ECOGRAPHY 27: 165–172, 2004

Effects of restricting environmental range of data to project current and future species distributions

Wilfried Thuiller, Lluis Brotons, Miguel B. Araújo and Sandra Lavorel

Thuiller, W., Brotons, L., Araújo, M. B. and Lavorel, S. 2004. Effects of restricting environmental range of data to project current and future species distributions. – Ecography 27: 165–172.

We examine the consequences of restricting the range of environmental conditions over which niche-based models are developed to project potential future distributions of three selected European tree species to assess first, the importance of removing absences beyond species known distributions ("naughty noughts") and second the importance of capturing the full environmental range of species. We found that restricting the environmental range of data strongly influenced the estimation of response curves, especially towards upper and lower ends of environmental ranges. This induces changes in the probability values towards upper and lower environmental boundaries, leading to more conservative scenarios in terms of changes in distribution projections.

Using restricted data analogous to not capturing the fun species' environmental range, reduces strongly the combinations of environmental conditions under which the models are calibrated, and reduces the applicability of the models for predictive purposes. This may generate unpredictable effects on the tails of the species response curves, yielding spurious projections into the future provided that probability of occurrence is not set to zero outside the environmental limits of the species. Indeed, as the restricted data does not capture the whole of the response curve, projections of future species distributions based of ecological niche modelling may be only valid if niche models are able to approach the complete response curve of environmental predictors.

W. Thuiller (thuiller@cefe.cnrs-mop.fr), L. Brotons and M. B. Araújo, Centre d'Ecologic Fonctionelle et Evolutive, CNRS, 1919 route de Mende, F-34293 Montpellier Cedex 5, France. – S. Lavorel, Lab. d'ecologie Alpine, CNRS, Univ. J. Fournier, BP 53X, F-38041 Grenoble Cedex, France.

There is an increasing interest in assessing potential impacts of climate change on species distributions. Amongst modelling techniques available, niche-based models, which are designed to approximate species' ecological niches (Austin et al. 1990), include some of the most popular methods to project future potential species distributions (Thuiller 2003).

In spite of the popularity of these models (Guisan and Zimmermann 2000, Thuiller et al. 2003a), there are difficulties in projecting species distributions into areas or times different from those used to calibrate models (Loehle and LeBlanc 1996, Woodward and Beerling 1997, Lawton 2000). This is because modelled species' realised niches capture many factors other than species limits of tolerance to environmental variables (e.g. competition and historical contingency, Araújo et al. 2001, Leathwick and Austin 2001) and these factors are not easily predictable (Loehle and LeBlanc 1996). However, it is possible that increasing the spatial resolution and extent of the studied area might help reducing the importance of unpredictable inter-specific competition in the projections (Beerling et al. 1995).

Accepted 30 September 2003

Copyright © ECOGRAPHY 2004
ISSN 0906-7590

One of the critical issues for making reasonable projections of species distributions into different spatial or temporal scenarios is to have appropriate descriptions of species realised niches. Realised niches are commonly approximated by series of curves describing the likelihood of species' occurrence following a set of environmental predictors using presence/absence or abundance data.

One of the problems with this procedure is that modellers often lack information on the overall species potential or realised distributions. Hence, response curves are often incomplete descriptions of the responses of species to environmental predictors. Different authors have suggested different approaches to tackle this problem. For example, Austin and Meyers (1996) argued that the inclusion of zeros beyond species known distributions ("the naughty noughts") might perturb the correct estimation of response curves and suggested restricting model calculations within every species known environmental range. Since statistical models can fit extreme zeros reasonably, others authors argued that such restrictions are unnecessary and may lead to ecologically unrealistic response curves (Oksanen and Minchin 2002).

On the other hand, biased availability of environmental conditions are often encountered when modelling distributions of a large number of species over large areas a common feature when assessing for instance biogeographic effects of climate change on communities (Bakkenes et al. 2002). This is mostly due to restrictions in data accessibility and sampling strategies, to geographic constraints (species at the edge of ranges) or both, preventing the correct capture of the entire environmental range of species (Thuiller et al. 2003b).

In this study, we examine the consequences of restricting the range of environmental conditions over which niche-based models are developed to project potential future distributions of three selected European tree species to asses first, the importance of removing "naughty noughts" and second the importance of capturing the fun environmental range of species.

Methods

Species and climate data

We modelled climate response curves, realised niches and current and future species distributions for three European tree species: *Pinus mugo*, *Salix appendiculata* and *Quercus crenata*. We selected species with distributions restricted to Europe according to Tutin et al. (1964–1993) and which occur primarily in central Europe. This was required because we wanted to maximise the probability of including species' complete realised niches and a central European location maximises the probability of species occurring in the centre of climatic gradients. Current distributions were taken from the Atlas Florae Europaeae (AFE) (Jalas and Suominen 1972–1996), which uses 2610 UTM 50×50 km grid cells and was digitised by Lahti and Lampinen (1999). Four environmental variables were selected and converted from 0.5° latitude – longitude maps to UTM 50×50 km grid cells. These included: mean ratio of annual actual evapotranspiration over annual potential evapotranspiration (AET/PET), mean annual growing degree days (5°) (GDD), mean temperature of the coldest month (MTC), and mean annual precipitation (MAP) (Michell pers. comm.).

Models

To describe species' response curves and approximate current realised niches we used generalised additive models (GAM) (Hastie and Tibshirani 1990, for a comparison with other methods, see Thuiller et al. 2003a). We evaluated the fit of the models with a calibration set of 70% randomly chosen points. Then we compared the fit with projections using the 30% remaining set using a threshold independent method, the area under the relative operating characteristic (ROC) curve (AUC, Hanley and McNeil 1982, Pearce and Ferrier 2000). Three models were run for every species according to the following protocol: 1) using all climatic conditions available for the study area; 2) using data restricted to the climate range of species but by setting a limit to include at least 100 observations of the species absence above and below the last presence (for more detail see Austin and Meyers 1996, Austin et al. 1994); 3) using restricted data without zeros spanning beyond the climatic range of species.

The first case corresponded to what is commonly used to project future species distributions (e.g. Bakkenes et al. 2002, Peterson et al. 2002, Thuiller 2003). We considered this data as the reference model for comparison with the remaining models. The second case was performed to test if removing naughty noughts could disturb or improve the projected future distributions. The third case was aimed at analysing the impact of not capturing the full environmental range of species and to project distributions outside the environmental limits used to calibrate models. We assume that cutting the data just after the last presence on each gradient would be similar to having species with just the edge of their distributions in the studied area, or to have insufficient sampling points beyond a given location. To compare the predictive accuracy of models, using a same range of environmental data, we evaluated the three models on the evaluation data from the restricted data.

Projected response curves

Response curves were then used to project species distributions on a larger European window using current and future climatic data at finer resolution (10′ grid) (Thuiller et al. in press). Projections for the reference model were carried out simply by using the model calibrated on "all data" and projecting on the new data set. For "restricted + 100", we followed Austin and Meyers (1996) approach, projecting models onto a new data set, but setting probabilities equal to zero above and below the climatic limits used to calibrate models.

Assuming that "restricted" data for a selected species could correspond to a sub-set of the full suitable environmental range of the species, we extrapolated the response curve outside the environmental conditions used to calibrate the model without setting probabilities equal to zero outside the environmental range used to calibrate model. This method is commonly used when researchers predict and project large dataset of species without knowing if they captured the full environmental range of each species (e.g. Bakkenes et al. 2002, Thuiller 2003). Alternatively, as previous projections could lead to less accurate distributions, we also projected models by setting probabilities to zero outside the environmental range used to calibrate model.

Mapping of projected and future distributions

We then mapped current projected species distributions from each model. We also estimated potential future distributions using the third Hadley Global Circulation Model (HadCM3) under one emission scenario (A2) at the same resolution (For details of emission scenarios, see the IPCC's Special Report on Emissions Scenarios, Nakicenovic and Swart 2000).

Results

The predictive accuracy of models was similar for all models (Table 1).

Calibrated response curves

The shapes of response curves were also similar across the three models, although the removal of zeros beyond the species current climatic ('restricted' model) range affected the tail of response curves. We illustrated this by analysing the response curve for the most significant variable among all species, i.e. mean temperature of the coldest month (MTC) (Fig. 1).

For the whole selected species, curves were similar between models but with marked differences in the tails. Left tails for the "restricted" models showed noticeable difference (higher probability values) with the two other models for all species. Response curve for *Salix appendiculata* using "restricted + 100" data had, however, higher probability values than response curve derived from "all data" (Fig. 1).

Table 1. Area under the curve (AUC) of the ROC approach describing the accuracy of each model projected on restricted evaluation data. *Salix app*. is the abbreviation of *Salix appendiculata*. AUC is ranging from 0.5 to 1. AUC = 0.5: null accuracy and AUC = 1: excellent accuracy.

	Models	AUC Evaluation	AUC SdDev
Pinus mugo	"all data"	0.87	0.027
	"restricted + 100"	0.87	0.026
	"restricted"	0.86	0.027
Salix app.	"all data"	0.90	0.022
	"restricted + 100"	0.91	0.022
	"restricted"	0.91	0.022
Quercus crenata	"all data"	0.86	0.074
	"restricted + 100"	0.87	0.070
	"restricted"	0.86	0.063

Projected response curves

Projected response curves, showed that the differences, between the approaches of setting probabilities equal to zero outside the environmental range used to calibrate the models and of letting the models estimate the response without any constraint, were important (Fig. 2). The projected response of "restricted data" model without setting probability values equal to zero outside the environmental range used to calibrated models exhibited an strong increase of probability of occurrence for cold temperatures while "all data" model exhibited the inverse trend. This had strong implications for the mapping.

Mapping of projected and future distributions

For instance, projections of *Quercus crenata* using "restricted" models differed noticeably according to the approach used. Setting probability values equal to zero outside the environmental limits used to calibrate models provided similar projections than "restricted + 100" (Fig. 3d). If probability values outside the calibrated environmental limits were not setting to zero, projections completely over-estimated the observed distribution of *Q. crenata*. This later method showed a strong monotonic increase of probability with a decrease of temperature resulting from a problem of the left tail of the fitted response curve (Fig. 2b) that increased for low values of temperature. So when the response curve was projected to the larger European window, the response curve for low temperature continued to increase while

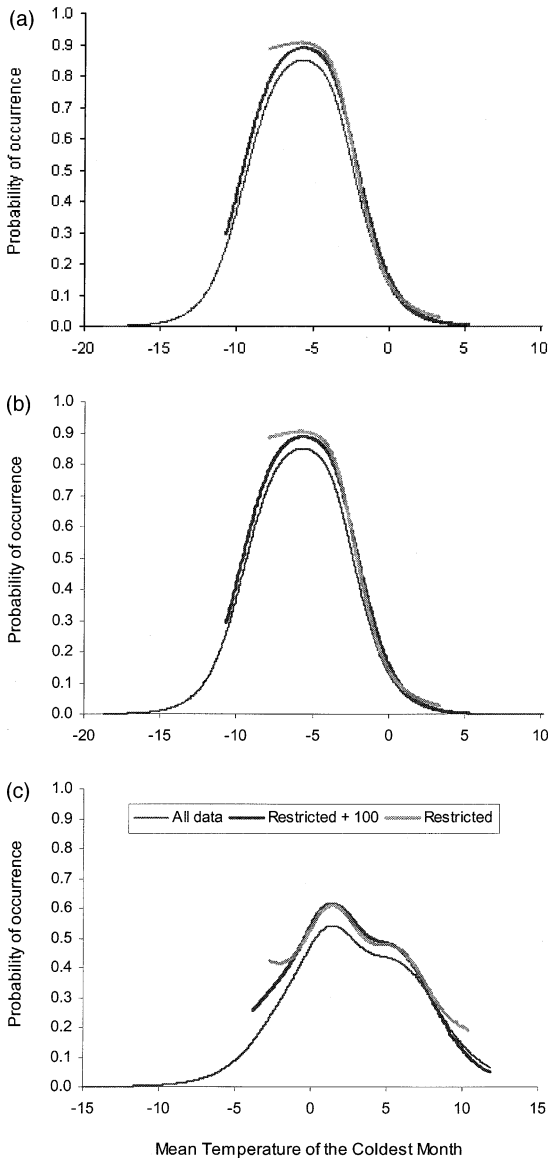

Fig. 1. Comparisons of the shapes of the response functions of MTC of the three different models: (a) *Pinus mugo*; (b) *Salix appendiculata*; and (c) *Quercus crenata*.

the response curve for "restricted+100" and "all data" had been set to 0 (Fig. 2b).

Projections of *Quercus crenata* distribution were very similar between "all data" and "restricted+100" models (Fig. 3b–c). Projections using "restricted+100" data and setting probability values equal to zero outside the environmental range used to calibrate models showed more conservative distribution than "all data" model.

Difference between methods to project "restricted" models was exacerbated in the future with spurious projections for models without setting zeros outside the environmental limits of calibrated model (Fig. 4c–d).

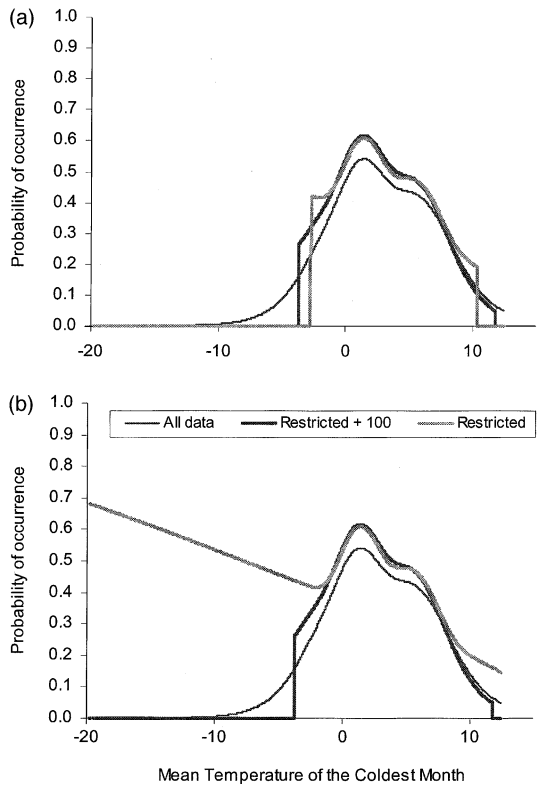

Fig. 2. Projections of response curve of *Quercus crenata* on larger spectrum of climate data at finer resolution. (a) The three models setting probability values equal to zero outside the environmental limits used to calibrate models; (b) Same as (a) but without setting probability values equal to zero outside the environmental limits used to calibrate models for the restricted model.

Again, "restricted+100" projections were still more conservative than projections using "all data" models (Fig. 4a–b).

Discussion

We found that restricting the environmental range of data strongly influenced the estimation of response curves, especially towards upper and lower ends of environmental ranges. This has a consequence of first modifying the probability values towards upper and lower environmental boundaries, and second leading to more conservative or liberal projections, depending on the approach.

Effect of restriction the range of environmental conditions on model accuracy

We applied the approach proposed by Austin and Meyers (1996) concerning the effects of naughty noughts

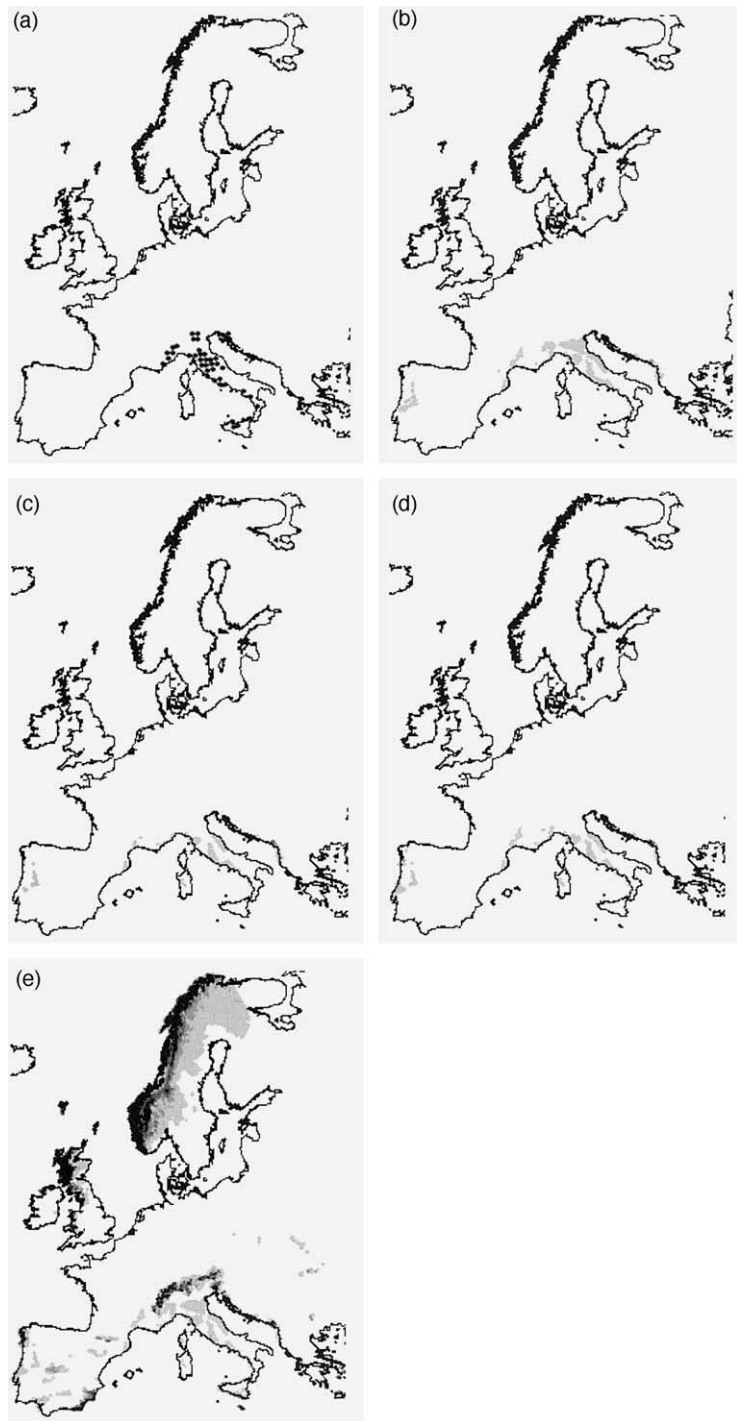

Fig. 3. Spatial distribution of *Q. crenata* projected on the whole Europe (lat-long $10' \times 10'$). (a) Observed distribution from Atlas Florae Europaeae; (b) Projected distribution using "all data" model on current climate; (c) Projected distribution using "restricted + 100" model on current climate with setting probability values equal to zero outside the environmental limits used to fit models; (d) Projected distribution using "restricted" model on current climate with setting probability values equal to zero outside the environmental limits used to fit models; (e) Projected distribution using "restricted" model on current climate without setting probability values equal to zero outside the environmental limits used to fit models.

to approximate the species' realised niche in a different context, that is of projecting distribution outside the geographic area used to calibrated models and modelling the future distribution of species under global change. We showed that in terms of accuracy, both "all data" and "restricted + 100" were quite similar which demonstrated that removing the naughty noughts does not improve or change the predictive performance of models.

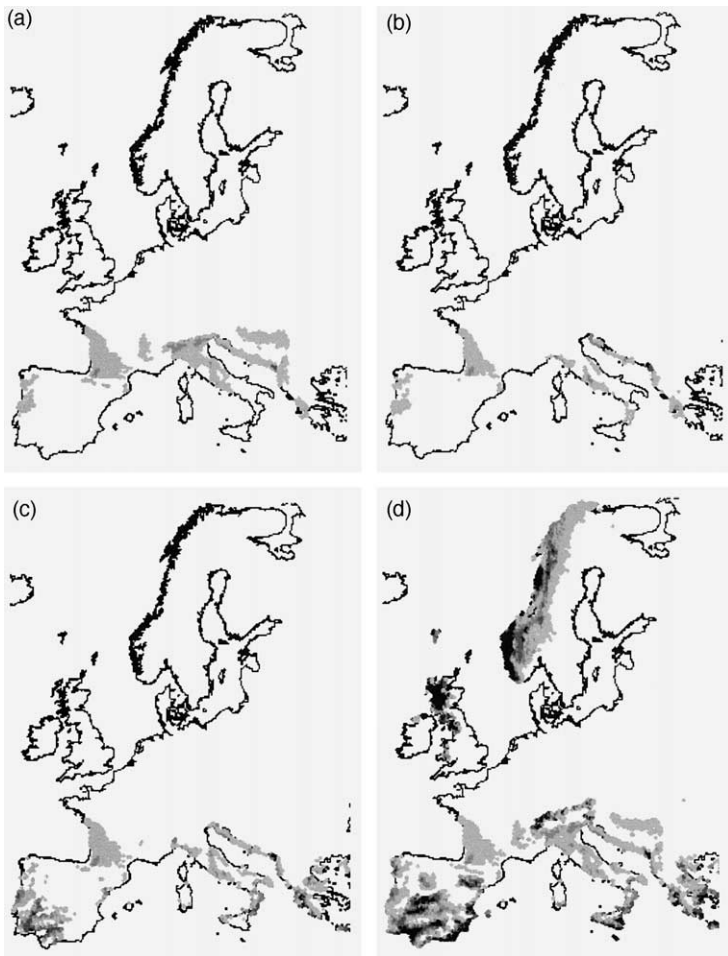

Fig. 4. Future spatial distribution of *Quercus crenata* according to Hadcm3 climate change scenario. Future distribution using: (a) "all data" model; (b) "restricted+100" model setting probability values equal to zero outside the environmental limits used to fit model; (c) "restricted" model setting probability values equal to zero outside the environmental limits used to fit model; (d) "restricted" model without setting probability values equal to zero outside the environmental limits used to fit model.

The projections of realised niche on 10′ grid and into future climatic conditions demonstrated that removing the naughty noughts to calibrate models and project distribution setting probability values equals to zero outside the environmental limits used, provided more conservative distributions than using all data without restrictions. These results were in total agreement with Austin and Meyers (1996). However, Austin and Meyers (1996) regarded over-predictions as a serious error whereas in a context of global change we tend to regard under-predictions as more serious. When considering using niche-based models to project future distributions, reasonable over-predictions are regarded as existing suitable habitats that species have yet not occupied because of history, constraints to dispersal or other ecological non-equilibrium reasons. From a conservation perspective, such suitable habitats are of interest as they locate the availability of potentially suitable areas, which may then used to plan reintroductions of species or relocate reserves.

Moreover, from a more ecological standpoint, one can see species environmental responses as a gradual continuum along ecological gradients and, as such, it seems more intuitive that species probability of occurrence decreases gradually beyond where it is found rather than assuming a truncated response in which the probability of species occurrence becomes zero beyond the current extent of occurrence.

Effect of restricting the range of environmental conditions beyond the limits used to calibrate models

Restricting data entails several implications and consequences: First, absences are often true absences providing potentially relevant information on species ecology. Using restricted data (similar to not capture the full species' environmental range) reduces strongly the combinations of environmental conditions under which the models are calibrated, and reduces the applicability

of the model for predictive purposes (Pearson and Dawson 2003). This problem has important implications when future projections of species distributions are sought. In particular, species niche from restricted data sets might be seen as analogous to the modelling of species niche from a limited geographic location not covering the complete range of environmental conditions in which species may occur. We have shown that in such cases, there were two possibilities in projecting models to novel environmental conditions: the first was to project distributions using the calibrated models and second was to project using the calibrated model but setting probabilities equal to zero outside the environmental limits used to calibrate them. The first approach provides liberal projections and second can underpredict the true distribution. The question remains as to which is the best approach: to model species niche from a limited geographic extend (edge of distribution)? Or to project distribution outside the environmental limits used to calibrate the models? In the absence of accurate information on the extent of environmental conditions used to calibrate niche-based models, it might be advisable to project to unknown environmental conditions by using a conservative approach as recommended by Austin and Meyers (1996) in another context.

Secondly, from the perspective of predictive modelling, restricting conditions on which models are calibrated may generate unpredictable effects on the tails of the species response curves, yielding spurious projections into the future (without setting probability values equal to zero outside the environmental limits). Indeed, as the restricted data does not capture the whole of the response curve, projections of future species distributions based of ecological niche modelling may be only valid if niche models are able to approach the complete response curve of environmental predictors. As an example of how considering such constraints into bioclimatic modelling, Pearson et al. (2002) carried out species niche models at the European scale to include the fun bioclimatic envelope of the species studied and then downscaled in its application to Great Britain, ensuring that when applied to future climate scenarios the model is not used to extrapolate outside its training data range. Our results support this method as being more suitable to extrapolate species distribution from bioclimatic models because it includes a large part of the range of species distribution and of the environmental combinations where the species currently occur or not.

Acknowledgements – This research was funded by the European Commission's ATEAM (Advanced Terrestrial Ecosystem Analysis and Modelling) project (EVK2-CT-2000-00075). Species data were kindly supplied by R. Leemans and M. B. Araújo. The authors thank particularly T. D. Mitchell of the Tyndall Centre for Climate Change Research for providing climate and scenario data and S. Zachle and M. Erhard of Potsdam-Institute for Climate Impact Research for aggregating climate data. MBA also thanks FCT (SPRH/BPD/5547/2001) and LB European Community (contract HPMF-CT-2002-01987) for funding.

References

Araújo, M. B. et al. 2001. Would environmental diversity be a good surrogate for species diversity? – Ecography 24: 103–110.

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

Austin, M. P. et al. 1990. Measurement of the realized qualitative niche: environmental niches of five *Eucalyptus* species. – Ecol. Monogr. 60: 161–177.

Austin, M. P. et al. 1994. Determining species response functions to an environmental gradient by means of a beta-function. – J. Veg. Sci. 5: 215–228.

Bakkenes, M. et al. 2002. Assessing effects of forecasted climate change on the diversity and distribution of European higher plants for 2050. – Global Change Biol. 8: 390–407.

Beerling, D. J. et al. 1995. Climate and the distribution of *Fallopia japonica*: use of an introduced species to test the predictive capacity of response surface. – J. Veg. Sci. 6: 269–282.

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

Hanley, J. A. and McNeil, B. J. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. – Radiology 143: 29–36.

Hastie, T. J. and Tibshirani, R. 1990. Generalized additive models. – Chapman and Hall.

Jalas, J. and Suominen, J. 1972–1996. Altas Florae Europaeae. – The Committee for Mapping the Flora of Europe and Societas Biologica Fennica Vanamo, Helsinki.

Lahti, T. and Lampinen, R. 1999. From dot maps to bitmaps – Altas Florae Europaeae goes digital. – Acta Bot. Fenn. 162: 5–9.

Lawton, J. H. 2000. Concluding remarks: a review of some open questions. – In: Hutchings, M. J., John, E. and Stewart, A. J. A. (eds), Ecological consequences of heterogeneity. Cambridge Univ. Press, pp. 401–424.

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

Loehle, C. and LeBlanc, D. 1996. Model-based assessments of climate change effects on forests: a critical review. – Ecol. Model. 90: 1–31.

Nakicenovic, N. and Swart, R. (eds) 2000. Emissions scenarios: a special report of working group III of the intergovernmental panel on climate change. – Cambridge Univ. Press.

Oksanen, J. and Minchin, P. T. 2002. Continuum theory revisited: what shape are species responses along ecological gradients. – Ecol. Model. 157: 119–129.

Pearce, J. and Ferrier, S. 2000. Evaluating the predictive performance developed using logistic regression. – Ecol. Model. 133: 225–245.

Pearson, R. G. and Dawson, T. E. 2003. Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? – Global Ecol. Biogeogr. 12: 361–372.

Pearson, R. G. et al. 2002. SPECIES: A Spatial Evaluation of Climate Impact on the Envelope of Species. – Ecol. Model. 154: 289–300.

Peterson, A. T. et al. 2002. Future projections for Mexican faunas under global climate change scenarios. – Nature 416: 626–629.

Thuiller, W. 2003. BIOMOD: Optimising predictions of species distributions and projecting potential future shifts under global change. – Global Change Biol. 9: 1353–1362.
Thuiller, W. et al. 2003a. Generalized models versus classification tree analysis: a comparative study for predicting spatial distributions of plant species at different scales. – J. Veg. Sci. 14: 669–680.
Thuiller, W. et al. in press. Do we need land-cover data to model species distributions in Europe? – J. Biogeogr.
Thuiller, W. et al. 2003b. Large-scale environmental correlates of forest tree distributions in Catalonia (NE Spain). – Global Ecol. Biogeogr. 12: 313–325.
Tutin, T. G. et al. 1964–1993. Flora Europaea. – Cambridge Univ. Press.
Woodward, F. I. and Beerling, D. J. 1997. The dynamics of vegetation change: health warnings for equilibrium 'dodo' models. – Global Ecol. Biogeogr. Lett. 6: 413–418.