Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2012, **3**, 327–338 doi: 10.1111/j.2041-210X.2011.00172.x

Selecting pseudo-absences for species distribution models: how, where and how many?

Morgane Barbet-Massin[1]*, Frédéric Jiguet[1], Cécile Hélène Albert[2,3] and Wilfried Thuiller[3]

[1]*Muséum National d'Histoire Naturelle, UMR 7204 MNHN-CNRS-UPMC, Centre de Recherches sur la Biologie des Populations d'Oiseaux, CP 51, 55 Rue Buffon, 75005 Paris, France;* [2]*Department of Biology, McGill University, 1205 Docteur Penfield, Montréal, QC, Canada; and* [3]*Laboratoire d'Ecologie Alpine, UMR-CNRS 5553, Université Joseph Fourier, Grenoble I, BP 53, 38041 Grenoble Cedex 9, France*

Summary

1. Species distribution models are increasingly used to address questions in conservation biology, ecology and evolution. The most effective species distribution models require data on both species presence and the available environmental conditions (known as background or pseudo-absence data) in the area. However, there is still no consensus on how and where to sample these pseudo-absences and how many.

2. In this study, we conducted a comprehensive comparative analysis based on simple simulated species distributions to propose guidelines on how, where and how many pseudo-absences should be generated to build reliable species distribution models. Depending on the quantity and quality of the initial presence data (unbiased vs. climatically or spatially biased), we assessed the relative effect of the method for selecting pseudo-absences (random vs. environmentally or spatially stratified) and their number on the predictive accuracy of seven common modelling techniques (regression, classification and machine-learning techniques).

3. When using regression techniques, the method used to select pseudo-absences had the greatest impact on the model's predictive accuracy. Randomly selected pseudo-absences yielded the most reliable distribution models. Models fitted with a large number of pseudo-absences but equally weighted to the presences (i.e. the weighted sum of presence equals the weighted sum of pseudo-absence) produced the most accurate predicted distributions. For classification and machine-learning techniques, the number of pseudo-absences had the greatest impact on model accuracy, and averaging several runs with fewer pseudo-absences than for regression techniques yielded the most predictive models.

4. Overall, we recommend the use of a large number (e.g. 10 000) of pseudo-absences with equal weighting for presences and absences when using regression techniques (e.g. generalised linear model and generalised additive model); averaging several runs (e.g. 10) with fewer pseudo-absences (e.g. 100) with equal weighting for presences and absences with multiple adaptive regression splines and discriminant analyses; and using the same number of pseudo-absences as available presences (averaging several runs if few pseudo-absences) for classification techniques such as boosted regression trees, classification trees and random forest. In addition, we recommend the random selection of pseudo-absences when using regression techniques and the random selection of geographically and environmentally stratified pseudo-absences when using classification and machine-learning techniques.

Key-words: background data, bias, BIOMOD, ecological niche modelling, sampling design, virtual species

Introduction

Species distribution models (SDM) are increasingly used to address numerous questions in conservation biology, ecology

*Corresponding author. E-mail: barbet@mnhn.fr
Correspondence site: http://www.respond2articles.com/MEE/

and evolution (Guisan & Thuiller 2005). They have been used to test biogeographical, ecological and evolutionary hypotheses (Graham *et al.* 2004a), to predict species' invasion and proliferation (Peterson & Vieglais 2001), to assess the impact of climate, land use and other environmental changes on species distributions (Thuiller *et al.* 2005), to improve surveys for rare species by identifying sites where the probability of occurrence is high (Engler, Guisan & Rechsteiner 2004) and to support conservation planning and reserve selection (Marini *et al.* 2009).

The SDM widely used in these studies can be categorised in two groups: methods that only require presence data vs. those that require both presence and absence data (Brotons *et al.* 2004). Contrary to popular belief, there are very few presence-only SDM, the most common being rectilinear envelope (e.g. BIOCLIM, Busby 1991) and distance-based envelope (e.g. Mahalanobis distance, Farber & Kadmon 2003). SDM such as Maxent or GARP, sometimes misleadingly referred to as presence-only methods, actually do require the use of background data or pseudo-absence data. As confirmed absences are very difficult to obtain, especially for mobile species, and require higher levels of sampling effort to ensure their reliability compared with presence data (Mackenzie & Royle 2005), presence-only models have often been used to cope with the lack of absence data (Graham *et al.* 2004b). However, comparisons of various SDM show that presence–absence models tend to perform better than presence-only models (Elith *et al.* 2006). Thus, presence–absence models are increasingly used when only presence data is available, by creating artificial absence data (usually called pseudo-absences or background data).

As false absence data can have negative effects on SDM (Gu & Swihart 2004), different strategies have been proposed to improve the selection of an appropriate pseudo-absence data set. Some studies have suggested using pseudo-absence data selected outside a pre-defined region based on a simple preliminary model or based on a minimum distance to the presence (Zaniewski, Lehmann & Overton 2002; Engler, Guisan & Rechsteiner 2004; Lobo, Jimenez-Valverde & Hortal 2010). If presences of the studied species have been collected during field surveys that also considered other species, such that bias in the sampling design is the same for all species, better results can be obtained by taking pseudo-absences within the presence points of these other species (Phillips *et al.* 2009). To our knowledge, the influence of the number of pseudo-absences selected has rarely been investigated. For the Maxent technique, Phillips & Dudik (2008) found that predictive accuracy was higher with around 10 000 background pseudo-absences. Nevertheless, prevalence (defined here as the ratio of the quantity of presence data to the quantity of absence data used to fit the model) has been shown to influence model accuracy (McPherson, Jetz & Rogers 2004). Although very informative, most of these previous studies used empirical data without knowing the true distribution of the species, the sampling design or presence data bias (for discussion on bias and sampling design, see Albert *et al.* 2010). Indeed, besides the obvious problems related to unreliable absence data, the presence data may also be biased or incomplete, depending on the sampling scheme, accuracy of the data and species detection probability (Barbet-Massin, Thuiller & Jiguet 2010). Generalisation and application of the conclusions of these empirical studies are therefore of limited interest in general compared with conclusions from virtual experiments where results or patterns can be compared with the known truth (Zurell *et al.* 2010).

The goal of this study is to systematically test the effect of known sources of variability related to the selection of pseudo-absence data to deliver a comprehensive guideline on how, where and how many pseudo-absences should be generated to build unbiased and reliable SDM. Here, we aimed to answer the following questions:

(a) Which ratio of presences/absences achieves the highest model accuracy?

(b) What is the optimal number of replicate sets of pseudo-absences?

(c) What is the optimal number and weighting scheme of pseudo-absences per replicate?

(d) Which method for generating pseudo-absences results in the most accurate models?

(e) How does bias in the sampling design influence the optimal use of pseudo-absences?

(f) Which parameters (number of pseudo-absences, method of generating pseudo-absences and weighting scheme) have the greatest influence on the models' predictive accuracy?

For each one of these six questions, we further tested for an effect of the number of presences available and the choice of the modelling technique, using seven different SDM. To do so, we performed a comparative analysis based on virtual data. We thus knew the species' true distribution and were able to simulate different realisations of this distribution that were either unbiased or purposely biased geographically or climatically. Geographically biased presence data could arise from sampling along main roads or railways, or within a subset of the countries where the species occurs (Kadmon, Farber & Danin 2004; Albert *et al.* 2010). Geographical bias matches some large-scale surveys like the North American Breeding Bird Survey with sampling sites along the main roads or some common data sets used for species distribution modelling which follow political boundaries (e.g. European breeding birds, Huntley *et al.* 2008). Climatically biased presence data can result either from a spatially biased sampling design, that is, when data from an area with climatically different characteristics are missing (Barbet-Massin, Thuiller & Jiguet 2010), or from sampling that was not carried out over the whole environmental range of a given species, which is often the case for species ranging from low to very high altitude, because the latter is usually less thoroughly surveyed.

Methods

CREATING VIRTUAL SPECIES

To make sure that our results were not influenced by the choice of a species and the peculiarities thereof, we created two geographically distinct virtual species (Fig. S1). To produce the simplest possible

potential distributions based on uncorrelated variables, we constrained the distributions of these virtual species by two explanatory variables. To include realistic environmental conditions, we chose these two uncorrelated environmental variables as the first two axes of a principal component analysis (PCA) conducted on eight variables related to temperature and precipitation at European scale (from the Worldclim data base at a 10 arc-min resolution): (i) annual mean temperature, (ii) mean temperature of the warmest month, (iii) mean temperature of the coldest month, (iv) temperature seasonality, (v) annual precipitation, (vi) precipitation of the wettest month, (vii) precipitation of the driest month, (viii) precipitation seasonality. For each species, we assumed a bell-shaped relationship between the probability of occurrence and each composite environmental variable. Each fundamental niche is therefore an ellipsoid in the principal component space, as previously used by Godsoe (2010) and Soberon & Nakamura (2009), although the geographical points falling within that environmental ellipsoid can result in a distorted ellipsoid, depending on its position in the environmental space cloud (Soberon & Nakamura 2009) (Fig. S1). Although Gaussian response curves might seem unrealistic at a first glance, this is what is expected from a theoretical point of view (Lawton 1999). Whilst the SDM accuracy (in absolute terms) may depend upon the response curves chosen to create the virtual species, this choice should not influence how different methods for generating pseudo-absences affect the quality of a given SDM (in relative terms). The virtual species reflect similar ecological constraints (same shape of response curves to the same environmental variables), to ensure our results reflect differences resulting from the methods used to generate pseudo-absences and not differences arising from species characteristics.

The probability of occurrence of each species in a given pixel was calculated by multiplying the probabilities linked to both variables. This final probability distribution was then rescaled so that the maximum probability was equal to 1. Finally, a binary realisation of the potential distribution was generated by applying an arbitrary probability of occurrence threshold of 0·25.

Given that in the real world, a species may not totally fill its potential distribution and is more likely to be present where the climate is most suitable, we computed an 'actual' distribution by generating presences following a binomial distribution, with a different probability of success for each pixel (i.e. the probability previously calculated). Each 'actual' distribution was made up of *c.* 2700 presence points.

SELECTING SETS OF PRESENCES USED FOR MODEL CALIBRATION

Sampling bias

To investigate the effects of sampling bias in presence data on the models' predictive accuracy (*question e*), we created three biased sub-distributions from the actual species distributions. Firstly, we created a climatically biased distribution by considering a probability surface whose Gaussian response curve means were slightly different from the means of the potential distribution (Fig. S2). Presence points of the climatically biased distribution were then sampled from the actual distribution following a binomial distribution, the probability of success for each pixel being extracted from the biased probability surface. As a result, the presence points from this sample did not include the full extent of the fundamental climatic niche of the virtual species. Secondly, we created two spatially biased samples. One was made by removing presences from several countries on one side of the distribution, and the second by only selecting presences along transport routes (roads or railways) (Fig. S1). It should be noted that the first spatial bias considered can also be interpreted as a species that does not fully occupy its potential distribution because of dispersal limitations, historical legacies and exclusion through biotic interactions. Each one of the biased samples contained approximately 1000 presence points (Fig. S1).

Number of presence points

To answer each question relative to the best use of pseudo-absences, we further tested what would be the influence of the amount of presence data. We used sample sizes of 30, 100, 300 or 1000 presence points randomly chosen from the actual distribution, the climatically biased distribution, and from each of the two spatially biased distributions (for each virtual species).

GENERATING ABSENCE DATA: TRUE ABSENCES AND PSEUDO-ABSENCES

Five different sample sizes of absence data were considered: 100, 300, 1000, 3000 or 10 000 absences. Depending on the question under consideration, we used either true absences or pseudo-absences as absence data. We considered as true absences all points located outside the potential distribution of the species, whereas pseudo-absences were always generated without considering the species potential distribution. True absences were randomly sampled among all true absences available. We used four different methods to generate the pseudo-absences (using the BIOMOD package in R, Thuiller *et al.* 2009): (i) random selection from all points within the studied area excluding available presence points ('random'), (ii) random selection of points from all points outside of the suitable area estimated by a rectilinear surface envelope from the presence sample (surface range envelope model using only presence-only data, Thuiller *et al.* 2009; hereafter, the 'SRE' method), (iii) random selection of any point located at least one degree in latitude or longitude from any presence point (the '1°far' method) and (iv) random selection of any available point located at least two degrees away from any presence point (the '2°far' method). Note that pseudo-absences can be presences that were not retained within the presence sample used to build the models (i.e. false absences).

FITTING AND ASSESSING DISTRIBUTION MODELS

For any given set of presences and absences, we used seven SDM (to detect a potential effect of the choice of the modelling method) as found in the BIOMOD package in R (see Thuiller *et al.* 2009 for further details on these modelling methods): three regression methods (GLM, GAM and MARS), two classification methods (MDA and CTA) and two machine-learning methods (BRT and RF). The models were fitted either by assigning an equal weight to each presence and absence point or by balancing the weight of presences vs. absences (*question c*), such that all presence data combined had the same weight as the total weight of the absence data (except for MARS and RF, which could not consider different weights for different data at the time of the analysis). Binary transformation was carried out using the threshold that maximised the true skill statistics (TSS; Allouche, Tsoar & Kadmon 2006). TSS corresponds to the sum of sensitivity and specificity minus one (the sensitivity is the proportion of presences correctly predicted, and the specificity is the proportion of absences correctly predicted). This threshold was shown to produce the most accurate predictions (Jimenez-Valverde & Lobo 2007). Models were evaluated using four different criteria: the area under the receiver operating characteristic (ROC) curve (AUC)

(Fielding & Bell 1997), sensitivity, specificity and TSS. These four predictive accuracy measures were calculated in reference to the potential distribution only.

(A) WHICH RATIO OF PRESENCES/ABSENCES ACHIEVED THE HIGHEST MODEL ACCURACY?

To investigate the effect of prevalence, we used four different numbers of presences (30, 100, 300 or 1000) and five different numbers of absences (100, 300, 1000, 3000 or 10 000). To make sure the results were not influenced by false positives or false negatives, presences were randomly selected from the 'actual' unbiased distribution and true absences were randomly selected as absence data. To account for the variability arising from the random selection of a set of presences, the models were fitted with 20 different random presence sets for each combination of sample size and each virtual species (Fig. 1). For each random presence set, accuracy measures were then calculated by considering the mean of the 20 distributions obtained using different random replicates of true absences as the result distribution.

(B) WHAT IS THE OPTIMAL NUMBER OF REPLICATE SETS OF PSEUDO-ABSENCES?

To investigate this issue and the four that follow, we used three different numbers of presences (30, 100 and 300), three different numbers of pseudo-absences (100, 1000 and 10 000), four methods to generate them and two different weighting schemes for all seven SDM and all pools of presences (Fig. 2). For each combination of parameters, 20 replicates with different presence data selections were performed to account for the variability in model accuracy because of the random sampling of presence data (Fig. 2). For each presence data sample, several replicates with different pseudo-absences selections were performed to further account for the variability because of the random

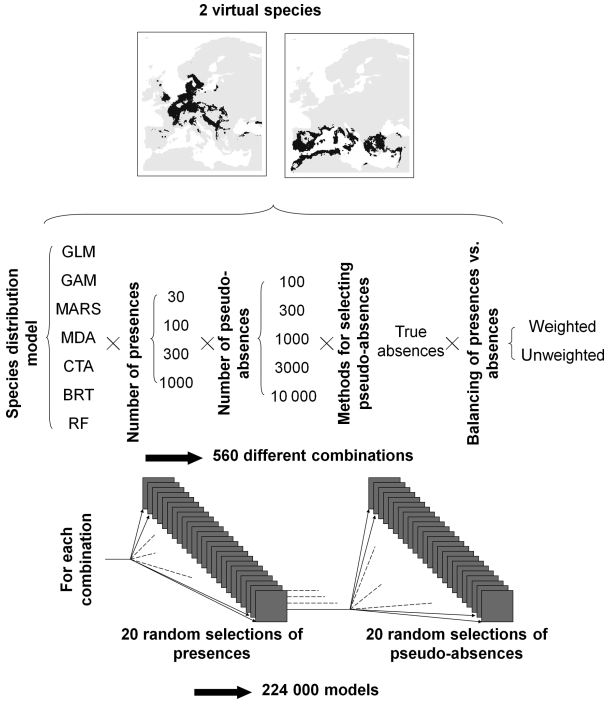

Fig. 1. General framework for data simulation and selection illustrating all factors tested to study the influence of prevalence.

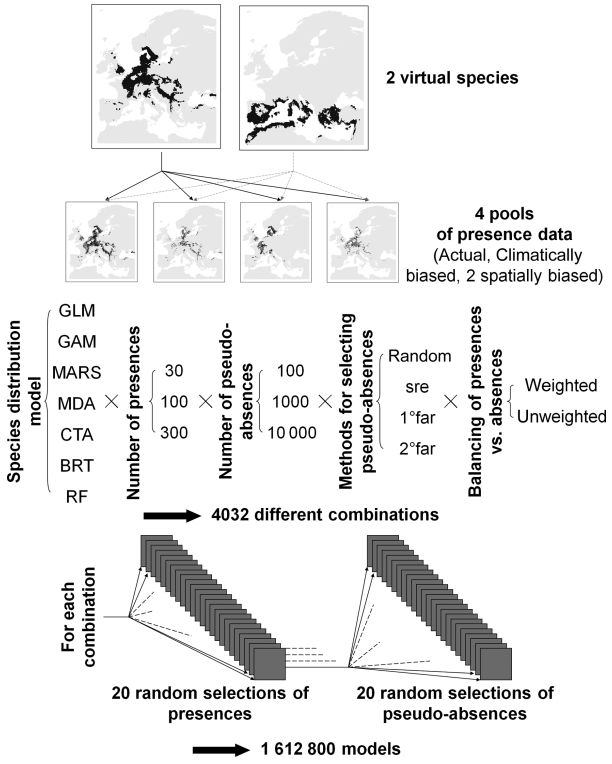

Fig. 2. General framework for data simulation and selection illustrating all factors tested to study the influence of pseudo-absence selection and biased presence data.

sampling of pseudo-absence data (Fig. 2).To investigate the optimal trade-off between the number of replicates, the number of pseudo-absences and the predictive accuracy, we calculated mean predicted distributions (hereafter called mean predictions) resulting from several (2–20) replicates of pseudo-absences selection. To estimate the number of replicates of pseudo-absences above which model quality does not increase significantly, we compared mean TSS across the number of replicates for each combination of pools of presence data × number of presences × number of pseudo-absences (Fig. 2).

(C) WHAT IS THE OPTIMAL NUMBER AND WEIGHTING SCHEME OF PSEUDO-ABSENCES PER REPLICATE?

We tested for an effect of the number/weighting scheme of pseudo-absences on model accuracy via a likelihood ratio test. This test compared the likelihood of two linear models: one that included as covariates both the method of generating pseudo-absences and the number/weighting scheme of pseudo-absences, and one that included only the former. The number/weighting scheme covariate was coded as a 6-level factor (100, 1000 or 10 000 pseudo-absences, with either equal or unequal weighting of presences vs. absences).

(D) WHICH METHOD OF GENERATING PSEUDO-ABSENCES RESULTS IN THE MOST ACCURATE MODELS?

For each number of presences considered, we tested how the method of generating pseudo-absences affected model accuracy. This was done via a likelihood ratio test that compared the likelihood of a linear model which included the method of generating pseudo-absences and

the number/weighting scheme of pseudo-absences as covariates with the likelihood of a model including only the latter.

(e) How do biases in the sampling design influence the optimal use of pseudo-absences?

Accuracy results from models run with spatially biased presences (countries or transportation biases) were aggregated because we did not detect any difference between them. Thus, two types of sampling bias were considered: climatically biased and spatially biased presence samples. For both sampling biases, tests similar to those described in (c) and (d) were computed.

(f) Which parameters have the greatest influence on the models' predictive accuracy?

For each SDM, we used an ANOVA to test the effects of the number of pseudo-absences, the method used for the selection of pseudo-absences, and the weighting scheme for presences vs. absences on model quality, for each combination of virtual species, pool of presence data and number of presences. In each case, the relative contribution of each effect was estimated as the ratio between the explained and the null deviances. Using the same approach, we also considered SDM as an additional effect to compare variability between SDM, that is, variations in model accuracy owing to differences in the way each SDM handles pseudo-absences.

Results

(a) Which ratio of presences/absences achieved the highest model accuracy?

The models could be separated into three groups according to the effect of prevalence on their predictive accuracy (Fig. 3). GAM behaved differently from the others given this technique was not influenced by prevalence. The accuracy of MARS and MDA increased with prevalence, whereas the accuracy increased until an asymptote when the number of presences reached one tenth of the number of absences for GLM, BRT and RF or reached the same amount as the number of absences for CTA. These trends were not influenced by the weighting scheme of presences vs. absences.

(b) What is the optimal number of replicate sets of pseudo-absences?

Model quality (i.e. TSS) increased with the number of replicates of pseudo-absences used to calculate the mean prediction until reaching an asymptote (Fig. 4). The number of replicates to reach the asymptote decreased significantly with the number of pseudo-absences selected per replicate. When 10 000 pseudo-absences (i.e. 20% of the study area) were used in each replicate, there was no effect of the number of replicates on model quality (i.e. no need for repetition). When 1000 pseudo-absences (i.e. 2% of the study area) were generated in each replicate, five replicates were enough to reach the asymptote with respect to model quality (TSS) for the GAM and CTA models, whereas the number of replicates did not affect model quality for the other five SDM (i.e. no need for repetition). When 100 pseudo-absences were generated in each replicate, model quality reached an asymptote at 12 replicates for the GAM model, seven replicates for GLM, MARS, MDA, CTA and RF, and four replicates for the BRT model. However, we noticed that with 100 pseudo-absences, the variability in TSS was substantial across the replicates, such that it was difficult to reliably identify an asymptote below 20 replicates (Fig. 4): even though accuracy was not significantly different between the mean prediction obtained with 15 replicates and the mean prediction obtained with 20 replicates, the former was lower than the latter. The use of the mean distribution obtained from 20 replicates of pseudo-absence selection for each selection of presences that was a priori chosen to reduce the variability resulting from pseudo-absence selection and answer all other questions was therefore conservative.

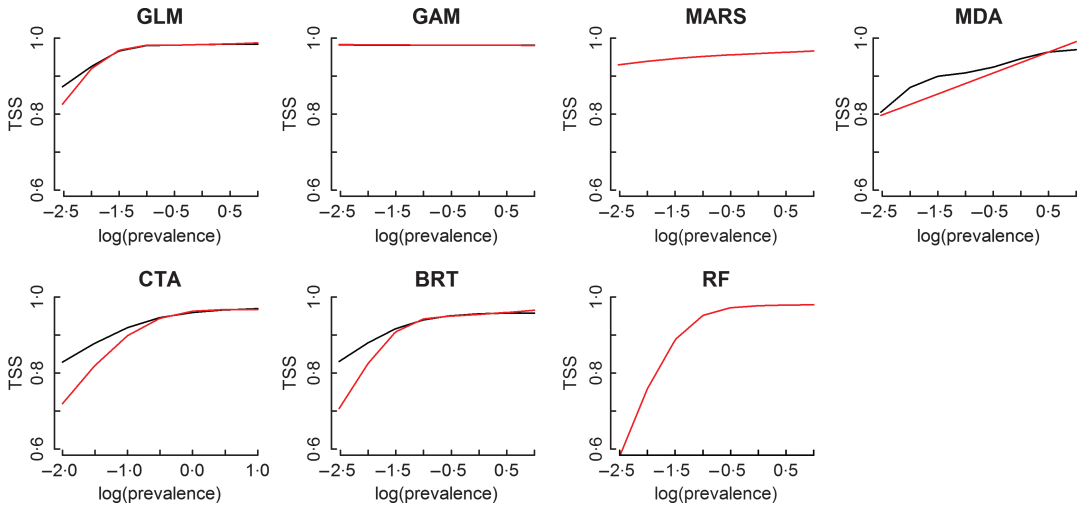

Fig. 3. Evaluation results (TSS) of the mean distribution according to the prevalence. The black and red curves represent results with a weighted and un-weighted scheme respectively.

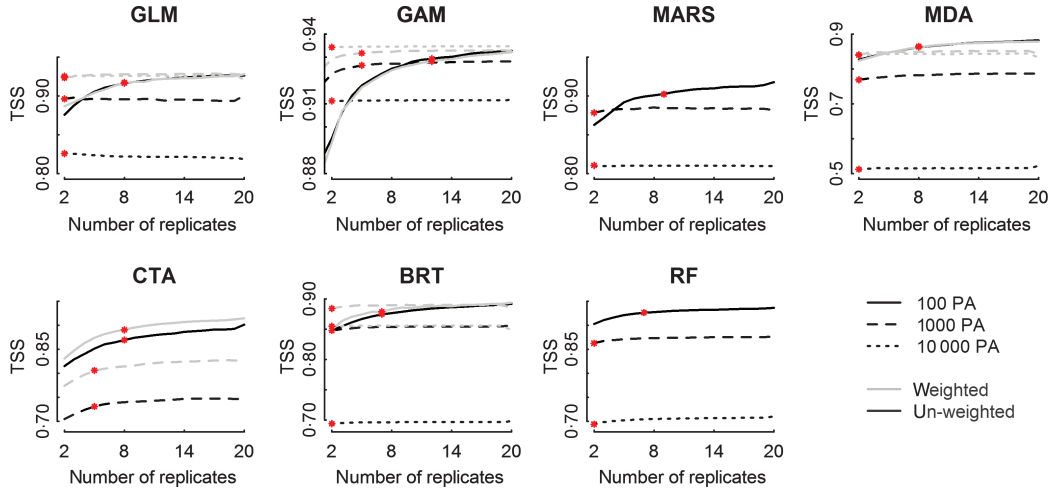

Fig. 4. Evaluation results (TSS) of the mean distribution according to the number of replicates with different pseudo-absences used to get that distribution. The different curves represent the results with 100, 1000, or 10 000 pseudo-absences selected in each replicate, as well as the weighting scheme. Red asterisks indicate that the TSS from the mean distribution with a larger number of replicates is not significantly better. These results were obtained with 100 climaticallybiased presences from the first virtual species (similar results were obtained with spatially biased presences and unbiased presences).

(c) What are the optimal number and weighting scheme of pseudo-absences per replicate?

Depending on the SDM used, the interaction between the number of pseudo-absences and weighting of presences vs. absences had different but significant effects on TSS. The models can be separated into three groups (Figs 5 and 6). Firstly, GLM and GAM showed little variation in predictive accuracy in response to the number of pseudo-absences, but the predictive accuracy increased when using pseudo-absences with equal weight for presences and absences. Secondly, for CTA, BRT and RF, predictive accuracy was highest when approximately the same number of pseudo-absences was used as the number of presences (Fig. 3). For CTA and BRT, when the number of pseudo-absences differed from the number of presences, an equal weight for presences and absences gave better model predictive quality. These results were mainly explained by the very low sensitivity of these two SDM when a large number of

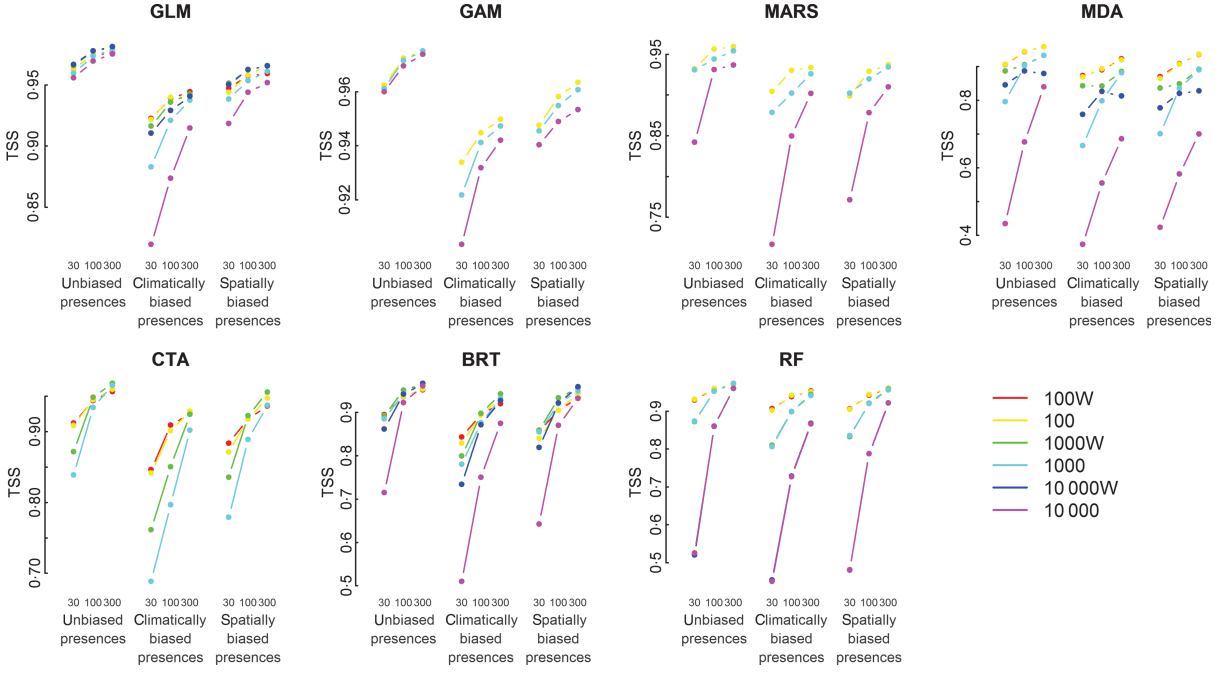

Fig. 5. Evaluation results (TSS) according to the modelling technique, the number of presences, to the quality of presences and the number and weighting scheme of pseudo-absences (mean over the method used to select pseudo-absences and the random selection of presences) (*W* stands for an equal weight of presences vs. absences).

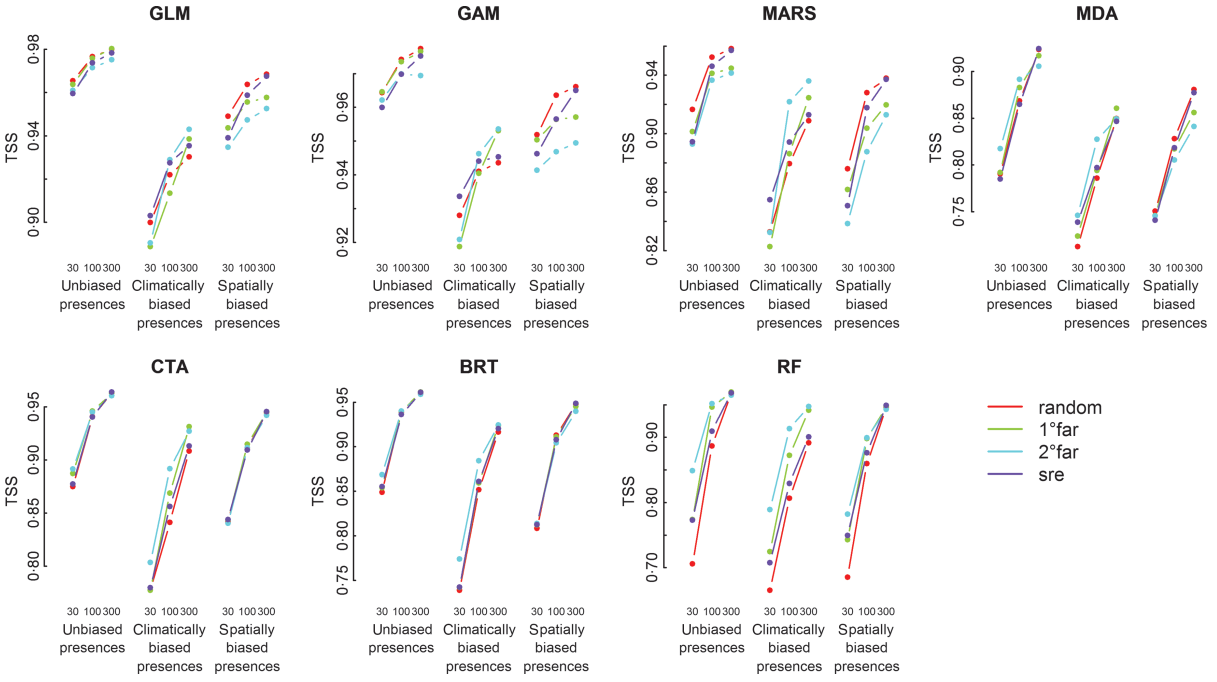

Fig. 6. Evaluation results (TSS) according to the modelling technique, the number of presences, to the quality of presences and the method used to select pseudo-absences (mean over the different numbers of pseudo-absences, the weighting scheme, and the random selection of presences).

pseudo-absences were generated (Fig. S3). Lastly, when using MARS and MDA, model quality was highest when 100 pseudo-absences were generated in each run, with equal weight given to presences and absences.

(d) Which method of generating pseudo-absences resulted in the most accurate models?

Model accuracy was affected by the method used to generate pseudo-absences for each SDM (Figs 5 and 6): likelihood ratio tests were significant in all cases excepted with spatially biased presences with CTA. For GLM, GAM and MARS, randomly selected pseudo-absences produced the most accurate models. For the other four SDM (MDA, BRT, CTA and RF), there was less variation in the results obtained for each different method used to select pseudo-absences, but pseudo-absences selected with geographical exclusion ('2°far') yielded significantly better models with few presences, whereas pseudo-absences selected with climatic exclusion ('SRE') yielded better models with more presences. Consistently across SDM and the number of presences, we found that pseudo-absences selected with geographical exclusion ('2°far' and '1°far') yielded predictions with higher sensitivities, whereas randomly selected pseudo-absences yielded predictions with higher specificities (Figs S3 and S4).

(e) How did potential biases in the presence sampling influence the optimal use of pseudo-absences?

The predictive accuracy of the models in relation to the number and weighting scheme of pseudo-absences was not influenced by the sampling biases of presence data (Fig. 5). Regarding the method used to generate pseudo-absences, the results obtained with spatially biased presences were similar to those obtained with unbiased presences (Fig. 6), except for MDA for which 'random' yielded better models with spatially biased presences. With the three regression techniques (GLM, GAM and MARS), 'random' did not perform well with climatically biased presences, but 'SRE' yielded better results when few presences were available from the actual distribution and '2°far' yielded better results when more presences were selected. For the other four SDM (MDA, CTA, BRT and RF), '2°far' performed better when presences were climatically biased (Fig. 6).

(f) Which parameters have the greatest influence on the models' predictive accuracy?

The relative contribution of each methodological choice to variations in model quality depended on the SDM used. GLM and GAM methods responded similarly: when 30 presences were selected, variation in TSS among distributions obtained from all models was only partly explained by the number of pseudo-absences, the method used for selecting pseudo-absences, and the weighting of presences vs. absences (Fig. 7). This pattern suggested that results were most influenced by the random set of presences from the actual species distribution. However, when the number of sampled presences increased, the contribution of the other factors to variability in TSS also increased: with 100 or 300 presences, the method used for selecting the pseudo-absences explained most of the variation in TSS for GLM and GAM. In contrast, for the five remaining SDM, the number of pseudo-absences selected for each run made the biggest

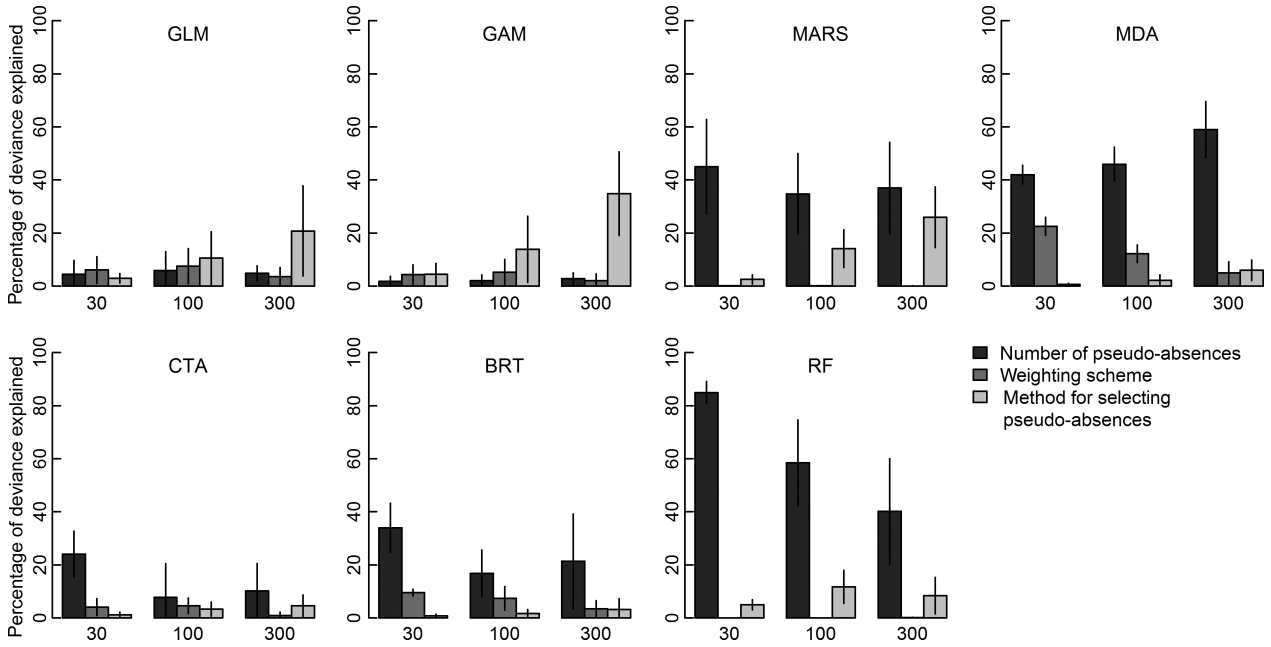

Fig. 7. Percentage of deviance explained by the three factors regarding the use of pseudoabsences according to the number of presences used for modelling [average across the two virtual species and the four pools of presence data (unbiased or biased presences)].

contribution to the variability in TSS regardless of the number of presences sampled. The method used for selecting pseudo-absences also partly explained the variation in TSS and its influence increased with the number of presences sampled.

Overall, the variability arising from each methodological choice regarding the use of pseudo-absences was lower than the variability arising from the use of different SDMs, especially when at least 100 presence data were sampled (Fig. 8).

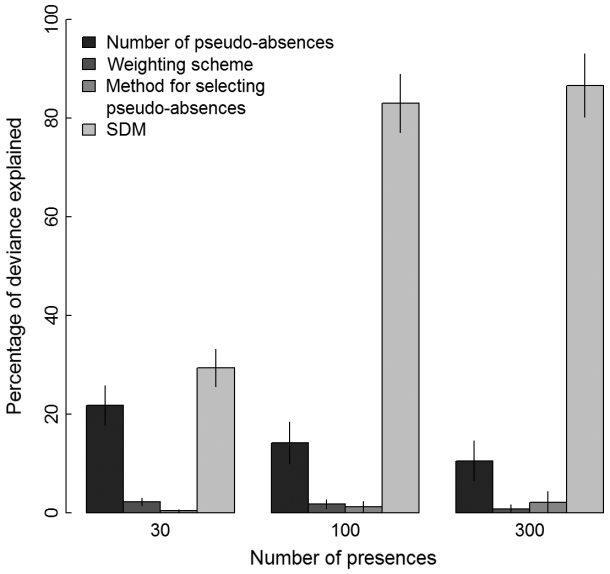

Fig. 8. Percentage of deviance explained by the three factors regarding the use of pseudoabsences and the modelling techniques according to the number of presences used for modelling [average across the two virtual species and the four pools of presence data (unbiased or biased presences)].

In addition, we found that AUC and TSS were highly correlated (using Pearson's product-moment correlation, $r = 0{\cdot}82 \pm 0{\cdot}10$ across all SDM). Therefore, the relative performance of the different methods used to select the pseudo-absences did not depend on the choice of the evaluation criterion. Although we presented results on the models' predictive accuracy, the results and conclusions were the same for the models' ability to correctly predict climatic suitability (assessed using a correlation test between the probability distribution obtained from the model and the probabilities of occurrence of the potential distribution chosen for a given species, Fig. S5).

Discussion

INFLUENCE OF THE MODELLING TECHNIQUE

In general, our results showed that the behaviour of the different SDM varied widely depending on how, where and how many pseudo-absences were used. First of all, although the model accuracy of regression techniques GLM and GAM was not influenced as much as other SDM by the number of pseudo-absences used in each replicate, the best results were obtained by using a large number of pseudo-absences (e.g. 10 000) with presences and absences weighted equally. These results are consistent with those obtained with Maxent (Phillips & Dudik 2008) for which more accurate results were also obtained with 10 000 background points. Conversely, for classification and machine-learning techniques including MARS, the models' predictive accuracy was greater when a moderate number of pseudo-absences per replicate were used (either few pseudo-absences or not more than the number of presences). For these models, the choice of the number of pseudo-absences

used in each replicate was the main influence on model accuracy, making it a key decision when setting up a modelling exercise. This difference in terms of the optimal number of pseudo-absences to use in each replicate for different SDM could not be solely attributed to the poor performance of classification and machine-learning techniques when the number of false absences increases (which is automatically the case when the number of pseudo-absences increases), because the study regarding the influence of prevalence over model accuracy, performed with true absences only, lead to the same conclusions. This difference could therefore be attributed to the intrinsic properties of the different SDM with regard to prevalence.

The different SDM investigated in this study also appeared to behave differently with regard to the method used to generate pseudo-absences. Indeed, regression techniques were more greatly influenced by the choice of the method than classification and machine-learning techniques, and different methods were found to optimise model accuracy. When using regression techniques (GLM, GAM and MARS), the best strategy was to randomly generate the pseudo-absences data, which supported results from Wisz & Guisan (2009). Indeed, their study using simulated data showed that randomly selected pseudo-absences yielded better results than pseudo-absences selected from low suitability areas predicted using ENFA or BIOCLIM (equivalent to SRE). For classification and machine-learning techniques, although the method used to generate pseudo-absences had little influence on the models' predictive accuracy, '2°far' yielded significantly better models with few presences, whereas 'SRE' yielded better models with more presences. We can assume the difference in the best method for generating pseudo-absences according to the number of available presences to be the consequence of different false negative rates. Indeed, with few available presences, it is very unlikely that these presences represent the full climatic niche of the species. Therefore, pseudo-absences selected with environmental exclusion ('SRE') may have a higher chance of being false absences than pseudo-absences selected with large geographical exclusion ('2°far'). However, as the amount of available presences increases, the probability of pseudo-absences selected with environmental exclusion being false absences decreases. With large amounts of presence data, although pseudo-absences selected with large geographical exclusion still have a better chance of being true absences, they are probably too different from the presence data to be as informative as the pseudo-absences selected with environmental exclusion. This may also depend in part on the level of spatial aggregation in species presences. Such differences regarding the best method of generating pseudo-absences indicate that regression techniques were less sensitive to false absences than classification and machine-learning techniques.

Finally, the optimal number of pseudo-absence replicates also differed between the different SDM. Some of these differences could be explained by the intrinsic properties of the SDM. For example, BRT and RF were the SDM that needed the lowest number of 100 pseudo-absences replicates, perhaps because both have internal replication procedures.

ENSEMBLE FORECAST PERSPECTIVES

As modelling a species distribution under current and future conditions can give different results according to the SDM used (Thuiller 2004; Elith *et al.* 2006) and as none of the widely used techniques performs universally better than the others (Elith *et al.* 2006), the use of an ensemble forecast framework has been recommended (Buisson *et al.* 2010). The ensemble forecast framework aims to consider the central trend of several SDM, using different methods (Marmion *et al.* 2009), and is now widely used amongst species distribution modellers, often with the same use of pseudo-absences across the different SDM used. However, we have shown here that the optimal way of creating and using pseudo-absences information differs widely across SDM. The best way of using pseudo-absences through an ensemble forecast technique could therefore be to use pseudo-absences differently for each SDM. However, most ensemble forecast techniques compare model accuracy either to select the best models or to weight their predictions differently, which can only be done in an unbiased way if the same data were used for all SDM. One way of overcoming this potential problem could be to group together SDM that share the same way of optimising the use of pseudo-absences (e.g. GLM and GAM; BRT and RF), compare their model accuracy, select the best one from each group and then obtain the median or mean distribution from all selected models.

SPATIAL EXTENT OF THE STUDY AREA

As well as being influenced by the number of pseudo-absences and the method used to generate them, model performance also relies on the spatial extent of the study. Indeed, model performance is lower when pseudo-absences are taken from either a restricted or particularly broad area (Van Der Wal *et al.* 2009). Pseudo-absences are meant to be compared with the presence data and help differentiate the environmental conditions under which a species can occur or not. Therefore, pseudo-absences taken too far from the presence data in the environmental space would not be very informative. As pseudo-absences that are very distant from all presence points (from a geographical point of view) are more likely to exhibit environmental conditions that are very different from those for the presence data, a larger spatial extent of the study will lead to the selection of a higher proportion of less informative pseudo-absences. The optimal number of pseudo-absences to generate in each run is therefore likely to depend on the spatial extent of the study, which influences environmental variability. At a given spatial resolution, a higher number of pseudo-absences may be needed to optimise model performance for a larger spatial extent of the study, to ensure the selection of enough informative pseudo-absences.

MAXIMISING SENSITIVITY OR SPECIFICITY

When the modelling goal is to identify potential presences of rare species for new survey efforts (Engler, Guisan &

Table 1. How to choose pseudo-absences according to the modelling technique from results of this study, the bold criteria being the most important for the considered modelling technique. 'Same as number of presences, at least 10 runs when less than 1000 PA' means that when more than 300 presence points were considered, 1000 PA should be selected and when 100 or less presences points were considered, a minimum of 10 runs with 100 PA gave the best results

	Method for selecting pseudo-absences	Number of pseudo-absences
GLM, GAM	**'random' performs consistently well, excepted when presences are climatically biased for which '2°far' is the best method**	10 000 PA or a minimum of 10 runs with 1000 PA with an equal weight for presences and absences
MARS	'random' performs consistently well, except when presences are climatically biased for which '2°far' is the best method	**A minimum of 10 runs with 100 PA**
MDA	'2°far' performs consistently better with few presences, 'SRE' performs better with a large number of presences; 'random' performs consistently well with spatially biased presences	**A minimum of 10 runs with 100 PA with an equal weight for presences and absences**
CTA, BRT, RF	'2°far' performs consistently better with few presences, 'SRE' performs better with a large number of presences	**Same as number of presences, 10 runs when less than 1000 PA with an equal weight for presences and absences**

Rechsteiner 2004), high sensitivity is preferred, even if it generates overprediction. High sensitivity ensures that the percentage of true presences predicted as absences will be minimised. In such studies, the 'SRE', '1 and 2°far' methods can be used as well as other methods for selecting pseudo-absences outside both spatially and climatically suitable areas (Hengl *et al*. 2009; Lobo, Jimenez-Valverde & Hortal 2010). The selection of fewer pseudo-absences in each replicate also yielded better sensitivity (except for GLM and GAM, for which large amounts of pseudo-absences with an equal weighting of presences vs. absences still yielded better sensitivity). In contrast, other studies may wish to maximise specificity, so that the predicted distribution of a species would only be the area where the species is highly likely to be present. This is particularly true for studies on reserve planning (Marini *et al*. 2009). High specificity ensures that the percentage of true absences predicted as presences will be minimised. In such cases, the random selection of pseudo-absences will maximise specificity. As for the number of pseudo-absences to generate in each replicate to maximise specificity, it depends on the number of presence points available, but overall a large number of pseudo-absences tends to yield better specificity for all SDM except GLM and GAM for which fewer pseudo-absences are better. All these results regarding sensitivity and specificity are dependent on the threshold used to produce binary distributions. The use of another commonly used threshold (minimising the difference between sensitivity and specificity) could yield slightly different results as it tends to favour specificity, whereas the threshold we used tends to favour sensitivity (Jimenez-Valverde & Lobo 2007).

Conclusion

Overall, we recommend the use of a large number (e.g. 10 000) of pseudo-absences with equal weighting for presences and absences when using GLM and GAM, averaging several runs with relatively fewer pseudo-absences (e.g. 100) with equal weighting for presences and absences with MARS and MDA, and using the same amount of pseudo-absences as the amount of available presences (averaging several runs if few pseudo-absences) for CTA, BRT and RF (Table 1). In addition, we recommend the random selection of pseudo-absences with regression techniques and the random selection of pseudo-absences with geographical and environmental exclusion with classification and machine-learning techniques. These recommendations further apply when using data likely to be biased (e.g. GBIF data). For all SDM, we recommend the random selection of pseudo-absences when high specificity is valued over high sensitivity (e.g. reserve planning). Nevertheless, in studies seeking to identify unsurveyed sites with a high probability of occurrence for rare species, pseudo-absences that are more likely to be true absences (outside the suitable area of the species and not too close to a presence point) are recommended.

Acknowledgements

We are grateful to Margaret K. Evans and Jean-Pierre Moussus for helpful corrections of the manuscript and to Romain Lorrillière whose computer was of great help for running part of the simulations. WT and CHA received support from the ANR DIVERSITALP (ANR-07-BDIV-014) and European Commission funded FP6 ECOCHANGE (GOCE-CT-2007-036866) projects. Most of the computations were performed using the CIMENT infrastructure (https://ciment.ujf-grenoble.fr) supported by the Rhône-Alpes region (GRANT CPER07-13 CIRA). Thanks also to Version Originale for checking and correcting the English in this article. Finally, the authors thank seven anonymous referees and Jana McPherson for insightful comments on earlier versions of the manuscript.

References

Albert, C.H., Graham, C.H., Yoccoz, N.G., Zimmermann, N.E. & Thuiller, W. (2010) Applied sampling in ecology and evolution – integrating questions and designs. *Ecography*, **33**, 1028–1037.

Allouche, O., Tsoar, A. & Kadmon, R. (2006) Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, **43**, 1223–1232.

Barbet-Massin, M., Thuiller, W. & Jiguet, F. (2010) How much do we overestimate future local extinction rates when restricting the range of occurrence data in climate suitability models? *Ecography*, **33**, 878–886.
Brotons, L., Thuiller, W., Araújo, M.B. & Hirzel, A.H. (2004) Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. *Ecography*, **27**, 437–448.
Buisson, L., Thuiller, W., Casajus, N., Lek, S. & Grenouillet, G. (2010) Uncertainty in ensemble forecasting of species distribution. *Global Change Biology*, **16**, 1145–1157.
Busby, J. R. (1991) BIOCLIM – a bioclimate analysis and prediction system. *Plant Protection Quarterly*, **6**, 8–9.
Elith, J., Graham, C.H., Anderson, R.P., Dudik, M., Ferrier, S., Guisan, A. *et al*. (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**, 129–151.
Engler, R., Guisan, A. & Rechsteiner, L. (2004) An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *Journal of Applied Ecology*, **41**, 263–274.
Farber, O. & Kadmon, R. (2003) Assessment of alternative approaches for bioclimatic modeling with special emphasis on the Mahalanobis distance. *Ecological Modelling*, **160**, 115–130.
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.
Godsoe, W. (2010) I can't define the niche but I know it when I see it: a formal link between statistical theory and the ecological niche. *Oikos*, **119**, 53–60.
Graham, C.H., Ron, S.R., Santos, J.C., Schneider, C.J. & Moritz, C. (2004a) Integrating phylogenetics and environmental niche models to explore speciation mechanisms in dendrobatid frogs. *Evolution*, **58**, 1781–1793.
Graham, C.H., Ferrier, S., Huettman, F., Moritz, C. & Peterson, A.T. (2004b) New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology & Evolution*, **19**, 497–503.
Gu, W.D. & Swihart, R.K. (2004) Absent or undetected? Effects of non-detection of species occurrence on wildlife-habitat models *Biological Conservation*, **116**, 195–203.
Guisan, A. & Thuiller, W. (2005) Predicting species distribution: offering more than simple habitat models. *Ecology Letters*, **8**, 993–1009.
Hengl, T., Sierdsema, H., Radovic, A. & Dilo, A. (2009) Spatial prediction of species' distributions from occurrence-only records: combining point pattern analysis, ENFA and regression-kriging. *Ecological Modelling*, **220**, 3499–3511.
Huntley, B., Collingham, Y.C., Willis, S.G. & Green, R.E. (2008) Potential impacts of climatic change on European breeding birds. *PLoS ONE*, **3**, e1469.
Jimenez-Valverde, A. & Lobo, J.M. (2007) Threshold criteria for conversion of probability of species presence to either-or presence-absence. *Acta Oecologica-International Journal of Ecology*, **31**, 361–369.
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.
Lawton, J. H. (1999) Are there general laws in ecology? *Oikos*, **84**, 177–192.
Lobo, J.M., Jimenez-Valverde, A. & Hortal, J. (2010) The uncertain nature of absences and their importance in species distribution modelling. *Ecography*, **33**, 103–114.
Mackenzie, D.I. & Royle, J.A. (2005) Designing occupancy studies: general advice and allocating survey effort. *Journal of Applied Ecology*, **42**, 1105–1114.
Marini, M.A., Barbet-Massin, M., Lopes, L.E. & Jiguet, F. (2009) Major current and future gaps of Brazilian reserves to protect Neotropical savanna birds. *Biological Conservation*, **142**, 3039–3050.
Marmion, M., Parviainen, M., Luoto, M., Heikkinen, R.K. & Thuiller, W. (2009) Evaluation of consensus methods in predictive species distribution modelling. *Diversity and Distributions*, **15**, 59–69.
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.
Peterson, A. T. & Vieglais, D. A. (2001) Predicting species invasions using ecological niche modeling: new approaches from bioinformatics attack a pressing problem. *BioScience*, **51**, 363–371.
Phillips, S.J. & Dudik, M. (2008) Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography*, **31**, 161–175.
Phillips, S.J., Dudik, 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.
Soberon, J. & Nakamura, M. (2009) Niches and distributional areas: concepts, methods, and assumptions. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 19644–19650.
Thuiller, W. (2004) Patterns and uncertainties of species' range shifts under climate change. *Global Change Biology*, **10**, 2020–2027.
Thuiller, W., Lavorel, S., Araújo, M.B., Sykes, M.T. & Prentice, I.C. (2005) Climate change threats to plant diversity in Europe. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 8245–8250.
Thuiller, W., Lafourcade, B., Engler, R. & Araújo, M.B. (2009) BIOMOD – a platform for ensemble forecasting of species distributions. *Ecography*, **32**, 369–373.
Van Der Wal, J., Shoo, L.P., Graham, C. & William, S.E. (2009) Selecting pseudo-absence data for presence-only distribution modeling: how far should you stray from what you know? *Ecological Modelling*, **220**, 589–594.
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.
Zaniewski, A.E., Lehmann, A. & Overton, J.M.C. (2002) Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecological Modelling*, **157**, 261–280.
Zurell, D., Berger, U., Cabral, J.S., Jeltsch, F., Meynard, C.N., Munkemuller, T. *et al*. (2010) The virtual ecologist approach: simulating data and observers. *Oikos*, **119**, 622–635.

Received 12 July 2010; accepted 11 November 2011
Handling Editor: Jana McPherson

Supporting Information

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

Fig. S1. (a) Response curves considered for the two virtual species along both composite environmental variables (species 1 in black and species 2 in grey) (species 1: mean = −0.5 and SD = 0.3 for both axes; species 2: mean = 0.8 on axis 1 and mean = −1 on axis 2, SD = 0.5 for both axes), (b) potential (light blue and purple) and actual (dark blue and purple) niches of both virtual species in the climatic space, (c) pools of presence data for species 1, (d) pools of presence data for species 2.

Fig. S2. Response curves considered for the climatically biased presences (grey) for species 1 (a) and 2 (b), compared to the response curves of its fundamental distribution (black).

Fig. S3. (a) Evaluation results (sensitivity) according to the modelling technique, the number of presences, to the quality of presences and the number and balancing of pseudo-absences (mean over the method used to select pseudo-absences and the random selection of presences) (*W* stands for an equal weight of presences vs. absences). (b) Evaluation results (sensitivity) according to the modelling technique, the number of presences, to the quality of presences and the method used to select pseudo-absences (mean over the different numbers of pseudo-absences, the balancing of presences vs. absences, and the random selection of presences).

Fig. S4. (a) Evaluation results (specificity) according to the modelling technique, the number of presences, to the quality of presences and the number and balancing of pseudo-absences (mean over the method used to select pseudo-absences and the random selection of presences) (*W* stands for an equal weight of presences vs. absences). (b) Evaluation results (specificity) according to the modelling technique, the number of presences, to the quality of presences and the method used to select pseudo-absences (mean over the different numbers of pseudo-absences, the balancing of presences vs. absences, and the random selection of presences).

Table S5. Pearson's product-moment correlation between the TSS and the correlation coefficient between the probability distribution obtained from modeling and the climatic suitability calculated to create the virtual species for all models.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising fromsupporting information (other than missing files) should be addressed to the authors.