Working Paper 2009-04

Variation in Spatial Predictions Among Species Distribution Modeling Methods

Alexandra D. Syphard and Janet Franklin

Variation in spatial predictions among species distribution modeling methods

Syphard, Alexandra D.* and Franklin, Janet

Department of Biology, San Diego State University, San Diego, CA 92182-4614, USA; E-mail: janet@sciences.sdsu.edu; *Corresponding author; E-mail: asyphard@yahoo.com

THIS MANUSCRIPT WAS SUBMITTED TO *ECOGRAPHY* OCT 2008. AFTER REVISIONS IT WAS ACCEPTED FOR PUBLICATION SEP 2009 AND IS IN PRESS. THE FINAL PUBLISHED VERSION WITH THE TITLE "Differences in spatial predictions among species distribution modeling methods vary with species traits and environmental predictors" WILL BE AVAILABLE FROM WILEY-BLACKWELL (WHO OWNS THE COPYRIGHT) doi: 10.1111/j.1600-0587.2009.05883.x. THIS EARLIER DRAFT IS POSTED WITH PERMISSION.

Abstract

Prediction maps produced by species distribution models (SDMs) influence decision-making in resource management or designation of land in conservation planning. Many studies have compared the prediction accuracy of different SDM modeling methods, but few have quantified the similarity among prediction maps. There has also been little systematic exploration of how the relative importance of different predictor variables varies among model types. Our objective was to expand the evaluation of SDM performance for 45 plant species in southern California to better understand how map predictions vary among model types, and to explain what factors may affect spatial correspondence, including the selection and relative importance of different environmental variables. Four types of models were tested. Correlation among maps was highest between generalized linear models (GLMs) and generalized additive models (GAMs) and lowest between classification trees and GAMs or GLMs. Correlation between Random Forests (RFs) and GAMs was the same as between RFs and classification trees. Spatial correspondence among maps was influenced the most by model prediction accuracy (AUC) and species prevalence; map correspondence was highest when accuracy was high and prevalence was intermediate. Species functional type and the selection of climate variables also influenced map correspondence. For most (but not all) species, climate variables were more important than terrain or soil in predicting their distributions. Environmental variable selection varied according to modeling method, but the largest differences were between RFs and GLMs or GAMs. Although prediction accuracy was equal for GLMs, GAMs, and RFs, the differences in spatial predictions suggest that it may be important to evaluate the results of more than one model to estimate a range of spatial uncertainty before making planning decisions based on map outputs. This may be particularly important if models have low accuracy or if species prevalence is not intermediate.

Introduction

Spatial prediction of species' geographic distributions has become a fundamental component of conservation planning, resource management, and environmental decision-making. Therefore, methodological issues related to species distribution models (SDMs) have been the focus of much discussion in the recent scientific literature. SDMs are quantitative, predictive models of the species-environment relationship that correlate observations of species occurrence or abundance with mapped environmental variables to make spatial predictions of habitat suitability or species occurrence (Franklin 1995, Guisan and Zimmermann 2000, Scott, et al. 2002, Guisan, et al. 2006). The methods are based on the assumption that species' distributions are correlated with environmental gradients represented by landscape variables that are distally or proximally related to their physiological tolerances or resource requirements, and thus, their realized niches (Austin 2002).

In recent years, a growing number of modeling methods has been applied to improve the performance and ecological validity of SDM, and the different approaches vary in terms of their complexity, assumptions, data requirements, and ease of use. For example, some newer statistical learning methods have been adopted because they are better than classical statistical methods at capturing the complex, nonlinear relationships between response variables and multiple predictors (Hastie, et al. 2001) that characterize species' responses to their environment (Austin 2002). Other methods have been applied because they are better suited to the unique characteristics of presence-only data (Stockwell and Peters 1999, Phillips, et al. 2006).

The recent SDM literature has emphasized comparison of these different model types to better understand their relative differences in performance (Bio, et al. 1998, Franklin 1998, Moisen and Frescino 2002, Segurado and Araújo 2004, Elith, et al. 2006, Maggini, et al. 2006, Guisan, et al. 2007). The majority of these comparisons have focused on prediction accuracy as a measure of model performance, in which the models are developed and the predictions are evaluated using either a single data set or two independent data sets, and one or more standard metrics are applied. For categorical prediction accuracy ("threshold-dependent"), common metrics include Kappa, Sensitivity, or Specificity. Alternatively, the area under the curve (AUC) of receiver-operating characteristic (ROC) plots (Fielding and Bell 1997) is a particularly useful metric for model comparison because it avoids the need to choose a threshold probability that separates "suitable" from "unsuitable" (or presence from absence) (i.e., it is "threshold-independent"). The AUC is also widely used because it describes the overall ability of the model to discriminate between two cases (but see Lobo, et al. 2008).

Although metrics like AUC are important components of model performance evaluation, there has been less emphasis in the literature on other methods of comparing and evaluating models. In particular, few studies have quantified the similarity among maps predicted by different model types. Yet, in many applications, the maps produced by SDMs are the key outputs that influence decision-making or designation of land, for example, for nature reserves (Mladenoff, et al. 1995, Johnson, et al. 2004). Therefore, measuring the amount of map overlap among predictions may provide important information about the strengths and limitations of different model types that may not be apparent from global measures such as AUC. For example, models that demonstrate equally high accuracy when assessed with test data could yield incongruent maps because the models use different assumptions, algorithms, and parameterizations.

Some studies have incorporated qualitative comparisons of predictive distribution maps resulting from different modeling methods and discussed their differences in the context of extent-based (non-spatial) accuracy measures, e.g., the tendency of some methods to over-predict or under-predict distributions (Loiselle, et al. 2003, Elith, et al. 2006). However, they did not quantify the spatial congruence of different predictions. Others have focused on the large variation among projections of species' future ranges under various climate change scenarios from different models (Thuiller, et al. 2004, Araújo, et al. 2005, Araújo and New 2007), or for invasive species introduced into new regions (Crossman and Bass 2008, Kelly, et al. 2008), but with an emphasis on using ensemble forecasting (i.e., combining predictions across multiple models) to address the spatial uncertainty associated with these projections.

The studies that have quantitatively compared prediction maps from different models have shown that spatial predictions may vary considerably depending on the method and other modeling decisions; however, the authors did not statistically relate spatial correspondence among predictive maps to potential explanatory variables. Prasad et al. (2006) concluded that maps produced using ensemble statistical learning methods (e.g., Random Forests), were more similar to each other (and more realistic) than to those produced using single models. In another study it was shown that global statistical models, such as generalized linear models (GLMs) and generalized additive models (GAMs), produced maps that were more similar to each other than they were to local models, that is, statistical methods that allow model parameters to vary spatially (Osborne and Suarez-Seoane 2007). That study used a partitioned Kappa statistic (Pontius 2000) to compare predictive maps. Johnson and Gillingham (2005), who developed four SDMs for caribou from presence-only observations, suggested that discrepancies in predicted maps of ranked habitat suitability may have been due to differences in the predictor variable sets used to build the models. Hernandez et al. (2006) showed, through quantitative map comparison using the Kappa statistic, that the spatial prediction of suitable habitat varied depending on the number of observations available to train the model.

Thuiller et al. (2004) proposed that discrepancies in model projections may be related to differences in the ways that model types make predictions under different environmental conditions. However, as with comparison of prediction maps, there has been little systematic exploration of environmental variable selection, and the relative importance of different predictor variables, among model types. Peterson and Nakazawa (2008) showed that, when using one model type (GARP), spatial predictions of native and introduced distributions of fire ants were sensitive to the environmental data sets used to develop the models. Considering this potential influence that different environmental data sets may have on spatial predictions, the authors called for further research on the topic.

The selection of environmental predictor variables (and the maps that represent them) in SDM is often a function of the scale of the analysis; but in general, the predictors describing the physical environment often fall into three classes: 1) climate, 2) terrain, and/or 3) substrate or landform (Franklin 1995). It is widely acknowledged that climate is a primary factor controlling plant species distributions due to controls over light, moisture, temperature and nutrient regimes (Mackey and Lindenmayer 2001), particularly at broad biogeographical scales (Busby 1986, Woodward and Williams 1987). While the predictive power of SDMs at broad scales may not be substantially improved by including variables other than climate (Thuiller, et al. 2004), terrain and geological variables that are related to direct and resource gradients may be more important at finer, landscape scales (Franklin 1995); in many cases, a combination of climate and edaphic

factors may produce the best models (Iverson and Prasad 1998). One recent meta-analysis found that those models that included environmental predictors from multiple, hierarchical scales yielded the most accurate predictions (Meyer and Thuiller 2006).While many individual SDM studies describe the correlations between predictors and species occurrence, e.g., the relative importance of different predictors, there is still little in the way of general guidelines about the relative importance of, e.g., climate, terrain and edaphic variables in different models and for different taxa. Also, just as there has been little exploration of how map predictions or selection of environmental variables may vary among model types, few studies have explored whether spatial correspondence of maps derived from different modeling methods may vary based on the relative importance of different environmental predictors.

Our objective in this study was to expand the evaluation of SDM model performance to better understand how mapped predictions may vary among model types, and to explain what factors may affect spatial correspondence. Furthermore, we evaluated the selection and relative importance of different environmental variables used to predict plant species distribution for 45 species in southern California using four types of models.

We asked these questions:

1). Do different SDM modeling methods produce similar spatial predictions?

We expected map correlation to be highest between similar model types, e.g., between those that used supervised, machine learning methods (classification trees, CT, and Random Forests, RF) and those model types that are extensions of linear multiple regression models (GLMs and GAMs).

2). How does the correlation of spatial predictions from different models vary in relation to prediction accuracy, species' prevalence, species' functional type, or type of environmental variables in the model?

We expected spatial correspondence among maps to be highest when models had greater prediction accuracy. We also expected species that occurred over smaller extents on the map to have better map correspondence because prevalence (the proportion of species' presences in the training data) is often negatively related to performance (Stockwell and Peterson 2002, Segurado and Araújo 2004, Luoto, et al. 2005, Elith, et al. 2006, McPherson and Jetz 2007). Because model accuracy was found to be strongly a function of plant functional type for the same study area (Syphard and Franklin in review), we also expected map correlation to vary among functional types. Finally, we expected higher map correspondence to occur when climate variables were selected as important because climate varies more slowly over space than terrain or soil variables.

3). Do different modeling methods select for different types of environmental variables?

Overall, we expected climate to be more important than terrain and soil for all methods due to climate's direct effect on plant species' requirements for or tolerance to heat, moisture, and light regimes. We also expected RF and CTs to select soil order (a categorical variable) more than GLMs or GAMs because categorical predictors are well handled by decision-tree methods (Breiman et al. 1984). We compared the selection and importance of soil order to three different continuous soil variables to determine whether those variables that should have a more direct physiological influence on plant species (continuous variables) would better explain their

distributions. This comparison of soil variables could be useful to modelers who only have access to certain types of soil data.

Methods

The species' distribution models examined in this study were developed as part of a larger project and the study area, species data, environmental data, and modeling methods are described in detail elsewhere (Syphard and Franklin in review). They will be summarized briefly here.

Study area and species data

Species data were developed for 45 plants typical of the Chaparral and Sage Scrub shrubland plant communities that dominate the foothills and coastal plain of southern California (Westman 1981, Schoenherr 1992, Hickman 1993, Keeley 2000) and represent a range of plant functional types. Species locations were acquired for 1,471 southern California shrubland locations (Taylor 2004) from a database (http://vtm.berkeley.edu/) of vegetation plots surveyed in the 1930s (Wieslander 1935, Kelly, et al. 2005, Barbour, et al. 2007). These species were found in at least 30 plots (prevalence > 0.02).

Environmental predictors

We used eight climate, terrain, and soil variables as predictors. Climate variables used were mean annual precipitation, mean minimum January temperature, and mean maximum July temperature interpolated to 1-km grids from 1966 to 1995 climate station data (Franklin, et al. 2001). Terrain-distributed solar radiation (Dubayah and Rich 1995) was modeled from U.S. Geological Survey 30-m resolution digital elevation model (DEMs) using the Solar Analyst 1.0 extension for ArcView™ (ESRI, Redlands, CA, USA) Geographic Information System (GIS). Daily insolation was calculated for two single days, the summer and winter solstice (using site latitude of N 33°, sky size of 200, and 0.2 clear sky irradiance) and used to represent the seasonal extremes of radiation on the landscape. The Topographic Moisture Index (TMI) represents relative soil moisture availability based on upslope catchment area and slope, which were derived from the DEM (Moore, et al. 1991, Wilson and Gallant 2000). We created a grid of soil order, a categorical variable, using the California State Soil Geographic Database (STATSGO). We also evaluated three continuous soil variables that are known to affect plant species distributions in the region: available water capacity (cm cm^{-1}), soil depth (m), and pH. These soil variables were also derived from the STATSGO GIS soil database (1: 250,000 scale) and a table from the Environmental Protection Agency (EPA) that described the map unit-level characteristics (Hannah, et al. 2008).

Species' Distribution models

We developed four models for each species using the following methods: generalized linear models (GLMs), generalized additive models (GAMs), classification trees (CTs), and Random Forests (RFs). GLMs in the form of logistic regression models are commonly used in species distribution modeling with species' presence/absence data (Guisan, et al. 2002). Although

GLMs allow for non-linear relationships to be accommodated using polynomial terms, they are nevertheless parametric models with distributions that do not always reflect complex species responses to the environment (Austin 2002, Austin, et al. 2006). GAMs (Yee and Mitchell 1991) have been widely used in species distribution modeling as an alternative to GLMs (Lehmann, et al. 2002) because global regression coefficients are replaced by local smoothing functions, allowing the structure of the data to determine the shape of the species response curves.

CTs are supervised classifiers that develop rules, based on binary recursive partitioning, that can be used to classify new observations (Breiman, et al. 1984). CTs iteratively split a full data set into partitions and evaluate how well the rules that determine these splits can separate the data into homogeneous classes. Typically, CTs are partitioned until a split no longer achieves a certain level of homogeneity, and then they are "pruned" back so that the model does not over-fit the data and can provide robust predictions for new data. Classification trees easily handle categorical predictors and interactions between variables (which do not have to be specified a priori) (De'ath and Fabricius 2000). On the other hand, CTs can be unstable, that is, they may produce very different models if the inputs are slightly varied (Prasad, et al. 2006). A newer ensemble modeling method, RFs, overcomes this instability by developing many (hundreds or thousands of) tree models using random subsets of the cases and the predictor variables and then averaging the predictions (Breiman 2001). Estimates of model error and variable importance for RF models are estimated via bootstrapping (Cutler, et al. 2007).

Based on exploratory data analysis (Syphard and Franklin in review), both linear and quadratic relationships were evaluated for all the continuous variables in the GLMs and we used three target degrees of freedom for smoothing splines in the GAMs. Backward stepwise variable selection has frequently been used in SDM (Wintle, et al. 2005) and was used here to provide a consistent and automated approach for selecting final GLMs and GAMs for all species, in spite of the acknowledged limitations of this approach (James and McCulloch 1990). Predictors were entered in the following order: climate, terrain, then soil variables, based on their relative importance determined in preliminary analyses. GLMs were further refined by manually removing quadratic terms if their coefficients were positive, e.g., if the response curve was inverted. Although a response curve can theoretically be bimodal in the presence of competition (Austin and Smith 1989), we considered this fitted form (increasing probability of occurrence at extremely high and low values of a predictor) to be a poor approximation of a bimodal response, and one that produced predictions that were ecologically unrealistic (Austin 2002). Thus, we only retained the linear term for that predictor if it remained significant.

Spatial autocorrelation (SA) of model residuals was tested for the GLMs because, among the model types used in this study, these global, parametric models are most susceptible to misspecification in the face of autocorrelation (Miller, et al. 2007). Moran's I (Moran 1948) was calculated for lag = 4000m. The distribution of nearest neighbor distances among the vegetation plots was 210-13,800 m (median 1600 m). Ninety percent of the plots had a nearest neighbor within 4000 m, and so 4000 m was examined as the lag distance. Monte Carlo simulation (1000 replicates) was used to estimate the significance of Moran's I because the residuals from a logistic regression are not normally distributed.

Full classification trees were built for each species and then pruned using an algorithm that automatically selected the complexity parameter associated with the smallest cross-validated error. If this algorithm selected only one split, we increased the number of splits to two so our pruned tree would include at least two decision rules. For Random Forests models, we averaged

the predictions from 500 trees. We evaluated three randomly selected variables for each tree based on the suggestion by Breiman (2001), that the square root of the number of variables gives optimum results.

The performance evaluation measure that we used to describe SDM prediction accuracy for each model was the area under the curve (AUC) of the receiver operating characteristic (ROC) plot (Hanley and McNeil 1982). ROC plots show the true positive predictions versus false positive predictions for all possible threshold values. Therefore, the AUC (ranging from 0 to 1) represents the probability that, for a randomly selected set of observations, the model prediction for a presence observation will be higher than the prediction for an absence observation. Although prediction accuracy as measured by AUC is only one measure of model performance (Rykiel 1996, Morrison, et al. 1998), often we lack adequate knowledge or data to evaluate other measures, such as correct selection of predictor variables and characterization of response curves (e.g., Austin, et al. 2006). Therefore, it is common to use a measure of prediction accuracy as a performance metric when comparing the relative performance of different species distribution models (e.g., Segurado and Araújo 2004, Elith, et al. 2006, and many other studies).

Bootstrapping was carried out to estimate AUC for GLM and GAM models (Wintle, et al. 2005). We created 500 bootstrapped models by iteratively resampling and partitioning the data so that some data were used to train the models and some were used to test them. In this way, the reduced prediction accuracy expected when a model is confronted with new data could be estimated. To calculate prediction accuracy with classification trees, we used 15-fold cross-validation using the same number of splits for pruning all cross-validated models. We calculated the average AUC based on the results of the cross-validation. To calculate the AUC for RF, we used the averaged "out-of-bag" predictions from the models.

Modeling was carried out in the R 2.7.0 statistical programming environment (R Development Core Team 2004) using the packages gam, rpart, randomForest, ROCR, spdep, yaImpute and model_functions.R (from Wintle, et al. 2005).

Ranking environmental variables

After bootstrapping for the GLMs and GAMs was complete, the model summaries listed percentages for how many times the environmental variables (and/or their polynomial terms) were used in the model, thereby providing a measure of their importance. For example, 75% for a certain variable would mean that, in the 500 models that were developed using the bootstrapping, that variable was retained 75% of the time. In Random Forests, variable importance is determined by comparing the misclassification error rate of a tree with the error rate that occurs if the values of a predictor variable are randomly permuted (Cutler, et al. 2007). The actual metric that we evaluated was the decrease in accuracy (i.e., after permuting the variable) averaged for the 500 model replicates. We did not assess variable importance for the classification tree models because we did not perform any bootstrapping or model averaging for this method.

Because measures of variable importance are calculated differently in Random Forests than in GLMs and GAMs, we developed a ranking system so we could compare environmental selection among the different model types. For each species in each model type (and for each type of soil variable), we evaluated all environmental variables and ranked them from 1 (most

important) to 8 or 10 depending on the soil variable(s). If two variables had the same importance, we assigned them both the same rank and then proceeded to rank the rest of the variables based on the order they would be in if there were no tie. We also averaged the ranks together for climate, terrain, and soil variables for some analyses.

Generating prediction maps

To specify quadratic relationships in our bootstrapped GLM models, we specified second-order polynomial terms using the poly() function in R. However, to create prediction maps, we re-estimated the GLMs using Identity() so that model estimates would be in the same units as the environmental predictor variables, which we needed for spatial extrapolation. To create prediction maps for all of our models, we used the R package yaImpute, version 1.0-3. The AsciiGridPredict() command in the yaImpute package works by applying the predict function for any model to every cell in the study area using ascii grid maps of the environmental predictor variables as input.

Correlation among maps

In previous studies, spatial overlap in predictions has been estimated using Kappa or Spearman rank correlations that are appropriate for categorical maps (for example, Prasad, et al. 2006, Termansen, et al. 2006). Because all of the methods we used generated a likelihood of species' presence on a scale of 0-1, we used a Pearson's correlation coefficient to calculate to correlation between prediction maps for each species, pairwise between models (e.g., Termansen, et al. 2006).

Analysis

After calculating pairwise map correlations to evaluate differences in spatial correspondence among model types, we averaged correlations among all model types for each species to use as the dependent variable in a regression analysis. We first developed simple regression models for each explanatory variable to explore the effects of model accuracy, species prevalence, species functional type, and environmental variable importance on map correlation.

To estimate the effect of model accuracy, we averaged the AUC from the predictions of all model types for each species to use as the predictor variable. Species prevalence was calculated as the proportion of plots in which the species was present. We developed the species functional type classification (Table 1)based on natural groupings of species' life form, demographic attributes, and fire response strategy (details in Syphard and Franklin in review). For the environmental variables, we used the average importance rank for the climate, terrain, and soil variables (that were developed from GLMs, GAMs, and RFs only). Because we developed models using two different types of soil variables, we also performed the regression analyses separately for the different soil data types (i.e., there were two regression models for every predictor variable we evaluated).

After developing the simple regression analyses, we estimated two multiple regression models, one for each set of models using different types of soil variables. We entered the

explanatory variables into the model in the order of the amount of variation they explained in the simple models, and we only retained those variables that were significant at $P \leq 0.05$.

Results

There was significant ($P < 0.05$) positive spatial autocorrelation (SA) in the residuals of GLMs for only 7 of 45, or less than 16%, of the species. There was no apparent relationship between SA in the residuals and species prevalence, model performance, or species traits. Because so few models showed significant SA in the residuals, and because the emphasis of this study was on prediction and not estimation of parameters, we did not treat SA further (e.g., by fitting a spatial autoregressive error model).

Mean correlation among prediction maps varied according to the method used to develop the models, but there was substantial variability in the correlations among species (Figure 1). The lowest correlation between maps occurred between CTs and GAMs or GLMs, and the highest correlation occurred between GLMs and GAMs. Correlation between RF and CTs was similar to the correlation between RF and GLMs or GAMs. Overall, the mean correlation between prediction maps was slightly higher for those models developed using the continuous soil variables.

Although the mean prediction accuracy of classification trees (AUC 0.69) was significantly lower than that of the other three methods (AUC 0.78 – 0.79) (Syphard and Franklin in review), there was little difference in prediction accuracy between those models developed with continuous soil versus those developed using soil order, regardless of model type (Figure 2). When different model types had reasonable accuracy (AUCs generally above 0.75), they predicted species to be distributed in the same general locations of the study area (Figure 3B, C). When model accuracy and species prevalence were both low, Random Forests predicted distributions to occur over a larger extent and to be more dispersed than GLMs or GAMs (Figure 3A, D).

Species prevalence and model accuracy explained more variation in map correlation than the other variables, although functional type and the importance of climate in model selection were also significant (Table 2). As was the case with AUC, the influence of the variables on map correlation did not vary depending on the type of soil variables used to develop the models. The relationship between map correlation and prediction accuracy was positive and linear, but the relationship with prevalence was quadratic. Prediction maps had the lowest correspondence when both species prevalence and model accuracy were lowest. With higher prediction accuracy and species prevalence, map correlation was also much higher; but for the species with the highest prevalence (< 0.2), the relationship with map correlation was negative (Figure 3). Map correlation was higher for species that experience fire-cued germination from a dormant seed bank (facultative seeder shrubs and obligate seeder shrubs) and lowest for perennial herbs and subshrubs that respond to fire through vegetative propagation (Figure 4). Although the importance of terrain and soil variables in the models did not influence the correspondence among prediction maps, those models for which climate was most important produced maps that had better map correlation.

When all of the significant explanatory variables (prevalence, AUC, functional type, and climate) were included in multiple regression models, only species prevalence and AUC

remained significant predictors of map correlation (Table 3). Although patterns of the results were similar for models developed using both types of soil variables, the multiple regression model (with prevalence and AUC only) based on models using soil order explained more variation ($R^2 = 0.76$) than the multiple regression based on models using continuous soil variables ($R^2 = 0.55$).

When averaged together across all model types (GLMs, GAMs and RFs), the climate variables were more important in the SDMs than terrain or soil, which were both similar in their relative importance (Figure 5). The relative importance of different variables when evaluated individually, however, was different depending on the model type (Figure 6). For the GLMs and the GAMs, regardless of the type of soil variable in the models, the three climate variables had nearly equal importance, which was higher than the importance than the other variables. For all three model types and both soil types, summer radiation was more important than the other terrain variables; and the relative difference was substantial for the GLMs and GAMs. Whereas the importance of the other three terrain variables was similar for GLMs and GAMs (although TMI was generally the lowest), winter radiation and (especially) TMI were substantially lower than slope in Random Forests. Soil order was substantially more important than the terrain variables for GLMs and GAMs. However, the continuous soil variables were similar in importance to terrain for the GLMs, but higher than terrain for the GAMs. Differences in importance between terrain and soil variables were insubstantial for Random Forests.

Discussion

The use of metrics such as AUC has become standard practice in evaluating the performance of species distribution models. AUC is a very useful measure of comparative model performance because it is threshold independent, but any measure of predictive performance is limited by the data available for model evaluation. The results of this study reinforce the notion that it is also important to consider additional criteria in model evaluation, depending on the objective of the application (e.g., Austin, et al. 2006, Hernandez, et al. 2006). If prediction maps will be used to make conservation or resource management decisions, the spatial distribution of model uncertainty may be particularly important. While correlation among map predictions in our study significantly improved with more accurate models, there were other factors that strongly affected spatial correspondence among predictions, especially species prevalence. Map correlation also varied depending on the modeling method used, species functional traits, and type of environmental variables that were important in the models. The effect of these factors on model performance should therefore be taken into consideration for any SDM application.

With regards to modeling methods, classification trees overall had lower accuracy than the other three methods (Syphard and Franklin in review), which is likely why the pairwise comparisons of map correlation were lowest with the CTs. However, because Random Forests is essentially developed using an ensemble of trees, we were surprised that the correlation between RFs and CTs was as low as the correlation between RFs and GLMs or GAMs. The relatively low accuracy and low map correlation using single CTs is consistent with other studies that found them to be somewhat unstable (Benito Garzón, et al. 2006, Prasad, et al. 2006). While there are some features of CTs that may be more desirable than RFs (e.g., ability to visualize the classification rules portrayed in single trees), RFs may be a better choice for conservation practitioners trying to create the most robust predictive maps.

Although prediction accuracy was highest with Random Forests, the spatial correspondence in predictions was lower between RF and GAMs or GLMs than it was between GAMs and GLMs. Although the Random Forests models predicted greater extents of suitable habitat than GLMs or GAMs for species with low prevalence, it is unknown, based on the data we had for model evaluation, whether this low map correlation was due to true errors of commission. Alternatively, the greater predicted extent (i.e., analogous environmental conditions) may have represented areas that were truly suitable for the species, and the species may not have been sampled in that area, or it may have previously occupied the area. Although we calculated our AUCs using bootstrapping or cross-validation (iteratively sub- sampling the data to test the models on independent observations), a unique characteristic of spatial prediction is that, when creating a prediction map from a model, most of the cells in the map can essentially be considered independent observations. However, there are no data to confirm whether the species is actually absent or present in most cells. This issue has become particularly challenging for the use of SDMs in climate change modeling, where there are no independent observations to evaluate prediction accuracy (Heikkinen, et al. 2006). Interestingly, however, when comparing different modeling methods for predicting future distributions under climate change, Prasad et al. (2006) felt that ensemble methods, including Random Forests, outperformed other modeling methods in spatial prediction.

While prevalence strongly affected map correlation in this study (with the highest correlation at intermediate prevalence), other studies have shown that prevalence may also be significantly related to model performance. In some cases, prediction accuracy was higher when prevalence was low (e.g., Segurado and Araújo 2004, Elith, et al. 2006, Hernandez, et al. 2006, Syphard and Franklin in review), but McPherson et al. (2004) found that models performed best when prevalence was intermediate. In this study, prevalence had no significant effect on AUC ($P = 0.46$); therefore, the effect of prevalence on map correlation can not be directly attributed to the effect of prevalence on model performance. However, the relationship between species prevalence and map correlation may be partly related to the way that different models approximate species-response functions and how those response functions translate into predicted probabilities. If models vary in the way that their distributions of predicted probabilities reflect the prevalence of the species (based on the number of sample locations used to build the models), then presumably these model differences would be manifested more apparently in map correlations if prevalence were either very low or very high.

Another reason that low prevalence may have affected map correlation is that species may have low prevalence because they are difficult to detect. One consequence of low detectability is that a species could actually be present in locations where it is predicted to be absent. This would likely affect the model's prediction accuracy, but may also affect spatial extrapolation, reflecting differences in ways that models characterize species' presences. Furthermore, if species have low prevalence, there are fewer samples to develop accurate characterization of species presence.

In this study, the relationship between plant functional type and map correlation appears to result from the observed correspondence between functional type and model performance. The functional types with the highest prediction accuracy tended to be those with high site fidelity – long-lived facultative and obligate seeders with poor dispersal and persistent seed banks (Syphard and Franklin in review). Those functional types that had higher AUC also had higher

map correlation among models. Therefore, in a multiple regression model, functional type was not selected as a significant predictor of map correlation if AUC was already in the model.

A source of uncertainty in our study is the error inherent in the historic VTM data, which could also affect spatial correspondence of predictions, particularly for those models that select terrain and/or soil. Terrain and soil variables are more heterogeneous than climate at landscape scales. Because climate varies slowly over space, there is greater certainty that those variables would be accurately calculated within the 300-m range of the VTM data that had an average positional error of ~130 m (Kelly, et al. 2008). Coarsening data resolution could potentially increase or decrease model performance such that performance may increase after smoothing errors in environmental or species data, but performance may decrease if there is a lack of spatial matching between species observations and their associated environmental predictors (Guisan, et al. 2007).

The primary influence of environmental variable selection on the spatial correspondence of predicted distributions was related to climate. In other words, the more that a species could be modeled through climate variables alone, the more likely the predictions were likely to overlay. Some SDMs only use climate variables (e.g., bioclimatic envelope models, Huntley, et al. 2004, Kueppers, et al. 2005, Heikkinen, et al. 2006), and our results support the notion that climate tends to be the overriding driver of distributional patterns at a landscape scale for plants in southern California.

However, there was substantial variation in the importance of different environmental predictors among species; and terrain and soil were also important for explaining that variation. One potential reason that, for some species, terrain or soil variables were selected over climate variables is that their climatic range may have been greater than that which was in the study area. Thus, the model(s) sought finer-scaled variables to explain what aspects of species distribution patterns the climate variables were missing. Therefore, while climate variables do explain more variation in distribution patterns than terrain or soil, we suggest that both terrain and soil should be considered in any SDM study for plants at a landscape scale. As we have already noted, a meta-analysis found that models that included environmental predictors from multiple scales showed the highest predictive performance (Meyer and Thuiller 2006). Further, while there are perceived trade-offs between model parsimony and model accuracy, Drake et al. (2006) found that the most accurate models were those that included the largest number of environmental predictors, even after optimizing the models to avoid overfitting.

Map correlation was consistently higher with models that used continuous soil variables instead of soil order, which could be partly because the resolution of the continuous variables was coarse (1-km, the same as the climate variables). Unlike the maps developed from models that used the slowly varying continuous variables, maps developed with soil order may have been sensitive to resolution because of the sharp transition between boundaries (Guisan and Zimmermann 2000). Although correlation was higher with the continuous variables, model accuracy was very similar for models using either type of soil variable. Therefore, depending on the application or objective of the modeling study, soil order may be just as useful as several continuous soil variables that may have a more direct physiological influence on different species. In fact, soil order, which reflects the soil development processes acting at a site, was slightly more important in environmental selection than continuous soil. Although we evaluated several variables, differences among soil orders may indirectly distinguish between certain variables that are more influential on the species distributions in the region.

In conclusion, average model performance (measured by AUC) was essentially the same for the GLMs, GAMs, and Random Forests models (although CTs had lower accuracy). Yet, despite these similar accuracies, our results show that prediction maps and the environmental variables selected varied substantially among the different methods. When the goal of the SDM study is to create prediction maps, we suggest that the model evaluation process should go beyond global accuracy measures and include some evaluation of the spatial pattern of predictions.

In the context of climate change modeling and other applications of SDM, some authors have suggested averaging model predictions due to high variability in their projections (e.g., Thuiller, et al. 2004, Marmion, et al. 2008). However, Araujo et al. (2005) cautioned that accuracy will most likely increase only if better models are considered as opposed to more models. There are a number of approaches to ensemble forecasting in SDM and other modeling fields in addition to model averaging or consensus methods (Araújo and New 2007). It might be prudent to evaluate spatial predictions from model types that tend to be different, such as GLMs vs. Random Forests, to determine a bracket of uncertainty. This might be particularly important for species that have either very low or very high prevalence. Nevertheless, it is important to consider that map correlation in this study was a function of those models that we selected to examine. While we chose common methods used in SDM, the differences in predictions are ultimately a function of how the different models handle prediction.

Acknowledgements – This work was supported by a grant from the National Science Foundation (NSF BCS-0452389) to JF, J. Miller and R. Fisher. We thank J. Miller and R. Taylor for their contributions, and J.E. Keeley for reviewing the species traits. We also thank F. Davis, L. Hannah, D. Stoms and L. Ries, who provided soil data.

References

Araújo, M. B. and New, M. 2007. Ensemble forecasting of species distributions. - Trends in Ecology & Evolution 22: 42-47.

Araújo, M. B., Pearson, R. G., Thuiller, W. and Erhard, M. 2005. Validation of species-climate impact models under climate change. - Global Change Biology 11: 1504-1513.

Araujo, M. B., Thuiller, W., Williams, P. H. and Reginster, I. 2005. Downscaling European species atlas distributions to a finer resolution: implications for conservation planning. - Global Ecology and Biogeography 14: 17-30.

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., Belbin, L., Meyers, J. A., Doherty, M. D. and Luoto, M. 2006. Evaluation of statistical models used for predicting plant species distributions: Role of artificial data and theory. - Ecological Modelling 199: 197-216.

Austin, M. P. and Smith, T. M. 1989. A new model for the continuum concept. - Vegetatio 83: 35-47.

Barbour, M., Keeler-Wolf, T. and Schoenherr, A. A. (eds.). 2007. Terrestrial Vegetation of California. - University of California Press.

Benito Garzón, M., Blazek, R., Neteler, M., Sánchez de Dios, R., Ollero, H. S. and Furlanello, C. 2006. Predicting habitat suitability with machine learning models: the potential area of *Pinus sylvestris* L. in the Iberian Peninsula. - Ecological Modelling 197: 383-393.

Bio, A., Alkemade, R. and Barendregt, A. 1998. Determining alternative models for vegetation response analysis: a non-parametric approach. - Journal of Vegetation Science 9: 5-16.

Breiman, L. 2001. Random forests. - Machine Learning 45: 15-32.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. 1984. Classification and regression trees. - Wadsworth.

Busby, J. R. 1986. A biogeoclimatic analysis of *Nothofagus cunninghamii* (Hook.) Oerst. in southeastern Australia. - Australian Journal of Ecology 11: 1-7.

Crossman, N. D. and Bass, D. A. 2008. Application of common predictive habitat techniques for post-border weed risk management. - Diversity and Distributions 14: 213-224.

Cutler, D. R., Edwards Jr., T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J. and Lawler, J. J. 2007. Random forests for classification in ecology. - Ecology 88: 2783-2792.

De'ath, G. and Fabricius, K. E. 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. - Ecology 81: 3178-3192.

Drake, J. M., Randin, C. and Guisan, A. 2006. Modelling ecological niches with support vector machines. - Journal of Applied Ecology 43: 424-432.

Dubayah, R. and Rich, P. M. 1995. Topographic solar radiation for GIS. - International Journal of Geographic Information Systems 9: 405-419.

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. and Zimmermann, N. E. 2006. Novel methods improve prediction of species' distributions from occurrence data. - Ecography 29: 129-151.

Fielding, A. and Bell, J. 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 modeling of biospatial patterns in relation to environmental gradients. - Progress in Physical Geography 19: 474-499.

Franklin, J. 1998. Predicting the distribution of shrub species in southern California from climate and terrain-derived variables. - Journal of Vegetation Science 9: 733-748.

Franklin, J., Keeler-Wolf, T., Thomas, K., Shaari, D. A., Stine, P., Michaelsen, J. and Miller, J. 2001. Stratified sampling for field survey of environmental gradients in the Mojave Desert Ecoregion. - In: Millington, A., Walsh, S. and Osborne, P. (eds.), GIS and remote sensing applications in biogeography and ecology. Kluwer Academic Publishers, pp. 229-253.

Guisan, A., Edwards, T. C., Jr. and Hastie, T. 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. - Ecological Modelling 157: 89-100.

Guisan, A., Lehmann, A., Ferrier, S., Austin, M., Overton, J. M. C., Aspinall, R. and Hastie, T. 2006. Making better biogeographical predictions of species' distributions. - Journal of Applied Ecology 43: 386-392.

Guisan, A. and Zimmermann, N. E. 2000. Predictive habitat distribution models in ecology. - Ecological Modelling 135: 147-186.

Guisan, A., Zimmermann, N. E., Elith, J., Graham, C. H., Phillips, S. and Peterson, A. T. 2007. What matters for predicting the occurrences of trees: Techniques, data, or species' characteristics? - Ecological Monographs 77: 615-630.

Hanley, J. A. and McNeil, B. J. 1982. The meaning and use of the area under a receiver operating characteristics curve. - Radiology 143: 29-36.

Hannah, L., Midgley, G., Davies, I., Davis, F., Ries, L., Thuiller, W., Thorne, J., Seo, C., Stoms, D. and Snider, N. 2008. BioMove-Improvement and Parameterization of a Hybrid Model for the Assessment of Climate Change Impacts on the Vegetation of California. - California Energy Commission, Public Interest Energy Research Program.

Hastie, T., Tibshirani, R. and Friedman, J. 2001. The elements of statistical learning: data mining, inference and prediction. - Springer-Verlag.

Heikkinen, R. K., Luoto, M., Araujo, M. B., Virkkala, R., Thuiller, W. and Sykes, M. T. 2006. Methods and uncertainties in bioclimatic envelope modelling under climate change. - Progress in Physical Geography 30: 751-777.

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

Hickman, J. D. (ed.). 1993. The Jepson manual: higher plants of California. - University of California Press.

Huntley, B., Green, R. E., Collingham, Y. C., Hill, J. K., Willis, S. G., Bartlein, P. J., Cramer, W., Hagemeijer, W. J. M. and Thomas, C. J. 2004. The performance of models relating species geographical distributions to climate is independent of trophic level. - Ecology Letters 7: 417-426.

Iverson, L. R. and Prasad, A. M. 1998. Predicting abundance of 80 tree species following climate change in the eastern United States. - Ecological Monographs 68: 465-485.

James, F. C. and McCulloch, C. E. 1990. Multivariate analysis in ecology and systematics: Panacea or Pandora's box? - Annual Review of Ecology and Systematics 21: 129-166.

Johnson, C. J. and Gillingham, M. P. 2005. An evaluation of mapped species distribution models used for conservation planning. - Environmental Conservation 32: 117-128.

Johnson, C. J., Seip, D. R. and Boyce, M. S. 2004. A quantitative approach to conservation planning: Using resource selection functions to identify important habitats for mountain caribou. - Journal of Applied Ecology 41: 238-251.

Keeley, J. E. 2000. Chaparral. - In: Barbour, M. G. and Billings, W. D. (eds.), North American terrestrial vegetation. Cambridge University Press, pp. 204-253.

Kelly, M., Allen-Diaz, B. and Kobzina, N. 2005. Digitization of a historic dataset: The Wieslander California vegetation type mapping project. - Madrono 52: 191-201.

Kelly, M., Ueda, K. I. and Allen-Diaz, B. 2008. Considerations for ecological reconstruction of historic vegetation: Analysis of the spatial uncertainties in the California Vegetation Type Map dataset. - Plant Ecology 194: 37-49.

Kueppers, L. M., Snyder, M. A., Sloan, L. C., Zavaleta, E. S. and Fulfrost, B. 2005. Modeled regional climate change and California enaemic oak ranges. - Proceedings of the National Academy of Sciences of the United States of America 102: 16281-16286.

Lehmann, A., Overton, J. M. and Leathwick, J. R. 2002. GRASP: generalized regression analysis and spatial prediction. - Ecological Modelling 157: 189-207.

Lobo, J. M., Jimenez-Valverde, A. and Real, R. 2008. AUC: a misleading measure of the performance of predictive distribution models. - Global Ecology and Biogeography 17: 145-151.

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

Luoto, M., Poyry, J., Heikkinen, R. K. and Saarinen, K. 2005. Uncertainty of bioclimate envelope models based on the geographical distribution of species. - Global Ecology and Biogeography 14: 575-584.

Mackey, B. G. and Lindenmayer, D. B. 2001. Towards a hierarchical framework for modeling the spatial distribution of animals. - Journal of Biogeography 28: 1147-1166.

Maggini, R., Lehmann, A., Zimmermann, N. E. and Guisan, A. 2006. Improving generalized regression analysis for the spatial prediction of forest communities. - Journal of Biogeography 33: 1729-1749.

Marmion, M., Parviainen, M., Luoto, M., Heikkinen, R. K. and Thuiller, W. 2008. Evaluation of consensus methods in predictive species distribution modeling. - Diversity and Distributions.

McPherson, J. M. and Jetz, W. 2007. Effects of species' ecology on the accuracy of distribution models. - Ecography 30: 135-151.

McPherson, J. M., Jetz, W. and Rogers, D. J. 2004. The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? - Journal of Applied Ecology 41: 811-823.

Meyer, C. B. and Thuiller, W. 2006. Accuracy of resource selection functions across spatial scales. - Diversity and Distributions 12: 288-297.

Miller, J., Franklin, J. and Aspinall, R. 2007. Incorporating spatial dependence in predictive vegetation models. - Ecological Modelling 202: 225-242.

Mladenoff, D. J., Sickley, T. A., Haight, R. G. and Wydeven, A. P. 1995. A regional landscape analysis and prediction of favorable gray wolf habitat in the northern great lakes region. - Conservation Biology 9: 279-294.

Moisen, G. G. and Frescino, T. S. 2002. Comparing five modelling techniques for predicting forest characteristics. - Ecological Modelling 157: 209-225.

Moore, I. D., Grayson, R. B. and Ladson, A. R. 1991. Digital terrain modeling: a review of hydrological, geomorphologic and biological applications. - Hydrological Processes 5: 3-30.

Moran, P. A. P. 1948. The interpretation of statistical maps. - Journal of the Royal Statistical Society B 10: 243-251.

Morrison, M. L., Marcot, B. G. and Mannan, R. W. 1998. Wildlife-habitat relationships: concepts and applications. - The University of Wisconsin Press.

Osborne, P. E. and Suarez-Seoane, S. 2007. Identifying core areas in a species' range using temporal suitability analysis: An example using little bustards Tetrax tetrax L. in spain. - Biodiversity and Conservation 16: 3505-3518.

Peterson, A. T. and Nakazawa, Y. 2008. Environmental data sets matter in ecological niche modelling: an example with Solenopsis invicta and Solenopsis richteri. - Global Ecology and Biogeography 17: 135-144.

Phillips, S. J., Anderson, R. P. and Schapire, R. E. 2006. Maximum entropy modeling of species geographic distributions. - Ecological Modelling 190: 231-259.

Pontius, R. G. 2000. Quantification error versus location error in comparison of categorical maps. - Photogrammetric Engineering and Remote Sensing 66: 1011-1016.

Prasad, A. M., Iverson, L. R. and Liaw, A. 2006. Newer classification and regression techniques: bagging and random forests for ecological prediction. - Ecosystems 9: 181-199.

R Development Core Team. 2004. R: A language and environment for statistical computing. - R Foundation for Statistical Computing.

Rykiel, E. J., Jr. 1996. Testing ecological models: the meaning of validation. - Ecological Modelling 90: 229-244.

Schoenherr, A. 1992. A Natural History of California. - University of California Press.

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

Segurado, P. and Araújo, M. B. 2004. An evaluation of methods for modelling species distributions. - Journal of Biogeography 31: 1555-1568.

Stockwell, D. and Peterson, A. T. 2002. Effects of sample size on accuracy of species distribution models. - Ecological Modelling 148: 1-13.

Stockwell, D. R. B. and Peters, D. P. 1999. The GARP modelling system: problems and solutions to automated spatial prediction. - International Journal of Geographical Information Science 13: 143-158.

Syphard, A. D. and Franklin, J. in review. Species' functional type affects the accuracy of species distribution models for plants in southern California. - Journal of Vegetation Science.

Taylor, R. S. 2004. A natural history of coastal sage scrub in southern California: regional floristic patterns and relations to physical geography, how it changes over time, and how well reserves represent its biodiversity. Geography Department. - University of California.

Termansen, M., McClean, C. J. and 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., Pearson, R. G., Whittaker, R. J., Brotons, L. and Lavorel, S. 2004. Biodiversity conservation - Uncertainty in predictions of extinction risk. - Nature 430.

Thuiller, W., Brotons, L., Araujo, M. B. and Lavorel, S. 2004. Effects of restricting environmental range of data to project current and future species distributions. - Ecography 27: 165-172.

Westman, W. E. 1981. Diversity relations and succession in Californian coastal sage scrub. - Ecology 62: 170-184.

Wieslander, A. E. 1935. A vegetation type map of California. - Madroño 3: 140-144.

Wilson, J. and Gallant, J. 2000. Terrain analysis: principles and applications. - John Wiley & Sons.

Wintle, B. A., Elith, J. and Potts, J. M. 2005. Fauna habitat modelling and mapping: A review and case study in the Lower Hunter Central Coast region of NSW. - Austral Ecology 30: 719-738.

Woodward, F. I. and Williams, B. G. 1987. Climate and plant distribution at global and local scales. - Vegetatio 69: 189-197.

Yee, T. W. and Mitchell, N. D. 1991. Generalized additive models in plant ecology. - Journal of Vegetation Science 2: 587-602.

Table 1. Species and functional types evaluated in southern California map overlay. Prevalence is the proportion of plots in which species was present. Ranges of correlation and AUC were derived from GLM, GAM, CT, and RF model types. Functional Types: shrubFac = facultative seeder shrub; shrubOS = obligate seeder shrub; shrubOR = obligate resprouters shrub; subshrFac = resprouting subshrub; subshrub S = post-fire seeding subshrub; perrherb = perennial herb. SO = soil order; CS = Continuous soil variables

Species Scientific Name	Functional Type	Prevalence	Range Correlation, SO	Range AUC, SO	Range Correlation, CS	Range AUC, CS
Adenostoma fasciculatum	shrubFac	0.53	0.59 - 0.82	0.73 - 0.79	0.62 - 0.91	0.73 - 0.80
Adenostoma sparsifolium	shrubFac	0.06	0.74 - 0.92	0.85 - 0.93	0.75 - 0.98	0.82 - 0.94
Arctostaphylos glauca	shrubOS	0.07	0.51 - 0.79	0.71 - 0.92	0.45 - 0.92	0.76 - 0.84
Arctostaphylos pungens	shrubOS	0.06	0.55 - 0.85	0.79 - 0.91	0.55 - 0.88	0.78 - 0.92
Arctostaphylos glandulosa	shrubOR	0.14	0.55 - 0.91	0.78 - 0.84	0.66 - 0.95	0.79 - 0.86
Artemisia californica	subshrFac	0.39	0.70 - 0.92	0.80 - 0.84	0.74 - 0.93	0.79 - 0.85
Artemisia tridentata	subshrubS	0.03	0.50 - 0.67	0.81 - 0.90	0.40 - 0.96	0.80 - 0.90
Ceanothus crassifolius	shrubOS	0.12	0.48 - 0.81	0.75 - 0.84	0.39 - 0.92	0.74 - 0.86
Ceanothus cuneatus	shrubOS	0.03	0.37 - 0.70	0.67 - 0.93	0.59 - 0.97	0.65 - 0.93
Ceanothus greggii	shrubOS	0.12	0.78 - 0.93	0.85 - 0.94	0.80 - 0.94	0.87 - 0.94
Ceanothus leucadermis	shrubFac	0.12	0.70 - 0.90	0.77 - 0.89	0.66 - 0.91	0.79 - 0.89
Ceanothus tomentosus	shrubOS	0.12	0.56 - 0.98	0.78 - 0.84	0.60 - 0.98	0.78 - 0.86
Ceanothus verrucosus	shrubOS	0.03	0.61 - 0.84	0.74 - 0.92	0.63 - 0.90	0.69 - 0.92
Cercocarpus betuloides	shrubOR	0.15	0.63 - 0.90	0.76 - 0.86	0.66 - 0.90	0.78 - 0.86
Cneoridium dumosum	shrubOR	0.03	0.44 - 0.68	0.66 - 0.84	0.34 - 0.90	0.53 - 0.83
Eriophyllum confertiflorum	perrherb	0.06	0.12 - 0.57	0.51 - 0.63	0.31 - 0.94	0.54 - 0.69
Eriodictyon crassifolium	shrubFac	0.01	0.21 - 0.63	0.55 - 0.76	0.09 - 0.71	0.53 - 0.77
Eriogonum fasciculatum	subshrFac	0.46	0.45 - 0.96	0.58 - 0.66	0.48 - 0.95	0.59 - 0.68
Galium angustifolium	perrherb	0.03	0.23 - 0.59	0.62 - 0.83	0.42 - 0.98	0.62 - 0.85
Garrya veatchii	shrubFac	0.04	0.45 - 0.70	0.76 - 0.89	0.44 - 0.74	0.78 - 0.90
Gutierrezia sarothrae	subshrubS	0.05	0.29 - 0.83	0.60 - 0.80	0.45 - 0.77	0.61 - 0.77

Species Scientific Name	Functional Type	Prevalence	Range Correlation, SO	Range AUC, SO	Range Correlation, CS	Range AUC, CS
Hazardia squarrosa	shrubOR	0.09	0.35 - 0.77	0.48 - 0.66	0.32 - 0.93	0.63 - 0.71
Heteromeles arbutifolia	shrubOR	0.12	0.53 - 0.81	0.64 - 0.77	0.52 - 0.83	0.65 - 0.77
Keckiella antirrhinoides	subshrOR	0.06	0.36 - 0.66	0.62 - 0.75	0.40 - 0.69	0.66 - 0.74
Lonicera subspicata	subshrOR	0.05	0.25 - 0.64	0.68 - 0.76	0.27 - 0.72	0.66 - 0.76
Lotus scoparius	shrubOS	0.31	0.47 - 0.82	0.56 - 0.66	0.46 - 0.88	0.61 - 0.68
Malacothamnus fasciculatus	subshrFac	0.02	0.01 - 0.51	0.52 - 0.61	0.20 - 0.80	0.56 - 0.64
Malosma laurina	shrubFac	0.3	0.78 - 0.93	0.79 - 0.83	0.79 - 0.94	0.79 - 0.82
Mimulus aurantiacus	subshrubS	0.11	0.60 - 0.83	0.60 - 0.71	0.70 - 0.97	0.65 - 0.71
Opuntia littoralis	subshrubS	0.01	0.09 - 0.59	0.78 - 0.88	0.23 - 0.84	0.79 - 0.90
Penstemon spectabilis	perrherb	0.02	0.36 - 0.67	0.72 - 0.81	0.27 - 0.86	0.73 - 0.81
Prunus ilicifolia	shrubOR	0.09	0.58 - 0.80	0.68 - 0.83	0.53 - 0.80	0.71 - 0.85
Quercus berberidifolia	shrubOR	0.37	0.77 - 0.97	0.76 - 0.81	0.80 - 0.97	0.76 - 0.82
Quercus wislizeni	shrubOR	0.04	0.52 - 0.75	0.79 - 0.93	0.56 - 0.77	0.85 - 0.93
Rhamnus ilicifolia	shrubOR	0.1	0.58 - 0.84	0.67 - 0.76	0.50 - 0.92	0.67 - 0.78
Rhamnus crocea	shrubOR	0.05	0.08 - 0.31	0.53 - 0.63	0.18 - 0.56	0.45 - 0.68
Rhus integrifolia	shrubOR	0.11	0.66 - 0.88	0.80 - 0.89	0.69 - 0.98	0.81 - 0.90
Rhus ovata	shrubFac	0.16	0.62 - 0.85	0.74 - 0.78	0.61 - 0.89	0.74 - 0.81
Salvia apiana	subshrFac	0.33	0.48 - 0.94	0.61 - 0.72	0.56 - 0.95	0.65 - 0.74
Salvia mellifera	subshrFac	0.27	0.34 - 0.86	0.69 - 0.75	0.38 - 0.98	0.71 - 0.76
Toxicodendron diversilobum	subshrOR	0.04	0.33 - 0.90	0.62 - 0.72	0.36 - 0.99	0.62 - 0.73
Trichostema lanatum	shrubFac	0.03	0.51 - 0.73	0.73 - 0.85	0.45 - 0.94	0.72 - 0.84
Viguiera laciniata	subshrOR	0.03	0.20 - 0.61	0.65 - 0.80	0.26 - 0.64	0.55 - 0.84
Xylococcus bicolor	shrubOR	0.12	0.42 - 0.83	0.73 - 0.83	0.51 - 0.96	0.74 - 0.86
Yucca whipplei	subshrOR	0.13	0.69 - 0.90	0.70 - 0.75	0.61 - 0.83	0.69 - 0.76

Table 2. Model coefficients, p-values, and R^2 for the explanatory variables in the simple regression models for map correlation in southern California.

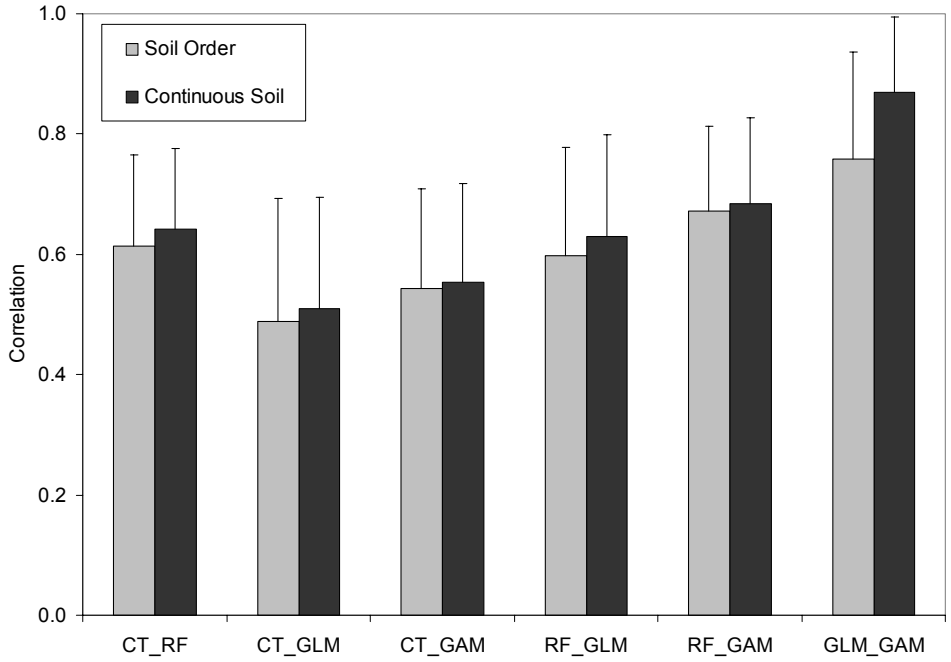
		Model Parameters		
	Variable	Coefficient	P-value	R^2
Soil Order	Prevalence	2.17	**<0.001**	0.34
	Prevalence ^2	-3.54	**0.003**	
	AUC	0.99	**<0.001**	0.34
	FunctionalType	NA	**0.049**	0.16
	Climate	0.11	**0.004**	0.16
	Terrain	-0.07	0.218	0.01
	Soil	-0.01	0.646	0
Continuous Soil	Prevalence	1.67	**0.002**	0.22
	Prevalence ^2	-2.78	**0.013**	
	AUC	0.74	**<0.001**	0.21
	FunctionalType	NA	0.062	0.14
	Climate	0.08	**0.051**	0.06
	Terrain	-0.07	0.089	0.04
	Soil	-0.01	0.868	0

Table 3. Coefficients and p-values for variables in the multiple regression models for map correlation in southern California.

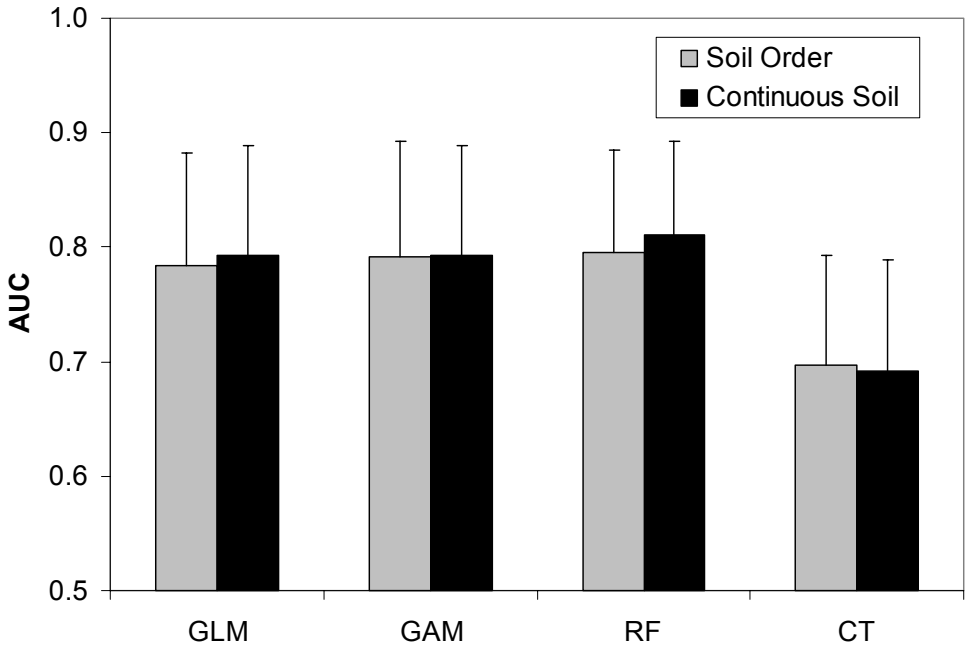
	Variable	Coefficient	p-value
Soil Order	(Intercept)	-0.51	**<0.001**
	Prevalence	2.15	**<0.001**
	Prevalence^2	-3.17	**<0.001**
	AUC	1.08	**<0.001**
			R^2- 0.76
Continuous Soil	(Intercept)	-0.505	0.13
	Prevalence	1.73	**<0.001**
	Prevalence^2	-2.59	**0.003**
	AUC	0.87	**<0.001**
			R^2- 0.55

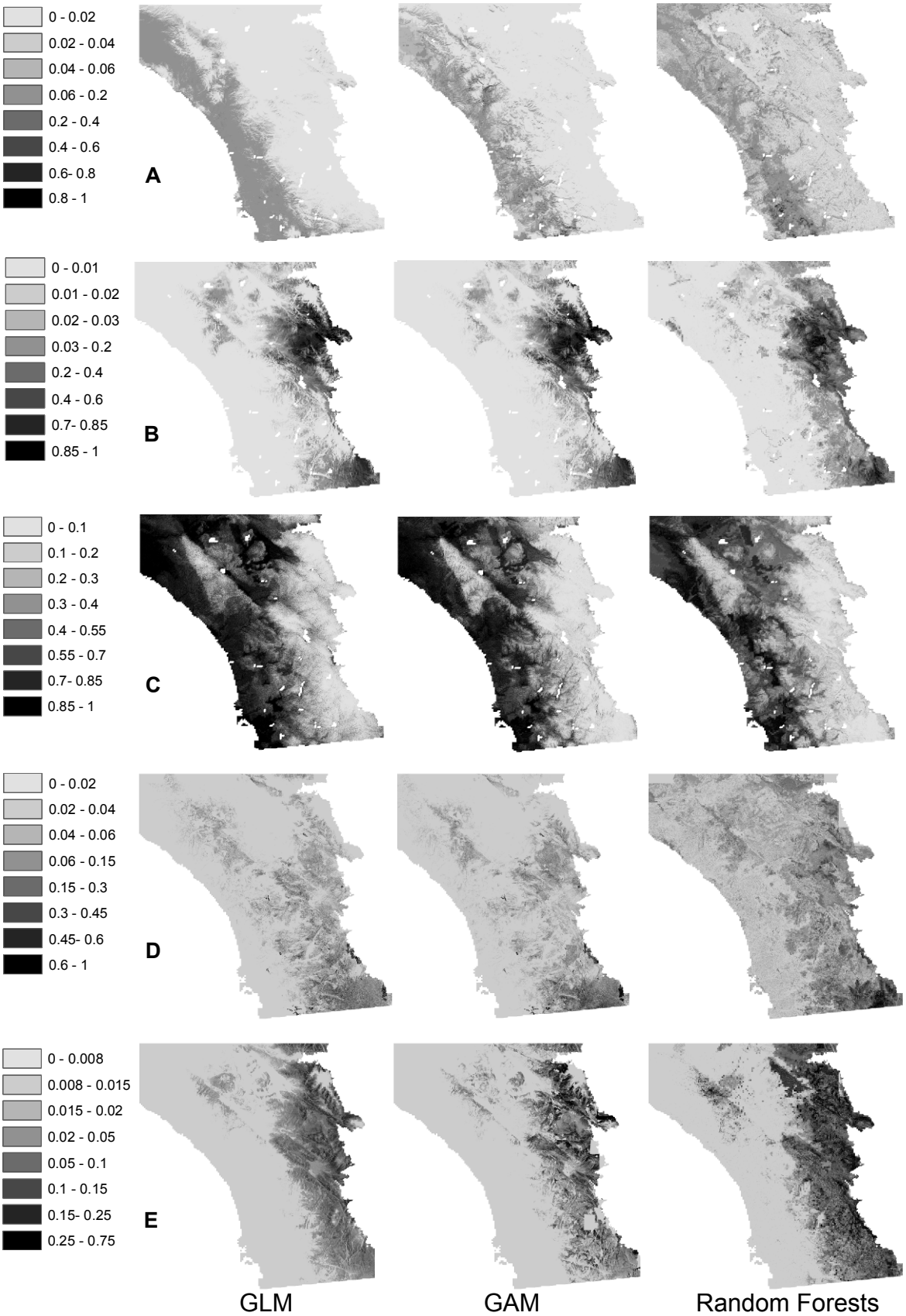
List of Figures

1


Soil Order
Continuous Soil
Correlation
1.0
0.8
0.6
0.4
0.2
0.0
CT_RF
CT_GLM
CT_GAM
RF_GLM
RF_GAM
GLM_GAM

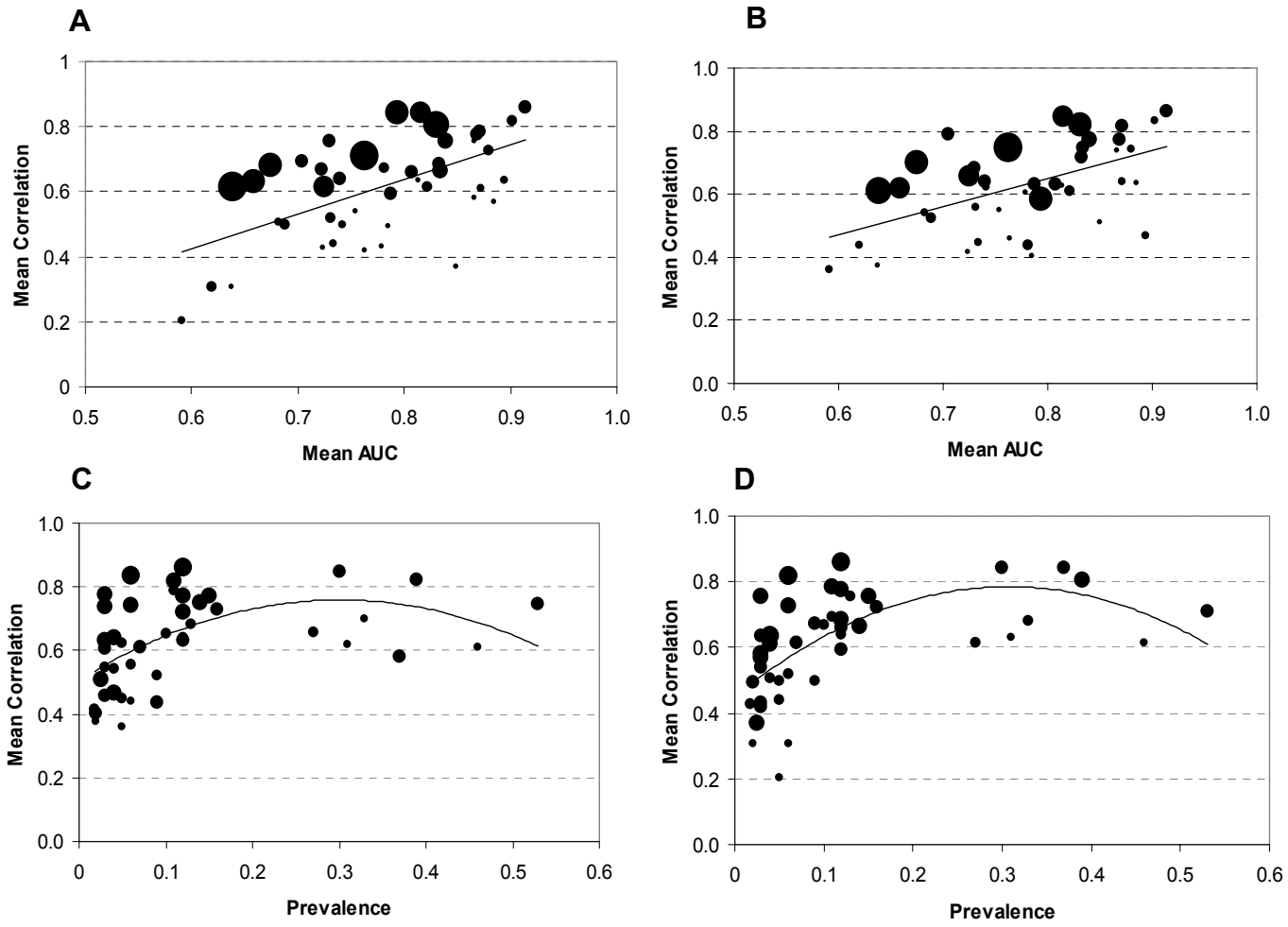
1




0 - 0.02
0.02 - 0.04
0.04 - 0.06
0.06 - 0.2
0.2 - 0.4
0.4 - 0.6
0.6- 0.8
0.8 - 1
A
0 - 0.01
0.01 - 0.02
0.02 - 0.03
0.03 - 0.2
0.2 - 0.4
0.4 - 0.6
0.7- 0.85
0.85 - 1
B
0 - 0.1
0.1 - 0.2
0.2 - 0.3
0.3 - 0.4
0.4 - 0.55
0.55 - 0.7
0.7- 0.85
0.85 - 1
C
0 - 0.02
0.02 - 0.04
0.04 - 0.06
0.06 - 0.15
0.15 - 0.3
0.3 - 0.45
0.45- 0.6
0.6 - 1
D
0 - 0.008
0.008 - 0.015
0.015 - 0.02
0.02 - 0.05
0.05 - 0.1
0.1 - 0.15
0.15- 0.25
0.25 - 0.75
E
GLM
GAM
Random Forests

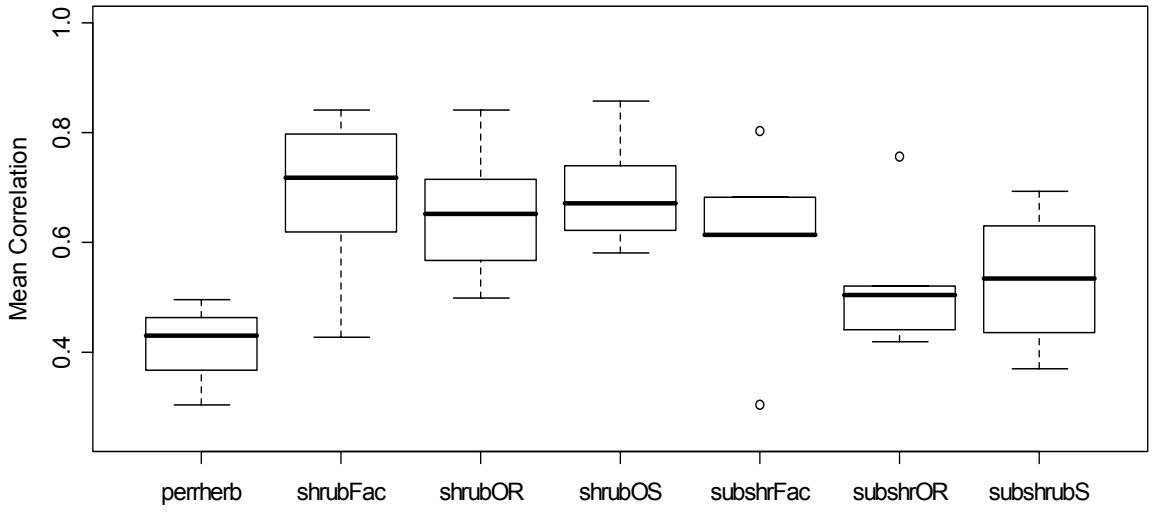
1

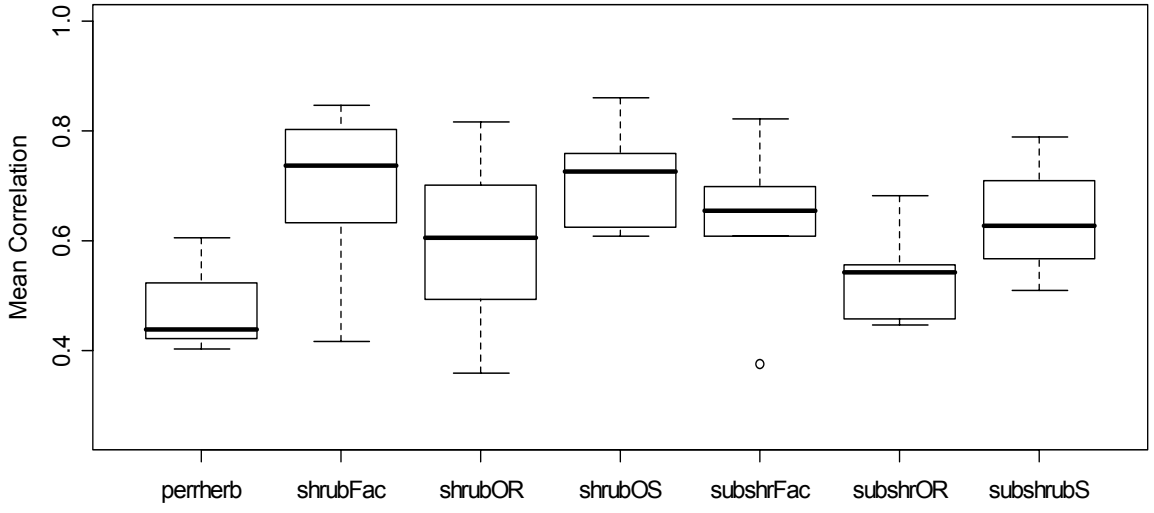
1
2
3


A
Mean Correlation
Mean AUC
B
Mean Correlation
Mean AUC
C
Mean Correlation
Prevalence
D
Mean Correlation
Prevalence

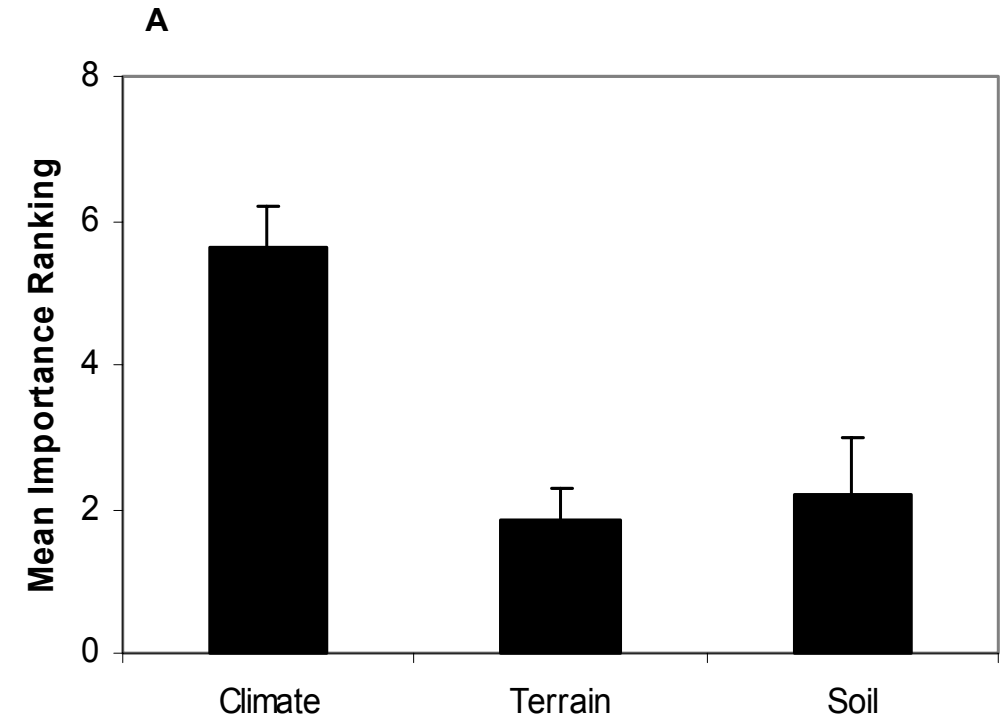
4
5

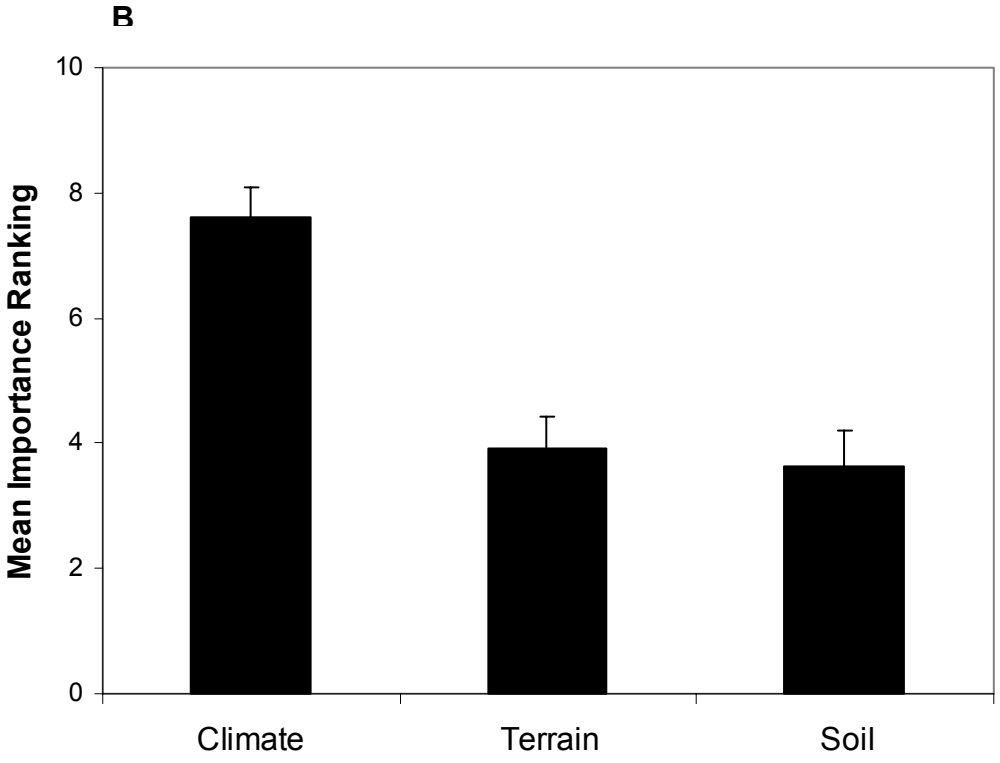
A


Mean Correlation
0.4
0.6
0.8
1.0
perrherb
shrubFac
shrubOR
shrubOS
subshrFac
subshrOR
subshrubS


Mean Correlation
0.4
0.6
0.8
1.0
perrherb
shrubFac
shrubOR
shrubOS
subshrFac
subshrOR
subshrubS

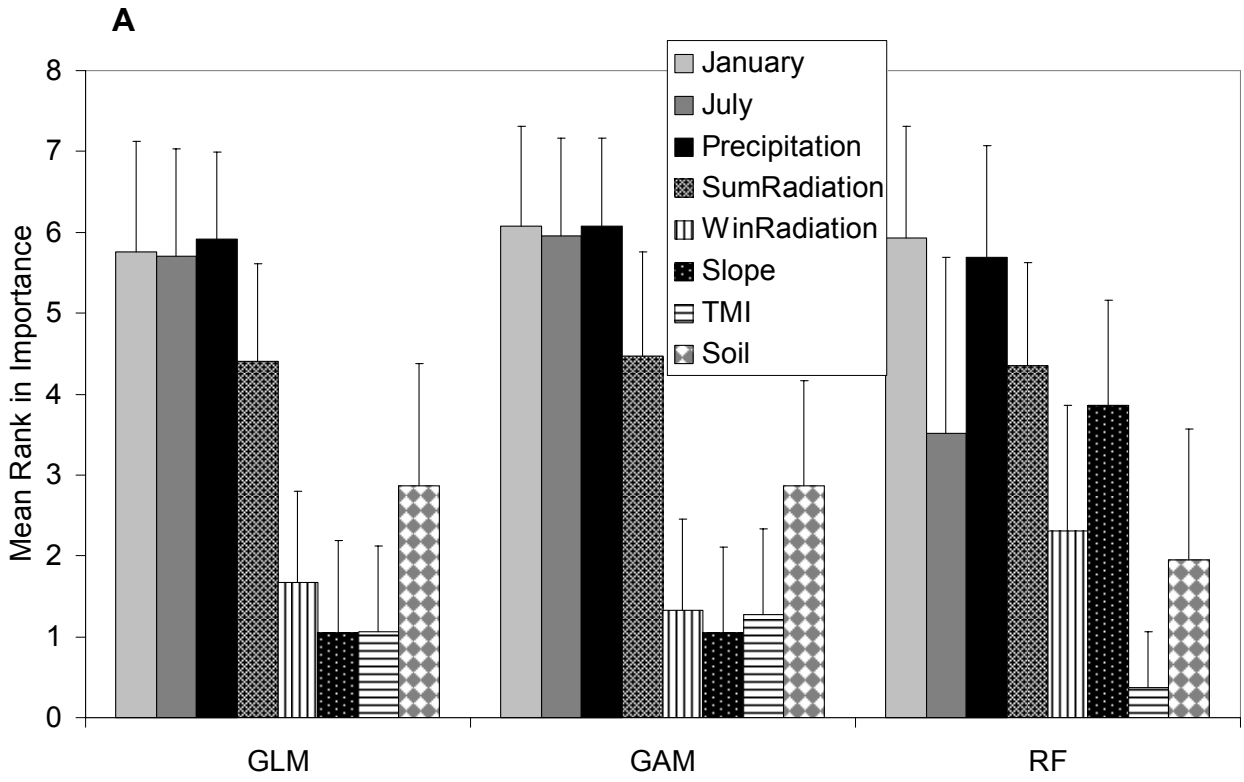
B

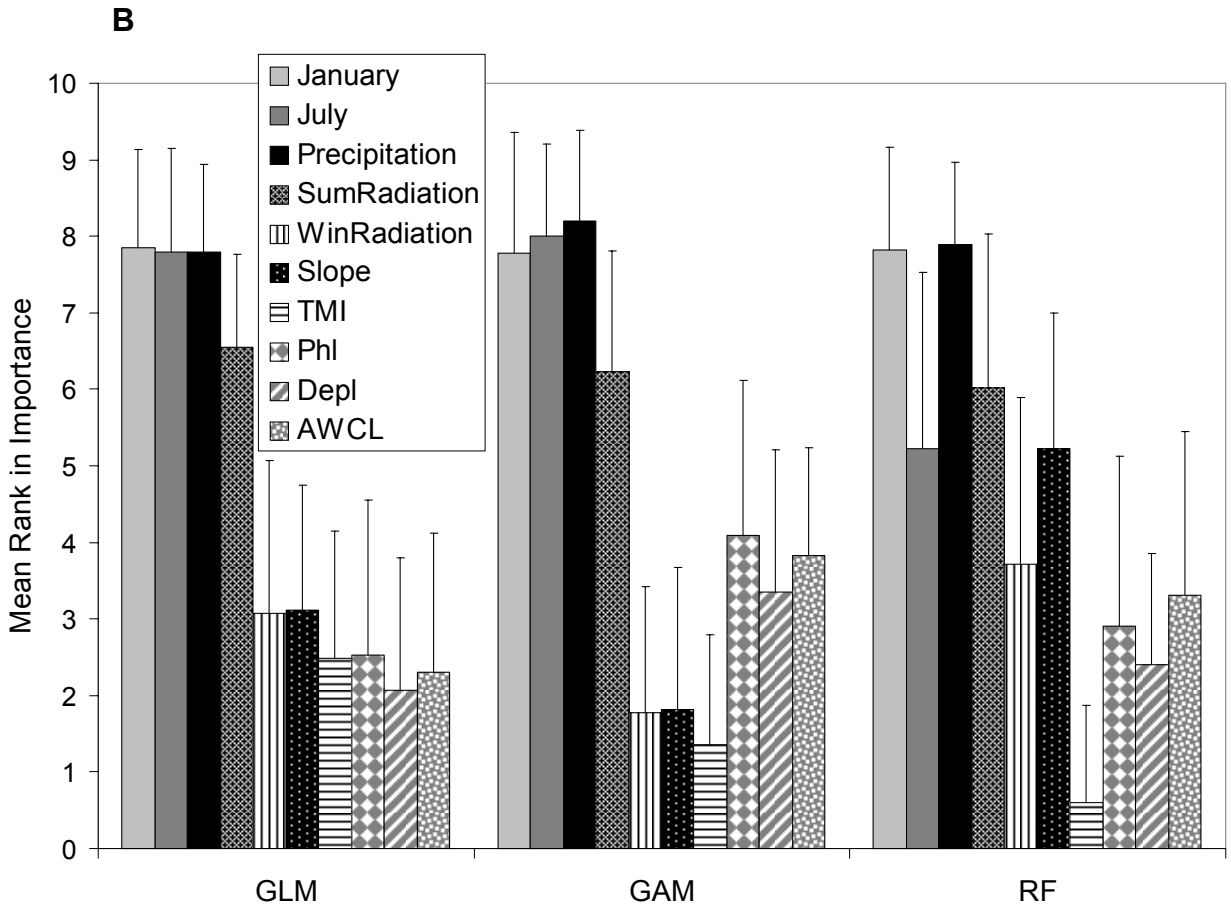

A
Mean Importance Ranking
8
6
4
2
0
Climate
Terrain
Soil


B
Mean Importance Ranking
10
8
6
4
2
0
Climate
Terrain
Soil

1

1
2
3
4
5


A
Mean Rank in Importance
0
1
2
3
4
5
6
7
8
GLM
GAM
RF
January
July
Precipitation
SumRadiation
WinRadiation
Slope
TMI
Soil


B
Mean Rank in Importance
0
1
2
3
4
5
6
7
8
9
10
GLM
GAM
RF
January
July
Precipitation
SumRadiation
WinRadiation
Slope
TMI
Phl
Depl
AWCL