REVIEW

Spatial modelling of biodiversity at the community level

SIMON FERRIER* and ANTOINE GUISAN†

**New South Wales Department of Environment and Conservation, PO Box 402, Armidale, New South Wales 2350, Australia; and* †*Department of Ecology and Evolution, University of Lausanne, CH-1015 Lausanne, Switzerland*

Summary

1. Statistical modelling is often used to relate sparse biological survey data to remotely derived environmental predictors, thereby providing a basis for predictively mapping biodiversity across an entire region of interest. The most popular strategy for such modelling has been to model distributions of individual species one at a time. Spatial modelling of biodiversity at the community level may, however, confer significant benefits for applications involving very large numbers of species, particularly if many of these species are recorded infrequently.

2. Community-level modelling combines data from multiple species and produces information on spatial pattern in the distribution of biodiversity at a collective community level instead of, or in addition to, the level of individual species. Spatial outputs from community-level modelling include predictive mapping of community types (groups of locations with similar species composition), species groups (groups of species with similar distributions), axes or gradients of compositional variation, levels of compositional dissimilarity between pairs of locations, and various macro-ecological properties (e.g. species richness).

3. Three broad modelling strategies can be used to generate these outputs: (i) 'assemble first, predict later', in which biological survey data are first classified, ordinated or aggregated to produce community-level entities or attributes that are then modelled in relation to environmental predictors; (ii) 'predict first, assemble later', in which individual species are modelled one at a time as a function of environmental variables, to produce a stack of species distribution maps that is then subjected to classification, ordination or aggregation; and (iii) 'assemble and predict together', in which all species are modelled simultaneously, within a single integrated modelling process. These strategies each have particular strengths and weaknesses, depending on the intended purpose of modelling and the type, quality and quantity of data involved.

4. *Synthesis and applications.* The potential benefits of modelling large multispecies data sets using community-level, as opposed to species-level, approaches include faster processing, increased power to detect shared patterns of environmental response across rarely recorded species, and enhanced capacity to synthesize complex data into a form more readily interpretable by scientists and decision-makers. Community-level modelling therefore deserves to be considered more often, and more widely, as a potential alternative or supplement to modelling individual species.

Key-words: biodiversity, community, composition, modelling, prediction, spatial, statistical

Journal of Applied Ecology (2006) **43**, 393–404
doi: 10.1111/j.1365-2664.2006.01149.x

Introduction

Predictive spatial modelling has been used increasingly over the past 20 years to relate sparse biological survey or collection data to remotely mapped environmental attributes, thereby allowing distributions of biological

Correspondence: Simon Ferrier, New South Wales Department of Environment and Conservation, PO Box 402, Armidale, New South Wales 2350, Australia (fax +61 267722424; e-mail simon.ferrier@environment.nsw.gov.au).

entities to be extrapolated across an entire region of interest (see reviews by Franklin 1995; Austin 1998; Guisan & Zimmermann 2000; Scott *et al.* 2002; Guisan & Thuiller 2005). Commonly used predictors include terrain indices, long-term average climate surfaces, edaphic variables, land-cover variables and spectral bands or indices from remote sensing. By far the most popular strategy has been to model distributions of individual species one at a time. Species-level modelling is, however, but one of a rich array of possible approaches to spatial modelling of biodiversity (Franklin 1995; Ferrier & Watson 1997; Guisan & Zimmermann 2000; Ferrier 2002).

In this review we focus on 'community-level' modelling strategies. We define such a strategy as one that both: (i) combines data from multiple species at some stage in the analytical process and (ii) produces information on spatial pattern in the distribution of biodiversity at a collective (or emergent) community level instead of, or in addition to, the level of individual species. The origins of community-level modelling go back as far, if not further, than those of species-level modelling (for early applications of the basic concepts discussed here see Kessell 1976; Strahler *et al.* 1978 ; for an overview of the history of this general approach see Franklin 1995).

The appropriateness of modelling biodiversity at the community level, as opposed to the species level, is likely to vary depending on the purpose of a given study and the type, quality and quantity of data involved. Community-level modelling may confer significant benefits for applications involving very large numbers of species, particularly where a sizeable proportion of these species is rarely recorded in the data set. Unlike species-level modelling, for which species with too little data are usually excluded from further analysis (for statistical reasons), many community-level modelling strategies make use of all available data across all species, regardless of the number of records per species. Hence, the data for more common species may help to support the modelling of less frequent species (Guisan *et al.* 1999). By producing information on spatial pattern in biodiversity in a collective sense (Austin 1999), community-level modelling also provides a means of synthesizing complex data on large numbers of species into a simpler form that may be more readily interpretable by both scientists and decision-makers.

Several community-level modelling strategies have emerged during the past 20 years, along with a variety of analytical techniques and software tools for implementing these strategies. Our aim in this review is to provide a comprehensive overview of the current 'state of the art' of such modelling. To our knowledge this type of review has never been attempted before. We start by defining the basic data inputs used in community-level modelling, and the various types of spatial output that can be produced. We then describe three broad modelling strategies, and review specific techniques that have been, or could be, used to implement each strategy. Next we examine the strengths and weaknesses of these approaches within the context of various applications. We finish by suggesting a number of future directions that we feel are worth pursuing to extend and refine current approaches to spatial modelling of biodiversity at the community level.

Data inputs and outputs

All of the strategies described in this review work with a regular spatial grid superimposed over the study area of interest (typically consisting of many thousands, or millions, of cells depending on the spatial extent and resolution of the study). We assume the existence of two types of input data relating to this grid: (i) biological survey or collection data, typically for scattered locations across the grid, and (ii) remotely derived environmental predictors, covering the entire grid. These data would normally be stored within, or linked directly to, a geographical information system (GIS).

Biological data are usually available for only a small proportion of grid cells in the study area. The best types of data for the modelling approaches discussed in this review are presence–absence or abundance data collected by systematic field surveys of species composition, preferably based on a well-designed environmentally stratified sampling scheme (Austin 1998). Each species within the biological group of interest (e.g. diurnal birds and vascular plants) is recorded as either present (optionally with a rating of relative abundance) or absent at each surveyed location. Here we assume that each sampling unit is wholly contained within a single grid cell, and that each cell contains no more than one sampling unit. Problems that can be caused by departures from this assumption (e.g. sampling plots smaller than modelling units) are discussed elsewhere (Guisan & Thuiller 2005).

Biological data derived from natural history collections may also be used with many of the modelling approaches discussed here. However these data are more problematic, as they typically consist of 'presence-only' records, i.e. locations where a species was collected, without any explicit information on other visited locations from which the species was not collected (Zaniewski *et al.* 2002; Graham *et al.* 2004). Presence-only data sets pose similar challenges for community-level modelling as they do for species-level modelling. Recent work by us and various collaborators suggests that 'pseudo-absences' generated from presences of other species in the same biological group provide the best available basis for building models from presence-only data. Thus we will generally assume throughout this review that those species not collected at a location, where one or more other species in the same biological group have been collected, are treated as being absent at that location.

Remotely derived environmental predictors are normally stored as a stack of GIS grid layers. However, it may help to envisage both the biological and environmental inputs structured as two simple data frames or
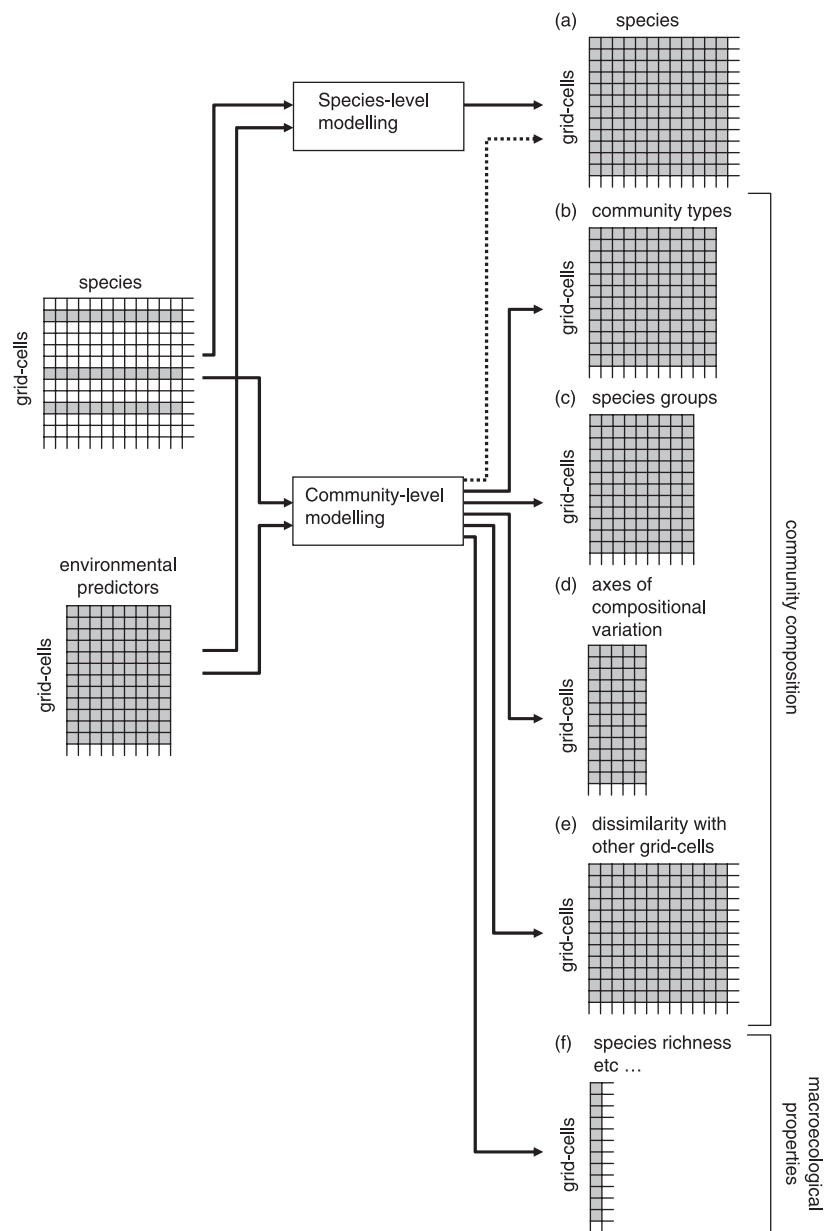
Fig. 1. Data inputs and types of spatial output for species-level and community-level modelling.

matrices, one for the biological data and the other for the environmental predictors (Fig. 1). The rows of both matrices correspond to grid cells, with the environmental matrix containing data for every cell in the study area and the biological matrix containing data only for those cells that have been surveyed. The biological data matrix has a column for each species, while the environmental matrix has a column for each predictor.

As for species-level modelling, the main objective of community-level modelling is to use observed relationships between biological data and remotely derived environmental predictors to extrapolate patterns across an entire study area. With species-level modelling this can be viewed as filling blank rows (grid cells) in the biological data matrix with the predicted occurrence or abundance of each species in each unsurveyed cell (Fig. 1a). With community-level modelling,

production of this complete cells-by-species matrix is supplemented or replaced by production of a matrix in which the columns correspond to some set of derived community-level entities or attributes.

In addition to distributions of individual species, the five main types of entities and attributes that can be predictively mapped by community-level modelling are: community types, species groups, axes of compositional variation, dissimilarities between pairs of cells and macro-ecological properties (definitions are provided in Table 1; Fig. 1b–e). The first four of these communicate something about the way in which community composition changes across a study area. This family of output types is the main focus of our review. However, for the sake of completeness, we also devote some attention to the fifth type of output, i.e. emergent macro-ecological properties, such as species richness,

Table 1. The six main types of spatial output that can be generated using community-level modelling

Spatial output	Description	Structure of derived grid layer(s)
Individual species	Predicted distributions of multiple species, as for species-level modelling	A separate layer for each species, indicating the predicted probability of occurrence or abundance of that species in each cell
Community types	Each 'community type' defined as a group of locations (grid cells) that closely resemble one another in terms of predicted species composition. Grouping normally achieved through some form of numerical classification	Either (i) a single layer with each cell assigned exclusively to one community type (depicted as a map with different colours indicating different types) or (ii) a separate layer for each community type, indicating the probability of that type occurring in each cell (depicted as multiple grey-scale or colour-ramp maps)
Species groups	Each 'species group' defined as a subset of species with similar predicted distributions. Grouping again achieved through numerical classification, but in this case the objects classified are species rather than locations	A separate layer for each species group, indicating the predicted prevalence or abundance of that group in each cell (depicted as multiple grey-scale of colour-ramp maps)
Axes of compositional variation	A set of continuous axes (or gradients) representing dimensions of a reduced space that summarizes the compositional pattern exhibited by multiple species. These axes most commonly derived through some form of ordination	A separate layer for each axis, indicating the predicted score for that axis in each cell (depicted either as multiple grey-scale or colour-ramp maps, or as a single map by assigning each of the first three axes to a different colour dimension, e.g. red, blue, green)
Levels of compositional dissimilarity between pairs of cells	The predicted level of dissimilarity in community composition between all possible pairs of grid cells in a region	In theory a complete matrix of pair-wise dissimilarities, but in practice these values are usually predicted dynamically as required by the application of interest (difficult to depict spatially without prior conversion to community types or axes of compositional variation)
Macro-ecological properties	Most commonly modelled property is species richness, either of a whole group (e.g. all vascular plants) or of a functional subgroup (e.g. annuals and trees). Many other macro-ecological properties (e.g. mean range size and endemism) can potentially be modelled	A separate layer for each macro-ecological property (depicted as a grey-scale or colour-ramp map)

that do not retain or communicate explicit information about composition (Gaston & Blackburn 1999).

Modelling strategies

Three broad modelling strategies exist for producing one or more of the community-level outputs described in the previous section (Fig. 2). All of these strategies employ the same types of biological and environmental input data.

STRATEGY 1: ASSEMBLE FIRST, PREDICT LATER

This strategy, also known as 'classification-then-modelling' (Ferrier *et al.* 2002), involves two distinct stages. In the first stage the biological survey data are subjected to some form of classification, ordination or aggregation, without any reference to the environmental data (Fig. 2). This stage therefore involves only those locations with biological data. In the second stage the community-level entities generated for these locations are modelled as a function of environmental predictors (example studies are listed in Table 2).

Classification or ordination of the biological survey data to generate community types, species groups or axes of compositional variation can be achieved using any of the very wide range of pattern analysis techniques routinely applied in community ecology (see Appendix S1 in the supplementary material). Derivation of macro-ecological properties, such as species richness, from the biological data for each surveyed location is a relatively straightforward procedure. The modelling approach used in the second stage of this strategy depends on the nature of the community-level entities generated in the first stage. For example, community types can be modelled one at a time, by relating the observed presence or absence of each community to available environmental predictors (e.g. using generalized linear, or additive, modelling), or a single model can be fitted to all communities simultaneously by treating community membership as a multinomial response (e.g. using classification and regression trees). Further detail on these, and related, modelling approaches is provided in Appendix S1 (see the supplementary material).

STRATEGY 2: PREDICT FIRST, ASSEMBLE LATER

In this strategy, also known as 'predict first, classify later' (Overton *et al.* 2002) and 'classification-then-modelling'
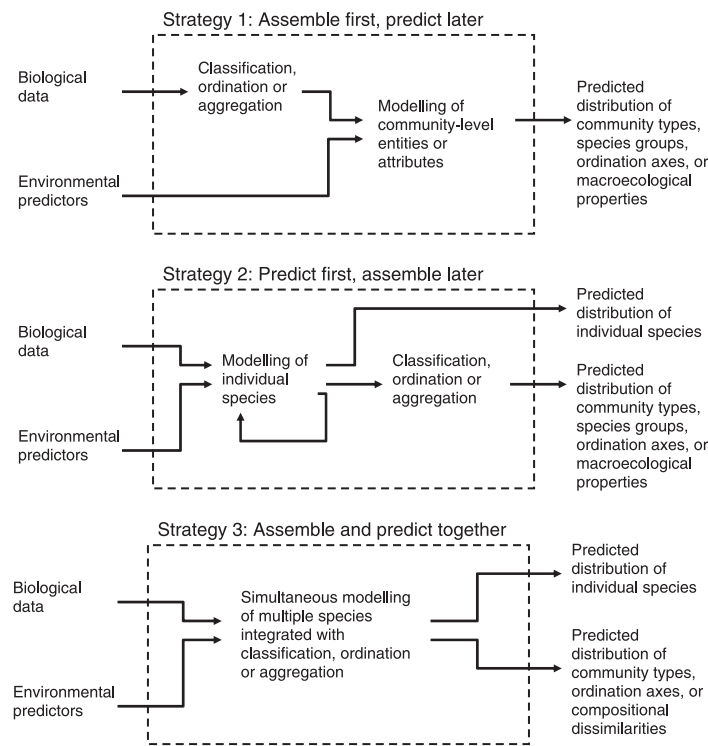
Fig. 2. Three broad strategies for spatial modelling of biodiversity at the community level.

Table 2. Examples of studies employing different approaches to spatial modelling of biodiversity at the community level. Studies predicting community-level responses to climate change are underlined

Broad strategy	Specific approach	References
1. Assemble first, predict later	1a. Modelling of pre-derived community types	Davis & Goetz (1990), Franklin & Wilson (1991), Lees & Ritman (1991), Moore *et al*. (1991), Fitzgerald & Lees (1992), Brzeziecki *et al*. (1993), Brown (1994), Brzeziecki *et al*. (1995), Lewis (1998), Zimmermann & Kienast (1999), Keith & Bedward (1999), Hilbert & Ostendorf (2001), Ferrier *et al*. (2002)
	1b. Modelling of pre-derived species groups	McKenzie *et al*. (1989), Bojórquez-Tapia *et al*. (1995), Ferrier *et al*. (2002)
	1c. Modelling of pre-derived ordination axes	Faith *et al*. (2003), Chang *et al*. (2004)
	1d. Modelling of pre-derived species richness levels	Pausas (1994), Heikkinen & Neuvonen (1997), Fraser (1998), Saetersdal *et al*. (1998); Wohlgemuth (1998); Leathwick *et al*. (1998); Grytnes *et al*. (1999); Guisan & Theurillat (2000), Lehmann *et al*. (2002); Lobo *et al*. 2004
2. Predict first, assemble later	2a. Derivation of community types from modelled species' distributions	Ferrier *et al*. (2002), Lenihan (1993), Leathwick *et al*. (1996), Ferrier & Watson (1997), Austin (1998), Zimmermann & Kienast (1999); Guisan & Theurillat (2000), Leathwick (2001); Cawsey *et al*. (2002), Overton *et al*. (2002), Peppler-Lisbach & Schröder (2004)
	2b. Derivation of species groups from modelled species' distributions	Nix (1991), Lehmann *et al*. (2002), Ferrier *et al*. (2002), Scotts & Drielsma (2003)
	2c. Derivation of ordination axes from modelled species' distributions	Peters & Thackway (1998), Peppler-Lisbach & Schröder (2004)
	2d. Derivation of richness levels from modelled species' distributions	Guisan & Theurillat (2000); Gioia & Pigott 2000; Lehmann *et al*. (2002); Peppler-Lisbach & Schröder 2004; Gelfand *et al*. 2005
3. Assemble and predict together	3a. Multiresponse modelling of multiple species	Olden (2003), Joy & Death (2004), Leathwick *et al*. (2005)
	3b. Constrained ordination	Ferrier & Watson (1997), Gottfried *et al*. (1998), Ohmann & Gregory (2002), Dirnbock *et al*. (2003)
	3c. Constrained classification	De'Ath (2002), Ferrier *et al*. (2002)
	3d. Modelling of compositional dissimilarity	Ferrier (2002), Ferrier *et al*. (2002), Ferrier *et al*. (2004)

© 2006 The Authors.
Journal compilation
© 2006 British
Ecological Society,
Journal of Applied Ecology, **43**,
393–404

(Ferrier *et al.* 2002), individual species are initially modelled one at a time as a function of environmental predictors, thereby generating a separate predicted distribution map for each species. The resulting stack of extrapolated species' distributions is then subjected to some form of classification, ordination or aggregation to derive the required community-level output (Fig. 2). The analytical techniques employed in this second stage are very similar to those used in the first stage of strategy 1. However, rather than applying them to biological data from the original surveyed locations, strategy 2 instead applies the techniques to predictions of species occurrence (or abundance) for all grid cells in a region. Each cell is therefore effectively treated as a virtual survey plot (Cawsey *et al.* 2002), with predicted data for each species in place of direct observations. Hence this strategy attempts to reconstruct community composition or macro-ecological properties from predicted species' distributions in a bottom-up manner, in contrast with the top-down approach of modelling pre-derived community-level entities (strategy 1).

As species are initially modelled individually, virtually any property of communities and ecosystems could theoretically be reconstructed with this strategy. However, while many recent studies have modelled large numbers of species (Bakkenes *et al.* 2002; Segurado & Araujo 2004), surprisingly few of these have proceeded to apply any community-level analysis to the resulting stack of species' distributions. Of those studies that have proceeded to this second stage most have concentrated on reconstructing either species richness or community types. Further detail on the techniques commonly employed to perform these two types of reconstruction is provided in Appendix S2 (see the supplementary material). This general modelling strategy has also been used less frequently to derive and map species groups and ordination axes (for examples see Table 2).

STRATEGY 3: ASSEMBLE AND PREDICT TOGETHER

Unlike the first two strategies, which treat the derivation of community-level entities or attributes and the modelling of biological–environmental relationships as two distinct steps, this final strategy performs these two tasks together (Fig. 2). The strategy therefore works with the data for all species simultaneously, within a single integrated modelling process (see example studies in Table 2).

A number of the techniques traditionally used in species-level modelling have been extended to allow a single 'multiresponse' model to be fitted to data for multiple species, rather than modelling each species separately. These extended techniques include multiresponse neural networks (Olden 2003), vector generalized linear (or additive) models (Yee & Mackenzie 2002) and a multiresponse implementation of multivariate adaptive regression splines (Leathwick *et al.*

2005). While the primary output is usually a set of predicted distributions (one for each species), these techniques may also provide valuable information on the relative importance of environmental predictors, or weighted combinations of these (e.g. hidden layers in multiresponse neural networks), in explaining overall patterns of species composition. There is therefore the potential to map important predictors revealed by such analysis as a means of visualizing major compositional gradients across a region, but we could not find any example of this in the literature.

Mapping of compositional gradients has been more often achieved using some form of constrained ordination (e.g. canonical correspondence analysis), in which ordination axes summarizing compositional variation in the biological data are constrained to be weighted combinations of the environmental predictors (see Appendix S3 in the supplementary material). While constrained ordination has been applied very widely as an analytical tool in community ecology, only a small proportion of these applications has used the approach to extrapolate beyond surveyed locations, and thereby map predicted patterns of community composition across a whole region. However, if environmental variables are mapped as GIS layers it is straightforward to map each ordination axis as a weighted combination of these variables (Ohmann & Gregory 2002). Such mapping provides a useful way of visualizing the main ecofloristic or ecofaunistic gradients across a region and can also provide a basis for mapping predicted distributions of individual species, or of pre-defined community types derived by a separate classification of surveyed locations. Such predictions of species or communities are usually based on some measure of the distance, in ordination space, of each unsurveyed grid cell from the centroid of the surveyed locations for a given species or community type (Guisan *et al.* 1999; Ohmann & Gregory 2002; Dirnbock *et al.* 2003).

An interesting variant of multiresponse modelling can also be used to derive constrained classifications (*sensu* Gordon 1996) from ecological data. This is an extension of the classification and regression tree approach often used to model pre-defined community types, which recursively splits a set of surveyed locations into nested subsets of sites using decision rules based on environmental predictors. However, in this particular application the explanatory power of candidate decision rules is assessed directly using the raw species data, rather than in terms of community types derived from a prior numerical classification of these data. The species data are used to evaluate and select environmental decision rules that minimize biological heterogeneity within resulting subsets of locations, while maximizing differences between these groups. Heterogeneity is usually assessed in terms of compositional dissimilarity between pairs of locations. This approach to constrained classification was first implemented by Ferrier (1992) and later described in detail by Ferrier *et al.* (2002). Another manifestation of the

approach, referred to as 'multivariate regression trees' (MRT), has been described independently by De'ath (2002). The general approach not only produces a community classification but also automatically provides the environmental rules for predictively mapping each of these community types (Ferrier *et al.* 2002).

If viewed in a slightly different way, constrained classification is effectively a means of modelling compositional dissimilarity between locations by partitioning environmental space into discrete, biologically homogeneous, classes (i.e. community types). An alternative approach is to treat compositional dissimilarity between locations as a continuous function of the separation of these locations in environmental space. This idea underpins the recently developed technique of generalized dissimilarity modelling (GDM; Ferrier 2002; Ferrier *et al.* 2002), a non-linear extension of permutational matrix regression (Legendre *et al.* 1994). GDM models the compositional dissimilarity observed between pairs of surveyed locations as a continuous non-linear function of the relative position of these sites along multiple environmental gradients.

Models fitted with GDM can be used to predict compositional dissimilarity between any pair of grid cells within a region, knowing only the environmental characteristics of these cells. The output from this modelling is therefore a complete matrix of predicted dissimilarities between all possible pairs of grid cells. This matrix is difficult to depict spatially in its raw form. However, it provides all the raw materials necessary to perform either a numerical classification of grid cells to derive mappable groups of cells with similar predicted composition (i.e. community types) or an ordination of grid cells to derive mappable axes of compositional variation. The predicted dissimilarities also provide a basis for predicting distributions of individual species, or of pre-defined community types (derived by a separate classification of surveyed locations), using the distance-based approach described above in relation to constrained ordination.

Applicability of approaches in different contexts

GENERAL STRENGTHS AND WEAKNESSES

As we have shown, approaches to community-level modelling are numerous and highly varied. Differences between approaches occur at three levels: (i) the broad analytical strategy employed (Fig. 2); (ii) the type of spatial output produced (Fig. 1 and Table 1); and (iii) the exact analytical technique (algorithm) used to produce a given spatial output employing a given broad strategy (e.g. using generalized linear modelling vs. neural networks to model pre-classified community types). The existing literature on community-level modelling devotes very little attention to discussing the relative strengths and weaknesses of available options across these three levels. Published examples have focused on detailed differences between analytical

techniques, rather than on differences between broad analytical strategies or types of spatial output. Thus readers of this literature may gain the impression that there is only one logical strategy for modelling biodiversity at the community level, and that the only choice that needs to be made concerns the exact analytical algorithm to be employed. This narrowness of focus on low-level differences between analytical techniques within a given strategy, rather than on high-level differences between alternative strategies, is consistent with a similar focus within the species-level modelling literature. We feel there is a real need for more broadly focused consideration of the strengths and weakness of major alternative strategies for modelling biodiversity, at both the species level and the community level. Decisions relating to the selection of a broad modelling strategy for any given study will probably have a much greater impact on the effectiveness of such modelling than decisions relating to the exact analytical algorithm employed.

No single approach to community-level modelling is likely to be optimal for all purposes and across all data sets. Different approaches may be better suited to different situations. The real challenge should therefore be seen not as one of searching for a single best approach, but rather of selecting the most appropriate approach in any given situation. Such decisions need to be informed by a good understanding of the respective strengths and weaknesses of available options. In Table 3 we present a first attempt at summarizing the relative strengths of the various community-level modelling approaches described earlier in this review. This evaluation considers only those approaches that retain and convey information on community composition.

Each approach is first rated in terms of its capacity to analyse rapidly very large numbers of species (strength 1 in Table 3). Processing time may be an important constraint when dealing with data sets containing many hundreds or thousands of species. Approaches within strategy 2 (predict first, assemble later) require that a separate model be fitted and extrapolated for every individual species, and are therefore likely to be more time-consuming to implement than approaches within the other two strategies. Strategy 3 (assemble and predict together) offers the best potential to minimize processing time, by performing all analyses simultaneously within a single integrated process. A related weakness of strategy 2 is that species occurring infrequently in a data set may not be modelled reliably, or may not be modelled at all, because of insufficient records (strength 2). These species therefore contribute little to the subsequent derivation of community-level entities or attributes from the stack of modelled species' distributions. This could be a significant problem for data sets in which a sizeable proportion of species is represented by very few records, particularly for conservation-related applications requiring an emphasis on the needs of rare species. The other two strategies (1, assemble first, predict later, and 3, assemble and predict

© 2006 The Authors.
Journal compilation
© 2006 British
Ecological Society,
Journal of Applied
Ecology, **43,**
393–404

Table 3. Relative strengths of different approaches to spatial modelling of community composition (approaches that predict species richness, or other macro-ecological properties, are not included). *Limited capacity, **moderately developed capacity, ***highly developed capacity

Broad strategy	Specific approach	Strengths							
		1. Rapidly analyses very large numbers of species	2. Adds value to data for rare species by 'pooling'	3. Addresses interactions between species	4. Allows individualistic species responses	5. Combines taxa surveyed at different sets of locations	6. Produces individual species' distributions	7. Enforces congruence with known 'communities'	8. Extrapolates beyond known 'communities'
1. Assemble first, predict later	1a. Modelling of pre-derived community types	**	***	**	*	*	*	***	*
	1b. Modelling of pre-derived species groups	**	***	**	*	*	*	**	**
	1c. Modelling of pre-derived ordination axes	**	***	**	*	*	*	*	**
2. Predict first, assemble later	2a. Derivation of community types from modelled species' distributions	*	*	*	***	***	***	*	**
	2b. Derivation of species groups from modelled species' distributions	*	*	*	***	***	***	*	**
	2c. Derivation of ordination axes from modelled species' distributions	*	*	*	***	***	***	*	**
3. Assemble and predict together	3a. Multiresponse modelling of multiple species	***	**	***	**	**	***	*	**
	3b. Constrained ordination	***	***	**	*	*	**	*	**
	3c. Constrained classification	***	***	**	**	*	**	**	*
	3d. Modelling of compositional dissimilarity	***	***	**	**	**	**	*	***

together) use data from all species, no matter how infrequently recorded, in deriving community-level entities or attributes. By pooling data from all species these strategies may provide more power to detect shared patterns of environmental response across infrequently recorded species than can be detected by analysing the data for each of these species independently. Combining data in this way also provides more scope to address interactions between the distributions of different species, such as those resulting from competition or predation (strength 3).

Despite the drawbacks just discussed, strategy 2 has some unique strengths. By modelling species one at a time this strategy provides maximum opportunity, or flexibility, for each species to respond to the environment in an individualistic manner (strength 4). In contrast, approaches within the other two strategies place various constraints on the flexibility of species–environment relationships, for example by assuming that all species are responding to the same set of environmental gradients or that the functional form (shape) of these responses is the same across all species. Another unique strength of strategy 2 is the potential ability to combine species' models derived from different biological survey or collection data sets (strength 5). Imagine, for a hypothetical region, that animal species have been surveyed at one set of sites while plant species have been surveyed at a different set of sites. Once models have been derived for individual species of plants and animals, the combined stack of extrapolated distributions can then be readily subjected to community-level classification, ordination or aggregation regardless of the fact that these distributions were originally modelled using different survey data sets (Scotts & Drielsma 2003). This capability also has benefits for the analysis of presence-only data from museum collections, in which data for different taxa are often derived from different sources, collectors or expeditions. In contrast, strategy 1 and, to a lesser extent strategy 3, assume that all species have been surveyed (i.e. recorded as present or absent) at the same set of sites. Another strength of strategy 2 is that modelled distributions of individual species are produced as a standard by-product, thereby complementing any community-level outputs derived from these models (strength 6). However, it should be noted that several approaches within strategy 3 also offer this capability.

The final two strengths evaluated in Table 3 are less closely aligned with any particular broad strategy and relate more to individual approaches within these strategies. The first of these strengths concerns the extent to which an approach constrains the community-level composition predicted for each unsurveyed grid cell to match the composition observed at one or more surveyed locations (strength 7). Such congruence is enforced most strictly by approach 1a (modelling of pre-derived community types) within strategy 1. This approach treats community types, generated by the initial classification of surveyed locations, as fixed entities.

The subsequent modelling stage therefore forces each unsurveyed grid cell to be assigned to one of these known communities. Whether this enforced congruence is regarded as a strength, rather than a weakness, is likely to depend on the purpose of modelling and the nature of the data involved. For example, this might be viewed as a strength for an application requiring that mapped entities concord directly with pre-defined community types, and for which sufficient survey work has been conducted to detect all community types within the region of interest. However, where field sampling of communities is incomplete, or sparse, then enforcement of a one-to-one congruence between surveyed community types and modelled entities may instead be seen as a weakness. This situation may be better served by approaches with an ability to extrapolate beyond sampled communities, thereby predicting the occurrence of other community types, or species assemblages, in as yet unsampled environments (strength 8). Most approaches other than 1a (modelling of pre-derived community types) offer at least some capacity for such extrapolation.

The main conclusion to be drawn from Table 3 is that the appropriateness of a given approach for a given application will depend on the relative importance of various strengths and weaknesses in relation to both the purpose of the application and the type, quality and quantity of data involved. We illustrate this point using the application of community-level modelling to predict distributional shifts in biodiversity in response to climate change.

A SPECIAL CHALLENGE: PREDICTING RESPONSES TO CLIMATE CHANGE

Palaeoecology provides clear evidence that community types of the past were different from those observed today (Huntley 1991; Ackerly 2003). Hence some community-level modelling approaches may face serious problems in predicting probable responses to climate change if they assume that the composition of communities will remain fixed over time, i.e. that the same community types will continue to exist and only the distributions of these types will change. Palaeoecology also provides evidence of an even more fundamental problem, i.e. the realized niches of some species appear to change over time (Ackerly 2003), probably as a result of changing interactions (e.g. competition and predation) between species (but see also Peterson *et al.* 1999). Any approach to predicting climate change responses that treats the currently realized niche of a species as if it were the fundamental niche is therefore also likely to encounter problems (Austin 1992).

All three modelling strategies described in this review have been used to predict community-level responses to climate change (Table 2). However, if future community types are likely to differ in composition from those observed today then strategy 1 will generate unreliable predictions. This strategy may

provide useful evidence that current community types will not be able to persist at their present locations, but cannot reliably predict where similar communities (if maintained as such), or new community types, will be distributed in the future. Strategy 2 may be a better option in this regard (Guisan & Theurillat 2000) because it can account for individual responses of species, and may even allow different migration rates to be incorporated into the modelling of distributional shifts for different species. However, a potential problem with most existing implementations of this strategy is that current species' interactions, and therefore the realized (as opposed to fundamental) niches of species, are assumed to remain constant into the future. This is unlikely and thus inclusion of species' (and other biotic) interactions into the species-modelling stage of this strategy is expected to improve the rigour with which shifts in community composition are predicted. Leathwick *et al*. (1996) is the only example found so far where interactions were explicitly included in species' models used to derive spatially explicit climate change projections at the community level. Strategy 3 has, to date, been employed much less frequently in predicting climate change responses. However, the potential for this strategy to provide an effective balance between the respective strengths of the other two strategies warrants further attention.

Future directions

There is a clear need for more attention to be directed to comparative evaluation of the efficacy of different community-level modelling approaches. Surprisingly little empirical testing of alternative approaches has occurred to date. A small number of studies has compared the performance of strategies 1 and 2 for modelling species richness (Guisan & Theurillat 2000; Lehmann *et al*. 2002). Ferrier & Watson (1997) and Ferrier *et al*. (2002) have also evaluated the performance of various approaches to modelling community composition, although this work was confined to a single region and a single application, the selection of conservation areas. There is considerable scope to extend this type of evaluation to cover a wider range of modelling approaches, environments, data types and applications. Much could also be gained by supplementing assessments using real data sets with testing based on artificial (simulated) data.

Several potential avenues are available for further refining and extending the modelling approaches reviewed here. For example, some of these approaches could be readily extended to accommodate expanded measures of community composition that consider not just the occurrence (or abundance) of species at a location, but also the attributes of (e.g. functional and behavioural traits; Legendre *et al*. 1997; Pavoine *et al*. 2004), or relationships between (e.g. taxonomic and phylogenetic links; Webb *et al*. 2002; Pavoine *et al*. 2004), these species. Another promising avenue of particular relevance to the predict first, assemble later strategy is the use of ecological assembly rules (Keddy

1992) to guide, or constrain, the derivation of communities from individual species' models (see Appendix S4 and Fig. S1 in the supplementary material).

CONCLUSION

Despite the current popularity of species-level modelling, spatial modelling of biodiversity at the community level may confer significant benefits for applications involving very large numbers of species, particularly if many of these species are recorded infrequently. Potential benefits in such situations include faster processing, increased power to detect shared patterns of environmental response across rarely recorded species, and enhanced capacity to synthesize complex data into a form more readily interpretable by scientists and decision-makers. Community-level modelling therefore deserves to be considered more often, and more widely, as a potential alternative or supplement to modelling individual species.

We have defined three very different strategies for modelling biodiversity at the community level: assemble first, predict later; predict first, assemble later; and assemble and predict together, and we have described a number of available approaches within each strategy. Each of these strategies and approaches has different strengths and weaknesses, and therefore no single approach is likely to be optimal for all purposes and across all data sets. Prospective users of community-level modelling should therefore exercise care in selecting a modelling approach that is best suited to their particular needs.

References

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

Austin, M.P. (1992) Modeling the environmental niche of plants: implications for plant community response to elevated CO_2 levels. *Australian Journal of Botany*, **40**, 615–630.

Austin, M.P. (1998) An ecological perspective on biodiversity investigations: examples from Australian eucalypt forests. *Annals of the Missouri Botanical Garden*, **85**, 2–17.

Austin, M.P. (1999) The potential contribution of vegetation ecology to biodiversity research. *Ecography*, **22**, 465–484.

Bakkenes, M., Alkemade, J.R.M., Ihle, F., Leemans, R. & Latour, J.B. (2002) Assessing effects of forecasted climate change on the diversity and distribution of European higher plants for 2050. *Global Change Biology*, **8**, 390–407.

Bojórquez-Tapia, L.A., Azuara, I., Ezcurra, E. & Flores-Villela, O. (1995) Identifying conservation priorities in Mexico through geographic information systems and modeling. *Ecological Applications*, **5**, 215–231.

Brown, D.G. (1994) Predicting vegetation types at treeline using topography and biophysical disturbance variables. *Journal of Vegetation Science*, **5**, 641–656.

Brzeziecki, B., Kienast, F. & Wildi, O. (1993) A simulated map of the potential natural forest vegetation of Switzerland. *Journal of Vegetation Science*, **4**, 499–508.

Brzeziecki, B., Kienast, F. & Wildi, O. (1995) Modeling potential impacts of climate-change on the spatial distribution of zonal forest communities in Switzerland. *Journal of Vegetation Science*, **6**, 257–268.

Cawsey, E.M., Austin, M.P. & Baker, B.L. (2002) Regional vegetation mapping in Australia: a case study in the practical use of statistical modelling. *Biodiversity and Conservation*, **11**, 2239–2274.

Chang, C.R., Lee, P.F., Bai, M.L. & Lin, T.T. (2004) Predicting the geographical distribution of plant communities in complex terrain: a case study in Fushian Experimental Forest, northeastern Taiwan. *Ecography*, **27**, 577–588.

Davis, F.W. & Goetz, S. (1990) Modeling vegetation pattern using digital terrain data. *Landscape Ecology*, **4**, 69–80.

De'ath, G. (2002) Multivariate regression trees: a new technique for modeling species–environment relationships. *Ecology*, **83**, 1105–1117.

Dirnbock, T., Dullinger, S., Gottfried, M., Ginzler, C. & Grabherr, G. (2003) Mapping alpine vegetation based on image analysis, topographic variables and canonical correspondence analysis. *Applied Vegetation Science*, **6**, 85–96.

Faith, D.P., Carter, G., Cassis, G., Ferrier, S. & Wilkie, L. (2003) Complementarity, biodiversity viability analysis, and policy-based algorithms for conservation. *Environmental Science and Policy*, **6**, 311–328.

Ferrier, S. (1992) *Development of a Predictive Modelling Module for E-RMS*. Consultancy report prepared by NSW National Parks and Wildlife Service. Australian National Parks and Wildlife Service, Canberra, Australia.

Ferrier, S. (2002) Mapping spatial pattern in biodiversity for regional conservation planning: where to from here? *Systematic Biology*, **51**, 331–363.

Ferrier, S. & Watson, G. (1997) An evaluation of the effectiveness of environmental surrogates and modelling techniques in predicting the distribution of biological diversity. Environment Australia, Canberra, Australia. http://www.deh.gov.au/biodiversity/publications/technical/surrogates/, accessed 23.02.06.

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. *Bioscience*, **54**, 1101–1109.

Fitzgerald, R.W. & Lees, B.G. (1992) The application of neural networks to the floristic classification of remote sensing and GIS data in complex terrain. *Proceedings of the XVII Congress of the International Society for Photogrammetry and Remote Sensing* (ed. Anongmous), pp. 570–573. American Society for Photogrammetry and Remote Sensing, Washington, D.C.

Franklin, J. (1995) Predictive vegetation mapping: geographic modelling of biospatial patterns in relation to environmental gradients. *Progress in Physical Geography*, **19**, 474–499.

Franklin, S.E. & Wilson, B.A. (1991) Vegetation mapping and change detection using SPOT MLA and Landsat imagery in Kluane National Park. *Canadian Journal of Remote Sensing*, **17**, 2–17.

Fraser, R.H. (1998) Vertebrate species richness at the mesoscale: relative roles of energy and heterogeneity. *Global Ecology and Biogeography Letters*, **7**, 215–220.

Gaston, K.J. & Blackburn, T.M. (1999) A critique for macroecology. *Oikos*, **84**, 353–368.

Gelfand, A.E., Schmidt, A.M., Wu, S., Silander, J.A., Latimer, A. & Rebelo, A.G. (2005) Modelling species diversity through species level hierarchical modelling. *Applied Statistics*, **54**, 1–20.

Gioia, P. & Pigott, J.P. (2000) Biodiversity assessment: a case study in predicting richness from the potential distributions of plant species in the forests of south-western Australia. *Journal of Biogeography*, **27**, 1065–1078.

Gordon, A.D. (1996) A survey of constrained classification. *Computational Statistics and Data Analysis*, **21**, 17–29.

Gottfried, M., Pauli, H. & Grabherr, G. (1998) Prediction of vegetation patterns at the limits of plant life: a new view of the alpine–nival ecotone. *Arctic and Alpine Research*, **30**, 207–221.

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.

Grytnes, J.A., Birks, H.J.B. & Peglar, S.M. (1999) Plant species richness in Fennoscandia: evaluating the relative importance of climate and history. *Nordic Journal of Botany*, **19**, 489–503.

Guisan, A. & Theurillat, J.P. (2000) Equilibrium modeling of alpine plant distribution: how far can we go? *Phytocoenologia*, **30**, 353–384.

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

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

Guisan, A., Weiss, S.B. & Weiss, A.D. (1999) GLM versus CCA spatial modeling of plant species distribution. *Plant Ecology*, **143**, 107–122.

Heikkinen, R.K. & Neuvonen, S. (1997) Species richness of vascular plants in the subarctic landscape of northern Finland: modelling relationships to the environment. *Biodiversity and Conservation*, **6**, 1181–1201.

Hilbert, D.W. & Ostendorf, B. (2001) The utility of artificial neural networks for modelling the distribution of vegetation in past, present and future climates. *Ecological Modelling*, **146**, 311–327.

Huntley, B. (1991) How plants respond to climate change: migration rates, individualism and the consequences for plant communities. *Annals of Botany*, **67**, 15–22.

Joy, M.K. & Death, R.G. (2004) Predictive modelling and spatial mapping of freshwater fish and decapod assemblages using GIS and neural networks. *Freshwater Biology*, **49**, 1036–1052.

Keddy, P.A. (1992) Assembly and response rules: two goals for predictive community ecology. *Journal of Vegetation Science*, **3**, 157–164.

Keith, D.A. & Bedward, M. (1999) Native vegetation of the South East Forests region, Eden, New South Wales. *Cunninghamia*, **6**, 1–218.

Kessell, S. (1976) Gradient modeling: a new approach to fire modeling and wilderness resource management. *Environmental Management*, **1**, 39–48.

Leathwick, J.R. (2001) New Zealand's potential forest pattern as predicted from current species–environment relationships. *New Zealand Journal of Botany*, **39**, 447–464.

Leathwick, J.R., Burns, B.R. & Clarkson, B.D. (1998) Environmental correlates of tree alpha-diversity in New Zealand primary forests. *Ecography*, **21**, 235–246.

Leathwick, J.R., Rowe, D., Richardson, J., Elith, J. & Hastie, T. (2005) Using multivariate adaptive regression splines to predict the distributions of New Zealand's freshwater diadromous fish. *Freshwater Biology*, **50**, 2034–2052.

Leathwick, J.R., Whitehead, D. & McLeod, M. (1996) Predicting changes in the composition of New Zealand's indigenous forests in response to global warming: a modelling approach. *Environmental Software*, **11**, 81–90.

Lees, B.G. & Ritman, K. (1991) Decision-tree and rule-induction approach to integration of remotely sensed and GIS data in mapping vegetation in disturbed or hilly environments. *Environmental Management*, **15**, 823–831.

Legendre, P., Galzin, R. & Harmelin-Vivien, M.L. (1997) Relating behavior to habitat: solutions to the fourth-corner problem. *Ecology*, **78**, 547–562.

Legendre, P., Lapointe, F.-J. & Casgrain, P. (1994) Modeling brain evolution from behavior: a permutational regression approach. *Evolution*, **48**, 1487–1499.

Lehmann, A., Leathwick, J.R. & Overton, J.M. (2002) Assessing New Zealand fern diversity from spatial predictions of species assemblages. *Biodiversity and Conservation*, **11**, 2217–2238.

Lenihan, J.M. (1993) Ecological response surfaces for North American boreal tree species and their use in forest classification. *Journal of Vegetation Science*, **4**, 667–680.

Lewis, M.M. (1998) Numerical classification as an aid to spectral mapping of vegetation communities. *Plant Ecology*, **136**, 133–149.

Lobo, J.M., Jay-Robert, P. & Lumaret, J.P. (2004) Modelling the species richness distribution for French Aphodiidae (Coleoptera, Scarabaeoidea). *Ecography*, **27**, 145–156.

McKenzie, N.L., Belbin, L., Margules, C.R. & Keighery, G.J. (1989) Selecting representative reserve systems in remote areas. *Biological Conservation*, **50**, 239–261.

Moore, D.M., Lees, B.G. & Davey, S.M. (1991) A new method for predicting vegetation distributions using decision tree analysis in a geographic information system. *Environmental Management*, **15**, 59–71.

Nix, H.A. (1991) Biogeography: pattern and process. *Rainforest Animals: Atlas of Vertebrates Endemic to Australia's Wet Tropics* (ed. H.A. Nix & M.A. Switzer), pp. 11–39. Australian National Parks and Wildlife Service, Canberra, Australia.

Ohmann, J.L. & Gregory, M.J. (2002) Predictive mapping of forest composition and structure with direct gradient analysis and nearest-neighbor imputation in coastal Oregon, USA. *Canadian Journal of Forest Research*, **32**, 725–741.

Olden, J.D. (2003) Species-specific approach to modelling biological communities and its potential for conservation. *Conservation Biology*, **17**, 854–863.

Overton, J.M., Stephens, R.T.T., Leathwick, J.R. & Lehmann, A. (2002) Information pyramids for informed biodiversity conservation. *Biodiversity and Conservation*, **11**, 2093–2116.

Pausas, J.G. (1994) Species richness in the understorey of Pyrenean *Pinus sylvestris* forest. *Journal of Vegetation Science*, **5**, 517–524.

Pavoine, S., Dufour, A. & Chessel, D. (2004) From dissimilarities among species to dissimilarities among communities: a double principal coordinate analysis. *Journal of Theoretical Biology*, **228**, 523–537.

Peppler-Lisbach, C. & Schröder, B. (2004) Predicting the species composition of *Nardus stricta* communities by logistic regression modelling. *Journal of Vegetation Science*, **15**, 623–634.

Peters, D. & Thackway, R. (1998) *A New Biogeographic Regionalisation for Tasmania.* Tasmanian Parks and Wildlife Service, Hobart, Tasmania. http://www.gisparks.tas.gov.au/dp/newibra/, accessed 23.02.06.

Peterson, A.T., Soberon, J. & Sanchez-Cordero, V. (1999) Conservatism of ecological niches in evolutionary time. *Science*, **285**, 1265–1267.

Saetersdal, M., Birks, H.J.B. & Peglar, S.M. (1998) Predicting changes in Fennoscandian vascular-plant species richness as a result of future climatic change. *Journal of Biogeography*, **25**, 111–122.

Scott, J.M., Heglund, P.J., Haufler, J.B., Morrison, M., Raphael, M.G., Wall, W.B. & Samson, F. (2002) *Predicting Species Occurrences: Issues of Accuracy and Scale*. Island Press, Covelo, CA.

Scotts, D. & Drielsma, M. (2003) Developing landscape frameworks for regional conservation planning: an approach integrating fauna spatial distributions and ecological principles. *Pacific Conservation Biology*, **8**, 235–254.

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

Strahler, A.H., Logan, T.L. & Bryant, N.A. (1978) Improving forest cover classification accuracy from Landsat by incorporating topographic information. *Proceedings of the 12th International Symposium on Remote Sensing of the Environment* (ed. Anonymous), pp. 927–942. Environmental Research Institute of Michigan, Ann Arbor, MI.

Webb, C.O., Ackerly, D.D., McPeek, M.A. & Donoghue, M.J. (2002) Phylogenies and community ecology. *Annual Review of Ecology and Systematics*, **33**, 475–505.

Wohlgemuth, T. (1998) Modelling floristic species richness on a regional scale: a case study in Switzerland. *Biodiversity and Conservation*, **7**, 159–177.

Yee, T.W. & Mackenzie, M. (2002) Vector generalized additive models in plant ecology. *Ecological Modelling*, **157**, 141–156.

Zaniewski, A.E., Lehmann, A. & Overton, J.M. (2002) Predicting species distribution using presence-only data: a case study of native New Zealand ferns. *Ecological Modelling*, **157**, 259–278.

Zimmermann, N.E. & Kienast, F. (1999) Predictive mapping of alpine grasslands in Switzerland: species versus community approach. *Journal of Vegetation Science*, **10**, 469–482.

Supplementary material

The following supplementary material is available as part of the online article (full text) from http://w.w.w.blackwell-synergy.com.

Appendix S1. Additional information on modelling strategy 1 (assemble first, predict later).

Appendix S2. Additional information on modelling strategy 2 (predict first, assemble later).

Appendix S3. Additional information on modelling strategy 3 (assemble and predict together).

Appendix S4. The potential role of ecological assembly rules.

Figure S1. Proposed use of ecological assembly rules (EAR), based on characteristics of species or of whole ecosystems (global), to constrain the derivation of communities from modelled species' distributions.