampling 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 the 'MASS' R package (Venables & 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, Pearson's r correlation was computed between pixel values of the KDE layer (the proxy for sampling effort) and the predicted probabilities of models after the KDE correction.

Spatial heterogeneities in data collection can also generate spatial autocorrelation (SAC) between presence records, which can violate model calibration assumptions and affect model accuracy with incorrect parameter estimations (Segurado et al. 2006, Dormann 2007, Crase et al. 2012). Several approaches have been developed to take SAC into account in SDMs (see Crase et al. 2012 for a review). They consist of including an additional term in the models (the auto-covariate) which represents the influence of neighboring records on modeling predictions. The significance of SAC was tested using the Moran *I* autocorrelation index computed on model residuals (Luoto et al. 2005, Crase et al. 2012) for both original and corrected models. Models were built using time-averaged environmental descriptors (1955–2012).

Testing for the effect of the chronological addition of new records on model performance

Our data set consisted of presence-only data collected during various scientific cruises with distinct sampling protocols, which may alter the performance of the models (Fig. 1). To test for model reliability, we separately analysed the influence of (1) the chronological addition of presence records, (2) data number alone and (3) sampling patterns (the distribution of data in space). The analyses were performed for *A. cordatus*, *C. nutrix* and *S. diadema*; not enough data were available for *B. antarcticus*. We used time-averaged environmental descriptors (1955–2012) to generate the models.

To test for the potential effect of the chronological addition of new data on model performance, we followed the protocol proposed by Aguiar et al. (2015). The data set was split into distinct subsets corresponding to main periods of sampling effort (1975, including Marion Dufresne campaigns; 1993, including ANARE campaigns; 2010, including POKER II 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 Schoener's *D* statistic. Schoener's *D* is a correlation metric adapted to the study of niche similarities (Warren et al. 2008, Rödder & Engler 2011). It evaluates the similarity of pixel values between 2 distribution grids. A *D* value of 0 means that the 2 maps are perfectly different, and 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 & 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 II, PROTEKER campaigns). All models were compared with each other.

Testing for the effect of temporal variations on model performance

To test for the effect of environmental shifts on the models, different distribution models were generated using distinct environmental descriptors for 4 periods (1955–1964; 1965–1974; 1975–1994; 2005–2012) and the complete set of presence data available. Similarities between models were measured using Schoener's *D* statistic.

RESULTS

Environmental shifts

Mean sea surface temperature and amplitude, mean seafloor temperature and amplitude and mean sea surface salinity and amplitude all differed significantly among all studied decades ($p < 0.001$). Only seafloor temperature amplitude did not significantly differ between the time periods 2005–2012 and 1955–1964. These results indicate that significant environmental shifts occurred during the studied time period, and this may induce important variations in the models since the data set extends over 145 yr.

Spatial bias

Null model #1 predicted higher suitability values in areas with the most intense sampling effort, corresponding to the northern part of the Kerguelen Plateau and the vicinity of the Kerguelen archipelago (Fig. 3A). In contrast, null model #2 predicted medium suitability values over the entire Kerguelen Plateau because presence data were sampled randomly in the area (Fig. 3B). The difference between null models #1 and #2 was significant for the 4 species (Fig. 3), showing that sampling bias has a significant impact on model outputs, which will overestimate environment suitability in areas with the highest number of sampling sites if no correction is applied.

Correlation between visited areas and predicted probability distribution decreased in models built with the KDE-correction compared to non-corrected models (Table 2), showing that the correction is efficient at reducing the influence of sampling bias on modeling performance. However, the correction proved less efficient in models of the coastal and narrow niche species *Abatus cordatus*, for which correlation values after the KDE correction remained high (r = 0.44) (Table 2).

SAC was 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* (see Table S2 and Fig. S3 in Supplement 2). This shows that the KDE procedure corrected for SAC in 3 of the 4 studied species.

Chronological addition of new records

The different models built with a chronological addition of new data showed high AUC values (mean ± SD) for *Ctenocidaris nutrix* and *A. cordatus* ($0.814 \pm 0.018 < AUC_{C.nutrix} < 0.883 \pm 0.024$ and $0.908 \pm 0.023 < AUC_{A.cordatus} < 0.909 \pm 0.018$ respectively), demonstrating the relevance of all models (Fig. 4, see Fig. S4 in Supplement 3). For these 2 species, Schoener's *D* correlation values were high (mean ± SD; $D_{A.cordatus} = 0.978 \pm 0.023$, $D_{C.nutrix} = 0.968 \pm 0.020$) and significant, showing that the models were similar to each other. (see Table S3 in Supplement 3)

In contrast, models generated for *Sterechinus diadema* significantly differed from each other with lower Schoener's *D* statistics ($D_{S.diadema} = 0.932 \pm 0.036$) (see Fig. S5 in Supplement 3). Therefore, the chronological addition of new data has contrasting impacts on model outputs in the studied species, which may be explained by the 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 points presents high and signifi-

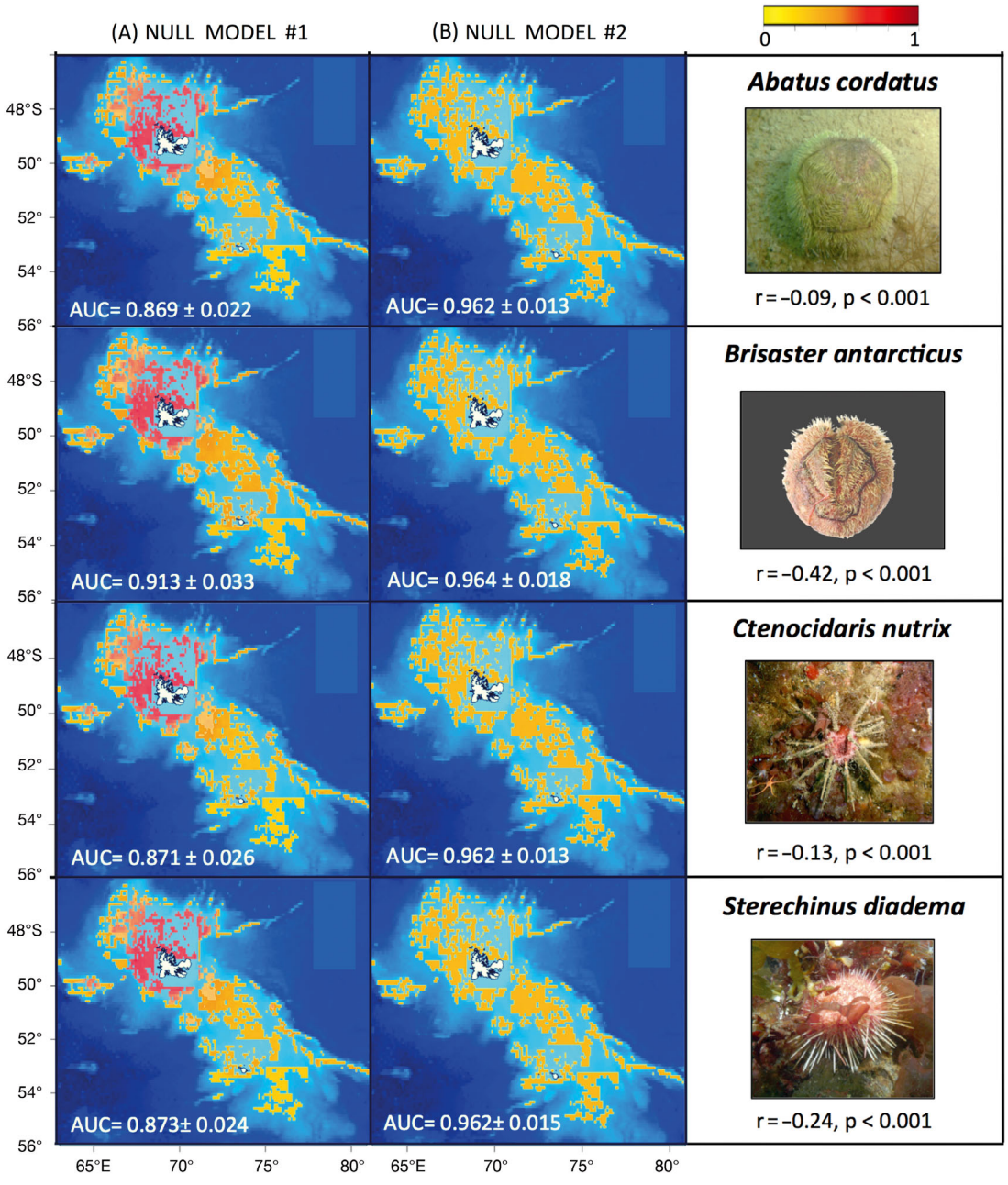

Fig. 3. (A) Null model #1 and (B) null model #2 for the different species under study. Mean (±SD) area under the receiver operating curve (AUC) values are given for the 100 replicates. Comparisons between models compiled with Pearson's r correlation values and associated probabilities (colour bar)

cant Schoener's *D* values (minimum $D = 0.979 \pm 0.031$ for *S. diadema*, maximum $D = 0.985 \pm 0.020$ for *C. nutrix*), showing that model outputs did 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 2 types of models presented low values, suggesting a significant influence of sampling pattern on model output (Table 3).

Environmental change and model performance

The different models generated with contrasting environmental descriptors were highly similar, as

Table 2. Pearson's r correlation of pixel values between the kernel density estimation (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

shown by high Schoener's *D* and low standard deviation values (*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 did not vary significantly among periods for the 4 species. However, *A. cordatus* seems to be less impacted by environmental shifts than the other species (Fig. 5).

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

Final species distribution models

Sampling bias analyses and model corrections showed that reliable distribution models can be built for *C. nutrix* only; this was the only data set in which spatial and temporal heterogeneities did 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 centuries 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 distributions on the Kerguelen Plateau faces significant limitations due to gaps and heterogeneities in the data (Guillaumot et al. 2016). Such limitations can seriously limit the relevance of modeling procedures, which are re-

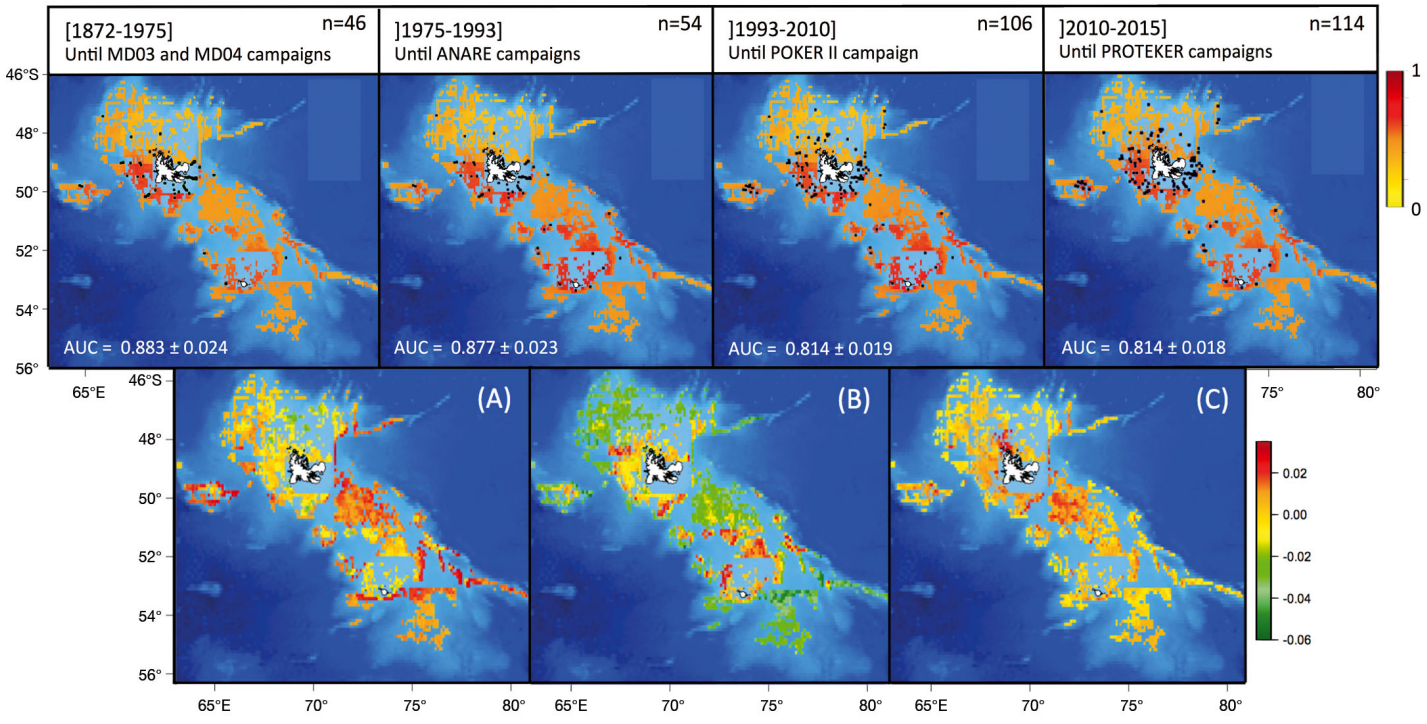

Fig. 4. First row: distribution models of *Ctenocidaris nutrix* for 4 periods, with increasing number of presence data points to build the model (averaged maps of 100 model replicates). Colour bar: probabilities of distribution predicted by the model (between 0 and 1). Second row: difference in probability distribution between (A) n = 54 and n = 46, (B) n = 106 and n = 54 and (C) n = 114 and n = 106. Colour bar represent differences in distribution probabilities between maps

Table 3. Influence of data addition and sampling patterns on models for *Abatus cordatus*, *Ctenocidaris nutrix* and *Sterechinus diadema*. Data addition: mean (±SD) Schoener's *D* and associated p-value computed between models (100 replicates) produced for the different species with n = 54, 76, 95, n = 46, 54, 106, 114 and n = 54, 66, 98 occurrences randomly sampled from the total dataset. Sampling pattern: 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	Data addition		Sampling pattern	
	D_{obs}	p	D_{obs}	p
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

quired by environmental managers for conservation purposes (Féral et al. 2016, Koubbi et al. 2016). In the present work, we followed a step-by-step protocol to assess, quantify and correct the potential effects of data scarcity and heterogeneity on SDMs, 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 SAC

Building SDMs for remote and little-accessed regions often requires the use of spatially biased data sets 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 frequently sampled sites, and thus model performance is reduced (Phillips et al. 2009). In the present work, a significant difference was measured between the 2 null models (generated by selecting presence data either from visited stations only or from random sites over the total investigated area), highlighting the strong heterogeneity of

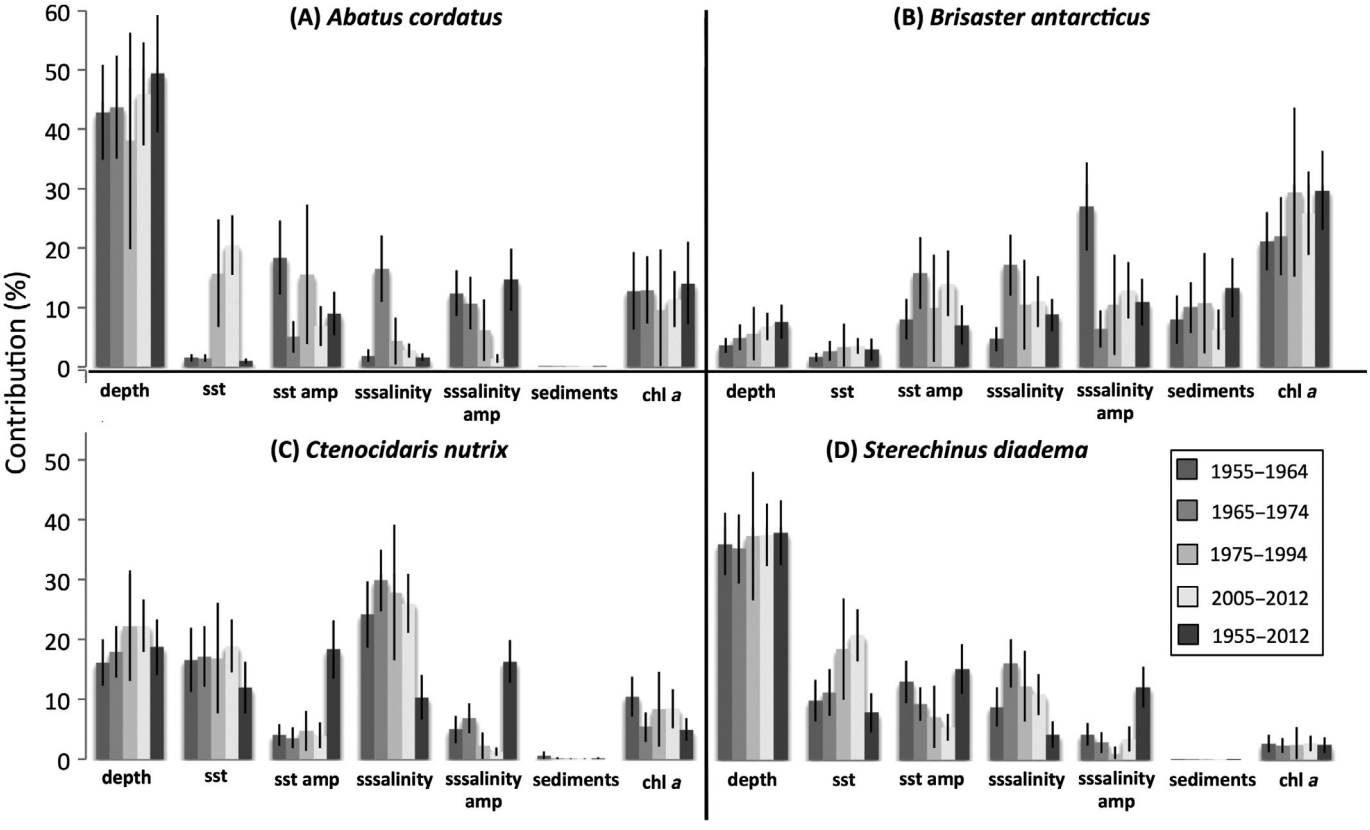

Fig. 5. Mean (±SD) contributions of environmental descriptors to the models for the 4 time periods and species under study. sst: sea surface temperature; sst amp: sea surface temperature amplitude; sssalinity: sea surface salinity; sst amp: sea surface salinity amplitude; chl a: chlorophyll *a* (see Guillaumot et al. 2016 for details)

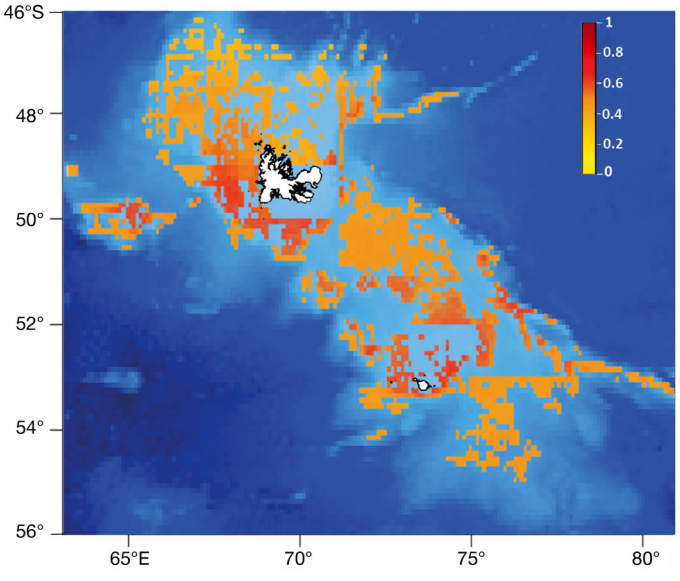

Fig. 6. Species distribution model generated for *Ctenocidaris nutrix* using all presence-only data available and present environmental descriptors (2005–2012). Mean (±SD) area under the receiver operating curve (AUC) = 0.813 ± 0.02

sampling effort with more data collected in the northern part of the Kerguelen Plateau and in shallow coastal areas.

The significant 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 SDMs, 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 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 previously used in studies based on presence-only and limited data sets (Mateo et al. 2010, Pokharel et al. 2016, Phillips et al. 2017). These methods allowed us 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 proven to be efficient to correct both for the influence of uneven sampling effort on predicted distributions (Table 2) and for SAC on all SDMs except for models of *Abatus 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 the most intense sampling efforts. This is in line with previous studies that 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 for broader-niche species. Reducing the extent of distribution modeling of narrow-niche species to the boundaries of their environmental limits could prove a good alternative.

Influence of record addition

The chronological addition of new data had a limited impact on certain model outputs, as demonstrated by high similarities between the chronological models generated for *A. cordatus* and *Ctenocidaris nutrix*. In contrast, chronological models of *Sterechinus diadema* differed significantly from each other. A detailed analysis of data increments proved that the increasing number of presences had no impact on modeling performance, which is not in line with previous works (Stockwell & Peterson 2002, Wisz et al. 2008). However, these results can be altered by our incomplete knowledge of full species distributions due to sampling bias and the limited number of data sets available (Hernandez et al. 2006, Bean et al. 2012). With *S. diadema*, differences between the chronological models were due to contrasted spatial patterns between data sets (transects versus random patterns).

Historical data and environmental change

Significant environmental shifts were measured for the descriptors analysed between 1955 and 2012 over the Kerguelen Plateau (i.e. mean sea surface temperature and amplitude, mean surface salinity and amplitude). However, for all species, distribution models built for each decade were highly similar to each other. These results confirm that temporal heterogeneities in data sets do not necessarily impact the robustness of the models, because spe-

cies preferences for their environment may be wider than the magnitude of changes in time. Working with both present and historical data to improve the completeness of occurrence records proved reliable when assuming that species niche and distribution have not significantly changed during the studied time period.

Between 1955 and 2012, 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* or *S. diadema*; variations between decades were more marked in models produced for *A. cordatus*. This near-shore species is found in shallow waters of the 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 rather to modeling limitations.

Sea surface temperature and salinity amplitudes contributed significantly to the models, contributing more than the averaged parameters (i.e. *A. cordatus* and *B. antarcticus*; Fig. 5). This is in line with the results of Bradie & Leung (2017), who tested for the contribution of several environmental descriptors across a wide panel of taxa. They showed the importance of including seasonal means and extremes in models to further depict species distributions, 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) may be considered the best approach to produce representative models, independent of short-term environmental variations. Unexpectedly, our results showed that for all species, contributions of time-averaged descriptors to the models were much more different than all differences between decadal descriptors (Fig. 5). This suggests that using time-averaged descriptors for long time periods does not necessarily improve model reliability compared to using descriptors averaged over shorter time periods. This also highlights the importance of the descriptor selection in modeling procedures, a critical issue for improving model performance as already stressed in previous studies (Bradie & Leung 2017). This is particularly relevant for certain regions of the Southern Ocean, such as the Western Antarctic Peninsula, which has experienced among the most significant environmental changes in the world's oceans during the last decades (Turner et al. 2014).

Influence of species niche width in modeling performances

Among the 4 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 encountered when correcting for spatial bias compared to other species models. Moreover, the limited environmental variability between coastal sampling sites of the different oceanographic surveys can also explain the absence of a 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 2 species, record data are more widely distributed and show contrasting sampling patterns (i.e. transect-like versus random patterns) that were 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 98 for *C. nutrix* and *S. diadema* respectively) that allowed a more complete survey of *C. nutrix* distribution. Finally, only the *C. nutrix* data set contained the quality and number of occurrence records that fulfilled all methodological requirements to produce a reliable distribution model.

Considering species niche width in order to cope with spatial and temporal bias in SDMs 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, number of data points and type of modeling technique used. However, they did not generate any analysis of species with broad ecological niches as a comparison. Our results are also in line with other modeling studies in which distribution models of species with broad niches were the least stable (Reiss et al. 2011, Guo et al. 2015, Qiao et al. 2015, Ranc et al. 2017).

CONCLUSIONS

The use of SDMs has gained importance during the last decades, providing complementary information

for environmental managers. Modeling results can help interpolate species distributions, identify the potential drivers of a species' distribution and predict the potential effects of environmental changes on habitat suitability. However, modeling species distributions over vast and remote marine areas like the Southern Ocean using poor and heterogeneous data sets remains challenging, and improvement of biological and environmental data sets is still required.

In the present study, we showed that reliable SDMs can be produced in such areas as long as the amount 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 distributions. Using time-averaged predictors over long time periods can generate unrealistic models.

Model selection is also crucial at this stage and the statistical performance of models is not the only criteria to be considered. Modeling procedures must be chosen with regards to the scientific issues that are being addressed. Two procedures (BRT and RF) performed best in our case study, but one of them (BRT) proved to be more relevant because it dealt better with transferability and data patchiness.

Modeling species distributions in data-poor areas poses the practical problem of the minimum number of presence-only data points required to run reliable models, although this is not the only or most critical issue. The number of occurrence records must be high enough for testing model robustness and reliability. In regions with limited access, sampling effort may be heterogeneous, which influences model performance. We showed that sampling bias can be corrected, but the efficiency of the correction depends on species niche width, with narrow-niche species models being more troublesome to correct. In our study, *A. cordatus* is a species limited to shallow coastal areas, which implies a strong correlation between species occurrence and sampling patterns. Restricting the model to a more reduced area could allow for correction of spatial bias and improve modeling performance.

There is also a crucial need for improving the quality of data sets (Kennicutt et al. 2014) and running more accurate models to better tackle conservation issues (Rodríguez et al. 2007, Guisan et al. 2013). For the time being, producing uncertainty maps can be an alternative (Rocchini et al. 2011, Tessarolo et al. 2014) and can provide additional information to environmental managers and stakeholders (Addison et al. 2013, Guisan et al. 2013).

Model reliability and performance also depend on the interaction between data set completeness and a species' intrinsic ecological properties. Hence, we showed that the type and width of ecological niches are important to consider, with the distribution of narrow-niche species being easier to model and less sensitive to incomplete data sets (Guo et al. 2015, Ranc et al. 2017). However, narrow niches usually imply that species are distributed over small areas, for which distribution models will be highly sensitive to extrapolations.

Our protocol showed that reliable SDMs can be produced when enough data are available and data set bias can be tested and corrected. In the present study, only one SDM (*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 distributions based on poor and spatially biased data sets, and should contribute to bringing new insights and enhancing modeling performance in future studies.

Acknowledgements. This work is a contribution to the IPEV program PROTEKER funded by the French polar institute (IPEV program no.1044) and contribution no. 21 to the vERSO project (www.versoproject.be) funded by the Belgian Science Policy Office (BELSPO, contract no. BR/132/A1/vERSO). We thank the 19 scientific cruises for the collection of the data used to realise this work (Table 1, Guillaumot et al. 2016, see also Supplement 3).

LITERATURE CITED

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, 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

Aguirre-Gutiérrez J, Carvalheiro LG, Polce C, van Loon EE, Raes N, Reemer M, Biesmeijer JC (2013) Fit-for-purpose: species distribution model performance depends on evaluation criteria–Dutch hoverflies as a case study. PLOS ONE 8:e63708

Araújo MB, Guisan A (2006) Five (or so) challenges for species distribution modelling. J Biogeogr 33:1677–1688

Barbet-Massin M, Jiguet F, Albert CH, Thuiller W (2012) Selecting pseudo-absences for species distribution models: How, where and how many? Methods Ecol Evol 3: 327–338

Bradie J, Leung B (2017) A quantitative synthesis of the importance of variables used in MaxEnt species distribution models. J Biogeogr 44:1344–1361

Breiman L (2001) Random forests. Mach Learn 45:5–32

Brotons L, Thuiller W, Araújo MB, Hirzel AH (2004) Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. Ecography 27: 437–448

Costa GC, Nogueira C, Machado RB, Colli GR (2010) Sampling bias and the use of ecological niche modeling in conservation planning: a field evaluation in a biodiver hotspot. Biodivers Conserv 19:883–899

Crase B, Liedloff AC, Wintle BA (2012) A new method for dealing with residual spatial autocorrelation in species distribution models. Ecography 35:879–888

Danis B, Van de Putte A, Renaudier S, Griffiths H (2013) Connecting biodiversity data during the IPY: the path towards e-polar science. In: Verde C, di Priso G (eds) Adaptation and evolution in marine environments, Vol 2. Springer, Berlin, p 21–32

David B, Choné T, Mooi R, De Ridder C (2005) Antarctic Echinoidea. Synopses of the Antarctic benthos, Vol 10. Koeltz Scientific, Königstein

De Broyer C, Koubbi P, Griffiths H, Grant A and others (2014) Biogeographic atlas of the Southern Ocean. Scientific Committee on Antarctic Research, Cambridge

Dormann CF (2007) Effects of incorporating spatial autocorrelation into the analysis of species distribution data. Glob Ecol Biogeogr 16:129–138

Dormann CF, Elith J, Bacher S, Buchmann C and others (2013) Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 36:27–46

Douglass LL, Turner J, Grantham HS, Kaiser S, Constable A and others (2014) A hierarchical classification of benthic biodiversity and assessment of protected areas in the Southern Ocean. PLOS ONE 9:e100551

Duque-Lazo J, Van Gils HAMJ, Groen TA, Navarro-Cerrillo RM (2016) Transferability of species distribution models: the case of *Phytophthora cinnamomi* in southwest Spain and southwest Australia. Ecol Modell 320:62–70

Elith J, Leathwick J (2014) Boosted regression trees for ecological modeling. www.lcis.com.tw/paper_store/paper_store/brt(5)-2015115131033846.pdf

Elith J, Leathwick JR, Hastie T (2008) A working guide to boosted regression trees. J Anim Ecol 77:802–813

Féral JP, Saucède T, Poulin E, Marschal C and others (2016) PROTEKER: implementation of a submarine observatory at the Kerguelen Islands (Southern Ocean). Underwat Technol 34:3–10

Ferrier S, Guisan A (2006) Spatial modelling of biodiversity at the community level. J Appl Ecol 43:393–404

Ficetola GF, Thuiller W, Padoa-Schioppa E (2009) From introduction to the establishment of alien species: bioclimatic differences between presence and reproduction localities in the slider turtle. Divers Distrib 15:108–116

Fielding AH, Bell JF (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ Conserv 24:38–49

Franklin J (2009) Mapping species distributions: spatial inference and prediction. Cambridge University Press, Cambridge

Greathead C, González-Irusta JM, Clarke J, Boulcott P, Blackadder L, Weetman A, Wright PJ (2015) Environmental requirements for three sea pen species: relevance to distribution and conservation. ICES J Mar Sci 72: 576–586

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

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

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

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

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

Gutt J, Zurell D, Bracegridle T, Cheung W and others (2012) Correlative and dynamic species distribution modelling for ecological predictions in the Antarctic: a cross-disciplinary concept. Polar Res 31:11091

Gutt J, Barnes D, Lockhart SJ, van de Putte A (2013) Antarctic macrobenthic communities: a compilation of circumpolar information. Nat Conserv 4:1–13

Hand DJ (2009) Measuring classifier performance: a coherent alternative to the area under the ROC curve. Mach Learn 77:103–123

Heikkinen RK, Marmion M, Luoto M (2012) Does the interpolation accuracy of species distribution models come at the expense of transferability? Ecography 35:276–288

Hernandez PA, Graham CH, Master LL, Albert DL (2006) The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography 29:773–785

Hijmans RJ, Phillips S, Leathwick J, Elith J (2016) dismo: species distribution modeling. R package version 1.1-1. https://CRAN.R-project.org/package=dismo

Hortal J, Lobo JM, Jiménez-Valverde A (2007) Limitations of biodiversity databases: case study on seed-plant diversity in Tenerife, Canary Islands. Conserv Biol 21:853–863

Hortal J, Jiménez-Valverde A, Gómez JF, Lobo JM, Baselga A (2008) Historical bias in biodiversity inventories affects the observed environmental niche of the species. Oikos 117:847–858

Jackson AL, Inger R, Parnell AC, Bearhop S (2011) Comparing isotopic niche widths among and within communities: SIBER–stable isotope Bayesian ellipses in R. J Anim Ecol 80:595–602

Jiménez-Valverde A (2012) Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. Glob Ecol Biogeogr 21:498–507

Kennicutt MC, Chown SL, Cassano JJ, Liggett D and others (2014) Six priorities for Antarctic science. Nature 512: 23–25

Koubbi P, Mignard C, Causse R, Da Silva O and others (2016) Ecoregionalisation of the Kerguelen and Crozet islands oceanic zone. Part I: Introduction and Kerguelen oceanic zone. Commission for the Conservation of Antarctic Marine Living Resources Report, Working Group on Ecosystem Monitoring and Management No.16/43 CCAMLR, Bologna

Lobo JM, Jiménez-Valverde A, Real R (2008) AUC: a misleading measure of the performance of predictive distribution models. Glob Ecol Biogeogr 17:145–151

Loiselle BA, Jørgensen PM, Consiglio T, Jiménez I, Blake JG, Lohmann LG, Montiel OM (2008) Predicting species distributions from herbarium collections: Does climate bias in collection sampling influence model outcomes? J Biogeogr 35:105–116

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

Mateo RG, Croat TB, Felicísimo AM, Muñoz J (2010) Profile or group discriminative techniques? Generating reliable species distribution models using pseudo-absences and target-group absences from natural history collections.

Divers Distrib 16:84–94
McCoy FW (1991) Southern Ocean sediments; circum-Antarctic to 30°S. In: Hayes DE (ed) Marine geological and geophysical atlas of the circum-Antarctic to 30° S. Antarctic Research Series, Vol 54. American Geophysical Union, Washington, DC, p 37–46
McCune JL (2016) Species distribution models predict rare species occurrences despite significant effects of landscape context. J Appl Ecol 53:1871–1879
Merckx B, Steyaert M, Vanreusel A, Vincx M, Vanaverbeke J (2011) Null models reveal preferential sampling, spatial autocorrelation and overfitting in habitat suitability modelling. Ecol Modell 222:588–597
Muscarella R, Galante PJ, Soley-Guardia M, Boria RA, Kass JM, Uriarte M, Anderson RP (2014) ENMeval: an R package for conducting spatially independent evaluations and estimating optimal model complexity for ecological niche models. Methods Ecol Evol 5:1198–1205
Naimi B, Hamm NA, Groen TA, Skidmore AK, Toxopeus AG (2014) Where is positional uncertainty a problem for species distribution modelling? Ecography 37:191–203
Newbold T (2010) Applications and limitations of museum data for conservation and ecology, with particular attention to species distribution models. Prog Phys Geogr 34: 3–22
Peterson AT, Pape M, Soberón J (2008) Rethinking receiver operating characteristic analysis applications in ecological niche modeling. Ecol Modell 213:63–72
Phillips SJ, Anderson RP, Schapire RE (2006) Maximum entropy modeling of species geographic distributions. Ecol Modell 190:231–259
Phillips SJ, Dudík M, Elith J, Graham CH, Lehmann A, Leathwick J, Ferrier S (2009) Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecol Appl 19:181–197
Phillips ND, Reid N, Thys T, Harrod C, Payne N and others (2017) Applying species distribution modelling to a data poor, pelagic fish complex: the ocean sunfishes. J Biogeogr 44:2176–2187
Pierrat B, Saucède T, Laffont R, De Ridder C and others (2012) Large-scale distribution analysis of Antarctic echinoids using ecological niche modelling. Mar Ecol Prog Ser 463:215–230
Pokharel KP, Ludwig T, Storch I (2016) Predicting potential distribution of poorly known species with small database: the case of four horned antelope *Tetracerus quadricornis* on the Indian subcontinent. Ecol Evol 6:2297–2307
Qiao H, Soberón J, Peterson AT (2015) No silver bullets in correlative ecological niche modelling: insights from testing among many potential algorithms for niche estimation. Methods Ecol Evol 6:1126–1136
Raes N, ter Steege H (2007) A null-model for significance testing of presence-only species distribution models. Ecography 30:727–736
Ranc N, Santini L, Rondinini C, Boitani L, Poitevin F, Angerbjörn, Maiorano L (2017) Performance tradeoffs in target-group bias correction for species distribution models. Ecography 40:1076–1087
Randin CF, Dirnböck T, Dullinger S, Zimmermann NE, Zappa M, Guisan A (2006) Are niche-based species distribution models transferable in space? J Biogeogr 33: 1689–1703
Reiss H, Cunze S, König K, Neumann H, Kröncke I (2011) Species distribution modelling of marine benthos: a North Sea case study. Mar Ecol Prog Ser 442:71–86
Reiss H, Birchenough S, Borja A, Buhl-Mortensen L and others (2015) Benthos distribution modelling and its relevance for marine ecosystem management. ICES J Mar Sci 72:297–315
Reutter BA, Helfer V, Hirzel AH, Vogel P (2003) Modelling habitat suitability using museum collections: an example with three sympatric *Apodemus* species from the Alps. J Biogeogr 30:581–590
Ridgeway G (2015) gbm: generalized boosted regression models. R package version 2.1.1. https://CRAN.R-project.org/package=gbm
Robinson LM, Elith J, Hobday AJ, Pearson RG, Kendall BE, Possingham HP, Richardson AJ (2011) Pushing the limits in marine species distribution modelling: lessons from the land present challenges and opportunities. Glob Ecol Biogeogr 20:789–802
Rocchini D, Hortal J, Lengyel S, Lobo JM and others (2011) Accounting for uncertainty when mapping species distributions: the need for maps of ignorance. Prog Phys Geogr 35:211–226
Rödder D, Engler JO (2011) Quantitative metrics of overlaps in Grinnellian niches: advances and possible drawbacks. Glob Ecol Biogeogr 20:915–927
Rodríguez JP, Brotons L, Bustamante J, Seoane J (2007) The application of predictive modelling of species distribution to biodiversity conservation. Divers Distrib 13:243–251
Segurado PA, Araújo MB, Kunin WE (2006) Consequences of spatial autocorrelation for niche-based models. J Appl Ecol 43:433–444
Stockwell DR, Peterson AT (2002) Effects of sample size on accuracy of species distribution models. Ecol Modell 148: 1–13
Tessarolo G, Rangel TF, Araújo MB, Hortal J (2014) Uncertainty associated with survey design in species distribution models. Divers Distrib 20:1258–1269
Thuiller W, Georges D, Engler R, Breiner F (2016) biomod2: ensemble platform for species distribution modeling. R package version 3.3-7. https://CRAN.R-project.org/package=biomod2
Turner J, Barrand NE, Bracegirdle TJ, Convey P and others (2014) Antarctic climate change and the environment: an update. Polar Rec 50:237–259
Van de Putte AP, Griffiths HJ, Raymond B, Danis B (2014) Data and mapping. In: De Broyer C, Koubbi P, Griffiths HJ, Raymond B and others (eds) Biogeographic atlas of the Southern Ocean. Scientific Committee on Antarctic Research, Cambridge, p 14–15
van Proosdij ASJ, Sosef MSM, Wieringa JJ, Raes N (2016) Minimum required number of specimen records to develop accurate species distribution models. Ecography 39:542–552
Venables WN, Ripley BD (2002) Modern applied statistics with S, 4th edn. Springer, New York, NY
Vierod AD, Guinotte JM, Davies AJ (2014) Predicting the distribution of vulnerable marine ecosystems in the deep sea using presence-background models. Deep Sea Res II 99:6–18
Warren DL, Glor RE, Turelli M (2008) Environmental niche equivalency versus conservatism: quantitative approaches to niche evolution. Evolution 62:2868–2883
Wenger SJ, Olden JD (2012) Assessing transferability of ecological models: an underappreciated aspect of statistical validation. Methods Ecol Evol 3:260–267
Wisz MS, Hijmans RJ, Li J, Peterson AT, Graham CH, Guisan A (2008) Effects of sample size on the performance of species distribution models. Divers Distrib 14:763–773
World Ocean Atlas (2013) v2. National Centers for Environmental Information, NOAA. https://www.nodc.noaa.gov/OC5/woa13/woa13data.html

Editorial responsibility: Jean-Sébastien Lauzon-Guay, Dartmouth, Nova Scotia, Canada

Submitted: July 27, 2017; Accepted: March 1, 2018
Proofs received from author(s): April 21, 2018