Journal of Applied Ecology 2006
43, 386–392

GUEST EDITORIAL

Making better biogeographical predictions of species' distributions

ANTOINE GUISAN,* ANTHONY LEHMANN,† SIMON FERRIER,‡ MIKE AUSTIN,§ JACOB MC. C. OVERTON,¶ RICHARD ASPINALL** and TREVOR HASTIE††

**University of Lausanne, Department of Ecology and Evolution, Biology Building, CH-1015 Lausanne, Switzerland; †Swiss Center for Faunal Cartography, Terreaux 14, CH-2000 Neuchâtel, Switzerland; ‡New South Wales Department of Environment and Conservation, PO Box 402 Armidale, NSW 2350, Australia; §CSIRO Sustainable Ecosystems, GPO Box 284, Canberra, ACT 2601, Australia; ¶Landcare Research, Private Bag 3127, Hamilton, New Zealand; **Arizona State University, Department of Geography, PO Box 870104, Tempe, AZ 85287–0104, USA; and ††Stanford University, Statistics Department, Sequoia Hall, Stanford, CA 94305, USA*

Summary

1. Biogeographical models of species' distributions are essential tools for assessing impacts of changing environmental conditions on natural communities and ecosystems. Practitioners need more reliable predictions to integrate into conservation planning (e.g. reserve design and management).

2. Most models still largely ignore or inappropriately take into account important features of species' distributions, such as spatial autocorrelation, dispersal and migration, biotic and environmental interactions. Whether distributions of natural communities or ecosystems are better modelled by assembling individual species' predictions in a bottom-up approach or modelled as collective entities is another important issue. An international workshop was organized to address these issues.

3. We discuss more specifically six issues in a methodological framework for generalized regression: (i) links with ecological theory; (ii) optimal use of existing data and artificially generated data; (iii) incorporating spatial context; (iv) integrating ecological and environmental interactions; (v) assessing prediction errors and uncertainties; and (vi) predicting distributions of communities or collective properties of biodiversity.

4. *Synthesis and applications*. Better predictions of the effects of impacts on biological communities and ecosystems can emerge only from more robust species' distribution models and better documentation of the uncertainty associated with these models. An improved understanding of causes of species' distributions, especially at their range limits, as well as of ecological assembly rules and ecosystem functioning, is necessary if further progress is to be made. A better collaborative effort between theoretical and functional ecologists, ecological modellers and statisticians is required to reach these goals.

Key-words: artificial data, autocorrelation, community and diversity modelling, errors and uncertainties, generalized regressions, interactions, niche-based model

Journal of Applied Ecology (2006) **43**, 386–392
doi: 10.1111/j.1365-2664.2006.01164.x

Introduction

In 2001, participants in a workshop on predicting geographical distributions of organisms (Guisan 2002; Lehmann, Overton & Austin 2002) concluded that more robust biogeographical models are essential for environmental management and for assessing impacts

Correspondence: Antoine Guisan, University of Lausanne, Department of Ecology and Evolution, Biology Building, CH-1015 Lausanne, Switzerland (fax +41 21 692 42 65; e-mail antoine.guisan@unil.ch).

of changing environmental conditions, including climate change, on natural communities and ecosystems. Most models used at that time still largely ignored, or inappropriately took into account, important features of species' distributions, such as spatial autocorrelation, dispersal, migration and biotic and environmental interactions. Whether distributions of natural communities or ecosystems were better modelled by assembling individual species' predictions in a bottom-up approach or modelled as collective entities was another important issue identified. Where are we now with these issues? How much remains to be done? Recent papers dealing with these issues have tended to focus on a single issue at a time, for example spatial autocorrelation or species' interactions, but not both. Despite the proliferation of studies modelling species' distributions, most models published in recent years still suffer from the same set of limitations identified in 2001.

Discussion at a follow-up workshop in 2004 focused particularly on: (i) strengthening the link between ecological theory and modelling tools; (ii) taking better advantage of existing data, including occurrence data from herbaria and museums, and artificially generated data; (iii) giving more consideration to spatial context in modelling; (iv) recognizing and integrating ecological and environmental interactions; (v) assessing errors and uncertainties associated with predictions; and (vi) extending species-level modelling approaches to predict distributions of higher-level entities (e.g. communities, ecosystems) or collective properties of biodiversity (e.g. species richness, beta diversity).

We discuss these issues primarily within the context of 'generalized regression' modelling methods. This broad family of methods is by far the most widely applied approach to species modelling and includes generalized linear models (GLM), generalized additive models (GAM), vector GLM and GAM (VGLM/VGAM), multiple additive regression splines (MARS) and generalized linear and additive mixed models (GLMM/GAMM). Concentrating on this one family of models helped to focus discussion at the workshop on the six major issues outlined above, rather than on detailed comparisons between individual modelling techniques. Other methods, such as classification and regression trees (CART), were considered only when they provided complementary solutions to regression methods.

Strengthening the link between ecological theory and modelling tools

One of the first steps in building predictive distribution models is to assume a conceptual model of the expected species–environment relationships, before then fitting parameters to this model using existing biological data and environmental variables. Strengthening the link between ecological theory and statistical models is thus an important step for improving biogeographical models (Austin 2002) and for making more effective use of these models in tackling conservation issues (Rushton, Ormerod & Kerby 2004). Some exploratory tools commonly used in ecology are based on ecologically unrealistic working assumptions. For instance, canonical correspondence analysis (CCA) relies on regular species packing and equal optima and amplitude of species. An alternative method, canonical Gaussian ordination (CGO; Yee 2004a), has been developed recently that has less constraints and thus is more likely to reflect ecological reality. It is similar to CCA but based on GLM rather than least squares, so that it can accommodate non-normally distributed errors (e.g. Poisson or binomial) and allows irregularly spaced species' optima along environmental gradients and unequal amplitude across species (Yee 2004a). Another promising exploratory tool to build conceptual models is structural equation modelling (SEM; Grace & Pugesek 1997), a modern version of path modelling that allows investigation of partial correlations between variables and the identification of more proximal (i.e. causal) relationships, thereby distinguishing between direct and indirect predictors. The output from SEM might thus be used to make better informed selection of predictors as input to predictive models, before applying an optimized selection procedure (e.g. shrinkage rules). Alternatively, multimodel inference can be applied to a set of competing models reflecting different biological hypotheses (Johnson & Omland 2004).

As another example, advances have been made in identification of ecologically meaningful species' response curves along environmental gradients (Austin 2002), suggesting that unimodal-skewed responses are common in ecology and can be adequately reproduced by semi-parametric methods such as GAM and their extensions (e.g. VGAM, Yee & Wild 1996; GAMM, Wood 2004). However, these unimodal responses are expected to hold only for regulator variables, such as temperature. As suggested by Liebig's law of the minimum, they can be further modulated by a species' response to resource variables, such as soil nutrients and light, which vary in space and time (e.g. being depleted locally by another species), resulting 'in a continual shifting of limitation from one factor to another' (Huston 2002). Thus, the true response curve of a species, measured in terms of probability of presence, abundance or fitness, to a given regulator or resource variable can be quantified only when all other factors occur at non-limiting levels, an unlikely situation with observations from the natural world, thus requiring the use of non-standard statistical methods (Huston 2002). The most promising non-standard method is quantile regression, where the lower- and upper-bound of data (e.g. 5 and 95 percentiles) are modelled rather than the mean. This method can be used to model any quantile, meaning that the full distribution of data points is modelled, whereas more standard regression methods only model one property of this distribution.

Statistical solutions for computing quantile regression are now available in standard software (e.g. Splus and R), as in Yee's VGAM package (Yee 2004b). Recent examples of the application of this approach to modelling species' responses to environmental variables are Knight & Ackerly (2002), relating species' DNA content to temperature, and Schroder, Anderson & Kiehl (2005), relating fen species to flooding and other variables.

Taking better advantage of existing data and artificially generated data

A wide array of data on species' distributions is now available to the scientific community, from local to global scales. Important sources include the many natural history collections (NHC) stored in museums, herbaria and national biological data banks world-wide (Graham *et al*. 2004). However, these are usually spatially and temporally heterogeneous samples without any absence information available and containing unknown levels of bias and error. A challenge is thus to take maximum advantage of these data by developing appropriate methods that accommodate their special characteristics. For instance, while several methods are now available to fit models using such 'occurrence' (i.e. presence-only) data, for example bioclimatic envelopes, ecological niche factor analysis and standard methods employing pseudo-absences (Pearce & Boyce 2006), problems remain in the evaluation of predictions from these models. Evaluating habitat suitability predictions with presence-only data is not possible using standard agreement measures [e.g. kappa or the area under the curve (AUC) of a ROC plot] and this is therefore an issue in need of further investigation. Improvements to existing presence-only evaluation techniques (e.g. the area-adjusted frequency index; Boyce *et al*. 2002; Pearce & Boyce 2006) are currently being investigated and new developments should soon be made available, for example with the BIOMAPPER software (Hirzel *et al*. 2002; Hirzel *et al*. in press).

Another important issue is the use of artificial data to answer specific questions, be they statistical, methodological or ecological. The basic principle here is to build a response variable and a set of predictor variables in such a way that the underlying relationships (truth) are known. Predictions from models fitted to training data generated from these known relationships (e.g. using probability sampling, with or without added noise) can then be evaluated against the truth. Early developments were made by Minchin with his COMPAS software in the late 1980s (Minchin 1987), allowing a set of artificial environmental gradients to be generated together with a set of virtual species with defined ecological response curves for each of these gradients. Austin *et al*. (in press) have used COMPAS to explore the relative performance of GLM and GAM regressions where 'truth' is known. Using simple theoretical models of species' responses to environmental variables by analysts unaware of truth, they demonstrate that the different methods achieve similar success. They concluded that ecological insight into the nature of environmental variables and statistical skill are more important than the precise method used.

COMPAS simulates virtual species in a virtual landscape, but intermediate solutions have become popular more recently, such as simulating virtual species in a real landscape (Hirzel, Helfer & Metral 2001; Moisen & Frescino 2002). However, an acute problem is to know just how realistic such virtual species and landscape are, and therefore the extent to which this approach can help answer specific questions related to species distribution modelling. This is likely to depend on the specific objective of the analysis. One approach is to start from a perfect model, where truth is entirely known, and to degrade the data progressively by adding noise (residuals) or spatial autocorrelation, to see how it affects the model and its properties. Further clarity of thinking is still needed in this field.

Giving more consideration to spatial context in our models

Even though current models of species' distributions are often said to be 'spatial', in most cases they are only partially spatial. The species–environment relationship is often fitted without explicit consideration of the neighbouring spatial context, for example without taking spatial autocorrelation or dispersal into account. These models are therefore 'spatially invariant (or neutral)', in that permuting the points and their associated species and environmental data in geographical space, by permuting the $<x;y>$ coordinates, would not alter the fitted species–environment relationship. In the 'overlay mode' employed here, geographical position is used simply as a means of attaching environmental attributes through GIS overlay, thereby locating each survey location in environmental space defined by the Hutchinsonian multidimensional niche (Hutchinson 1975). Predictions generated by models fitted within environmental space are then projected back onto the geographical space. Spatially invariant models do not address a number of important spatial factors, for example whether a species is more likely to occur at a location that is surrounded by other occurrences of the species, or whether a species might be prevented from migrating to a location because of geographical or environmental barriers (e.g. landscape fragmentation) or because of insufficient speed of migration (Pulliam 2000).

How can such spatial factors be more effectively addressed in species modelling? We propose a multiscale hierarchical framework, based on pioneering work by Legendre (1993), that takes spatial patterns and processes into account. First, the framework starts from the neutral spatially invariant model. Secondly, strong geographical gradients at large extent and coarse grain are removed, for example through

incorporating trend surface analysis in the model. Thirdly, possible interactions between the strictly environmental (neutral) and geographical components are considered. Finally, the local neighbouring context around each site (i.e. smaller extent and fine grain) is taken into account, for example through fitting a spatial autoregressive model (Segurado, Araújo & Kunin 2006). More direct methods for fitting spatial autoregressive models in GLM or GAM now exist, such as GLMM and GAMM (Wood 2004); the application of these to species modelling was discussed during the workshop. Consideration of these different spatial levels will allow further investigation of the respective role of large-scale migration vs. local dispersal processes (Ronce 2001), which might then be implemented in more dynamic models of species dispersal and migration (using, for example, cellular automaton; Carey 1996). From a different perspective, spatial structures in the data can also generate bias in predictive models by inflating the significance level, and various solutions have been proposed to overcome the problem (Segurado *et al.* 2006).

Recognizing and integrating ecological and environmental interactions

Most models consider species' distributions as being shaped purely by environmental constraints, ignoring ecological interactions. Two types of ecological interaction are of interest here: (i) biotic interactions, in which the distribution of one species is influenced by the distribution of other species; and (ii) predictor interactions, in which the effect of one environmental predictor on a species varies according to the levels of other predictors. Biotic interactions can be within the same group (e.g. competition, facilitation and parasitism in plants) or between groups, as best exemplified by relationships in food webs (herbivory, predation and symbiosis). Such interactions can be addressed by adding one or more biotic predictors to the model formula alongside the environmental predictors (Leathwick & Austin 2001). A major issue is therefore identifying which biotic predictors to include in models, and at which spatial and temporal scales (e.g. competitive interactions may act at a more local scale than other types of interaction; Huston 2002). While a basic approach to selecting biotic predictors is to rely on ecological theory (e.g. assembly rules or food webs), many modellers may wish to adopt an automated procedure. Again, SEM can provide good support here, by elucidating positive or negative correlations between species. When many species need to be modelled jointly, and each of them can potentially act as a predictor of the others, a system of simultaneous parallel regressions is required, using VGAM or the SME approach (a challenge already addressed by Guisan & Zimmerman 2000).

Most generalized regression approaches can incorporate pairwise (or higher-order) interactions between environmental/biotic predictors. However, an obvious problem is that the number of possible combinations of predictors increases rapidly with increasing numbers of predictors and quickly becomes unmanageable using classical selection approaches, for example stepwise selection. Thus automatic methods, identifying significant interaction terms, are highly desirable. An elegant solution, first proposed by Hastie (Hastie, Tibshirani & Friedman 2001) and recently implemented in the GRASP modelling package (Lehmann, Overton & Leathwick 2002), is to fit a weighted classification tree (CART) to the residuals of a GLM or GAM model, using the same predictor variables (biotic and/or abiotic), then define a factor variable from the terminal nodes and refit the GLM or GAM model with that factor additionally included. Each class of the factor then represents a set of rules that can be interpreted by looking at the path(s) leading to that particular class in the tree.

Assessing errors and uncertainties associated with predictions

A key issue for environmental managers wanting to use predictions from these models is reliability. What is the error and uncertainty associated with a model and its spatial prediction, for example in the form of a habitat suitability map? How does the level of uncertainty vary across a study area? Uncertainty is a difficult issue, primarily because there are many different types of error and associated uncertainty. These can be: (i) measurement error, for instance caused by low detectability, which itself can be a function of environmental predictors; (ii) systematic error, as, for instance, caused by an accidental shift of an environmental grid in the GIS; (iii) model error, as, for instance, resulting from the choice of an inappropriate probability distribution in a GLM or GAM; or (iv) natural variation and subjective judgement (Elith, Burgman & Regan 2002; Barry & Elith 2006). Errors also propagate throughout the modelling process and intermingle into an overall compound error. This overall error is the one assessed in most studies, when final predictions are compared with independent observations, or semi-independent observations in the case of resampling schemes such as bootstrapping and cross-validation. However, partitioning out the different error components might provide more useful information for further improving models. Such partitioning could help in deciding where to direct the greatest energy to reduce overall error, for example in improving the accuracy of predictors or improving the measurement of the biological response. Developing an integrated framework for assessing uncertainty, and tracing error propagation throughout the modelling process, is another research direction worthy of more attention. Pending the results of such research, the best practice in the interim is to provide at least a proper, spatially explicit assessment of model predictions, for instance by considering confidence intervals around predictions and drawing maps of 95%

confidence limits or standard error across the study area (Aspinall 1992). This can provide the support required to interpret prediction maps and reduces the risk of these maps being misused, for example by practitioners, or in further analyses or meta-analyses combining these predictions to reveal, or test, collective properties of communities or ecosystems (e.g. diversity, community structure and ecosystem function).

Modelling higher-level entities or collective properties of biodiversity

Three possible strategies exist for modelling higher-level entities such as assemblages, communities and ecosystems, or collective properties of biodiversity such as species richness and beta diversity (Ferrier & Guisan 2006), all of which rely potentially on regression techniques: (i) assemble first, predict later; (ii) predict first, assemble later; (iii) assemble and predict simultaneously. In the first strategy, assemblages or communities are derived directly from a biological survey data set using classical reduction techniques (classification or ordination) commonly applied in community ecology and phytosociology. The distribution of each entity (e.g. community) derived in this first stage is then modelled as a function of environmental predictors. This approach therefore assumes that the composition of species' assemblages is stable over time and geographical space. The second strategy starts by predicting the distribution of large numbers of individual species and then subjects this 'stack' of predicted distributions to classification or ordination, thereby building communities in a bottom-up manner. The third strategy employs environmental constraints directly in the classification or ordination of biological survey data, and thus it encompass all constrained (canonical) ordination and classification techniques, like CCA and CGO, and constrained classification via recursive partitioning (Ferrier *et al*. 2002).

The second strategy, assembling communities from individual species' predictions, is the newest of the three and represents a growing field of interest in ecology (Ferrier & Guisan 2006). Previous applications of this approach have raised many questions, especially in relation to predicting responses to environmental change such as global warming. A first, basic question with this approach is how to assemble species to predict diversity and community composition. Should we develop more realistic assembly rules or constraints that limit how assemblages are selected from a larger species pool, based on functional traits and competitive hierarchies, and use them to simulate communities? Another, more fundamental, question is how far can species models based on the realized niche (but see Pulliam 2000 for other possible cases) succeed in predicting future changes in community composition, knowing that the physiology of species, and interactions between species, which currently define realized niches might themselves change in response to a changing environment (Ackerly 2003). These questions are still open and deserve more investigation.

Furthermore, all previously discussed limitations and possible improvements of species distribution models need to be taken into consideration when interpreting predicted assemblages. For instance, how do all of the uncertainties associated with individual species distribution models combine into an overall level of uncertainty for each derived assemblage? Does consideration of a large number of species in a single process result in a 'bootstrap effect' that might reduce the overall amount of uncertainty in predicted patterns of community distribution?

All three strategies described above can also be used to model collective properties of biodiversity such as species richness and beta diversity or compositional turnover. One example is general dissimilarity modelling (GDM), a new non-linear extension of matrix regression, that models turnover in community composition between sites as a direct function of separation in both environmental and geographical space (Ferrier *et al*. 2004; Ferrier & Guisan 2006).

New statistical tools available to ecologists

Several new packages dedicated to GLM and GAM modelling have recently been made available to ecologists. The GRASP package, developed for Splus by A. Lehmann and colleagues (Lehmann, Overton & Leathwick 2002) and for R by F. Fivaz (www.cscf.ch/grasp), includes many new developments and provides a significant support to ecologists for building spatial predictions in a generalized regression framework. The BIOMOD package (Thuiller 2003) is another such ecological GLM/GAM tool for R and Splus that additionally allows fitting CART (classification and resgression trees), ANN (artificial neural network), BRT (boosted regression trees) and a few additional techniques. BRT, in particular, was shown recently to be among the most powerful techniques (Elith *et al.* in press). Two new libraries now allow building GAM models in R (http://cran.r-project.org; last visited 16 March 2006), an improved one by S. Wood (2004, MGCV; the original in R) and a new one by T. Hastie (2005, GAM). MGCV additionally allows fitting GLMM and GAMM (e.g. to fit autocorrelative models). Another R library – VGAM – allows fitting VGLM and VGAM (Yee & Wild 1996), and also offers an other means of performing quantile regressions, CGO (Yee 2004b) and other advanced modelling methods (e.g. fitting several species models at once). New powerful predictive methods other than generalized regressions include MAXENT (Maximum Entropy; Phillips *et al.* 2005) or SVM (Support Vector Machines; Drake *et al.* 2006).

Conclusions

There remains great scope for further improvements to biogeographical modelling of species' distributions, a

task facilitated by new statistical tools recently made available to ecologists (see Guisan & Thuiller 2005). The challenges involved in making better predictions of species' distributions are both applied and theoretical. Practitioners need reliable predictions of species' distributions to evaluate properly the impact of climate and land-use changes on the distribution, composition, structure and functioning of community and ecosystems. These applications are required to assess, for example, the value of current reserve networks and the ability of ecosystems to provide human societies with expected goods and services, both now and in the future. From a theoretical point of view, better predictions of biological communities and ecosystems can emerge only from (i) more robust species' distribution models and better documentation of the uncertainty associated with these models; and (ii) an improved understanding of causes of species' distributions, especially at their range limits, as well as of ecological assembly rules and ecosystem functioning. A better collaborative effort between theoretical and functional ecologists, ecological modellers and statisticians is required to achieve these goals.

Acknowledgements

This paper greatly benefited from discussions at the Riederalp workshop. The following scientists additionally attended (alphabetic order) and should be acknowledged: M. Araùjo (P, UK), S. Barry (AU), M. Boyce (CA), C. Coudun (F), T. Dirnböck (AUS), T. C. Edwards (USA), J. Elith (AU), E. Fleishmann (USA), E. Heegaard (NO), A. Hirzel (Switz.), J. Leathwick (NZ), G. Le Lay (Switz.), S. Manel (F), J. Miller (USA), G. Moisen (USA), P. Osborne (UK), J. Pearce (Canada), K. Van Neil (AU), M. Wisz (DK), S. Wood (UK), T. Yee (NZ), N. G. Yoccoz (NO), N. E. Zimmermann (Switz.) and seven staff PhD students. The workshop was financially supported by the 'Fondation Herbette', the 'Fondation du 450e', the Faculty of Geoscience, and the Department of Ecology and Evolution at the University of Lausanne (Switz.) and by the Swiss Academy of Science. Two other special issues from the workshop are currently in press in the *Journal of Biogeography* (Araújo & Guisan, in press) and *Ecological Modelling* (Moisen, Osborne & Edwards, in press).

References

Ackerly, D.D. (2003) Community assembly, niche conservatism, and adaptive evolution in changing environments. *International Journal of Plant Science*, **164**, S165–S184.

Aspinall, R. (1992) An inductive modeling procedure based on Bayes theorem for analysis of pattern in spatial data. *International Journal of Geographical Information Systems*, **6**, 105–121.

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

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

Barry, S.C. & Elith, J. (2006) When things go wrong: error and uncertainty in habitat models. *Journal of Applied Ecology*, **43**, 413–423.

Boyce, M.S., Vernier, P.R., Nielsen, S.E. & Schmiegelow, F.K.A. (2002) Evaluating resource selection functions. *Ecological Modelling*, **157**, 281–300.

Carey, P.D. (1996) DISPERSE: a cellular automaton for predicting the distribution of species in a changed climate. *Global Ecology and Biogeography Letters*, **5**, 217–226.

Drake, J.A., Randin, C. & Guisan, A. (2006) Modelling ecological niches with support vector machines. *Journal of Applied Ecology*, **43**, 424–432.

Elith, J., Burgman, M.A. & Regan, H.M. (2002) Mapping epistemic uncertainties and vague concepts in predictions of species distribution. *Ecological Modelling*, **157**, 313–329.

Ferrier, S. & Guisan, A. (2006) Spatial modelling of biodiversity at the community level. *Journal of Applied Ecology*, **43**, 393–404.

Ferrier, S., Drielsma, M., Manion, G. & Watson, G. (2002) Extended statistical approaches to modelling spatial pattern in biodiversity in north-east New South Wales. II. Community-level modelling. *Biodiversity and Conservation*, **11**, 2309–2338.

Ferrier, S., Powell, G.V.N., Richardson, K.S., Manion, G., Overton, J.M., Allnutt, T.F., Cameron, S.E., Mantle, K., Burgess, N.D., Faith, D.P., Lamoreux, J.F., Kier, G., Hijmans, R.J., Funk, V.A., Cassis, G.A., Fisher, B.L., Flemons, P., Lees, D., Lovett, J.C. & Van Rompaey, R.S.A.R. (2004) Mapping more of terrestrial biodiversity for global conservation assessment: a new approach to integrating disparate sources of biological and environmental data. *Bioscience*, **54**, 1101–1109.

Grace, J.B. & Pugesek, B.H. (1997) A structural equation model of plant species richness and its application to a coastal wetland. *American Naturalist*, **149**, 436–460.

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

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

Hastie, T., Tibshirani, R. & Friedman, J. (2001) *The Elements of Statistical Learning*. Springer Verlag, Berlin.

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

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

Hirzel, A., Le Lay, G., Helfer, V., Randin, C. & Guisan, A. (in press) Evaluating predictive habitat distribution models with presence-only data: a new method. *Ecological Modelling*.

Huston, M.A. (2002) Introductory essay: critical issues for improving predictions. *Predicting Species Occurrences: Issues of Accuracy and Scale* (eds J.M. Scott, P.J. Heglund, M.L. Morrison, J.B. Haufler, M.G. Raphael, W.A. Wall & F.B. Samson), pp. 7–21. Island Press, Covelo, CA.

Johnson, J.B. & Omland, K.S. (2004) Model selection in ecology and evolution. *Trends in Ecology and Evolution*, **19**, 101–108.

Knight, C.A. & Ackerly, D.D. (2002) Variation in nuclear DNA content across environmental gradients: a quantile regression analysis. *Ecology Letters*, **5**, 66–76.

Leathwick, J.R. & Austin, M.P. (2001) Competitive interactions between tree species in New Zealand's old-growth indigenous forests. *Ecology*, **82**, 2560–2573.

Legendre, P. (1993) Spatial autocorrelation: trouble or new paradigm? *Ecology*, **74**, 1659–1673.

Lehmann, A., Overton, J.M. & Austin, M.P. (2002) Regression models for spatial prediction: their role for biodiversity and conservation. *Biodiversity and Conservation*, **11**, 2085–2092.

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

Minchin, P.R. (1987) Simulation of multidimensional community patterns: toward a comprehensive model. *Vegetatio*, **71**, 145–156.

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

Pearce, J. & Boyce, M. (2006) Modelling distribution and abundance with presence-only data. *Journal of Applied Ecology*, **43**, 405–412.

Phillips, S.J., Dudik, M. & Schapire, R.E. (2005) Maximum entropy modeling of species geographic distributions. *Ecological Modeling*, **190**, 231–259.

Pulliam, H.R. (2000) On the relationship between niche and distribution. *Ecology Letters*, **3**, 349–361.

Ronce, O. (2001) Understanding plant dispersal and migration. *Trends in Ecology and Evolution*, **16**, 663.

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

Schroder, H.K., Anderson, H.E. & Kiehl, K. (2005) Rejecting the mean: estimating the response of fen plant species to environmental factors by non-linear quantile regression. *Journal of Vegetation Science*, **16**, 373–382.

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

Thuiller, W. (2003) BIOMOD: optimizing predictions of species distributions and projecting potential future shifts under global change. *Global Change Biology*, **9**, 1353–1362.

Wood, S.N. (2004) *Low Rank Scale Invariant Tensor Smooths for Generalized Additive Mixed Models*. Department of Statistics, University of Glasgow, Glasgow, UK.

Yee, T.W. (2004a) A new technique for maximum likelihood canonical Gaussian ordination. *Ecological Monographs*, **74**, 685–701.

Yee, T.W. (2004b) Quantile regression via vector generalized additive models. *Statistics in Medicine*, **23**, 2295–2315.

Yee, T.W. & Wild, C.J. (1996) Vector generalized additive models. *Journal of the Royal Statistical Society, B*, **58**, 481–493.

Received 18 November 2005; final copy received 20 January 2006
Editor: Rob Freckleton