Research

Ecography 39: 542–552, 2016
doi: 10.1111/ecog.01509

Subject Editor: Miguel Nakamura. Editor-in-Chief: Miguel Araújo. Accepted 25 May 2015

Minimum required number of specimen records to develop accurate species distribution models

André S. J. van Proosdij, Marc S. M. Sosef, Jan J. Wieringa and Niels Raes

A. S. J. van Proosdij (andrevanproosdij@hotmail.com), M. S. M. Sosef and J. J. Wieringa, Biosystematics Group, Wageningen Univ., Droevendaalsesteeg 1, NL-6708 PB Wageningen, the Netherlands. – N. Raes, ASJvP and JJW, Naturalis Biodiversity Center (Botany section), Darwinweg 2, NL-2333 CR Leiden, the Netherlands. MSMS also at: Botanic Garden Meise, Nieuwelaan 38, BE-1860 Meise, Belgium.

Species distribution models (SDMs) are widely used to predict the occurrence of species. Because SDMs generally use presence-only data, validation of the predicted distribution and assessing model accuracy is challenging. Model performance depends on both sample size and species' prevalence, being the fraction of the study area occupied by the species. Here, we present a novel method using simulated species to identify the minimum number of records required to generate accurate SDMs for taxa of different pre-defined prevalence classes. We quantified model performance as a function of sample size and prevalence and found model performance to increase with increasing sample size under constant prevalence, and to decrease with increasing prevalence under constant sample size. The area under the curve (AUC) is commonly used as a measure of model performance. However, when applied to presence-only data it is prevalence-dependent and hence not an accurate performance index. Testing the AUC of an SDM for significant deviation from random performance provides a good alternative. We assessed the minimum number of records required to obtain good model performance for species of different prevalence classes in a virtual study area and in a real African study area. The lower limit depends on the species' prevalence with absolute minimum sample sizes as low as 3 for narrow-ranged and 13 for widespread species for our virtual study area which represents an ideal, balanced, orthogonal world. The lower limit of 3, however, is flawed by statistical artefacts related to modelling species with a prevalence below 0.1. In our African study area lower limits are higher, ranging from 14 for narrow-ranged to 25 for widespread species. We advocate identifying the minimum sample size for any species distribution modelling by applying the novel method presented here, which is applicable to any taxonomic clade or group, study area or climate scenario.

Despite globally increasing investments in biodiversity research, our knowledge of the biodiversity on our planet is still limited, especially for data-sparse areas like the tropics (Whittaker et al. 2005, Costello et al. 2013). We can only guess at the total number of extant species, let alone that we know their spatial distribution (Mora et al. 2011). Rare species, those with either a small range or a low abundance (Rabinowitz 1981), represent the vast majority of species (Longino et al. 2002, ter Steege et al. 2013) and are consequently represented by few samples in natural history collections, our primary source of distributional data. Typically, these collections have a long-tailed relative abundance distribution as illustrated by ter Steege et al. (2011) for the Guianas. Species distribution models (SDMs) have been developed to overcome this lack of information (Guisan and Zimmermann 2000, Araujo and Peterson 2012) as they are able to predict the probability of occurrence of species for non-sampled areas too. SDMs relate recorded species presences to abiotic factors that are thought to determine the species' distribution (Araujo and Peterson 2012). SDMs are thereby built on the assumption that the sample data cover the species' full ecological range (Sánchez-Fernández et al. 2011, Raes 2012).

The effect of sample size on model accuracy is an aspect that is often neglected (Wisz et al. 2008, Mateo et al. 2010). However, ignoring this effect results in increased levels of error in distribution models for species represented by (too) few records, which are mostly rare species. In addition to sample size, the species' prevalence has a strong impact on model performance; species' prevalence is defined as the fraction of the study area occupied by a species (McPherson et al. 2004). Model performance for ecologically and geographically narrow-ranged species is significantly better compared to widespread species found in a wider range of habitats (Hernandez et al. 2006, Mateo et al. 2010, Lobo and Tognelli 2011, Tessarolo et al. 2014). Identifying the lower limit of the number of records that is required to develop accurate SDMs in relation to the species' prevalence

is therefore highly topical. Despite the large number of studies using SDMs and the results from recent studies on the negative effects of small sample sizes on SDM performance (Loiselle et al. 2008, Wisz et al. 2008, Mateo et al. 2010, Tessarolo et al. 2014), few studies actually address the minimum number of unique records required to generate an accurate SDM. Model performance is known to rapidly decrease for sample sizes smaller than 20 (Stockwell and Peterson 2002) or 15 (Papeş and Gaubert 2007), and is dramatically poor for samples sizes smaller than 5 records (Pearson et al. 2007). Contrary to this, high model accuracy was observed using several modelling techniques for models based on samples as small as 5, 10 and 25 compared to models based on 100 samples (Hernandez et al. 2006). Given that the true distribution of a species is unknown, model evaluation in these studies is based on in-sample test data and not on independent test data.

Standard evaluation parameters are based on a confusion matrix measuring both sensitivity (correctly predicted presences) and specificity (correctly predicted absences) (Fielding and Bell 1997). Consequently, these metrics require absence data, which are typically not available when presence-only data from herbaria or zoological collections are used. To remedy the lack of absence data, random background data or pseudo-absences are used instead (Phillips et al. 2009). Commonly, the number of background points is high compared to the number of presences, resulting in a low sampling prevalence; where sampling prevalence is defined as the number of presences relative to the entire sample. From the suite of model accuracy measures, the area under the curve (AUC) of the receiver operator characteristic (ROC) is the only one shown to be largely independent of sampling prevalence when applied to presence–absence data (McPherson et al. 2004). This renders the AUC the only useful indicator of model accuracy applicable to SDMs based on low-prevalence data, typical for presence-only data samples (Metz 1978, Fielding and Bell 1997). The AUC value translates to the chance that a randomly chosen presence has a higher predicted probability of occurrence than a randomly chosen absence. However, when applied to presence-only data, the use of AUC values is strongly criticized for the above mentioned imbalance between presences and absences, where including more absences that are environmentally more distant from the species' presences increases the fraction of correctly predicted absences (specificity), resulting in higher AUC values (Lobo et al. 2008, Jiménez-Valverde 2012). In addition, when used on unbiased presence-only data, the maximum achievable value of the AUC is not 1, but $1 - a/2$, where a represents the fraction of the area covered by the species' true distribution, which is typically not known (Phillips et al. 2006, Raes and ter Steege 2007, Jiménez-Valverde 2012, Smith 2013). Hence, when applied to presence-only data, the maximum AUC value is species' prevalence sensitive after all, and the commonly applied AUC value of 0.7 indicative for an SDM with acceptable accuracy is flawed (Raes and ter Steege 2007). To overcome this problem, Raes and ter Steege (2007) developed a null-model test to assess whether the AUC value of an SDM deviates significantly from random expectation. However, a null-model test neither assesses how accurate the species' real distribution is modelled, nor how many records are required to obtain high model accuracy. Here, we introduce the use of simulated species with defined occurrence probability to rigorously assess how many records are required to develop accurate SDMs.

In a virtual environment using simulated species, the species' response to environmental variables is fully controlled and thereby its reciprocal spatial distribution is defined and known (Hirzel et al. 2001, Austin et al. 2006, Zurell et al. 2010, Duan et al. 2015). The use of simulated species has been advocated to systematically evaluate how specific aspects of data, sampling strategy, model building and model evaluation affect SDMs (Saupe et al. 2012, Miller 2014). Studies using simulated species offer unique opportunities to assess SDM accuracy for different sample sizes and species' prevalence classes (Meynard and Quinn 2007, Jiménez-Valverde et al. 2009). AUC values can be calculated on defined presence and absence data derived from a probability of occurrence distribution and compared with AUC values of SDMs based on presence-only and background data. In addition, the defined probability of occurrence distribution can be compared with the predicted probability of occurrence distribution using Schoener's D and Hellinger distance I metrics (Schoener 1970, Warren et al. 2008), which are widely used to measure niche overlap (Rödder and Engler 2011). Once tested in a fully controlled virtual environment, the same method can be applied to a real environment. As a pilot we selected tropical Africa as the real environment, focusing on the country of Gabon, as we prepare a botanical diversity assessment using SDMs for this country.

Specifically, we assess the effects of sample size and species' prevalence on SDM accuracy using simulated species in a virtual as well as an African study area. We present a novel method to rigorously identify the minimum number of records relative to the species' prevalence that is required to generate SDMs with high model accuracy.

Material and methods

We used the following procedures to define simulated species for different prevalence classes in a virtual as well as an African study area. To increase readability, 'simulated species' are referred to as 'species', unless stated otherwise. All analyses were performed in R (R Core Team) using functions described below and provided in six separate R scripts (Supplementary material Appendix 4–9). A brief manual explaining the application of the method presented here is provided (Supplementary material Appendix 3).

Virtual study area and simulated species

The virtual study area is defined as a square of 100 by 100 raster cells in which two orthogonal gradients of equal length shaped the ecological landscape: the first linearly increasing from west to east, the second linearly increasing from north to south with the mean value for both located in the center of the study area (0, 0) (Fig. 1a, b). We defined species' prevalence as the fraction of raster cells where the species is present and used six prevalence classes: 0.05, 0.1, 0.2, 0.3, 0.4, and 0.5. Recent studies on trees (Boucher-Lalonde et al. 2012)

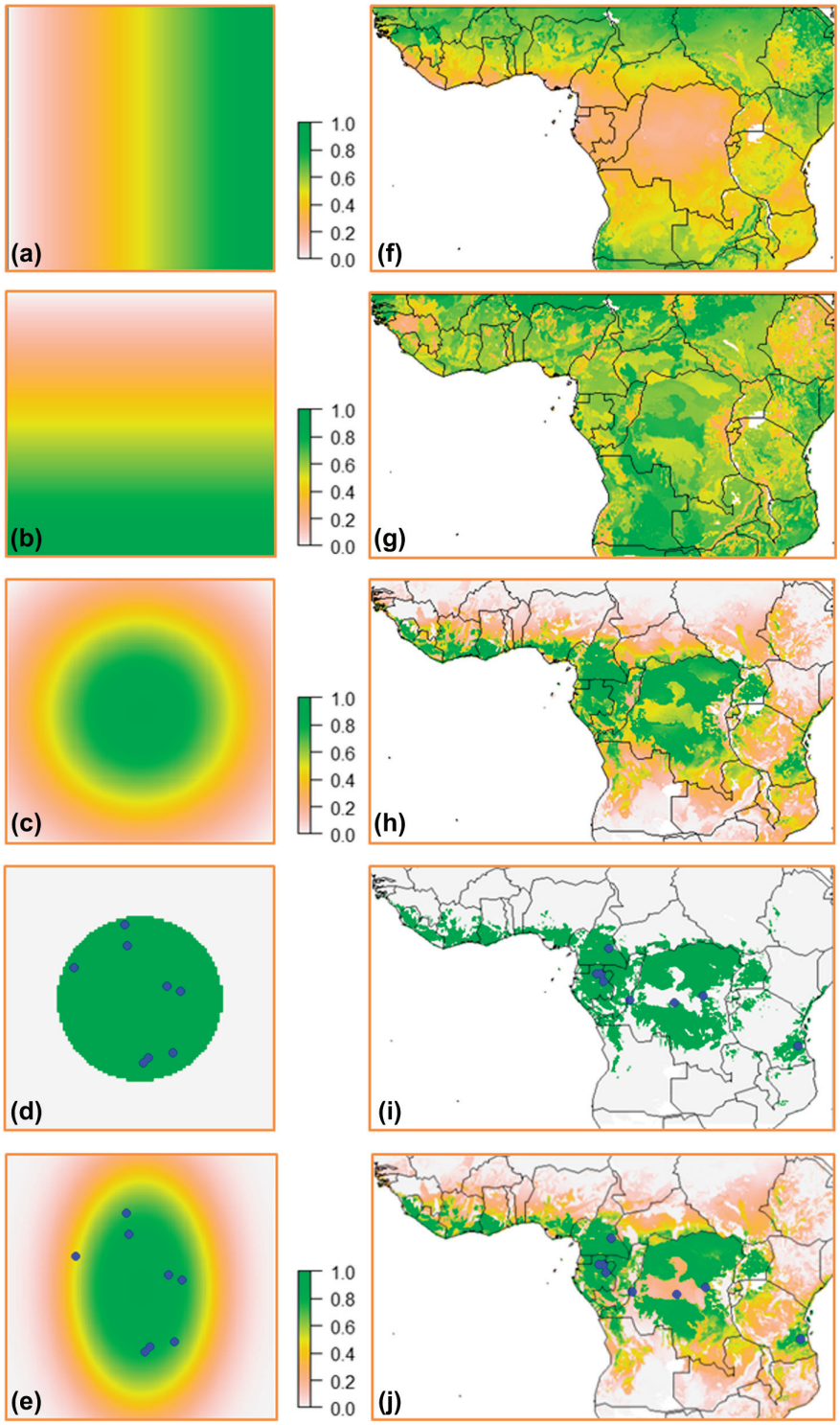

Figure 1. Methodological steps illustrated by examples of simulated species with prevalence 0.2 in the virtual (a–e) and in the real African study area (f–j): 2 orthogonal variables shaping the study area (a and b); defined habitat suitability (c); defined presence (green areas) and absence (white areas) of simulated species and sampled locations (blue dots, sample size 8) (d); predicted habitat suitability and sampled locations (e); 2 orthogonal variables (PCA1 and PCA2) shaping the study area (f and g); defined habitat suitability (h); defined presence (green areas) and absence (white areas) of simulated species and sampled locations (blue dots, sample size 8) (i); predicted habitat suitability and sampled locations (j).

and on birds and mammals (Boucher-Lalonde et al. 2014) using temperature and precipitation, showed that in general a species' niche can be described as a Gaussian function to these environmental variables. Based on these findings, we used a bivariate normal function as did others for the same reasons (Broennimann et al. 2011, Varela et al. 2014, Duan et al. 2015). We defined our simulated species by computing their habitat suitability or probability of occurrence as a

bivariate normal response to the two orthogonal gradients using the 'dmvnorm' function of the R-library 'mvtnorm' (Gentz et al. 2014). We assumed that the virtual species' distribution is shaped by these two orthogonal factors only and that niche filling is complete. We defined the different prevalence classes by increasing the standard deviation (SD) of the bivariate normal response (Fig. 1c). This procedure resulted in a defined habitat suitability score for each raster cell for each species. In our virtual study area, the optimum of the ecological niche of each species was set at the center (0, 0). We defined a species to be present in raster cells whose environmental bivariate variables are within the central region of the bivariate normal density that has probability 68%. Here, this region is represented by a circle as the two axes represent fully orthogonal, normalized variables with the same variance. Hence, the 68% circle cuts the axes at the points (optimum – 1 SD, 0), (0, optimum + 1 SD), (optimum + 1 SD, 0), and (0, optimum – 1 SD) (Fig. 1d). For each prevalence class, a small initial SD was iteratively increased until the desired prevalence value was approximated by less than 1% difference (for details see Supplementary material Appendix 5).

African study area and simulated species

Our real world study area encompassed most of tropical Africa ranging from 15°N to 19°S and from 18°W to 43°E. The African study area covered 179 994 raster cells with environmental data at 5 arc-minutes spatial resolution, excluding oceans and other large water bodies. Similar to our virtual study area following Broennimann et al. (2011), we used two orthogonal gradients, that were constructed by means of a principal components analysis (PCA) on fifteen selected environmental variables. These included bioclimatic variables (<www.worldclim.org>) (Hijmans et al. 2005), soil variables (Harmonized World Soil Database) (FAO/IIASA/ISRIC/ISSCAS/JRC 2012), and 90 m resolution elevation data (<www.srtm.csi.cgiar.org>) (Supplementary material Appendix 1). From 39 original variables we selected fifteen variables that had a Spearman's $|rho| < 0.7$ (Dormann et al. 2013) (for details see Supplementary material Appendix 4). We used the first two standardized PCA axes that together explained 43% of the variance in multivariate environmental parameter space (Fig. 1f, g). In our African study area, the niche optimum was different for each simulated species, reflecting that each species has a unique ecological niche (Aguirre-Gutiérrez et al. 2014). For each species, the species' optimum was defined by randomly selecting one raster cell from the area delineating Gabon extended by a 5 degree buffer zone. The values of the two PCA-based predictors at this randomly selected location were used as the means of the species' bivariate normal response curve, thus defining the species' optimum (Fig. 1h, for details see Supplementary material Appendix 7). Again, we defined a species to be present in raster cells whose environmental bivariate variables are within the central region of the bivariate normal density that has probability 68% (Fig. 1i), following the same procedure as in the virtual study area. Species' prevalence classes in the African study area were the same as in the virtual study area.

Sampling and replications

For both study areas and for each prevalence class we defined twenty-four sample sizes: 3–20, 25, 30, 35, 40, 45 and 50. For each study area, species and sample size, presences were drawn from the defined presence cells. Sampling probability was equal to the defined habitat suitability score, reflecting higher abundance and therefore higher detectability of species in areas with optimal environmental values (Lomolino et al. 2010). In our virtual study area, where the optimum of every species was (0, 0), we created six species (one for each of 6 prevalence classes). These species were sampled, with each sample size replicated 100 times (6 prevalence classes, 24 sample sizes, 100 replications each), summing to a total of 14 400 species samples from the virtual study area (Fig. 1d). For the African study area, where each species has a unique optimum, for each of the prevalence classes the species definitions were replicated 100 times, resulting in 600 species (6 prevalence classes, 100 replications each). Subsequently, we sampled each species for the 24 different sample sizes (Fig. 1i), also resulting in 14 400 species samples from the African study area.

Species distribution modelling

All SDMs were developed with MaxEnt (Phillips et al. 2006). MaxEnt estimates the species potential geographic distribution by finding the distribution of maximum entropy (closest to uniform) subject to the constraint that the expected value of each feature under this estimated distribution matches its empirical average. MaxEnt was developed to use presence-only data and has shown to outperform other algorithms, including when applied to small data sets (Elith et al. 2006, Hernandez et al. 2006, Aguirre-Gutiérrez et al. 2013). Default MaxEnt settings were adjusted to include linear and quadratic features for all sample sizes, while hinge, product and threshold features were excluded to prevent over-parameterization of the models (Merow et al. 2013). Restricting MaxEnt to only use linear features for sample sizes smaller than 10, disables MaxEnt to fit a model on data that demonstrate other responses such as the bivariate normal response of our simulated species. This illustrates the need to adjust default MaxEnt settings based on biologically motivated modelling decisions (Merow et al. 2013). All above mentioned samples were subsequently modelled resulting in 28 800 SDMs (Fig. 1e, j, for details see Supplementary material Appendix 6 and 7).

Testing model accuracy

For each SDM, we calculated the real AUC value (real AUC) by cross-validating the predicted MaxEnt habitat suitability scores with our defined presences and absences using the 'evaluate' function of the R-library 'dismo' (Hijmans et al. 2013). Due to computational limitations of R, the real AUC for the African study area was calculated using a 10% random subsample of the presences and absences. Second, for each SDM we obtained the internal AUC value calculated by MaxEnt (MaxEnt AUC), which is based on the predicted habitat suitability scores of the sampled presences

and background sites. Third, for each sample size, null models were generated by randomly selecting the same number of background sites as sample size records from the entire study area, replicated 99 times. These 99 sets of random points were treated as presences and modelled similarly as the species, resulting in 99 AUC values of MaxEnt models based on randomly drawn points for each sample size. The species' SDM is regarded significantly better than random expectation if its AUC value exceeds rank number 95, when ranked with the 99 null model AUC values, corresponding to a one-sided significance level of 0.05 (Raes and ter Steege 2007). Both the real AUC and MaxEnt AUC values of each SDM were tested against their corresponding 95th AUC value of the null-distribution (Supplementary material Appendix 6 and 7). Fourth, we calculated the Spearman rank correlation rho values between the defined and predicted habitat suitability based on a cell-by-cell comparison. Finally, for reasons of comparison, we included an analysis of Schoener's *D* and Hellinger distance *I* (Schoener 1970, Warren et al. 2008). Here, high overlap indicates that SDMs produce an accurate prediction of our defined species distribution. Both *D* and *I* were calculated with the 'niche.overlap' function of the R-library 'phyloclim' (Heibl and Calenge 2013) applied to the defined and predicted habitat suitability for each sample.

Identifying the required minimum number of records

We used the upper 95% range values from the 100 replications for each combination of prevalence class and sample size of the real AUC, MaxEnt AUC, real AUC rank, MaxEnt AUC rank, Spearman rank correlation, Schoener's *D* and Hellinger distance *I* values. This effectively excludes the 5% worst performing SDMs. To mask out stochastic effects, the lower and upper limits of this upper 95% range values were smoothed by applying the 'loess' function of the R-library 'stats' (R Core Team) with default settings.

To identify for each prevalence class the minimum number of records that is required to accurately model a species' distribution, we evaluated model performance using three decision rules applied to the smoothed lower range limit values for both study areas separately. First, for each prevalence class, we identified the sample size for which the lower range limit of the SDM's real AUC values exceeds 0.9. An AUC value of 0.9 is commonly used as indicative for a 'very good model' performance (Pearce and Ferrier 2000, Manel et al. 2001), although the original author did not explicitly state so (Swets 1988). Second, for each prevalence class, we identified the sample size for which the lower range limit of the SDM's real AUC rank values exceeds 95, corresponding to a performance significantly better than random expectation based on a significance level of $p < 0.05$. Finally, for each prevalence class, we identified the sample size for which the lower range limit of the SDM's Spearman rank correlation values exceeds 0.9, indicating strong correlation between defined and modelled species distributions. The behavior of the MaxEnt AUC, Schoener's *D*, and Hellinger distance *I* values as a function of sample size and species' prevalence are discussed in the context of identifying the required minimum sample sizes.

Data available from the Dryad Digital Repository: <http://dx.doi.org/10.5061/dryad.8sb8v> (van Proosdij et al. 2015).

Results

For both the virtual and the real African study area and for each prevalence class, model performance defined as the lower range limit of the upper 95% range values of the real AUC, MaxEnt AUC, real AUC rank, MaxEnt AUC rank, Spearman rank correlation, Schoener's *D*, and Hellinger distance *I* increased with increasing sample sizes (Fig. 2–5). As expected, the mean and maximum MaxEnt AUC values of our simulated species decreased with increasing sample size, which is in line with the observations of others using real species (Raes and ter Steege 2007, Merckx et al. 2011).

Our results show a strong effect of species' prevalence on model performance: SDMs for species with a small prevalence perform better than SDMs for species with a large prevalence when using the same number of records to train the model (Fig. 2–4), which is in line with results reported by others (Manel et al. 2001, McPherson and Jetz 2007, Mateo et al. 2010, Lobo and Tognelli 2011). In contrast, values for Schoener's *D* and Hellinger distance *I* increase with increasing prevalence using the same number of records (Fig. 5).

Figure 2 shows that the MaxEnt AUC values approach an asymptote with increasing sample sizes for each prevalence class. The value to which this asymptote converges strongly decreases with prevalence (Fig. 2). This difference in maximum possible MaxEnt AUC value underlines the importance of being cautious when using the AUC value based on presence-only data (Phillips et al. 2006, Raes and ter Steege 2007, Lobo et al. 2008, Jiménez-Valverde 2012). In our results the MaxEnt AUC values slightly exceed the expected value of $1 - a/2$, where a represents the species' prevalence (Fig. 2: red horizontal line), which is caused by our sampling strategy favoring locations with a higher given habitat suitability (Jiménez-Valverde 2012, Smith 2013).

The large spread in AUC values at small sample sizes (Fig. 2 and 3) illustrates the low accuracy of SDMs based on small sample sizes. The spread in AUC values, rank AUC values and Spearman rank correlation values decrease with decreasing prevalence and with increasing sample sizes (Fig. 2–4). The spread in observed values was largest at small sample sizes and for widespread species, which illustrates the low accuracy of SDMs based on sample sizes smaller than the required minimum number of records to obtain an accurate SDM.

The required minimum number of records or lower limits of sample size for each prevalence class were identified for both the virtual and the African study area based on our three pre-defined decision rules: the lower limit of the upper 95% range values of 1) real AUC > 0.9; 2) real AUC ranks > 95; 3) Spearman rank correlation > 0.9 (Table 1). Note that although we do show results for prevalence class 0.05 (Table 1, Fig. 2–5), we discuss minimum sample sizes for prevalence classes 0.1–0.5 only, as prevalences below 0.1 should be avoided (see Discussion). For the most ideal situation, that of our virtual study area with balanced, orthogonal gradients, we observed that for criterion 1

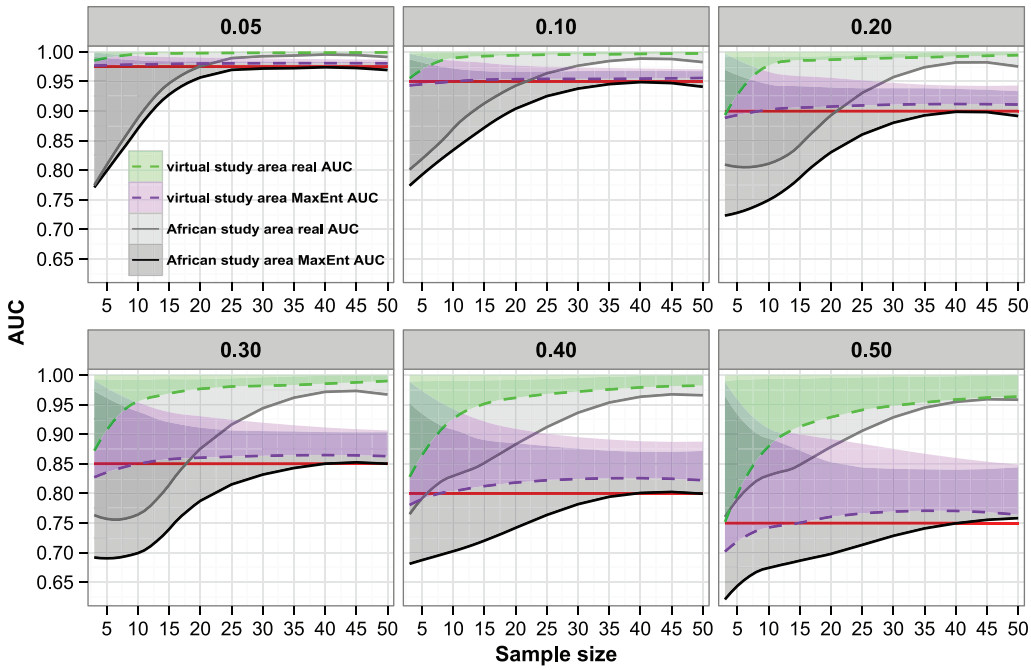

Figure 2. Model performance as a function of species' prevalence class and sample size in a virtual and in an African study area based on AUC values with separate panels for each prevalence class. Ribbons show the upper one-sided 95% range of the observed values, excluding the 5% worst performing models; darker lines show the lower range limit ; dashed lines for the virtual study area; solid lines for the African study area. Red, horizontal lines show the AUC values of $1 - a/2$ (where a is the species' prevalence).

(real AUC > 0.9, Fig. 2), the minimum sample size ranged from 3 for narrow-ranged species (prevalence class 0.1) to 13 for widespread species (prevalence class 0.5). However, when using criterion 2 (real AUC rank > 95, Fig. 3), the required minimum numbers of records were substantially lower for species of larger prevalence classes: 3 for species in prevalence class 0.1 to 8 for species in prevalence class 0.5. In contrast, based on criterion 3 (Spearman > 0.9, Fig. 4), the required minimum numbers of records were considerably higher and ranged from 10 for species of prevalence class 0.1 to 30 for

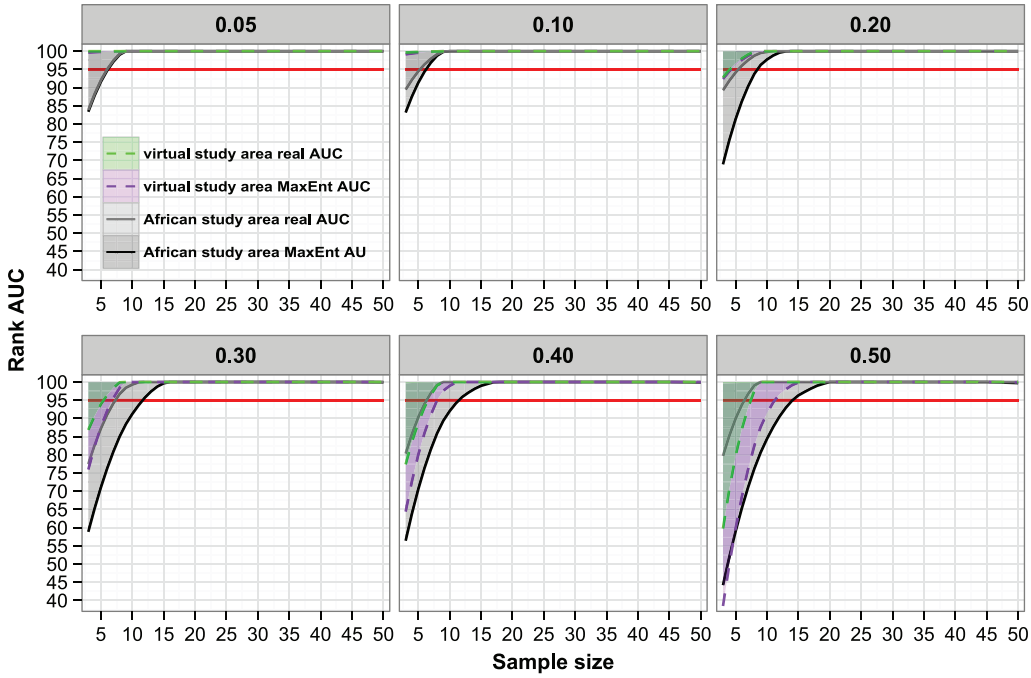

Figure 3. Model performance as a function of species' prevalence class and sample size in a virtual and in an African study area based on rank numbers of AUC values with separate panels for each prevalence class. Ribbons show the upper one-sided 95% range of the observed values, excluding the 5% worst performing models; darker lines show the lower range limit; dashed lines for the virtual study area; solid lines for the African study area; red, horizontal lines show the critical AUC rank value of 95.

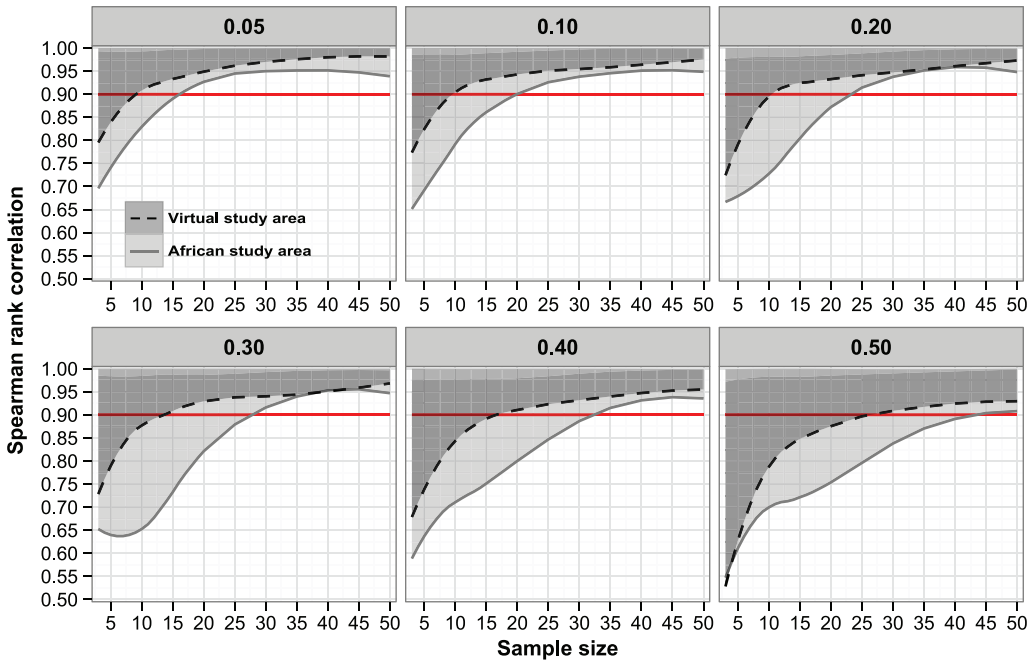

Figure 4. Model performance as a function of species' prevalence class and sample size in a virtual and in an African study area based on Spearman rank correlation values between defined and predicted habitat suitability with separate panels for each prevalence class. Ribbons show the upper one-sided 95% range of the observed values, excluding the 5% worst performing models; darker lines show the lower range limit; dashed lines for the virtual study area; solid lines for the African study area. Red, horizontal lines show the Spearman rank correlation value of 0.9.

species in prevalence class 0.5. For the African study area, the minimum sample size based on criterion 1 (real AUC > 0.9) ranged from 14 for species with a prevalence of 0.1 to 25 for species with a prevalence of 0.5. When using criterion 2 (real AUC rank > 95), minimum sample sizes ranged from 6 for species in prevalence class 0.1 to 7 for species in prevalence class 0.5 (the observed sample size of 8 for prevalence class 0.3 might be caused by a stochastic effect). Here again,

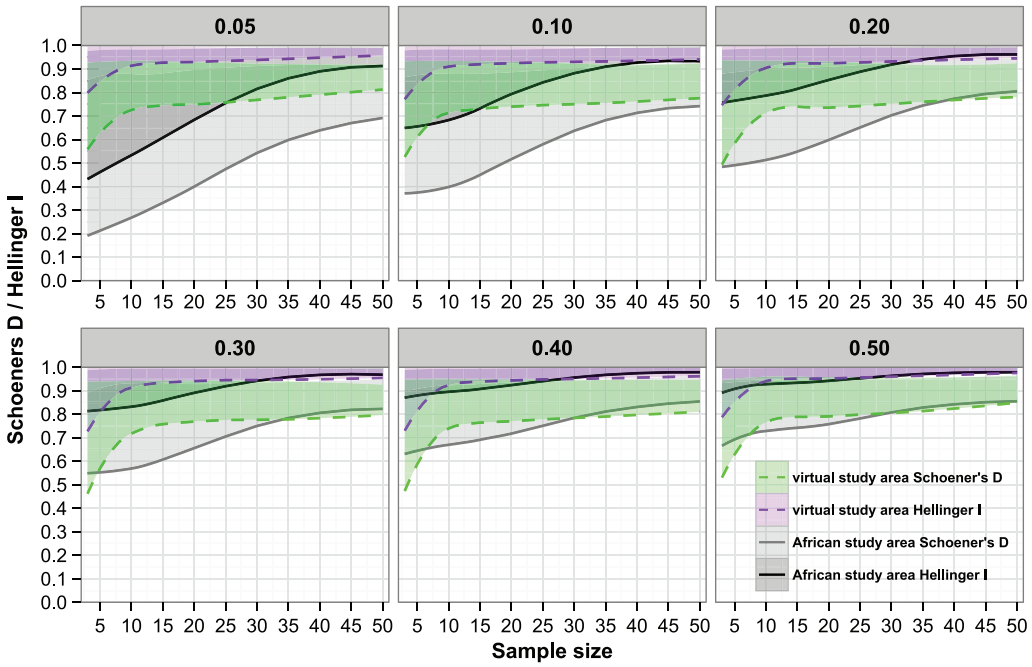

Figure 5. Model performance as a function of species' prevalence class and sample size in a virtual and in an African study area based on Schoener's *D* and Hellinger distance *I* values with separate panels for each prevalence class. Ribbons show the upper one-sided 95% range of the observed values, excluding the 5% worst performing models; darker lines show the lower range limit; dashed lines for the virtual study area; solid lines for the African study area.

Table 1. The minimum number of records required for building accurate species distribution models for a virtual study area (a) and an African study area (b) based on critical minimal values of model performance using the following indicators of model performance: real AUC, real AUC rank, MaxEnt AUC rank, and Spearman rank correlation. Minimum sample sizes are based on the lower limit of the upper one-sided 95% range of the model performance values.

Species prevalence	Real AUC > 0.90	Real AUC rank > 95	MaxEnt AUC rank > 95	Spearman *rho* > 0.90
(a) Virtual study area				
0.05	3	3	3	10
0.10	3	3	3	10
0.20	4	5	5	11
0.30	5	6	7	14
0.40	8	7	9	17
0.50	13	8	12	30
(b) African study area				
0.05	11	6	7	17
0.10	14	6	7	20
0.20	25	6	9	25
0.30	25	8	12	30
0.40	25	7	12	35
0.50	25	7	15	45

based on criterion 3 (Spearman > 0.9), minimum sample sizes were considerably higher and ranged from 20 for species with prevalence 0.1 to 45 for species in prevalence class 0.5.

Discussion

Novel method to identify the required minimum number of presence records

The novel method presented here enables users to rigorously identify the required minimum number of records to generate accurate SDMs for species of different species' prevalence classes under ideal conditions. It does so by using simulated species for which presence and absence is defined and quantifies the effect of sample size and species' prevalence on model performance. As such, this study supplies handles for anyone using SDMs to assess whether their data allows generating accurate SDMs. Our results (Table 1) corroborate two main aspects addressed by others before. First, model performance strongly depends on sample size and a small increase in the smallest sample sizes results in a large increase of model performance (Stockwell and Peterson 2002, Papeş and Gaubert 2007, Pearson et al. 2007, Loiselle et al. 2008, Wisz et al. 2008). Second, model performance decreases with increasing species' prevalence when using the same number of records to train the model (Stockwell and Peterson 2002, Hernandez et al. 2006, Mateo et al. 2010, Lobo and Tognelli 2011, Tessarolo et al. 2014).

Minimum sample sizes required for model calibration

The minimum number of records required to generate accurate SDMs differs between our virtual and real African study areas (Table 1) and can be different for other study areas. Our virtual study area represents an ideal situation with balanced, orthogonal gradients, and the required minimum sample size for each species' prevalence class should thus be regarded as a theoretical absolute lower limit. In our African study area the lower limits are considerably higher (Fig. 2–4; Table 1), as a result of the non-uniform frequency distribution of environments in this study area. For both study areas, the minimum required sample size is higher for widespread than for narrow-ranged species, as in general the ecological niche of the latter is comparatively better covered by the samples. Therefore, studies applying a generalized a priori defined minimum number of records (Schmidt et al. 2005, Algar et al. 2009, Raes et al. 2009, 2013) will lead to erroneous exclusion of models for narrow-ranged species when setting the limit too high and/or erroneous acceptance of those for widespread species for which too few records are used. Where scientists are usually concerned about data-deficiency for rare, narrow-ranged species, data quantity for widespread species appears to be crucial too. Therefore, estimating the prevalence of a modelled species is essential to determine the required minimum number of records. Although typically unknown, the species' prevalence can be estimated e.g. by calculating the extent of occurrence (EOO) or area of occupancy (AOO) (IUCN 2001) or by calculating the predicted presence fraction based on an exploratory SDM to which a threshold is applied (Syfert et al. 2014).

The differences in required minimum number of records based on the indicators of model performance applied here are the result of the different nature of these indicators. At the top end of the scale of model performance are thresholds such as the real AUC value of 0.9 and a Spearman rank correlation value of 0.9, which both classify a model as 'good'. Note that a Spearman correlation test compares absolute ranks, whereas the AUC only compares the relative ranks between presence and absence. Testing an SDM against null models (rank AUC) informs us if the SDM is 'significantly better than random expectation', which is not the same as 'good'. Obviously, the desired model accuracy strongly depends on the application (Guisan and Zimmermann 2000, Peterson 2006, Jiménez-Valverde et al. 2011, Liu et al. 2011). One may choose to accept a reasonably accurate model as an indication for where a species occurs, but rely exclusively on highly accurate models when more precise predictions are needed.

Factors that increase the minimum required sample size

We stress that the minimum numbers of records listed in Table 1 increases when working with real species and real spatial data related to factors addressed here. First, real species possibly correlate to a larger and more complex set of environmental predictors, whose signal is not fully represented by the two orthogonal PCA-based variables that were used as a proxy for environmental conditions here. To test the effect of the number of included variables on the required minimum number of records, we repeated our simulations using the first three and first four PCA axes, which together explained resp. 55% and 65% of the variance instead of the 43% explained by the first two PCA axes only. These analyses gave similar threshold values for the minimum required number of records, indicating that the thresholds will not substantially change by including more environmental

variables (Supplementary material Appendix 2). Related to this and deserving future research is the question on the effect of including a variable important for the occurrence of a species that is not part of the model variables, e.g. a variable not included in global climate and soil data sets. In addition, multicollinearity of environmental predictors will have a negative effect on model accuracy (Dormann et al. 2013), which was not an issue in our study using orthogonal variables. Third, we defined all our simulated species to be in equilibrium with climate (Araújo et al. 2005), and therefore to demonstrate niche stability and complete niche-filling. The effects of time scale and biotic interactions on niche stability, as well as the effects of dispersal limitations on niche-filling have serious impact on the correctness of an SDM considered to reflect a species' distribution (Nogués-Bravo 2009, Saupe et al. 2012). Fourth, natural history specimens are commonly subject to a collecting bias (Reddy and Davalos 2003, ter Steege et al. 2011), often representing a geographical bias (Kadmon et al. 2004, Loiselle et al. 2008). When a collecting bias translates into an environmental bias – an unbalanced or partial coverage of the ecological niche – models show falsely inflated AUC values, but are actually performing worse compared to models based on unbiased data sets (Merckx et al. 2011, Bean et al. 2012, Raes 2012). Consequently, the minimum required number of records increases (Feeley and Silman 2011). Finally, other factors with a negative impact on model performance based on real data include misidentifications, incorrect georeferencing (Graham et al. 2008, Moudry and Simova 2012), and uncertainty in environmental variables and accuracy of climate data (Hijmans et al. 2005). Although these aspects warrant future research, we feel safe to ignore them here, when specifically addressing the questions of our current study working with simulated species under optimal orthogonal bivariate environmental conditions.

AUC values based on presence-only data

Our results show that MaxEnt AUC values based on sampled presence vs. background data, differ from real AUC values based on true presence vs. absence data (Fig. 2). This supports previous critics on the use of AUC as indicator of model performance without further analysis (Raes and ter Steege 2007, Lobo et al. 2008, Jiménez-Valverde 2012). Applying a generalized AUC threshold value – commonly set at 0.7 – results in the erroneous acceptance of SDMs based on small sample sizes as these have inflated MaxEnt AUC values and dismissal of good SDMs for widespread species that theoretically can never reach an AUC of 0.7. The aspect of the unknown maximum value of the MaxEnt AUC due to its dependence on the unknown real species' prevalence disqualifies the MaxEnt AUC as a reliable indicator of model accuracy if treated without further evaluation such as i.e. a null model test (Raes and ter Steege 2007). The exceptionally high values of both real AUC and MaxEnt AUC for models based on small sample sizes of narrow-ranged species (Fig. 2) should be treated with caution. Species with a prevalence of 0.05 show a strong imbalance between true presences and absences: 5% presences vs 95% absences. The chance that a random presence has a higher probability of occurrence than a random absence for such species is high, resulting in high AUC values. These statistical artefacts inflate AUC values for SDMs of species with a prevalence below 0.1 (McPherson et al. 2004, McPherson and Jetz 2007, Lobo and Tognelli 2011), which is confirmed by our results. It is therefore recommended to choose the study area proportionally to the presence area of the assessed species, so that a species' prevalence between 0.1 and 0.9 is achieved (McPherson et al. 2004). Note that the species' dispersal capacity should be leading in defining the width of the border around the presence localities (Barve et al. 2011). Unfortunately, the true distribution and prevalence of a species is usually not known – after all, that is what we are trying to assess. In the current study, we assessed the reliability of an alternative: to evaluate SDMs using the MaxEnt AUC in a comparative way by testing it against a null model. For the virtual study area, our results show a strong congruence between the behavior of the real AUC and the MaxEnt AUC rank for species of all prevalence classes as well as required minimum sample sizes based on them (Fig. 2 and 3; Table 1). In contrast, for the real African study area, the behavior of these indicators of model performance differs and the minimum sample sizes required to obtain an accurate SDM based on the lower limit of the upper 95% range values of real AUC $>$ 0.9 are on average twice as high as those based on the lower range limit of MaxEnt AUC rank values $>$ 95. This difference is the result of the different nature of these indicators of model performance as addressed above ('good' vs 'significantly better than random expectation'). We conclude that the use of a null model test (Raes and ter Steege 2007) is an appropriate method to evaluate SDMs using the MaxEnt AUC value in a comparative way, provided that the nature of this indicator is respected: to identify SDMs that perform 'significantly better than random expectation'.

Aspects of modelling narrow-ranged species

The difficulties with accurate modelling of species with a narrow ecological niche are illustrated in our results by the very low Schoener's *D* and Hellinger distance *I* values for narrow-ranged species in the African study area, indicating that SDMs for narrow-ranged species perform worse than those of widespread species when using the same sample size (Fig. 5). In contrast, AUC values for narrow-ranged species are high, although a large spread in values is visible for smaller sample sizes (Fig. 2). This contrast between low *D* and *I* values and high AUC values can be explained by the large number of absences of which many are correctly predicted as absences (high specificity), but small numbers of presences of which only few are predicted correctly as presences (low sensitivity). The above-mentioned statistical artefact for species with a prevalence below 0.1 should be noted too. Other explanations for the high AUC values and AUC rank values of these SDMs could be spatial autocorrelation (Merckx et al. 2011) and collecting bias (Phillips et al. 2009) in the training samples, although the latter only applies to most situations when working with datasets of real species. Null models are based on randomly sampled records from the entire study area, with less or no spatial autocorrelation and no sampling bias. Consequently, a random null-model test results in a too optimistic acceptance of SDMs based on few samples. To counter the effect of collecting bias in real datasets, we recommend evaluating SDMs by using

bias-corrected null-models (Raes and ter Steege 2007) based on target-group background sampling, which has been shown to be effective (Phillips et al. 2009).

Conclusions

We conclude that applying a lower limit to the sample size used in SDMs is essential for generating accurate SDMs. The lower limit strongly depends on the species' prevalence and the specific features of the targeted study area. The required minimum numbers of records for species of different prevalence classes based on analyses in our virtual study area apply only to an ideal, balanced, orthogonal world. These numbers strongly increase for an irregular real study area like our African study area. MaxEnt AUC values cannot be used for model evaluation as such, but testing these against random or bias-corrected null models provides a reliable alternative method. Generating and evaluating SDMs for narrow-ranged species, those with a prevalence below 0.1, is difficult and should be avoided by selecting a study area proportionally to the species' presence area and with respect to the species' dispersal capacity.

The novel method presented here is applicable to any taxonomic clade or other group, study area and past, current or future climate scenario. The R-scripts with detailed stepwise methodology and a brief manual on how to apply these scripts to given data are provided in the Supplementary material Appendix 3–9. We advocate the use of our method as a routine procedure prior (or in retrospect) to any SDM study. This will aid in verifying if required levels of data quantity and quality are met and will improve the reliability of SDMs as well as the results of all future studies involving SDMs.

Acknowledgements – We thank the subject editor for his valuable comments on the manuscript. NR was supported by NWO-ALW grant 819.01.014. All authors declare to have no conflict of interest.

References

Aguirre-Gutiérrez, J. et al. 2013. Fit-for-purpose: species distribution model performance depends on evaluation criteria – dutch hoverflies as a case study. – PLoS One 8: e63708.

Aguirre-Gutiérrez, J. et al. 2014. Similar but not equivalent: ecological niche comparison across closely-related Mexican white pines. – Divers. Distrib. doi: 10.1111/ddi.12268

Algar, A. C. et al. 2009. Predicting the future of species diversity: macroecological theory, climate change, and direct tests of alternative forecasting methods. – Ecography 32: 22–33.

Araujo, M. B. and Peterson, A. T. 2012. Uses and misuses of bioclimatic envelope modeling. – Ecology 93: 1527–1539.

Araújo, M. B. et al. 2005. Validation of species–climate impact models under climate change. – Global Change Biol. 11: 1504–1513.

Austin, M. P. et al. 2006. Evaluation of statistical models used for predicting plant species distributions: role of artificial data and theory. – Ecol. Model. 199: 197–216.

Barve, N. et al. 2011. The crucial role of the accessible area in ecological niche modeling and species distribution modeling. – Ecol. Model. 222: 1810–1819.

Bean, W. T. et al. 2012. The effects of small sample size and sample bias on threshold selection and accuracy assessment of species distribution models. – Ecography 35: 250–258.

Boucher-Lalonde, V. et al. 2012. How are tree species distributed in climatic space? A simple and general pattern. – Global Ecol. Biogeogr. 21: 1157–1166.

Boucher-Lalonde, V. et al. 2014. A consistent occupancy–climate relationship across birds and mammals of the Americas. – Oikos 123: 1029–1036.

Broennimann, O. et al. 2011. Measuring ecological niche overlap from occurrence and spatial environmental data. – Global Ecol. Biogeog. 21: 481–497.

Costello, M. J. et al. 2013. Can we name Earth's species before they go extinct? – Science 339: 413–416.

Dormann, C. F. et al. 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. – Ecography 36: 27–46.

Duan, R.-Y. et al. 2015. SDMvspecies: a software for creating virtual species for species distribution modelling. – Ecography 38: 108–110.

Elith, J. et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. – Ecography 29: 129–151.

FAO/IIASA/ISRIC/ISSCAS/JRC 2012. Harmonized World Soil Database (version 1.2). – <http://webarchive.iiasa.ac.at/Research/LUC/External-World-soil-database/HTML/index.html?sb=1>.

Feeley, K. J. and Silman, M. R. 2011. Keep collecting: accurate species distribution modelling requires more collections than previously thought. – Divers. Distrib. 17: 1132–1140.

Fielding, A. H. and Bell, J. F. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. – Environ. Conserv. 24: 38–49.

Gentz, A. et al. 2014. mvtnorm: multivariate normal and t distributions. – R package ver. 0.9-9997, <http://CRAN.R-project.org/package=mvtnorm>.

Graham, C. H. et al. 2008. The influence of spatial errors in species occurrence data used in distribution models. – J. Appl. Ecol. 45: 239–247.

Guisan, A. and Zimmermann, N. E. 2000. Predictive habitat distribution models in ecology. – Ecol. Model. 135: 147–186.

Heibl, C. and Calenge, C. 2013. phyloclim: integrating phylogenetics and climatic niche modeling. – R package ver. 0.9-4, <http://CRAN.R-project.org/package=phyloclim>.

Hernandez, P. A. et al. 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. – Ecography 29: 773–785.

Hijmans, R. J. et al. 2005. Very high resolution interpolated climate surfaces for global land areas. – Int. J. Climatol. 25: 1965–1978.

Hijmans, R. J. et al. 2013. dismo: species distribution modeling. – R package ver. 0.8-17, <http://CRAN.R-project.org/package=dismo>.

Hirzel, A. H. et al. 2001. Assessing habitat-suitability models with a virtual species. – Ecol. Model. 145: 111–121.

IUCN 2001. IUCN Red List categories and criteria, version 3.1. – IUCN Species Survival Commision, Cambridge, UK.

Jiménez-Valverde, A. 2012. Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. – Global Ecol. Biogeogr. 21: 498–507.

Jiménez-Valverde, A. et al. 2009. The effect of prevalence and its interaction with sample size on the reliability of species distribution models. – Community Ecol. 10: 196–205.

Jiménez-Valverde, A. et al. 2011. Use of niche models in invasive species risk assessments. – Biol. Invasions 13: 2785–2797.

Kadmon, R. et al. 2004. Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. – Ecol. Appl. 14: 401–413.

Liu, C. R. et al. 2011. Measuring and comparing the accuracy of species distribution models with presence–absence data. – Ecography 34: 232–243.

Lobo, J. M. and Tognelli, M. F. 2011. Exploring the effects of quantity and location of pseudo-absences and sampling biases on the performance of distribution models with limited point occurrence data. – J. Nat. Conserv. 19: 1–7.
Lobo, J. M. et al. 2008. AUC: a misleading measure of the performance of predictive distribution models. – Global Ecol. Biogeogr. 17: 145–151.
Loiselle, B. A. et al. 2008. Predicting species distributions from herbarium collections: does climate bias in collection sampling influence model outcomes? – J. Biogeogr. 35: 105–116.
Lomolino, M. V. et al. 2010. Biogeography. – Sinauer Associates.
Longino, J. T. et al. 2002. The ant fauna of a tropical rain forest: estimating species richness three different ways. – Ecology 83: 689–702.
Manel, S. et al. 2001. Evaluating presence–absence models in ecology: the need to account for prevalence. – J. Appl. Ecol. 38: 921–931.
Mateo, R. G. et al. 2010. Effects of the number of presences on reliability and stability of MARS species distribution models: the importance of regional niche variation and ecological heterogeneity. – J. Veg. Sci. 21: 908–922.
McPherson, J. M. and Jetz, W. 2007. Effects of species' ecology on the accuracy of distribution models. – Ecography 30: 135–151.
McPherson, J. M. et al. 2004. The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? – J. Appl. Ecol. 41: 811–823.
Merckx, B. et al. 2011. Null models reveal preferential sampling, spatial autocorrelation and overfitting in habitat suitability modelling. – Ecol. Model. 222: 588–597.
Merow, C. et al. 2013. A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. – Ecography 36: 1058–1069.
Metz, C. E. 1978. Basic principles of ROC analysis. – Sem. Nucl. Med. VIII: 283–298.
Meynard, C. N. and Quinn, J. F. 2007. Predicting species distributions: a critical comparison of the most common statistical models using artificial species. – J. Biogeogr. 34: 1455–1469.
Miller, J. A. 2014. Virtual species distribution models: using simulated data to evaluate aspects of model performance. – Prog. Phys. Geogr. 38: 117–128.
Mora, C. et al. 2011. How many species are there on Earth and in the ocean? – PLoS Biol. 9: e1001127.
Moudry, V. and Simova, P. 2012. Influence of positional accuracy, sample size and scale on modelling species distributions: a review. – Int. J. Geogr. Inform. Sci. 26: 2083–2095.
Nogués-Bravo, D. 2009. Predicting the past distribution of species climatic niches. – Global Ecol. Biogeogr. 18: 521–531.
Papeş, M. and 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. – Divers. Distrib. 13: 890–902.
Pearce, J. and Ferrier, S. 2000. Evaluating the predictive performance of habitat models developed using logistic regression. – Ecol. Model. 133: 225–245.
Pearson, R. G. et al. 2007. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. – J. Biogeogr. 34: 102–117.
Peterson, A. T. 2006. Uses and requirements of ecological niche models and related distributional models. – Biodivers. Inform. 3: 59–72.
Phillips, S. J. et al. 2006. Maximum entropy modeling of species geographic distributions. – Ecol. Model. 190: 231–259.
Phillips, S. J. et al. 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. – Ecol. Appl. 19: 181–197.
Rabinowitz, D. 1981. Seven forms of rarity. – In: Synge, H. (ed.), The biological aspects of rare plants conservation. Wiley, pp. 205–217.
Raes, N. 2012. Partial versus full species distribution models. – Nat. Conserv. 10: 127–138.
Raes, N. and ter Steege, H. 2007. A null-model for significance testing of presence-only species distribution models. – Ecography 30: 727–736.
Raes, N. et al. 2009. Botanical richness and endemicity patterns of Borneo derived from species distribution models. – Ecography 32: 180–192.
Raes, N. et al. 2013. Legume diversity as indicator for botanical diversity on Sundaland, south east Asia. – S. Afr. J. Bot. 89: 265–272.
Reddy, S. and Davalos, L. M. 2003. Geographical sampling bias and its implications for conservation priorities in Africa. – J. Biogeogr. 30: 1719–1727.
Rödder, D. and Engler, J. O. 2011. Quantitative metrics of overlaps in Grinnellian niches: advances and possible drawbacks. – Global Ecol. Biogeogr. 20: 915–927.
Sánchez-Fernández, D. et al. 2011. Species distribution models that do not incorporate global data misrepresent potential distributions: a case study using Iberian diving beetles. – Divers. Distrib. 17: 163–171.
Saupe, E. E. et al. 2012. Variation in niche and distribution model performance: the need for a priori assessment of key causal factors. – Ecol. Model. 237: 11–22.
Schmidt, M. et al. 2005. Herbarium collections and field data-based plant diversity maps for Burkina Faso. – Divers. Distrib. 11: 509–516.
Schoener, T. W. 1970. Nonsynchronous spatial overlap of lizards in patchy habitats. – Ecology 51: 408–418.
Smith, A. B. 2013. On evaluating species distribution models with random background sites in place of absences when test presences disproportionately sample suitable habitat. – Divers. Distrib. 19: 867–872.
Stockwell, D. R. B. and Peterson, A. T. 2002. Effects of sample size on accuracy of species distribution models. – Ecol. Model. 148: 1–13.
Swets, J. A. 1988. Measuring the accuracy of diagnostic systems. – Science 240: 1285–1293.
Syfert, M. M. et al. 2014. Using species distribution models to inform IUCN Red List assessments. – Biol. Conserv. 177: 174–184.
ter Steege, H. et al. 2011. A model of botanical collectors' behavior in the field: never the same species twice. – Am. J. Bot. 98: 31–37.
ter Steege, H. et al. 2013. Hyperdominance in the Amazonian tree flora. – Science 342 doi: 10.1126/science.1243092
Tessarolo, G. et al. 2014. Uncertainty associated with survey design in species distribution models. – Divers. Distrib. doi: 10.1111/ddi.12236
van Proosdij, A. S. J. et al. 2015. Data from: Minimum required number of specimen records to develop accurate species distribution models. – Dryad Digital Repository, <http://dx.doi.org/10.5061/dryad.8sb8v>.
Varela, S. et al. 2014. Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. – Ecography 37: 1084–1091.
Warren, D. L. et al. 2008. Environmental niche equivalency versus conservatism: quantitative approaches to niche evolution. – Evolution 62: 2868–2883.
Whittaker, R. J. et al. 2005. Conservation biogeography: assessment and prospect. – Divers. Distrib. 11: 3–23.
Wisz, M. S. et al. 2008. Effects of sample size on the performance of species distribution models. – Divers. Distrib. 14: 763–773.
Zurell, D. et al. 2010. The virtual ecologist approach: simulating data and observers. – Oikos 119: 622–635.

Supplementary material (Appendix ECOG-01509 at <www.ecography.org/appendix/ecog-01509>). Appendix 1–9.