Journal of Biogeography (J. Biogeogr.) (2014) **41**, 629–643

SPECIAL PAPER

Making better MAXENT models of species distributions: complexity, overfitting and evaluation

Aleksandar Radosavljevic[1]* and Robert P. Anderson[1,2,3]

[1]*Department of Biology, City College of the City University of New York, New York, NY 10031, USA,* [2]*Graduate Center of the City University of New York, New York, NY 10016, USA,* [3]*Division of Vertebrate Zoology (Mammalogy), American Museum of Natural History, New York, NY 10024, USA*

*Correspondence and present address: Aleksandar Radosavljevic, Plant Biology and Conservation, Northwestern University, Evanston, IL 60208 USA.
E-mail: aleks.rado@u.northwestern.edu

ABSTRACT

Aim Models of species niches and distributions have become invaluable to biogeographers over the past decade, yet several outstanding methodological issues remain. Here we address three critical ones: selecting appropriate evaluation data, detecting overfitting, and tuning program settings to approximate optimal model complexity. We integrate solutions to these issues for MAXENT models, using the Caribbean spiny pocket mouse, *Heteromys anomalus*, as an example.

Location North-western South America.

Methods We partitioned data into calibration and evaluation datasets via three variations of *k*-fold cross-validation: randomly partitioned, geographically structured and masked geographically structured (which restricts background data to regions corresponding to calibration localities). Then, we carried out tuning experiments by varying the level of regularization, which controls model complexity. Finally, we gauged performance by quantifying discriminatory ability and overfitting, as well as via visual inspections of maps of the predictions in geography.

Results Performance varied among data-partitioning approaches and among regularization multipliers. The randomly partitioned approach inflated estimates of model performance and the geographically structured approach showed high overfitting. In contrast, the masked geographically structured approach allowed selection of high-performing models based on all criteria. Discriminatory ability showed a slight peak in performance around the default regularization multiplier. However, regularization levels two to four times higher than the default yielded substantially lower overfitting. Visual inspection of maps of model predictions coincided with the quantitative evaluations.

Main conclusions Species-specific tuning of model parameters can improve the performance of MAXENT models. Further, accurate estimates of model performance and overfitting depend on using independent evaluation data. These strategies for model evaluation may be useful for other modelling methods as well.

Keywords
Cross validation, evaluation, *Heteromys*, Maxent, niche, overfitting, rodent, smoothing, South America, tuning.

INTRODUCTION

Three challenges in ecological niche modelling

In recent years, many techniques for modelling species' niches and distributions have been developed and applied extensively throughout biogeography (Guisan & Zimmermann, 2000; Guisan & Thuiller, 2005; Peterson, 2006; Kozak *et al.*, 2008; Peterson *et al.*, 2011; whose terminology we follow). Among the algorithms available, MAXENT has come into particularly common use (Phillips *et al.*, 2006; Elith *et al.*, 2011). MAXENT has performed well (Elith *et al.*, 2006; Hernandez *et al.*, 2006; Wisz *et al.*, 2008), but its output depends critically on model complexity and how closely data

doi:10.1111/jbi.12227

Journal of Biogeography

match assumptions (Phillips & Dudík, 2008; Elith *et al.*, 2010; Anderson & Gonzalez, 2011; Warren & Seifert, 2011). We address three outstanding issues for MAXENT modelling: (1) quantifying overfitting (to detect overly complex models); (2) tuning program settings (to determine those that lead to optimal model complexity); and (3) acquiring independent evaluation data (to quantify overfitting and allow for proper tuning of program settings). Throughout, we emphasize principles relevant for modelling the environmental conditions and areas suitable for a species, which is necessary for transferring a model across space or time (Peterson *et al.*, 2011; Anderson, 2012, 2013). Many of the approaches examined here should also be relevant to other modelling techniques.

First, there has been insufficient treatment of producing models with an appropriate balance between simplicity and complexity, thereby avoiding underfitting or overfitting (Elith *et al.*, 2010; Warren & Seifert, 2011). Typically, studies evaluate model quality via quantitative measures of performance by dividing occurrence data into calibration and evaluation datasets. Overfitting occurs when a model fits the calibration data too closely (in environmental space) and, therefore, fails to predict independent evaluation data accurately. Similarly, underfitted models (those that fail to include sufficient complexity) do not provide adequate discrimination and, hence, predict poorly as well. Research suggests that underfitting is less frequently a problem than overfitting, at least for techniques that can fit complex responses, such as MAXENT (Elith *et al.*, 2006; Anderson & Gonzalez, 2011; Warren & Seifert, 2011). Both overfitted and underfitted models lack generality, which hinders studies that involve model transfer to another region or time period or that aim to compare species niches (Peterson, 2003; Araújo *et al.*, 2005a; Hijmans & Graham, 2006; Randin *et al.*, 2006; Peterson *et al.*, 2007, 2011; Phillips, 2008; Jezkova *et al.*, 2009).

Second, the process of tuning (or smoothing) involves varying model parameters to approximate the optimal level of model complexity: that which best predicts calibration data without overfitting (by 'tuning' the settings of a program, or 'smoothing' a species' response curves to particular predictor variables; Elith *et al.*, 2011). MAXENT software provides default settings, based on the average values determined as optimal in extensive empirical tuning (Phillips & Dudík, 2008). However, optimal settings are likely to vary according to species, occurrence localities, study region and environmental data. Furthermore, due to several methodological issues, we predict that the current default settings lead to overly complex models (see Materials and Methods). Species-specific tuning of program settings shows promise, particularly when general and transferable models are desired (Phillips & Elith, 2010; Anderson & Gonzalez, 2011; Warren & Seifert, 2011).

Third, to approximate optimal model complexity via tuning experiments, it is necessary to use truly independent evaluation data (Peterson *et al.*, 2011). Such data allow reliable estimates of model performance, generality and transferability. Most studies evaluate performance based on random partitioning of occurrence data into calibration and evaluation datasets (split-sample approach of Guisan & Zimmermann, 2000; e.g. Anderson *et al.*, 2002a; Hernandez *et al.*, 2006; Raxworthy *et al.*, 2007; Jezkova *et al.*, 2009). Unfortunately, random partitioning has proven problematic (Araújo *et al.*, 2005b). First, because calibration and evaluation localities often lie close to each other, localities used to evaluate the model are not truly independent of those used to calibrate it. Therefore, as a result of spatial autocorrelation of the environment, they do not provide realistic tests of model quality, typically leading to overestimates of performance (Veloz, 2009; Hijmans, 2012; Bahn & McGill, 2013). Second, geographical biases in the occurrence data associated with frequent sampling near roads, rivers and population centres often lead to environmental biases (Reddy & Dávalos, 2003; Hortal *et al.*, 2008; Loiselle *et al.*, 2008; Boakes *et al.*, 2010). Environmental biases can affect model calibration adversely (Wintle *et al.*, 2005; Araújo & Guisan, 2006; Anderson & Gonzalez, 2011). Under random partitioning, any environmental biases in the original dataset will be preserved in both calibration and evaluation datasets and the latter will not be able to detect any overfitting to the biases (but rather, only to random noise in the calibration dataset; Peterson *et al.*, 2011, pp. 160–161). Therefore, environmental biases also lead to inflated estimates of performance for randomly partitioned occurrence data (Veloz, 2009).

For these reasons, evaluation data should be spatially independent from the calibration data and not contain any environmental bias found in them. Ideally, evaluation data would come from another time period and/or geographical region. Because investigators typically lack occurrence data from other time periods, evaluation across *space* has been proposed as the most reasonable approach for achieving realistic evaluations (Araújo & Rahbek, 2006). Specifically, spatially independent evaluations should be used to identify models that avoid overfitting (Bahn & McGill, 2013). However, in spatially independent evaluations, the comparison dataset (e.g. absence or background) should derive only from regions corresponding to the occurrence localities used in calibration (Phillips, 2008; Anderson & Raza, 2010). Here, we implement a practical way of evaluating across space, in the context of tuning experiments aimed at identifying optimal model complexity.

MATERIALS AND METHODS

Study species and occurrence records

A large high-quality occurrence dataset exists for the Caribbean spiny pocket mouse, *Heteromys anomalus* (Thompson, 1815), along with natural history information that facilitates interpretation of model predictions in geography (Anderson, 2003a; Anderson & Gutiérrez, 2009). Typically, the species ranges from sea level to *c.* 1600 m a.s.l. and inhabits mature or secondary deciduous and evergreen forests, but it has also

been collected in gallery forests in the *Llanos* (savannas) of Venezuela. Most records lie across northern Colombia and Venezuela, as well as on the islands of Trinidad, Tobago and Margarita (Fig. 1). Although a geographically distinct (and possibly disjunct) distributional area occurs in the upper Río Magdalena valley to the south, we exclude that part of the species' distribution because it is poorly characterized and the assumptions of stationarity are less likely to be true.

Georeferenced occurrence localities came from recent taxonomic revisions (Anderson, 2003a; Anderson & Gutiérrez, 2009; 208 unique localities, excluding those in the upper Río Magdalena valley). Because these localities based on museum specimens derive from multiple unplanned surveys typically biased in geography, the resulting localities are likely to exhibit spatial autocorrelation and suffer from environmental biases (Reddy & Dávalos, 2003; Araújo & Guisan, 2006; Hortal *et al.*, 2008; Loiselle *et al.*, 2008; Boakes *et al.*, 2010). To lessen such problems, we filtered data spatially. Such filtering should lead to better locality data – both for model calibration and for model evaluation (Veloz, 2009; Anderson, 2012; Hijmans, 2012). We conducted one test designed to assess the impact of spatial filtering and then used the spatially filtered localities for all other analyses. Specifically, we filtered localities to obtain the maximum number that were at least 10 km apart (Anderson & Raza, 2010). Although the 10-km rule is arbitrary (Hidalgo-Mihart *et al.*, 2004; Iguchi *et al.*, 2004; Pearson *et al.*, 2007), given the topographic and environmental heterogeneity of this system, we chose it with the aim of satisfying the above goals without unduly reducing the number of localities. For each cluster of localities less than 10 km from each other, we determined the maximum number of localities that could be retained. When more than one co-optimal solution existed for a given cluster, we selected one randomly. After filtering, 124 unique localities remained (Fig. 1).

Environmental data

For the environmental data, we used 19 bioclimatic variables from WorldClim 1.4 (http://www.worldclim.org/) at a resolution of 30 arc-seconds. These variables have predicted the abiotically suitable areas of other small non-volant mammals successfully in this region (Anderson & Raza, 2010; Anderson & Gonzalez, 2011). We chose this dataset to determine the behaviour of MAXENT with a set of variables that are likely to predict the abiotically suitable area of this species and that show characteristics typical of those employed by many current modelling studies.

As the study region, we delimited a rectangle that surrounded the full extent of the known occurrences of the northern distribution of the species (i.e. excluding records from the upper valley of the Río Magdalena). The limits were the nearest even half degree that was at least a half degree from the nearest locality after filtering (7–13° N, 60–78° W). This area seems reasonable for approximating the assumptions of background selection by not including large regions that the species does not inhabit because of limitations to dispersal or because of biotic interactions (Anderson & Raza, 2010; but note relatively small areas inhabited by the congeners *H. australis*, *H. catopterius* and *H. oasicus* in this region; Anderson, 2003b; Anderson & Gutiérrez, 2009). Whereas other shapes (such as minimum convex polygons) could have been reasonable as well, we used a rectangular region to simplify creation of data partitions.

Geographically structured evaluations

We implement three variations of *k*-fold cross-validation (Fig. 2; Peterson *et al.*, 2011, pp. 157–159). First, we use standard *k*-fold cross validation in our *randomly partitioned approach*. In *k*-fold cross-validation (= *k*-fold cross

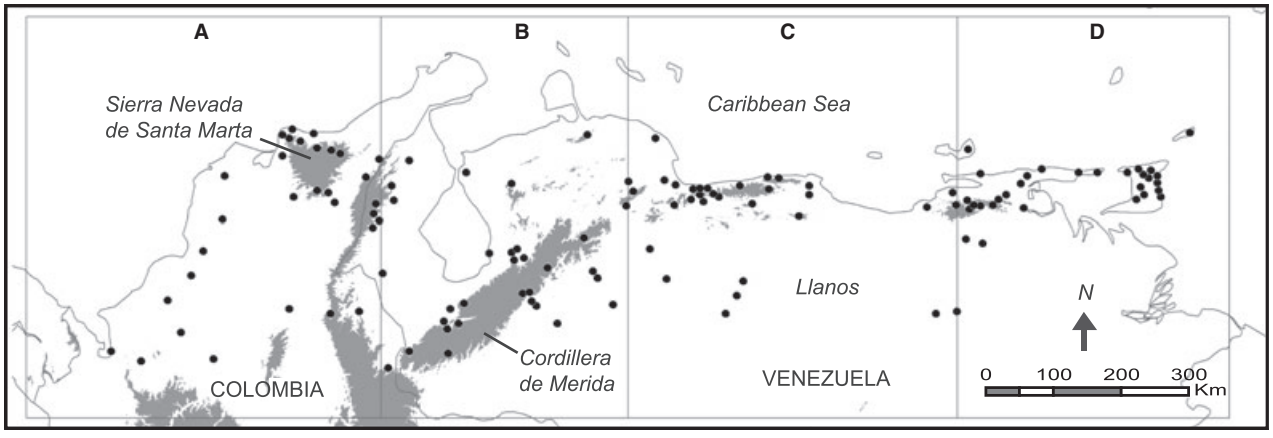

Figure 1 Filtered localities of the Caribbean spiny pocket mouse (*Heteromys anomalus*) in north-western South America. Boxes indicate the four geographical bins used in the present experiments (bins A, B, C and D). This area corresponds to the principal occupied distributional area of the species. The species also occurs in the upper Río Magdalena valley to the south (not shown). Shaded areas correspond to elevations above 1000 m.

partitioning), occurrence localities are divided randomly into k bins (subsets), each of equal sample size (Boyce *et al.*, 2002; Lehmann *et al.*, 2002). Then, models are built in an iterative manner, using $(k - 1)$ bins for calibration in each iteration, with the remaining bin withheld for evaluation. This is repeated until all bins have been used once for evaluation – i.e. until k models are produced. In essence, this procedure constitutes an $(n - 1)$ jackknife of bins, where $n = k$ (Peterson *et al.*, 2011). The evaluation measure/s can be averaged over the iterations. This method holds the drawback that, in each iteration, both calibration and evaluation datasets will hold the same environmental biases. Hence, even when calibration and evaluation localities do not lie close to each other in space, random partitioning can lead to overestimates of performance (Peterson *et al.*, 2011, pp. 160–161). In this approach, Maxent samples background data from the entire study region.

Therefore, following the call for cross-space evaluations, we modify k-fold cross-validation in our *geographically structured approach* by segregating localities into bins spatially (geographically; Fig. 2; Araújo & Rahbek, 2006; Peterson *et al.*, 2011, pp. 161–162; Jiménez-Valverde *et al.*, 2011). Each bin provides spatially independent evaluation data (except for localities very near an adjacent bin). Although any environmental biases present in the overall dataset still exist, this approach segregates such biases geographically, allowing for evaluations capable of detecting overfitting to any corresponding environmental biases (i.e. in addition to overfitting to noise). While using geographically restricted occurrence data can truncate niche estimates, such evaluations assess, in a general sense, the model's transferability across space (Thuiller *et al.*, 2004; Barbet-Massin *et al.*, 2010). However, if any environmental bias in sampling exists uniformly across the geographical bins, this approach (as well as the modification outlined below) will not be able to detect any overfitting to it. As in the former approach, Maxent here samples background data from the entire study region.

Because geographical structuring of calibration and evaluation localities artificially increases geographical biases, it requires another modification. We do so in our *masked geographically structured approach*, by masking out environmental data (see Appendix S1 in Supporting Information) for background sampling from the area corresponding to the localities used for model evaluation (Fig. 2; Bahn & McGill, 2013). When calibration localities are selected from only some portions of the study region, they represent a geographically biased sample (that may be biased in environmental space as well; Peterson *et al.*, 2011, pp. 161–162). Hence, it mimics the natural processes of dispersal limitation and geographical heterogeneity in biotic interactions that can cause a species to inhabit less than its abiotically suitable distribution (Anderson & Raza, 2010, p. 1389). For modelling approaches that use a background, pseudoabsence or absence sample from the study region in model calibration, that region should not include areas where the species is absent

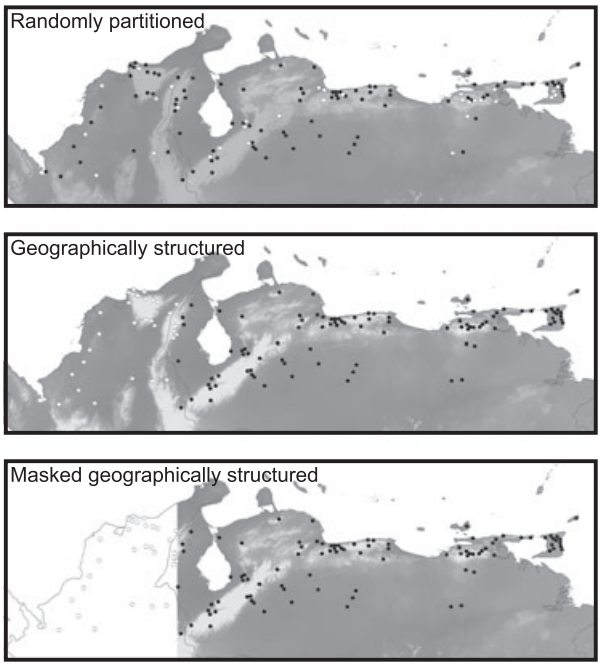

Figure 2 Example of data-partitioning approaches and corresponding regions for background selection used in tuning experiments of Maxent models of the Caribbean spiny pocket mouse, *Heteromys anomalus*. Black circles represent localities used for model calibration and white ones denote localities employed in model evaluation. Shaded areas correspond to one environmental variable (annual mean temperature) for the respective regions used for background sampling in each approach.

because of dispersal limitations or biotic interactions (Anderson & Raza, 2010; Barve *et al.*, 2011). Therefore, theory suggests that when employing geographically structured k-fold cross-validation (or any other geographically structured data-partitioning scheme; e.g. Peterson *et al.*, 2007), background, pseudoabsence or absence data should not be drawn from areas from which known localities were excluded in model calibration (Phillips, 2008; Bahn & McGill, 2013).

Masked geographically structured evaluations constitute a test of transferability in the strict sense, as the models are projected onto an evaluation region that was not included in the calibration process (Randin *et al.*, 2006; Phillips & Dudík, 2008; Peterson *et al.*, 2011). Such transfer across space requires assumptions: essentially the same ones necessary for any spatial or temporal transfer (Anderson, 2013). As an overreaching principle, the species' response should be stationary (Osborne & Suárez-Seoane, 2002). A process is considered stationary if the statistics that define it and that are measured within any subset accurately describe the entire dataset (Osborne *et al.*, 2007). Here, stationarity requires the following assumptions. First, populations across the species' range should not differ in inherited niche characteristics (an assumption of any niche-modelling analysis; Murphy & Lovett-Doust, 2007); similarly, cross-time transfers assume

no niche evolution (Pearson & Dawson, 2003; Nogués-Bravo, 2009). Additionally, relevant biotic interactions should not differ between the two regions or time periods (Anderson *et al.*, 2002b). Furthermore, the second region or time period should not include abiotic environments that are beyond the range of those available in the calibration region (especially in cases of strict transferability). When the latter assumption is violated, additional assumptions are required in order to make a prediction in such cells of the region to which the model is transferred (Phillips *et al.*, 2006; Anderson & Raza, 2010; see below). As is typical for most species, we have no data regarding the possibility of local adaptations across the range of *H. anomalus*. The study region used here only includes relatively small areas inhabited by congeneric species (which may represent competitors; see above). Finally, we use maps of the effect of clamping to assess any environmental heterogeneity across bins (see below). Taking into account these requirements, we proceed with the following experiments.

MAXENT, regularization and model complexity

To make models, we employed MAXENT 3.2.1, with logistic output (see Appendix S2; Phillips *et al.*, 2006; Phillips & Dudík, 2008). MAXENT produces a model based on a series of 'features' (an environmental variable or function thereof). At the sample size of localities used here, MAXENT suggests use of all feature classes (linear, hinge, quadratic, product, threshold and discrete); we used all except for discrete, which is only relevant for categorical variables (Phillips & Dudík, 2008). In cases where a model is transferred to a study region that contains environmental conditions outside the range of those existing in the calibration study region, the species' response curve is said to be truncated (Thuiller *et al.*, 2004; Williams & Jackson, 2007). MAXENT addresses this issue via the assumption of clamping ('clamping' the species' response at that of the most-similar conditions in the calibration region; Phillips *et al.*, 2006; Anderson & Raza, 2010). In cases of spatial or temporal transfer, it is necessary to examine maps that indicate the degree of clamping to determine the effect (if any) that it had on model predictions; we did so by inspection (an alternative route could be via multivariate environmental similarity surfaces; Elith *et al.*, 2010).

MAXENT limits model complexity – and, hence, protects against overfitting – by regularization: a penalty for each term included in the model and for higher weights given to a term (Phillips *et al.*, 2006; Anderson & Gonzalez, 2011). This penalty occurs in the form of a β regularization parameter specific to each feature class (see the 'lasso' for generalized linear and generalized additive models; Phillips *et al.*, 2006). Current releases of MAXENT implement a regularization multiplier, a user-specified coefficient that is applied to the value of the respective β parameter of each feature class, altering the overall level of regularization rather than changing the β parameters individually.

We expect that regularization multipliers higher than default will be necessary to achieve optimal model complexity. First, the default regularization values of MAXENT (determined by Phillips & Dudík, 2008) were based on tuning experiments using *random* partitioning of calibration and evaluation localities, which should lead to overestimates of performance (see above). Furthermore, because overfitted models excel in predicting non-independent evaluation data, random partitioning should tend to select inappropriately low regularization values. Second, other than removing duplicate localities that fell into the same map pixel, no spatial filtering was employed by Phillips & Dudík (2008). Spatial filtering (implemented here) would have reduced the negative effects of spatial autocorrelation, which leads to problems explained above. Third, the measures of performance used to select the default regularization values were AUC (area under the curve of the receiver operating characteristic plot) and log loss (Phillips & Dudík, 2008). While AUC (see below) reflects the discriminatory ability of the model, it does not directly quantify overfitting, which was not considered as an optimality criterion by the authors above. Model selection based solely on discriminatory ability, without consideration of overfitting, tends to result in overly complex models (corresponding to low regularization values).

Effect of spatial filtering

Before conducting the main experiments, we carried out preliminary analyses to test for the expected effects of spatial autocorrelation (and non-independence of calibration and evaluation datasets) on estimates of model performance for randomly partitioned datasets. To do so, we made models using the randomly partitioned approach and the default regularization multiplier. We employed three datasets: (1) all 208 unfiltered localities; (2) the 124 filtered localities; and (3) a rarefied dataset of 124 localities randomly selected from the unfiltered localities. Comparison of the 208 unfiltered localities with the 124 filtered ones assessed the role that this degree of filtering played in reducing the proclivity of the randomly partitioned approach to inflate estimates of model performance. Models calibrated with the third (rarefied) dataset of 124 unfiltered localities served as a control regarding sample size. We predict that the models made with unfiltered localities will lead to higher (inflated) estimates of performance than that calibrated with filtered localities.

As in the main experiments (see below), we made models via k-fold cross validation ($k = 4$). For each dataset (unfiltered, filtered and rarefied), we used the respective evaluation localities to calculate AUC, a measure of the overall discriminatory ability of the model (see below). We averaged those values for each dataset and compared the averages.

Tuning experiments and data partitions

For the main experiments, we used filtered localities and influenced the level of model complexity. We calibrated models with different values for the regularization multiplier

(0.25, 0.50, 1.00, 1.50, 2.00, 4.00, 6.00, 8.00 and 10.00; default setting is 1.00). We again divided the localities into four bins of equal sample size (31 localities in each bin; Figs 1 & 2). For the randomly partitioned approach, we divided localities into bins randomly (note that later versions of Maxent automate implementation of *k*-fold cross validation as in the randomly partitioned approach). In contrast, for the geographically structured and masked geographically structured approaches, we partitioned data *spatially* with four bins (each corresponding to a rectangle) arranged longitudinally from west to east. Each bin had equal sample size, but the corresponding geographical rectangles differed in area, together matching the extent of the full study region (longitudes: Bin A, 72.70–78.00° W; Bin B, 69.00–72.70° W; Bin C, 64.07–69.00° W; Bin D, 60.00–64.07° W).

Quantitative evaluations

We assessed model performance using threshold-independent and threshold-dependent measures (see Appendix S3). As a threshold-independent assessment of overall model performance (discriminatory ability), we used AUC. For presence–background evaluations, AUC quantifies the probability that the model correctly orders (ranks) a random presence locality higher than a random background pixel (Phillips *et al.*, 2006). AUC values calculated with presence–background evaluation data vary according to the proportion of the study region that is suitable for the species and, hence, are not comparable among species or across study regions. Because the present models (produced with different program settings) all correspond to the same species and are evaluated in the same study region, their evaluation AUC values are appropriate for comparison (Lobo *et al.*, 2008; Peterson *et al.*, 2008, 2011). For each data-partitioning approach, we averaged AUC across the four iterations for each regularization multiplier. The average value for the evaluation AUC reflects the relative-ranking ability of the models; however, it does not provide direct information regarding overfitting.

We quantified overfitting directly in several ways. The first, a threshold-independent measure, was by calculating the difference between the calibration and evaluation AUCs (Warren & Seifert, 2011). In our data-partitioning experiments, the magnitude of the difference between calibration and evaluation AUCs quantifies the degree of overfitting to noise in the randomly partitioned approach and overfitting to noise and/or environmental bias in the geographically structured and masked geographically structured approaches (assuming stationarity of the species' response across geography; see above). Valid comparisons with the other approaches (and subtraction of evaluation AUC from calibration AUC) require a non-traditional modification of calibration AUC for the masked geographically structured approach, in which the calibration regions differ among the *k* iterations as well as from those of the other two approaches. Therefore, we calculated both evaluation and calibration AUCs over the full study region and calculated averages as above.

Additionally, we quantified overfitting by comparing threshold-dependent omission rates with theoretically anticipated levels of omission. To do so, we employed thresholding rules with clear expectations: the lowest presence threshold and the 10th percentile presence threshold. Under either thresholding rule, pixels with values equal to or higher than the threshold are considered suitable, yielding a binary prediction. For each, we determined the value of the threshold based on the observed omission of calibration localities – and then employed that threshold to calculate the omission rate for evaluation localities. The lowest presence threshold (= minimum training presence threshold of Maxent software; Pearson *et al.*, 2007) is the lowest value of the prediction for any of the pixels that hold calibration localities; it indicates the least-suitable environmental conditions for which a locality was available in the calibration data set. Similarly, the 10th percentile presence threshold (= 10 percentile training omission threshold of Maxent software) sets as the threshold the value that excludes the 10 percent of the localities having the lowest predicted values. It constitutes a stricter (less permissive) criterion for converting a continuous prediction to a binary one, leading to a smaller geographical prediction. We averaged omission rates as described above for AUC.

To use these omission rates as estimates of overfitting, we compared the observed rates to theoretical expectations. For an ideal model, we expect zero omission of evaluation localities using the lowest presence threshold and approximately 10 percent omission for the 10th percentile presence threshold. Omission rates higher than the theoretical expectation for a given threshold indicate overfitting (assuming stationarity of the species' response across geography; see above). The lowest presence threshold is sensitive to the particular locality that is least suitable (which may often have a substantially lower value than the next-least suitable one). Therefore, it may often lead to an overly extensive prediction when many calibration localities exist. The 10th percentile presence threshold should not be nearly as sensitive to particular extreme localities.

With the goal of identifying the settings that led to optimal model complexity, we used the quantitative measures of performance to select the optimal value(s) of the regularization multiplier. Presence-only datasets provide concrete information regarding the species' presence but no direct data regarding absence, leading to asymmetric errors (Peterson *et al.*, 2011; see also apparent commission error in Anderson, 2003b). Therefore, we considered low overfitting as the primary criterion and secondarily took into account discriminatory ability (Shcheglovitova & Anderson, 2013). Specifically, we interpreted as optimal the regularization multiplier/s that: (1) reduced omission rates to the lowest observed value (or near it) and minimized the difference between calibration and evaluation AUC; and (2) still led to maximal or near maximal observed values for the evaluation

AUC (which assesses discriminatory ability). When more than one regularization multiplier fulfilled these criteria equally well, we chose the lowest one, to promote discriminatory ability (and hence, counter any tendency towards underfitting). Models with an optimal level of complexity can also be estimated through techniques such as the Akaike information criterion (AIC) or nonlinear generalized cross validation (GCV) – which each penalize increasingly complex models – but recently proposed use of such approaches for ecological niche models requires further empirical testing (Warren & Seifert, 2011; Renner & Warton, 2013).

Qualitative evaluations

We also evaluated models by qualitative visual examination of the resulting maps, based on expert knowledge of the distribution of vegetation and habitat types in which the species is known to occur (Huber, 1997). For brevity we present interpretations only for selected comparisons: four regularization multipliers (0.25, 1.00, 2.00 and 6.00) using the randomly partitioned approach; and one bin (C) and regularization multiplier (2.00) for each of the three data-partitioning approaches. We observed: (1) whether the model showed signs of overfitting to the environmental conditions found at calibration localities; (2) the strength of the prediction in the region of the excluded bin (not relevant for the randomly partitioned approach); (3) the overall discriminatory ability of the model; and (4) details of the predictions in particular regions where strong differences were apparent among regularization multipliers and/or data-partitioning approaches. As signs of overfitting, we searched for very small regions of high prediction (lying close to calibration localities) that do not correspond to recognized vegetation types that the species is known to inhabit. In addition, where relevant, we examined maps of clamping, to assess the degree to which it may have affected predictions.

RESULTS

Spatial filtering

In the preliminary experiments, models calibrated using unfiltered localities led to much higher quantitative estimates of performance than those made with filtered localities, but visual interpretations of resulting maps indicated the opposite result. Using the default regularization multiplier and random partitions, the models calibrated using all 208 (unfiltered) localities produced average evaluation AUC scores much higher than those made using the 124 filtered ones (Fig. 2a; unfiltered, 0.81; filtered, 0.73). Similarly, the rarefied dataset of 124 localities randomly chosen from the unfiltered ones yielded models with substantially higher average evaluation AUC values than did those calibrated using the 124 filtered localities (rarefied unfiltered, 0.80). On the contrary, visual inspections of the corresponding maps indicated less realistic models for both of the analyses using unfiltered localities. Specifically, the models calibrated with unfiltered localities showed signs of strong overfitting: areas of highest prediction primarily restricted to regions close to calibration localities (not shown). In contrast, in the maps corresponding to the models calibrated with filtered localities, overfitting was substantially lower (see below).

Tuning experiments: quantitative evaluations

In all three approaches, average evaluation AUC (hereafter, AUC) remained relatively flat across the range of values for the regularization multiplier (Fig. 3a). However, each approach showed the highest AUC value at the default regularization multiplier (1.00) and performance decreased slightly as the regularization multiplier was increased or decreased from the default. Across all values of the regularization multiplier, the geographically structured approach showed substantially lower AUC values than did the randomly partitioned one. However, the masked geographically structured approach yielded values similar to those of the random partitions.

All three approaches displayed similar trends regarding the *difference* between calibration and evaluation AUC values. The difference (which indicates overfitting) was moderately high at low levels of the regularization multiplier but rapidly decreased approaching the default setting (1.00) and levelled off at 4.00 (Fig. 3b). Across all regularization multiplier values, the geographically structured approach displayed a notably higher difference than did the randomly partitioned or masked geographically structured approaches.

Average omission rate for the evaluation localities (hereafter omission rate) using the lowest presence threshold was very high for all three approaches at low regularization values but quickly declined for intermediate and high ones (Fig. 3c). The three curves were virtually flat above a regularization multiplier of 1.50, where rates were only slightly above the zero omission rate expected without overfitting (omission rate at regularization multiplier of 1.50: randomly partitioned, 0.065; geographically structured, 0.073; masked geographically structured, 0.032). The geographically structured approach displayed a higher average omission rate than the random one at regularization multiplier values of 0.25 to 1.00, but the two yielded similar estimates beyond that. The masked geographically structured approach yielded values similar to those of the randomly partitioned one, but at regularization multipliers above 1.00, the omission rate was slightly lower for the former.

Using the 10th percentile presence threshold, all three approaches showed a pattern similar to but more pronounced than that for the lowest presence threshold (Fig. 3d). High omission rates occurred at low regularization multipliers. Omission rates decreased markedly as the regularization multiplier increased; however, here they did not level off until a regularization multiplier of 4.00.

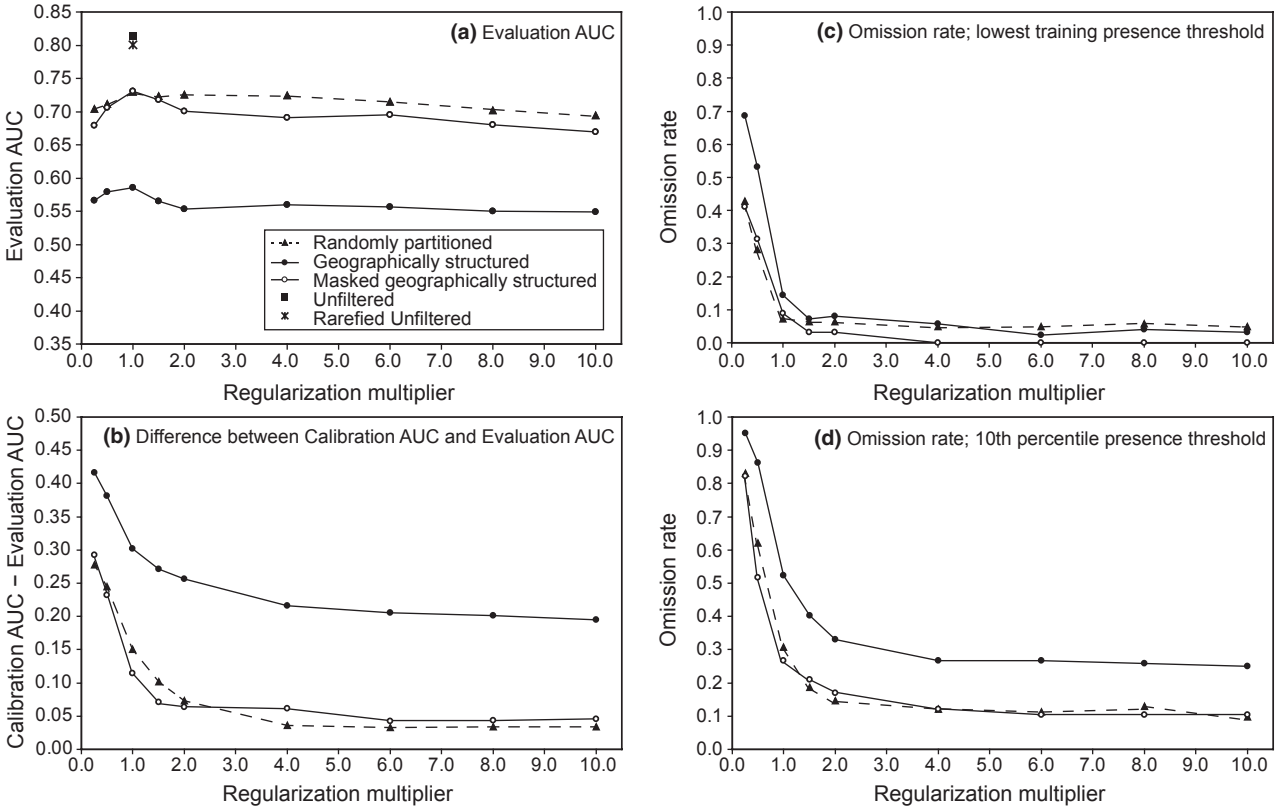

Figure 3 Results of threshold-independent and threshold-dependent evaluations in tuning experiments of Maxent models of the Caribbean spiny pocket mouse, *Heteromys anomalus*: (a) evaluation AUC, (b) calibration AUC minus evaluation AUC, (c) omission rate using the lowest presence threshold, and (d) omission rate using the 10th percentile presence threshold. For each data-partitioning approach, the respective evaluation measure was averaged across the *k* iterations of each value of the regularization multiplier. Whereas evaluation AUC reflects the overall discriminatory ability of the model (with higher values denoting better performance), the other three measures reflect the degree of overfitting (with lower values indicating better performance). Because the randomly partitioned approach provides inflated estimates of performance due to the lack of independence between calibration and evaluation datasets, it appears here as a dashed line. Note the high performance (high AUC and low omission) of the masked geographically structured approach and the decrease in overfitting as the regularization multiplier increases (in b, c and d).

Furthermore, the lowest omission rates achieved were substantially higher than those for the lowest presence threshold for all approaches (Fig. 3c,d; at regularization multiplier 4.00: randomly partitioned, 0.097; geographically structured, 0.250; masked geographically structured, 0.105). Across all regularization multipliers, the geographically structured approach led to a higher omission rate than the randomly partitioned or masked geographically structured ones. At regularization multipliers of 4.00 and above, the omission rates of the randomly partitioned and masked geographically structured approaches were only slightly higher than expected (0.10, or 10%), but that for the geographically structured approach was substantially higher.

Tuning experiments: qualitative assessments

Viewed in geographical space, the maps of model predictions differed substantially among regularization values and, to a lesser degree, among data-partitioning approaches. Although trends were similar for all bins, we present and interpret only those for Bin C and selected regularization multipliers. For all approaches, signs of overfitting decreased markedly with increased regularization, but the very highest regularization values led to models that failed to capture important aspects of the species' abiotically suitable area (based on expert knowledge). We illustrate these patterns for one iteration of the cross-validation experiment for the randomly partitioned approach (Fig. 4). Models made with the lowest regularization multiplier (0.25; Fig. 4a) suffered from extreme overfitting, with the strongest predictions largely restricted to areas near calibration localities. At the default regularization multiplier (1.00; Fig. 4b), overfitting was substantially lower. At regularization multiplier 2.00 (Fig. 4c), the areas strongly predicted for the species generally corresponded to vegetation types where it is known to occur. Overall, good discrimination between suitable and unsuitable environments was found at high elevations (e.g. Cordillera de Mérida, Sierra Nevada de Santa Marta; see also Fig. 1). Although models made using regularization multiplier 6.00 (Fig. 4d) appear broadly similar to those for regularization multiplier

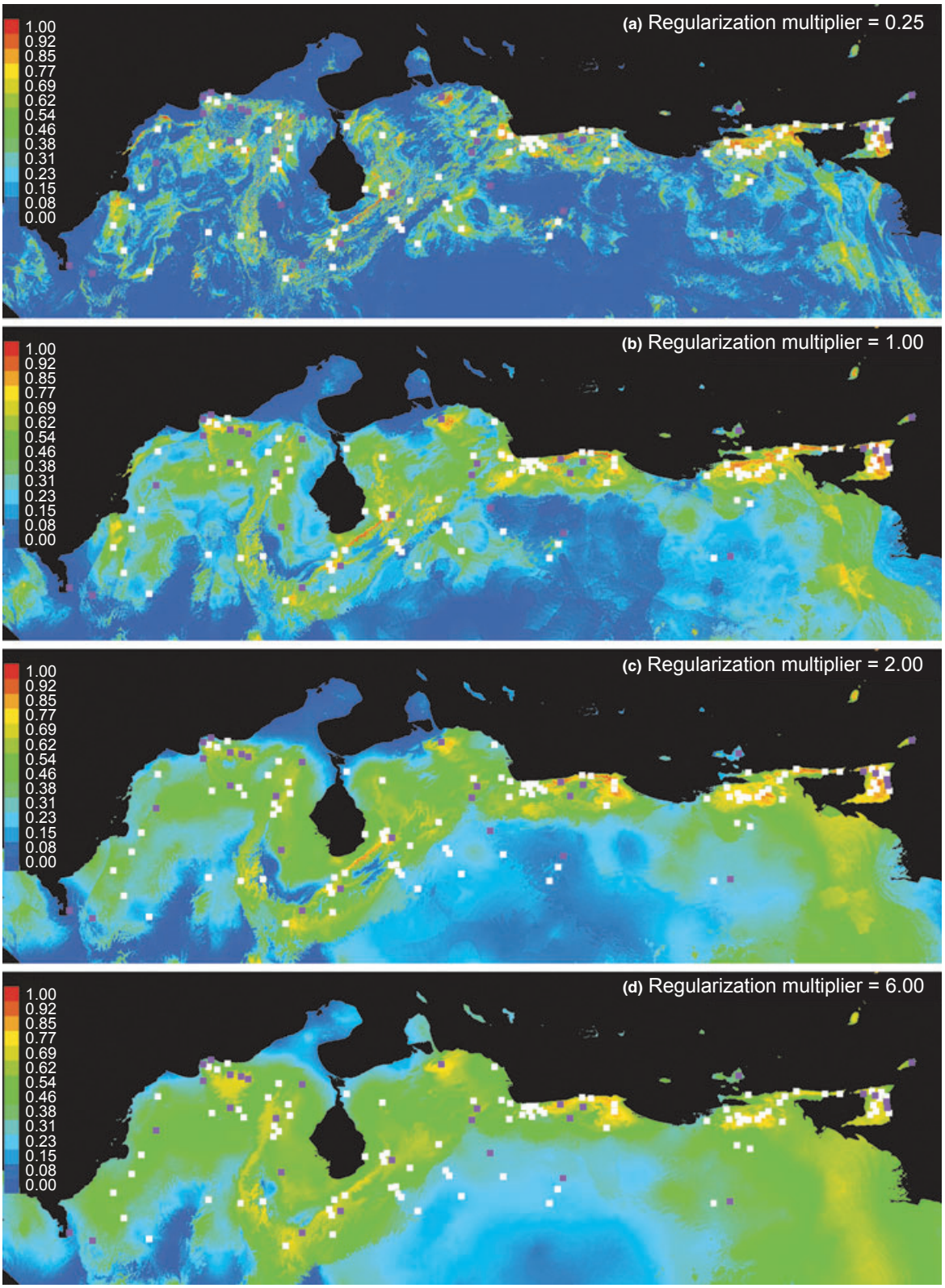

Figure 4 MAXENT models of the abiotically suitable area of the Caribbean spiny pocket mouse (*Heteromys anomalus*), showing the effect of changes in the regularization multiplier. Examples appear here for the randomly partitioned approach, one iteration of the cross-validation experiment and selected regularization multipliers: (a) 0.25, (b) 1.00, (c) 2.00 and (d) 6.00. The predictions show a suitability gradient from low (0, blue) to high (1, red). Squares correspond to calibration (white) and evaluation (purple) localities. In these examples, note signs of overfitting at the two lower values of the regularization multiplier and a loss of discriminatory ability at the highest regularization value (see Fig. 1).

2.00, they lack substantial discrimination in general; furthermore, they do not reflect the species' tolerances accurately, showing unreasonably strong predictions in the high-elevation areas mentioned above.

The masked geographically structured approach led to more realistic predictions than the other data-partitioning approaches. We illustrate these patterns for Bin C and regularization multiplier 2.00 (Fig. 5). At most regularization multipliers, the prediction in the area corresponding to Bin C (the evaluation bin for these predictions; Fig. 1) was weaker for the geographically structured approach (Fig. 5b) than for the randomly partitioned one (Fig. 5a). In comparison with these first two approaches, the masked geographically structured approach (Fig. 5c) showed a notably

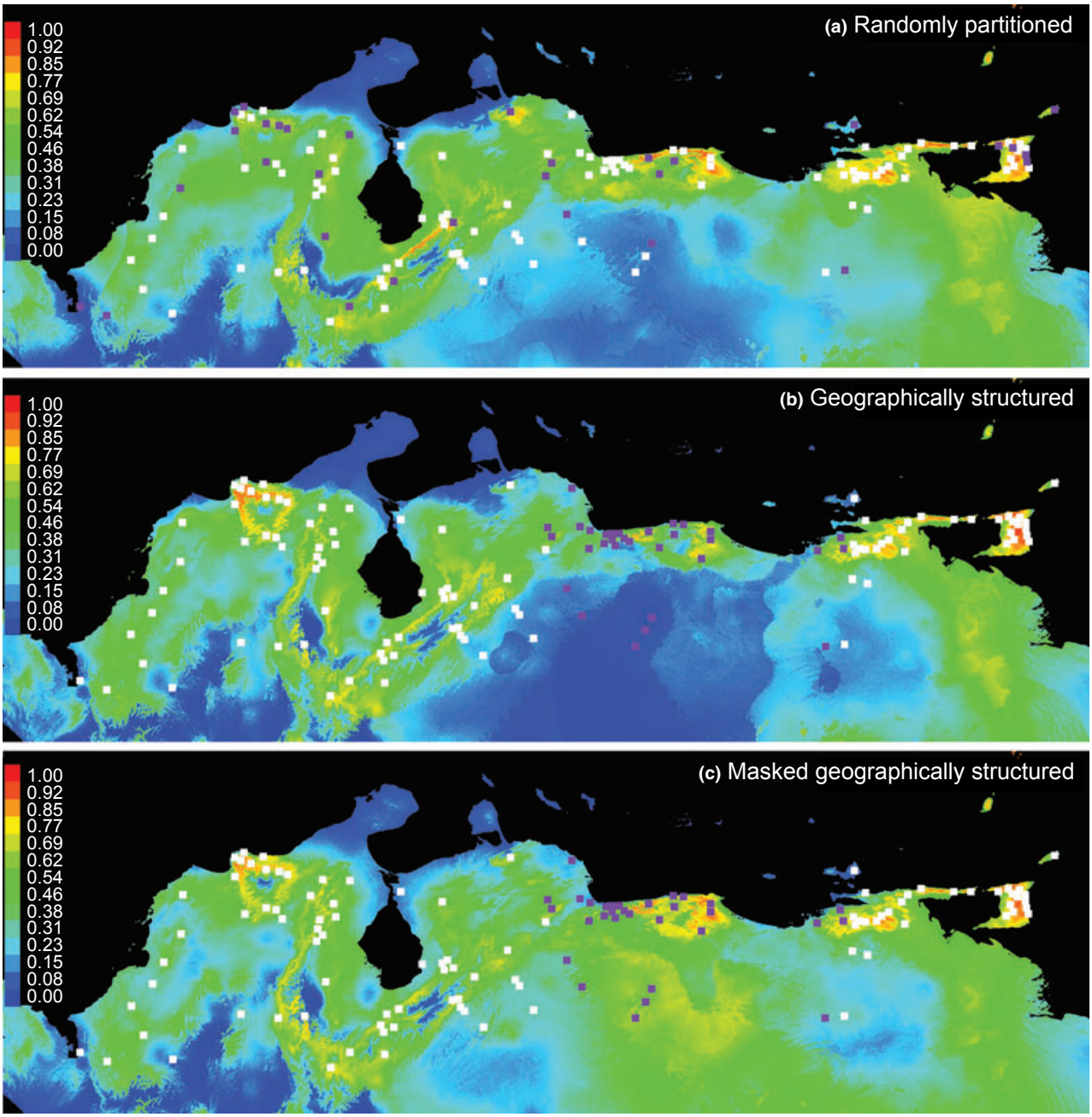

Figure 5 Maxent models of the abiotically suitable area of the Caribbean spiny pocket mouse (*Heteromys anomalus*), showing variation among the three data-partitioning approaches: (a) randomly partitioned, (b) geographically structured and (c) masked geographically structured. Examples appear here for regularization multiplier 2.00 and models for which Bin C constituted the evaluation data. The predictions display a suitability gradient from low (0, blue) to high (1, red). Squares correspond to calibration (white) and evaluation (purple) localities. Overall, the masked geographically structured approach led to the most realistic predictions (see Fig. 1).

stronger prediction in the area of Bin C. The broader prediction in that region matches more closely to the species' known occurrence records in the mixed savanna areas of Bin C. In regions corresponding to the bins used for calibrating the model, the masked geographically structured approach (Fig. 5c) was generally similar to the other two approaches, except for in high-montane areas. In such regions, the masked geographically structured approach (and to a lesser degree the geographically structured approach) showed good discrimination and low predictions for the species in the highest areas (see also Fig. 1). In contrast, the randomly partitioned approach (Fig. 5a) overestimated suitability, especially in the Sierra Nevada de Santa Marta. For all regularization multipliers, clamping was minimal in the masked geographically structured approach – appreciable only in small areas along the Caribbean coast and at the southern end of the Lago de Maracaibo (not shown).

DISCUSSION

Interpretation of experiments

The preliminary analyses using unfiltered versus filtered localities for the randomly partitioned approach yielded substantially higher AUC scores for models calibrated and evaluated with unfiltered localities. This discrepancy existed even when controlling for sample size. Visual inspections of the predictions in geographical space indicated that models made with unfiltered localities exhibited strong signs of overfitting. These results confirm that, as expected, the non-independence between the calibration and evaluation localities in the randomly partitioned approach led to inflated estimates of performance. Although the spatial filtering implemented here lessened these problems, the results of this analysis support interpretation of measures of performance as inflated in the subsequent experiments.

In the main experiments, the threshold-independent evaluations using AUC indicate strong differences among data-partitioning approaches (Fig. 3a). First, the randomly partitioned models showed higher AUC values than the geographically structured ones and this difference probably derives from both artefactual and real causes. Some of the observed difference in estimates of performance between the two approaches derives from an inflated AUC for the randomly partitioned approach (see filtering experiment). The rest of the difference probably reflects the effects of overfitting to any environmental biases in the geographically structured approach (including those resulting from the artificial spatial bias inserted in that approach; Phillips, 2008; Anderson & Raza, 2010).

In contrast, the masked geographically structured models enjoy high and realistic estimates of performance. The difference in performance between the geographically structured and masked geographically structured approaches emphasizes the importance of selecting calibration regions that match modelling assumptions (Anderson & Raza, 2010; Barve *et al.*, 2011; Peterson *et al.*, 2011). As expected, including background data from a region that corresponds to evaluation but not calibration localities (in the geographically structured approach; Barbet-Massin *et al.*, 2010) provides a false negative signal that interferes with successful modelling of the species' existing fundamental niche, here decreasing model performance dramatically.

The difference between calibration and evaluation AUC detects strong overfitting at low regularization values (Fig. 3b). Once again, we interpret that the estimates of overfitting for the randomly partitioned approach are overly optimistic. The higher overfitting indicated here for the geographically structured models probably corresponds to its ability to detect overfitting to any environmental biases (including those resulting from the artificial spatial bias). On the contrary, the masked geographically structured models show performance nearly identical to that of the randomly partitioned ones here and this low estimate of overfitting for the former approach should be realistic.

The threshold-dependent evaluations of omission rate indicate similar differences in performance among regularization multipliers. All three data-partitioning approaches show a striking decline from the lowest regularization multipliers to a value of 2.00, which we attribute to a reduction in overfitting to noise. Both thresholding rules illustrate this marked decrease in omission rate. However, the respective curves level off at different regularization multipliers and the best (lowest) omission rate achieved in a given analysis varies. Whereas the three data-partitioning approaches show virtually identical performance using the lowest presence threshold, the 10th percentile presence threshold indicates notable differences (Fig. 3d). The lack of discrimination among approaches for the lowest presence threshold may be related to its sensitivity to the particular locality that is least suitable. Although the lowest presence threshold may actually lead to an underestimate of the suitable areas for species with very few records, more restrictive thresholding rules are likely to be more appropriate for with species with many occurrence records, as here.

The lowest omission rates achieved are reasonable for both thresholding rules (at least for two of the approaches). Using the lowest presence threshold, all three approaches yield rates at or only slightly higher than the expected zero omission rate (Fig. 3c). Using the 10th percentile presence threshold, both the random and masked geographically structured approaches yield omission rates near the expected 10%, but the higher rate mentioned above for the geographically structured approach is quite high (almost 30%; Fig. 3d). Recall, however, that the randomly partitioned approach produces artefactually low omission rates regardless of the thresholding rule adopted.

The visual interpretations of maps of the predictions in geographical space match patterns observed in the quantitative measures of performance. As judged by visual interpretations for all approaches, low regularization multipliers produce problematic levels of overfitting, intermediate ones

yield satisfactory predictions, and the highest multipliers lead to underfitted models that provide unrealistic predictions in some regions (Fig. 4). As expected, the three data-partitioning approaches differ strongly in their predictions with regard to the area corresponding to the bin used for evaluation, with the masked geographically structured approach leading to the most realistic predictions overall (Fig. 5).

Conclusion and recommendations

We interpret that optimal performance for the present models corresponds to regularization multipliers higher than the default, echoing the findings of studies for other species (Elith *et al.*, 2010; Anderson & Gonzalez, 2011). Although a slight peak occurs in AUC at the default regularization value, all other measures indicate much better performance at slightly to substantially higher regularization multipliers. Specifically, regularization multipliers as high as 2.00 to 4.00 are necessary to reduce overfitting to low levels. Qualitative assessments of the geographical predictions reiterate this conclusion. Although AUC values and omission rates do not worsen with regularization multipliers above 4.00, qualitative visual assessments of models in geographical space show a decline in model quality and overall discriminatory ability. Had these experiments been conducted using unfiltered localities for model calibration, we predict that even higher regularization multipliers would have been necessary to achieve optimal performance on spatially independent evaluation data.

The masked geographically structured approach showed clear advantages over the other two data-partitioning strategies. As predicted, the randomly partitioned approach produced inflated estimates of performance and led to overfitted models. In the geographically structured approach, increasing the regularization multiplier was insufficient to counteract the effects of the strong spatial bias in the localities used for model calibration (artificially inserted in that approach). In contrast, the masked geographically structured approach sidestepped the problem of the artificial spatial bias that we inserted (and any corresponding environmental biases) and allowed for detection of overfitting to environmental biases that differed among the spatial partitions (Phillips, 2008). In conjunction with tuning experiments, this approach can allow selection of model settings likely to avoid overfitting to noise as well as to the latter class of biases.

The current results lead to recommendations regarding the use of tuning to identify optimal model complexity in Maxent for a given species and dataset. Ideally, both the regularization multiplier and the feature classes considered should be subjected to tuning experiments (Shcheglovitova & Anderson, 2013). Future research should also determine if varying the regularization multiplier is sufficient to achieve optimal regularization (i.e. rather than tuning β individually for each feature class; Anderson & Gonzalez, 2011). To reach general conclusions and guidelines regarding model complexity in Maxent, comprehensive experiments are necessary with multiple species. Such research should examine the effects of sample size and spatial autocorrelation in the localities (e.g. with different levels of filtering of calibration localities) and of the level of correlation among environmental variables (Elith *et al.*, 2010; Hijmans, 2012). While we used a simple west-to-east partitioning tactic with only four bins, both the geographical arrangement and the number of bins should be tailored to the project at hand (Peterson *et al.*, 2011; see also a checkerboard approach, Pearson *et al.*, 2013). The approach suggested here should also be compared with that of correcting for the effects of sampling bias when it can be quantified directly or estimated using a suitable target group (which has the potential to avoid overfitting to environmental biases that are uniform across the study region Anderson, 2003b; Phillips *et al.*, 2009). More generally, results based on the tuning approach should be compared with model selection based on information criteria (e.g. the AIC corrected for small sample size, AIC_c; Warren & Seifert, 2011) and generalized cross validation (GCV), which is similar in intent (Renner & Warton, 2013). This overall research agenda may allow for a complex set of rules for estimating new optimal settings for Maxent.

ACKNOWLEDGEMENTS

This research was supported by the US National Science Foundation (NSF DEB-0717357 and DEB-1119915) and International Biogeography Society (Student Travel Award, to A.R.). Darla M. Thomas assisted with data preparation and preliminary modelling. Amy C. Berkov, Robert A. Boria, Ana C. Carnaval, Eliécer E. Gutiérrez, David J. Lohman, Ali Raza, Jhanine L. Rivera, Mariya Shcheglovitova, Mariano Soley-G., Sara Varela and anonymous referees offered insightful comments on various drafts of the manuscript. Steven J. Phillips provided instruction in running Maxent from the command line and answered several queries about other Maxent functionalities.

REFERENCES

Anderson, R.P. (2003a) Taxonomy, distribution, and natural history of the genus *Heteromys* (Rodentia: Heteromyidae) in western Venezuela, with the description of a dwarf species from the Península de Paraguaná. *American Museum Novitates*, **3396**, 1–43.

Anderson, R.P. (2003b) Real vs. artefactual absences in species distributions: tests for *Oryzomys albigularis* (Rodentia: Muridae) in Venezuela. *Journal of Biogeography*, **30**, 591–605.

Anderson, R.P. (2012) Harnessing the world's biodiversity data: promise and peril in ecological niche modeling of species distributions. *Annals of the New York Academy of Sciences*, **1260**, 66–80.

Anderson, R.P. (2013) A framework for using niche models to estimate impacts of climate change on species distribu-

tions. *Annals of the New York Academy of Sciences*, **1297**, 8–28.

Anderson, R.P. & Gonzalez, I., Jr (2011) Species-specific tuning increases robustness to sampling bias in models of species distributions: an implementation with Maxent. *Ecological Modelling*, **222**, 2796–2811.

Anderson, R.P. & Gutiérrez, E.E. (2009) Taxonomy, distribution, and natural history of the genus *Heteromys* (Rodentia: Heteromyidae) in central and eastern Venezuela, with the description of a new species from the Cordillera de la Costa. *Bulletin of the American Museum of Natural History*, **331**, 33–93.

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., Gómez-Laverde, M. & Peterson, A.T. (2002b) Using niche-based GIS modeling to test geographic predictions of competitive exclusion and competitive release in South American pocket mice. *Oikos*, **11**, 131–141.

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

Araújo, M.B. & Rahbek, C. (2006) How does climate change affect biodiversity? *Science*, **313**, 1396–1397.

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

Araújo, M.B., Whittaker, R.J., Ladle, R.J. & Erhard, M. (2005b) Reducing uncertainty in projections of extinction risk from climate change. *Global Ecology and Biogeography*, **14**, 529–538.

Bahn, V. & McGill, B.J. (2013) Testing the predictive performance of distribution models. *Oikos*, **122**, 321–331.

Barbet-Massin, M., Thuiller, W. & Jiguet, F. (2010) How much do we overestimate future local extinction rates when restricting the range of occurrence data in climate suitability models? *Ecography*, **33**, 878–886.

Barve, N., Barve, V., Jiménez-Valverde, A., Lira-Noriega, A., Maher, S.P., Peterson, A.T., Soberón, J. & Villalobos, F. (2011) The crucial role of the accessible area in ecological niche modeling and species distribution modeling. *Ecological Modelling*, **222**, 1810–1819.

Boakes, E.H., McGowan, P.J.K., Fuller, R.A., Ding, C.-Q., Clark, N.E., O'Connor, K. & Mace, G.M. (2010) Distorted views of biodiversity: spatial and temporal bias in species occurrence data. *PLoS Biology*, **8**, e1000385.

Boyce, M.S., Vernier, P.R., Nielsen, S.E. & Schmiegelow, F.K. (2002) Evaluating resource selection functions. *Ecological Modelling*, **157**, 281–300.

Elith, J., Graham, C.H., Anderson, R.P. *et al.* (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **2**, 129–151.

Elith, J., Kearney, M. & Phillips, S. (2010) The art of modelling range-shifting species. *Methods in Ecology and Evolution*, **1**, 330–342.

Elith, J., Phillips, S.J., Hastie, T., Dudík, M., Chee, Y.E. & Yates, C.J. (2011) A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, **17**, 43–57.

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

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

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*, **5**, 773–785.

Hidalgo-Mihart, M.G., Cantú-Salazar, L., González-Romero, A. & López-González, C.A. (2004) Historical and present distribution of coyote (*Canis latrans*) in Mexico and Central America. *Journal of Biogeography*, **31**, 2025–2038.

Hijmans, R.J. (2012) Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model. *Ecology*, **93**, 679–88.

Hijmans, R.J. & Graham, C.H. (2006) The ability of climate envelope models to predict the effect of climate change on species distributions. *Global Change Biology*, **12**, 2272–2281.

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.

Huber, O. (1997) Ambientes fisiograficos y vegetales de Venezuela. *Vertebrados acutales y fosiles de Venezeula* (ed. by E. Lar Marca), pp. 279–298. Museuo de Ciencia y Tecnologia de Merida, Merida, Venezuela.

Iguchi, K., Matsuura, K., McNyset, K.M., Peterson, A.T., Scachetti-Pereira, R., Powers, K.A., Vieglais, D.A., Wiley, E.O. & Yodo, T. (2004) Predicting invasions of North American basses in Japan using native range data and a genetic algorithm. *Transactions of the American Fisheries Society*, **133**, 845–854.

Jezkova, T., Jaeger, J.R., Marshall, Z.L. & Riddle, B.R. (2009) Pleistocene impacts on the phylogeography of the desert pocket mouse (*Chaetodipus penicillatus*). *Journal of Mammology*, **90**, 306–320.

Jiménez-Valverde, A., Peterson, A.T., Soberón, J., Overton, J.M., Aragón, P. & Lobo, J.M. (2011) Use of niche models in invasive species risk assessments. *Biological Invasions*, **13**, 2785–2797.

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.

Lehmann, A., Overton, J.M. & Leathwick, J.R. (2002) GRASP: generalized regression analysis and spatial prediction. *Ecological Modelling*, **157**, 189–207.

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.

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

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

Nogués-Bravo, D. (2009) Predicting the past distribution of species climatic niches. *Global Ecology and Biogeography*, **18**, 521–531.

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

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.

Pearson, R.G. & Dawson, T.P. (2003) Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? *Global Ecology and Biogeography*, **12**, 361–371.

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.

Pearson, R.G., Phillips, S.J., Loranty, M.M., Beck, P.S.A., Damoulas, T., Knight, S.J. & Goetz, S.J. (2013) Shifts in Arctic vegetation and associated feedbacks under climate change. *Nature Climate Change*, **3**, 673–677.

Peterson, A.T. (2003) Predicting the geography of species' invasions via ecological niche modeling. *The Quarterly Review of Biology*, **78**, 419–33.

Peterson, A.T. (2006) Uses and requirements of ecological niche models and related distributional models. *Biodiversity Informatics*, **3**, 59–72.

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. (2011) *Ecological niches and geographic distributions*. Princeton University Press, Princeton, NJ.

Phillips, S.J. (2008) Transferability, sample selection bias and background data in presence-only modelling: a response to Peterson *et al.* (2007). *Ecography*, **31**, 272–278.

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. & Elith, J. (2010) POC plots: calibrating species distribution models with presence-only data. *Ecology*, **91**, 2476–84.

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**, 181–97.

Randin, C.F., Dirnböck, T., Dullinger, S., Zimmermann, N.E., Zappa, M. & Guisan, A. (2006) Are niche-based species distribution models transferable in space? *Journal of Biogeography*, **33**, 1689–1703.

Raxworthy, C.J., Ingram, C.M., Rabibisoa, N. & Pearson, R.G. (2007) Applications of ecological niche modeling for species delimitation: a review and empirical evaluation using day geckos (*Phelsuma*) from Madagascar. *Systematic Biology*, **56**, 907–23.

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

Renner, I.W. & Warton, D.I. (2013) Equivalence of MAXENT and Poisson point process models for species distribution modeling in ecology. *Biometrics*, **69**, 274–81.

Shcheglovitova, M. & Anderson, R.P. (2013) Estimating optimal complexity for ecological niche models: a jackknife approach for species with small sample sizes. *Ecological Modelling*, **269**, 9–17.

Thuiller, W., Brotons, L., Araújo, M.B. & Lavorel, S. (2004) Effects of restricting environmental range of data to project current and future species distributions. *Ecography*, **2**, 165–172.

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. & 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.

Williams, J.W. & Jackson, S.T. (2007) Novel climates, no-analog communities, and ecological surprises. *Frontiers in Ecology and the Environment*, **5**, 475–482.

Wintle, B., Elith, J. & Potts, J. (2005) Fauna habitat modelling and mapping: a review and case study in the Lower Hunter Central Coast region of NSW. *Austral Ecology*, **30**, 719–738.

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.

SUPPORTING INFORMATION

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

Appendix S1 Procedure for masking environmental variables.
Appendix S2 Instructions for running MAXENT using .bat files.
Appendix S3 AUC and omission rate calculations.

BIOSKETCHES

Aleksandar Radosavljevic is a biogeographer and systematist particularly interested in the evolution and biogeography of tropical legumes. His research aims to integrate phylogenetics and niche modelling to better understand patterns of legume evolution and diversification in the Neotropics.

Robert P. Anderson conducts biogeographical studies at the interface between ecology and evolution. His research focuses on the development of methods for modelling species niches and distributions. In addition to these techniques of general application, his taxonomic and geographical specialty is Neotropical mammals.

Editor: Miguel Araújo