Diversity and Distributions, (Diversity Distrib.) (2008) **14**, 763–773

BIODIVERSITY RESEARCH

Effects of sample size on the performance of species distribution models

M. S. Wisz[1]*, R. J. Hijmans[2], J. Li[3], A. T. Peterson[4], C. H. Graham[5], A. Guisan[6] and NCEAS Predicting Species Distributions Working Group†

[1]*Department of Arctic Environment, National Environmental Research Institute, University of Aarhus, Frederikborgvej 399, Roskilde, Denmark,* [2]*International Rice Research Institute, Los Baños, Laguna, Philippines,* [3]*Department of Marine and Coastal Environment, Geoscience, Canberra, ACT, Australia,* [4]*University of Kansas Natural History Museum and Biodiversity Research Center, Lawrence, KS, USA,* [5]*Department of Ecology and Evolution, Stony Brook University, NY 11794, USA,* [6]*Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland*

*Correspondence: M. S. Wisz, Department of Arctic Environment, National Environmental Research Institute, University of Aarhus, Frederikborgvej 399, Roskilde, Denmark. E-mail: msw@dmu.dk
†NCEAS Predicting Species Distributions Working Group: J. Elith School of Botany, University of Melbourne, Parkville, Victoria 3010, Australia; R. P.; M. Dudík, Princeton University, Princeton, NJ, USA; S. Ferrier, Department of Environmental and Climate Change, Armidale, NSW, Australia; F. Huettmann, University of Alaska Fairbanks, AK, USA; J. R. Leathwick, NIWA, Hamilton, New Zealand; A. Lehmann, Swiss Centre for Faunal Cartography (CSCF), Neuchâtel, Switzerland; L. Lohmann, Universidade de São Paulo, Brazil; B. A. Loiselle, University of Missouri, St. Louis, USA; G. Manion, Department of Environmental and Climate Change, Armidale, NSW, Australia; C. Moritz, The University of California, Berkeley, USA; M. Nakamura, Centro de Investigación en Matematicas (CIMAT), Mexico; Y. Nakazawa, University of Kansas, Lawrence, KS, USA; J. McC. Overton, Landcare Research, Hamilton, New Zealand; S. J. Phillips, AT&T Labs-Research, Florham Park, NJ, USA; K. S. Richardson, McGill University, QC, Canada; R. Scachetti-Pereira, Centro de Referência em Informacão Ambiental, Brazil; R. E. Schapire, Princeton University, Princeton, NJ, USA; J. Soberón, University of Kansas, Lawrence, KS, USA; S. E. Williams, James Cook University, Queensland, Australia; N. E. Zimmermann, Swiss Federal Research Institute WSL, Birmensdorf, Switzerland.

ABSTRACT

A wide range of modelling algorithms is used by ecologists, conservation practitioners, and others to predict species ranges from point locality data. Unfortunately, the amount of data available is limited for many taxa and regions, making it essential to quantify the sensitivity of these algorithms to sample size. This is the first study to address this need by rigorously evaluating a broad suite of algorithms with independent presence–absence data from multiple species and regions. We evaluated predictions from 12 algorithms for 46 species (from six different regions of the world) at three sample sizes (100, 30, and 10 records). We used data from natural history collections to run the models, and evaluated the quality of model predictions with area under the receiver operating characteristic curve (AUC). With decreasing sample size, model accuracy decreased and variability increased across species and between models. Novel modelling methods that incorporate both interactions between predictor variables and complex response shapes (i.e. GBM, MARS-INT, BRUTO) performed better than most methods at large sample sizes but not at the smallest sample sizes. Other algorithms were much less sensitive to sample size, including an algorithm based on maximum entropy (MAXENT) that had among the best predictive power across all sample sizes. Relative to other algorithms, a distance metric algorithm (DOMAIN) and a genetic algorithm (OM-GARP) had intermediate performance at the largest sample size and among the best performance at the lowest sample size. No algorithm predicted consistently well with small sample size ($n < 30$) and this should encourage highly conservative use of predictions based on small sample size and restrict their use to exploratory modelling.

Keywords

Ecological niche model, MAXENT, model comparison, OM-GARP, sample size, species distribution model.

INTRODUCTION

Accurate descriptions of species ecological and geographical distributions are fundamental to understanding patterns of biodiversity and the processes that shape them (Ferrier *et al.*, 2002; Rushton *et al.*, 2004). Species distribution models use quantitative methods to infer species environmental requirements from conditions at known occurrences and are increasingly used to predict distributions. A wide range of methods has been used, including factor analysis (Hirzel *et al.*, 2002), distance metrics (Carpenter *et al.*, 1993) bounding boxes (Busby, 1991), logistic regression (Buckland *et al.*, 1996), artificial neural networks (Manel *et al.*, 1999), genetic algorithms (Stockwell & Peters, 1999), and Bayesian approaches (Hepinstall & Sader, 1997). Different methods can produce rather different predictions (Ladle *et al.*, 2004; Elith *et al.*, 2006; Pearson *et al.*, 2006).

Despite the frequent use of distribution models, the number of occurrence records available for individual species from which to generate predictions is often quite limited. In addition to the various types of species rarity that might limit the availability of

DOI: 10.1111/j.1472-4642.2008.00482.x
www.blackwellpublishing.com/ddi

locality records (Kunin & Gaston, 1993), this paucity exists because some species are difficult to sample, or because available data are not yet available electronically (Graham *et al.*, 2004). Data paucity is of concern because model quality is clearly influenced by the number of records used in model building (Carroll & Pearson, 1998; Cumming, 2000; Pearce & Ferrier, 2000; Stockwell & Peterson, 2002; Kadmon *et al.*, 2003; Hernandez *et al.*, 2006). Few records may suffice to characterize distributions of species with narrow environmental tolerances, especially compared to those with broader tolerances (Kadmon *et al.*, 2003). However, in general, predictions based on few records are unlikely to be as good as those based on a large number of samples (Pearce & Ferrier, 2000; Kadmon *et al.*, 2003; Hernandez *et al.*, 2006).

There is a series of reasons why model performance generally decreases with sample size. First, levels of uncertainty associated with parameter estimates (e.g. means, modes, medians, predicted probabilities of occurrence) decrease with increasing sample size (Crawley, 2002). When sample sizes are small, outliers carry more weight in analyses than if more data were available to buffer their effects. Furthermore, given the highly dimensional, complex nature of ecological niches of species (Hutchinson, 1957), large numbers of samples may be needed to allow for accurate description of the range of conditions over which a species occurs. Moreover, empirical studies have shown that species responses to environmental gradients can be skewed or multimodal (Austin, 2002). Finally, interactions among environmental variables are often important in describing species–environment relationships and the number of parameters to be estimated for interactive effects increases exponentially with number of predictor variables (Rushton *et al.*, 2004). As a consequence, larger amounts of data might be needed to describe complex relationships and interactions, however, algorithms that perform well with a large sample sizes will not necessarily perform well with fewer samples. This necessitates the investigation of possible trade-offs between sample size and model complexity.

Previous studies have evaluated sample size effects on distributional models for only a few algorithms each and most did not test with data collected independently of the training data. For example, Stockwell & Peterson (2002) explored the effects of sample size on the performance of a genetic algorithm and a logistic regression method on North American bird predictions. Cumming (2000) evaluated the effects of sample size in the predictive performance of logistic regression models built for African ticks, and Kadmon *et al.* (2003) examined this for BIOCLIM models built for woody plant species from Israel. Hernandez *et al.* (2006) evaluated sample size relationships to model performance in BIOCLIM, DOMAIN, DK-GARP, and MAXENT on 17 species of vertebrates and one species of insect from California. In the most comprehensive study to date comparing species distribution model performance, Elith *et al.* (2006) demonstrated that several novel modelling methods yielded particularly good predictions. As part of the same cooperative effort, we have evaluated effects of experimentally manipulated/controlled sample sizes on predictive accuracy of 12 modelling algorithms in five geographical regions on four continents, using independent evaluation data and threshold-independent statistics. Models were based on occurrence records from data associated with natural history collections, a rapidly growing data source (Graham *et al.*, 2004). Natural history collection data typically reflect opportunistic sampling, which may introduce sampling biases (Hijmans *et al.*, 2000), for example, if sampling is concentrated only near roads, high elevation or particularly wet areas may be underrepresented. Such biases can exacerbate statistical problems when models are based on small sample sizes.

We anticipated a continuum of possible responses of modelling algorithms to sample size manipulations, but in general, models could be relatively insensitive (cases A, D, and E Fig. 1) or sensitive (cases B and C) to sample size, perhaps with a threshold at which model quality breaks down (case B). The quality of model predictions may be independent of sample size sensitivity, but models that are sensitive would produce poorer predictions at smaller sample sizes.

METHODS

Experimental framework

This study forms part of a larger project examining performance of species distribution models using locality data from natural history collections (Graham *et al.*, 2004; Elith *et al.*, 2006; Guisan *et al.*, 2007). We generated predictions from 12 modelling algorithms for 46 species at high (100 records), medium (30 records), and low (10 records) sample sizes to allow assessment of sample size effects on algorithm performance. We trained the models on

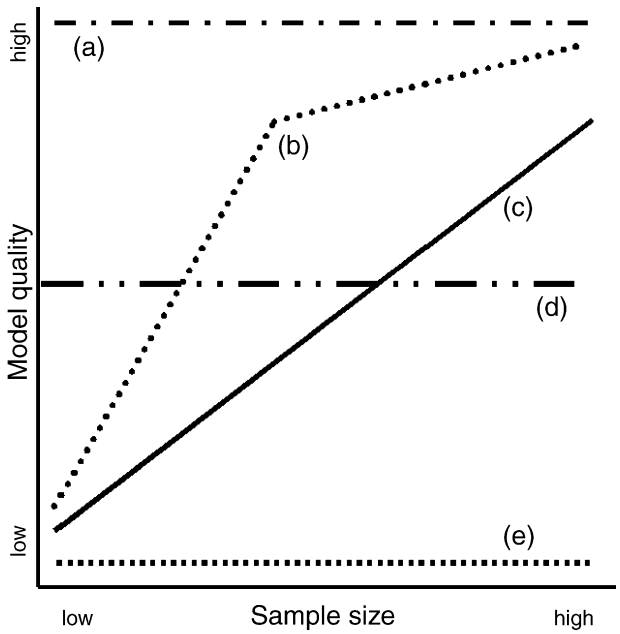

Figure 1 Five potential responses of modelling algorithms to sample size manipulations. (a) High-quality models, no sensitivity to sample size; (b) sensitivity to particularly small sample sizes; (c) sensitivity to sample size over whole range; and (d) intermediate-quality models, no sensitivity to sample size; (e) low-quality models, no sensitivity to sample size.

presence-only (PO) records from natural history collections and non-systematic sampling. We subsequently evaluated the predictions with independent testing data from strategically collected presence and absence (PA) data.

Occurrence data

Presence only data

In each of five regions, we chose 10 species that had > 100 unique occurrence localities (for one region, data available for only six species met this sample size requirement). The resulting data set consisted of 10 species each and included breeding birds of southern Ontario, Canada (CAN) based on nest records from the Royal Ontario Museum; small vertebrates of New South Wales, Australia (NSW) based on incidental records reported in the Atlas of NSW Wildlife; trees of New Zealand (NZ) based on herbarium data; and alpine plants of Switzerland (SWI) based on non-systematic surveys of forest vegetation. We also included six species of plants from the Amazon basin, South America (SA) based on herbarium data. Species were selected in consultation with regional experts to span a range of geographical distributional sizes and (where possible) life-history characteristics. Full details on the data used are provided in Elith *et al.* (2006).

Evaluation data

The independent presence–absence data were from planned surveys, with accurate location records. Full details of the evaluation data are provided in Elith *et al.* (2006), but briefly, predictions from CAN were evaluated with data from the Breeding Bird Atlas for Ontario; predictions from NSW were evaluated with data from designed surveys (Ferrier & Watson, 1996), as were those for NZ trees; evaluation of SA predictions were made with Al Gentry's transect data at 152 sites (source: Missouri Botanic Gardens), and SWI predictions were evaluated with data from forest inventory plots surveyed on a regular lattice.

Range overlap

The occurrence data for each species were subsampled randomly to provide data sets of 10, 30, and 100 points, with smaller samples selected at random from the full data set (i.e. not nested within larger samples). To quantify the amount of information about the species preserved in each random sample, we computed an environmental 'range overlap'. We computed 'range overlap' as the amount of the environmental space represented in the sample used to train a given model relative to the environmental space represented in the relatively complete evaluation data. Specifically, we calculated this as the ratio of (1) the arithmetic mean of the fractional overlap of the range of each environmental variable in each sample of training data (i.e. 10, 30, or 100 records) with (2) the arithmetic mean of the fractional overlap of these environmental variables in the presence data of the complete evaluation (testing) data set. Range overlap values thus ranged from 0 to 1 with smaller values indicating that less of the species environmental space is represented in the given sample. We compared arcsine-transformed range overlap values for each subsample to each other using paired-sample *t*-tests (two-tailed) with appropriate Bonferroni corrections.

Modelling species predictions

Twelve predictive techniques were used to fit the species distribution models (Table 1). These were: (1) the DIVA-GIS implementation of BIOCLIM (Busby, 1991); (2) DOMAIN (Carpenter *et al.*,

Table 1 Algorithms examined in this paper.

Method	Class of model (see Elith *et al.*, 2006 for full details of implementation)	Data	Software	Example
BIOCLIM	Envelope model	PO	DIVA-GIS (www.diva-gis.org)	(Busby, 1991)
BRUTO	Regression, a fast implementation of a GAM	PA	R and S-Plus, mda package	(Hastie *et al.*, 2001)
DK-GARP	Rule sets derived with genetic algorithms; desktop version	PA	DesktopGarp	(Stockwell & Peters, 1999)
DOMAIN	Multivariate distance	PA	DIVA-GIS	(Carpenter *et al.*, 1993)
GAM	Regression: GAM	PA	S-Plus, GRASP add-on	(Guisan *et al.*, 2002)
GBM	Boosted decision trees	PA	R, gbm package	(Friedman *et al.*, 2000)
GLM	Regression; generalized linear model	PA	S-Plus, GRASP add-on	(Guisan *et al.*, 2002)
LIVES	Multivariate distance	PA	Specialized program not yet publicly released	(Elith *et al.*, 2006)
MARS	Regression; multivariate adaptive regression splines	PA	R, mda package plus new code to handle binomial responses	(Moisen & Frescino, 2002)
MARSINT	As above; interactions allowed	PA	As above	(Moisen & Frescino, 2002)
MAXENT	Maximum entropy with threshold features	PE	Maxent	(Phillips *et al.*, 2006)
OM-GARP	Rule sets derived with genetic algorithms; open modeller version	PA	New version of GARP not yet available	(Anderson *et al.*, 2002; Elith *et al.*, 2006)

PO, only presence data used; PE, presence compared against the entire region; PA, presence and some form of absence required. For these analyses, we randomly selected 10,000 pseudo-absences from each region. GAM, generalized additive model.

1993); (3) GLM: generalized linear model (Guisan *et al.*, 2002); (4) GAM: generalized additive model (Guisan *et al.*, 2002); (5) BRUTO: a fast implementation of GAM (Hastie *et al.*, 2001); (6) MARS: multivariate adaptive regression splines (Moisen & Frescino, 2002); (7) MARSINT: similar to (6) but incorporating interactions between predictors (Moisen & Frescino, 2002); (8) GBM: generalized boosting methods/boosted decision trees (Friedman *et al.*, 2000); (9) DK-GARP: genetic algorithm for rule-based predictions based on Stockwell & Peters (1999); (10) OM-GARP: new implementation of a genetic algorithm for rule-based predictions that minimizes errors of omission and balances errors of commission (Anderson *et al.*, 2002); (11) LIVES: based on distances in multidimensional space (Elith *et al.*, 2006); and (12) MAXENT: based on maximum entropy and L-1 regularization (Phillips *et al.*, 2006). Some of these methods required both presence and absence data (Table 1). For these methods we randomly generated 10,000 pseudo-absences within each region. These were intended as a sample of the whole region, and though possible, there was a very low probability that a background sample coincided with a presence record. All data, modelling techniques, specific fitting details, and modelling implementations are described in full in Elith *et al.* (2006). The models were fitted by those among the authors who knew the technique best. DK-GARP could not be run for NZ owing to computer memory limitations; therefore, the total number of model runs made for this study was 46 species × 3 sample sizes × 12 modelling algorithms minus the 10 NZ species not modelled by DK-GARP. This resulted in 1646 spatial predictions of species distributions.

Evaluating predictions

We evaluated our predictions using independently collected presence and absence data. These data were often more precise and accurate than the data used to train the models. To assess model discriminatory power for each prediction, we computed the area under the receiver operating characteristic curve (AUC) and correlation (COR).

AUC evaluates how well model predictions discriminate between locations where observations are present and absent, and is one of the most widely used threshold-independent evaluators of model discriminatory power (Fielding & Bell, 1997). These curves are plotted as model sensitivity versus (1 – specificity) for a range of increasing, predictive threshold values. AUC can range from 0 to 1. An AUC = 0.5 indicates that model performance is equal to that of a random prediction, while an AUC of 0.8 means that in places where a species is present in 80% of cases the predicted values will be higher than where the species has not been recorded. Furthermore, in interpreting AUC in terms of correct ranking of random suitable sites versus random unsuitable sites, a model with AUC = 0.66 ranks the suitability of the site correctly 66% of the time. AUC is not an absolute measure and is sensitive to the method in which absences in the evaluation data are selected (Lobo *et al.*, 2008). As long as the presence data are predicted reasonably well, it is very easy to obtain high AUC values if the evaluation data contain absence points selected from a very large area (e.g. an entire continent or the whole world). Nevertheless, AUC remains valid as a measure of relative model performance between models and between sample size for the same species and study area.

The correlation, COR, between the observation in the PA data set (a dichotomous variable) and the continuous prediction can be calculated as a Pearson correlation coefficient (COR). It is similar to AUC, but carries with it extra information: instead of being rank based, it takes into account how far the prediction varies from the observation (Elith *et al.*, 2006). Our results for AUC and COR were highly correlated, and consequently, we present results on AUC only.

We examined factors influencing model performance using linear mixed-effect (LME) models using the LME function in S-Plus. We modelled the response (arcsine transformed AUC) as a function of the following fixed effects: sample size (10, 30, 100), algorithm (the 12 algorithms), and their two-way interaction (sample size × algorithm), with species as random effects nested within blocks identified by region.

To examine the ecological effects of extremes in dispersal ability (volant birds and sedentary plants and trees) on model performance, we then excluded the three reptile species from NSW and performed separate, similar LME analyses that included additional fixed effects for spatial grain: (100-m resolution in CAN and SA versus 1000-m resolution in NSW, NZ, and SWI) and major taxonomic groups: birds (CAN and NSW) versus plants (NZ, SA, and SWI).

After confirming that algorithm and sample size had an important effect on model performance in the LME analysis, we summarized median AUC in relation to sample size by method in interaction plots, and also compared arcsine-transformed AUC of individual methods in Wilcoxon's paired tests. As AUC is best-suited for comparing relative performance of models within species rather than across species (Lobo *et al.*, 2008), we ranked the algorithms individually for each species to assess the position of each algorithm relative to others and analysed the ranks. We evaluated significant differences between pairs of ranked methods in Wilcoxon's paired tests to determine differences in model performance. We ran all possible combinations of tests and used a Bonferroni correction to adjust the significance level threshold, as done in Graham *et al.* (2008).

RESULTS

Three paired-sample *t*-tests confirmed that smaller samples indeed exhibited significantly lower environmental range overlap than larger samples (Fig. 2). The first *t*-test compared the range overlap mean computed from 10-record samples (mean = 0.64, standard deviation (SD) = 0.110) to those of 30-record samples (mean = 0.80; SD = 0.097, $P < 0.001$, $t = -11.41$, d.f. = 45). The second *t*-test compared the range overlap mean of 10-record samples to 100-record samples (mean = 0.91, SD = 0.07, $P < 0.001$, $t = -10.69$, d.f. = 45). The last test compared 30-record mean to the 100-record mean ($P < 0.001$, $t = -22.27$, d.f. = 45). In each case, the significant test result confirmed that our experimental manipulation indeed challenged the modelling

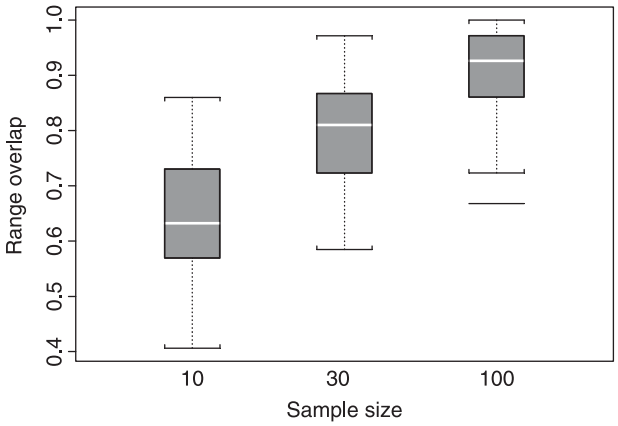

Figure 2 Range overlap values for the three sample size manipulations. The horizontal line shows the median response per sample size. The top and bottom of the box show the 25 and 75 percentiles, respectively. The horizontal line joined to the box by the dashed line shows 1.5 times the interquartile range of the data. Outliers are indicated by horizontal lines.

Table 2 Results of a linear mixed-effect (LME) analysis investigating determinants of area under the receiver operating characteristic curve (AUC) scores. Arcsine-transformed AUC scores were modelled as a function of the main fixed effects: sample size (10, 30, or 100 records) and model (12 algorithms listed in Table 1), and their two-way interaction. Each unique species was treated as a random effect, nested within each of the five regions. Aikake's Information Criterion (AIC) for the full model (including the significant two-way interaction term) was –2720.16, while the less parsimonious model which lacked the interaction term had AIC = −2693.42.

	Degrees of freedom	*F*-value	*P*-value
Intercept	1	1334.39	< 0.0001
Sample size	2	88.49	< 0.0001
Model	11	18.51	< 0.0001
Sample size × model	22	3.21	< 0.0001

algorithms by requiring them to predict species distributions based on less overall information.

Our first linear mixed-effect model indicated a strong response to the experimental manipulation of sample size; most importantly, the response was model dependent. Significant effects on AUC of sample size, algorithm, and their interactions were all observed (*F*-test, $P < 0.0001$, Table 2). The significant interaction between sample size and algorithm term exists because performance of some algorithms was much more sensitive to sample size than others (Fig. 3). For example, DOMAIN and LIVES were insensitive to sample size relative to other algorithms while MARS and BRUTO were highly sensitive to sample size (Fig. 3).

No algorithm performed better than all others across sample sizes, and only MAXENT approached a type A (Fig. 1) pattern. MAXENT had moderate sample size sensitivity combined with excellent predictive ability. It was the second best performer at the high and intermediate sample sizes and best at low sample sizes. GBM was the best performing algorithm at sample sizes 30 and 100. However, not all of these relationships were significantly

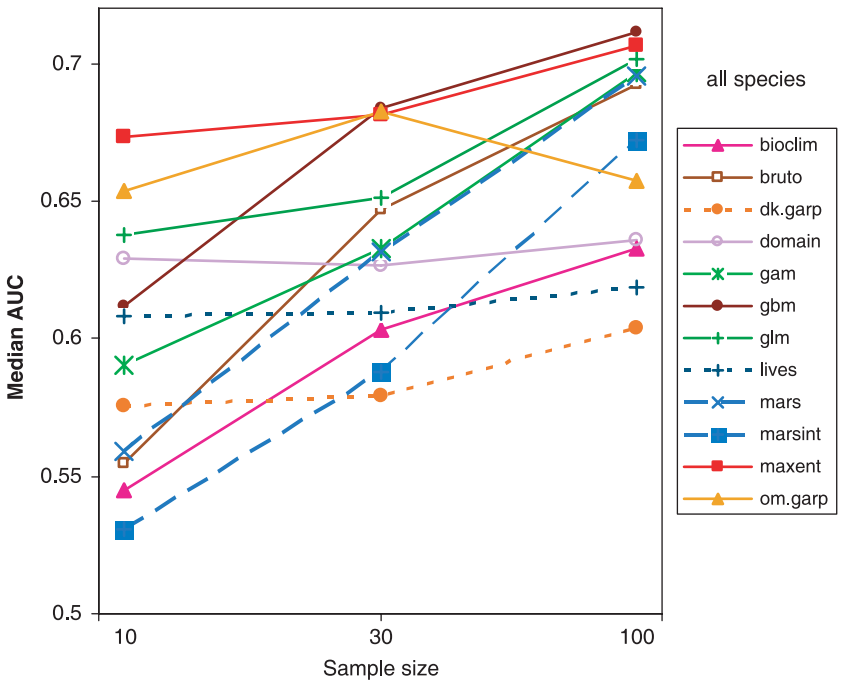

Figure 3 Median area under the receiver operating characteristic curve (AUC) versus sample size (all regions pooled).

different, and even the 'best' algorithms were significantly better than only a few other algorithms at the lowest sample size (see Appendix S1 in Supplementary Material). At a sample size of 100, GBM was significantly better than all but GAM and MAXENT, and MAXENT was significantly better than all but GAM, GBM, and MARS. Among the two best performing algorithms at a sample size of 10, OMGARP was significantly better than BIOCLIM, BRUTO, GAM, and MARSINT, while MAXENT was significantly better than only BIOCLIM, BRUTO, GAM, MARS, and MARSINT (Appendix S1).

Median AUCs of most algorithms resembled a type B or C pattern (or something in between), in which nearly all algorithms performed better with more records. GBM and GLM were among the top 25% of algorithms for 100 records (Fig. 3) but these algorithms showed sensitivity to sample size (type B to C pattern); at 10 records, AUCs for GBM and GLM were considerably lower than at larger sample sizes (Fig. 3). MARS, MARSINT, GAM, and BRUTO exhibited a clear type C pattern, in which performance decreased steadily with sample size (Fig. 3). Among these, BRUTO dropped from intermediate performance at 100 and 30 records to the lowest quartile of algorithms at 10 records. DOMAIN and LIVES showed insensitivity to sample size (type D pattern), with intermediate AUC at low sample size, but relatively poor performance at high sample size (Fig. 3).

Regional differences

The main effects described above hide major differences between algorithm quality and response to sample size among regions (see Appendix S2 in Supplementary Material). In some regions all algorithms performed much better than in other regions. Median AUCs were highest in NZ (0.68 across sample sizes and methods), and consistent low-quality predictions across sample sizes (median AUC = 0.573) were obtained with all algorithms in CAN, paralleling results of previous analyses (Elith *et al.*, 2006).

Variance in AUC

Ideally, algorithms should yield predictions with high AUCs with low variability across species (Elith *et al.*, 2006). Across sample sizes, BIOCLIM exhibited the lowest variance, but also among the lowest AUC, while BRUTO, GAM, and GLM yielded the highest variances, with intermediate AUC values (Fig. 4). At 10 records, MAXENT, OM-GARP, and DOMAIN had the highest AUCs combined with intermediate variances. At 30 and 100 records, MAXENT maintained intermediate variance with relatively high AUCs. GBM had the highest AUCs combined with intermediate variance at 30 and 100 records.

Ranking of algorithms

According to ranked median AUCs across species, certain algorithms performed better at the largest sample size than at smaller sample sizes. At 100 records, GBM and MAXENT outranked all other algorithms according to median ranks (Fig. 5). GBM significantly outranked all except MAXENT and GLM at 100 records, and outranked all except these and OMGARP at 30 records (see Appendix S3 in Supplementary Material). At 100 records MAXENT significantly outranked only BIOCLIM, LIVES, and MARSINT, but at 30 records it outranked these as well as LIVES and BRUTO (Appendix S3). Although GBM ranked best at 100 and 30 records, it ranked fifth at 10 records (Fig. 5). At the smallest sample size, MAXENT, DOMAIN, and OM-GARP performed best when ranked against other algorithms (Fig. 5). MAXENT significantly outranked all algorithms except DOMAIN, OMGARP, GAM, GBM, and GLM. OMGARP significantly outranked only BIOCLIM and BRUTO, while DOMAIN significantly outranked only MARSINT, LIVES, and BIOCLIM (Appendix S3)

Grain and taxon

Environmental data from CAN and SA had a spatial resolution of approximately 1000 m, while those from NSW, NZ, and SWI had a resolution of 100 m. CAN and NSW were vertebrate data sets consisting mainly of birds, while the other regions consisted of plants. Sample size and algorithm had significant effects on mean AUC, even after controlling for grain size and taxon. A linear mixed effect model fitting mean AUC as a function of the fixed effects experiment, algorithm, and grain size (100 m or 1000 m) and with species nested within regions as random effects revealed that AUC tended to be slightly higher in regions analysed with 100-m environmental data than in those regions analysed with 1000-m resolution data ($P = 0.48$, $t = -3.22747$). A similar model that coded species as plants or birds (the three NSW reptile species were excluded from this analysis) showed that differences in AUC between plants and birds were not significant ($P > 0.05$) (see Appendix S4 in Supplementary Material).

DISCUSSION

Our results are consistent with previous studies that found model performance increases while variability in predictive accuracy decreases with increasing sample size (Cumming, 2000; Pearce & Ferrier, 2000; Stockwell & Peterson, 2002; Kadmon *et al.*, 2003; Reese *et al.*, 2005; Hernandez *et al.*, 2006). Among these studies, only Hernandez *et al.* (2006) included and evaluated MAXENT predictions, finding that it yielded high-quality predictions that outperformed DOMAIN, DK-GARP, and BIOCLIM. The present study is the first to evaluate the predictive power of those as well as many other algorithms, including other novel methods (e.g. OMGARP, BRUTO, and GBM) that yielded particularly robust predictions in our more general evaluation (Elith *et al.*, 2006). Unlike previous studies, we analysed data from multiple continents and diverse taxa, and used independent data for model evaluation.

Models that included interactions or other complex relationships to predictors (e.g. GBM, GAM, MARSINT) performed better at larger sample size than at smaller sample size. MAXENT and OMGARP were among the least sensitive to sample size, and generally outperformed other methods at the smallest sample size. However, no algorithm predicted consistently well across all

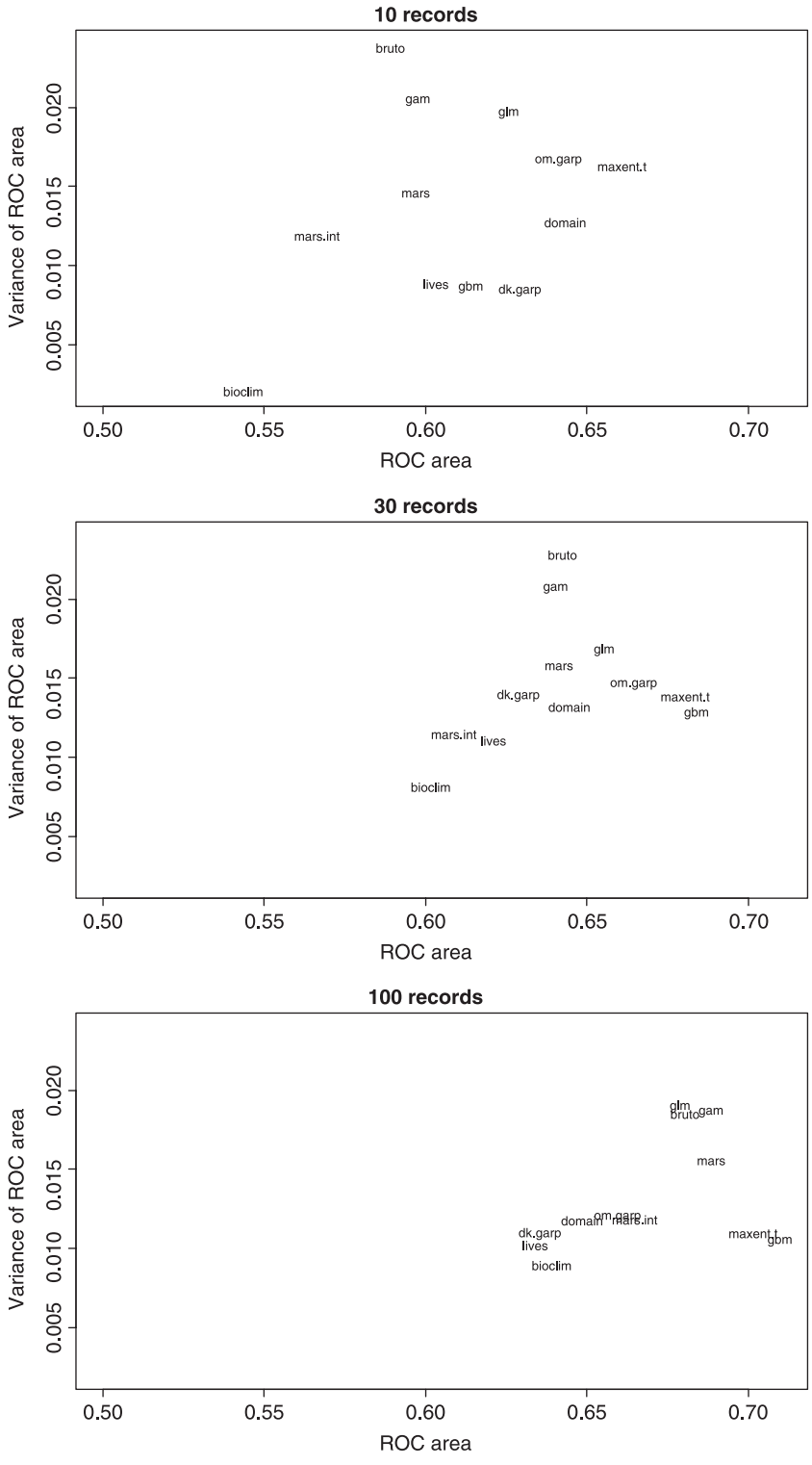

Figure 4 Performance measured by mean area under the receiver operating characteristic curve (AUC) and its variance; for consistent and good performance, algorithms should yield predictions with high AUC and low variance (i.e. lower right in plot).

species and regions. Any prediction should be considered preliminary until it can be confirmed, but particular caution should be applied to predictions made from small sample sizes (Figs 3 and 4). Because many species are known by relatively few records, our results highlight the need for accelerated development of high-quality data bases of occurrence information associated with specimens in museums and herbaria, and through new, high-quality field surveys that produce records vouchered by specimens.

Algorithms that model complex relationships of predictors and interactions are typically considered to be 'data-hungry', and consequently have seldom been used in predicting species

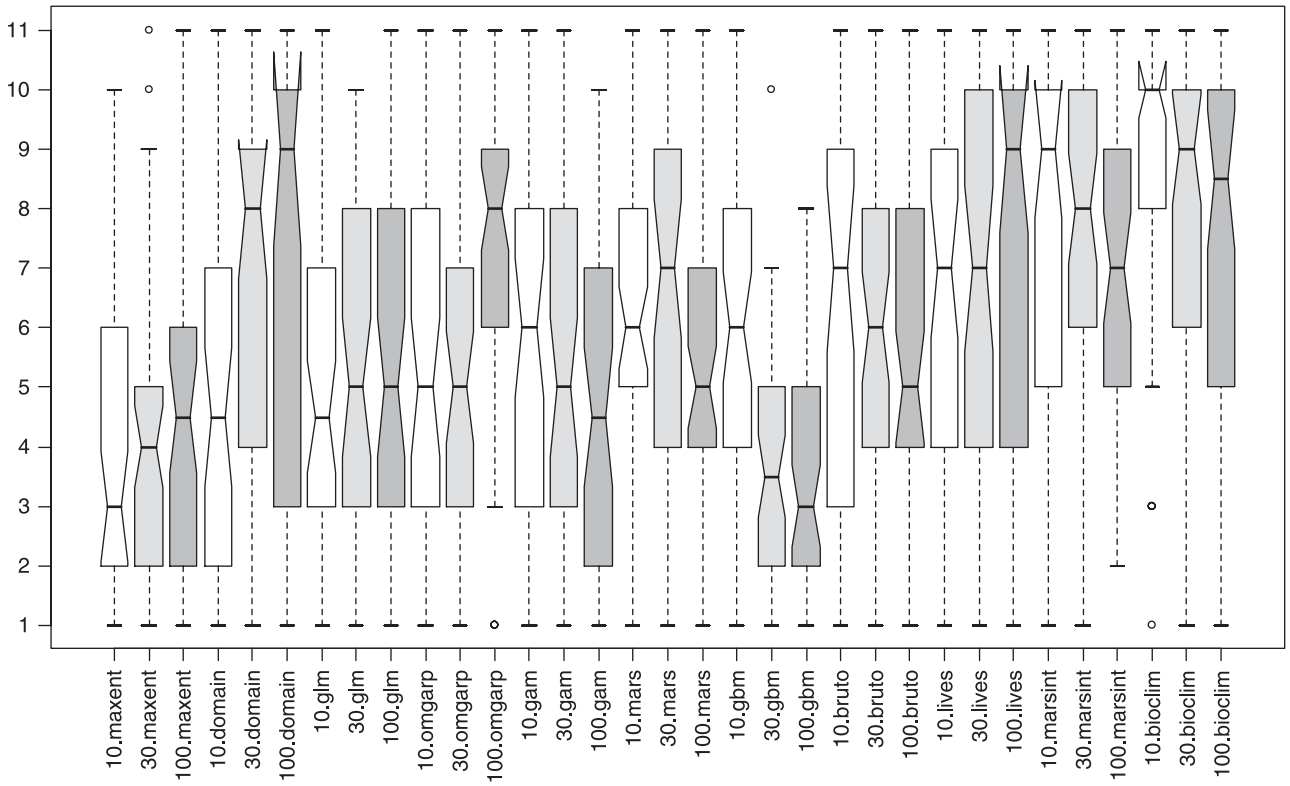

Figure 5 Notched boxplots of rank area under the receiver operating characteristic curve by method and sample size. If the notches of two plots do not overlap this is strong evidence that the two medians differ at the 95% confidence interval (Chambers *et al.*, 1983). Due to the large number of paired comparisons, this figure should be read in connection with Appendix S3 to identify the pairs that are significantly different after Bonferonni correction in Wilcoxon paired tests.

distributions when data are limited (Guisan & Thuiller, 2005) Three otherwise-powerful algorithms (Elith *et al.*, 2006) showing clear sensitivity to sample size are GAM, GBM, and BRUTO. These algorithms can model complex relationships to predictors and/or interactions, but their performance suffers as data are rarified. MAXENT has the capacity to model complex relationships and interactions, but our implementation did not include interactions.

In order to establish why some methods performed better than others for different species in different regions, ideally we would like to have known *a priori* which species had complex relationships to predictor variables, and which should have best been modelled by interactions in order to control for their influence on predictive power, but this was not possible. If we had used a simulated species with a known distribution (as applied in Austin *et al.*, 2006) we could have specified the complexity of the species relationship to predictor variables *a priori*. Further to this, it would have been interesting to use a simulated region with controlled complexity (e.g. predefined landscape heterogeneity) to evaluate the effects of this on model performance. Such controlled simulation would have helped us exclude some of the noise from our analyses that hampered our efforts to discover why some methods performed consistently better than others with small sample size. However, although simulated species and regions can be useful in controlled experiments to explore bias in modelling, they are, of course, unrealistically simple, and results obtained from such experiments may not always be applicable to real species and regions. On the other hand, even though our study species were nested in regions represented by different major taxonomic groups (birds, reptiles, herbaceous plants, or trees), different spatial grain (100-m resolution in NSW, NZ, and SWI to 1000 m in CAN and SA), and different environmental predictors, many results were consistent across regions and are therefore clearly robust.

Strong and consistent performance across sample size manipulations probably depends on an algorithm's ability to generate inferences from small amounts of environmental information. MAXENT's performance may be explained by the way it uses regularization to avoid over-fitting. The amount of regularization varies flexibly with sample size to ensure consistent performance, while the type of regularization (L-1) employed tends to cause irrelevant variables to be omitted from the model L-1 regularization has also been applied to GLMs and GAMs, and is called the 'lasso' in that context (Phillips *et al.*, 2006). L-1 regularization is an alternative to stepwise model selection procedures and uses a shrinkage parameter, typically selected by cross-validation but in this case it is specially tuned for each given sample size to determine coefficients in the model. This parameter shrinks the coefficients of terms in the model towards zero, while setting the terms with

limited explanatory power to zero in the process. This has the effect of reducing the variance of the fit of the model, while increasing the bias. Striking a balance between these two quantities results in a parsimonious model that best predicts unseen observations (Guisan *et al.*, 2002). OM-GARP's consistent performance across sample size manipulations is likely to result from use of the best-subsets procedure which forces resulting models to be general in their predictive abilities, while striking a balance between errors of omission and commission(Anderson *et al.*, 2002). MAXENT and OM-GARP have only recently been applied in ecology (Elith *et al.*, 2006; Phillips *et al.*, 2006), so their solid performance here, particularly with small data sets, should further encourage their use.

No algorithm predicted all 46 species well at the smallest sample size, but predictions made for large sample sizes generally outperformed those at the smallest sample size. Species distributions not successfully modelled by any method may be determined by factors other than climate, possess complex relationships among predictors, and/or interaction terms may be important. Several studies have demonstrated the importance of interactions and nonlinear relationships in explaining species ecological and geographical distributions (Austin, 2002; Guisan & Thuiller, 2005). Though beyond the scope of this study, it could be helpful to examine influences of interaction terms and the complexity of response shapes for a large sample size (e.g. > 100 records) in a diversity of modelling algorithms. If a species distribution is best represented by the inclusion of interaction terms and/or complex response shapes at large sample size, we might expect these algorithms to perform poorly at reduced sample sizes, as less information is available to support complex models.

Typical data sets may be characterized by sampling biases. All of our species had at least 100 samples and smaller samples were drawn from random throughout the species range. As a consequence, we cannot know what results could be expected from a species that was 'naturally' data-depauperate (fewer than 30 records) due to biological rarity, though this is a topic that deserves further research.

The degree of predictive accuracy necessary depends on the intended use of the model (Araujo *et al.*, 2005), and our results have important implications for applications of distribution modelling. We have shown that predictions based on small samples are generally unlikely to be suitable for conservation planning and other complex applications. What is more, unless suitable evaluation data are available (in which case, low sample sizes are probably not an issue), it may not be possible to know which models are appropriately trained and will yield robust predictions. Rare or poorly known species are of particular interest to conservation practitioners, and models are often used to fill in gaps in information. If, on the other hand, the intended use of a model is to explore the data available, predictions based even on small sample sizes may yield results useful in prioritizing future data collection efforts for rare species (Raxworthy *et al.*, 2003; Engler *et al.*, 2004; Guisan *et al.*, 2006) or exploring macroecological patterns in poorly known regions or taxa (Wisz *et al.*, 2007). We have shown that no modelling approach tested was fully robust to small sample sizes, but that for exploratory modelling with sample sizes between 10 and 30 records, MAXENT, OM-GARP, and possibly DOMAIN may be the best available.

Distribution models are increasingly used to forecast geographical range shifts of floras and faunas in response to climate and land-use change. For example (Thomas *et al.*, 2004) presented a compilation of a distribution modelling predictions across multiple regions to evaluate potential climate change effects. While some models appear intrinsically capable of making such predictions (Hijmans & Graham, 2006), results can remain controversial because of uncertainty associated with predicting distributions for current and hence future situations (Thuiller, 2004). Here, we have shown that very different predictions can be obtained depending on region, sample size, and the algorithm used. Model-based studies should thus consider whether the uncertainty associated with the choice of modelling algorithm exceeds the tolerance for uncertainty required by the question at hand. Our results should encourage further, though cautious, use of predictions based on small sample size.

ACKNOWLEDGEMENTS

This research was initiated in a working group at the National Center for Ecological Analysis and Synthesis (NCEAS), Santa Barbara, California, USA 'Testing Alternative Methodologies for Modelling Species' Ecological Niches and Predicting Geographic Distributions', conceived and led by A.T. Peterson and C. Moritz.

This manuscript benefited from the helpful suggestions of three anonymous referees.

REFERENCES

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

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

Austin, M.P. (2002) Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological Modelling*, **157**, 101–118.

Austin, M.P., Belbin, L., Meyers, J.A., Doherty, M.D. & Luoto, M. (2006) Evaluation of statistical models used for predicting plant species distributions: role of artificial data and theory. *Ecological Modelling*, **199**, 197–216.

Buckland, S.T., Elston, D.A. & Beaney, S.J. (1996) Predicting distributional change, with application to bird distributions in northeast Scotland. *Global Ecology and Biogeography Letters*, **5**, 66–84.

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

Carpenter, G., Gillison, A.N. & Winter, J. (1993) DOMAIN: a flexible modelling procedure for mapping potential distributions

of plants and animals. *Biodiversity and Conservation*, **2**, 667–680.

Carroll, S.S. & Pearson, D.L. (1998) The effects of scale and sample size on the accuracy of spatial predictions of tiger beetle (Cicindelidae) species richness. *Ecography*, **21**, 401–414.

Chambers, J.M., Cleveland, W.S., Kleiner, W.S. & Tukey, P.A. (1983) *Graphical methods for data analysis*. Wadsworth and Brooks/Cole, Pacific Grove, CA.

Crawley, M. (2002) *Statistical computing: an introduction to data analysis using S-plus*. John Wiley & Sons, Ltd, Chichester, U.K.

Cumming, G.S. (2000) Using between-model comparisons to fine-tune linear models of species ranges. *Journal of Biogeography*, **27**, 441–455.

Elith, J., Graham, C.H., Anderson, R.P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L.G., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J.M., Peterson, A.T., Phillips, S.J., Richardson, K., Scachetti-Pereira, R., Schapire, R.E., Soberon, J., Williams, S., Wisz, M.S. & Zimmermann, N.E. (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**, 129–151.

Engler, R., Guisan, A. & Rechsteiner, L. (2004) An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *Journal of Applied Ecology*, **41**, 263–274.

Ferrier, S. & Watson, G. (1996) An evaluation of the effectiveness of environmental surrogates and modelling techniques in predicting the distribution of biological diversity. Consultancy report prepared by the NSW National Parks and Wildlife Service for Department of Environment, Sport and Territories, Canberra, ACT, Australia.

Ferrier, S., Watson, G., Pearce, J. & Drielsma, M. (2002) Extended statistical approaches to modelling spatial pattern in biodiversity: the north-east New South Wales experience. I. Species-level modelling. *Biodiversity and Conservation*, **11**, 2275–2307.

Fielding, A.H. & Bell, J.F. (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, **24**, 38–49.

Friedman, J., Hastie, T. & Tibshirani, R. (2000) Additive logistic regression: a statistical view of boosting. *Annals of Statistics*, **28**, 337–374.

Graham, C.H., Elith, J., Hijmans, R.J., Guisan, A., Peterson, A.T., Loiselle, B.A. & 'NCEAS, Predicting Distributions Working Group'. (2008) The influence of spatial errors in species occurrence data used in distribution models. *Journal of Applied Ecology*, **45**, 239–247.

Graham, C.H., Ferrier, S., Huettman, F., Moritz, C. & Peterson, A.T. (2004) New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology and Evolution*, **19**, 497–503.

Guisan, A., Broennimann, O., Engler, R., Vust, M., Yoccoz, N.G., Lehmann, A. & Zimmermann, N.E. (2006) Using niche-based models to improve the sampling of rare species. *Conservation Biology*, **20**, 501–511.

Guisan, A., Edwards, T.C. & Hastie, T. (2002) Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling*, **157**, 89–100.

Guisan, A., Graham, C.H., Elith, J. & Huettmann, F. (2007) Sensitivity of predictive species distribution models to change in grain size. *Diversity and Distributions*, **13**, 332–340.

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

Hastie, T., Tibshirani, R. & Friedman, J.H. (2001) *The elements of statistical learning: data mining, inference, and prediction*. Springer-Verlag, New York.

Hepinstall, J.A. & Sader, S.A. (1997) Using Bayesian statistics, Thematic Mapper satellite imagery, and breeding bird survey data to model bird species probability of occurrence in Maine. *Photogrammetric Engineering and Remote Sensing*, **63**, 1231–1237.

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

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

Hijmans, R.J., Garrett, K.A., Huaman, Z., Zhang, D.P., Schreuder, M. & Bonierbale, M. (2000) Assessing the geographic representativeness of Genebank collections: the case of Bolivian wild potatoes. *Conservation Biology*, **14**, 1755–1765.

Hirzel, A.H., Hausser, J., Chessel, D. & Perrin, N. (2002) Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? *Ecology*, **83**, 2027–2036.

Hutchinson, G.E. (1957) Concluding remarks. *Cold Springs Harbor Symposia on Quantitative Biology*, **22**, 415–427.

Kadmon, R., Farber, O. & Danin, A. (2003) A systematic analysis of factors affecting the performance of climatic envelope models. *Ecological Applications*, **13**, 853–867.

Kunin, W.E. & Gaston, K.J. (1993) The biology of rarity – Patterns, causes and consequences. *Trends in Ecology and Evolution*, **8**, 298–301.

Ladle, R.J., Jepson, P., Araujo, M.B. & Whittaker, T.J. (2004) Dangers of crying wolf over risk of extinctions. *Nature*, **428**, 799.

Lobo, J.M., Jiménez-Valverdel, A. & Real, R. (2008) AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, **17**, 145–151.

Manel, S., Dias, J.M. & Ormerod, S.J. (1999) Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a Himalayan river bird. *Ecological Modelling*, **120**, 337–347.

Moisen, G.G. & Frescino, T.S. (2002) Comparing five modeling techniques for predicting forest characteristics. *Ecological Modelling*, **157**, 209–225.

Pearce, J. & Ferrier, S. (2000) An evaluation of alternative algorithms for fitting species distribution models using logistic regression. *Ecological Modelling*, **128**, 127–147.

Pearson, R.G., Thuiller, W., Araujo, M.B., Martinez-Meyer, E., Brotons, L., McClean, C., Miles, L., Segurado, P., Dawson, T.P. & Lees, D.C. (2006) Model-based uncertainty in species range prediction. *Journal of Biogeography*, **33**, 1704–1711.

Phillips, S.J., Anderson, R.P. & Schapire, R.E. (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231–259.

Raxworthy, C.J., Martinez-Meyer, E., Horning, N., Nussbaum, R.A., Schneider, G.E., Ortega-Huerta, M.A. & Peterson, A.T. (2003) Predicting distributions of known and unknown reptile species in Madagascar. *Nature*, **426**, 837–841.

Reese, G.C., Wilson, K.R., Hoeting, J.A. & Flather, C.H. (2005) Factors affecting species distribution predictions: a simulation modeling experiment. *Ecological Applications*, **15**, 554–564.

Rushton, S.P., Ormerod, S.J. & Kerby, G. (2004) New paradigms for modelling species distributions? *Journal of Applied Ecology*, **41**, 193–200.

Stockwell, D. & Peters, D. (1999) The GARP modelling system: problems and solutions to automated spatial prediction. *International Journal of Geographyraphic Information Science*, **13**, 143–158.

Stockwell, D.R.B. & Peterson, A.T. (2002) Effects of sample size on accuracy of species distribution models. *Ecological Modelling*, **148**, 1–13.

Thomas, C.D., Cameron, A., Green, R.E., Bakkenes, M., Beaumont, L.J., Collingham, Y.C., Erasmus, B.F.N., de Siqueira, M.F., Grainger, A., Hannah, L., Hughes, L., Huntley, B., van Jaarsveld, A.S., Midgley, G.F., Miles, L., Ortega-Huerta, M.A., Peterson, A.T., Phillips, O.L. & Williams, S.E. (2004) Extinction risk from climate change. *Nature*, **427**, 145–148.

Thuiller, W. (2004) Patterns and uncertainties of species' range shifts under climate change. *Global Change Biology*, **10**, 2020–2027.

Wisz, M.S., Walther, B.A. & Rahbek, C. (2007) Using potential distributions to explore determinants of Western Palaearctic migratory songbird species richness in sub-Saharan Africa. *Journal of Biogeography*, **34**, 828–841.

Editor: David Richardson

SUPPLEMENTARY MATERIAL

The following supplementary material is available for this article:

Appendix S1 Probability that the mean area under the receiver operating characteristic curve (AUC) of a pair of methods is significantly different based on paired *t*-tests comparing arcsine transformed AUC.

Appendix S2 Median area under the receiver operating characteristic curve (AUC) versus sample size by region.

Appendix S3 Wilcoxon paired tests on ranks of AUC scores.

Appendix S4 Boxplots of area under the receiver operating characteristic curve by region and sample size for three methods: (a) MAXENT, (b) GBM, and (c) GLM.

This material is available as part of the online article from:
http://www.blackwell-synergy.com/doi/abs/10.1111/j.1472-4642.2008.00482.x
(This link will take you to the article abstract).

Please note: Blackwell Publishing is not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.