Diversity and Distributions, (Diversity Distrib.) (2009) **15**, 565–576

BIODIVERSITY RESEARCH

Using species distribution models to predict new occurrences for rare plants

John N. Williams[1]*, Changwan Seo[2], James Thorne[3], Julie K. Nelson[4], Susan Erwin[4], Joshua M. O'Brien[5] and Mark W. Schwartz[1]

[1]*Department of Environmental Science & Policy, University of California, Davis, CA 95616, USA,* [2]*University of Seoul, 90 Jeonnong-dong, Dongdaemun-gu, Seoul 130-743, Korea,* [3]*Information Center for the Environment, University of California, Davis, CA 95616, USA,* [4]*United States Forest Service, Shasta-Trinity National Forest, 3644 Avtech Parkway, Redding, CA 96002, USA,* [5]*Department of Veterinary Medicine and Epidemiology, Center for Animal Disease Modeling and Surveillance, University of California, Davis, CA 95616, USA*

*Correspondence: John N. Williams, Department of Environmental Science & Policy, University of California, Davis, CA 95616, USA.
E-mail: jnwill@ucdavis.edu

ABSTRACT

Aim To evaluate a suite of species distribution models for their utility as predictors of suitable habitat and as tools for new population discovery of six rare plant species that have both narrow geographical ranges and specialized habitat requirements.

Location The Rattlesnake Creek Terrane (RCT) of the Shasta-Trinity National Forest in the northern California Coast Range of the United States.

Methods We used occurrence records from 25 years of US Forest Service botanical surveys, environmental and remotely sensed climate data to model the distributions of the target species across the RCT. The models included generalized linear models (GLM), artificial neural networks (ANN), random forests (RF) and maximum entropy (ME). From the results we generated predictive maps that were used to identify areas of high probability occurrence. We made field visits to the top-ranked sites to search for new populations of the target species.

Results Random forests gave the best results according to area under the curve and Kappa statistics, although ME was in close agreement. While GLM and ANN also gave good results, they were less restrictive and more varied than RF and ME. Cross-model correlations were the highest for species with the most records and declined with record numbers. Model assessment using a separate dataset confirmed that RF provided the best predictions of appropriate habitat. Use of RF output to prioritize search areas resulted in the discovery of 16 new populations of the target species.

Main conclusions Species distribution models, such as RF and ME, which use presence data and information about the background matrix where species do not occur, may be an effective tool for new population discovery of rare plant species, but there does appear to be a lower threshold in the number of occurrences required to build a good model.

Keywords
Edaphic specialist, endemism, Maxent, random forests, rarity, serpentine.

INTRODUCTION

Species distribution models (SDMs) have emerged as an effective tool in spatial ecology, conservation and land management (Raxworthy *et al.*, 2003; Rushton *et al.*, 2004; Elith *et al.*, 2006). Species with narrow geographical distributions and specialized habitat requirements represent a particular challenge for statistical range representation in these models for three reasons. First, such species frequently have both small distributions and small sample sizes, creating power issues that may compromise model robustness (Stockwell & Peterson, 2002a,b; Pearson *et al.*, 2007; Wisz *et al.*, 2008). Second, SDMs model realized, not fundamental, niches (Malanson *et al.*, 1992; Hijmans & Graham, 2006) incorporating whatever sampling bias is inherent in the data (Wisz *et al.*, 2008). For narrowly distributed species, the opportunity to identify mistakenly a narrow climatic distribution as a fundamental niche limitation when it actually reflects other

DOI: 10.1111/j.1472-4642.2009.00567.x
www.blackwellpublishing.com/ddi

ecological constraints (e.g. dispersal barriers, biotic interactions, edaphic constraints) is high (Schwartz *et al.*, 2006; Wisz *et al.*, 2008). Finally, narrowly distributed species that are habitat specialists often have patchy distributions of occurrences. Thus, defining a general range extent becomes less useful from a management perspective compared with understanding habitat occupancy. This is a challenge for SDMs because they are meant to identify the overall extent of a species' range and may perform poorly if that range is heterogeneous or not well-sampled (Pearce *et al.*, 2001; Seoane *et al.*, 2005; McPherson & Jetz, 2007).

In this study, we explored the utility and output of four types of SDMs in predicting habitat and occurrence locales for six rare plant species endemic to the serpentine soils of the northern California Coastal Range. Each of our focal taxa is narrowly endemic. The most broadly distributed is limited to three counties in northern California (*c.* 26,000 km^2), while the smallest range is approximately 4 km^2 (CNPS 2006). There is naturally some uncertainty in the accuracy of these range estimates because forest inventories are frequently conducted for specific objectives, such as timber sales or road building (Stohlgren *et al.*, 1995; Kadmon *et al.*, 2004). Species distribution models may help correct such biases because they require the modeller to be explicit about the known and unknown parameters for the species' range, habitat type and occupancy (Elith *et al.*, 2002). Regardless of this beneficial feature of SDMs, the limited ranges and few known occurrences of the species in this study test the limits of how small a sample size can be used to construct a useful predictive model (Stockwell & Peterson, 2002b; Schwartz *et al.*, 2006; Pearson *et al.*, 2007). Our goals were to: (1) examine how limited occurrence data, model selection and habitat characterization affect the prediction of habitat occupancy, (2) determine which model or models are most useful for new population discovery, and (3) provide land managers with a strategy for biodiversity management in terms of population discovery, prioritizing conservation sites and identifying potential restoration sites.

Species distribution models are already used for a variety of ecological applications such as biodiversity discovery (Raxworthy *et al.*, 2003; Bourg *et al.*, 2005; Guisan *et al.*, 2006), conservation management (Cabeza *et al.*, 2004; Zacharias & Gregr, 2005) and global warming response modelling (Iverson *et al.*, 2004; Ballesteros-Barrera *et al.*, 2007; Gomez-Mendoza & Arriaga, 2007). Methodological advances in mathematical models, machine learning and statistical tools have resulted in significant improvements in model performance (Scott *et al.*, 2002; Guisan & Thuiller, 2005; Olden *et al.*, 2008). While no single best approach has emerged, several more recent models have consistently outperformed simpler, earlier models (Hirzel *et al.*, 2006). In particular, models that characterize the background environment where target species *do not* occur have generally performed better than those that do not (Stockwell & Peterson, 2002a; Engler *et al.*, 2004). This can be done by providing the models with environmental data gathered on true absences, i.e. places where the species is known not to occur or by generating a set of pseudo-absences when data on true absences are not available.

For modelling rare species without true absence data, pseudo-absences may be particularly appropriate, given the high probability that most points selected will be absences. Although small, there is a risk that if a pseudo-absence is assigned to a point occupied by the species, the negative value could have a disproportionate effect because there are few presence points with which to train the model. Similarly, in cases of low habitat saturation by the target species, a pseudo-absence occurring on appropriate habitat that is unoccupied could disproportionately train the model away from that habitat.

Recent research on how SDMs perform with respect to limited occurrence data has shed some light on what is important to make the models function with such constraints on input (Stockwell & Peterson, 2002a,b; Edwards *et al.*, 2004; Engler *et al.*, 2004; Hernandez *et al.*, 2006; Pearson *et al.*, 2007). We have taken these findings to heart, using commonly used models, adding remotely sensed data and designing our analysis to meet the primary goal of predicting new occurrences. Our main concern was to identify models that performed well given the challenging characteristics of our focal species, particularly: (1) restricted range, for which we might not understand all of the parameters; (2) the nature of the specific habitat, which may reflect either a species' environmental preferences or a suboptimal refuge that allows the species to avoid competition or predation (Naves *et al.*, 2003; Sanders & McGraw, 2005); and (3) presence where, even in optimal habitat, saturation may be low and/or a species may be cryptic (Brose *et al.*, 2003; Linkie *et al.*, 2006; Hare *et al.*, 2007).

METHODS

Study area

The study area is the Rattlesnake Creek Terrane (RCT), a 15 by 80-km tract of land on the south-western edge of the Klamath Mountains (Fig. 1). The RCT is characterized by a distinct geological history (hence the geological term 'terrane') that has resulted in a mélange of serpentine soils overlain by volcanic and sedimentary rocks of the Upper Triassic and Lower Jurassic periods (Wright & Wyld, 1994). In particular, serpentinite and peridotite units occur as a patchwork across the RCT and other parts of the northern coast range (Kruckeberg, 1984). To analyse the study site spatially, we divided the RCT polygon into roughly 50,000 grid cells, each of which was characterized by the presence or absence of the target species and a suite of environmental and climatic factors. The cells measured 150 m on a side – a size selected as a compromise between the large grain size of the climate data (*c.* 1 km^2) and the finer scale of soils (5–25 m), elevation (30 m) and occurrence data (5–10 m). This grain size is on the fine end of the SDM grain scale for landscape-level modelling, and given there is evidence suggesting that these models are

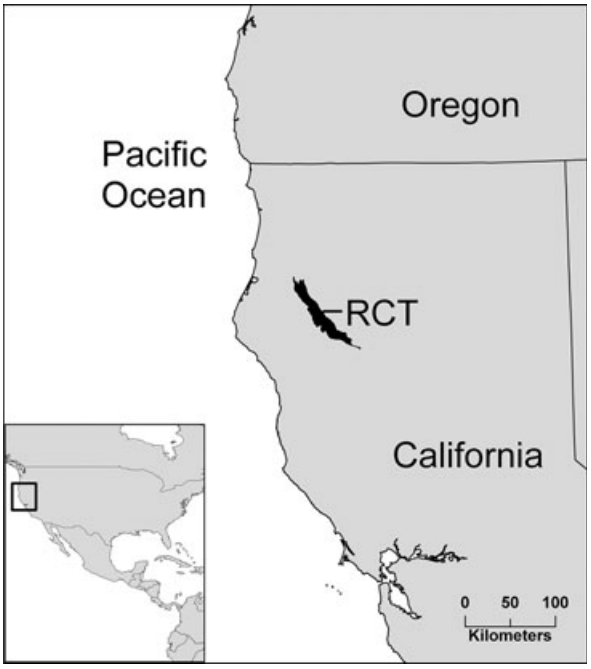

Figure 1 Location of the Rattlesnake Creek Terrane (RCT) study area.

relatively insensitive to even a 10-fold increase in grain size (Guisan *et al.*, 2007), we felt that our choice of size was reasonable.

Species

Our study focused on six plant taxa that are all endemic or nearly so to the northern Coast Range of California (Table 1). These taxa include two annual tarweeds (*Harmonia doris-nilesiae*, *Harmonia stebbinsii*), three perennial herbs [*Eriogonum libertini*, *Leptosiphon nuttallii* subsp. *howellii* (hereafter *L. nuttallii*) and *Minuartia rosei*] and a low-statured woody perennial (*Ericameria ophitidis*). All taxa are known from just one to three counties (http://www.calflora.org), are associated with ultramafic soils (Hickman, 1993; Nakamura & Nelson, 2001) and are tracked by the California Native Plant Society as species of special concern (CNPS 2006).

The USDA Forest Service (USFS) has occurrence data on rare plant populations in the Shasta-Trinity National Forest of north-western California (Fig. 1). These occurrences are based on rare plant surveys conducted by USFS botanists over the past 25 years, along with California Natural Diversity Database (CNDDB) occurrences for the region (http://www.dfg.ca.gov/biogeodata/cnddb). The number of georeferenced occurrences for our target taxa ranges from 9 to 129.

Models

The primary inputs for all of the models were the spatial coordinates of the known occurrences for the target species and a suite of environmental variables characterizing those locales. Three of the four models also required explicit information about where a species *does not* occur so that the model can discriminate between appropriate and inappropriate habitat. Because absence data for the target species was not recorded in the original field surveys, we used randomly selected pseudo-absences in lieu of actual data.

Pseudo-absences were generated by randomly assigning unoccupied grid cells within a polygon containing the collective known distribution of each species within the study region. We defined this polygon by drawing a 5-km buffer (Jenness, 2003) around the convex hull defined by the perimeter of known occurrences. The fourth model, Maxent, does not use pseudo-absences *per se*, but distinguishes between presences and random points from a background area using a probability distribution (Phillips *et al.*, 2006). For comparability among models, we made the background area for Maxent the same as the polygons from which the pseudo-absences were drawn from the other three models. Each model therefore used the extent of these polygons as the training region to predict occurrences over the entirety of the larger study region.

The four statistical models we used include: generalized linear models (GLM); artificial neural networks (ANN); a classification and regression tree model called random forests (RF) and a machine learning algorithm called maximum entropy or Maxent (ME). This study was not meant to be an

Table 1 Names and attributes of rare plant species modelled in Rattlesnake Creek Terrane; Shasta, Tehama and Trinity counties, California.

Species name*	Elevation (m)*, life history	Counties of occurrence*	No. occurrences	No. cells
Ericameria ophitidis (Serpentine goldenbush)	± 1600, Peren.	Shasta, Tehama, Trinity†	129	688
Eriogonum libertini (Dubakella Mountain buckwheat)	1200–1600, Peren.	Shasta, Tehama, Trinity	96	341
Harmonia doris-nilesiae (Niles' harmonia)	850–1400, Ann.	Trinity	28	81
Harmonia stebbinsii (Stebbins' harmonia)	1100–1400, Ann.	Shasta, Tehama, Trinity	18	100
Minuartia rosei (Peanut sandwort)	750–1350, Peren.	Shasta, Tehama, Trinity†	34	249
Leptosiphon nuttallii subsp. *howellii* (Mount Tedoc leptosiphon)	1500–1800, Peren.	Tehama‡	9	49

*Information from Hickman (1993) unless otherwise cited.
†CNPS (2006).
‡Shasta-Trinity National Forest records.

exhaustive test of the relative merits of modelling frameworks as others have done (e.g. Elith *et al.*, 2006). Instead, we wanted to evaluate how four commonly used models responded to the particular constraints of our target species and which model or combination of models would be most helpful for guiding new population discovery. We used these models to generate probability surfaces for the six target species, individually and for all species combined.

We included GLM because of its incorporation as a building block in many of the more sophisticated newer models and for its continued use in species distribution modelling (Austin, 2002; Brotons *et al.*, 2004; Austin, 2007; Gibson *et al.*, 2007). This method conforms well to presence–absence data given its threshold-type response and assumes additive or linear relationships between data.

Artificial Neural Networks provides a flexible generalization of GLM. In particular, ANN is considered to perform better than GLM when modelling nonlinear relationships (Lek *et al.*, 1996). Artificial Neural Networks makes use of intermediate nodes in what is referred to as a 'hidden layer', where each node contributes differentially with respect to the variables included in the model. It can accommodate interaction effects that are fed into the exponent term of logistic regression. The iterative model is constantly doing sensitivity analyses on minute variations of the included predictor variables, looking for optimal solutions. Our model had seven nodes in the hidden layer and through trial and error we chose a decay term of 0.01 to avoid over-fitting.

Recent studies have demonstrated the utility of RF compared to other techniques for modelling rare and invasive species, habitats and changes in species distribution under climate change (Prasad *et al.*, 2006; Cutler *et al.*, 2007; Benito Garzon *et al.*, 2008). Random forests create a suite of models using a classification and regression tree (CART) approach (Breiman, 2001). A general criticism of the CART method is its instability, especially at finer scales such as those considered here (Thuiller *et al.*, 2003); a small change in the data can make a large difference in the predictions of the fitted model. The RF method addresses this shortcoming by growing a collection of regression trees, each trained on a bootstrap sample of the original data. Each tree's output selection acts like a vote in an election and the model selects its outcome based on the maximum vote-getter after all the trees in the forest have voted. We populated our forest with 1000 trees for each species.

Maxent is a general purpose, machine learning model that searches for the target probability (predicted occurrence) based on the probability distribution that is closest to uniform using a set of constraints/variables imposed by the modeller (Phillips *et al.*, 2006). The model is flexible with respect to the types of variables used and the form of their relationship to a species' presence (e.g. linear, nonparametric, etc.). Maxent does not require the user to choose pseudo-absences. A review comparing 16 models on over 200 taxa found that newly emerging models, ME among them, consistently outperformed traditional linear methods (Elith *et al.*, 2006).

Table 2 Principle components analysis (PCA) of climatic variables (http://www.worldclim.org) used to generate summary climatic variables 1 and 2. The PCA as run on the 49,619 cells in the Rattlesnake Terrane using JMP 5.1 (SAS 2006).

Principal components: on correlations	PC1	PC2
Eigenvalue	6.667	3.498
Per cent	60.613	31.799
Cumulative per cent	60.613	92.412
Eigenvectors		
Total annual precipitation	−0.363	−0.049
Isothermality	−0.326	0.265
Annual mean temperature	0.055	0.517
Mean diurnal range	0.359	0.059
Precipitation of driest month	−0.038	−0.512
Precipitation seasonality (CV)	0.322	0.086
Precipitation of wettest month	−0.366	−0.048
Temperature annual range	0.378	−0.094
Maximum temperature of warmest month	0.311	0.314
Minimum temperature of coldest month	−0.110	0.508
Temperature seasonality (SD*100)	0.373	−0.140
Correlation with total precipitation (R^2)	−0.932	−0.087
Correlation with elevation (R^2)	0.277	−0.934
Correlation with slope (R^2)	−0.085	0.112
Correlation with aspect (R^2)	0.020	0.009

Predictor variables

For each grid cell in the study area, we calculated values for seven predictor variables: three topographic variables, two soils variables and two summary climate variables (Table 2). We used USGS digital ortho-quadrangles to generate the topographic variables of elevation, slope and aspect for each cell. These are proxy measures that indirectly assess gradients in climatic variables such as temperature, precipitation and incident radiation (Franklin, 1995).

We used a soil survey to identify all ultramafic soils and to classify the degree of serpentine syndrome for the soil units in the study area (Alexander, 2002). We used two soil classification variables: (1) a binary classification of soils as serpentine or not, and (2) the degree of severity of serpentine syndrome on a scale of 1–40. We then calculated distance from each grid cell to the nearest edge of an ultramafic site to estimate whether target species populations tended to occur near serpentine edges where the degree of serpentine syndrome might be less than in the centre of a unit due to mixing with non-serpentine soils. In addition, as in some cases, occurrence maps and soils units were mapped prior to GPS technology, incorporating the distance to the nearest serpentine unit provided a variable that could potentially correct for mapping or registration errors.

The climate variables were derived from WORLDCLIM, a 30 arc-second global climate data model (http://www.worldclim.org). WORLDCLIM provides information on 11 climatic variables that are thought to be potentially biologically important (Table 2). We distilled the information in these climate variables into two principal components: PC1 was strongly correlated with precipitation ($R^2 = 0.932$) and

explained 60.6% of the climatic variation; PC2 explained 31.8% of the remaining climatic variation and correlated strongly with elevation ($R^2 = 0.934$).

Using the digital elevation model (DEM), we constructed topographic roughness and aspect variables to evaluate the impact of variable terrain and solar radiation on model success. The roughness value (or topographic ruggedness index) for a cell was calculated as the square root of the average of the squared differences in elevation between the target centre cell and the eight cells immediately surrounding it (Riley *et al.*, 1999). While there are a variety of ways to transform aspect to represent incident radiation (Beers *et al.*, 1966; McCune & Keon, 2002), we chose a simple transformation that uses the compass value given by the DEM (0–360°), normalizes it between 0 and 1 and then takes the absolute value, which serves to fold the aspect, giving equal value to aspects that are equidistant east or west of the meridian. The calculation is given by:

$$\text{Aspect} = \left|180 - X\right|/180.$$

We also included the normalized difference vegetation index (NDVI) as a variable to highlight barren areas that are often signs of the serpentine syndrome. Vegetation on open rocky substrates has dried and browned by late summer. We used an NDVI measure from September (MODIS 16-day Albers Equal Area Conic for contiguous United States, 29-8-2006 to 13-9-2006) to get an alternate measure of low canopy cover and low-productivity sites where the target species tend to grow.

Analysis

For each of the six target species, we ran the three SDMs (GLM, ANN and RF) using the statistical software R (Version 2.6.0, http://www.r-project.org) and ME using Maxent (Version 3.1.0, Phillips *et al.* 2006) with the same suite of predictor variables. Each model run generated a probability of occurrence for each of the 49,619 RCT grid cells. Multiple model runs ($n = 30$) gave a probability distribution for each cell and the final output for a set of runs was a mean predictive cell value ranging from 0 to 1.

To evaluate the relative performances of the models, we compared SDM outputs using the area under the curve (AUC) of the receiver operating characteristic (Hanley & McNeil, 1982). This technique requires the modeller to divide the occurrence data into training and testing batches, where the first batch is the input for the predictive cell values that the model will generate and the second batch serves as the standard for evaluating how well the model performs. We selected a 70–30% training–testing mix. We then ran the models 100 times, with the training and testing cells selected at random each time – a variation of bootstrap resampling. Sensitivity (how often the model predicts true presences) was plotted against one-minus-specificity (the false-positive rate measuring how often the model predicts presences where none occurs) for the set of runs. The performance score is the AUC measured on a scale of 0–1, where 1 is a perfect score (no errors of omission or commission) and 0.5 is what we would expect from random selection.

We used Cohen's Kappa statistic as an alternative evaluation criterion because of its recognized value in identifying how well models predict species presence (Fielding & Bell, 1997; Prasad *et al.*, 2006). To assess whether the four models used provided consistent predictions in terms of environmental variables identified as important, model fit and geographical distribution of favourable habitat, we performed a simple linear correlation of predicted values for all grid cells, using pair-wise comparisons for all models for each species (Prasad *et al.*, 2006; Termansen *et al.*, 2006).

Pseudo-absence sensitivity

Because the presence-to-absence data ratio can affect model performance, we examined model sensitivity to the number of pseudo-absences included, looking for the point at which model output stabilized. To determine the point at which there were minimal or no improvements to confidence interval values by the addition of more pseudo-absences, we examined model results for 100 runs of each of the three presence–absence model methods (excluding ME) using 1, 2, 4, 8 and 16 times the number of presences as pseudo-absences for each species occurrence. We selected a 1 : 2 presence to pseudo-absence ratio based on minimal, non-significant improvements in AUC values (and in some cases a worsening) with increasing pseudo-absences beyond that. Our selected ratios are consistent with other studies (Kvamme, 1985; Zaniewski *et al.*, 2002; Olivier & Wotherspoon, 2006).

Population discovery

We estimated SDM identification of new population discovery in two ways. First, we used data collected from the USDA Forest Service's Klamath-Shasta-Trinity vegetation plots (the KST dataset) to test model specificity. These plots represent an independent assessment of occurrences for our target species from data not used in model development. We tested the models using this dataset by examining what threshold value was needed to include all occurrences.

As an additional method for evaluating model utility, we conducted field surveys of 11 sites with high predictive scores for multiple species using the best performing model, RF. Sites were cruised on foot and visually inspected for species presence. Site locations with occurrences (36 cells) were recorded using a hand-held GPS unit (Garmin 76S). We digitized the cruised pathways on a GIS (ArcMap 9.0, ESRI, Redlands, CA, USA), and identified the number of cells inspected, recording their predictive scores. For each species, we compared the predicted values for the cells with that species to those without, but containing one or more of the other target species.

Finally, we used the 935 KST plots to examine cell values for species presences versus absences. To assess scores of

unoccupied cells, we used the 911 cells from the KSF dataset that contained none of the target species. We compared these values to the 24 KSF cells that contained one or more of our target taxa. Again, we used the AUC values from the RF model and then calculated the percentile ranking for those cells relative to the entire study area. The KSF plots were not explicitly selected to survey for the target species or serpentine soil plants, but were instead designed to characterize the plant communities of the study region. The botanists conducting these surveys may not have spotted the target species if conditions were poor (e.g. non-flowering or desiccated annuals), but were familiar with the local flora and, in principle, could identify the target species.

RESULTS

Model fit

We evaluated the relative success of the four model methods in predicting target species occurrences by comparing mean AUC values and Kappa statistics (Table 3). Model fit was generally good for all models and species, with no relationship between the number of known occurrences (Table 1) and model fit (Table 3). Among the models, RF yielded results with the highest mean AUC and Kappa values for each species, followed by ME. For the AUC comparison, RF was the most consistent in its performance, as indicated by small standard deviations, although ME had the smallest standard deviations with respect to Kappa.

Important variables in creating model fit were consistent across species (Table 4). Elevation and distance to serpentine were the two most important variables. Distance to serpentine was an important variable for all species except *L. nuttallii*. Distance to serpentine is likely to be a better predictor of occurrence than serpentine syndrome because of the high probability of small registration errors in location (many records were made prior to GIS technology and the mapping of serpentine units, and some units were too small and intermixed to be classified as serpentine). Climate PC1 and PC2 were also relatively strong predictors of occurrence, and NDVI, a measure of canopy cover, was moderately important variable for *L. nuttallii* and *H. doris-nilesiae*.

Model comparisons

The cross model correlations were variable, depending on species and models compared. Six of the seven correlations that were < 0.3 involved ANN (Table 5). The highest average correlations involved RF with ME and GLM. Through examination of maps and correlation statistics we observed that RF and ME generated similar results, both in terms of AUC scores and the physical locations where occurrences were predicted. No methods correlated consistently well across all species. Instead, correlation loosely tracked the number of occurrences, performing best for *E. libertini* and *E. ophitidis*, followed by moderate correlation for *H. doris-nilesiae* and *M. rosei*, and generally poor agreement for *H. stebbinsii* and *L. nuttallii*.

Population discovery

Of the 935 KSF plots that fell within the RCT, 24 contained one or more of our target species and only three species occurred

Table 3 Comparison of four model methods used to predict occurrences for rare serpentine-endemic plant species. The models include: general linear models (GLM); artificial neural networks (ANN); random forests (RF); and Maxent (ME). Values presented are: (a) the average scores for 100 runs of the area under the curve (AUC) statistic of the receiver operating characteristic; and (b) Cohen's Kappa statistic, which measures the proportion of presences and absences predicted correctly after accounting for chance. Numbers in parentheses are standard deviations; bold indicates the model with the highest score.

	GLM	ANN	RF	ME	Average
(a) AUC scores					
Eriogonum libertini	0.927 (0.012)	0.910 (0.019)	**0.949** (0.011)	0.945 (0.006)	0.933
Ericameria ophiditis	0.875 (0.012)	0.879 (0.059)	**0.951** (0.009)	0.910 (0.007)	0.904
Harmonia doris-nilesiae	0.843 (0.043)	0.830 (0.051)	**0.956** (0.023)	0.939 (0.020)	0.892
Harmonia stebbinsii	0.829 (0.045)	0.827 (0.066)	**0.954** (0.021)	0.948 (0.012)	0.890
Leptosiphon nuttallii	0.952 (0.041)	0.891 (0.071)	**0.976** (0.024)	0.968 (0.011)	0.947
Minuartia rosei	0.737 (0.031)	0.766 (0.075)	**0.942** (0.015)	0.904 (0.016)	0.837
Six species	0.881 (.010)	0.888 (0.059)	**0.947** (0.006)	0.901 (0.006)	0.903
Mean AUC	0.863	0.856	0.954	0.931	
(b) Cohen's Kappa					
Eriogonum libertini	0.698 (0.035)	0.673 (0.074)	**0.756** (0.037)	0.742 (0.014)	0.717
Ericameria ophiditis	0.568 (0.030)	0.592 (0.067)	**0.761** (0.028)	0.650 (0.013)	0.643
Harmonia doris-nilesiae	0.548 (0.080)	0.576 (0.094)	**0.809** (0.076)	0.808 (0.025)	0.685
Harmonia stebbinsii	0.542 (0.073)	0.586 (0.107)	**0.822** (0.062)	0.765 (0.022)	0.679
Leptosiphon nuttallii	0.836 (0.094)	0.766 (0.104)	**0.869** (0.072)	0.834 (0.029)	0.826
Minuartia rosei	0.370 (0.050)	0.438 (0.084)	**0.753** (0.048)	0.687 (0.026)	0.562
Six species	0.587 (.022)	0.601 (0.065)	**0.740** (0.020)	0.656 (0.009)	0.646
Mean Kappa	0.593	0.605	0.787	0.735	

Table 4 Relative variable importance for each target species plus all species combined (RARE) using random forests. Bold indicates relative variable importance of > 0.15; values > 0.10 are italicized. See text for an explanation of the two climate variable principle components and aspect. Species include: *Eriogonum libertini* (ERLI); *Ericameria ophiditis* (EROP); *Harmonia stebbinsii* (HAST); *Harmonia doris-nilesiae* (HADO); *Leptosiphon nuttallii* (LENU); and *Minuartia rosei* (MIRO).

	ERLI	EROP	HADO	HAST	LENU	MIRO	RARE
Elevation	**0.21**	**0.23**	*0.12*	**0.21**	**0.35**	0.10	**0.18**
Slope	0.04	0.03	0.05	0.04	0.02	0.04	0.04
Bioclim PC1	*0.11*	*0.14*	*0.11*	**0.24**	0.08	**0.26**	**0.16**
Bioclim PC2	*0.14*	*0.13*	*0.11*	*0.15*	**0.27**	*0.13*	*0.12*
Serpentine syndrome	0.02	0.05	0.02	0.01	0.00	0.05	0.04
Degree of serpentine	0.03	0.06	0.03	0.01	0.01	0.07	0.05
Distance to serpentine	**0.23**	*0.15*	**0.22**	**0.18**	0.04	*0.13*	**0.18**
Incident radiation	0.05	0.05	0.09	0.03	0.01	0.06	0.05
Aspect	0.04	0.04	0.09	0.03	0.02	0.05	0.05
Roughness	0.04	0.04	0.05	0.03	0.03	0.04	0.04
NDVI	0.10	0.08	*0.11*	0.07	*0.17*	0.08	0.10

Table 5 Cross model correlations (Method 1 compared to Method 2) for six rare endemic plant species in the Rattlesnake Terrane of northern California. Models include: general linear models (GLM); artificial neural networks (ANN); random forest (RF); and Maxent (ME). Species include: *Eriogonum libertini* (ERLI); *Ericameria ophiditis* (EROP); *Harmonia stebbensii* (HAST); *Harmonia doris-nilesiae* (HADO); *Minuartia rosei* (MIRO); and *Leptosiphon nuttallii* (LENU). All pair-wise correlations > 0.7 are bold-faced for emphasis.

Method 1	GLM	GLM	GLM	ANN	ANN	RF	
Method 2	ANN	RF	ME	RF	ME	ME	Average
Species							
ERLI	**0.861**	**0.900**	**0.805**	**0.827**	**0.761**	**0.816**	0.828
EROP	0.512	**0.794**	**0.755**	0.495	0.489	**0.762**	0.635
HAST	0.361	0.519	0.383	0.288	0.208	0.650	0.402
HADO	0.504	**0.725**	0.306	0.501	0.122	0.483	0.440
MIRO	0.611	0.380	0.659	0.344	0.495	0.342	0.472
LENU	0.140	0.426	0.206	0.225	0.245	**0.719**	0.327
Six species	**0.939**	**0.808**	**0.884**	**0.826**	**0.860**	**0.812**	0.855
Average	0.561	0.650	0.571	0.501	0.454	0.655	

three or more times. Results suggest that GLM was the most discriminating model and ME the least. In total, all plot occurrences fell within the top 8.9% of cells (cells with the highest predicted probability of occurrence) using GLM, compared to the top 21% of cells for ME (Table 6). The results varied considerably according to model used and species evaluated. Therefore, we concentrated on a more general analysis of these data and arbitrarily selected the top 25% as a cut-off threshold to limit targeted species searches.

Table 6 An assessment of model performance by examining the minimum predicted cell values that capture all occurrences of the target species in the Klamath-Shasta-Trinity data set. For the three species with multiple occurrences, we recorded the percentage of cells that would need to be surveyed to capture all KSF plot occurrences. Models include: generalized linear models (GLM); artificial neural networks (ANN); random forest (RF); and maximum entropy (ME).

	GLM	ANN	RF	ME	*n*
Eriogonum libertini	10.2	13.9	14.7	22.1	6
Ericameria ophiditis	6.40	3.30	4.09	23.2	10
Minuartia rosei	10.1	21.9	14.3	17.7	8
Mean fraction	8.9	13.0	11.0	21.0	

Table 7 Two sets of data are used to ascertain differences between cells with species occurrences and those without. (a) In a search area covering approximately four miles, 36 grid cells were found to be occupied by one or more target species. The average percentile ranking (maximum = 100) for occurrences of each taxa exceeded the average ranking for cells that were found lacking that target species but occupied by other target species. Rankings are based the random forests model. (b) The USFS used 935 plots (KSF plots) to assess ecological attributes in the study region. Among these, 24 plots contained an occurrence of at least one species. The average cell value for the KSF presences significantly ($P < 0.001$) exceeded the average cell value for the 911 cells in which no target species was found. The target species found include: *Eriogonum libertini* (ERLI); *Ericameria ophiditis* (EROP); *Harmonia stebbinsii* (HAST); and *Minuartia rosei* (MIRO). Numbers in parentheses refer to number of occurrences.

	ERLI	EROP	HAST	MIRO
(a)				
Population search presences	96.4 (10)	93.9 (11)	97.0 (2)	98.4 (11)
Population search absences	93.2* (26)	92.4 (25)	96.3 (34)	95.8* (25)
(b)				
KSF plot presences	92.8 (6)	98.4 (10)	N/A	92.8 (8)
KSF plot absences	79.0†	79.1†	80.5†	72.5†

*Marginally significant difference from population search presence, $P < 0.1$ using a one-tailed *t*-test.
†Highly significant difference ($P << 0.001$) from population search presence using single factor ANOVA.

The second method used to assess the utility of the models for new population discovery entailed field visits to the sites ranked highest by the RF model. This exercise resulted in the discovery of 16 new populations at 11 different sites in 36 cells (a site was defined as a grouping of adjacent cells, all of which contained individuals of at least one target species). These sites also ranked highly according to the ME output. The cells in which new species presences were found consistently had

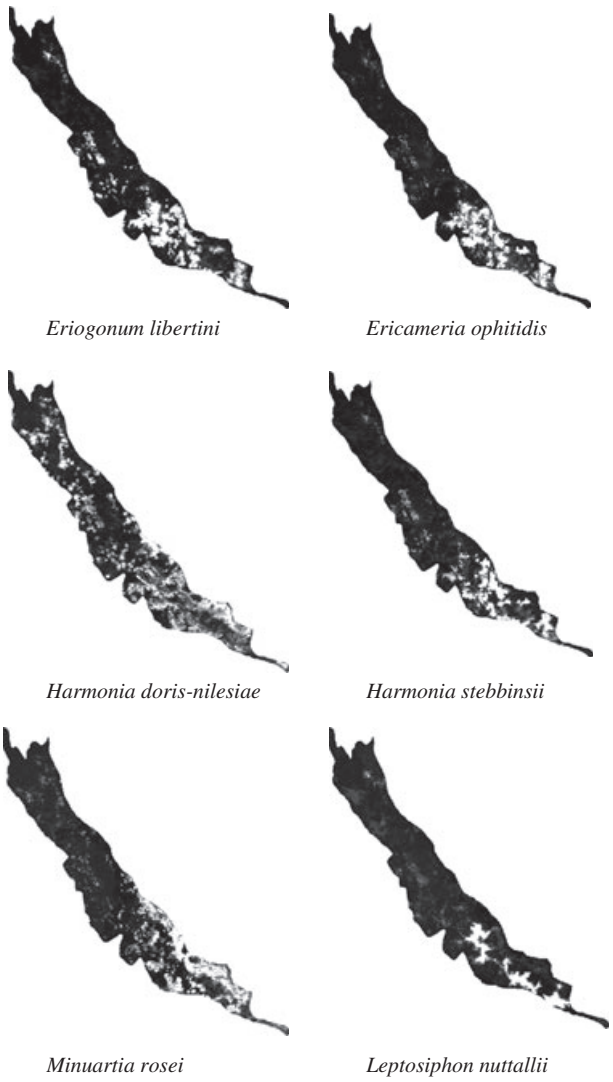

Figure 2 Predicted occurrence maps for six target species based on probability of occurrence for individual grid cells of the Rattlesnake Creek Terrane. White areas indicate higher probability of a rare species occurrence decreasing to grey then black for low probability of occurrence. Values were calculated using the random forests model trained with all occurrence data.

higher model probability values than cells where the species were absent (Table 7). For each of the 36 cells where at least one target species was found, we considered these to be true absences for any target species not found in that cell. The predictive values of those absences were not significantly different from the values for the remaining cells where that species was found. In contrast, the KSF plots with new occurrences scored significantly higher for target taxa than in the 911 cells where no occurrences were found. Our searches did not reveal any new populations of *H. doris-nilesiae* or *L. nuttallii* – the two most geographically restricted taxa. The successful model-driven searches for the other species suggest both that RF and ME can be effective tools for population discovery and that *E. libertini*, *E. ophiditis* and *M. rosei* may be less rare than previously supposed.

Figure 3 Predicted occurrence map for all six target species combined using the random forests model. White areas indicate higher probability of a rare species occurrence decreasing to grey then black for low probability of occurrence.

We used to generate the predictive map for each species because it consistently gave the best fits (highest AUC and Kappa statistics for each species) using the entire data set (Fig. 2). Having assessed model fit by withholding occurrences, we used all test occurrences to identify potential locales for species occurrence. This was accomplished by generating AUC probabilities for each of the grid cells in the study area. Two or more target species were found to co-occur on forest plots mapped for a single species on multiple occasions. We hypothesized that there may be shared habitat preferences among the species that are not defined explicitly, but that could potentially wield predictive power for an anonymous target species using the larger training and testing data set. Therefore, we also used RF to generate a prediction map for all of the target species together (Fig. 3).

DISCUSSION

Habitat specialist species with narrow geographical ranges test the limits of SDMs. For our study species and for many other restricted range plants, it is not well understood to what extent distributions are limited by suitable habitat availability or by barriers to movement and stochastic processes (Wiser *et al.*, 1998). Furthermore, while we may be able to define rough range limits for these species at a landscape scale, conservation management actions may require planning at the scale of habitat occupancy within the defined range. At a conceptual level, range and habitat occupancy are fundamentally different concepts (Gaston, 2003) and SDMs have mostly been applied to predict species ranges. However, the better a SDM, the more it can distinguish between the characteristics of places where species occur and the surrounding background matrix where they are absent, i.e. a species' habitat occupancy. This study highlights some of the strengths and limitations of using SDMs for habitat occupancy modelling. One primary utility is for new population discovery across a geographically small but heterogeneous landscape. Given that we lacked true absence data for this study, we do not claim that our results provide

generalizable assessments of the modelling platforms for small samples. Instead, we assert that the model performances reported here represent a case study that may in turn be used to inform a more general comparison under similar parameters with true absence data.

The fact that new species occurrences were discovered at locations where models indicated high probability is encouraging and suggests that SDMs can discriminate between the background matrix and potential habitat even at fine scales. Effective new population discovery using our SDMs also suggests that the differential habitat values described by the model might also be used to guide other types of conservation management decisions. These conservation decisions include identifying potential restoration sites or scoring aggregate conservation values based on high densities of cells with high habitat values for multiple species.

If a species fails to occupy many suitable sites within its distribution (i.e. has low habitat saturation), then non-occurrence data become less informative and a model may give high habitat suitability scores and good AUC values without offering much in the way of on-the-ground utility (Gibson *et al.*, 2007). While our model outputs led to the discovery of new populations of some target species, our finding in the field was that a great deal of seemingly appropriate habitat did not contain these plants. An approach to improving SDM output accuracy for such applications would be to run SDMs on an iterative basis, where model performance is monitored as inputs are updated with the addition of new populations and real absences are determined from field surveys after the first model outputs have been field tested. There is a recognized need for such iterative efforts in landscape ecology (Gardner & Urban, 2007) and its application to rare species modelling could help refine predictions of suitable habitat for rare species. Such iterative work could be achieved by collaborating with resource management agencies to develop the capacity to run SDMs with new data being incorporated at the end of each field season.

A second issue that may reduce the ability of SDMs to discriminate suitable habitat from background matrix relates to the selection of non-occurrences for edaphic specialist species, such as our focal species that are largely restricted to ultramafic soils. If edaphic constraints are strong, but pseudo-absences are not restricted to these edaphic features, then the resulting model can have good fit, but may represent little more than a soils map. Alternatively, restricting pseudo-absences to the edaphic features may test whether variation within soils plays a strong role in predicting occurrences, but it is likely to result in a model with poor overall fit and could eliminate possible occurrences not determined by soils. There may be no clear solution to this problem. We addressed this issue by limiting pseudo-absences to areas within 5 km of the convex polygon defining the suite of known locations of rare species. These areas contain, but are not exclusively composed of, ultramafic soils.

An additional concern is that few known occurrences make it difficult to model actual and projected population distributions with confidence (Stockwell & Peterson, 2002b; Schwartz *et al.*, 2006; Wisz *et al.*, 2008). Sparse data make it difficult for models to distinguish important predictor variables. Although there has been a proliferation of new modelling techniques (Guisan & Thuiller, 2005; Elith *et al.*, 2006), relatively few have focused on treating very restricted data sets (Papes & Gaubert, 2007; Pearson *et al.*, 2007), despite the fact that this is a common characteristic of rare species. We found a lower bound in the utility of modelling geographically restricted species with few occurrences. In this study, a small number of occurrences ($n = 9$) in a confined area (*c.* 4 km^2) combined with relatively homogeneous environmental conditions did not provide the models with enough spatial variation to differentiate between appropriate and inappropriate habitat for the species in question (*L. nuttallii*). The models did return high AUC values for this species, but the mapped results simply confirmed that the species is narrowly endemic, identified few sites outside the known range and led to the discovery of no new populations. Although this may be an entirely accurate portrayal of the species, our overall lack of confidence in the robustness of the output along with its low utility encouraged us to discount the result.

In contrast, *H. stebbinsii*, with 18 occurrences and a distribution of at least 15 km^2, had sufficient environmental and climatic variation to generate enough pseudo-absences and get better results. The discovery of two new populations of this species during field validation was possibly the most impressive result of the modelling. Thus, while there is no fixed lower-bound in occurrence number and range size, we found an empirical difference in the contrasting results for *H. stebbinsii* and *L. nuttallii* that may arguably be attributed to thresholds for these variables.

For restricted-range species, it can be difficult to define geographical boundaries. This ambiguity can make the selection of modelling area a subjective process. We restricted the placement of pseudo-absences to within a 5-km convex buffer around the known occurrences for each species, based on the average maximum distance between populations (we found no data on the maximum dispersal distances of our target species). There was nothing to suggest, however, that habitat quality dropped markedly outside the buffered boundaries. For example, the presence of serpentine soils, environmental variables and climatic conditions in the northern part of the RCT appear to be similar to the southern part, yet there are no recorded occurrences in the north. Although model output was applied to the entire region, all of the data input for training the models came from the buffered polygons and the model outputs predicted few occurrences north of these boundaries. The KSF dataset provides encouraging justification for this constraint, with strong evidence of the lack of target populations in the north. For modelling future climate scenarios of these species, we would recommend a broader geographical domain, which might identify suitable locations outside the known historic distributions (McLachlan *et al.*, 2007).

Model selection may be important for rare species with sparse data. Although the target species in this study are poorly studied, their edaphic restrictions, limited ranges and low densities make them interesting species to model. In general, the models performed similarly and where they diverged could be at least partially predicted by the input (fewer occurrences generally resulted in lower correlations among models). In contrast, the places on the map where the models did converge in their prediction of high suitability were the places we went and found several new populations.

Nevertheless, there were clear differences in the models. The Kappa statistic confirmed what the AUC values indicated and what others have also found in SDM comparisons (Hernandez *et al.*, 2008): namely, that RF and ME were the best performing models and had the highest mean correlation with one another. While GLM and ANN were similar in terms of statistics, we found ANN to be the least useful model, giving predictive maps that often did not match the other models evaluated or corroborate our field experience. Although more consistent than ANN, GLM did not perform as well as the other two models overall, which may be an example of its inconsistent performance with low-prevalence species that others have also reported (Meynard & Quinn, 2007). Fitting a polynomial term to the GLM (which we did not do) or replacing it with a generalized additive model may have improved performance by making it less restrictive (Yee & Mitchell, 1991; Austin *et al.*, 2006), yet Elith *et al.* (2006) still found ME to outperform both of these variations on the linear GLM. Of the model types we did use, we found the inclusion of pseudo-absences (by GLM, ANN and RF) or implied absences (by ME) enhanced the ability of the models to discriminate among habitat types.

Other than *L. nuttallii*, whose nine occurrences and 4-km^2 range seemed to fall below the lower limits of data required for adequate modelling, our results suggest that there are SDMs that can effectively and accurately model species with multiple forms of rarity to the point of leading to new population discovery. Given sufficient resolution of predictor variables, we recommend the development of SDMs for rare plant species, particularly if there is an opportunity for iterative model and search phases of future studies.

ACKNOWLEDGEMENTS

We thank the people at the Shasta-Trinity National Forest for making this study possible. *E. Alexander* helped with all questions soil-related. *H. Safford* put us in touch with the USFS team. *C.K. Lamar* provided much help in the field. *A. Fremier*, *J. Viers*, *S. Waddell* and *D. Juhn* helped with GIS issues. The Schwartz laboratory provided helpful commentary. The study was funded by the USDA Forest Service.

REFERENCES

Alexander, E. (2002). *Trinity serpentine soil survey: a detail survey of serpentine soils in the southern exposure of the Rattlesnake Creek Terrane, Klamath Mountains*. United States Forest Service, Redding, CA.

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

Austin, M. (2007) Species distribution models and ecological theory: a critical assessment and some possible new approaches. *Ecological Modelling*, **200**, 1–19.

Austin, M.P., Belbin, L., Meyers, J.A., Doherty, M.D. & Luoto, M. (2006) Evaluation of statistical models used for predicting plant species distributions: role of artificial data and theory. *Ecological Modelling*, **199**, 197–216.

Ballesteros-Barrera, C., Martinez-Meyer, E. & Gadsden, H. (2007) Effects of land-cover transformation and climate change on the distribution of two microendemic lizards, genus Uma, of northern Mexico. *Journal of Herpetology*, **41**, 733–740.

Beers, T.W., Dress, P.E. & Wensel, L.C. (1966) Aspect transformation in site productivity research. *Journal of Forestry*, **64**, 691.

Benito Garzon, M., Sanchez De Dios, R. & Sainz Ollero, H. (2008) Effects of climate change on the distribution of Iberian tree species. *Applied Vegetation Science*, **11**, 169–178.

Bourg, N.A., Mcshea, W.J. & Gill, D.E. (2005) Putting a cart before the search: successful habitat prediction for a rare forest herb. *Ecology*, **86**, 2793–2804.

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

Brose, U., Martinez, N.D. & Williams, R.J. (2003) Estimating species richness: sensitivity to sample coverage and insensitivity to spatial patterns. *Ecology*, **84**, 2364–2377.

Brotons, L., Thuiller, W., Araujo, M.B. & Hirzel, A.H. (2004) Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. *Ecography*, **27**, 437–448.

Cabeza, M., Araujo, M.B., Wilson, R.J., Thomas, C.D., Cowley, M.J.R. & Moilanen, A. (2004) Combining probabilities of occurrence with spatial reserve design. *Journal of Applied Ecology*, **41**, 252–262.

CNPS (2006) *Inventory of rare and endangered plants*. California Native Plant Society, Sacramento, CA. URL: http://www.cnps.org/inventory

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

Edwards, T.C., Cutler, D.R., Geiser, L., Alegria, J. & Mckenzie, D. (2004) Assessing rarity of species with low detectability: lichens in Pacific Northwest forests. *Ecological Applications*, **14**, 414–424.

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

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, C., Nakamura, M., Nakazawa, Y., Overton, J.M.,

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.

Engler, R., Guisan, A. & Rechsteiner, L. (2004) An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *Journal of Applied Ecology*, **41**, 263–274.

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: geographic modelling of biospatial patterns in relation to environmental gradients. *Progress in Physical Geography*, **19**, 474–499.

Gardner, R.H. & Urban, D.L. (2007) Neutral models for testing landscape hypotheses. *Landscape Ecology*, **22**, 15–29.

Gaston, K.J. (2003). *The structure and dynamics of geographic ranges*. Oxford Series in Ecology and Evolution, Oxford.

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.

Gomez-Mendoza, L. & Arriaga, L. (2007) Modeling the effect of climate change on the distribution of oak and pine species of Mexico. *Conservation Biology*, **21**, 1545–1555.

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

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

Guisan, A., Graham, C.H., Elith, J. & Huettmann, F. (2007) Sensitivity of predictive species distribution models to change in grain size. *Diversity and Distributions*, **13**, 332–340.

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

Hare, K.M., Hoare, J.M. & Hitchmough, R.A. (2007) Investigating natural population dynamics of Naultinus manukanus to inform conservation management of New Zealand's cryptic diurnal geckos. *Journal of Herpetology*, **41**, 81–93.

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

Hernandez, P.A., Franke, I., Herzog, S.K., Pacheco, V., Paniagua, L., Quintana, H.L., Soto, A., Swenson, J.J., Tovar, C., Valqui, T.H., Vargas, J. & Young, B.E. (2008) Predicting species distributions in poorly-studied landscapes. *Biodiversity and Conservation*, **17**, 1353–1366.

Hickman, J.C. (1993). *The Jepson manual: higher plants of California*. University of California Press, Berkeley, CA.

Hijmans, R.J. & Graham, C.H. (2006) The ability of climate envelope models to predict the effect of climate change on species distributions. *Global Change Biology*, **12**, 2272–2281.

Hirzel, A.H., Le Lay, G., Helfer, V., Randin, C. & Guisan, A. (2006) Evaluating the ability of habitat suitability models to predict species presences. *Ecological Modelling*, **199**, 142–152.

Iverson, L.R., Schwartz, M.W. & Prasad, A.M. (2004) Potential colonization of newly available tree-species habitat under climate change: an analysis for five eastern US species. *Landscape Ecology*, **19**, 787–799.

Jenness, J. (2003) *Random point generator 1.27, p. Shareware written in Avenue for ArcGIS 3.2* Available at http://www.jennessent.com/arcview/random_points.htm.

Kadmon, R., Farber, O. *et al.* (2004) Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. *Ecological Applications*, **14**, 401–413.

Kruckeberg, A.R. (1984) *California serpentines: flora, vegetation, geology, soils and management problems*. University of California Press, Berkeley, CA.

Kvamme, K.L. (1985) Determining empirical relationships between the natural environment and prehistoric site location: a hunter-gatherer example. *For concordance in archaeological analysis: bridging data structure, quantitative technique, and theory* (ed. by C. Carr), pp. 208–238. Westport Publishers, Kansas City, KS.

Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J. & Aulagnier, S. (1996) Application of neural networks to modelling nonlinear relationships in ecology. *Ecological Modelling*, **90**, 39–52.

Linkie, M., Chapron, G., Martyr, D.J., Holden, J. & Leader-Williams, N. (2006) Assessing the viability of tiger subpopulations in a fragmented landscape. *Journal of Applied Ecology*, **43**, 576–586.

Malanson, G.P., Westman, W.E. & Yan, Y.L. (1992) Realized versus fundamental niche functions in a model of chaparral response to climatic change. *Ecological Modelling*, **64**, 261–277.

McCune, B. & Keon, D. (2002) Equations for potential annual direct incident radiation and heat load. *Journal of Vegetation Science*, **13**, 603–606.

McLachlan, J.S., Hellmann, J.J. & Schwartz, M.W. (2007) A framework for debate of assisted migration in an era of climate change. *Conservation Biology*, **21**, 297–302.

McPherson, J.M. & Jetz, W. (2007) Effects of species' ecology on the accuracy of distribution models. *Ecography*, **30**, 135–151.

Meynard, C.N. & Quinn, J.F. (2007) Predicting species distributions: a critical comparison of the most common statistical models using artificial species. *Journal of Biogeography*, **34**, 1455–1469.

Nakamura, G. & Nelson, J.K. (2001). *Illustrated field guide to selected rare plants of Northern California*. University of California Agriculture and Natural Resources, Los Angeles, CA.

Naves, J., Wiegand, T., Revilla, E. & Delibes, M. (2003) Endangered species constrained by natural and human factors:

the case of brown bears in northern Spain. *Conservation Biology*, **17**, 1276–1289.

Olden, J.D., Lawler, J.J. & Poff, N.L. (2008) Machine learning methods without tears: a primer for ecologists. *Quarterly Review of Biology*, **83**, 171–193.

Olivier, F. & Wotherspoon, S.J. (2006) Modelling habitat selection using presence-only data: case study of a colonial hollow nesting bird, the snow petrel. *Ecological Modelling*, **195**, 187–204.

Papes, M. & Gaubert, P. (2007) Modelling ecological niches from low numbers of occurrences: assessment of the conservation status of poorly known viverrids (Mammalia, Carnivora) across two continents. *Diversity and Distributions*, **13**, 890–902.

Pearce, J., Ferrier, S. & Scotts, D. (2001) An evaluation of the predictive performance of distributional models for flora and fauna in north-east New South Wales. *Journal of Environmental Management*, **62**, 171–184.

Pearson, R.G., Raxworthy, C.J., Nakamura, M. & Peterson, A.T. (2007) Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. *Journal of Biogeography*, **34**, 102–117.

Phillips, S.J., Anderson, R.P. & Schapire, R.E. (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231–259.

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.

Rabinowitz, D., Cairns, S. & Dillon, T. (1986) Seven forms of rarity and their frequency in the flora of the British Isles. *Conservation Biology: The Science of Scarcity and Diversity. Xiii+584p*, (ed. by M. E. Soule), pp. 182–204. Sinauer Associates, Inc., Sunderland, Mass., USA. Illus. Paper. Maps.

Raxworthy, C.J., Martinez-Meyer, E., Horning, N., Nussbaum, R.A., Schneider, G.E., Ortega-Huerta, M.A. & Peterson, A.T. (2003) Predicting distributions of known and unknown reptile species in Madagascar. *Nature*, **426**, 837–841.

Riley, S.J., Degloria, S.D. & Elliot, R. (1999) A terrain ruggedness index that quantifies topographic heterogeneity. *Intermountain Journal of Sciences*, **5**, 1–4.

Rushton, S.P., Ormerod, S.J. & Kerby, G. (2004) New paradigms for modelling species distributions? *Journal of Applied Ecology*, **41**, 193–200.

Sanders, S. & McGraw, J.B. (2005) *Hydrastis canadensis* L. (Ranunculaceae) distribution does not reflect response to microclimate gradients across a mesophytic forest cove. *Plant Ecology*, **181**, 279–288.

SAS (2006) *JMP*. SAS Institute, Cary, NC.

Schwartz, M.W., Iverson, L.R., Prasad, A.M., Matthews, S.N. & O'connor, R.J. (2006) Predicting extinctions as a result of climate change. *Ecology*, **87**, 1611–1615.

Scott, J.M., Heglund, P.J., Morrison, M.L., Haufler, J.B., Raphael, M.G., Wall, W.A. & Samson, F.B. (2002) *Predicting species occurrences: issues of accuracy and scale*. Island Press, Washington, DC.

Seoane, J., Carrascal, L.M., Alonso, C.L. & Palomino, D. (2005) Species-specific traits associated to prediction errors in bird habitat suitability modelling. *Ecological Modelling*, **185**, 299–308.

Stockwell, D.R. & Peterson, A.T. (2002a) Controlling bias in biodiversity. *Predicting species occurrences: issues of accuracy and scale.*(ed. by) J.M. Scott, P.J. Heglund, M.L. Morrison, J.B. Haufler, M.G. Raphael, W.A. wall and F.B. Samson, pp. 537–546. Island Press, Washington, DC.

Stockwell, D.R.B. & Peterson, A.T. (2002b) Effects of sample size on accuracy of species distribution models. *Ecological Modelling*, **148**, 1–13.

Stohlgren, T.J., Quinn, J.F., Ruggiero, M. & Waggoner, G.S. (1995) Status of biotic inventories in US national parks. *Biological Conservation*, **71**, 97–106.

Termansen, M., Mcclean, C.J. & Preston, C.D. (2006) The use of genetic algorithms and Bayesian classification to model species distributions. *Ecological Modelling*, **192**, 410–424.

Thuiller, W., Araujo, M.B. & Lavorel, S. (2003) Generalized models vs. classification tree analysis: predicting spatial distributions of plant species at different scales. *Journal of Vegetation Science*, **14**, 669–680.

Wiser, S.K., Peet, R.K. & White, P.S. (1998) Prediction of rare-plant occurrence: a southern appalachian example. *Ecological Applications*, **8**, 909–920.

Wisz, M.S., Hijmans, R.J., Li, J., Peterson, A.T., Graham, C.H. & Guisan, A. (2008) Effects of sample size on the performance of species distribution models. *Diversity and Distributions*, **14**, 763–773.

Wright, J.E. & Wyld, S.J. (1994) The Rattlesnake Creek Terrane, Klamath Mountains, California – an early Mesozoic volcanic arc and its basement of tectonically disrupted oceanic crust. *Geological Society of America Bulletin*, **106**, 1033–1056.

Yee, T.W. & Mitchell, N.D. (1991) Generalized additive models in plant ecology. *Journal of Vegetation Science*, **2**, 587–602.

Zacharias, M.A. & Gregr, E.J. (2005) Sensitivity and vulnerability in marine environments: an approach to identifying vulnerable marine areas. *Conservation Biology*, **19**, 86–97.

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

Editor: Janet Franklin