ECOGRAPHY 29: 773–785, 2006

The effect of sample size and species characteristics on performance of different species distribution modeling methods

Pilar A. Hernandez, Catherine H. Graham, Lawrence L. Master and Deborah L. Albert

Hernandez, P. A., Graham, C. H., Master, L. L. and Albert D. L. 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. – Ecography 29: 773–785.

Species distribution models should provide conservation practioners with estimates of the spatial distributions of species requiring attention. These species are often rare and have limited known occurrences, posing challenges for creating accurate species distribution models. We tested four modeling methods (Bioclim, Domain, GARP, and Maxent) across 18 species with different levels of ecological specialization using six different sample size treatments and three different evaluation measures. Our assessment revealed that Maxent was the most capable of the four modeling methods in producing useful results with sample sizes as small as 5, 10 and 25 occurrences. The other methods compensated reasonably well (Domain and GARP) to poorly (Bioclim) when presented with datasets of small sample sizes. We show that multiple evaluation measures are necessary to determine accuracy of models produced with presence-only data. Further, we found that accuracy of models is greater for species with small geographic ranges and limited environmental tolerance, ecological characteristics of many rare species. Our results indicate that reasonable models can be made for some rare species, a result that should encourage conservationists to add distribution modeling to their toolbox.

P. A. Hernandez (pilar_hernandez@natureserve.org), L. L. Master and D. L. Albert, NatureServe, 1101 Wilson Blvd., 15th Floor Arlington, VA 22209, USA. – C. H. Graham, Dept of Ecology and Evolution, State Univ. of New York at Stony Brook, 650 Life Sciences Building, Stony Brook, NY 11794-5245, USA.

Effective conservation plans require accurate estimates of the spatial distributions of the species they are trying to protect. With such information conservationists can predict how a species' distribution will respond to landscape alteration and environmental (climate) change. Species distribution modeling can provide a measure of a species' occupancy potential in areas not covered by biological surveys and consequently is becoming an indispensable tool to conservation planning (Guisan and Zimmermann 2000, Corsi et al. 2000, Elith and Burgman 2003, Loiselle et al. 2003). These models combine points of known occurrence with spatially continuous environmental layers to infer ecological requirements of a species, generally using a statistical algorithm. The geographic distribution of a species is then predicted by mapping the area where these environmental requirements are met (Elith et al. 2006). Depending on data quality and the application at hand, these models can assist in identifying previously unknown populations, determining sites of high candidacy for reintroductions, guiding additional surveys, and informing selection and management of protected areas (Graham et al. 2004).

Of particular interest to conservation biologists are rare species. By definition, rare species have sparse and/or restricted spatial distribution patterns (Rabinowitz et al. 1986, Kattan 1992, Gaston 1997), which often means they are habitat specialists and that there is a

Accepted 5 June 2006

ISSN 0906-7590

limited number of sites of known occurrence. Species ecological characteristics (i.e. range sizes and ecological specialization) have been shown to influence model performance (Segurado and Araujo 2004, Brotons et al. 2004, McPherson et al. 2004, Elith et al. 2006). Generally, models for species with broad geographic ranges and environmental tolerances tend to be less accurate than those for species with smaller geographic ranges and limited environmental tolerance (Manel et al. 2001, Boone and Krohn 2002, Kadmon et al. 2003, Thuiller et al. 2004, Luoto et al. 2005, Elith et al. 2006).

Small sample sizes pose challenges to any statistical analyses and result in decreased predictive potential when compared to models developed with more occurrences (Stockwell and Peterson 2002, McPherson et al. 2004). As sample size increases, accuracy should also increase until achieving its maximum accuracy potential thereby reaching an asymptote. The maximum accuracy potential and the sample size at which the asymptote is reached will depend on the study area and species, the quality and spatial resolution of the environmental and species occurrence data used to develop the model, and the modeling method itself. While many previous researchers have investigated the effect of sample size on model accuracy, most have not explicitly manipulated sample size by species (Segurado and Araujo 2004, Brotons et al. 2004), making it difficult to evaluate the effects of species ecological characteristics versus sample size on model accuracy. For instance, if a species has a limited range it is likely that proportionally more of its environmental space is sampled with fewer points, than a species with a large range. Two studies, Stockwell and Peterson (2002) and McPherson et al. (2004) manipulated sample size and determined that models built with fewer points were generally less accurate, however the latter used sample sizes (50, 100, 300, and 500) much larger than those typically available for species of conservation concern. To further explore the variation among sample size and model performance, we generated models for 18 California taxa with varying degrees of habitat specialization using four modeling methods and a variety of sample sizes characteristic of rare species ($n = 5$, 10, 25, 50, 75, or 100).

The potential for a predictive species' distribution model to aid conservation planning will depend on the model's ability to accurately depict the species' occupancy potential in the geographic region in question. Evaluating these models with presence/absence data is challenging (Fielding and Bell 1997, Pearce and Ferrier 2000, Manel et al. 2001, Elith and Burgman 2003). This task is further complicated when only presence occurrence data are available because it is very difficult to evaluate false-positive prediction (commission) errors. Given this limitation, many modelers opt to evaluate only omission errors and ignore commission errors. This is insufficient because a model that has no omission errors can also have high commission errors because as omission errors decrease, commission errors tend to increase and vice versa (Fielding and Bell 1997). Evaluating a model for omission alone will fail to identify models that balance commission and omission errors (or weights one slightly above the other depending on which is considered by the modeler to be a more serious problem) and therefore fail to identify models with the highest predictive ability. Given these complexities, research on model evaluation for presence-only modeling is vital. Here we take a multifaceted approach using three different evaluation approaches that vary in how they measure omission and commission errors. To evaluate overall model fit we use 1) receiver operating curves (ROC), which are threshold independent and include both omission and commission error. 2) We use predictive success to measure omission error while 3) the spatial comparison of predictions by models generated with the full species occurrence data to those of models generated with fewer observations provides an assessment of both commission error and model stability. This final measure is essential for evaluating which modeling method performs well with incomplete species occurrence data sets, a common condition when modeling rare species distributions.

Numerous species distribution modeling methods exist, each unique with regard to their data requirements, statistical methods and overall ease of use (Guisan and Zimmermann 2000, Elith and Burgman 2003, Elith et al. 2006). These different modeling methods can produce clearly different geographic predictions and therefore resultant conservation strategies, even when using the same data (Loiselle et al. 2003). We chose four modeling methods potentially useful to conservation planning, Bioclim (Nix 1986), Domain (Carpenter et al. 1993), GARP (Stockwell and Peters 1999), and Maxent (Phillips et al. 2004, 2006). These methods were chosen because they are easy to use, batchable, produced useful predictions in other research (Lindenmayer et al. 1991, Gillison 1997, Anderson and Martínez-Meyer 2004, Phillips et al. 2004), and do not require an explicit quantification of absence to formulate a predicted distribution model. Further, they varied in predictive performance in a recent comprehensive model study (Elith et al. 2006) in which Bioclim performed relatively poorly, Domain and GARP had intermediate performance, and Maxent performed very well.

In sum, we test models that use presence occurrence data, which is of great utility because the vast majority of biotic data available to modelers are presence-only. We manipulate sample sizes to include those very small samples typical of rare species and quantify ecological characteristics of species to evaluate the relative influence of both species ecology and sample size on model

performance. Further, we use a suite of evaluation procedures that complement each other to obtain as complete a picture as possible of the predictive ability of each model. Our comprehensive analyses should stimulate continued use of distributional modeling in conservation management, particularly when only limited occurrence data are available.

Materials and methods

Species occurrence data

Species occurrence data were extracted from the 'RareFind' dataset of the California Natural Diversity Database (CNDDB). CNDDB maintains information about the natural history and locations of rare, threatened, endangered, and special status species and natural communities of California. Eighteen terrestrial taxa (Table 1) were selected from CNDDB that have habitat requirements that lend themselves to climatic modeling (i.e. are not restricted to a geological formation or micro-habitat that is unlikely to be detected using climatic information) and had at least 150 recorded occurrences with an estimated mapping precision of no larger than 1 km. To eliminate a potential bias of clustered occurrences, the datasets were filtered so that there was only one record per 1 km^2 cell for each species. We selected species from a variety of taxonomic groups that had much of their range within California. Two of these are easily identified subspecies of widespread species. We qualitatively compared occurrence data with known distributional ranges for each species to ensure that the known occurrences represented each species' entire geographic range within California.

Environmental data

We compiled climate and topographic geographic information systems (GIS) layers covering the geographic extent of California. The original data were at a resolution of 1 km or finer, and the fine scale data were resampled to the coarser 1 km resolution. We obtained climatic layers from Daymet, an eighteen year daily surface weather database (<www.daymet.org>). The Daymet monthly layers were further summarized into 36 biologically relevant climatic variables following Nix (1986). We derived slope from a digital elevation model (DEM) of 30 m resolution for California, acquired from the USGS national elevation dataset. We eliminated correlated environmental variables using a Pearson correlation test to obtain the 10 variables used as predictors in our models (Table 2; Johnson et al. 2002).

Modeling methods

We randomly selected 150 occurrences from the CNDDB dataset for each of the 18 study species so that we had the same base number of occurrences for each species. We partitioned the data from each species into training (occurrences used to develop the prediction models) and evaluation datasets. Training datasets with sample sizes of 5, 10, 25, 50, 75, or 100 were generated by randomly selecting the required number of observations from each taxon's full 150 occurrence dataset. Then we used 50 randomly selected records from those remaining to create the evaluation dataset. We repeated the process for each sample size category and taxon to obtain 10 replicates of each species.

Predictive distribution models were formulated using the four different modeling techniques, entering the

Table 1. Californian taxa modeled.

Class	Taxa	Common name
Insect	*Danaus plexippus*	Monarch butterfly
Amphibian	*Ambystoma californiense*	California tiger salamander
Amphibian	*Scaphiopus hammondii*	Western spadefoot
Amphibian	*Rana aurora*	Red-legged frog
Amphibian	*Rana boylii*	Foothill yellow-legged frog
Amphibian	*Rana muscosa*	Mountain yellow-legged frog
Reptile	*Clemmys marmorata*	Western pond turtle
Reptile	*Phrynosoma coronatum*	Coast horned lizard
Reptile	*Cnemidophorus hyperythrus*	Orange-throated whiptail
Bird	*Accipiter gentilis*	Northern goshawk
Bird	*Buteo swainsoni*	Swainson's hawk
Bird	*Grus canadensis tabida*	Greater sandhill crane
Bird	*Athene cunicularia*	Burrowing owl
Bird	*Strix occidentalis caurina*	Northern spotted owl
Bird	*Polioptila californica*	California gnatcatcher
Bird	*Agelaius tricolor*	Tricolored blackbird
Mammal	*Spermophilus mohavensis*	Mohave ground squirrel
Mammal	*Arborimus pomo*	Red tree vole

Table 2. Environmental predictor variables used in each model.

Variable
Annual temperature range
Isothermality (mean diurnal range/temperature annual range)
Annual mean precipitation
Precipitation of the warmest quarter
Coefficient of variation of monthly precipitation
Annual total radiation
Annual radiation range
Coefficient of variation of monthly relative humidity
Elevation
Slope

occurrence datasets as the dependent variable and the selected environmental variables as the predictors. Hence, for each species we generated 244 models (10 models for each modeling method for each of the six sample sizes and one model for each modeling method with the full dataset of 150 occurrences). The modeling methods are briefly described below.

1) Bioclim: this "boxcar" environmental envelope algorithm identifies locations that have environmental values that fall within the range of values measured from the occurrence dataset (Nix 1986, Busby 1991). The area, often termed the "core bioclimate", represents the 5–95 percentile limits and is calculated by disregarding 5% of the lower and higher values of each climatic index thereby attempting to reduce the impact of outliers (Carpenter et al. 1993, Farber and Kadmon 2003). For the purposes of this study and to facilitate analysis as well as comparison to other models, the output of the standard Bioclim algorithm was altered to obtain ten prediction classes (minimum–maximum range, within 2.5–97.5; 5–95; 7.5–92.5; 10–90; 15–85; 20–80; 25–75; 30–70; and 35–65 percentile limits).

2) Domain: this method derives a point-to-point similarity metric to assign a classification value to a potential site based on its proximity in environmental space to the most similar occurrence. The Gower metric, which is the sum of the standardized distance between two points for each predictor variable, is used to quantify the similarity between two sites. The standardization is achieved using the predictor variable range at the presence sites to equalize the contribution from each predictor variable. Similarity is then calculated by subtracting the distance from 1. The maximum similarity between a candidate point and the set of known occurrences is assigned to each grid cell within the study area; these similarity values are degrees of classification confidence (Carpenter et al. 1993).

3) Genetic algorithm for rule-set prediction (GARP): the desktop version (<http://beta.lifemapper.org/desktopgarp>) of this artificial intelligence-based approach employs four distinct modeling methods: atomic, logistic regression, bioclimatic envelope, and negated bioclimatic envelope rules to derive several different rules (Stockwell and Peters 1999). GARP uses these rules to iteratively search for non-random correlations between the presence and background absence observations and the environmental predictors. GARP prepares the occurrence data by resampling the occurrence points in environmental space into 1250 presence and 1250 non-presence pixels randomly selected from the background. A GARP run begins by using 50% of these occurrence observations to train the model and then tests the resulting model with the remaining observations. It then resamples the observation points again, dividing the dataset into new training and test datasets and attempts to improve on the first model created. This process is repeated iteratively generating a set of "rules" that are altered in a genetic fashion until the best possible model is achieved or a set number of iterations are performed. The output for a GARP run is a binary map of predicted presence and absence. Since GARP's output is stochastic and can often produce different models for the same set of observations (Anderson et al. 2003) we ran GARP 500 times for each occurrence dataset. Using the "best subsets" feature, the 10 best models were selected based on internal GARP evaluation measures of omission and commission error rates. These models were merged to produce a final predicted distribution map having values between 0 and 10, 10 being cells that all 10 models predicted present.

4) Maximum entropy (Maxent): Maxent utilizes a statistical mechanics approach called maximum entropy to make predictions from incomplete information. Maxent estimates the most uniform distribution (maximum entropy) across the study area given the constraint that the expected value of each environmental predictor variable under this estimated distribution matches its empirical average (average values for the set occurrence data) (Phillips et al. 2004, 2006). Continuous environmental data can also be used to define quadratic features and product features (for this study only quadratic terms were considered), thereby adding further constraints to the estimated probability distribution by restricting the variance of each environmental predictor and covariance of each pair of environmental predictors to match the variance and covariance on the occurrence dataset.

Similar to logistic regression, Maxent weights each feature (environmental variable or its square, in this study) by a constant. The estimated probability distribution is exponential in the sum of the weighed features, divided by a scaling constant to ensure that the probability values range from 0–1 and sum to 1. The program starts with a uniform probability distribution and iteratively alters one weight at a time to maximize the likelihood of the occurrence dataset. The algorithm is guaranteed to converge to the optimum probability distribution and because the algorithm does not use randomness, the outputs are deterministic.

Given that the traditional implementation of maximum entropy is prone to over fitting, Maxent employs a relaxation. It constrains the estimated distribution so that the average value for a given predictor is close to the empirical average (within empirical error bounds) rather than equal to it. This smoothing procedure is called regularization and the user has the option to alter the parameters of this procedure to potentially compensate for small sample sizes. In this study we maintained a constant regularization parameter throughout.

Maxent's predictions for each analysis cell are "cumulative values", representing as a percentage, the probability value for the current analysis cell and all other cells with equal or lower probability values. The cell with a value of 100 is the most suitable, while cells close to 0 are the least suitable within the study area (Phillips et al. 2004).

Model evaluation

The models developed using the occurrence datasets of various sample sizes were queried spatially to determine their predictions at the 50 locations of the occurrences set aside for evaluation. The results were processed using the following three evaluation methods.

We selected the receiver operating characteristic (ROC) plot as one method to evaluate the predictive ability of the generated distribution models. A ROC plot is created by plotting the sensitivity values, the true-positive fraction against 1–specificity, the false-positive fraction for all available probability thresholds (Fielding and Bell 1997, Manel et al. 2001). A curve that maximizes sensitivity for low values of the false-positive fraction is considered a good model and is quantified by calculating the area under the curve (AUC). The AUC can be used as a measure of the model's overall performance and has values usually ranging from 0.5 (random) to 1.0 (perfect discrimination) but can have values below this range indicating a model that is worse than random (Engler et al. 2004). The ROC plot method has an advantage over confusion matrix-derived evaluation methods (for examples see below and Fielding and Bell 1997) because it does not require an arbitrary selection of a threshold above which prediction is considered positive, a procedure that can bias evaluations (Fielding and Bell 1997). While generally used when presence and absence data are available, ROC plots can also be generated with presence and background absence data (Phillips et al. 2006). To implement the ROC evaluation procedure in this study we randomly generated 50 background observations within the California boundary and entered them into the ROC curve procedure in place of absences to accompany the 50 presence observations. The AUC derived from the ROC plot of this study can be interpreted as a measure of the ability of the algorithm to discriminate between a suitable environmental condition and a random analysis pixel (background), rather than between suitable and unsuitable conditions, as an AUC developed with measured absences is interpreted (Phillips et al. 2006). The ROC curves were plotted and the AUC value calculated in SPSS (Anon. 2001).

Given that only presence occurrences are available, it is useful to use an evaluation method that does not require absence occurrences. We selected prediction success, which is the percentage of positive evaluation occurrences correctly classified as positive. This evaluation method requires a threshold to convert continuous model predictions to dichotomous classifications of presence/absence. A threshold should not be chosen arbitrarily but selected based on the objectives for generating the species distribution model (Wilson et al. 2005). As larger thresholds are selected, commission errors tend to decrease while omission errors will increase (Fielding and Bell 1997). If commission errors are considered to be more serious, then a larger threshold should be selected thereby minimizing this error at the expense of greater omission errors. When there is no inclination as to which type of error is most critical, then a threshold, or "cut-off" for each model can be obtained by identifying the point on the ROC curve where the sum of the sensitivity and specificity is maximized (Manel et al. 2001). This value was determined for the 100 sample size dataset models and averaged over the 10 replicate models to obtain the threshold for each taxon and modeling method. While this is a reasonable strategy it may put some modeling methods at a disadvantage. Models generated with small sample sizes could potentially be evaluated as poor models by the threshold dependent measures if the thresholds selected for the 100 sample models are inappropriate for models generated with fewer occurrences. We acknowledge this possible limitation, however selecting an appropriate threshold for models generated with small sample sizes is difficult using standard techniques because of the lack of power in the presence data (prevalence, Manel et al. 2001). Hence, we used the thresholds based on largest sample size (100) category models to reclassify each taxon's models generated with the various sample sizes to binary maps of predicted presence and absence. Predictive success was the percentage of the 50 evaluation points that were correctly classified by a given model/sample size combination. Prediction success gives an estimate of the number of true-positive predictions (measure of the omission error rate) but it does not give an estimate the other type of error, commission, i.e. the false-positive predictions. In an attempt to identify the magnitude of commission errors, the total spatial area predicted present was recorded. The area predicted present can be used as a surrogate for commission errors, assuming that models with larger prediction

areas are also more likely to have higher rates of commission errors.

As a final evaluation measure, the binary prediction maps were spatially compared to the full 150-occurrence dataset model for the same taxa and modeling method. Here we assume that the models based on the complete dataset (150 localities) are the most representative of the true distribution of the species given the limitations of the modeling method, and the species occurrence and environmental data available. The full 150-occurrence dataset models were reclassified into binary prediction maps using the same thresholds used to convert the sample size category model replicates for the prediction success evaluation. We then determined the spatial overlap between each of the 60 models (i.e. 10 per each of the six sample size categories) per species and the model created with all 150 occurrences for a given modeling method. A confusion matrix (Fig. 1) was populated for each model with the spatial area where the model was in accordance or discordance with its full 150-sample size model for the two possible classifications of state of occurrence. We used kappa (Fielding and Bell 1997) to summarize the overall agreement between the spatial predictions of each replicate to its full 150-sample size model. A value of one would indicate a perfect agreement between the two predictive maps. This evaluation was not a measure of a modeling method's overall prediction accuracy potential, but did provide an assessment of the methods overall stability in its predictions when presented with incomplete species occurrence data sets.

Species level characteristics

To obtain an approximation of each taxon's distributional spatial extent and ecological characteristics within California we used the full set of occurrence points available in CNDDB for each taxon. For the estimated distributional spatial extent we used a kernal density estimator (Elith and Burgman 2003, Fortin et al. 2005), to generate polygons encompassing probability density functions of 0.75 around the occurrence points and then calculated the polygon's total area to characterize a taxon's distributional spatial extent. We calculated two measures of the environmental niche: marginality and tolerance, using ecological niche factor analyses (ENFA in BIOMAPPER; Hirzel et al. 2002). Marginality is the difference between the species optimum and the mean environmental conditions in the study area and is therefore representative of the species' ecological niche position. Tolerance describes the species' niche breadth by comparing the variability in the environmental conditions where the species occurs to the range of environmental conditions in the study area. We conducted Spearman's Rank non-parametric correlations to evaluate the relationship between model performance and ecological characteristics.

		Full Model: +	Full Model: -
Replicate Model	+	a	b
	-	c	d

Fig. 1. Confusion matrix used to spatially compare the replicate model to the full 150-occurrence dataset model for the same taxon and modeling method.

Results

The distribution of AUC values for the four modeling methods for each of the six sample size categories are represented in the box plots of Fig. 2. AUC was generally highest for the 100-sample size models for all methods except GARP where models generated with a sample size of 50 performed very well. AUC generally scores increased with sample size for all modeling methods. The AUC values reached near maximal levels at 75 observations for Bioclim; 50 for both Domain and Maxent; 10 for GARP whereby they remained relatively consistent for that sample size category and all categories above it. In some instances the AUC scores decreased at larger sample sizes but for the most part these changes were insignificant. Overall, Domain and Maxent achieved the highest AUC values followed by GARP and then Bioclim across most sample sizes. The 18 taxa modeled displayed considerable variation in model performance as evaluated by AUC. This result was mirrored in the other two evaluation methods, which similarly had a large spread of evaluation measure values.

Similar box plots for prediction success using thresholds identified using the ROC curve are presented in Fig. 3. Again prediction success increased with sample size though it did not reach a maximum achievable value except possibly for GARP and Maxent with data sets of 75 samples. At samples sizes between 5 and 25, the method performance from highest to lowest was: Maxent, GARP, Bioclim, and Domain. This order changed for sample size categories between 50 and 100 to be GARP, Maxent, Domain, and Bioclim, though differences between the performances of the four modeling methods were more marked at the smaller sample sizes. The average range in values for prediction success between models built with 5 and 100 occurrences was smallest for Maxent (sample size 5 mean: 61.2; sample size 100 mean: 90.9), followed by GARP (sample size 5 mean: 32.1; sample size 100 mean: 95.6), Bioclim (sample size 5 mean: 11.6; sample size 100 mean: 85.4)

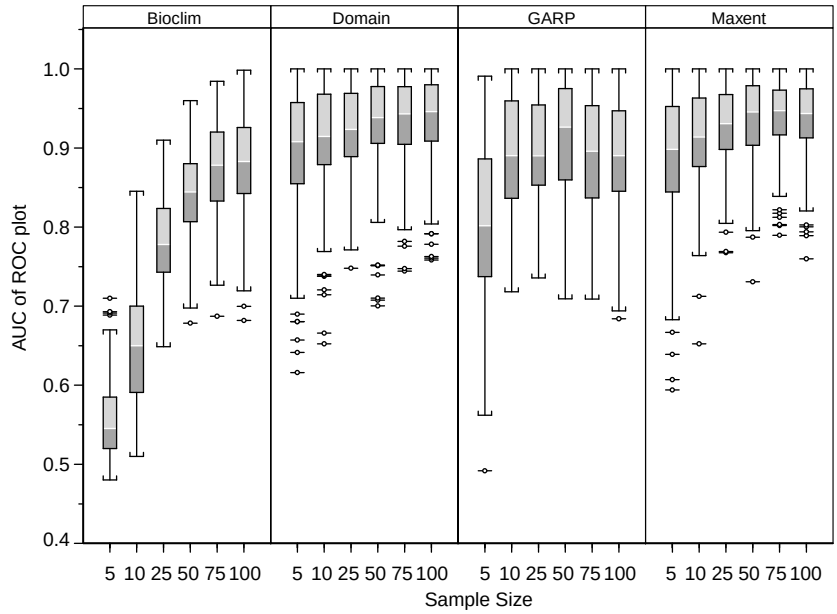

Fig. 2. Box plot displaying the interquantile range and outliers around the median AUC values of ROC plots for each modeling method by sample size category.

and then Domain (sample size 5 mean: 8.5; sample size 100 mean: 92.3).

These values should be considered in conjunction with the area predicted present as a percentage of the total study area (Fig. 4). Given that we are evaluating the models in a presence-only framework we should be concerned with minimizing the potential of gross commission errors. Therefore when absence data is not available to quantitatively evaluate commission errors a good model could be considered one that achieves low omission errors (i.e. high prediction success) while still generating the most parsimonious model with regard to the total area (i.e. number of 1 km^2 cells) predicted positive. In general, Domain and Bioclim predicted the smallest area, followed by Maxent and then GARP, though this relationship varied with sample size. The area predicted as suitable for a given species by Maxent remained fairly level at sample sizes of 25 and above while the other methods predicted more area with increasing sample sizes. The most noticeable difference among methods across samples sizes was the placement of Maxent, which predicted the largest area of all modeling methods at sample size 5 (mean: 8.0) and the smallest area using 100 occurrences (mean: 13.2) and

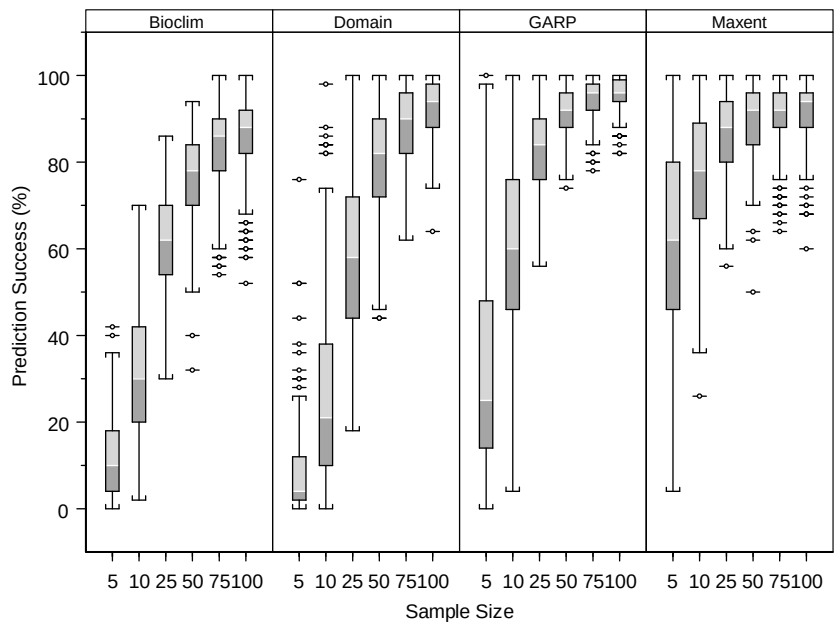

Fig. 3. Box plot displaying the interquantile range and outliers around the median prediction success values for each modeling method by sample size category.

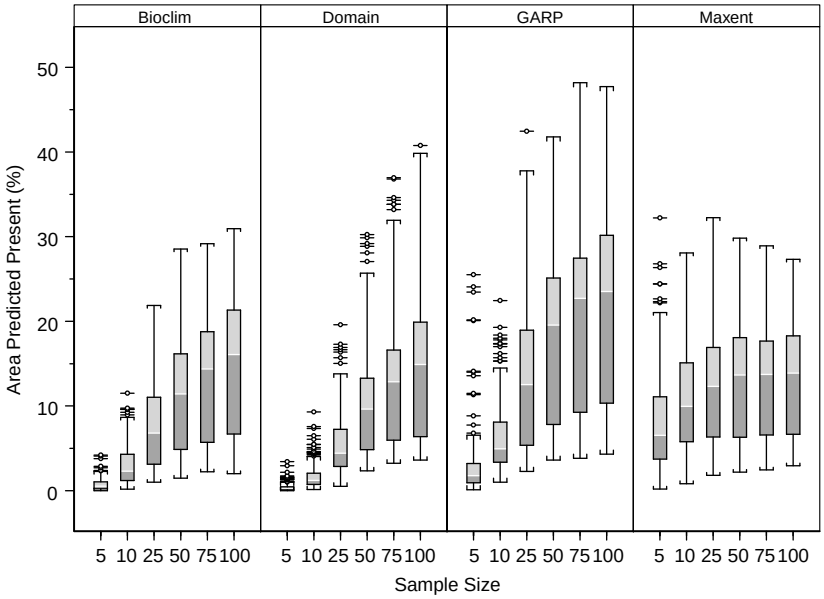

Fig. 4. Box plot displaying the interquantile range and outliers around the median percent area predicted present for each modeling method by sample size category.

therefore had the smallest range of average area predicted of all modeling methods. Bioclim followed with the next smallest range of average area predicted (sample size 5 mean: 0.8; sample size 100 mean: 14.7), followed by Domain (sample size 5 mean: 0.4; sample size 100 mean: 14.7), and the range was largest for GARP (sample size 5 mean: 3.0; sample size 100 mean: 21.6). At the 100 sample size category GARP predicted on average 64 percent more area than Maxent at the same sample size category.

The spatial comparisons of each replicate to its full 150-sample size model summarized using kappa, are displayed in the box plots of Fig. 5. The kappa coefficient increased with larger sample size categories for all modeling methods and an asymptote was not reached for any method, suggesting that maximal concordance was not achieved and will likely continue to increase with sample size until the full data set is used. The order of the four modeling methods level of concordance as evaluated by the kappa coefficient again differed by sample size category, but Maxent achieved the highest values for every size category.

Correlations among the three ecological characteristics, marginality, tolerance and distributional spatial extent were as follows: extent and tolerance were positively related (Spearman's $R = 0.74$, $p < 0.05$); extent

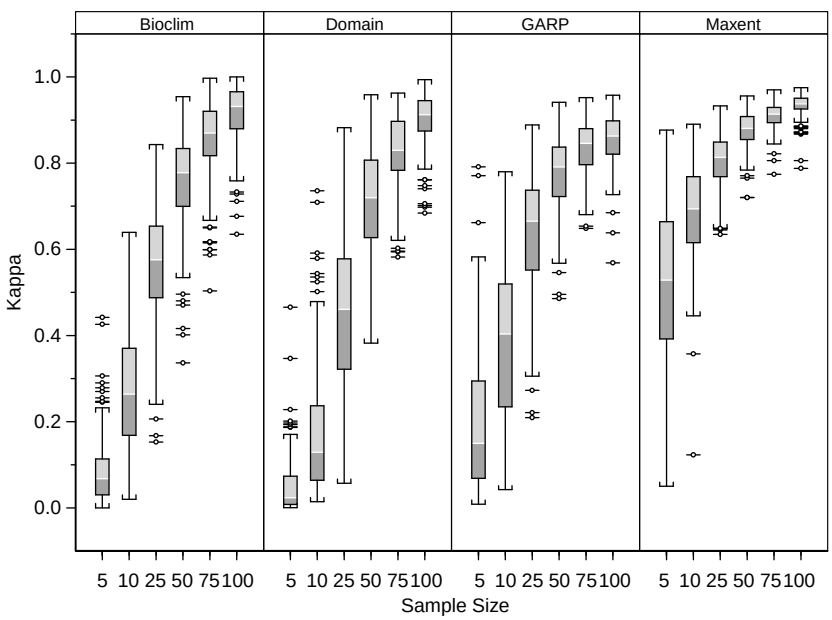

Fig. 5. Box plot displaying the interquantile range and outliers around the median Kappa for each modeling method by sample size category when considering the full 150 model as the observed.

Fig. 6. Mean AUC values for models built with a sample size of 100 occurrences for each taxon and modeling method (n = 10). Taxa are sorted in ascending order by (a) marginality (b) tolerance and (c) estimated spatial extent in California.

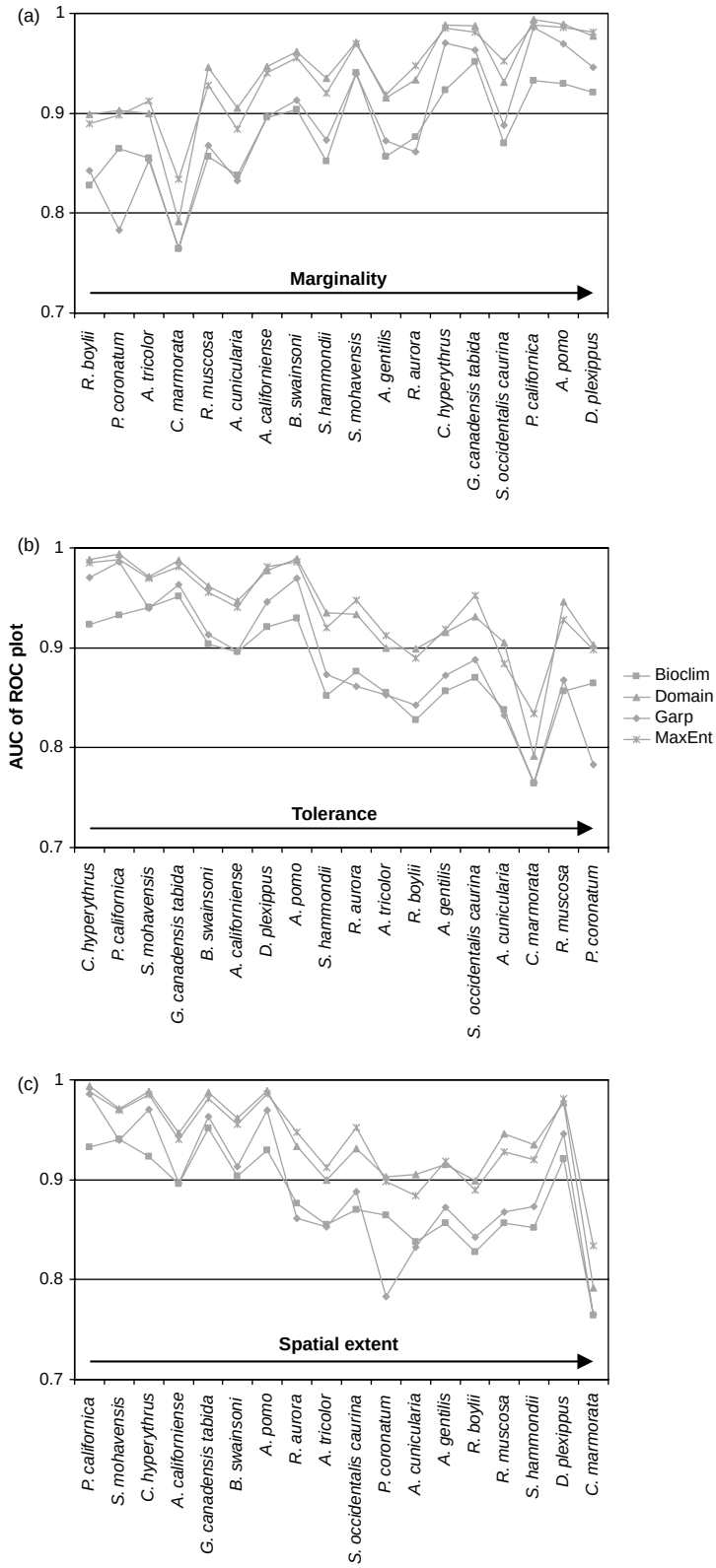

Table 3. Spearman's R correlations between model performance as measured by AUC and species characteristics for models built with 3 different sample sizes; 5, 50 and 100 samples.

	Bioclim			Domain			GARP			Maxent		
Sample size	5	50	100	5	50	100	5	50	100	5	50	100
Extent	−0.72	−0.69	−0.76	−0.68	−0.65	−0.62	−0.60	−0.59	−0.64	−0.77	−0.57	−0.63
Tolerance	−0.55	−0.84	−0.78	−0.86	−0.81	−0.80	−0.84	−0.75	−0.86	−0.82	0.74	−0.79
Marginality	0.14*	0.67	0.72	0.62	0.77	0.78	0.67	0.76	0.81	0.76	0.86	0.85

*Indicates not significant; all other values are significant at a level of $p > 0.05$.

and marginality were negatively correlated, though not significantly (Spearman's $R = -0.34$); and tolerance and marginality were negatively correlated (Spearman's $R = 0.57$, $p < 0.05$). Graphical representation of the relationship between the taxa's marginality, tolerance and distributional spatial extent, and AUC for models generated with sample sizes of 100 occurrences are presented in Fig. 6. These relationships were consistent across a range of sample sizes (see Table 3 for sample sizes 5, 50, and 100), and therefore only data for the largest sample size are displayed graphically and a subset in Table 3. Marginality was positively related to AUC (Fig. 6a, Table 3). Generally as tolerance (breadth of environmental space used) increased, model predictive accuracy decreased (Fig. 6b, Table 3). Likewise, as taxa spatial distributional extent increased, model prediction accuracy decreased (Fig. 6c, Table 3), though Spearman's R was generally higher for tolerance than for spatial distributional extent. Similar results were obtained for the relationship between the three ecological characteristics and predicted success (results not shown), although the response was not as strong as for AUC.

Discussion

Model accuracy increased with larger sample sizes for all modeling methods across the 18 California taxa tested. Nonetheless, useful models were produced with as few as 5–10 positive observations, and models produced with 50 observations were similar to those created with twice as many locations. This result, along with the indication that ecologically specialized species are easier to model than wide ranging species, is especially encouraging for modeling rare species. Also, given that distribution modeling can be used for a variety of different objectives ranging from guiding future exploration of a species range to creating an accurate model for conservation planning, models built with few points, while not as accurate as those built with large datasets and potentially not appropriate for all applications, are still useful. Our results increase the relevance of data housed in museum or herbarium collections, or similar databases such as those maintained by NatureServe and its network of natural heritage programs (Stein et al. 2000, Graham et al. 2004). As such occurrence databases become more widely available, thereby making species distribution modeling more accessible to conservation planners, research such as that presented in this paper is imperative to guide modelers.

Maxent had the strongest performance of the methods tested here because it performed well and remained fairly stable in both prediction accuracy and the total area predicted present across all sample size categories. Further, it often had the highest accuracy and spatial concordance, especially for the two smallest sample size categories. These results indicate that Maxent can somewhat compensate for incomplete, small species occurrence data sets and perform near maximal accuracy level in these conditions. The success of Maxent is likely due to its regularization procedure that counteracts a tendency to over-fit models when using few species occurrences (Phillips et al. 2006). Our results support those obtained by Elith et al. (2006) who also found that Maxent was one of the strongest performers in a large model comparison study.

In contrast, Bioclim does not appear to be capable of maximizing its accuracy potential with small sample sizes and did not perform as well as the other modeling methods using the datasets of larger samples. For these reasons we would not suggest its use when modeling with small numbers of species observations. It is interesting to remark that Bioclim did attain relatively high concordance at the larger sample size categories when spatially compared to models generated with all 150-occurrences. This result is not surprising given that the Bioclim algorithm does not extrapolate beyond the bounds of the environmental conditions at known locations of occurrence. As additional observations are included in the development of the Bioclim model, the envelope defining the environmental conditions at known occurrences will by default expand from defining a small portion of the species' full environmental envelope (here developed with all 150-occurrences) towards defining a larger portion of that full envelope.

In some respects Domain and GARP performed fairly similarly, achieving relatively high prediction accuracy values at large sample size categories with low prediction success and spatial concordance at small sample sizes. However, the two evaluation measures of prediction accuracy (AUC and prediction success) revealed conflicting assessments of GARP's performance. The AUC

evaluation indicated that GARP models reached near maximum accuracy (ca 10% lower AUC than the full model AUC) when using sample sizes of 10 observations, a result that is supported by Stockwell and Peterson (2002), but when evaluating the models using prediction success, the maximum accuracy was reached at the 75-sample size category. The inconsistent assessment of model prediction accuracy supports the necessity to evaluate models with multiple evaluation metrics. Specifically, metrics such as prediction success should be reviewed with caution when not accompanied with total suitable area predicted. GARP had the largest prediction success values for the larger sample size categories but predicted a spatial area that far exceeds the other three modeling methods, thereby increasing its chances of correctly classifying positive occurrences but as a result most likely increasing its commission error rate as well. Manel et al. (2001) also cautioned against basing evaluations of presence-absence model performance solely on prediction success.

All things considered we interpret our results to indicate that overall Domain performed better than GARP because Domain had higher AUC values for all sample size categories. The AUC of a ROC plot generated with presence and background data evaluates a model based on its prediction success but also penalizes it for predicting proportionately larger spatial areas (Phillips et al. 2006), thereby evaluating both omission and commission errors simultaneously. Hence, the low AUC values obtained by GARP at low sample size categories likely reflect commission error, while the Domain models appear to have less commission error. Further, the threshold selection strategy most likely artificially decreased Domain's prediction potential for the smaller sample size categories as assessed by both the prediction success and spatial comparison evaluations. Since Domain derives a point-to-point distance for each pixel based on its proximity in environmental space to the most similar occurrence, it follows that when models are built with fewer occurrences this distance will be greater.

We confirm the results of other researchers that the ecological characteristics of model species affect model accuracy potential, where species widespread in both geographic and environmental space are generally more difficult to model than species with compact spatial distributions (Araujo and Williams 2000, Stockwell and Peterson 2002, Thuiller et al. 2003, Segurado and Araujo 2004). In all but one case (Table 3), significant relationships existed between model performance (AUC) and spatial extent of a species distribution, tolerance, and marginality. These relationships were consistent across all sample size treatments indicating that the ability to model species effectively is strongly influenced by species ecological characteristics independent of sample size. Tolerance generally had the highest correlation with AUC, indicating that environmental space occupied by a given species might be a better measure than geographic space occupied, although the kernel density estimator used to estimate the species' spatial ranges likely overestimated the area occupied for some species creating artificial outliers. In particular the range size estimator undoubtedly overestimated the spatial range of the monarch butterfly (*Danaus plexippus*, range size estimated as 105 874 km^2, tolerance 1.64) a mostly coastal and patchily distributed wintering butterfly species.

In this study the species with the smallest geographic extent of occurrence and very low tolerance (small niche breadth), the California gnatcatcher *Polioptila californica* generally had the highest AUC and prediction success values, whereas the opposite was found for the western pond turtle *Clemmys marmorata*, which has the widest geographic range of the study species within California and a very high ecological tolerance. Our results provide support for the explanation offered by Stockwell and Peterson (2002) that local ecological adaptation by sub-populations is more likely to occur for widely distributed species resulting in different habitat preferences in discrete parts of the species' range. In climatic modeling each sub-population would have a distinct climatic range in which it occurs and therefore when the species is modeled as a whole over its entire geographic range, the total climatic range encompasses climatic conditions not suitable for occupancy, thereby overestimating the species' ecological climatic breadth. The fact that the models for the two species in this study that have considerable taxonomic confusion regarding their Californian distributions, the western pond turtle and the coast horned lizard *Phrynosoma coronatum*, performed poorly provides support to notion that local ecological adaptation results in a decrease in model accuracy. It would be interesting to partition the data for these two species based on the geographic boundaries of the proposed subspecies or genetic lineages to determine whether the resulting models do indeed result in an increase in model accuracy.

Other possible explanations for variation in model performance not related to geographic range size or ecological niche breadth are that some species are just not suited for climatic modeling and/or the spatial grain (pixel resolution) was inappropriate for modeling some taxa's distribution in the geographic study area of California. Models for the red-legged frog *Rana aurora* would likely have benefited from the inclusion of a variable describing the distribution pattern of introduced bullfrogs and the western pond turtle models may have been improved with a description of the amount of wetlands present within an area surrounding an occurrence. These are examples of cases where the climatic variables may be insufficient to model the species' distribution and where important variables that either positively or negatively contributed to the observed

spatial distribution pattern are missing from model formulation, thereby resulting in relatively poor predictive distribution models.

Of course, as with any comparative modeling method exercise, these results may differ in a new study area, at a different spatial scale (extent and/or grain), with varying qualities of model data (species and environmental), and for study species of different ecological characteristics. Our results clearly indicate that future studies should use multiple evaluation measures, because each measure provides only a portion of the elusive "truth" of the predictive ability of a species distribution model. Further, while we found that reasonable models could be generated with low sample sizes, we sub-sampled from a larger set of data and presumably obtained a relatively representative, albeit small, number of points. However, if decreasing sample size increases bias, which may be the case with data collected in an ad-hoc fashion, then models built with small samples may be quite poor. The fact that we could develop models with small samples for some species does not mean this will be possible for all species.

In general, practitioners should remember that models are simply an estimate of a species' potential distribution. Species distribution modeling cannot replace fieldwork intended to collect more distributional data but can be a useful tool for data exploration to help identify potential knowledge gaps and provide direction to fieldwork design (Engler et al. 2004). By carefully evaluating models and including both species characteristics and sample size in our analyses our results indicate considerable promise for modeling rare species. This result should encourage conservation practitioners to explore the use of distribution modeling across a variety of applications.

Acknowledgements – We thank the Seaver Inst. for funding this research and the California Natural Diversity Database for making their species occurrence data available to this project. We are also most appreciative of Steven Phillips for his insightful comments and corrections to this manuscript.

References

Anderson, R. P. and 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: 167–179

Anderson, R. P. et al. 2003. Evaluating predictive models of species' distributions: criteria for selecting optimal models. – Ecol. Modell. 162: 211–232.

Anon. 2001. SPSS for Windows. – SPSS, Chicago.

Araujo, M. B. and Williams, P. H. 2000. Selecting areas for species persistence using occurrence data. – Biol. Conserv. 96: 331–345.

Boone, R. B. and Krohn, W. B. 2002. Modeling tools and accuracy assessment. – In: Scott, J. M. et al. (eds), Predicting species occurrences: issues of accuracy and scale. Inland Press, pp. 265–270.

Brotons, L. et al. 2004. Presence–absence versus presence–only modeling methods for predicting bird habitat suitability. – Ecography 27: 437–448.

Busby, J. R. 1991. BIOCLIM – a bioclimate analysis and prediction system. – In: Margules, C. R. and Austin, M. P. (eds), Nature conservation: cost effective biological surveys and data analysis. CSIRO, pp. 64–68.

Carpenter, G. et al. 1993. DOMAIN: a flexible modeling procedure for mapping potential distributions of plants and animals. – Biodiv. Conserv. 2: 667–680.

Corsi, F. et al. 2000. Modelling species distribution with GIS. – In: Boitani, L. and Fuller, T. K. (eds), Research techniques in animal ecology; controversies and consequences. Columbia Univ. Press, pp. 389–434.

Elith, J. and Burgman, M. A. 2003. Habitat models for PVA. – In: Brigham, C. A. and Schwartz, M. W. (eds), Population viability in plants. Conservation, management and modeling of rare plants. Springer, pp. 203–235.

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

Engler, R. et al. 2004. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. – J. Appl. Ecol. 41: 263–274.

Farber, O. and Kadmon, R. 2003. Assessment of alternative approaches for bioclimatic modeling with special emphasis on the Mahalanobis distance. – Ecol. Modell. 160: 115–130.

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

Fortin, M.-J. et al. 2005. Species ranges and distributional limits: pattern analysis and statistical issues. – Oikos 108: 7–17.

Gaston, K. J. 1997. What is rarity? – In: Kunin, W. E. and Gaston, K. J. (eds), The biology of rarity. Causes and consequences of rare-common differences. Chapman and Hall, pp. 30–47.

Gillison, A. N. 1997. Mapping the potential distribution of plants and animals for wildlife management: the use of the DOMAIN software package. – In: Romimoharto, K. et al. (eds), Proceedings of the national seminar on the role of wildlife conservation and its ecosystem in national development. The Indonesian Wildlife Fund (IWF), Jakarta, pp. 114–119+two maps.

Graham, C. H. et al. 2004. New developments in museum-based informatics and application in biodiversity analysis. – Trends Ecol. Evol. 19: 497–503.

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

Hirzel, A. H. et al. 2002. Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? – Ecology 83: 2027–2036.

Johnson, C. M. et al. 2002. Predicting the occurrence of amphibians: an assessment of multiple-scale models. – In: Scott, J. M. et al. (eds), Predicting species occurrences: issues of accuracy and scale. Inland Press, pp. 157–170.

Kadmon, R. et al. 2003. A systematic analysis of factors affecting the performance of climatic envelope models. – Ecol. Appl. 13: 853–867.

Kattan, G. H. 1992. Rarity and vulnerability: the birds of the cordillera central Columbia. – Conserv. Biol. 6: 64–70.

Lindenmayer, D. B. et al. 1991. The conservation of Leadbeater's possum, *Gymnobelideus leadbeateri* (McCoy): a case study of the use of bioclimatic modeling. – J. Biogeogr. 18: 371–383.

Loiselle, B. A. et al. 2003. Avoiding pitfalls of using species distribution models in conservation planning. – Conserv. Biol. 17: 1591–1600.

Luoto, M. et al. 2005. Uncertainty of bioclimate envelope models based on the geographical distribution of species. – Global Ecol. Biogeogr. 14: 575–584.

Manel, S. et al. 2001. Evaluating presence-absence models in ecology: the need to account for prevalence. – J. Appl. Ecol. 38: 921–931.

McPherson, J. M. et al. 2004. The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artifact? – J. Appl. Ecol. 41: 811–823.
Nix, H. 1986. A biogeographic analysis of Australian elapid snakes. – In: Longmore, R. (ed.), Atlas of elapid snakes of Australia. Bureau of Flora and Fauna, Canberra, pp. 4–15.
Pearce, J. and Ferrier, S. 2000. Evaluating the predictive performance of habitat models developed using logistic regression. – Ecol. Modell. 133: 225–245
Phillips, S. J. et al. 2004. A maximum entropy approach to species distribution modeling. – In: Brodley, C. E. (ed.), Machine learning. Proc. of the Twenty-first Century International Conference on Machine Learning, Banff, Canada, 2004. ACM Press, p. 83.
Phillips, S. J. et al. 2006. Maximum entropy modeling of species geographic distributions. – Ecol. Modell. 190: 231–259.
Rabinowitz, D. et al. 1986. Seven forms of rarity and their frequency in the flora of the British Isles. – In: Soulé, M. E. (ed.), Conservation biology: the science of scarcity and diversity. Sinauer, pp. 182–204.
Segurado, P. and Araujo, M. B. 2004. An evaluation of methods for modelling species distributions. – J. Biogeogr. 31: 1555–1568.
Stein, B. A. et al. 2000. Precious heritage. – Oxford Univ. Press.
Stockwell, D. and Peters, D. 1999. The GARP modeling system problems and solutions to automated spatial prediction. – Int. J. Geogr. Inform. Sci. 13: 143–158.
Stockwell, D. R. B. and Peterson, A. T. 2002. Effects of sample size on accuracy of species distribution models. – Ecol. Modell. 148: 1–13.
Thuiller, W. et al. 2003. Generalized models vs. classification tree analysis: predicting spatial distributions of plant species at different scales. – J. Veg. Sci. 14: 669–680.
Thuiller, W. et al. 2004. Relating plant traits and species distributions along bioclimatic gradients for 88 *Leucadendron* species in the Cape Floristic Region. – Ecology 85: 1688–1699.
Wilson, K. A. et al. 2005. Sensitivity of conservation planning to different approaches to using predicted species distribution data. – Biol. Conserv. 122: 99–112.

Subject Editor: Miguel Araújo.