*Journal of Vegetation Science 21: 908–922, 2010*
*DOI: 10.1111/j.1654-1103.2010.01198.x*


# Effects of the number of presences on reliability and stability of MARS species distribution models: the importance of regional niche variation and ecological heterogeneity

**Rubén G. Mateo; Ángel M. Felicísimo & Jesús Muñoz**

## Abstract

**Question:** What are the effects of the number of presences on models generated with multivariate adaptive regression splines (MARS)? Do these effects vary with data quality and quantity and species ecology?

**Location:** Spain and Ecuador.

**Methods:** We used two data sets: (1) two trees from Spain, representing high-occurrence number data sets with real absences and unbalanced prevalence; (2) two herbs from Ecuador, representing low-occurrence number data sets without real absences and balanced prevalence. For model quality, we used two different measures: reliability and stability. For each sample size, different replicates were generated at random and then used to generate a consensus model.

**Results:** Model reliability and stability decrease with sample size. Optimal minimum sample size varies depending on many factors, many of which are unknown. Regional niche variation and ecological heterogeneity are critical.

**Conclusions:** (1) Model predictive power improves greatly with more than 18-20 presences. (2) Model reliability depends on data quantity and quality as well as species ecological characteristics. (3) Depending on the number of presences in the data set, investigators must carefully distinguish between models that should be treated with skepticism and those whose predictions can be applied with reasonable confidence. (4) For species combining few initial presences and wide environmental range variation, it is advisable to generate several replicate models that partition the initial data and generate a consensus model. (5) Models of species with a narrow environmental range variation can be highly stable and reliable, even when generated with few presences.

**Keywords:** AUC; Consensus model; MARS; Pearson correlation coefficient; Regional niche variation; Sample size.

**Abbreviations:** SDM = Species distribution model; MARS = Multivariate adaptive regression splines; NHC = Natural history collections.

## Introduction

Predictive modelling is a powerful tool in many fields where direct observation or experimentation is not easy. Species distribution models (SDM) have increased in importance in recent years (Guisan & Zimmermann 2000; Araújo & Guisan 2006). These models have wide relevance in conservation biology (Araújo et al. 2005b; Rissler et al. 2006), biogeography (Lobo et al. 2001; Luoto et al. 2006; Richards et al. 2007), reserves design (Araújo & Williams 2000; Margules & Pressey 2000; Ortega-Huerta & Peterson 2004) and climate change (Iverson 2004; Araújo et al. 2006; Botkin et al. 2007; Pearman et al. 2008). Depending on the target species to be modelled, the number of presences used in published papers ranges from a few – occasionally just one in rare organism modelling (Pearson et al. 2007) or risk assessment (Beguería 2006) – to (very) a high number (Zaniewski et al. 2002).

The total amount of information available in natural history collections (NHC) is enormous, but very few taxa occurrences can be counted in the hundreds (Loiselle et al. 2008). In fact, most taxa have rarely been collected and their full distribution areas are highly hypothetical, which makes them prime targets for conservation biology and ecological modelling (Graham et al. 2004).

Some review papers have paved the way with regard to modelling methods, model parameterization or selection, and comparative accuracy of the available methods (e.g. Guisan & Zimmermann

**Mateo, R.G.** (corresponding author, Ruben.GMateo@uclm.es) & **Muñoz, J.** (jmunoz@rjb.csic.es): Real Jardín Botánico (CSIC), Plaza de Murillo 2, Madrid, Spain.
**Felicísimo, A.M.** (amfeli@unex.es): Escuela Politécnica (Universidad de Extremadura), C/Santa Teresa de Jornet s/n, Mérida, Spain.
**Mateo, R.G.**: Universidad de Castilla-La Mancha, Av/Carlos III s/n, Toledo, Spain.

2000; Guisan & Thuiller 2005; Araújo & Guisan 2006; Elith et al. 2006); others have dealt with the fact that for most organisms there are data only on presence but not on absence (e.g. Zaniewski et al. 2002; Brotons et al. 2004; Pearce & Boyce 2006; Mateo et al. 2010); and others have also studied why widespread species are usually harder to model than narrowly distributed organisms (Manel et al. 2001; Luoto et al. 2005; Elith et al. 2006; McPherson & Jetz 2007). However, there are issues remaining on the modelling processes that have mostly been obviated in many works, such as the influence of sample size (total number of data points, presence plus absence) and prevalence (proportion of presences in the data set) in the generated models.

Although there is a consensus that sample size affects the modelling outcome, most authors have ignored this issue, and a number of SDMs have even been generated using an extremely low sample size without considering the potential consequences (model accuracy and reliability). For example, the two presences used by Ortega-Huerta & Peterson (2004) and by McClean et al. (2005), four by Loiselle et al. (2003) and by Cuesta-Camacho et al. (2006), or the seven to 12 presences used by Anderson & Martinez-Meyer (2004). Others have merely mentioned the potential drawbacks of small data sets (Stockwell & Peters 1999; Reese et al. 2005), while some, whose research aims were different, expressed concern about minimum sample size (Cumming 2000b; Pearce & Ferrier 2000; Drake et al. 2006; Guisan et al. 2007). Several papers have dealt in depth with these topics, e.g. the study of sample size effects on results given by three different methods, logistic regression, GARP and Bioclim (Stockwell & Peterson 2002); the exploration of prevalence effects on SDM accuracy (McPherson et al. 2004); the study of Hernandez et al. (2006) of sample size effects on rare species modelled using four common methods (Bioclim, Domain, GARP and Maxent); or the comparison between GARP and Maxent (Papeş & Gaubert 2007; Pearson et al. 2007); and finally comparison of 12 different methods realized by Wisz et al. (2008). Some of these works documented a minimum sample size required to generate reliable SDMs, although the results, even for the same method, are very variable: five using Maxent (Hernandez et al. 2006; Pearson et al. 2007), ten using GARP (Stockwell & Peterson 2002), 15 using GARP and Maxent (Papeş & Gaubert 2007), 20 using logistic multiple regression (Stockwell & Peterson 2002), 40 using support vector machines (Drake et al. 2006), more than 30 using 12 different methods included GARP and Maxent (Wisz et al. 2008), between 50 and 75 using Bioclim (Kadmon et al. 2003), and 300 using logistic multiple regression (Cumming 2000a).

According to Stockwell & Peterson (2002), predictive power of models generally improves with additional information; although "plateaus" commonly exist and then any additional data adds little to model performance. Moreover, an increase in the number of observations may even reduce model accuracy due to over-fitting (Verbyla 1986; Verbyla & Litvaitis 1989). Therefore, sample size has a potential influence on model accuracy, reliability and stability (Hernandez et al. 2006), which combined with sometimes extreme differences in outcomes predicted by different methods (Loiselle et al. 2003; Elith et al. 2006), becomes an interesting avenue of research. In theory, model stability (the extent to which a model yields the same results on repeated trials), accuracy (quality and predictive ability of a model) and reliability (capacity of the model to be credible, not spurious) should decrease as sample size decreases (Stockwell & Peterson 2002; McPherson et al. 2004; Hernandez et al. 2006; Wisz et al. 2008), and therefore a minimum number of presences is needed to generate a robust model. On the other hand, to limit sampling effort to a minimum size would allow generation of accurate SMDs without wasting valuable resources, as data on species distribution can be extremely difficult or expensive to obtain, particularly in tropical areas (Raven & Wilson 1992; Cayuela et al. 2009). Moreover, several studies have shown that beyond a threshold – dependent on the organism being modelled – the predictive accuracy of models may remain constant (Pearce & Ferrier 2000; Hjort & Marmion 2008).

Following the above line of research, this paper is focused on how the number of presences affects SDM reliability and stability on multivariate adaptive regression spline (MARS) models, a method not previously tested on these grounds. Here, we are not interested in validation of the accuracy of SDMs. Most published studies refer to model accuracy, but here we use reliability and stability to measure model performance for several reasons: (1) data on NHCs are usually (very) scarce and do not allow splitting the data set into training and testing portions without losing precious information, and also cannot be considered as fully independent data sets (Araújo et al. 2005a; McPherson & Jetz 2007); (2) they lack data on absences, which precludes most orthodox validation techniques; (3) the AUC (area under the ROC curve) value is the only measure of SDM accuracy that is prevalence- and

threshold-independent and therefore appropriate in this exercise, but requires amounts of data not always available, as in this study; (4) model verification offers no information in this study, as sample size is very small (Mateo 2008); (5) models idealize prediction of ecological niche, but their accuracy can only be measured using data on actual distribution ranges, which are also shaped by biological interactions not considered when generating the potential environmental model (Kadmon et al. 2003); (6) performance values (AUC, kappa, etc.) can be highly dependent on the random split of the original data set (Phillips et al. 2006; Raes & Steege 2007); (7) accuracy assessment of presence-only SDMs cannot alone be sufficient without testing them against a null model (Anderson et al. 2002; Raes & Steege 2007); and finally, (8) model reliability must be used in similar studies with satisfactory results (Hernandez et al. 2006).

Model reliability depends on numerous issues, as well as methodological aspects, such as the method used (Thuiller 2003; Segurado & Araújo 2004) and diverse aspects of data accessible for model training (Kadmon et al. 2003; McPherson et al. 2004). According to Kadmon et al. (2003), reliability mainly depends on (a) properties of the data and (b) properties of the organism being modelled.

Regarding *properties of the data*, the present study focuses on the number of available presences, and we therefore replicate the analyses by varying the number of presences. We also deal with data quality, in the sense that the *Anthurium* data do not cover the full distribution area of the species, while the Spanish tree data set does represent the real distribution of both species in that area.

Regarding *properties of the species*, ecological characteristics are potentially important, since they can affect the reliability of SDMs. We therefore select, within each of the two data sets, species with different ecological requirements (see species data). For widely distributed species, ecological conditions may vary significantly between diverse areas inside the range (Murphy & Lovett-Doust 2007) and so it is reasonable to suppose that the most accurate predictions would be achieved for narrowly distributed species, characterized by well-defined niches (Kadmon et al. 2003; Papeş & Gaubert 2007; Pearson et al. 2007).

In short, in this work, we test the influence of quality and quantity of input data on the output models using the same method and independent variables. To do so, we use two different data sets that represent the two principal options of data available for ecological modelling: (1) natural history collections without real absences and (2) massive data sets with real presences and absences. These data sets were prepared with the aim of covering, as much as possible, the different quality and quantity data sets that are usually available: (1) presence-only data and randomly generated pseudo-absences versus real presence–absence data; (2) balanced versus unbalanced prevalence; and (3) low versus high number of presences. The main aims were to evaluate: (1) if models generated with incomplete information from a region are a fair representation of the model generated with all available data from the same region; (2) the influence of quality and quantity of input data on model reliability and stability; (3) if low-number data in NHCs allow generation of stable and robust MARS models; (4) the likely combined influence of regional niche variability and sample size on reliability and stability of the models; and (5) if a consensus model can recover information from individual models to become more stable and robust.

## Methods

### *Species data*

As dependent variables, we selected four plant species, grouped into two types of data set that differ sufficiently to cover as much as possible the difficulties faced by modellers, and also to represent very different biogeographic and climatic areas (tropical, Eurosiberian and Mediterranean) and highly unequal sampling efforts. The four species analysed present a relatively wide ecological niche and a widespread distribution. For all variables, both target and environmental, pixel size was 0.00833 degrees.

#### ***Large data sets with real absences and unbalanced prevalence***

We used presence/absence of European beech (*Fagus sylvatica* L.) and Pyrenean oak (*Quercus pyrenaica* Wild.) in Spain (Fig. 1). The former species grows in the mountains of northern Spain and in a disjunct locality in the Central Range. Its distribution covers Central and Western Europe, from the Iberian Peninsula to Poland and from Scandinavia to Sicily. The latter species grows in central and northern Spain; its distribution spans western and southwestern France, northern Morocco and the Iberian Peninsula.

Both taxa show regional niche differences. In the core of its European distribution, including

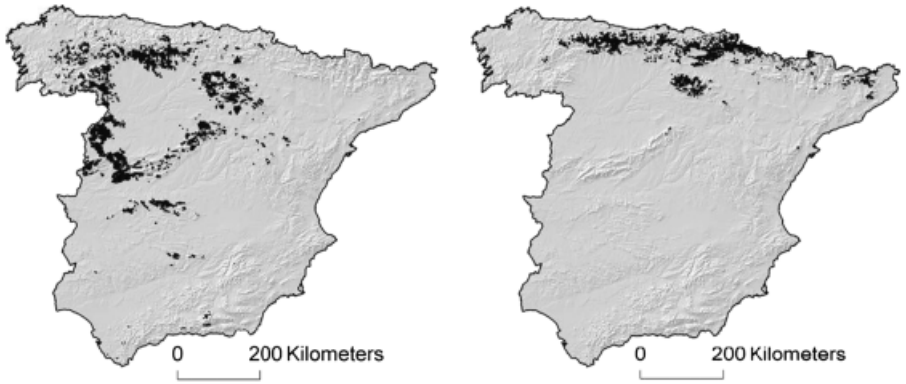


**Fig. 1.** Original data set (presence data) for *Fagus sylvatica* (right) and *Quercus pyrenaica* (left) (Ceballos 1966). WGS84 projection.

northernmost Spain, *Fagus sylvatica* grows in a wide spectrum of ecological conditions, both edaphic and climatic. However, at its southern Iberian Peninsula limit, it grows exclusively in mountain areas on siliceous substrates (Costa Tenorio et al. 1998). In contrast, *Quercus pyrenaica* can be considered a submediterranean species, growing in more humid and less hot areas in the central Iberian Peninsula than other strictly Mediterranean members of *Quercus* (e.g. *Q. rotundifolia*). Its presence in cool and very humid Eurosiberian areas next to the sea in northern Spain is considered to be a consequence of interglacial occupation of suitable areas, followed by extinction in less sheltered areas (Costa Tenorio et al. 1998).

Table 1 presents details of stratified sampling done on a digital version of the Forest Map of Spain (Ceballos 1966). For the two study species, we sampled this digital map to generate 20 replicates per sample size (15%, 5%, 1%, 0.1% and 0.05% of the total cover of each of these forest formations). Twenty-five per cent (3734 presences for *F. sylvatica* and 14 661 for *Q. pyrenaica*) of the original data was enough to be considered as a fair representation of the "true" distribution of these taxa; it should be noted that most modelling exercises barely use more than a handful of presences, and very few count presences by the thousand, as used here. As in other published studies, the selection of this number of sites is somewhat arbitrary, but is more than enough to fit models (Elith & Graham 2009). Following Hernandez et al. (2006), downsized replicate models were compared with this reference model, considered to be ". . . the most representative of the true distribution of the species given the limitations of the modelling method, the species occurrence and environmental data available."

**Table 1.** Sampling details for *Fagus sylvatica* (Fs) and *Quercus pyrenaica* (Qp). Average number of presences and absences data per replicate and number of replicates per sample size.

| | Presences | Absences | Replicates |
|---|---|---|---|
| Fs | | | |
| 25% | 3734 | 106 330 | 1 |
| 15% | 2240 | 63 780 | 20 |
| 5% | 744 | 21 275 | 20 |
| 1% | 149 | 4250 | 20 |
| 0.1% | 15 | 430 | 20 |
| 0.05% | 8 | 215 | 20 |
| Qp | | | |
| 25% | 14 661 | 132 000 | 1 |
| 15% | 8800 | 79 200 | 20 |
| 5% | 3090 | 27 800 | 20 |
| 1% | 590 | 5300 | 20 |
| 0.1% | 84 | 770 | 20 |
| 0.05% | 29 | 260 | 20 |

In order to obtain real absences, absence data were extracted from mature forests other than the *F. sylvatica* or *Q. pyrenaica* formations, so as to prevent false negative records (i.e. no sampling was done on crop or deforested areas that could constitute part of the potential distribution area). Table 1 also presents the number of pixels sampled for each sample size down-sized category.

### *Small data sets without real absences and balanced prevalence*

We used presence data from the TROPICOS database (Missouri Botanical Garden; http://mobot.mobot.org/W3T/Search/vast.htlm) for two endemic species from Ecuador: *Anthurium dolichostachyum* and *A. mindense*. The former has a western distribution area, while the latter grows on both slopes of the Andean range (Fig. 2). This genus has

been thoroughly studied and collected in Ecuador by its specialist, Thomas B. Croat (783 collections and 236 localities in TROPICOS; see Croat 1992, 1999). All the specimens used in this study have been checked both for taxonomic identification and spatial accuracy (Mateo 2008).

Lacking real absences, we generated random pseudo-absences in an approximately equal number to the presences to avoid problems associated with unbalanced prevalence (Titeux 2006). To reduce the number of false negatives, we imposed the spatial restriction of a minimum distance of 30 km between the generated pseudo-absences and any of the known presences (Mateo et al. 2010). This distance could be arbitrarily set, or could be established according to a particular characteristic of the species, such as dispersion capacity (Graham & Hijmans 2006). In our case, 30 km is the maximum pixel size containing the same information as a pixel of size 1 km, calculated according to the Shannon entropy formula. This radius was calculated by doubling the pixel size from the original map (1 km pixel spatial resolution) and calculating the information contained in both the original pixel size map and the doubled pixel size map. Models of 1, 2, 4, 8, 16 and 32 km pixel size basically contain the same information per pixel, and so we decided to set the buffer size to 30 km (Mateo et al. 2010).

Both *Anthurium* species full-size data sets have 72 pixel presences. They were sampled at random to generate ten replicates, each of size 60, 50, 40, 30, 25, 18 and nine presences. The reference model to compare with these replicates (see above) was generated using the 72 presences.

*Environmental variables*

We used the 19 WorldClim 1.3 bioclimatic variables to build the models. These are described in Hijmans et al. (2005) and are freely available on the web (http://www.worldclim.org). There are 19 variables: annual mean temperature, mean diurnal range, isothermality, temperature seasonality, maximum temperature of warmest month, minimum temperature of coldest month, temperature annual range, mean temperature of wettest quarter, mean temperature of driest quarter, mean temperature of warmest quarter, mean temperature of coldest quarter, annual precipitation, precipitation of wettest month, precipitation of driest month, precipitation seasonality, precipitation of wettest quarter, precipitation of driest quarter, precipitation of warmest quarter, and precipitation of coldest quarter.

*Modelling technique*

Multivariate adaptive regression splines – MARS – (Friedman 1991; Hastie et al. 2001) have been applied in previous ecological modelling exercises (Muñoz & Felicísimo 2004; Leathwick et al. 2006; Elith & Leathwick 2007). MARS combine classical linear regression, mathematical construction of splines and binary recursive partitioning to produce a local model in which relationships between responses and predictors are either linear or not linear. MARS approximate the underlying function through a set of adaptive piece-wise linear regressions, termed *basis functions*, the slope of

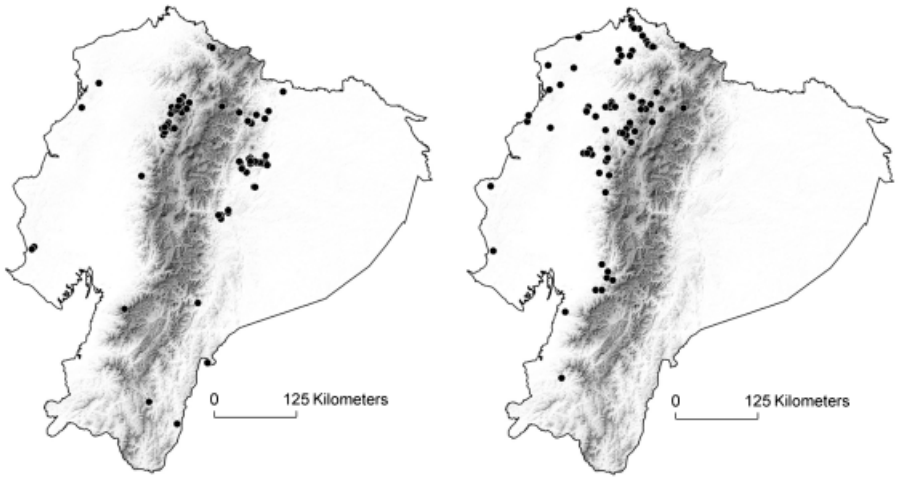


**Fig. 2.** Original data set (presences in TROPICOS database) for *Anthurium mindense* (left) and *Anthurium dolichostachyum* (right). WGS84 projection.

which changes at points called *knots*. The regression line is thus allowed to bend at the knots, which mark the end of one region of data and the beginning of another with different behaviour of the function. Knots are established in a forward/backward stepwise way. A model that clearly over-fits the data is produced first. In subsequent steps, knots that contribute least to the efficiency of the model are discarded by backward pruning steps. The best model is selected via cross-validation, a process that applies a penalty to each term (knot) added to the model to keep low complexity values.

We ran all the models using MARS 2.0 software (http://www.salford-systems.com). For each replicate, we ran 30 models using different parameters, mainly (1) maximum number of basic functions (15, 20, 25, 26, 27, 28, 29, 30, 45); (2) interactions (to allow or prevent interactions between basic functions); and (3) independent variables (to include all the variables or to eliminate mean annual temperature and mean annual precipitation, to reduce multicollinearity). Given the small sample sizes of the *Anthurium* data sets and because we were not interested in evaluating the individual models, we measured AUC using the same data set from which the models had been generated.

### *Model comparison*

In total, 101 models for each species of tree (one reference model, 20 replicates × five sample sizes) and 71 models for each species of *Anthurium* (one reference model, ten replicates × seven sample sizes) were built. To explore if the models generated with few presences were stable and reliable, they were compared to the reference model (full occurrence model in the *Anthurium* species, and the 25% total occurrence model in the tree species) using the Pearson product-moment correlation coefficient ($r$) in two ways: (1) for each sample size, each individual model (10 or 20 replicates, see above) was compared to the reference model, and then the average Pearson correlation coefficient was estimated for that particular sample size; and (2) for each sample size, an average model of the replicates was generated (here named consensus model), and correlated to the reference model. Following Hernandez et al. (2006), we consider that the reference model is the closest to the potential distribution of the species for the given modelling method.

In addition, we evaluated model stability. We calculated the correlation between all pairs of replicates per sample size. The inverse of the standard deviation of the average of $r$ was considered to be the stability indicator.

## Results

### *Reliability*

The AUC values obtained ranged from 0.941 to 1.000, which are well within what is considered a good measure of accuracy (Swets 1988).

When the downsized data sets are compared to the reference model using the Pearson correlation coefficient, three main results were highlighted (Fig. 3): (1) $r$ values decreased as number of presences decreased; (2) in all four species the correlation coefficient of the consensus model (average model for each sample size) was greater than the mean Pearson correlation coefficient (Fig. 3); and (3) the correlation coefficient of the consensus model was more stable in the two species with the greater sample size (Fig. 3).

In the two species of tree (European beech and Pyrenean oak) a tendency was observed towards a decrease of the mean correlation coefficient of the responses when sample size was decreased (Fig. 3). This value was much smaller (0.34) in the case of European beech, where the sample size was smaller (eight presences). For Pyrenean oak, with a sample size of 29 presences (0.01% of total occurrences data set), a correlation coefficient of 0.62 was obtained. In the case of European beech, 15 presences seem to be sufficient to obtain a model comparable to the reference model, since a correlation coefficient of 0.70 was obtained.

On the other hand, in both *Anthurium* species (Fig. 3), for which the original amount of data was quite limited compared to data for the tree species, the effect of the number of presences on the final models was much more dramatic. In every case, the mean correlation was much lower, even for a number of presences similar to that of the trees; for Pyrenean oak a value of 0.62 was obtained with 29 presences (0.01% of the total occurrences data set), while for the two *Anthurium* species values of 0.34 and 0.27 were obtained with 30 presences. For European beech a value of 0.70 was obtained with 15 presences (0.05% of the total occurrences data set), but for the two species of *Anthurium* the values were 0.31 and 0.22 with 18 presences.

### *Stability*

SDM stability decreased as sample size decreased (Table 2). The models of the two species of *Anthurim* were more unstable and showed more erratic behaviour than those of the two trees. Figures 4 and 5 show how the ten models of *Anthurium mindense* obtained with nine presences are

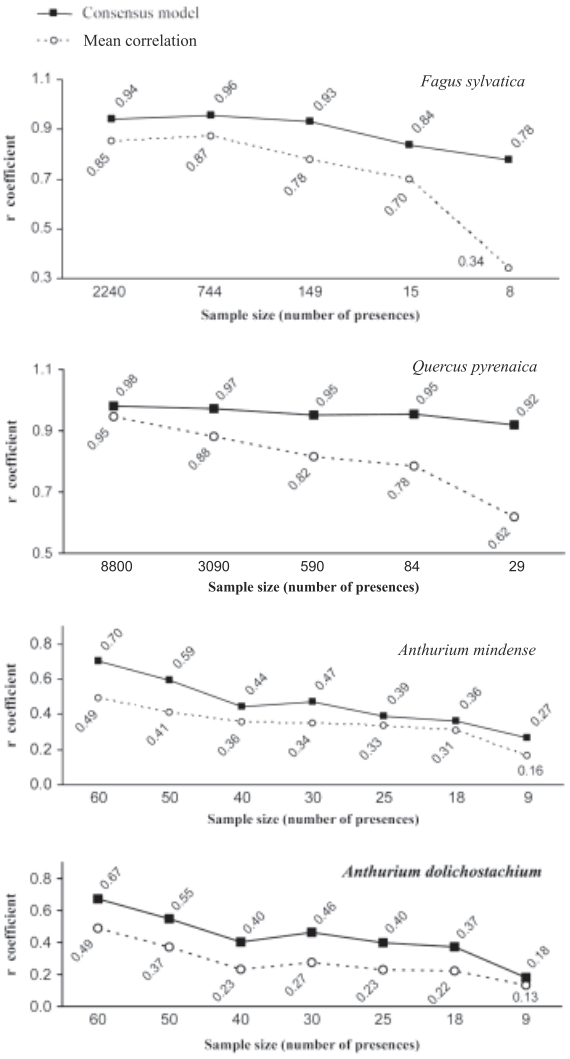


**Fig. 3.** Relation between the *r* Pearson coefficient and number of presences for the four species (*Fagus sylvatica, Quercus pyrenaica, Anthurium dolichostachyum* and *Anthurium mindense*). The continuous line with black squares represents the consensus model of the replicates (10 or 20) for each different sample size. The discontinuous line with a white circumference represents the mean Pearson correlation coefficient for each sample size.

much more unstable than models obtained with 18 presences.

## Discussion

*Comparison between data sets (small data sets without real absences and balanced prevalence versus large data sets with real absences and unbalanced prevalence)*

The models generated with data from the genus *Anthurium* were less reliable and less stable than models generated with data from European beech and Pyrenean oak. This effect was mainly related to five factors: (1) The quantity and quality of data on the initial presences. (2) Lower reliability of the pseudo-absences with respect to real absences used in the case of the two tree species. (3) The precise knowledge of distribution of the species in the case of the two tree species. (4) The larger number of precise absences used in the tree species provided extra information for the models, which helped to delimit areas not really suitable for development of these species. (5) Ecology of the species (range size and regional niche variation). Therefore, the important conclusion that can be drawn from this analysis is that the minimum sample size is closely linked to the ecology and distribution area of a species, as well as the quality of the initial data, both for presences as well as absences. We will now discuss some of these aspects further.

*Reliability*

The two species of tree, with a large starting sample size, showed similar modelling behaviour. The slope of the graph remained relatively constant until it reached the minimum sample size (0.01%), where an important drop in the correlation coefficient occurred. However, it is notable that for *Fagus sylvatica* a relatively reliable model was obtained with a smaller sample size (15 presences) than for *Quercus pyrenaica* (29 presences). This was likely due to the ecological requirements of *Fagus sylvatica*, which, despite its broad distribution, is characterized by a more sharply delimited ecological range, making it easier to model its distribution from smaller sample sizes.

Regarding the two *Anthurium* species, the effect of sample size was much more dramatic. In every case, the mean correlation values were lower, even for similar numbers of presences, compared to the two trees. At the beginning of this study, an opposite effect was expected, because, in the case of the trees, the reference models were generated with a much larger amount of data, and consequently the difference between reference model and downscaled models was expected to be larger. We consider that this unexpected result is mostly due to errors that occurred when geolocating specimens in natural history collections without full locality information (Margules & Pressey 2000; Soberon & Peterson 2004; Rowe 2005; Edwards et al. 2006; Papeş & Gaubert 2007), and also perhaps to uncertainty associated with using pseudo-absences (Mateo 2008).

**Table 2.** Comparison between all pairs of replicates per number of presences; *r* (mean): average correlation coefficient; standard deviation: standard deviation of the average of *r* (stability indicator); stability: 1/standard deviation.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| *Fagus sylvatica* | | | | | | | |
| Number of presences | 2240 (15%) | 745 (5%) | 150 (1%) | 15 (0.05%) | 8 (0.01%) | | |
| *r* (mean) | 0.821 | 0.825 | 0.674 | 0.588 | 0.145 | | |
| Standard deviation | 0.105 | 0.072 | 0.122 | 0.153 | 0.255 | | |
| Stability (1/SD) | 9.524 | 13.888 | 8.197 | 6.536 | 3.922 | | |
| *Quercus pyrenaica* | | | | | | | |
| Number of presences | 8800 (15%) | 3090 (5%) | 590 (1%) | 85 (0.05%) | 30 (0.01%) | | |
| *r* (mean) | 0.917 | 0.811 | 0.714 | 0.664 | 0.413 | | |
| Standard deviation | 0.043 | 0.114 | 0.162 | 0.219 | 0.205 | | |
| Stability (1/SD) | 23.256 | 8.772 | 6.173 | 4.566 | 4.878 | | |
| *Anthurium mindense* | | | | | | | |
| Number of presences | 60 | 50 | 40 | 30 | 25 | 18 | 9 |
| *r* (mean) | 0.441 | 0.426 | 0.667 | 0.457 | 0.706 | 0.706 | 0.192 |
| Standard deviation | 0.190 | 0.233 | 0.156 | 0.269 | 0.131 | 0.125 | 0.340 |
| Stability (1/SD) | 5.263 | 4.292 | 6.411 | 3.717 | 7.633 | 8.000 | 2.941 |
| *Anthurium dolichostachyum* | | | | | | | |
| Number of presences | 60 | 50 | 40 | 30 | 25 | 18 | 9 |
| *r* (mean) | 0.463 | 0.394 | 0.270 | 0.271 | 0.304 | 0.251 | 0.209 |
| Standard deviation | 0.191 | 0.255 | 0.233 | 0.228 | 0.203 | 0.237 | 0.311 |
| Stability (1/SD) | 5.236 | 3.922 | 4.292 | 4.386 | 4.926 | 4.219 | 3.215 |

False presences and false absences have a negative effect on the reliability of models, an effect that increases as sample size decreases (Carroll & Pearson 1998). If we start from a large number of presences and absences, the bias introduced by inaccurate (taxonomically or spatially) data will be compensated by the information contributed from the correct data. On the other hand, a single inaccurate data point in a small data set will have a dramatic effect on the resulting model.

### *Stability*

There were two quite different patterns of stability for the two groups of data used. For the two tree species, the stability decreased as sample size decreased (Table 2), while in the two *Anthurium* species this relation is not that clear for the stability, which behaves in a much more irregular manner (Table 2). In the case of the two *Anthurium* species, the stability values are relatively low at all sample sizes and, although stability values also behave irregularly for these two species, they increase as sample size decreases, in contrast to what happens in the case of the trees. We interpret this behaviour as a consequence of a likely collection bias in the *Anthurium* data, which seem to be spatially and, consequently, ecologically biased (Reddy & Dávalos 2003; Kadmon et al. 2004; Hortal et al. 2007). Such collection bias might thus result in certain localities being systematically under-represented by the SDM, which can only be resolved by additional collections (Raes & Steege 2007).

### *Consensus model*

The consensus model (mean of replicates per sample size) was shown to be more reliable in the two tree species than in the two *Anthurium* species (Fig. 3). The tree data sets had reliable information about the real distribution of the species, and we also generated a larger number of responses for each sample size than for the *Anthurium* data sets, which have a considerably smaller number of presences. As in the case of the tree species, *Anthurium* consensus models always showed a higher correlation with the reference model, although they were less stable than those of tree consensus models. In species poorly represented in NHCs, such as the *Anthurium* species, collection data likely offer highly biased information of the niche of the species, and therefore the consensus models cannot reconstruct their complete niche, unlike the case with the consensus models generated for well-collected organisms (see above for the trees). This becomes evident from a visual inspection of the maps (Figs. 4 and 5), which show that inclusion or not of certain presence data has a dramatic effect on the final results of the SDMs. Random selection of the presences from which the SDMs are generated leads to large variations in the final result of the model (Pearson et al. 2007), with more striking effects for decreases in sample size (Fig. 4).

### *Minimum number of presences*

The construction of reliable and stable SDMs depends on many factors, some of which are well known. Sample size is one element that can drama-

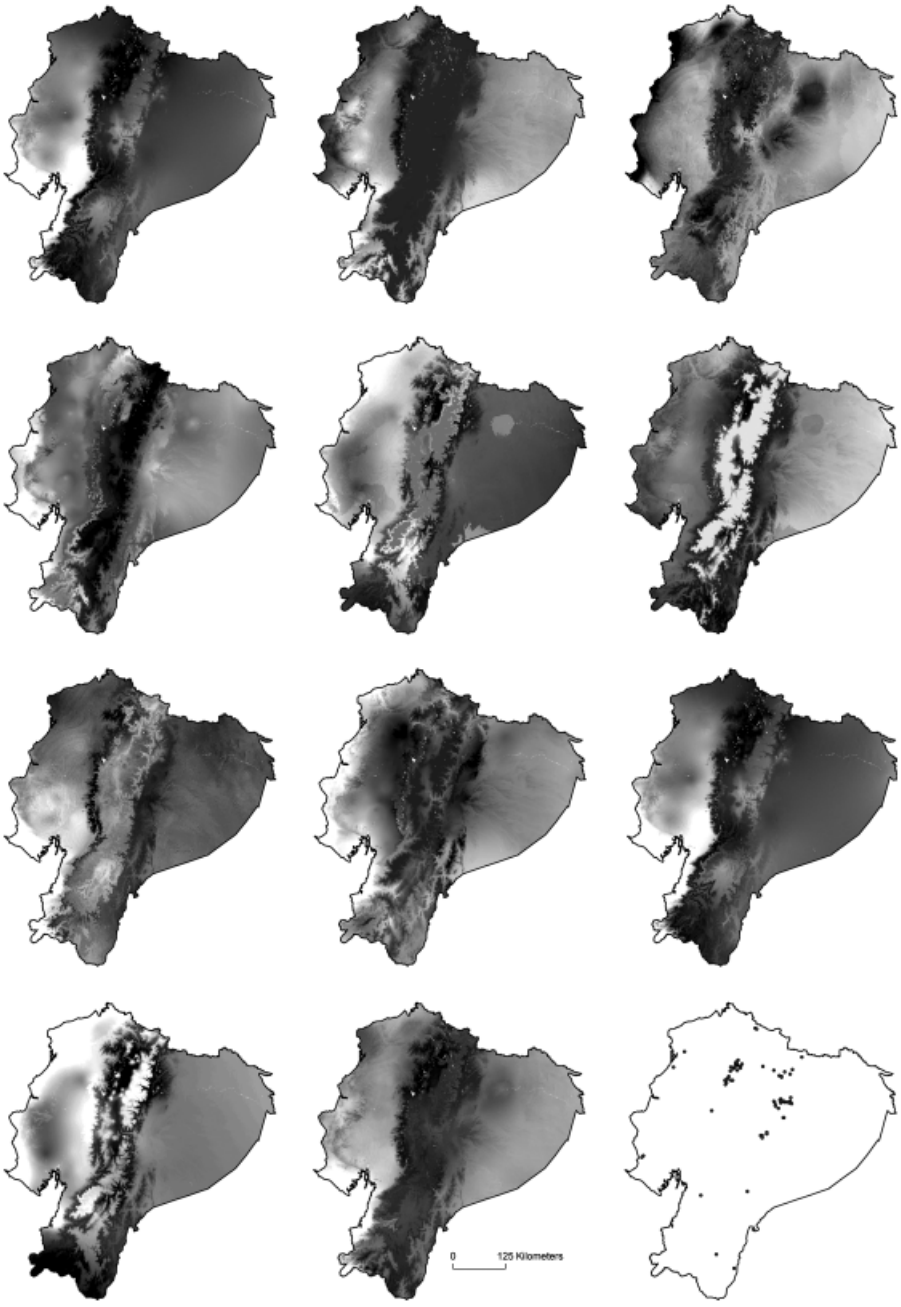


**Fig. 4.** The ten models generated with nine presences for *Anthurium mindense*. The consensus model of these ten models (map with the scale bars in the middle of the last line). All presences available for *A. mindense* (map on the right in last line).

tically affect the results of any study of ecological modelling. Several authors have documented a minimum sample size required to generate reliable SDMs. The results, even for the same method, are very variable (see Introduction), and according to results obtained in this study, we believe that a general rule regarding the minimum number of presences cannot be provided.

This minimum sample size can vary depending on many factors, such as the quality of the original data, method of ecological modelling, independent variables, pixel size, area being studied and ecology of the species (widespread or narrow distribution). Furthermore, the appropriate sample size depends on the objectives of each project, since, for example, in studies of very rare species there is no other op-

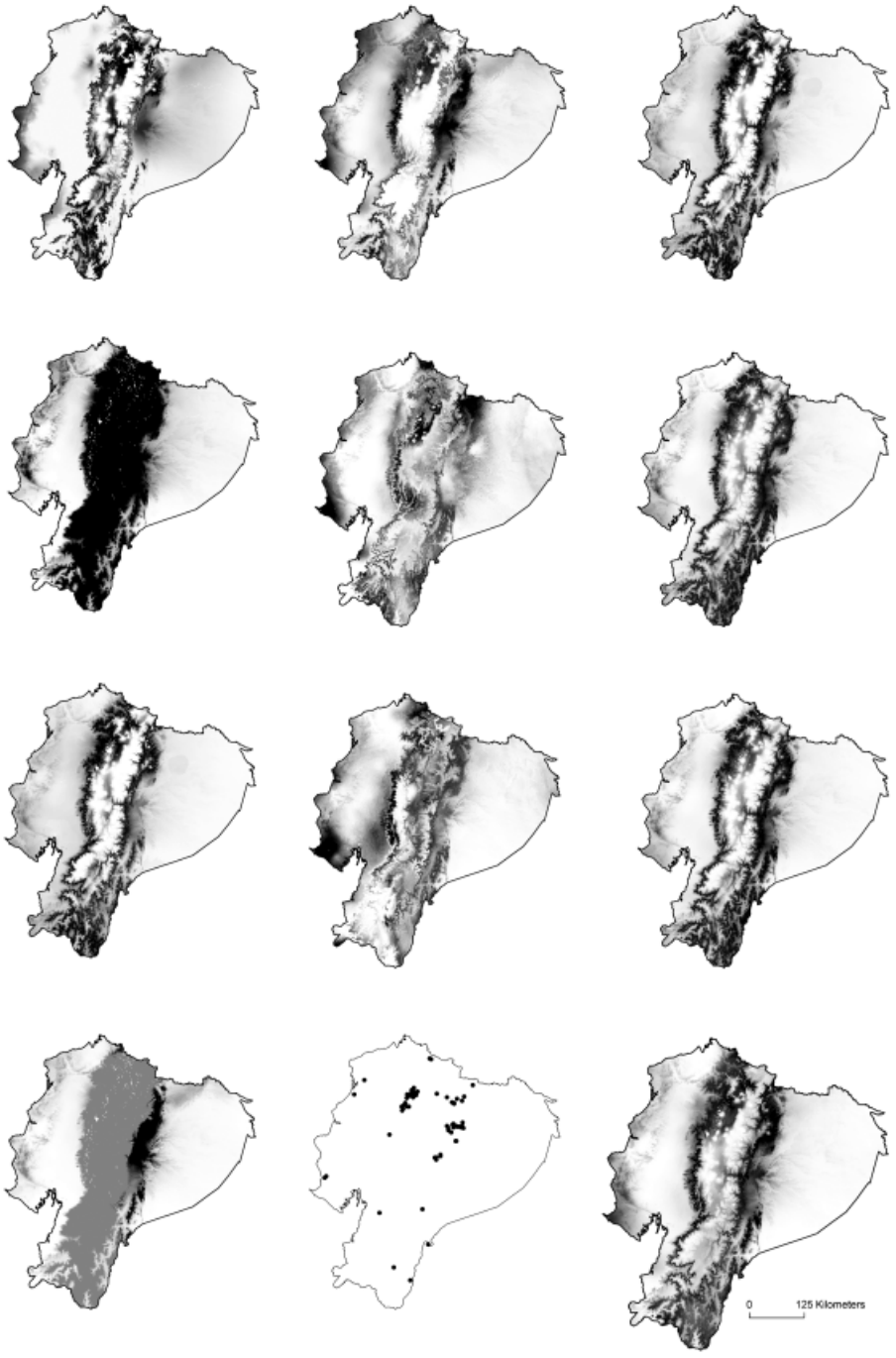


**Fig. 5.** The ten models generated with 18 presences for *Anthurium mindense*. The consensus model of these ten models (map with the scale bars in the middle of the last line). All the presences available for *A. mindense* (map on the right in the last line).

tion except to create the SDMs with very few presence data. In this situation where a species is closely associated with a particular habitat or geographical area, the SDMs may be precise, even with a very small sample size, as these few presences reliably represent the ecological variability in which the species grows and, therefore, generate reliable SDMs (Hernandez et al. 2006; Pearson et al. 2007). But this cannot be considered as a general rule, and we cannot apply this minimum sample size to species with a wider ecological niche. In this study, the four species analysed represent a relatively wide ecological niche and a widespread distribution, and therefore can be considered a more general application study.

Although, as already stated, a minimum sample size valid for all studies could not be established, we can evaluate the general trends and draw conclusions applicable to specific cases. In the first place, we can confirm that the number of presences used to generate SDMs with MARS dramatically affects their reliability and stability. This is consistent with previous studies using others methods. We observed a similar trend for all four species, but the effect is much more striking in the case of the two *Anthurium* species.

In general, the two *Anthurium* species show a decrease in mean correlation coefficient; in models created with 60, 50 and 40 presences this decrease is noticeable, although the correlation coefficient stabilizes in models with 30, 25 and 18 presences. Finally, there was an important drop for models obtained using nine presences. For both *Anthurium* species, very similar values were obtained in all of the sample sizes. These results indicate that there was practically no difference between using 40 and 18 presences, and also that it seems senseless to use less than around 18 presences to create an SDM, as this is the value after which the correlation coefficient drops significantly, which would indicate that the reliability of the generated models would also drop radically. In natural history collections, the number of presences for many species is usually less than 20 (Stockwell & Peterson 2002; Loiselle et al. 2008; Mateo 2008), and our results show that collecting more than around this number will have hardly any effect on the reliability of SDMs generated with MARS, except at levels exceeding 70 presences distributed more or less uniformly throughout the study area, which in most cases would imply an exceedingly high cost that is generally unaffordable.

### *Quality or quantity of data?*

In most cases, the "quality" (ecological information) of the presence/absence data is by far more important than the "number". Indeed, model predictions are influenced by the properties of the data, both quality and quantity, and distribution properties of the modelled species (Kadmon et al. 2003).

In the case of widely distributed species, the potential distribution for species may be difficult to model due to factors such as: (1) identification of important niche dimensions (Anderson et al. 2002); (2) ecological adaptation for subpopulations (Peterson & Holt 2003); (3) training data, source/sink dynamics (Pulliam 2000), or (4) ranges that might only be incompletely represented in the present databases. Additionally, spatially biased presence-only data sets could represent regional niche variation or only part of the ecological heterogeneity of the species. The fewer the number of presence data used to generate the model, the smaller the environmental universe sampled, and therefore the more difficult it becomes to capture the total niche dimensions of the species, which in turn implies that reliability and stability of the models will decrease. This problem can be partially solved by generating a consensus model among the different responses of a given sample size, combining information from various partial models (Araújo & New 2007; Marmion et al. 2008), which could generate even better results than a single model generated with all the available presences. Spatial partitioning of the data is necessary to improve predictions of models where regional niche variation occurs (Osborne & Suárez-Seoane 2002), and here consensus models may play a crucial role. However, this approach is not without risk, as regional portioning of the data may not necessarily be the most appropriate approach in all circumstances (Murphy & Lovett-Doust 2007). In the case of narrowly distributed species, there is no such niche variation, or it is weak, and as a consequence stable and reliable models can be generated from data sets with few presences (McPherson & Jetz 2007; Pearson et al. 2007).

### *Multivariate adaptive regression splines*

MARS, like other algorithms that create models of complex relationships and interactions between variables, require a large number of presences to generate optimal results (Guisan & Thuiller 2005; Wisz et al. 2008). If working with a limited number of presences, other methods, especially Maxent, can generate better models (Hernandez et al. 2006; Phillips et al. 2006; Papeş & Gaubert 2007; Pearson et al. 2007; Wisz et al. 2008). The results obtained in this study suggest that in order to generate optimal models, MARS need to be trained with at least 15 presences, although the results are not equally reliable in all cases. For example, European beech models generated with 15 presences are more stable and reliable than *Anthurium* species models generated with 18 presences. The obvious conclusion, indeed applicable to all ecological modelling methods, is that we cannot establish a valid criterion for all options; this will depend on initial data and the objectives of the study.

*Temperate versus tropical areas*

Most methodological ecological modelling studies are performed with data from temperate areas, from where more good quality data are available. However, most studies related to prioritization of areas for conservation and many studies on biodiversity patterns are focused on tropical areas, where data are generally scarce and highly biased. As a consequence, it is highly risky to extrapolate conclusions of a methodological study performed in temperate areas to tropical areas. The present study also shows that it is necessary to continue collecting data from areas with few samples obtained from natural history collections, like most tropical areas, which are indeed priority areas for biodiversity conservation (Myers et al. 2000; Deutsch et al. 2008). This implies that the current trend of decreasing the number of collecting expeditions should be reversed, as well as continuing to create or reinforce existing biodiversity information systems (Bisby 2000; Soberon & Peterson 2004; Guralnick et al. 2007). As already pointed out by Lobo (2008), we think that adding new data will be far more beneficial than developing more complex techniques (Lobo 2008).

*Conclusions*

At present, many conservation strategies, reserve designs and studies on the effects of climatic change are based on results generated through ecological modelling. However, in some of these studies the consequences of sample size on resulting models are not taken into consideration, which can lead to conclusions that are weakly supported.

An obvious use of SDMs generated with few presences is to direct fieldwork that aims to collect more data, which can be used to generate better, more reliable and stable SDMs (Pearson et al. 2007). However, this strategy is not always possible due to the usually very high associated costs, both in terms of time and money, of collecting new data, especially in tropical areas. Other uses of models generated with few presences or with collection-biased data must be used with caution, and always in association with species characteristics that help in the interpretation of such models.

This study is based on four species, which arguably could limit the generalization of the results, but still allows generation of a number of conclusions. (1) In general, model predictive power improves greatly when sample size exceeds 18-20 unique presences. (2) The reliability of models depends on the properties of the data, both the quantity and quality, as well as species ecological characteristics. (3) Depending on the number of unique presences in the initial data set, investigators must carefully distinguish between models that should be treated with skepticism and those whose predictions can be applied with reasonable confidence. (4) For species combining few initial presences and having wide environmental range variation, it is advisable to generate several replicate models that partition the initial data and generate a consensus model, as indicated in Araújo & New (2007). (5) Models of species with a narrow environmental range variation can be highly accurate in terms of stability and reliability, even when generated with few presences.

**Acknowledgements.** We thank the BBVA Foundation for financial support and the Missouri Botanical Garden for generously providing the *Anthurium* data. Tania Delgado checked the *Anthurium* data set for locality errors and helped us to understand *Anthurium* species distributions in Ecuador. We are indebted to A. Guisan and two anonymous reviewers for their insightful comments on the manuscript.

## References

Anderson, R.P. & Martinez-Meyer, E. 2004. Modeling species' geographic distributions for preliminary conservation assessments: an implementation with the spiny pocket mice (Heteromys) of Ecuador. *Biological Conservation* 116: 167–179.

Anderson, R.P., Gómez-Laverde, M. & Peterson, A.T. 2002. Geographical distributions of spiny pocket mice in South America: insights from predictive models. *Global Ecology and Biogeography* 11: 131–141.

Araújo, M.B. & Guisan, A. 2006. Five (or so) challenges for species distribution modelling. *Journal of Biogeography* 33: 1677–1688.

Araújo, M.B. & New, M. 2007. Ensemble forecasting of species distributions. *Trends in Ecology and Evolution* 22: 42–47.

Araújo, M.B. & Williams, P.H. 2000. Selecting areas for species persistence using occurrence data. *Biological Conservation* 96: 331–345.

Araújo, M.B., Pearson, R.G., Thuiller, W. & Erhard, M. 2005a. Validation of species–climate impact models under climate change. *Global Change Biology* 11: 1504–1513.

Araújo, M.B., Thuiller, W., Williams, P.H. & Reginster, I. 2005b. Downscaling European species atlas distributions to a finer resolution: implications for conservation planning. *Global Ecology and Biogeography* 14: 17–30.

Araújo, M.B., Thuiller, W. & Pearson, R.G. 2006. Climate warming and the decline of amphibians and

reptiles in Europe. *Journal of Biogeography* 33: 1712–1728.

Beguería, S. 2006. Validation and evaluation of predictive models in hazard assessment and risk management. *Natural Hazards* 37: 315–329.

Bisby, F.A. 2000. The quiet revolution: biodiversity informatics and the internet. *Science* 289: 2309–2312.

Botkin, D.B., Saxe, H., Araújo, M.B., Betts, R., Bradshaw, R.H.W., Cedhagen, T., Chesson, P., Dawson, T.P., Etterson, J.R., Faith, D.P., Ferrier, S., Guisan, A., Hansen, A.S., Hilbert, D.W., Loehle, C., Margules, C., New, M., Sobel, M.J. & Stockwell, D.R.B. 2007. Forecasting the effects of global warming on biodiversity. *Bioscience* 57: 227–236.

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.

Carroll, S.S. & Pearson, D.L. 1998. The effects of scale and sample size on the accuracy of spatial predictions of tiger beetle (Cicindelidae) species richness. *Ecography* 21: 401–414.

Cayuela, L., Golicher, D.J., Newton, A.C., Kolb, M., Alburquerque, F.S., Arets, E.J.M.M., Alkemade, J.R.M. & Pérez, A.M. 2009. Better species distribution modeling needed for the tropics. *Tropical Conservation Science* 2: 319–352.

Ceballos, L. 1966 Mapa forestal de España. pp.50+20 maps; 82×53 cm. In: Ministerio de Agricultura. Dirección General de Montes, Caza y Pesca Fluvial, Madrid, ES.

Costa Tenorio, M., Morla Juaristi, C. & Sainz Ollero, H. 1998. *Los bosques ibéricos. Una intrepetación geobotánica*. Editorial Planeta, Barcelona, ES.

Croat, T.B. 1992. Species diversity of Araceae in Colombia: a preliminary survey. *Annals of the Missouri Botanical Garden* 79: 17–28.

Croat, T.B. 1999. Araceae. In: Jørgensen, P.M. & León-Yánez, S. (eds.) *Catalogue of the vascular plants of Ecuador*. pp. 227–246. Missouri Botanical Garden, St. Louis, MO, US.

Cuesta-Camacho, F., Ganzenmueller, A. & Baquero, F. 2006. Predicting distribution of Andean-centered taxa using ecological niche modelling methods. *Lyonia* 9: 19–33.

Cumming, G.S. 2000a. Using between-model comparisons to fine-tune linear models of species ranges. *Journal of Biogeography* 27: 441–455.

Cumming, G.S. 2000b. Using habitat models to map diversity: pan-African species richness of ticks (Acari: Ixodida). *Journal of Biogeography* 27: 425–440.

Deutsch, C.A., Tewksbury, J.J., Huey, R.B., Sheldon, K.S., Ghalambor, C.K., Haak, D.C. & Martin, P.R. 2008. Impacts of climate warming on terrestrial ectotherms across latitude. *Proceedings of the National Academy of Sciences USA* 105: 6668–6672.

Drake, J.M., Randin, C. & Guisan, A. 2006. Modelling ecological niches with support vector machines. *Journal of Applied Ecology* 43: 424–432.

Edwards, J.T.C., Cutler, D.R., Zimmermann, N.E., Geiser, L. & Moisen, G.G. 2006. Effects of sample survey design on the accuracy of classification tree models in species distribution models. *Ecological Modelling* 199: 132–141.

Elith, J. & Graham, C.H. 2009. Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. *Ecography* 32: 66–77.

Elith, J. & Leathwick, J.R. 2007. Predicting species distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. *Diversity and Distributions* 13: 265–275.

Elith, J., Graham, C.H., Anderson, R.P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L.G., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J.M., Peterson, A.T., Phillips, S.J., Richardson, K., Scachetti-Pereira, R., Schapire, R.E., Soberón, J., Williams, S., Wisz, M.S. & Zimmermann, N.E. 2006. Novel methods improve prediction of species' distributions from occurrence data. *Ecography* 29: 129–151.

Friedman, J.H. 1991. Multivariate adaptive regression splines. *Annals of Statistics* 19: 1–141.

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

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

Guisan, A. & Zimmermann, N.E. 2000. Predictive habitat distribution models in ecology. *Ecological Modelling* 135: 147–186.

Guisan, A., Zimmermann, N.E., Elith, J., Graham, C.H., Phillips, S. & Peterson, A.T. 2007. What matters for predicting the occurrences of trees: techniques, data, or species' characteristics? *Ecological Monographs* 77: 615–630.

Guralnick, P., Hill, A.W. & Lane, M. 2007. Towards a collaborative, global infrastructure for biodiversity assessment. *Ecology Letters* 10: 663–672.

Hastie, T., Tibshirani, R. & Friedman, J. 2001. *The elements of statistical learning: data mining, inference, and prediction*. Springer, New York, NY, US.

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

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G. & Jarvis, A. 2005. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology* 25: 1965–1978.

Hjort, J. & Marmion, M. 2008. Effects of sample size on the accuracy of geomorphological models. *Geomorphology* 102: 341–350.

Hortal, J., Lobo, J.M. & Jiménez-Valverde, A. 2007. Limitations of biodiversity databases: case study on seed-plant diversity in Tenerife, Canary Islands. *Conservation Biology* 21: 853–863.

Iverson, L.R. 2004. How fast and far might tree species migrate in the eastern United States due to climate change? *Global Ecology and Biogeography* 13: 209–219.

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

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

Leathwick, J.R., Elith, J. & Hastie, T. 2006. Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. *Ecological Modelling* 199: 188–196.

Lobo, J.M. 2008. More complex distribution models or more representative data? *Biodiversity Informatics* 5: 14–19.

Lobo, J.M., Castro, I. & Moreno, J.C. 2001. Spatial and environmental determinants of vascular plant species richness distribution in the Iberian Peninsula and Balearic Islands. *Biological Journal of the Linnean Society of London* 73: 233–253.

Loiselle, B., Howell, C.A., Graham, C.H., Goerck, J.M., Brooks, T., Smith, K.G. & Williams, P.H. 2003. Avoiding pitfalls of using species distributions models in conservation planning. *Conservation Biology* 17: 1591–1600.

Loiselle, B.A., Jørgensen, P.M., Consiglio, T., Jiménez, I., Blake, J.G., Lohmann, L.G. & Montiel, O.M. 2008. Predicting species distributions from herbarium collections: does climate bias in collection sampling influence model outcomes? *Journal of Biogeography* 35: 105–116.

López González, G. 1990. Fagaceae. In: Castroviejo, S., Laínz, M., González, G.L., Montserrat, P., Garmendia, F.M., Paiva, J. & Villar, L. (eds.) *Flora iberica. Plantas vasculares de la Península Ibérica e Islas Baleares. Vol. II [Platanaceae–Plumbaginaceae (patim)]*. Real Jardín Botánico (C.S.I.C.), Madrid, ES.

Luoto, M.Luoto, M., Heikkinen, R.K. & Saarinen, K. 2005. Uncertainty of bioclimate envelope models based on the geographical distribution of species. *Global Ecology and Biogeography* 14: 575–584.

Luoto, M., Heikkinen, R.K., Poyry, J. & Saarinen, K. 2006. Determinants of the biogeographical distribution of butterflies in boreal regions. *Journal of Biogeography* 33: 1764–1778.

Manel, S., Williams, H.C. & Ormerod, S.J. 2001. Evaluating presence–absence models in ecology: the need to account for prevalence. *Journal of Applied Ecology* 38: 921–931.

Margules, C.R. & Pressey, R.L. 2000. Systematic conservation planning. *Nature* 405: 243–252.

Marmion, M., Parviainen, M., Luoto, M., Heikkinen, R.K. & Thuiller, W. 2008. Evaluation of consensus methods in predictive species distribution modelling. *Diversity and Distributions* 15: 59–69.

Mateo, R.G. 2008. Modelos predictivos de riqueza de diversidad vegetal. Comparación y optimización de métodos de modelado ecológico. PhD thesis, Universidad Complutense de Madrid, Madrid, ES.

Mateo, R.G., Croat, T.B., Felicísimo, Á.M. & Muñoz, J. 2010. Profile or group discriminative techniques? Generating reliable pseudo-absences and target-group absences from natural history collections. *Diversity and Distributions* 16: 84–94.

McClean, C.J., Lovett, J.C., Küper, W., Hannah, L., Sommer, J.H., Barthlott, W., Termansen, M., Smith, G.F., Tokumine, S. & Taplin, J.R.D. 2005. African plant diversity and climate change. *Annals of the Missouri Botanical Gardens* 92: 135–152.

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

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.

Muñoz, J. & Felicísimo, A.M. 2004. A comparison between some statistical methods commonly used in predictive modelling. *Journal of Vegetation Science* 15: 285–292.

Murphy, H.T. & Lovett-Doust, J. 2007. Accounting for regional niche variation in habitat suitability models. *Oikos* 116: 99–110.

Myers, N., Mittermeler, R.A., Mittermeler, C.G., Fonseca, G.A.B. & Kent, J. 2000. Biodiversity hotspots for conservation priorities. *Nature* 403: 853–858.

Ortega-Huerta, M.A. & Peterson, A.T. 2004. Modelling spatial patterns of biodiversity for conservation prioritization in North-eastern Mexico. *Diversity and Distributions* 10: 39–54.

Osborne, P.E. & Suárez-Seoane, S. 2002. Should data be partitioned spatially before building large-scale distribution models? *Ecological Modelling* 157: 249–259.

Papeş, M. & Gaubert, P. 2007. Modelling ecological niches from low numbers of occurrences: assessment of the conservation status of poorly known viverrids (Mammalia, Carnivora) across two continents. *Diversity and Distributions* 13: 890–902.

Pearce, J. & Boyce, M. 2006. Modelling distribution and abundance with presence-only data. *Journal of Applied Ecology* 43: 405–412.

Pearce, J. & Ferrier, S. 2000. Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling* 133: 225–245.

Pearman, B., Randin, C.F., Broennimann, O., Vittoz, P., Knaap, W.O.v.d., Engler, R., Lay, G.L., Zimmermann, N.E. & Guisan, A. 2008. Prediction of plant species distributions across six millennia. *Ecology Letters* 11: 357–369.

Pearson, R.G., Raxworthy, C.J., Nakamura, M. & Peterson, A.T. 2007. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. *Journal of Biogeography* 34: 102–117.

Peterson, A.T. & Holt, R.D. 2003. Niche differentiation in Mexican birds: using point occurrences to detect ecological innovation. *Ecology Letters* 6: 774–782.

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

Pulliam, R.H. 2000. On the relationship between niche and distribution. *Ecology Letters* 3: 349–361.

Raes, N. & Steege, H. 2007. A null-model for significance testing of presence-only species distribution models. *Ecography* 30: 727–736.

Raven, P.H. & Wilson, E. 1992. A fifty-year plan for biodiversity surveys. *Science* 258: 1099–1100.

Reddy, S. & Dávalos, L.M. 2003. Geographical sampling bias and its implications for conservation priorities in Africa. *Journal of Biogeography* 30: 1719–1727.

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

Richards, L., Carstens, B.C. & Knowles, L.L. 2007. Distribution modelling and statistical phylogeography: an integrative framework for generating and testing alternative biogeographical hypotheses. *Journal of Biogeography* 34: 1833–1845.

Rissler, L.J., Hijmans, R.J., Graham, C.H., Moritz, C. & Wake, D.B. 2006. Phylogeographic lineages and species comparisons in conservation analyses: a case study of California Herpetofauna. *The American Naturalist* 167: 655–666.

Rowe, R.J. 2005. Elevational gradient analyses and the use of historical museum specimens: a cautionary tale. *Journal of Biogeography* 32: 1883–1897.

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

Soberon, J. & Peterson, A.T. 2004. Biodiversity informatics: managing and applying primary biodiversity data. *Philosophical Transactions of the Royal Society of London Series B Biological Sciences* 359: 689–698.

Stockwell, D. & Peters, D. 1999. The GARP modelling system: problems and solutions to automated spatial prediction. *International Journal of Geographical Information Science* 13: 143–158.

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

Swets, J.A. 1988. Measuring the accuracy of diagnostic systems. *Science* 240(4857): 1285–1293.

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

Titeux, N. 2006 Modelling species distribution when habitat occupancy departs from suitability. Application to birds in a landscape context. PhD thesis, Université Catholique de Louvain, Louvain-la-Neuve. BE.

Verbyla, D.L. 1986. Potential prediction bias in regression and discriminant analysis. *Canadian Journal of Forest Research* 16: 1255–1257.

Verbyla, D.L. & Litvaitis, J.A. 1989. Resampling methods for evaluating class accuracy of wildlife habitat models. *Environmental Management* 13: 783–787.

Wisz, M.S., Hijmans, R.J., Li, J., Peterson, A.T., Graham, C.H., Guisan, A. & Group, N.P.S.D.W. 2008. Effects of sample size on the performance of species distribution models. *Diversity and Distributions* 14: 763–773.

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

Received 17 December 2008;
Accepted 5 May 2010.
Co-ordinating Editor: Dr. Bastow Wilson.