

# Incorporating distance constraints into species distribution models

Omri Allouche\*, Ofer Steinitz, Dotan Rotem, Arik Rosenfeld and Ronen Kadmon

Department of Evolution, Systematics and Ecology, Institute of Life Sciences, The Hebrew University of Jerusalem, Givat-Ram, Jerusalem 91904, Israel

---

## Summary

1. Species distribution models (SDM) are increasingly applied as predictive tools for purposes of conservation planning and management. Such models rely on the concept of the ecological niche and assume that distribution patterns of the modelled species are at some sort of equilibrium with the environment. This assumption contrasts with empirical evidence indicating that distribution patterns of many species are constrained by dispersal limitation.

2. We demonstrate that the performance of SDM based on presence-only data can be significantly enhanced by incorporating distance constraints (functions relating the likelihood of species' occurrences at a site to the distance of the site from known presence locations) to the modelling procedure. This result is highly consistent for a variety of niche-based models (ENFA, DOMAIN and Mahalanobis distance), distance functions (nearest neighbour distance, cumulative distance and Gaussian filter) and taxonomic groups (plants, snails and birds, a total of 226 species).

3. Distance constraints are expected to enhance the accuracy of niche-based models even in the absence of strong dispersal limitation by accounting for mass effects and spatial autocorrelation in environmental factors for which data are not available.

4. While distance-based methods outperformed niche-based models when all data were used, their accuracy deteriorated sharply with smaller sample sizes. Niche-based methods are shown to cope better with small sample sizes than distance-based methods, demonstrating the potential advantage of niche-based models when calibration data are limited.

5. *Synthesis and applications.* Incorporating distance constraints in SDM provides a simple yet powerful method to account for spatial autocorrelation in patterns of species distribution, and is shown empirically to improve significantly the performance of such models. We therefore recommend incorporating distance constraints in future applications of SDM.

**Key-words:** bioclimatic models, climatic envelope, dispersal limitation, mass effect, predictive accuracy, spatial autocorrelation

## Introduction

Species distribution models (SDM; *sensu* Guisan & Thuiller 2005) are increasingly applied for purposes of conservation planning and ecosystem management. In spite of substantial differences among modelling methods (for contemporary reviews see Guisan & Zimmermann 2000; Scott *et al.* 2002), all SDM rely, explicitly or implicitly, on the concept of the ecological niche (Hutchinson 1957). Practically, this hypothesis implies that species are at some sort of equilibrium with their environment, and that distribution ranges can be predicted

based on the environmental characteristics of locations where the species were observed to occur (Guisan & Thuiller 2005). This assumption contrasts with empirical evidence indicating that patterns of species distribution are often constrained by dispersal limitation (for a review see Nathan 2001). Surprisingly, while the idea that dispersal may limit distribution ranges of species has been recognized for decades, most SDM ignore the potential consequences of dispersal limitation (Guisan *et al.* 2006).

We propose a simple methodology that allows for dispersal limitation in SDM without explicitly incorporating dispersal mechanisms into the model. The essence of our approach is the incorporation of distance constraints in the modelling algorithm. We use the term distance constraints to denote

\*Correspondence author. E-mail: omri.allouche@gmail.com

mathematical expressions that relate the likelihood of a species' occurrence at a site to the distance of the site from locations where the species has been documented to occur. Thus distance constraints can be considered as a means of incorporating spatial autocorrelation into SDM. Autocorrelated patterns of species' distributions have previously been modelled using autoregressive models (Smith 1994; Augustin, Muggleston & Buckland 1996; Gelfand *et al.* 2005; Betts *et al.* 2006; Dormann 2007) but these models require both presence and absence data for model construction. In contrast, distance constraints can be applied easily with presence-only data, like those provided by museum collections and observational databases (Ponder *et al.* 2001; Hirzel *et al.* 2002; Zaniwski, Lehmann & Overton 2002; Reutter *et al.* 2003). This capability is a significant advantage because presence-only data are much more available and easier to collect than presence-absence data. Moreover, taking into account that factors affecting species' distributions always show some level of autocorrelation, distance constraints can be expected to enhance the accuracy of SDM even if dispersal is not a strong limiting factor.

We demonstrate the applicability of distance constraints to species distribution modelling using three simple distance functions: nearest neighbour distance, cumulative distance and Gaussian filter (Davis 1986; Davies 1990; Snell, Gopal & Kaufmann 2000). Each distance function was applied alone (hereafter distance-based models) as well as in combination with niche-based models (hereafter hybrid models) to predict patterns of species distribution. Three different niche-based models were examined: ecological niche factor analysis (ENFA; Hirzel *et al.* 2002), DOMAIN (Carpenter, Gillison & Winter 1993) and the Mahalanobis distance (MD; Farber & Kadmon 2003). We also evaluated the performance of each niche-based SDM without incorporating distance constraints.

Each of the models was applied using presence-only data from existing databases on the distribution of 226 species of woody plants, land snails and nesting land birds in Israel. The predictions of the models were validated using presence-absence data obtained independently from an extensive sampling project that was designed for this purpose. This extensive design enabled us to evaluate our approach under a wide spectrum of modelling techniques and to assess the robustness of our results to differences in the modelling algorithms and the modelled taxa.

We tested three predictions. First, we verified our assumption that species tend to occur in sites that are both environmentally similar and geographically close to other occupied sites. Secondly, we tested the prediction that incorporating distance constraints into niche-based models enhances the accuracy of model predictions. Thirdly, we tested whether niche-based methods differ from distance-based methods in their sensitivity to variation in the amount of data available for model construction. Specifically we predicted that both methods would perform well when large data sets are available for model construction (i.e. when the entire distribution range is well represented by the available records) but that niche-based methods would perform better when only a small amount of data are available for model calibration. This

hypothesis lies at the heart of niche-based modelling but has rarely been put into an explicit test.

## Methods

### CONSTRUCTION OF SDM

SDM were constructed using data obtained from collections and observational databases of all relevant institutions in Israel. We performed field sampling in 27 sites for validation purposes. Only species documented in our sampling sites were selected for the analysis. The plant data were compiled from five sources: the herbarium of the Hebrew University, the Database Unit of the Israel Nature and Parks Authority, the database of the Society for the Protection of Nature in Israel, the Israeli Gene Bank and a database of plant observations developed by A. Danin. The snail data were obtained from the mollusc collection of the Hebrew University. Records of bird distribution were obtained from the Zoological Collections of Tel Aviv University and the Database Unit of the Israel Nature and Parks Authority. A total of 59 457 geo-referenced records of plants (49 466 records, 174 species), land snails (2507 records, 23 species) and nesting land birds (7484 records, 29 species) were compiled from the different sources. Records positioned within any of the field sampling sites were removed from the data set in order to increase independence between the calibration and validation data sets.

Three climatic variables were used as predictors in the SDM: mean annual rainfall, mean daily temperature of the hottest month (August) and mean minimum temperature of the coldest month (January). These variables were chosen because they showed high correlations with other climatic variables in the study area but relatively low correlations among themselves (Steinitz *et al.* 2005). Together, these variables captured the main climatic gradients of Israel and previous studies have shown that they are important determinants of distribution ranges of plants (Kadmon & Danin 1999), land snails (Heller 1988; Kadmon & Heller 1998; Steinitz *et al.* 2005) and birds (Shirihai 1996; Steinitz *et al.* 2005) in the study area. Adding additional variables (to a total of 23 variables) to the SDM did not improve the accuracy of model predictions. Digital maps of rainfall (Kadmon & Danin 1997) and temperature (Kurtzman & Kadmon 1999) were obtained from the GIS Center of the Hebrew University and rescaled into a spatial resolution of 1 km<sup>2</sup> to fit the spatial resolution of the field sampling.

Niche-based SDM were constructed using three different presence-only methods, ENFA (Hirzel *et al.* 2002), DOMAIN (Carpenter, Gillison & Winter 1993) and MD (Farber & Kadmon 2003). Each method applies a different algorithm to define the ecological niche of the species. ENFA calculates a measure of habitat suitability based on an analysis of marginality (how the species' mean differs from the mean of all sites in the study area) and environmental tolerance (how the species' variance compares with the global variance of all sites). DOMAIN uses a point-to-point similarity metric (based on the Gower distance) to assign a value of habitat

suitability to each potential site based on its proximity in the environmental space to the closest (most similar) occurrence location. MD ranks potential sites by their Mahalanobis distance to a vector expressing the mean environmental conditions (i.e. the centroid) of all the species' records in the environmental space. All models were implemented within the MATLAB environment (The MathWorks Inc.). ENFA was simulated using the Medians algorithm to produce results equal to those calculated by BIOMAPPER (Hirzel, Hausser & Perrin 2004). Box-Cox transformation of the environmental variables produced slightly poorer results and was therefore not used.

To enable a standardized evaluation of all models, all predictions were expressed as binary (presence-absence) predictive maps. Threshold values were applied to transform the predictions generated by the various models to binary predictions. For each model, the threshold that maximized the overall value of Kappa for all species was selected (Collingham *et al.* 2000; Pearson *et al.* 2002).

DISTANCE CONSTRAINTS

Various methods can be applied to calculate the likelihood of species' occurrences at a site based on the geographical locations of known occurrences. In this study we chose to consider three basic functions that can easily be incorporated into presence-only SDM, nearest-neighbour distance, cumulative distance and Gaussian filter. These basic methods were chosen to demonstrate the applicability of incorporating distance constraints into SDM. Future research might consider more complex methods, for example indices that take into account the dispersal properties of the modelled species. The nearest-neighbour distance scores a potential site by its Euclidean distance to the closest presence site in the calibration data set (Snell, Gopal & Kaufmann 2000). The cumulative distance sums the inverse of the squared Euclidean distances of a potential site to all presence sites in the calibration data set (Davis 1986). The Gaussian filter creates a Gaussian with a selected standard deviation around each presence site in the calibration data set and assigns to each site the sum of all Gaussians in that site (Davies 1990). The formulae for the three methods are given in Table 1. Distance thresholds were applied to transform predictions produced by each distance-based model into binary presence-absence maps. As with the niche-based approaches, the threshold that maximized the value of the Kappa statistic over all species was selected for each method.

**Table 1.** Distance functions used in this study. *P* denotes the group of records used for model calibration,  $D(i, j)$  is the Euclidean (geographical) distance between sites *i* and *j*.  $\sigma^2$  denotes the variance of the Gaussian filter

Nearest neighbour	$S_i = \min_{j \in P} D(i, j)$
Cumulative distance	$S_i = \sum_{j \in P} \frac{1}{(D(i, j))^2}$
Gaussian filter	$S_i = \sum_{j \in P} G(i, j) = \sum_{j \in P} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(D(i, j))^2}{2\sigma^2}}$

Hybrid models were produced using all combinations of niche-based and distance-based methods. In these models, a species was predicted to occur if both the niche-based method and the distance-based method predicted so. Thus the predicted distribution range was the intersection of the niche-based and distance-based predictions. The combination of environmental and geographical thresholds applied to each model was selected to maximize the value of Kappa over all species. The selected thresholds for all models are summarized in Table 2.

FIELD SAMPLING OF VALIDATION DATA

Data for model validation were collected in 27 sampling sites of 1 × 1 km covering the main climatic gradients of Israel. Of these, woody plants were sampled at 25 sites, land snails were sampled at all sites and nesting land birds were sampled at 21 sites. The spatial distribution of the sampling sites was determined using a novel approach that was designed to minimize the correlation of environmental similarity and geographical distance between the selected sites (Steinitz *et al.* 2005). This approach enabled us to separate better the effects of environmental similarity and geographical distance on patterns of species distribution.

Within-site sampling protocols were determined based on sample-based rarefaction analyses (Gotelli & Colwell 2001) of preliminary data obtained from several representative ecosystems (O. Steinitz, D. Rotem & A. Rozenfeld, unpublished data). Woody plants were documented in nine regularly spaced plots of 0.1 ha at each site. All species of woody plants were documented within each plot independently of their abundance, age and stage of development. Snails were recorded based on empty shells in nine plots of 0.01 ha, nested within the plots used for plant sampling. A searching

**Table 2.** Thresholds used to convert model outputs to predictions of presence-absence, for the different distance-based, niche-based and hybrid models. Values in bold type are niche-based thresholds; values in Roman type are distance-based thresholds

Gaussian filter	Cumulative distance		Nearest neighbour		No geographical distance		
0.014		0.38		4.5			No environmental distance
0.0131	<b>0.15</b>	0.25	<b>0.15</b>	4.5	<b>0.15</b>	<b>0.4</b>	ENFA
0.0138	<b>17.5</b>	0.37	<b>17</b>	4.5	<b>10</b>	<b>3.5</b>	MD
0.0138	<b>0.86</b>	0.37	<b>0.905</b>	4.5	<b>0.905</b>	<b>0.975</b>	DOMAIN

effort of 12 min was allocated for snail sampling in each plot. Only snail species with shells larger than 5 mm were documented because accurate sampling of microsnails required sampling effort that was not feasible at such scales. Nesting land birds were sampled in five regularly spaced plots within each site using point counts (Bibby *et al.* 2000). At each point, birds seen or heard within a 60-m radius were recorded for 10 min. Each site was sampled twice during the main breeding season. The first sampling period was during the spring (March–April) and the second during the summer (July–August). The sampling plots of each site were located in the field using GPS with a spatial accuracy of 12 m (Garmin 12XL; Garmin International Inc., Olathe, Kansas, USA). The data obtained from all plots of each site were pooled, and a site-by-species matrix of presence–absence data was constructed for each taxonomic group.

#### ACCURACY ASSESSMENT

Predictive maps produced by the various models were compared with the data set obtained from the field sampling for estimating their accuracy. The field data set consisted of 25, 27 and 21 validation sites for plants, snails and birds, respectively. Taking into account the number of species sampled in each taxonomic group (174, 23 and 29, respectively), the resulting validation data set contained a total of 5580 validation values. Predictions of each model were compared with the validation data set to form one confusion matrix with a total of 5580 cases from which Cohen's Kappa (Cohen 1960) was calculated. The Kappa statistic defines the accuracy of prediction, relative to the accuracy that might have resulted by chance alone. It ranges from  $-1$  to  $+1$ , where  $+1$  indicates perfect agreement between predictions and observations and values of 0 or less indicate agreement no better than random classification. By using a single confusion matrix for all species we reduced the bias caused by low prevalence of some species (McPherson, Jets & Rogers 2004; Allouche, Tsoar & Kadmon 2006). The area under the ROC (Receiver Operating Characteristic) curve (AUC; Fielding & Bell 1997; Manel, Williams & Ormerod 2001), a popular threshold-independent statistic, could not be computed for our hybrid models because these models integrated two different types of thresholds (environment and geographical distance). However, we also evaluated the accuracy of the predictive maps with the true skill statistics (TSS), a measure that is highly correlated with AUC and is not biased by prevalence (Allouche, Tsoar & Kadmon 2006). The results were similar to those obtained for our calculation of Kappa, so only the Kappa values are reported here. Averaging model performance for all species without lumping them into one confusion matrix did not change the results of our analysis and is therefore not reported.

#### DATA ANALYSIS

The assumption that species tend to occur in sites that are both environmentally similar and geographically close to other occupied sites was verified by comparing the environ-

mental and geographical characteristics of species records in the validation and calibration data sets. First, a random set of 12 validation sites was selected for each species, and each site in the selected set was classified as presence or absence. This procedure was repeated for each species. Then, for each site–species combination, we calculated a measure of geographical distance (the Euclidean distance to the nearest record of the species in the calibration data set) and a measure of environmental distance (the MD between the selected site and the centroid of all records of the relevant species). The latter measure can be considered the distance of the site from the environmental 'optimum' of the relevant species. By plotting the sites selected for each species in a two-dimensional space defined by the geographical and environmental distances, we were able to evaluate our assumption that species tend to be present in sites that are both environmentally similar and geographically close to other occupied sites.

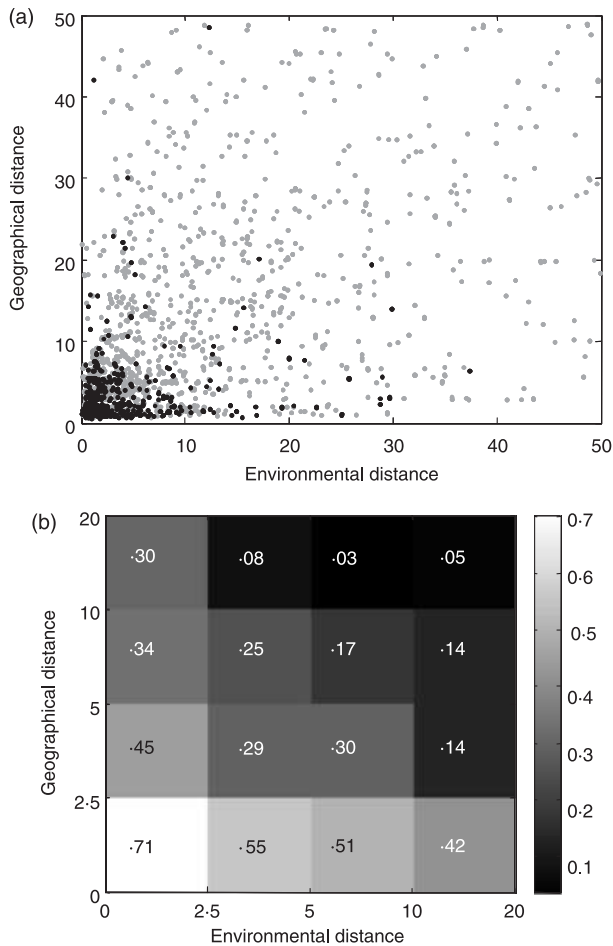
The prediction that incorporating geographical constraints into SDM enhances the accuracy of model predictions was tested by comparing the performance of hybrid models with that of the corresponding niche-based models. The significance of levels of differences between Kappa values of hybrid and niche-based models was determined based on the asymptotic standard errors of Kappa, as a single value of Kappa was generated for each model (Kraemer 1982; Blackman & Koval 2000).

The prediction that niche-based methods are less sensitive to small sample sizes than distance-based methods was tested using Monte Carlo simulations by resampling data from 103 species with more than 200 records in the calibration data set. For each species we produced distribution maps by all niche-based and distance-based methods using randomly selected sets of 5, 10, 20, 30, 50, 75, 100 and 200 records from the calibration data set. Two-hundred predictions (repetitions) were generated for each sample size. The accuracy of each map was determined using the Kappa statistic based on the data obtained from the field validation sites. As each random selection yielded a different prediction with different accuracy, we used the mean value of Kappa (averaged across all species) to characterize the performance of each model at each sample size.

## Results

#### ENVIRONMENTAL AND GEOGRAPHICAL DISTANCES

The environmental and geographical characteristics of the validation sites (12 random sites for each of the 226 species) are shown in Fig. 1. As expected, sites containing presences tended to have small environmental distances, as well as small geographical distances, while sites containing absences showed much higher values of both distances (Fig. 1a). Calculation of the relative frequency of presence records within different ranges of environmental and geographical distances indicated that both types of distances had a negative effect on the frequency of presence records (Fig. 1b). It could also be seen that there was a trade-off between the effects of environmental and geographical distances: small geographical distances compensated for higher environmental distances in determining

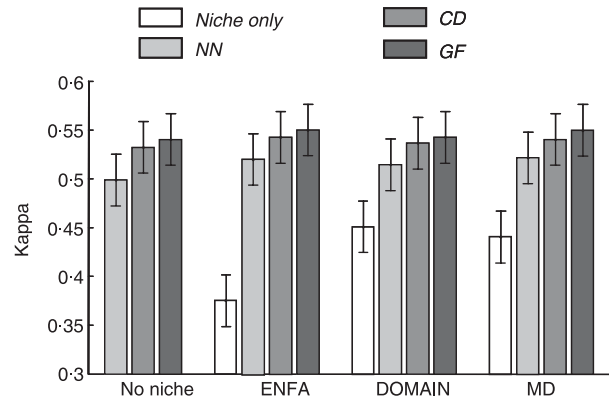


**Fig. 1.** Environmental and geographical characteristics of the validation sites. Twelve random sites were sampled for each species. Each site was classified as ‘presence’ or ‘absence’ with respect to the relevant species, and was assigned a measure of environmental distance (the MD between the site and the centroid of all the species records in the environmental space) and geographical distance (the Euclidean distance between the site and the closet species record in the calibration database). (a) Distribution of sites classified as presences (black) and absences (grey) in a two-dimensional space defined by the two distances. (b) Relative frequency of sites classified as presences within different categories of the two distances.

the frequency of presences and vice versa (Fig. 1b). Another pattern emerging from this analysis was that even sites that were environmentally close (in terms of their MD) to the ‘optimum’ of the relevant species were often characterized as absences. For example, of the 633 sites showing an MD lower than 2.5, almost half of the sites (46.9%) were classified as absences. This result indicated that species were often absent from sites that were environmentally suitable to them in terms of the variables included in our analysis.

#### EFFECT OF DISTANCE CONSTRAINTS ON PREDICTIVE ACCURACY

Kappa values obtained for the niche-based models ranged from 0.375 (ENFA) to 0.451 (DOMAIN). Incorporating distance constraints increased the accuracy of all models to



**Fig. 2.** Differences in predictive accuracy between distance-based models (NN, nearest neighbour; CD, cumulative distance; GF, Gaussian filter), niche-based models (ENFA, DOMAIN, and MD) and corresponding hybrid models. Each model was applied to 226 species of plants, snails and birds in Israel and its accuracy was determined using Cohen’s Kappa based on field validation data. Error bars indicate 95% confidence levels.

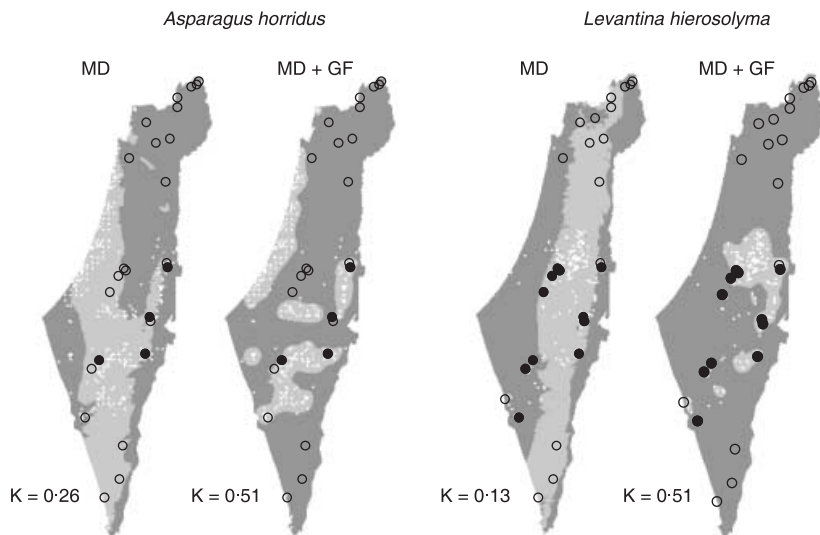
values of 0.514 and above (Fig. 2). In all cases, hybrid models were significantly more accurate than their corresponding niche-based models (Fig. 2). Adding distance constraints to niche-based models also improved the predictive accuracy for each taxonomic group separately (plants, snails and birds) for all combinations of modelling algorithms. These results supported the prediction that adding distance constraints to niche-based models improves the accuracy of model predictions.

Some differences were observed in the contribution of alternative distance functions to predictive accuracy. In general, hybrid models based on the cumulative distance and Gaussian filter performed better than corresponding models based on nearest neighbour distance (Fig. 2).

The consequences of adding distance constraints to niche-based models are illustrated using predictive maps produced for two species: the plant *Asparagus horridus* and the snail *Levantina hierosolyma* (Fig. 3). *Asparagus horridus* is a geophyte species that grows in rocky desert habitats but penetrates into the Mediterranean region along the coastal sands. *Levantina hierosolyma* is endemic to the southern Levant and occupies crevices in rocky habitats of the Judean mountains and the northern Negev desert in Israel. It can be seen that niche-based models overestimated the distribution of both species, resulting in low values of Kappa (Fig. 3). Incorporating distance constraints reduced errors caused by overestimation and, thus, increased the values of Kappa.

#### EFFECT OF SAMPLE SIZE ON PREDICTIVE ACCURACY

The size of the data set used for model calibration had a strong positive effect on the accuracy of all modelling methods (Fig. 4). However, different methods responded differentially to changes in the number of records included in the calibration data set. ENFA and MD showed a steep increase in predictive accuracy with increasing sample size and reached their asymptotic Kappa values at sample sizes of 20–50 records



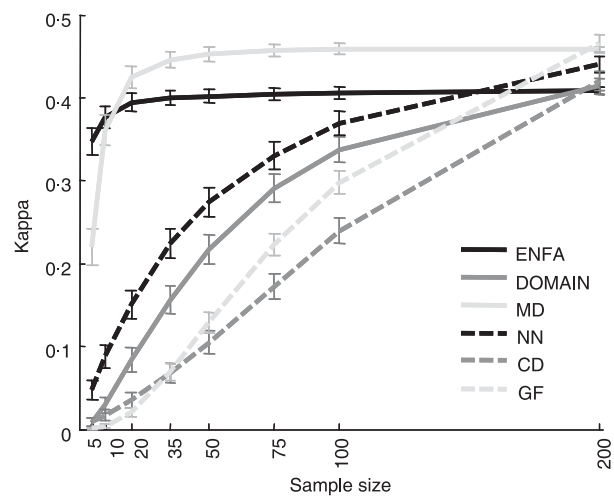
**Fig. 3.** Examples for predictive maps produced using niche-based vs. hybrid models. Maps shown were produced for the plant *Asparagus horridus* and the snail *Levantina hierosolyma* using MD and a hybrid model combining MD and a Gaussian filter (MD + GF). White dots are the records used for calibrating the model. The light grey area is the predicted distribution range. Circles are the validation sites (black, presence; open, absence). Kappa values are shown for each map.

(Fig. 4). DOMAIN and all three distance-based methods showed much more gradual responses and did not approach asymptotic levels of Kappa even at sample sizes of 200 records. The performance of the distance-based models was comparable with that of the niche-based models when large data sets were used for model calibration, but it declined sharply when smaller data sets were used for model calibration. Thus, except for DOMAIN, the results support the prediction that niche-based models cope better with small calibration data sets than distance-based models.

Predictive maps based on the entire set of data available for each species revealed Kappa values of 0.5 or higher for all distance-based methods (Fig. 2). It is interesting to note that when the entire data set was used for model construction, distance-based methods performed better than niche-based models in all cases, and the accuracy of their predictions was only slightly lower than that of the corresponding hybrid models (Fig. 2).

## Discussion

Our results demonstrate that incorporating distance constraints in SDM strongly improves the accuracy of model predictions (Fig. 2). This finding was consistent over all taxonomic groups, niche-based models and distance functions. It is interesting that the thresholds of geographical distances that maximized the values of Kappa in the hybrid models were relatively small. For example, the threshold of the nearest neighbour distance that maximized the value of Kappa across all species was 4.5 km. As a result, predictive maps generated by the hybrid models were highly patchy, a pattern that is very different from the continuous pattern typical of niche-based SDM (see examples in Fig. 3). These results suggest that patterns of species distribution are much patchier than is usually assumed, and that incorporating distance constraints enhances the ability of niche-based SDM to cope with such patchiness.



**Fig. 4.** Results of resampling experiments testing the effect of sample size (number of observations in the calibration data set) on predictive accuracy of niche-based models (ENFA, DOMAIN and MD) and distance-based models (nearest neighbour distance, cumulative distance and Gaussian filter). Analyses were performed using data for 103 species with > 200 calibration records. Two-hundred predictions (repetitions) were generated for each species at each sample size and the accuracy of each map was determined using Cohen's Kappa based on the field validation data. Error bars represent 1 SE.

## ENVIRONMENTAL AND GEOGRAPHICAL DISTANCES

As we expected, validation sites classified as presences were both environmentally similar and geographically close to sites in which the relevant species was previously documented to occur (Fig. 1a). Even within narrow ranges of environmental distances, the frequency of validation sites classified as presences decreased with increasing geographical distance from known occurrences (Fig. 1b). These results are consistent with previous studies showing that patterns of species distribution are characterized by strong spatial autocorrelation (Rushton, Ormerod & Kerby 2004; Heikkinen *et al.* 2004;

Karst, Gilbert & Lechowicz 2005; Luoto *et al.* 2005; Sanderson, Eyre & Rushton 2005; Betts *et al.* 2006; Segurado, Araújo & Kunin 2006) and may explain the observed superiority of the hybrid models over the niche-based models.

Of course, documenting spatial autocorrelation in distribution ranges does not provide much information about the mechanisms that generate these patterns. Basically, patterns of spatial autocorrelation in species' distributions reflect the combined effects of dispersal processes and spatial autocorrelation in the environment (Sokal & Oden 1978; Legendre 1993; Lichstein *et al.* 2002; Miller & Franklin 2002). Distinguishing between these two effects and evaluating the relative importance of each mechanism was beyond the scope of this study. However, a previous analysis of data collected in our study sites indicated that compositional similarity between any two sites was negatively correlated with the geographical distance between the sites. Such a negative correlation was obtained for snails and birds even after controlling for spatial autocorrelation in a wide spectrum of rainfall, temperature, lithology and vegetation variables (Steinitz *et al.* 2005). Further analyses indicated that the rate of decay in species' similarity with increasing geographical distance differed significantly between snails and birds, with snails showing much steeper distance decays than birds (Steinitz *et al.* 2006). All these findings are consistent with the hypothesis of dispersal limitation. It should also be noted that increasing the number of environmental variables used by the SDM to 23 variables did not lead to a significant improvement in predictive accuracy (data not shown). A similar result was obtained in a previous analysis of plant distribution in Israel (Farber & Kadmon 2003). These findings suggest that dispersal limitation plays an important role in determining the patterns of species' distributions observed in this study.

#### EFFECT OF DISTANCE CONSTRAINTS ON PREDICTIVE ACCURACY

All distance functions examined in this study improved the accuracy of predictions generated by pure niche-based models (Fig. 2). The fact that a significant improvement was documented for a wide range of modelling algorithms and taxonomic groups suggests that this finding is not unique to the specific combinations of algorithms and/or organisms examined in this study. We therefore recommend incorporating distance constraints in future applications of SDM. The approach proposed in this study is one method by which such constraints can be incorporated, and it has the advantage of being very simple and applicable to any niche-based model.

One unexpected result of our study was that even pure distance-based models outperformed niche-based models in all cases (Fig. 2). This finding contrasts with the common belief that niche-based models provide more accurate predictions of species' distributions than methods based on simple spatial interpolation. Similar results were obtained by Araújo & Williams (2000) for models based on presence-absence data that were applied to 174 native tree species and subspecies distributed across Europe. In their study distance constraints

were modelled by contagion, an index relating the likelihood of species' occurrences in a site to the number and distance of presence records in the neighbourhood of that site. As in our study, predictions of distance-based models were more accurate than those of niche-based models and the best predictions were obtained using hybrid models (niche-based logistic regressions incorporating the effect of contagion as an additional predictor). Even in the hybrid models, contagion showed a higher frequency of significant effects than any of the environmental variables (Araújo & Williams 2000). Incorporating the contagion index as a distance function in our models produced poor results (data not shown), probably because our analysis was carried out at a much higher spatial resolution.

Incorporating distance constraints may improve the accuracy of SDM by accounting for dispersal limitation, which prevents species colonizing isolated or remote sites even if these are ecologically suitable (Guisan *et al.* 2006). However, distance constraints may also improve the performance of niche-based models by accounting for spatially autocorrelated factors that are not included as predictors in the model (Legendre 1993; Lichstein *et al.* 2002; Miller & Franklin 2002; Barry & Elith 2006; Guisan *et al.* 2006). Taking into account the fact that environmental factors are always autocorrelated to some extent, and that predictive models never include the full range of variables involved in determining distribution patterns, distance constraints can be expected to improve the performance of niche-based models even in the absence of any limitation by dispersal. Yet the degree to which incorporating distance constraints actually improves predictive accuracy can be expected to depend on the reliability of the calibration data. If the data available for model calibration do not represent the actual distribution of the modelled species, incorporating distance constraints may not contribute and may even deteriorate the accuracy of model predictions by underestimating the true distribution range. Obviously, distance-based methods cannot be applied to regions with no data while niche-based methods can, given that the calibration area represents the entire environmental space of the species. Therefore, some useful applications of niche-based models, such as evaluation of the spreading potential of invading species (Higgins *et al.* 1999; Peterson & Vieglais 2001; Peterson & Robins 2003; Rouget *et al.* 2004) and the identification of potential areas for successful re-introduction of endangered species (Engler, Guisan & Rechsteiner 2004; Bourg, McShea & Gill 2005), cannot be enhanced by incorporating distance constraints. It should also be emphasized that the simple distance-based functions we used do not account explicitly for dispersal processes and are therefore of limited use for studies of species' responses to climate change. Models based on cellular automata are more useful in such cases (Carey 1996; Iverson, Prasad & Schwartz 1999; Iverson, Schwartz & Prasad 2004; Midgley *et al.* 2006).

#### EFFECT OF SAMPLE SIZE

Our simulations demonstrate that the effect of sample size on the accuracy of model predictions depends on the specific



modelling method (Fig. 4). ENFA and MD reached asymptotic values of Kappa at sample sizes of 50 records, a result consistent with previous analyses of the performance of niche-based models (Nix 1986; McKenney *et al.* 1998; Stockwell & Peterson 2002; Kadmon, Farber & Danin 2003). Distance-based methods produced predictions that were comparable with those of niche-based models at the highest sample size (200 records) but their accuracy deteriorated sharply at smaller sample sizes (Fig. 4). These findings support our prediction that niche-based methods cope better with small sample sizes than distance-based methods, and demonstrate the potential advantage of niche-based models when calibration data are limited.

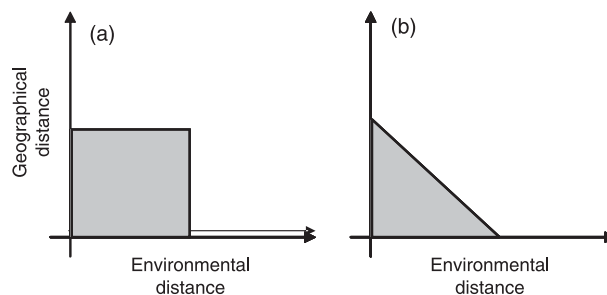
The results obtained for DOMAIN were exceptional and showed a pattern that was more similar to the distance-based methods (Fig. 4). We attribute this anomaly to the fact that DOMAIN defines the suitability of a potential site based on its distance to the closest species record in the environmental hyperspace. As the distance to a single record is used, exclusion of this record from the data set affects greatly the performance of the model. The two other niche-based methods (ENFA and MD) use all records in the data set for scoring each potential site, and are therefore more robust to exclusion of individual records from the calibration data set. It is interesting to note that, in contrast with its poor performance in simulations based on small sample sizes (Fig. 4), DOMAIN was the most accurate niche-based model in tests based on the entire calibration data set (Fig. 2). This result is consistent with the fact that DOMAIN was the only niche-based model that did not reach saturation at the highest sample size (Fig. 4).

The fact that distance-based methods performed better than niche-based methods when applied to the entire data set, in spite of being more sensitive to simulated small sample sizes, may indicate the high quality of the calibration data used in this study. However, our calibration data set was far from being optimal. A naive distance-based model, predicting presence if and only if the calibration data set contained a record at the relevant site and absence otherwise, performed no better than random (Kappa = 0.054).

Our results suggest that comparative analyses of alternative SDM should pay more attention to the sensitivity of model predictions to variation in the size of the calibration data set. According to our results, conclusions derived from such comparisons (and thus the selection of 'optimal' models for specific applications) may be biased by the size (and maybe other characteristics) of the specific data set used in the evaluation. Such interactions between properties of the data and properties of the model have rarely been investigated (but see Stockwell & Peterson 2002; Hernandez *et al.* 2006) and should receive more attention in future applications of SDM.

#### METHODOLOGICAL CONSIDERATIONS

The method we used to combine the effects of geographical and environmental thresholds in our hybrid models implicitly assumes independence of the two thresholds (Fig. 5a). More sophisticated approaches may assume compensation of one



**Fig. 5.** Two possible approaches for determining environmental and distance thresholds for transforming continuous predictions of hybrid SDM to binary presence–absence maps. (a) Rectilinear envelope (the two thresholds are assumed to be independent). (b) Triangular envelope (a trade-off is assumed between the two thresholds). Grey areas indicate prediction of presence.

threshold for the other to reflect better possible interactions between the effects of geographical and environmental distances (Fig. 5b). If, for example, a species is highly prevalent in the neighbourhood of a relatively unsuitable site, it is also likely to be present in the unsuitable site because of a mass effect (Shmida & Wilson 1985). On the other hand, if a site is highly suitable in its ecological conditions, it is likely to be occupied even if it is relatively distant from other presence locations, as rare colonization events might be enough for the species to establish and maintain itself. Our finding that small geographical distances may compensate for large environmental distances in determining the frequency of occupied sites (Fig. 1b) is consistent with these assumptions. Another indication for such compensation is that the thresholds of geographical distances that maximized the accuracy of hybrid models were larger than those maximizing the accuracy of pure distance-based models. Further support for the assumption of threshold compensation comes from an analysis of factors affecting the degree of similarity in species composition among our study sites (Steinitz *et al.* 2006), which indicated that a small geographical distance between sites may compensate for large environmental distances in determining the degree of between-site similarity in species composition. We therefore recommend that future developments of hybrid models should take into account the potential implications of a trade-off between the environmental and geographical thresholds.

When interpreting predictive maps produced by SDM, one should take into account source–sink population dynamics (Pulliam 1988, 2000). Source–sink dynamics may deteriorate the predictive power of SDM by two contrasting effects. First, including records from sink populations in the calibration data set may lead to overestimation of the fundamental niche. Secondly, and in contrast, predictions based on records from source populations may underestimate the actual distribution range if part of the distribution range consists of sink populations. Metapopulation dynamics further deteriorate the predictive power of SDM because species may be absent from source sites as a result of stochastic local extinctions. Incorporating distance constraints into SDM may reduce the



magnitude of errors generated by both mechanisms because sites close to established populations have a higher probability of being colonized if empty and a lower probability of extinction if they are inhabited (because of the rescue effect; Brown & Kodric-Brown 1977) than sites distant from established populations.

Scale is another major factor that should be considered in interpreting the results of SDM (Pearson & Dawson 2003; Thuiller, Araújo & Lavorel 2004; Guisan & Thuiller 2005). At relatively coarse spatial scales, climatic factors play an important role in determining patterns of species' distributions, which may explain the relatively high success of climate-based SDM at such scales (Venier *et al.* 1999; Pearson *et al.* 2002). At smaller scales, factors such as land cover, disturbance and biotic interactions become increasingly important (Willis & Whittaker 2002; Pearson & Dawson 2003). The idea that factors affecting species' distributions are scale-dependent has long been accepted by ecologists (Wiens, Rotenberry & Vanhorne 1987; Levin 1992; Willis & Whittaker 2002) but has rarely been implemented in SDM. Recently, Pearson, Dawson & Liu (2004) offered a hierarchical framework for SDM, where climate determines species' distributions at relatively large scales, and additional factors are incorporated to predict distribution patterns at finer scales. Our modelling approach can be considered as an implicit implementation of this approach, with distance constraints functioning as surrogates for dispersal limitation and for environmental factors operating at relatively small spatial scales.

Soberon & Peterson (2005; see also Peterson 2006) proposed a conceptual framework for distribution modelling that combines the idea that factors affecting distribution ranges are scale-dependent with the concept of the ecological niche. According to their approach, a species is expected to occur at a site only if three conditions are satisfied: (i) the abiotic conditions at the site are favourable; (ii) appropriate biotic conditions are satisfied (presence of hosts, food plants, pollinators, etc., and absence of competitors, diseases and predators); and (iii) the site is accessible to dispersal by the species from established populations. They further argued that the first condition defines the fundamental niche of the species while its intersection with the second condition defines the realized niche. The intersection of the three conditions defines the actual geographical distribution of the species (although sink populations may occur even if the first two conditions are not satisfied). According to this framework, niche-based SDM predict the fundamental niche (Soberon & Peterson 2005) or the realized niche (Pearson & Dawson 2003). Still, in many practical applications of SDM, particularly those applied for conservation and management purposes, the main aim is to predict the actual distribution of the modelled species. Incorporating distance constraints as modelled in this study is one possible approach for approximating actual distribution ranges better in such applications.

In a recent review of species distribution models, Guisan *et al.* (2006) highlighted the need for strengthening the link between ecological theory and modelling tools and called for more consideration of spatial aspects in SDM. Most SDM

still largely ignore issues such as spatial autocorrelation, dispersal limitation and biotic interactions, or deal with a single issue at a time (Guisan *et al.* 2006). Our method provides a simple yet powerful tool to account for all of the above issues simultaneously, and connects better SDM to ecological theory.

## CONCLUSIONS

Almost any textbook in ecology emphasizes the importance of dispersal as one of the main factors that limit the distribution of species. Yet most applications of SDM ignore the potential consequences of dispersal limitation. In this study we propose a simple approach by which dispersal limitation can be incorporated into SDM and provide empirical evidence that the proposed approach significantly enhances the accuracy of model predictions. Even if dispersal is not a limiting factor, our approach can be expected to enhance the predictive power of SDM by accounting for spatial mass effects and spatial autocorrelation in environmental factors that are not included as predictors in the model. The fact that a significant improvement in predictive accuracy was obtained for different taxonomic groups, different modelling techniques and different distance functions is encouraging, and suggests that our approach can be applied successfully to a wide range of systems.

## Acknowledgements

We thank A. Tsoar for valuable comments on the paper. The research was supported by The Israel Science Foundation (grant no. 545/03) and the Ministry of Environment.

## References

- Allouche, O., Tsoar, A. & Kadmon, R. (2006) Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, **43**, 1223–1232.
- Araújo, M.B. & Williams, P.H. (2000) Selecting areas for species persistence using occurrence data. *Biological Conservation*, **96**, 331–345.
- Augustin, N.H., Muggleston, M.A. & Buckland, S.T. (1996) An autologistic model for the spatial distribution of wildlife. *Journal of Applied Ecology*, **33**, 339–347.
- Barry, S. & Elith, J. (2006) Error and uncertainty in habitat models. *Journal of Applied Ecology*, **43**, 413–423.
- Betts, M.G., Diamond, A.W., Forbes, G.J., Villard, M.A. & Gunn, J.S. (2006) The importance of spatial autocorrelation, extent and resolution in predicting forest bird occurrence. *Ecological Modelling*, **191**, 197–224.
- Bibby, C.J., Burgess, N.D., Hill, D.A. & Mustoe, S.H. (2000) *Bird Census Techniques*. Academic Press, London, UK.
- Blackman, N.J.M. & Koval, J.J. (2000) Interval estimation for Cohen's kappa as a measure of agreement. *Statistics in Medicine*, **19**, 723–741.
- Bourg, N.A., McShea, W.J. & Gill, D.E. (2005) Putting a cart before the search: successful habitat prediction for a rare forest herb. *Ecology*, **86**, 2793–2804.
- Brown, J.H. & Kodric-Brown, A. (1977) Turnover rates in insular biogeography: effect of immigration on extinction. *Ecology*, **58**, 445–449.
- 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.
- Carpenter, G., Gillison, A.N. & Winter, J. (1993) Domain: a flexible modelling procedure for mapping potential distributions of plants and animals. *Biodiversity and Conservation*, **2**, 667–680.
- Cohen, J. (1960) A coefficient of agreement of nominal scales. *Educational and Psychological Measurement*, **20**, 37–46.
- Collingham, Y.C., Wadsworth, R.A., Huntley, B. & Hulme, P.E. (2000) Predicting the spatial distribution of non-indigenous riparian weeds: issues of spatial scale and extent. *Journal of Applied Ecology*, **37**, 13–27.

- Davies, E. (1990) *Machine Vision: Theory, Algorithms and Practicalities*. Academic Press, San Diego, CA.
- Davis, J.C. (1986) *Statistics and Data Analysis in Geology*, 2nd edn. John Wiley and Sons, New York, NY.
- Dormann, C.F. (2007) Effects of incorporating spatial autocorrelation into the analysis of species distribution data. *Global Ecology and Biogeography*, **16**, 129–138.
- Engler, R., Guisan, A. & Rechsteiner, L. (2004) An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *Journal of Applied Ecology*, **41**, 263–274.
- Farber, O. & Kadmon, R. (2003) Assessment of alternative approaches for bioclimatic modelling with special emphasis on the Mahalanobis distance. *Ecological Modelling*, **160**, 115–130.
- Fielding, A.H. & Bell, J.F. (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, **24**, 38–49.
- 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. *Journal of the Royal Statistical Society Series C, Applied Statistics*, **54**, 1–20.
- Gotelli, N.J. & Colwell, R.K. (2001) Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters*, **4**, 379–391.
- 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., Lehmann, A., Ferrier, S., Austin, M., Overton, J.M.C., Aspinall, R. & Hastie, T. (2006) Making better biogeographical predictions of species' distributions. *Journal of Applied Ecology*, **43**, 386–392.
- Heikkinen, R.K., Luoto, M., Virkkala, R. & Rainio, K. (2004) Effects of habitat cover, landscape structure and spatial variables on the abundance of birds in an agricultural–forest mosaic. *Journal of Applied Ecology*, **41**, 824–835.
- Heller, J. (1988) The biogeography of the land snails of Israel. *The Zoogeography of Israel* (eds Y. Yom-Tov & E. Tchernov), pp. 325–353. Dr W. Junk Publishers, Dordrecht, the Netherlands.
- Hernandez, P.A., Graham, C.H., Master, L.L. & Albert, D.L. (2006) The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography*, **29**, 773–785.
- Higgins, S.I., Richardson, D.M., Cowling, R.M. & Trinder-Smith, T.H. (1999) Predicting the landscape-scale distribution of alien plants and their threat to plant diversity. *Conservation Biology*, **13**, 303–313.
- 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., Hausser, J. & Perrin, N. (2004) *Biomapper 3-0*. Division of Conservation Biology, Bern, Switzerland. <http://www.unil.ch/biomapper/>, accessed June 19, 2005.
- Hutchinson, G.E. (1957) Concluding remarks. *Cold Spring Harbor Symposium on Quantitative Biology*, **22**, 415–457.
- Iverson, L.R., Prasad, A. & Schwartz, M.W. (1999) Modeling potential future individual tree-species distributions in the eastern United States under a climate change scenario: a case study with *Pinus virginiana*. *Ecological Modelling*, **115**, 77–93.
- Iverson, L.R., Schwartz, M.W. & Prasad, A.M. (2004) How fast and far might tree species migrate in the eastern United States due to climate change? *Global Ecology and Biogeography*, **13**, 209–219.
- Kadmon, R. & Danin, A. (1997) Floristic variation in Israel: a GIS analysis. *Flora*, **192**, 341–345.
- Kadmon, R. & Danin, A. (1999) Distribution of plant species in Israel in relation to spatial variation in rainfall. *Journal of Vegetation Science*, **10**, 421–432.
- Kadmon, R. & Heller, J. (1998) Modelling faunal responses to climatic gradients with GIS: land snails as a case study. *Journal of Biogeography*, **25**, 527–539.
- Kadmon, R., Farber, O. & Danin, A. (2003) A systematic analysis of factors affecting the performance of climatic envelope models. *Ecological Applications*, **13**, 853–867.
- Karst, J., Gilbert, B. & Lechowicz, M.J. (2005) Fern community assembly: the roles of chance and the environment at local and intermediate scales. *Ecology*, **86**, 2473–2486.
- Kraemer, H.C. (1982) Kappa coefficient. *Encyclopedia of Statistical Sciences* (eds S. Kotz & N.L. Johnson), pp. 352–354. John Wiley and Sons, New York, NY.
- Kurtzman, D. & Kadmon, R. (1999) Mapping of temperature variables in Israel: a comparison of different interpolation methods. *Climate Research*, **13**, 33–43.
- Legendre, P. (1993) Spatial autocorrelation: trouble or new paradigm. *Ecology*, **74**, 1659–1673.
- Levin, S.A. (1992) The problem of pattern and scale in ecology. *Ecology*, **73**, 1943–1967.
- Lichstein, J.W., Simons, T.R., Shiner, S.A. & Franzreb, K.E. (2002) Spatial autocorrelation and autoregressive models in ecology. *Ecological Monographs*, **72**, 445–463.
- Luoto, M., Poyry, J., Heikkinen, R.K. & Saarinen, K. (2005) Uncertainty of bioclimate envelope models based on the geographical distribution of species. *Global Ecology and Biogeography*, **14**, 575–584.
- McKenney, D.W., Mackey, B.G., Bogart, J.P., Mckee, J.E., Oldham, M.J. & Chek, A. (1998) Bioclimatic and spatial analysis of Ontario reptiles and amphibians. *Ecoscience*, **5**, 18–30.
- McPherson, J.M., Jetz, W. & Rogers, D.J. (2004) The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? *Journal of Applied Ecology*, **41**, 811–823.
- Manel, S., Williams, H.C. & Ormerod, S.J. (2001) Evaluating presence–absence models in ecology: the need to account for prevalence. *Journal of Applied Ecology*, **38**, 921–931.
- Midgley, G.F., Hughes, G.O., Thuiller, W. & Rebelo, A.G. (2006) Migration rate limitations on climate change-induced range shifts in Cape Proteaceae. *Diversity and Distributions*, **12**, 555–562.
- Miller, J. & Franklin, J. (2002) Modeling the distribution of four vegetation alliances using generalized linear models and classification trees with spatial dependence. *Ecological Modelling*, **157**, 226–247.
- Nathan, R. (2001) Dispersal biogeography. *Encyclopedia of Biodiversity* (ed. S.A. Levin), pp. 127–152. Academic Press, San Diego, CA, USA.
- Nix, H.A. (1986) A biogeographic analysis of Australian elapid snakes. *Atlas of Elapid Snakes of Australia* (ed. R. Longmore), pp. 4–15. Australian Government Publications Service, Canberra, Australia.
- Pearson, R.G. & Dawson, T.P. (2003) Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? *Global Ecology and Biogeography*, **12**, 361–371.
- Pearson, R.G., Dawson, T.P. & Liu, C. (2004) Modelling species distributions in Britain: a hierarchical integration of climate and land-cover data. *Ecography*, **27**, 285–298.
- Pearson, R.G., Dawson, T.P., Berry, P.M. & Harrison, P.A. (2002) SPECIES: a spatial evaluation of climate impact on the envelope of species. *Ecological Modelling*, **154**, 289–300.
- Peterson, A.T. (2006) Uses and requirements of ecological niche models and related distributional models. *Biodiversity Informatics*, **3**, 59–72.
- Peterson, A.T. & Robins, C.R. (2003) Using ecological-niche modeling to predict barred owl invasions with implications for spotted owl conservation. *Conservation Biology*, **17**, 1161–1165.
- Peterson, A.T. & Vieglais, D.A. (2001) Predicting species invasions using ecological niche modeling: new approaches from bioinformatics attack a pressing problem. *Bioscience*, **51**, 363–371.
- Ponder, W.F., Carter, G.A., Flemons, P. & Chapman, R.R. (2001) Evaluation of museum collection data for use in biodiversity assessment. *Conservation Biology*, **15**, 648–657.
- Pulliam, H.R. (1988) Sources, sinks, and population regulation. *American Naturalist*, **132**, 652–661.
- Pulliam, H.R. (2000) On the relationship between niche and distribution. *Ecology Letters*, **3**, 349–361.
- Reutter, B.A., Helfer, V., Hirzel, A.H. & Vogel, P. (2003) Modelling habitat-suitability using museum collections: an example with three sympatric *Apodemus* species from the Alps. *Journal of Biogeography*, **30**, 581–590.
- Rouget, M., Richardson, D.M., Nel, J.L., Le Maitre, D.C., Egoh, B. & Mgidi, T. (2004) Mapping the potential ranges of major plant invaders in South Africa, Lesotho and Swaziland using climatic suitability. *Diversity and Distributions*, **10**, 475–484.
- Rushton, S.P., Ormerod, S.J. & Kerby, G. (2004) New paradigms for modelling species' distributions? *Journal of Applied Ecology*, **41**, 193–200.
- Sanderson, R.A., Eyre, M.D. & Rushton, S.P. (2005) Distribution of selected macroinvertebrates in a mosaic of temporary and permanent freshwater ponds as explained by autologistic models. *Ecography*, **28**, 355–362.
- Scott, J.M., Heglund, P.J., Morrison, M.L., Haufler, J.B., Raphael, M.G., Wall, W.A. & Samson, F.B. (2002) *Predicting Species Occurrences: Issues of Scale and Accuracy*. Island Press, Covelo, CA.
- Segurado, P., Araújo, M.B. & Kunin, W.E. (2006) Consequences of spatial autocorrelation for niche-based models. *Journal of Applied Ecology*, **43**, 433–444.
- Shirihai, H. (1996) *The Birds of Israel*. Academic Press, London, UK.

- Shmida, A. & Wilson, M.V. (1985) Biological determinants of species diversity. *Journal of Biogeography*, **12**, 1–20.
- Smith, P.A. (1994) Autocorrelation in logistic regression modelling of species' distributions. *Global Ecology and Biogeography Letters*, **4**, 47–61.
- Snell, S.E., Gopal, S. & Kaufmann, R.K. (2000) Spatial interpolation of surface air temperatures using artificial neural networks: evaluating their use for downscaling GCMs. *Journal of Climate*, **13**, 886–895.
- Soberon, J. & Peterson, A.T. (2005) Interpretation of models of fundamental ecological niches and species' distribution areas. *Biodiversity Informatics*, **2**, 1–10.
- Sokal, R.R. & Oden, N.L. (1978) Spatial autocorrelation in biology. I. Methodology. *Biological Journal of the Linnean Society*, **10**, 199–228.
- Steinitz, O., Heller, J., Tsoar, A., Rotem, D. & Kadmon, R. (2005) Predicting regional patterns of similarity in species composition for conservation planning. *Conservation Biology*, **19**, 1978–1988.
- Steinitz, O., Heller, J., Tsoar, A., Rotem, D. & Kadmon, R. (2006) Environment, dispersal and patterns of species similarity. *Journal of Biogeography*, **33**, 1044–1054.
- Stockwell, D.R.B. & Peterson, A.T. (2002) Effects of sample size on accuracy of species distribution models. *Ecological Modelling*, **148**, 1–13.
- Thuiller, W., Araújo, M.B. & Lavorel, S. (2004) Do we need land-cover data to model species distributions in Europe? *Journal of Biogeography*, **31**, 353–361.
- Venier, L.A., McKenney, D.W., Wang, Y. & Mckee, J. (1999) Models of large-scale breeding-bird distribution as a function of macro-climate in Ontario, Canada. *Journal of Biogeography*, **26**, 315–328.
- Wiens, J.A., Rotenberry, J.T. & Vanhorne, B. (1987) Habitat occupancy patterns of North American shrub steppe birds: the effects of spatial scale. *Oikos*, **48**, 132–147.
- Willis, K.J. & Whittaker, R.J. (2002) Ecology – species diversity – scale matters. *Science*, **295**, 1245–1248.
- Zaniewski, A.E., Lehmann, A. & Overton, J.M.C. (2002) Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecological Modelling*, **157**, 261–280.

Received 1 January 2007; accepted 6 November 2007

Handling Editor: Marc Cadotte