Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2014, **5,** 947–955 doi: 10.1111/2041-210X.12203

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, Lausanne, Switzerland;* [2]*Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland; and* [3]*Institute of Earth Surface Dynamics, University of Lausanne, Lausanne, Switzerland*

Summary

1. Species distribution models are increasingly used to address conservation questions, so their predictive capacity requires careful evaluation. Previous studies have shown how individual factors used in model construction can affect prediction. Although some factors probably have negligible effects compared to others, their relative effects are largely unknown.

2. We introduce a general 'virtual ecologist' framework to study the relative importance of factors involved in the construction of species distribution models.

3. We illustrate the framework by examining the relative importance of five key factors – a missing covariate, spatial autocorrelation due to a dispersal process in presences/absences, sample size, sampling design and modelling technique – in a real study framework based on virtual plants in a mountain landscape at regional scale, and show that, for the parameter values considered here, most of the variation in prediction accuracy is due to sample size and modelling technique. Contrary to repeatedly reported concerns, spatial autocorrelation has only comparatively small effects.

4. This study shows the importance of using a nested statistical framework to evaluate the relative effects of factors that may affect species distribution models.

Key-words: linear mixed-effects model, relative importance, spatial autocorrelation, virtual ecologist

Introduction

Predictions based on species distribution models (SDMs) have become widespread (Guisan & Thuiller 2005; Elith & Leathwick 2009; Franklin 2010; Peterson *et al.* 2011), stimulated by the pressing need identified a decade ago for more predictive models (Clark *et al.* 2001; Côté & Reynolds 2002) and by their use in global change assessments (e.g. Schröter *et al.* 2005; Pereira *et al.* 2010). SDMs are empirical models based on observed occurrences or abundances of species. By fitting the observed environmental conditions in which a species can grow and survive, these models are rooted in the realized environmental niche concept, that is, the envelope of environmental conditions constrained by competitive interactions and dispersion limitations (Guisan & Thuiller 2005; Soberón 2007).

Many methodological decisions must be made during the construction of an SDM (Guisan & Thuiller 2005; Franklin 2010), introducing various sources of uncertainty for model building and predictions (Beale & Lennon 2012), and it is important to understand these. Some factors can be partly controlled (e.g. sample size, sampling design, modelling technique), but others are harder to handle (e.g. species dispersal and distribution patterns). Most previous studies have assessed the effects of such factors only one at a time (Table S8, but see Dormann *et al.* 2008; Diniz-Filho *et al.* 2009; Garcia *et al.* 2012), thus preventing the calculation of their relative effects. Most such studies have used real data, but a few have used simulated, or so-called virtual, data (Zurell *et al.* 2010; Miller 2014).

Studies using real data have mostly proceeded by degrading initial conditions individually – for example, adding spatial autocorrelation, coarsening the grain or decreasing sample size – and then measuring the resulting increase in prediction error and decrease in predictive power. Even if much has been learned from them, these single-factor analyses cannot allow us to gauge the relative effects of different factors. Other studies have assessed how variation in factors may affect predictions (Thuiller *et al.* 2004; Baselga & Araújo 2009; Buisson *et al.* 2010), but are uninformative about the relative effect sizes because the truth is unknown and only uncertainty can be quantified. Assessing the relative effects of several factors in model building requires the truth to be known. Using a 'virtual ecologist' approach (Zurell *et al.* 2010; Miller 2014) with simulated virtual species distribution data and virtual observers offers such a truth-proofing perspective and allows full control

*Correspondence author. E-mail: emeric.thibaud@epfl.ch

of how factors are combined. There is no consensus on how best to simulate virtual species distributions (see Miller 2014): proposed approaches include simulating from a model (e.g. a generalized linear model, GLM; McCullagh & Nelder 1989) fitted to real data (as in Wisz & Guisan 2009; Albert *et al.* 2010), defining theoretical response functions (e.g. Hirzel & Guisan 2002), delimiting theoretical ranges along environmental gradients (e.g. Saupe *et al.* 2012) or using more dynamic approaches (e.g. Zurell *et al.* 2010) and then treating the resulting spatial predictions as the true distributions.

This study proposes a systematic approach to assessing how the factors involved in the construction of an SDM affect its prediction performance, using virtual species simulations. Our approach is based on environmental covariates that reflect natural situations and is illustrated by a comparison of the effects of sample size, sampling design, modelling technique and spatial autocorrelation resulting from a dispersal process or a missing covariate. Unlike previous studies, the present simulations include all possible combinations of the factors and use replication to assess the variability due to different presence–absence and sampling patterns.

Materials and methods

We decompose our simulation scheme into three parts, corresponding to three steps of species distribution modelling. The first step is the simulation of virtual species, with or without a missing covariate, and with or without spatial autocorrelation among the simulated presences and absences. In practice, the choice of species is dictated by the goals of the study and is not under the control of the modeller. The second step is to sample these virtual species at different locations, controlling the sample size and the sampling design. In practice, this step is determined by the resources available for data gathering. The third step is to apply different modelling techniques to the data obtained at the sampling step and to evaluate their performances using the root-mean-squared error. The modelling technique is chosen by the analyst, but in practice, its performance cannot be directly evaluated, as no ground truth is available. Our approach is implemented in version 3.0.1 of the open-source software R (R Core Team 2013). The code is given in the Supporting Information.

VIRTUAL SPECIES SIMULATION

Our simulation uses a real Alpine landscape in the western Swiss Alps, an area of about 700 km^2 of the canton of Vaud, comprising 1 127 599 pixels at a resolution size of 25 m. Plants were intensively sampled in this region, and many environmental maps were assembled (see Supporting Information). This helps to provide realistic estimates of parameters needed for the simulation. We used five real climatic and topographic predictors, labelled $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_5$: degree days above three (*ddeg300*), a moisture index between June and August (*mmind68*), daily average global potential shortwave radiation per month (*sumradyy*), the annual average number of frost days during the growing season (*sfroyy*) and topographic position (*topos*). Figure 1 shows the spatial patterns of these predictors.

We generated virtual species distributions (see Fig. 1) based on $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_5$ by fitting a GLM to real data, projecting it throughout the study area and then treating the resulting spatial predictions as the true distributions (see Wisz & Guisan 2009; Albert *et al.* 2010). Let $\boldsymbol{X}$ denote the $N \times 5$ matrix of predictors, each row of which corresponds to a different pixel. To create a surface $\{p_1,\ldots,p_N\}$ of presence probabilities from $\boldsymbol{X}$, we used the probit function to relate the predictors to the presence probabilities through

$$p(\boldsymbol{X}) = \Phi\{\eta(\boldsymbol{X})\} \qquad \text{eqn 1}$$

where Φ is the standard Gaussian cumulative distribution function. For simplicity, we included only quadratic terms for each predictor, and we did not consider interactions, that is, we let

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

with coefficients α_j and β_j chosen to produce a realistic distribution for each virtual species. We used different values of α and β to construct $S = 10$ surfaces, thereby creating S virtual species with different responses to the predictors. These species were generated using parameters estimated by fitting model (1) to data on ten real species distributed in the study area, chosen to represent different taxonomic groups, distributions and habitat types (see the Supporting Information). Once a presence probability function $p(\boldsymbol{X})$ corresponding to a virtual species was defined, we simulated a presence or absence by generating independent standard normal variables $\gamma_1,\ldots,\gamma_N$: the random variables

$$Y_i = \begin{cases} 1, & \gamma_i \leq \eta_i, \\ 0, & \gamma_i > \eta_i, \end{cases} \qquad \text{eqn 3}$$

that indicate presence or absence at a site whose predictors $\boldsymbol{X}_i$ yield $\eta_i = \eta(\boldsymbol{X}_i)$ are equal to 1 with probabilities $p_i = \Phi(\eta_i)$ given by eqn (1). We used the probit link in (1) because its representation (3) in terms of normal variables simplifies the simulation of correlated presences–absences; see below. Using a logit link instead had almost no effect on our results.

Spatial autocorrelation (SAC) represents the clustering tendency of presences and absences (Cablk, White & Kiester 2002; Segurado, Araújo & Kunin 2006; Dormann 2007; Dormann *et al.* 2008). It may be explained by the effect of spatial covariates such as altitude or by biological processes of species dispersal or interactions with other species. Although these are two different types of SAC, they both usually result in spatially correlated residuals from the fit of a SDM and are difficult to distinguish based on data (Guisan & Thuiller 2005; Dormann *et al.* 2007; Beale *et al.* 2010). The distinction between these types of SAC is important for the simulation: SAC due to missing covariates affects the function $\eta(\boldsymbol{X})$ and thus is present at the probability level but does not add further randomness to the simulations, whereas SAC due to a biological process does not change the marginal probabilities p_i but gives different clusterings of the presences and absences in different simulations. We simulated these two types of SAC as follows.

SAC due to an unobserved spatially varying covariate was generated by excluding a predictor from our virtual data set, so that the spatial variation of the presences could not be entirely explained using the remaining predictors. For each virtual species, the predictor to be excluded was determined by fitting a probit model (1) to the simulated presence–absence data, measuring the reduction in the likelihood due to excluding one predictor at a time from eqn (2) and then excluding the second most important predictor. Correlation between the spatial predictors means that this type of SAC will generally alter the estimates of the remaining parameters. The range of SAC for the excluded variable is not controlled.

SAC representing a stochastic colonization or dispersal process was simulated by using correlated γ_i in (3): we replaced the independent γ_i by a spatial Gaussian process $\gamma(\boldsymbol{s})$, where $\boldsymbol{s}$ is the pixel location. The process $\gamma(\boldsymbol{s})$ has zero mean, unit variance and correlation function ρ;

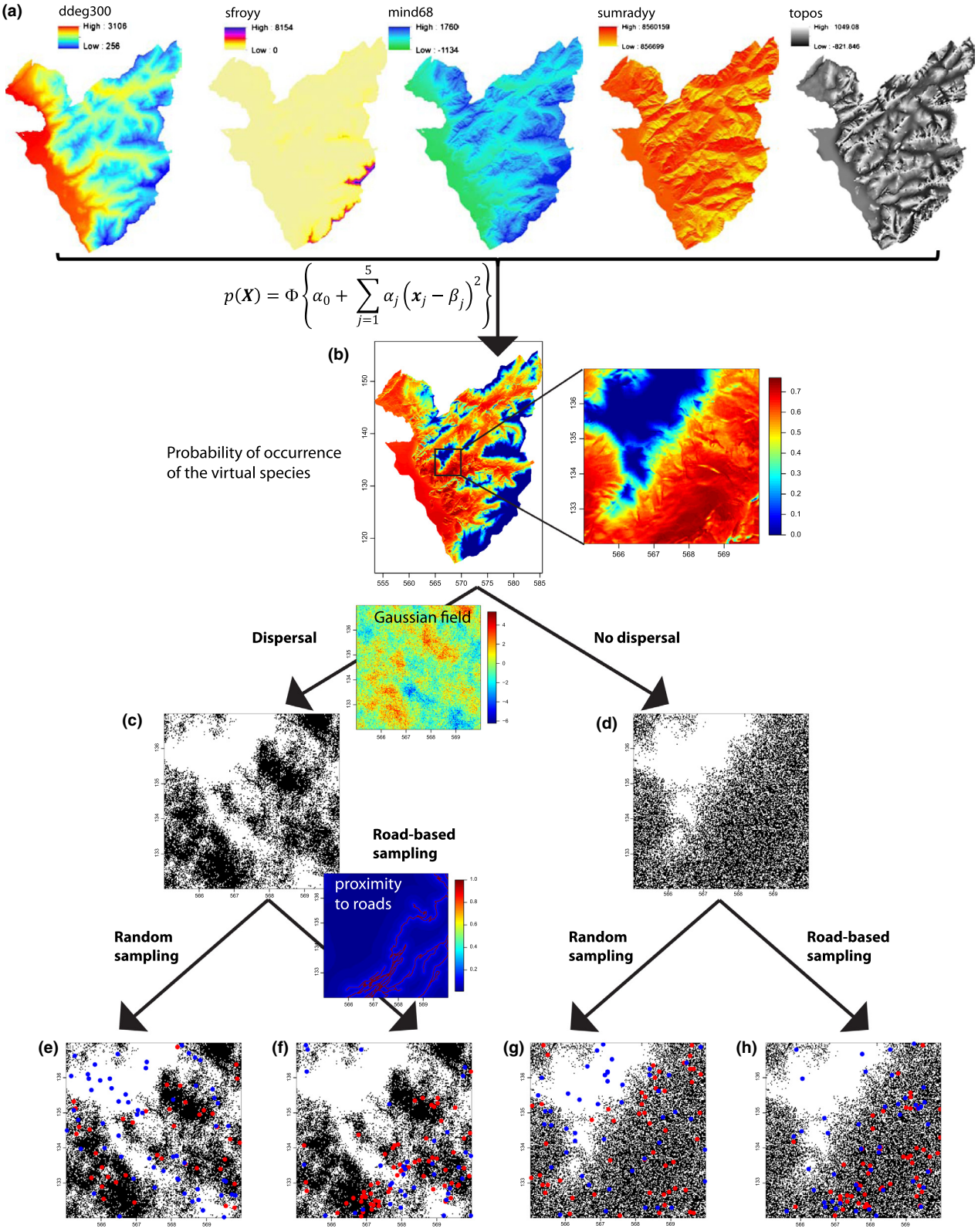

Fig. 1. Simulation of the first virtual species. We use the five real predictors in (a) and the probit function to define the map of probability of occurrence in (b). The presence–absence process is generated with SAC in (c) using a Gaussian random field having an effective range of 1·5 km, and in (d) without SAC using independent normal variables. For each configuration of dispersal (true or false), a total of 100 presences (red dots) and absences (blue dots) are sampled using a simple random design in (e) and (g), and a road-based design in (f) and (h).

see Fig. 1. This type of SAC is easy to implement, and its range can be precisely controlled using the parameters of the correlation function. The marginal probability of presence at location $\boldsymbol{s}$ is still given by eqn (1), but with presences and absences tending to cluster. This type of SAC increases the variance of SDM predictions but does not bias them. We used the correlation function $\rho(h) = \exp(-h/\lambda)$, where h is the distance (km) between locations and λ is a range parameter. We estimated λ by fitting a spatial probit model to the ten real species that were used to create our virtual species, giving estimates of λ that were all smaller than 0·5 (see Table S1). Although we cannot be sure that this SAC corresponds to a real dispersal process, we fixed $\lambda = 0{\cdot}5$ in our simulation for comparability with the real data. This is justified by the plausible short-range dependence in plant species distributions; here, the correlation of $\gamma(\boldsymbol{s})$ is lower than 0·05 beyond 1·5 kilometres, an upper bound for the maximum distance reported for most dispersal strategies among the plant species in the study area (Vittoz & Engler 2007).

For each of the $S = 10$ virtual species, and for each combination of missing covariate (true or false) and dispersal (true or false), we independently simulated the spatial presence–absence process $R = 10$ times to account for stochastic variability.

SAMPLING VIRTUAL SPECIES

When sampling virtual species, we controlled the sample size n and the design. We created training samples with $n = 100$ and $n = 500$, sampling both presences and absences. The number of presences in the training samples varied between 20 and 80% in both cases. We used two sampling designs: a simple random (or uniform) design (Albert *et al.* 2010), for which every pixel has the same probability of being sampled, and a design under which locations that are close to roads are more likely to be sampled, as occurs frequently in practice (also called sampling bias, see Graham *et al.* 2004) ; see Fig. 1. With each training sample of size n, we also uniformly sampled an independent test sample $\mathcal{T}$ of size $n' = 5000$ from all pixels not in the training sample; the test sample is used to evaluate the fitted models.

This sampling of the virtual species is repeated $R' = 5$ times on each of the $R = 10$ previously simulated presences and absences for each of the $S = 10$ simulated species. This hierarchical simulation scheme permits us to separate the effects of species variation and sampling procedure. For computational reasons, simulation of the dispersal process takes place during the random sampling stage, as it is infeasible to generate the Gaussian process at 10^6 sites.

FITS AND EVALUATIONS

We used four different modelling techniques to estimate the presence–absence probabilities for each simulated species, using five or four predictors depending on whether some SAC was due to a missing predictor. The first was a generalized linear model (GLM, McCullagh & Nelder 1989) for binary responses (the presence–absence data) with the probit link function, quadratic terms for each predictor and no interactions. If no covariate is missing, this is the true model that generates our data via eqn (1). The second, which was fitted using the R package gam (Hastie 2013), was a generalized additive model (GAM; Hastie & Tibshirani 1990) with binary response, the probit link, smoothing spline terms and no interactions between predictors. The third was a maximum entropy (MaxEnt; Phillips, Anderson & Schapire 2006) model. MaxEnt models presence-only data using background 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. MaxEnt was fitted using the R package dismo (Hijmans *et al.* 2013) with default settings, with the logistic output and interactions modelled only when there were over 80 presences; see the Supporting Information. Finally, random forests (RF; Breiman 2001) were fitted (using the R package randomForest, Liaw & Wiener 2002) with default settings, that is, 500 trees and with two variables at each node of each tree. Each of these four modelling techniques provides estimates $\hat{p}_i$ at the locations $i \in \mathcal{T}$, 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.

There are many measures of model accuracy. We focused on the root-mean-squared error (RMSE; Caruana & Niculescu-Mizil 2004; Liu, White & Newell 2011) owing to its advantages in our context: it compares probabilities to probabilities, it is easily interpretable, and the distribution of log(RMSE) was found to be appropriate for the analysis of variance. The RMSE is computed by comparison of the estimated presence probabilities with the true marginal probabilities (unknown in real applications):

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

with $n' = \mathrm{Card}(\mathcal{T})$, and where the true probabilities p_i come from eqn (1). This RMSE corresponds to the accuracy related to the estimation of the environmental niche of the species, excluding possible undesirable effects of the SAC at the presence–absence level. The RMSE may be interpreted as the mean distance between predicted and true probabilities over the locations of the test sample. As the RMSE involves the true p_i, it can be computed in simulation settings, although not for real data. Other measures, such as the area under the receiver operating characteristic curve (AUC; Mason & Graham 2002) or the point-biserial correlation (COR; Tate 1954), popular in SDM, can also be computed, but using them in the same framework is more complicated. The Supporting Information contains more discussion on accuracy measures.

SDMs can typically be evaluated on landscapes other than those used to fit the models (Randin *et al.* 2006), where different interactions between the predictors can affect their prediction accuracy. The procedure presented here easily extends to validation in external landscapes. In addition to measuring the importance of the factors in the original landscape, we projected the true distributions of the ten virtual species into three other regions of Switzerland, sampled test localities and calculated the RMSE in these regions; see the Supporting Information. The RMSEs for these regions were then analysed separately.

STATISTICAL ANALYSIS

The simulation scheme described above requires the application of four modelling techniques to each of $2^4 \times R \times R' \times S$ samples. With $S = 10$ simulated species, $R = 10$ presence-absence simulations, and $R' = 5$ sampling patterns, we have 8000 data sets, to each of which we apply four techniques. The entire procedure, from the simulation of species to the evaluation of predictions, takes less than a day using the statistical environment R with parallel computing on eight CPUs. Although there is no limit on the number of presence–absence simulations R or the number of species S, it is more difficult to increase the number R' of sampling patterns, as the simulation of dispersal using the Gaussian process involves the calculation of inverse matrices of side approximately $1200 \times R'$; this can be computer-intensive.

To assess the importance of the different factors, we propose to study variation in the values of log(RMSE) via a linear mixed-effects model.

The log transformation stabilizes the variance of the RMSEs; see Fig. 2. For each simulated species, there are five factors: *missing* for the presence of a missing influential covariate (T/F); *dispersal* for the presence of a dispersal process (T/F); *n* the sample size (100 or 500); *design* for the sampling design (simple random/road-based); and finally, *technique* for the different modelling techniques.

As we built our simulation hierarchically, values from the same species, from the same random simulation or from the same sampling pattern, are expected to be similar. Hence, we use a model with three nested random effects, corresponding to the species, the simulation and the sampling pattern:

$$\log(\mathrm{RMSE})_{kmsindjt} = \alpha_0 + \alpha_{m*s*n*d*t} + \varepsilon_k + \varepsilon_{kmsi} + \varepsilon_{kmsindj} + \varepsilon_{kmsindjt}, \quad \text{eqn 4}$$

with α_0 the intercept and $\alpha_{m*s*n*d*t}$ the 32 fixed-effect parameters for *missing* (m), *dispersal* (s), *n* (n), *design* (d), *technique* (t) and all their interactions. The indices $k = 1,\ldots,10$, $i = 1,\ldots,10$ and $j = 1,\ldots,5$, respectively, represent the ten simulated species, the ten presence–absence simulations and the five sampling patterns. The ε_k, ε_{kmsi} and $\varepsilon_{kmsindj}$ are independent zero-mean random variables that account for the species, the presence–absence simulations and the sampling patterns. The $\varepsilon_{kmsindjt}$ are independent centred normal random variables that correspond to the finest level of variation, representing the errors in the linear mixed-effects model. This is a classical split-plot design with nested random effects; see section 10.2 of Venables & Ripley (2002). Inspection of the residuals indicated that (4) was appropriate for modelling the log(RMSE).

Different procedures have been proposed to examine the relative importances of factors in linear models. Hierarchical partitioning (Chevan & Sutherland 1991; Mac Nally 2000) has been widely used, and in a similar context by Dormann *et al.* (2008) in particular. In a balanced design such as ours, this method essentially compares the sums of squares corresponding to each factor in the analysis of variance (ANOVA, Table 1). Although the importances of different factors are easily compared when the sums of squares for interactions are tiny compared with those for main effects, this is harder when some interactions have large effects, as in the present case. Thus, in addition to ANOVA, we compute the marginal and conditional coefficients of determination, R^2, which were defined in the context of mixed-effects models by Nakagawa & Schielzeth (2013). We use the R package nlme (Pinheiro *et al.* 2013) to fit model (4) by restricted maximum likelihood (REML; Venables & Ripley 2002), with each factor and all its

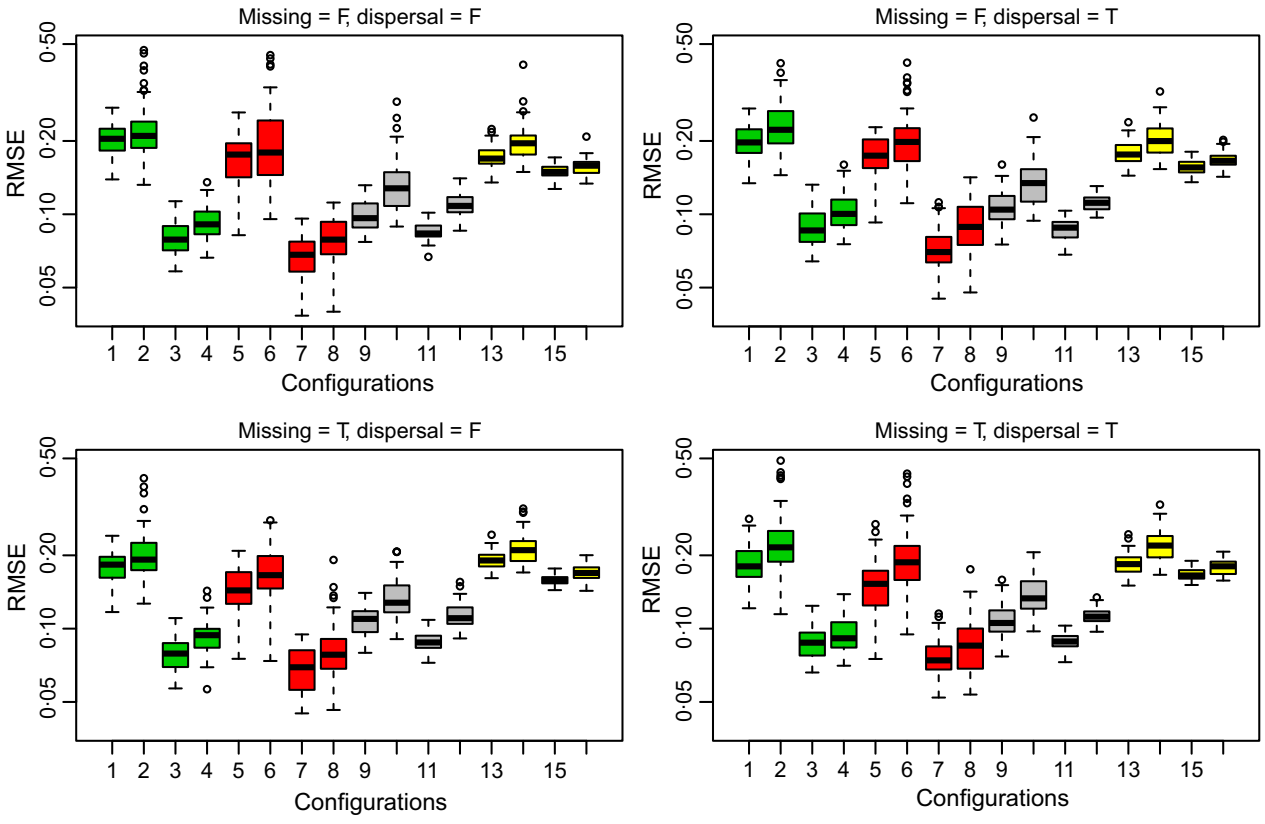

Fig. 2. Boxplots of RMSE for the first simulated species, with a log scale on the vertical axis. Each box shows the variation due to the five samples for each of the 10 simulations and for one particular configuration of factors (missing covariate, dispersal, modelling technique, sample size *n* and 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 of the four panels corresponds to a different combination of the factors *missing* and *dispersal*. In each panel, the four modelling techniques are distinguished by colours: generalized additive models (GAMs) in green, generalized linear models (GLMs) in red, maximum entropy (MaxEnt) in grey and random forests (RF) in yellow. Inside the colour 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 in the top left panel correspond to the 50 RMSE values for GAMs with (sample size, sampling design) settings (100, simple random), (100, road-based), (500, simple random) and (500, road-based), respectively, for data with no missing covariate and no dispersal process. Configurations 5–8 in the same panel show the RMSEs for GLMs, configurations 9–12 for MaxEnt, and 13–16 for RF, all with the corresponding settings.

Table 1. Analysis of variance for the linear mixed-effects model (4). Only factors for which the corresponding *P*-value is smaller than 0·05 are shown. A colon denotes interaction

Error level	Factor	Df	Sum Sq	Mean Sq	*F*-value	$\log_{10}(p)$
Species	Residual	9	124·686	13·854		
Simulation in species	Missing	1	15·835	15·835	159	−30
	Dispersal	1	12·363	12·363	124	−24
	Residual	387	38·582	0·1		
Sampling in simulation	*n*	1	1723·328	1723·328	20 074	−2133
	Design	1	373·194	373·194	4347	−748
	Missing:*n*	1	15·62	15·62	182	−40
	Dispersal:*n*	1	4·69	4·69	55	−13
	Missing:Design	1	2·449	2·449	29	−7
	n:Design	1	0·581	0·581	7	−2
	Missing:*n*:Design	1	1·618	1·618	19	−5
	Residual	7588	651·414	0·086		
Within sampling	Technique	3	480·571	160·19	4636	−2380
	Missing:Technique	3	7·964	2·655	77	−49
	Dispersal:Technique	3	3·529	1·176	34	−21
	n:Technique	3	702·992	234·331	6782	−3196
	Design:Technique	3	60·178	20·059	581	−364
	Missing:*n*:Technique	3	11·517	3·839	111	−71
	Dispersal:*n*:Technique	3	1·003	0·334	10	−6
	Missing:Design:Technique	3	0·929	0·31	9	−5
	n:Design:Technique	3	1·346	0·449	13	−8
	Missing:*n*:Design:Technique	3	1·662	0·554	16	−10
	Residual	23 952	827·58	0·035		

interactions excluded, and compare the resulting marginal R^2, which indicates the proportion of variance explained by the fixed effects in the model, with that of the full model. Thus, excluding one factor leads to the exclusion of 16 terms, allowing us to measure the full contribution of that factor to the model (4), including any interactions with other factors. Although R^2 is easy to interpret and has desirable properties, there are some difficulties associated with it, and we discuss these and compare it with likelihood in the Supporting Information.

Results

Figures 2 and 3 show that sample size and modelling technique are the largest sources of variability in prediction accuracy, followed by the sampling design; the effects of the missing covariate and the dispersal process, that is, the two possible sources of SAC, are less visible. Although there are differences among the species, all the graphs exhibit the same general pattern (see Supporting Information).

The ANOVA (Table 1) and the marginal R^2 (Table 2) were used to quantify these visual impressions. The sample size n and the modelling technique are the most important factors. Varying n from 100 to 500 produces the greatest reduction in RMSE. The variation among modelling techniques can be huge, as seen in Fig. 3, but this factor appears to be slightly less important than sample size in the ANOVA and in terms of reduction of the marginal R^2. There is a strong interaction between modelling technique and sample size, as can be seen in the ANOVA and in Fig. 3. Although GAMs show generally poor performance in small samples, MaxEnt and RF are much less affected by small n. For the largest sample size, $n = 500$, GLMs are generally the best models, but for $n = 100$, they tend to predict worse than MaxEnt. In most cases, using MaxEnt on presence–absence data yields the best predictions in small samples. The effect of the sampling design is important, with a strong interaction with modelling technique: GLM and GAM are more sensitive to the choice of sampling design than are MaxEnt and RF (see the boxplots for the ten species in the Supporting Information). The dispersal process and the missing covariate are the least important factors in our study.

For the three other regions of Switzerland, the ranking of the five factors changed only slightly. Sample size and modelling technique still had the largest impacts, while SAC due to the dispersal process remained least important (see the Supporting Information).

Discussion

This paper proposes a simulation method for assessing the relative importance of factors in SDMs. The value of our general approach is that it provides a framework for simulation experiments, with a great variety of species, landscapes and ranges of factors, that allow the assessment of the effects of SAC and further complications in other settings. In an application to five key factors in a real data framework – virtual plants in a mountain region – sample size and modelling technique had the largest relative effects on predictions, followed by the sampling design. In this illustration, the presence of SAC in the residuals, whether due to the dispersal process or a missing covariate, appeared to be of lower importance, despite being seen by many authors as a major issue (e.g. Segurado, Araújo & Kunin 2006; Dormann *et al*. 2007). This suggests that the effect of SAC on prediction accuracy of SDMs may be relatively minor when ecological and topographic configurations are similar to

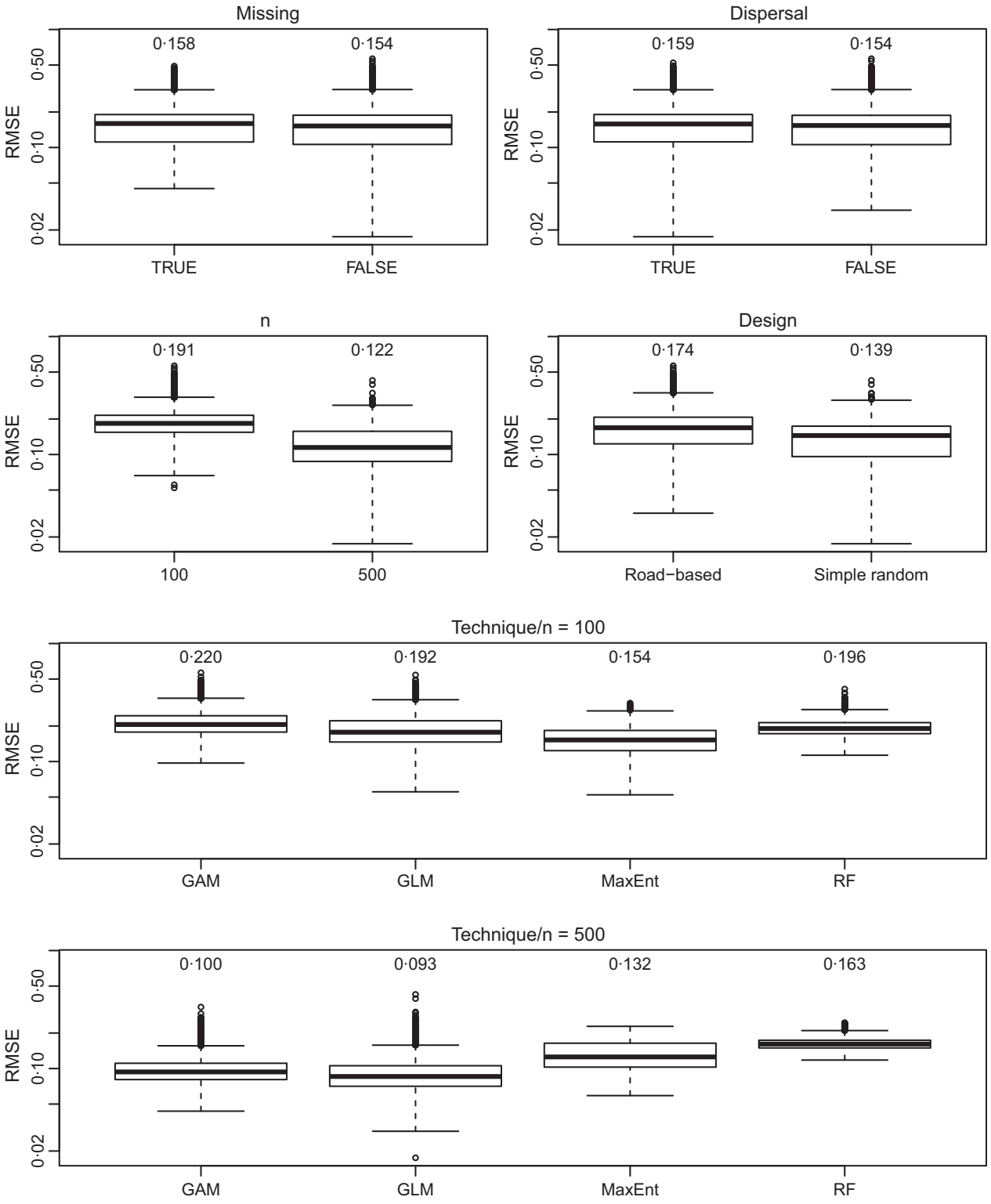

Fig. 3. Boxplots of all RMSE values (for the ten virtual species) for each configuration of the five factors. The mean of the RMSE values for each configuration is printed over the corresponding boxplot.

our study area. Our analysis and virtual case study also confirmed the findings of Elith *et al.* (2006) and Wisz *et al.* (2008) regarding the performance of machine learning techniques such as MaxEnt, which generally outperform GLMs and GAMs in small samples. For the larger sample size, we found that GLMs were generally best, which is not surprising as our virtual species were generated using a GLM, and that MaxEnt and RF, which both model interactions between predictors, gave poorer predictions. The good performance of MaxEnt in small presence-absence samples merits further investigation.

Table 2 . Marginal and conditional R^2 (Nakagawa & Schielzeth 2013) for the full model (4) and the five submodels with one factor, and all its interactions with the other factors, excluded at a time

Model	Full model	Missing	Dispersal	*n*	Design	Technique
Marginal R^2	0·674	0·662	0·669	0·188	0·587	0·423
Conditional R^2	0·782	0·777	0·781	0·594	0·766	0·450

We chose values for factors, such as the sample size, to agree with applications related to our particular data set, but of course, these values are not universal. In particular, our simulation of SAC at the presence–absence level is intended to represent species dispersal at distances of around 1·5 kilometres. For the plant species of our data set, this seems realistic (Vittoz & Engler 2007). A longer-range correlation increases the relative importance of dispersal (see the Supporting Information), but seems less plausible. SAC resulting from an unobserved spatial covariate can yield dependence at much larger distances, but can also be removed by including the corresponding predictor, so its importance is unclear. It can be argued that this type of SAC does not affect the independence of the observations (Lichstein *et al.* 2002; Diniz-Filho, Bini & Hawkins 2003; Guisan & Thuiller 2005). Furthermore, we have only investigated the relative effects of the factors in terms of prediction accuracy, but other properties such as model selection, coefficient estimation and uncertainty estimation may be of interest (see e.g. Dormann 2007). Although here SAC had relatively low impact on prediction accuracy, its effect on model selection and on the estimation of standard errors may be larger, as in a strong dispersal process, the degrees of freedom may be lower than the number of observations.

The individual effects of different factors on SDM predictions have been the object of many previous studies, but, to our knowledge, only Dormann *et al.* (2008), Diniz-Filho *et al.* (2009) and Garcia *et al.* (2012) considered the relative effects of factors on SDMs jointly. Diniz-Filho *et al.* (2009) and Garcia *et al.* (2012) evaluated the relative uncertainty due to the use of different modelling techniques and climate models to predict species responses to global change. Dormann *et al.* (2008) considered the relative effects of modelling technique, data uncertainty, collinearity correction and variable selection method, but did not include SAC. This last study only used a single species and did not use as comprehensive a statistical framework as that used here. In contrast, our approach uses virtual data from a systematic experimental design that appropriately includes the three steps of species simulation, sampling procedure and modelling. Our framework is ecologically realistic and integrates randomness into the species distributions and sampling procedure. Other factors, such as location error, predictor error, multicollinearity or the effect of using pseudo-absences, could also be assessed using the same framework.

Possibilities for future work include the study of the effects of niche complexity on prediction accuracy. Techniques such as MaxEnt or RF might be expected to be more efficient for modelling complex responses to predictors and thus might give more accurate predictions than GLMs or GAMs in realistic situations, where species responses to predictors can be highly nonlinear and depend on interactions between predictors. We acknowledge that the simulation of our virtual species may not be fully ecologically realistic, and we encourage users of our methodological framework to explore alternative ways to simulate virtual species; no universal method currently exists, and more investigation would be valuable. Another topic requiring further research is the relations between the different measures of accuracy used to assess the predictive power of SDMs. The RMSE appeared to be a natural choice in the framework of our simulations, but its relation to other metrics would benefit from further clarification.

We hope that our paper will pave the way towards more systematic analyses of factors affecting predictive models, especially when these and their predictions are subsequently to be used to derive scenarios of global change impact on biodiversity.

Acknowledgements

We thank M.G. Genton, P.J. Solomon, N.G. Yoccoz, R.B. O'Hara and anonymous reviewers for comments on the manuscript. This study was funded by the Swiss National Science Foundation in the context of the NCCR Plant Survival.

Data accessibility

Species data: uploaded as online supporting information.

R scripts: uploaded as online supporting information.

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.

Baselga, A. & Araújo, M.B. (2009) Individualistic vs. community modelling of species distributions under climate change. *Ecography*, **32**, 55–65.

Beale, C.M. & Lennon, J.J. (2012) Incorporating uncertainty in predictive species distribution modelling. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **367**, 247–258.

Beale, C.M., Lennon, J.J., Yearsley, J.M., Brewer, M.J. & Elston, D.A. (2010) Regression analysis of spatial data. *Ecology Letters*, **13**, 246–264.

Breiman, L. (2001) Random forests. *Machine Learning*, **45**, 5–32.

Buisson, L., Thuiller, W., Casajus, N., Lek, S. & Grenouillet, G. (2010) Uncertainty in ensemble forecasting of species distribution. *Global Change Biology*, **16**, 1145–1157.

Cablk, M., White, D., Kiester, A.R. (2002) Assessment of spatial autocorrelation in empirical models in ecology. *Predicting Species Occurrences: Issues of Accuracy and Scale* (eds J. ScottP. HeglundM. MorrisonJ. HauflerM. RaphaelW. Wall & F. Samson), pp. 429–440. Island Press, Covelo, California, USA.

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, New York, USA.

Chevan, A. & Sutherland, M. (1991) Hierarchical Partitioning. *The American Statistician*, **45**, 90–96.

Clark, J.S., Carpenter, S.R., Barber, M., Collins, S., Dobson, A., Foley, J.A., *et al.* (2001) Ecological forecasts: an emerging imperative. *Science*, **293**, 657–60.

Côté, I.M. & Reynolds, J.D. (2002) Conservation biology. Predictive ecology to the rescue? *Science*, **298**, 1181–1182.

Diniz-Filho, J.A.F., Bini, L.M. & Hawkins, B.A. (2003) Spatial autocorrelation and red-herrings in geographical ecology. *Global Ecology and Biogeography*, **12**, 53–64.

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., *et al.* (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. & 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., Graham, C.H., Anderson, R.P., Dudík, M., Ferrier, S., Guisan, A., *et al.* (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**, 129–151.

Franklin, J. (2010) *Mapping Species Distributions. Spatial Inference and Prediction*. Cambridge University Press, Cambridge, UK.

Garcia, R.A., Burgess, N.D., Cabeza, M., Rahbek, C. & Araújo, M.B. (2012) Exploring consensus in 21st century projections of climatically suitable areas for African vertebrates. *Global Change Biology*, **18**, 1253–1269.

Graham, C.H., Ferrier, S., Huettman, F., Moritz, C. & Peterson, A.T. (2004) New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology & Evolution*, **19**, 497–503.

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

Hastie, T. (2013) *gam: Generalized Additive Models*. R package version 1.08.

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

Hastie, T.J. & Tibshirani, R.J. (1990) *Generalized Additive Models*. Chapman & Hall, London.

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

Hirzel, A. & Guisan, A. (2002) Which is the optimal sampling strategy for habitat suitability modelling. *Ecological Modelling*, **157**, 331–341.

Liaw, A. & Wiener, M. (2002) Classification and regression by randomForest. *R News*, **2**, 18–22.

Lichstein, J.W., Simons, T.R., Shriner, S.A. & Franzreb, K.E. (2002) Spatial autocorrelation and autoregressive models in ecology. *Ecological Monographs*, **72**, 445–463.

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

Mac Nally, R. (2000) The distinction between –and reconciliation of –'predictive' and 'explanatory' models. *Biodiversity & Conservation*, **9**, 655–671.

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.

McCullagh, P. & Nelder, J.A. (1989) *Generalized Linear Models*, 2nd edn. Chapman & Hall, London.

Miller, J.A. (2014) Virtual species distribution models: using simulated data to evaluate aspects of model performance. *Progress in Physical Geography*, **38**, 117–128.

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.

Pereira, H., Leadley, P., Proenca, V., Alkemade, R., Scharlemann, J., Fernandez-Manjarres, J., *et al.* (2010) Scenarios for global biodiversity in the 21st century. *Science*, **330**, 1496–1501.

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*. Princeton University Press, Princeton, USA.

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

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. & R Development Core Team (2013) *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-109.

R Core Team (2013) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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.

Saupe, E.E., Barve, V., Myers, C.E., Soberón, J., Barve, N., Hensz, C.M., Peterson, A.T., Owens, H.L. & Lira-Noriega, A. (2012) Variation in niche and distribution model performance: The need for a priori assessment of key causal factors. *Ecological Modelling*, **237–238**, 11–22.

Schröter, D., Cramer, W., Leemans, R., Prentice, I.C., Araújo, M.B., Arnell, N.W., *et al.* (2005) Ecosystem service supply and vulnerability to global change in Europe. *Science*, **310**, 1333–1337.

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

Soberon, J. (2007) Grinnellian and Eltonian niches and geographic distributions of species. *Ecology Letters*, **10**, 1115–1123.

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., Araújo, M., Pearson, R., Whittaker, R., Brotons, L. & Lavorel, S. (2004) Uncertainty in predictions extinction risk. *Nature*, **430**, 34.

Venables, W.N. & Ripley, B.D. (2002) *Modern Applied Statistics with S*, 4th edn. Springer, New York. ISBN 0-387-95457-0.

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 Perdicting Species Distributions Working Group (2008) Effects of sample size on the performance of species distribution models. *Diversity and Distributions*, **14**, 763–773.

Zurell, D., Berger, U., Cabral, J.S., Jeltsch, F., Meynard, C.N., Munkemüller, T., *et al.* (2010) The virtual ecologist approach: simulating data and observers. *Oikos*, **119**, 622–635.

Received 5 March 2014; accepted 24 April 2014
Handling Editor: Robert B. O'Hara

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Appendix S1. Supporting Information.

Appendix S2. Zip file with R scripts and data for the simulations.