Diversity and Distributions, (Diversity Distrib.) (2009) **15**, 59–69

BIODIVERSITY RESEARCH

Evaluation of consensus methods in predictive species distribution modelling

Mathieu Marmion[1,2]*, Miia Parviainen[1,2], Miska Luoto[1,2], Risto K. Heikkinen[3] and Wilfried Thuiller[4]

[1]*Department of Geography,* [2]*Thule Institute, University of Oulu, PO Box 3000, FIN-90014 Oulu, Finland,* [3]*Finnish Environment Institute, Research Program for Biodiversity, PO Box 140, FIN-00251 Helsinki, Finland,* [4]*Laboratoire d'Ecologie Alpine, UMR CNRS 5553, Université Joseph Fourier, BP 53, 38041 Grenoble Cedex 9, France*

*Correspondence: Mathieu Marmion, Department of Geography, University of Oulu, PO Box 3000, FIN-90014 Oulu, Finland.
E-mail: mathieu.marmion@oulu.fi

ABSTRACT

Aim Spatial modelling techniques are increasingly used in species distribution modelling. However, the implemented techniques differ in their modelling performance, and some consensus methods are needed to reduce the uncertainty of predictions. In this study, we tested the predictive accuracies of five consensus methods, namely Weighted Average (WA), Mean(All), Median(All), Median(PCA), and Best, for 28 threatened plant species.

Location North-eastern Finland, Europe.

Methods The spatial distributions of the plant species were forecasted using eight state-of-the-art single-modelling techniques providing an ensemble of predictions. The probability values of occurrence were then combined using five consensus algorithms. The predictive accuracies of the single-model and consensus methods were assessed by computing the area under the curve (AUC) of the receiver-operating characteristic plot.

Results The mean AUC values varied between 0.697 (classification tree analysis) and 0.813 (random forest) for the single-models, and from 0.757 to 0.850 for the consensus methods. WA and Mean(All) consensus methods provided significantly more robust predictions than all the single-models and the other consensus methods.

Main conclusions Consensus methods based on average function algorithms may increase significantly the accuracy of species distribution forecasts, and thus they show considerable promise for different conservation biological and biogeographical applications.

Keywords
Distribution modelling, ensemble, machine learning methods, model selection, predictive accuracy, regression and classification methods.

INTRODUCTION

Predictive species distribution models have an important role in ecology and biogeography (Guisan & Zimmermann, 2000; Scott *et al.*, 2002; Guisan & Thuiller, 2005), and are increasingly used in a range of applications including regional biodiversity assessments, conservation biology, wildlife management and conservation planning (Elith *et al.*, 2006; Elith & Leathwick, 2007; Rodriguez *et al.*, 2007; Thuiller, 2007). The increase in applications of species distribution models is based on the growth in the availability of remotely sensed (RS) data and development of GIS techniques integrated with novel statistical methods (Guisan & Zimmermann, 2000). Carefully generated predictive models can effectively contribute to the insufficient field survey and museum data (Muñoz *et al.*, 2005; Guisan *et al.*, 2006; Rodriguez *et al.*, 2007), and occasionally even provide a more useful basis for biodiversity assessments than existing published range maps and national atlases (Bustamante & Seoane, 2004).

However, alongside the growth in the use of species distribution models, a number of studies have addressed the errors and uncertainties embedded in such models (Elith *et al.*, 2002; Barry & Elith, 2006; Heikkinen *et al.*, 2006; Hernandez *et al.*, 2006). The sources of uncertainty are diverse and range from measurement errors, small sample size, missing covariates and biased samples (Edwards *et al.*, 2006) to uncertainties in model building procedures. Recently much attention has been paid to investigation of the model-based uncertainty in species range prediction. This attention is of utmost importance because the performance of different modelling techniques has been shown to vary considerably in predicting both broad-scale biogeographical (Thuiller, 2004; Lawler *et al.*, 2006; Pearson *et al.*, 2006) and regional species distributions (Manel *et al.*, 1999). There are two main approaches to reduce the model-based uncertainty in species range simulations: (i) gathering understanding, via extensive model comparisons, concerning which of the methods will generally provide the best predictive performance and in what conditions

DOI: 10.1111/j.1472-4642.2008.00491.x

(Segurado & Araújo, 2004; Elith *et al.*, 2006; Prasad *et al.*, 2006), and (ii) the use of consensus methods or ensemble forecasting of species distributions (Thuiller, 2004; Thuiller *et al.*, 2005). In this paper we focus on consensus methods, which provide means to combine ensembles of species range forecasts and in this way overcome the problem of variability in predictions.

Consensus methods are based on combinative algorithms of the predictions provided by different single-models (Gregory *et al.*, 2001; Thuiller, 2003; Thuiller, 2004; Araújo & New, 2007). These techniques have earlier been employed in economics (Gregory *et al.*, 2001), biomedicine (Nilsson *et al.*, 2000; Nilsson *et al.*, 2002), meteorology (Sanders, 1963), climatology (Benestad, 2004) and hydrology (Goswami & O'Connor, 2007). They have recently been applied in broad-scale conservation studies, particularly to examine the impacts of climate change on various species (Thuiller, 2004; Araújo *et al.*, 2005b, 2006; Thuiller *et al.*, 2005). The consensus approach is based on the idea that different predictions are copies of possible states of the real distributions, and they form an ensemble. Already in the 19th century, Laplace (1820) emphasized the relevance of combinative algorithms in increasing the accuracy of an ensemble of predictions: 'In combining the results of these two methods, one can obtain a result whose probability law of error will be more rapidly decreasing'. In other words, a relevant combination of several unbiased (i.e. with a fair accuracy) model outputs will result in a more accurate prediction. Each sample taken into account contains some information that will be *de facto* transmitted to the resulted estimate. Some grid squares may be well classified by some methods and misclassified by others, even if all the methods have similar global accuracy. The term 'consensus' refers to a majority view or to an agreement of different model outputs (Gregory *et al.*, 2001; Thuiller, 2004). The matter resides in finding a relevant algorithm, the output of which follows a majority trend. However, although the consensus approach clearly has a number of attractive features, our understanding of its merits is still limited. There are different ways to build consensus predictions, and it has rarely if ever been tested which of the consensus methods provide the best predictive performance and whether these methods are able to consistently generate more accurate species range predictions than recent novel single-model methods available for species distribution modelling (cf. Elith *et al.*, 2006).

In this study, the predictive performances of five consensus methods were tested. As the basis for the consensus methods we used outputs from eight state-of-the-art modelling techniques, including generalized linear models (GLM), generalized additive models (GAM), multivariate adaptive regression splines (MARS), artificial neural networks (ANN), general boosting method (GBM), random forests (RF), classification tree analysis (CTA), and mixture discriminant analysis (MDA). These eight single-models provided the ensemble of predictions, which contained the eight separate predictive distributions generated for 28 threatened plant species in north-eastern Finland. The five consensus methods employed here form a representative sample of the most commonly used techniques. Two methods (Median(All) and Mean(All)) are based on global (i.e. output of all eight single-models) median and mean functions, whereas Weighted Average (WA), Best, and Median(PCA) methods preselect the single-models based on certain predefined criteria. In WA, half of the single-model outputs are preselected on the basis of the AUC values. The selected single-models are combined using an average function. Best proceeds via picking up the most accurate single-model, and Median(PCA) is based on the median of half of the single-models outputs, preselected by a principal component analysis. The main aims of this study were to investigate (i) which of the consensus algorithms can improve the accuracy of predictions from single-models, and how much, and (ii) the statistical differences in predictive ability of the eight single-model techniques.

METHODS

Study area

The study area (41 750 km^2) is located between 31°–32°45′ E and 65°–67°50′ N in north-eastern Finland. Phytogeographically, the area lies within the northern boreal zone (Ahti *et al.*, 1968), where pine- and spruce-dominated forests prevail. Numerous wetlands, lakes, and rivers characterize the landscape. The bedrock is calcium-rich, providing favourable conditions for species-rich plant communities. The climate is more continental than in most of northern Europe but with a humid element added (Atlas of Finland, 1987).

The species data consisted of presence records of 28 threatened vascular plant species with 10 or more records in 1677 grid squares with a resolution of 25 ha (Parviainen *et al.*, 2008). According to the IUCN classification (Gärdenfors *et al.*, 2001), 24 (86%) of these plant species were defined as vulnerable and four (14%) as endangered species. The flora in the study area is relatively well known because it has traditionally been a target for numerous studies of vascular plant species. Consequently, we assumed that the absence of a record in any of the 1677 grid squares corresponded to true absence of the species (Eyre *et al.*, 2004), given the quasi-exhaustive sampling strategy.

Calibration – evaluation

The data set of 1677 25-ha grid squares was first randomly split into two main subsets: the model calibration data set including 70% of the grid squares, and the evaluation data set containing the remaining 30% of the grid squares. The calibration data set was further divided randomly into two subsets, which were called 'inner-calibration' and 'inner-validation' data sets. In summary, we used four different data sets in the subsequent analysis: inner-calibration, inner-validation, calibration and evaluation, which contained 821, 352, 1173 and 504 grid squares, respectively (see Fig. 1a). These four data sets have different functions in the study design, which consists of two steps. First, the inner-calibration and inner-validation data sets were used in consensus preselective algorithms (see section *Consensus methods*). The inner-calibration data set was used to calibrate the single-models before implementing them in the inner-validation

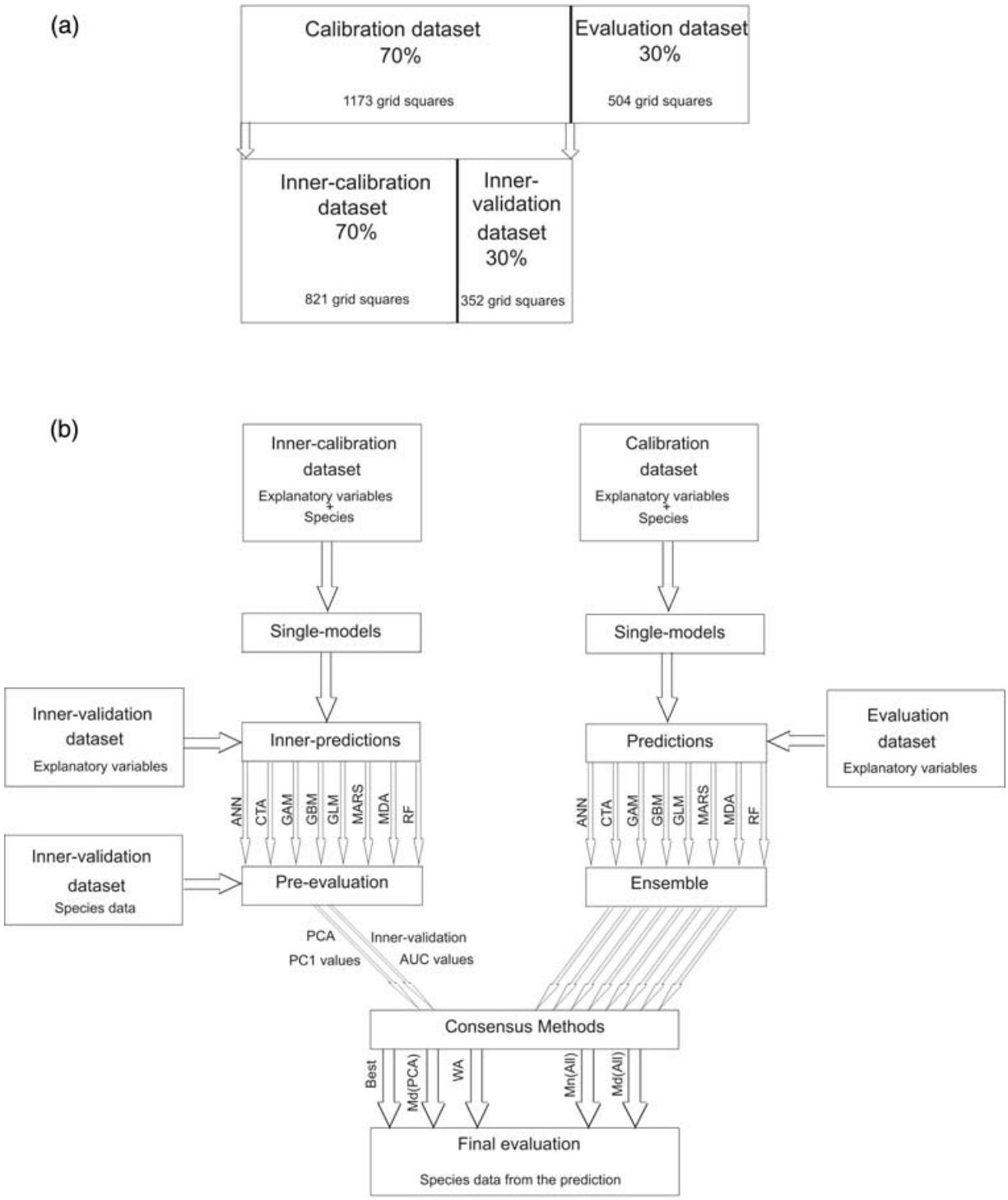

Figure 1 (a) Presentation of the four data sets and their relationships. (b) Schematic representation of the study design.

data set. Both data sets were used in pre-evaluating the predictive performances of the eight individual single-models (Fig. 1b). This pre-evaluation constituted an obligatory element in the consensus preselective algorithms. At the second step, the eight single-models and five consensus methods that were built using the calibration data set were then fitted in the evaluation data set. This procedure yielded the assessment of the predictive performance of both the single-models and the consensus methods. We used area under the curve (AUC) of the receiver-operating characteristic (ROC) plot as the means to evaluate the performance of the models (Fielding & Bell, 1997).

We acknowledge here that our evaluation data set does not represent a totally independent test set for assessing the predictive abilities of different models (cf. Araújo *et al.*, 2005a; Randin

et al., 2006; Heikkinen *et al.*, 2007). However, as the 25-ha grid cells in both our model calibration and evaluation data sets were distributed rather sparsely across the whole study area (grid cells used in modelling covered only *c.* 1% of the whole study area; see Fig. 3), we assume that the predictive capabilities of different models were assessed in our case rather well. For illustrative purposes, we projected the simulated distribution based on the model calibration data for selected species and selected models over the whole study area, which consisted of 166 968 25-ha grid cells (see Fig. 3).

Explanatory variables

The single-models were run with 16 explanatory variables: three climate, four topography, four geology and five land-cover variables were calculated for each grid square. The climate data were derived from the Finnish Meteorological Institute climate data sets (Venäläinen & Heikinheimo, 2002) and averaged for the time period 1961–90. These data were downscaled from the original 10-km grid to 0.5-km (25 ha) grid by using kriging interpolation (Parviainen *et al.*, 2008). Climate variables included growing degree days (> 5 °C), mean temperature of the coldest month (January; °C), and water balance (mm). The topography variables included mean elevation (m), mean topographical wetness index, mean radiation (kj/cm^2/a), and proportion of steep topography (> 15°). These variables were derived from the digital elevation model (DEM) at 25-m resolution using the ArcGIS and ArcView software (ESRI, 1991). The geology variables, related in the study as percentage covers for each study square, were sand/gravel soil, calcareous rock, quartzite rock and rock terrain. These variables were derived from digital maps of Quaternary deposit and pre-Quaternary rocks (Atlas of Finland) using ArcGIS software (ESRI, 1991). The land-cover variables selected and employed were the percentage covers of open mire, forested peatland, deciduous–mixed forest on mineral soil, rivers, and alpine area. We utilized European land-cover and land-use classification CORINE (Coordination of Information on the Environment) as land-cover information in our analysis (European Commission, 1994).

Single-models

We simulated the distribution of 28 threatened plant species using the BIOMOD tool (Thuiller, 2003), as implemented for R software. Eight techniques were used in modelling analyses: GLM, GAM, MARS (constituting the three regression methods), ANN, GBM, RF (the three machine learning methods), CTA, and MDA (the two classification methods). GLM, GAM, CTA and, ANN are described and discussed in the original BIOMOD paper (Thuiller, 2003). MARS represents a relatively new technique that utilizes classical linear regression (Friedman, 1991), and was recently tested in an extensive study comparing 16 predictive techniques (Elith *et al.*, 2006). MDA (Hastie & Tibshirani, 1996) and RF (Breiman, 2001; Cutler *et al.*, 2007) were also used as promising modelling methods. GBM is a machine learning method which was only recently introduced in ecology. GBM is highly efficient in fitting the data and combines the strengths of different modern statistical techniques (Ridgeway, 1999; Thuiller *et al.*, 2006). It was classified as one of the most predictive methods by Elith *et al.* (2006).

Consensus

The main aim of consensus methods is to decrease the predictive uncertainty of single-models by combining their predictions, as illustrated in Fig. 1b (Araújo *et al.*, 2005b). Some consensus methods contain a preselective algorithm. Such selective algorithms are based on various approaches such as PCA (Thuiller, 2004; Araújo *et al.*, 2005b) and statistical criteria (Johnson & Omland, 2004), or on basic mathematical functions such as averages and medians of ensembles of predictions (Gregory *et al.*, 2001; Araújo & New, 2007).

This study employs five of the most commonly used consensus methods (Gregory *et al.*, 2001; Johnson & Omland, 2004; Thuiller, 2004; Araújo *et al.*, 2005b; Araújo & New, 2007; Goswami & O'Connor, 2007): two non-selective approaches and three using a preselection of the modelling techniques. Eight single-models were first generated separately for each of 28 threatened vascular plant species. The combining of the outputs of the single-models then provided the ensemble of predictions, which contained eight forecasted distributions (probability values) for each species.

Median(all) consensus method is the median value of the outputs of all eight single-models. Median(all) has been used in an ecological context by Araújo *et al.* (2005b). Gregory *et al.* (2001) considered this method to belong to the same class as computing the mean value of the whole predictions ensemble (*Mean(all)*). In our understanding, Median(all) is less frequently used than Mean(all). Examples of studies using Mean(all) can be found in McNees (1987) and Araújo & New (2007).

The *WA* consensus method utilizes pre-evaluation of the predictive performance of the single-models. In this approach, half (i.e. four) of the eight single-models with highest accuracy are selected first, and then a WA is calculated based on the pre-evaluated AUC of the single-models as described by Eqn 1

$$WA_i = \frac{\sum_j (AUC_{mj_i} \times mj_i)}{\sum_j AUC_{mj_i}} \quad (1)$$

where mj_i are the probability-of-occurrence values of the *i*th threatened plant species in a given grid cell for the *j*-selected single-models for which pre-evaluation AUC values were the highest. There are different algorithms to assign weights to single-models (see Hartley *et al.*, 2006). Goswami & O'Connor (2007) used weighted averages in hydrology, and in their study, the weights of the single-models were computed via a least square procedure. Araújo & New (2007) also presented this method with a Bayesian approach.

Median(PCA) approach was based on calculating the median value of only four single-models out of eight for each threatened plant species. Half (i.e. four) of the eight single-models were selected via a PCA. A PCA provides for each single-model a rate reflecting its ability to follow the general trend of projections of

the eight single-models (see Thuiller, 2004). The first principal component (PC1) reflects the general trend followed by the eight single-models, for each threatened plant species. The selection algorithm of this consensus method was based on this approach. The four models for which the variance of predictions along PC1 was the greatest were selected. Then the median value of the output of these four single-models was computed. The models that did not follow the general trend were then rejected and not taken into account. This method was used by Thuiller (2004), Araújo *et al*. (2005b), and Thuiller *et al*. (2005) in a biogeographical context.

The fifth consensus method, Best, proceeds via picking up the best of the eight separate single-models for each plant species. Here, the best model was selected based on the highest pre-evaluated AUC value. Overall, various approaches for selecting the best model exist and the approach of picking up the best single-model based on some predefined criterion is commonly used in biogeography (Thuiller, 2003; Segurado & Araújo, 2004; Elith *et al*., 2006). Johnson & Omland (2004) presented various selective criteria to choose the best single-models, such as the Akaike information criterion (AIC), the Bayesian information criteria, and the likelihood ratio tests (LRT).

RESULTS

The predictive accuracies of the eight single-models and five consensus methods are summarized in Table 1 and Figs 1 and 2 (see also Table S1 in Supporting Information). The mean AUC values of the eight single-models ranged from 0.697 (CTA) to 0.813 (RF) and from 0.757 (Median(PCA)) to 0.850 (WA) for the consensus methods. A total of six species were classified excellently with RF, whereas four were classified poorly. The five consensus methods showed varying levels of predictive accuracy. WA and Mean(All) consensus methods provided significantly more robust predictions than all the single-models and the other consensus methods (Wilcoxon signed rank test; P-value < 0.05; Table 2). The most superior consensus method was WA, with a mean AUC value of 0.850 and classifying nine plant species excellently. WA showed higher predictive performance for 21 plant species models out of 28 than did the single-models.

Table 1 Distribution of the 28 threatened vascular plant species in different accuracy classes of area under the curve (AUC) with respect to the eight modelling techniques and the five consensus methods. Rating of the model accuracy: excellent = AUC > 0.9, fair = 0.7 < AUC < 0.9 and poor = AUC < 0.7 (Swets 1988).

	ANN	CTA	GAM	GBM	GLM	MARS	MDA	RF	Mn(All)	WA	Best	Md(All)	Md(PCA)
Excellent	7	3	1	3	0	1	3	6	9	9	6	4	5
Fair	16	10	19	17	21	18	12	18	17	16	17	21	16
Poor	5	15	8	8	7	9	13	4	2	3	5	3	7

ANN, artificial neural networks; CTA, classification tree analysis; GAM, generalized additive models; GBM, general boosting method; GLM, generalized linear models; MARS, multivariate adaptive regression splines; MDA, mixture discriminant analysis; RF, random forests; Mn(All), Mean(All); WA, Weighted Average; Md(All), Median(All); Md(PCA), Median(PCA).

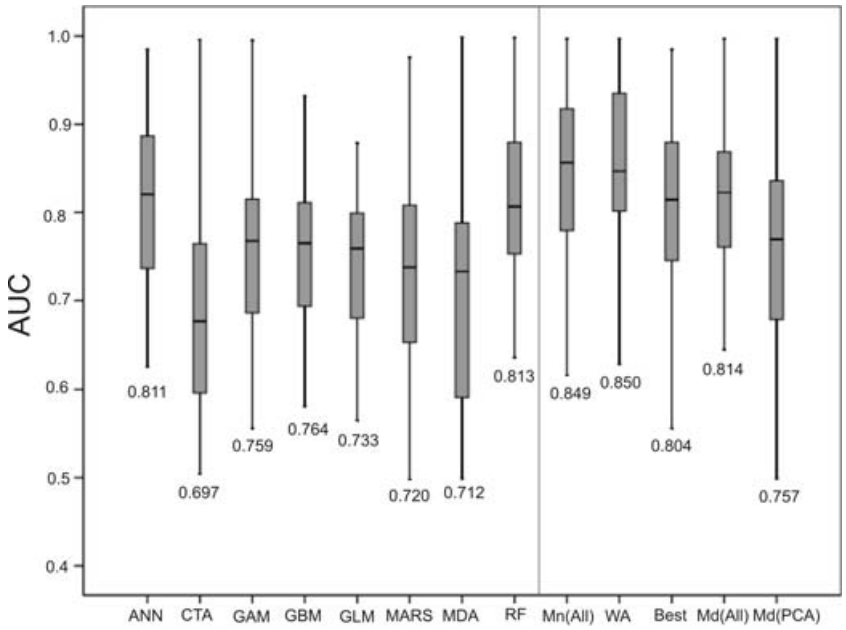

Figure 2 A box-whisker plot illustrating the predictive accuracies of the eight single models (left part of the Figure) and five consensus methods (right). The boxes show median and 1st and 3rd quartile values. The mean area under the curve (AUC) value out of the 28 threatened plant species is indicated under the boxes for each single model and consensus method.

Table 2 Statistical differences in the predictive performance of eight different single-models and five consensus methods for 28 threatened plant species occurrences. Statistical tests of the differences among the predictive accuracies of different methods were tested by Wilcoxon signed rank test. In the upper part of the table *P*-values are presented, whereas comparisons of the predictive performance of the single-models and consensus methods are presented in the lower part of the table. Ranks: (positive;negative;tied). Positive or negative ranks refer to models located in the left column (for example: RF-ANN: (12;15;1) means AUCRF > AUCANN 12 times, AUCRF < AUCANN 15 times, AUCRF = AUCANN once).

	ANN	RF	GBM	GAM	GLM	MARS	MDA	CTA	Mn(All)	WA	Best	Md(All)	Md(PCA)
ANN	–	0.773	0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	0.001	0.001	0.501	0.909	0.113
RF	(12;15;1)	–	< 0.001	0.003	0.007	< 0.001	< 0.001	< 0.001	0.007	0.004	0.872	0.882	0.199
GBM	(5;23;0)	(6;22;0)	–	0.577	0.136	0.017	0.020	< 0.001	< 0.001	< 0.001	0.001	< 0.001	0.387
GAM	(5;23;0)	(6;22;0)	(13;15;0)	–	0.079	0.014	0.013	< 0.001	< 0.001	< 0.001	0.004	< 0.001	0.487
GLM	(3;25;0)	(9;19;0)	(9;19;0)	(11;17;0)	–	0.473	0.439	0.067	< 0.001	< 0.001	0.002	< 0.001	0.161
MARS	(4;24;0)	(4;24;0)	(7;21;0)	(6;22;0)	(12;16;0)	–	0.733	0.400	< 0.001	< 0.001	0.001	< 0.001	0.210
MDA	(3;25;0)	(4;23;1)	(7;19;2)	(8;18;2)	(12;16;0)	(14;14;0)	–	0.531	< 0.001	< 0.001	< 0.001	< 0.001	0.021
CTA	(1;27;0)	(2;26;0)	(5;23;0)	(6;22;0)	(8;20;0)	(11;16;1)	(12;16;0)	–	< 0.001	< 0.001	< 0.001	< 0.001	0.015
Mn(All)	(24;4;0)	(20;8;0)	(26;2;0)	(27;1;0)	(26;2;0)	(27;1;0)	(27;1;0)	(28;0;0)	–	0.594	0.001	0.003	< 0.001
WA	(22;6;0)	(21;7;0)	(25;3;0)	(27;1;0)	(26;2;0)	(27;1;0)	(27;1;0)	(28;0;0)	(15;11;2)	–	< 0.001	0.007	< 0.001
Best	(5;11;12)	(9;10;9)	(21;4;3)	(22;5;1)	(22;5;1)	(22;4;2)	(22;5;1)	(26;2;0)	(5;23;0)	(3;25;0)	–	0.829	0.280
Md(All)	(11;15;2)	(14;14;0)	(23;5;0)	(25;3;0)	(25;3;0)	(23;5;0)	(24;3;1)	(26;2;0)	(4;22;2)	(4;22;2)	(13;15;0)	–	0.046
Md(PCA)	(11;17;0)	(11;16;1)	(18;10;0)	(18;10;0)	(19;9;0)	(17;11;0)	(20;8;0)	(23;5;0)	(1;24;3)	(3;23;2)	(12;15;1)	(9;16;3)	–

ANN, artificial neural networks; RF, random forests; GBM, general boosting method; GAM, generalized additive models; GLM, generalized linear models; MARS, multivariate adaptive regression splines; MDA, mixture discriminant analysis; CTA, classification tree analysis; Mn(All), Mean(All); WA, Weighted Average; Md(All), Median(All); Md(PCA), Median(PCA).

Figure 3 illustrates the predictions based on WA compared to single-models based on MARS and RF for two threatened plant species, *Cypripedium calceolus* and *Dactylorhiza incarnata* ssp. *cruenta*.

Table S3 in Supporting Information shows the summary information of the explanatory variables selected in the single-models, i.e. how many times a given variable was selected in the distribution models developed for the 28 species by each of the eight modelling techniques. CTA led to the simplest single-models, with a mean of 4.4 explanatory variables, whereas ANN, MDA, and RF produced the most complex single-models including all 16 variables. Among the explanatory variables, temperature of the coldest month (mean 24.1 °C) and growing degree days (mean 23.3 °C) were the most often selected variables in all single-models. Additionally, the cover of open mire (mean 22.3%), water balance (mean 22.1 mm), and elevation (mean 21.3 m) were often significantly related to the distribution patterns of the studied 28 vascular plant species. These variables appear to reflect the main biophysical gradients with a recognized, physiological influence on plant species in the taiga forest landscapes.

DISCUSSION

Predictive species distribution modelling can provide a valuable and cost-effective tool for conservation planning and biodiversity management, especially in poorly surveyed regions that are under accelerating pressure of habitat loss and degradation (Austin, 2002; Bustamante & Seoane, 2004; Araújo & Guisan, 2006; Rodriguez *et al.*, 2007). However, in order to generate as useful and accurate models as possible, researchers should have a thorough understanding of the limitations and uncertainties embedded in species distribution modelling (e.g. Elith *et al.*, 2002; Loiselle *et al.*, 2003; Barry & Elith, 2006; Gibson *et al.*, 2007). Loiselle *et al.* (2003) and Wilson *et al.* (2005) argued that in present-day conservation planning, too little attention has been paid to how robust the different species distribution models are, or to the discrepancies among the outputs from different models. Our results echoed these concerns with respect to the regional conservation planning of threatened vascular plants in northern landscapes, which due to the limited resources may often need to be based on simulated distribution maps.

In agreement with earlier studies on climate change impacts on species distributions (Thuiller, 2004; Araújo *et al.*, 2005b; Pearson *et al.*, 2006), model comparison studies based on present-day distributions of species have reported important differences in the spatial predictions from different models (Loiselle *et al.*, 2003; Elith *et al.*, 2006; Heikkinen *et al.*, 2007). The first strategy to reduce the inconsistencies between different species distribution models is to conduct thorough model comparison evaluations and adopt the most promising techniques for modelling (Elith *et al.*, 2006; Lawler *et al.*, 2006; Prasad *et al.*, 2006). A general outcome of the model comparisons has been that novel modelling techniques, such as RF and GBM, consistently outperform more established techniques (Cutler *et al.*, 2007). Our results also provide support for arguments of the excellent performance of RF, at least in such cases where the projections of the model are generated under similar ecological–geographical conditions and space that were used when calibrating the model.

However, the second option, i.e. the use of consensus methods (Laplace, 1820; Thuiller, 2004; Araújo & New, 2007), provides

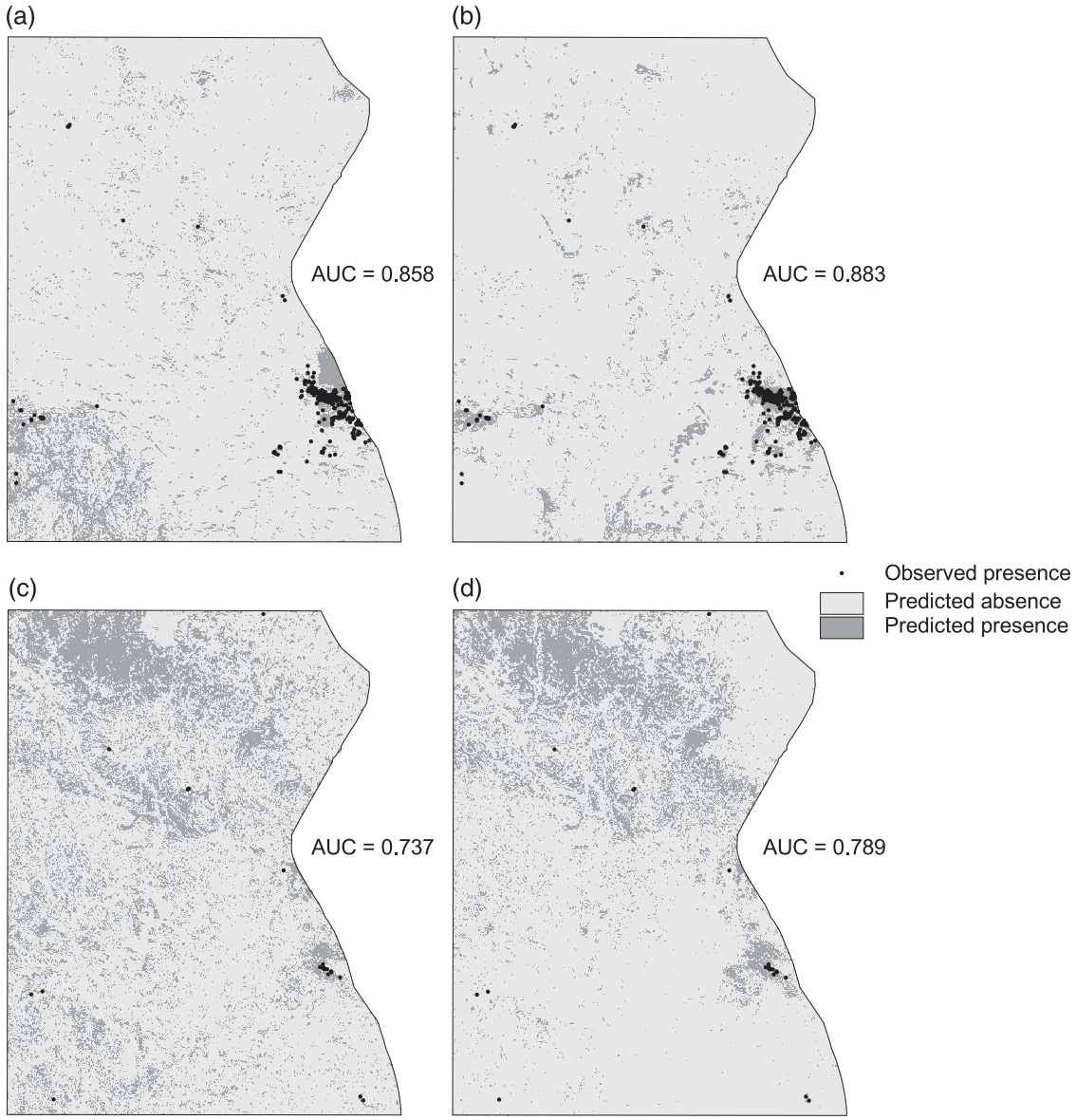

Figure 3 Predicted distribution of *Cypripedium calceolus* provided by multivariate adaptive regression splines (MARS) (a) and the weighted average (WA) consensus method (b), and of *Dactylorhiza incarnata* ssp. *cruenta* provided by random forest (RF) (c) and the WA consensus method (d). The black dots emphasize the observations of these threatened species in the study area. The area under the curve (AUC) values reflect the accuracy of the models based on the evaluation data set.

two attractive features that are lacking from single-models. First, it is not certain that the model with the highest accuracy with the species data at hand will provide the most realistic simulations of the species distribution in a new area or under future climate conditions (Thuiller, 2004; Araújo *et al.*, 2005a). Such observations have increased interest in the careful selection of the choice of most appropriate models to carry out interpolation or extrapolation modelling. Some single-models such as RF are accurate in interpolation modelling (Cutler *et al.*, 2007). However, models based on continuous curves (e.g. GLM and GAM) may be more reliable than those based on classification trees (Thuiller *et al.*, 2004) in order to transfer the calibrated models into new areas or time periods via extrapolation modelling (where inputs may be outside the range within which the models were built). Curves can still be defined out of the range of calibration, even if the interval of confidence becomes large. The consensus method based on combinative algorithms circumvents this problem by summarizing agreements among projections generated by different models. Second, the consensus method may be intuitively a more realistic approach than the search for a superior single-model technique because many model comparison studies have been unable to report a superior performance for any of the techniques. Instead, the model with most accurate predictions often varies from species to species (Thuiller, 2003).

In this study, we performed an extensive evaluation of the predictive ability of five consensus methods and eight single-model

techniques. In the field of biogeography and ecology, the only comparable previous study that focused on predictive modelling is that of Araújo *et al.* (2005b), which was based on four single-models and two consensus methods. The other studies in the field dealing with consensus methods (e.g. Thuiller, 2004; Araújo & New, 2007) have not addressed the evaluation of predictive performance of the methods.

The explanatory variables included in this study were chosen so that they covered a broad spectrum of the possible ecological determinants of the distributions of the modelled vascular plant species. The variables that were most often selected in the single-models were reasonable with respect to the *a priori* understanding of the ecological characteristics of the study species (for a more detailed discussion see Parviainen *et al.*, 2008). Certain climate variables, particularly growing degree-days, temperature of the coldest month, and water balance, appeared as significant determinants of the 28 studied threatened plant species across all eight modelling techniques. Such variables represent primary environmental regimes related to the physiological tolerances of organisms (Austin *et al.*, 1990; Box *et al.*, 1993), and factors with a significant effect on plant growth and reproduction, as well as over-wintering survival (Skov & Svenning, 2004). A number of land-cover variables were also often selected in the different models. Links between a certain species distribution and a given land-cover variable were often based on clear biological inferences. For example, open mire contributed significantly to the distribution models for the studied species, which was only logical as many of our study species rarely occur outside peat soils (see Parviainen *et al.*, 2008).

However, although certain environmental variables were important across all the modelling techniques, there were nevertheless differences in which variables were selected in the models. Such differences in final model variables may be a function of underlying design and model form (Edwards *et al.*, 2006), because different methods combine responses to individual predictors in different ways (Elith *et al.*, 2005). All the single-models selected the environmental variables that best discriminate the suitability of habitats to the presence of a certain species, except for ANN and MDA that do not have a selective algorithm. Thus ANN and MDA may lead to exceedingly complex models, and thus to overfitting problems (Guisan & Thuiller, 2005). The consensus algorithms also inherently include many explanatory variables because they include all the explanatory variables selected by the combined single-models.

In our study, the most efficient consensus methods were the WA consensus method based on AUC values and Mean(All). They improved significantly the predictive accuracy of all single-models. The high potential of these methods has been advocated in a number of studies (Gregory *et al.*, 2001; Johnson & Omland, 2004; Araújo & New, 2007; Goswami & O'Connor, 2007), and our results support these arguments. The good performance of consensus methods based on average function may be explained by the low-pass filtering ability of the average function. In signal processing, a low-pass filter can be obtained by calculating the neighbour average (e.g. spatial average in image processing, temporal average in signal processing; (Araújo *et al.*, 2005b; Russ, 2006)). In cases in which several methods (less than half of the whole number of single-models, i.e. four in this study) overestimate the distribution of some species compared to other methods, isolated predicted occurrences may be removed. In other words, there is a 'cleaning' effect, which was also visible in our results (Fig. 3). The southern part of the predicted distribution of both plant species based on WA is much 'cleaner' and less sparse than the prediction based on a single-model in the same area. The use of WA prevented the appearance of isolated predicted occurrences which apparently resulted from the overestimated predictions of single-models. However, if most of the single-models underestimate the distribution of a given species, the cleaning effect may be too strong and affect the accuracy of the species distribution forecasts. Average methods will thus work best when single-models that over- or underestimate the spatial distributions of a given species are in the minority among the used panel of models. To fulfil this condition, the use of several single-models based on various algorithms is recommended.

The consensus method based on Median(PCA) has been applied in species–climate impact studies by Thuiller (2004) and Araújo *et al.* (2005b) to select the most consensual single-models. The study of Araújo *et al.* (2005b) also provided a test of transferring single-models and two consensus methods from the models calibrated with climate and bird data from the UK at one point in time to evaluation data collected at another point in time. The outcome of this test was that the Median(PCA) method was able to increase the accuracy of the model projections compared to four single-models considered. In contrast to these results, Median(PCA) did not provide more accurate predictions than half of the single-models in our study. This discrepancy between our results and those of the earlier studies may be caused by the differences in the study settings. In our study, we investigated the transferability of the models at one point in time and between ecologically similar grid squares in one region, whereas the study by Araújo *et al.* (2005b) included a temporal transferring of the models, which is often a more demanding test for the models (Araújo *et al.*, 2005a). However, more research is clearly required to identify the study settings in which the different consensus methods are likely to perform best in different conditions or times.

The two remaining consensus methods used in this study, Best and Median(All), did not improve significantly the predictive accuracies of the two most accurate single-models, RF and ANN. According to Araújo & New (2007), ANN and RF incorporate the notion of ensemble forecasting. This may well explain the good predictive performances of these two modelling techniques compared to the other single-models, and their similarity in accuracies with Best and Median(All). However, inclusion of Best among the consensus methods should be made with caution. Thuiller (2004), Araújo *et al.* (2005b), and Gregory *et al.* (2001) considered a consensus method to be a method based on an algorithm detecting a consensual trend of predictions. In this study, Best did not contain such an algorithm, because the used model selection criterion (AUC) did not reflect the consensual trend of predictions. However, when Best is based on a PCA,

such as choosing the single-model for which the PC1 were the highest, the method belongs clearly to the group of consensus methods. However, an apparent deficiency in the consensus models applied here is that it is rather difficult to conclude with certainty the number of single-models that should be preselected. The predictive accuracy of Best suggests that the selection of a single-model is not efficient. WA preselected half of the single-models, whereas Mean(All) included all the eight modelling techniques. In addition to the presented results of this study, we evaluated the predictive performance of Mean(Best 4) and Median(Best 4), which compute the mean and median values of the output of the four single-models preselected via the same preselective algorithm of WA. However, there were no significant differences in the predictive AUC values of Mean(Best 4) and Median(Best 4) and their 'all-models-count-partners', Mean(All) and Median(All). We also tested yet another alternative to WA, a consensus model into which different single-models were preselected if their inner validation AUC value was higher than 95% of the AUC of the best single-model. This approach is, to some extent, analogous to the method described by Burnham & Anderson (2002). However, the predictive AUC value of the 'highest-95%-AUC' WA method was 0.840, without significant differences between Mean(All) and WA. Thus, we can conclude that preselecting only a part of the most accurate single-models for consensus methods significantly improved the predictive accuracy of the consensus methods in our study material.

CONCLUSIONS

In this study, we provide support for the argument that significant improvements in the accuracy of species distribution predictions can be achieved by applying consensus methods, especially those based on the average function. Our findings may have important consequences for regional conservation and management planning studies in which biased or insufficient field data should be complemented as accurately as possible using modelling of species distributions. However, our results also showed that consensus methods do not necessarily always improve the predictive accuracy of the single-models. In sum, the accuracy of the consensus methods is always dependent on the accuracy of the single-models on which they are based (Araújo *et al.*, 2005b; Araújo & New, 2007). Therefore, equally important in developing reliable forecasts of species distributions is also to pay attention to different critical underlying issues in single-models (e.g. Barry & Elith, 2006; Heikkinen *et al.*, 2006), and to aim at generating better models with improved data.

ACKNOWLEDGEMENTS

Different parts of this study were funded by the Academy of Finland (project grant 116544) and the EC FP6 Integrated Project ALARM (GOCE-CT-2003-506675; Settele *et al.*, 2005). WT received support from European Commission's FP6 MACIS (Minimization of and Adaptation to Climate change Impacts on biodiversity no. 044399) and ECOCHANGE (Challenges in assessing and forecasting biodiversity and ecosystem changes in Europe, no. 066866 GOCE) projects. A study of this nature would not have been possible without the hundreds of volunteers who contributed their data to the threatened plant species data base. Michael Bailey helped with correction of the English text. The comments by T.C. Edwards and three anonymous referees helped greatly in improving this paper.

REFERENCES

Ahti, T., Hämet-Ahti, L. & Jalas, J. (1968) Vegetation zones and their sections in northwestern Europe. *Annales Botanici Fennici*, **5**, 169–211.

Araújo, M.B. & Guisan, A. (2006) Five (or so) challenges for species distribution modelling. *Journal of Biogeography*, **33**, 1677–1688.

Araújo, M.B. & New, M. (2007) Ensemble forecasting of species distributions. *Trends in Ecology and Evolution*, **22**, 42–47.

Araújo, M.B., Pearson, R.G., Thuiller, W. & Erhard, M. (2005a) Validation of species-climate impact models under climate change. *Global Change Biology*, **11**, 1504–1513.

Araújo, M.B., Whittaker, R.J., Ladle, R.J. & Erhard, M. (2005b) Reducing uncertainty in projections of extinction risk from climate change. *Global Ecology and Biogeography*, **14**, 529–538.

Araújo, M.B., Thuiller, W. & Pearson, R.G. (2006) Climate warming and the decline of amphibians and reptiles in Europe. *Journal of Biogeography*, **33**, 1712–1728.

Atlas of Finland (1987) *Climatology, folio 131*. National Board of Survey and Geographical Society of Finland, Helsinki, Finland.

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

Austin, M.P., Nicholls, A.O. & Margules, C.R. (1990) Measurement of the realized qualitative niche: environmental niches of five Eucalyptus species. *Ecological Monographs*, **60**, 161–177.

Barry, S. & Elith, J. (2006) Error and uncertainty and habitat models. *Journal of Applied Ecology*, **43**, 413–423.

Benestad, R.E. (2004) Tentative probabilistic temperature scenarios for northern Europe. *Tellus*, **56A**, 89–101.

Box, E.O., Crumpacker, D.W. & Hardin, E.D. (1993) A climatic model for location of plant species in Florida, U.S.A. *Journal of Biogeography*, **20**, 629–644.

Breiman, L. (2001) Random forests. *Machine Learning*, **45**, 5–32.

Burnham, K.P. & Anderson, D.R. (2002) *Model selection and multimodel inference: a practical information-theoretic approach*. Springer-Verlag, New York.

Bustamante, J. & Seoane, J. (2004) Predicting the distribution of four species of raptors (Aves: Accipitridae) in southern Spain: statistical models work better than existing maps. *Journal of Biogeography*, **31**, 295–306.

Cutler, D.R., Edwards, T.C. Jr, Beard, K.H., Cutler, A., Hess, K.T., Gibson, J. & Lawler, J.J. (2007) Random forests for classification in ecology. *Ecology*, **88**, 2783–2792.

Edwards, T.C. Jr, Cutler, D.R., Zimmermann, N.E., Geiser, L. & Moisen, G.G. (2006) Effect of sample survey design on the accuracy of classification tree models in species distribution models. *Ecological Modelling*, **199**, 132–141.

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

Elith, J., Ferrier, S., Huettmann, F. & Leathwick, J.R. (2005) The evaluation strip: a new and robust method for plotting predicted responses from species distribution models. *Ecological Modelling*, **186**, 280–289.

Elith, J., Graham, C.H., Anderson, R.P., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L.G., Loiselle, B.A., Manion, G., Moritz, G., Nakamura, M., Nakazawa, Y., McC Overton, J., Peterson, A.T., Phillips, S.J., Richardson, K., Scachetti-Pereira, R. Schapire, R.E., Soberon, J., Williams, S., Wisz, M.S. & Zimmermann, N.E. (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**, 129–151.

Elith, J. & Leathwick, J. (2007) Predicting species distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. *Diversity and Distributions*, **13**, 265–275.

Environment Systems Research Institute (ESRI) (1991) *ARC/INFO user's guide. Cell-based modelling with GRID. Analysis, display and management*. ESRI, Inc., Redlands, California.

European Commission (1994) *EUR 12585 – CORINE land cover – technical guide*. Office for Official Publications of the European Communities, Luxemburg.

Eyre, M., Rushton, S., Luff, M. & Telfer, M. (2004) Predicting the distribution of ground beetle species (Coleoptera, Carabidea) in Britain using land cover variables. *Journal of Environmental Management*, **72**, 163–174.

Fielding, A. & Bell, J. (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, **24**, 38–49.

Friedman, J. (1991) Multivariate adaptive regression splines. *Annals of Statistics*, **19**, 1–141.

Gärdenfors, U., Hilton-Taylor, C., Mace, G.M. & Rodríguez, J.P. (2001) The application of IUCN Red List criteria at regional levels. *Conservation Biology*, **15**, 1206–1212.

Gibson, L., Barrett, B. & Burbidge, A. (2007) Dealing with uncertain absences in habitat modelling: a case study of a rare ground-dwelling parrot. *Diversity and Distributions*, **13**, 704–713.

Goswami, M. & O'Connor, K.M. (2007) Real-time flow forecasting in the absence of quantitative precipitation forecasts: a multi-model approach. *Journal of Hydrology*, **334**, 125–140.

Gregory, A.W., Smith, G.W. & Yetman, J. (2001) Testing for forecast consensus. *Journal of Business and Economic Statistics*, **19**, 34–43.

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., Broennimann, O., Engler, R., Yoccoz, N.G., Vust, M., Zimmermann, N.E. & Lehman, A. (2006) Using niche-based models to improve the sampling of rare species. *Conservation Biology*, **20**, 501–511.

Hartley, S., Harris, R. & Lester, P.J. (2006) Quantifying uncertainty in the potential distribution of an invasive species: climate and the Argentine ant. *Ecology Letters*, **9**, 1068–1079.

Hastie, T. & Tibshirani, R. (1996) S Archive: (mda). StatLib. http://lib.stat.cmu.edu/S/.

Heikkinen, R.K., Luoto, M., Araújo, M.B., Virkkala, R., Thuiller, W. & Sykes, M.T. (2006) Methods and uncertainties in bioclimatic envelope modelling under climate change. *Progress in Physical Geography*, **30**, 751–777.

Heikkinen, R.K., Luoto, M., Virkkala, R., Pearson, R.G. & Körber, J.H. (2007) Biotic interactions improve prediction of boreal bird distributions at macro-scales. *Global Ecology and Biogeography*, **16**, 754–763.

Hernandez, P.A., Graham, C., Master, L.L. & Albert, D.L. (2006) The effect of sample size and species characteristics on performance of different species distribution modelling methods. *Ecography*, **29**, 773–785.

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

Laplace, P.S. (1820) *Théorie analytique des probabilités*. Courcier, Paris.

Lawler, J.J., White, D., Neilson, R.P. & Blaustein, A.R. (2006) Predicting climate-induced range shifts: model differences and model reliability. *Global Change Biology*, **12**, 1–17.

Loiselle, B.A., Howell, C.A., Graham, C.H., Goerck, J.M., Brooks, T., Smith, K.G. & Williams, P.H. (2003) Avoiding pitfalls of using species distribution models in conservation planning. *Conservation Biology*, **17**, 1591–1600.

Manel, S., Dias, J.M., Buckton, S.T. & Ormerod, S.J. (1999) Alternative methods for predicting species distribution: an illustration with Himalayan river birds. *Journal of Applied Ecology*, **36**, 734–747.

McNees, S.K. (1987) Consensus forecasts: tyranny of the majority? *New England Economic Review* **Nov/Dec**, 15–21.

Muñoz, A.R., Real, R., Borbosa, A.M. & Vargas, J.M. (2005) Modelling the distribution of Bonelli's eagle in Spain: implications for conservation planning. *Diversity and Distributions*, **11**, 477–486.

Nilsson, J., Persson, B. & von Heijne, G. (2000) Consensus prediction of membrane protein topology. *FEBS Letters*, **486**, 267–269.

Nilsson, J., Persson, B. & Von Heijne, G. (2002) Prediction of partial membrane protein topologies using a consensus approach. *Protein Science*, **11**, 2974–2980.

Parviainen, M., Luoto, M., Ryttäti, T. & Heikkinen, R.K. (2008) Modelling the occurrence of threatened plant species in taiga landscapes: methodological and ecological perspectives. *Journal of Biogeography*, doi: 10.1111/j.1365-2699.2008.01922.x.

Pearson, R.G., Thuiller, W., Araujo, M.B., Martinez-Meyer, E., Brotons, L., McClean, C., Miles, L., Segurado, P., Dawson, T.P. & Lees, D.C. (2006) Model-based uncertainty in species range prediction. *Journal of Biogeography*, **33**, 1704–1711.

Prasad, A.M., Iverson, L.R. & Liaw, A. (2006) Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems*, **9**, 181–199.

Randin, C.F., Dirnböck, T., Dullinger, S., Zimmerman, N.E., Zappa, M. & Guisan, A. (2006) Are species distribution models transferable in space? *Journal of Biogeography*, **33**, 1689–1703.

Ridgeway, G. (1999) The state of boosting. *Computing Sciences and Statistics*, **31**, 172–181.

Rodriguez, J.P., Brotons, L., Bustamante, J. & Seoane, J. (2007) The application of predictive modelling of species distribution to biodiversity conservation. *Diversity and Distributions*, **13**, 243–251.

Russ, J.C. (2006) *The image processing handbook*. CRC Press, Boca Raton, Florida.

Sanders, F. (1963) On subjective probability forecasting. *Journal of Applied Meteorology*, **2**, 191–201.

Scott, J.M., Heglund, P.J., Samson, F., Haufler, J., Morrison, M., Raphael, M. & Wall, B. (2002) *Predicting species occurrences. Issues of accuracy and scale*. Island Press, Washington.

Segurado, P. & Araújo, M. (2004) An evaluation of methods for modelling species distributions. *Journal of Biogeography*, **31**, 1555–1569.

Settele, J., Hammen, V., Hulme, P., Karlson, U., Klotz, S., Kotarac, M., Kunin, W., Marion, G., O'Connor, M., Petanidou, T., Peterson, K., Potts, S., Pritchard, H., Pysek, P., Rounsevell, M., Spangenberg, J., Stefan-Dewenter, I., Sykes, M., Vighi, M., Zobel, M. & Kühn, I. (2005) ALARM – Assessing LArge-scale environmental Risks for biodiversity with tested Methods. *GAIA*, **14**, 69–72.

Skov, F. & Svenning, J.-C. (2004) Potential impact of climatic change on the distribution of forest herbs in Europe. *Ecography*, **27**, 366–380.

Swets, K. (1988) Measuring the accuracy of diagnostic systems. *Science*, **240**, 1285–1293.

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

Thuiller, W. (2004) Patterns and uncertainties of species' range shifts under climate change. *Global Change Biology*, **10**, 2020–2027.

Thuiller, W. (2007) Biodiversity – climate change and the ecologist. *Nature*, **448**, 550–552.

Thuiller, W., Araújo, M.B. & Lavorel, S. (2004) Generalized models vs. classification tree analysis: predicting spatial distributions of plant species at different scales. *Journal of Vegetation Science*, **14**, 669–680.

Thuiller, W., Lavorel, S., Araújo, M.B., Sykes, M.T. & Prentice, I.C. (2005) Climate change threats to plant diversity in Europe. *Proceedings of the National Academy of Sciences*, **102**, 8245–8250.

Thuiller, W.F., Midgley, G., Rougeti, M. & Cowling, R. (2006) Predicting patterns of plant species richness in megadiverse South Africa. *Ecography*, **29**, 733–744.

Venäläinen, A. & Heikinheimo, M. (2002) Meteorological data for agricultural applications. *Physics and Chemistry of the Earth*, **27**, 1045–1050.

Wilson, K.A., Westphal, M.I., Possingham, H.P. & Elith, J. (2005) Sensitivity of conservation planning to different approaches to using predicted species distribution data. *Biological Conservation*, **122**, 99–112.

Editor: Dr Mark Robertson

SUPPORTING INFORMATION

The following Supporting Information is available for this article:

Table S1 Variance projection of each single-model on the first component obtained by conducting a principal component analysis (PCA).

Table S2 Inner-validation AUC values for each single-model for the 28 threatened plant species.

Table S3 Frequency of the explanatory variables selected into the eight single-models.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.