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

Charlène Guillaumot*, Alexis Martin, Marc Eléaume, Thomas Saucède

*Corresponding author: charleneguillaumot21@gmail.com

Supplement 1: Evaluation and choice of the model

In the present work, we ran ensemble models as a decision tool to select algorithms that are the most appropriate to the type of data to be analysed (Scales et al. 2016). The performance of 10 different algorithms was compared using the default parametrization settings proposed in the 'biomod2' R package (see Thuiller et al. (2016) for calibration details and Marmion et al. (2009) for modeling documentation). The compared algorithms include Artificial Neural Network (ANN), Boosted Regression Trees (BRT), Classification Tree Analysis (CTA), Flexible Discriminant Analysis (FDA), Generalized Additive Model (GAM), Generalized Linear Models (GLM), Multivariate Adaptive Regression Splines (MARS), Maximum Entropy (MaxEnt), Random Forest (RF), and Surface Range Envelope (SRE).

Two analyses were realised to compare the respective performance of the models. First, for each algorithm, AUC values of 100 model replicates were computed. Models were performed using all occurrence data available for the species *Ctenocidaris nutrix* and *Sterechinus diadema* only (Fig S1A, S1C) because there were not enough data to perform the analysis for *Abatus cordatus* and *Brisaster antarcticus*.

In a second step, standard deviation of the 100 replicates were compared between models as the number of data was progressively increased between runs to represent the improvement of sampling effort through time (Fig S1B, S1D).

Presence-only records associated to non-informative environmental data (NA/, no data values) were removed as required to perform the *biomod2* analysis. Occurrence duplicates located on one single 0.1° grid cell were removed to reduce spatial weighting. 200 pseudo-absences were selected to perform the analysis.

Results show that Boosted Regression Trees (BRT) and Random Forest (RF) are the algorithms that perform best to model the distribution of *C. nutrix* and *S. diadema* (Fig. S1), with relatively stable (SD < 0.025) and high AUC values varying between [0.976,1] and [0.994,1] respectively of the analysis that studies data addition. Unexpectedly, algorithms previously shown to be well suited to presence-only data and small datasets (e.g. SRE or MaxEnt, see Araújo and Peterson 2012, Yackulic et al. 2013) did not perform well in our case study. Low performances of SRE have already been reported (Elith et al. 2006). The low number of pseudo-absences used to calibrate the model could explain the low performance of MaxEnt (Barbet-Massin et al. 2012, Phillips and Dudik 2008).
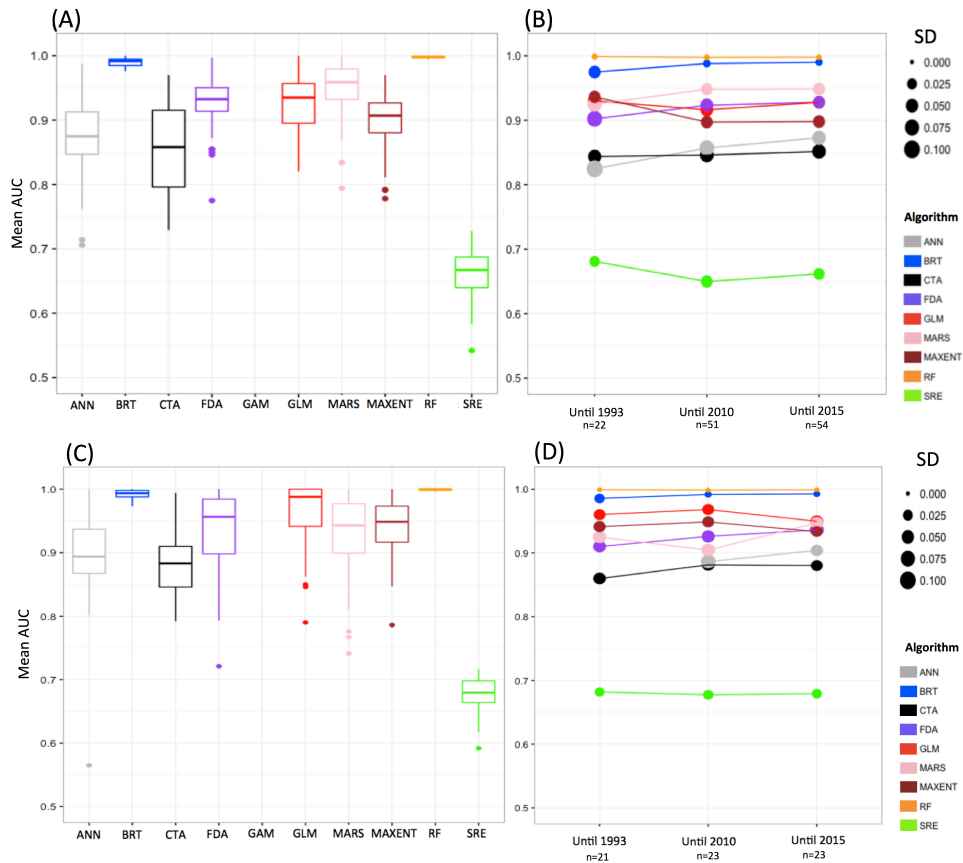
Figure S1: Compared performances of the different models for the species *Ctenocidaris nutrix* (A-B) and *Sterechinus diadema* (C-D). (A, C) Mean AUC values of model replicates for each algorithm. (B, D) Variation of mean AUC values and Standard Deviation (SD) of model replicates with data addition (n=22, n=51 or n=54 occurrences for *C. nutrix* and n=21, n= 23 and n=23 for *S. diadema*). For each analysis, 200 background data were randomly sampled in the studied area. Environmental descriptors correspond to [1955-2012].

The respective performance of BRT and RF (Fig. S1) was tested for spatial transferability following a non-random three-fold cross-validation procedure (Fig. S2, Wenger and Olden 2012). Model transferability is defined as the "extrapolative accuracy" of a model that is, the model ability to extrapolate in space and time (Randin et al. 2006, Wenger and Olden 2012). Three models were computed simultaneously using three different subsets of occurrences for *C. nutrix* (Fig. S2) alternatively used as training and test data (50 replicates). The three averaged models were compared with each others using the Schoener's D similarity index. D mean and standard deviation values were computed for all comparisons. All analyses were performed using time-averaged environmental parameters for the total period under study [1955-2012]. We considered that the most similar the distribution maps are the better the transferability performance is (Fig. S2).
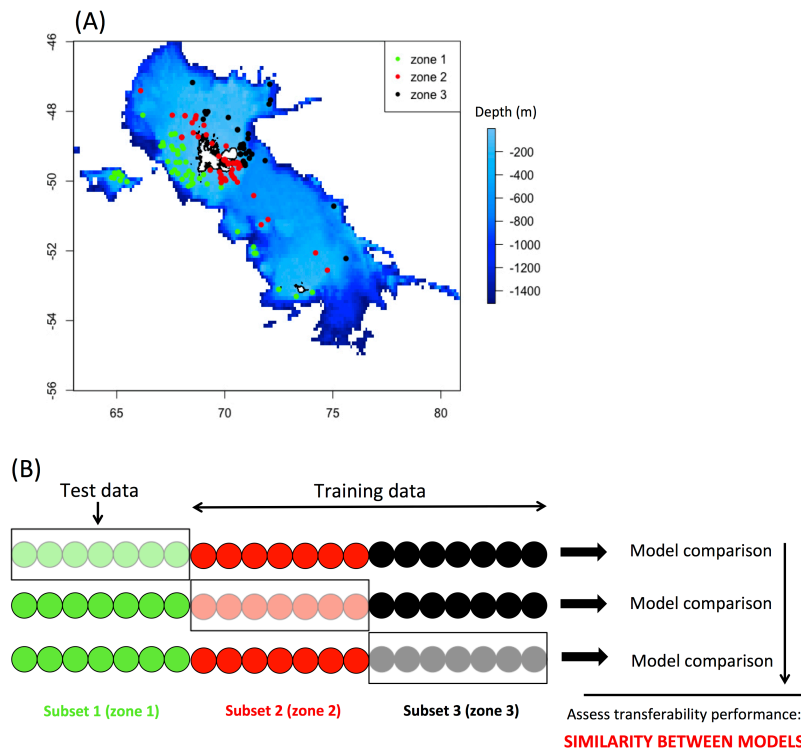
Figure S2: (A) Map showing the distribution of presence-only data in the three subsets defined for the cross-validation procedure. (B) Non-random three-fold cross-validation procedure performed to test for the transferability performance of models. Zones 1, 2 and 3 refere to (A).

Comparison between maps shows higher similarity values between the different models run with BRT (Schoener's $\overline{D}$=0.867± 0.034) than with RF ($\overline{D}$=0.761± 0.036), which highlights that BRT performs best for spatial transferability. Because transferability performance is a central criterion of model selection in our study (Araújo and Guisan 2006, Wenger and Olden 2012), BRT was selected for the further analyses. This result is in line with previous studies that highlight the high performance of BRT for prediction (Elith and Graham 2009, Guo et al. 2015) and transferability (Heikkinen et al. 2012, Wenger and Olden 2012, Crimmins et al. 2013) while RF has been shown to generate geographically restricted models with high accuracy (Guo et al. 2015, Qiao et al. 2015, Beaumont et al. 2016).

Table S1: Overall comparison between Random Forest, Boosted Regression Tree (BRT) and MaxEnt respective performances with reference works.

	Random Forest	**BRT**	**MaxEnt**
DATASET			
● *Type of data*	Presence-only/ presence-absence	Presence-only/ presence-absence	Presence-only/ presence-absence
● *Missing biological values*	Interpolation required first (Breiman 2001)	allowed	allowed
● *Categorical descriptors*	Biased if different categorical levels (Duan et al. 2014)	allowed	allowed
Robustness to spatially biased data	**More sensible than BRT to patchy patterns** (Marmion et al. 2009, Barbet-Massin et al. 2012) More adapted to bias correction methods than BRT (Barbet-Massin et al. 2012)		Not adapted (Royle et al. 2012) Unstable predictions (this study)
Overall modelling performance	**High performance and interpolation accuracy** (Wenger and Olden 2012, Guo et al. 2015) Biological responses often unrealistic (Beaumont et al. 2016)	Medium performance (Qiao et al. 2015) Performed better than RF in previous works on benthic marine species (Reiss et al. 2011)	High performance even with complex environmental interactions (Elith et al. 2011)
● *Transferability performance*	**Poor** (Wenger and Olden 2012, Crimmins et al. 2013).	**Good** (Heikkinen et al. 2012)	One of the highest (Heikkinen et al. 2012, Duque-Lazo et al. 2016)
● *Extrapolation performance*	**Not suitable** (Qiao et al. 2015, Beaumont et al. 2016) due to **overfitting** (Wenger and Olden 2012, Aguirre-Gutiérrez et al. 2013)	Good (Heikkinen et al. 2012). **High prediction performance** (Elith et al. 2006, Elith and Graham 2009, Guo et al. 2015)	Perform worse than BRT (this study) Tend to overpredict (Duan et al. 2014)
Required computation time	Long (Elith and Graham 2009, García-Callejas and Araújo 2016)	Medium (this study)	Medium (this study)

Supplement 2: Spatial autocorrelation (SAC)

Table S2: Moran I SAC index computed from mean residuals of the 100 model replicates and the associated significance for each species before and after spatial bias correction.

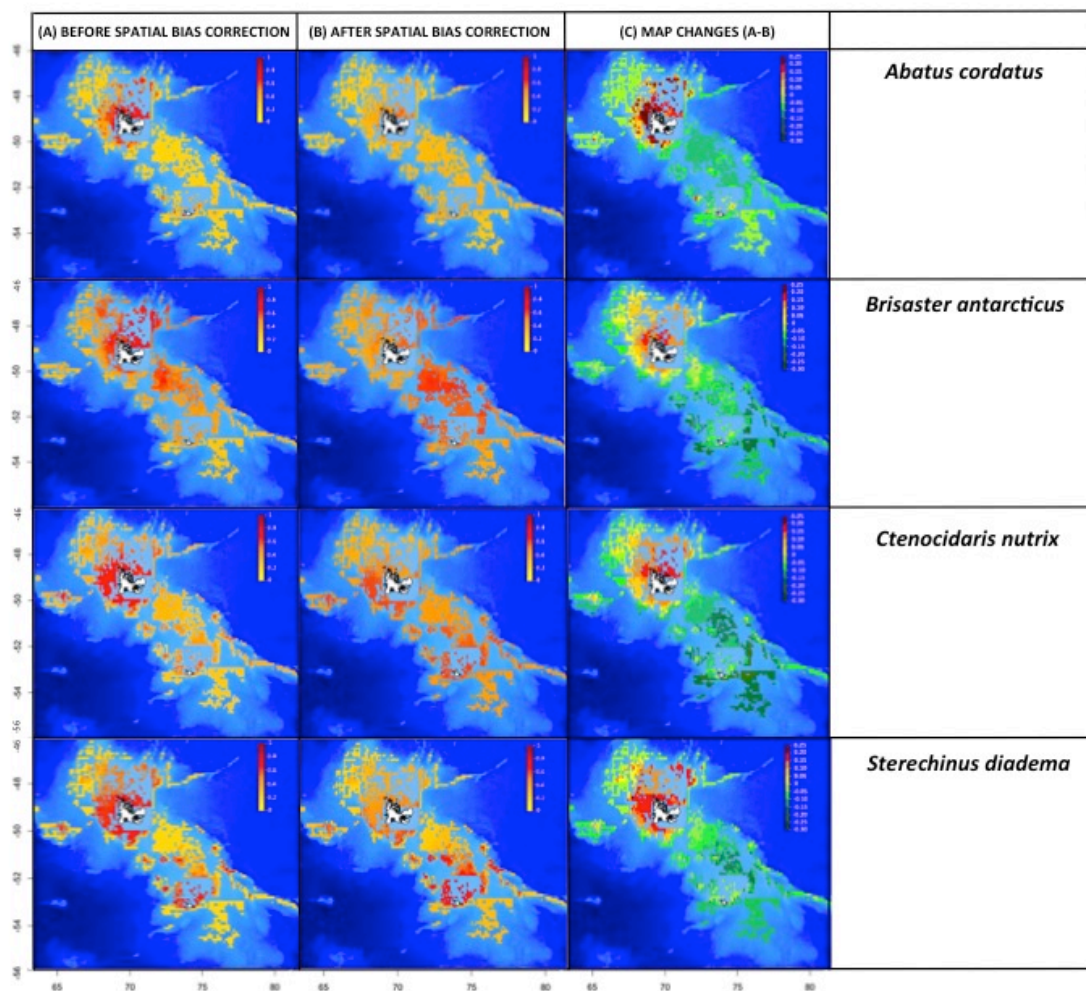|	Before correction		After correction	
	I_{obs}	p-value	I_{obs}	p-value
Abatus cordatus	0.16	1.19e-9	0.06	5.85e-4
Brisaster antarcticus	0.05	0.04	0.04	0.08
Ctenocidaris nutrix	0.07	7.37e-8	0.01	0.17
Sterechinus diadema	0.06	3.90e-3	0.02	0.13

Figure S3: Maps showing species distribution models computed before and after correcting for spatial bias by background sampling.

Supplement 3: Testing the influence of chronological addition of occurrences
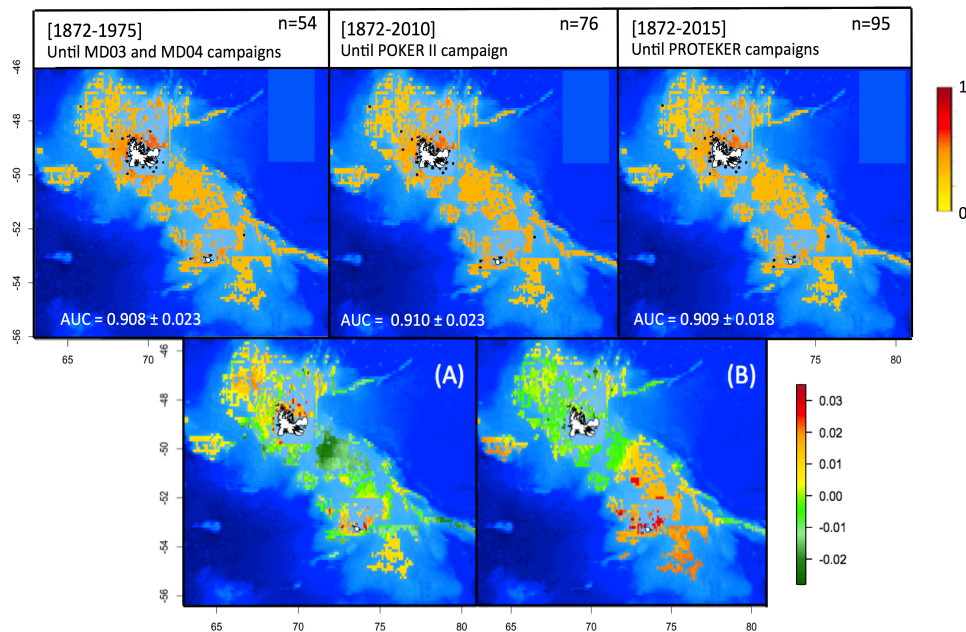
Figure S4: First row: distribution models of *Abatus cordatus* with increasing number of occurrences. Averaged maps of 100 model replicates. Second row: (A) Difference in probability distribution between n=76 and n=54, (B) between n=95 and n=76.
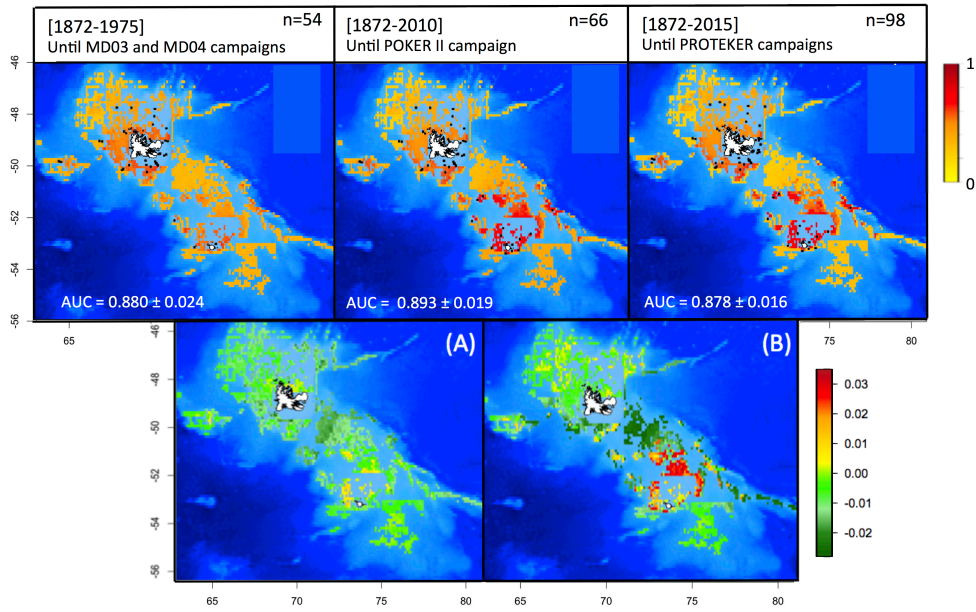
Figure S5: First row: distribution models of *Sterechinus diadema* with increasing number of occurrences. Averaged maps of 100 model replicates. Second row: (A) Difference in probability distribution between n=66 and n=54, (B) between n=98 and n=66.

Table S3: Effect of chronological addition of new data on model performance. Comparison between distribution maps. Upper diagonal: Schoener's D correlation between averaged maps. Lower diagonal: significance p-value of the associated Schoener's D correlation.

Abatus cordatus	←1975] n=54	-	←2010] n=76	←2015] n=95
←1975] n=54	-	-	0.972±0.025	0.980±0.021
-	-	-	-	-
←2010] n=76	0.002	-	-	0.981±0.023
←2015] n=95	0	-	0	-
Ctenocidaris nutrix	←1975] n=46	←1993] n=54	←2010] n=106	←2015] n=114
←1975] n=46	-	0.964±0.026	0.969±0.020	0.967±0.020
←1993] n=54	0.017	-	0.960±0.020	0.961±0.020
←2010] n=106	0.005	0.037	-	0.988±0.013
←2015] n=114	0.010	0.028	0	-
Sterechinus diadema	←1975] n=54	-	←2010] n=66	←2015] n=98
←1975] n=54	-	-	0.930±0.030	0.928±0.037
-	-	-	-	-
←2010] n=66	0.369	-	-	0.937±0.042
←2015] n=98	0.411	-	0.262	-

References

Aguirre-Gutiérrez J, Carvalheiro LG, Polce C, van Loon EE and others (2013) Fit-for-purpose: species distribution model performance depends on evaluation criteria–Dutch hoverflies as a case study. PloS one. 8: e63708.

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

Araújo MB and Peterson AT (2012) Uses and misuses of bioclimatic envelope modeling. Ecology. 93: 1527–1539.

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

Beaumont LJ, Graham E, Duursma DE, Wilson PD and others (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.

Breiman L (2001) Random forests. Machine learning. 45: 5–32.

Crimmins SM, Dobrowski SZ and Mynsberge AR (2013) Evaluating ensemble forecasts of plant species distributions under climate change. Ecol. Model. 266: 126–130.

Duan RY, Kong XQ, Huang MY, Fan WY and Wang ZG (2014) The predictive performance and stability of six species distribution models. PloS one. 9: e112764.

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

Elith JH, Graham R, Anderson M, Dudík S and others (2006) Novel methods improve prediction of species' distributions from occurrence data. Ecography. 29: 129–151.

Elith J and Graham CH (2009) Do they? How do they? Why do they differ? On finding reasons for differing performances of species distribution models. Ecography. 32: 66–77.

Elith J, Phillips SJ, Hastie T, Dudík M, Chee YE and Yates CJ (2011) A statistical explanation of MaxEnt for ecologists. Divers. Distrib. 17: 43–57.

García-Callejas D and Araújo MB (2016) The effects of model and data complexity on predictions from species distributions models. Ecol. Model. 326: 4–12.

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

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

Hijmans RJ, Phillips S, Leathwick J and Elith J (2016) dismo: Species Distribution Modeling. R package version 1.1-1. https://CRAN.R-project.org/package=dismo.

Marmion M, Luoto M, Heikkinen RK and Thuiller W (2009) The performance of state-of-the-art modelling techniques depends on geographical distribution of species. Ecol. Model. 220: 3512–3520.

Merow C, Smith M and Silander J (2013) A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. Ecography. 36: 1058–1069.

Phillips S and Dudík M (2008) Modeling of species distributions with MaxEnt: new extensions and a comprehensive evaluation. Ecography. 31: 161–175.

Qiao H, Soberón J and Peterson AT (2015) No silver bullets in correlative ecological niche modelling: insights from testing among many potential algorithms for niche estimation. Method. Ecol. Evol. 6: 1126–1136.

Randin CF, Dirnböck T, Dullinger S, Zimmermann NE and others (2006) Are niche-based species distribution models transferable in space? J. Biogeogr. 33: 1689–1703.

Reiss H, Cunze S, König K, Neumann H and Kröncke I (2011) Species distribution modelling of marine benthos: a North Sea case study. Mar. Ecol. Prog. Ser. 442: 71–86.

Royle JA, Chandler RB, Yackulic C and Nichols JD (2012) Likelihood analysis of species occurrence probability from presence- only data for modelling species distributions. Method. Ecol. Evol. 3: 545–554.

Scales KL, Miller PI, Ingram SN, Hazen EL, Bograd SJ and Phillips RA (2016). Identifying predictable foraging habitats for a wide-ranging marine predator using ensemble ecological niche models. Divers. Distrib. 22: 212-224.

Thuiller W, Georges D, Engler R and Breiner F (2016) biomod2: Ensemble Platform for Species Distribution Modeling. R package version 3.3-7. https://CRAN.R-project.org/package=biomod2.

Yackulic C, Chandler R, Zipkin E, Royle J and others (2013) Presence-only modelling using MaXent: when can we trust the inferences? Meth. Ecol. Evol. 4: 236-243.

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