Supporting Information for "Measuring the relative effect of factors affecting species distribution model predictions"

Emeric Thibaud,[1*] Blaise Petitpierre,[2] Olivier Broennimann,[2]
Anthony C. Davison,[1] and Antoine Guisan[2,3]

[1]Chair of Statistics, Ecole Polytechnique Fédérale de Lausanne,
EPFL-FSB-MATHAA-STAT, Station 8, 1015 Lausanne, Switzerland
[2]Department of Ecology and Evolution, University of Lausanne, 1015 Lausanne, Switzerland
[3]Institute of Earth Surface Dynamics, University of Lausanne, 1015 Lausanne, Switzerland

[*]To whom correspondence should be addressed; E-mail: emeric.thibaud@epfl.ch.

Contents

S1 Data

Our simulation uses a real Alpine landscape in the western Swiss Alps ($7°2'$–$7°14'$ E; $46°28'$–$46°31'$ N). This landscape represents an area of about 700 km^2 of the canton of Vaud, comprising $1,127,599$ pixels at a resolution size of 25 meters. Plants have been intensively sampled in this region (Pottier *et al.*, 2013) and a large set of environmental maps has been assembled (see http://www.unil.ch/rechalpvd). This helps in providing realistic estimates for parameters for the simulation. We used five real climatic and topographic predictors known for their importance in shaping plant distributions (Zimmermann & Kienast, 1999; Randin *et al.*, 2006), labelled $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_5$:

- the annual sum of degree days above three degrees (*ddeg300*);
- the annual average number of frost days during the growing season (*sfroyy*) from existing maps (Zimmermann & Kienast, 1999);
- a moisture index between June and August (*mind68*);
- daily average global potential shortwave radiation per month (*sumradyy*); and
- topographic position (*topos*, Randin *et al.*, 2006).

We simulated our virtual species based on real species data. The species dataset comprises presence-absence records at 912 locations for 260 plant species. Our simulation was based on parameters from ten representative species in the study area (Table S1). These are herbs, growing from the lowland to the alpine levels with different ecological tolerances. No trees or shrubs were included because forestry and pasturing may have truncated their distribution, especially at higher altitudes. These species were used as realistic references for the α and β parameters of the probit models used to build our virtual species distributions, and were also used to estimate a realistic spatial autocorrelation coefficient; see §S4.

S2 MaxEnt

For modelling presence-absence data in the training samples, we used a maximum entropy (MaxEnt, Phillips *et al.*, 2006) model fitted using the R package dismo (Hijmans *et al.*, 2013) with default settings. MaxEnt models presence-only data using background data (Elith *et al.*, 2011; Renner & Warton, 2013; Merow *et al.*, 2013). The relationship between the data and the predictors is defined through functions called "features" (Phillips *et al.*, 2006; Phillips & Dudík, 2008) that control the shape of the response the model can produce. MaxEnt uses five classes of features for continuous predictors and, by default, those used depend on the number of presence records; when this is larger than 15, MaxEnt uses linear, quadratic and hinge features, while threshold and product (which models interactions between predictors) features are only used when there are over 80 presence records (Phillips & Dudík, 2008). In our application, the number of presences in the training samples varied between 17 and 78 for $n = 100$, and 105 and 379 for $n = 500$, so that interactions between predictors were modelled only for the larger sample size. Renner & Warton (2013) show that MaxEnt is equivalent to Poisson regression, and link it to a Poisson point process model, thus giving insight into how MaxEnt deals with presence-only data. Because absence data were available to us, we used MaxEnt in a non-standard manner, using the absence data in place of a background sample. We used the logistic output of MaxEnt (Yackulic *et al.*, 2013) which under general SDM assumptions (Guisan & Thuiller, 2005; Elith & Leathwick, 2009; Araújo & Peterson, 2012) and algorithm-specific assumptions (Hastie & Fithian, 2013), estimate the true presence probabilities p_i.

S3 The root mean squared error

Many measures of model accuracy (Fielding & Bell, 1997; Caruana & Niculescu-Mizil, 2004; Liu *et al.*, 2011) can be used in SDMs. Because in practice the true probabilities of presence p_i are unknown, accuracy measures used in SDMs compare the observed presences and absences in the test sample with the predicted probabilities $\hat{p}_i$. Two popular measures are the area under the receiver operating characteristic curve (AUC, Mason & Graham, 2002) and the point-biserial correlation (COR, Tate, 1954).

For our simulation, the true distribution of the virtual species is known, which suggests comparing the p_i with the predictions $\hat{p}_i$. We used the root mean squared error (RMSE, Caruana & Niculescu-Mizil, 2004; Liu *et al.*, 2011):

$$\text{RMSE} = \sqrt{\frac{1}{n'} \sum_{i \in \mathcal{T}} (\hat{p}_i - p_i)^2}.$$

The RMSE has a simple interpretation as the mean distance between predicted and true probabilities over the locations of the test sample. The decomposition of the expectation of its square as

$$\mathrm{E}\left\{ \frac{1}{n'} \sum_{i \in \mathcal{T}} (\hat{p}_i - p_i)^2 \right\} = \frac{1}{n'} \sum_{i \in \mathcal{T}} \left[\mathrm{var}(\hat{p}_i) + \{\mathrm{E}(\hat{p}_i) - p_i\}^2 \right],$$

shows that it penalizes both the variability and the bias of the estimator.

Because the RMSE uses only the true p_i, it corresponds to the accuracy related to the estimation of the fundamental niche of the species, excluding possible undesirable effects of SAC at the presence-absence level. In practice, when using cross-validation and the AUC or COR, modelers must use independent training and test samples to avoid bias in estimation of prediction accuracy. In presence of SAC, modelers must use well-separated pixels for the training and test samples in order to reduce the correlation between these two samples.

The AUC and COR can be calculated in our simulation framework by generating presences and absences at each location of the test sample based on the true p_i. Two approaches to generate these presences and absences are possible. One approach generates correlated presences and absences using correlated normal variables that mimic the SAC present in the training sample. To avoid bias in the estimation of these accuracy measures, these presences and absences must be independent of those in the training sample. This approach corresponds to estimation on a test sample independent of the training sample but within which the records are dependent because of SAC. A second approach uses independent normal variables to generate independent presences and absences without SAC. The degrees of freedom of the test sample for the first approach are smaller than for the second approach and the variability of the measure of prediction accuracy is larger. Compared to the RMSE, which uses the true, non-random, p_i, measures based on simulated presences and absences are more variable.

In our application, the distribution of log(RMSE) was found to be approximately normal, which was appropriate for the use of the linear mixed-effect model and the analysis of variance. We also calculated the AUC and COR but their distributions makes further analysis more complicated: a transformation that makes their distributions approximately normal is required, and this does not seem straightforward. The use of measures other than log(RMSE) within our framework seems quite feasible, at least in principle, but requires further investigation.

S4 Estimating spatial autocorrelation

Here we discuss the choice of the strength of spatial autocorrelation (SAC) corresponding to a dispersal process. To determine a sensible value for the range parameter λ chosen in our simulation (see main article), we estimated the range of spatial autocorrelation on the real species we used (see Section S1) by fitting a spatial probit model similar to those used to simulate the virtual species.

Given predictors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_5$ we let $\boldsymbol{X}$ denote the $N \times 5$ matrix of predictors, in which each row corresponds to a different location. We use the probit function to relate the predictors to the presence probabilities, i.e.,

$$p(\boldsymbol{X}) = \Phi\{\eta(\boldsymbol{X})\}, \tag{S1}$$

where Φ is the standard Gaussian cumulative distribution function. We include quadratic terms for each predictor and we do not consider interactions, so that

$$\eta(\boldsymbol{X}) = \alpha_0 + \sum_{j=1}^{5} \alpha_j (\boldsymbol{x}_j - \beta_j)^2$$

with coefficients α_j and β_j to be estimated. The standard GLM probit model assumes independence of the outcomes (in our case, presences–absences) but here we consider a spatial model in which the marginal distributions are unchanged but presences and absences tend to cluster.

In the standard probit model, presences and absences are obtained by generating independent standard normal variables $\gamma_1, \ldots, \gamma_N$ for the N locations, writing η_i for the linear predictor at the ith location, and defining the binary responses at those locations by

$$Y_i = \begin{cases} 1, & \gamma_i \leq \eta_i, \\ 0, & \gamma_i > \eta_i. \end{cases} \tag{S2}$$

The spatial probit model considered here presupposes that presences and absences are obtained using correlated γ_i in (S2): we replace the independent γ_i by a Gaussian process $\gamma(\boldsymbol{s})$ with zero mean, unit variance and correlation function ρ, where $\boldsymbol{s}$ is the location of the record. The range of SAC can be precisely controlled, as the parameters of the correlation function ρ control the range of dependence. The marginal probability of presence at location $\boldsymbol{s}$ is still defined from Equation (S1) but now the presences–absences tend to cluster. Pairwise marginal distributions are given by

$$\Pr(Y_i = y_i, Y_j = y_j) = \begin{cases} \Phi_2(\eta_i, \eta_j; \rho_{ij}), & y_i = 1, y_j = 1, \\ \Phi(\eta_i) - \Phi_2(\eta_i, \eta_j; \rho_{ij}), & y_i = 1, y_j = 0, \\ \Phi(\eta_j) - \Phi_2(\eta_i, \eta_j; \rho_{ij}), & y_i = 0, y_j = 1, \\ 1 - \Phi(\eta_i) - \Phi(\eta_j) + \Phi_2(\eta_i, \eta_j; \rho_{ij}), & y_i = 0, y_j = 0, \end{cases} \tag{S3}$$

where $\Phi_2(\ ,\ ; \rho)$ denotes the bivariate standard Gaussian cumulative distribution function with correlation ρ. Here we use $\rho_{ij} = \exp(-h_{ij}/\lambda)$, where h_{ij} is the distance (in km) between locations $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ and λ is a range parameter to be chosen.

We use a composite likelihood approach to estimate the parameters of the spatial probit model. Pairwise distributions (S3) are used to construct the pairwise log-likelihood (Varin *et al.*, 2011) function

$$\ell(\boldsymbol{\psi}) = \sum_{1 \leq i < j \leq N} w_{ij} \log \Pr(Y_i = y_i, Y_j = y_j), \tag{S4}$$

where the sum is over all distinct pairs (i, j), the w_{ij} are weights to be chosen, and $\boldsymbol{\psi}$ is the vector of model parameters. Properties of composite likelihood estimation are discussed by Varin *et al.* (2011). In particular, under similar assumptions to that of classical likelihood estimation, it can be shown that the maximum composite likelihood $\hat{\boldsymbol{\psi}} = \operatorname{argmax}_{\boldsymbol{\psi}} \ell(\boldsymbol{\psi})$ estimator is consistent and that its distribution is asymptotically normal, with mean $\boldsymbol{\psi}$ and covariance matrix of the standard "sandwich" form (Varin *et al.*, 2011). Using composite likelihood estimation instead of the full likelihood usually results in a loss in efficiency. No closed form expression exist for $\hat{\boldsymbol{\psi}}$ but the pairwise log-likelihood function $\ell(\boldsymbol{\psi})$ can be maximized using the R function optim. Because the number of pairs can be large, maximizing this likelihood can be burdensome, so we used two shortcuts to accelerate the optimization. First, we estimated the regression parameters in the η_i by using a probit model that wrongly assumes that the responses are all independent (i.e., a GLM with a probit link). Such estimators are consistent, but their variances are mis-estimated. The estimates of η_i are then plugged into the likelihood (S4), so that then only only the dependence parameter λ need be estimated. Second, to reduce the number of terms involved, we excluded some pairs, by setting $w_{ij} = 0$ for pixels more than 2.4 km apart.

We use the spatial probit model and pairwise likelihood estimates to find estimates of λ for each species. The results are shown in Table S1. These estimates are sensitive to the choice of the subset of pairs used in the likelihood, so we chose to fix $\lambda = 0.5$ in our simulation, though the estimated values of λ are all smaller than 0.39. Moreover, choosing $\lambda = 0.5$ implies correlation lower than 0.05 at distances over 1.5 km. Similar species dispersal values have been found by Vittoz & Engler (2007).

Table S1: Names of the real species used for the simulation and estimated values of spatial autocorrelation using a spatial probit model.

	Species name	$\hat{\lambda}$	Effective range (km)
1	Festuca pratensis sl.	0.13	0.39
2	Prunella vulgaris	0.27	0.79
3	Veronica chamaedrys	0.19	0.57
4	Taraxacum officinale aggr.	0.03	0.08
5	Plantago lanceolata	0.39	1.18
6	Cerastium fontanum sl.	0.02	0.05
7	Agrostis capillaris	0.26	0.79
8	Alchemilla xanthochlora aggr.	0.30	0.90
9	Leontodon hispidus sl.	0.23	0.68
10	Festuca rubra aggr.	0.29	0.88

S5 Varying the strength of spatial autocorrelation

We investigate how the conclusions of the main paper on the relative importance of the five factors change when varying the strength of SAC in the dispersal process. We choose different values of the range parameter λ of the correlation function of the underlying Gaussian process used to mimic a dispersal process, representing weaker or stronger SAC at the presence–absence simulation level. We then calculate the marginal R^2 (Nakagawa & Schielzeth, 2013) for the full model and each sub-model to determine the relative importance of the factors. The marginal R^2 values obtained are shown in Table S2. For a Gaussian process having an effective range of 5 km, the effect of dispersal is comparable to that of sampling design, for 10 km it is comparable to that of modeling technique and for 15 km it is comparable to that of sample size n. Not surprisingly, increasing the strength of SAC for the dispersal process increases its relative effect on prediction accuracy, and it becomes one of the most important factors for large SAC, but such large dispersal values are unrealistic for the species considered in our study, see Table S1 and Vittoz & Engler (2007).

Table S2: Marginal R^2 for the full model (4) and the five sub-models with one factor (and all its interactions with the other factors) excluded at a time, and for various strengths of spatial autocorrelation for the dispersal process. Different values for the range parameter λ correspond to different effective ranges (distance after which the correlation drops below 5%).

Effective range (km)	λ	full model	−missing	−dispersal	−n	−design	−technique
1.5	0.5	0.674	0.662	0.669	0.188	0.587	0.423
5	1.67	0.644	0.632	0.566	0.233	0.559	0.412
10	3.34	0.651	0.641	0.446	0.315	0.579	0.428
15	5	0.670	0.661	0.372	0.381	0.612	0.451

S6 External validation

We investigated the ranking of the factors in terms of the accuracy of model predictions estimated on different landscapes. We measured the accuracy of predictions obtained by the different combinations of the factors using the RMSE calculated by predicting species distributions in other regions where different correlations among the predictors occur. In the main paper, we used the landscape VD of the Vaud Alps and ten real species present in this region. For external validation we use three other regions of Switzerland where the same species are present but that present different topo-climatic conditions: the lower Engadine (EN), the canton of Neuchâtel (NE), and the south part of the canton of Ticino (TI). Table S3 summarises the main characteristics of the four regions. We use the same climatic and topographic predictors as for VD, with the same pixel resolution of 25 meters.

Table S3: Coordinates, number of pixels, and mean values for the predictors for the four regions of Switzerland used in our simulation (VD, EN, NE and TI).

	Pixels	Coordinates	ddeg300	sfroyy	mind68	sumradyy	topos
VD	$1,127,599$	$7°2'$–$7°14'$ E; $46°28'$–$46°31'$ N	1788	21	177	199866	-6.2
EN	$1,597,016$	$9°56'$–$10°29'$ E; $46°36'$–$47°0'$ N	735	204	244	194969	-3.8
NE	$1,146,942$	$6°25'$–$7°5'$ E; $46°50'$–$47°9'$ N	1861	15	-176	217983	0.8
TI	$874,996$	$8°45'$–$9°9'$ E; $45°49'$–$46°11'$ N	2693	4	233	199602	-0.8

The theoretical probability maps of presence for the ten virtual species in the landscapes EN, NE and TI are obtained via the probit models used to generate the virtual species in VD. Figures S10–S19 show the maps of presence probabilities for the ten virtual species in the four landscapes.

Predictions from modeling techniques are obtained using exactly the same procedure as for internal validation. One RMSE value is computed by comparing predicted probabilities of presence with the truth over 5000 locations of a test sample.

The results of the fit of mixed-effects models for each validation landscape are shown in Table S4. Using the R^2 for the full models and all sub-models with one factor excluded at a time, we rank the importance of the factors for each landscape. For VD, NE and TI, the sample size is the most important factor followed by the modeling technique. For EN, the effect of modeling technique is larger than the effect of the sample size. For all landscapes except TI, the effect of the sampling design is larger than that of the missing covariate. Dispersal is the least important factor for all the landscapes, though it ties with sampling design in TI.

Correlation matrices between the five predictors *ddeg300*, *sfroyy*, *mind68*, *sumradyy* and *topos*

(in this order) for the four landscapes are:

$$\mathrm{cor}_{\mathrm{VD}} = \begin{pmatrix} 1.00 & -0.49 & -0.85 & 0.19 & -0.27 \\ -0.49 & 1.00 & 0.55 & -0.18 & 0.23 \\ -0.85 & 0.55 & 1.00 & -0.54 & 0.21 \\ 0.19 & -0.18 & -0.54 & 1.00 & 0.09 \\ -0.27 & 0.23 & 0.21 & 0.09 & 1.00 \end{pmatrix},$$

$$\mathrm{cor}_{\mathrm{EN}} = \begin{pmatrix} 1.00 & -0.79 & -0.70 & 0.00 & -0.38 \\ -0.79 & 1.00 & 0.77 & 0.04 & 0.47 \\ -0.70 & 0.77 & 1.00 & -0.27 & 0.36 \\ 0.00 & 0.04 & -0.27 & 1.00 & 0.11 \\ -0.38 & 0.47 & 0.36 & 0.11 & 1.00 \end{pmatrix},$$

$$\mathrm{cor}_{\mathrm{NE}} = \begin{pmatrix} 1.00 & -0.67 & -0.81 & 0.10 & -0.28 \\ -0.67 & 1.00 & 0.49 & -0.08 & 0.07 \\ -0.81 & 0.49 & 1.00 & -0.43 & 0.36 \\ 0.10 & -0.08 & -0.43 & 1.00 & 0.08 \\ -0.28 & 0.07 & 0.36 & 0.08 & 1.00 \end{pmatrix},$$

$$\mathrm{cor}_{\mathrm{TI}} = \begin{pmatrix} 1.00 & -0.62 & -0.58 & 0.11 & -0.32 \\ -0.62 & 1.00 & 0.37 & 0.01 & 0.30 \\ -0.58 & 0.37 & 1.00 & -0.70 & 0.05 \\ 0.11 & 0.01 & -0.70 & 1.00 & 0.11 \\ -0.32 & 0.30 & 0.05 & 0.11 & 1.00 \end{pmatrix}.$$

Table S4: Marginal R^2 for the full model (4) and the five sub-models with one factor (and all its interactions with the other factors) excluded at a time, and for various external validation landscapes.

Landscape	full model	−missing	−dispersal	−n	−design	−technique
VD (internal)	0.674	0.662	0.669	0.188	0.587	0.423
EN	0.250	0.247	0.250	0.162	0.166	0.146
NE	0.434	0.415	0.428	0.085	0.403	0.265
TI	0.325	0.248	0.319	0.082	0.319	0.223

S7 Measures for the relative importance of factors

Here we discuss the use of the ANOVA, the R^2, and the likelihood to measure the relative effect of the factors in the linear mixed-effects model (4). The sum of squares in the ANOVA shows the contribution of each factor to the variability of the log(RMSE). This was used by Dormann *et al.* (2008) and Diniz-Filho *et al.* (2009) in a similar context to ours. However, when the sums of squares for interactions have large effects (as in the present case), the full variability that can be attributed to a factor is difficult to quantify because part of it may be included in the interactions with other factors. Here we discuss some ways to measure the full contribution of a factor and its interactions.

A first possibility is to add up the sums of squares (sum of SS) for the factor and all its interactions with the other factors in the ANOVA. However these values will not correspond to the variability due to one factor only, as the sums of squares for interactions correspond also to the other factors.

A second possibility is to fit sub-models in which one factor, and all its interactions with other factors, are excluded, and then to compute goodness-of-fit measures for the full model and sub-models. The differences in goodness-of-fit measures indicate the effect of excluding the factors from the full model. This approach differs from summing the sum of squares in the ANOVA, because excluding a factor and its interactions transfers the variation due to this factor to the residuals (at the corresponding level) and affect the fits of the linear mixed-effects sub-models. Several goodness-of-fit measures can be used. We computed the likelihood, which is a universal and well-known goodness-of-fit measure, and the marginal and conditional R^2 introduced by Nakagawa & Schielzeth (2013). The marginal R^2 represents the proportion of variance in the model explained by the fixed effects only, while the conditional R^2 also accounts for the random effects. The interpretation of R^2 in terms of the percentage of variance explained by the model makes its values easy to understand. However, there may be difficulties in interpreting the variation in R^2, as the inclusion of a fixed effect in the model may occasionally reduce the marginal R^2 (Nakagawa & Schielzeth, 2013). For computing the R^2, the models were fitted with REML to obtain unbiased estimates of the variance parameters. For computing the likelihoods, the models were fitted with maximum likelihood to obtain comparable likelihoods. Table S5 shows the results obtained with the different measures for the simulation of the paper (dispersal with $\lambda = 0.5$ and validation in VD). The sum of SS of the ANOVA, the marginal R^2 and the log-likelihoods give the same rankings of the factors. The conditional R^2 gives a larger effect for technique than for n, and also accounts for the variance explained by the random effects, which is not of interest in our case. The rankings obtained by the three other measures leads to similar conclusions in our case but this may not always occur. When the sum of squares of interactions are large, we recommend first looking at the main effects in the ANOVA, then looking at the interactions, and finally computing several goodness-of-fit measures for sub-models.

We computed the sum of SS and the log-likelihood differences for the simulations with various spatial autocorrelation strengths (§S5) and for external validations (§S6). The rankings based on the sum of SS are the same as those obtained with the marginal R^2. For the log-likelihood differences, the results (Tables S6 and S7) are consistent with those obtained with the marginal R^2 (Tables S2 and S4) in most cases; for varying the strength of SAC, dispersal appears to have a weaker effect in terms of likelihood than R^2.

Table S5: First row is the sum of SS in the ANOVA for each factor and all its interactions with the other factors. Other rows are marginal and conditional R^2 and log-likelihoods, given in terms of differences for the value of the full model, for the five sub-models with one factor, and all its interactions with the other factors, excluded at a time. A larger value corresponds to a larger effect of the factor.

Measure	missing	dispersal	n	design	technique
ANOVA (sum of SS)	58	23	2465	442	1272
Marginal R^2	0.012	0.005	0.486	0.087	0.251
Conditional R^2	0.005	0.001	0.188	0.016	0.332
Log-likelihood	502	158	12543	2647	11176

Table S6: Log-likelihood differences for the full model (4) and the five sub-models with one factor (and all its interactions with the other factors) excluded at a time, and for various strengths of spatial autocorrelation for the dispersal process (see Table S2).

Effective range (km)	λ	full model	−missing	−dispersal	−n	−design	−technique
1.5	0.5	4776	502	158	12543	2647	11176
5	1.67	5313	467	1739	10735	2466	10261
10	3.34	5465	499	3968	9885	2289	10411
15	5	5938	497	5745	9927	2219	11236

Table S7: Log-likelihood differences for the full model (4) and the five sub-models with one factor (and all its interactions with the other factors) excluded at a time, and for various external validation landscapes.

Landscape	full model	−missing	−dispersal	−n	−design	−technique
EN	−22164	107	24	1880	1155	3951
NE	−4827	532	99	7417	711	6194
TI	−7866	701	144	5380	138	4061

S8 R scripts and data

We provide the R scripts and data to run all analyses presented in the article. See the separate file entitled "Scripts_Thibaud_etal_MEE.zip", which contains:

1. main.R: The main R script for running the simulation, plotting the results and calculating the coefficients R^2.
2. functionsFI.R: Some functions for the simulation.
3. create_species.R: R scripts for creating the virtual species.
4. species_VD.R: R scripts for calculating the spatial autocorrelation for VD species.
5. fit_spatialprobit.R: Function to fit the spatial probit model described in §S4.
6. save_matrixresults.R: Function to save the results of the simulation (RMSE) and the corresponding configurations in a matrix.
7. plot_results.R: Function to plot the results.
8. anova.R: Function to do the ANOVA for the log(RMSE) values.
9. factor_importance_R2.R: Function to calculate the coefficients R^2.
10. factor_importance_nllik.R: Function to calculate the log-likelihood differences.
11. extern_valid.R: Fonction for validation on other landscapes.
12. pred_VD.Rdata, pred_EN.Rdata, pred_NE.Rdata, pred_TI.Rdata: Data file for the real predictors in VD, EN, NE and TI.
13. distroad_VD.Rdata: Data file for the sampling design (distance to the nearest-road).
14. datasp_VD.Rdata: Data file for the 10 real species used in VD.
15. FI-VD.Rdata, FI-VD-extEN.Rdata, FI-VD-extNE.Rdata, FI-VD-extTI.Rdata: The results for the simulations of the paper.

To run the simulation process, only the file main.R need be open. It must be run line by line and it will load the data and source the necessary other files. The script species_VD.R is independent of the simulation and allows estimation of the spatial autocorrelation for the real species using a spatial probit model. Finally, the data files "FI-VD.Rdata", "FI-VD-extEN.Rdata", "FI-VD-extNE.Rdata" and "FI-VD-extTI.Rdata" contain the results of the simulations presented in the paper.

S9 Supporting Table and Figures

Additional Tables and Figures:

Table S8: Some references treating the effects of modeling factors on SDM predictions.

Figures S1–S9: Boxplots of RMSE values for the simulated species 2–10 and for each combination of the factors we considered. Validation is performed on the same landscape (VD).

Figures S10–S19: Maps of presence probability for the simulated species 1–10 in the landscapes VD, EN, NE and TI.

Table S8: References treating the effects of modeling factors on SDM predictions.

Modeling factor	References
Modeling techniques	Hirzel *et al.* (2001);Moisen & Frescino (2002);Thuiller (2003);Segurado & Araújo (2004); Elith *et al.* (2006);Guisan *et al.* (2007a);Meynard & Quinn (2007)*;Tsoar *et al.* (2007); Elith & Graham (2009)*
Grain	Hortal *et al.* (2006);McPherson *et al.* (2006);Guisan *et al.* (2007b)
Location error	Graham *et al.* (2008);Johnson & Gillingham (2008);Naimi *et al.* (2011)*
Pseudo-absence selection	Zaniewski *et al.* (2002);Phillips *et al.* (2009);Wisz & Guisan (2009)*;Vanderwal *et al.* (2009)
Sample size	Hirzel *et al.* (2002);Stockwell & Peterson (2002);Kadmon *et al.* (2003);Reese *et al.* (2005)*; Hernandez *et al.* (2006);Wisz *et al.* (2008);Jiménez-Valverde *et al.* (2009)*
Multi-collinearity	Graham (2003)†
Geographic extent	Thuiller *et al.* (2004)
Threshold criteria for binarizing predictions	Freeman & Moisen (2008);Liu *et al.* (2005)
Data completeness	Peterson & Cohoon (1999);Kadmon *et al.* (2003)
Data characteristics	Kadmon *et al.* (2003);Guisan *et al.* (2007a)
Species characteristics	McPherson *et al.* (2004)†;Guisan *et al.* (2007a);M McPherson & Jetz (2007); Broennimann *et al.* (2012)†
Sampling design	Austin & Adomeit (1991)*;Hirzel *et al.* (2002);Kadmon *et al.* (2003);Kadmon *et al.* (2004); Reese *et al.* (2005)*;Phillips *et al.* (2009);Albert *et al.* (2010)*
Spatial autocorrelation	Cablk *et al.* (2002);Segurado *et al.* (2006)†;Dormann (2007);Dormann *et al.* (2007)*

*use simulated data

† use simulated and real data

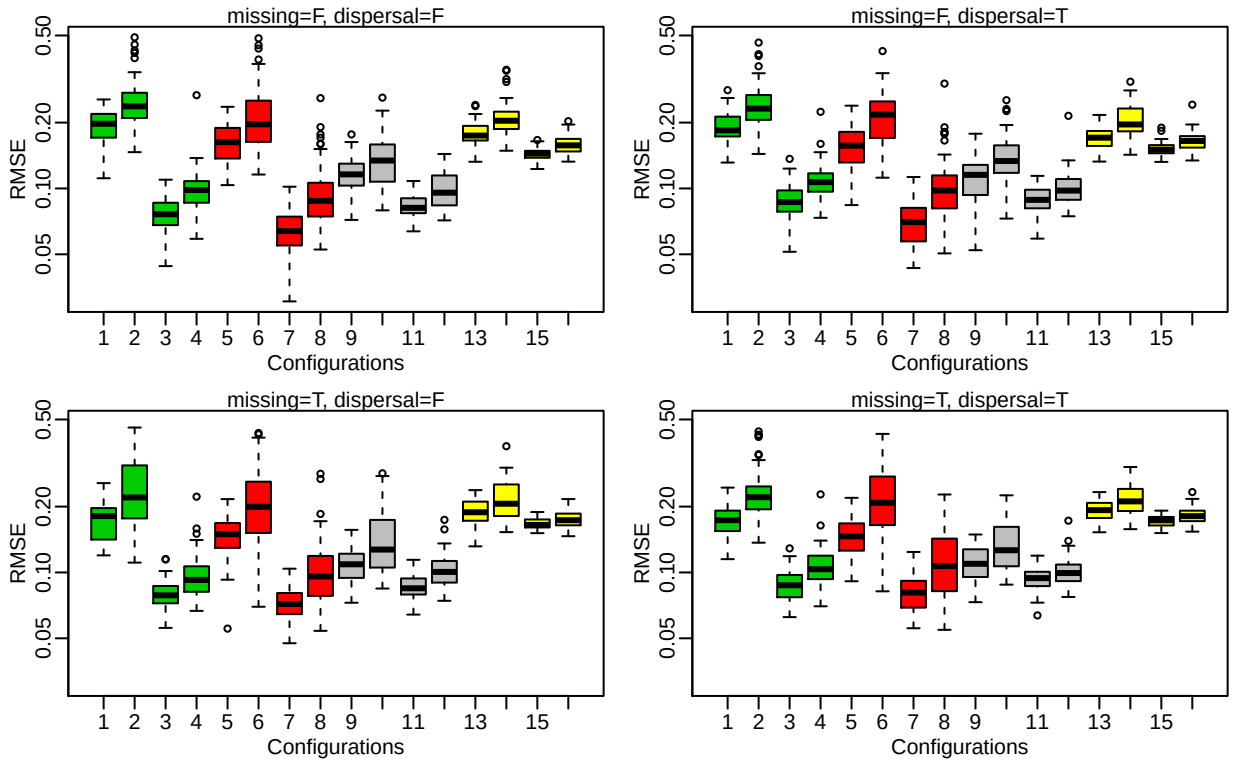

Figure S1: Boxplots of RMSE for simulated species 2, on a log scale. Each boxplot shows the variation due to the 5 samples for each of the 10 simulations and for one particular configuration of factors (missing covariate, dispersal, modeling technique, sample size n and sampling design), and thus contains 50 RMSE values, each equal to the root mean squared difference of the estimated and the true probabilities of presence over the 5000 sites of the test samples. Each panel corresponds to a different configuration of the factors *missing* and *dispersal*. In each panel colors correspond to the different modeling techniques: GAMs in green, GLMs in red, MexEnt in grey, and RF in yellow. Inside the color groups, boxplots are first separated by the sample size n (100 or 500, to the left or right) and within sample size by the sampling design (simple random or road-based, to the left or right). Thus, configurations 1–4 correspond to results for GAMs with (sample size, sampling design) settings (100, simple random), (100, road-based), (500, simple random) and (500, road-based), respectively.

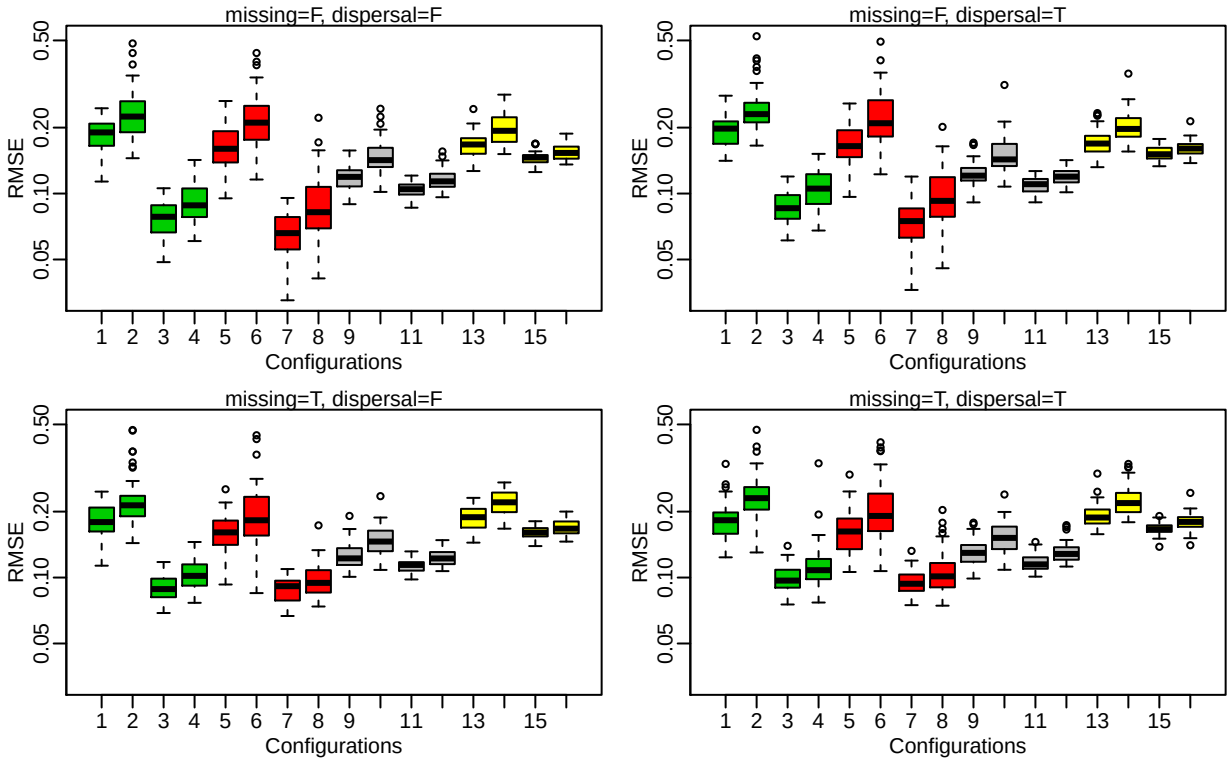

Figure S2: Boxplots of RMSE for simulated species 3. Same caption as Figure S1.

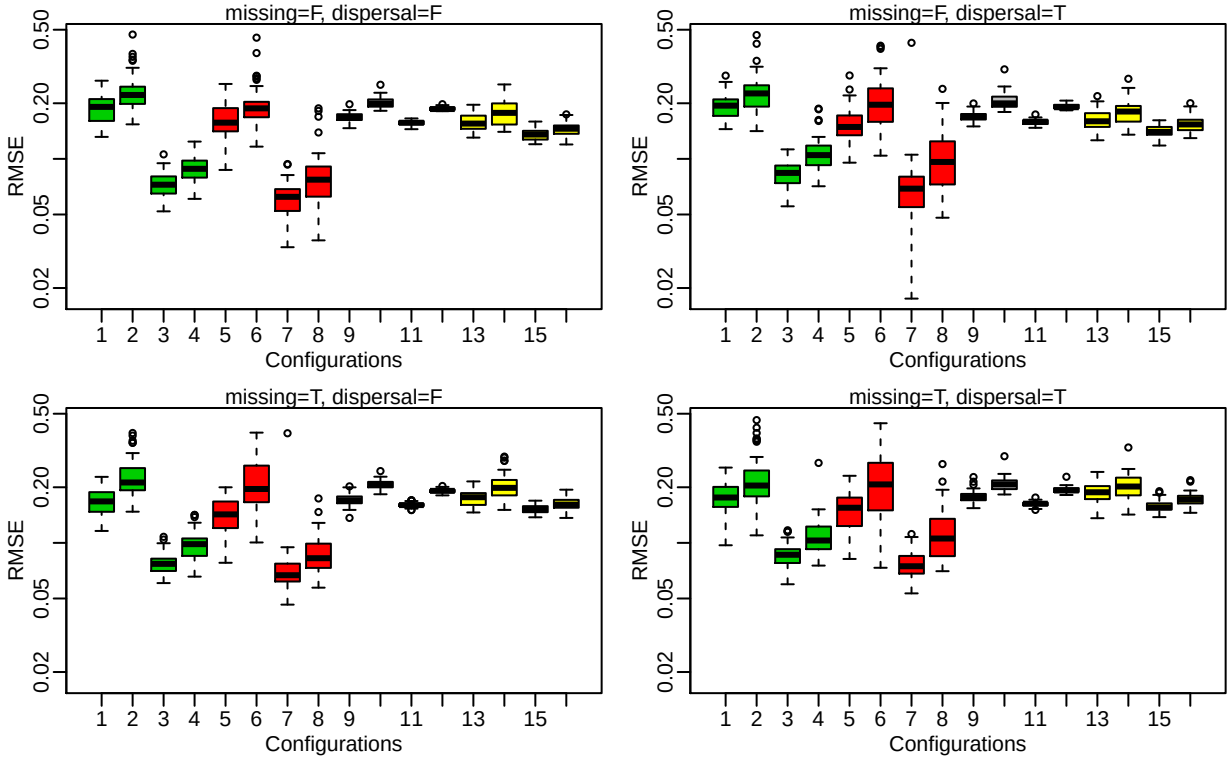

Figure S3: Boxplots of RMSE for simulated species 4. Same caption as Figure S1.

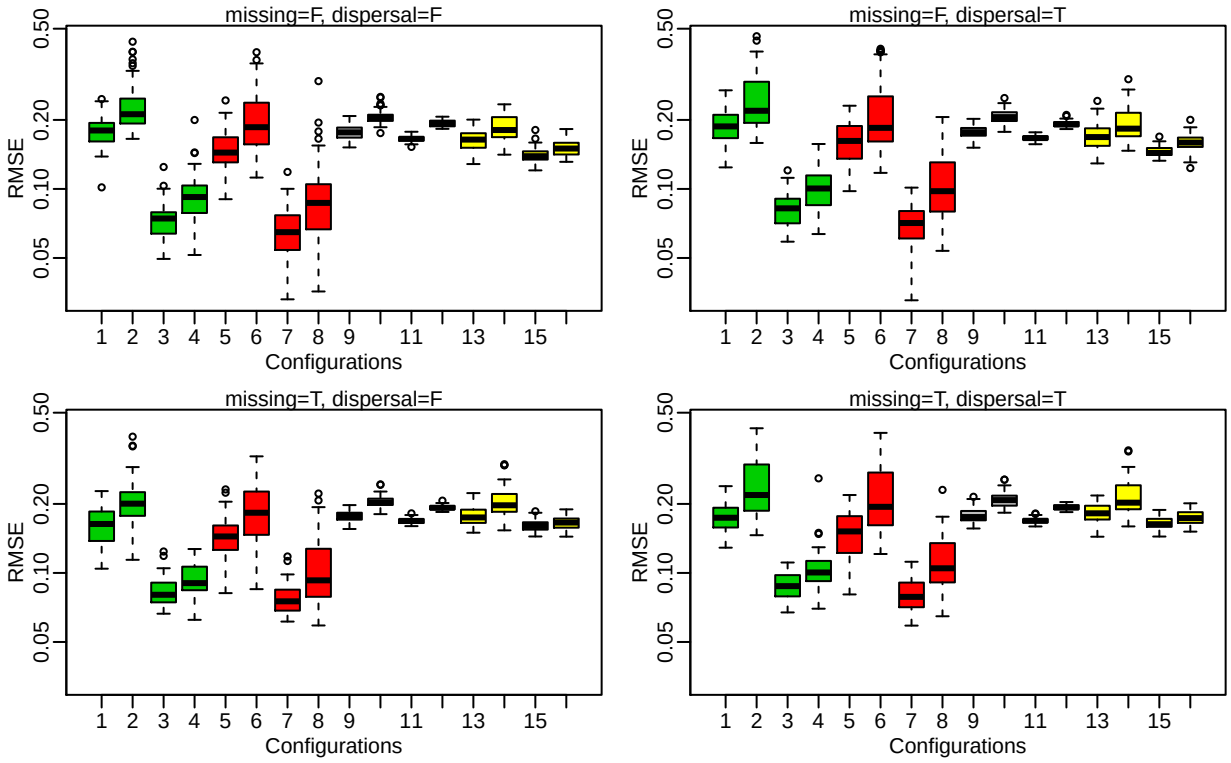

Figure S4: Boxplots of RMSE for simulated species 5. Same caption as Figure S1.

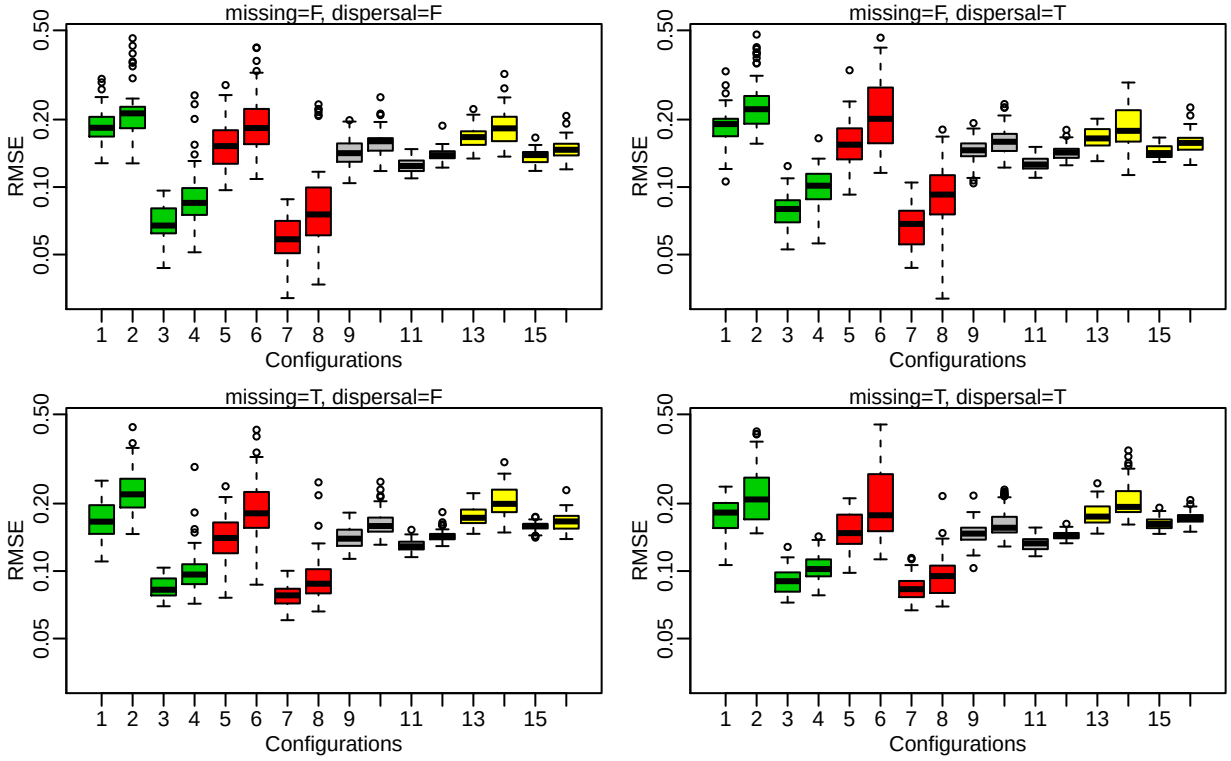

Figure S5: Boxplots of RMSE for simulated species 6. Same caption as Figure S1.

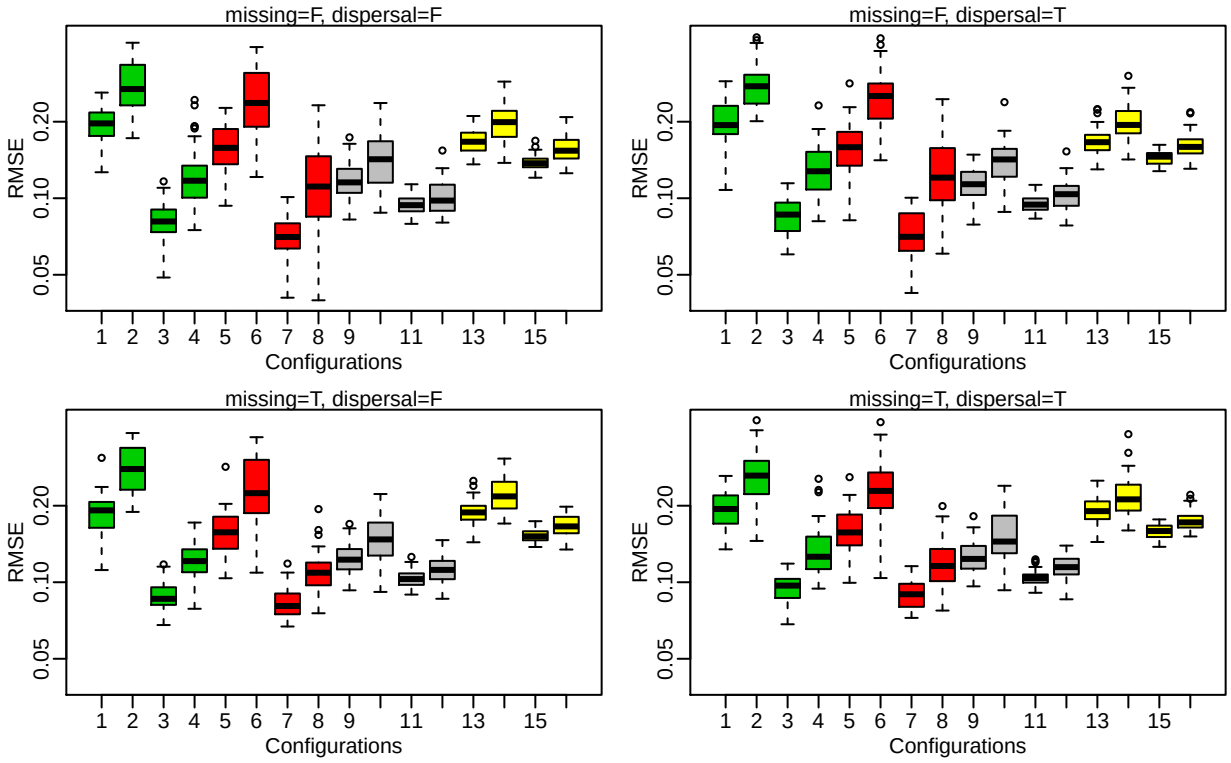

Figure S6: Boxplots of RMSE for simulated species 7. Same caption as Figure S1.

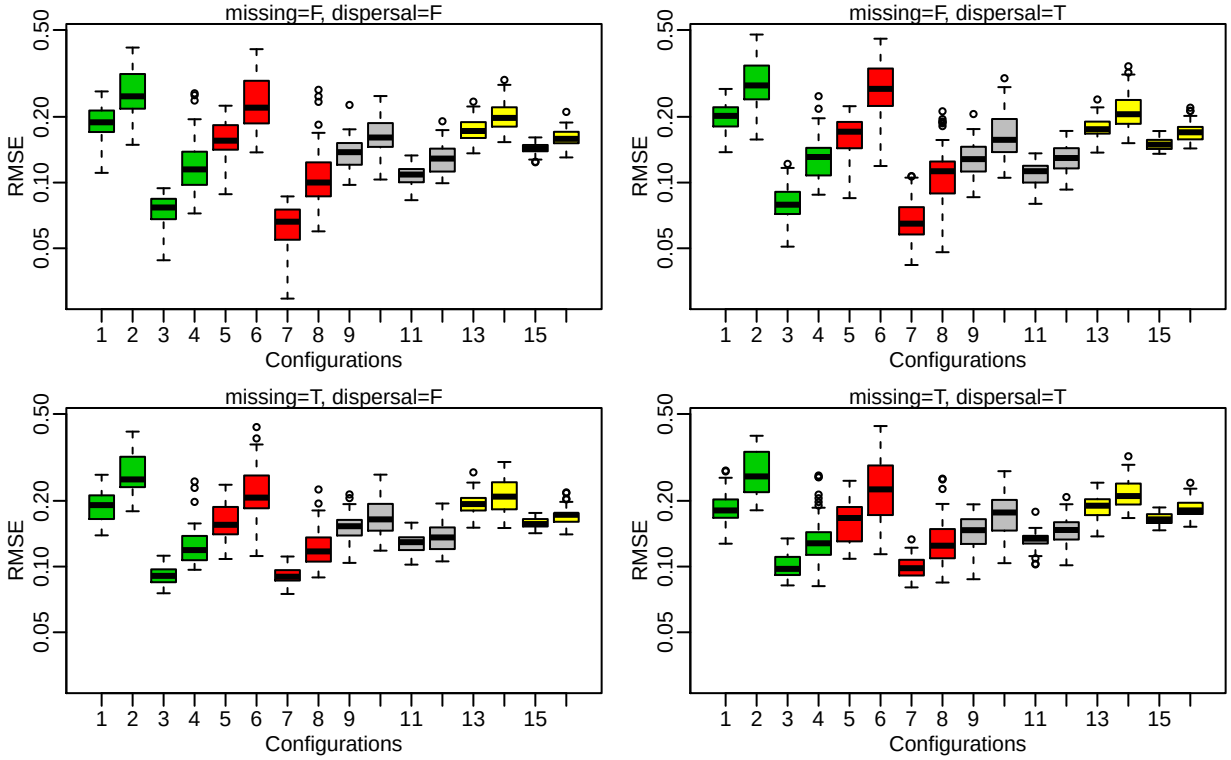

Figure S7: Boxplots of RMSE for simulated species 8. Same caption as Figure S1.

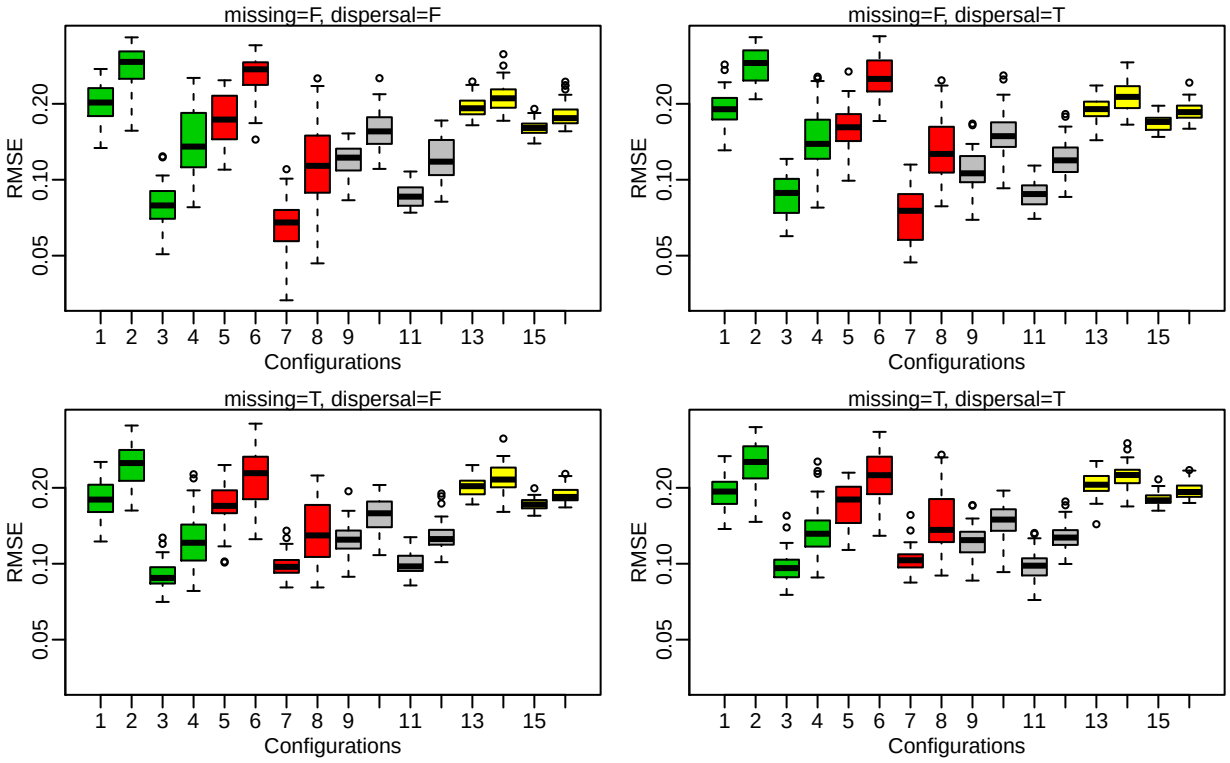

Figure S8: Boxplots of RMSE for simulated species 9. Same caption as Figure S1.

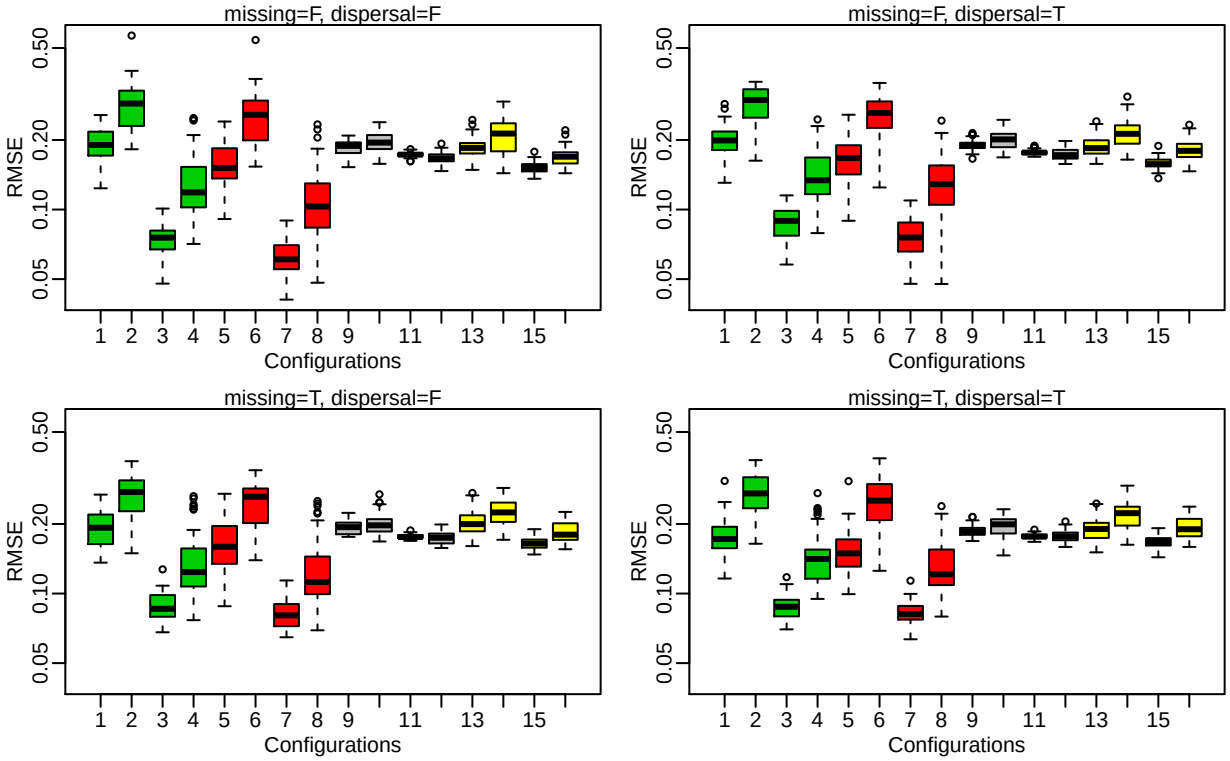

Figure S9: Boxplots of RMSE for simulated species 10. Same caption as Figure S1.

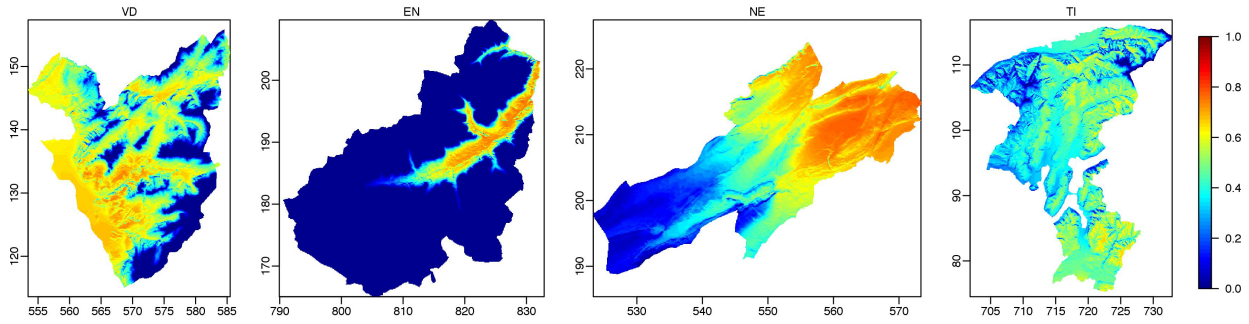

Figure S10: Theoretical probabilities of presence for simulated species 1.

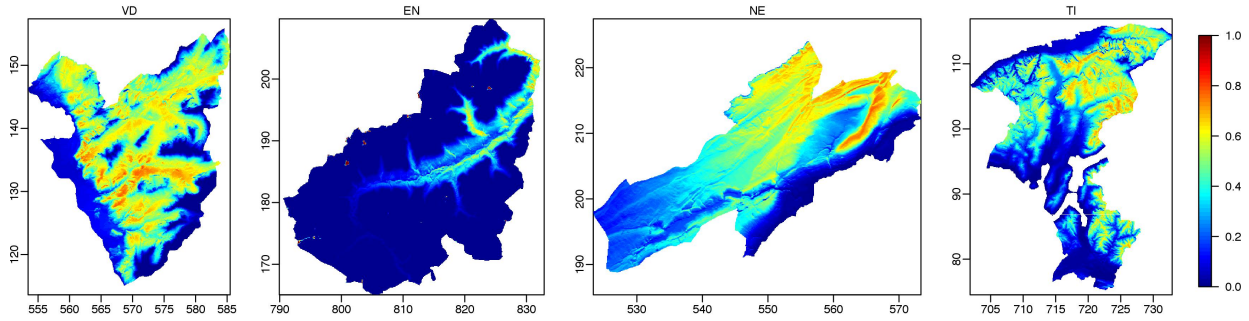

Figure S11: Theoretical probabilities of presence for simulated species 2.

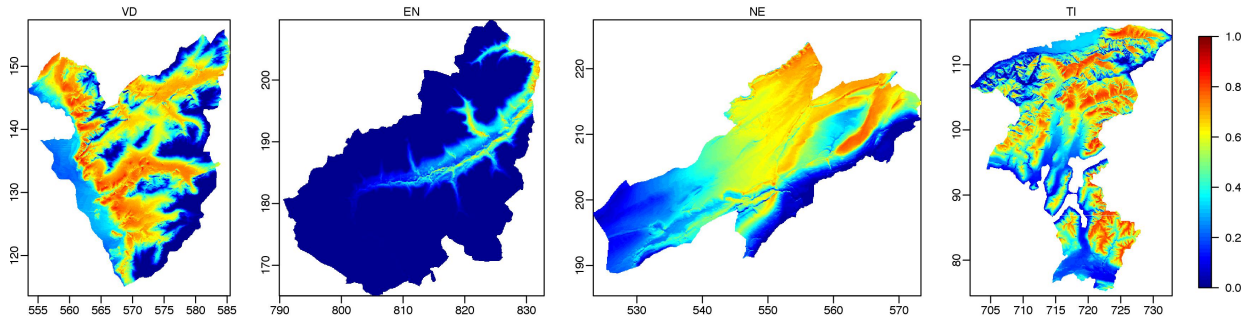

Figure S12: Theoretical probabilities of presence for simulated species 3.

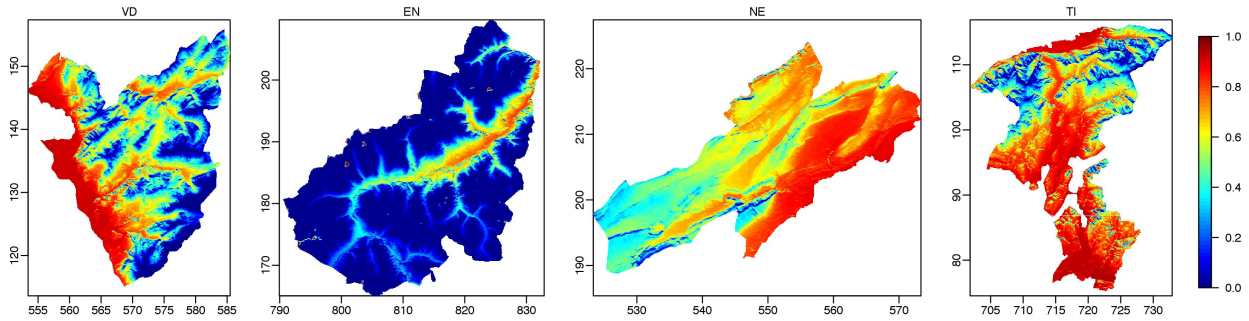

Figure S13: Theoretical probabilities of presence for simulated species 4.

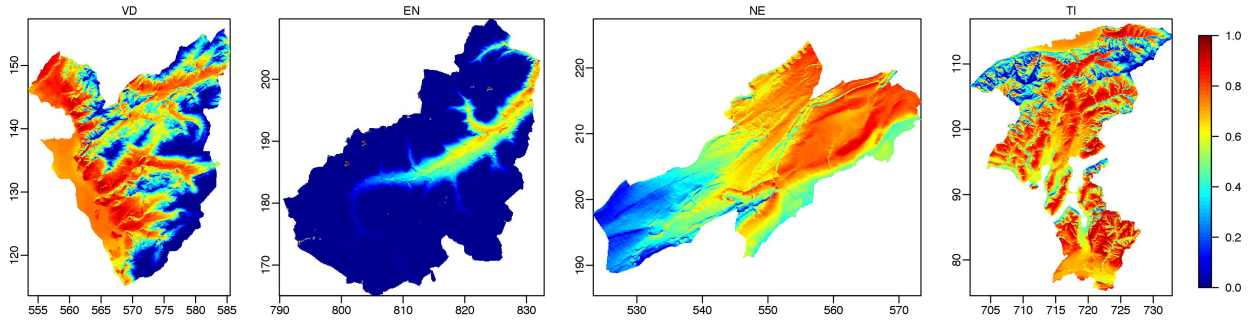

Figure S14: Theoretical probabilities of presence for simulated species 5.

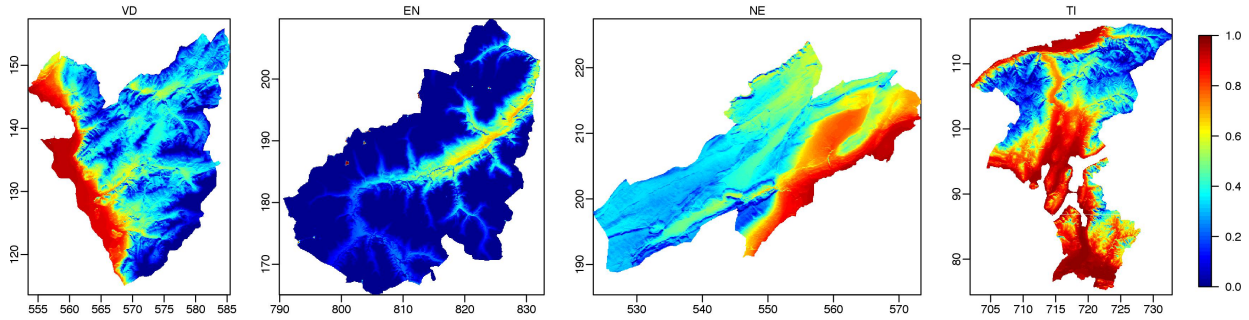

Figure S15: Theoretical probabilities of presence for simulated species 6.

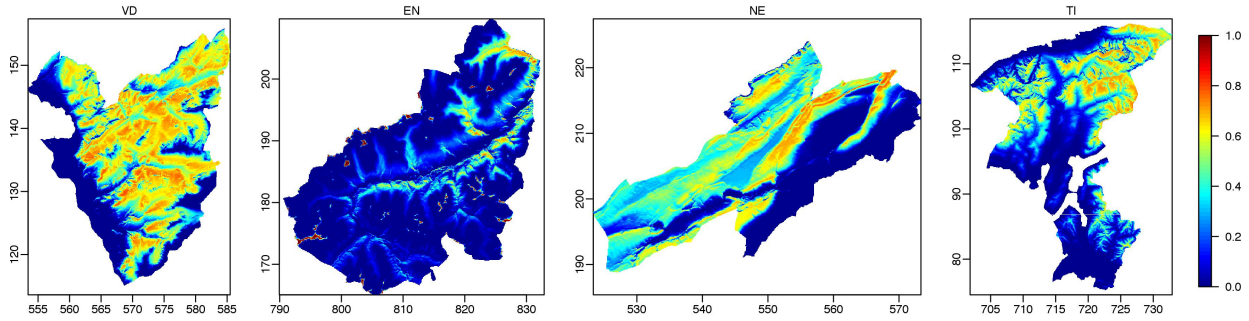

Figure S16: Theoretical probabilities of presence for simulated species 7.

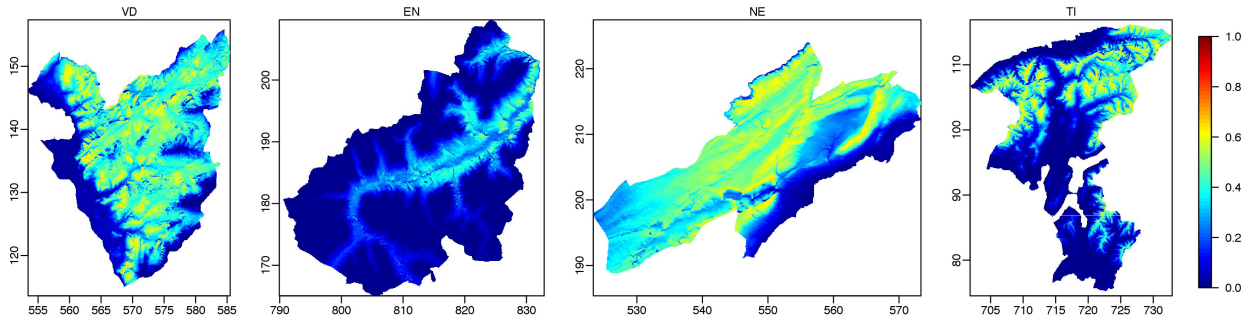

Figure S17: Theoretical probabilities of presence for simulated species 8.

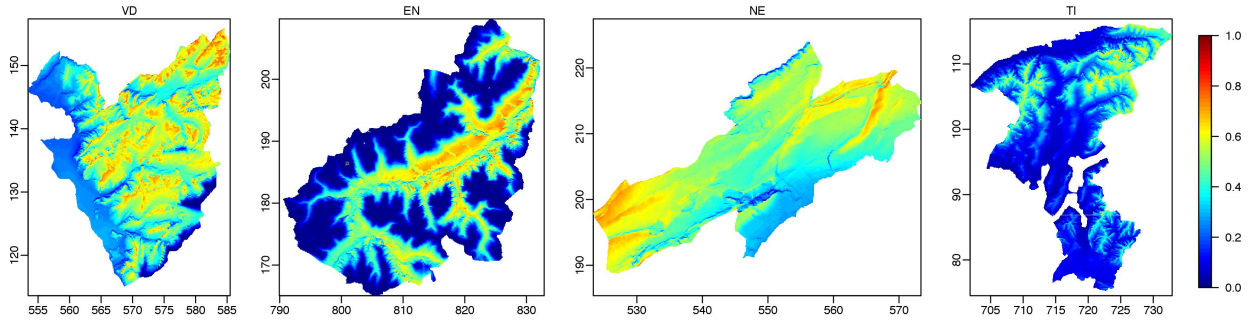

Figure S18: Theoretical probabilities of presence for simulated species 9.

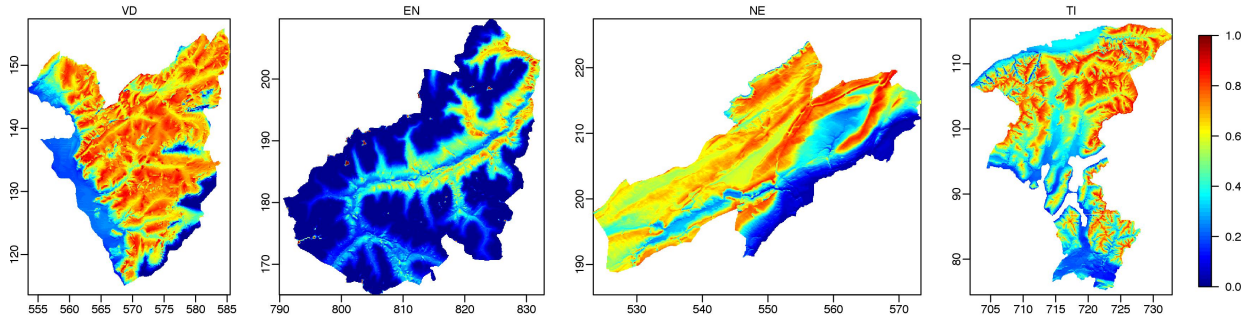

Figure S19: Theoretical probabilities of presence for simulated species 10.

References

Albert, C.H., Yoccoz, N.G., Edwards, T.C., Graham, C.H., Zimmermann, N.E. & Thuiller, W. (2010) Sampling in ecology and evolution – bridging the gap between theory and practice. *Ecography*, **33**, 1028–1037.

Araújo, M.B. & Peterson, A.T. (2012) Uses and misuses of bioclimatic envelope modeling. *Ecology*, **93**, 1527–1539.

Austin, M.P. & Adomeit, E.M. (1991) Sampling strategies costed by simulation. C.R. Margules & M.P. Austin, eds., *Nature Conservation: cost effective biological surveys and data analysis*, pp. 167–175. Commonwealth Scientific & Industrial Research (CSIRO), Dickson, Australia.

Broennimann, O., Fitzpatrick, M.C., Pearman, P.B., Petitpierre, B., Pellissier, L., Yoccoz, N.G., Thuiller, W., Fortin, M.J., Randin, C., Zimmermann, N.E., Graham, C.H. & Guisan, A. (2012) Measuring ecological niche overlap from occurrence and spatial environmental data. *Global Ecology and Biogeography*, **21**, 481–497.

Cablk, M., White, D. & Kiester, A.R. (2002) Assessment of spatial autocorrelation in empirical models in ecology. J. Scott, P. Heglund, M. Morrison, J. Haufler, M. Raphael, W. Wall & F. Samson, eds., *Predicting Species Occurrences: Issues of Accuracy and Scale*, pp. 429–440. Island Press, Covelo, CA.

Caruana, R. & Niculescu-Mizil, A. (2004) Data mining in metric space: an empirical analysis of supervised learning performance criteria. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, pp. 69–78. ACM, New York, NY, USA.

Diniz-Filho, J.A.F., Mauricio Bini, L., Fernando Rangel, T., Loyola, R.D., Hof, C., Nogués-Bravo, D. & Araújo, M.B. (2009) Partitioning and mapping uncertainties in ensembles of forecasts of species turnover under climate change. *Ecography*, **32**, 897–906.

Dormann, C.F. (2007) Effects of incorporating spatial autocorrelation into the analysis of species distribution data. *Global Ecology and Biogeography*, **16**, 129–138.

Dormann, C.F., McPherson, J.M., Araújo, M.B., Bivand, R., Bolliger, J., Carl, G., Davies, R.G., Hirzel, A., Jetz, W., Kissling, W.D., Kuhn, I., Ohlemuller, R., Peres-Neto, P.R., Reineking, B., Schroder, B., Schurr, F.M. & Wilson, R. (2007) Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, **30**, 609–628.

Dormann, C.F., Purschke, O., García Márquez, J.R., Lautenbach, S. & Schröder, B. (2008) Components of uncertainty in species distribution analysis: a case study of the Great Grey Shrike. *Ecology*, **89**, 3371–3386.

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**, 66–77.

Elith, J., H. Graham, C., P. Anderson, R., Dudík, M., Ferrier, S., Guisan, A., J. Hijmans, R., Huettmann, F., R. Leathwick, J., Lehmann, A., Li, J., G. Lohmann, L., A. Loiselle, B., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., McC. M. Overton, J., Townsend Peterson, A., J. Phillips, S., Richardson, K., Scachetti-Pereira, R., E. Schapire, R., Soberón, J., Williams, S., S. Wisz, M. & E. Zimmermann, N. (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**, 129–151.

Elith, J. & Leathwick, J.R. (2009) Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology Evolution and Systematics*, **40**, 677–697.

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

Fielding, A.H. & Bell, J.F. (1997) A review of methods for the assessment of prediction errors in conservation presence / absence models. *Environmental Conservation*, **24**, 38–49.

Freeman, E.A. & Moisen, G.G. (2008) A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. *Ecological Modelling*, **217**, 48–58.

Graham, C.H., Elith, J., Hijmans, R.J., Guisan, A., Townsend Peterson, A. & Loiselle, B.A. (2008) The influence of spatial errors in species occurrence data used in distribution models. *Journal of Applied Ecology*, **45**, 239–247.

Graham, M.H. (2003) Confronting Multicollinearity in Ecological Multiple Regression. *Ecology*, **84**, 2809–2815.

Guisan, A. & Thuiller, W. (2005) Predicting species distribution: offering more than simple habitat models. *Ecology Letters*, **8**, 993–1009.

Guisan, A., Zimmermann, N.E., Elith, J., Graham, C.H., Phillips, S. & Peterson, A.T. (2007a) What Matters for Predicting the Occurrences of Trees: Techniques, Data, or Species' Characteristics? *Ecological Monographs*, **77**, 615–630.

Guisan, A., Graham, C.H., Elith, J. & Huettmann, F. (2007b) Sensitivity of predictive species distribution models to change in grain size. *Diversity and Distributions*, **13**, 332–340.

Hastie, T. & Fithian, W. (2013) Inference from presence-only data; the ongoing controversy. *Ecography*, **36**, 864–867.

Hernandez, P.A., Graham, C.H., Master, L.L. & Albert, D.L. (2006) The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography*, **29**, 773–785.

Hijmans, R.J., Phillips, S., Leathwick, J. & Elith, J. (2013) *dismo: Species distribution modeling*. R package version 0.8-17.

Hirzel, A.H., Hausser, J., Chessel, D. & Perrin, N. (2002) Ecological-Niche Factor Analysis: How to compute habitat-suitability maps without absence data? *Ecology*, **83**, 2027–2036.

Hirzel, A.H., Helfer, V. & Metral, F. (2001) Assessing habitat-suitability models with a virtual species. *Ecological Modelling*, **145**, 111–121.

Hortal, J., Borges, P.A.V. & Gaspar, C. (2006) Evaluating the performance of species richness estimators: sensitivity to sample grain size. *Journal of Animal Ecology*, **75**, 274–287.

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

Johnson, C.J. & Gillingham, M.P. (2008) Sensitivity of species-distribution models to error, bias, and model design: An application to resource selection functions for woodland caribou. *Ecological Modelling*, **213**, 143–155.

Kadmon, R., Farber, O. & Danin, A. (2003) A systematic analysis of factors affecting the performance of climatic envelope models. *Ecological Applications*, **13**, 853–867.

Kadmon, R., Farber, O. & Danin, A. (2004) Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. *Ecological Applications*, **14**, 401–413.

Liu, C., Berry, P.M., Dawson, T.P. & Pearson, R.G. (2005) Selecting thresholds of occurrence in the prediction of species distributions. *Ecography*, **28**, 385–393.

Liu, C., White, M. & Newell, G. (2011) Measuring and comparing the accuracy of species distribution models with presence–absence data. *Ecography*, **34**, 232–243.

M McPherson, J. & Jetz, W. (2007) Effects of species' ecology on the accuracy of distribution models. *Ecography*, **30**, 135–151.

Mason, S.J. & Graham, N.E. (2002) Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Quarterly Journal of the Royal Meteorological Society*, **128**, 2145–2166.

McPherson, J., Jetz, W. & Rogers, D. (2006) Using coarse-grained occurrence data to predict species distributions at finer spatial resolutions–possibilities and limitations. *Ecological Modelling*, **192**, 499–522.

McPherson, J.M., Jetz, W. & Rogers, D.J. (2004) The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? *Journal of Applied Ecology*, **41**, 811–823.

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*, pp. no–no.

Meynard, C.N. & Quinn, J.F. (2007) Predicting species distributions: a critical comparison of the most common statistical models using artificial species. *Journal of Biogeography*, **34**, 1455–1469.

Moisen, G. & Frescino, T. (2002) Comparing five modelling techniques for predicting forest characteristics. *Ecological Modelling*, **157**, 209–225.

Naimi, B., Skidmore, A.K., Groen, T.A. & Hamm, N.A.S. (2011) Spatial autocorrelation in predictors reduces the impact of positional uncertainty in occurrence data on species distribution modelling. *Journal of Biogeography*, **38**, 1497–1509.

Nakagawa, S. & Schielzeth, H. (2013) A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, **4**, 133–142.

Peterson, A.T. & Cohoon, K.P. (1999) Sensitivity of distributional prediction algorithms to geographic data completeness. *Ecological Modelling*, **117**, 159–164.

Phillips, S.J., Anderson, R.P. & Schapire, R.E. (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231–259.

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

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. *Ecological Applications*, **19**, 181–197.

Pottier, J., Dubuis, A., Pellissier, L., Maiorano, L., Rossier, L., Randin, C.F., Vittoz, P. & Guisan, A. (2013) The accuracy of plant assemblage prediction from species distribution models varies along environmental gradients. *Global Ecology and Biogeography*, **22**, 52–63.

Randin, C.F., Dirnböck, T., Dullinger, S., Zimmermann, N.E., Zappa, M. & Guisan, A. (2006) Are niche-based species distribution models transferable in space? *Journal of Biogeography*, **33**, 1689–1703.

Reese, G.C., Wilson, K.R., Hoeting, J.A. & Flather, C.H. (2005) Factors affecting species distribution predictions: a simulation modeling experiment. *Ecological Applications*, **15**, 554–564.

Renner, I.W. & Warton, D.I. (2013) Equivalence of MAXENT and Poisson Point Process Models for Species Distribution Modeling in Ecology. *Biometrics*, **69**, 274–281.

Segurado, P., Araújo, M.B. & Kunin, W.E. (2006) Consequences of spatial autocorrelation for niche-based models. *Journal of Applied Ecology*, **43**, 433–444.

Segurado, P. & Araújo, M.B. (2004) An evaluation of methods for modelling species distributions. *Journal of Biogeography*, **31**, 1555–1568.

Stockwell, D.R.B. & Peterson, A.T. (2002) Effects of sample size on accuracy of species distribution models. *Ecological Modelling*, **148**, 1–13.

Tate, R.F. (1954) Correlation Between a Discrete and a Continuous Variable. Point-Biserial Correlation. *The Annals of Mathematical Statistics*, **25**, 603–607.

Thuiller, W. (2003) BIOMOD - optimizing predictions of species distributions and projecting potential future shifts under global change. *Global Change Biology*, **9**, 1353–1362.

Thuiller, W., Brotons, L., Araújo, M.B. & Lavorel, S. (2004) Effects of restricting environmental range of data to project current and future species distributions. *Ecography*, **27**, 165–172.

Tsoar, A., Allouche, O., Steinitz, O., Rotem, D. & Kadmon, R. (2007) A comparative evaluation of presence-only methods for modelling species distribution. *Diversity and Distributions*, **13**, 397–405.

Vanderwal, J., Shoo, L., Graham, C. & Williams, S. (2009) Selecting pseudo-absence data for presence-only distribution modeling: How far should you stray from what you know? *Ecological Modelling*, **220**, 589–594.

Varin, C., Reid, N. & Firth, D. (2011) An overview of composite likelihood methods. *Statistica Sinica*, **21**, 5–42.

Vittoz, P. & Engler, R. (2007) Seed dispersal distances: a typology based on dispersal modes and plant traits. *Botanica Helvetica*, **117**, 109–124.

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

Wisz, M.S., Hijmans, R.J., Peterson, A.T., Graham, C.H., Guisan, A. & NCEAS Predicting Species Distributions Working Group (2008) Effects of sample size on the performance of species distribution models. *Diversity and Distributions*, **14**, 763–773.

Yackulic, C.B., Chandler, R., Zipkin, E.F., Royle, J.A., Nichols, J.D., Campbell Grant, E.H. & Veran, S. (2013) Presence-only modelling using MAXENT: when can we trust the inferences? *Methods in Ecology and Evolution*, **4**, 236–243.

Zaniewski, A., Lehmann, A. & Overton, J. (2002) Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecological Modelling*, **157**, 261–280.

Zimmermann, N.E. & Kienast, F. (1999) Predictive mapping of alpine grasslands in Switzerland: Species versus community approach. *Journal of Vegetation Science*, **10**, 469–482.