Journal of Biogeography (J. Biogeogr.) (2007) **34**, 1455–1469

ORIGINAL ARTICLE

Predicting species distributions: a critical comparison of the most common statistical models using artificial species

Christine N. Meynard[1,2*] and James F. Quinn[1]

[1]*Department of Environmental Science and Policy, University of California, 1 Shields Avenue, Davis, CA 95616, USA and* [2]*Núcleo Científico Milenio FORECOS, 4 Piso Facultad de Ciencias Forestales, Universidad Austral de Chile, Casilla 567, Valdivia, Chile*

ABSTRACT

Aim To test statistical models used to predict species distributions under different shapes of occurrence–environment relationship. We addressed three questions: (1) Is there a statistical technique that has a consistently higher predictive ability than others for all kinds of relationships? (2) How does species prevalence influence the relative performance of models? (3) When an automated stepwise selection procedure is used, does it improve predictive modelling, and are the relevant variables being selected?

Location We used environmental data from a real landscape, the state of California, and simulated species distributions within this landscape.

Methods Eighteen artificial species were generated, which varied in their occurrence response to the environmental gradients considered (random, linear, Gaussian, threshold or mixed), in the interaction of those factors (no interaction vs. multiplicative), and on their prevalence (50% vs. 5%). The landscape was then randomly sampled with a large ($n = 2000$) or small ($n = 150$) sample size, and the predictive ability of each statistical approach was assessed by comparing the true and predicted distributions using five different indexes of performance (area under the receiver-operator characteristic curve, Kappa, correlation between true and predictive probability of occurrence, sensitivity and specificity). We compared generalized additive models (GAM) with and without flexible degrees of freedom, logistic regressions (general linear models, GLM) with and without variable selection, classification trees, and the genetic algorithm for rule-set production (GARP).

Results Species with threshold and mixed responses, additive environmental effects, and high prevalence generated better predictions than did other species for all statistical models. In general, GAM outperforms all other strategies, although differences with GLM are usually not significant. The two variable-selection strategies presented here did not discriminate successfully between truly causal factors and correlated environmental variables.

Main conclusions Based on our analyses, we recommend the use of GAM or GLM over classification trees or GARP, and the specification of any suspected interaction terms between predictors. An expert-based variable selection procedure was preferable to the automated procedures used here. Finally, for low-prevalence species, variability in model performance is both very high and sample-dependent. This suggests that distribution models for species with low prevalence can be improved through targeted sampling.

Keywords

Artificial species, classification trees, conservation biogeography, GAM, GARP, GLM, species-distribution modelling.

*Correspondence: Christine Meynard, Núcleo Científico, Milenio FORECOS, 4° piso Facultad de Ciencias Forestales, Universidad Austral de Chile, Casilla 567 Valdivia, Chile.
E-mail: christinemeynard@uach.cl

INTRODUCTION

In past decades, ecologists have used a variety of statistical techniques to predict species occurrences over broad geographical areas. In general, these models employ correlations between point-location data on species occurrences, and environmental predictors from GIS or other mapped data. These models have wide management applications in the context of conservation biology, biogeography and climate change studies (Guisan & Zimmermann, 2000; Guisan & Thuiller, 2005; Araújo & Rahbek, 2006). Despite the large amount of work on this topic, we typically lack knowledge about the mechanisms that drive species distributions (Gaston, 2003; Holt, 2003).

When modelling species distributions, predictors can be classified into three main categories (Guisan & Zimmermann, 2000): resource variables, which represent consumed matter or energy; direct gradients, which represent variables that have some physiological influence on organisms; and indirect gradients, which do not have a direct relationship to the species physiology but have a strong correlation to other direct or resource gradients and are easily measured. Dr Mike Austin has argued that using predictors that have a direct relationship to species responses is important to link theory and statistical modelling, and may also increase a model's predictive power (Austin *et al.*, 2006; Austin, 2007). For example, species abundances along an environmental gradient are predicted to be symmetrical or skewed bell-shaped responses under different theoretical frameworks (Austin, 2002, 2007). However, species distributions are often modelled using presence–absence data rather than abundance, because this information is easier to gather from museum collections and rapid field surveys (Latimer *et al.*, 2006). A further complication comes from translating theoretical predictions based on abundance responses (Austin *et al.*, 2006; Austin, 2007) into patterns of species presences and absences (He & Gaston, 2003; Hui *et al.*, 2006), especially if occurrence and abundance patterns respond by different mechanisms to the same environmental controls (Potts & Elith, 2006).

Numerous studies have compared the performance of statistical techniques to predict species distributions, resulting in a variety of recommendations regarding model use (Ferrier *et al.*, 2002; Segurado & Araújo, 2004; Elith *et al.*, 2006). Comparing models with real data poses several problems beyond the lack of knowledge on the empirical distributions discussed above. Interpreting data collected in the real world is often confounded by problems of differing species detectability (Boulinier *et al.*, 1998; Royle, 2004; Buckland *et al.*, 2005); differences in species prevalence (Segurado & Araújo, 2004); variability among observers and habitat types (Buckland *et al.*, 2005; Royle *et al.*, 2005); biased sampling intensity due to limited access of certain areas (Austin & Heyligers, 1989; Wessels *et al.*, 1998); effects of factors not considered in the modelling process (Guisan & Zimmermann, 2000; Austin, 2002); or effects of individuals actively aggregating in space (Keitt *et al.*, 2002; Lichstein *et al.*, 2002). Species with low prevalence impose a particular challenge due to the usually limited number of occurrence observations available, and the lower performance of most models when faced with small sample sizes (Bourg *et al.*, 2005; Nielsen *et al.*, 2005). Estimating the empirical performance of models to determine whether certain models perform better in some situations and not others requires (usually unattainable) knowledge of all of these issues.

In contrast, generating artificial data on species distributions provides the advantage of giving us perfect knowledge and control over the causal factors of interest (Hirzel *et al.*, 2001; Austin *et al.*, 2006). Previous simulation studies have generated artificial data on environmental gradients as well as abundance distributions for several species along a gradient, where the position of a species is constrained by others in the environmental space (Austin *et al.*, 2006 and references therein). Other studies have used one species type with a particular combination of occurrence–environment relationships to compare performance of different modelling strategies (Austin *et al.*, 1995; Hirzel *et al.*, 2001), while others focused on the effects of spatial autocorrelation and sampling design on model fit (Hirzel & Guisan, 2002; Reese *et al.*, 2005). In general, these simulations assumed additive effects between environmental gradients, suggesting that these predictors were mutually substitutable. However, in the real world, many species may present non-substitutable responses to a particular environmental gradient (Prasad *et al.*, 2006; Termansen *et al.*, 2006).

In this study we generated 18 artificial species, which differed in their response shapes (random, linear, Gaussian, threshold and mixed) to three environmental gradients, in the way these environmental variables interacted in determining species occurrences (additive vs. multiplicative), and in their prevalence (50% vs. 5%). To our knowledge, this is the first simulation study to address the effects of differing occurrence–environment relationships on model performance. We generated probabilities of occurrence, rather than abundance responses, using different spatially explicit, rule-based mechanisms on three direct environmental variables. Given the large number of modelling strategies available (e.g. Elith *et al.*, 2006), we restricted ourselves to comparing four commonly used methods (Guisan & Thuiller, 2005; Latimer *et al.*, 2006). We contrasted the simulated and modelled distributions, focusing on three questions: (1) Is there a statistical technique that has a consistently higher predictive ability than others for all kinds of occurrence–environment relationship? (2) How does species frequency of occurrence influence each model's relative performance? (3) When an automated stepwise selection procedure is used, does it improve predictive modelling, and are the relevant variables being selected?

MATERIALS AND METHODS

We started our modelling process by generating 18 artificial species that differed in their occurrence responses to three

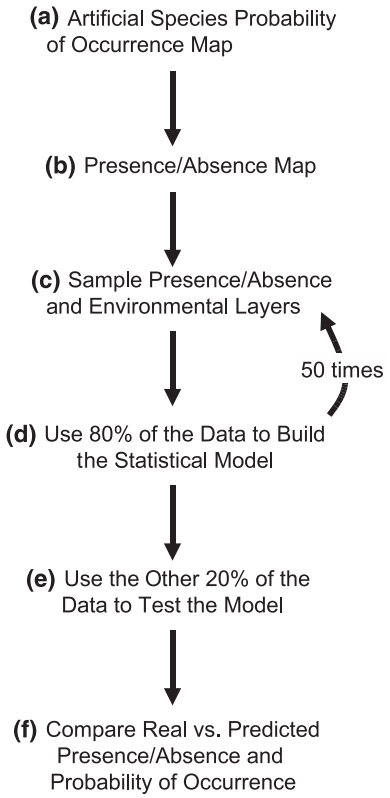

Figure 1 General modelling process. (a) We started by generating a map of probability of occurrence of an artificial species with known environment–occurrence relationships. (b) The probability of occurrence was translated into a presence and absence map. (c) We then sampled the landscape in 2000 or 150 random points, where we registered the nine environmental variables used in the statistical modelling building, as well as presences and absences and true probabilities of occurrence. This sampling was repeated 50 times, yielding 50 iterations per species. (d) From this sample, 80% of the points were used to create the statistical models and (e) the other 20% to test the model performance; (f) we could then compare the predictions of the models with the real probability of occurrence and presence/absence patterns.

environmental gradients: mean temperature, oxygen partial pressure, and net primary productivity (Fig. 1). Nine species were generated with a prevalence of 50%, and a second set of nine species were generated with a prevalence of 5%. We then simulated a statistical modelling process for those species occurrences using direct, indirect and unrelated variables (Table 1). By doing this, we assumed lack of knowledge about the relevant variables and their relationship to the species-occurrence patterns. In what follows, we explain the details of each step.

Simulation of species presences/absences

In total, we used 10 environmental variables for the state of California (Table 1). Three were used to generate the artificial species, and nine were used in the statistical model building process (Table 1). All the layers were clipped to a California coverage, and converted to a geographical projection raster with a resolution of 30 degree seconds (*c.* 1 km resolution). All necessary layers were then sampled, and each sample was exported into the R statistical package ver. 2.1.0. (R Development Core Team, 2005) for the statistical modelling. The rest of the geographical data processing was done using ArcGIS ver. 9.0.

Table 1 Identity of environmental gradients used in the modelling process.

Variable	Type	Modelling	Source
Annual mean temperature	Direct	Yes	Worldclim*
Oxygen partial pressure	Direct	No	Worldclim†
Net primary productivity	Direct	Yes	Modis‡
Altitude	Indirect	Yes	Worldclim*
Annual temperature range	Indirect	Yes	Worldclim*
Annual precipitation	Unrelated	Yes	Worldclim*
Precipitation seasonality	Unrelated	Yes	Worldclim*
Temperature seasonality	Indirect	Yes	Worldclim*
Normalized difference vegetation index	Indirect	Yes	Modis‡
Presence of forest	Unrelated	Yes	GAP Analysis§

'Type' refers to whether the variable was used to generate the artificial species (direct), was highly correlated with direct variables (indirect), or was unrelated to direct variables (unrelated). 'Modelling' indicates whether or not the variable was used as a predictor when building the statistical models.

*Climatic variables for California, downloaded from Worldclim (Hijmans *et al.*, 2004):http://www.worldclim.org.

†Generated from altitude using a standard atmosphere model (West, 1996).

‡Available at http://modis.gsfc.nasa.gov, described by Running *et al.* (2004).

§Land cover extracted from the California GAP Analysis and converted to a forest/non-forest layer.

We used data from three empirical environmental variables to generate rule-based species distributions: annual mean temperature (mtemp), oxygen partial pressure (pO_2) and estimated net primary productivity (NPP). These three environmental factors have been widely described as having a direct effect on the ecophysiology and habitat selection of many vertebrate species (Prosser & Brown, 1961; McNab, 1980; Spicer & Gaston, 1999). Surrogate variables such as altitude are often indirectly related to species distributions through causal mechanisms that involve their relationship to physiological constraints (thermoregulation and oxygen consumption at higher elevation) and productivity (food and water availability, quantity and quality of resources) (Guisan & Zimmermann, 2000). Altitude could, of course, have been used instead as a distribution-generating variable, and would undoubtedly have led to qualitatively similar results. We chose oxygen partial pressure as having a direct mechanistic meaning to large mobile animals that motivate this study, while altitude was used in the statistical model fitting as an indirect variable that is strongly correlated to the real causal gradients (see Guisan & Zimmermann, 2000 for a discussion of elevation as a

surrogate variable). Similarly, we used an estimate of primary productivity generated from the normalized difference vegetation index (NDVI), rather than NDVI itself.

For each virtual species, the probability of occurrence, P_{occ}, was calculated as:

$$P_{occ} = f(\text{mtemp}, pO_2, \text{NPP}) + \varepsilon \quad (1)$$

where f (mtemp, pO_2, NPP) represents a function of three environmental factors (annual mean temperature, oxygen partial pressure and net primary productivity, respectively), and ε represents a normally distributed random error. We generated two kinds of species with respect to the relationship between the three environmental variables considered. In the first species type, which we call additive species, the probability of occurrence was calculated as the sum of the independent probabilities of occurrence given each one of the environmental factors, a method comparable with that used by Hirzel *et al.* (2001). This additive species mimics a case in which the three environmental factors considered are substitutable and independent. In other words, if conditions regarding one factor are poor, this could be compensated by the other two factors being favourable:

$$P_{occ} = f_1(\text{mtemp}) + f_2(pO_2) + f_3(\text{NPP}) + \varepsilon \quad (2)$$

The second kind of species, which we call multiplicative species, show a multiplicative effect between the three environmental variables considered:

$$P_{occ} = f_1(\text{mtemp}) \times f_2(pO_2) \times f_3(\text{NPP}) + \varepsilon \quad (3)$$

We can think about this model species as one in which there is interaction between environmental factors. Here the three environmental variables are essential and not replaceable for the species subsistence. If one of the factors is very unfavourable at a site, the species will have a low probability of occurrence even though the other two factors may be near the species optimum. This may be a more realistic assumption for many species and their limiting resources, since it is likely that many environmental factors interact in determining species-occurrence patterns (Prasad *et al.*, 2006).

The functions f_1, f_2 and f_3 determine a suitability score for each environmental variable, and represent the particular shape of the occurrence–environment relationship for each environmental layer. To make sure each factor had the same weight in the artificial species occurrence, each function was rescaled to have the same range of values, and the final probability of occurrence was rescaled between 0 and 1 by using a linear transformation between the unconstrained and constrained values. We preferred this transformation to a logit, for two reasons. First, a linear transformation will preserve the shape of the relationships between occurrence and the environmental gradient originally created. Second, the generalized models that we tested later use a logit-link function to constrain probability values (see below), and we wanted to avoid biasing our analysis by creating species that behaved in an ideal way for a subset of the models tested.

More complex variations on both additive and multiplicative species are readily constructed, and it might be reasonable to do so for particular taxa where there are sound mechanistic reasons to infer particular environmental controls on species distributions (Tilman, 1980). Here we limit ourselves to simple models, which may be viewed as limiting cases, to study how multiple causal factors with and without interactions affect the performance of our statistical estimation procedures. However, statistical models are likely to behave in a similar fashion, given a different set of direct and indirect predictors combined in a similar way. Therefore we see no reason to think our findings are specific either to the California landscape or to a particular taxon.

Shape of the occurrence–environment relationship

The species responses to the environmental factors considered (f_1, f_2 and f_3 in equations 2 and 3) were allowed to take one of three shapes: Gaussian, linear or threshold. A Gaussian species responded to the particular environmental factor considered by having a symmetrical and decreasing probability of occurrence around an optimum value, while a linear response was characterized by a steady increase or decrease in the probability of occurrence. A threshold response was characterized by a simple 'all-or-nothing' response to a threshold in the environmental gradient.

In pure species, the shape of f_1, f_2 and f_3 was the same for all three factors. In addition, we created a mixed species, which had a Gaussian response to mean temperature (f_1), a threshold response to oxygen partial pressure (f_2), and a linear relationship to NPP (f_3) (Table 2). We also generated a species that had a random distribution with respect to the environment. However, this species resulted in a model performance that was not significantly different from a random prediction for all statistical models compared, and is therefore excluded from any further discussion.

The combination of different shapes of the occurrence–environment relationships and the additive/multiplicative species results in eight different types of simulated species, here called additive-linear, additive-Gaussian, additive-threshold, additive-mixed, multiplicative-linear, multiplicative-Gaussian, multiplicative-threshold and multiplicative-mixed.

Generating species presences and absences

First, California was divided into a regular grid populated with random numbers drawn from a uniform distribution. If the random number in a particular cell was smaller than the probability of occurrence for that cell, the species was considered present. For example, if the probability of occurrence in a particular cell is equal to 0.8, and the range of the uniform random number is 0–1, 80% of the time the random number will be < 0.8. The desired species prevalence can be adjusted by modifying the range of the probabilities of occurrence generated.

Table 2 Different types of artificial species generated.

Feature	Type
Shape of environment–occurrence relationships	Linear
	Gaussian
	Threshold
Interaction between factors	No interactions: environmental gradients are independent – these are 'additive species'
	Multiplicative effects between environmental gradients
Combination of relationships	None, all environmental gradients have the same type of relationship – these are 'pure species'
	Mixed: the occurrence–environment relationship is linear for net primary productivity, Gaussian for mean annual temperature and threshold for oxygen partial pressure
Species prevalence	50%, high-prevalence species
	5%, low-prevalence species

Eight species types were generated with respect to occurrence–environment relationships: additive-linear, additive-Gaussian, additive-threshold, additive-mixed, multiplicative-linear, multiplicative-Gaussian, multiplicative-threshold, multiplicative-mixed. The same species types were generated with two levels of prevalence. Two additional species were randomly distributed across the landscape, with low and high prevalence.

We started by modelling each type of species with a prevalence of 0.5. We tested the statistical approaches by sampling 2000 random locations across California and generating predictions of probabilities of occurrence and presence and absence that we could then compare with their true values. We then generated the same species types with a prevalence of 0.05. In this case, we tested the statistical models using a large sample size (2000 random samples), and a small sample size (150 random samples). An example is shown in Fig. 2.

Statistical models to be compared

Six statistical models were used and compared to predict species distributions: logistic regression (general linear models, GLM), generalized additive models (GAM), classification trees, the genetic algorithm for rule-set production (GARP), logistic regression with a stepwise variable selection (sGLM), and generalized additive models with a flexible smooth term and flexible degrees of freedom (sGAM). Table 3 summarizes the basic differences between these statistical modelling strategies, and more detailed descriptions can be found elsewhere (McCullagh & Nelder, 1989; Hastie & Tibshirani, 1990; Stockwell, 1999; Venables & Ripley, 2002). In all analyses, each predictor was included both untransformed and squared in order to consider quadratic relationships between predictors and species probability of occurrence. The one exception was the forest layer, and all variables for the GAM analysis. No interaction terms between variables were included.

We ran the GLM with a logit-link function and binomial distribution to model presence/absence data with all nine predictors included in the modelling process. For sGLM, a variable selection based on Akaike's information criterion (Burnham & Anderson, 2002) was used to reduce the number of predictors further. To implement GAM, we also assumed a binomial distribution and logit link with nine predictors. The *mgcv* library within the statistical package R was used to create the GAM models, with 4 d.f. on all variables except the forest layer, which was modelled as a linear relationship. A second model strategy, described by Wood (2004), involves allowing for some flexibility in the degrees of freedom of GAM for each variable included in the model. In this approach, which we call sGAM, the algorithm runs several iterations in which it chooses the lower level of complexity for each variable that optimizes model fitting by minimizing the unbiased risk estimator (UBRE), a criterion similar to AIC (Venables & Ripley, 2002; Wood, 2004). It also reduces the importance of irrelevant variables by changing its effective degrees of freedom. Classification trees were implemented with the

Table 3 Summary of basic assumptions in each statistical modelling strategy.

Model	Assumptions
GLM	Additive and linear relationship between predictors
	Logit relationship between predictors and response (probability of occurrence)
	No variable selection
sGLM	Same as GLM, but with variable selection – models with lower AIC values are selected in a stepwise procedure
GAM	Additive relationship between predictors, smooth (nonlinear) terms are allowed with 4 d.f. for continuous variables
	Logit relationship between sum of predictors and response (probability of occurrence)
sGAM	Same as GAM, but with variation in complexity of smooth terms – models with a lower UBRE value (similar to AIC) will be preferred, irrelevant variables will have little importance in the model (Wood & Augustin, 2002)
Classification trees	No assumption on shape of relationships between predictors
	Predicts a hierarchical threshold response
	No variable selection, but more important variables will appear higher in the hierarchy of the tree
GARP	Combination of all previous strategies (and assumptions) according to model performance: envelope rules (similar to classification trees), logit rules (similar to GLM and GAM), negation rules

GLM, logistic regression (general linear models); sGLM, logistic regression with a stepwise variable selection; GAM, generalized additive models; sGAM, generalized additive models with a flexible smooth term and flexible degrees of freedom; GARP, genetic algorithm for rule-set production.

function *rpart* in the R statistical package (Venables & Ripley, 2002). This function grows the trees by partitioning the data into sequentially homogenous groups, as described by Breiman *et al.* (1984). Then a pruning method based on minimizing a cost–complexity measure is applied to reduce the number of leaves in the tree to an optimal size to minimize overfitting (Breiman *et al.*, 1984; Venables & Ripley, 2002). Finally, GARP is a computer-intensive method that can use a variety of techniques and combinations of variables to find the best empirical predictions. Here we used DESKTOP GARP ver. 1.1.6. (Stockwell, 1999), although there is now a newer version available through OPENMODELLER (Elith *et al.*, 2006). Each run results in a presence-and-absence map, and the probability of occurrence is calculated by averaging all the rasters of presences and absences generated for each case. Due to computer and timing constraints, we ran GARP using the default options of choosing one rule type at a time, and 20 runs per case.

Assessment of model performance

For each artificial species generated, we took 50 equally sized random samples of the landscape. For each sample, we generated a statistical model of species distribution using 80% of the data (training set). Each statistical model resulted in a prediction of probability of occurrence which was used in the evaluation of the model. We used the other 20% of the data to test the model (testing set), mimicking a data-splitting strategy often used to evaluate distribution models applied to real species (Fig. 1). As both sampling instances were totally random (e.g. there is no bias in the sampling process, as there would be in the real world), and represented a small proportion of a large area (< 0.5% of available grids), this is equivalent to having two independent data sets to build the models on one hand, and test them on the other: within the subsamples there was no spatial autocorrelation in the environmental gradients used (Moran's I, $P > 0.05$), and there was no temporal autocorrelation involved as our species distributions are static. However, this data-splitting approach has some other limitations that are considered in the Discussion.

Five measures of model performance were used. The area under the receiver–operator characteristic curve (AUC) provides several advantages over other performance indices, as it is less sensitive to species prevalence and it measures overall model performance (Zweig & Campbell, 1993; Fielding & Bell, 1997; Webb & Ming Ting, 2005). The maximum Kappa value is used to assess the improvement over chance given by the predictive model (Fielding & Bell, 1997). We also calculated the Pearson correlation value between the true and predicted probability of occurrence. Finally, on occasions it might be useful to know if the models are predicting the presences or absences more successfully (Fleishman *et al.*, 2001; Bulluck *et al.*, 2006). For this reason, we also calculated sensitivity and specificity for each model. Sensitivity is the proportion of presences that are predicted by the model and that are in fact presences, and specificity is the proportion of true absences (Fielding & Bell, 1997). These were calculated based on the probability threshold that corresponded to the maximum value of Kappa (Liu *et al.*, 2005). This method has been used frequently to predict species presence due to its simplicity, although methods based on the receiver–operator characteristic curve and cost–benefit analysis have been recommended over Kappa-based methods (Liu *et al.*, 2005).

To explore whether different modelling strategies showed better predictive ability with particular species types, we used a two-way ANOVA with species type and statistical model as factors. Tukey's honestly significant difference (Tukey's HSD) test for multiple comparisons (Steel *et al.*, 1997) was used to look for significant differences between groups.

RESULTS

Large sample size, high species prevalence

A two-way ANOVA shows that both the main effects and the interaction terms are significant for all measures of performance ($P < 0.05$). Among the main effects, an important result is that GAM, GLM and sGLM do not differ significantly among themselves in terms of the AUC, Kappa, specificity or sensitivity. sGAM presents higher performance values with regard to different indices, but does not differ significantly from GAM, except for Kappa and the correlation value (Fig. 3; Table 4).

Classification trees appear to perform particularly well in predicting species presences, as shown by a higher sensitivity (Fig. 3). However, classification trees are often poor predictors of presences and absences combined, as shown by all other indices of performance. The only exception is for the additive-threshold case, where classification trees perform significantly better than any other statistical model. GARP provides particularly poor predictions of the probability of occurrence, but otherwise behaves similarly to the classification trees for other indices of performance (Fig. 3; Table 4). GAM significantly outperforms the logistic regression approach (GLM) only for the additive and multiplicative Gaussian species and for the multiplicative-mixed species (Tukey's HSD, $P < 0.05$). In all other cases, although GAM performs slightly better than GLM, this difference is not statistically significant (Tukey's HSD, $P > 0.05$).

Surprisingly, all statistical models show a similar pattern, in that they tend to perform better for the same set of species. For each additive species, the presence and absence predictive ability (AUC and Kappa indices) is higher than for its multiplicative counterpart (Fig. 4). In other words, all statistical models provide better overall predictions when they are presented with additive effects rather than multiplicative effects, especially when the species has a threshold or mixed response to the environmental gradients. On the contrary, sensitivity shows higher values for multiplicative species, suggesting that all models tend to estimate presences better than absences when the overall predictive performance is low,

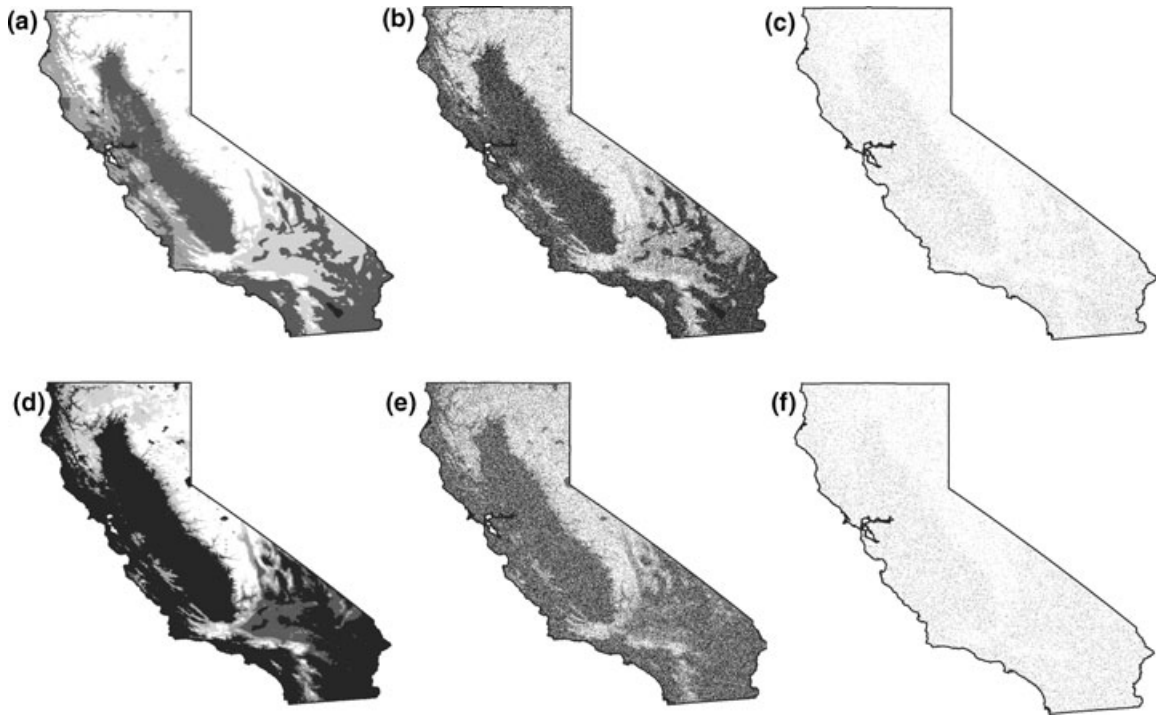

Figure 2 Example of a species-distribution simulation. The upper row of graphs represents an additive-mixed species: (a) probability of occurrence; (b) simulated presence for the species with 50% prevalence; (c) presence with 5% prevalence. The lower row represents a multiplicative-mixed species: (d) probability of occurrence; (e) simulated presence with 50% prevalence; (f) simulated presence with 5% prevalence. Dark grey in (a,d) represents areas of higher suitability.

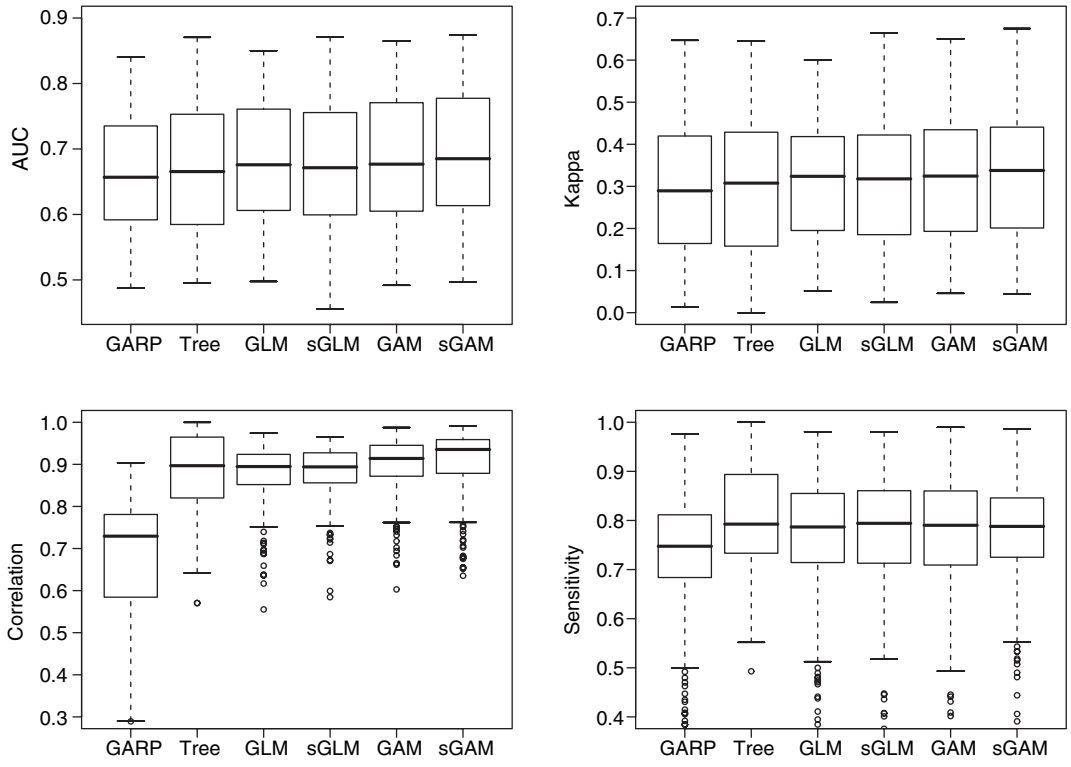

Figure 3 Predictive performance of the different modelling strategies compared when using a sample size of 2000. For simplicity, all species types are grouped together. A two-way ANOVA with species type and statistical models as factors shows significant differences in the main effects and their interactions. See Table 4 for statistically significant differences between groups.

Table 4 Results from a two-way ANOVA with two factors (model type and species type) were significant ($P < 0.001$) for AUC, Kappa, correlation and sensitivity.

	Significance of model comparisons					
Index	GLM	sGLM	GAM	sGAM	Tree	GARP
AUC						
GLM		ns	ns	**	***	***
sGLM	ns		ns	***	***	***
GAM	ns	ns		‘	***	***
sGAM	**	***	‘		***	***
Tree	***	***	***	***		ns
GARP	***	***	***	***	ns	
Kappa						
GLM		ns	ns	***	***	***
sGLM	ns		ns	***	***	**
GAM	ns	ns		**	***	***
sGAM	***	***	**		***	***
Tree	***	***	***	***		ns
GARP	***	**	***	***	ns	
Correlation						
GLM		ns	***	***	ns	***
SGLM	ns		***	***	ns	***
GAM	***	***		*	*	***
sGAM	***	***	*		***	***
Tree	ns	ns	*	***		***
GARP	***	***	***	***	***	
Sensitivity						
GLM		ns	ns	ns	***	***
SGLM	ns		ns	ns	***	***
GAM	ns	ns		ns	***	***
SGAM	ns	ns	ns		***	***
Tree	***	***	***	***		***
GARP	***	***	***	***	***	

Tukey's honestly significant difference (Tukey's HSD) tests for multiple comparisons were carried out for all measures of model performance between all models. Significance levels: ns = non-significant; ‘, $0.10 < P < 0.5$; *$P < 0.5$; **$P < 0.01$; ***$P < 0.001$.

GLM, logistic regression (general linear models); sGLM, logistic regression with a stepwise variable selection; GAM, generalized additive models; sGAM, generalized additive models with a flexible smooth term and flexible degrees of freedom; Tree, Classification trees; GARP, genetic algorithm for rule-set production.

therefore overestimating species range sizes (Fig. 4). Presences and absences of species with linear relationships are particularly poorly modelled (Fig. 4; Table S1 in Supplementary Material).

Large and small sample size, low species prevalence

When examining species with 5% prevalence and with a large sample size (2000 samples), results are similar to those for species with 50% prevalence. As in the previous case, a two-way ANOVA shows highly significant differences in the main terms (statistical model and species type) and their interaction (Table S2). However, we can note a few important differences. First, all models decreased their performance in every measure compared with the species with the larger prevalence (Fig. 5). Second, results of different performance indices became more variable, and differences between modelling strategies became less significant. Classification trees are heavily affected by the lower prevalence, and perform worse than any other model regardless of the species type or index considered. However, classification trees typically perform better in terms of sensitivity than the other models. In other words, when the species is infrequent, classification trees tend to have good predictions of the species presence, but very poor predictions of species absences, systematically overpredicting the species occupancy range.

Another important difference compared with the high-prevalence species is that GARP performed in a similar way to GLM, sGLM, GAM and sGAM in terms of AUC and the Pearson correlation between true and predicted probability of occurrence. In fact, it outperforms GAM and sGAM in the additive-Gaussian, multiplicative-Gaussian and multiplicative-linear species in terms of AUC and sensitivity (Tukey's HSD, $P < 0.05$). However, when we examine Kappa, specificity and the correlation with the true probability of occurrence, GLM and GAM still outperform GARP in all species types.

With a small sample size and species of low prevalence, these tendencies are only accentuated (Table S3), although the general trends regarding models' relative performance are the same. A species prevalence of 5% on a sample size of 2000 will generate, on average, 100 presences, while on a sample size of 150 it will generate only 7.5 presences. Model performance on every index becomes more variable when species prevalence decreases from 50% to 5%, so much so that all statistical models generate equally poor predictions that are, on average, not significantly different from a random prediction (AUC = 0.5). However, GAM and GARP generate the best predictions in all species types, with GAM usually outperforming GARP in all indexes except the correlation with the true probability of occurrence in the additive-Gaussian and multiplicative-Gaussian cases. This suggests that, under such conditions, the performance of each particular model will be highly dependent on the sample, some of them resulting in very good predictions, and some others in very poor predictions. When the expected number of occupied locations is this low (e.g. < 10), the average behaviour of the models is unsatisfactory in all cases.

Variable selection

For simplicity, we present here only the results for low-prevalence species (e.g. prevalence of 5%), although results for the high-prevalence species are qualitatively similar. The three variables used to generate mechanistically the distributions of the artificial species (altitude through pO_2, mean temperature and NPP) are identified by standard statistical methods as significant empirical predictors of those distributions in fewer than 50% of the iterations (Fig. 6).

When using sGAM with a large sample size, the three variables used to build the artificial species are identified as significant

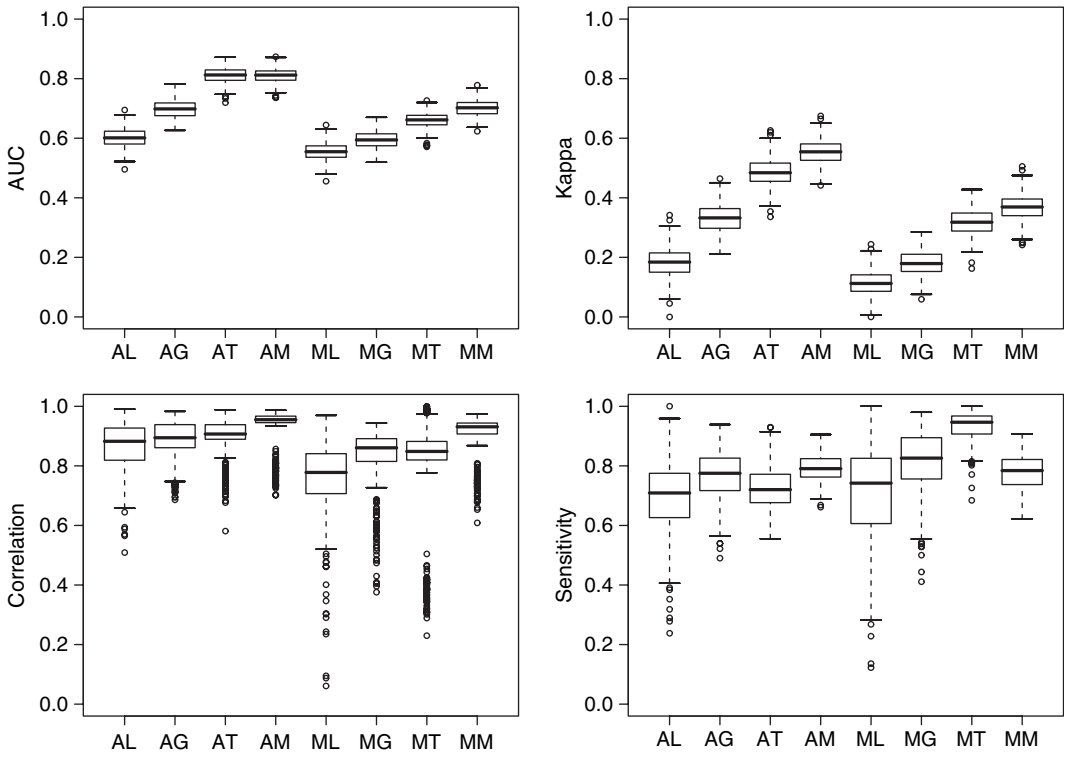

Figure 4 Predictive performance of the different species types for all statistical models used with a sample size of 2000 and species of 50% prevalence. Significant differences between groups can be seen in Table S3. AL, additive-linear; AG, additive-Gaussian; AT, additive-threshold; AM, additive-mixed; ML, multiplicative-linear; MG, multiplicative-Gaussian; MT, multiplicative-threshold; MM, multiplicative-mixed.

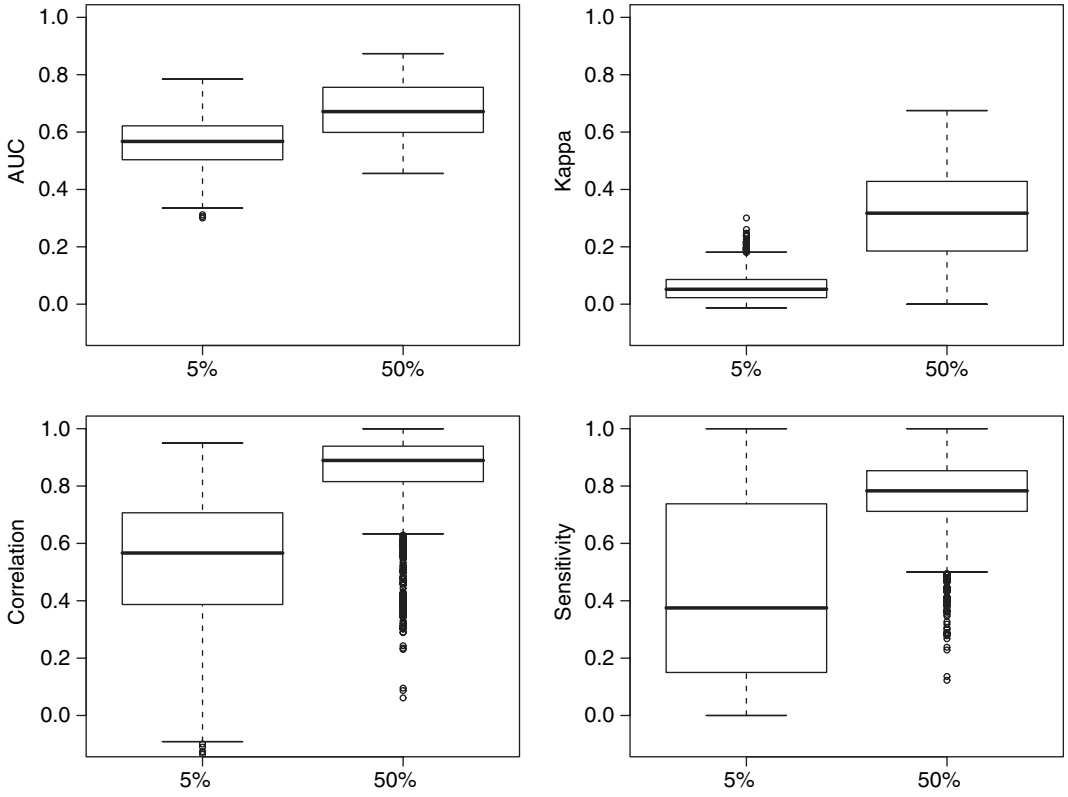

Figure 5 Comparison of model performance for common (prevalence = 50%) and rare (prevalence = 5%) species with a large sample size ($n = 2000$). Model predictive performance is significantly higher for all indices for the species with higher prevalence (Tukey's HSD, $P < 0.05$).

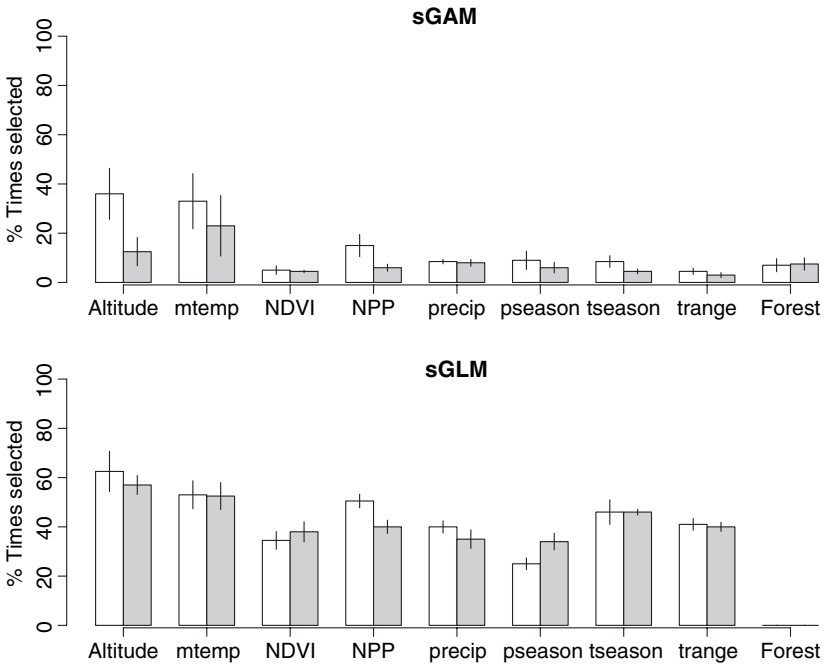

Figure 6 Variable selection for sGAM and sGLM with a sample size of 2000. Top panel shows percentage of iterations in which a particular variable was selected as significant ($P < 0.05$) in sGAM. Bottom panel shows percentage of iterations in which each variable was selected by the AIC criterion in sGLM. mtemp, annual mean temperature; NDVI, normalized difference vegetation index; NPP, net primary productivity; pseason, precipitation seasonality; tseason, temperature seasonality; trange, yearly temperature range; precip, annual mean precipitation.

from 6% to 36% of the time, while other indirect and unrelated variables were identified as significant in < 10% of the cases. Multiplicative species result in poor variable selection (Fig. 6), with only the mixed species showing significance of altitude and mean temperature in slightly more than 25% of the iterations. Additive species, especially for threshold and mixed responses, result in the direct variables used to generate the occurrence patterns being selected more often, compared with other species types. When the sample size is small, the success at selecting relevant variables is even lower, with NDVI and precipitation being significant more often than the other variables (*c.* 20% of the time for most cases), and altitude, mean temperature and NPP being significant < 10% of the time.

When using sGLM and a large sample size, altitude and mean temperature are selected as important variables more often than NPP. With additive species, these three variables are selected in > 50% of the iterations. However, other variables such as temperature seasonality and range are also selected in *c.* 40% of the cases (Fig. 6). The variable selection performs better for threshold and mixed species, and worse for linear species. Multiplicative species show a lower performance of the variable selection procedure as many other variables are selected as frequently as NPP, and only slightly less frequently than altitude and mean temperature. When the sample size is small, all variables, including altitude, mean temperature and NPP, are selected > 50% of the time. Thus the stepwise selection procedure shows little discrimination ability between causal and merely correlated data on candidates for predictor variables.

DISCUSSION

Our modelling framework examines the applicability of a number of the most widely used models to species with intermediate and low prevalence, the distributions of which are largely set by habitat and the physical environment in a controlled setting. From the results presented here, we can draw several recommendations for future use of these statistical techniques to predict species distributions. First, GAM and GLM performed better overall than classification trees and GARP. As with other comparative studies (Franklin, 1998; Elith & Burgman, 2002; Thuiller *et al.*, 2003; Segurado & Araújo, 2004), these two modelling strategies seem to represent a good trade-off between model complexity and performance for a diverse set of species. We found that, despite the differences in model assumptions, all statistical approaches seem to provide the best predictions for additive species, in particular those that respond to a threshold in the environmental gradients or that have a mixed functional response. Therefore, if there is mechanistic evidence of nonlinear effects and strong interactions, specifying that the model includes those variables and interactions should improve the predictive power of the model. We found the worst predictions when the species had a linear relationship to the environmental gradient, in particular in multiplicative species. This could be due to the structure of the modelling strategies *per se*. The logit-link function used in GAM and GLM is a nonlinear function, and the grouping strategy used in classification trees represents better a threshold response. GARP, being essentially a combination of the regression (e.g. GLM) and envelope rules, would also respond better to nonlinear and threshold cases.

We show here that GAM and sGAM perform better overall than GLM, sGLM, classification trees and GARP, although these differences are not significant when compared with GLM. However, in our modelling strategy, both the training and testing data sets are taken from the same geographical area (e.g. California). This validation procedure is often used, but has some drawbacks. Environmental layers in both data

sets have the same structure and relationships, and therefore these results may not be transferable to a different geographical region (Araújo *et al.*, 2005; Randin *et al.*, 2006). In other words, it is possible that when the testing set is taken from an area with a different environmental structure, the relative performances of the models changes. Randin *et al.* (2006), for example, found that GAM performed better than GLM when the testing data set was a subset of the data in the same geographical region as was used to generate the statistical model. However, when the testing data set was taken from a different region, simpler models such as GLM tend to perform better (Randin *et al.*, 2006). This difference could be due to more flexible models, such as GAM, overfitting the data. This may not be an issue when the interest in modelling species distribution comes from characterizing present-day distributions. Guisan *et al.* (2006) for example used species-distribution modelling to identify efficient sampling strategies for rare species within their potential ranges. This is highly relevant to generating sound conservation strategies for endangered species, but it requires using a model strategy that works best for the region of study, the same one as is used to generate the initial statistical model. However, when studying climate change scenarios, more flexible models, such as GAM, may perform poorly in predicting distribution shifts, due to their reduced performance under new environmental conditions (Araújo *et al.*, 2005; Araújo & Rahbek, 2006; Randin *et al.*, 2006).

Regarding variable selection, sGAM and sGLM did not improve predictions with respect to GAM and GLM. sGAM may provide some improvement over GAM for some indexes of performance, especially for the most common species, and would therefore be preferable over sGLM. However, sGAM and sGLM failed to discriminate mechanistically relevant variables from correlated environmental factors. This limits the interpretation of selected variables as biologically significant, and suggests that an expert-based selection of potentially relevant variables would be preferable to the variable selection procedures presented here. Other authors have reached similar conclusions for several automated stepwise variable selection methods (Derksen & Keselman, 1992; Maggini *et al.*, 2006; Segurado *et al.*, 2006). These results are consistent with previous empirical studies showing that an implementation of GAM with flexible degrees of freedom, equivalent to sGAM here, performs similarly to other traditional methods (e.g. GAM, GLM) in terms of predictive ability, and fails to select the meaningful model factors into the models (Elith *et al.*, 2006; Maggini *et al.*, 2006).

In our modelling framework, classification trees perform poorly and tend to overpredict the area of occupancy, especially when the species has low prevalence. However, this technique presents some advantages when the species is common, and responds to a threshold in the environmental gradients. Given that some empirical studies have shown a better performance of classification trees compared with GLM and GAM in some of the species modelled (Franklin, 1998; Segurado & Araújo, 2004), it is worth asking whether these species respond to thresholds in the environmental gradients, displaying an on–off response rather than continuous variations, for which GLM and GAM seem more effective. This could have important consequences in the context of climate change and habitat fragmentation, as a threshold response would create abrupt changes in species distributions as opposed to gradual shrinkages as climate and landscapes change in an unfavourable manner. Newer tree-based methods may increase the potential of classification trees. For example, Maggini *et al.* (2006) apply classification trees on residuals of a regression to identify interactions between predictors. This method takes advantage of the properties of both modelling strategies, and may outperform any one of them applied separately (Maggini *et al.*, 2006). Prasad *et al.* (2006) proposed to combine different tree-based methods in order to get the best predictions with newer techniques, but interpret the results with the more traditional ones.

The fact that all strategies performed poorly when the species being modelled was uncommon is not surprising, as it has been shown previously that sample size and number of presence records greatly influence model performance (McPherson *et al.*, 2004; Reese *et al.*, 2005; Guisan *et al.*, 2006). McPherson *et al.* (2004) showed that this could be an effect of sample prevalence rather than species prevalence. To avoid this problem, they suggest weighting absences by the ratio of number of presences to number of absences. Even by doing so, the authors show that the variance of performance as measured by different indices increases when the species is rare or very common (prevalence < 0.20 or > 0.75). This is consistent with our findings that different models generate better predictions (on average and with less variance) with species of 50% prevalence vs. 5% prevalence.

Some recent studies have shown that DESKTOP GARP performs poorly when compared with simpler and computationally less intensive methods, and that a newer version available through OPENMODELLER (GARP 3.0) generates better results than DESKTOP GARP 1.1.6, used here (Elith *et al.*, 2006). We show here that GARP may actually present some advantages over GAM when the species is rare and there are extremely few presence records, a situation that may arise frequently for endangered species. For species with fewer than 100 presence records (5% prevalence in a sample size of 2000), variability in model performance in general is very high for all modelling strategies. In practice, this translates into some samples generating very poor predictions, while others generate very good predictions. Strategies such as those exemplified by Guisan *et al.* (2006), where predictive modelling is used iteratively in conjunction with a stratified sampling of potentially suitable habitat, may be the only way to characterize rare species' geographical ranges. Increasing sampling size guided by habitat modelling in an iterative process would certainly increase model performance (Engler *et al.*, 2004; Reese *et al.*, 2005; Guisan *et al.*, 2006).

To our knowledge, this is the first study to compare species with different functional environment–occurrence relationship shapes in a controlled setting. Results appear to be fairly

consistent across different shapes of the environment–occurrence relationship and species prevalence, and the type of interaction between factors (additive or no interaction vs. multiplicative), validating previous empirical studies in their general recommendations (Segurado & Araújo, 2004; Elith *et al.*, 2006).

There are a few previous studies that use virtual species in a real landscape to study species-distribution models. Austin *et al.* (1995) and Hirzel *et al.* (2001) modelled artificial species that are equivalent to the additive species used here, and used only one species type, with a combination of functional shapes, to test statistical models. We believe that the multiplicative species are more realistic in that environmental factors are likely to interact in determining most species-occupancy patterns, and modellers tend to ignore these potential interactions. Other studies, reviewed by Austin *et al.* (2006) and Austin (2007), used artificial data to generate abundance responses along environmental gradients. Austin *et al.* (2006) generated artificial data both on species abundance and on environmental gradients. There is obviously a relationship between species abundance and probability of occurrence (Brown, 1995; Holt *et al.*, 2002; Gaston *et al.*, 2006). However, this relationship is not always easy to characterize (Holt, 2003; Gaston *et al.*, 2006; Hui *et al.*, 2006). While bell-shaped curves have been widely advocated when studying variations in species relative abundances along environmental gradients (Austin, 2007), other types of response, such as environmental thresholds and linear relationships, have been used more often when studying species presence–absence patterns (Hirzel *et al.*, 2001; Guisan & Thuiller, 2005). As argued by Austin *et al.* (1995, 2006) and Austin (2007), different statistical models are often used without clear reference to ecological theory. There is an urgent need to create a theoretical framework for species distributions that includes occurrence–abundance relationships. He & Gaston (2003) recently proposed a model that links species abundance, variance and occupancy patterns (He & Gaston, 2003), but the model has not been tested (Gaston *et al.*, 2006). Potts & Elith (2006) compared different abundance models and found that the hurdle model, which separates the occurrence pattern from the relative abundance pattern, performed better than other techniques at recovering species distributions. This suggests that occurrence and abundance may respond differently to environmental gradients, and incorporating both would aid better understanding of species distributions.

Finally, we suggest that this simulation strategy opens the door for testing a variety of hypotheses regarding species distributions. Some of the questions that could be further studied using this approach relate to the effects of sampling strategies, population structure, species spatial aggregation and habitat patchiness, among other practical issues. For example, we did not consider here the effect of active spatial aggregation of individuals, a factor that can add significant complexities to the environment–occurrence patterns (Lichstein *et al.*, 2002; Reese *et al.*, 2005; Hoeting *et al.*, 2006). Previous studies suggest that this kind of spatial autocorrelation can change significance levels of predictors and affect variable selection techniques, but do not necessarily substantially affect predictive ability (Reese *et al.*, 2005). The artificial species created here did not aggregate actively, therefore any spatial autocorrelation in their distribution can be explained by the environmental predictors (Haining, 2003). Other simulation studies have been applied to these questions (Hirzel & Guisan, 2002; Dark, 2004; Reese *et al.*, 2005; Wintle & Bardos, 2006) and could guide our empirical studies in a more systematic fashion. We suggest that artificial species populating real landscapes could be used more frequently to guide new sampling efforts in particular geographical areas, and could become especially relevant in guiding studies of rare and endangered species.

ACKNOWLEDGEMENTS

This work was supported in part by a Fulbright fellowship, a dissertation year fellowship and several other funding sources from the University of California to C.N.M., and by the National Biological Information Infrastructure. We thank in particular David M. Kaplan for his assistance with programming issues. Steve Greco, Alan M. Hastings, Mark Schwartz, Art M. Shapiro, David M. Kaplan, Miguel Araújo and two anonymous reviewers provided useful comments on early drafts of the manuscript.

REFERENCES

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

Araújo, M.B., Pearson, R.G., Thuiller, W. & Erhard, M. (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. (2007) Species distribution models and ecological theory: a critical assessment and some possible new approaches. *Ecological Modelling*, **200,** 1–19.

Austin, M.P. & Heyligers, P.C. (1989) Vegetation survey design for conservation – gradsect sampling of forests in north-eastern New South Wales. *Biological Conservation*, **50,** 13–32.

Austin, M.P., Meyers, J.A., Belbin, I. & Doherty, M.D. (1995) *Modelling of landscape patterns and processes using biological data. Subproject 5: Simulated data case study*. Division of Wildlife and Ecology, Commonwealth Scientific and Industrial Research Organisation, Canberra.

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.

Boulinier, T., Nichols, J.D., Sauer, J.R., Hines, J.E. & Pollock, K.H. (1998) Estimating species richness: the importance of

heterogeneity in species detectability. *Ecology*, **79,** 1018–1028.

Bourg, N.A., McShea, W.J. & Gill, D.E. (2005) Putting a cart before the search: successful habitat prediction for a rare forest herb. *Ecology*, **86,** 2793–2804.

Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J. (1984) *Classification and regression trees*. Chapman & Hall/CRC Press, New York.

Brown, J.H. (1995) *Macroecology*. Chicago University Press, Chicago.

Buckland, S.T., Anderson, D.N., Burnham, K.P., Laake, J.L., Borchers, D.L. & Thomas, L. (2005) *Introduction to distance sampling: estimating abundance of biological populations*, 3rd edn. Oxford University Press, Oxford.

Bulluck, L., Fleishman, E., Betrus, C. & Blair, R. (2006) Spatial and temporal variations in species occurrence rate affect the accuracy of occurrence models. *Global Ecology and Biogeography*, **15,** 27–38.

Burnham, K.P. & Anderson, D.R. (2002) *Model selection and multimodel inference: a practical information-theoretic approach*, 2nd edn. Springer-Verlag, New York.

Dark, S.J. (2004) The biogeography of invasive alien plants in California: an application of GIS and spatial regression analysis. *Diversity and Distributions*, **10,** 1–9.

Derksen, S. & Keselman, H.J. (1992) Backward, forward and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*, **45,** 265–282.

Elith, J. & Burgman, M.A. (2002) Predictions and their validation: rare plants in the Central Highlands, Victoria, Australia. *Predicting species occurrences: issues of scale and accuracy* (ed. by J.M. Scott, P.J. Heglund, M.L. Morrison, J.B. Haufler, M.G. Raphael, W.A. Wall and F.B. Samson), pp. 303–313. Island Press, Covelo, CA.

Elith, J., Graham, C.H., Anderson, R.P., Dudik, 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., Pearce, J. & Drielsma, M. (2002) Extended statistical approaches to modeling spatial pattern in biodiversity in northeast New South Wales. 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.

Fleishman, E., MacNally, R., Fay, J.P. & Murphy, D.D. (2001) Modeling and predicting species occurrence using broad-scale environmental variables: an example with butterflies of the Great Basin. *Conservation Biology*, **15,** 1674–1685.

Franklin, J.F. (1998) Predicting the distribution of shrub species in southern California from Climate and terrain-derived variables. *Journal of Vegetation Science*, **9,** 733–748.

Gaston, K. (2003) *The structure and dynamics of geographic ranges*, 1st edn. Oxford University Press, Oxford.

Gaston, K.J., Borges, P.A.V., He, F.L. & Gaspar, C. (2006) Abundance, spatial variance and occupancy: arthropod species distribution in the Azores. *Journal of Animal Ecology*, **75,** 646–656.

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

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

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.

Haining, R. (2003) *Spatial data analysis: theory and practice*, 1st edn. Cambridge University Press, Cambridge.

Hastie, T. & Tibshirani, R.J. (1990) *Generalized additive models*. Chapman & Hall/CRC Press, London.

He, F.L. & Gaston, K.J. (2003) Occupancy, spatial variance, and the abundance of species. *The American Naturalist*, **162,** 366–375.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G. & Jarvis, A. (2004) *The WorldClim interpolated global terrestrial climate surfaces. Version 1.3.* http://biogeo.berkeley.edu

Hirzel, A. & Guisan, A. (2002) Which is the optimal sampling strategy for habitat suitability modelling? *Ecological Modelling*, **157,** 331–341.

Hirzel, A.H., Helfer, V. & Metral, F. (2001) Assessing habitat-suitability models with a virtual species. *Ecological Modelling*, **145,** 111–121.

Hoeting, J.A., Davis, R.A., Merton, A.A. & Thompson, S.E. (2006) Model selection for geostatistical models. *Ecological Applications*, **16,** 87–98.

Holt, R.D. (2003) On the evolutionary ecology of species' ranges. *Evolutionary Ecology Research*, **5,** 159–178.

Holt, A.R., Gaston, K.J. & He, F.L. (2002) Occupancy–abundance relationships and spatial distribution: a review. *Basic and Applied Ecology*, **3,** 1–13.

Hui, C., McGeoch, M.A. & Warren, M. (2006) A spatially explicit approach to estimating species occupancy and spatial correlation. *Journal of Animal Ecology*, **75,** 140–147.

Keitt, T.H., Bjornstad, O.N., Dixon, P.M. & Citron-Pousty, S. (2002) Accounting for spatial pattern when modeling organism–environment interactions. *Ecography*, **25,** 616–625.

Latimer, A.M., Wu, S., Gelfand, A.E. & Silander, J.A. (2006) Building statistical models to analyze species distributions. *Ecological Applications*, **16,** 33–50.

Lichstein, J.W., Simons, T.R., Shriner, S.A. & Franzreb, K.E. (2002) Spatial autocorrelation and autoregressive models in ecology. *Ecological Monographs*, **72,** 445–463.

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

Maggini, R., Lehmann, A., Zimmermann, N.E. & Guisan, A. (2006) Improving generalized regression analysis for the spatial prediction of forest communities. *Journal of Biogeography*, **33,** 1729–1749.

McCullagh, P. & Nelder, J.A. (1989) *Generalized linear models*, 2nd edn. Chapman & Hall/CRC Press, London.

McNab, B.K. (1980) Food-habits, energetics, and the population biology of mammals. *The American Naturalist*, **116,** 106–124.

McPherson, J.M., Jetz, W. & Rogers, D.J. (2004) The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? *Journal of Applied Ecology*, **41,** 811–823.

Nielsen, S.E., Johnson, C.J., Heard, D.C. & Boyce, M.S. (2005) Can models of presence–absence be used to scale abundance? Two case studies considering extremes in life history. *Ecography*, **28,** 197–208.

Potts, J.M. & Elith, J. (2006) Comparing species abundance models. *Ecological Modelling*, **199,** 153–163.

Prasad, A.M., Iverson, L.R. & Liaw, A. (2006) Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems*, **9,** 181–199.

Prosser, C.L. & Brown, F. (1961) *Comparative animal physiology*. W. B. Saunders, Philadelphia, PA.

R Development Core Team (2005) *r: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. http://www.R-project.org

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

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.

Royle, J.A. (2004) N-mixture models for estimating population size from spatially replicated counts. *Biometrics*, **60,** 108–115.

Royle, J.A., Nichols, J.D. & Kery, M. (2005) Modelling occurrence and abundance of species when detection is imperfect. *Oikos*, **110,** 353–359.

Running, S.W., Nemani, R.R., Heinsch, F.A., Zhao, M., Reeves, M. & Hashimoto, H. (2004) A continuous satellite-derived measure of global terrestrial primary production. *Bioscience*, **54,** 547–560.

Segurado, P. & Araújo, M.B. (2004) An evaluation of methods for modelling species distributions. *Journal of Biogeography*, **31,** 1555–1568.

Segurado, P., Araújo, M.B. & Kunin, W.E. (2006) Consequences of spatial autocorrelation for niche-based models. *Journal of Applied Ecology*, **43,** 433–444.

Spicer, J.I. & Gaston, K. (1999) *Physiological diversity and its ecological implications*, 1st edn. Blackwell Science, Oxford.

Steel, R.G.D., Torrie, J.H. & Dickey, D.A. (1997) *Principles and procedures of statistics: a biometric approach*, 3rd edn. McGraw-Hill, New York.

Stockwell, D. (1999) Genetic algorithms II: species distribution modelling. *Machine learning methods for ecological applications* (ed. by A.H. Fielding), pp. 123–144. Kluwer Academic, Dordrecht.

Termansen, M., McClean, C.J. & Preston, C.D. (2006) The use of genetic algorithms and Bayesian classification to model species distributions. *Ecological Modelling*, **192,** 410–424.

Thuiller, W., Araújo, M.B. & Lavorel, S. (2003) Generalized models vs. classification tree analysis: predicting spatial distributions of plant species at different scales. *Journal of Vegetation Science*, **14,** 669–680.

Tilman, D. (1980) Resources – a graphical-mechanistic approach to competition and predation. *The American Naturalist*, **116,** 362–393.

Venables, W.N. & Ripley, B.D. (2002) *Modern applied statistics with S*, 4th edn. Springer-Verlag, New York.

Webb, G.I. & Ming Ting, K. (2005) On the application of ROC analysis to predict classification performance under varying class distributions. *Machine Learning*, **58,** 25–32.

Wessels, K.J., Van Jaarsveld, A.S., Grimbeek, J.D. & Van der Linde, M.J. (1998) An evaluation of the gradsect biological survey method. *Biodiversity and Conservation*, **7,** 1093–1121.

West, J.B. (1996) Prediction of barometric pressures at high altitudes with the use of model atmospheres. *Journal of Applied Physiology*, **81,** 1850–1854.

Wintle, B.A. & Bardos, D.C. (2006) Modeling species–habitat relationships with spatially autocorrelated observation data. *Ecological Applications*, **16,** 1945–1958.

Wood, S.N. (2004) Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, **99,** 673–686.

Wood, S.N. & Augustin, N.H. (2002) GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecological Modelling*, **157,** 157–177.

Zweig, M.H. & Campbell, G. (1993) Receiver–operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, **39,** 561–577.

SUPPLEMENTARY MATERIAL

The following supplementary material is available for this article online:

Table S1 Multiple comparisons between species types ($n = 2000$, prevalence = 50%).

Table S2 Multiple comparisons between models ($n = 2000$, prevalence = 5%).

Table S3 Multiple comparisons between models ($n = 150$, prevalence = 5%).

This material is available as part of the online article from http://www.blackwell-synergy.com/doi/abs/10.1111/j.1365–2699.2007.01720.x

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.

BIOSKETCHES

Christine N. Meynard has recently completed her PhD at the University of California, Davis and is now a postdoctoral researcher at the Universidad Austral, Valdivia, Chile. She has focused her research on strategies for modelling species distributions and their use in assessing areas of conservation interest. Other areas of interest include biogeography and landscape ecology.

James F. Quinn is a professor in the Department of Environmental Science and Policy and is co-director of the Information Center for the Environment at the University of California, Davis. His fields of interest include conservation biology and the use of new technologies for biodiversity data access.

Editor: Miguel Araújo