Journal of Applied Ecology 2004
41, 263–274

An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data

ROBIN ENGLER, ANTOINE GUISAN and LUCA RECHSTEINER
Laboratoire de Biologie de la Conservation (LBC), Département d'Ecologie et d'Evolution, Université de Lausanne, BB, CH-1015 Lausanne, Switzerland

Summary

1. Few examples of habitat-modelling studies of rare and endangered species exist in the literature, although from a conservation perspective predicting their distribution would prove particularly useful. Paucity of data and lack of valid absences are the probable reasons for this shortcoming. Analytic solutions to accommodate the lack of absence include the ecological niche factor analysis (ENFA) and the use of generalized linear models (GLM) with simulated pseudo-absences.
2. In this study we tested a new approach to generating pseudo-absences, based on a preliminary ENFA habitat suitability (HS) map, for the endangered species *Eryngium alpinum*. This method of generating pseudo-absences was compared with two others: (i) use of a GLM with pseudo-absences generated totally at random, and (ii) use of an ENFA only.
3. The influence of two different spatial resolutions (i.e. grain) was also assessed for tackling the dilemma of quality (grain) vs. quantity (number of occurrences). Each combination of the three above-mentioned methods with the two grains generated a distinct HS map.
4. Four evaluation measures were used for comparing these HS maps: total deviance explained, best kappa, Gini coefficient and minimal predicted area (MPA). The last is a new evaluation criterion proposed in this study.
5. Results showed that (i) GLM models using ENFA-weighted pseudo-absence provide better results, except for the MPA value, and that (ii) quality (spatial resolution and locational accuracy) of the data appears to be more important than quantity (number of occurrences). Furthermore, the proposed MPA value is suggested as a useful measure of model evaluation when used to complement classical statistical measures.
6. *Synthesis and applications.* We suggest that the use of ENFA-weighted pseudo-absence is a possible way to enhance the quality of GLM-based potential distribution maps and that data quality (i.e. spatial resolution) prevails over quantity (i.e. number of data). Increased accuracy of potential distribution maps could help to define better suitable areas for species protection and reintroduction.

Key-words: ecological niche factor analysis (ENFA), *Eryngium alpinum*, generalized linear model (GLM), habitat suitability (HS) model, minimal predicted area (MPA), spatial resolution vs. data size.

Journal of Applied Ecology (2004) **41**, 263–274

Introduction

A variety of predictive models is currently in use to simulate the spatial distribution of plant and animal species (Franklin 1995; Guisan & Zimmermann 2000; Scott *et al*. 2002). Most of the models rely on adjusting a quantitative relationship between a taxon and its direct environment. Models have been developed for plant communities (Brzeziecki, Kienast & Wildi 1993; Brown 1994; Zimmermann & Kienast 1999), for individual plant species (Guisan, Theurillat & Kienast 1998; Guisan, Weiss & Weiss 1999; Peterson 2001; Bakkenes

Correspondence: Antoine Guisan, Laboratoire de Biologie de la Conservation (LBC), Département d'Ecologie et d'Evolution, Université de Lausanne, BB, CH-1015 Lausanne, Switzerland (e-mail antoine.guisan@ie-bsg.unil.ch).

et al. 2002) and for plant species assemblages and biodiversity reconstructed from superimposing individual species' predictions (Guisan & Theurillat 2000; Lehmann, Overton & Leathwick 2002).

These models result in spatial predictions indicating locations of the most suitable (and unsuitable) habitats for a target species, community or biodiversity (i.e. indicating 'hotspots'). Generalized linear and generalized additive models (GLM and GAM), implemented within a geographical information system (GIS), have become very popular for predicting such distributions (Guisan, Edwards & Hastie 2002).

However, as yet relatively few predictive models have been applied to rare and endangered species (Miller 1986; Myatt 1987; Carey & Brown 1995; Godown & Peterson 2000; Elith & Burgman 2002), despite their potential in conservation management, for instance in identifying sites with high potential for colonization. This may be because (i) data for rare and endangered taxa very often consist of a set of observed occurrences without sites of observed absences (hereafter called presence-only data); (ii) data for a single taxon are usually scarce (few observations); and (iii) often observations are not associated with any defined sampling unit (of known surface area) or they lack sufficient locational accuracy.

The first problem is commonly associated with data stored in large biological data bases. Such data have often been recorded by volunteers, usually without recourse to any predefined sampling strategy. Scarcity of data is specific to uncommon and rare species, for which prevalence in a data bank is, by definition, very low. Historical records, such as herbarium or museum collections, often lack precise details of location: at best they show proximity to a common site, a valley or village at a scale of a kilometre or more. These two problems make it more difficult to apply the usual statistical approaches. Such data contrast unfavourably with recent observations (≤ 10 years) sampled using a global positioning system (GPS) with a much higher spatial accuracy.

This highlights the dilemma of quantity (number of occurrences) vs. quality (locational accuracy). When the spatial accuracy associated with the geographical location of each observation site is known (e.g. the true site has a 95% probability of being within a 100-m radius), it becomes a major consideration in choosing the cell size (grain) of the study.

The choice of cell size may be determined by other criteria. A larger cell size might result in a more manageable data set or might be chosen if spatial autocorrelation is measured within the species' data and, as a result, observations that cannot be considered independent need to be aggregated. In contrast, a smaller cell size might better represent the ecological processes. Here, we will focus mainly on situations where spatial accuracy is known and can vary from one observation to the other.

Data of varying spatial accuracy can be manipulated to avoid propagating measurement errors in the model

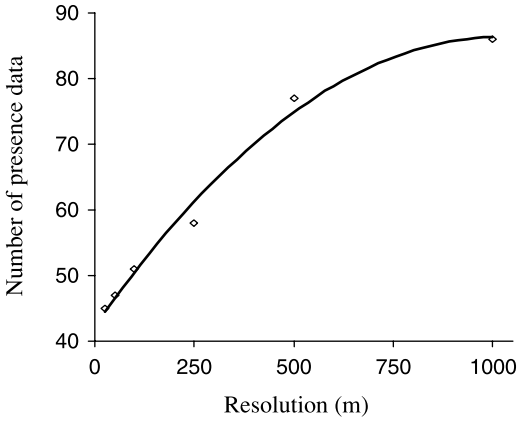

Fig. 1. Number of occurrences of *Eryngium alpinum* at each resolution.

(Elith, Burgman & Regan 2002) by either (i) aggregating all data in regular grid cells (or possibly other cell shapes) whose size still matches the poorest locational accuracy of observed occurrences, or (ii) dropping the most inaccurate data. A balance offering the best sample size vs. accuracy is usually found between these two options. This is illustrated in Fig. 1, which shows the decrease in the number of occurrences of the rare species *Eryngium alpinum* L. (Apiaceae) as the spatial resolution increases (i.e. decreasing cell size). This is due to the fact that fewer occurrences have a high locational accuracy associated with grid cells at high resolution (fine grain). Lowering the spatial resolution (coarser grain) allows less precise observations to be made, thus increasing their overall number. However, such a decrease is not straightforward because, when lowering the resolution (i.e. increasing cell size), distinct occurrences can also be aggregated in the same grid cell. Hence, the choice of method depends on the resolution of environmental layers available in the GIS, on the biology of the focus species and on the spatial distribution of its recorded occurrences.

Such data configuration results in severe limitations to the fitting of many statistical models, such as linear models (Guisan, Edwards & Hastie 2002). However, one alternative is to use models based on presence-only data. These are called profile techniques, as opposed to group discrimination approaches that need presence–absence or abundance data (Robertson, Caithness & Villet 2001). A well-known example of a profile-type model is the climatic envelope approach developed largely in the late 1980s by Australian scientists and implemented in the BIOCLIM package (Busby 1991; now ANUCLIM; Houlder *et al*. 2000). Another, more recent, example is the ecological niche factor analysis (ENFA) implemented in the BIOMAPPER package (Hirzel *et al*. 2002; Hirzel, Hausser & Perrin 2002). However, a common problem of profile methods is that they tend to generate over-optimistic predictions, i.e. they predict the species at too many locations. This is easily understood by the fact that, from a quantitative evaluation perspective, a

'perfect' model with such data would be a model predicting the species everywhere (i.e. '1' would be attributed to all cells in the area), as all observations would be correctly predicted as '1' and no discriminating absence would be available to restrict the predictions to zero where needed (i.e. at environmentally inappropriate locations).

In this regard, GLM constitute a better choice because they can deal with many types of predictors (continuous, binary, qualitative, ordinal), but on the other hand they must have presence and absence data. In order to use GLM when no absence data are available, one approach is to generate 'pseudo-absences' (Zaniewski, Lehmann & Overton 2002) and to use them in the model as absence data for the species. The manner in which pseudo-absences are generated is particularly important because it can have a significant influence on the final quality of the model (Zaniewski, Lehmann & Overton 2002).

The easiest way to choose pseudo-absences is simply to generate them totally at random over the study area (Hirzel, Helfer & Métral 2001; Zaniewski, Lehmann & Overton 2002). However, this method runs the risk of generating an absence in an area that is, in fact, favourable to the species. Indeed, when dealing with common species, choosing such a 'wrong absence' may not be too problematical because the numerous presence records will counteract its effect. However, when working with rare species, data are often scarce and choosing a wrong absence could significantly reduce the quality of a model.

To avoid, or at least reduce, this problem, more subtle methods can be employed to generate the pseudo-absences. For example, Zaniewski *et al*. (2002) first create a habitat suitability (HS) map of all fern species (a presence can be the occurrence of any species) using a GAM with totally random pseudo-absences. Then, a second set of pseudo-absences are randomly selected proportionally to the predictions by the first HS map and used to fit GAM models for every species. Selecting pseudo-absences proportionally to the overall sampling effort aims at avoiding sampling pseudo-absences in sites that were under-sampled in the field. However, multi-species data are not always available. In such situation, the first map – based on purely random pseudo-absences – is specific to the modelled species and pseudo-absences can be selected in areas below a certain threshold, in order to maximize the discriminating ability of the second model. The choice of this threshold must be defined as objectively as possible, for instance as the lowest value still encompassing 95% of observed species' occurrences.

In this study, we propose another way to generate pseudo-absences, which combines the respective strengths of ENFA and GLM. It is also a two-step approach, but uses ENFA instead of a GLM with totally random pseudo-absences to calculate the first HS map that is used to weight the selection of pseudo-absences. The calculation of this first model is particularly straightforward with ENFA (e.g. no need to select predictors).

The aims of this study were twofold. The first was to evaluate different methods for predicting rare species distribution, using ENFA with presence-only data, GLM with presence and random pseudo-absences, and a combination of both approaches. The second aim was to assess the dilemma between quality and quantity, trying more specifically to answer the question: is it preferable to have a large number of observations, which is better from a statistical point of view, or should one favour locational accuracy of observations (dropping all inaccurate ones, thus using a reduced set to calibrate the model) to ensure a better correspondence with environmental predictors used to predict the observations? This part of the study was conducted by building models at two different resolutions (25 and 500 m) having a different number of occurrences associated with each (Fig. 1). *Eryngium alpinum* (Apiaceae), a flagship threatened species in the European Alps, was chosen as an illustration. Finally, results from field investigations demonstrate the usefulness of such a model for suggesting new observation sites for rare and endangered species.

Methods

STUDY AREA

Switzerland covers an area of 41 293 km^2 and consists essentially of two mountain chains with a west–east orientation: the Jura (highest peak in Switzerland 1607 m a.s.l.) in the north and the Alps (highest peak in Switzerland 4634 m a.s.l.) in the south, separated by a lowland corridor, 50–100 km wide, generally referred to as the Swiss midlands and ranging from about 360 to 900 m a.s.l.

The whole country belongs to a single floristic unit, the medio-european or subatlantic domain (Ozenda 1982), which also corresponds to a single general climate of suboceanic type. A large proportion of the rain carried from the Atlantic Ocean and North Sea is stopped by the mountains and thus highest annual rainfall occurs in the Jura and in the northern part of the Alps. Mean annual temperature is also unequally distributed, with warm, wet areas in the south-eastern part of the country (Ticino, southern Alps), warm, dry areas in the south-eastern part (Valais), and a cooler and wetter climate characterizing the midlands and the Jura.

TEST SPECIES

Eryngium alpinum, commonly called Alpine Eryngo or Queen of the Alps, is a perennial hemi-cryptophyte about 30–100 cm high and easily recognizable by its cylindrical capitulum surrounded by a large involucre of huge blue-violet bracts. According to Landolt (1977), *E. alpinum* prefers habitat with moist, deep, alkaline soils rich in nutrients. It also needs full sunlight and therefore almost never occurs in forested areas.

Eryngium alpinum often grows on steep slopes of the alpine and subalpine belts, in a megaphorb-like vegetation. In Switzerland, it is found between 1400 and 2100 m, essentially in the pre-Alps (the northern foothills of the Alps) of the cantons of Vaud, Valais and Fribourg but some populations are also found in the cantons of Grisons, Uri and Unterwald (eastern part of Switzerland). Due to its ornamental features, this species is also often grown in gardens and cemeteries. Although the exact reasons for the recent decline of this species are still unknown, major threats might be picking, and changes in pasturing practices. The species is threatened in the whole of Switzerland and appears on the Swiss red lists of endangered plant species (Moser *et al*. 2002). The genetic structure and reproductive biology of *E. alpinum* are the focus of research to understand better the causes of its rarity (Gaudeul *et al*. 2002).

SPECIES DATA

The Swiss Floristic Network (CRSF) in Geneva provided a large part of the data used in this study. These data consist of observations of presences only (hereafter referred to as presence-only data) of *E. alpinum* at known geographical locations. These observations were aggregated within regular square units of two grids covering the country, resulting in 46 and 77 occurrences for the resolutions of 25 and 500 m, respectively (Fig. 1). No geographical location of observed absence was available at the start of the study.

ENVIRONMENTAL DATA

Hereafter, environmental variables used to predict species distribution are termed the predictors. On a mesoscale such as the whole of Switzerland, direct environmental predictors such as climate should prove more powerful than indirect predictors like altitude (Guisan & Hofer 2003). This is because different climates, not all of which might be suitable for the target species, can occur at the same altitude throughout Switzerland. Due to the need for normality in the ENFA method, only quantitative predictors were used. As a comparison, the same predictors were used to fit the GLM, although the latter statistical method can theoretically deal with all kind of predictors.

The following pool of quantitative environmental predictors was selected according to ecology and data availability for *E. alpinum* in the ArcGIS software (ESRI Inc., Redlands, CA, USA) in the case of GLM-based predictions and in the IDRISI software (Eastman 1997) in the case of ENFA-based predictions (Table 1).

The term pool is used here to indicate that not all of these environmental predictors were necessarily used to fit the different models. The original resolution of all environmental maps was 25 m. That is, over the map of Switzerland, the data layers were divided into pixels of 25×25 m. To predict the species distribution at the wider resolution of 500×500 m, an aggregation of the 25×25-m data was performed in ArcGIS by calculating the average value of the 400 25-m pixels enclosed within each 500×500-m pixel.

In addition, two qualitative environmental variables providing information, respectively, on main land-use classes (land-use) and on geology and soil types (substratum) associated with each pixel were used separately in a last step, to filter predictions made by the models.

STATISTICAL METHODS

The two statistical methods used within this study were the ENFA and GLM.

Ecological niche factor analysis (ENFA)

The ENFA (Hirzel *et al*. 2002) is a method based on a comparison between the environmental niche of the species and the environmental characteristics of the entire study area (stored as GIS layers), hereafter termed the background data. Hence, ENFA only needs a set of

Table 1. Descriptions and abbreviations of the quantitative environmental variables forming the initial pool of predictors. They were all originally prepared at a resolution of 25 m. The last two rows represent two qualitative variables. A few selected classes of these were used to filter the final predictions (see text)

Abbreviation	Description	Unit
slope	Average slope of each quadrate	%
srad3	Sum of March solar radiation	kJ day^{-1}
srad7	Sum of July solar radiation	kJ day^{-1}
tave7	Average July temperature	°C
rain48	Sum of rainfall from April to August	mm
rain49	Sum of rainfall from April to September	mm
ddeg300	Sum of daily average temperature above 3 °C	°C
topo500	Topographical position with a radius of 500 m	m
topo1000	Topographical position with a radius of 1 km	m
Landuse	Land-use classes (forests, agricultural areas, roads, buildings, rivers, open areas, screes, glaciers, etc.), rasterized from a 1 : 25 000 topographic vector map (Vector 25).	43 classes
Substratum	Geotechnical map providing information on geology and soil, rasterized from a 1 : 200 000 vector map (Geotech)	30 classes

presence data (no absences are required) and a set of background GIS predictors. In contrast to many other predictive methods that can be fitted outside the GIS (like GLM), ENFA thus requires a dynamic access to the ecogeographical predictors.

ENFA is similar to a kind of principal component analysis (PCA) in that it also transforms the original ecogeographical variables into new, uncorrelated, axes. However, whereas in PCA the successive axes are simply selected to match the direction of maximum variance in the multidimensional ecogeographical space, ENFA's principal components all possess a true ecological meaning for the modelled species.

The first component is called the marginality factor (MF). It passes through the centroid of all species' observations (multidimensional optimum) and the centroid of all background cells in the study area (mean environmental conditions). Hence, a high marginality value indicates that the species' requirements differ considerably from the average habitat conditions in the study area.

Several specialization factors (SF) are then successively extracted from the $n-1$ residual dimensions. Each SF is calculated in order to (i) ensure its orthogonality to the marginality factor and to the other SF, while at the same time (ii) maximize the ratio between the residual variance of the background data and the variance of the species' occurrences. A high specialization indicates that a species has a restricted ecological tolerance compared with the overall range of conditions that prevail in the study area.

Because most of the information is usually contained in a few first factors (usually the marginality factor and up to three or four specialization factors), only these are kept to compute the final HS map. All cells in the map obtain a HS value that is proportional to the distance between their position and the position of the species optimum in the new factorial space.

All ENFA analyses were performed within the BIOMAPPER software (version 2·1; Hirzel, Hausser & Perrin 2002). Correlations between all variables in the initial pool of predictors (Table 1) were calculated prior to ENFA analyses, in order to determine which variables should preferably be used in the ENFA. When two or more predictors had a correlation coefficient greater than 0·5, only the most proximal (in the sense of Austin 2002) was kept for the ENFA. As ENFA requires normally distributed data, all environmental layers were normalized through the 'box–cox' algorithm (Sokal & Rohlf 1981). Although several variables did not recover normality after the box–cox normalization, ENFA is not considered too sensitive to such violation (Hirzel *et al*. 2002).

GENERALIZED LINEAR MODELS (GLM)

GLM (McCullagh & Nelder 1989; for HS application of GLM see Guisan, Edwards & Hastie 2002) are an extension of the classical multiple regression, allowing non-normal response variables to be modelled. GLM were used in our case to model presence–absence of the species. As absences were not available in the original data set, pseudo-absences were generated in various ways (see below). All GLM were fitted within the R software (R 1·4·0; A Language for Data Analysis and Graphics ©2002), by specifying a binomial distribution and a logistic link function, as similarly done for other presence–absence data in ecological studies (Guisan, Weiss & Weiss 1999; Manel, Dias & Ormerod 1999; Guisan & Hofer 2003; also called logit regression).

Selection of predictors (and their possible transformations, e.g. polynomial terms) is certainly the most important and difficult step when fitting a GLM. As the number of combinations is too great to test all of them, we used a custom stepwise selection procedure programmed in R that offers the best-suited combination of predictors, even with large data sets. A first exploratory analysis of the different predictors based on univariate GLM (i.e. fitting a single predictor at a time but allowing polynomial terms to be considered) showed that all response curves were at most of a $y = x + x^2$ type, providing a basis for ignoring polynomial terms higher than quadratic. In a second step, GLM were fitted to all possible pairs of uncorrelated predictors (and their square term if significant) and the reduction of deviance was tested and recorded in each case. The pair of predictors causing the highest deviance reduction was kept. In the third step, deviance reduction was tested on each pair of predictors previously selected, adding the remaining predictors one at a time. Again, the trio expressing the greatest deviance was kept. Finally, step three was repeated until the addition of any predictor, and possibly its square term, was no longer significant. This procedure is close to forward stepwise selection, although a significant difference is that the pair of predictors (among all possible pairs) causing the highest deviance reduction is entered first in the model, and the same rule further applies to the selection of the following predictors in the formula.

The best combination of predictors was considered to be the one that expressed the greatest amount of deviance while having all terms in the equation below a significant level ($P \leq 0{\cdot}05$) of deviance reduction (chi-test in the case of binomial models).

THE MODELLING FRAMEWORK

Three methods were used to model species' distributions from presence-only data (i.e. without observed absences): (i) ENFA; (ii) GLM with randomly chosen pseudo-absences, hereafter called GLM-R; and (iii) GLM with ENFA-weighted pseudo-absences, hereafter called GLM-ENFA (Fig. 2). To assess the importance of quantity vs. quality, these models were all fitted with data prepared at the two resolutions of 25 m and 500 m.

The generation of pseudo-absences was done in two ways. (i) Totally at random: geographic coordinates were chosen at random over the Swiss territory. (ii) Weighted by ENFA predictions: in this case, coordinates were

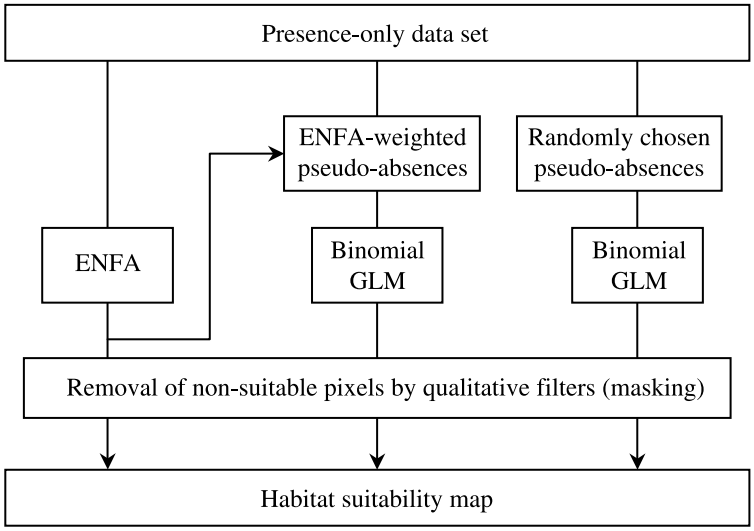

Fig. 2. The three methods used to produce habitat suitability maps: ENFA, GLM with totally random selected pseudo-absences, and the combined GLM-ENFA method where the pseudo-absences are weighted by the result of the ENFA.

also chosen at random, but only in areas where predictions by the prior ENFA model for *E. alpinum* were lower than 0·3. This threshold was chosen because it was slightly lower than the lowest ENFA prediction of 0·33 associated with observed presences.

In each case, the number of generated pseudo-absences was the same as the number of real presences, namely 45 for the 25-m resolution and 77 for the 500-m resolution (Fig. 1). In addition, pseudo-absences were never chosen in pixels having a value for land-use or substratum that was incompatible with the presence of *E. alpinum*, because these areas were removed later on during the filtering procedure. Moreover, as particular environmental factor values can sometimes be associated with these incompatible areas (e.g. glaciers usually have lower temperatures than surrounding areas) this might reduce the discriminating power of the model (i.e. the model will differentiate between suitable and very unsuitable locations but will not highlight finer variants in more or less suitable areas). Pseudo-absences were then combined with real presences into a single presence–absence data set ready to use in a binomial GLM.

Because chance plays a part in the choice of the pseudo-absences, each modelling procedure was repeated 1500 times (for both types of pseudo-absence and for both resolutions, 25 and 500 m). For each of these procedures ($4 \times 1500 = 6000$ in total) the custom stepwise selection was performed, ensuring that the best predictor combination was selected for each data set (as they all differed in their pseudo-absences).

Finally, qualitative environmental data (i.e. land-use and substratum layers) were used to filter the ENFA-based and GLM-based predictions in order to eliminate those areas where *E. alpinum* is unlikely to grow, e.g. urban areas, forests or glaciers. Such end-process filtering was performed by setting the model predictions to zero where unsuitable land-use and substratum classes occurred.

MODEL EVALUATION

The following evaluation measures were calculated for each of the 1500 GLM fitted on each presence/pseudo-absence data set, at each resolution.

Explained deviance (adjusted-D^2)

This is the percentage of deviance (i.e. variance in GLM) explained by the GLM. This measure expresses the fit of the model, weighted by the effective number of degrees of freedom (i.e. taking into account the number of predictors and the number of observations) used to build the model (Guisan, Weiss & Weiss 1999).

Best kappa (B-kappa)

The kappa coefficient (Cohen 1960; Fielding & Bell 1997) was calculated for all thresholds between zero and one by increments of 0·05. The greatest value was kept as the 'best kappa' value (Elith 2002). This measure expresses the best possible agreement not obtained randomly between two qualitative variables (of which a binary variable is a particular case).

Gini coefficient (equation 1)

This is a transformation of the area under the curve value (AUC), obtained by the receiver-operating characteristic plot method (ROC-plot; Fielding & Bell 1997), so that values have a wider range than the AUC (Hand & Henley 1997; Copas 1999):

$$\text{Gini coefficient} = 2 \times (\text{AUC} - 0{\cdot}5) \qquad \text{eqn 1}$$

The Gini coefficient as used here is an extension of its original use to describe income disparity. It will usually vary between zero (for an uninformative model) and one (for a model with perfect discrimination), but exceptionally it could be negative for cases where the model tends to make higher predictions at absence sites (i.e. the model is worse than chance; J. Elith, personal communication).

Minimal predicted area (MPA)

The minimal predicted area (MPA) is the minimal surface obtained by considering all pixels with predictions above a defined probability threshold (e.g. 0·7) that still encompasses 90% of the species' occurrences (see more explanation below).

To our knowledge, the MPA measure has never been used before and requires more explanation. Indeed, when evaluating HS maps using presence-only data, a map predicting presence of the species everywhere would show the best evaluation (because all presences would then effectively be predicted as presences), but such a map would be useless. The idea behind MPA is based on the assumption that a good HS map drawn from presence-only data should predict a potential area that is as small as possible (rule of parsimony) while still including a maximum number of the species' occurrences. For each model, predicted values were attributed to all occurrence sites and the value (e.g. 0·7) above which 90% of occurrences were observed was used as a threshold to reclassify the probability HS map into presence/absence (1/0). The positively predicted area (defined by the '1' only, when greater than 0·7) corresponds to the 'minimal predicted area'. All MPA values were calculated before the filtering operations because we wanted to assess performances of the statistical models only.

As no independent presence–absence data set was available, a common situation with rare species, best kappa and Gini coefficient were calculated on pseudo-independent presence/pseudo-absence data sets generated through a 'leave-one-out' jack-knife procedure (Jaberg & Guisan 2001). These two measures could not be calculated in the case of simple ENFA models, because no pseudo-absences were necessary for building these models. This is a recurrent problem encountered when evaluating model predictions with presence-only data. As a result, MPA was the only evaluation measure available for ENFA models.

The evaluation of GLM-based methods (i.e. GLM-ENFA and GLM-R) was performed on each of the 1500 generated data at both resolutions using the four evaluation measures, whereas the evaluation of the ENFA was performed on the single HS map calculated at each resolution. Hence, evaluation values of GLM-ENFA and GLM-R approaches are best provided in the form of box-plots while ENFA evaluations are represented by single values (bars) in the MPA graphs (Fig. 3).

COMPARING THE DIFFERENT EVALUATION MEASURES

We further tested whether those models (out of the 1500 runs for each of the four GLM experiments) that obtained the best evaluation for explained deviance, best kappa and Gini coefficient, were also those that obtained the best MPA evaluation. To test this, correlation coefficients (Spearman) were calculated between the ranks given to each model by the different evaluations for both GLM-ENFA and GLM-R at both resolutions.

USING HS MAPS TO SUPPORT FIELD INVESTIGATIONS

The final goal of our study was to predict and discover new populations of *E. alpinum*. For this purpose, we selected the GLM-ENFA model that obtained the best average value for explained deviance, the best kappa and Gini coefficient (i.e. a rank classification was made for each of these evaluation measurements and the model with the best average rank was chosen). This model was implemented in the ArcGIS software and the resulting probability map was filtered with the discriminating classes of qualitative predictors (filters) to remove areas where the presence of *E. alpinum* was highly improbable (i.e. with no occurrence ever recorded).

The predictions of the models that fell within the following types of land-use classes were set to zero: forests, urban and agricultural areas, lakes and rivers, glaciers, swamps and other areas transformed by humans (i.e. gravel pits, dams, etc.). The same operation was performed using the substratum map, and only seven classes of limestone and marly substratum were considered suitable for the species, all others being set to zero.

Results

Correlations between environmental predictors calculated at the 25-m resolution were very similar to those calculated at the 500-m resolution, so that the predictors retained by both ENFA analyses were the same. They were: slope, srad3, ddeg300, rain49 and topo500 (for their descriptions see Table 1).

The HS map was obtained by considering the first two components of the ENFA, which expressed, respectively, 92·8% and 88·1% of the variance at the resolutions of 25 m and 500 m. MPA values obtained for the two ENFA HS maps are given in Fig. 3d. For the two types of GLM models, box-plots in Fig. 3 show the distribution of (a) explained deviance (D^2), (b) best kappa value (B-kappa), (c) Gini coefficient (Gini) and (d) MPA.

A Wilcoxon rank test confirmed that, for all these evaluation values, the averages were significantly different ($P < 0{\cdot}01$) between the following pairs of models: GLM-ENFA 25 : GLM-R 25, GLM-ENFA 500 : GLM-R 500, GLM-ENFA 25 : GLM-ENFA 500 and GLM-R 25 : GLM-R 500. The number indicates

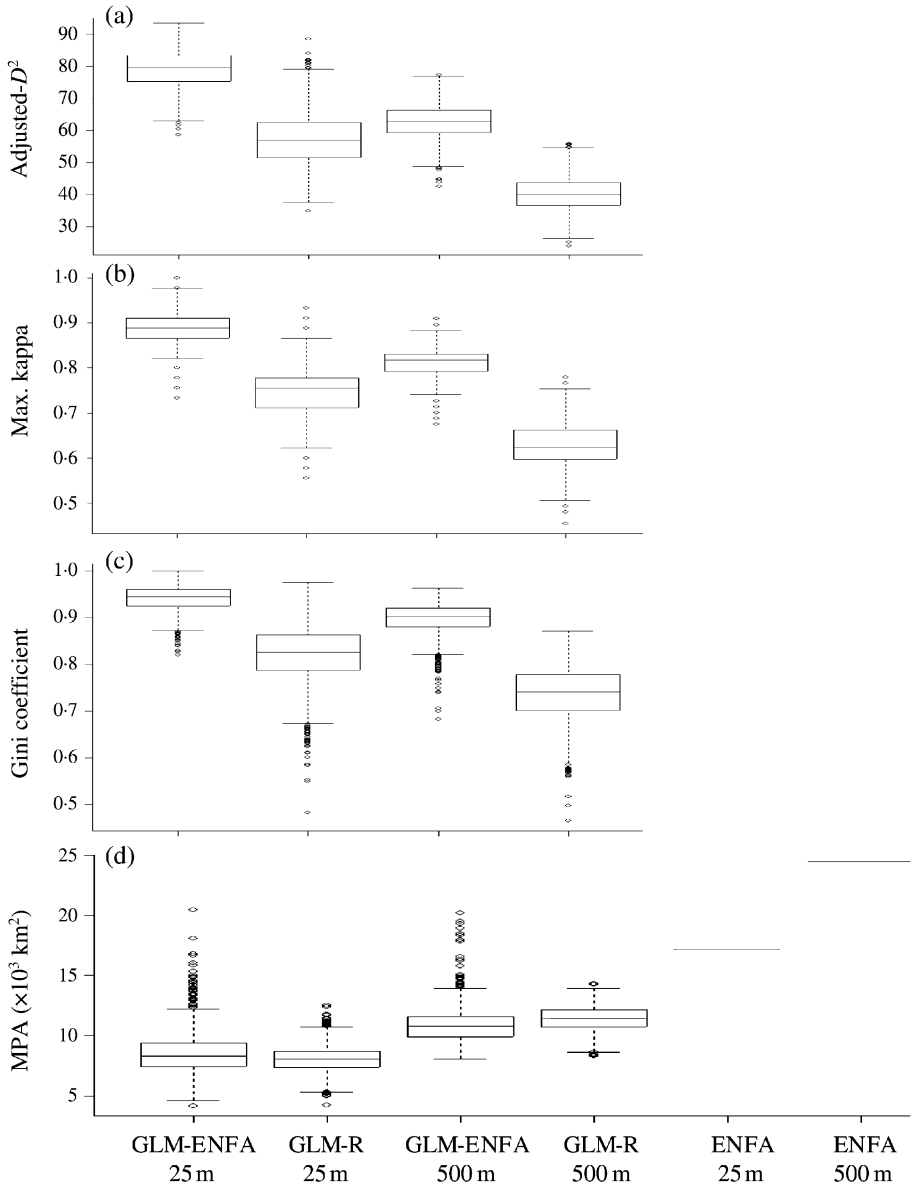

Fig. 3. Box-plots of the four evaluation measures calculated for each method of creating HS maps: (a) explained deviance in per cent (adjusted for the effective number of degrees of freedom); (b) maximum kappa value; (c) Gini coefficient; (d) MPA value (in 103 km^2). Values for ENFA are only shown for MPA as other values could not be calculated because of the lack of real absences. On graphs (a–c), the higher the value of evaluation, the better the model. On graph (d), the lower the MPA the better, but in conjunction with the highest possible value for the evaluation measures (a–c).

the spatial resolution considered. Correlations between the different measures of evaluation calculated for each fitted model at each resolution are given in Table 2.

Based on the cartographic implementation (potential map; Fig. 4) of the GLM-ENFA model at a resolution of 25 m, four new populations of the species were discovered in the field, all of them in pixels characterized by a high to very high habitat suitability (probability values of 0·98, 0·93, 0·92 and 0·79).

Discussion

The first goal was to compare two existing and one new approach to predicting rare species with presence-only (occurrence) data. Due to the lack of true absences, a formal comparison between ENFA and GLM-based methods (i.e. GLM-ENFA and GLM-R) is difficult. Indeed, in our study, MPA is the only evaluation measure available for comparison. Results show that the HS maps obtained with ENFA predict a MPA value that is approximately twice the mean of the GLM-based methods, both at the 25-m and 500-m resolutions (Fig. 3d). This result seems to confirm the tendency of ENFA to over-predict species distribution (Zaniewski, Lehmann & Overton 2002), due to the lack of discriminating absences, as discussed in the introduction. Another possible explanation of the apparent (but not proved) lack of accuracy of ENFA models could result from a violation of the assumption of normality of predictors that is required by the ENFA algorithm, as

Table 2. Spearman rank correlation coefficients between the different evaluation values calculated for each fitted model. The correlation values are averages of the correlations obtained for GLM-ENFA and GLM-R methods at both 25-m and 500-m resolutions ($n = 2 \times 2 \times 1500 = 6000$ models)

	Explained deviance	Best kappa value	Gini coefficient	MPA
Explained deviance		0·76	0·87	−0·01
Best kappa value			0·74	−0·21
Gini coefficient				−0·12
MPA				

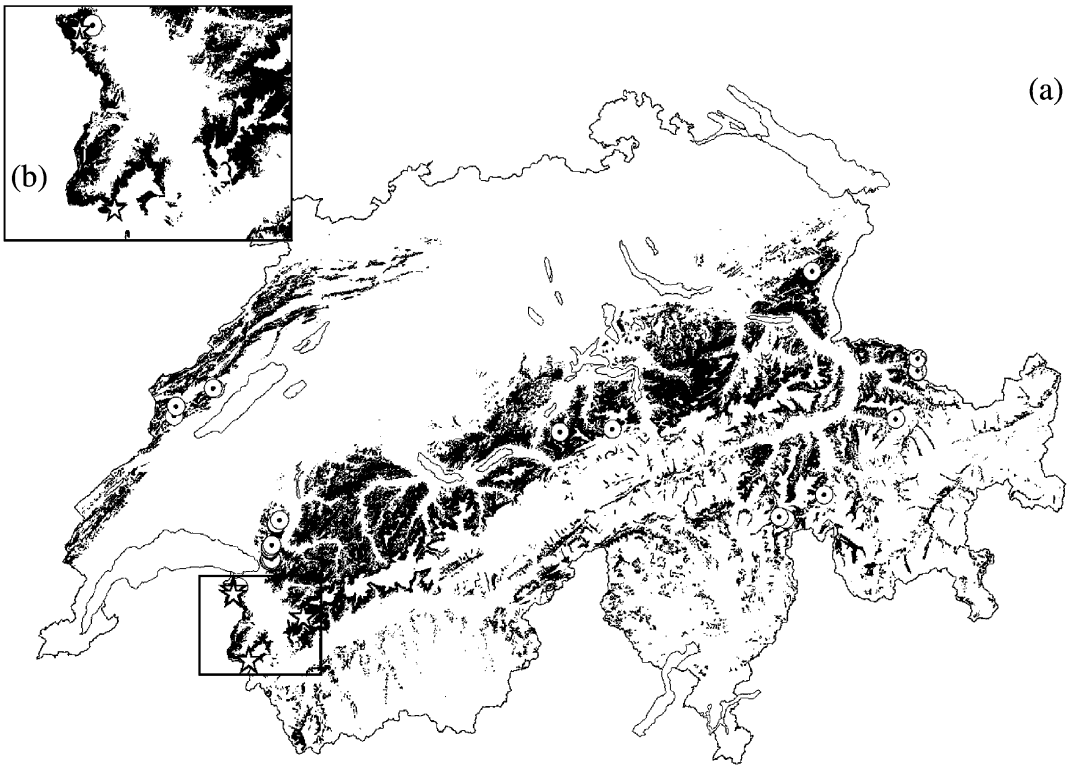

Fig. 4. (a) Potential distribution map for *Eryngium alpinum* in Switzerland drawn from one of the GLM-ENFA models at a resolution of 25 m. Black and dark grey tones indicate highly predicted areas, white circles indicate real presence points used to generate the map and white stars represent two new populations of the species discovered in the field. Highly predicted areas tend to be located in mountainous regions with higher rainfall (Jura and northern part of the Alps), which is consistent with the ecology of *E. alpinum*. (b) Magnified map showing the locations (star symbols) of the four new populations. Note that these are located within highly predicted areas (see text).

many of our predictors were actually not normally distributed (Kolmogorov–Smirnov goodness-of-fit test). Further testing would be needed to assess the robustness of ENFA (Hirzel, Helfer & Métral 2001, Hirzel *et al*. 2002) in such circumstances. This situation is likely to prevail in many other similar studies.

However, although we observed a large difference in MPA values between ENFA and GLM-based methods, it should not be concluded that the latter methods always prove better than the former. For instance, Hirzel, Helfer & Métral (2001), using a virtual species with known absences in a real landscape, have shown that ENFA can prove superior to GLM in the specific case of invading species (for their quantity of seed, expansion power and spread, as well as considering the virtual species' predefined niche), i.e. when species do not yet occupy all their potential habitats in the landscape. This might be less likely to be the case for many rare and threatened species, which tend to occupy most of their potential habitats, as these have usually been drastically reduced and, as a result, cover only a small proportion of the territory. However, although *E. alpinum* is truly a rare species (in the sense of Rabinovitz *et al.* 1986), it always yielded rather large predictions (compared with other species; O. Broennimann & A. Guisan, unpublished results), which might suggest either that its habitat is spatially not so restricted and that other reasons (like cutting) have limited its occurrence in the past, and/or that important environmental predictors are missing. None the less, the results suggest that the performance of these methods also depends on the type of organisms being modelled, on the type of

environmental predictors that are being used, and on the grain and extent considered.

Further comparisons were possible between GLM-ENFA and GLM-R because absence data were included in their evaluations. However, care should be taken when interpreting these results, as such evaluation measures are based on pseudo-absences and not on observed absences. This is a recurrent limitation of evaluating models based on occurrence data and a research area where progress is still required.

The three first evaluation measures (Fig. 3a–c) are consistent with each other, showing that GLM using ENFA-weighted pseudo-absences provide significantly better results on average (Wilcoxon rank test) and less deviance than those using randomly chosen pseudo-absences. This is true at both the 25-m and 500-m resolutions, which confirms that choosing pseudo-absences in an ENFA-weighted way rather than totally at random will enhance the accuracy of an HS map.

Interestingly, results from MPA measures were not consistent with the other evaluation measures. Indeed, we did not expect such consistency because the MPA concept is based on the parsimony criterion that 'the smaller the potential map the better the related model', which does not necessarily fit the mathematical criterion of statistical evaluations. At the higher resolution (25 m) GLM-R models provided a smaller MPA average value (remember that for MPA, the smaller the value the better the model) and a smaller deviance, whereas at the lower resolution (500 m) GLM-ENFA models obtained a similar average MPA value as GLM-R models, but showed a much larger deviance.

Furthermore, the comparison of the four evaluation measures, based on all individual model predictions (1500) in each GLM experiment (two types of GLM at two resolutions), did not show any correlation between MPA and any of the three quantitative measures (D^2, B-kappa or Gini). We do not believe that these results imply that MPA is a non-reliable value because the rule of parsimony used in MPA has a practical justification in conservation ecology, especially in the case of rare species that are, by definition, not widely distributed. Hence, further research is needed, at least in the case of these species, to assess whether the best HS map would not be the one that maximizes quantitative evaluation statistics while at the same time minimizing the predicted area. One identified limitation to the generalized use of MPA might be that it fails to evaluate appropriately those potential maps produced by certain modelling techniques, such as BIOCLIM (Busby 1991), as this type of model always fits the maximum possible predictions at observed presence sites (J. Elith, personal communication).

The second goal of this study was to determine whether it is preferable to use (i) less data with higher spatial accuracy or (ii) more data with lower accuracy. Comparing the evaluations of the 25-m and 500-m resolution HS maps reveals that averages for all these evaluation values are always better with the 25-m resolution. Overall, this conclusion is still valid when considering the three different types of models, GLM-ENFA, GLM-R and ENFA, although their deviances do not differ significantly.

The lower performances observed at the 500-m resolution could result from the combination of three factors. First, a loss of information is inevitable when aggregating environmental maps. Secondly, the low accuracy associated with some species' occurrences used at this resolution might still be underestimated and a greater measurement error (as defined by Elith, Burgman & Regan 2002) might actually prevail in the data. This is less likely to occur in the case of observations that have a higher spatial accuracy. Thirdly, plants being fixed organisms, they are highly influenced by the local microclimate. Therefore, relating species data that have a high geographical precision (of site location) with high-resolution environmental data should have a better predictive power because they reflect very local ecological conditions, while aggregated data reflect smoother environmental gradients in the area (e.g. mesoclimate). Furthermore, some important combinations of environmental predictors might not be appropriately expressed in such aggregated data.

In turn, such superiority of higher resolution predictors and less data might not be true for non-fixed organisms, as the required resolution for these is certainly dependent on a larger home range of target species, as suggested for bats by Jaberg & Guisan (2001). None the less, many of our potential maps have spatial predictions that cover a large proportion of the mountainous areas of Switzerland, even those with a good evaluation. This primarily reflects the fact that large areas are probably truly suitable sites for the target species *E. alpinum*, from the single perspective of those predictors that were used to build the model. Other factors, not included in the model but which might potentially have a more direct influence on plants (i.e. more proximal in the sense of Austin 2002), probably account for the unexplained spatial variation, and thus could enhance map precision. The problem remains, however, that data on many of these very important environmental factors, such as nutrient content in soils or precise physiologically meaningful microclimatic measurements, are still difficult to obtain in a spatially explicit way.

The best option for improving the HS maps would certainly be to obtain additional data for the target species, but this is difficult in the case of rare species. HS maps prove particularly useful in this regard, by suggesting new sampling sites for the species, such as those pixels of high prediction values that are not in the close vicinity of an observed population of the species. Visiting such suggested sites in the field at the optimum flowering time for the species may produce new presences or, at least, attested absences. This was done at the end of our study and four new populations of *E. alpinum* were found at sites of high predicted probability of presence. Such new data should then optimally be used to refine the models and generate new predictions

that will need to be verified in the field. Theoretically, reiterating such processes should lead to equilibrium, when new data (presences or absences) no longer contribute to improvement of the model (reaching a plateau).

Another solution for improving the accuracy of HS maps could be to refine the choice of pseudo-absences. In this study, the GLM-ENFA method was indeed used in a relatively simple manner. We used the ENFA predictions to divide the study area into two parts, one with probabilities of presence greater than 30% and the other with lower probabilities. Pseudo-absences were then randomly chosen in the latter category. A more subtle way of choosing the pseudo-absences could be, for instance, to stratify their distribution along a suitability gradient, like mean annual temperature. This could be a more precise method for choosing suitable areas.

However, an alternative exists to the process of selecting as many pseudo-absences as there are presences (usually a very limited number), and ideally repeating the process a number of times (e.g. 1500). This might be to sample, once only, a larger number of pseudo-absences (say 10 000) and to assign a weight, $w = k/10\,000$, in the GLM to each, so that the sum of the weight of all pseudo-absences adds up to give the number of presences k (i.e. ensuring a prevalence of 0·5) (M. Wisz, personal communication). This could prevent the inherent risk of inappropriately choosing a limited number of pseudo-absences (i.e. providing a low fit) when only one selection run is made.

Our third goal was to use these maps for suggesting new sampling sites for rare species. Although this study was not focused on this particular application of predictive models, a preliminary field campaign based on the selected HS map (Fig. 4) led to the discovery of four new populations of this highly endangered species. Indeed, this is a very promising result that strongly supports the use of predictive habitat distribution models for nature conservation planning.

Acknowledgements

To Jane Elith, Alan Fielding, Steve Rushton, Mary Wisz, Anthony Lehmann and an anonymous referee for their useful comments. To Simone Peverelli and Christophe Randin for their technical help with GIS. To Beat Bäumler at the Center of the Swiss Floristic Network (CRSF) in Geneva for providing most species data. We are also very thankful to Niklaus E. Zimmermann and Felix Kienast for providing the climatic data and Yamama Naciri-Graven, Olivier Broennimann, Mathias Vust and Charlotte Trippi for providing additional species data. Finally, to Julie Warrillow and Sophie Rickebush for kindly making the linguistic revision.

References

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

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

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.

Busby, J.R. (1991) BIOCLIM – a bioclimate analysis and prediction system. *Nature Conservation: Cost Effective Biological Surveys and Data Analysis* (eds C.R. Margules & M.P. Austin), pp. 64–68. CSIRO, Melbourne, Australia.

Carey, P.D. & Brown, N.J. (1995) The use of GIS to identify sites that will become suitable for a rare orchid, *Himantoglossum hircinum* L., in future changed climate. *Biodiversity Letters*, **2**, 117–123.

Cohen, J. (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **41**, 687–699.

Copas, J. (1999) The effectiveness of risk scores: the logit rank plot. *Journal of the Royal Statistical Society Series C-Applied Statistics*, **48**, 165–183.

Eastman, J.R. (1997) *Idrisi for Windows*. Clark University, Worcester, UK.

Elith, J. (2002) *Predicting the distribution of plants*. PhD Thesis. University of Melbourne, Melbourne, Australia.

Elith, J. & Burgman, M. (2002) Predictions and their validation: rare plants in the central highlands, Victoria, Australia. *Predicting Species Occurrences: Issues of Accuracy and Scale* (eds J.M. Scott, P.J. Heglund, J.B. Haufler, M. Morrison, M.G. Raphael, W.B. Wall & F. Samson), pp. 303–313. Island Press, Covelo, CA.

Elith, J., Burgman, M.A. & Regan, H.M. (2002) Mapping epistemic uncertainties and vague concepts in predictions of species distribution. *Ecological Modelling*, **157**, 313–330.

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.

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

Gaudeul, M., Naciri-Graven, Y., Gauthier, P. & Pompanon, F. (2002) Isolation and characterization of microsatellite markers in a perennial Apiaceae, *Eryngium alpinum* L. *Molecular Ecology Notes*, **2**, 107–109.

Godown, M.E. & Peterson, A.T. (2000) Preliminary distributional analysis of US endangered bird species. *Biodiversity and Conservation*, **9**, 1313–1322.

Guisan, A. & Hofer, U. (2003) Predicting reptile distributions at mesoscale: relation to climate and topography. *Journal of Biogeography*, **30**, 1233–1243.

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

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

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

Guisan, A., Theurillat, J.-P. & Kienast, F. (1998) Predicting the potential distribution of plant species in an alpine environment. *Journal of Vegetation Science*, **9**, 65–74.

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

Hand, D.J. & Henley, W.E. (1997) Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society A*, **160**, 523–541.

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

Hirzel, A., Hausser, J. & Perrin, N. (2002b) *Biomapper 2·1 Manual*. Laboratory for Conservation Biology, University of Lausanne, Lausanne, Switzerland [http://www.unil.ch/biomapper].

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

Houlder, D.J., Hutchinson, M.F., Nix, H.A. & McMahon, J.P. (2000) *ANUCLIM User Guide, Version 5·1*. Centre for Resource and Environmental Studies, Australian National University, Canberra, Australia [http://cres.anu.edu.au/outputs/anuclim.html].

Jaberg, C. & Guisan, A. (2001) Modelling the distribution of bats in relation to landscape structure in a temperate mountain environment. *Journal of Applied Ecology*, **38**, 1169–1181.

Landolt, E. (1977) *Ökologische Zeigerwerte zur Schweizer Flora*. Veröffentlichtes Geobotanisches Institute ETH Stiftung Rübel, Zürich, Switzerland.

Lehmann, A., Overton, J.McC. & Leathwick, J. (2002) GRASP: generalized regression analysis and spatial prediction. *Ecological Modelling*, **157**, 189–208.

McCullagh, P. & Nelder, J.A. (1989) *Generalized Linear Models*, 2nd edn. Chapman & Hall, London, UK.

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

Miller, R.I. (1986) Predicting rare plant distribution patterns in the southern Appalachians of the south-eastern U*SA*. *Journal of Biogeography*, **13**, 293–311.

Moser, D.M., Gygax, A., Bäumler, B., Wyler, N. & Palese, R. (2002) *Liste rouge des espèces menacées de Suisse. Fougères et plantes à fleurs*. Office Fédéral de l'Environnement, des Forêts et du Paysage (OFEFP), Bern, Switzerland.

Myatt, M.M. (1987) Predicting the habitat geography of sensitive plants and community types. *Conservation and Management of Rare and Endangered Plants* (ed. T.S. Elias), pp. 173–179. The California Native Plants Society, Sacramento, CA.

Ozenda, P. (1982) *Les Végétaux Dans la Biosphère*. Nathan, Paris, France.

Peterson, A.T. (2001) Predicting species' geographic distributions based on ecological niche modelling. *Condor*, **103**, 599–603.

Rabinowitz, D., Cairns, S. & Dillon, T. (1986) Seven forms of rarity and their frequency in the flora of the British Isles. *Conservation Biology* (ed. M. Soulé), pp. 182–204. Sinauer Associates, Sunderland, MA.

Robertson, M.P., Caithness, N. & Villet, M.H. (2001) A PCA-based modelling technique for predicting environmental suitability for organisms from presence records. *Diversity and Distribution*, **7**, 15–27.

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.

Sokal, R.R. & Rohlf, F.J. (1981) *Biometry*. Freeman, New York, NY.

Zaniewski, A.E., Lehmann, A. & Overton, J.McC. (2002) Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecological Modelling*, **157**, 261–280.

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

Received 24 April 2003; final copy received 2 November 2003