Progress in Oceanography 175 (2019) 198–207

Contents lists available at ScienceDirect

Progress in Oceanography

journal homepage: www.elsevier.com/locate/pocean

Broad-scale species distribution models applied to data-poor areas

Guillaumot Charlène[a,c,*], Artois Jean[b], Saucède Thomas[c], Demoustier Laura[a], Moreau Camille[a,c], Eléaume Marc[d], Agüera Antonio[a,e], Danis Bruno[a]

[a] Université Libre de Bruxelles, Marine Biology Lab., Avenue F.D. Roosevelt, 50, CP 160/15, 1050 Bruxelles, Belgium
[b] Université Libre de Bruxelles, Spatial Epidemiology Lab. (SpELL), Avenue F.D. Roosevelt, 50, CP 160/15, 1050 Bruxelles, Belgium
[c] UMR 6282 Biogéosciences, Univ. Bourgogne Franche-Comté, CNRS, 6 bd Gabriel, F-21000 Dijon, France
[d] Muséum national d'Histoire naturelle, Département Systématique et Évolution, UMR ISYEB 7205, 57 rue Cuvier, F-75231 Paris Cedex 05, France
[e] Danish Shellfish Center, DTU-aqua, Ørroddevej 80, 7900 Nykøbing Mors, Denmark

ARTICLE INFO

Keywords:
Boosted Regression Trees (BRTs)
Presence-only
Cross-validation
Extrapolation
Modelling evaluation

ABSTRACT

Species distribution models (SDMs) have been increasingly used over the past decades to characterise the spatial distribution and the ecological niche of various taxa. Validating predicted species distribution is important, especially when producing broad-scale models (i.e. at continental or oceanic scale) based on limited and spatially aggregated presence-only records. In the present study, several model calibration methods are compared and guidelines are provided to perform relevant SDMs using a Southern Ocean marine species, the starfish *Odontaster validus* Koehler, 1906, as a case study. The effect of the spatial aggregation of presence-only records on modelling performance is evaluated and the relevance of a target-background sampling procedure to correct for this effect is assessed. The accuracy of model validation is estimated using k-fold random and spatial cross-validation procedures. Finally, we evaluate the relevance of the Multivariate Environmental Similarity Surface (MESS) index to identify areas in which SDMs accurately interpolate and conversely, areas in which models extrapolate outside the environmental range of occurrence records.

Results show that the random cross-validation procedure (i.e. a widely applied method, for which training and test records are randomly selected in space) tends to over-estimate model performance when applied to spatially aggregated datasets. Spatial cross-validation procedures can compensate for this over-estimation effect but different spatial cross-validation procedures must be tested for their ability to reduce over-fitting while providing relevant validation scores. Model predictions show that SDM generalisation is limited when working with aggregated datasets at broad spatial scale. The MESS index calculated in our case study show that over half of the predicted area is highly uncertain due to extrapolation. Our work provides methodological guidelines to generate accurate model assessments at broad spatial scale when using limited and aggregated presence-only datasets. We highlight the importance of taking into account the presence of spatial aggregation in species records and using non-random cross-validation procedures. Evaluating the best calibration procedures and correcting for spatial biases should be considered ahead the modelling exercise to improve modelling relevance.

1. Introduction

Species Distribution Models (SDMs) have been increasingly used during the past decades. The diversity of applications has widened to include a vast panel of topics from studies of invasive species distribution range shifts to assessment of species responses to environmental drivers and conservation issues from local to global scales (Guisan and Thuiller, 2005, Ficetola et al., 2007, Guisan et al., 2013, Beaumont et al., 2016, Phillips et al., 2017). In vast and remote areas such as the Southern Ocean, modelling species distributions is challenged by (1) the paucity of biotic data available (a serious constraint when describing species realised niche), (2) by the heterogeneous quality of environmental data describing environmental conditions (e.g. missing data in coastal areas, low resolution of environmental layers, limited number of environmental descriptors available), and (3) by the sampling bias (spatial and temporal aggregation of data collection) (Barry and Elith, 2006, Robinson et al., 2011, Hortal et al., 2012, Tessarolo et al., 2014, Guillaumot et al., 2018). Sampling effort has mostly been carried out offshore or in the vicinity of research stations during the austral summer while remote shallow areas are seldom accessed and dense winter sea ice conditions limit oceanographic studies (Gutt et al., 2012).

* Corresponding author.
E-mail address: charleneguillaumot21@gmail.com (G. Charlène).

https://doi.org/10.1016/j.pocean.2019.04.007
Received 25 June 2018; Received in revised form 31 March 2019; Accepted 20 April 2019
Available online 22 April 2019
0079-6611/

Several studies have proposed model corrections or alternatives to separately mitigate the induced impacts of spatial and temporal biases on modelling performance (Phillips et al., 2009, Newbold, 2010, Barbet-Massin et al., 2012, Hijmans, 2012, Tessarolo et al., 2014, Guillera-Arroita et al., 2015, Guillaumot et al., 2018, Valavi et al., 2018). However, to our knowledge, no study has yet proposed methodological guidelines to address such issues when dealing with data-poor and broad spatial areas (i.e. at continental or oceanic scales).

Several statistical tools such as the Area Under the Curve of the Receiver Operating characteristic (AUC), the True Skill Statistic, or the Point Biserial Correlation are commonly used to evaluate the relevance of SDM predictions (Fielding and Bell, 1997, Allouche et al., 2006). Using these indices for models performed with presence-only data has been widely discussed because background-data are usually considered as absences, leading to confusion in model interpretation and violating most test assumptions (i.e. computing AUC and TSS statistics requires the use of true absences) (Jiménez-Valverde, 2012, Li and Guo, 2013). These methods can also be biased when applied to limited and broadly distributed data. Machine-learning algorithms are widely used in SDMs to fit complex relationships between species occurrences and environmental data (Elith et al., 2006). The resulting models may be highly complex and poorly efficient under changing environmental conditions as they may fit a response to any variation including the random noise (=model overfitting), (Wenger and Olden, 2012). Models' ability to predict in new environmental conditions is described as the generalisation performance by Friedman et al. (2001).

Producing reliable SDMs implies finding a good trade-off between model complexity and predictive and generalisation performances (Anderson and Gonzalez, 2011, Radosavljevic and Anderson, 2014). The relevance of modelling and generalisation performance, and the optimal level of model complexity can be tested using independent data. The method has been commonly applied and referred to as the cross-validation procedure (Araujo and Guisan, 2006, Valavi et al., 2018). The cross-validation procedure uses a training subset of occurrence data to fit the model and a separate test subset to validate the predictions and the statistical relationships between the studied variables (Fielding and Bell, 1997). 'Random cross-validation' procedures are widely used and randomly split the occurrence dataset into training and test subsets. However, the spatial aggregation of occurrence data can lead to the violation of the independence assumption between training and test data randomly sampled, and in turn to false confidence in modelling validation performances (Hijmans, 2012). The violation of the independence assumption can also lead to generate highly complex and overfitted models (Boria et al., 2014, Merow et al., 2013, Radosavljevic and Anderson, 2014). Therefore, the cross-validation procedure should be adapted to each given dataset and case study, so that, different 'spatial cross-validation' procedures have been developed and compared in this study. The spatial cross-validation procedures aim at spatially splitting the occurrence dataset into a training and a test subset by increasing the geographical distance between the two subsets (Veloz, 2009, Brenning, 2012, Muscarella et al., 2014, Radosavljevic and Anderson, 2014, Brown et al., 2017, Valavi et al., 2018). The spatial cross-validation reduces spatial correlation between training and test data in situations where spatial autocorrelation is significant in the occurrence dataset, a common issue in ecology (Roberts et al., 2017).

Uncertainties in SDMs represent another limitation to model usage that should be quantified and the effects must be specifically assessed or taken into account during model interpretation (Barry and Elith, 2006, Carvalho et al., 2011, Beale and Lennon, 2012, Guisan et al., 2013). Model extrapolation outside the range of the known species environmental conditions leads to misinterpretation of SDM outputs and can be a real issue when using SDM predictions as a support tool for conservation decisions. Therefore, areas of optimal predictions and limited uncertainties must be identified. This can be achieved using indicators such as the Multivariate Environmental Similarity Surface (MESS). Developed for SDMs, the MESS index highlights areas where environmental conditions are outside the range of conditions observed in data (Elith et al., 2010).

In the present study, model uncertainties and the performance of several spatial cross-validation procedures were analysed using the case study of the sea star *Odontaster validus* Koehler, 1906. Distributed over the entire Southern Ocean (< 45°S), *O. validus* is a common and abundant species in shallow-water benthic habitats (McClintock et al., 2008, Lawrence, 2013), characterised by an opportunistic feeding behaviour (from suspension-feeding to algivory, deposit-feeding and predation). It has been shown to play a significant role in structuring benthic communities and regulating populations of other benthic taxa (McClintock et al., 2008). The species physiology was recently modelled using the Dynamic Energy Budget approach (Agüera et al., 2015) which allows for the assessment of the metabolic performance of the species under different environmental conditions. Here, SDMs were produced to interpolate the known distribution of *O. validus* over its entire geographic range using an available dataset of environmental descriptors. The influence of spatial data aggregation on model outputs was analysed and the performance of correction procedures evaluated. In a second step, several cross-validation procedures were assessed and compared to test for modelling accuracy, optimal level of complexity and predictive performance. A final 'optimum' model is proposed, which takes into account uncertainty estimates. Results are generalised and formalised as guidelines for further SDM works, showing the relevance of the approach when working at broad spatial scale with a limited number of spatially aggregated presence-only records.

2. Material and methods

2.1. Model selection and calibration procedures

SDMs were generated using the Boosted Regression Trees (BRTs) algorithm. BRTs were selected for their ability to fit complex relationships between species records and the related environment, while guarding against over-fitting (Elith et al., 2008, Reiss et al., 2011). BRTs are also adapted to deal with incomplete datasets (Elith et al., 2008), can perform well with low prevalence datasets (Barbet-Massin et al., 2012), are weakly sensitive to species niche width (Qiao et al., 2015) and were recognised to transfer well in space and time (Elith et al., 2006, Elith and Graham, 2009, Heikkinen et al., 2012).

BRTs were calibrated using the method proposed by Elith et al. (2008) to select the optimal number of trees in the final model (Appendix A). The combination of parameters that minimises the optimal number of trees to build the model (reduction of complexity) while reaching a minimum predictive deviance to the test data (reduction of error) was selected. The following parameters were used to calibrate the models: tree complexity = 4, bag fraction = 0.75 and learning rate = 0.007 (Fig. S2). The number of background data sampled in the area was set at 1000 sampled points after evaluating the optimal number of data points to be sampled (see Appendix A for details). This number constitutes the best trade-off between describing environmental conditions and being as close as possible to the number of species presence records available (Barbet-Massin et al., 2012). All background sampling was restricted in space to areas shallower than 1500 m depth, which corresponds to the species deepest record, in order to avoid model extrapolation at depths known as unsuitable for the species survival based on knowledge of the species ecology (McClintock et al., 2008, Lawrence, 2013). Sampling was restricted to a single background data per pixel. Similarly, presence records falling on a same 0.1° grid-cell pixel were filtered before model calibration in order to reduce spatial over-weighting (Segurado et al., 2006, Boria et al., 2014).

2.2. Occurrence dataset

SDMs were generated using presence-only data made available for the sea star *O. validus* by Moreau et al. (2018). Presence-only records of *O. validus* are strongly aggregated in space (i.e. concentrated in "easily" accessible and frequently visited areas characterised by relatively low sea ice concentrations), a condition also prevailing in the total dataset available for Southern Ocean benthic taxa (updated from Griffiths et al. (2014), Fig. S3), making *O. validus* a representative case study for Southern Ocean benthic studies.

Models were generated using the environmental descriptors published as raster layers by Fabri-Ruiz et al. (2017). They were collected from different sources and modified to fit modelling requirements at the scale of the Southern Ocean (from 45°S latitude to Antarctica coasts). Collinearity between environmental descriptors was tested using the Variance Inflation Factor (VIF) stepwise procedure of the 'usdm' R package (Naimi et al., 2014) and Spearman correlations (rs) (R Core Team, 2017). Surface temperature and roughness, a depth-derived variable, were respectively correlated to ice cover and depth. They were omitted according to the commonly used thresholds of VIF > 5 and rs > 0.85 (Pierrat et al., 2012, Dormann et al., 2013, Duque-Lazo et al., 2016). A final set of 16 environmental descriptors at 0.1° resolution was compiled to build the models (Table S5).

2.3. Evaluation and correction of spatial aggregation

The significance of spatial aggregation of occurrence data was tested by measuring spatial autocorrelation (Legendre and Fortin, 1989) on model residuals using the Moran's I index (Segurado et al., 2006, Dormann, 2007, Crase et al., 2012). A positive Moran's I value (between 0 and 1) indicates that spatially close residuals will share similar values. A negative (close to −1) or null value respectively indicates a maximal dispersion or a random dispersion of residuals in space (Cliff and Ord, 1981). Detecting significant spatial autocorrelation in presence-only records will assess the degree of aggregation of species records in the studied area.

Two null models were generated and their respective outputs compared to each other in order to evaluate the importance of spatial aggregation in the total Southern Ocean benthic dataset (Fig. S3). Null model #1 was produced to evaluate the overall spatial aggregation of benthic records in the Southern Ocean due to sampling effort. It was generated by randomly sampling n = 309 occurrence records (corresponding to the number of non-duplicate presence-only data available for *O. validus*) in the total Southern Ocean benthic dataset (Fig. S3). 1000 background records were randomly sampled in the entire Southern Ocean. The Moran's I score was calculated by comparing model #1 predictions to the distribution of the total Southern Ocean benthic dataset (Fig. S3). Null model #2 was built to compute a reference Moran's I score for a model generated with randomly distributed records. 309 presence data and 1000 background data were randomly sampled in the entire Southern Ocean. Null model #2 will provide a reference value for spatial autocorrelation scores due to the intrinsic structure of environmental data. It will serve as a reference model for comparison with Moran's I scores of model null #1 and to assess the degree of spatial aggregation due to sampling effort.

To correct for the effect of spatial aggregation on modelling performance, a target-background correction method was applied (Phillips et al., 2009). The total Southern Ocean benthic dataset (Fig. S3) was used to create a Kernel Density Estimation layer that provides an estimate of the probability to find a benthic presence data for each pixel. The Kernel Density Estimation was calculated with the 'kde2d' function of the *MASS* R package (Ripley, 2015) on the extent of the Southern Ocean (n and lims parameters defined to fit a raster layer of extent (−180, 180, −80, −45) and 0.1° resolution). Null model #1 was corrected by randomly sampling 1000 background records according to the weighting scheme of the Kernel Density Estimation layer.

After evaluating spatial aggregation in the total Southern Ocean benthic dataset, spatial autocorrelation was specifically assessed for *O. validus*. Spatial autocorrelation was measured for two models generated without (model A) and with (model B) Kernel Density Estimation correction. Comparison between the two models aimed at assessing the efficiency of the Kernel Density Estimation correction for *O. validus*. Model A (without correction) was built using all presence-only data available for *O. validus* and 1000 background records randomly sampled in the Southern Ocean. Model B (with correction) was built using all presence-only data available for *O. validus* and 1000 background records that were sampled following the weighting scheme of the Kernel Density Estimation layer. Each model was generated 100 times and the two averaged models (average models A and B) were compared to each other. Differences between models A and B quantify the importance of spatial aggregation on model outputs.

Finally, model relevance was assessed using three statistics: the Area Under the Receiver Operating Curve (AUC) (Fielding and Bell, 1997), the Point Biserial Correlation between predicted and observed values (COR, Elith et al., 2006) and the True Skill Statistic (TSS, Allouche et al., 2006).

2.4. Testing different cross-validation procedures

SDMs validation was performed using different cross-validation procedures. Background data were first sampled in the entire area following the Kernel Density Estimation scheme and the compilation of presence-only and background data was then split into a training and a test subset to build the cross-validation procedure. Two splitting procedures were followed; they differ between each other in the spatial independence between the training and the test subset. (1) The random cross-validation procedure, commonly used in SDMs, aims at randomly splitting the dataset into training and test subsets (Fielding and Bell, 1997, Hijmans, 2012) which may lead to close spatial vicinity between the two datasets (Hijmans, 2012), and, (2) the spatial cross-validation procedure that aims at spatially spitting the dataset in order to reduce spatial correlation and may improve independence between the two subsets (Hijmans, 2012, Muscarella et al., 2014).

The random procedure was therefore compared to four different spatial cross-validation procedures. (1) In the 'BLOCK' method developed by Muscarella et al. (2014), different subsets of equal occurrence numbers are created. For each replicate, this k-fold procedure divides the dataset into four equal subsets according to the mean latitude and mean longitude positions of occurrence data (Fig. 1C), then three of these four subsets are randomly selected to train the model (75%) and the last one is used to test the model (25%). (2) In the 'CLOCK' methods, the dataset was divided according to random longitudinal transects, splitting the Antarctic Circle into two parts (2-fold 'CLOCK' method, Fig. 1B), (3) three parts (3-fold 'CLOCK' method, Fig. 1D) or (4) four parts (4-fold 'CLOCK" method, Fig. 1E). In the 2-fold 'CLOCK' method, one subset was considered as the training subset, the second one as the test subset; in the 3-fold 'CLOCK' method, two subsets were defined for training and the third one for testing; in the 4-fold 'CLOCK' method, three subsets were considered for training and one for testing (Fig. 1). Different cross-validation procedures were tested using the 'gbm.step' procedure available in the *dismo* R package (Elith et al., 2008, Hijmans et al., 2016). Once the dataset is split in different folds, Elith et al. (2008) apply an iterative procedure that enable to find the minimum deviance to the test data, and relates it to the optimal number of trees (optimal model complexity) to generate the model. If test and training data are spatially correlated, the number of trees required to build BRTs will be overestimated. Therefore, the use of Elith et al. (2008) procedure will enable to accurately interpret and compare optimal complexity and performance scores of models calibrated with either randomly or spatially segregated folds (i.e. with contrasting distances between training and test subsets), and thus will help explain the influence of occurrence spatial aggregation on model complexity and

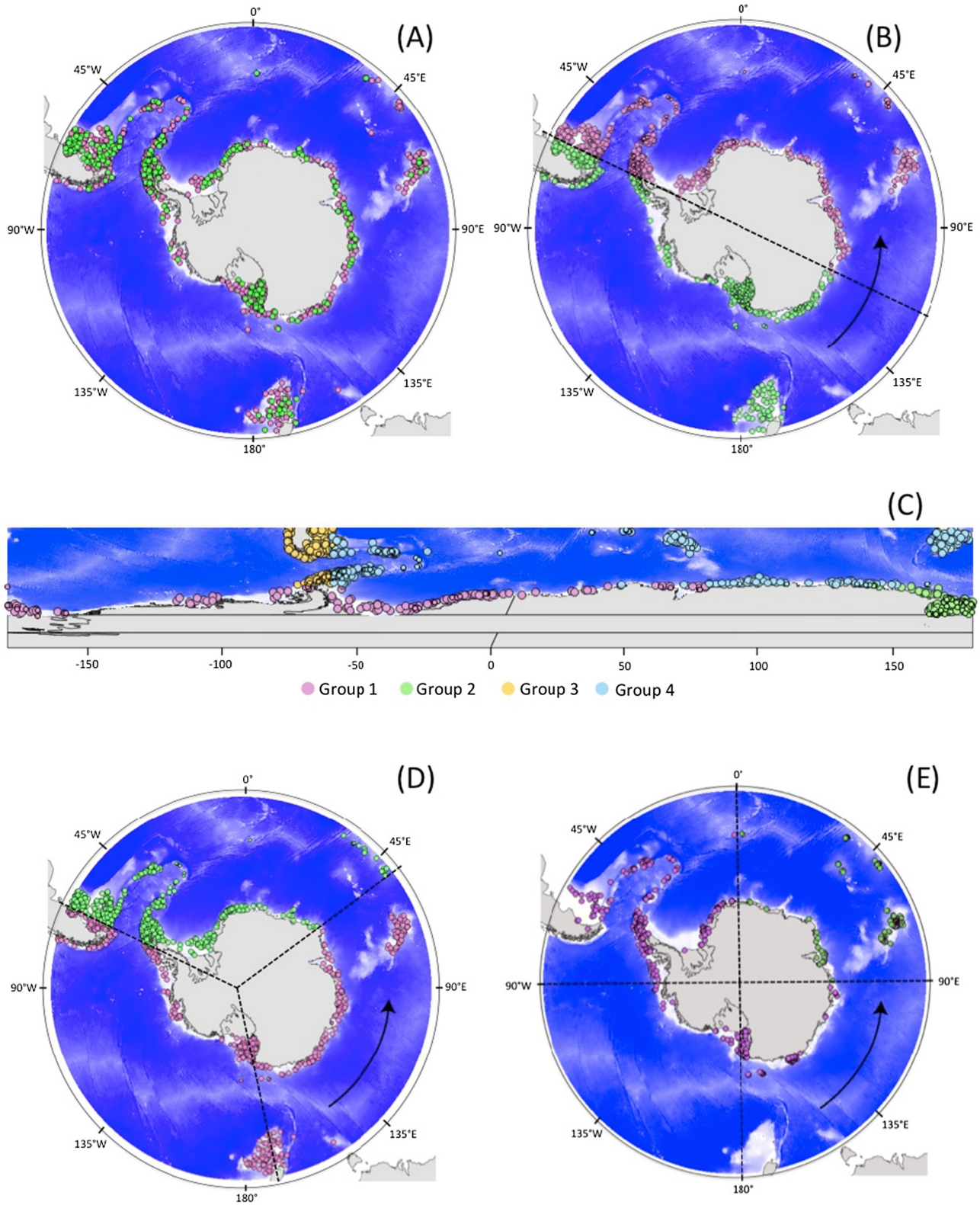

Fig. 1. Comparison of the different cross-validation procedures. Dots represent *Odontaster validus* presence-only records and a random set of 1000 background data, sampled according to the Kernel Density Estimation weighting scheme. Colors indicate data splitting into training (pink) and test (green) subsets. Blue background corresponds to bathymetry and grey areas to emerged lands. For each case, 100 replicates of random background-data sampling and transects partitioning are performed, symbolised by the arrows on the figure. (A) Random cross-validation procedure, with a random splitting into 75% training and 25% test data. (B) '2-fold CLOCK' clustering by random spatial partition of the dataset into two groups (one training, one test). (C) 'BLOCK' splitting, generated according to the median latitudinal and longitudinal values (Muscarella et al., 2014). After generation of four groups (corresponding to the four colors), one group is randomly defined as the test subset, the other three groups as the training subset. A different system of projection was used to represent this map to highlight the latitudinal and longitudinal definition of the transects. (D) '3-fold CLOCK' clustering by random spatial partition of the dataset into three groups (2 training, 1 test). (E) '4-fold CLOCK' clustering by random spatial partition of the dataset into four groups (3 training, 1 test). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

performance.

R scripts written to generate the models and the different cross-validation procedures are provided online at: https://github.com/charleneguillaumot/THESIS/.

Independence between training and test subsets was evaluated using the Spatial Sorting Bias index (SSB) (Hijmans, 2012). SSB compares the distance between training-presence and testing-presence data with the distance between training-presence and training-background. SSB ~ 0 (non independence) means that the ''distance between training-presence and test-presence sites will tend to be smaller than the distance between training-presence and test-background sites'' (Hijmans, 2012). SSB ~ 1 indicates that the two distances are comparable (independent enough) (Hijmans, 2012). SSB was calculated with the *dismo* R package (Hijmans et al., 2016).

SDMs evaluation was generated by computing the percentage of test data that fall on grid-cell pixels predicted as suitable. Suitable pixels were defined using the Maximum sensitivity plus specificity threshold (MaxSSS) that splits models into suitable (> MaxSSS value) and unsuitable areas (< MaxSSS value). MaxSSS is accepted as a relevant threshold for presence-only SDMs (Liu et al., 2013). The averaged optimal number of trees required to generate BRTs was compared between models and used as a proxy of model complexity.

Statistical differences between models generated with the different cross-validation procedures (AUC, TSS, COR, percentage of correctly classified test data, number of trees) were tested using the non-parametric Mann-Whitney Wilcoxon pairwise comparison.

2.5. Assessment of model uncertainty

The Multivariate Environmental Similarity Surface (MESS) index was estimated following the procedure described by Elith et al. (2010) using the *dismo* R package (Hijmans et al., 2016). The MESS calculation consists in extracting the environmental conditions where presence-only data were recorded and determining for each pixel of the model projection layer if environmental conditions are covered by presence-only records. Negative MESS values indicate areas of model extrapolation in which the value of at least one environmental descriptor is beyond the environmental range covered by available presence-only records. Conversely, positive MESS values indicate areas of model projection in which values of environmental descriptors are within the environmental range covered by presence-only records. According to the number of environmental descriptors that are not included inside the range of presence records values, MESS outcome can strongly vary. The MESS evaluation deals with each environmental descriptor equally (unweighted analysis) and in this study, a pixel was considered as unsuitable as soon as a single descriptor value does not match the environmental range of presence-only records. On a projection map, SDM predictions were darkened according to the MESS extrapolation range to visualise the uncertain area due to extrapolation. Extrapolation performance of SDMs was assessed by comparing the proportion of the environment predicted as suitable by the model with the total set of environmental conditions.

3. Results

3.1. Available data and spatial autocorrelation

Distribution records available for *Odontaster validus* display a circumpolar and patchy spatial pattern (Fig. 2A). The niche occupied by *O. validus* does not cover the entire range of environmental conditions prevailing in the projection area (Fig. 2B). *O. validus* is recorded in conditions close to zero and sub-zero seafloor temperatures (Fig. 2B) and is mainly distributed in shallow and coastal areas. Most of *O. validus* presence records are aggregated in regions where scientific benthic surveys are most often led and where sampling effort was privileged due to access facilities (e.g. the Ross Sea and the Antarctic Peninsula). Overall, this holds true for presence records of all benthic Southern Ocean taxa as well (Fig. S3), although, in this case, most environmental conditions are covered by the total benthic samples (Fig. 2B).

Spatial autocorrelation was measured for both the total Southern Ocean benthic dataset (null models) and for *O. validus* alone (models A and B) (Table 1). Moran's I scores were tested significant for all models, null model #2 excepted. The absence of spatial autocorrelation ($I = 0.005 \pm 0.004$; $p = 0.19$) in null model #2 shows that environmental data are not strongly aggregated in space. In contrast, presence-only records of the total Southern Ocean benthic dataset are spatially aggregated. The degree of spatial aggregation due to sampling effort is evidenced by the comparison between null model #1 and #2, scores of model #1 being 10 times higher than those of null model #2 (Moran's $I = 0.050 \pm 0.011$ and 0.005 ± 0.004, respectively).

Values of Moran's I computed for models of *O. validus* (models A and B) are higher than those computed for the total Southern Ocean benthic dataset (null model #1 and #1 with Kernel Density Estimation). The sampling bias is therefore more pronounced for *O. validus* than for the majority of other benthic species.

Model correction by the Kernel Density Estimation procedure was shown to reduce spatial autocorrelation with Moran's I values decreasing from 0.050 to 0.034 for null model #1, and from 0.085 to 0.069 for *O. validus* models A and B (Table 1). However, although lower, Moran's I values remain significant after correction, indicating that the applied corrections do not entirely remove the presence of spatial autocorrelation.

3.2. Comparison of cross-validation procedures

For the BRTs fitted with the random cross-validation procedure, all overall goodness-of-fit metrics (AUC, TSS, COR) were good with predictive accuracy Area Under the Curve (AUC) values higher than 0.9 (Table 2). However, when evaluated through spatial cross-validation procedure, the AUC scores decreased in all BRTs. These results show that BRTs tend to overfit the data if the independence between training and test data is not ensured. Indeed, the random cross-validation procedure presents SSB values close to zero, indicating that training and test subsets may be highly correlated (Fig. 1A). In contrast, all spatial cross-validation procedures have SSB values close to 1, indicating a better spatial independence between training and test data (Table 2).

The generalisation performance (AUC and correctly classified test data) are very high for the random cross-validation procedure, with more than 89.4% of test-presence records falling correctly in areas predicted as suitable by the model (Table 2).

The random cross-validation procedure generates more complex BRTs compared to the spatial methods (significantly higher number of trees for the random cross-validation procedure compared to the spatial cross-validation procedures). As the model closely fits the dataset used for its construction, high AUC, TSS and COR scores were obtained but these results may be misleading and overestimated. In contrast, spatial cross-validation procedures generate less complex models (more general), which could account for lower AUC, TSS and COR scores.

3.3. Proposed model and uncertainty map

We decided to maximise the spatial independence between training and test subsets, minimise model complexity and optimise generalisation performances in *O. validus* model. Using these criteria, we found that the '2-fold CLOCK' modelling method was well adapted to *O. validus* dataset (second highest TSS and COR scores; high proportion of test data being correctly classified, with the lowest standard deviation score ($80.04 \pm 3.49\%$); an important proportion of the total dataset used a test subset [19–81%] and the lowest model complexity ($ntrees = 375 \pm 91.9$)).

The MESS index was calculated in order to define the part of this extrapolated area, that is, the part of the geography for which at least

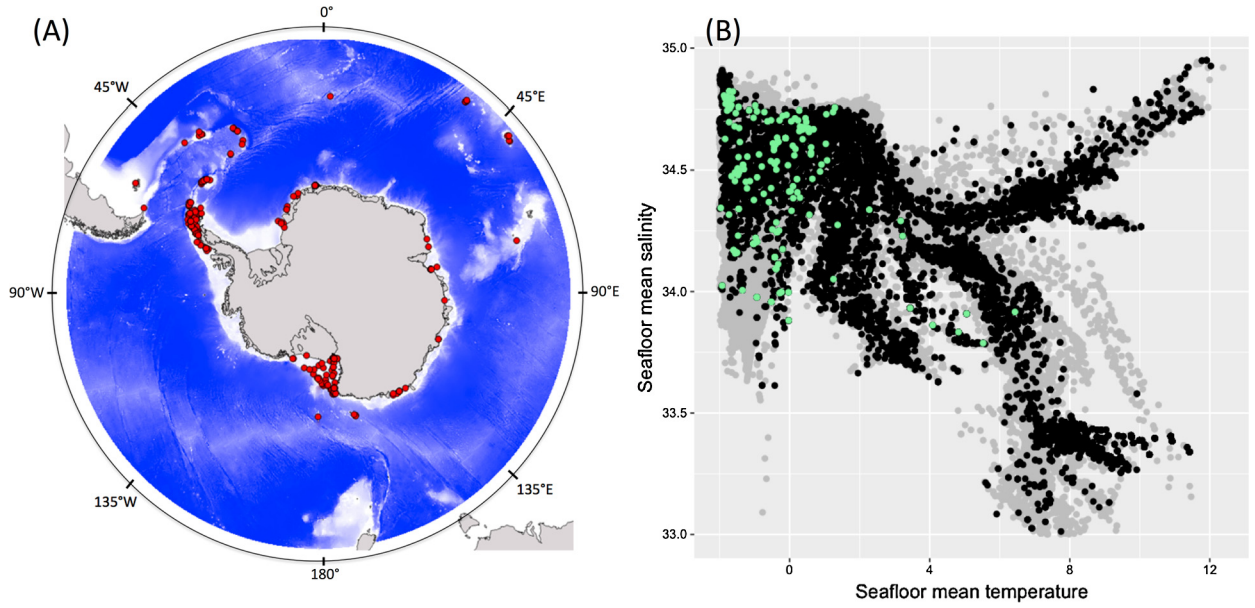

Fig. 2. (A) Presence-only records of the sea star *Odontaster validus* in the Southern Ocean. Duplicates (occurrences falling on a same 0.1° resolution pixel) were removed from the display. (B) Values of the environmental range covered by the entire benthos sampling dataset presented in Fig. S3 (black dots), by presence-only records of *O. validus* (green dots) in comparison with a set of 1000 background dots randomly sampled according to the Kernel Density Estimation scheme (grey dots) for two environmental descriptors: mean seafloor temperature (°C) and mean seafloor salinity (PSU). A part of the environment (grey dots) does not contain benthic occurrence samples (black dots), illustrating that sampling effort is not geographically exhaustive. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
Comparison between models of spatial autocorrelation values measured on model residuals (average and standard deviation of Moran's I values computed for 100 model replicates). Moran's I significance is indicated by p-values; for $p < 0.05$, the absence of spatial autocorrelation (null hypothesis) is rejected. Null model #1: 309 presence records were randomly sampled among occurrences of the total Southern Ocean benthic dataset (Fig. S3) and background data are composed of 1000 points randomly sampled in the entire Southern Ocean; model #2: 309 records (to define presence records) and 1000 background data both randomly sampled in the entire Southern Ocean; model #1 with Kernel Density Estimation: similar to model null #1 but with 1000 background data randomly sampled following the Kernel Density Estimation weighting scheme; model A: 309 presence records of *Odontaster validus* and 1000 background data were randomly sampled in the entire Southern Ocean; model B: similar to model A but with the 1000 background data sampled following the Kernel Density Estimation weighting scheme. AUC: Area Under the Receiver Operating Curve, TSS: True Skill Statistic, COR: Point Biserial Correlation.

	Null model #1	Null model #2	Null model #1 with KernelDensity Estimation	Model A	Model B
Spatial autocorrelation (Moran's I)	0.050 ± 0.011 *p < 0.001*	**0.005 ± 0.004** ***p = 0.19***	0.034 ± 0.011 ***p < 0.001***	0.085 ± 0.009 *p < 0.001*	0.069 ± 0.006 *p < 0.001*
AUC	0.976 ± 0.010	0.710 ± 0.014	0.964 ± 0.015	0.997 ± 0.001	0.948 ± 0.003
TSS	0.674 ± 0.013	0.331 ± 0.020	0.660 ± 0.019	0.698 ± 0.002	0.696 ± 0.003
COR	0.850 ± 0.028 *p < 0.001*	0.336 ± 0.018 *p < 0.001*	0.801 ± 0.037 *p < 0.001*	0.944 ± 0.011 *p < 0.001*	0.923 ± 0.015 *p < 0.001*

one environmental descriptor is outside the environmental conditions of the sampled presence records. The MESS index was compiled as a raster layer and projected on the probability distribution map by darkening uncertain areas (Fig. 3). Uncertain areas due to extrapolation represent 64.2% of the entire projected surface, the major part being also predicted by the model as unsuitable (Table 3). Almost 9.5% of the area was however predicted as suitable by the model although considered as an extrapolated area.

Table 2
Average Spatial Sorting Bias (SSB) and standard deviation values for the 100 model replicates (background sampling + test/training clustering). AUC: Area Under the Receiver Operating Curve; Correctly classified test data (%): percentage of presence-test and background-test records falling on predicted suitable areas (prediction > maximum sensitivity plus specificity (maxSSS) threshold); TSS: True Skill Statistic; COR: Point Biserial Correlation; ntrees: averaged optimal number of trees required to generate BRTs. Stars are indicated for spatial cross-validation groups significantly different from the random cross-validation procedure (non-parametric pairwise Mann-Whitney Wilcoxon test, p-value < 0.01).

	Random cross-validation Random splitting	Spatial cross-validation Block method	Spatial cross-validation 2-fold Clock method	Spatial cross-validation 3-fold Clock method	Spatial cross-validation 4-fold Clock method
Mean SSB	0.101 ± 0.04	0.802 ± 0.37	0.832 ± 0.09	0.803 ± 0.23	0.848 ± 0.32
AUC	0.947 ± 0.013	0.854* ± 0.06	0.811* ± 0.053	0.818* ± 0.078	0.824* ± 0.089
Correctly classified test data (%)	89.452 ± 1.523	80.946* ± 7.504	80.039* ± 3.489	80.713* ± 5.421	79.471* ± 8.538
Test data (% of total dataset)	25%	[13–38]%	[19–81%]	[1–68%]	[1–66%]
TSS	0.715 ± 0.041	0.542* ± 0.188	0.465* ± 0.088	0.490* ± 0.136	0.576* ± 0.165
COR	0.792 ± 0.029	0.632* ± 0.126	0.584* ± 0.089	0.591* ± 0.12	0.483* ± 0.197
ntrees	1580 ± 251.058	543.5* ± 88.9	375* ± 91.9	424.5* ± 131.1	379* ± 98.5

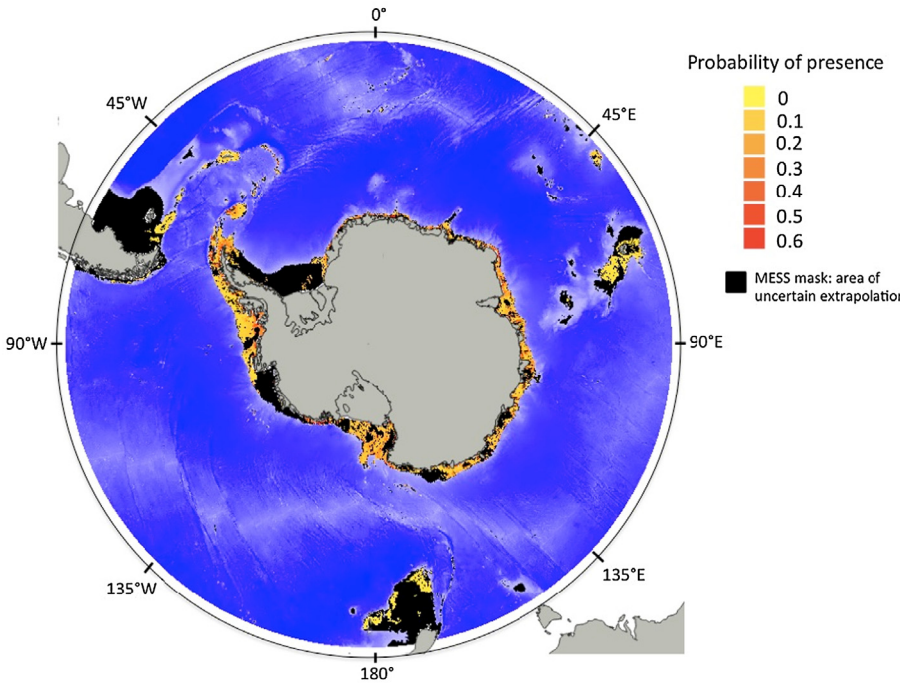

Fig. 3. SDMs performed with the spatial cross-validation '2-fold CLOCK' method. Average of 100 model replicates. Distribution probabilities are darkened according to the Multivariate Environmental Similarity Surface (MESS) layer, with dark pixels corresponding to regions where the model extrapolates outside of the environmental conditions in which the species was sampled. Dark pixels represent 64.2% of the entire projected area. Probabilities of presence are contained between 0 and 1 but the colorbar was scaled until 0.6 to enhance visual contrast.

Table 3
Proportion of interpolated and extrapolated pixels according to the averaged SDM predictions. Interpolation (or uncertain extrapolation respectively) refers to areas where environmental conditions within the pixel are inside (or outside, respectively) of the species ecological range, as defined by the Multivariate Environmental Similarity Surface (MESS). Suitable pixels were defined using the MaxSSS threshold that splits model predictions into suitable (> maxSSS mean score) or unsuitable areas (< maxSSS mean score).

MESS classification	Model prediction	
	Suitable pixels	Unsuitable pixels
Interpolation	10.24%	25.57%
Uncertain extrapolation	9.42%	54.77%

4. Discussion

4.1. Evaluating SDM performance

Using independent datasets to test SDM performance is a prerequisite for relevant validation analyses (Peterson et al., 2011). At broad spatial scale and in data-poor areas, the number of available data is limited and data distribution often patchy, which really challenges the success of validation procedures. Estimating the performance of SDM predictions and the level of extrapolation in such areas is a necessity. The cross-validation procedure has been proposed as a reliable approach to evaluate modelling performances (Fielding and Bell, 1997, Hijmans, 2012, Dhingra et al., 2016, Roberts et al., 2017). Cross-validation procedures must however be adapted to spatially aggregated data because training and test subsets may be sampled in close areas, violating the independence assumption (Segurado et al., 2006, Hijmans, 2012). Such a potential bias is rarely taken into account. In the present work, we compared SDM performance using five different cross-validation procedures for modelling, at broad spatial scale, the distribution of a species for which available data are limited in number and are spatially aggregated. Results show strong differences between procedures, which highlights the importance of testing and selecting the most appropriate method when evaluating model performance.

4.2. Correction for spatial autocorrelation and spatial bias

Strong significant Moran's I scores were measured on model residuals, revealing the presence of spatial autocorrelation in the total Southern Ocean benthic dataset (Fig. S3). The difference between null models #1 and #2 evidences the influence of sampling aggregation on spatial autocorrelation values (Table 1) as discussed by Guillaumot et al. (2018). *O. validus* presence-only dataset follows the same pattern, with records aggregated in coastal areas where sampling effort has been mostly concentrated (Table 1, Fig. 2). A target-group background sampling was applied and proved to be efficient to reduce spatial autocorrelation (as assessed using Moran's I statistic), although it still remains at a significant level. Spatial autocorrelation scores are strongly dependent on the resolution of environmental raster layers. The coarse resolution of environmental data used in the present study may be responsible for the over-estimation of spatial autocorrelation scores. This could account for spatial autocorrelation remaining significant even after the Kernel Density Estimation correction.

4.3. Selection of cross-validation procedures

The random cross-validation procedure has been widely used in ecological modelling to evaluate model predictions (Fielding and Bell, 1997, Merow et al., 2013, Mainali et al., 2015, Torres et al., 2015, Phillips et al., 2017) but the method has been rarely compared to alternative procedures. The present study shows that contrasting model assessments are obtained when using different cross-validation procedures (Radosavljevic and Anderson, 2014, Roberts et al., 2017). Applying a random cross-validation to an aggregated dataset at a broad spatial scale can result in training and test subsets being sampled in the same area, and leads to an inflation of modelling performances (Veloz, 2009, Hijmans, 2012, Radosavljevic and Anderson, 2014, Wenger and Olden, 2012). In the context of this study, SDMs produced with a broad-scale and spatially aggregated occurrence dataset and a random cross-validation procedure are more complex and likely over-fit the training dataset. This also may account for the high evaluation scores obtained (AUC, TSS, COR) and may also explain the apparent high generalisation performance of BRTs fitted with random cross-validation. The lack of model generality can *a posteriori* lead to strong caveats and unreliable

models with poor transferability performance when projected on a new environmental space (Wenger and Olden, 2012, Crimmins et al., 2013). Methods that select the most parsimonious BRT, combine low model complexity and high modelling performance should therefore be preferred.

The spatial cross-validation procedures tested in this study were shown to produce less complex models than the random cross-validation procedure. Increased model generality (i.e. decrease in model overfitting) and forced spatial segregation between training and test subsets result in decreasing SDM validation scores. These results show that applying a random cross-validation procedure for a patchy dataset can lead to over-estimation of SDM predictive performance if training and test subsets are not independent. This is in line with several works (Brenning, 2005, Elith et al., 2010, Anderson, 2013, Muscarella et al., 2014) in which a decrease of AUC scores can be reported when using a spatial cross-validation procedure instead of a random procedure. Machine-learning algorithms have been reported to be the best approaches to generate SDMs but the influence of over-fitting on model evaluation are underestimated (Reiss et al., 2011, Duan et al., 2014, Beaumont et al., 2016, Thuiller et al., 2016, Guillaumot et al., 2018) although its effect has been pointed out in several works (Elith et al., 2008, Jiménez-Valverde, 2008, Wenger and Olden, 2012). Our results show that the evaluation of SDM performance can be strongly influenced by the choice of the evaluation procedure.

In this work, several spatial cross-validation procedures were compared with each other but no single and best procedure emerged, a common case in ecological modelling (Qiao et al., 2015). The appropriate method to be used is highly dependent on the species and dataset under study. For instance, the 'BLOCK' method introduced by Muscarella et al. (2014) should not be used at broad spatial scale, where too important latitudinal contrasts in environmental conditions are present. In this study, such contrasting environmental conditions (due to the presence of an environmental latitudinal gradient between sub-Antarctic and Antarctic regions, with occurrence aggregation in the two regions) lead to higher variability in generalisation performance during model projection, depending on the data subsets selected to train and test the model (Roberts et al., 2017). The 'BLOCK' method favors the independence between training and test subsets but models are slightly more complex because they are calibrated on contrasting environmental conditions (sub-Antarctic vs. Antarctic areas) and over-fit the training dataset that could also present a patchy distribution. The 'BLOCK' method is therefore more adapted to case studies without strong patchy and contrasting environmental conditions. The 'CLOCK' procedures developed in this study helped reduce the effect of latitudinal patchy occurrences distribution by mixing presence records sampled in Antarctic and sub-Antarctic regions to define training and test subsets. The 'CLOCK' methods generate less complex models and were proved more efficient to define spatially independent training and test subsets. However, the number of training and test records sampled between model replicates is not constant, which contributes to an important variability in validation performance scores. The selection of the different 'CLOCK' methods also depends on the importance of data aggregation and patchy patterns within environmental conditions. For strong data aggregation, the '2-fold CLOCK' approach will help reduce the influence of patchy patterns during model calibration and will help generalise the model and decrease its complexity. '3 or 4-fold CLOCK' methods present close modelling performances but the proportion of occurrence records used to test the model can be very low.

Alternative SDM evaluation procedures can be found in the literature: for instance, calibrated cross-validation procedures aim at removing occurrences from the test subset when considered too close to the training subset (and considered as non-informative according to a statistical threshold) (Hijmans, 2012). For limited presence-only datasets, removing a part of the available occurrence data may lead to the removal of a proportion of informative records, which does not constitute a reasonable option (Bean et al., 2012, van Proosdij et al., 2016). The leave-one-out method can also provide a relevant estimate of model goodness-of-fit, even for spatially aggregated datasets (Olden et al., 2002, Wenger and Olden, 2012). The method aims at randomly excluding a single record from the total dataset. The model is trained on the remaining data and predicts the model response on the single removed point to test for model prediction. The procedure is replicated several times, providing a powerful evaluation of model accuracy. However, assessment of generalisation performances is not permitted with this approach (Wenger and Olden, 2012).

In addition to cross-validation procedures, the relevance of model validation performance is also strongly dependent on the quality of environmental descriptors available. The number of no-data pixels as well as grid-cell resolution can critically affect model evaluation. This is especially true in the present study because environmental variables, measured or interpolated, rarely extend to coastal areas, and resolution in the Southern Ocean can rarely be better than 10 km^2. Good quality datasets are needed and such limitations must be taken into account when interpreting model outputs.

4.4. Uncertainty assessment in SDMs predictions

SDM uncertainty assessment has been a widely discussed topic (Barry and Elith, 2006, Carvalho et al., 2011, Beale and Lennon, 2012, Guisan et al., 2013). Uncertainty in model predictions has been often assessed as the variation among the predicted distribution probabilities (Buisson et al., 2010) but this approach does not provide precise information on the origin of uncertainty (Tessarolo et al., 2014).

The MESS metric is a relevant indicator of SDM extrapolation performance (Elith et al., 2010, Dhingra et al., 2016). The Mobility Oriented Parity (MOP) introduced by Owens et al. (2013) was recently proposed as an alternative to the MESS index. MESS considers extrapolation on a pixel as uncertain when at least one environmental value falls outside the environmental range of presence records. In contrast, MOP offers more flexibility by defining an extrapolated area when all environmental values fall outside the sampled environmental range. Therefore, MESS is more conservative than MOP to define species ecological envelope.

Here, MESS was used to assess the proportion of the projected area for which models extrapolate. Our results show that more than half of the area corresponds to environmental conditions for which presence records have not been sampled. 9.42% of this extrapolated area is even predicted as a suitable environment. This highlights the weakness of SDMs for spatial generalisation and the risk of providing inaccurate SDMs for conservation purposes, especially if the communication between modellers and environmental managers is neglected (Guisan et al., 2013). Our results show the importance of providing uncertainty maps along with SDM outputs in order to help interpret models with the necessary caution.

5. Conclusion

This work highlights the importance of assessing the relevance of SDM evaluation procedures. When applied to occurrence datasets, spatially autocorrelated and broad-scale presence-only datasets, the random cross-validation procedure may over-estimate model validation scores due to the violation of independence between training and test subsets. Applying a spatial cross-validation procedure that spatially segregates training and test data was shown to be effective to provide a reliable analysis of model performance. Spatial cross-validation methods also help reduce model complexity and therefore improve generalisation performances. The 'CLOCK' methods developed in this paper were proved to be appropriate to our Southern Ocean case study and could be applied to other non-polar case studies. This study proves the importance of testing and comparing several spatial cross-validation procedures to identify the procedure most adapted to each case study.

The MESS index was used to visualise areas where SDMs extrapolate

outside the range of the environmental conditions where presence records were sampled. Such results show the importance of providing information on model uncertainty to correctly interpret SDM outputs.

Acknowledgements

This work was supported by a “Fonds pour la formation à la Recherche dans l’Industrie et l’Agriculture” (FRIA) grant to C. Guillaumot. This is contribution no. 26 to the vERSO project (http://www.versoproject.be), funded by the Belgian Science Policy Office (BELSPO, contract n°BR/132/A1/vERSO). We are also thankful to the anonymous reviewers that help improve this work with relevant remarks and advices.

Authors’ contributions

CG, JA, TS conceived the ideas and designed the methodology; LD provided a part of the data; CM, ME, AA and BD contributed to the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.pocean.2019.04.007.

References

Allouche, O., Tsoar, A., Kadmon, R., 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). J. Appl. Ecol. 43 (6), 1223–1232.
Agüera, A., Collard, M., Jossart, Q., Moreau, C., Danis, B., 2015. Parameter estimations of Dynamic Energy Budget (DEB) model over the life history of a key Antarctic species: the Antarctic sea star *Odontaster validus* Koehler, 1906. Plos One 10 (10), e0140078.
Anderson, R.P., Gonzalez Jr, I., 2011. Species-specific tuning increases robustness to sampling bias in models of species distributions: an implementation with Maxent. Ecol. Model. 222 (15), 2796–2811.
Anderson, R.P., 2013. A framework for using niche models to estimate impacts of climate change on species distributions. Ann. N. Y. Acad. Sci. 1297 (1), 8–28.
Araujo, M.B., Guisan, A., 2006. Five (or so) challenges for species distribution modelling. J. Biogeogr. 33 (10), 1677–1688.
Barbet-Massin, M., Jiguet, F., Albert, C.H., Thuiller, W., 2012. Selecting pseudo-absences for species distribution models: how, where and how many? Methods Ecol. Evol. 3 (2), 327–338.
Barry, S., Elith, J., 2006. Error and uncertainty in habitat models. J. Appl. Ecol. 43 (3), 413–423.
Beale, C.M., Lennon, J.J., 2012. Incorporating uncertainty in predictive species distribution modelling. Philos. Trans. Roy. Soc. B 367 (1586), 247–258.
Bean, W.T., Stafford, R., Brashares, J.S., 2012. The effects of small sample size and sample bias on threshold selection and accuracy assessment of species distribution models. Ecography 35 (3), 250–258.
Beaumont, L.J., Graham, E., Duursma, D.E., Wilson, P.D., Cabrelli, A., Baumgartner, J.B., Laffan, S.W., 2016. Which species distribution models are more (or less) likely to project broad-scale, climate-induced shifts in species ranges? Ecol. Model. 342, 135–146.
Boria, R.A., Olson, L.E., Goodman, S.M., Anderson, R.P., 2014. Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. Ecol. Model. 275, 73–77.
Buisson, L., Thuiller, W., Casajus, N., Lek, S., Grenouillet, G., 2010. Uncertainty in ensemble forecasting of species distribution. Glob. Change Biol. 16 (4), 1145–1157.
Brenning, A., 2005. Spatial prediction models for landslide hazards: review, comparison and evaluation. Nat. Hazards Earth Syst. Sci. 5 (6), 853–862.
Brenning, A., 2012. Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The R package sperrorest. In: Geoscience and Remote Sensing Symposium (IGARSS), 2012 IEEE International. IEEE, pp. 5372–5375.
Brown, J.L., Bennett, J.R., French, C.M., 2017. SDMtoolbox 2.0: the next generation Python-based GIS toolkit for landscape genetic, biogeographic and species distribution model analyses. PeerJ 5, e4095.
Carvalho, S.B., Brito, J.C., Crespo, E.G., Watts, M.E., Possingham, H.P., 2011. Conservation planning under climate change: toward accounting for uncertainty in predicted species distributions to increase confidence in conservation investments in space and time. Biol. Conserv. 144 (7), 2020–2030.
Cliff, A., Ord, J.K., 1981. Spatial Processes Models and Application. Pion Ltd.
Crase, B., Liedloff, A.C., Wintle, B.A., 2012. A new method for dealing with residual spatial autocorrelation in species distribution models. Ecography 35 (10), 879–888.
Crimmins, S.M., Dobrowski, S.Z., Mynsberge, A.R., 2013. Evaluating ensemble forecasts of plant species distributions under climate change. Ecol. Model. 266, 126–130.
Dhingra, M.S., Artois, J., Robinson, T.P., Linard, C., Chaiban, C., Xenarios, I., Von Dobschuetz, S., 2016. Global mapping of highly pathogenic avian influenza H5N1 and H5Nx clade 2.3. 4.4 viruses with spatial cross-validation. elife 5, e19571.
Dormann, C.F., 2007. Effects of incorporating spatial autocorrelation into the analysis of species distribution data. Glob. Ecol. Biogeogr. 16 (2), 129–138.
Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Münkemüller, T., 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 36 (1), 27–46.
Duan, R.Y., Kong, X.Q., Huang, M.Y., Fan, W.Y., Wang, Z.G., 2014. The predictive performance and stability of six species distribution models. PLoS One 9 (11), e112764.
Duque-Lazo, J., Van Gils, H.A., Groen, T.A., Navarro-Cerrillo, R.M., 2016. Transferability of species distribution models: The case of *Phytophthora cinnamomi* in Southwest Spain and Southwest Australia. Ecol. Model. 320, 62–70.
Elith, J., Anderson, R., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R., Loiselle, B., 2006. Novel methods improve prediction of species’ distributions from occurrence data. Ecography 29 (2), 129–151.
Elith, J., Leathwick, J.R., Hastie, T., 2008. A working guide to boosted regression trees. J. Anim. Ecol. 77 (4), 802–813.
Elith, J., Graham, C.H., 2009. Do they? How do they? Why do they differ? On finding reasons for differing performances of species distribution models. Ecography 32 (1), 66–77.
Elith, J., Kearney, M., Phillips, S., 2010. The art of modelling range-shifting species. Methods Ecol. Evol. 1 (4), 330–342.
Fabri-Ruiz, S., Saucède, T., Danis, B., David, B., 2017. Southern Ocean Echinoids database–An updated version of Antarctic, Sub-Antarctic and cold temperate echinoid database. ZooKeys 697, 1.
Ficetola, G.F., Thuiller, W., Miaud, C., 2007. Prediction and validation of the potential global distribution of a problematic alien invasive species—the American bullfrog. Divers. Distrib. 13 (4), 476–485.
Fielding, A.H., Bell, J.F., 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ. Conserv. 24 (1), 38–49.
Friedman, J., Hastie, T., Tibshirani, R., 2001. The elements of statistical learning, Vol. 1, No. 10. Springer series in statistics, New York.
Griffiths, H.J., Van de Putte, A.P., Danis, B., 2014. CHAPTER 2.2. Data distribution: Patterns and implications. In: De Broyer, C., Koubbi, P., Griffiths, H.J., Raymond, B., Udekem d’Acoz, C.d’. (Eds.), Biogeographic Atlas of the Southern Ocean. Scientific Committee on Antarctic Research, Cambridge, pp. 16–26.
Guillaumot, C., Martin, A., Eléaume, M., Saucède, T., 2018. Methods for improving species distribution models in data-poor areas: example of sub-Antarctic benthic species on the Kerguelen Plateau. Mar. Ecol. Prog. Ser. 594, 149–164.
Guillera-Arroita, G., Lahoz-Monfort, J.J., Elith, J., Gordon, A., Kujala, H., Lentini, P.E., Wintle, B.A., 2015. Is my species distribution model fit for purpose? Matching data and models to applications. Glob. Ecol. Biogeogr. 24 (3), 276–292.
Guisan, A., Thuiller, W., 2005. Predicting species distribution: offering more than simple habitat models. Ecol. Lett. 8 (9), 993–1009.
Guisan, A., Tingley, R., Baumgartner, J.B., Naujokaitis-Lewis, I., Sutcliffe, P.R., Tulloch, A.I., Martin, T.G., 2013. Predicting species distributions for conservation decisions. Ecol. Lett. 16 (12), 1424–1435.
Gutt, J., Zurell, D., Bracegridle, T., Cheung, W., Clark, M., Convey, P., Griffiths, H., 2012. Correlative and dynamic species distribution modelling for ecological predictions in the Antarctic: a cross-disciplinary concept. Polar Res. 31 (1), 11091.
Heikkinen, R.K., Marmion, M., Luoto, M., 2012. Does the interpolation accuracy of species distribution models come at the expense of transferability? Ecography 35 (3), 276–288.
Hijmans, R.J., Phillips, S., Leathwick, J., Elith, J., 2016. Package ‘dismo’. Available online at: http://cran.r-project.org/web/packages/dismo/index.html.
Hijmans, R.J., 2012. Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model. Ecology 93 (3), 679–688.
Hortal, J., Lobo, J.M., Jiménez-Valverde, A., 2012. Basic questions in biogeography and the (lack of) simplicity of species distributions: putting species distribution models in the right place. Natureza & Conservação 10 (2), 108–118.
Jiménez-Valverde, A., Lobo, J.M., Hortal, J., 2008. Not as good as they seem: the importance of concepts in species distribution modelling. Divers. Distrib. 14 (6), 885–890.
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 (4), 498–507.
Lawrence, J.M., 2013. Starfish: Biology and Ecology of the Asteroidae. JHU Press, Baltimore, pp. 267p.
Legendre, P., Fortin, M.J., 1989. Spatial pattern and ecological analysis. Vegetatio 80 (2), 107–138.
Li, W., Guo, Q., 2013. How to assess the prediction accuracy of species presence–absence models without absence data? Ecography 36 (7), 788–799.
Liu, C., White, M., Newell, G., 2013. Selecting thresholds for the prediction of species occurrence with presence-only data. J. Biogeogr. 40 (4), 778–789.
Mainali, K.P., Warren, D.L., Dhileepan, K., McConnachie, A., Strathie, L., Hassan, G., Parmesan, C., 2015. Projecting future expansion of invasive species: comparing and improving methodologies for species distribution modeling. Glob. Change Biol. 21 (12), 4464–4480.
McClintock, J.B., Angus, R.A., Ho, C., Amsler, C.D., Baker, B.J., 2008. A laboratory study of behavioral interactions of the Antarctic keystone sea star *Odontaster validus* with three sympatric predatory sea stars. Mar. Biol. 154 (6), 1077–1084.
Merow, C., Smith, M.J., Silander, J.A., 2013. A practical guide to MaxEnt for modeling species’ distributions: what it does, and why inputs and settings matter. Ecography 36 (10), 1058–1069.
Moreau, C., Mah, C., Agüera, A., Améziane, N., Barnes, D., Crokaert, G., Jażdżewska, A.,

2018. Antarctic and sub-Antarctic Asteroidea database. ZooKeys 747, 141.
Muscarella, R., Galante, P.J., Soley-Guardia, M., Boria, R.A., Kass, J.M., Uriarte, M., Anderson, R.P., 2014. ENMeval: An R package for conducting spatially independent evaluations and estimating optimal model complexity for Maxent ecological niche models. Methods Ecol. Evol. 5 (11), 1198–1205.
Naimi, B., Hamm, N.A., Groen, T.A., Skidmore, A.K., Toxopeus, A.G., 2014. Where is positional uncertainty a problem for species distribution modelling? Ecography 37 (2), 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 (1), 3–22.
Olden, J.D., Jackson, D.A., Peres-Neto, P.R., 2002. Predictive models of fish species distributions: a note on proper validation and chance predictions. Trans. Am. Fish. Soc. 131 (2), 329–336.
Owens, H.L., Campbell, L.P., Dornak, L.L., Saupe, E.E., Barve, N., Soberón, J., Peterson, A.T., 2013. Constraints on interpretation of ecological niche models by limited environmental ranges on calibration areas. Ecol. Model. 263, 10–18.
Peterson, A.T., Soberón, J., Pearson, R.G., Anderson, R.P., Martínez-Meyer, E., Nakamura, M., Araújo, M.B., 2011. Ecological Niches and Geographic Distributions (MPB-49). Princeton University Press.
Phillips, S.J., Dudík, M., Elith, J., Graham, C.H., 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 (1), 181–197.
Phillips, N.D., Reid, N., Thys, T., Harrod, C., Payne, N.L., Morgan, C.A., Houghton, J.D., 2017. Applying species distribution modelling to a data poor, pelagic fish complex: The ocean sunfishes. J. Biogeogr. 44 (10), 2176–2187.
Pierrat, B., Saucède, T., Laffont, R., De Ridder, C., Festeau, A., David, B., 2012. Large-scale distribution analysis of Antarctic echinoids using ecological niche modelling. Mar. Ecol. Prog. Ser. 463, 215–230.
Qiao, H., Soberón, J., Peterson, A.T., 2015. No silver bullets in correlative ecological niche modelling: insights from testing among many potential algorithms for niche estimation. Methods Ecol. Evol. 6 (10), 1126–1136.
R Core Team, 2017. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, pp. 2016.
Radosavljevic, A., Anderson, R.P., 2014. Making better Maxent models of species distributions: complexity, overfitting and evaluation. J. Biogeogr. 41 (4), 629–643.
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.
Ripley, B. (2015). MASS: Support Functions and Datasets for Venables and Ripley's MASS. 2015. https://CRAN. R-project. org/package= MASS. R package version, pp. 7–3.
Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Warton, D.I., 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40 (8), 913–929.
Robinson, L.M., Elith, J., Hobday, A.J., Pearson, R.G., Kendall, B.E., Possingham, H.P., Richardson, A.J., 2011. Pushing the limits in marine species distribution modelling: lessons from the land present challenges and opportunities. Glob. Ecol. Biogeogr. 20 (6), 789–802.
Segurado, P.A., Araújo, M.B., Kunin, W.E., 2006. Consequences of spatial autocorrelation for niche-based models. J. Appl. Ecol. 43 (3), 433–444.
Tessarolo, G., Rangel, T.F., Araújo, M.B., Hortal, J., 2014. Uncertainty associated with survey design in Species Distribution Models. Divers. Distrib. 20 (11), 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.
Torres, L.G., Sutton, P.J., Thompson, D.R., Delord, K., Weimerskirch, H., Sagar, P.M., Phillips, R.A., 2015. Poor transferability of species distribution models for a pelagic predator, the grey petrel, indicates contrasting habitat preferences across ocean basins. PLoS One 10 (3), e0120014.
Valavi, R., Elith, J., Lahoz-Monfort, J.J., Guillera-Arroita, G., 2018. blockCV: an R package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. bioRxiv, 357798.
van Proosdij, A.S., Sosef, M.S., Wieringa, J.J., Raes, N., 2016. Minimum required number of specimen records to develop accurate species distribution models. Ecography 39 (6), 542–552.
Veloz, S.D., 2009. Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. J. Biogeogr. 36 (12), 2290–2299.
Wenger, S.J., Olden, J.D., 2012. Assessing transferability of ecological models: an underappreciated aspect of statistical validation. Methods Ecol. Evol. 3 (2), 260–267.