Ecological Applications, 16(1), 2006, pp. 33–50

BUILDING STATISTICAL MODELS TO ANALYZE SPECIES DISTRIBUTIONS

ANDREW M. LATIMER,[1,4] SHANSHAN WU,[2] ALAN E. GELFAND,[3] AND JOHN A. SILANDER, JR.[1]

[1]*Department of Ecology and Evolutionary Biology, University of Connecticut, 75 North Eagleville Road, Unit 3043, Storrs, Connecticut 06269 USA*
[2]*Department of Statistics, University of Connecticut, 215 Glenbrook Road, Unit 4120, Storrs, Connecticut 06269 USA*
[3]*Institute for Statistics and Decision Sciences, Box 90251, Duke University, Durham, North Carolina 27708 USA*

Abstract. Models of the geographic distributions of species have wide application in ecology. But the nonspatial, single-level, regression models that ecologists have often employed do not deal with problems of irregular sampling intensity or spatial dependence, and do not adequately quantify uncertainty. We show here how to build statistical models that can handle these features of spatial prediction and provide richer, more powerful inference about species niche relations, distributions, and the effects of human disturbance. We begin with a familiar generalized linear model and build in additional features, including spatial random effects and hierarchical levels. Since these models are fully specified statistical models, we show that it is possible to add complexity without sacrificing interpretability. This step-by-step approach, together with attached code that implements a simple, spatially explicit, regression model, is structured to facilitate self-teaching. All models are developed in a Bayesian framework. We assess the performance of the models by using them to predict the distributions of two plant species (Proteaceae) from South Africa's Cape Floristic Region. We demonstrate that making distribution models spatially explicit can be essential for accurately characterizing the environmental response of species, predicting their probability of occurrence, and assessing uncertainty in the model results. Adding hierarchical levels to the models has further advantages in allowing human transformation of the landscape to be taken into account, as well as additional features of the sampling process.

Key words: Bayesian; Cape Floristic Region; hierarchical; Proteaceae; spatial dependence; spatial random effects; spatially explicit; species distribution model; uncertainty.

INTRODUCTION

Ecologists increasingly use species distribution models to address theoretical and practical issues, including predicting the response of species to climate change (e.g., Midgley et al. 2002), identifying and managing conservation areas (e.g., Austin and Meyers 1996), finding additional populations of known species or closely related sibling species (e.g., Raxworthy et al. 2003), and seeking evidence of competition among species (e.g., Leathwick 2002). In all of these applications, the core problem is to use information about where a species occurs (and where it does not) and about the associated environment to predict how likely the species is to be present or absent in unsampled locations.

Spatial prediction of species distributions is thus directly related to the concept of the environmental niche, a specification of a species' response to a suite of environmental factors (MacArthur et al. 1966, Austin et al. 1990, Brown et al. 1995). But are contemporary environmental factors alone sufficient to account for species distributions? Other ecological processes including dispersal, reproduction, competition, and the dynamics of large and small populations may also affect the spatial arrangement of species distributions (Gaston 2003). Most species-distribution models ignore spatial pattern and thus are based implicitly on two assumptions: (1) that environmental factors are the primary determinants of species distributions; and (2) that species have reached or nearly reached equilibrium with these factors. These assumptions underlie both types of modeling approaches that are currently dominant: generalized linear and generalized additive regression models (GLM and GAM), and climate envelope models (see Guisan and Zimmermann 2000). These assumptions may or may not be adequate approximations, depending on the relative influence of environmental change, including both climate change and direct human transformation of landscapes, microevolution, and dispersal-related lags in species movement across landscapes (Peterson 2003). To the extent

Manuscript received 2 April 2004; revised 27 September 2004; accepted 27 September 2004; final version received 2 December 2004. Corresponding Editor: D. S. Schimel. For reprints of this Invited Feature, see footnote 1, p. 3.

[4] E-mail: andrew.latimer@uconn.edu

that model assumptions are violated, or that species distribution data are inadequate, simple species distribution models may fail to provide adequate predictive power, or may underestimate the degree of uncertainty in their predictions, resulting in poor characterization of the environmental response of the species. In addition to these fundamental ecological issues, spatial modeling also raises practical complications relating to sampling, including observer error, variable sampling intensity, and gaps in sampling. There may also be a spatial mismatch between different data sources (e.g., Agarwal et al. 2002). Here we present models that address these problems, show how they are related to simpler models, and demonstrate how to build them.

The literature on species distribution modeling covers many modeling approaches and applications. Fortunately, there are useful review papers that organize and categorize model approaches (Guisan and Zimmermann 2000, Ferrier et al. 2002, Guisan et al. 2002). This paper aims to complement these reviews by focusing on Bayesian regression models for species presence/absence and taking a model-building approach.

We start with a brief overview of the problem of prediction of species distributions in space, and discuss the features of a statistical model suitable for this problem. As an initial step, we introduce a familiar generalized linear model that relates distribution data to environmental attributes. This basic model, and variants of it, are currently in wide use in ecological research (Guisan and Zimmermann 2000). Then we consider what can be done to improve this model. We build from this familiar starting point toward more complex models. The final model we consider is a spatially explicit hierarchical regression model implemented in a Bayesian framework (Wikle 2003, Gelfand et al. 2005*a*). Hierarchical models are statistical models in which data can enter at different stages, where these stages describe conceptual but unobservable latent processes (see, e.g., Carlin and Louis [2000] for a general discussion of hierarchical modeling and Bannerjee et al. [2003] in the context of spatial data). As we construct the models, we consider the advantages of different model features, such as including spatial features in the mean vs. using spatial random effects, specifying spatial relationships (point vs. areal), and modeling with hierarchical levels. The goal is to give ecologist readers a coherent framework for understanding the basic regression approach, and simultaneously to show how to implement newer statistical techniques in their own modeling. To facilitate this self-teaching, we include code for some of the models that readers can run with the free, publicly available software package WinBUGS (*available online*).[5]

All of the models presented here are implemented as Bayesian models. The simpler models are equally easy to implement through classical maximum likelihood methods. However, the Bayesian approach has strong advantages when building complex, hierarchical models (Link and Sauer 2002, Clark 2003, Hooten et al. 2003), so we construct all the models in this framework. Complexity can then be added without switching the basic inferential paradigm. A Bayesian model uses Bayes' theorem to combine the information in the data with additional, independently available information (the prior) to produce a full probability distribution (posterior distribution) for all parameters (Gelman et al. 1995, Carlin and Louis 2000, Congdon 2001). The Bayesian perspective allows us to ask directly how probable are the hypotheses, given the data. Bayes Theorem can be expressed as

$$P(\text{Parameters}\,|\,\text{Data}) = \frac{P(\text{Data}\,|\,\text{Parameters}) \times P(\text{Parameters})}{P(\text{Data})} \quad (1)$$

where the notation $P(x\,|\,y)$ means the probability of x conditional on y. This theorem, introduced posthumously by Thomas Bayes in 1763, enables drawing inferences directly about the parameters of interest (the left-hand side of the equation). This ''posterior probability distribution,'' $P(\text{Parameters}\,|\,\text{Data})$, provides a full picture of what is known about each parameter based on the model and the data, together with any prior information, $P(\text{Parameters})$. For a particular parameter, this posterior distribution, unlike the mean and confidence interval produced by classical analyses, enables explicit probability statements about the parameter. Thus the region bounded by the 0.025 and 0.975 quantiles of the posterior distribution has an intuitive interpretation: under the model, the unknown parameter is 95% likely to fall within this range of values.

One feature of Bayesian models, and also a source of much debate within the statistics community, is their ability to incorporate already-known or ''prior'' information about parameters into the models. In some applications, this can be a critical advantage, particularly when data are sparse, decisions must be made, and expert opinion is available (Gelman et al. 1995). But here we do not put informative priors on the parameters, because, though there may be previous studies relating presence/absence to various environmental factors, it is not straightforward to transform this information into prior specifications for regression coefficients in the complex models we consider. Moreover, we are primarily interested in learning what information is contained in the data. Thus, all priors in these models are vague or uninformative, and thus the posterior distributions for all parameters are driven by the data.

This paper does not attempt to be a primer in Bayesian statistics; the books mentioned above cover that ground (see also Ellison [2004] for an introduction to

[5] ⟨http://www.mrc-bsu.cam.ac.uk/bugs/winbugs⟩

the Bayesian perspective for ecologists). We do not address the use of informative priors in Bayesian models, because we use only vague priors here; discussion is available in the foregoing texts on Bayesian statistics (e.g., Gelman et al. 1995). We also do not consider formal statistical model comparison. The current state of the art is that there are tools to effectively compare models that share a common hierarchical specification. However, for comparing models having a different number of hierarchical levels or different stochastic mechanisms at these levels, as would be the case across the models we are examining here, criteria for model comparison are not well developed, and there is little theory to provide guidance. See the paper by Spiegelhalter et al. (2002) and the associated discussion section for more on this issue. Moreover, customary model comparison that reduces a model to a single number may be misleading. We argue that examination of predicted species distributions across models relative to observed species presence and absence patterns will be more illuminating. Indeed, this is what we do, providing maps of predicted distributions and using receiver operating characteristics (ROC) curves and related criteria to assess relative model performance.

Finally, this paper does not address in detail the mechanics of fitting these models using Markov chain Monte Carlo (MCMC) methods. For more on implementing MCMC and assessing its convergence, sources include Carlin and Louis (2000), Gelman et al. (1995), and Gilks et al. (1995).

Problems in spatial prediction in ecology

Most problems in spatial ecological prediction arise from three features of species distributions and the data that are gathered to describe them. First, there are problems that relate to sampling. In general, the spatial area covered by the species distributions is vastly larger that the area sampled, and correspondingly, the spatial unit of prediction is usually larger than the sites sampled on the ground. An additional sampling problem relates to the heterogeneity of sampling intensity: while large parts of a domain may be unsampled, other parts may be relatively heavily sampled. Finally, the environmental data for the region of interest are typically available at a much coarser spatial resolution than the scale at which species distribution data may be collected. Relating the species distribution data to the environmental data and to the region where prediction is sought thus often presents problems of spatial misalignment. For example, some data may be available at point locations (point data), while other data are in the form of a regular grid (e.g., climate data) or irregular polygons (e.g., edaphic or geological data). These different kinds of data sources do not line up (e.g., Agarwal et al. 2002). There may also be sample bias in the available data, so that sampling locations and/or intensity are unevenly distributed with respect to relevant characteristics of the region sampled (Mugglin et al. 2000, Gelfand et al. 2002). These problems are often ignored in spatial modeling, or are addressed indirectly by attempting to minimize bias through stratified sampling across major environmental gradients (e.g., Ferrier et al. 2002).

Second, there is the problem of spatial dependence or spatial autocorrelation. Data are spatially dependent or autocorrelated when the degree of correlation among observations depends on their relative locations. In ecological data, pairs of observations that are closer together often tend to be more similar than pairs of observations that are farther apart, and this produces positive autocorrelation. Because ecological processes such as reproduction and dispersal typically generate spatial autocorrelation in species occurrences, and because some residual autocorrelation in environmental factors tends to remain even when many environmental factors included in a model, the ''residuals'' of a purely environment-based predictive model will usually exhibit some degree of autocorrelation. Using models that ignore this dependence can lead to inaccurate parameter estimates and inadequate quantification of uncertainty (Ver Hoef et al. 2001). Equally important, to ignore this spatial dependence is to throw out meaningful information (Wikle 2003). Ecological prediction often ignores this problem or deals with it in a less than satisfactory way (see Guisan and Zimmerman 2000). One might, for example, include latitude and longitude through a trend surface in the mean to improve prediction, but this approach may still miss spatial dependence which explicit modeling of spatial association can capture. Another alternative provided by the generalized regression analysis and spatial prediction (GRASP) modeling package treats spatial autocorrelation at the data stage, by feeding into the model a new data layer that reflects neighborhood values in model predictions (software *available online*).[6] The model is then iterated to convergence (Augustin et al. 1996, Lehmann et al. 2002). This approach does deal with autocorrelation, but fails to quantify explicitly the strength of spatial pattern in the residuals and associated uncertainty.

Third, spatial modeling presents problems in quantifying uncertainty. Because the spatial domain of prediction is typically very large relative to the domain in which the data were collected, environmental data are not available on a scale as fine as that experienced by individual organisms. Predictions frequently involve extrapolation to unobserved parts of the study region and to larger-scale areal units. Assessment of uncertainty in such predictions is crucial when they are used to set conservation policy or to evaluate the impact of climate change on species (e.g., Thomas et al. 2004). But again, ecological distribution models that

[6] ⟨http://www.cscf.ch/grasp/⟩

focus exclusively on the mean may yield false confidence in the predictions made (Ver Hoef et al. 2001).

The tools presented here can be used to deal with these three kinds of problems in a straightforward, transparent way. Within the Bayesian framework, full inference about uncertainty, given what we have observed (the data) and what we know or assume about the process (the model), comes ''free'' with the model predictions. Spatial autocorrelation can be incorporated into a regression model through random effects that capture spatial dependence in the data. Since the random effects are model parameters, they also emerge with a full posterior distribution that allows quantification of uncertainty. By adding hierarchical levels to a regression model, issues of sampling intensity, holes in the data, and human transformation of the landscape can be dealt with explicitly (Gelfand et al. 2005*b*). Hierarchical models are statistical models in which data can enter at various levels and, thus, model parameters or unknowns are themselves functions of other model parameters and data. These hierarchical stages can describe conceptual but unobservable latent processes that are ecologically important, such as error in the observation process, or potential suitability of a site in a transformed landscape for a particular species. In this way, uncertainty attached to unknowns at different model stages is propagated across model levels to more accurately reflect overall inferential uncertainty.

Crucial to the approach presented here is the notion of transparency. One reaction of ecological modelers to the problems inherent in species distribution modeling is to adopt more flexible methods like neural networks, genetic algorithms, and discriminant analysis (Manel et al. 1999, Moisen and Frescino 2002, D'Heygere et al. 2003). These methods offer the advantages of responding flexibly to interactions and nonlinearities in data relationships, but often at the expense of interpretability or mechanistic insights. We hope to show here that many of the difficulties involved in ecological distribution modeling can be dealt with in a very general way through hierarchical modeling without sacrificing the interpretability of simpler statistical models.

Methods

Data sources and preparation

We illustrate the modeling through prediction of distributions of plant species in the Cape Floristic Region of South Africa (CFR), a global hotspot of plant diversity, endemism, and rarity (Cowling et al. 1997, Meyers et al. 2000). The data used here consist of observations of species presence/absence and environmental characteristics at grid-cell level. More precisely, each species presence/absence observation is referenced to a particular point in space, although, in fact, sampling was conducted in a small areal unit. The sample locations are displayed with a map of the CFR in Fig. 1. Fig. 1 also presents a close-up view of a small part of the CFR to illustrate the data in more detail, and shows in particular that the sampling intensity is quite variable and that there are many cells in which the species have been observed to be both present and absent at different sample locations in a single cell. While these sample locations in fact represent compact areal surveys, not points, they are smaller than a grid cell and vastly smaller than the entire CFR, so it is reasonable to take these units as points. The species data were collected by the Protea Atlas Project of South Africa's National Botanical Institute. The grid-cell referenced environmental data are data layers listed in Table 1 and displayed in Appendix B. See also Gelfand et al. (2005*b*). Here the grid cells are $1' \times 1'$ rectangles which, at this latitude, translates to about 1.55×1.85 km. The data layers cover various aspects of the environment potentially important for plant species persistence, including temperature, precipitation, topography (elevation, roughness), and edaphic features (soil fertility, texture, pH).

For illustration, we selected two species in the family Proteaceae, a diverse group of flowering woody plants that includes the king protea (*Protea cynaroides*), the national flower of South Africa. These species, *Protea mundii* and *P. punctata*, are relatively widespread within the region and are closely related, with abutting but nonoverlapping distributions. The species exhibit some ecological differences, with *P. punctata* occupying higher elevation, inland sites and *P. mundii* nearer the coast and at lower elevations. The abrupt transition between the two species' distributions, however, is not associated with any obvious sudden environmental transition. These two species thus present a good challenge for models, both to predict distribution boundaries correctly and to identify environmental variables that distinguish the environmental responses of the two species.

A critical aspect of environment that is not routinely considered in species distribution modeling is the extent to which the landscape has been transformed from a natural state by human intervention or by introduced alien species. Since most areas of the world, including the Cape region, have been significantly affected by urbanization, agriculture, alien invasive species, and other impacts, for most purposes it would seem necessary to consider the effect of these changes on the occurrence of species. The final model presented here, the Bayesian hierarchical spatial model, enables this information to be taken into account and predictions to be made for untransformed conditions as well as current, transformed conditions. To make these predictions, we incorporate a data layer that combines the major classes of human impact and specifies the proportion of the landscape transformed for each grid cell (Rouget et al. 2003).

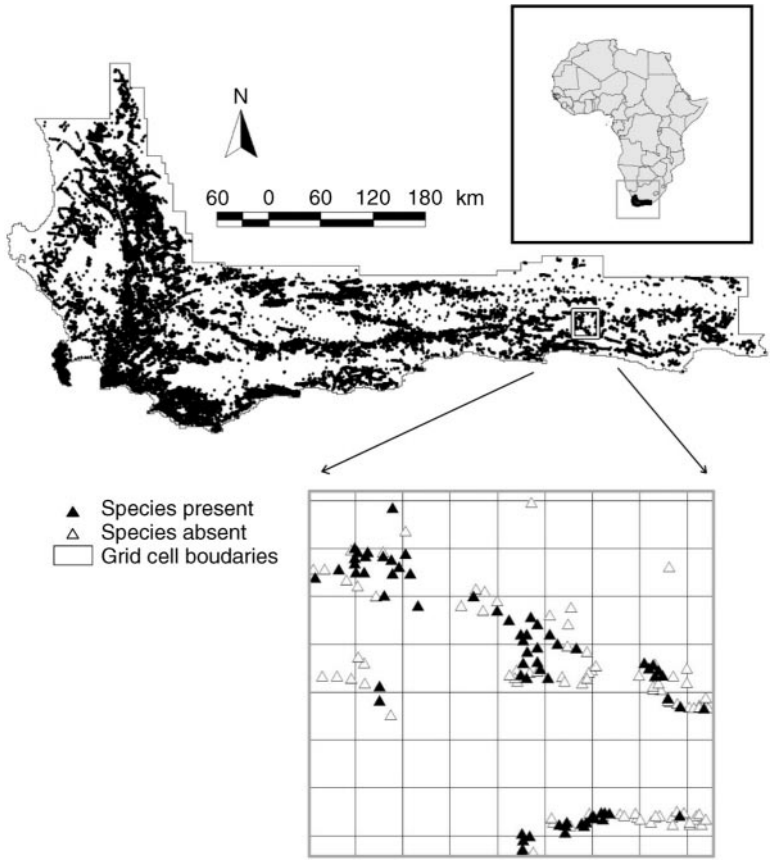

FIG. 1. The Cape Floristic Region with Protea Atlas Project sample points overlaid. The thumbnail shows the region's location on the African continent. The pull-out box displays, for a small (12 × 13 km) part of the region, the grid cells used for modeling and sample locations scattered across these grid cells, including both sample locations at which *Protea mundii* was observed (solid triangles) and the ''null sites'' at which it was not (open triangles).

TABLE 1. Posterior summary of significant/suggestive coefficients (β's) for *Protea mundii*.

Variable	Abbreviation	Model 1	Model 2	Model 3	Model 4
Roughness	ROUGH	−	+	(+)	+
Elevation	ELEV	+	+	NS	NS
Potential evapotranspiration	POTEVT	+	NS	NS	NS
Interannual CV precipitation	PPCTV	(−)	NS	NS	NS
Frost season length	FROST	−	NS	NS	NS
Heat units	HEATU	−	−	−	−
January maximum temperature	JANMAXT	+	+	(+)	+
July minimum temperature	JULMINT	NS	+	NS	NS
Seasonal concentration of precipitation	PPTCON	−	−	(−)	−
Summer soil moisture days	SUMSMD	+	NS	+	(+)
Winter soil moisture days	WINSMD	−	NS	−	NS
Enhanced vegetation index	EVI	+	+	+	+
Low fertility	FERT1	+	NS	+	NS
Moderately low fertility	FERT2	NS	NS	(+)	NS
Moderately high fertility	FERT3	−	NS	NS	NS
Fine texture	TEXT1	+	NS	NS	NS
Acidic soil	PH1	−	NS	−	NS
Alkaline soil	PH3	−	NS	−	NS

Notes: For Tables 1 and 2, + and − denote positive and negative coefficients with 95% credible intervals that do not overlap 0; (+) and (−) denote positive and negative coefficients with 90% credible intervals that do not overlap 0; NS, not statistically significant.

A simple generalized linear model

Let us begin with the simplest GLM that directly relates presence/absence data to environmental explanatory variables. Here, we have two options to deal with the misalignment between the species presence/absence data, which are referenced to point locations, and the environmental data, which are referenced to grid cells. We can work at the scale of the responses, i.e., the sample sites, and assign to each sample site the values of the environmental factors for the grid cell in which the site falls. Alternatively, we can work at the grid cell level, assigning presence to the grid cell if any sample site in that cell showed presence, or assigning absence if it was not found at any sample site. The latter choice seems less attractive in that it fails to account for the sampling intensity attached to a grid cell. Absence at four sample sites in a cell, for example, should be viewed differently from absence based only on one sample site. Expressed in a different way, the latter reduces the response probability for a grid cell to a single Bernoulli trial (i.e., a coin toss) while the former results in a binomial variable reflecting the number of trials on the grid cell. Since the models are equally routine to fit, we choose the former here.

Let $Y(s)$ be the presence/absence (1/0) of the modeled species at sample location s in cell i. Summing up $Y(s)$ over the number of sample sites in cell i (n_i) yields grid-cell level counts: $Y_{i+} = \Sigma_{s\in \text{grid}\,i} Y(s)$. Since all of these sites are assigned the same levels for the environmental factors, if we assume independence for the trials, a binomial distribution results for Y_{i+}:

$$Y_{i+} \sim \text{Binomial}(n_i, p_i). \tag{2}$$

We assume that the probability that the species occurs in cell i, p_i, is related functionally to the environmental variables. A logistic (logit) function is convenient to relate p_i to the linear predictor $\mathbf{w}_i'\beta$ (other link functions, e.g., the probit, could be considered but would not be expected to affect subsequent predictions of spatial distribution), yielding

$$\log\left(\frac{p_i}{1 - p_i}\right) = \mathbf{w}_i'\beta. \tag{3}$$

In this logistic regression model, $\mathbf{w}_i'$ is a vector of explanatory environmental variables associated with cell i and β is a vector of the associated coefficients. Note that here the first entry in $\mathbf{w}_i'$ is a 1, which provides an overall intercept for the regression, so that $\mathbf{w}_i'\beta = \beta_0 + \beta_1\mathbf{w}_{i1} + \cdots + \beta_t\mathbf{w}_{it}$. For an unsampled cell ($n_i = 0$), there will be no contribution to the likelihood. For a sampled cell ($n_i \geq 1$), there will be a contribution to the likelihood and the likelihood component can be written as

$$P(Y_{i+} = y) = \binom{n_i}{y}\frac{(e^{\mathbf{w}_i'\beta})^y}{(1 + e^{\mathbf{w}_i'\beta})^{n_i}}. \tag{4}$$

Model 1 can be fitted easily by classical or Bayesian approaches. Available software includes SAS (SAS Institute, Cary, North Carolina, USA), S-Plus (Insightful Corporation, Seattle, Washington, USA), and WinBUGS (see footnote 5). The first two are widely used in classical framework while the later is the most user-friendly software available for Bayesian modeling. Note that, to fully specify a Bayesian model, a prior distribution has to be assigned for every random parameter. The only parameters in this model are the β's. We use normal priors centered at 0 with a fixed large variance, $\beta \sim N(0, \sigma_\beta^2)$, where σ_β^2 is a hyperprior whose value is assigned by the modeler. These priors are almost flat; they contain very little information about the location of the parameters, and are thus examples of ''vague priors'' (see, e.g., Gelman et al. [1995] for prior specification methods). A script for implementing Model 1 in WinBUGS is provided, along with data and initial values, in Appendix A.

A more detailed discussion of the model results appears in the next section, but as we work through the models we present species distribution maps created from the model output to allow for an immediate visual comparison. Fig. 2 shows this model's prediction of the distributions of *Protea mundii* and *P. punctata*. Because of the large number of grid cells in the modeled region (36 907), pull-out boxes are used in this and the other model prediction figures to display detailed views of small, representative parts of the grid. Though the model correctly picks out the core part of *P. mundii*'s distribution, there are also areas of apparent underprediction, where the model is not predicting a high probability of presence even though the species has been observed repeatedly. This problem is particularly noticeable in the western populations of this species, which are separated from the rest by hundreds of kilometers and might thus be expected to present a difficult challenge for the model.

A simple spatially explicit model

Spatial structure or autocorrelation in ecological pattern and process is pervasive. In the context of species distribution patterns, we would anticipate that the presence/absence of a species at one location may be associated with presence/absence at neighboring locations. A variety of different mechanisms could generate such association (e.g., dispersal of propagules among neighboring cells, similar ecological attributes among neighboring cells, etc.). We select a model that can respond flexibly to this spatial dependence, while retaining the environmental response coefficient (i.e., $\mathbf{w}_i'\beta$) structure. This can be achieved by adding spatial random effects to the model, which yields our Model 2. Modeling at the grid-cell level, under Eq. 2, a spatial term ρ_i associated with grid cell i is added in Eq. 3

$$\log\left(\frac{p_i}{1 - p_i}\right) = \mathbf{w}_i'\beta + \rho_i. \tag{5}$$

In this equation, each grid cell has an associated ran-

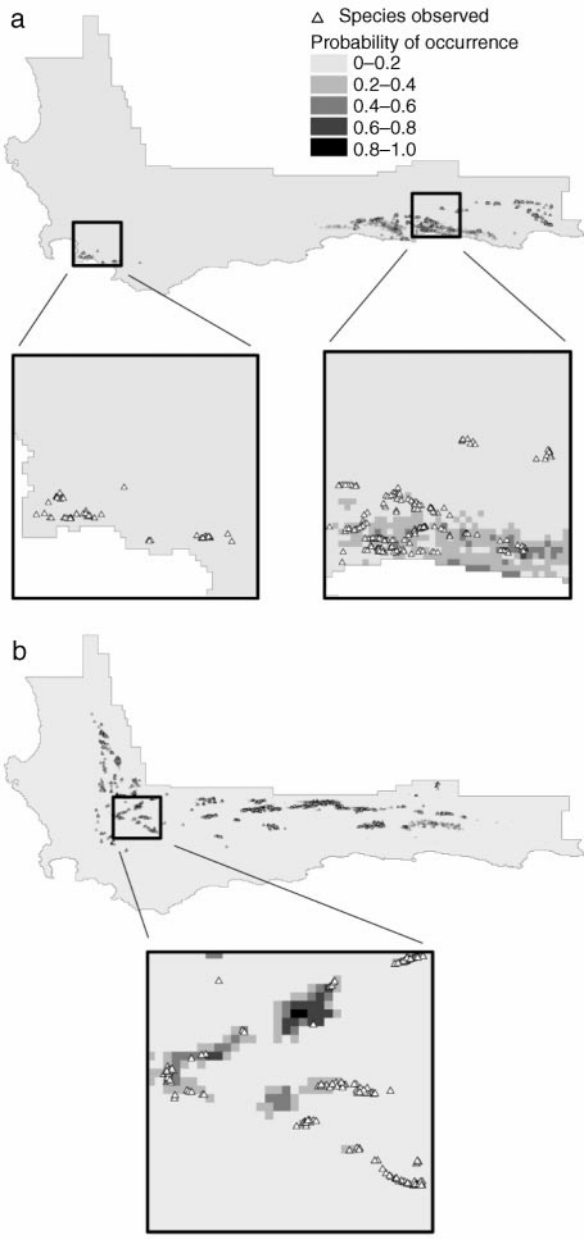

FIG. 2. Predicted distribution for (a) *Protea mundii* and (b) *P. punctata* from Model 1. For this and all other distribution maps (Figs. 4–6), each grid cell is assigned the mean of the posterior distribution for p_i, the predicted probability of occurrence of the species in cell *i*. Because the full region includes 37 000 grid cells, for purposes of visualization, pull-out boxes are used to present close-ups of selected portions of the predicted region. Sample locations at which the species were observed are overlaid as triangles.

dom effect ρ_i that adjusts the probability of presence of the modeled species up or down, depending on the values of ρ in cell *i*'s spatial neighborhood. To capture this process, we employ a Gaussian (intrinsic) conditional autoregressive (CAR) model (Besag 1974). To specify this model, we assume the conditional distribution of the spatial random effect in cell *i*, given values for the spatial random effect in all other cells $j \neq i$, depends only on the spatial random effect of the neighboring cells of *i*, $\delta_i i$. Here, we specify that cell *i* is a neighbor of *j* if their boundaries intersect (see Fig. 3). In this version, the spatial effect for any given cell depends only on the values of ρ for the cells in its neighborhood, and the neighborhood encompasses only the eight immediately adjacent cells. The neighborhood could alternatively be defined to be larger, and different weights could be assigned to cells at different distances.

Formally, the Gaussian CAR model for the spatial random effect at cell *i* can be presented by a conditional distribution (conditioned on all the other cells)

$$\rho_i \,|\, \rho_j \approx N\left(\frac{\sum_{j\in\delta i} a_{ij}\rho_j}{a_{i+}}, \frac{\sigma_\rho^2}{a_{i+}}\right) \qquad j \neq i \tag{6}$$

where a_{i+} denotes the total number of cells which are neighbors of *i*, and $a_{ij} = 1$ if sites *i* and *j* share the same boundary and 0 otherwise. The conditional variance σ_ρ^2 is a hyper-parameter (a parameter that is a prior for another parameter) and requires a prior distribution. As σ_ρ^2 is a variance, we assign an inverse gamma prior, $\sigma_\rho^2 \sim \text{IG}(2, b_\rho)$. This distribution has mean b_ρ with infinite variance. Again, for the βs, we assign normal priors centered at 0 with a fixed large variance. This spatially explicit model can be directly fitted in WinBUGS using Bayesian simulation-based methods. WinBUGS code is provided in Appendix A.

Fig. 4 shows predicted distributions for *Protea mundii* and *P. punctata* using Model 2. It is immediately apparent that this model improves over the nonspatial model. For example, unlike Model 1, this model is able to predict the presence of *P. mundii* in its isolated western populations, and also shows a tighter fit in the main eastern populations. Two methods of quantifying model performance are presented in *Results*, and both confirm

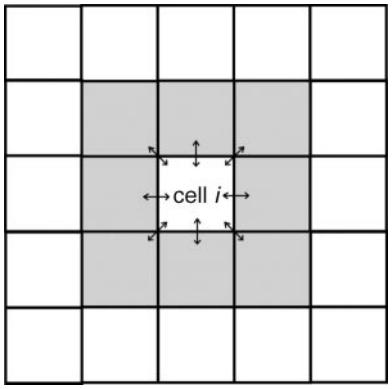

FIG. 3. Diagram of the grid cell neighborhood used in the conditional autoregressive (CAR) models employed in Models 2 and 4.

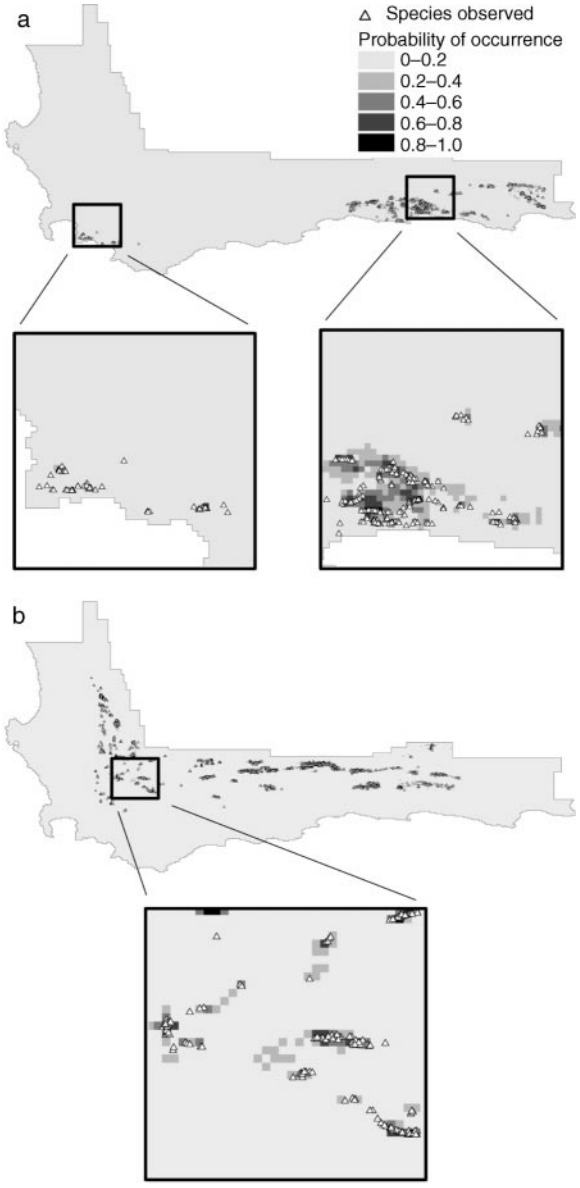

FIG. 4. Predicted distribution for (a) *Protea mundii* and (b) *P. punctata* from Model 2. See the explanation in the Fig. 2 legend.

that making the model spatially explicit dramatically improves performance.

A point-level spatial model

Models 1 and 2 were built at the grid-cell scale, so spatial association was specified as a relationship between grid cells in a neighborhood. Point-level models provide an alternative way of building spatially explicit models for species distributions. Point-level models use the (geo-coded) locations of the observation points themselves, rather than referencing the points to discrete grid cells, to specify spatial relationships. For example, the Protea Atlas Project data we are using here includes abundance information as well as presence/absence; these abundance estimates are only for the point locations, not a continuous grid. So if we were modeling abundance, it might be preferable to model at the point level. Point-level models have been widely used to explain ecological patterns and processes, such as animal movement and metapopulation dynamics (e.g., Turchin 1998, Ettema et al. 2000, Stoyan et al. 2000). To build a point-level model, spatial dependence can be modeled directly between the points based on the distances among them, with the degree of autocorrelation decreasing as some function of distance. This modeling of dependence as a continuous function of distance contrasts with the neighborhood approach typically used in grid-level or lattice models, in which spatial dependence is modeled between discrete sets of neighboring cells (compare the individual points in Fig. 1 with the grid-cell neighborhood in Fig. 3). Kriging is perhaps the most familiar use of point-level modeling: kriging is a way of predicting the values of a response variable (like species abundance or soil characteristics) at new, unmeasured locations (Bannerjee et al. 2003). Our goal, however, is not simply to predict but also to investigate the relationships between species distributions, environmental variables, and spatial pattern, so our modeling also includes environmental covariates as well as spatial random effects.

We introduce a point-level spatial model as Model 3. In this model, data are referenced to the sample locations themselves rather than to grid cells in which they occur. Because there is only one observation of presence or absence of the species per sample site, the data point $Y(\mathbf{s})$ is the outcome of a single Bernoulli trial:

$$Y(\mathbf{s}) \sim \text{Bernoulli}[p(\mathbf{s})]. \quad (7)$$

The probability that the species occurs in site s, $p(\mathbf{s})$, can be related to the set of environmental variables through the logistic regression model shown in Eq. 3. To do this, of course, requires that we have the environmental data for each sample site. For this illustration, we can use the same grid-based environmental data, and let $\mathbf{w}(\mathbf{s}) = \mathbf{w}_i$ when s is within grid i.

The point-level spatial model can now be specified as

$$\log\frac{p(\mathbf{s})}{1 - p(\mathbf{s})} = \mathbf{w}'(\mathbf{s})\beta + \rho(\mathbf{s}) \quad (8)$$

where $\rho(\mathbf{s})$ is the spatial random effect associated with point s. Because there are an uncountable number of locations s it is not appropriate to model the $\rho(\mathbf{s})$ using a neighbor-based specification such as a CAR. Rather, we describe the spatial dependence at the point level through a spatial process model that directly models pairwise association using a parametric covariance

function: $\text{cov}(\mathbf{s}_i, \mathbf{s}_j) = f(d_{ij}; \theta)$, where f is a decreasing function of d_{ij}, the distance between $\mathbf{s}_i$ and $\mathbf{s}_j$, and θ is an unknown parameter of this function. In such a model, the $\rho(\mathbf{s})$'s are spatial random effects whose spatial structure is specified by the covariance function. For example, a widely used choice for the covariance function is an exponential function $\exp(-(\varphi d_{ij})^{\kappa})$, where φ is a ''decay'' parameter and κ is a ''power'' parameter applied to the distance. Together, these parameters control the rate of decay of covariance with distance d_{ij}. For a small number of sample sites, such models can be implemented in WinBUGS. But while overcoming the disadvantages of the CAR model, this model introduces its own disadvantage, computational limitation. The computational demand of the MCMC sampling algorithm for this model scales as the third power of the number of sample points, so it rapidly becomes unfeasible to run as the number of points increases. Currently, WinBUGS can handle about 100 sample points; individually tailored code can accommodate perhaps up to 1000 points, but with more than that, some form of approximation will be required.

In our case, there are 52 275 sample locations, which is more than an order of magnitude greater, so we offer as an alternative a kernel smoothing model to capture spatial association at the point level (Higdon et al. 1999). This approach reduces the cost of modeling the spatial association to a function of the number of data points, rather than a function of the number of points cubed, as in the spatial process model. We select K locations over the region, t_i, $i = 1, 2, \ldots, K$, which correspond to the centroids of the 36 907 grid cells used in Models 1 and 2, and define

$$\rho(\mathbf{s}) = \sum_i g(\|\mathbf{s} - t_i\|)Z_i \qquad (9)$$

where the Z_i are assumed to be normally distributed, $N(o, \sigma^2)$, and $\|\cdot\|$ denotes straight-line interpoint (Euclidean) distance. Under this specification we have $\text{cov}(\mathbf{s}_i, \mathbf{s}_j) = \sigma^2 \Sigma_i g(\|\mathbf{s} - t_i\|)g(\|\mathbf{s}_j - t_i\|)$. $g(\|\cdot\|)$ should be a function which decreases in $\|\cdot\|$ to 0, for example, an inverse distance function, $g(\|\mathbf{s} - t_i\|) = \exp\{-\gamma\|\mathbf{s} - t_i\|\}$. This inverse distance function is the ''kernel'' that defines the spatial structure in $\rho(\mathbf{s})$. The degree of local spatial smoothing decreases with γ; in other words, a larger gamma makes the spatial random effects in neighboring cells more strongly correlated. For the illustration presented here, we chose γ such that the weight function $g(\|\mathbf{s} - t_i\|)$ approaches zero when $\|\mathbf{s} - t_i\|$ is three times the length of the side of a grid cell (here 1′).

Fairly sophisticated code is required to fit this model; no packaged software is available. Though we show the results of fitting this model below and in *Results*, our primary purpose in presenting it is to show that there are spatial modeling alternatives to the grid-cell (areal) versions we have focused on.

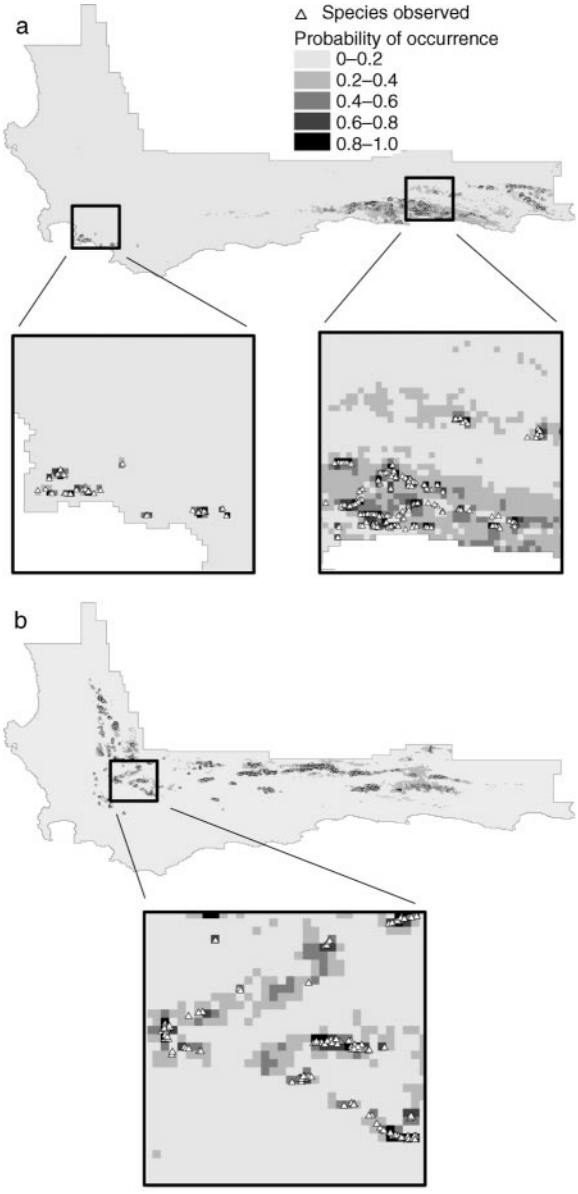

FIG. 5. Predicted distribution for (a) *Protea mundii* and (b) *P. punctata* from Model 3. See the explanation in the Fig. 2 legend.

The distributions predicted by this model are quite different from those in the previous models, with many more cells assigned high probabilities of occurrence. Fig. 5 shows that the model predicts reasonably high probability of occurrence for both species where they have been observed, but also that the model predicts them to occur in many adjacent areas where they have not been observed, i.e., the model overpredicts the species' distributions. This behavior appears to result from the stronger smoothing applied by the Gaussian kernels in this model, and results in a smoother, generally high-

er-intensity probability surface. A larger γ would impart less spatial smoothing.

A hierarchical spatially explicit model

In the Cape Floristic Region, as in many other parts of the world, including those with highest conservation priorities, the landscape has already been substantially altered by human disturbance, including agriculture, urbanization, forestry, and invasion by exotic plant species. The data on current species presence or absence represent species distributions within this transformed landscape. Some of the sample points inevitably fall within areas that have been transformed to some degree. The single-level regression models do not allow us to deal explicitly with this, and so the interpretation of the model predictions is somewhat ambiguous. Should they be interpreted as predicting the natural distribution of the species in unchanged conditions, or as predicting distributions given human disturbance? With a hierarchical model, here presented as Model 4, we can resolve this problem by incorporating data on transformation of the landscape into the model, so that the contribution of each sample point to the model prediction reflects the degree to which the cell in which it falls has been transformed. Then the model will be able to predict explicitly what the distribution would have been in the absence of transformation (i.e., potential distribution) and what the distribution would be in the current transformed landscape (i.e., transformed distribution). To do this, we will add a second hierarchical level to the model.

We have been treating the sample points as Bernoulli trials, and specifying the probability of observing a species in a grid cell as a function of environment and a spatial random effect. But the sampling data are actually more complex. Whether a species is observed in a cell is a function of both the ecological attributes of the site, or site suitability (which influences whether a species is in fact there), and of the sampling process (whether, in the inventory process, the species was observed). These distinct but related processes can be investigated by adding a third hierarchical level to the model.

To demonstrate the advantages of hierarchical modeling, we develop here a multilevel hierarchical model that considers not only irregular sampling intensity and spatial dependence, but also incorporates the influence of land transformation and the data collection process. In particular, for each grid cell i, we introduce two unobservable (or latent) binary variables. X_i denotes the event that a randomly selected location in grid cell i is environmentally suitable (1) or unsuitable (0) for the species, that is, that the species could exist there or not. V_i denotes the same event, adjusted for transformation. For a particular grid cell, the probability that a randomly selected point in the cell is suitable is $p_i = P(X_i = 1)$. Thus, cells with high values of p_i are predicted to be highly environmentally suitable for the species, that is, there is a high probability that the species can grow there, or a high potential presence. A single p_i can be viewed as the relative suitability of a particular grid cell, scaled 0 to 1. Across the grid-cell domain, the suite of p_is can be seen as representing the potential distribution of the species in the absence of landscape transformation, while $P(V_i = 1)$ represents the transformed distribution, given extant human alterations of the landscape.

In modeling the $p_i = P(X_i = 1)$, or the relative suitability of the grid cells, we begin with environmental variables that are expected to affect the suitability of grid cell i for the modeled species. The variables used in the models described here are listed in Table 1. For p_i, we use a logistic regression model conditional on cell-level environmental variables and on cell-level spatial random effects just as in Eq. 6, where $\mathbf{w}_i$ is again a vector of grid cell level environmental characteristics, and β is a vector of the associated coefficients, including an intercept. As in Model 2, the ρ_i denote spatially associated random effects which are modeled using a CAR specification.

Next, at the second hierarchical level, we model $P(V_i \mid X_i)$. Let U_i denote the proportion of area in the ith cell which is transformed, $0 \leq U_i \leq 1$. Intuitively, we model $P(V_i = 0 \mid X_i = 1) = U_i$, which captures the idea that the probability of absence caused by land transformation given the potential presence is proportional to the transformed percentage. Of course, $P(V_i = 1 \mid X_i = 0) = 0$. Marginalizing over X_i (that is, summing $P(V_i = 1 \mid X_i = 0)$ and $P(V_i = 1 \mid X_i = 1)$) we get

$$P(V_i = 1) = (1 - U_i)p_i. \tag{10}$$

This equation has the interpretation that the probability of transformed presence is adjusted by $(1 - U_i)\cdot 100\%$ of the probability of the potential presence. A theoretical justification, viewing the probabilities as averages of binary processes, is presented in Gelfand et al. (2005*b*).

At the third hierarchical level of the model, we address the relationship between the observed presence/absence data and the environmental suitability variables. Assume that cell i has been visited n_i times (in untransformed areas) within the cell. Further, let Y_{ij} be the presence/absence status of the species in the ith cell at the jth sample location within that cell. Given $V_i = 1$, we view the Y_{ij} as independent, identically distributed, Bernoulli trials with success probability q_i. In other words, q_i represents, for an arbitrarily selected location in cell i, the probability that the species is observed given that it could grow there. There are various reasons the species might not have been observed in a highly suitable grid cell: there may be observation error, or low sampling intensity, or the species may simply not have established a population in the cell.

Of course, given $V_i = 0$ (i.e., the landscape is 100% transformed), $Y_{ij} = 0$ with probability 1.

Marginalizing over V_i and using Eq. 10, we get $P(Y_{ij} = 1) = q_i(1 - U_i)p_i$. Given $V_i = 1$, $Y_{i+} = \Sigma_{j=1}^{n_i} Y_{ij} \sim$ Binomial(n_i, q_i). For an unsampled cell ($n_i = 0$) there will be no contribution to the likelihood. For a sampled cell ($n_i \geq 1$) there will be a contribution to the likelihood and, in fact, again marginalizing over V_i, for $y > 0$,

$$P(Y_{i+} = y) = \binom{n_i}{y}(q_i)^y(1 - q_i)^{n_i - y}(1 - U_i)p_i$$

and for $y = 0$,

$$P(Y_{i+} = 0) = (1 - q_i)^{n_i}(1 - U_i)p_i + [1 - (1 - U_i)p_i].$$

The two components of this latter expression have specific interpretation. The first, for an arbitrary location in cell *i*, provides the probability that the species is not present though the location is suitable while the second provides the probability that the location is not suitable (so it could not be present). For the transformed cell, the contribution to the likelihood is adjusted by the percentage of transformation, U_i. If a cell is 100% transformed, then it has no contribution to the likelihood.

In modeling for q_i, the probability that the species was observed at the *j*th site in cell *i*, we use a logistic regression model:

$$\log\left(\frac{q_i}{1 - q_i}\right) = \mathbf{w}_i\beta \tag{11}$$

where $\mathbf{w}_i$ are cell level environmental characteristics that may affect q_i. For the current model, we allow the same set of environmental characteristics at this level as were included in the first level of the model ($\mathbf{w}_i$). There is also an intercept term in the model.

For fitting such a complex model, simulation-based (MCMC) Bayesian model fitting is currently the only option. The computation is demanding, but in return, we can obtain the full posterior distributions of all parameters, which makes possible full inference for quantities of interest. A disadvantage is that this model cannot be implemented in current releases of WinBUGS, and therefore the model must be custom coded.

Fig. 6 presents predictions from this model of the potential probability of occurrence of both *P. mundii* and *P. punctata*, that is, their potential distributions in the absence of transformation. Because these species, particularly *P. punctata*, tend to occur in areas of high elevation and poor soils that are not likely to be used for agriculture or housing, the probability surfaces for transformed probability of occurrence are nearly identical and are not shown. Essentially all points where the species were observed lie in areas of predicted moderate or high probability of occurrence, with less apparent overprediction than Model 3 (compare Figs. 5 and 6).

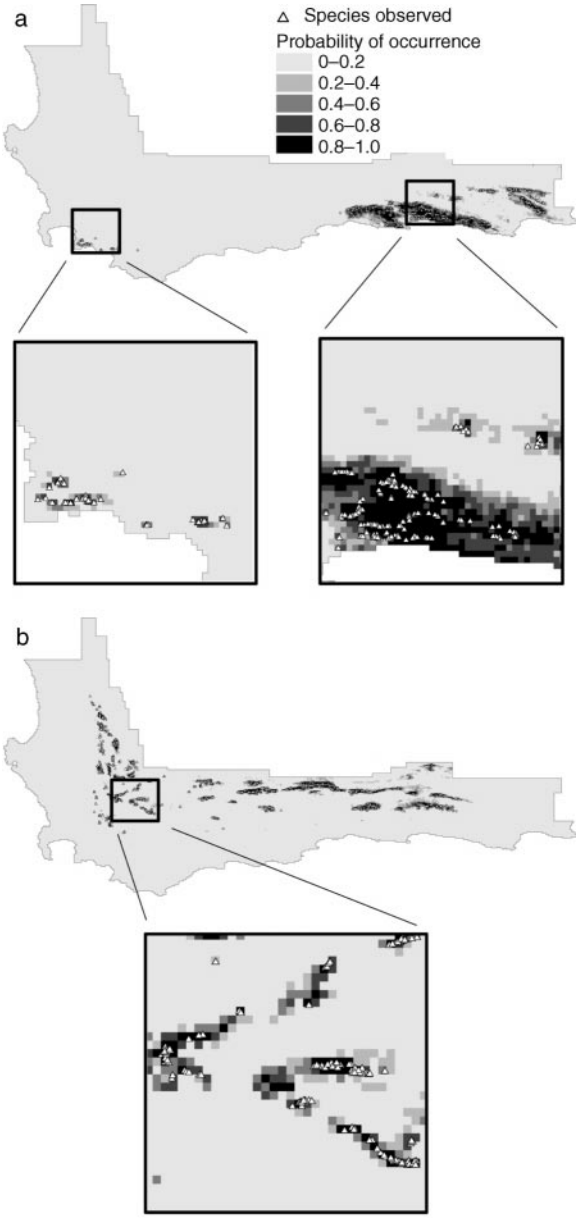

FIG. 6. Predicted potential distribution for (a) *Protea mundii* and (b) *P. punctata* from Model 4. See the explanation in the Fig. 2 legend.

RESULTS

We have seen already that adding complexity to basic generalized linear models improved the models' characterization of the distributions of our two example species. We now turn to a more thorough evaluation of the model output, including the estimates for the environmental coefficients, the spatial random effect variables, and the uncertainty associated with model parameters.

Table 1 summarizes the estimates from all four models for the βs, or coefficients for the explanatory en-

TABLE 2. Posterior summary of significant/suggestive coefficients (β's) for *Protea punctata*.

Variable	Abbreviation	Model 1	Model 2	Model 3	Model 4
Roughness	ROUGH	+	+	+	+
Elevation	ELEV	+	+	+	+
Potential evapotranspiration	POTEVT	+	+	+	+
Interannual CV precipitation	PPCTV	−	−	−	−
Frost season length	FROST	+	+	+	+
Heat units	HEATU	−	NS	−	NS
January maximum temperature	JANMAXT	−	NS	−	NS
July minimum temperature	JULMINT	+	NS	NS	(−)
Mean annual precipitation	MAP	+	NS	(+)	NS
Seasonal concentration of precipitation	PPTCON	−	NS	−	NS
Winter soil moisture days	WINSMD	−	NS	NS	NS
Enhanced vegetation index	EVI	−	NS	−	+
Moderately high fertility	FERT3	+	+	+	+
Fine texture	TEXT1	(+)	NS	NS	NS
Medium coarse texture	TEXT3	−	NS	−	NS
Coarse texture	TEXT4	−	NS	−	NS
Alkaline soil	PH3	−	NS	NS	NS

Notes: For Tables 1 and 2, + and − denote positive and negative coefficients with 95% credible intervals that do not overlap 0; (+) and (−) denote positive and negative coefficients with 90% credible intervals that do not overlap 0; NS, not statistically significant.

vironmental variables (**w**) for *Protea mundii*. Table 2 provides the same information for *P. punctata*. It is immediately apparent that the different models present somewhat different assessments of the relative contributions of environmental explanatory variables. The most striking and consistent pattern is that the nonspatial model, Model 1, has many more significant coefficients than the other three models. This is intuitive; with a large sample including more than 52 000 sample sites, in the absence of flexible random effects, more environmental factors would be expected to emerge as significant. We elaborate this point further in the *Discussion* section. Generally, the edaphic coefficients (i.e., fertility, texture, pH) drop out more often than climate coefficients (variables 3–12 in Tables 1 and 2) when spatial random effects are added to the model. Though the parameter estimates may differ across models, they are generally consistent in sign across the models, so all the models convey broadly the same picture about the species' relationship with environmental factors.

From those parameters that are significant across more than one model, a picture emerges of the species' ecological characteristics. Both are associated with higher elevation, with *P. punctata* having significantly larger coefficients; it is typically found at higher elevations. *P. punctata* is also favored by a longer frost season and higher potential evapotranspiration, and discouraged by higher January (summer) maximum temperature. *P. mundii* is conversely positively associated with higher January maximum temperature, though negatively with higher mean summer heat levels (heat units, the sum of the number degrees by which each summer month's mean daily maximum temperature exceeds 18°C), and is also significantly associated with a more even year-round rainfall pattern (negative coefficient for seasonal concentration of precipitation, and positive coefficient for summer soil moisture availability, SMDSUM). The edaphic coefficients show that *P. punctata* is positively associated with moderately fertile soils, while *P. mundii* is associated, though less consistently, with the least fertile soils.

Posterior distributions are available for the spatial random effects, and they can be summarized in a map that displays the mean value of the effects. Fig. 7 presents the mean value of the spatial random effect for *Protea mundii* for each grid cell for each of the three spatial models (Models 2, 3, and 4). The spatial effects for all models show strong spatial patterning at multiple scales. The spatial effects for Models 2 and 4 are qualitatively similar, including larger regions and local areas where the species are encouraged (+) or discouraged (−). The spatial effect surface for Model 3 appears quite different, with values for individual cells that are noticeably higher or lower than cells in their immediate neighborhood. Further, these spatial effects appear rather uniform away from the areas in which the species were observed, which may contribute to the higher overall level of predicted probability of occurrence from this model (see Fig. 5).

The posterior distributions for probability of species presence/absence can also be summarized to give a picture of the level of uncertainty about parameter values. Fig. 8 displays the variance in probability of occurrence for the four models for *P. mundii*. The figures show the lowest variance for Model 1, with the highest overall level for Model 3. The nonspatial model, Model 1, produces parameter estimates with relatively low variance. Further, since the model cannot respond to spatial relationships, the areas of slightly higher variance are not associated spatially with patterns in the sample data. The variance for all spatial models, by

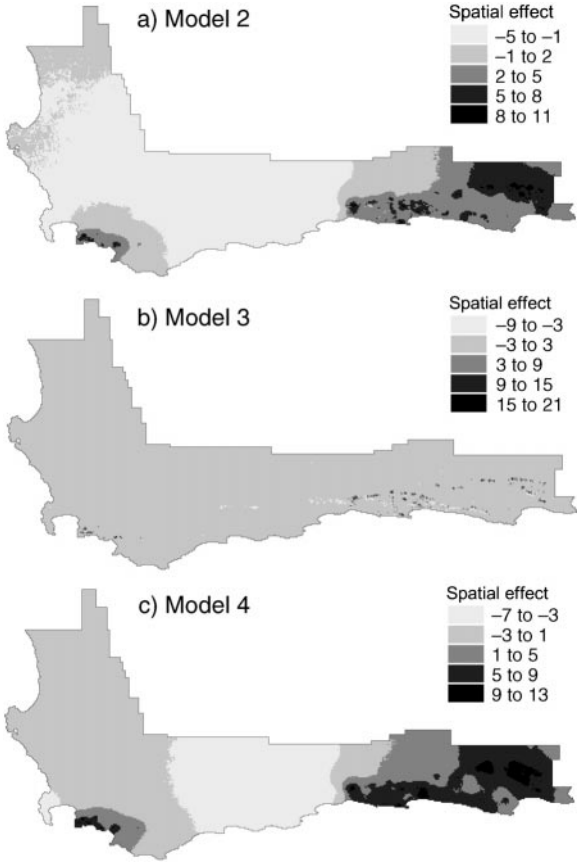

FIG. 7. Comparison of spatial random effects from the three spatially explicit models for *Protea mundii*.

contrast, is much higher, and has strong spatial patterning. In particular, the variance tends to be high in areas where the species has not been observed, but that are close to grid cells where it has been found. In these cells, the model predicts a relatively high probability of occurrence, but also a higher level of uncertainty.

Table 3 gives the ''prevalence,'' or mean probability of occurrence across the region, i.e., $\Sigma_i p_i$, for the two species for all models. Prevalence can be seen as a measure of overall commonness of occurrence, and provides one simple way of comparing across models. Adding spatial random effects to the models does not in itself greatly change the predicted prevalence for either species. Note, however, that the confidence interval is slightly wider in both cases for the spatial model; as in the case of the environmental coefficients, Model 1 appears to be underestimating the degree of uncertainty about the p_i. The point-level model, by contrast, produces far higher prevalences for each species than any of the other models. This result is consistent with the overprediction that is clearly visible in the distribution predictions from this model; see Fig. 5. The hierarchical model produces an intermediate level of prevalence for both species; this is not simply due to the effects of transformation, since even the transformed prevalence estimates from this model are significantly higher than those of Models 1 and 2. Notice that whereas the distribution of *Protea mundii* is substantially affected by transformation, *P. punctata* is predicted to be almost completely unaffected (Table 3).

Receiver operating characteristics (ROC) analysis provides a quantitative complement to visual assessments of the fit of model predictions. ROC plots summarize graphically the performance of a model as a tradeoff between sensitivity, in this context the probability of correctly predicting a true presence, and specificity, the probability of correctly predicting a true absence (see Fielding and Bell 1997). A species distribution model that fits the data well will predict high probability of presence where the species was observed, while minimizing the probability of presence where the species was observed to be absent. ROC plots display sensitivity on the *y*-axis and (1 − selectivity) on the *x*-axis. The area under the curve (AUC) then provides an integrated measure of the performance of the model, with high values reflecting better performance (i.e., sensitivity can be increased without losing much selectivity, and vice versa). For species distribution predictions, ROC curves can be generated using only sampled cells, and assigning each cell presence or absence according to what was observed, or can also be generated under the assumption that the species is absent in all cells except those in which it was observed, including unsampled cells. We present results using both approaches in Table 4. Clearly, by the AUC criterion, the spatially explicit GLM (Model 2) outperforms the nonspatial GLM (Model 1), while the hierarchical model (Model 4) performs best. Only when the AUC calculated for *P. mundii* using all cells ranks Model 4 as slightly worse than the other spatial models. This result reflects a disagreement between the model's prediction and the assumption that species are absent from unsampled cells—given that we do not know the truth about these cells, this result is difficult to interpret. Model 3's tendency to overpredict is reflected in a curve that rises slowly as selectivity decreases, giving this model an AUC between those of Models 1 and 2.

A recently proposed measure, ''minimum predicted area'' (MPA; Engler et al. 2004) is closely related to ROC curves: it counts the number of cells in the predicted distribution that are above some threshold value corresponding to a specified sensitivity level. Because it corresponds to a distribution prediction that can be visualized on a map, it has a more intuitive spatial interpretation than the AUC. To find the MPA of a model prediction, we select a threshold value p_t that gives the desired sensitivity level, say 0.95 (that is, 95% of cells in which the species was observed have predicted probability of occurrence greater than p_t). Then we count how many sampled cells in the entire predicted region have predicted probability of presence

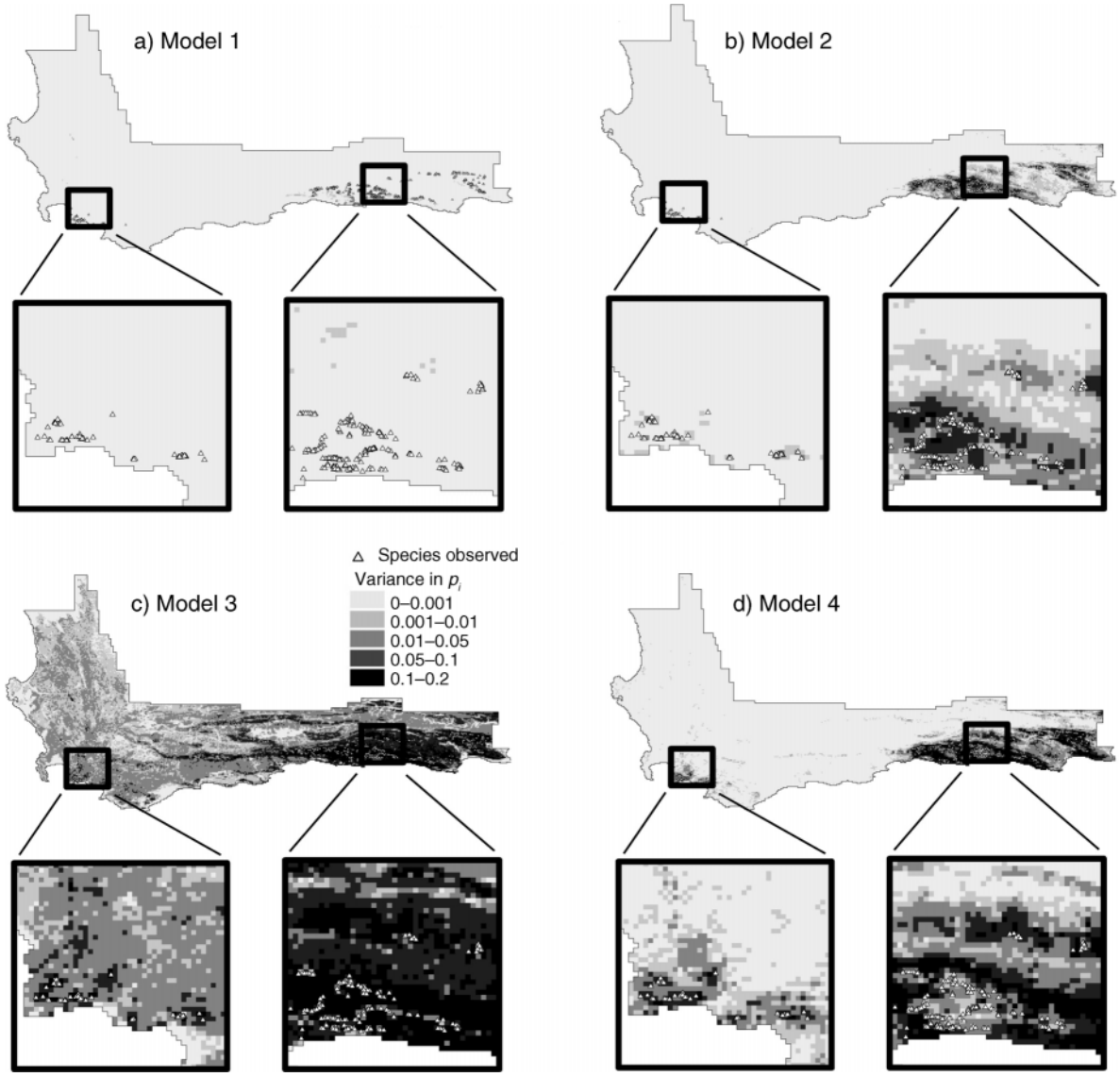

FIG. 8. Variance in the predicted probability of occurrence of *P. mundii* for Models 1–4. Each grid cell is assigned the variance of the posterior distribution of p_i, the predicted probability of occurrence of the species in cell *i*.

TABLE 3. Predicted prevalence.

	Protea mundii quantile			*Protea punctata* quantile		
Model	0.025	0.50	0.975	0.025	0.50	0.975
Model 1	0.011	0.012	0.013	0.018	0.019	0.020
Model 2	0.011	0.012	0.014	0.017	0.018	0.021
Model 3	0.051	0.057	0.065	0.057	0.064	0.071
Model 4 (potential)	0.040	0.046	0.051	0.042	0.047	0.052
Model 4 (adjusted)	0.031	0.035	0.039	0.041	0.046	0.051

Notes: Potential prevalence is the mean probability of occurrence in an untransformed landscape; adjusted prevalence is the mean probability of occurrence given human transformation.

TABLE 4. Model performance.

	Protea punctata			*Protea mundii*		
Model	AUC (sampled)	AUC (all)	MPA	AUC (sampled)	AUC (all)	MPA
Model 1	0.964	0.969	4487	0.940	0.948	8367
Model 2	0.995	0.992	2128	0.996	0.993	2013
Model 3	0.978	0.974	6127	0.992	0.992	2950
Model 4	0.996	0.995	1180	0.998	0.989	1180

Notes: Table 4 displays the area under the receiver operating characteristics (ROC) curve (AUC) and minimum predicted area (MPA) for each model. MPA was calculated for a sensitivity value of 0.95. Larger AUC and smaller MPA reflect better model performance.

greater than p_t. With this criterion, smaller is better. Note that because it requires setting an arbitrary sensitivity level, MPA is the equivalent of only a single point on the ROC curve. In any case, at a sensitivity level of 0.95, the spatial models outperform the nonspatial model, with one exception, and Model 4 always does best by a large margin (see Table 4).

DISCUSSION

We have seen that the basic nonspatial logistic regression model can be improved by adding features to the model that reflect major known attributes of the data, including variable sampling intensity, spatial autocorrelation, and land use impacts. There is a strong case for making these models spatially explicit. Indeed, Gaston (2003) and others have called for the incorporation of spatially explicit methods in determining the structure and dynamics of species geographic ranges, but progress has been limited. Further, one of the problems in comparative biogeography is that sampling and data gathering are conducted with a multitude of different methodologies (Gaston 2003, Graham et al. 2004). The consequence is uncertainty in comparing and interpreting spatial patterns in species distributions. Are the patterns real or are they simply artifacts of differences in sampling effort? The full, hierarchical model developed here allows one to incorporate variation in sampling effort directly in the model.

Our model results confirm that the spatial pattern of presence and absence of a species includes more information than can be explained through just the mean effect of a suite of environmental variables. There are two main explanations for this. One is that biological processes tend to generate spatial pattern. In the case of species distributions, the probability that a site contains a species depends on not only on its climatic and edaphic characteristics, but also on its neighborhood. Such spatial dependence can arise from biological processes at a number of levels. Processes in the life history of individual organisms, including reproduction, territoriality, and dispersal, can generate clustering or evenness in species distributions. Interactions of species with each other and with resources (e.g., effects of grazers on vegetation, the role of nurse plants in recruitment) can likewise cause and perpetuate spatial association. The particular occupancy history of a site can also exert a long-term spatial influence on its neighborhood. These spatial patterns are not mere epiphenomena, but rather can strongly influence individual species distributions, as well as interspecific interactions and thus community composition and potentially ecosystem processes. The second explanation for autocorrelation is the influence of unobserved environmental variables, and of nonlinearities in interactions among sets of (observed and unobserved) factors, all of which may have some degree of spatial dependence and interdependence (Ver Hoef et al. 2001). Moreover, it is inevitable that the identity of critical explanatory variables may change from one part of a species' geographical range to another (Gaston 2003). Since models cannot include all important variables, and may include some unimportant ones, there will usually be some degree of autocorrelation in model residuals. Models (such as Models 2, 3, and 4) that can fit environmental responses despite spatial dependence in the data, and that identify and quantify this spatial dependence, are thus essential tools for investigating and predicting species distributions.

Critically, without spatial structure in the model, the level of uncertainty about model parameters can be dramatically underestimated and poorly characterized. One effect of this is that a model will identify more explanatory variables as significantly related to species presence/absence (Ver Hoef et al. 2001). Since a nonspatial model like Model 1 will tend to underestimate the uncertainty about the environmental response of the species (i.e., the βs), the credible intervals for these parameters are smaller and are more likely to exclude zero. The nonspatial model can thus be seen to be overstating the contribution of some environmental parameters to the species distributions. It is perhaps not surprising that many of the coefficients for edaphic variables drop out in the spatial models, since these variables reflect underlying features of the terrain with strong local spatial association, pattern that will largely be taken into account by the spatial random effects. From this perspective the spatial models are preferable, since their spatial random effects take into account both spatial heterogeneity and spatial autocorrelation in the model residuals and thereby reduce the confounding of spatial autocorrelation with species niche relations.

Comparing the uncertainty surfaces for the four models graphically illustrates how nonspatial models may poorly characterize uncertainty. The variance in predicted probability of occurrence from Model 1 is dramatically lower than that for the three spatial models, and in particular shows no clear spatial pattern. Intuitively, it seems obvious that knowing whether a species has been observed nearby should affect one's ex-

pectation of encountering a species, and also one's level of uncertainty about this. But a nonspatial model, since it has no way of referencing grid cells or points in space, cannot respond to neighborhood information. Accordingly, whereas the spatial models produce variance surfaces that respond to the intensity and outcome of local sampling, the nonspatial model produces a fairly flat, low-intensity surface.

The spatial random effects themselves provide a characterization of spatial pattern in the data (see Fig. 7). As discussed above in Methods, the adjustment in the spatial random effects (ρ_i's) is based solely on the values of the spatial effect in neighboring cells (or sites). Where the value of the spatial effect is positive, those cells (or sites) are more suitable for the species than would be indicated by climate and edaphic features alone. The converse applies when the effect is negative. The spatial effects that were implemented as CARs (Models 1 and 3) show spatial structure both at very fine scales (sharp peaks and valleys in the surface), and at larger, subregional scales (e.g., an east–west trend for *Protea mundii*). The spatial effects surface for the kernel-smoothing model (Model 3) appears quite different, though it also indicates some areas of strong spatial association.

Model comparison for the four types of models we have presented is a difficult issue. In general, with binary data, there is no widely accepted criterion for choosing between hierarchical and nonhierarchical models (see Spiegelhalter et al. 2002, including the associated discussion). It is not entirely clear how to ascribe appropriate penalties for model complexity in such cases. Even more important in our case is the fact, though the *Y*'s are always Bernoulli trials with regard to presence/absence, the *p*'s we are defining are different across the models. That is, for Models 1 and 2, we ignore location for a trial in a given grid cell, for Model 3 we envision a *p* at every location in the region with pairwise dependence as a function of distance, for Model 4 the *p*'s are interpreted through environmental suitability as a surrogate for presence/absence. So, comparison among these models is a bit like comparing ''apples and oranges.'' Moreover, it is not necessarily sensible to reduce a model to a single number, as needed for model selection. It is arguably preferable to compare ''explanation'' across models with regard to environmental factors and ''prediction'' across the models with regard to species distribution patterns.

By the two quantitative performance measures presented here, the area under ROC curves and minimum predicted area, the spatial models clearly outperform the nonspatial Model 1, with the simple spatially explicit GLM, Model 2, generally falling somewhat short of the hierarchical Model 4 (see Table 4). The point-level Model 3 overpredicts the extent of the species distribution at high sensitivity levels, while Model 4 overall performs best at high sensitivity levels, as reflected in the MPA scores (see Table 4). In addition to improving the quantification of uncertainty, then, and identifying important environmental factors, making a model spatially explicit can increase its predictive power. The more complex hierarchical model has advantages in interpretability and also provides better prediction, giving the least overprediction at high sensitivity levels. If it is important to capture accurately all occurrences of a species, including peripheral ones that are difficult to capture with simpler models (e.g., *Protea mundii* in the extreme west), a model like Model 4 may be worth the extra work. These performance evaluations still involve an ''apples-to-oranges'' comparison in that the interpretation of p_i differs across the models, but from a purely functional perspective of comparing predictive ability, they provide a reasonable approach.

Using these models to predict the distributions of example species from the CFR has demonstrated that the spatial regression models can predict species distributions relatively well, and clearly better than a nonspatial model. These models also identify environmental variables that are strongly associated with their presence or absence. Here, for example, the two example species *Protea punctata* and *P. mundii* are shown to be distinguished by temperature, rainfall seasonality, and topographical variables, and less clearly by soil fertility. Of course, correlation is not causation, but such results provide hypotheses for testing; for example, the species can be transplanted across gradients of temperature and seasonal moisture availability to test whether the responses are consistent with model prediction, something that is now being done for these and a set of other, closely related species (A. M. Latimer, *unpublished data*).

For practical applications as well as for ecological inference, it is important to be able to quantify uncertainty. Obtaining the full probability distributions for parameter estimates is a major advantage of the Bayesian framework. Some potential applications for this model output include screening sites for reintroduction of a species or for reserve expansion. In these cases, the model could be used to select areas in which there is a high degree of certainty that the species can occur; this can be done by identifying cells that have credible intervals for predicted probability of occurrence (p_i) that lie above some threshold value, say 0.5. Using the posterior distributions enables the level of uncertainty to be specified, even when the distribution of the parameter estimate is skewed or multimodal. The most complex model presented here, Model 4, predicts species distributions in the absence of transformation as well as distributions under current conditions. The potential distribution prediction has obvious conservation planning applications, and by comparing the potential and transformed predictions, it is possible to present a precise estimate, again with uncertainty, of the extent

to which the species has been affected by transformation. Where transformation can be subdivided into different categories (e.g., agriculture, urbanization, alien invasive species), it is also possible to compare the effects of these various forms of transformation (see Latimer et al. 2004).

Characterizing species distributions, their limits, and their driving causes remains elusive, despite their close links to fundamental biogeographical questions relating to rarity, species richness and turnover (Gaston 2003). Going beyond ad hoc range maps to capture such features as holes in species distributions and uncertainty at edges requires probabilistic models that can specify a species range as a probability surface. The reality of species distributions is that species are never actually continuously distributed; the edge of a geographic range does not exist; rather ranges are continuously in flux in space and time. The spatial models we have presented can encapsulate aspects of this process via probability surfaces, along with a full specification of associated uncertainty. By allowing rigorous probabilistic inference about species distributions, these models should enable progress on problems relating to species distributions, including distinguishing different components of species distributions, such as area of occupancy, extent of occupancy, and prevalence, and testing relationships among prevalence, local abundance, and range size (see Gaston 2003). We have shown here that building such models in a Bayesian framework by adding spatial random effects to a GLM is relatively straightforward. With increasing computer power, such models can be fitted even to very large and rich data sets. Such models hold promise for untangling key questions about the causes and features of distributions of species as well their joint distribution, biodiversity.

Acknowledgments

This research was supported in part by NSF Grant DEB-008901 to J. A. Silander and A. E. Gelfand, and by NSF Grant DEB-0091961, supporting the Statistics in Ecology Workshop. The research was also conducted as part of the Bayesian Macro-Ecology Working Group and supported by the National Center for Ecological Analysis and Synthesis, a Center funded by NSF (Grant DEB-0072909), the University of California, and the Santa Barbara campus. A. M. Latimer's work was also funded by an NSF Graduate Research Fellowship. We thank Dr. Anthony G. Rebelo, director of the Protea Atlas Project of the South African National Botanical Institute (NBI), for providing the species distribution data and sharing his unparalleled knowledge of protea biogeography. We also thank Dr. Richard M. Cowling of the University of Port Elizabeth, Dr. Henri Laurie of the University of Cape Town, Dr. Alexandra Schmidt of the University of Rio de Janeiro, and Walter Smit of the National Botanical Institute, as well as Dr. Rebelo, for their invaluable participation the Bayesian Macro-Ecology Working Group and collaboration on the Protea Biogeography Project. Dr. Guy F. Midgley, head of NBI's climate change program, provided some of the climate data layers. Finally, we acknowledge the two anonymous reviewers whose comments have contributed greatly to the readability and, we hope, usefulness of this article.

Literature Cited

Agarwal, D. K., A. E. Gelfand, and J. A. Silander, Jr. 2002. Investigating tropical deforestation using two-stage spatially misaligned regression models. Journal of Agricultural Biological and Environmental Statistics **7**:420–439.

Augustin, N. H., M. A. Mugglestone, and S. T. Buckland. 1996. An autologistic model for the spatial distribution of wildlife. Journal of Applied Ecology **33**:339–347.

Austin, M. P., and J. A. Meyers. 1996. Current approaches to modelling the environmental niche of eucalypts: implications for management of forest biodiversity. Forest Ecology and Management **85**:95–106.

Austin, M. P., A. O. Nicholls, and C. R. Margules. 1990. Measurement of the realized qualitative niche: environmental niches of five *Eucalyptus* species. Ecological Monographs **60**:161–177.

Bannerjee, S., B. P. Carlin, and A. Gelfand. 2003. Hierarchical modeling and analysis for spatial data. Chapman and Hall, Boca Raton, Florida, USA.

Besag, J. 1974. Spatial interaction and the statistical analysis of lattice systems (with discussion). Journal of the Royal Statistical Society Series B **36**:192–236.

Brown, J. H., D. W. Mehlman, and G. C. Stevens. 1995. Spatial variation in abundance. Ecology **76**:2028–2043.

Carlin, B. P., and T. A. Louis. 2000. Bayes and empirical bayes methods for data analysis. Second edition. Chapman and Hall, New York, New York, USA.

Clark, J. S. 2003. Uncertainty and variability in demography and population growth: a hierarchical approach. Ecology **84**:1370–1381.

Congdon, P. 2001. Bayesian statistical modelling. John Wiley and Sons, Chichester, UK.

Cowling, R. M., D. M. Richardson, R. J. Schultze, M. T. Hoffman, J. J. Midgley, and C. Hilton-Taylor. 1997. Species diversity at the regional scale. Pages 447–473 *in* R. M. Cowling, D. M. Richardson, and S. M. Pierce, editors. Vegetation of southern Africa. Cambridge University Press, Cambridge, UK.

D'heygere, T., P. L. M. Goethals, and N. De Pauw. 2003. Use of genetic algorithms to select input variables in decision tree models for the prediction of benthic macroinvertebrates. Ecological Modelling **160**:189–300.

Ellison, A. M. 2004. Bayesian inference in ecology. Ecology Letters **7**:509–520.

Engler, R., A. Guisan, and L. Rechsteiner. 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.

Ettema, C. H., S. L. Rathbun, and D. C. Coleman. 2000. On spatiotemporal patchiness and the coexistence of five species of *Chronogaster* (Nematoda: Chronogasteridae) in a riparian wetland. Oecologia **125**:444–452.

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

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

Gaston, K. J. 2003. The structure and dynamics of geographic ranges. Oxford University Press, New York, New York, USA.

Gelfand, A., D. Agarwal, and J. A. Silander, Jr. 2002. Investigating tropical deforestation using two stage spatially misaligned regression models. Journal of Agricultural, Biological and Environmental Statistics **7**:420–439.

Gelfand, A. E., A. M. Schmidt, S. Wu, J. A. Silander, A. Latimer, and A. G. Rebelo. 2005*a*. Explaining species diversity through species level hierarchical modeling. Applied Statistics **65**:1–20.

Gelfand, A. E., J. A. Silander, Jr., S. Wu, A. Latimer, P. O. Lewis, A. G. Rebelo, and M. Holder. 2005*b*. Explaining species distribution patterns through hierarchical modeling. Bayesian Analysis **1**:41–92.

Gelman, A., J. B. Carlin, and D. B. Rubin. 1995. Bayesian data analysis. CRC Press, Boca Raton, Florida, USA.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter. 1995. Markov chain Monte Carlo in practice. CRC Press, Boca Raton, Florida, USA.

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

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

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

Higdon, D., J. Swall, and J. Kern. 1999. Nonstationary spatial modeling. Pages 761–768 *in* J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors. Bayesian statistics 6. Oxford University Press, Oxford, UK.

Hooten, M. B., D. R. Larsen, and C. K. Wikle. 2003. Predicting the spatial distribution of ground flora on large domains using a hierarchical Bayesian model. Landscape Ecology **5**:487–502.

Latimer, A. M., J. A. Silander, Jr., A. E. Gelfland, A. G. Rebelo, and D. M. Richardson. 2004. Quantifying threats to biodiversity from invasive alien plants and other factors: a case study from the Cape Floristic region. South African Journal of Science **100**:81–86.

Leathwick, J. R. 2002. Intra-generic competition among *Nothofagus* in New Zealand's primary indigenous forests. Biodiversity and Conservation **11**:2177–2187.

Lehmann, A., J. M. Overton, and J. R. Leathwick. 2002. GRASP: generalized regression analysis and spatial prediction. Ecological Modelling **157**:189–207.

Link, W. A., and J. R. Sauer. 2002. A hierarchical analysis of population change with application to Cerulean Warblers. Ecology **83**:2832–2840.

MacArthur, R. H., H. F. Recher, and M. L. Cody. 1966. On the relation between habitat selection and species diversity. American Naturalist **100**:319–332.

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

Midgley, G. F., L. Hannah, D. Millar, M. C. Rutherford, and L. W. Powrie. 2002. Assessing the vulnerability of species richness to anthropogenic climate change in a biodiversity hotspot. Global Ecology and Biogeography **11**:445–451.

Moisen, G. G., and T. S. Frescino. 2002. Comparing five modelling techniques for predicting forest characteristics. Ecological Modelling **157**:209–225.

Mugglin, A. S., B. P. Carlin, and A. Gelfand. 2000. Fully model based approaches for spatially misaligned data. Journal of the American Statistical Association **95**:877–887.

Peterson, A. T. 2003. Predicting the geography of species' invasions via ecological niche modeling. Quarterly Review of Biology **78**:419–433.

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

Rouget, M., D. M. Richardson, R. M. Cowling, J. W. Lloyd, and A. T. Lombard. 2003. Current patterns of habitat transformation and future threats to biodiversity in terrestrial ecosystems of the Cape Floristic Region, South Africa. Biological Conservation **112**:63–85.

Spiegelhalter, D. J., N. Best, B. P. Carlin, and A. van der Linde. 2002. Bayesian measures of model complexity and fit (with discussion). Journal of the Royal Statistical Society Series B **64**:583–639.

Stoyan, H., H. De-Polli, S. Bohm, G. P. Robertson, and E. A. Paul. 2000. Spatial heterogeneity of soil respiration and related properties at the plant scale. Plant and Soil **222**: 203–214.

Thomas, C. D., et al. 2004. Extinction risk from climate change. Nature **427**:145–148.

Turchin, P. 1998. Quantitative analysis of movement: measuring and modeling population redistribution in animals and plants. Sinauer, Sunderland, Massachusetts, USA.

Ver Hoef, J. M., N. Cressie, R. N. Fisher, and T. J. Case. 2001. Uncertainty and spatial linear models for ecological data. Pages 214–237 *in* C. T. Hunsaker, M. F. Goodchild, M. A. Friedl, and T. J. Case, editors. Spatial uncertainty in ecology. Springer-Verlag, New York, New York, USA.

Wikle, C. K. 2003. Hierarchical Bayesian models for predicting the spread of ecological processes. Ecology **84**: 1382–1394.

APPENDIX

A table describing environmental data layers (*Ecological Archives* A016-003-A1).

SUPPLEMENT

Model code for WinBUGS (*Ecological Archives* A016-003-S1).