Ecological Applications, 21(2), 2011, pp. 335–342

Ecological niche modeling in Maxent: the importance of model complexity and the performance of model selection criteria

Dan L. Warren[1,3] and Stephanie N. Seifert[2]

[1]*Section of Integrative Biology, University of Texas at Austin, Austin, Texas 78712 USA*
[2]*Department of Entomology, University of California, Davis, California 95616 USA*

Abstract. Maxent, one of the most commonly used methods for inferring species distributions and environmental tolerances from occurrence data, allows users to fit models of arbitrary complexity. Model complexity is typically constrained via a process known as L_1 regularization, but at present little guidance is available for setting the appropriate level of regularization, and the effects of inappropriately complex or simple models are largely unknown. In this study, we demonstrate the use of information criterion approaches to setting regularization in Maxent, and we compare models selected using information criteria to models selected using other criteria that are common in the literature. We evaluate model performance using occurrence data generated from a known "true" initial Maxent model, using several different metrics for model quality and transferability. We demonstrate that models that are inappropriately complex or inappropriately simple show reduced ability to infer habitat quality, reduced ability to infer the relative importance of variables in constraining species' distributions, and reduced transferability to other time periods. We also demonstrate that information criteria may offer significant advantages over the methods commonly used in the literature.

Key words: Akaike information criterion; AUC; Bayesian information criterion; environmental niche modeling; Maxent; maximum entropy; model complexity; model transferability; niche shifts; species distribution modeling.

Introduction

Environmental (or ecological) niche models (ENMs) are a class of methods that use occurrence data in conjunction with environmental data to make a correlative model of the environmental conditions that meet a species' ecological requirements and predict the relative suitability of habitat. ENMs are most often used in one of four ways: (1) to estimate the relative suitability of habitat known to be occupied by the species, (2) to estimate the relative suitability of habitat in geographic areas not known to be occupied by the species, (3) to estimate changes in the suitability of habitat over time given a specific scenario for environmental change, and (4) as estimates of the species niche. While transferability of ENMs (i.e., uses 2 and 3) and use of ENMs as niche estimates (4) are known to be accompanied by a host of conceptual and practical problems (see also Hampe 2004, Soberón and Peterson 2005, Menke et al. 2009), in many cases ENM methods are employed because they are quite simply the only tools available.

In this study, we will look at the effects of model complexity on ENMs constructed using one of the most commonly employed tools for this purpose, Maxent (Phillips et al. 2006). Maxent uses the principle of maximum entropy on presence-only data to estimate a set of functions that relate environmental variables and habitat suitability in order to approximate the species' niche and potential geographic distribution (Phillips et al. 2006). In principle maximum entropy seeks a marginal suitability function for each variable that matches the empirical data, is maximally uninformative elsewhere, and has a mean equal to that from the empirical data. However, strict adherence to this requirement can lead to models that overfit input data. For this reason, Maxent uses a process called L_1 regularization to constrain modeled distributions to lie within a certain interval around the empirical mean rather than matching it exactly. L_1 regularization operates by minimizing the following expression:

$$\tilde{\pi}(-\ln[q_\lambda]) + \sum_j \beta_j|\lambda_j|$$

(Phillips et al. 2006). The first term represents the log loss, while the second term consists of a set of weights (λ_j) for the j features used to build the model and a set of weighting penalties, β_j. Users of Maxent can specify values of β_j that penalize the addition of extra parameters to suit their particular application, data, and biological intuition. Recent releases of the Maxent software use different default β values for each type of variable (linear, quadratic, hinge, and so on; Phillips and Dudík 2008), which are set to values that were obtained

Manuscript received 9 June 2010; revised 6 July 2010; accepted 15 July 2010. Corresponding Editor: T. J. Stohlgren.

[3] E-mail: dan.l.warren@gmail.com

using empirical data for 226 species from six different geographic regions. The ability to change the settings for each variable type is available, but users typically adjust regularization via a single β setting that acts as a multiplier for the default values, if they explore regularization at all. For the present study we manipulate only this regularization multiplier and refer to it simply as β.

While L_1 regularization allows users to constrain over-parameterization, they must still decide on a criterion by which to choose an appropriate value of β for their data. Phillips and Dudík (2008) suggest that the default settings in Maxent are likely to be appropriate for many modeling efforts, and the empirical data presented in that study lend credibility to this statement. However, those default settings were obtained using criteria that did not penalize model complexity per se, and models were evaluated only in their performance on independent test data. This type of model evaluation is one of the cornerstones of a school of statistical modeling known as algorithmic modeling (AM; Breiman 2001). While the more traditional "data modeling" (DM) approach starts with a family of models that are specified a priori and attempts to select the model or set of models that best fit the data, AM treats the true model as an unknown (and potentially quite complex) reality that is difficult or impossible to truly estimate. Consequently, while DM investigators usually evaluate models using goodness-of-fit metrics and place a high priority on model simplicity, AM users more typically focus on the ability of the model to predict independent test data and tend to be less directly focused on controlling over-parameterization. As normally applied, the Maxent software falls firmly under the AM approach, and previous studies evaluating its performance draw primarily from a model selection toolbox typical of that school of thought. This way of thinking about ENMs may be entirely appropriate for users whose primary objective is to predict the distribution of their species within the same set of environments available for data collection. However, some modelers using Maxent may have more cause to be concerned with model complexity per se, and to be less satisfied with using predictive accuracy on test data as a criterion for model selection. For applications 2, 3, and 4 above, the predictions of suitability on training and test data are only a projection of the true item of interest (i.e., the estimate of the species' niche) onto one distribution of environmental variables, and those predictions often cannot be tested in the environmental space of primary interest (e.g., a new geographic region [2], a different time period [3], or the space of all possible environments [4]). In addition to these issues, some authors have expressed concern regarding whether randomly withheld test data can truly be considered independent when training and test data are both subject to similar spatial sampling biases, as is usually the case for the occurrence data used in ENM construction (Veloz 2009; A. Radosavljevic and R. P. Anderson, *personal communication*).

In this study, we examine the effects of regularization on model performance using a simulation approach in which the underlying model generating species occurrences (the "true" environmental niche) is known, and evaluate a broader range of model performance and selection criteria than has previously been used. The purpose of these model selection criteria is not to replace the regularization currently available in Maxent, but rather to determine what methods users might employ to help them best use Maxent's existing regularization functions. We acknowledge that the use of $\mathrm{AIC_c}$ and BIC here is somewhat at odds with common practice for these criteria (Burnham and Anderson 2002); rather than specifying a set of models a priori, we simply specify a range of levels of complexity and allow the Maxent algorithm to control parameterization of the models. This is therefore still an AM approach to model construction, but with the addition of model selection heuristics from the DM literature. We therefore make no attempt to justify $\mathrm{AIC_c}$ and BIC from first principles. Instead, we simply focus on testing their utility as alternative heuristics to the more commonly used $\mathrm{AUC_{Train}}$ and $\mathrm{AUC_{Test}}$ (see *Methods: Criteria for model selection* for a description of these heuristics). However, we note that the philosophy underlying information criterion approaches and L_1 regularization are the same: we should reward models that fit data while penalizing unnecessary parameters. The primary conceptual difference is that $\mathrm{AIC_c}$ and BIC provide explicit criteria for selecting models of appropriate complexity, while L_1 regularization leaves that choice to the user.

Methods

In order to reliably assess the ability of Maxent to estimate species' environmental tolerances or to estimate suitability of habitat, we must start by knowing their true values. However, environmental niche modeling primarily exists because these are difficult, and frequently impossible, quantities to estimate. For that reason, we present a simulation approach which is now available in ENMTools under the name "resample from raster" (Warren et al. 2010). In this approach, we start with a Maxent ENM built using occurrence data from a real biological species. We treat that model as "truth" for the sake of simulation, and sample geographic localities with a probability proportional to the raw estimate of habitat suitability in the grid cell they represent. These occurrence data are then used as input for Maxent, and the ability of the program to infer the underlying model is evaluated. Since the original model was generated from Maxent, we know a priori that the software is capable of inferring that model precisely and that a perfect mapping of the true model to the geographic distribution of habitat suitability scores is possible. As such, this represents a relatively simple test of Maxent's ability to infer biological truth. By varying the β parameter used to construct the "true" ENM from empirical data, we can generate models and distributions of similarity scores for a variety of scenarios in which the "true" niche varies in complexity. Data simulated in this fashion are different from those obtained from field studies in that they are sampled with no spatial bias. Because of this, Maxent must be run with the "addsamplestobackground" option disabled in order to achieve statistically consistent results. A graphical representation of the simulation and testing process is given in Appendix D.

Data

We present analyses for 51 different species from California. All occurrence data were obtained from the Museum of Vertebrate Zoology at the University of California, Berkeley. Environment layers at a resolution of 30 arc seconds were obtained from Worldclim (Hijmans et al. 2005) and the California Gap Analysis Project (Davis et al. 1998), and trimmed to the state boundaries of California using ArcGIS (ESRI 2006). Layers used included slope, altitude, gap vegetation type, and the 19 layers commonly referred to as the "Bioclim" layers, which represent various aspects of temperature, precipitation, and seasonality. Although many of these variables are spatially correlated, we retained the entire set in order to determine the effects of different modeling approaches on variable selection. In order to estimate the transferability of models, we also projected true and inferred models onto future climate predictions for California under the CSIRO model, a2a climate scenario. These data were also obtained from Worldclim (Hijmans et al. 2005). Because no gap vegetation projections were available under this climate scenario, we treated the vegetation as unchanging over this time period.

Analyses

For each of the 51 species, we generated a "true" model at ten different levels of complexity by setting β at 1, 3, 5, 7, 9, 11, 13, 15, 17, and 19. We sampled 100 occurrence points from each of those "true" models, and constructed models from those occurrence points using the same ten values for β. We then repeated this process using 1000 simulated occurrence points in order to determine the effects of sample size on model performance and model selection. Twenty percent of occurrence records were withheld from each model to be used as independent test data. All other Maxent settings relating to model construction (with the exception of "add samples to background," as mentioned above) were left at their default values.

Evaluating model performance

We evaluate model performance using a variety of metrics (Appendix A). The I metric measures the ability of the model to estimate the true suitability of habitat (Warren et al. 2008), while the RR metric measures its ability to estimate the relative ranking of habitat patches. The M metric measures the ability of Maxent to determine the relative importance of environmental variables. The I_{Proj} and $\mathrm{RR}_{\mathrm{Proj}}$ metrics measure the effectiveness of Maxent at estimating true suitability or relative ranking of habitat in a different time period, respectively. All metrics range from 1, where the truth is estimated perfectly, to 0, where the inferred model shows no similarity to the true model. We note that these metrics may be broadly useful for comparing models, but they are generally not applicable to model selection using real data as they can only be calculated when the true niche and suitability of habitat are known.

Criteria for model selection

Finding that under- or over-parameterization are problematic is of great concern, but limited utility, in a world where the true complexity of the environmental niche is unknown and perhaps unknowable. We therefore test the performance of different methods of selecting from a set of models of varying complexity. They are:

1) Information criteria. Here we implement the Bayesian (BIC), and sample size corrected Akaike information criteria ($\mathrm{AIC_c}$) for Maxent ENMs (Akaike 1974, Burnham and Anderson 2002:284). These metrics are assessed by standardizing raw scores for each ENM so that all scores within the geographic space sum to 1 and then calculating the likelihood of the data given each ENM by taking the product of the suitability scores for each grid cell containing a presence. Both training and test localities are used in calculating likelihoods. The number of parameters is measured simply by counting all parameters with a nonzero weight in the lambda file produced by Maxent, a small text file containing model details that Maxent produces as part of the modeling process. We exclude models with zero parameters (which occur in some cases with small sample sizes and extremely high values for β) and models with more parameters than data points (which violate the assumptions of $\mathrm{AIC_c}$). We note that Maxent is capable of producing marginal suitability functions that take on a variety of shapes. Some of these functions may involve many more parameters than others (e.g., hinge features as compared to linear features), and as such are likely to be penalized more severely by information criteria. Functions for calculating $\mathrm{AIC_c}$ and BIC are now available in ENMTools (Warren et al. 2010).

2) Maximum training AUC. This method selects the model that produces the maximum value for the area under the receiver operating characteristic curve (AUC) calculated using the data used in model construction. This value is often reported in the literature as an estimate of model quality, but this interpretation is questionable as $\mathrm{AUC_{Train}}$ is generally expected to favor models with more parameters.

3) Maximum test AUC. This method selects the model that produces the maximum AUC value on randomly

selected test data that was withheld from model construction. This method is generally thought not to suffer from the same overfitting problems as $\mathrm{AUC}_{\mathrm{Train}}$, because overfitting the model to the training data should not necessarily improve the fit to independent test data.

4) Minimum difference between training and test data ($\mathrm{AUC}_{\mathrm{Diff}}$). This metric is based on the intuitive notion that overfit models should generally perform well on training data but poorly on test data (S. Sarkar, S. E. Strutz, D. M. Frank, C.-L. Rivaldi, B. Sissel, and V. Sanchez-Cordero, *unpublished manuscript*). By minimizing the difference between training and test data, we minimize the risk that our model is over-parameterized in such a way as to be overly specific to the training data.

Results

The effects of over- and under-parameterization can be seen in Fig. 1. Models constructed using the same value for β may differ in the number of parameters they include due to differences in sample size, the true distribution being estimated, and stochastic effects of sampling. Comparisons of models in this study are therefore made using the number of parameters rather than using β directly. Models with approximately the same number of parameters as the true model exhibited better performance on average than did models with too few or too many parameters, both in the present day and when projected onto a future climate scenario. When over- and under-parameterized models are analyzed separately, linear regression shows a significant relationship between degree of over- (or under-) parameterization and each of the performance metrics, although with model similarity (M) at a sample size of 1000 the effect size is small (slope $= 0.65 \times 10^{-3}$ for under-parameterized and -0.45×10^{-3} for over-parameterized models). Further, the absolute value of the slope of the regression was always greater for under- than for over-parameterized models, indicating that on a per-parameter basis under-parameterization has a stronger negative effect on model performance. ANCOVA showed that these differences in slope were highly significant ($P < 0.001$) for I, RR, I_{Proj}, and $\mathrm{RR}_{\mathrm{Proj}}$ when 1000 occurrence points were used and for I and I_{Proj} when 100 points were used, but were not significant for RR or $\mathrm{RR}_{\mathrm{Proj}}$ with 100 points ($P = 0.06$ and 0.08, respectively) or for M ($P = 0.158$ for 100 points and 0.139 for 1000). Although these results show that under-parameterization is generally worse for Maxent models than over-parameterization, we also examine separately the special case in which the number of parameters in the true model exceeds the number of occurrence points used in model construction (Appendix B). Under these conditions, we find that Maxent exhibits better performance on all metrics when the inferred model is simpler than the true model. Comparison of performance of model selection criteria is given in Fig. 2. All differences in ranking of models selected by different criteria were statistically significant (Friedman test, all $P < 10^{-5}$). Both of the information criterion-based methods outperformed all AUC-based methods for all metrics of model quality regardless of sample size with one exception: $\mathrm{AUC}_{\mathrm{Diff}}$ outperformed $\mathrm{AIC}_{\mathrm{c}}$ on selecting the appropriate number of parameters when $N = 1000$ (Fig. 2A), which is also the only comparison in which BIC outperformed $\mathrm{AIC}_{\mathrm{c}}$, and the only comparison where $\mathrm{AUC}_{\mathrm{Diff}}$ outperformed $\mathrm{AUC}_{\mathrm{Test}}$. $\mathrm{AUC}_{\mathrm{Train}}$ was the worst-performing method of model selection in all comparisons except for M and I when $N = 100$; for these comparisons, $\mathrm{AUC}_{\mathrm{Diff}}$ exhibited the poorest performance. We conducted a separate set of analyses to study the effects of model over- and under-parameterization on inferences of niche breadth and changes in niche breadth over time. These analyses are presented in Appendix S3 and discussed below, and should be of interest to investigators using Maxent to predict range contraction or expansion as a function of climate change.

Discussion

The prospect of using occurrence points and environmental data to estimate species' ecological tolerances opens up the possibility of conducting studies in both conservation and basic science that are difficult, and sometimes impossible, to address using other methods. However, this potential comes at a price: the inferences made from the models are only as good as the models themselves. This is particularly problematic given our inability in many cases to assess how well the models estimate biological truth. In this paper, we have generated simulated occurrence data from a known Maxent model and used those to examine the effects of model complexity on our ability to infer biological parameters of interest and on model transferability. The most commonly used metric for model quality ($\mathrm{AUC}_{\mathrm{Test}}$) relies on our ability to predict the distribution of independent occurrences in the training area. This metric is intuitive and useful, but it can lead to a level of confidence in the underlying model that may be unjustified: we found many situations in which I, RR, $\mathrm{AUC}_{\mathrm{Test}}$, $\mathrm{AUC}_{\mathrm{Train}}$, and $\mathrm{AUC}_{\mathrm{Diff}}$ were all high for a model while the M score was still quite low, indicating that even high predictive accuracy on species geographic distributions may be achieved by models that do a poor job of estimating the underlying biology (for a more thorough discussion of this point see Godsoe 2010).

Several interesting things are apparent from the comparison shown in Fig. 2. First, we note that all model selection criteria except $\mathrm{AUC}_{\mathrm{Train}}$ perform better when given more data. When $\mathrm{AUC}_{\mathrm{Train}}$ is used as a model selection criterion, it performs worse on the data sets containing 1000 occurrences than on data sets containing 100, regardless of which metric for model performance is used (panels B–E). This is likely due to its tendency to favor over-parameterized models, as seen in panel A. For this reason we advocate that AUC values on training data be approached with considerable

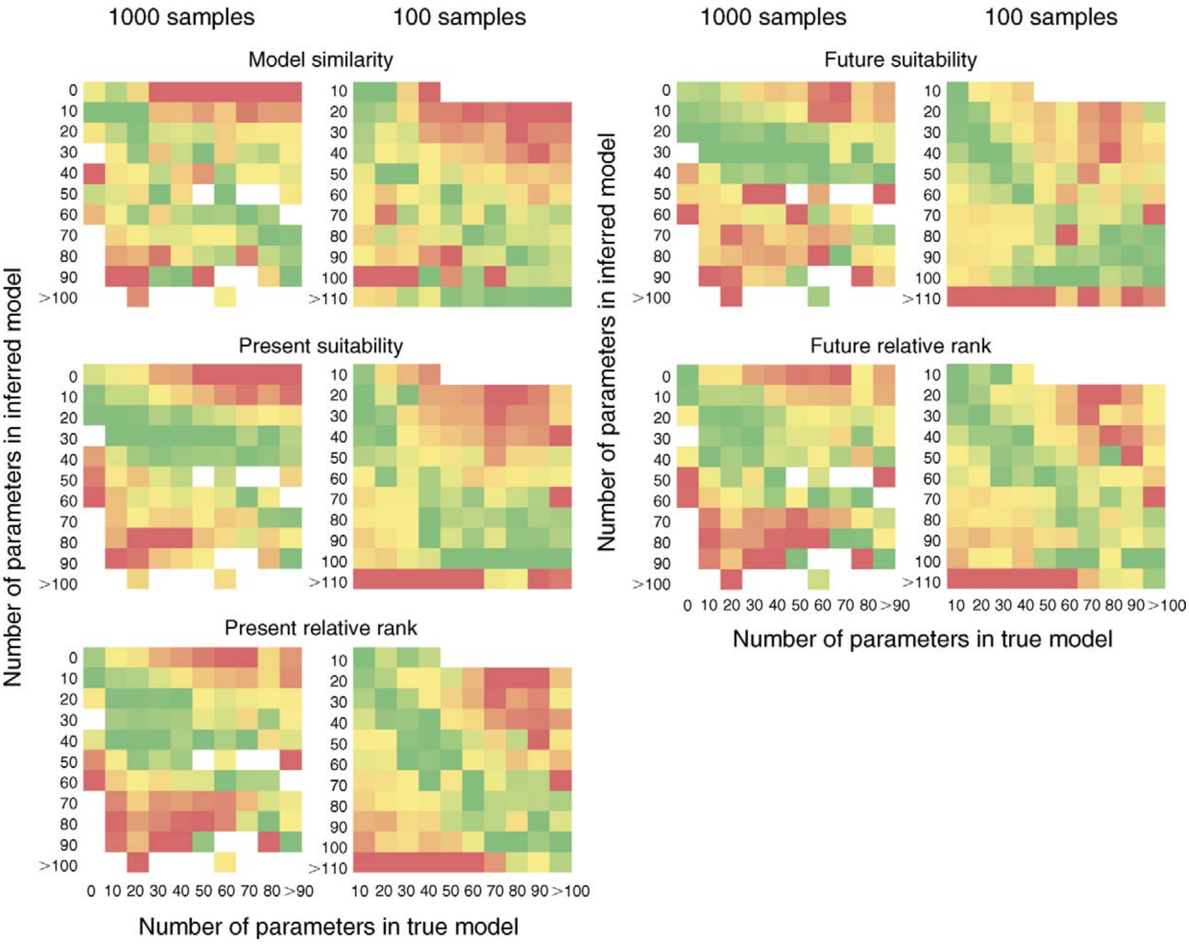

FIG. 1. Extent of over- or under-parameterization and model performance for 100 (left column) and 1000 (right) simulated occurrence points. Cooler colors indicate better performance. Color scales are calibrated individually for each column. In each comparison, the number of parameters in the true model is indicated on the x-axis while the number of parameters in the inferred model is given on the y-axis. The top left panel illustrates the ability of models to infer the relative importance of environmental variables. The second and third left panels demonstrate the effects of complexity on the ability to determine the true suitability and relative ranking of habitat in the present day, while the right panels show the effects of complexity on our ability to infer the distribution of suitable habitat in a different time period.

skepticism. We find that AIC_c exhibits the best average performance on selecting models that estimate the true model complexity when $N = 100$ (panel A), although not when $N = 1000$. In addition, models preferred by AIC_c more accurately estimate the relative importance of variables (panel B) as well as the suitability of habitat both in the training region (panels C and D) and when models are transferred to a different time period (panels E and F). The average performance of the information criteria (AIC_c, BIC) is slightly greater than that of AUC_{Test} and AUC_{Diff} in the larger data sets, but the difference is considerably greater when fewer data points are available, indicating that information criterion-based approaches to model selection may be particularly useful when sample sizes are small. In addition, we see that the amount of variation in the performance of the information criterion methods is much smaller than that seen in the AUC-based methods, particularly in the lower tail. This indicates that AIC_c and BIC are selecting the worst models from each analysis less often than are the AUC-based methods. In this study, model selection criteria were given a total of 1020 trials, selecting from a set of 10 models in each replicate. The worst-performing model selection criterion, AUC_{Train}, chose the default behavior of Maxent ($\beta = 1$) in all but ten replicates. Therefore the poor behavior of AUC_{Train} is approximately what would be seen if no model selection criterion was applied at all. Although AIC and AIC_c have been used for ENMs before (Hao et al. 2007, Dormann et al. 2008, Hengl et al. 2009), little information about their actual performance has been available. Here we demonstrate that they may make a valuable contribution to the toolbox of investigators using Maxent to model species distributions. Whether these benefits extend to other modeling methods is unknown.

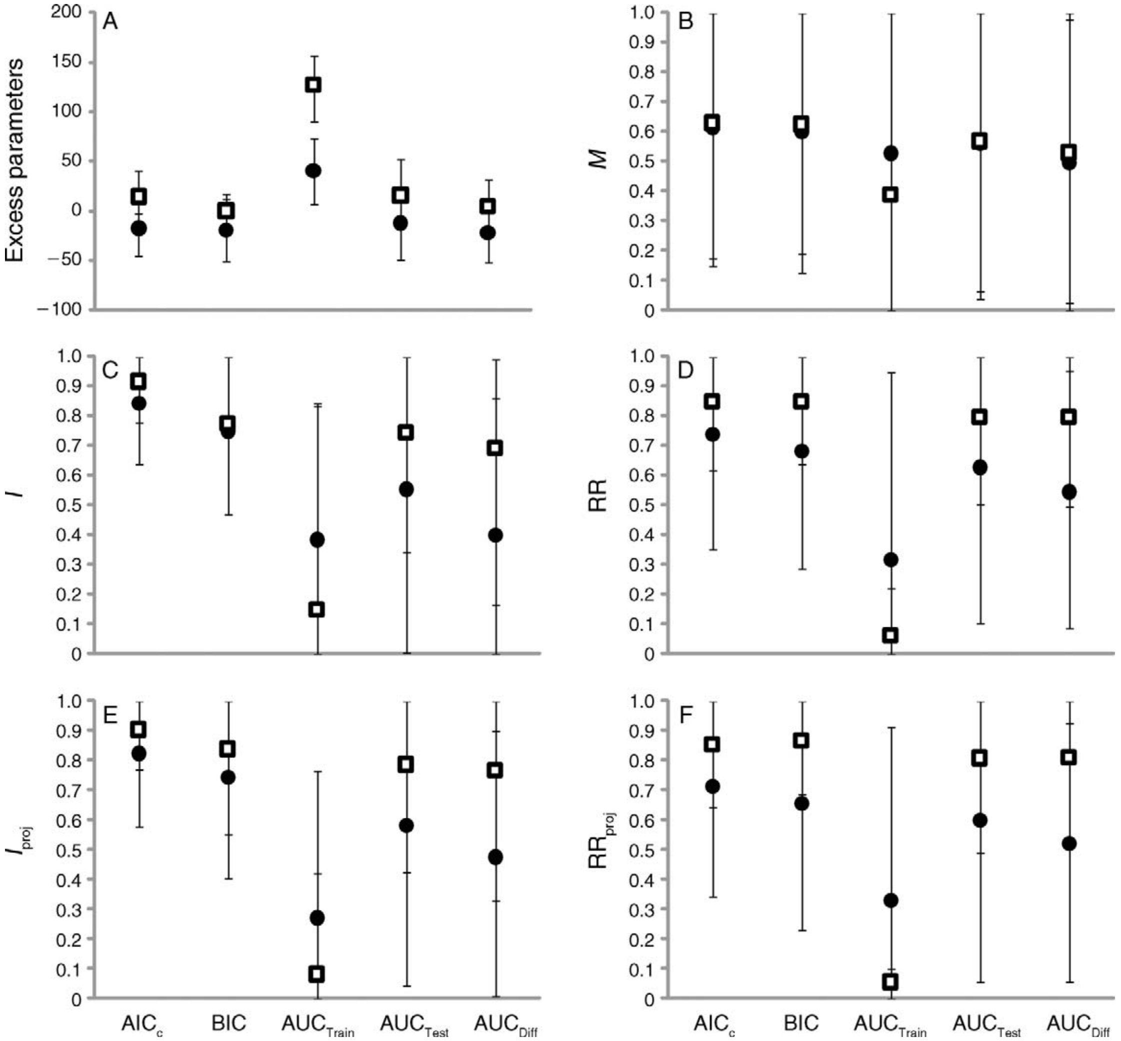

FIG. 2. Relative performance of model selection criteria for 1000 (open squares) and 100 (solid circles) simulated occurrence points. (A) The performance of each criterion in regard to over- or under-parameterization; (B–E) the performance of the selected models using a variety of metrics. Because "true" models were built for many different species using different levels of complexity, we must scale the performance of each selected model by the models that were available for that combination of species and complexity. For this reason, values on the y-axis in panels B–E are given as $[A - \min(A)]/[\max(A) - \min(A)]$, where A is the metric in question and $\max(A)$ and $\min(A)$ are the maximum and minimum values of that metric over all models constructed using that "true" model. In this formulation each model gets a score of 1 if it is the best performing of that set of models and a score of 0 if it is the worst. Metrics are: I, the ability of the model to estimate the true suitability of habitat; RR, the ability of the model to estimate the relative ranking of habitat patches; M, the ability of the Maxent model to determine the relative importance of environmental variables; I_{Proj} and RR_{Proj}, the effectiveness of Maxent at estimating true suitability or relative ranking of habitat in a different time period, respectively. See *Methods: Criteria for model selection* for descriptions of the criteria listed on the x-axes. Error bars indicate the 10th and 90th percentiles of the distribution of performance scores.

Transferring models between geographic regions or periods of time represents an additional set of challenges for ENM methods. The presence of combinations of climate variables in the projected region for which there is no analog in the training region necessitates extrapolation of those models into a set of conditions for which no presence or pseudo-absence data were available during model construction. Nevertheless, ENMs are frequently used to study the effects of climate change on habitat suitability or to estimate the ability of species to expand into new habitat. Using our simulation approach, we are able to assess the transferability of models in comparison to the true change in habitat suitability. We find that over-parameterized models tend to underestimate the availability of suitable habitat when transferred into a new time period, while under-parameterized models tend to overestimate it (Appendix C, middle row). However, the same models show identical behavior in the present day (Appendix C, top row), and the two cancel each other to a great extent so that the bias in the inferred level of change is comparatively minor (Appendix C, bottom row) although still statistically highly significant ($P < 0.001$). Although it is tempting to conclude that inferences of niche expansion and contraction as a function of climate change are therefore fairly robust to over- or under-parameterization, the magnitude of the effect is small only because the substantial biases caused by inappropriately complex or simple models mostly cancel each other out when subtracting present from future niche breadth. The actual predictions of habitat suitability, relative rank, and niche breadth in the future climate scenario are still of poor quality compared to models of appropriate complexity. It is only by comparison to similarly biased models in the training region that they appear to yield an approximately correct signal of environmental niche contraction or expansion. We note that the breadth metric used here (Levins 1968) relies on standardized suitability scores without the application of a threshold, and as such it may obscure changes in the average suitability of habitat over time. It is therefore possible that significant range expansions and contractions, or artifactual inferences of these phenomena, could be obscured by this method. Spurious inferences of range contraction or expansion may therefore be more widespread and severe than this study indicates.

Many investigators have treated model complexity in Maxent as unimportant, assuming that unnecessary parameters were of small effect and could safely be ignored. Our results demonstrate that model complexity affects model performance for many, if not all, applications. We find that over-parameterization is less problematic than under-parameterization in most cases. It is tempting to conclude that models should be allowed to be arbitrarily complex, but we suggest that the reduced penalty for over-parameterization is less important than the observation that models of appropriate complexity perform best. The one exception is when model complexity approaches or exceeds the number of occurrences available for model construction, in which case overly simplistic models perform better (Appendix B). We suggest that the effects of regularization on model structure and performance should always be evaluated, if for no other reason than that parsimony dictates that our models should be no more complicated than need be to explain our observations.

Ideally, investigators should attempt to construct models of approximately the true level of complexity. However, the very absence of this information is one of the primary reasons that ENMs are used: it is difficult to argue that they are superior to experimental physiological studies in any sense except for convenience and tractability. We therefore need criteria by which to select models of appropriate levels of complexity when the truth is unknown, and in this study we demonstrate that information criterion-based metrics perform well with Maxent ENMs.

Maxent (Phillips et al. 2006) is rapidly becoming one of the most widely used software packages for environmental niche modeling for a variety of good reasons. However, as with any analytical method that is easy to use, it is easy for users to become complacent in their modeling efforts and uncritically accept the models it produces without exploring the effects of the modeling process. To date, this has been the case with respect to model complexity for many users of Maxent. Here we demonstrate conclusively that model complexity affects users' ability to infer the suitability of habitat both with and without thresholds, the relative importance of environmental variables to determining species' distributions, estimates of the breadth of species' environmental niches, and the transferability of models. We also demonstrate that information-theoretic approaches to selecting models offer significant advantages over the methods that are commonly employed by Maxent users. Although it may be premature to suggest that these methods replace the more traditional model selection methods used with Maxent, the current study makes a clear case for their inclusion in the Maxent modeler's toolbox.

Acknowledgments

The authors thank the following people for comments on early drafts of this manuscript: Steven Phillips, A. Townsend Peterson, William Godsoe, Sam Veloz, Robert Anderson, Sean Maher, Sahotra Sarkar, Stavana Strutz, Ophelia Wang, Blake Sissel, Matthew Moskwik, Rich Glor, Robert Hijmans, and Michael Turelli. We also thank two anonymous reviewers for their helpful and rapid feedback. This work was funded by grants from the UCD Center for Population Biology, the California Department of Fish and Game, NSF postdoctoral fellowship DBI-0905701 to D. Warren, and NSF grant DEB-0815145.

Literature Cited

Akaike, H. 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19:716–723.

Breiman, L. 2001. Statistical modeling: the two cultures. Statistical Science 16:199–231.

Burnham, K. P., and D. R. Anderson. 2002. Model selection and multimodel inference: a practical information-theoretic approach. Second edition. Springer-Verlag, Berlin, Germany.

Davis, F. W., D. M. Stoms, A. D. Hollander, K. A. Thomas, P. A. Stine, D. Odion, M. I. Borchert, J. H. Thorne, M. V. Gray, R. E. Walker, K. Warner, and J. Graae. 1998. The California gap analysis project: final report. University of California, Santa Barbara, California, USA.

Dormann, C. F., O. Purschke, J. R. García Márquez, S. Lautenbach, and B. Schröder. 2008. Components of uncertainty in species distribution analysis: a case study of the Great Grey Shrike. Ecology 89:3371–3386.

ESRI. 2006. ArcGIS 9.2. Environmental Systems Research Institute, Redlands, California, USA.

Godsoe, W. 2010. I can't define the niche but I know it when I see it: a formal link between statistical theory and the ecological niche. Oikos 119:53–60.

Hampe, A. 2004. Bioclimatic models: what they detect and what they hide. Global Ecology and Biogeography 11:469–471.

Hao, C., C. LiJun, and T. P. Albright. 2007. Predicting the potential distribution of invasive exotic species using GIS and information-theoretic approaches: a case of ragweed (*Ambrosia artemisiifolia* L.) distribution in China. Chinese Science Bulletin 52:1223–1230.

Hengl, T., H. Sierdsema, A. Radovic, and A. Dilod. 2009. Spatial prediction of species' distributions from occurrence-only records: combining point pattern analysis, ENFA and regression-kriging. Ecological Modeling 24:3499–3511.

Hijmans, R. J., S. E. Cameron, J. L. Parra, P. G. Jones, and A. Jarvis. 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25:1965–1978.

Levins, R. 1968. Evolution in changing environments. Monographs in population biology. Volume 2. Princeton University Press, Princeton, New Jersey, USA.

Menke, S. B., D. A. Holway, R. N. Fisher, and W. Jetz. 2009. Characterizing and predicting species distributions across environments and scales: Argentine ant occurrences in the eye of the beholder. Global Ecology and Biogeography 18: 50–63.

Phillips, S. J., R. P. Anderson, and R. E. Schapire. 2006. Maximum entropy modeling of species geographic distributions. Ecological Modeling 190:231–259.

Phillips, S. J., and M. M. Dudík. 2008. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography 31:161–175.

Soberón, J., and A. T. Peterson. 2005. Interpretation of models of fundamental ecological niches and species' distributional areas. Biodiversity Informatics 2:1–10.

Veloz, S. 2009. Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. Journal of Biogeography 36:2290–2299.

Warren, D. L., R. E. Glor, and M. Turelli. 2008. Environmental niche equivalency versus conservatism: quantitative approaches to niche evolution. Evolution 62:2868–2883.

Warren, D. L., R. E. Glor, and M. Turelli. 2010. ENMTools: a toolbox for comparative studies of environmental niche models. Ecography. [doi: 10.1111/j.1600-0587.2009.06142.x]

APPENDIX A

Metrics of model performance (*Ecological Archives* A021-018-A1).

APPENDIX B

Parameters exceeding data points (*Ecological Archives* A021-018-A2).

APPENDIX C

Niche breadth and transferability (*Ecological Archives* A021-018-A3).

APPENDIX D

Simulation and testing process (*Ecological Archives* A021-018-A4).