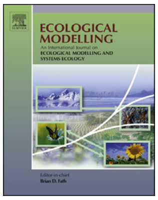
Ecological Modelling 269 (2013) 9–17

Contents lists available at ScienceDirect

Ecological Modelling

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



Estimating optimal complexity for ecological niche models: A jackknife approach for species with small sample sizes

Mariya Shcheglovitova[a,*,1], Robert P. Anderson[a,b,c]

[a] Department of Biology, City College of the City University of New York, 160 Convent Avenue, New York, NY 10031, USA
[b] Graduate Center, City University of New York, 365 5th Avenue, New York, NY 10016, USA
[c] 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 16 May 2013
Received in revised form 19 August 2013
Accepted 20 August 2013

Keywords:
Ecological niche model
Feature class
Jackknife
Maxent
Species distribution model
Tuning

ABSTRACT

Algorithms for producing ecological niche models and species distribution models are widely applied in biogeography and conservation biology. However, in some cases models produced by these algorithms may not represent optimal levels of complexity and, hence, likely either overestimate or underestimate the species' ecological tolerances. Here, we evaluate a delete-one jackknife approach for tuning model settings to approximate optimal model complexity and enhance predictions for datasets with few (here, <10) occurrence records. We apply this approach to tune two settings that regulate model complexity (feature class and regularization multiplier) in the presence-background modeling program Maxent for two species of spiny pocket mice in Ecuador and southwestern Colombia. For these datasets, we identified an optimal feature class parameter that is more complex than the default. Highly complex features are not typically recommended for use with small sample sizes in Maxent. However, when coupled with higher regularization, complex features (that allow more flexible responses to environmental variables) can obtain models that out-perform those built using default settings (employing less complex feature classes). Although small sample sizes remain a serious limitation to model building, this jackknife optimization approach can be used for species with few localities (<approximately 20–25) to produce models that maximize the utility of the little information available.

1. Introduction

Ecological niche models (ENMs) and species distribution models (SDMs) based on presence-only occurrence data constitute widely used tools for many areas of biogeographic research, as well as for conservation planning (Papeş and Gaubert, 2007; Wilting et al., 2010; Lawler et al., 2011; Anderson, 2013). Here, we follow the paradigm of ecological niche modeling of the conditions suitable for the species in model calibration, evaluation, and interpretation (Peterson et al., 2011; Anderson, 2012). However, the methodological advances we apply are equally applicable to models aimed at characterizing the species' occupied distribution (SDMs, *sensu stricto*). ENMs examine associations between known occurrences of a species and abiotic environmental (often climatic) data in the geographic region of interest. The resulting model approximates the environmental conditions that the species can inhabit (the species' existing fundamental niche, subject to clear assumptions); that model then can be applied to geography, yielding estimates of the corresponding areas with suitable environmental conditions (its abiotically suitable distribution; see Peterson et al. (2011) for terminology and assumptions regarding the characteristics of occurrence and environmental data).

Despite their broad appeal, ENMs may be especially problematic when implemented with species for which few occurrence records exist; nevertheless, such situations often correspond to precisely the species most in need of predictive models for conservation-based initiatives (Gaubert et al., 2006). Specifically, model accuracy decreases and model variability increases with decreasing sample size (Wisz et al., 2008). If possible, the paucity of occurrence data should be rectified by increasing efforts put into field surveys and data sharing (Cayuela et al., 2009). However, this seldom is feasible in the time frame within which conservation decisions need to be made. As an alternative, optimizing or tuning model settings (sometimes called "smoothing") to estimate optimal model complexity can result in higher-quality output than employing default settings (Elith et al., 2010; Anderson and Gonzalez, 2011; Warren and Seifert, 2011; Radosavljevic and Anderson, in press). Furthermore, optimal settings likely vary among species as well as for different combinations of the occurrence localities, study region, and environmental data at hand. Therefore, we explored model tuning as a

[*] Corresponding author Tel.: +1 347 410 3228; fax: +1 202 994 6100.
E-mail address: mshcheg@gwu.edu (M. Shcheglovitova).
[1] Present address: Department of Biological Sciences, George Washington University, 2023 G St. NW, Washington, DC 20052, USA.

0304-3800/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.ecolmodel.2013.08.011

way of improving ENMs for datasets with few occurrence records. In particular, we used a delete-one jackknife approach suggested for model evaluation recently (a form of *k*-fold cross validation where *k* is equal to the number of occurrence localities in the original dataset; Peterson et al., 2011; see also Pearson et al., 2007). Although this approach may also be useful for higher sample sizes (e.g., up to ca. 25 records), we here employ it for species with very few records (<10).

As an assessment of this approach, we used the presence-background modeling software Maxent (Phillips et al., 2006) to generate ENMs for two species of spiny pocket mice across a range of program settings (Supplementary Fig. 3). We compared the performance of default settings to a variety of user-specified settings. Maxent identifies geographic areas of suitable conditions for a species, based on known occurrence records, by applying a maximum entropy model to estimate the species' response given a set of constraints (environmental variables). We chose Maxent because it: (1) is in common use; and (2) has been found to perform well for small sample sizes in previous studies (Wisz et al., 2008); yet, (3) is sensitive to model settings that affect model complexity (Elith et al., 2010; Anderson and Gonzalez, 2011; Warren and Seifert, 2011; Syfert et al., 2013). In the tuning experiments that led to the current default settings, Phillips and Dudík (2008) stated that for datasets unlike those used in that study, it may be necessary to use further tuning to optimize the program's performance. Even though we tested our approach using Maxent, this jackknife approach for model tuning with small sample sizes is general and can be extended to other modeling methods. We assessed models based on quantitative evaluations of performance, and compared optimal to default models using measures of similarity. Independently, we evaluated model output qualitatively.

2. Materials and methods

2.1. Study species and region

We used two species of spiny pocket mice, *Heteromys australis* and *Heteromys teleus* (Rodentia: *Heteromyidae*), to conduct our tuning experiments. These species represent suitable entities for the current study for several reasons. Recent taxonomic research provides high-quality (although limited) occurrence data, as well as general natural-history information regarding the habitats occupied by the species. Furthermore, strong climatic gradients exist in the regions occupied by these species, facilitating both model calibration and interpretation.

Both species inhabit western Ecuador. In addition, the range of *H. australis* extends into Colombia, eastern Panama, and western Venezuela (Anderson and Jarrín-V, 2002 Fig. 1). In northwestern Ecuador and southwestern Colombia, *H. australis* can be found in very wet and unseasonal evergreen forests, while *H. teleus* inhabits slightly drier and markedly seasonal, but still evergreen forests in central–western Ecuador (Anderson and Jarrín-V, 2002). Both species occur in a wide range of altitudes on the Pacific coastal lowlands and western slopes of the Andes (from up to ca. 2000 m; Anderson and Jarrín-V, 2002). Preliminary conservation assessments were undertaken several years ago for these species in Ecuador using cruder climatic data and a different modeling method (Anderson and Martínez-Meyer, 2004). Our aim here is to explore model complexity with Maxent, leaving conservation-related questions for these species to other ongoing studies (Burneo, pers. comm.).

We modeled the environmental requirements of *H. teleus* in its full known distribution and those for *H. australis* in part of its range (Ecuador and southwestern Colombia). We did so for *H. australis* because high-quality occurrence data exist for it in this region. We acknowledge that the present dataset may underestimate the species' full environmental tolerances. For each species, we delimited a rectangle that surrounds the occurrence records, specifically one whose borders were the nearest half degree from the most-peripheral occurrence record after filtering (0.5°S to 2°N, 77.5–80°W for *H. australis*; 2.5°S to 1°N, 78.5–81.5° W for *H. teleus*; see Section 2.2 for a description of filtering).

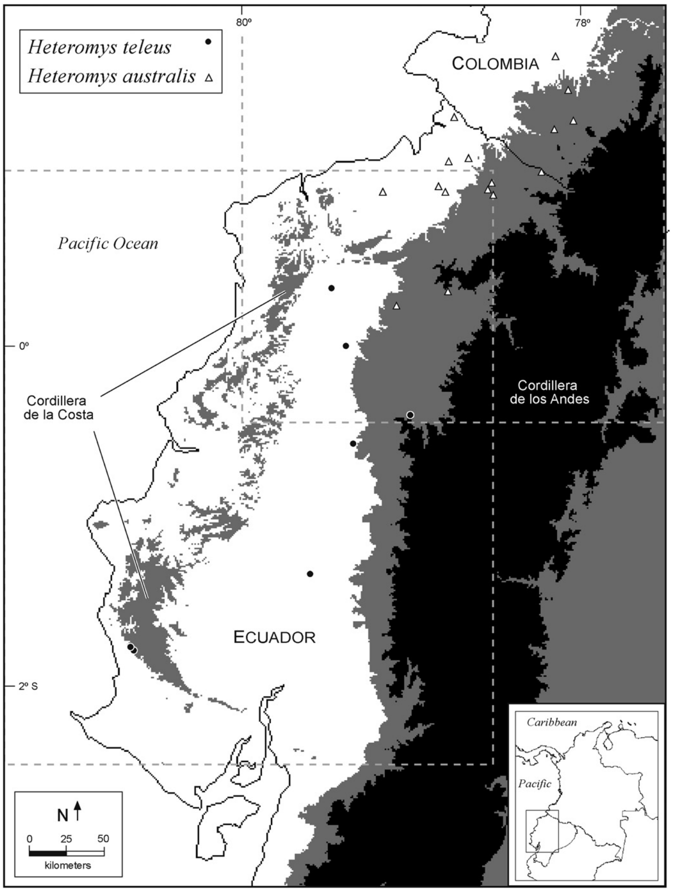


Fig. 1. Map showing all (unfiltered) occurrence records for *Heteromys australis* (triangles) and *H. teleus* (circles) in Ecuador and southwestern Colombia (data from Anderson and Jarrín-V, 2002). Regions above 300 m are shown in gray, and areas above 2000 m appear in black. The dashed boxes indicate the study regions used here in modeling the abiotically suitable areas for each species using spatially filtered localities.

2.2. Occurrence and environmental data

We created jackknife sets for each species after filtering occurrence records to reduce the likely effects of spatial autocorrelation due to biased sampling typical of museum and herbarium data (Peterson et al., 2011; Anderson, 2012). The probable environmental bias introduced by spatially autocorrelated occurrence records has been observed to affect model complexity (Anderson and Gonzalez, 2011). Additionally, Maxent tends to produce overfit predictions when used with biased occurrence records (Peterson et al., 2007). An overfit model is more complex than the true relationships between the species' niche and the examined environmental variables (Peterson et al., 2011). Had we not filtered occurrence records, geographically proximate records with similar environmental characteristics may have led to inflated estimates of performance (Veloz, 2009) and, therefore, to selection of overly complex models as optimal.

To filter occurrence records, we only retained those with a linear distance more than 30 km to neighboring records, such that the final dataset consisted of the maximum number of occurrence records (see Anderson and Raza, 2010). To do so, we calculated the distance between each pair of localities and identified all clusters of localities containing pairwise distances less than or equal to 30 km. For each such cluster, we determined (by inspection) all possible deletions that would yield a smaller cluster with pairwise distances of at least 30 km. Of those, we selected the one that maintained the largest number of records; if multiple co-optimal solutions existed, we chose one randomly. Sixteen occurrence records existed for *H. australis*, and seven for *H. teleus* (Anderson and Jarrín-V, 2002). This led to the removal of one occurrence record for *H. teleus* and seven for *H. australis*, producing respective totals of six and nine after filtering. We created delete-one jackknife sets for each of the filtered datasets. Each jackknife dataset contained one evaluation record and $n-1$ calibration records (where n equals the total number of filtered occurrence records). In this way, the spatial filtering ensured that evaluation records were at least 30 km away from the records used to calibrate the model.

We used 19 environmental variables from WorldClim (version 1.4; Hijmans et al., 2005) for modeling. The WorldClim variables have been broadly used for generating ENMs, including successful use with small non-volant mammals in the northern Neotropics (Anderson and Gonzalez, 2011; Anderson and Raza, 2010). These variables constitute derivatives of interpolated climatic data, in particular precipitation, temperature, and their seasonality. Maxent has been found to be relatively robust to correlated environmental variables (Elith et al., 2011). Although all 19 variables were considered by the algorithm, regularization (described in detail in Section 2.3) is employed to reduce the number of variables actually selected for inclusion in the final model (Phillips and Dudík, 2008; Elith et al., 2011).

2.3. Experimental design

To investigate the possibility of making better models with non-default settings, we apply the delete-one jackknife approach (Pearson et al., 2007; also called "$n-1$ jackknife" or "leave-one-out-jackknife" Peterson et al., 2011). Pearson et al. (2007) used a delete-one jackknife to develop a test assessing the statistical significance of ENMs made with small numbers of occurrence records. In contrast, we employ it as a means of quantifying measures of performance, as suggested recently (Peterson et al., 2011). Specifically, we use two common metrics of model performance to compare the effects of program settings and to approximate optimal model complexity.

In the present tuning experiments, we examined two settings specific to Maxent: feature class and regularization (Supplementary Fig. 3). In this aspect, we followed Anderson and Gonzalez (2011) (see Elith et al. (2010), Warren and Seifert (2011), Radosavljevic and Anderson (in press) for tuning Maxent models by varying only regularization and Syfert et al. (2013) for tuning Maxent models by varying only feature class). Feature class determines the kinds of constraints allowed in a model. A feature is a function of an environmental variable and in Maxent can be any single one or various combination of six classes: linear (L), quadratic (Q), product (P), threshold (T), hinge (H) or category indicator (C) (Phillips et al., 2006; Phillips and Dudík, 2008). The constraints placed on the model by features result in models of varying complexities. For instance, a model built with L features is less complex than one built with L and Q features. Hinge features model a piece-wise linear response to the environmental variable. This allows for parts of the response curve to be defined by a linear relationship while other parts can be defined by a more complex, non-linear relationship (Phillips and Dudík, 2008). Thus, L features represent a special (restrictive) case of H features and result in less complex models (Phillips and Dudík, 2008). Note that even if multiple feature classes are allowed for model-building, not all classes will necessarily be incorporated in the final model.

The default Maxent setting for feature class, called "auto features," applies the class or classes estimated to be appropriate for the particular sample size of occurrence records, according to a previous extensive tuning experiment (Phillips and Dudík, 2008). Phillips and Dudík (2008) selected the following feature classes for continuous variables as default for the corresponding sample sizes: all feature classes for at least 80 occurrence records; L, Q and H for sample sizes 15 to 79; L and Q for 10 to 14 records; only L for below 10 records (Phillips and Dudík, 2008).

While using complex feature settings allows Maxent to produce a model that is more sensitive to details of a species' environmental tolerance, it is possible that complex feature classes can lead to overfit models. Regularization is a penalty for including additional constraints (e.g., variables) in the model, with increasing penalties for higher weights applied to a given constraint. Hence, higher regularization decreases the chance that the model will be overly complex, or overfit to noise or bias in the occurrence data (Phillips et al., 2006). Regularization implicitly affects environmental variable selection by making it more likely that the value of some variables will be zero in the model (Elith et al., 2011). The default regularization value determined by Phillips and Dudík (2008) based on their previous tuning experiment is specific to each feature class and depends on the sample size, with higher values (stronger protection against overfitting) at lower sample sizes (Phillips and Dudík, 2008; Phillips et al., 2006). The Maxent software employs a convenient regularization multiplier that at once controls the intensity of regularization across all feature classes used to produce a model. The default regularization multiplier is 1; a regularization multiplier lower than the default is likely to result in a more restricted and potentially overfit prediction (environmentally and geographically), while a larger regularization multiplier should result in a broader, less discriminating, prediction (Phillips et al., 2006).

The previous tuning experiment performed by Phillips and Dudík (2008) involved a range of organisms and sample sizes (between 6 and 3162). It relied on measures of the area under the curve of the receiver operating characteristic plot (AUC) and log loss (which is highly correlated with AUC) to determine optimal regularization values for each feature class at each examined sample size. Phillips and Dudík (2008) used a randomly selected 60% of the presence records for calibration and 40% of the presence records for evaluation for each dataset. The optimal settings determined by their tuning experiment were presented as default settings for Maxent with the caveat that for datasets unlike those used for tuning, further experiments may be necessary to optimize the program's performance.

For several reasons, the extensive experiment conducted by Phillips and Dudík (2008) may have selected overly complex settings as optimal (and, hence, recommended these settings as default). First, only AUC and the highly correlated log loss were optimized, without considering omission rate (OR; see Section 2.5). AUC reflects the discriminatory ability of the model (favoring complex models) but, in contrast to OR, it does not quantify overfitting. Model selection based on OR would tend to select simpler models (Radosavljevic and Anderson, in press). Second, the extensive study regions used for most species likely led to violations of modeling assumptions regarding abiotic and dispersal-related drivers (Anderson and Raza, 2010), tending to reward overly complex models (Anderson, 2012). Finally, other than removing duplicate localities falling into the same map pixel, no spatial filtering was employed (in contrast to the present study; see Section 2.2), likely overestimating model performance due to spatial autocorrelation (Veloz, 2009), again favoring complex models.

2.4. Modeling

We used Maxent version 3.2.1 (Phillips et al., 2006; http://www.cs.princeton.edu/~schapire/maxent/) to calibrate models for each jackknife iteration (leaving the respective evaluation record to test the resulting model). To explore a broad swath of parameter space likely to be reasonable for species with very few records, we created models for all combinations of the following feature classes and regularization multipliers: L, H, LQ, and LQH; and 0.5, 0.75, 1.0, 1.25, 1.5, 1.75, and 2.0, for a total of 28 combinations for each species. We chose feature classes starting with the default (L) for the sample sizes involved here and increasing in complexity to include H and the combinations that represent default settings for slightly higher sample sizes (LQ and LQH). The selected regularization multipliers represent a narrower range of values than those explored in some previous studies (Warren and Seifert (2011): from 1 to 19; Radosavljevic and Anderson (in press): from 0.25 to 10). By selecting regularization values closer to the default, 1, we explore the possibility of relatively small changes in this parameter affecting model output.

2.5. Optimality criteria

In our evaluations of model performance for each combination of feature class and regularization multiplier, we employed two quantitative measures: omission rate (OR) (threshold-dependent) and AUC (threshold-independent). For each combination of settings, we calculated the value for the respective evaluation record of each jackknife iteration and then averaged across those iterations. Because presence-only occurrence data provide concrete information regarding the species' presence, omitted evaluation records represent true error given the following assumptions: correct species identification, accurate georeferencing, occurrence records only from source habitat, and correct selection of threshold. In contrast, various factors can lead to inflated estimates of commission (see apparent commission error and asymmetric errors; Anderson et al., 2003; Peterson et al., 2011). Hence, we used OR as the primary criterion for selecting optimal combinations of feature class and regularization multiplier. We employed the lowest presence threshold rule (LPT of Pearson et al., 2007; equal to the minimum training presence threshold (MTP) in Maxent) for converting continuous models to binary predictions (and determining whether an evaluation record falls into or out of the predicted area when calculating OR). With few records, this represents a conservative rule for thresholding, unlikely to overestimate the areas suitable for the species (which can be a problem for this rule with larger sample sizes; Radosavljevic and Anderson, in press). By using the LTP rule, we expect 0% test omission. Although this conservative rule may underestimate suitable areas with small sample sizes, evaluation ORs higher than 0% should tend to indicate overfitting.

The AUC calculated with presence-background evaluation data represents a threshold-independent measure of a model's discriminatory ability (Phillips et al., 2006). Some studies have identified the presence-background AUC as a questionable measure of model performance (Lobo et al., 2008; Warren and Seifert, 2011), but it does provide valid and useful information under some circumstances (see clarifications in Peterson et al. (2011)). For example, it is relevant and appropriate for comparisons among model settings for a single study species in a single study region, as is the case in our current study. In cases where the lowest OR (the primary optimality criterion) corresponded to multiple feature–class–regularization–multiplier combinations, we selected from among those and designated as optimal the combination with the highest evaluation AUC (the secondary optimality criterion). We extracted evaluation AUC values from the Maxent output for each jackknife iteration and averaged them as for OR.

2.6. Assessing model quality and similarity

In addition to our quantitative comparisons, we inspected model predictions in geography visually to assess quality and concordance with known geographic features, climatic patterns, vegetation zones, and natural history/habitat information for the species (Anderson and Jarrín-V, 2002). We did this for models produced with the default settings as well as those made using settings determined as optimal via the quantitative evaluations. Specifically, for each suite of models (produced with either default or optimal settings), we averaged the respective *n* predictions, creating a composite prediction. The logistic output was used for all visualizations (Phillips and Dudík, 2008).

To compare optimal and default models quantitatively in geographic space, we used the niche similarity metrics Schoener's *D* and the corrected modified Hellinger distance (*I*, Warren et al., 2008; Rödder and Engler, 2011). *I* has been identified as often overestimating model similarity, whereas *D* is a more conservative measure of this metric (Rödder and Engler, 2011). These metrics range from 0 (no overlap) to 1 (models identical). *I* and *D* were calculated using the formulas presented by Rödder and Engler (2011) and implemented in the Python programming language (http://www.python.org) using the functions available in NichePy version 1.1 (http://www.purl.org/NichePy; Bentlage and Shcheglovitova, 2012; the code implementing the tests performed in this study is available at https://github.com/mshcheg/Maxent-Jackknife-Scripts.git). For each species, we generated distributions for *I* and *D* by calculating each metric for all corresponding pairs of jackknife iterations (1 to *n*) for models built using optimal settings vs. models built using default settings: Optimal_1 vs. Default_1, Optimal_2 vs. Default_2, ..., Optimal_n vs. Default_n.

3. Results

3.1. Omission rate

H. australis generally suffered from higher ORs than *H. teleus* (Fig. 2). Observed average ORs were variable across regularization multipliers for each feature class, with higher regularization multipliers generally leading to lower ORs within a given feature class. However, the default feature class (L) displayed notably less variability across changing regularization multipliers. The majority of feature–class–regularization–multiplier combinations omitted the evaluation record in four or more of the nine jackknife iterations, including the default settings (L/1.0; evaluation record omitted in five out of nine jackknife iterations). For this species, two potentially complex settings, LQH and H, exhibited a large range of ORs across the examined regularization multipliers. Notably, they showed a similar pattern for OR when using low regularization multipliers. For both H/0.5–1.0 and LQH/0.5–1.0, the evaluation record was omitted in eight out of nine jackknife iterations (the highest observed OR). In contrast, the H/1.75–2.0 combinations showed the lowest OR, omitting the evaluation record in only two out of nine jackknife iterations.

The measures of OR for *H. teleus* were generally lower and varied less across regularization multipliers for each feature class (Fig. 2). The H/1.75 and H/2.0 combinations displayed the lowest OR, omitting the evaluation record in only one out of six jackknife iterations. Again, within each feature class, higher regularization multipliers generally corresponded to lower ORs. The default settings L/1.0 resulted in omission of the

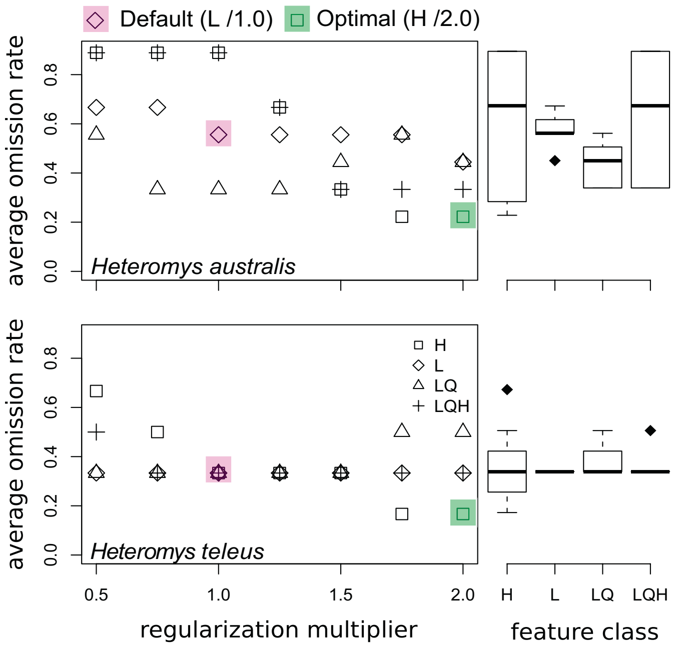


Fig. 2. Omission rates (ORs) for jackknife tuning experiments of *Heteromys australis* and *H. teleus* in Ecuador and southwestern Colombia. ORs were averaged over jackknife iterations for each set of feature classes for each of the examined regularization multipliers. Box plots show the median, first and third quartiles, and outliers for average ORs compared across regularization multipliers for each feature class.

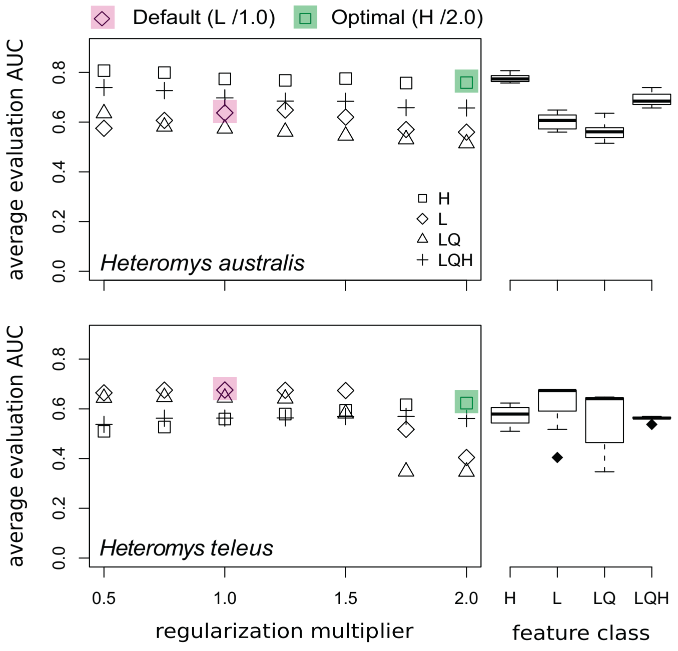


Fig. 3. Evaluation AUCs for jackknife tuning experiments of *Heteromys australis* and *H. teleus* in Ecuador and southwestern Colombia. Evaluation AUCs were averaged over jackknife iterations for each set of feature classes for each of the examined regularization multipliers. Box plots show the median, first and third quartiles, and outliers for average AUCs compared across regularization multipliers for each feature class.

evaluation record in two of the six jackknife iterations. Most of the feature–class–regularization–multiplier combinations showed ORs identical to the default settings, with seven combinations that had a higher OR than the default settings representing notable exceptions. The H/0.5 combination had the highest omission rate, omitting the evaluation record in four out of six jackknife iterations.

3.2. AUC

Experiments for *H. australis* showed strong differences in average evaluation AUCs among feature classes, but rather consistent AUC values across regularization multipliers within a feature class (Fig. 3). Hinge features displayed the highest AUC values of all feature classes across all regularization multiplier values. Linear–Quadratic–Hinge had the second highest AUC values, L the third, and LQ the fourth (except for the lowest regularization multiplier, where the last two switched positions).

For *H. teleus*, evaluation AUCs across feature classes responded to increased regularization in different ways (Fig. 3). Linear–Quadratic–Hinge features performed nearly constant across all regularization multiplier values. Linear and LQ features displayed similar patterns, obtaining the highest AUC values at low regularization multipliers and then dropping in performance at high regularization multiplier values. In contrast, H features showed low AUC values at low regularization multipliers, but AUC values rose steadily for H features as regularization multipliers increased. The highest AUC was observed for the L feature class at low to intermediate regularization multipliers (including the default settings).

3.3. Optimal models

Based on the employed criteria, we did not find the default settings to be optimal for either dataset. We identified the H feature class and the highest examined regularization multiplier (2.0) as the optimal combination for both datasets (Table 1). Because of the sequential nature of the optimality criteria, we did not necessarily select as optimal the combination with the highest evaluation AUC. However, the difference between the average evaluation AUC of the optimal combination and that of the highest average evaluation AUC was small: *H. australis*: difference of 0.05 (maximum AUC 0.81 for H/0.5; 0.76 for optimal settings H/2 and 0.64 for default settings); *H. teleus*: difference of 0.06 (maximum AUC 0.68 for L/1.0 (default settings); 0.62 for optimal settings).

3.4. Model similarity

Despite the differences in OR and AUC between default and optimal settings, we observed high levels of similarity in geographic space between comparisons of jackknife datasets for both *H. australis* and *H. teleus* (supplementary Figs. 1 and 2). For both species, the average *I* value was 0.988. The average *D* value for *H. australis* was 0.877, whereas that for *H. teleus* was 0.911.

3.5. Qualitative evaluations of models

In general, models built with optimal settings agreed more with our current knowledge of the natural history/habitat information

Table 1
Summary of quantitative evaluation metrics for jackknife tuning experiments of *Heteromys australis* and *H. teleus* in Ecuador and southwestern Colombia. Average evaluation omission rates and average evaluation AUCs are provided for default and optimal feature classes and regularization multiplier combinations. Omission rates were calculated using the least presence threshold rule.

	Heteromys australis	*Heteromys teleus*
Sample size	9	6
Default combination	L/1.0	L/1.0
Optimal combination	H/2.0	H/2.0
Default average evaluation OR	0.56	0.33
Optimal average evaluation OR	0.22	0.17
Default average evaluation AUC	0.64	0.68
Optimal average evaluation AUC	0.76	0.62

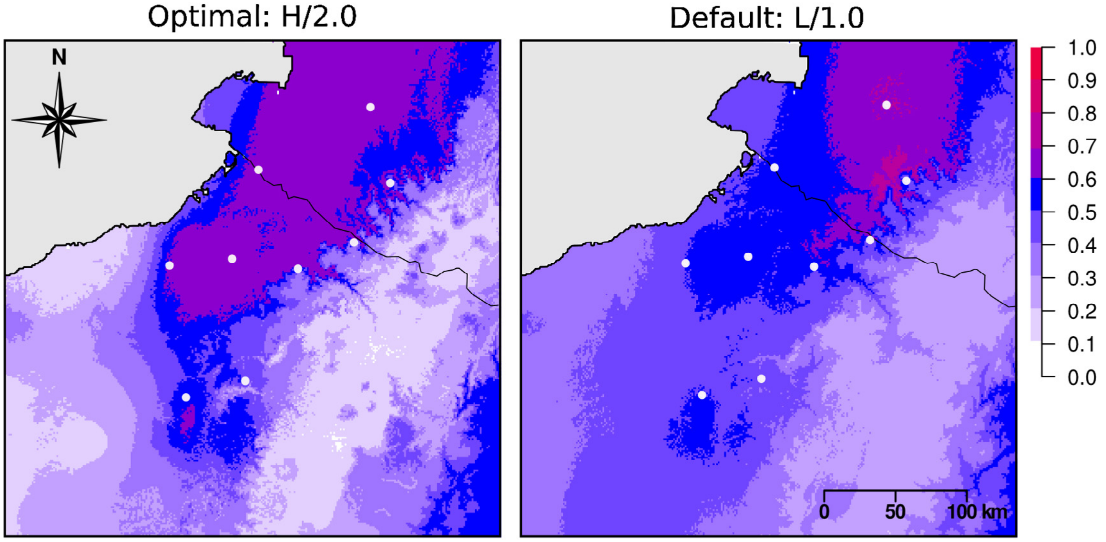


Fig. 4. Composite Maxent models from tuning experiments for *Heteromys australis* in western Ecuador and southwestern Colombia. Predictions (logistic output) were averaged over jackknife iterations for default and optimal combinations of feature classes and regularization multiplier.

and environmental preferences of the species than did models made with default settings. For *H. australis*, the composite model for optimal settings showed simpler yet more realistic geographic patterns (Fig. 4). This model identified a much larger region of high prediction (above logistic value 0.7) in extreme northwesternmost parts of Ecuador and into southwestern Colombia (contiguous areas with similar uniformly non-seasonal wet rain forest where the species is known). In contrast, in the composite default-settings model, the area of similarly high prediction was restricted mostly to an oval-shaped region in southwestern Colombia. In addition, the optimal settings led to a much sharper definition of highly suitable conditions in the low valley of the Río Mira, which holds conditions that should be suitable for the species. Finally, the composite model for optimal settings indicated a much stronger decrease in prediction strength in more seasonal areas between 79°00′ and 79°30′W. This region represents a known transition zone of rapidly increasing precipitation seasonality to the west, in which *H. australis* has never been recorded (Anderson and Jarrín-V, 2002).

Similarly, the composite model for *H. teleus* made with optimal settings comprised simpler and more reasonable geographic patterns than did that for the default settings (Fig. 5). This suite of models for optimal settings predicted no areas with an average strength greater than 0.7 (logistic value). In contrast, the composite model for default settings indicated a ribbon-like area, expanding to the north, of higher prediction (>0.7) abutting the western slopes of the Andes, culminating in a small area of very high prediction. Those areas of high prediction were followed to the west by a series of areas decreasing in prediction strength. In contrast, the model corresponding to optimal settings indicated as fairly suitable (logistic value greater than 0.6) a broad region closely matching areas of evergreen vegetation. Most of that region corresponds to highly seasonal yet evergreen forests matching the conditions of

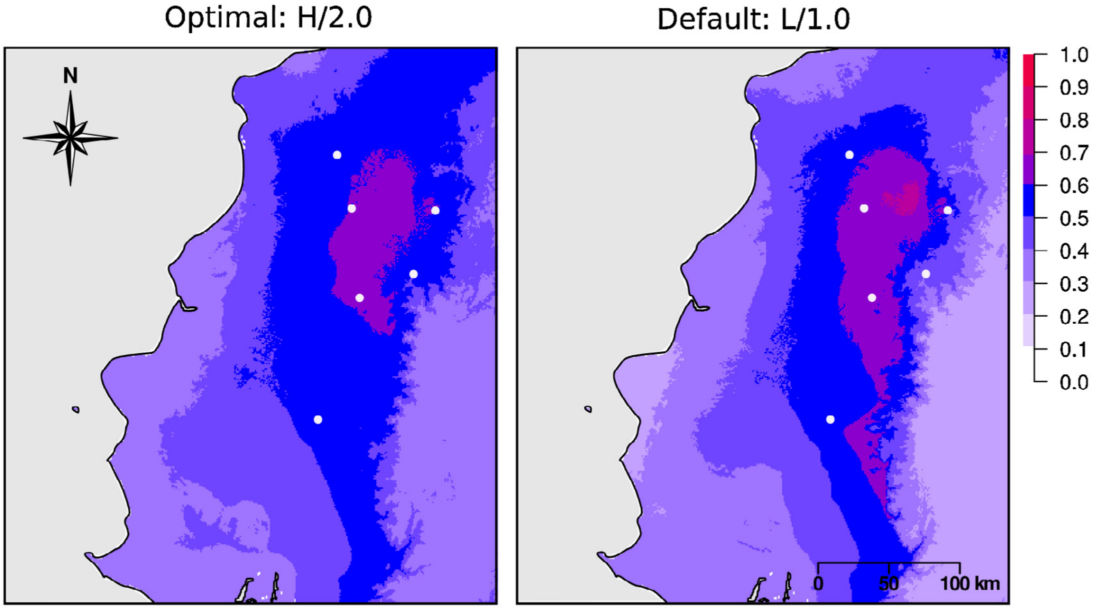


Fig. 5. Composite Maxent models from tuning experiments for *Heteromys teleus* in western Ecuador. Predictions (logistic output) were averaged over jackknife iterations for default and optimal combinations of feature class and regularization multiplier.

most known localities of *H. teleus* (although the northern portion that holds records of *H. australis* is unseasonal evergreen forest; see Anderson and Martínez-Meyer (2004) for possible competitive exclusion between the species). Neither composite model adequately delimited small areas of likely suitable conditions for the species in the southern portion of the Cordillera de la Costa. This is not surprising, given that those regions hold evergreen cloud forest due to highly localized horizontal precipitation (mist and fog) coming off the Pacific, rather than to vertical precipitation reflected in the environmental variables used to make the model (Anderson and Jarrín-V, 2002).

4. Discussion

4.1. Optimal settings for the examined species

Several notable patterns emerge from the quantitative evaluations. Although the species differed in details, a few prevailing trends existed for OR and AUC across regularization multipliers. Most feature classes showed lower ORs at higher regularization multipliers. This result is consistent with the higher protection against overfitting provided by higher regularization multipliers—which should lead to simpler, less restricted predictions. In contrast, AUC values varied comparatively less across regularization multipliers. Rather, stark differences in AUC emerged among feature classes. For *H. australis*, the orderings among feature classes remained consistent across regularization multipliers, whereas their performance inverted at high regularization multipliers for *H. teleus*. Interestingly, at high regularization multipliers, both species showed the same ordering of feature classes: H (highest), LQH, L, and LQ (lowest). These results indicate generally higher performance for more complex sets of feature classes at high regularization multipliers, especially for those including H. Hinge features are more complex than L features, which are the default for fewer than 10 occurrence records (recall, L features are a special and very restrictive case of H features; Phillips and Dudík, 2008). We note the similar ORs observed for H and LQH features and the large range these ORs displayed across regularization multipliers. The similar behavior of these two combinations of feature classes could be due to the presence of H features in both. We speculate that the large range of OR values for these combinations could be due to the greater effect that protection against overfitting has on feature classes that allow modeling of a more complex response.

Furthermore, the optimal regularization multipliers were higher than those suggested as default by the software. The finding of increased model performance at higher regularization multipliers agrees with at least four studies assessing model complexity in Maxent in different contexts: climate change (Elith et al., 2010), evaluating model complexity and predictivity using information criteria (AICc; Warren and Seifert, 2011), spatially independent evaluations (Radosavljevic and Anderson, in press), and effects of sampling bias on model performance (Anderson and Gonzalez, 2011). Notably, the former three studies examined high sample sizes, whereas the latter one dealt with a species with few occurrence records, as here.

The optimal settings determined here produce an apparent paradox regarding feature classes and regularization multipliers, which can be resolved by considering the findings together. The tuning exercises selected as optimal more complex feature class combinations, yet higher regularization multipliers (which tend to produce simpler models). While it may seem counter-intuitive to use more complex feature classes with few occurrence records, coupling complex features with a high regularization may prevent the model from overfitting to the input data. More complex feature classes (and sets of features) allow more flexibility in the shape (and complexity) of the modeled response of the species to an environmental variable (e.g., H vs. L). Here, this flexibility led to better models, if coupled with greater protection against overfitting (i.e., higher regularization). This conclusion echoes that found by another study addressing model complexity with Maxent for species with few occurrence records (Anderson and Gonzalez, 2011).

Additionally, H features with high regularization multipliers have been found to perform well when used for Maxent models of well sampled species (over 1000 occurrence records; Elith et al., 2010). In that case, H features combined with a high regularization multiplier (2.5) resulted in a simpler (and more realistic) model than the default settings, which specified a lower regularization multiplier and allowed use of more feature classes. Syfert et al. (2013) also selected simpler features (LQ) for modeling species with a large number of occurrence records (over 400). In their study, Syfert et al. (2013) found that simpler features led to similar performance to the complex default feature set, when data were corrected for sampling bias. Future research with many species of varying sample sizes is necessary to address the possible generality of these intriguing patterns.

4.2. Evaluations of models in geographic space

Despite high levels of quantitative similarity, the models made with optimal settings showed more reasonable patterns than did those produced using default settings. We focus our interpretation on *D*, a more conservative and realistic measure of similarity (*I* often overestimates model similarity; Rödder and Engler, 2011). The respective average values for *D* comparing optimal vs. default settings (0.88, *H. australis*; 0.91, *H. teleus*) both quantified high similarity in geographic space but fell short of identical predictions (1.00). These differences, due only to changes in software settings, correspond to approximately 10% difference in the predictions. Visual inspection of the predictions in geography indicated that the differences correspond to sensible patterns given known natural history/habitat information for the study species, with the models based on optimal settings providing more realistic predictions. The observed differences could be due to overfitting to the conditions present at documented localities (in the default settings), more realistic shapes for the response curves (in the optimal settings), or both.

4.3. Recommendations and future directions

The present results indicate that a jackknife tuning approach holds promise and point to complementary avenues for future research.

Our analyses were conducted on relatively unbiased occurrence records, here, via spatial filtering to reduce the effects of biased geographic sampling (Anderson and Gonzalez, 2011; Anderson, 2012). Most datasets, especially those from natural history museums and herbaria, will require some sort of filtering, or, alternatively, correction for sampling bias (when it can be quantified via direct information or an index created using the results of sampling for an overall target group; Anderson, 2003; Phillips et al., 2009; Syfert et al., 2013). Otherwise, biased records will tend to inflate estimates of performance and lead to the selection of overfit models. The degree of filtering necessary remains an open research question (Anderson, 2012).

Several issues related to ORs should be considered carefully in future studies. Here, we chose a low OR as the primary criterion and utilized AUC (overall ranking ability) as a secondary criterion.

Future experiments could weight these in another manner but should take into account principles of asymmetrical loss functions for presence vs. absence or background information (Peterson et al., 2011). By using feature class and regularization multiplier combinations that lead to low ORs, we generated models that were not restricted to the most common environments in the input dataset of occurrence records. Such models could provide better insights for discovering unknown populations or potentially for determining the extent of protected areas for species facing habitat loss. Additionally, other thresholding rules should be explored to determine ORs. Threshold selection was doubly difficult here. First, choosing an appropriate threshold becomes difficult with limited occurrence records (Bean et al., 2012). Furthermore, most of the thresholding rules commonly employed for presence–absence datasets (Liu et al., 2005) are not valid for presence–pseudoabsence or presence–background situations (due to apparent commission error; Anderson et al., 2003; Peterson et al., 2011). Other possibilities include thresholding rules based on the cumulative Maxent output (which have theoretical expectations of omission rates on independent data; Phillips et al., 2006).

Varying both feature class and the regularization multiplier emerged as a key element of model tuning. Because the highest regularization multipliers employed led to the highest performance in this study, even higher regularization multipliers should be included in future experiments. However, future studies should consider the association between increases in the regularization multiplier and decreases in ORs. This relationship indicates a need to choose ranges of regularization multipliers carefully, so that they do not lead to a deflation of ORs via unrealistic models (e.g., by predicting nearly the entire study region as suitable). We recognize that our experiments examined only a subset of the possible reasonable combinations of feature class and regularization multipliers. In addition to a larger range of regularization multipliers, the category indicator, product and threshold feature classes represent interesting additions to future tuning experiments. The current results highlight the H feature class as promising for use with small sample sizes. We recommend it as one of several options to be compared in tuning experiments for datasets with few occurrence records.

Additional research is necessary to provide further guidelines regarding optimal complexity for species with few occurrence records. A previous study indicated that application of the $n-1$ jackknife was reasonable in determining statistical significance of models for datasets with fewer than 25 records (Pearson et al., 2007;see also Anderson and Raza, 2010). Pearson et al. (2007) generated significantly predictive models with as few as five occurrence records, and here we conduct tuning experiments for as few as six records. We suggest that researchers should acknowledge the likely simplistic nature of models produced with such small datasets, while keeping in mind that such models still could be potentially useful for some applications, such as discovering additional populations of extremely poorly known species (Peterson et al., 2011). As an alternative to species-specific tuning based on performance on withheld data (here with a jackknife for species with few occurrence records), future research should compare this approach to selection of model complexity via information criteria, using all localities to calibrate and evaluate the model (i.e., AICc; Warren and Seifert, 2011). Nevertheless, a benefit of the present jackknife approach is that it allows for quantification of uncertainty due to variation in the environmental information among occurrence records, as well as detection of records that hold substantially different environmental conditions. Finally, the generality of the jackknife approach could be explored by using simulated data for similar tuning experiments.

Acknowledgments

This research was made possible by funding from the U. S. National Science Foundation (NSF DEB-1119915 and DEB-0717357, including a Research Experiences for Undergraduates supplement to support MS) and the City College of the City University of New York. MS was supported by awards from the City College Fellowship, Gerald S. Brenner Endowed Science Scholarship, and the City College Academy for Professional Preparation. Funds to present preliminary results were provided by the International Biogeography Society (Student Travel Award, to MS). We would like to thank the members of the Anderson lab, past and present – especially Aleksandar Radosavljevic, Darla M. Thomas, Eliécer E. Gutiérrez, Mariano Soley-Guardia, Peter J. Galante, and Robert A. Boria – as well as Robert Muscarella and two anonymous reviewers for thoughtful comments that helped greatly improve this manuscript. We would also like to thank Dan L. Warren for his insights and advice on jackknife variance estimates.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ecolmodel.2013.08.011.

References

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

Anderson, R.P., 2012. Harnessing the world's biodiversity data: promise and peril in ecological niche modeling of species distributions. Ann. N.Y. Acad. Sci. 1260, 66–80.

Anderson, R.P., 2013. A framework for using niche models to estimate impacts of climate change on species distributions. Ann. N.Y. Acad. Sci., in press.

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

Anderson, R.P., Jarrín-V, P., 2002. A new species of spiny pocket mouse (*Heteromyidae*: *Heteromys*) endemic to western Ecuador. Am. Mus. Novit. 3382, 1–26.

Anderson, R.P., Lew, D., Peterson, A.T., 2003. Evaluating predictive models of species' distributions: criteria for selecting optimal models. Ecol. Model. 162, 211–232.

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. Biol. Conserv. 116 (2), 167–179.

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. J. Biogeogr. 37, 1378–1393.

Bean, W.T., Stafford, R., Brashares, J.S., 2012. The effects of small sample size and sample bias on threshold selection and accuracy assessment of species distribution models. Ecography 35 (3), 250–258.

Bentlage, B., Shcheglovitova, M., 2012. NichePy: modular tools for estimating the similarity of ecological niche and species distribution models. Methods Ecol. Evol. 3 (3), 484–489.

Cayuela, L., Golicher, D.J., Newton, A.C., Kolb, M., de Alburquerque, F.S., Arets, E.J.M.M., Alkemade, J.R.M., Pérez, A.M., 2009. Species distribution modeling in the tropics: problems, potentialities, and the role of biological data for effective species conservation. Trop. Conserv. Sci. 2 (3), 319–352.

Elith, J., Kearney, M., Phillips, S., 2010. The art of modelling range-shifting species. Methods Ecol. Evol. 1, 330–342.

Elith, J., Phillips, S., Hastie, T., Dudík, M., Chee, Y.E., Yates, C., 2011. A statistical explanation of MaxEnt for ecologists. Divers. Distrib. 17 (1), 43–57.

Gaubert, P., Papeş, M., Peterson, A.T., 2006. Natural history collections and the conservation of poorly known taxa: ecological niche modeling in central African rainforest genets (*Genetta* spp.). Biol. Conserv. 130 (1), 106–117.

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. Int. J. Climatol. 25, 1965–1978.

Lawler, J.J., Wiersma, Y.F., Huettmann, F., 2011. Using species distribution models for conservation planning and ecological forecasting. In: Drew, C.A., Wiersma, Y.F., Huettmann, F. (Eds.), Predictive Species and Habitat Modeling in Landscape Ecology: Concepts and Applications. Springer, New York, pp. 271–290.

Liu, C., Berry, P.M., Dawson, T.P., Pearson, R.G., 2005. Selecting thresholds of occurrence in the prediction of species distributions. Ecography 28 (3), 385–393.

Lobo, J.M., Jiménez-Valverde, A., Real, R., 2008. AUC: a misleading measure of the performance of predictive distribution models. Glob. Ecol. Biogeogr. 17 (2), 145–151.

Papeş, M., Gaubert, P., 2007. Modelling ecological niches from low numbers of occurrences: assessment of the conservation status of poorly known viverrids (Mammalia, Carnivors) across two continents. Divers. Distrib. 13, 890–902.

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. J. Biogeogr. 34, 102–117.

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., Soberón, J., Pearson, R.G., Anderson, R.P., Martínez-Meyer, E., Nakamura, M., Araújo, M.B., 2011. Ecological niches and geographic distributions. Monographs in Population Biology, Princeton University Press, Princeton, NJ.

Phillips, S.J., Anderson, R.P., Schapire, R.E., 2006. Maximum entropy modeling of species geographic distributions. Ecol. Model. 190, 231–259.

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., 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. Ecol. Appl. 19, 181–197.

Radosavljevic A. and Anderson R.P., Making better Maxent models of species distributions: complexity, overfitting, and evaluation, J. Biogeogr. in press.

Rödder, D., Engler, J.O., 2011. Quantitative metrics of overlap in Grinnellian niches: advances and possible drawbacks. Glob. Ecol. Biogeogr. 20 (6), 915–927.

Syfert, M.M., Smith, M.J., Coomes, D.A., 2013. The effects of sampling bias and model complexity on the predictive performance of MaxEnt species distribution models. PLoS One 8 (2), e55158, http://dx.doi.org/10.1371/journal.pone.0055158.

Veloz, S.D., 2009. Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. J. Biogeogr. 36, 2290–2299.

Warren, D.L., Glor, R.E., Turelli, M., 2008. Environmental niche equivalency versus conservatism: quantitative approaches to niche evolution. Evolution 62 (11), 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. Ecol. Appl. 21 (2), 335–342.

Wilting, A., Cord, A., Hearn, A.J., Hesse, D., Mohamed, A., Traeholdt, C., Cheyne, S.M., Sunarto, S., Jayasilan, M., Ross, J., Shapiro, A.C., Sebastian, A., Dech, S., Breitenmoser, C., Sanderson, J., Duckworth, J.W., Hofer, H., 2010. Modelling the species distribution of flat-headed cats (*Prionailurus planiceps*), an endangered southeast Asian small felid. PLoS One 5 (3), e9612.

Wisz, M.S., Hijmans, R.J., Peterson, A.T., Graham, C.H., Guisan, A., 2008. Effects of sample size on the performance of species distribution models. Divers. Distrib. 14, 763–773.