Ecological Modelling 222 (2011) 2796–2811

Contents lists available at ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel

Species-specific tuning increases robustness to sampling bias in models of species distributions: An implementation with Maxent

Robert P. Anderson [a,b,*], Israel Gonzalez Jr. [a,1]

[a] Department of Biology, 526 Marshak Science Building, City College of the City University of New York, 160 Convent Avenue, New York, NY 10031, USA
[b] Division of Vertebrate Zoology (Mammalogy), American Museum of Natural History, Central Park West at 79th Street, New York, NY 10024, USA

ARTICLE INFO

Article history:
Received 23 November 2010
Received in revised form 31 March 2011
Accepted 7 April 2011
Available online 26 May 2011

Keywords:
Complexity
Maximum entropy
Niche
Noise
Overfitting
Sample size

ABSTRACT

Various methods exist to model a species' niche and geographic distribution using environmental data for the study region and occurrence localities documenting the species' presence (typically from museums and herbaria). In presence-only modelling, geographic sampling bias and small sample sizes represent challenges for many species. Overfitting to the bias and/or noise characteristic of such datasets can seriously compromise model generality and transferability, which are critical to many current applications – including studies of invasive species, the effects of climatic change, and niche evolution. Even when transferability is not necessary, applications to many areas, including conservation biology, macroecology, and zoonotic diseases, require models that are not overfit. We evaluated these issues using a maximum entropy approach (Maxent) for the shrew *Cryptotis meridensis*, which is endemic to the Cordillera de Mérida in Venezuela. To simulate strong sampling bias, we divided localities into two datasets: those from a portion of the species' range that has seen high sampling effort (for model calibration) and those from other areas of the species' range, where less sampling has occurred (for model evaluation). Before modelling, we assessed the climatic values of localities in the two datasets to determine whether any environmental bias accompanies the geographic bias. Then, to identify optimal levels of model complexity (and minimize overfitting), we made models and tuned model settings, comparing performance with that achieved using default settings. We randomly selected localities for model calibration (sets of 5, 10, 15, and 20 localities) and varied the level of model complexity considered (linear versus both linear and quadratic features) and two aspects of the strength of protection against overfitting (regularization). Environmental bias indeed corresponded to the geographic bias between datasets, with differences in median and observed range (minima and/or maxima) for some variables. Model performance varied greatly according to the level of regularization. Intermediate regularization consistently led to the best models, with decreased performance at low and generally at high regularization. Optimal levels of regularization differed between sample-size-dependent and sample-size-independent approaches, but both reached similar levels of maximal performance. In several cases, the optimal regularization value was different from (usually higher than) the default one. Models calibrated with both linear and quadratic features outperformed those made with just linear features. Results were remarkably consistent across the examined sample sizes. Models made with few and biased localities achieved high predictive ability when appropriate regularization was employed and optimal model complexity was identified. Species-specific tuning of model settings can have great benefits over the use of default settings.

1. Introduction

Many techniques exist for modelling species ecological niches and geographic distributions, and these approaches hold myriad applications in environmental biology. Occurrence localities for the study species and environmental variables for the study region are used to calibrate the model. Modelling with data from planned presence–absence surveys is well established (Guisan et al., 2002; Scott et al., 2002), but high-quality presence–absence data are available for only relatively few species worldwide. In contrast, vast data documenting species' *presence* exist in natural history museums and herbaria (Graham et al., 2004). Accordingly, numerous techniques have been developed or modified recently to allow for modelling based on presence-only data (Peterson, 2003; Wiens

* Corresponding author at: 526 Marshak Science Building, City College of the City University of New York, 160 Convent Avenue, New York, NY 10031, USA.
Tel.: +1 2126508504; fax: +1 2126508585.
E-mail address: anderson@sci.ccny.cuny.edu (R.P. Anderson).
[1] Present address: College of Medicine, Howard University, Washington, DC, USA.

0304-3800/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2011.04.011

and Graham, 2005; Elith et al., 2006; Kozak et al., 2008; Peterson et al., in press).

Unfortunately, two data-related problems affect such approaches. Museum and herbarium data typically derive from many individual surveys conducted across several decades by numerous researchers sampling with different techniques, intensity, and periods of time (Soberón and Peterson, 2004). When some areas have been sampled more than others, the resulting localities for a species reflect this bias. Such geographic sampling bias can lead to sampling bias in environmental space, which represents a major problem for modelling (Hortal et al., 2008; see also Veloz, 2009 for the effects of sampling bias on model evaluation). Geographic and environmental biases violate an explicit or implicit assumption of many modelling techniques, namely that the localities represent a random sample from the entity being modelled (Phillips et al., 2006). When it is possible to quantify sampling bias, corrections for it can be made in the modelling process (Zaniewski et al., 2002; Phillips et al., 2009), but such cases remain rare. In addition to these problems related to sampling bias, most species are known from only a few localities. Such datasets are less likely to reflect the species' full niche and are often plagued by random vagaries of sampling (noise), both of which can hinder the calibration of realistic models (Wisz et al., 2008).

These issues represent challenges for modelling in large part because, all other things being equal, datasets with biased and/or few localities are especially prone to overfitting. An overfit model is more complex than the real relationships between the species' niche and the examined variables (Peterson et al., in press). A model can be overfit to sampling bias and/or noise, especially when employing many predictor variables and when the researcher allows a high level of model complexity. Clearly, if no sampling bias exists, a model cannot be overfit to it. In contrast, noise (due to random chance in sampling) plagues datasets of all sample sizes but should have a stronger effect for species with few localities – with other factors, such as the number of predictor variables and the allowed level of complexity, held constant. For a species with many localities, the pervasive environmental signal reflective of its niche more likely dominates any random (apparent but false) "signal" due to noise. In contrast, a species with few localities lacks a large enough body of localities to establish the true niche-based signal with sufficient inertia to protect against the false "signal" of random noise. Because of this, such a species will be more likely to suffer from overfitting to noise.

Hence, especially for datasets with biased and/or few localities, successful techniques must achieve an appropriate balance between too simple versus excessively complex models (e.g. those including too many parameters and/or overly high weights for them). We desire an optimal model that provides the maximal explanation of calibration data while still maintaining generality – the ability to predict independent data. To avoid or at least ameliorate overfitting, internal safeguards can be applied that penalize the production of complex models (e.g. Phillips et al., 2006; Phillips and Dudík, 2008; see below). Alternatively, models with an optimal level of complexity can be identified through techniques such as the Akaike Information Criterion (AICc), but extensions of such approaches only recently have been proposed for these models (Warren and Seifert, 2011).

Achieving models with an optimal level of complexity and showing high generality is important for all uses, especially several current applications requiring high transferability (applicability to other situations). These include applying a model to another time period (e.g. after climatic change) or region (e.g. prediction of an invasive species; Araújo and Rahbek, 2006; Williams and Jackson, 2007). Furthermore, studies of niche evolution (Wiens and Graham, 2005; Warren et al., 2008) also require models that are not overfit. However, even without the need for transferring a model to another time period or place, conservation assessments, macroecological analyses, distributional studies of zoonotic diseases, and many other uses need models with an optimal level of complexity (Anderson and Martínez-Meyer, 2004; Peterson et al., in press).

Table 1
Default settings for the β regularization parameter for linear and quadratic feature classes (Version 2.2.0 of Maxent). Values employed by the algorithm are interpolated for sample sizes in between those given here. Note that the default settings for this version suggest the use of only linear features for datasets with 2–9 localities and both linear and quadratic features for 10–79 localities (with the addition of hinge features for 15 localities and higher; see Phillips and Dudík, 2008).

Linear features				
Number of localities	0	10	30	100+
β regularization parameter	1.0	1.0	0.2	0.05

Linear and quadratic features					
Number of localities	0	10	17	30	100+
β regularization parameter	1.3	0.8	0.5	0.25	0.05

We address these issues with the maximum entropy (Maxent) approach. Maximum entropy is an area of machine learning that has been applied widely in recent years to model species niches and distributions based on presence-only occurrence data (Phillips et al., 2006; Phillips and Dudík, 2008). Maxent has performed well in comparisons with other presence-only methods (Elith et al., 2006; Hernandez et al., 2006; Wisz et al., 2008; but see Peterson et al., 2007). The current contribution not only provides insights into the general issues of sampling bias, small sample sizes, model complexity, and model performance (relevant to many modelling techniques) but also affords specific results and recommendations for Maxent.

Specifically, we address the role that species-specific tuning or "smoothing" of model settings – rather than employing default settings – can play in reducing overfitting and achieving optimal Maxent models for a dataset with few and biased localities. The protection against overfitting employed in Maxent is termed l_1 regularization and is built into the modelling process. Equivalent to the "lasso" applied to generalized linear and generalized additive models (GLM/GAMs; Guisan et al., 2002), regularization applies a penalty for each term that is included in the model; the strength of the penalty is determined by a parameter (β) that is multiplied by the weight given to that term in the model (Phillips et al., 2006; see Section 2.2).

Suggested β values have been determined for Maxent based on extensive empirical tuning, but using those values presents two drawbacks (see default settings in Version 2.2.0 and subsequent releases, Table 1; Phillips and Dudík, 2008). Those tuning experiments were conducted with random splits of localities into calibration and evaluation datasets (split-sample approach of Guisan and Zimmermann, 2000). Unfortunately, that approach does not yield truly independent evaluation data, and the resulting measures of performance are likely inflated, in part due to spatial autocorrelation between calibration and evaluation localities (Araújo et al., 2005; Veloz, 2009). In addition, because the effects of any sampling bias are preserved in both datasets, such evaluations cannot detect overfitting to bias (but rather, only to noise). Finally, the default settings correspond to those that were the best on average for many species using different sets of environmental variables in various regions of the world, but they will not necessarily lead to the optimal level of complexity for the species (and occurrence localities) at hand with the environmental variables and study region of a given analysis. Indeed, recent empirical work showed that species-specific tuning or smoothing (rather than the use of default regularization) greatly increased model quality and

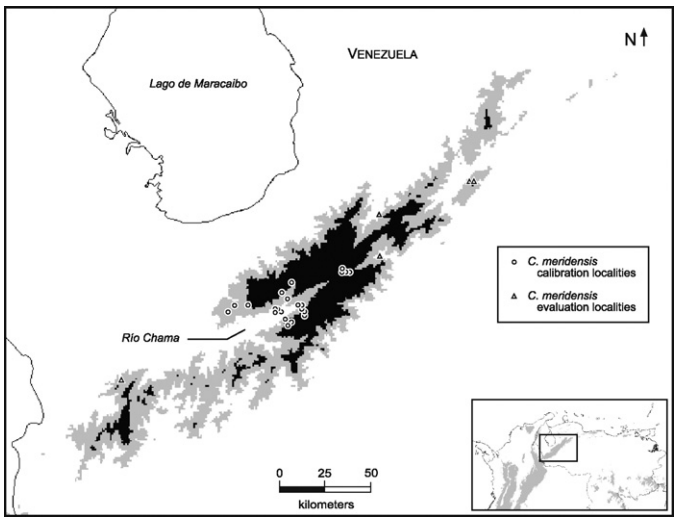

Fig. 1. Localities of *C. meridensis* used in this study (from Woodman, 2002; Appendix A). Circles indicate localities from the portion of the species' range in the vicinity of the city of Mérida (where sampling effort has been relatively high); subsets of these localities were used in model calibration. Triangles denote localities in other parts of the species' range (which have received much lower levels of sampling); this second group constituted the evaluation data. The species is endemic to the Cordillera de Mérida and inhabits montane forests and páramo habitats at ca. 1640–3950 m. In the map of the Cordillera de Mérida, areas shaded grey correspond to elevations from 2000 to 2999 m, and those shown in black indicate regions over 3000 m (in the inset map, shaded areas represent elevations above 1000 m).

transferability with Maxent (Elith et al., 2010), calling for similar efforts with other species and study systems.

As a possible solution to these problems, we conduct species-specific tuning via model evaluations that use localities from other geographic areas (as suggested by Araújo and Rahbek, 2006), rather than with localities geographically intermixed with the ones used in model calibration. This allows researchers the potential to detect overfitting to sampling bias (in addition to overfitting to noise) and identify model settings that achieve an optimal level of model complexity. This approach assumes that the species' response is "stationary" across space. For example, the respective areas used for calibration versus evaluation should not differ in the available environments or biotic contexts, and the populations in each area should have the same inherited niche-related characteristics (especially regarding physiology; Osborne et al., 2007; Williams and Jackson, 2007; Anderson and Raza, 2010).

With that caveat, we conduct a set of experiments using a convenient species, the Mérida shrew, *Cryptotis meridensis* (Mammalia: Soricomorpha: Soricidae). This species is endemic to a single range of the Andes in northwestern Venezuela (the Cordillera de Mérida) where no other species of shrew is present (Fig. 1; Woodman and Díaz de Pascual, 2004). These factors simplify both modelling and interpretation because limitations to dispersal and competitive interactions with related species are unlikely to restrict the species' distribution within this mountain range (Anderson et al., 2002a,b). In addition, they increase the likelihood of stationarity. Furthermore, known localities for *C. meridensis* lend themselves to addressing the effects of sampling bias, with plentiful localities available from one part of the species' range (near the large city of Mérida) and a few available from other, less-studied areas that hold similar environmental conditions.

Our experimental design includes calibrating models under differing conditions and evaluating performance quantitatively and qualitatively. We vary sample size, using random sets of 5–20 localities to make the models. We take these calibration localities from the portion of the species' range that has been intensely sampled. We then evaluate the models with localities from other parts of the species' range – those that have received little sampling. To assess whether this geographic bias in sampling corresponds to an environmental bias as well, we compare the environments occupied by the species in the two datasets. Our evaluations of model performance include threshold-independent (AUC) and threshold-dependent (omission rate) measures, as well as well as visual interpretations of predictions in geography (see Section 2.4). We also alter the level of regularization, to determine what value leads to the highest performance with these datasets that are highly biased geographically. Finally, we change the level of model complexity allowed (the feature classes considered; see Section 2.3). These experiments are designed to help understand the effects of species-specific tuning and the performance of Maxent with few and geographically biased localities, facilitating its use for academic studies and conservation efforts and making a contribution to the literature regarding these general issues in presence-only distributional modelling.

2. Methods

2.1. Data sources

The Mérida small-eared shrew, *C. meridensis* is a small terrestrial mammal endemic to the Cordillera de Mérida in Venezuela (Fig. 1; Hutterer, 2005). It inhabits a known elevational range of 1640–3950 m and occurs in primary montane forest (including cloud forest) and páramo habitats above tree line, as well as disturbed forest and secondary vegetation along streams (Woodman and Díaz de Pascual, 2004). Although the species can tolerate a short dry season in seasonal montane forests, it is absent from lower and drier vegetation types, such as dry thorn scrub (Durant and Péfaur, 1984; Díaz et al., 1997; Soriano et al., 1999). We examined its potential geographic distribution in northwestern Venezuela, including the Cordillera de Mérida, a few associated low mountain chains, a small portion of the Tamá region of the Cordillera Oriental, and extensive lowland areas (Fig. 1; 7°36′–10°06′N, 69°12′–72°30′W). In this study region, the species' potential distribution likely corresponds closely to its actual distribution, with the exception of the small Tamá region (to which the species likely has not dispersed) and areas that have been heavily impacted by humans.

For the occurrence data, we used 27 localities of *C. meridensis* from a recent taxonomic study (Woodman, 2002). These localities correspond to museum specimens examined by that author as well as other records of the species reported in the literature by other workers (all cited in Woodman, 2002). Because *C. meridensis* is the only shrew known from the Cordillera de Mérida, these identifications should be reliable. We determined latitude and longitude for these localities using a variety of sources, including gazetteers, topographic maps, and original publications by collectors (Appendix A).

For the environmental data, we used 19 climatic variables from WorldClim (http://biogeo.berkeley.edu/worldclim/worldclim.htm). The WorldClim project interpolated data from weather stations (generally from 1950 to 2000) using a splining technique (thin-plate smoothing) to produce high-resolution (ca. 1 km^2) maps of average monthly climatic data. These raw monthly data were then processed to yield 19 bioclimatic variables that reflect various aspects of temperature, precipitation, and seasonality and are likely important in determining species distributions (Hijmans et al., 2005). They consist of annual mean temperature, mean diurnal range (mean of monthly values of maximum temperature minus minimum temperature), isothermality, temperature seasonality, maximum temperature of warmest month, minimum temperature of coldest month, temperature annual range, mean temperature of wettest quarter, mean temperature of driest quar-

ter, mean temperature of warmest quarter, mean temperature of coldest quarter, annual precipitation, precipitation of wettest month, precipitation of driest month, precipitation seasonality, precipitation of wettest quarter, precipitation of driest quarter, precipitation of warmest quarter, and precipitation of coldest quarter.

2.2. Model calibration

We calibrated models of the potential distribution of *C. meridensis* using the maximum entropy (Maxent) method. Maximum entropy is an area of machine learning used to make inferences from incomplete information and has been applied recently to modelling species distributions using presence-only occurrence data (Phillips et al., 2006; Phillips and Dudík, 2008). In addition to using environmental data from the localities of known presence, Maxent also takes a sample of 10,000 pixels from the study region used in model calibration in order to characterize the "background" of environments available to the species (hence, it is termed a presence–background technique). Maxent has performed well in many comparative studies (e.g. Elith et al., 2006; Guisan et al., 2007; Wisz et al., 2008; but see Peterson et al., 2007). Version 2.2.0 was used here, and the current basic research is relevant to subsequent releases of Maxent as well (http://www.cs.princeton.edu/~schapire/maxent/). Maxent calculates a raw probability value for each pixel of the study region, indicating the suitability of a given pixel relative to all other pixels. These raw probabilities are scaled to sum to 1 and do not represent probability of occurrence but rather an index of relative suitability. Maxent presents suitability in an intuitive cumulative representation, in which the value for a pixel represents its raw suitability plus the sum of the suitabilities of all pixels of lesser or equal suitability, multiplied by 100 to yield a percentage (Phillips et al., 2006; for information regarding the alternative logistic output, see Phillips and Dudík, 2008).

Maxent produces a model of the species' requirements based on "features," which are environmental variables and functions thereof (Phillips et al., 2006). A linear feature uses the variable itself and models the mean value of the variable for localities occupied by the species (optimal conditions for the species). Similarly, a quadratic feature (the square of an environmental variable) models the variance of that variable for the species' localities (its tolerance for variation from optimal conditions, when used along with linear features). Other feature classes are available for use by Maxent with continuous variables (such as product, threshold, and hinge features; Phillips and Dudík, 2008). However, the default settings for Version 2.2.0 suggest the use of only linear and product features with small sample sizes of localities (below 15 localities, and just linear features for sample sizes below 10). To compare performance according to the classes of features that Maxent was allowed to consider, we conducted our experiments once using just linear features and then employing both linear and quadratic features (see Section 2.3).

To reduce the tendency to create models that are overfit, Maxent employs what is termed regularization to penalize strong weights given to features. This penalizes complex models that include many features and/or have strong weights for them, in effect forcing Maxent to concentrate on only the most important features – those with the highest explanatory ability (Phillips et al., 2006). This process is similar in intent to selecting optimal models using the Akaike Information Criterion (AICc; Warren and Seifert, 2011). The regularization value used by Maxent for a particular feature is determined as follows:

$$\text{regularization} = \frac{\beta \times \text{standard deviation}}{(\text{sample size})^{1/2}}$$

where β is a constant, the standard deviation used is that of the values of the feature for the localities, and sample size is the number of localities. Note that the standard deviation/(sample size)$^{1/2}$ is the standard error (calculated from the values of that feature for the localities; = standard error of the mean). The β parameter is multiplied by the standard error because it measures how much the sample mean is expected to vary from the true mean (S. J. Phillips, personal communication.). Higher values for the β parameter yield higher penalties for use of a feature and for higher weights being given to a feature. The value of β can differ for different feature classes and sample sizes (and does so in the default settings implemented in the software, based on previous empirical tuning; Phillips and Dudík, 2008). We overrode the default settings and varied regularization values directly in our experiments (see Section 2.3). It is important to note that the β regularization parameter is not the same as the regularization multiplier found in later releases of Maxent. That multiplier is a user-specified number that is multiplied to the value of the β parameter of each respective feature class (the default regularization multiplier is 1).

2.3. Experimental design

We designed these experiments with the primary goal of determining the robustness of Maxent to geographically biased localities (which are common in museum databases due to unequal sampling effort across geography). To do so, we made the models based on localities from one portion of the species' range (an area surrounding the city of Mérida that has received substantial collection effort) and then evaluated them based on their ability to predict localities in other parts of the species' known range (areas that have received very little collection effort, e.g. Paynter, 1982: p. 216; Fig. 1). Twenty-two localities exist from the well-sampled portion of the species' range, for calibration (=training), and five occur in other areas, for evaluation (=testing; Fielding and Bell, 1997; Guisan and Zimmermann, 2000).

To assess whether any environmental bias accompanies this geographic bias, we compared climatic data for localities from the well-sampled portion of the species' range with data from the localities from other areas. We addressed both differences in central tendency (here, median) and dispersion (here, range). First, because these data are unlikely to be normally distributed, we conducted non-parametric tests (two-tailed Mann–Whitney *U*-tests) comparing the median value for each climatic variable between the two datasets. Then, we calculated the minimum and maximum value for each variable in each dataset and inspected the results to determine whether the values for either dataset fell far beyond the observed range of the other dataset. Similarly, to compare the environmental information for the full set of 22 localities from the well-sampled portion of the species' range with that contained in a random sample of them, we conducted the same analyses for the full set versus the first random sample of 5 localities. To accomplish these comparisons, we extracted climatic data for each locality from each of the 19 bioclimatic variables using DIVA-GIS (version 5.4.0.1; www.diva-gis.org) and then performed the statistical analyses in Minitab (2003), release 14.1. Because the small sample sizes here likely afforded low power, we determined significance by comparing probabilities to an $\alpha = 0.05$ but also inspected and commented on tests showing $p < 0.10$.

Secondarily, to assess the effects of sample size on model performance, we varied the number of localities in the calibration datasets. To do so, we created calibration datasets of various sample sizes (5, 10, 15, and 20 localities) by randomly selecting from the 22 localities in the well-sampled portion of the species' range. To produce results more robust to chance events regarding the selection of localities, we made 10 random sets of calibration localities for

each sample size. Each resulting model was evaluated using the 5 localities from other parts of the species' known range (areas having received very little collection effort).

For each experiment, we also varied the level of regularization. We used eight different fixed values of the β parameter (0.01, 0.1, 0.5, 1.0, 1.5, 2.0, 3.0, 5.0; each specified via code), as well as the default recommendation of Version 2.2.0. The default settings vary depending on the sample size and are based on empirical tuning performed in previous studies (Phillips and Dudík, 2008); L: 1.0 for 5 localities, 1.0 for 10 localities, 0.8 for 15 localities, 0.6 for 20 localities; LQ: 1.050 for 5 localities, 0.800 for 10 localities, 0.586 for 15 localities, 0.442 for 20 localities.

We evaluated regularization in two ways. The first was to calibrate models of a particular sample size by using the β parameter that we designated; the software then multiplied the value of the parameter by the standard error of the sample for a given variable to yield the regularization value for that variable (see Section 2.2). We term this the sample-size-dependent approach because the square root of the sample size is the denominator of the standard error. The second way was by multiplying the β parameter by the square root of the sample size. This effectively eliminated any mathematical dependence on sample size (counteracting the fact that the software multiplies β by the standard error; in effect, β is here multiplied by the standard deviation). We refer to this as the sample-size-independent approach.

Finally, we also compared model performance according to classes of features used. Specifically, we conducted our experiments first using just linear features and then employing both linear and quadratic features. In summary, the use of 4 sample sizes, 10 datasets per sample size, 9 β values, 2 regularization approaches, and 2 sets of feature classes led to the production of 1440 total models.

2.4. Model evaluation

We evaluated models using threshold-independent and threshold-dependent means. A threshold-independent evaluation provides a single measure of model performance across all possible thresholds (levels of strength of the prediction). In contrast, application of a threshold divides a continuous prediction into a binary one (i.e., those areas predicted suitable versus not suitable for the species).

In the threshold-independent analysis, we evaluated model performance using the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve (Fielding and Bell, 1997; Elith and Burgman, 2002). Use of AUC/ROC analyses with presence-only evaluation datasets has been clarified and justified recently for the classification problem of presence versus random, using information regarding the background of the study region instead of absence data (Wiley et al., 2003; Phillips et al., 2006). Although AUC is a non-parametric measure (and, hence, cannot assess goodness-of-fit) and varies according to the proportion of the study region that is suitable for the species (and, therefore, should not be compared across species or across study regions, at least with presence–background evaluation data), it is appropriate for our comparisons of the relative ranking ability of models produced with different settings but for the same species in the same study region (Lobo et al., 2008; Peterson et al., 2008).

The AUC from a ROC analysis varies from 0 to 1 and represents a measure of overall model performance (ability to discriminate suitable versus unsuitable areas) independent of threshold. However, the maximum achievable AUC with a presence–background evaluation data set is less than 1.0 (Wiley et al., 2003; Phillips et al., 2006). Maxent provides AUC values based on the evaluation localities for each model. We averaged the AUC of the models produced using each of the 10 data sets for each experiment (e.g. particular sample size, β value, etc.) for graphical presentation.

Complementarily, in the threshold-dependent approach, we divided the continuous prediction into a binary prediction of presence or absence using two different thresholding rules and then assessed omission rates and model significance. High-quality models should show zero or low omission of evaluation localities and predict evaluation localities statistically better than a random prediction. We used one-tailed binomial probabilities to determine whether evaluation localities fell into regions of predicted presence more often than expected by chance (model significance; Anderson et al., 2002a; Fielding, 2002; Phillips et al., 2006). For simplicity (but to maintain a wide range of settings and allow comparisons with the most-commonly used ones), we performed these threshold-dependent evaluations only for sample sizes 5 and 20; selected values of the regularization parameter (β = 0.01, 1.5, 5.0, and default); the sample-size-dependent regularization approach; and for models made with only linear features as well as those using both linear and quadratic feature classes. For the first rule, we used the lowest-presence threshold (LPT) (Pearson et al., 2007; = minimum training presence threshold of Maxent software), which sets the threshold at the lowest value of the prediction for any of the presence localities in the calibration dataset. This yields a binary prediction that includes all pixels at least as suitable (according to the model) as those where the species is known to be present (in the calibration dataset). The values for these thresholds varied by model.

For the second rule, we used a fixed threshold for all models, specifically the value of 10 (out of 100) for the cumulative output. A convenient interpretation exists for the output of cumulative probabilities provided by Maxent, where the expected omission rate for localities of the species is equal to the threshold employed. For example, an ideal model and a threshold of 10 would be expected to yield approximately 10% omission in an independent, unbiased sample of localities of the species. Hence, use of the fixed threshold of 10 is expected to lead to omission levels of approximately 10%.

For the same subset of experiments, we also evaluated model performance by qualitative visual examination of the maps of the species' predicted potential distribution. For each experimental combination examined, we conducted visual inspections of the predictions for the model with the highest AUC (out of the 10 datasets). We observed whether the model successfully predicted the species' potential distribution throughout the full extent of the Cordillera de Mérida, versus just in the centre of the Cordillera (the area from which the calibration localities were drawn). In addition, we noted the model's predictions in the extensive lowlands, the dry, rain-shadowed valley of the Río Chama (which drains to the southwest between two major sections of the Cordillera; Fig. 1), and the highest peaks of the study region (all areas with environments that the species is not known or expected to inhabit).

3. Results

3.1. Environmental comparisons between datasets

Localities from the well-sampled portion of the species' distribution (used for calibration) differed from those from other areas (the evaluation localities) for some variables, but the patterns varied between the tests for central tendency (median) and the inspection of the observed ranges of values (Table 2). Three temperature variables differed significantly in median between the two datasets (p = 0.039–0.049), and five others nearly did (p = 0.053–0.081). None of the other three temperature variables nor or any of the eight pre-

Table 2
Environmental comparisons for three datasets of localities of *C. meridensis*. The first analyses determined whether environmental bias accompanies geographic bias in sampling by comparing values of climatic variables at localities from a portion of the species' range that has received relatively high sampling effort (those used for model calibration; left column) with values at localities from other areas (with low sampling effort; evaluation localities; right column). The second set of analyses determined whether environmental differences occur with random sampling (rarefaction) by comparing values of climatic variables at localities from the portion of the species' range that has received relatively high sampling effort (left column) with the values at 5 localities randomly sampled from that full dataset (centre column). For each dataset, median and minimum–maximum are provided. The results of respective comparisons with the full dataset from the highly sampled area (left column) are indicated in the centre and right columns. The Mann–Whitney *U*-tests for comparisons with the area having received low sampling effort indicate significant or nearly significant differences median for several temperature variables (All tests with $p \leq 0.10$ are indicated in bold; $^{*} = p \leq 0.05$ n.s. = $p > 0.10$). Additionally, minima and maxima that differ markedly between the two respective datasets appear in bold. Following the source data in WorldClim (Hijmans et al., 2005), the units for precipitation variables are mm (except for precipitation seasonality, which is the unitless coefficient of variation), and those for temperature variables are °C multiplied by 10 (except for isothermality, which is a unitless ratio, and temperature seasonality, which is the standard deviation [of the values in °C multiplied by 10] multiplied by 100).

Climatic variable	Dataset		
	High sampled area (all 22 localities)	High sampled area (random sample of 5 localities)	Low sampled area (all 5 localities)
Annual mean temperature	108	76 *n.s.*	**147** (***p* = 0.053**)
	32–185	70–150	**127**–179
Mean diurnal range	97	94 *n.s.*	96 *n.s.*
	87–107	93–103	90–105
Isothermality	80	80 *n.s.*	81 *n.s.*
	77–84	79–83	77–83
Temperature seasonality	519.5	515 *n.s.*	**429*** (***p* = 0.049**)
	467–592	495–**524**	**401**–605
Maximum temperature of warmest month	163.5	128 *n.s.*	**204** (***p* = 0.081**)
	82–245	122–207	**179**–237
Minimum temperature of coldest month	43	11 *n.s.*	**86*** (***p* = 0.039**)
	−30–118	5–83	**63**–111
Temperature annual range	120.5	117 *n.s.*	119 *n.s.*
	112–127	117–124	116–126
Mean temperature of wettest quarter	111.5	79 *n.s.*	**148** (***p* = 0.066**)
	34–189	73–153	**129**–181
Mean temperature of driest quarter	103	72 *n.s.*	**142*** (**0.049**)
	26–182	65–145	**123**–174
Mean temperature of warmest quarter	113	81 *n.s.*	**153** (***p* = 0.061**)
	38–190	75–154	**130**–182
Mean temperature of coldest quarter	101	71 *n.s.*	**139** (***p* = 0.057**)
	25–179	64–143	**121**–174
Annual precipitation	1209	1227 *n.s.*	1360 *n.s.*
	1128–1292	1150–1236	**978–1433**
Precipitation of wettest month	161	168 *n.s.*	167 *n.s.*
	147–179	154–171	**127–210**
Precipitation of driest month	22	21 *n.s.*	31 *n.s.*
	20–41	21–**27**	17–**62**
Precipitation seasonality	50	52 *n.s.*	53 *n.s.*
	36–54	**46**–53	31–58
Precipitation of wettest quarter	415	437 *n.s.*	470 *n.s.*
	380–457	394–443	352–**582**
Precipitation of driest quarter	87	83 *n.s.*	106 *n.s.*
	78–139	81–**100**	64–**191**
Precipitation of warmest quarter	380	408 *n.s.*	351 *n.s.*
	333–431	358–413	330–**524**
Precipitation of coldest quarter	97	95 *n.s.*	116 *n.s.*
	78–170	83–**113**	77–**238**

cipitation ones approached significance ($p = 0.180$–0.975) except precipitation of wettest quarter (0.098). Most of the significant or nearly significant temperature variables showed a common pattern: localities from the well-sampled portion of the species' distribution exhibited a lower median than did those from other areas. Temperature seasonality constituted the lone exception, showing the opposite trend.

In contrast to the results regarding central tendency, inspection of observed ranges (minima and maxima) indicated important differences between the datasets for both some temperature and some precipitation variables (Table 2). For the temperature variables (with one exception, temperature seasonality), the localities from the well-sampled portion of the species' distribution showed a wider climatic range, encompassing the full gamut of conditions found at localities from the other areas. The most striking discrepancies in range for temperature variables corresponded to the minimum values, with the localities from the well-sampled portion of the species' distribution exhibiting much lower minima for most variables. Conversely, for precipitation variables, the localities from the well-sampled portion of the species' distribution showed a narrower range than did the localities from other areas. Here, the more conspicuous differences concerned the maximum values for most variables, although two variables showed substantially lower minima as well.

In comparison with the full dataset from the well-sampled portion of the species' distribution, the first random sample of 5 localities showed no differences in central tendency and only a few notable differences in the observed range of climatic variables (Table 2). No variable differed significantly in median between the full dataset and the random sample of 5 localities ($p = 0.236$–0.950). For every variable, the full dataset showed a wider climatic range than the random sample, but most differences in minima and maxima were small. Substantial differences occurred only for one variable for the minimum value and four variables for the maximum value.

3.2. Threshold-independent evaluations

3.2.1. Comparisons among β regularization values

For experiments with the sample-size-dependent approach and the use of linear features, the highest performance corresponded to intermediate and high values of the β regularization parameter (see Fig. 2a and c for selected results). Regardless of sample size, the lowest average AUC values occurred at the lowest values of the regularization parameter (β = 0.01 and 0.1; average AUC = 0.72–0.77). Average AUC values generally increased with increasing values of β until reaching a plateau at intermediate-to-high values of β (1.5–3.0), where performance was high (average AUC = 0.88–0.90). Average AUC values were slightly lower at β = 5.0 (average AUC = 0.83–0.89). The default regularization settings led to high performance for sample sizes 5 and 10 (average AUC 0.87–0.88), similar to that of the respective optimal regularization setting (the value of the β parameter that yielded the highest average AUC value). In contrast, for sample sizes 15 and 20, the default setting yielded lower performance (average AUC 0.83–0.87) than did the corresponding optimal value. Overall, patterns of performance across values of the β parameter were very similar for all sample sizes (with the exception of default β, see above; see also Section 3.2.4).

Experiments with the sample-size-dependent approach and the use of both linear and quadratic features showed results that were very similar to those with just linear features, but with performance consistently somewhat higher (Fig. 2a and c). Again, intermediate and high values of regularization produced the highest performance. The lowest average AUC values corresponded to the lowest levels of regularization (β = 0.01 and 0.1; average AUC = 0.78–0.82). Average AUC values then increased dramatically and reached a plateau at intermediate-to-high values of β (1.0–5.0; average AUC = 0.90–0.92). In contrast to the results with just linear features, the results showed no tendency towards a decrease in performance for the highest value of β (5.0). These patterns emerged consistently for all sample sizes. The default regularization settings yielded high performance for sample sizes 5 and 10 (average AUC = 0.89–0.91), similar to that of the respective optimal value, but performance slightly lower than that of the corresponding optimal value for sample size 15 and substantially lower for sample size 20 (average AUC = 0.84–0.89). As with the use of just linear features, performance here (with the addition of quadratic features) was very similar across values of β for most sample sizes (again with the exception of default β).

In contrast to experiments with the sample-size-dependent regularization approach, results with the sample-size-independent approach and the use of linear features showed highest performance at intermediate regularization levels, with much lower performance for both low and high values of the β parameter (Fig. 2b and d). The lowest values of that parameter (β = 0.01 and 0.1) showed very low performance (average AUC = 0.74–0.81), but performance increased to maximal levels (average AUC = 0.82–0.90) at β values of 0.5–2.0, depending on the sample size. Performance then dropped precipitously and was very low at the two highest values of β (3.0 and 5.0; average AUC = 0.71–0.75). These patterns remained consistent for all sample sizes. The default regularization settings led to very high performance for all sample sizes (average AUC = 0.88–0.90), similar to that of the respective optimal value. Here, patterns of performance across values of the β parameter were very similar for each sample size; in addition, average AUC for particular values of β was very similar across sample sizes.

Experiments with the sample-size-independent approach and the use of both linear and quadratic features showed patterns similar to those based on just linear features but achieved much higher performance, especially at higher levels of β (Fig. 2b and d). Again, the highest performance occurred at intermediate values of the β regularization parameter, with lower performance at both low and (to a lesser degree) high β values. The lowest values of the regularization parameter (β = 0.01 and usually 0.1) showed low performance (average AUC = 0.79–0.86). Performance again increased substantially to a plateau at intermediate and moderately high regularization levels (generally β = 0.5–3.0, depending on the sample size; average AUC = 0.86–0.91). These patterns remained consistent for all sample sizes. Performance was slightly to moderately lower at β = 5.0, depending on the sample size (for sample sizes 5 and 10, average AUC = 0.83; for sample sizes 15 and 20, average AUC = 0.87–0.88) but did not decrease at these higher levels of regularization as much as it did for the sample-size-independent approach and the use of just linear features. The default regularization settings achieved very high performance for all sample sizes (average AUC = 0.90–0.91), always similar to that of the respective optimal setting. The various sample sizes led to consistent patterns of performance across values of β. In addition, except for β = 5.0, average AUC values were very similar across sample sizes for given β values in these analyses.

3.2.2. Comparisons between regularization approaches

Curves of performance versus regularization level showed different patterns for the sample-size-dependent and sample-size-independent regularization approaches, but each approach led to similar optimal levels of performance. Using linear features, the curves for each approach showed peaks at intermediate-to-high values, with a tendency to decrease at the highest levels of regularization (see Section 3.2.1). Despite this similarity in form, the curves of the two approaches showed a shift relative to each other (Fig. 3a and c), with the peak for the sample-size-independent approach appearing at a lower value of the β parameter than that for the sample-size-dependent approach. For default settings, the sample-size-independent approach generally out-performed the sample-size-dependent one. Patterns of performance between the two approaches for experiments using both linear and quadratic features were similar, but much more muted (Fig. 3b and d).

3.2.3. Comparisons between feature classes

For comparisons between models made with just linear features versus those made using both linear and quadratic features, linear and quadratic features always achieved higher average AUC than did just linear features. This was true for all sample sizes (5–20), all values of the β regularization parameter (including the default settings), and for both regularization approaches (Fig. 2).

3.2.4. Comparisons among sample sizes

Average AUC varied little among the sample sizes for most experiments. In the experiment with linear features and the sample-size-dependent regularization approach (Fig. 4a), average performance differed moderately among sample sizes for the lowest and highest vales of the β regularization parameter, with slightly higher performance for the higher sample sizes. In contrast, the default regularization settings led to lower performance at sample size 20 than at the other three sample sizes. Models made using both linear and quadratic features and the sample-size-dependent approach (Fig. 4c) showed little difference in performance among sample sizes for the fixed values of the regularization parameter but substantially lower performance for sample size 20 for the default regularization settings. In the sample-size-independent regularization approach, performance was very similar among sample sizes across all regularization values (including the default settings) when only linear features were used (Fig. 4b). With that approach and the use of both linear and quadratic features (Fig. 4d), the four

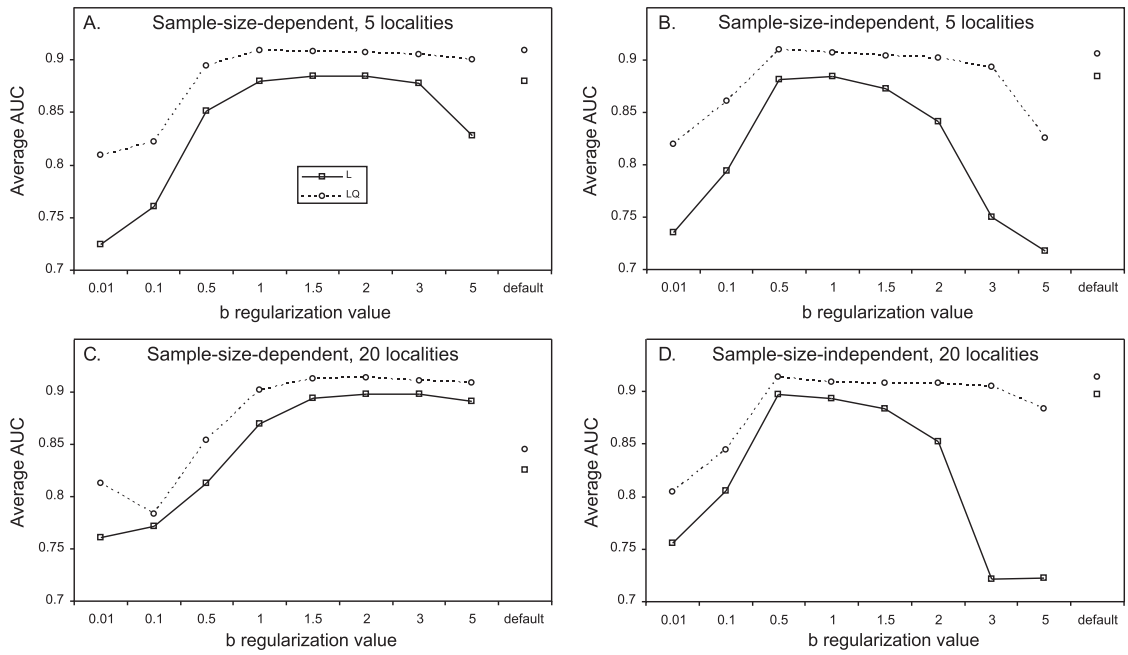

Fig. 2. Results of threshold-independent evaluations of Maxent models of the potential distribution of *C. meridensis*, comparing performance between models made with only linear features versus those made with both linear and quadratic features. Results correspond to analyses with 5 and 20 localities, plotting the Area Under the Curve (AUC) of a Receiver Operating Characteristic plot versus values of the β regularization parameter. For each experiment, the average AUC of the10 datasets indicates the ability of models made with localities from one portion of the species' distribution (which has received relatively high sampling effort) to predict evaluation localities in other parts of its range (where sampling effort has been lower; Fig. 1). Analyses with 10 and 15 localities (not shown) led to patterns similar to those with 5 and 20 localities, respectively. Default values of the β regularization parameter follow – L: 1.0 for 5 localities, 1.0 for 10 localities, 0.8 for 15 localities, 0.6 for 20 localities; LQ: 1.050 for 5 localities, 0.800 for 10 localities, 0.586 for 15 localities, 0.442 for 20 localities.

sample sizes led to very similar performance for the default regularization settings and all but the highest fixed value of the β parameter (5.0), where the higher sample sizes (15 and 20) led to higher average performance.

3.3. Threshold-dependent evaluation

Omission rates and model significance varied greatly across the examined values for the β regularization parameter for mod-

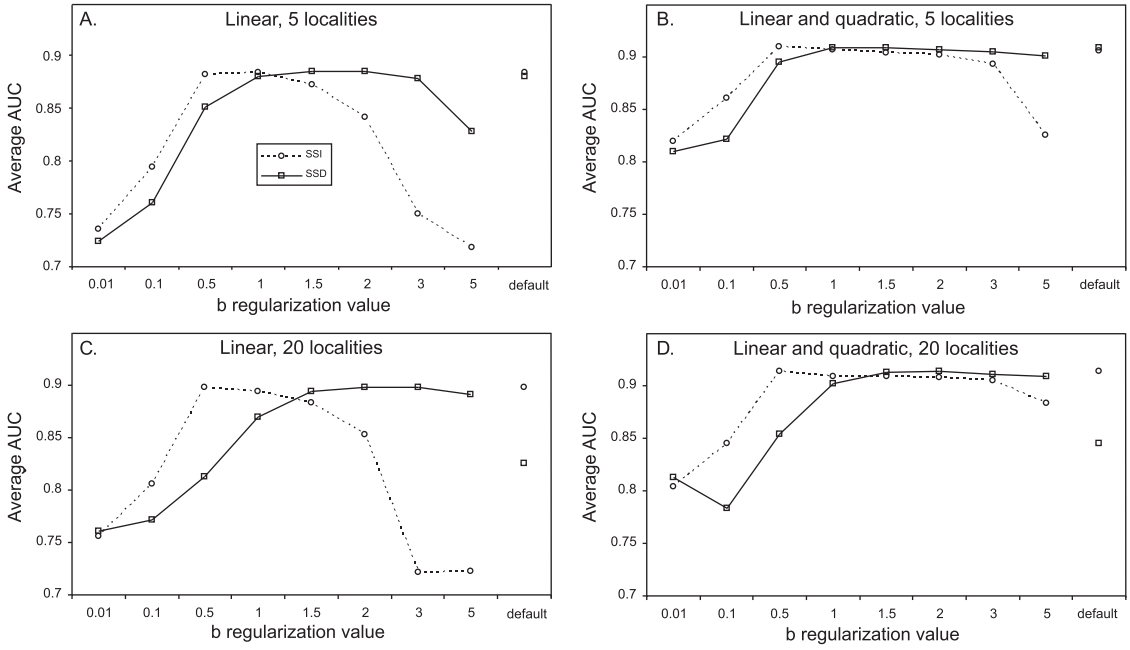

Fig. 3. Results of threshold-independent evaluations of Maxent models of the potential distribution of *C. meridensis*, comparing performance between models made with the sample-size-dependent regularization approach (implemented automatically in Maxent) versus those produced via the sample-size-independent regularization approach. Results correspond to analyses with 5 and 20 localities, plotting the Area Under the Curve (AUC) of a Receiver Operating Characteristic plot versus values of the β regularization parameter. For each experiment, the average AUC of the 10 datasets indicates the ability of models made with localities from a portion of the species' distribution that has received relatively high sampling effort to predict evaluation localities in other parts of its range, where sampling effort has been lower (Fig. 1). Analyses with 10 and 15 localities (not shown) led to patterns similar to those with 5 and 20 localities, respectively. Default values of the β regularization parameter follow – L: 1.0 for 5 localities, 1.0 for 10 localities, 0.8 for 15 localities, 0.6 for 20 localities; LQ: 1.050 for 5 localities, 0.800 for 10 localities, 0.586 for 15 localities, 0.442 for 20 localities.

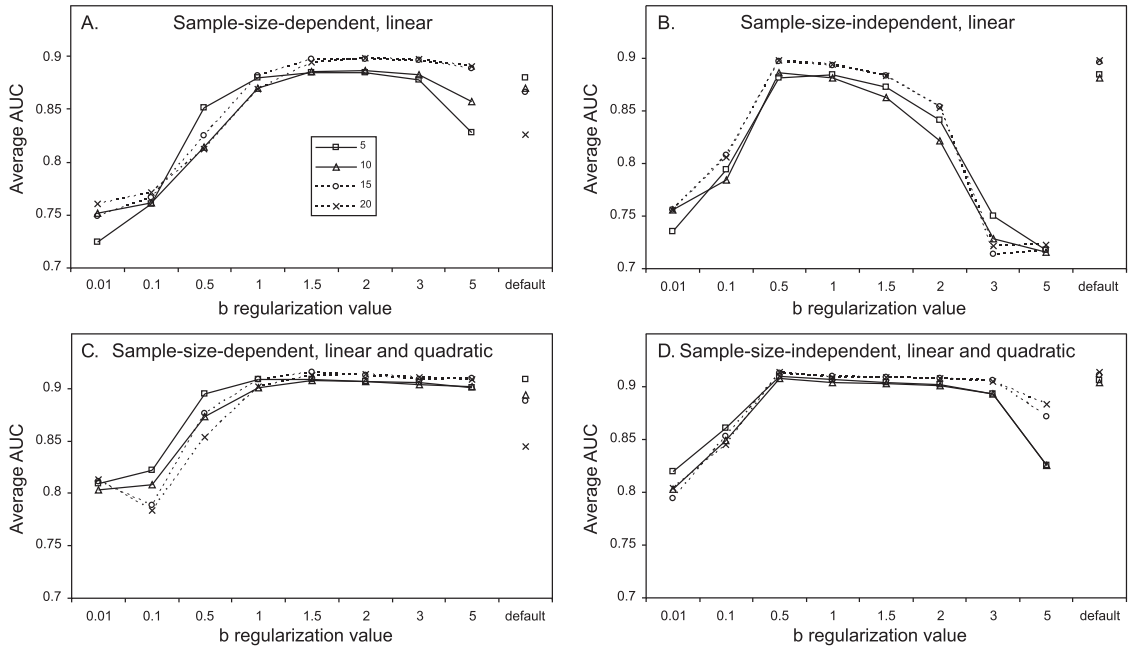

Fig. 4. Results of threshold-independent evaluations of Maxent models of the potential distribution of *C. meridensis*, comparing performance among models made with sample sizes of 5, 10, 15, and 20 localities. Results correspond to analyses with only linear features in comparison with those employing both linear and quadratic features, as well as for both the sample-size-dependent and sample-size-independent regularization approaches. The Area Under the Curve (AUC) of a Receiver Operating Characteristic plot is plotted versus values of the β regularization parameter. Results indicate the average AUC of the10 datasets for each experiment, reflecting the ability of models made with localities from the portion of the species' distribution that has received relatively high sampling effort to predict evaluation localities in other parts of its range, where sampling effort has been lower (Fig. 1). Default values of the β regularization parameter follow – L: 1.0 for 5 localities, 1.0 for 10 localities, 0.8 for 15 localities, 0.6 for 20 localities; LQ: 1.050 for 5 localities, 0.800 for 10 localities, 0.586 for 15 localities, 0.442 for 20 localities.

els made with the sample-size-dependent regularization approach (analyses were not conducted for the sample-size-independent approach). Using a set threshold of 10 (10% predicted omission) as the first thresholding rule, omission rates and significance varied among regularization levels, and patterns in one were usually associated with trends in the other (Table 3). Models made with linear features and low levels of regularization ($\beta = 0.01$) led to high average omission rates for both examined sample sizes (5 and 20). In contrast, omission rates were low for other values of the β parameter (1.5 and 5.0). Default regularization settings yielded low omission rates for models made with 5 localities but high omission rates for models made with 20 localities. Model with high omission rates were never significantly better than random ($P > 0.05$). Inversely, in most cases, low omission rates corresponded to models that were significantly better than random ($P \leq 0.05$). However, models made with 5 localities and $\beta = 5.0$ showed low omission rates (omission rate = 0.06) but rarely were significant (3 out of 10).

For models made with both linear and quadratic features, application of a threshold of 10 showed patterns similar to those for models made with only linear features. Again, low levels of regularization produced high average omission rates (Table 3; $\beta = 0.01$; with both sample sizes 5 and 20). Correspondingly, omission rates were low for other values of the β parameter (1.5 and 5.0). Default regularization settings led to a moderate omission rate for models made with 5 localities, but a high omission rate for models made with 20 localities. Similar to the results above, model significance generally matched patterns in omission rate for both sample sizes. Models with high omission rates never were significantly better than random ($P > 0.05$), but those with low omission rates were consistently so ($P \leq 0.05$). A more complicated situation existed for models made using default regularization settings and sample size 5. Overall, those models showed a moderate average omission rate (0.30), with only 6 of the 10 significantly better than random. Inspection of individual models indicated that they showed either a low omission rate (0.0–0.2) and were significant, or suffered a high omission rate (0.6–0.8) and were not significant.

In contrast, use of the second thresholding rule (LPT) generally led to high omission rates and few significant models. Models made with linear features showed high average omission rates in almost all cases (Table 3). The models made with the highest value of the regularization parameter ($\beta = 5.0$) and 20 localities constituted the one exception, with a moderate omission rate (0.34). Models made with default regularization settings yielded high omission rates for both sample sizes. Very few models made with linear features were significantly better than random. Models made with high regularization ($\beta = 5.0$) and 20 localities constituted the major exception, with all better than random.

Use of the lowest-presence threshold for models made with both linear and quadratic features showed patterns very similar to those for models made with only linear features. Again, average omission rates were usually high, and few models were significant (Table 3; for all regularization values, including default settings). Once more, the models made with the highest value for the regularization parameter ($\beta = 5.0$) and 20 localities constituted the major exception, showing no omission of evaluation localities and all models significantly better than random. The twenty models made with 5 localities and intermediate and high regularization levels ($\beta = 1.5$ and 5.0) produced mixed results. Some models showed low omission (0.0–0.2) and were significant, others had high omission (0.8–1.0) and were not significant, and the rest had intermediate omission rates (0.4–0.6) and varied in being significant or not.

3.4. Visual interpretations

Based on observation of the geographic predictions for the models made using the sample-size-dependent regularization approach, the models varied greatly across different values of the β regularization parameter. These visual interpretations corre-

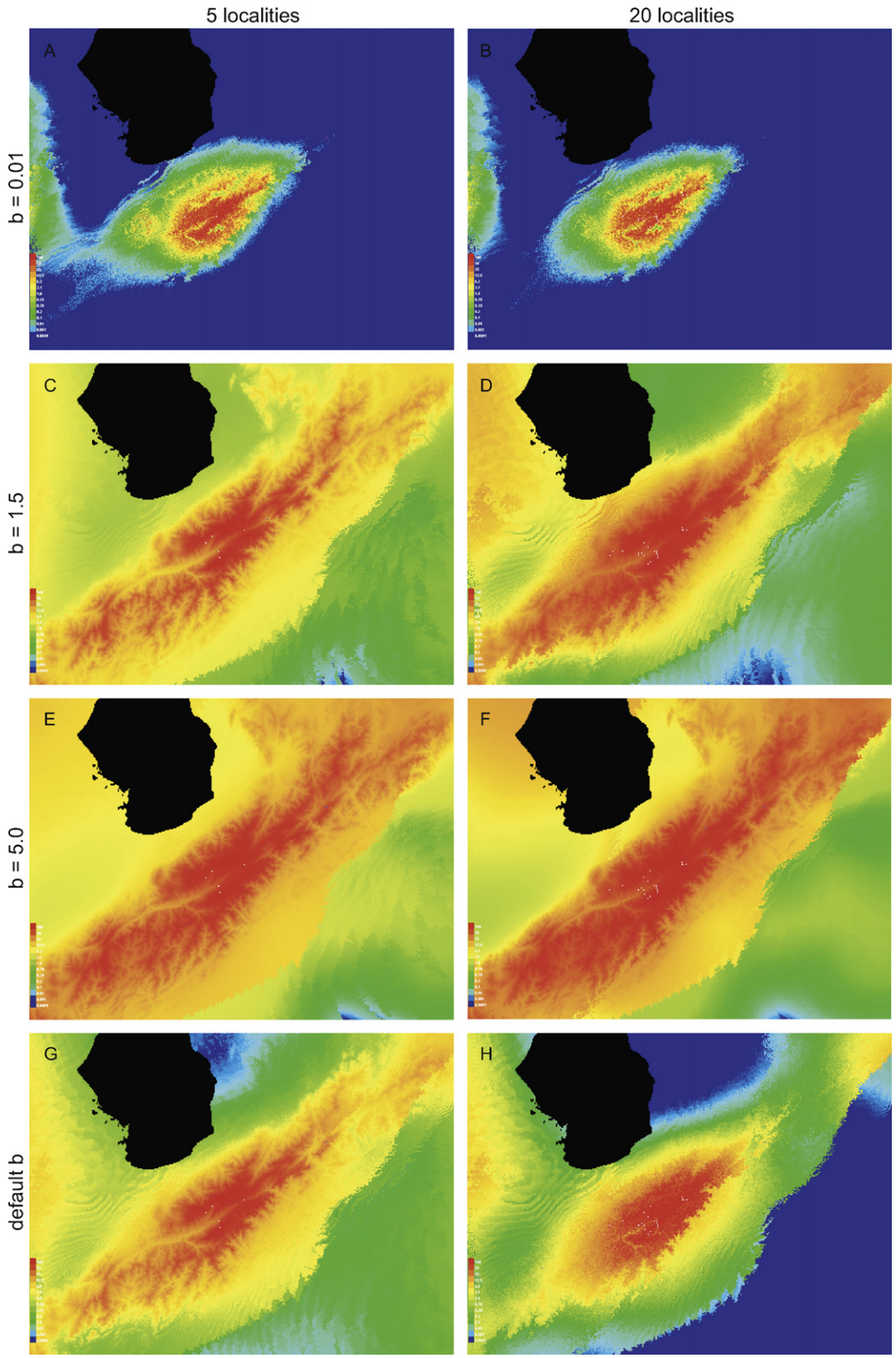

Fig. 5. Selected Maxent models of the potential geographic distribution of *C. meridensis* made using calibration localities from a portion of the species' distribution that has received relatively high sampling effort (Fig. 1) in conjunction with linear features and the sample-size-dependent regularization approach. The predictions show a suitability gradient from low (blue) to high (red) relative environmental suitability (cumulative output from 0 to 100; Phillips et al., 2006). Predictions are presented only for two sample sizes (5 and 20 localities) and four regularization values (0.01, 1.5, 5.0, and default). Out of the 10 datasets for each experiment, predictions correspond to the dataset with the highest AUC value from the threshold-independent evaluations. Default values of the β regularization parameter are 1.0 for 5 localities and 0.6 for 20 localities.

Table 3
Results of threshold-dependent evaluations of selected Maxent models of the potential distribution of *C. meridensis* calibrated using localities from a portion of the species' range that has received relatively high sampling effort. For the sets of 10 datasets for each experiment, models were evaluated by calculating the omission rate of localities from other parts of the species' range, where sampling effort has been lower, and significance was determined by one-tailed binomial probabilities. Continuous predictions of cumulative probability were converted to binary predictions of presence or absence by applying two thresholding rules: the lowest-presence threshold (LPT) and the fixed threshold of 10 (out of 100). These evaluations were conducted for two sample sizes (5 and 20 localities), four values of the β regularization parameter (0.01, 1.5, 5.0, and default), the sample-size-dependent regularization approach, and for models made with the use of only linear features as well as those employing both linear and quadratic features.

β regularization value	Sample size	Lowest-presence threshold				Fixed threshold of 10			
		Linear features		Linear and quadratic features		Linear features		Linear and quadratic features	
		Omission rate	Number of significant models	Omission rate	Number of significant models	Omission rate	Number of significant models	Omission rate	Number of significant models
0.01	5	1.0	0	1.0	0	1.0	0	0.96	0
0.01	20	1.0	0	1.0	0	1.0	0	1.0	0
1.5	5	0.6	4	0.68	4	0	10	0.12	10
1.5	20	0.92	0	0.84	0	0	10	0	10
5.0	5	0.74	1	0.48	5	0.02	3	0.02	10
5.0	20	0.34	10	0	10	0	10	0	10
Default (1.0 for linear; 1.050 for linear and quadratic)	5	0.82	1	0.82	1	0.06	10	0.3	6
Default (0.6 for linear; 0.442 for linear and quadratic)	20	1.0	0	1.0	0	0.62	0	1.0	0

sponded to the model with the highest AUC for each experiment (see Section 2.4). With the use of only linear features, the models gave strikingly different predictions at different β regularization values (Fig. 5). At the lowest value of the β parameter (0.01), the model was overfit for both 5 and 20 localities, with strongly predicted areas corresponding to the area where from which the calibration localities were drawn (Fig. 5a and b). For the intermediate β value of 1.5 and 5 localities, the model (correctly) showed strong prediction for the species throughout the Cordillera de Mérida and indicated low suitability in lowland areas (including the dry Río Chama valley). However, it overpredicted the highest regions (Fig. 5c). For the β value of 1.5 and 20 localities, the model again showed strong prediction throughout the Cordillera de Mérida. However, it seemed to overpredict many of the lowland regions adjacent to the Cordillera as well as the dry valley and the highest peaks (Fig. 5d). For the highest value of β (5.0) and 5 localities, the model indicated a strong prediction in highland regions throughout the study area, but the discrimination was not especially sharp (Fig. 5e). The prediction was too strong in many of the lowland regions and in the dry valley, and even more so in the highest regions. The model for 20 localities and a β value of 5.0 was similar to that with 5 localities but provided even less discrimination in the strength of the prediction between highland and lowland areas (Fig. 5f). The model made with default regularization settings and 5 localities gave a strong prediction throughout the Cordillera and a weak prediction in lowland regions and the dry valley. However, it overpredicted the highest regions (Fig. 5g). Use of the default regularization settings with 20 localities produced a model severely overfit to the area of the calibration localities (Fig. 5h).

For the analyses using both linear and quadratic features, the models again varied greatly among regularization values, but the predictions were generally better than those made with linear features alone (Fig. 6). For the lowest β value (0.01), use of 5 and 20 localities led to severely overfit models, with strongly predicted areas restricted to the centre of the species' known distribution (Fig. 6a and b). For the intermediate level of β (1.5) and 5 localities, the model showed strong prediction throughout most of the Cordillera de Mérida, as well as sharp discrimination between those areas versus areas of low suitability in the surrounding lowlands, the dry Río Chama valley, and the highest regions (Fig. 6c). For the β value 1.5 and 20 localities, the model had the same general characteristics as the one made with 5 localities, but the prediction was not as sharp (Fig. 6d). For the highest β value (5.0) and 5 localities, the model indicated strong prediction throughout the Cordillera de Mérida and weak prediction in the surrounding lowlands and in the dry valley. However, it showed an overly strong prediction in the highest regions (Fig. 6e). The model for 20 localities and a β value of 5.0 had similar characteristics to that for 5 localities but was not as sharp (Fig. 6f). The model produced with default regularization settings and 5 localities successfully indicated strong prediction across most of the Cordillera and weak prediction in the surrounding lowlands, the dry valley, and the highest regions (Fig. 6g). In contrast, the model made using default regularization settings and 20 localities was overfit to the centre of the species' known range (Fig. 6h).

4. Discussion

4.1. Environmental comparisons between datasets

The analyses comparing the environmental conditions at localities from the well-sampled portion of the species' range with data from the localities from other areas document that environmental biases indeed accompany the geographic biases in sampling (Table 2). Regarding most temperature variables, localities from the well-sampled portion of the species' distribution (used for calibration) encompass the full range of conditions found at localities in other areas (evaluation localities). However, for several temperature variables, the calibration localities differed in median (significantly or nearly so) from the evaluation localities and/or showed much lower minima. This derives from the fact that many of the localities from the well-sampled portion of the species' range correspond to higher elevations than any of those from the other areas. Because the calibration localities include the range of conditions that the species occupies in the evaluation localities for these variables, this bias should not be insurmountable for modelling the species' potential distribution.

On the contrary, for all precipitation variables, the evaluation localities correspond to conditions that are more extreme than the range of those for localities in the well-sampled portion of the species' distribution. Although no significant difference in median existed between the two datasets for any precipitation variable,

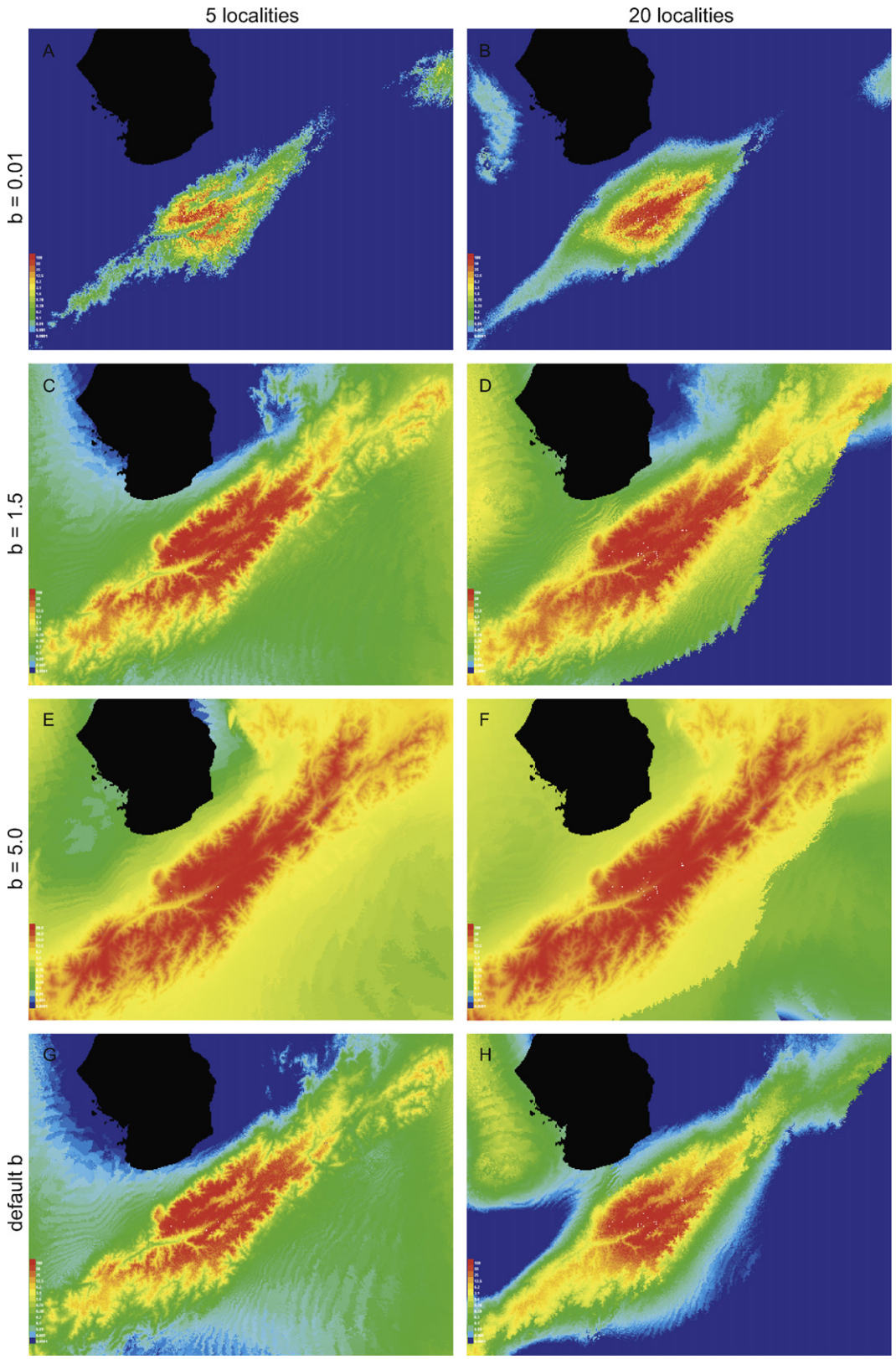

Fig. 6. Selected Maxent models of the potential geographic distribution of *C. meridensis* made using calibration localities from a portion of the species' distribution that has received relatively high sampling effort (Fig. 1) in conjunction with both linear and quadratic features and the sample-size-dependent regularization approach. The predictions show a suitability gradient from low (blue) to high (red) relative environmental suitability (cumulative output from 0 to 100; Phillips et al., 2006). Predictions are presented only for two sample sizes (5 and 20 localities) and four regularization values (0.01, 1.5, 5.0, and default). Out of the 10 datasets for each experiment, predictions correspond to the dataset with the highest AUC value from the threshold-independent evaluations. Default values of the β regularization parameter are 1.050 for 5 localities and 0.442 for 20 localities.

the evaluation localities frequently exhibited substantially higher maxima and, occasionally, notably lower minima. Because this environmental bias relates to conditions at evaluation localities not found in the calibration dataset, it represents the kind of differences to which a model can be overfit.

In contrast, little environmental difference existed between the full dataset from the well-sampled portion of the species' distribution and the examined random sample of 5 localities from it. As expected (since localities were chosen randomly), the two did not differ in median for any variable (Table 2). Similarly, the range of conditions present in the full dataset *must* include all those present in a sample taken from it. Interestingly, the random sample only differed from the full dataset substantially in minimum or maximum a few times. These instances represent an underestimation of the species' niche in the smaller sample due to random noise. In this case, even a small random sample appears to contain virtually the same environmental information as the full dataset. Models made with such a sample are expected to predict a slightly reduced portion of the species' niche and potential distribution. Although we did not conduct similar tests for the larger random samples (of 10, 15, and 20 localities), we expect that the present (non-significant) results for the tests of median would be the same for such samples as well, and that differences in minima and maxima between the samples and the full dataset would be less prominent with increased sample size.

4.2. Comparisons among regularization values

In the threshold-independent evaluations, performance varied greatly among regularization values. Optimal values of the β regularization parameter differed between the sample-size-dependent and sample-size-independent analyses (Figs. 2–4). For the sample-size-dependent approach, the highest performance corresponded to intermediate and high regularization values. In contrast, for the sample-size-independent one, the highest performance occurred at intermediate levels of regularization (but always lower ones than those that were optimal for the sample-size-dependent approach). The performance of the default regularization settings varied between the two approaches and according to sample size. For the sample-size-dependent approach, the default settings showed performance as good as the optimal values for small sample sizes (5 and 10) but lower performance for sample sizes 15 and 20, where the additional localities likely reinforced the bias in sampling (see Section 4.5). In contrast, for the sample-size-independent approach, the default settings were as good as the optimal regularization values for all sample sizes.

Similarly, the threshold-dependent evaluations indicated that omission rates differed substantially among values of the β regularization parameter using both thresholding rules (Table 3). For the fixed threshold of 10, omission rates were high with the lowest value of the regularization parameter ($\beta = 0.01$) and low for other values ($\beta = 1.5$ and 5.0). This makes intuitive sense, because models with only weak protection against overfitting are expected to produce overly restrictive predictions (with high omission of evaluation localities). In contrast, with the lowest-presence threshold, all regularization levels ($\beta = 0.01$, 1.5, and 5.0) showed generally high omissions rates. This probably relates to the fact that the threshold that results from this rule typically varies according to sample size. With larger sample sizes of calibration localities (used both to build the model and determine the threshold), the data should be more representative of the full range of the species' environmental tolerances, leading to a lower threshold and correspondingly more-extensive binary geographic prediction for the species. In contrast, with small sample sizes (that are less likely to encompass the species' full tolerances), the LPT rule is likely to lead to an overly restrictive binary prediction, as here. The default regularization settings led to high omission rates for most sample size and threshold-rule combinations. Across all of these analyses, significance generally corresponded to omission rates, with models showing low omission being significantly better than random, and models with high omission rates not achieving significance.

Results of the visual interpretations indicated that model quality varied tremendously among the regularization values and, to a lesser degree, among the sample sizes (Figs. 5 and 6; see Section 4.5). For models made with just linear features (regardless of sample size), the use of a low value of the regularization parameter ($\beta = 0.01$) led to overfit models, with the areas of strongest predictions concentrated around the calibration localities. The intermediate value of regularization ($\beta = 1.5$) yielded good models whose main flaw was slight overprediction of the highest regions. Finally, models made with the highest level of regularization ($\beta = 5.0$) corresponded to underfit (diffuse or blurry) predictions. These results match theoretical expectations (Phillips et al., 2006; Phillips and Dudík, 2008). Notably, the default regularization settings produced a good model at sample size 5 (similar to the prediction for $\beta = 1.5$), but that at sample size 20 was overfit; this pattern of poorer models at a larger sample size likely results from a reinforcement of spatial bias in sampling (see Section 4.6). The models made with both linear and quadratic features showed patterns similar to those for models made with only linear features, but they tended to be sharper, with better discrimination between suitable and unsuitable areas.

4.3. Comparisons between regularization approaches

Performance varied between the two regularization approaches (sample-size-dependent and sample-size-independent) according to the level of the β regularization parameter and the sample size (Fig. 3). Whereas the sample-size-dependent analyses yielded poor performance at low regularization values but generally good performance at both intermediate and high levels of regularization, the sample-size-independent approach showed lower performance at both low and high regularization values (and high performance at intermediate ones). For most regularization values, performance was similar across all sample sizes. Patterns of performance between the two regularization approaches for linear and quadratic features generally matched those found in the analyses employing just linear features, but the trends were not as marked. In sum, although equally high optimal performance was achieved for each regularization approach, the particular regularization value that led to the best models differed between the approaches and was higher for the sample-size-dependent one. Furthermore, the sample-size-independent approach more-consistently led to high-quality models with the default settings.

4.4. Comparisons between feature classes

The results for comparisons between the analyses based on linear features alone versus those for linear and quadratic features together varied according to the type of evaluation. Threshold-independent evaluation indicated consistently higher average AUC for models made allowing the use of both linear and quadratic features than for those calibrated with just linear ones (Fig. 2). In contrast, in the threshold-dependent evaluations, most experiments yielded similar omission rates and numbers of models that were significant (Table 3). Finally, for most regularization values, the visual comparisons indicated sharper (more discriminating) predictions for models made using both feature classes than for those produced using just linear features (Figs. 5 and 6). Overall, the use of both linear and quadratic feature classes was preferable.

4.5. Comparisons among sample sizes

Threshold-independent evaluations indicated that performance varied little among sample sizes for most experiments (Fig. 4). Surprisingly, however, the highest sample size led to lower average performance using the default regularization settings and the sample-size-dependent approach. This was true both for experiments using linear features alone and for those employing both linear and quadratic features. In contrast, we observed no noteworthy differences in the sample-size-independent analyses.

With only a few exceptions, the threshold-dependent evaluations indicated similar performance between sample sizes 5 and 20 (Table 3). One exception was for the fixed threshold of 10, where the default regularization settings led to higher omission rates (and fewer significant models) for sample size 20 than for sample size 5. Conversely, the other concerned the lowest-presence threshold, where the high regularization value ($\beta = 5.0$) led to lower omission (and more significant models) at sample size 20 than at sample size 5.

Visual interpretations indicated that the effect of sample size on model quality depended on the level of regularization. First of all, sample size did not affect the models much at low regularization values (Figs. 5 and 6). In contrast, at intermediate and high levels of regularization, models made with 5 localities yielded predictions that were sharper (more discriminating) than those made with 20 localities. Finally, models made with the default regularization settings yielded good-to-excellent predictions at sample size 5 but overfit ones at sample size 20. Below, we discuss these surprising results of equivalent or lower performance with increased sample size (in threshold-independent, threshold-dependent, and visual evaluations).

4.6. Conclusions and recommendations

Our experiments lead to several recommendations for implementation of Maxent and for future methodological research. Although the following conclusions may apply to other situations as well, the present results document them for occurrence data with few and highly biased localities. Foremost, models can vary greatly depending on the regularization employed, and regularization values other than the default settings may lead to higher performance. However, this appears to depend on which regularization approach is used. With the sample-size-dependent regularization approach (that implemented automatically by Maxent), intermediate (and sometimes high) regularization values led to the best performance, often better than that with the default (generally lower) values (similar to Elith et al., 2010). In contrast, under the sample-size-independent regularization approach, intermediate values led to the best performance, similar to that achieved based on the default values.

Models also vary according to the feature classes employed. Models made when allowing both linear and quadratic features generally performed better than those considering just linear features. This suggests that researchers can use both feature classes with few and highly biased localities, as long as appropriate regularization is employed. Species-specific tuning of feature classes may increase model performance substantially. The use of other feature classes, such as product, threshold, and hinge features, also should be considered (Phillips and Dudík, 2008). In general, when allowing Maxent to consider additional feature classes (and, hence, increasing the possibility for higher model complexity), even more care must be given to ensuring that the model is not overfit. When more feature classes are employed, theory suggests that higher regularization will be necessary (Phillips and Dudík, 2008).

Surprisingly, performance did not vary much between sample sizes 5 and 20 with these biased data. With appropriate regularization, even extremely few localities (5) achieved high performance. However, with the addition of more localities, performance sometimes increased slightly but occasionally decreased considerably. The failure to perform considerably better with more localities (contrary to expectations) may be due to the fact that these localities were clumped in space. Furthermore, the instances of decreased performance likely derive from a reinforcement of the sampling bias reflected in the occurrence data (Hortal et al., 2008; Veloz, 2009). When adding more localities (with similar environmental information; see Sections 3.1 and 4.1), the information regarding the species' environmental tolerances was not substantially increased. However, as a by-product of increased sample size without additional independent environmental information (i.e., with the inclusion of additional biased localities), the effective level of regularization decreases. This occurs because, for a given feature class, regularization equals the β value multiplied by the standard deviation and then divided by the square root of the sample size (i.e., regularization = $\beta \times$ standard deviation/(sample size)$^{1/2}$ = $\beta \times$ standard error; see Section 2.2). With the inclusion of additional localities with similar environmental information, the standard deviation of samples of a given variable stays approximately the same but the sample size increases, leading to division by a larger number (the square root of the sample size); together, these cause β to be multiplied by a smaller standard error. In effect, when biased samples violate the independence of localities (as random samples from the species' distribution; see Phillips et al., 2006), decreased regularization occurs, promoting overfitting.

The present results indicate that even with data that are highly biased geographically (and to a lesser extent, environmentally), high-quality models can be made with few localities when appropriate regularization is used. Overall, changes in regularization settings and feature classes led to much more variation in model output than did the number of localities used to make the model. Models sometimes can be improved by including more localities, but researchers should use caution, especially if localities are clumped and known or suspected to result from biased sampling, as higher sample sizes can exacerbate problems associated with these issues. When clumping exists and sampling bias is suspected as the causal agent, spatial filtering of localities (or simple rarefaction, as here) before modelling may ameliorate its negative effects (Anderson and Raza, 2010; M. Shcheglovitova and R.P. Anderson, unpublished data).

With highly biased occurrence data or when sampling bias is unknown but suspected to be substantial, we recommend tuning experiments that vary the feature classes employed as well as the level of regularization, in order to determine the optimal settings for the species and environmental data at hand (see also Elith et al., 2010). The default settings of Maxent were established previously based on tuning experiments with random splits of localities into calibration and evaluation datasets; hence, those experiments could not detect overfitting to any sampling bias present in the data. Because of this, they likely overestimated performance and led to suggested regularization values that are lower than the ones that would be optimal for typical datasets that suffer from the effects of sampling bias.

Two possible solutions seem feasible. One strategy could be to use default regularization settings with the sample-size-independent approach, which together led to high performance in this study. However, this drastic departure from all implementations of Maxent to-date would require substantial research before general usage. Alternatively, using the sample-size-dependent approach, species-specific tuning with spatially distant evaluation localities – as here – should increase the robustness of models to sampling bias (A. Radosavljevic, D.M. Thomas, and R.P. Anderson, unpublished data). Appropriate regularization settings may depend not only on the number of localities (and the level of

bias in them) but also the feature classes employed and the number and characteristics of the environmental variables used, an issue not yet considered in the literature. Future research should address whether varying the regularization multiplier present in later releases of Maxent (which preserves the relative strengths of β across feature classes) is sufficient to identify optimal settings. Such a situation would greatly simplify tuning exercises, eliminating the need to tune β individually for each feature class. Finally, when conducting tuning experiments, researchers should take into account principles for selecting an appropriate study region where biotic interactions and dispersal limitations do not (or are less likely to) lead to non-equilibrium distributions that violate assumptions of modelling (Anderson and Raza, 2010).

Future methodological experiments examining these factors with Maxent are necessary to reach recommendations applicable to a wide range of situations. Specifically, analyses should be conducted using multiple species with varied sample sizes, a range of bias in sampling effort, and differing number and composition of environmental variables. Ideally, such research would yield general conclusions and recommendations relating bias, sample size, and optimal regularization values and feature classes. Alternatively, species-specific tuning of feature class and regularization values may remain necessary, in which case automation of that process would be of great benefit.

These findings are relevant to the use of niche models in myriad areas. Foremost, perhaps, studies applying a model to another time period (e.g., after climatic change) or region (e.g. for invasive species) require high transferability (Araújo and Rahbek, 2006; Williams and Jackson, 2007). Similarly, assessments of niche evolution assume models with high generality (Wiens and Graham, 2005; Warren et al., 2008). However, even studies not involving transferability desire general models that are not overfit; these include conservation assessments, macroecological analyses, predictions regarding zoonotic diseases, and applications to many other areas of basic and applied environmental science (Wiens and Graham, 2005; Kozak et al., 2008; Peterson et al., in press). Among these, consideration of the effects of bias and low sample size seems especially germane for assessments of the conservation status of rare species, for which only few (and likely biased) localities may exist (Anderson and Martínez-Meyer, 2004). Although the present conclusions correspond directly to the commonly used Maxent, the same principles hold relevancy for other modelling techniques as well.

Acknowledgements

This research was possible via funding from the U.S. National Science Foundation (NSF DEB-0717357) and the City College of the City University of New York (City College Academy for Professional Preparation, Department of Biology, Office of the Dean of Science, and Office of the Provost). Martha R. Perez and Tiffany L. Johnson assisted in determining latitude and longitude for localities of *C. meridensis*. Steven J. Phillips and Miroslav Dudík answered many queries regarding Maxent. Catherine H. Graham, Eliécer E. Gutiérrez, Aleksandar Radosavljevic, Mariano Soley-G., Sara Varela, Dan L. Warren, and an anonymous reviewer provided insightful comments.

Appendix A. Gazetteer of localities of *Cryptotis. meridensis* used in this study (from Woodman, 2002). Boldface type indicates the place to which geographic coordinates correspond. The source for the coordinates follows the latitude and longitude

VENEZUELA: MÉRIDA: **Area 37, near Middle Refugio, 2 km S, 5.5 km E of Tabay**, 2630 m, 8°37′N, 71°02′W (DCN, 1977b; Woodman, 2002; but see Handley, 1976); **Area 51, near Middle Refugio, 1 km S, 5 km E of Tabay**, 2640 m, 8°38′N, 71°02′W (DCN, 1977b; Woodman, 2002; not Handley, 1976); El Tambor [= **Páramo Tambor**], 8800 ft [2682 m], 8°36′N, 71°24′W (DCN, 1977b; Paynter, 1982); La Aguada, near Laguna **La Fría**, 7 km SE of Mérida, 3600 m, 8°32′N, 71°06′W (DCN, 1977b; not Paynter, 1982); **La Coromoto**, 4 km S, 6.5 km E of Tabay, 3160–3175 m, 8°36′N, 71°01′W (Handley, 1976; DCN, 1977b); La Montaña, **3.1 km SE of Mérida**, 2250 m, 8°34′N, 71°07′W (DCN, 1977b); **Mérida**, 2165 m, 8°36′N, 71°08′W (DCN, 1977b; Paynter, 1982); **Monte Zerpa** cloud forest, Sierra del Norte de La Culata, 2000–2800 m, 8°37′N, 71°10′W (DCN, 1977b; Díaz de Pascual and de Ascenção, 2000); Montes de **La Culata**, 2000 m, 8°45′N, 71°05′W (DCN, 1977a; Paynter, 1982) and **Páramo de [La] Culata**, Río Mucujún, 9000 ft [2743 m], same coordinates (DCN, 1977a; Paynter, 1982); Montes del **Valle**, 2125–2165 m, 8°40′N, 71°06′W (DCN, 1977a; Paynter, 1982); **near La Mucuy**, 2.9 km E of Tabay, 2450 m, 8°38′N, 71°03′W (DCN, 1977b; Woodman, 2002; not Paynter, 1982); near **Laguna Mucubají**, 3.25 km ESE of Apartaderos, 3600 m, 8°48′N, 70°50′W (DCN, 1976b; not Paynter, 1982); near **Laguna Negra**, 5.75 km ESE of Apartaderos, 3500 m, 8°48′N, 70°48′W (DCN, 1976b; not Paynter, 1982) and **Río Chama, site E–I**, 3950 m, same coordinates (DCN, 1976b; Durant and Díaz, 1995); **near Laguna Verde**, 7.5 km E, 6 km S of Tabay, 3533–3545 m, 8°35′N, 71°01′W (DCN, 1977b; but see Handley, 1976); near Loma Redonda cable car station, **8.8 km SE of Mérida**, 4100 m, 8°33′N, 71°05′W (DCN, 1977b); **near Santa Rosa, 1 km N, 2 km W of Mérida**, 1980–1990 m, 8°37′N, 71°09′W (Handley, 1976; DCN, 1977b; but see Woodman, 2002); **Páramo de los Conejos**, 9600 ft [2926 m], 8°38′N, 71°18′W (H.E. Anthony field notes, 1932, Department of Mammalogy archives, American Museum of Natural History; DCN, 1977a; not Paynter, 1982 or Woodman, 2002); **Páramo de Mucubají**, Distrito Rangel, 3420–3800 m, 8°48′N, 70°49′W (DCN, 1976b; Durant and Péfaur, 1984); Páramo Tambor [at Hacienda **La Carbonera** near head of a western branch of Río Guachi], 8°38′N, 71°22′W (DCN, 1977b; Paynter, 1982; not Woodman, 2002); **right bank of the Río Santo Domingo**, near Quebrada de la Virgen, km 36 of the Apartadero–Barínas Highway, Distrito Rangel, 1640 m, 8°53′N, 70°39′W (DCN, 1976b); **Río Chama, site E-II**, 1890 m, 8°36′N, 71°10′W (DCN, 1977b; Durant and Díaz, 1995); **Río Motatán, site E–I**, 3890 m, 8°49′N, 70°50′W (DCN, 1976b; Durant and Díaz, 1995; not Woodman, 2002); **Río Mucujún** [see also **Páramo San Antonio**; Paynter, 1982], 9000–12,500 ft [2743–3810 m], 8°42′N, 71°08′W (DCN, 1977a; not Paynter, 1982 or Woodman, 2002); TÁCHIRA: **Río Escalante, site E–I**, Páramo de Mariño, 2600 m, 8°16′N, 71°56′W (DCN, 1976a; Durant et al., 1994; not Durant and Díaz, 1995 or Woodman, 2002); TRUJILLO: **Parque Nacional General Cruz Carrillo (Guaramacal)**, páramo, ca. 3100 m [for Locality 6 of Soriano et al., 1990], 9°15′N, 70°11′W (Soriano et al., 1990); **Parque Nacional General Cruz Carrillo (Guaramacal)**, selva nublada montana alta, ca. 2470 m [for Locality 4 of Soriano et al., 1990; but see Woodman, 2002], 9°15′N, 70°12′W (Soriano et al., 1990; note SW corner of their map should read 70°17′W); **Río Motatán, site E-II**, 1890 m, 9°05′N, 70°39′W (SAGCN, 1996; Durant and Díaz, 1995).

References

Anderson, R.P., Raza, A., 2010. The effect of the extent of the study region on GIS models of species geographic distributions and estimates of niche evolution: preliminary tests with montane rodents (genus *Nephelomys*) in Venezuela. Journal of Biogeography 37, 1378–1393.

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

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

Anderson, R.P., Peterson, A.T., Gómez-Laverde, M., 2002b. Using niche-based GIS modeling to test geographic predictions of competitive exclusion and competitive release in South American pocket mice. Oikos 98, 3–16.
Araújo, M.B., Rahbek, C., 2006. How does climate change affect biodiversity? Science 313, 1396–1397.
Araújo, M.B., Pearson, R.G., Thuiller, W., Erhard, M., 2005. Validation of species–climate impact models under climate change. Global Change Biology 11, 1504–1513.
DCN, 1976. Hoja 5840 (La Grita), escala 1:100.000. Dirección de Cartografía Nacional, Ministerio de Obras Públicas, Caracas.
DCN, 1976. Hoja 6042 (Timotes), escala 1:100.000. Dirección de Cartografía Nacional, Ministerio de Obras Públicas, Caracas.
DCN, 1977. Hoja 5942 (La Azulita), escala 1:100.000. Dirección de Cartografía Nacional, Ministerio del Ambiente y de los Recursos Naturales Renovables, Caracas.
DCN, 1977. Hoja 5941 (Mérida), escala 1:100.000, edición preliminar. Dirección de Cartografía Nacional, Ministerio del Ambiente y de los Recursos Naturales Renovables, Caracas.
Díaz, A., Péfaur, J.E., Durant, P., 1997. Ecology of South American páramos with emphasis on the fauna of the Venezuelan páramos. In: Wielgolaski, F.E. (Ed.), Ecosystems of the World: Polar and Alpine Tundra. Elsevier, New York, pp. 263–310.
Díaz de Pascual, A., de Ascenção, A.A., 2000. Diet of the cloud forest shrew *Cryptotis meridensis* (Insectivora: Soricidae) in the Venezuelan Andes. Acta Theriologica 45, 13–24.
Durant, P., Díaz, A., 1995. Aspectos de la ecología de roedores y musarañas de las cuencas hidrográficas Andino-Venezolanas. Caribbean Journal of Science 31, 83–93.
Durant, P., Péfaur, J.E., 1984. Sistemática y ecología de la musaraña de Mérida. Soricidae: Insectivora. *Cryptotis thomasi*. Revista de Ecología, Conservación y Ornitología Latinoamericana 1, 3–14.
Durant, P., Díaz, A., Díaz de Pascual, A., 1994. Pequeños mamíferos alto-andinos Mérida-Venezuela. Revista Forestal Latinoamericana 14–94, 103–131.
Elith, J., Burgman, M., 2002. Predictions and their validation: rare plants in the central highlands, Victoria, Australia. In: Scott, J.M., Heglund, P.J., Morrison, M.L., Haufler, J.B., Raphael, M.G., Wall, W.A., Samson, F.B. (Eds.), Predicting Species Occurrences: Issues of Accuracy and Scale. Island Press, Washington, DC, pp. 303–313.
Elith, J., Graham, C.H., Anderson, R.P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J.R., Leahmann, A., Li, J., Lohmann, L.G., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J.M., Peterson, A.T., Phillips, S.J., Richardson, K.S., Scachetti-Pereira, R., Schapire, R.E., Soberón, J., Williams, S., Wisz, M.S., Zimmermann, N.E., 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29, 129–151.
Elith, J., Kearney, M., Phillips, S., 2010. The art of modelling range-shifting species. Methods in Ecology and Evolution 1, 330–342.
Fielding, A.H., 2002. What are the appropriate characteristics of an accuracy measure? In: Scott, J.M., Heglund, P.J., Morrison, M.L., Haufler, J.B., Raphael, M.G., Wall, W.A., Samson, F.B. (Eds.), Predicting Species Occurrences: Issues of Accuracy and Scale. Island Press, Washington, DC, pp. 271–280.
Fielding, A.H., Bell, J.F., 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation 24, 38–49.
Graham, C.H., Ferrier, S., Huettman [sic], F., Moritz, C., Peterson, A.T., 2004. New developments in museum-based informatics and application in biodiversity analysis. Trends in Ecology and Evolution 19, 497–503.
Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. Ecological Modelling 135, 147–186.
Guisan, A., Edwards Jr., T.C., Hastie, T., 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecological Modelling 157, 89–100.
Guisan, A., Graham, C.H., Elith, J., Huettmann, F., the NCEAS Species Distribution Modelling Group, 2007. Sensitivity of predictive species distribution models to change in grain size. Diversity and Distributions 13, 332–340.
Handley Jr., C.O., 1976. Mammals of the Smithsonian Venezuelan Project. Brigham Young University Science Bulletin, Biological Series 20 (5), 1–91.
Hernandez, P.A., Graham, C.H., Master, L.L., Albert, D.L., 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography 29, 773–785.
Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A., 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25, 1965–1978.
Hortal, J., Jiménez-Valverde, A., Gómez, J.F., Lobo, J.M., Baselga, A., 2008. Historical bias in biodiversity inventories affects the observed environmental niche of the species. Oikos 117, 847–858.
Hutterer, R., 2005. Order Soricomorpha. In: Wilson, D.E., Reeder, D.M. (Eds.), Mammal Species of the World: A Taxonomic and Geographic Reference, volume 1, 3rd ed. Johns Hopkins University Press, Baltimore, pp. 220–311.
Kozak, K.H., Graham, C.H., Wiens, J.J., 2008. Integrating GIS-based environmental data into evolutionary biology. Trends in Ecology and Evolution 23, 141–148.
Lobo, J.M., Jiménez-Valverde, A., Real, R., 2008. AUC: a misleading measure of the performance of predictive distribution models. Global Ecology and Biogeography 17, 145–151.
Minitab, 2003. Minitab, Release 14.1 for Windows. Minitab, Inc, State College, PA.
Osborne, P.E., Foody, G.M., Suárez-Seoane, S., 2007. Non-stationarity and local approaches to modelling the distributions of wildlife. Diversity and Distributions 13, 313–323.
Paynter Jr., R.A., 1982. Ornithological Gazetteer of Venezuela. Museum of Comparative Zoology. Harvard University, Cambridge, MA.
Pearson, R.G., Raxworthy, C.J., Nakamura, M., Peterson, A.T., 2007. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. Journal of Biogeography 34, 102–117.
Peterson, A.T., 2003. Predicting the geography of species' invasions via ecological niche modeling. Quarterly Review of Biology 78, 419–433.
Peterson, A.T., Papeş, M., Eaton, M., 2007. Transferability and model evaluation in ecological niche modeling: a comparison of GARP and Maxent. Ecography 30, 550–560.
Peterson, A.T., Papeş, M., Soberón, J., 2008. Rethinking receiver operating characteristic analysis applications in ecological niche modeling. Ecological Modelling 213, 63–72.
Peterson, A.T., Soberón, J., Pearson, R.G., Anderson, R.P., Martínez-Meyer, E., Nakamura, M., Araújo, M.B. Ecological niches and geographic distributions. Monographs in Population Biology. Princeton University Press, Princeton, in press.
Phillips, S.J., Dudík, M., 2008. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography 31, 161–175.
Phillips, S.J., Anderson, R.P., Schapire, R.E., 2006. Maximum entropy modeling of species geographic distributions. Ecological Modelling 190, 231–259.
Phillips, S.J., Dudík, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J., Ferrier, S., 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecological Applications 19, 181–197.
SAGCN, 1996. Mapa del Estado Trujillo, edición 4. Servicio Autónomo de Geografía y Cartografía Nacional, Ministerio del Ambiente y de los Recursos Naturales Renovables, Caracas.
Scott, J.M., Heglund, P.J., Morrison, M.L., Haufler, J.B., Raphael, M.G., Wall, W.A., Samson, F.B. (Eds.), 2002. Predicting Species Occurrences: Issues of Accuracy and Scale. Island Press, Washington, DC.
Soberón, J., Peterson, A.T., 2004. Biodiversity informatics: managing and applying primary biodiversity data. Philosophical Transactions of the Royal Society of London B 359, 689–698.
Soriano, P.J., Díaz de Pascual, A., Ochoa-G., J., Aguilera, M., 1999. Biogeographic analysis of the mammal communities in the Venezuelan Andes. Interciencia 24, 17–25.
Soriano, P.J., Utrera, A., Sosa, M., 1990. Inventario preliminar de los mamíferos del Parque Nacional General Cruz Carrillo (Guaramacal), Estado Trujillo, Venezuela. Biollania 7, 83–99.
Veloz, S.D., 2009. Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. Journal of Biogeography 36, 2290–2299.
Warren, D.L., Glor, R.E., Turelli, M., 2008. Environmental niche equivalency versus conservatism, quantitative approaches to niche evolution. Evolution 62, 2868–2883.
Warren, D.L., Seifert, S.N., 2011. Ecological niche modeling in Maxent: the importance of model complexity and the performance of model selection criteria. Ecological Applications 21, 335–342.
Wiens, J.J., Graham, C.H., 2005. Niche conservatism: integrating evolution, ecology, and conservation biology. Annual Review of Ecology, Evolution, and Systematics 36, 519–539.
Wiley, E.O., McNyset, K.M., Peterson, A.T., Robins, C.R., Stewart, A.M., 2003. Niche modeling and geographic range predictions in the marine environment using a machine-learning algorithm. Oceanography 16, 120–127.
Williams, J.W., Jackson, S.T., 2007. Novel climates, no-analog communities, and ecological surprises. Frontiers in Ecology and the Environment 5, 475–482.
Wisz, M.S., Hijmans, R.J., Li, J., Peterson, A.T., Graham, C.H., Guisan, A., NCEAS Predicting Species Distributions Working Group, 2008. Effects of sample size on the performance of species distribution models. Diversity and Distributions 14, 763–773.
Woodman, N., 2002. A new species of small-eared shrew from Colombia and Venezuela (Mammalia: Soricomorpha: Soricidae: genus *Cryptotis*). Proceedings of the Biological Society of Washington 115, 249–272.
Woodman, N., Díaz de Pascual, A., 2004. *Cryptotis meridensis*. Mammalian Species 761, 1–5.
Zaniewski, A.E., Lehmann, A., Overton, J.M., 2002. Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. Ecological Modelling 157, 261–280.