Natureza & Conservação
Brazilian Journal of Nature Conservation

Essays & Perspectives

Natureza & Conservação 10(2):127-138, December 2012

Handling Editor: Paulo de Marco Jr
http://dx.doi.org/10.4322/natcon.2012.020

Partial versus Full Species Distribution Models

Niels Raes[1,2]*

[1] *Naturalis Biodiversity Center, Leiden, the Netherlands*
[2] *Leiden University, Section National Herbarium of the Netherlands, Leiden, the Netherlands*

Abstract

In this essay I assess the impact of generating species distribution models (SDMs), also known as ecological niche models (ENMs), within artificial geographical or political boundaries by comparing them with SDMs that use the complete distribution of species. I illustrate the differences between the paired SDMs on the plant genus *Inga* modelled within the political boundaries of Brazil (Partial SDM) compared to SDMs developed for the entire Neotropical humid tropics biome (Full SDM). Partial SDMs portray range contractions, or under-prediction, at the artificial boundaries and have different patterns of predicted presence and absence. It is therefore advisable that SDMs use presence data from the complete distribution ranges of species. Furthermore, it should be kept in mind that any SDM essentially has a partial extent in space and time.

Key words: Ecological Niche Modelling, Species Distribution Modelling, *Inga*, Neotropics, Partial Geographic Ranges.

Setting the Scene

The wide use of species distribution models (SDMs), was boosted by the seminal review paper of Guisan & Zimmermann (2000) on 'Predictive habitat distribution models in ecology' and has since grown explosively (Cayuela *et al.* 2009; Lobo *et al.* 2010). To date, this has resulted in two textbooks on the principles and applications of SDMs by Franklin (2009) and Peterson *et al.* (2011), and in numerous review and perspectives papers. The popularity can be ascribed to the application of SDMs in the fields of species discovery (Raxworthy *et al.* 2003), mapping biodiversity (Raes *et al.* 2009; van Welzen *et al.* 2011), conservation planning (Zhang *et al.* 2012), climate change effects (Hsu *et al.* 2011), species' invasions (Broennimann & Guisan 2008), evolution of niches (Yesson & Culham 2006; Evans *et al.* 2009), to list but a few (see Araújo & Peterson (2012) for an extensive list).

SDMs identify correlations between aspects of abiotic conditions and known occurrences of species across 'landscapes of interest' to define sets of conditions under which species are likely to be able to maintain viable populations (Araújo & Peterson 2012). This essay focuses on the impact of the extent of the 'landscapes of interest' on predicted distributions of species, for which I provide a worked out example. The focus lies on over- and under-prediction of SDMs fitted on an artificially constrained geographic space (i.e. political boundaries) compared to SDMs fitted on the total range of occurrence (sensu Maiorano *et al.*(2012) for time slices). To my knowledge this territory is largely unexplored (except Barbet-Massin *et al.* 2010; Sánchez-Fernández *et al.* 2011; and conceptually by Godsoe 2012).

Before getting into the subject of 'landscapes of interest', it is important to clarify the differences in the definitions of the terms: 'Bioclimatic envelope models', 'Ecological niche models (ENMs)', 'Habitat suitability models (HSMs)' and 'Species distribution models (SDMs)', as proposed by Araújo and Peterson (2012). All these terms are being used alternately, and not always in the correct context. Bioclimatic envelope models estimate the "multivariate space of climatic variables (the envelope) best matching the observed species' distribution". Instead of simply estimating the bioclimatic envelope, ENMs "link the envelope to elements of ecological niche theory rooted in the early work of Grinnell (1917) and Hutchinson (1957)", and also in the later work of Tilman (1982). I interpret ENMs as restricting the bioclimatic envelope to variables that are meaningful to the ecological niche of the species, without inferring any geographic projection. HSMs refer to "the suitability of area for a species to occur, its habitat; as such the physical space where the species lives and the available resources it can use are emphasized". This is a rather broad definition. Lastly, SDMs "characterize the multivariate environmental space delimiting species' distributions, and project this subset of environmental space back onto geography". SDMs directly build on Hutchinson's duality,

***Send correspondence to:** Niels Raes
Naturalis Biodiversity Center, Leiden, the Netherlands
E-mail: niels.raes@naturalis.nl

or the reciprocal correspondence between ecological niche space and geographic space. It should be noted, however, that any defined ecological niche space derived from the observed distribution of species in geographical space is, at best a realized niche; unless demonstrated otherwise (Colwell & Rangel 2009). The full extent of a species' fundamental niche cannot be revealed by the environmental conditions at observed collection localities. Estimation of the fundamental niche can only be achieved by experimental studies and physiological models (Colwell & Rangel 2009). This limitation should be kept in mind while interpreting any correlative model derived from observed collection localities and the abiotic conditions at those localities. Here, I prefer to use the term SDM because this unifies the niche concept with its geographical projection.

Question is: what do SDMs model or estimate? The presence of a species is determined by three factors that can be visualized by three overlapping circles, each representing a factor in the 'BAM'- framework (Soberón & Peterson 2005; Soberón 2007; Godsoe 2010). In the 'BAM'- framework, the first circle 'A' represents the geographic region with the appropriate set of abiotic conditions for the species, and may be regarded as the geographic expression of the fundamental abiotic niche; the second circle 'B' is the geographic region where the right combination with interacting species occurs, which may or may not overlap extensively with 'A'. The intersection of 'A' and 'B' represents the geographic extent of the realized niche of the species. And the third circle 'M' is a representation of the geographic region that is "accessible" to the species in some ecological sense, without barriers to movement and colonization. The intersection of the three circles is equivalent to the observed geographic distribution of the species. Given that most SDMs are fitted on a set of abiotic predictors, the output is an approximation of the realized abiotic niche (Colwell & Rangel 2009). Because dispersal limitation is (mostly) not taken into account when plotting the realized abiotic niche in its reciprocal geographic space, the result is the geographic representation of a species' potential distribution within the 'landscape of interest'. The degree to which the three factors overlap determines to what extent the observed geographic distribution is estimated by the realized abiotic niche. Efforts are being made to include dispersal limitation and biotic interactions in SDMs (Boulangeat *et al*. 2012), but this requires additional high quality data on dispersal mechanisms, life history traits, and species co-occurrences which are not available for many species and regions in the world.

Furthermore, the application of SDMs builds on number of assumptions (Araújo & Peterson 2012). When the intention is to predict presence of species for other regions or time periods than the 'landscape of interest' used to fit the SDM i.e. to predict the potential invasiveness, or impacts of climate change, it is assumed that species' niches are conserved over relevant time periods, known as niche conservatism (Wiens *et al*. 2010). Wiens *et al*. (2010) define niche conservatism as the retention of niche-related ecological traits over time. They provide an extensive list with examples supporting the existence of conservatism of the fundamental niche that provides predictability across environmental dimensions and time frames using SDMs; the same was concluded by Araújo & Peterson (2012). Nonetheless, examples of rapid niche evolution have been reported (Broennimann *et al*. 2007; Pearman *et al*. 2008). Holt (2009) provides a comprehensive framework to study the evolution of the niche. Although, the provisional conclusion can be that niches are conserved, which is relevant to the reliable use of SDMs; this conclusion is of less importance to the assessment of the impact of modelling partial versus full SDMs, because the models are not projected in time nor space.

Probably the most problematic and controversial for the reliable use of SDMs is the assumption that species' distributions are in equilibrium with climate. This was shown to be incorrect for European trees which are still filling their potential distribution since the last glacial maximum, 21 kyr before present (Svenning & Skov 2004). Similarly, expansions and contractions of the Amazonian rain forest under the influence of glacial cycles have been reported (Mayle *et al*. 2000). It is therefore advised that SDMs are calibrated across the broadest spatial, environmental and/or temporal extents that are biologically and biogeographically justifiable to capture a species' niche in its broadest sense (Barve *et al*. 2011; Araújo & Peterson 2012).

Nonetheless, SDMs are often used to model the distribution of species within the artificial boundaries of countries, and even provinces (Loiselle *et al*. 2008; Pineda & Lobo 2009; Zhang *et al*. 2012; among many others), covering a subset of species' niches. Here I assess the impact of modelling species' partial niches on their predicted distributions within the artificial boundaries of the 'landscape of interest', by comparing them with their 'expected' distributions (within the artificial boundaries) derived from a full niche model that takes all available collection localities in account. For reasons of clarity; this is different from testing how well models fitted within artificial boundaries are capable of predicting a species' full extent of occurrence, known as transferability studies (Wenger & Olden 2012; Zurell *et al*. 2012).

Partial versus Full Distribution Models

There are several reasons why it is important to include as many collections as possible and not to restrict SDMs to artificial (political) boundaries. First, the subset likely does not include the full environmental variation under which a species is known to occur. Second, even within the entire range of occurrence, collection localities tend to be biased to more accessible areas which can result in environmentally biased collections (Reddy & Davalos 2003; Hortal *et al*. 2007; Schulman *et al*. 2007). The use of environmentally biased collections to fit an SDM, in turn,

might result in under predicted species' distributions, and is essentially similar to modelling a partial niche. Environmental bias is also known to occur within country boundaries, as was reported for Ecuador (Loiselle *et al.* 2008); but that this is not necessarily the case, was shown for Israel (Kadmon *et al.* 2004). Third, it is common knowledge that the majority of species is rare (Hubbell *et al.* 2008), hence represented by a few collection records in herbaria and Natural History Museums. To capture the widest possible environmental variation under which a species is known to occur, it is important to include as many geographically unique collections as possible when constructing an SDM (Beaumont *et al.* 2009; Sánchez-Fernández *et al.* 2011).

The Inga Example

To illustrate that partial SDMs predict different extents of occurrence than full SDMs I worked out an example on 36 species of the plant genus *Inga* modelled for the entire Neotropical humid tropics (hereafter HT) biome and the Brazilian subset of the HT biome. Brazil covers the central subset of the entire HT ecological space expressed on the first two axes of a PCA analysis on eight least correlated environmental variables (Figure 1; see Environmental variables section). From Figure 1 it is clear that *Inga* collections (crosses) also occur outside the Brazilian ecological envelope (light grey dots). To model the species' distributions I used the maximum entropy algorithm – MaxEnt (Phillips *et al.* 2006; Elith *et al.* 2011), because this algorithm is performing among the best in comparative tests (Elith *et al.* 2006; Graham *et al.* 2008; Wisz *et al.* 2008), and also because it was specifically developed to model with presence-only data. Although MaxEnt uses presence-only data, it still needs to compare the predicted occurrence distribution against a background- or pseudo-absence sample. To prevent over-fitting of models in relation to the extent of the geographical background from where the pseudo-absences are drawn (Lobo *et al.* 2008; VanDerWal *et al.* 2009; Acevedo *et al.* 2012), I restricted the study area to the HT biome as defined by WWF (Figure 1b – all grey areas; Olson *et al.* 2001).

First, I developed 49 *Inga* SDMs for both the entire HT biome and the Brazilian subset. After testing all SDMs for significant deviation from random expectation (Raes & ter Steege 2007), the SDMs for 36 species pairs were retained. Secondly, I thresholded the maps to convert the continuous MaxEnt predictions to discrete presence-absence maps. Thirdly, I clipped the Brazilian extent from the HT biome SDMs, resulting in pairs of presence-absence maps both covering the Brazilian extent; one generated within the artificial political boundaries of Brazil, and one generated for the HT biome and clipped to the Brazilian extent. Finally, I assessed map similarities between the 36 paired maps using the kappa statistic (Visser & De Nijs 2006), AUC values, fraction correct prediction, and percentage difference in predicted extent. By subtracting the Brazilian maps from their paired clipped HT maps, I was able to identify regions with the highest dissimilarities in both geographical and environmental space.

Inga collection data

I selected the genus *Inga* for the following reasons; a) the genus was monographed in 1997 (Pennington *et al.* 1997), b) has a distribution largely restricted to the HT

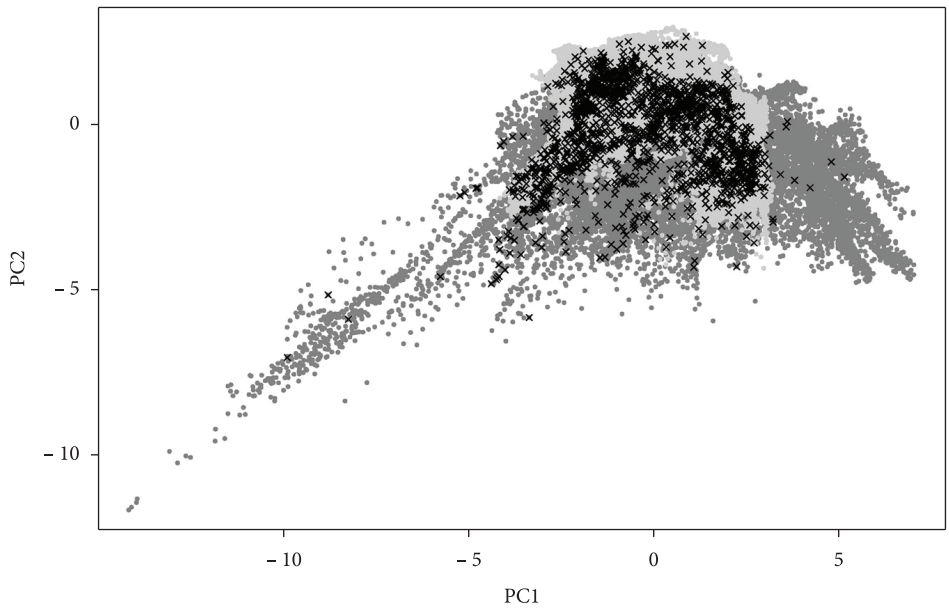

Figure 1. Ecological space, plotted on the first two principal components derived from 8 selected and standardized bioclimatic variables, of the HT biome (dark grey dots; Figure 2b), the Brazilian subset (light grey dots; Figure 2a), and *Inga* collections of the 36 species used in the analysis (black crosses).

biome (Richardson *et al*. 2001), and c) I could make use of Pennington's *Inga* occurrences dataset containing 9,379 collection records. Additionally, I downloaded all *Inga* records from SpeciesLink (2012) containing 5,842 records. The two datasets were merged and cleaned with GoogleRefine, and all unique species records per raster cell occurring in the HT biome were retained. From this dataset I selected all records of *Inga* species which were represented by at least 5 records in Brazilian subset of the HT biome, and with a maximum of 75% of their records within the political boundaries of Brazil. The latter assures that partial SDMs are modelled when they are restricted to the Brazilian subset. This procedure resulted in 3,607 unique collections covering 49 *Inga* species. After significance testing of the SDMs (see below) the SDMs of 36 *Inga* species were retained which were represented by 3,005 unique *Inga* collections.

Environmental variables

Although edaphic conditions can be very important to the definition of a species' fundamental niche (Tuomisto 2006; Bertrand *et al*. 2012), most of the variation in the geographic distribution of species relates to climate (Lalonde *et al*. 2012). Therefore, I downloaded the 19 bioclimatic variables, plus altitude, at 5 arc-minute spatial resolution downloaded from the Worldclim dataset (worldclim.org; Hijmans *et al*. 2005). To restrict the analysis to the broadest spatial extent that is biologically and biogeographically justifiable, I clipped the Neotropical humid tropics (HT) extent from this dataset with Manifold GIS (Manifold Ltd.).

To prevent problems with multi-collinearity and unnecessary model complexity, I tested the 20 variables for correlations with a Pearson's r correlation test after standardization (mean = 0, sd = 1) of the data. Simultaneously, I performed a principal component analysis (PCA) using the function 'dudi.pca' from the R-library 'ade4' (Dray & Dufour 2007; R Development Core Team 2012). From clusters of correlated variables (Pearson's $r > 0.7$) I retained one variable with the highest eigenvalue on one of the first two PCA axes. This resulted in an environmental dataset of eight selected variables for the entire HT biome covering 114,904 raster cells (Figure 2b – all grey areas; Table 1 – bottom triangle). To visualize the HT biome in ecological space I plotted the

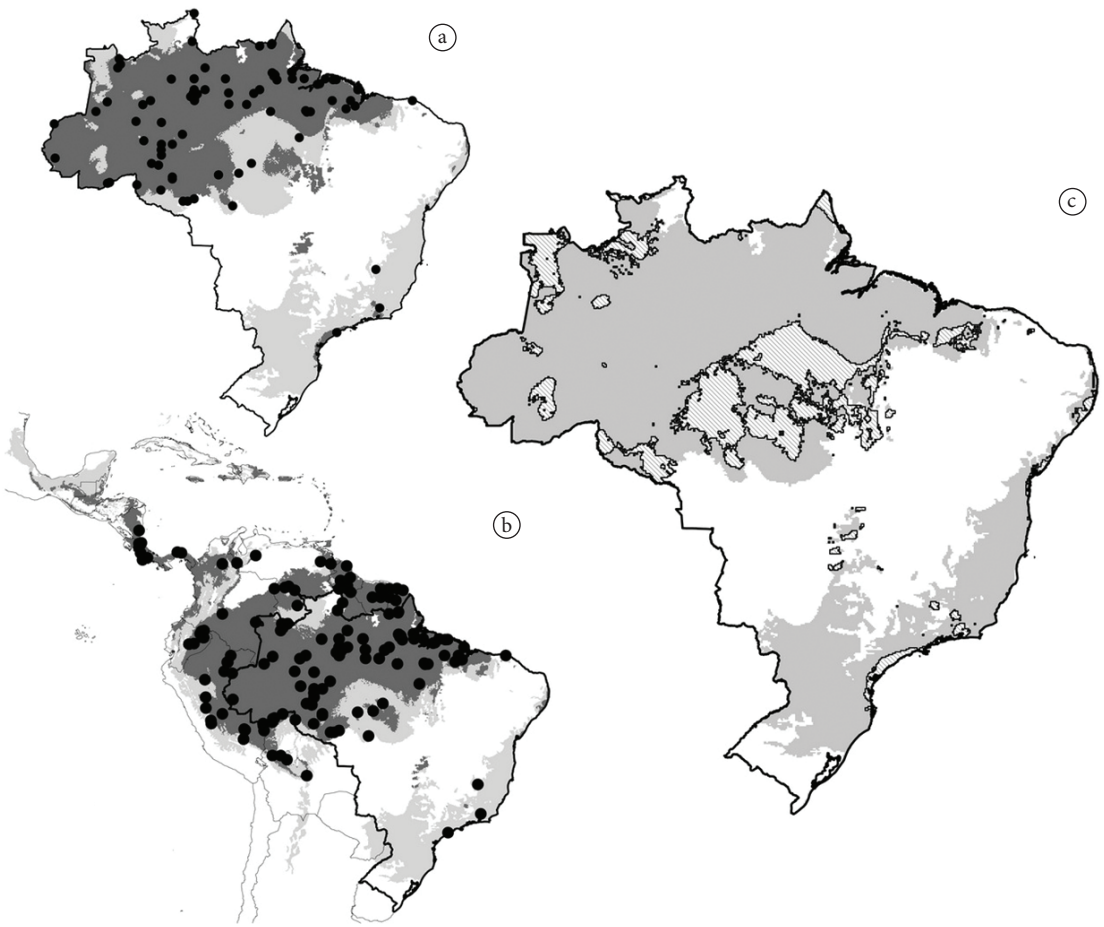

Figure 2. Map a) shows the partial SDM (dark grey = present/light grey = absent) for *Inga alba* modelled within the political boundaries of Brazil. Black points indicate collection localities. Map b) shows the full SDM (dark grey = present/light grey = absent) for *Inga alba* modelled for the entire Neotropical humid tropics biome; and map c) shows the dissimilarity between both predictions (hatched areas) for the Brazilian subset of the Neotropical humid tropics biome (all maps in geographic projection).

Table 1. Pearson's r correlation for the eight standardized bioclim variables used by the SDMs.

	bio02	bio03	bio05	bio06	bio12	bio17	bio18	bio19
bio02		–0.409	0.206	–0.593	–0.320	–0.494	0.181	–0.551
bio03	–0.337		0.390	***0.854***	0.600	0.323	–0.153	0.594
bio05	–0.068	–0.008		0.591	0.305	–0.329	–0.353	0.158
bio06	–0.648	0.563	0.692		0.568	0.187	–0.356	0.609
bio12	–0.399	0.480	0.244	0.534		0.591	0.263	0.560
bio17	–0.437	0.408	–0.081	0.281	***0.705***		0.447	0.409
bio18	0.001	0.030	–0.145	–0.099	0.488	0.559		–0.282
bio19	–0.491	0.526	0.183	0.549	0.675	0.496	–0.040	

The bottom triangle (grey cells) represents the Neotropical humid tropics biome (Figure 1b) and the top triangle the Brazilian subset (Figure 1a). Highest values printed in bold. *bio02 = Mean diurnal range* (*Mean of monthly* (*max temp – min temp*)); *bio03 = Isothermality; bio5 = Maximumtemperature of warmest month; bio06 = Minimumtemperature of coldest month; bio12 = Annual precipitation; bio17 = Precipitation of driest quarter; bio18 = Precipitation of warmest quarter; bio19 = Precipitation of coldest quarter.*

raster cells on the first two principal component (PC) axes of a PCA on the eight selected variables (Figure 1). PC1 and PC2 explain 46% and 21%, respectively, of the variance in the eight selected variables.

Since my intention is to assess whether a partial SDM results in the same predicted distribution as the full SDM, I clipped the Brazilian subset from the entire HT biome dataset. This resulted in the second environmental dataset covering the Brazilian extent (64,464 raster cells, or 56%) of the HT biome (Figure 2a – all grey areas). The Pearson's r test for the Brazilian subset indicated that bio03 and bio06 had a correlation of 0.854 (Table 1 – top triangle; caption gives the definition of the variables). For reasons of consistency I retained all eight variables in the Brazilian subset. To visualize the Brazilian subset in ecological space I plotted the Brazilian raster cells over the HT raster cells in the PCA graph (Figure 1; light grey dots). Crosses in Figure 1 represent the *Inga* collection localities in ecological space.

Species Distribution Models (SDMs) and significance testing with a null-model

SDMs were generated for all 49 *Inga* species on datasets of both the partial- and full HT biome. The AUC values (Fielding & Bell 1997) of all 98 SDMs were tested for significant deviation from random expectation with a null-model (Olden *et al.* 2002; Gotelli & McGill 2006; Raes & ter Steege 2007). I recognize that the AUC value as measure of model accuracy, when applied to presence-only data has flaws, caused by the fact that the maximum achievable AUC value is no longer 1, but 1-*a*/2; where *a* stands for the species' real distribution, which is typically not known (Phillips *et al.* 2006). However, testing the SDM AUC value against a null-distribution of AUC values, identifies those SDMs that have a correlation with one, or more, of the environmental variables that cannot be expected by random chance.

Testing against a null-model works as follows; for each number of records by which the modelled species are represented, a series of 99 times equally many records as species records is drawn randomly from the environmental dataset. These randomly drawn sets are modelled similar as the species in MaxEnt. Finally, the SDM AUC values of the *Inga* models are tested against the 95th ranked AUC values of the 99 models on sets of equally many random points as records of the *Inga* species which is tested. For example, the AUC value of a species represented by 11 records in the Brazilian subset of the HT biome is tested against the 95th ranked AUC value derived from 99 times 11 randomly drawn and modelled records from the Brazilian environmental dataset. A species' AUC value that is larger than the 95th ranked AUC value, indicates that the chance that a random set of 11 points results in an equally high AUC value is less than 5%, hence significantly better than random expectation with $p < 0.05$ (for details see Raes & ter Steege 2007). I only retained the species that had a significant SDM for both environmental datasets. This was the case for 36, or 73%, of the *Inga* species. The continuous MaxEnt SDMs were converted to discrete presence-absence maps by applying the 10% percentile training threshold, one of the more conservative thresholds methods.

Data analyses

To assess the impact of modelling species' partial distributions within artificial (political) boundaries compared to what is expected based on SDMs fitted on their full distribution I subtracted the thresholded map derived from the clipped full SDM from the thresholded partial SDM, for each of the 36 *Inga* species' paired maps. This resulted in negative values where the partial SDM predicts absence and the clipped full SDM presence, or under-prediction by the partial SDM; and in positive values where the partial SDM predicts presence and the full SDM absence, or over-prediction by the partial SDM. At the north-western border of Brazil, in the Amazonas province, distributions are under-predicted for 19 of the 36 *Inga* species (Figure 3a); and in central Brazil, in the western Pará province, there is an over-prediction for 14 species (Figure 3c) when the distributions of partial SDMs are compared to what is expected based on the full SDMs. When these values are plotted on the first two PCA axes

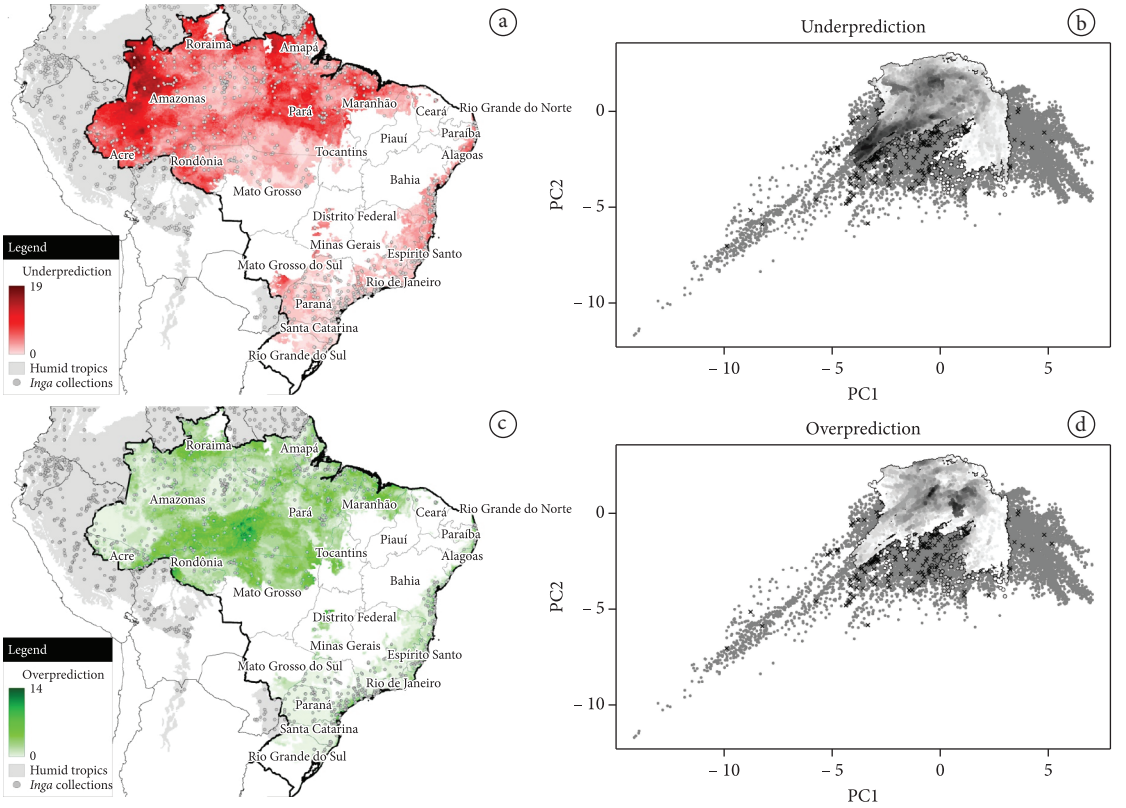

Figure 3. a) Number of partial SDMs that under-predict in Brazil when compared to the full HT biome SDMs (n = 36). Light gray area shows the extent of the HT biome; points are *Inga* collection sites; b) Raster cells representing under-predicted species presence (max. 19 – dark gray to white) in Brazil and their position in environmental space of the HT biome plotted on the first 2 PCA axes. Crosses indicate *Inga* collection sites outside Brazil; c) Number of partial SDMs that over-predict in Brazil when compared to the full HT biome SDMs (n = 36). Light gray area shows the extent of the HT biome; points are Inga collection sites, and; d) Raster cells representing over-predicted species presence (max. 14 – dark gray to white) in Brazil and their position in environmental space of the HT biome plotted on the first 2 PCA axes. Crosses indicate *Inga* collection sites outside Brazil.

the under-predicted raster cells are found in the lower left corner of the partial Brazilian ecological space (Figure 3b; dark grey color), which is close to a region where many collections are found just outside the boundary of the Brazilian ecological space (Figure 3b; crosses). Vice versa, raster cells representing over-prediction are found in the centre of the partial Brazilian ecological space (Figure 3d; dark grey color).

From the 36 *Inga* species there were 26 species with a smaller (under-)predicted presence range for partial SDMs, compared to what would be expected based on the clipped full SDMs (Table 2); and 10 species where the partial SDM over-predicted the expected presence extent (Table 2; grey bars). For these two groups separately I first assessed the similarity between the thresholded presence-absence maps of the partial - versus the full SDMs using the *Kappa* statistic implemented in the Map Comparison Kit (Visser & De Nijs 2006). The *Kappa* statistic measures the fraction of agreement, corrected for the fraction of agreement statistically expected from randomly relocating all cells in the compared maps (Hagen 2002). Both Figure 4a and Table 2 show that partial SDMs of under-predicted maps are on average approx. 60% similar to the full SDMs; and that the over-predicted maps on average have a slightly higher *Kappa* value. Secondly, I report the *Fraction correct*. The *Fraction correct* measure of map similarity is the uncorrected *Kappa* value. The average *Fraction correct* for both groups is approx. 85% (Figure 4c; Table 2). Thirdly, I assessed the difference in model accuracy based on AUC values. The AUC value for the Brazilian extent of the full SDM was calculated on the 'logistic' MaxEnt predictions clipped to the Brazilian extent. The presence localities used to calculate the AUC values were the same as the ones used for the paired partial Brazilian SDM. The AUC values were calculated with the function 'colAUC' in the R-library 'caTools' (Tuszynski 2012). Figure 4b shows that the AUC values for both groups were slightly higher for the partial SDMs (>0). This can at least partly be explained by the fact that the partial SDMs were fitted to the collection localities of Brazil alone. This can lead to over-fitting as can be concluded for the larger group of under- than over-predicting SDMs compared the full SDMs. This group has on average an approximate 20% reduction in their predicted presence compared to the full SDMs, as is illustrated by the 'Percentage difference Brazil (partial) vs. HT (full)' (Figure 4d; Table 2). Note that Figure 4d shows the absolute value of the 'percentage range difference' for the under-predicted models.

Table 2. Number of records used for the Brazilian Humid Tropics (partial) and the Neotropical Humid Tropics (full) SDMs and various measures of SDM accuracy and similarity. In all comparisons the Brazilian subset is compared to the clipped extent of the full SDM.

	Species	# records HT	# records Brazil	% Brazil vs. HT	AUC HT (clip Brazil)	AUC Brazil	AUC difference	Kappa	Fraction correct	# raster cells Brazil	# raster cells HT (clip Brazil)	% difference Brazil vs. HT
1	Inga alba	154	81	53	0.713	0.759	0.046	0.630	0.824	38559	40050	–4
2	Inga bourgonii	56	24	43	0.791	0.850	0.059	0.716	0.859	26750	31048	–14
3	Inga brachyrhachis	29	10	34	0.908	0.922	0.013	0.589	0.855	9569	18733	–49
4	Inga brevipes	9	6	67	0.804	0.943	0.139	0.140	0.539	5282	34974	–85
5	Inga cayennensis	64	26	41	0.852	0.901	0.049	0.460	0.770	14488	23651	–39
6	Inga cecropietorum	13	6	46	0.899	0.925	0.027	0.645	0.915	8944	9049	–1
7	Inga chartacea	41	12	29	0.823	0.893	0.071	0.478	0.802	18300	14274	28
8	Inga chrysantha	17	9	53	0.906	0.813	–0.093	0.415	0.736	26604	14609	82
9	Inga cinnamomea	45	25	56	0.779	0.826	0.048	0.719	0.863	42336	34944	21
10	Inga cordatoalata	23	12	52	0.845	0.868	0.023	0.566	0.817	13810	23323	–41
11	Inga disticha	51	35	69	0.819	0.854	0.034	0.751	0.883	26893	20667	30
12	Inga edulis	285	171	60	0.796	0.817	0.021	0.558	0.786	35853	41071	–13
13	Inga heterophylla	126	81	64	0.747	0.764	0.017	0.563	0.802	37901	49335	–23
14	Inga huberi	25	17	68	0.883	0.867	–0.016	0.612	0.809	30578	22100	38
15	Inga ingoides	115	57	50	0.823	0.834	0.010	0.561	0.784	33910	41712	–19
16	Inga lateriflora	57	37	65	0.812	0.864	0.052	0.548	0.774	25579	32523	–21
17	Inga lomatophylla	25	17	68	0.885	0.915	0.030	0.798	0.932	13863	13730	1
18	Inga macrophylla	67	33	49	0.843	0.888	0.045	0.571	0.802	17088	27102	–37
19	Inga marginata	432	283	66	0.852	0.861	0.009	0.717	0.859	28334	30908	–8
20	Inga melinonis	18	8	44	0.926	0.909	–0.017	0.636	0.885	12850	12665	1
21	Inga microcoma	10	6	60	0.859	0.893	0.035	0.680	0.886	12169	17434	–30
22	Inga nobilis ssp. nobilis	144	48	33	0.746	0.824	0.078	0.729	0.866	28426	29791	–5
23	Inga pezizifera	81	29	36	0.812	0.860	0.048	0.513	0.788	22035	19130	15
24	Inga pilosula	97	58	60	0.769	0.824	0.055	0.663	0.830	30531	37159	–18
25	Inga punctata	220	51	23	0.770	0.842	0.072	0.404	0.714	19562	29018	–33
26	Inga rubiginosa	46	30	65	0.870	0.884	0.014	0.889	0.959	16431	15307	7
27	Inga sertulifera ssp. sertulifera	30	15	50	0.882	0.925	0.043	0.606	0.850	11278	20415	–45
28	Inga stenoptera	84	39	46	0.791	0.839	0.048	0.774	0.889	24578	29256	–16
29	Inga stipularis	58	41	71	0.864	0.886	0.022	0.619	0.840	14453	22875	–37
30	Inga tenuistipula	33	13	39	0.899	0.956	0.057	0.539	0.902	6046	9381	–36
31	Inga thibaudiana ssp. thibaudiana	164	61	37	0.768	0.839	0.070	0.544	0.769	27855	35023	–20
32	Inga umbellifera	126	59	47	0.798	0.820	0.021	0.726	0.862	30243	37144	–19
33	Inga umbratica	55	35	64	0.833	0.858	0.026	0.659	0.832	36103	35707	1
34	Inga vera ssp. affinis	180	131	73	0.834	0.842	0.008	0.561	0.774	28404	40382	–30
35	Inga virgultosa	9	5	56	0.981	0.983	0.002	0.851	0.985	3238	3579	–10
36	Inga yacoana	16	10	63	0.947	0.974	0.027	0.598	0.926	4753	8317	–43

Grey records (n = 10) indicate a larger (over–) predicted presence range for partial SDMs. Bold table headers are also shown in Figure 4.

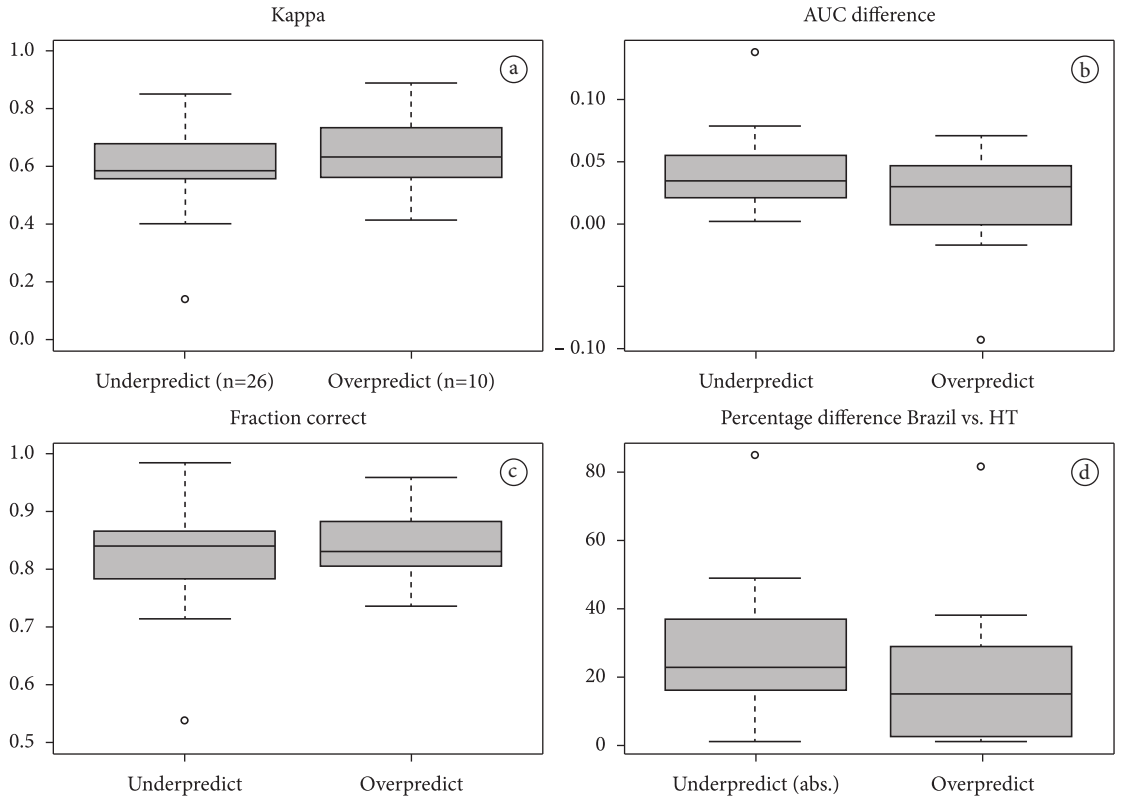

Figure 4. Different measures of model similarity and accuracy for SDMs developed for the partial Brazilian SDM compared to full HT biome SDMs for under- and over-predicted species separately; (abs = absolute value).

The impact of modelling partial SDMs

The *Inga* example illustrates that modelling the partial niche of species by setting artificial geographical or political boundaries results in patterns of predicted presence that are different from what can be expected from a full SDM. I take the position that full SDMs – taking all possible collection localities into account, and fitted within the Neotropical humid tropics biome as the biologically and biogeographically justifiable 'landscape of interest' – as the correct predictions to which the partial SDMs are compared. Importantly, all SDMs used in the comparisons were significantly different from random expectation and the lowest AUC value reported was 0.713 (Table 2).

The Kappa values indicate that similarities between the partial – and full SDMs are only 60-65% (Figure 4a); and when not corrected for the relative contribution of presence and absence area – the *Fraction correct* (Figure 4c), values of similarity average around 85%. Although the percentage difference in presence cells can be low, as is the case for *Inga alba* (Figure 2a, b; Table 2, –4%), the patterns of predicted presence-absence between partial – and full SDMs can be very different, which can be concluded from the kappa value of 0.630 (Table 2) and the areas of dissimilarity between the partial – and full SDM of *Inga alba* (Figure 2c). The AUC values of full SDMs were slightly lower than those of partial SDMs (Figure 4c). This can at least partly be attributed to the behaviour of the AUC value when applied to presence-only data. From the 36 partial SDMs, 26 had a smaller (under-)predicted range compared to the full SDMs. For the 26 under-predicted models the proportional area predicted present is reduced with 25% percent on average (Table 2; Figure 4d). This is equivalent to a proportional expansion of the 'landscape of interest', which also results in reduced percentages predicted presence. When AUC values are calculated with a background sample drawn from a proportional larger 'landscape of interest' automatically leads to AUC values that tend to be higher (Lobo *et al.* 2008, 2010). Therefore, it cannot be concluded from the slightly higher AUC values of the 26 under-predicted partial SDMs, that these models are more accurate than their full SDM counterparts.

This behaviour of the AUC value was also demonstrated by null-models, where larger sets of random points result in larger predicted presence areas and lower AUC values (Raes & ter Steege 2007). It is exactly this behaviour of AUC values when applied to presence-only data why all SDMs used in this example were tested for significance against a null-model (Raes & ter Steege 2007), instead of relying on subjective interpretation of AUC values, i.e. AUC > 0.8 as a reliable model. An explanation for the slightly higher average AUC of the 10 over-predicted partial SDMs compared to the full SDMs (Figure 4b) requires further study, and challenges the above discussion.

From the *Inga* example it can be concluded that modelling partial SDMs results in the contraction of many predicted distributions to the centre of ecological space (Figure 3d), which results in over-prediction in central Brazil when plotted in geographic space (Figure 3c); and in under-prediction at the artificially set boundaries (Figure 3a), there where the ecological gradients extend beyond the set boundary (Figure 3b). The under-predicted region in western Brazil corresponds with the region with the highest annual precipitation in the country (data not shown). Many *Inga* collections originate from localities just across the Brazil-Colombia/Peru border (Figure 3a, c, grey dots). Furthermore, the eastern side of the Ecuadorian Andes was also quite heavily sampled and is known to be humid. These conditions cannot be taken into account by the partial Brazilian SDMs and therefore result in predicted absence from the wetter side of the Brazilian precipitation gradient. The vector loading of annual precipitation (bio12) to PC1 (Figure 3b) was –0.86, what indicates that annual precipitation likely plays a role in the under-prediction of the partial SDMs in western Brazil. Partial SDMs, which do not take regions with high annual precipitation adequately into account in their presence – and background samples, result in predicted absence from these regions.

Similar contractions at artificial borders of predicted distributions based on partial SDMs were reported for the Iberian Peninsula (Sánchez-Fernández *et al*. 2011). Here I show that the geographic region of contraction corresponds with an artificial delimitation in ecological space in a direction where collections are found to occupy ecological space across this artificial boundary (Figure 3b – crosses). The over-prediction by partial SDMs in central Brazil is likely caused by interpolated environmental conditions between the reduced numbers of collections that are available to train the partial SDMs. To confirm these suggestions would require detailed analyses of species' individual response curves to the environmental gradients; a topic of further/future study and beyond the scope of this essay.

One of the few studies examining the effects of restricting the environmental range of data on the projection, or transferability, of SDMs to future climatic conditions (Thuiller *et al.* 2004) concluded that data restriction strongly influenced the estimation of the response curves. Notably, the effects were strongest towards the upper and lower ends of the environmental ranges. Thuiller *et al.* (2004) state that 'using restricted data is analogous to not capturing the full species' environmental range, reduces strongly the combinations of environmental conditions under which the models are calibrated, and reduces the applicability of the models for predictive purposes. This may generate unpredictable effects on the tails of the species response curves'. That data limitations can lead to truncated niches and unrealistic fits leading to spurious extrapolation to novel environments was also reported by Barbet-Massin *et al.* (2010) and Zurell *et al.* (2012). These findings are supported by the *Inga* example. Problems with transferability of partial SDMs not only apply to future projections but also extend into the past. Veloz *et al.* (2012) point out that 'a realized niche at any one time often only represents a subset of climate conditions in which a taxon can persist'. These problems directly relate to the non-analogue climatic contemporary conditions when SDMs are projected to the past or future (Roberts & Hamann 2011). SDMs fitted on contemporary climatic conditions therefore always are partial SDMs, with the possibility to represent truncated niches.

Based on the findings of others reported above and the *Inga* example presented here, I advise that SDMs use presence data from the complete distribution range of species, or at least from biogeographic instead of political boundaries. Furthermore, it should be kept in mind that any SDM is partial by nature, which is of special relevance when SDMs are projected into the past, present and future.

Acknowledgements

I like to thank Terence D. Pennington for allowing me to make use of his *Inga* dataset, Hans ter Steege for useful comments and improvements to the manuscript, and Jesus Aguire Gutierrez for his advice on the Map Comparison Kit. This research was made possible by NWO – ALW grant 819.01.014.

References

Acevedo P *et al*., 2012. Delimiting the geographical background in species distribution modelling. *Journal of Biogeography*, 39(8):1383-1390. http://dx.doi.org/10.1111/j.1365-2699.2012.02713.x

Araújo MB & Peterson AT, 2012. Uses and misuses of bioclimatic envelope modeling. *Ecology*, 93:1527-1539. PMid:22919900. http://dx.doi.org/10.1890/11-1930.1

Barbet-Massin M, Thuiller W & Jiguet F, 2010. How much do we overestimate future local extinction rates when restricting the range of occurrence data in climate suitability models? *Ecography*, 33:878-886. http://dx.doi.org/10.1111/j.1600-0587.2010.06181.x

Barve N *et al*., 2011. The crucial role of the accessible area in ecological niche modeling and species distribution modeling. *Ecological Modelling*, 222:1810-1819. http://dx.doi.org/10.1016/j.ecolmodel.2011.02.011

Beaumont LJ *et al*., 2009. Different climatic envelopes among invasive populations may lead to underestimations of current and future biological invasions. *Diversity and Distributions*, 15:409-420. http://dx.doi.org/10.1111/j.1472-4642.2008.00547.x

Bertrand R, Perez V & Gégout J-C, 2012. Disregarding the edaphic dimension in species distribution models leads to the omission of crucial spatial information under climate change: the case of *Quercus pubescens* in France. *Global Change Biology*, 18:2648-2660. http://dx.doi.org/10.1111/j.1365-2486.2012.02679.x

Boulangeat I, Gravel D & Thuiller W, 2012. Accounting for dispersal and biotic interactions to disentangle the drivers of species distributions and their abundances.

Ecology Letters, 15:584-593. PMid:22462813. http://dx.doi.org/10.1111/j.1461-0248.2012.01772.x

Broennimann O & Guisan A, 2008. Predicting current and future biological invasions: both native and invaded ranges matter. *Biology Letters*, 4:585-589. PMid:18664415 PMCid:2610080. http://dx.doi.org/10.1098/rsbl.2008.0254

Broennimann O *et al*., 2007. Evidence of climatic niche shift during biological invasion. *Ecology Letters*, 10:701-709. PMid:17594425. http://dx.doi.org/10.1111/j.1461-0248.2007.01060.x

Cayuela L *et al*., 2009. Species distribution modeling in the tropics: problems, potentialities, and the role of biological data for effective species conservation. *Tropical Conservation Science*, 2:319-352.

Colwell RK & Rangel TF, 2009. Hutchinson's duality: The once and future niche. *Proceedings of the National Academy of Sciences*, 106:19651-19658. PMid:19805163 PMCid:2780946. http://dx.doi.org/10.1073/pnas.0901650106

Dray S & Dufour AB, 2007. The ade4 Package: Implementing the Duality Diagram for Ecologists. *Journal of Statistical Software*, 22:20.

Elith J *et al*., 2006. Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29:129-151. http://dx.doi.org/10.1111/j.2006.0906-7590.04596.x

Elith J *et al*., 2011. A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, 17:43-57. http://dx.doi.org/10.1111/j.1472-4642.2010.00725.x

Evans, MEK *et al*., 2009. Climate, Niche Evolution, and Diversification of the "Bird-Cage" Evening Primroses (Oenothera, Sections Anogra and Kleinia). *The American Naturalist*, 173:225-240. PMid:19072708. http://dx.doi.org/10.1086/595757

Fielding AH & Bell JF, 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, 24:38-49. http://dx.doi.org/10.1017/S0376892997000088

Franklin J, 2009. *Mapping Species Distributions*: Spatial Inference and Prediction. Cambridge: Cambridge University Press.

Godsoe W, 2010. I can't define the niche but I know it when I see it: a formal link between statistical theory and the ecological niche. *Oikos*, 119:53-60. http://dx.doi.org/10.1111/j.1600-0706.2009.17630.x

Godsoe W, 2012. Are comparisons of species distribution models biased? Are they biologically meaningful? *Ecography*, 35:769-779. http://dx.doi.org/10.1111/j.1600-0587.2012.07456.x

Gotelli NJ & McGill BJ, 2006. Null versus neutral models: What's the difference? *Ecography*, 29:793-800. http://dx.doi.org/10.1111/j.2006.0906-7590.04714.x

Graham CH *et al*., 2008. The influence of spatial errors in species occurrence data used in distribution models. *Journal of Applied Ecology*, 45:239-247. http://dx.doi.org/10.1111/j.1365-2664.2007.01408.x

Grinnell J, 1917. The niche relationships of the California thrasher. *Auk*, 34:427-433. http://dx.doi.org/10.2307/4072271

Guisan A & Zimmermann NE, 2000. Predictive habitat distribution models in ecology. *Ecological Modelling*, 135:147-186. http://dx.doi.org/10.1016/S0304-3800(00)00354-9

Hagen A, 2002. Multi-method assessment of map similarity. In: *Proceedings of the 5th AGILE Conference on Geographic Information Science*; 2002; Palma. Mallorca.

Hijmans RJ *et al*., 2005. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, 25:1965-1978. http://dx.doi.org/10.1002/joc.1276

Holt RD, 2009. Bringing the Hutchinsonian niche into the 21st century: Ecological and evolutionary perspectives. *Proceedings of the National Academy of Sciences*, 106:19659-19665. PMid:19903876 PMCid:2780934. http://dx.doi.org/10.1073/pnas.0905137106

Hortal J, Lobo JM & Jiménez-Valverde A, 2007. Limitations of biodiversity databases: Case study on seed-plant diversity in Tenerife, Canary Islands. *Conservation Biology*, 21:853-863. PMid:17531062. http://dx.doi.org/10.1111/j.1523-1739.2007.00686.x

Hsu RCC *et al*., 2011. Simulating climate change impacts on forests and associated vascular epiphytes in a subtropical island of East Asia. *Diversity and Distributions*, 18(4):334-347.

Hubbell SP *et al*., 2008. How many tree species are there in the Amazon and how many of them will go extinct? *Proceedings of the National Academy of Sciences*, 105:11498-11504. PMid:18695228 PMCid:2556410. http://dx.doi.org/10.1073/pnas.0801915105

Hutchinson GE, 1957. Concluding remarks. *Proceedings of the Cold Spring Harbor Symposia on Quantitative Biology*, 22:415-427.

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. http://dx.doi.org/10.1890/02-5364

Lalonde VB, Morin A & Currie DJ, 2012. How are tree species distributed in climatic space? A simple and general pattern. *Global Ecology and Biogeography*. In press.

Lobo JM, Jiménez-Valverde A & Hortal J, 2010. The uncertain nature of absences and their importance in species distribution modelling. *Ecography*, 33:103-114. http://dx.doi.org/10.1111/j.1600-0587.2009.06039.x

Lobo JM, Jimenez-Valverde A & Real R, 2008. AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, 17:145-151. http://dx.doi.org/10.1111/j.1466-8238.2007.00358.x

Loiselle BA *et al*., 2008. Predicting species distributions from herbarium collections: does climate bias in collection sampling influence model outcomes? *Journal of Biogeography*, 35:105-116.

Maiorano L *et al*., 2012. Building the niche through time: using 13,000 years of data to predict the effects of climate change on three tree species in Europe. *Global Ecology and Biogeography*. In press. http://dx.doi.org/10.1111/j.1466-8238.2012.00767.x

Mayle FE, Burbridge R & Killeen TJ, 2000. Millennial-Scale Dynamics of Southern Amazonian Rain Forests.

Science, 290:2291-2294. PMid:11125139. http://dx.doi.org/10.1126/science.290.5500.2291

Olden JD, Jackson DA & Peres-Neto PR, 2002. Predictive Models of Fish Species Distributions: A Note on Proper Validation and Chance Predictions. *Transactions of the American Fisheries Society*, 131:329-336. http://dx.doi.org/10.1577/1548-8659(2002)131<0329:PMOFSD>2.0.CO;2

Olson DM *et al*., 2001. Terrestrial ecoregions of the world: A new map of life on earth. *Bioscience*, 51:933-938. http://dx.doi.org/10.1641/0006-3568(2001)051[0933:TEOTWA]2.0.CO;2

Pearman PB *et al*., 2008. Niche dynamics in space and time. *Trends in Ecology & Evolution*, 23:149-158. PMid:18289716. http://dx.doi.org/10.1016/j.tree.2007.11.005

Pennington TD *et al*., 1997. *The genus Inga*: Botany. London: Royal Botanical Gardens, Kew.

Peterson AT *et al*., 2011. *Ecological Niches and Geographic Distributions*. Princeton: Princeton University Press.

Phillips SJ, Anderson RP & Schapire RE, 2006. Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190:231-259. http://dx.doi.org/10.1016/j.ecolmodel.2005.03.026

Pineda E & Lobo JM, 2009. Assessing the accuracy of species distribution models to predict amphibian species richness patterns. *Journal of Animal Ecology*, 78:182-190. PMid:18771504. http://dx.doi.org/10.1111/j.1365-2656.2008.01471.x

R Development Core Team, 2012. *R*: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Available from: <http://www.R-project.org>.

Raes N *et al*., 2009. Botanical richness and endemicity patterns of Borneo derived from species distribution models. *Ecography*, 32:180-192. http://dx.doi.org/10.1111/j.1600-0587.2009.05800.x

Raes N & ter Steege H, 2007. A null-model for significance testing of presence-only species distribution models. *Ecography*, 30:727-736. http://dx.doi.org/10.1111/j.2007.0906-7590.05041.x

Raxworthy CJ *et al*., 2003. Predicting distributions of known and unknown reptile species in Madagascar. *Nature*, 426:837-841. PMid:14685238. http://dx.doi.org/10.1038/nature02205

Reddy S & Davalos LM, 2003. Geographical sampling bias and its implications for conservation priorities in Africa. *Journal of Biogeography*, 30:1719-1727. http://dx.doi.org/10.1046/j.1365-2699.2003.00946.x

Richardson JE *et al*., 2001. Rapid Diversification of a Species-Rich Genus of Neotropical Rain Forest Trees. *Science*, 293:2242-2245. PMid:11567135. http://dx.doi.org/10.1126/science.1061421

Roberts DR & Hamann A, 2011. Predicting potential climate change impacts with bioclimate envelope models: a palaeoecological perspective. *Global Ecology and Biogeography*, 21:121-133. http://dx.doi.org/10.1111/j.1466-8238.2011.00657.x

Sánchez-Fernández D, Lobo JM & Hernández-Manrique OL, 2011. Species distribution models that do not incorporate global data misrepresent potential distributions: a case study using Iberian diving beetles. *Diversity and Distributions*, 17:163-171. http://dx.doi.org/10.1111/j.1472-4642.2010.00716.x

Schulman L, Toivonen T & Ruokolainen K, 2007. Analysing botanical collecting effort in Amazonia and correcting for it in species range estimation. *Journal of Biogeography*, 34:1388-1399. http://dx.doi.org/10.1111/j.1365-2699.2007.01716.x

Soberón J, 2007. Grinnellian and Eltonian niches and geographic distributions of species. *Ecology Letters*, 10:1115-1123. PMid:17850335. http://dx.doi.org/10.1111/j.1461-0248.2007.01107.x

Soberón J & Peterson AT, 2005. Interpretation of models of fundamental ecological niches and species' distributional areas. *Biodiversity Informatics*, 2:1-10.

SpeciesLink. Sistema de informação distribuído para coleções biológicas. Centro de Referência em Informação Ambiental-CRIA Available from: <http://www.splink.cria.org.br>. Access in: 29 July 2012.

Svenning J-C & Skov F, 2004. Limited filling of the potential range in European tree species. *Ecology Letters*, 7:565-573. http://dx.doi.org/10.1111/j.1461-0248.2004.00614.x

Thuiller W *et al*., 2004. Effects of restricting environmental range of data to project current and future species distributions. *Ecography*, 27:165-172. http://dx.doi.org/10.1111/j.0906-7590.2004.03673.x

Tilman D, 1982. *Resource Competition and Community Structure*. Princeton: Princeton University Press. PMid:7162524.

Tuomisto H, 2006. Edaphic niche differentiation among Polybotrya ferns in western Amazonia: implications for coexistence and speciation. *Ecography*, 29:273-284. http://dx.doi.org/10.1111/j.2006.0906-7590.04390.x

Tuszynski J, 2012. *caTools*: Tools: moving window statistics, GIF, Base64, ROC AUC, etc.

Van Welzen PC *et al*., 2011. The current and future status of floristic provinces in Thailand. In: Trisurat Y, Shrestha RP & Alkemade R, editors. *Land Use*, *Climate Change and Biodiversity Modeling*: Perspectives and Applications. Hershey: IGI Globa. p. 219-247. http://dx.doi.org/10.4018/978-1-60960-619-0.ch011

VanDerWal J *et al*., 2009. Selecting pseudo-absence data for presence-only distribution modeling: How far should you stray from what you know? *Ecological Modelling*, 220:589-594. http://dx.doi.org/10.1016/j.ecolmodel.2008.11.010

Veloz SD *et al*., 2012. No-analog climates and shifting realized niches during the late quaternary: implications for 21st-century predictions by species distribution models. *Global Change Biology*, 18:1698-1713. http://dx.doi.org/10.1111/j.1365-2486.2011.02635.x

Visser H & De Nijs T, 2006. The Map Comparison Kit. *Environmental Modelling & Software*, 21:346-358. http://dx.doi.org/10.1016/j.envsoft.2004.11.013

Wenger SJ & Olden JD, 2012. Assessing transferability of ecological models: an underappreciated aspect of statistical validation. *Methods in Ecology and Evolution*, 3:260-267. http://dx.doi.org/10.1111/j.2041-210X.2011.00170.x

Wiens JJ *et al*., 2010. Niche conservatism as an emerging principle in ecology and conservation biology. *Ecology Letters*, 13:1310-1324. PMid:20649638. http://dx.doi.org/10.1111/j.1461-0248.2010.01515.x

Wisz MS *et al*., 2008. Effects of sample size on the performance of species distribution models. *Diversity and Distributions*, 14:763-773. http://dx.doi.org/10.1111/j.1472-4642.2008.00482.x

Yesson C & Culham A, 2006. Phyloclimatic Modeling: Combining Phylogenetics and Bioclimatic Modeling. *Systematic Biology*, 55:785-802. PMid:17060200. http://dx.doi.org/10.1080/1063515060081570

Zhang M-G *et al*., 2012. Using species distribution modeling to improve conservation and land use planning of Yunnan, China. *Biological Conservation*, 153:257-264. http://dx.doi.org/10.1016/j.biocon.2012.04.023

Zurell D, Elith J & Schröder B, 2012. Predicting to new environments: tools for visualizing model behaviour and impacts on mapped distributions. *Diversity and Distributions*, 18:628-634. http://dx.doi.org/10.1111/j.1472-4642.2012.00887.x

Received: August 2012
First Decision: September 2012
Accepted: October 2012