2011

THE INFLUENCE OF ENVIRONMENTAL VARIABLES ON PREDICTING RARE-PLANT HABITAT IN THE NEZ PERCE NATIONAL FOREST

Thor Burbach
The University of Montana

THE INFLUENCE OF ENVIRONMENTAL VARIABLES ON PREDICTING RARE-PLANT HABITAT IN THE NEZ PERCE NATIONAL FOREST

By

Thor Burbach

B.A. Geology, University of Montana, Missoula, MT, 1998

Thesis

presented in partial fulfillment of the requirements
for the degree of

Master of Science
in Geography, GIS/Cartography

The University of Montana
Missoula, MT

Fall 2011

Approved by:

Dr. Stephen Sprang, Associate Provost for Graduate Education
Graduate School

Dr. Anna Klene, Chair
Department of Geography

Dr. Zachary Holden, Member
United States Forest Service

Dr. Soloman Dobrowski, Outside Member
Department of Forestry Management

Susan Rinehart, Outside Member
United States Forest Service

Burbach, Thor, M.S., Spring 2011 Geography

The Influence of Environmental Variables on Predicting Rare-Plant Habitat in the Nez Perce National Forest

Chairperson: Dr. Anna Klene

Habitat suitability modeling is widely-used in both biogeography and ecology to characterize the biophysical requirements and distribution of plant and animal species. Many of these modeling efforts use different variants of essentially the same topo-climatic variables (elevation, slope, aspect, precipitation, and temperature). However, these commonly used variables may not sufficiently explain the distribution of rare-plant species, which may have additional habitat needs. The aim of this project was to determine guidelines for selection of variables to include in statistical modeling efforts to predict suitable rare-plant habitat. Additionally, how background extent, data resolution, sample size, and various ranking criteria effect environmental variable selection were considered. For this case study, Broad-fruit Mariposa (*Calochortus nitidus Dougl.*) a rare-plant species found within the 2.2 million acre Nez Perce National Forest of north-central Idaho was used. The study area is dominated by mountainous terrain, with elevations ranging from 500 to 2800 m (~1500 to 9000 ft). The widely used MAXENT model and additional methods were used to statistically determine the relative importance of more than 30 environmental variables considered in the analysis and yield recommendations about the most effective way of utilizing these often highly correlated variables. Study area extent and the sample size of occurrence data had by far the greatest impact. Sensitivity to these factors resulted in variables being ranked differently, but the majority of the models ranked elevation, May precipitation, vegetation type, April minimum temperature, NDMI, September precipitation, and July maximum temperature as highly important for Broad-fruit mariposa. Vegetation type, NDMI and NDVI tended to be ranked highest when modeled at the 30×30 m resolution, suggesting that these fine resolution datasets may be extremely valuable in predicting the habitat of Broad-fruit mariposa.

ACKNOWLEDGMENTS

This study was made possible by the effort and support of a number of individuals whose expertise and encouragement made this research possible. First I would like to thank the US Forest Service for funding this project, without which I would not have been able to further advance my understanding in field of study for which I am deeply interested. I would like to thank the members of my committee, Susan Rinehart, Dr. Zachary Holden, and Dr. Soloman Dobrowski for their valuable suggestions and guidance. I deeply appreciate the dedication, expertise and professionalism of botanists Linda Pietarinen of the Bitterroot National Forest, Lois Olsen of the Helena National Forest, and Mike Hays of the Nez Perce National Forest. I would also like to thank Chris Walters, Master's degree candidate at the University of Montana, for our numerous brainstorming sessions that helped this research take shape.

To Dr. Anna Klene, my advisor, I am particularly grateful for the immense investment of time she has devoted to editing and revising my work, and for her guidance and excellent suggestions that kept me from straying too far off track.

I offer a special thank you to my family for their endless support and encouragement. To my wife Brooke, who over the last two years has shouldered much more than her fair share of the parental and household duties, I am extremely grateful, and to my daughter Laney, I promise to make up for the play time we have missed.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

The Endangered Species Act of 1973 provides special legal protections for plant and animal species determined by the United States Fish and Wildlife Service (USFWS) to be in a threatened, endangered, or sensitive status (TES). Sections 4 and 7 of this Act specifically designate the critical habitat requirements of these species and requires all federal agencies to insure that any action authorized, funded, or carried out by them will not likely jeopardize the continued existence of these species or result in destruction or adverse modification of their critical habitat (USFWS, 1988).

Region One of the United States Forest Service has the responsibility to manage more than 200 TES plant species across four states (USDA, 2011; USDA, 2005; NEPA, 1986). The effective management of TES plant species is made particularly difficult due to their lack of abundance. Surveys to find new occurrences are often prohibitively expensive, time consuming, and often not very productive (Parviainen et al., 2008). With few occurrences by which to determine specific habitat requirements, managers are charged with the difficult task of planning for the conservation of these species and their habitat without this important information.

Habitat suitability models (HSM's) are a valuable tool that can help to overcome some of the management difficulties associated with the sparse data available for many of these rare plant species. HSM's have been successfully used to in a variety of ecological applications, by assisting the discovery of new populations, identifying sites suitable for reintroduction, predicting possible climate change effects, and aiding in conservation planning and management (Hirzel and Le Lay, 2008). HSM's produce detailed

geographic predictions of a species potential distribution by analyzing the environmental conditions in those areas the species is known to be present or absent (Elith et al., 2006).

Geographical modeling in ecology relies on quantifying the species–environment relationship. HSMs can help reveal the influence of environmental factors in defining the distribution of plant and animal species (Guisan and Zimmerman, 2000). These modeling efforts are made possible by the availability of digital maps of environmental variables (Franklin, 1995). The ever increasing availability of environmental variables in digital format, along with advances in GIS technology, promise to greatly improve the mapping of species distributions (Brotons et al., 2004). The choice of predictor variables used in HSMs has a strong influence on the model's output (Guisan and Araujo, 2006). Thus, the selection of those key environmental variables that approximate a species ecological niche is a crucial element of successful Habitat Suitability Models (Guisan and Zimmermann, 2000; Hirzel and Le Lay, 2008).

This research represents the second phase of an ongoing effort to improve habitat modeling methods and to develop a better understanding of the distribution of rare-plant habitats. One of the goals of the project was to develop a set of basic guidelines to assist in the selection of environmental variables to be used by statistical models to map and predict suitable rare-plant habitat. Broad-fruit Mariposa (*Calochortus nitidus Dougl.*) was selected from the 21 rare-plant species included in the first phase of this project (Nock, 2008) based on its relatively abundant occurrence within the study area.

To capture fine-scale variation in habitat distribution, the size of the study area was limited to a portion of the Nez Perce National Forest rather than half of U.S. Forest Service (USFS) Region One as was previously considered by Nock (2008). However,

with only minor modification the general procedures developed here should be applicable to other National Forests within USFS Region One. This study area was selected due to the large quantity and variety of rare plant occurrence data available (Idaho Fish and Game, 2010); however, the accuracy of this type of data can vary considerably depending on the collection methods (Elith et al., 2006). The locations of the most recent records collected using the Global Positioning System (GPS) are accurate to within 25 m whereas the locations of the older records in the dataset have accuracies of 200 m or more.

This research specifically addresses questions pertaining to the selection and relative predictive value of the environmental variables available for use in HSMs in USFS Region One. Specifically, how study area extent, data resolution, sample size, and variable rank criteria effect which environmental variables are most statistically influential, and which subsets of variables provide the most accurate habitat predictions for the rare plant species under consideration? In addition this project explored the influence of herbaria data precision on variable importance, by comparing models utilizing training data at 30×30 m resolution to those with data available at 200×200 m resolution. It provides guidelines for variable selection in HSMs in order to assist them in applying these techniques to their particular management goals.

2 BACKGROUND

2.1 Previous Work

The first phase of this project completed by Nock (2008) examined 21 rare-plant species as identified in the USFS Region One Regional Forester's Sensitive Species List. That study focused on the western half of Region One, an area of roughly 11 million hectares (Nock 2008). The goal of that phase of the research project was to examine the availability of rare-plant occurrence data and test the suitability of the DOMAIN algorithm (Carpenter et al., 1993). Rare-plant surveys were also conducted in the Beaverhead-Deerlodge, Bitterroot, and Nez Perce National Forests in order to refine survey techniques, obtain new rare-plant occurrence data, and to develop a deeper understanding of these rare-plants habitats (Nock, 2008).

The DOMAIN algorithm (Carpenter et al., 1993) was used to model suitable habitats for 21 rare-plant species using seven environmental variables: annual precipitation, mean May temperature, slope, aspect, elevation, geologic material, and dominant vegetation type (Nock, 2008). The DOMAIN model output was evaluated by comparison of the percentage of occurrences found within areas predicted as potential habitat, whether Forest Service botanists identified potential habitat within those areas, and whether new occurrences were discovered within areas predicted by the DOMAIN model (Nock, 2008).

Nock (2008) found that future research may find it beneficial to work at finer resolutions than 60 × 60 m in an attempt to include microhabitat. The DOMAIN algorithm generally performed well but did have its limitations: it did not address correlations and possible interactions between environmental variables nor did it isolate

the effect each environmental predictor had on the species' distribution to allow determination of which variables were the most important (Nock, 2008).

Of the 21 rare-plant species considered in the first phase of this project, Broad-fruit mariposa, was selected based on its relatively abundant occurrence within the Nez Perce NF. Broad-fruit mariposa is a perennial herb ranging from 20 to 41 cm (8 to 16 in) in height. It typically produces up to four large lavender flowers (Figure 1). The plant emerges in June, and grows rapidly through July, flowering for 7-10 days in early July (Hitchcock et al., 1969). Broad-fruit Mariposa can primarily be found in the Palouse grasslands of Eastern Washington and Central Idaho. It is generally associated with loess and alluvium-dominated soils, and can inhabit flat to moderately steep slopes in an elevation range from 500 to 2000 m above sea level. It is associated with landscapes dominated by perennial grasslands and deciduous shrub-lands; in Idaho it is also known to inhabit open woodland areas adjacent to Palouse grasslands (Hitchcock et al., 1969).

2.2 **Study Area**

The Nez Perce National Forest was selected from the area considered by Nock (2008) due to the highly variable environmental conditions and large number and variety of rare plant species present in this region. This National Forest covers more than 2.2 million acres of north-central Idaho, stretching from the Oregon border on the west to the Montana border on the east (Figure 2). It is bordered by the Clearwater National Forest to the north, the Bitterroot National Forest to the east, and the Payette National Forest to the south. Sixty-five percent of the Nez Perce NF (877,000 acres) is designated wilderness or inventoried road-less area including sections of the Selway-Bitterroot, Frank Church-River of No Return, and Gospel-Hump Wildernesses (USDA, 2007). The

study area is dominated by steep slopes, with elevations ranging from less than 460 to 2740 m, (1500 to >9000 ft). The Nez Perce NF contains several large rivers including the Snake, Salmon, Selway, and South Fork of the Clearwater. Located approximately 480 km (300 mi) from the Pacific Ocean, the area's climate is influenced by maritime air masses and prevailing westerly winds, resulting in increased precipitation and more moderate temperatures than those found at the same combination of latitude and altitude further inland (USDA, 2007). The soils are moderately productive; an ash layer covers many of the soils, adding nutrients, water-holding capacity, and soil stability. The Nez Perce NF supports a variety of vegetation types. The lower elevations and southerly aspects are dominated by ponderosa pine forests interspersed with native grass and shrub-lands. Locations at high elevation or with northerly aspects are heavily forested, containing fir, lodgepole pine, Ponderosa pine, western larch, western red cedar, and Engelmann spruce (USDA, 2007).

2.3 **Habitat Suitability Modeling**

2.3.1 Historical and Theoretical Basis of HSM

An understanding of plant distribution is of academic importance to numerous scientific disciplines, including ecology, botany, and geography; however, before it was the subject of academic interest, this understanding was a critical survival skill in hunter/ gather societies (Kelly, 1983). Some of the earliest written records of the plant/ environment relationship come from the ancient Greeks. Of the ancient Greeks, the greatest contributions to ecology were made by Aristotle and his student, Theophrastus, who made extensive botanical observations. The influence of sun, exposure, elevation,

aspect, soil, water, temperature, and even other plants and animals were common environmental factors that he examined (Hughes, 1975).

Joseph Grinnell's (1917) observations of California birds led to the formulation of the "Ecological Niche" as the fundamental concept governing the spatial distribution of plants and animals. Hutchinson (1957) refines the Grinnellian ecological niche concept through the incorporation of a mathematical function that links the fitness of individuals to their environment; specifically, as an *n*-dimensional volume in environmental space that defines a species' habitat, wherein each of the *n* axes represents an environmental variable that is critical in defining the habitat of that species. Modern habitat suitability models rely on this mathematical formulation of the niche theory to delineate those conditions that best define suitable habitats through statistical correlation between environmental variables at areas of known habitat occupation (Hirzel et al., 2002). Hutchinson (1957) also postulated an important distinction between the "fundamental niche," (the range of theoretically suitable habitat), and the "realized niche" (that part of the fundamental niche which is actually inhabited). The realized niche is limited by competition, physical barriers, and random extinction events (Pulliam, 2000; Guisan and Thuiller, 2005). The realized niche acknowledges that factors other than current environmental conditions - barriers, competition, past catastrophic events, etc. - can and often do influence a species' observed distribution.

Ecologists, botanists and biogeographers have long questioned whether species are distributed randomly within their environmentally defined range or systematically in response to geographically-varying environmental gradients. Whittaker's work in the Appalachian Mountains, correlating the occurrence of species to environmental gradients,

provided strong support for the systematic response theory (Whittaker, 1967). Biogeography was founded on the observed relationships between plant distribution and environmental characteristics. It is the quantification of these relationships that forms the basis for modeling the geographic distribution of plant habitats. These models predict the distribution of suitable plant habitats based on the environmental factors found at a particular location occupied by the plant species being modeled (Guisan and Zimmermann 2000).

2.3.2 Modern HSMs

The first habitat suitability models were primarily based on environmental envelope techniques. The BIOCLIM model calculates the smallest rectilinear envelope in a multi-dimensional climatic space that contains the occurrence data (Busby, 1991). Walker and Cocks (1991) in an effort to improve upon BIOCLIM developed the HABITAT model that uses a convex envelope to define the environmental space enclosing the occurrence records. The DOMAIN model (Carpenter et al., 1993) utilizes a point-to-point similarity metric (a measure of multivariate distances in environmental space) rather than classification trees and generally outperforms the BIOCLIM and HABITAT models particularly when the number of occurrence records is limited. DOMAIN was used by Nock (2008).

Canonical correspondence analysis (CCA) is a technique that analyzes species distribution by examination of principal ordination axes constructed from linear combination of environmental variables. The assumption of a Gaussian species response to the environmental variables is the primary limitation of this method (ter Braak, 1986). Ecological niche-factor analysis (ENFA) is the modeling algorithm implemented in the

BIOMAPPER package (Hirzel et al., 2000). It has many features that make it particularly suitable for use in ecological modeling, it requires only presence data, it transforms data much like principle component analysis (PCA) but on axes that represent indices of species *marginality* and *tolerance*, that are more easily interpreted in an ecological context than PCA (Guisan and Zimmermann, 2000).

Generalized Linear Models (GLM) and Generalized Additive Models (GAM) are multiple-regression techniques (Hastie and Tibshirani, 1987). Because of their ability to model complex ecological relationships and similarity with widely used and well understood multiple-regression techniques, (GLMs) and (GAMs) have seen extensive use in habitat-suitability modeling efforts (Austin, 2002). GLMs fit parametric terms: combinations of linear, quadratic, and cubic terms. GAMs further improve model flexibility through the use of data-defined non-parametric functions and cubic-splines to fit non-linear response curves (Elith et al., 2006; Hastie and Tibshirani, 1987). Multivariate adaptive regression-splines (MARS) is another regression-based method that is capable of fitting complex ecological response curves by utilizing piecewise linear functions (Leathwick et al., 2005).

Recently, there have been a number of artificial neural network and machine learning techniques applied to habitat-suitability modeling. The General Algorithm for Rule-Set Prediction (GARP) is a machine-learning technique that produces a set rules that when combined give a binary prediction of habitat. The rules developed from presence pixels are ranked by the algorithm based their significance as compared with random predictions based on sampled background pixels (Stockwell and Peters, 1999). Boosted regression trees (BRT), is a relatively new machine-learning technique that

combines two algorithms: a boosting algorithm that iteratively accesses the regression-tree algorithm to build a set of trees. The boosting procedure overcomes many of the path dependent inaccuracies commonly encountered in single classification tree models. Boosting develops the regression model iteratively, with each iteration modifying sections of the tree to better fit the data. The advantages of this method lie primarily in its ability to model variable interactions and in its proficiency in selecting relevant environmental variables (Friedman et al., 2000). MAXENT is also a recently developed machine-learning algorithm that models a species' distribution by finding the distribution of maximum entropy or that which is the most uniform distribution subject to the constraints imposed by the environmental variables (Phillips et al., 2006).

Of the modeling techniques discussed here, MARS, BRT, and MAXENT performed best over the wide range of conditions considered in a comprehensive comparison of 16 commonly employed models (Elith et al., 2006), however most of the models, including the simple DOMAIN algorithm performed fairly well. A number of other model comparison studies have also demonstrated MAXENT's high level of performance (Phillips et al., 2006; Sergio et al., 2007; Parolo et al., 2008; Phillips, 2008; Williams et al., 2009).

2.4 **Environmental variable selection**

Modern plant-distribution modeling is based upon the same environmental factors that were considered by those first ecologists and biogeographers. The environmental variables utilized in plant distribution modeling are often limited to those deemed most important, prompting critics to question whether these small sets of environmental

variables can truly describe the complex plant-environment relationships and approximate their distribution (Guisan and Araujo, 2006). Critics also make the argument that the distribution of plants may be determined by factors other than environmental relationships, such as those relating to species competition. These criticisms serve to highlight the need for cautious interpretation of any model's results.

Austin (1980) classified environmental variables based on their biophysical importance, identifying three basic types of environmental variables: "resource variables, direct variables, and indirect variables." Resource variables such as light, water, and soil nutrients are those that are consumed by plant species. Direct variables are those that are not consumed but have a direct physiological influence on a plant species. Indirect variables have no physiological effect on the plant species but are in some way correlated with variables that do. Models based on direct and resource variables will be the most robust and widely applicable, but it is very difficult to provide a continuous digital representation of these types of variables (Austin and Smith, 1989). Using these variables for predictive mapping of species distribution is difficult and is still extremely rare (Austin and Smith, 1989).

Models are exclusively driven by the data given to the algorithm. The choice of which environmental variables are used to characterize the ecological niche is therefore a critical step in any habitat suitability modeling endeavor (Guisan and Araujo, 2006; Guisan and Zimmermann, 2000; Heikkinen et al., 2006; Hirzel and Le Lay, 2008). Topo-climatic gradients such as elevation, slope, aspect, mean temperature, and precipitation are readily available or easily approximated from Digital Elevation Models (DEMs). For this reason, most habitat modeling efforts have focused on these types of

indirect variables. Inclusion of more proximal variables such as soil type, water availability, radiation, and surface temperature should greatly improve model performance (Austin and Smith, 1989), but until recently this information has not generally been available in the required format. Fortunately, advances in both satellite and airborne remote sensing technology and data analysis techniques continue to increase the availability and usefulness of this type of data (Zimmermann et al., 2007).

The question still remains, are these commonly used variables the right variables? Critics of the HSM approach would say that there are many factors as, or equally, important that the method does not consider. It is true that no model can hope to capture every variable that may influence the distribution of plants, but inclusion of complex inter-species relations also requires detailed knowledge which is very seldom available. That leaves these HSMs trying to assess what variables can get the best, albeit limited results. Are the variables which are generally used (latitude, elevation, slope, aspect, evapotranspiration, mean temperature, precipitation, soil type, water availability, radiation, cloud cover, and vegetation type) adequate? That question is difficult to answer, but HSM's do generally predict species' distribution fairly accurately, and the variables most often used are generally related to those that naturalists, botanists, and biogeographers have been identifying as important for more than 2000 years.

The list of variables that have been historically used in HSM efforts is extensive. But computationally and statistically, it is usually necessary to restrict the variables used to those deemed most important (Guisan and Araujo, 2006; Guisan and Zimmermann, 2000; Heikkinen et al., 2006; Hirzel and Le Lay, 2008). It is generally desirable to pick the most parsimonious model by restricting the variables included to a small

biophysically meaningful set. Parsimony is advantageous in that it simplifies the modeling process, and while the inclusion of fewer variables might decrease model accuracy slightly, it reduces the propagation of error into the output which can be multiplicative based upon the number of input layers (Wainwright and Mulligan, 2004). It also usually results in models that are easier to interpret and are more readily transferable to different geographic areas. Inclusion of multiple correlated variables (for instance, an annual mean and two mean monthly temperature layers) can lead to increases the uncertainty in estimates of which is most important (Phillips 2008).

An often-recommended method is to use expert knowledge of the species requirements to develop these lists (Guisan and Zimmerman, 2000). However, when knowledge of the species requirements is unknown or the important variables are not available in an appropriate format, automated algorithms are often used to minimize the number of variables considered while still fitting the data well (Hirzel and Le Lay, 2008).

Stepwise procedures such as Akaike Information Criterion (AIC; Akaike, 1974), Bayesian Information Criterion (BIC; Schwartz, 1978), and BRUTO (Hastie and Tibshirani, 1990) are often employed to identify the most parsimonious model while preventing the loss of important information. Sophisticated techniques such as ridge regression or lasso that is very similar to the regularization method used in MAXENT (Dudik et al., 2004; Tibshirani, 1996) are very powerful as they penalize over-fitting, using only those parameters that have the greatest contribution to model performance. Some modeling algorithms have incorporated these types of procedures, they often do this by running each variable separately or dropping out one variable at a time which may be inadequate if there are more than two correlated variables included as input layers, as

they are not fully step-wise and fail to consider all combinations of variables. Another method is to compare the variables selected by a number of competing models (Johnson and Omland, 2004). Simple Analysis of Variance (ANOVA) tests can reveal whether reduced and full model performance is significantly different.

3 METHODS

3.1 Data Acquisition

3.1.1 Occurrence Data

This study required both Broad-fruit Mariposa occurrence data and environmental data for the Nez Perce NF. The presence data was acquired from the August 2010 update of the Idaho Fish and Game Plant Conservation Database (Idaho Fish and Game, 2010). This species was selected because of its relatively large number of records and to maintain continuity with the original study conducted by Nock (2008) and allow comparison to results of that work.

The presence data came as polygon files and all occurrence records falling outside of the Nez Perce NF were removed from consideration. The extent of these polygons represented the estimated spatial error of the presence location, and this error was significantly larger than the desired modeling resolution of 30×30 m for the majority of the records. In order to explore rare plant micro-habitat association recommended by (Nock, 2008) two modeling resolutions were utilized, one at 30×30 m (fine) and one at 200×200 m (coarse). After removing all polygons with error envelopes greater than the 200 m cutoff, a total of 139 occurrence records remained; these remaining records were split into two separate groups. 119 records had sufficient precision to be included in the 200×200 m model while only 20 records had sufficient precision to be included in the 30×30 m model. In some instances a single polygon represented numerous distinct populations. These polygon features were converted from multi-part to single-part features and all occurrence polygons were then converted to point data by taking the centroid of each polygon.

This type of opportunistically collected ecological data often suffers from both spatial autocorrelation and sampling bias (Legendre, 1993; Phillips et al., 2009). The effect of sampling bias was not specifically addressed here, although road proximity was included as an environmental variable to assess how access may be related to occurrences. There are specific methods available (and they are easily implemented in MAXENT) to address the issue of sampling bias (Phillips et al., 2009) but they were not considered here due to a lack of sufficient data. Two approaches were used to assess the degree of spatial autocorrelation present within the occurrence records. First, a simple measure of Euclidean distance between occurrence points from the 30×30 m and 200×200 m datasets were calculated to determine the minimum, mean, and maximum distances between the data points (Figure 3) and second, a calculation of the average nearest-neighbor statistic that characterizes the degree of clustering in spatial data (values less than one indicate clustering, values greater than one indicates dispersion, and one indicates a random pattern). The 200 m occurrence records of Broad-fruit Mariposa had an average nearest-neighbor value of 0.65 (z-score of -2.95 and one sided p-value of -0.0031) indicating that it is highly unlikely that the data is randomly distributed (Ebdon, 1985).

3.1.2 Variable Selection

To identify which variables to consider for inclusion, it is typical to start with the biological requirements of any plant (radiation, heat, and moisture) and then add known species-specific needs. Expert-opinion was sought in this study, through examination of the literature, field notes from and personal communication with botanist Mike Hays, Nez

Perce National Forest, and personal communication with Steve Shelly, USFS Region One Rare Plant Program manager, to identify variables likely important to Broad-fruit Mariposa. Those included soil type, moisture availability, and an association with grass and shrub-dominated areas within a relatively narrow elevation band (Hitchcock et al., 1969; Hays (field notes) 2010; Shelly (personal communication), 2011). However, generally once a list of desired variables has been produced, the lack of appropriately scaled, spatially complete data layers for each variable requires substitutions of related layers. The list of environmental variables initially considered thus also considered variables readily accessible to USFS analysts and managers in Region One and those commonly used in other plant modeling efforts (Engler et al., 2004; Franklin, 1995; Guisan and Zimmermann, 2000; Parolo et al., 2008).

Forty-two environmental variables were initially considered in order to evaluate MAXENT's ability to select the most important variables from a large initial set. Unfortunately, several layers were not included (ecological subsections, climate zones, geomorphology, geologic parent material, and soil type) because they were not available at the desired resolutions. Soil type in particular may be a very important predictor of Broad-fruit Mariposa habitat, its future inclusion may significantly improve modeling results (Shelly (personal communication), 2011; Hays (personal communication), 2011). Hays also suggested that potential vegetation type might be more effective as a predictor than the dominate vegetation type used here (Hays (personal communication), 2011). The remaining 37 environmental variables were included as inputs to MAXENT to predict Broad-fruit Mariposa habitat. Each layer and its source are shown in (Table 4).

3.1.3 Climatic Data

The climate data used consisted of 30-yr mean monthly temperature and precipitation data for the months of April through September (1971-2000) and was downloaded from the PRISM climate group (www.prism.oregonstate.edu/products/). This data is gridded by an algorithm that produces a continuous raster grid of estimated climatic parameters from point measurements of precipitation, temperature, and other climatic factors (PRISM, 2010). This data was available at 800×800 m resolution and was down-scaled to 30×30 m resolution using a Digital Elevation Model (DEM) aided interpolation technique. That involved removing the elevation effects by using the initial resolution DEM and a standard lapse rate of 6.5°C per 1000 m, then re-sampling the remaining temperature signal to 30×30 m using linear interpolation, and then re-scaled up to DEM elevations using the finer resolution DEMs and the same lapse rate (Willmott, 1984). Temperature anomalies related to non-standard atmospheric conditions and topographic convergence not accounted for in the PRISM data would not be corrected for by this procedure. The 800×800 m resolution precipitation data was simply re-sampled to a 30×30 m by linear interpolation. Spring climate variables were calculated as the mean of the April, May, and June values, and the summer climate variables were calculated as the mean of the July, August, and September values. These variables were included as they are related to the species moisture availability and heat requirements (Table 4). However, the monthly and seasonal variables are highly correlated and a jackknife or some other selection procedure should be employed to isolate the most influential variable from each of these sets for inclusion in the final model.

3.1.4 Topographic Data

The topographic variables elevation, slope, aspect, were derived from a 30×30 m DEM (USDA, 2007). Beers aspect is a transformation of aspect into a continuous variable that ranges from 0.0 to 2.0 with 0 = SW, 1=NW & SE, and 2=NE aspects as:

$$\text{Beers aspect} = 1 + \cos\left((45° - \text{aspect})/\text{degrees}\right)$$

Beers aspect has greater ecological significance than standard aspect in that it better delineates cool and warm aspects (Beers and Wensel, 1966). The "Topographic Wetness Index" was also derived from the DEM to estimate the variability of soil moisture over the study area (Beven and Kirkby, 1979; Moore et al., 1991). The TWI value for each pixel will be calculated as the natural logarithm of the result of area draining into that pixel (a) divided by the tangent of the slope at that pixel ($\tan\beta$). Average annual solar radiation was calculated using the algorithm of van Manen (2010) that calculates average solar insolation using hill-shade functions derived from the DEM at the solstices and the spring equinoxes. The potential evapotranspiration data was calculated using mean solar insolation and mean temperature for the month of August calculated using the Solar Analyst tool in ArcGIS® 9.3 (ESRI, 2009) as inputs into the algorithm of Jensen and Haise (1963) that was specifically designed for use in the Intermountain West. Elevation was specifically identified by USFS botanists as being an important predictor for Broad-fruit mariposa (Table 4). The other topographic variables were included due to their common inclusion in other modeling efforts (Guisan and Zimmerman, 2000). These variables are, however, strongly correlated to elevation and likely add little new

information to the model and therefore could be excluded without adversely affecting the model results.

3.1.5 Vegetation, Road, and Disturbance Data

The categorical variables of road proximity, timber harvest history, fire history, and vegetation type were acquired from the USFS Region One geospatial database. The road data represent all public and USFS roads in the Nez Perce NF inventory. The roads line layer was used to calculate a 30×30 m raster representing the simple linear distance to the nearest road. The timber harvest layer was constructed by converting all polygons in which mechanical harvesting techniques were used from 1980 through 2010 and converting them to a single raster with integer values ranging from 0-30 that represent years since harvest. The fire history data was constructed in the same manner as the timber harvest data with values ranging from 0-20 years (USDA, 2010). These layers were included in an attempt to assess the possible effects of access and disturbance, which are often significant for rare species (Rinehart (personal communication), 2008). The vegetation data comes from the USFS VMAP depicting dominant vegetation types; it was available at a 30×30 m resolution and was originally derived from Landsat imagery (USDA, 2006). Vegetation type was included as it was identified by USFS botanists as being an important predictor for Broad-fruit mariposa (Table 4).

3.1.6 NDVI and NDMI layers

More difficult to gather was data for the satellite-based input layers. The Normalized Difference Vegetation Index (NDVI) is the most widely used of the vegetation indices due to its simplicity and direct link to physical processes in plants (Liu

and Huete, 1995), as shown by the reflectance in the red and near-infrared (NIR) portions of the spectrum. It is calculated:

$$NDVI = (NIR-RED) / (NIR+RED)$$

The Normalized Difference Moisture Index (NDMI) is similar to the NDVI but uses the mid-infrared (MIR) band that is strongly absorbed by the water contained within plants, and is strongly associated with water availability. NDMI is calculated:

$$NDMI = (NIR-MIR) / (NIR+MIR)$$

While it is difficult to make direct interpretations about the biophysical meaning of NDMI information, it has been shown to be highly correlated to the "wetness" feature of Kauth and Thomas', (1976) "Tasseled Cap" transformation (Wilson and Sader, 2002).

The data to calculate the NDVI and NDMI were obtained from the USGS Landsat archive at glovis.usgs.gov. Data from row 28 and paths 41 and 42 were needed to cover the entire Nez Perce NF. The USGS archive was queried for all available July scenes with cloud cover less than ten percent. Based on these constraints 19-years of Landsat TM data were used with collection dates ranging from 1986 through 2008 (USGS, 2010). The July images for 1987, 1993, 2001, and 2007 had cloud cover that exceeded 10% and were therefore excluded from the analysis. Nineteen-year mean July NDVI and NDMI values were calculated in order to make these datasets as comparable as possible to the 30-yr PRISM climate data. The data was downloaded having undergone pre-processing including standard geo-rectification and radiometric corrections (Chandler et. al, 2007). These variables were included as proxies for summer moisture availability, but may also serve to delineate changes in vegetation types, both of which were identified by botanists as important (Table 4).

3.1.7 Data Processing

All data was projected to the UTM 11N coordinate system, geo-referenced and clipped to the USGS geo-rectified NDVI and NDMI extents. Some of the environmental variables acquired from the USFS were only available for the National Forest and immediately adjacent areas only, requiring all of the data to be clipped to this smaller geographic extent. All data sets discussed above were re-sampled to 200×200 m, for use in models that used occurrence records with accuracies less than 30 m, using the Spatial Analyst tool in ArcGIS® 9.3.

While it was a useful exercise to present a wide range of environmental variables to experts for consideration, the initial set (Table 4) was much too large and needed to be constrained in order to be parsimonious and have meaningful results as described in section 2.4. Therefore, the large set initially considered was reduced to an intermediate (10 variable) and reduced (4 variable) subset using the procedures described below. Even the intermediate subset is considered to be too many variables by many researchers (Guisan and Zimmerman, 2000; Johnson and Omland, 2004).

3.2 **Algorithm Implementation**

The concept of maximum entropy lends itself well to modeling the distribution of rare-plant species based on presence data alone. The major advantage of this approach is that "it agrees with everything that is known, but carefully avoids assuming anything that is not known" (Jaynes, 1957). The MAXENT algorithm develops a probability distribution that is as uniform as possible while being constrained by the empirical mean of features developed from the environmental variables at the occurrence locations. Berger et al., (1996) have shown that a distribution of this type is equivalent to a Gibbs

probability distribution that minimizes the negative log likelihood (log loss) of the sample points. Using this method each feature is multiplied by a weighting factor (λ). The appropriate maximum entropy distribution is then found by starting with a uniform distribution in which all weighting factors (λ) equal zero and iteratively varies them such that the negative log likelihood decreases to a minimum (Dud´ık et al., 2004). In order to prevent over-fitting, the weighting factor (λ) is constrained by a "regularization factor" β (that determines how close the modeled values must be to the sample means). The "regularization factor" forces the MAXENT distribution to focus on the most important environmental variables, reduces the tendency to over-fit the available data (Dud´ık et al., 2004).

Of particular interest for this study, MAXENT incorporates a variety of tools for analyzing the relative importance of environmental input variables. During the model training process, the algorithm tracks of which environmental variables make the most significant contribution to the final habitat prediction and produces a table which ranks the importance for each environmental variable (Phillips, 2008). The built-in jackknife tests that first iteratively excludes each variable and also considers each variable in isolation allows easy comparison of the relative impact each variable has on the overall distribution (Phillips, 2008). The output of the jackknife test is a chart and corresponding data table that shows the effect that each environmental variable has on overall training gain, test gain, and the Area Under the [Receiver Operator Characteristic] Curve (AUC; Figure 3). This research utilized small presence-only datasets, and mainly focused on assessing the relative importance of each environmental variable. As such, MAXENT was a convenient model for this effort. It offered proven performance, the capability of

handling the available data, and had the tools needed to address the research questions. However, it is believed that most algorithms being used in the literature today would have produced similar maps of suitable habitat if driven by the same input variables and thresholds.

Technically, running MAXENT is a simple matter of identifying the file containing the occurrence data, the directory containing the environmental variables, and the directory where model out output will be stored (Phillips, 2008). The output is a probability distribution over the entire study area. Computationally, it starts with a uniform distribution and in an iterative fashion fits the modeled distribution to the training data. A randomly selected percentage of the occurrence data can be set aside to test model performance. The algorithm outputs a variety of variable contribution tables, as well as a probability surface, however, as this distribution is developed from only presence data it should be interpreted as an index of relative habitat suitability rather than the probability of species occurrence (Phillips 2008; Figure 4). These outputs were used to compare the effect of selecting different sets of environmental variables had on the predictive performance of the models. It is important to note that these comparisions were based on AUC, a measure that has known limitations but without absence data was the only option available, which limits the strengths of the analysis?

3.3 **Addressing Correlation of Occurrence Data**

The examination of the occurrence data discussed above indicated that the records were highly clustered and therefore likely to exhibit some degree of spatial autocorrelation (Legendre, 1993). Parolo et al. (2008) addressed this problem by dichotomously splitting those occurrences that were closest together into the training and

test datasets. This procedure would very likely artificially increase any measures of model performance. In order to avoid this pitfall, the procedure of Parolo et al. (2008) was modified by taking 30 random bootstrap samples of the occurrence records, splitting them into 75% training sites and 25% test sites, and running the model on each of these random partitions of the data. This bootstrapping procedure is easily implemented in MAXENT through the available optional settings. The mean response of these 30 predictions was then used in all further analyses. While this procedure does not remove the correlation in the data it does guard against generating by chance a model highly influenced by these correlations. In addition, examination of the 30 individual models created by this procedure may provide some insight into the effect that this correlation has on model performance (Figures 5 and 6). In addition to spatial auto-correlation in the occurrence data, many of the environmental variables were highly correlated to one another (Table 5). This correlation can confound the interpretation of MAXENT's measures of variable importance, making it difficult to accurately rank variable importance (Phillips et al., 2006).

3.4 Comparing Model Accuracy and Variable Importance

A variety of models were developed using different background extents, data resolution, sample sizes, occurrence data precision, environmental data resolution, , number of variables included and the variable ranking outputs available. All of these models were built using the 30-bootstrap replicate procedure described above. For each of the models used to examine these factors, both accuracy and the impact on which variables were statistically most important were examined.

3.4.1 Comparing Accuracy

The Area Under the [Receiver Operator Characteristic] Curve (AUC) is a threshold-independent statistic of model performance. AUC is a convenient index because it provides a single measure of overall prediction accuracy that is not dependent upon a particular threshold. However, AUC as implemented in MAXENT measures the model's ability to correctly rank sites with respect to their relative suitability. This performance measure does not directly assess omission and commission rates as would the use of a confusion matrix and Kappa statistics commonly used in presence/ absence modeling techniques. AUC values between range between 0.0 and 1.0. A value of 0.5 indicates that the model performs no better than random, while a score of 1.0 indicates perfect prediction (Fielding and Bell, 1997). AUC values greater than 0.7 are generally thought to be useful for conservation planning (Elith et al., 2006). The AUC is MAXENT's standard measure of model performance and is automatically generated for each model run (Phillips, 2008).

AUC values were computed for each of the various scenarios described below. The resulting mean-AUC values were compared using standard ANOVA and Welch t-test procedures to determine whether there was statistical evidence that these factors produced better performing models. When it was necessary to compare more than two mean-AUC values, and analysis of variance indicated significant differences, linear contrast techniques were used to make statistical comparisons of the group mean AUCs, using the R statistics package, in an effort to identify the variable selection method that produced on average the highest AUC (Ramsey and Schafer, 2002; Figures 11 and 12).

3.4.2 Comparing Variable Importance

Using the procedure described above, a variety of models were produced to examine the sensitivity of HSMs to a variety of inputs and the resulting differences in variable importance. Variable importance was ranked within MAXENT by examining variable rank output tables which show the influence of each variable. Most of these variable ranking procedures use training or test gain. Each iteration of the model attempts to improve model fit, measured as gain, by varying the coefficients (λ) associated with each feature (function of environmental variables). Gain is basically a likelihood or deviance statistic that is used to maximize the probability of the occurrence data relative to the background (see section 2.4 for more detail; Phillips, 2006).

The first of the variable importance measures considered was the percent contribution table which is a cumulative measure of training gain. The next three variable importance measures are developed using jackknife tests for training gain, test gain and contribution to AUC (which is the only measure not directly based on gain; Figure 3 and Table 6). The last measure of variable importance was calculated by the author using a manual combination of all of the automated outputs, by ranking each of the four outputs, summing across the rankings for each variable, and re-ranking the order (Table 7). Comparisons of these various scenarios based on the variables selected by the above procedures were then examined to determine which scenarios performed best in terms of AUC.

3.5 **Factors**

The specific factors considered included background extent, data resolution, sample size, and variable rank criteria. Systematic combinations of these factors were produced and are referred to as scenarios for comparison.

3.5.1 Study Area Extent

The study area mentioned above included the area within a one mile buffer of the Nez Perce NF boundary. However, the known occurrences of Broad-fruit Mariposa are confined to the south-westernmost portion of the Nez Perce NF. The steep environmental gradients found in this area were vastly different ecological conditions than in the northeast portion of the Nez Perce NF. This prompted concerns regarding the ability to extrapolate the modeling effort into these disparate environments that was addressed by others (Walter, in prep). As the aim of this project was to assess variable importance, the analysis was constrained to that area known to support populations of Broad-fruit Mariposa. Figure 2 shows the full extent of the Nez Perce NF and the smaller area used for most of the analysis. This interpolative approach should have avoided some of the uncertainty that may have been induced by considering those areas with different environmental gradients (Guisan and Zimmermann, 2000; Hirzel and Le Lay, 2008).

Full models including all 37 environmental variables were produced at the coarse resolution for the initial, large study area as well as for the reduced study area. While calculated, AUC is dependent upon study area extent and thus comparisons of AUC from different extents are not valid. These models were compared in terms of variable importance at the two extents (Figures 7 and 8). More refined models with fewer

variables are discussed separately to examine the effects of methods to reduce the number of variables.

3.5.2 Fine vs. Coarse Resolution

As discussed in section 3.1.1, two modeling resolutions were utilized, one fine (30×30 m) and one coarse (200×200 m). A series of runs were produced for the reduced study area at both the fine and coarse resolutions using the full set of 37 potential variables.

3.5.3 Using Intermediate Variable Importance to Reduce the Number of Variables

The intermediate models allow for evaluation of the variable importance and reduction of the number included. Two subsets of variables were selected from the full model results from both the fine and coarse resolution output for further development.

The first was based upon the top ten environmental variables, from the full model, as ranked by just the variable contribution ranking (scenario 1). The second was determined by ranking the 37 environmental variables based on their influence on training gain, test gain, and effect on model AUC, when a variable is removed from the model and when considered alone as determined by the variable jackknife procedure (Table 6). The ten variables with the highest ranks from the combination of these three jackknife outputs represent the second model (scenario 2).

The same procedure was used to build intermediate models at the coarse resolution (scenario 3) and (scenario 4). The performance of these models with an intermediate number of variables, were compared to assess the effect of these types of variable-ranking procedures on model accuracy (Figures 9 and 10).

Five additional models were constructed at both resolutions using the same procedure described above, but using only the top four variables from the full model as identified by the variable rank outputs. At the fine resolution, the models developed using test gain and the combination of all outputs both identified the same four variables ranked in the same order. This yielded four unique models for comparison based on the following outputs: percent contribution (scenario 5), training gain (scenario 6), test gain and combined (scenario 7), and contribution to AUC (scenario 8; Figure 11).

At the course resolution, the models developed using test gain, contribution to AUC, and the combination of all outputs all identified the same four variables ranked in the same order. This yielded three unique models for comparison based on the following outputs: percent contribution (scenario 9), training gain (scenario 10), test gain, contribution to AUC and combined (scenario 11; Figure 12).

Given the number of scenarios developed, descriptions of each scenario are .included in the comprehensive Table 8, which also shows the resulting statistics.

3.5.4 Occurrence Data Resolution

One of the goals of modeling at two resolutions was to explore the effect of resolution on model accuracy and variable importance. This comparison could not be fairly made using scenarios discussed earlier as the fine and coarse models sample locations were not the same. Addressing this question required building a model that utilized the 200 m occurrence data (n=119) and the 30 m environmental data (scenario 12). Comparing this model's output with the original coarse model's output (scenario 13) should isolate the effect of scale and allow or the evaluation of the effect of grain size on

variable importance and model accuracy (Figure 14; Engler et al., 2004; Guisan et al., 2007).

3.6 **Additional Evaluations**

3.6.1 Logistic Regression

Thus far in this research, MAXENT's variable selection criteria had only been evaluated in isolation because the literature indicates that the regularization method performs very well (Dudik et al., 2004; Tibshirani, 1996). Elith et al. (2011) indicate that variable selection utilizing regularization is much better than other commonly used step-wise selection procedures. However, in order to further test the assortment of variable selection procedures, a simple logistic-regression model with no interaction was built for the 200 m occurrence data that included all 37 variables. This logistic-regression model was built using 119 presence records and 10,000 absence records selected randomly from the background pixels (Engler et al., 2004; VanDerWal et al, 2009). The environmental variable values were sampled at each presence and absence location and written to tabular form. This tabular data was used to perform the logistic regression in the statistical analysis package R (www.r-project.org). From the resulting logistic regression, R's stepwise AIC function was used to select the most significant variables based on both a forward and backward stepwise AIC, (Akaike, 1974). A MAXENT model built using the variables identified using the logistic regression and AIC stepwise procedure (scenario 14), was then compared to a model with the same number of variables but identified by MAXENT using only linear features to make it comparable in complexity to the logistic model as possible (scenario 15). The goal of this comparison was to identify

which of these two variable selection methods produced the best model in terms of AUC (Figure 15).

3.6.2 Comparison of MAXENT to Domain

A comparison of the total area predicted as suitable Broad-fruit Mariposa habitat by Nock (2007) using the Domain algorithm to the area predicted by MAXENT was conducted to assess the relative effectiveness of the two procedures. The Domain model produces a binary output identifying areas as either suitable or unsuitable habitat. MAXENT's logistic output produces a continuous probability surface ranging from 0 to 1 with pixel values close to 1 indicating the most suitable habitat, that allows the user more flexibility in establishing particular binary thresholds (Phillips et al., 2008). The Domain model was run with seven environmental variables at a 60×60 m resolution. The model used in this comparison used the four best variables from the 200×200 m resolution for the small study area (Scenario 9). This particular scenario was selected because it uses a comparable number of occurrence records and fewer environmental variables. The predicted areas used in this comparison correspond to the minimum and maximum extents produced by MAXENT's recommended default threshold levels.

4 RESULTS AND DISCUSSION

Results are discussed in the order they were presented in the Methods chapter. For comparison, Table 8 gives a brief description of the parameters used for each scenario while Figure 16 graphs the AUC for each. Note that these only contain the results of runs involving the reduced study area as those are not easily comparable to those done on the full study area (Figure 2).

4.1 Spatial autocorrelation of occurrence data

It was difficult to specifically address how much of the variability present in the 30 bootstrapped model runs was attributable to spatial autocorrelation. Figures 5 and 6 demonstrate the effects of spatial autocorrelation on the fine and coarse resolution models. According to Phillips (2008), the omission and predicted area plot (Figures 5c and 6d) for independent samples should be close to predicted omission (black line) due to the definition of the cumulative threshold. As expected, the test omission is greater than predicted for the run with least spatially auto correlated sample split, and much lower for the most auto correlated run, resulting in inflated accuracy estimates for models in which all the test data is located near clusters of training locations. The worst performing run, of all the scenarios considered, (AUC 0.827) is the result of training/test partitions that grouped the most clustered occurrence records into training sites and the dispersed occurrences into the test sites. Conversely, the run that performed best, of all the scenarios considered, (AUC 0.998) split highly clustered occurrences into training and test sites such that test sites have high spatial autocorrelation with training sites. These

effects are most pronounce in the runs produced using the 30 m occurrences (n=20) but still present in those built from the 200 m occurrence data (n=119, AUC 0.967 and 0.869), for the high and low autocorrelation models respectively.

Many of the environmental variables that were commonly ranked as highly important are also the variables that were determined to be highly correlated with one another. The correlation between these variables not only increases uncertainty regarding the relative importance of these variables but also detracts from the model as they essentially contain no new information.

These results support the need for caution when interpreting accuracy of models with non-random sampling, small sample sizes and highly correlated environmental data. In addition, these result show that increased sample sizes may insulate predictions somewhat from the effect of spatial autocorrelation. While all of the models presented here performed relatively well (AUC better than 0.8). This may be the result of the study area still being overly large or the species prevalence low, and this level of performance will likely vary depending on the species considered, geographical location, and available variables.

4.2 **Study Area Extent**

Comparing the models performance on the large and reduced study areas (Figure 2) demonstrates a limitation of the AUC statistic. The AUC calculated for the full Nez Perce NF (0.990) is larger than that calculated for the reduced study area (0.980) because of the way that AUC is calculated with presence-only data (Fielding and Bell, 1997; Phillips et al., 2006). The use of AUC in evaluating the relative performance of two

presence-only models should be restricted to models having the same geographical extent.

Comparison of the most important variable rankings was very different at these two extents (Figures 7 and 8). The modeled species response to a particular environmental variable appears to be very sensitive to the range of the environmental gradients. When selecting the geographical extent of a study area it is prudent to compare the range of the environmental gradients present in the study area with the range sampled by occurrence data. Failure to do this can result in misleading interpretations of variable importance. In this study, the road proximity layer provides a good example of this. At the large study area extent road proximity is ranked as highly important where at the small extent it has very little influence on the model. The proximity of sites to roads is very different between the road-less wilderness in the northern portion of the Nez Perce and the relatively heavily roaded areas found further south. This is an extreme example of how inclusion of areas with vastly different environmental gradients from those occupied by the species can result in erroneous assessments of variable importance.

It was not surprising that the AUC value was higher for the model of the large study area as this is a well documented characteristic of the AUC statistic. However, the substantial difference in the variables selected as important between models at the two extents was unexpected.

4.3 **Using Intermediate Model Variable Rankings to Reduce Variables Included**

4.3.1 Ten Variable Results

The fine-resolution models created using the variable contribution output and a rank based on combining the training, test, and AUC jackknife outputs to identify the top

ten variables, resulted in models that performed statistically very similarly (scenario 1 and scenario 2, Table 8). A two sample t-test with the null hypothesis being that the there is no difference in mean AUC between the two models (Ramsey and Schafer, 2002) produced a t-statistic of 1.326 and a corresponding one sided p-value of 0.19, indicating that there is no statistical evidence that one model performed better than the other.

The models both ranked elevation, vegetation type, and April minimum temperature as being important, and the areas predicted as suitable are very similar (Figure 9).

The coarse-resolution models created using the variable contribution output and a rank based on combining the training, test, and AUC jackknife outputs to identify the top ten variables, produced similar results as well. A two sample t-test with the null hypothesis being that the there is no difference in mean AUC between the two models produced a t-statistic of 2.092 and a corresponding one sided p-value of 0.04, indicating there is weak statistical evidence that the scenario 3 performed slightly than scenario 4.

Both scenarios ranked elevation, vegetation type, and NDMI as being important, and the areas predicted as suitable are also very similar (Figure 10).

4.3.2 Four Variable Results

Five additional fine-resolution models were created using the variable contribution output and variable jackknife procedures to identify the top four variables. The jackknife of test gain identified the same four variables in the same order as the combined procedure so only that model was used in the analysis for a total of four scenarios: contribution output (scenario 5), training (scenario 6), test and combined (scenario 7), and AUC contribution (scenario 8). Of the resulting models (Table 8),

scenario 7 performed best, with a mean AUC of 0.968, while the worst was scenario 8 with a mean AUC of 0.929. A f-test with the null hypothesis being that the there is no difference in mean AUC between the five models produced a f-statistic of 24.995 and a corresponding p-value much less than 0.00001, indicating that there is convincing statistical evidence that at least one of the means was not equal. The specific hypothesis that scenario 7 was no better than the other four models was tested by the linear combination of group means procedure (Ramsey and Schafer, 2002), resulting in a t-statistic of 5.821 and a corresponding one sided p-value much less than 0.00001, providing convincing evidence that scenario 7 performed best (Table 8). The largest difference in predicted area of these 4-variable models was between scenarios 6 and 8, scenario 8 predicted 6.57% more area as suitable even though the AUC scores were very similar 0.938 and 0.929 respectively (Table 8 and Figure 11b and d).

From a practical standpoint, the three models that incorporated elevation and vegetation type performed well and predicted reasonably similar areas as having suitable habitat. This illustrates some of the problems identified earlier regarding the use of AUC as the sole means of model validation. Specifically demonstrating how two models with very similar AUC scores can predict substantially different areas as suitable.

Five coarse-resolution models were created using various variable importance outputs to identify the top four environmental variables. The jackknife of test gain, AUC, and the combined model identified the same four variables in the same order as the combined jackknife procedure so only the combined model was considered in the analysis. Of the resulting three unique scenarios contribution output (scenario 9), training (scenario 10), and test-combined-AUC (scenario 11), scenario 9 performed best with a

mean AUC of 0.931, while worst was scenario 11 with a mean AUC of 0.923. A f-test with the null hypothesis being that the there is no difference in mean AUC between the three models produced a f-statistic of 4.94 and a corresponding p-value of 0.0092, indicating that there is strong statistical evidence that at least one of the means was not equal. The specific hypothesis that the scenario 9 was no better than the other two models was tested by the linear combination of group means procedure (Ramsey and Schafer, 2002), resulting in a t-statistic of 3.10 and a corresponding one sided p-value of 0.0026, providing statistical evidence that the scenario 9 performed best (Figure 12).

4.3.3 Summary of Variable Rank Impacts

In summary, for the various variable selection methods described above, elevation, May precipitation, NDMI, vegetation type, and April minimum temperature were most commonly ranked as one of the top four important variables. Elevation appears to be highly important in all models and vegetation type is highly important only in fine resolution scenarios

MAXENT's default percent contribution output performed the more complicated jackknife procedure in all cases except for the reduced (4 variable) fine resolution model comparison. In that comparison, scenario 7 performed marginally better statistically than the default percent contribution output but the mean AUC difference was just 0.006. In addition to performing well, the percent contribution output is also the easiest of the outputs to interpret, and should be the default source for variable selection. The jackknife variable importance output can be used to modify the list selected variables, if the objective of the modeling effort is to maximize prediction of test data, or to identify

highly correlated variables, or identify those variables that may not add much to training gain but contain significant information not present in the other variables (Phillips, 2008).

4.4 Occurrence Data Resolution

To examine what the modeling result would be if all of the occurrence records had been collected with a known accuracy of 30 m or better, a comparison of two models that both used the 119 occurrence records, and the 10 most important environmental variables, was conducted. One model utilized environmental variables at 30×30 m resolution (scenario 12) and the other used environmental variables at 200×200 m resolution (scenario 13).

A two sample t-test with the null hypothesis being that the there is no difference in mean AUC between the two models produced a t-statistic of 1.685 and a corresponding one sided p-value of 0.0978, indicating that there is very little statistical evidence that one model performed better than the other (Table 8).

The results of this comparison indicate that the resolution had little impact on the type of variables selected, as the top five variables were the same for both scenario 12 and 13 though they were ranked differently (Figure 14). The Landsat-derived NDMI and NDVI were much more significant contributors in the finer resolution run; this is likely because at this resolution the total variability of these indices was preserved. NDMI, in particular, seems was an important contributor in models at both scales and was often very highly ranked in the fine-resolution runs.

Interestingly, vegetation type was consistently ranked higher in the runs developed using 30 m (n=20) occurrence records and fine-resolution input data (scenario 12). This may be due to the corresponding vegetation type data being classified from

Landsat imagery at that resolution or it could be that the vegetation characteristics are fundamentally more important to this species at that scale. Re-sampling to a coarse resolution also caused grassland pixels near forest pixels to be reclassified as forest. Examination of the vegetation type data at the 30×30 and 200×200 m resolutions revealed many instances of this kind of resampling-induced data degradation. It may also be that due to the few and closely grouped occurrence records available at the fine precision; the majority of the Broad-fruit Mariposa populations fell into one of two vegetation classes (grass and shrub dominated) by chance. While at the coarse resolution there were larger numbers of occurrence records that were more evenly distributed among more classes (grass, shrub, ponderosa, and lodgepole pine dominated) of vegetation making it less valuable as a predictor.

4.5 **Logistic Regression**

The comparison of the model using MAXENT-selected variables (scenario 14) versus the model using AIC-selected variables (scenario 15) suggests that MAXENT did a better job of removing correlated variables and retaining variables that contained unique information (Figure 15). Scenario 14 also performed significantly better than scenario 15, t-statistic 7.398 corresponding one-sided p-value much less than 0.00001, although both performed well (Table 8). This supports the findings of Dudik et al. (2004), Tibshirani (1996), and Elith et al. (2011) that the regularization method of variable selection generally out performs other methods. However, both scenarios included a large number of environmental variables (21), because the AIC-selection procedure could not remove any more variables without the AIC value increasing. Due to the large number of variables that needed to be considered and the relatively small presence

dataset (119) this comparison relied on a very simple implementation of the logistic regression model a more sophisticated comparison might reach different conclusions. However the results presented here suggest that the regularization method of variable selection outperforms AIC-based methods in both flexibility and model performance.

4.6 **Comparison of MAXENT and Domain**

The Domain model using seven environmental variables at the 60×60 m resolution produced a binary output identifying 123,094 ha of suitable habitat for Broad-fruit Mariposa (Nock, 2008). The MAXENT model using only the four best variables, predicted a minimum and maximum area of 23,517 and 126,898 ha, respectively (Table 9). The MAXENT algorithm using a more parsimonious model, in general performed better, predicting much smaller areas as suitable than did Domain. This is advantageous in the case of rare plant species as they usually do not occupy large areas and the more precisely those areas can be delineated the more effective the output will be in aiding conservation managers. Qualitatively the area predicted by the most inclusive MAXENT model and Domain are fairly similar, however the most restrictive MAXENT model identifies a much more specific area as suitable habitat. All of the MAXENT scenarios presented here performed very well with AUCs ranging from 0.987 to 0.923 well above the 0.70 value determined by Elith et al. (2006) to be of value in conservation planning (Figure 16).

5 CONCLUSIONS AND RECOMMENDATIONS

The number of environmental variables available in digital formats appropriate for habitat modeling continues to grow and this fuels the increasing efforts to find new ways to leverage this information into SDMs for a variety of management objectives. This makes the ability to effectively identify important subsets of environmental variables to be considered in predicting rare-plant habitat suitability even more critical. When possible the first step in this type of endeavor should be to gather as much information as possible about the relationships between the potential environmental variables and the target species. With some modification a method similar to the one presented here can provide some insight into the effects of factors such as correlated variables, sample size and sampling, and spatial extents may have on statistically derived variable selections. This is not to suggest that the procedure used here should replace variable selection based on sound ecological understanding of biological requirements of the species under consideration, when those requirements are known and the data is available, rather this method should be used to supplement that approach. It can be used to assist researchers narrow the list of important variables, particularly in those cases where there is a lack of definitive ecological understanding of the species/environment relationship as is often the case with rare species.

Often the environmental variables that have the most direct and proximal effect on rare plant distribution are precisely those variables that are unavailable in the geographical format required by statistical modeling efforts (Austin and Smith, 1989). In these cases, there may be a number of other indirect variables that are strongly associated with a particular rare plant species and a procedure similar to the one presented here, can

be used to ascertain which are statistically most predictive, allowing the development of an initial description of suitable habitat. However, through iterative cycles of data collection and model updating, the resulting output should become progressively better.

5.1 **Conclusions**

Validation of the results presented here has relied heavily on the AUC statistic. For the reasons discussed above, it would be advantageous to supplement this approach with some other validation procedure. Unfortunately these more robust validation procedures generally require presence/absence data. The importance of high quality presence/ absence data for the purpose of validating habitat suitability models has been well documented (Guisan and Zimmerman, 2000; Elith et al., 2006), and it can be particularly important in presence-only modeling efforts where the number of initial assumptions is necessarily large (Phillips et al., 2006), such as with rare species. It would therefore be extremely useful for agencies to develop at least a small set of presence/absence data for those species of rare plants that will likely be included in future modeling efforts. Optimally, these could be used not only to independently test the presence only model's accuracy, but many of the other assumptions as well.

The inclusion of highly correlated variables greatly complicates the interpretation of variable importance (Table 7; Figure 11d). In addition it results in a less parsimonious model that may include many redundant variables. It would be advantageous to identify and include only the most influential of the correlated variables, and exclude the others. The results presented here would have benefitted from the early exclusion of correlated variables that proved to be the least influential in identifying suitable habitat for Broad-fruit mariposa. This would have resulted in a more parsimonious model that took

advantage of as much unique information as possible, while reducing redundancy. The inclusion of additional variables, with the goal of "data mining" for new, unexpected biophysiological relationships, is strongly discouraged in the literature (Guisan and Zimmerman, 2000) and increases the difficulty in achieving accurate, interpretable results from a parsimonious model.

The specific variables identified as being important in defining the distribution of Broad-fruit mariposa habitat were sensitive to a number of factors: study area extent, sample size of occurrence data, variable ranking procedure (contribution, training gain, test gain, AUC maximized), environmental variable resolution, and to a lesser degree, model complexity. Study area extent and the sample size of occurrence data had by far the greatest impact (Table 8; Figure 16). Sensitivity to these factors resulted in output with important variables ranked differently, but the majority of the models rank the following variables as highly important for Broad-fruit Mariposa: elevation, May precipitation, vegetation type, April minimum temperature, NDMI, September precipitation, and July maximum temperature (Figures 8-12). Of these, elevation, vegetation type, and NDMI were among the variables identified as potentially important by USFS botanists (Table 4), strongly supporting the benefit of utilizing expert knowledge when available. It is worth noting that the strong correlation between elevation and predicted habitat may largely be the result of constraints imposed by modern land use practices as agriculture in the lower elevations may restrict Broad-fruit mariposa form inhabiting those areas.

The 30×30 m resolution output tended to rank those environmental variables that were collected at that resolution (vegetation type, NDMI and NDVI) higher than models

run at the coarse resolution, suggesting that vegetation type and NDMI may be extremely valuable in predicting the habitat of Broad-fruit Mariposa. Unfortunately, these results are based on only a few spatially correlated occurrence records and so collection of more data at this fine resolution is needed to confirm this apparent association. Predictions at this fine resolution may also increase the overall accuracy of the results (Engler et al., 2004), yielding predictions with smaller areas being defined as "suitable" that will permit identifying specific locations to conduct future surveys.

The MAXENT algorithm used here allowed many of the recommendations of Nock (2008) to be addressed. In addition, MAXENT appears to have outperformed Domain and offers users greater flexibility of implementation (Table 9). While MAXENT has many features that facilitated its use in this project, there are many other methods which could be used that should perform similarly well; BRT, MARS, and GLM/GAMs using pseudo-absences performed similarly in model comparison studies (Elith et al., 2006). In fact, a number of studies have shown that the selection of environmental variables, the quality of occurrence records, and amount of sampling bias are more important considerations than the type of algorithm used (Guisan and Araujo, 2006; Guisan and Zimmermann, 2000; Hirzel and Le Lay, 2008; Phillips et al., 2008).

The primary benefit of utilizing the bootstrapping to produce a range of models that are trained and tested on different subsets of the occurrence data is that it provides insight into the extent and effects of spatial autocorrelation (Figures 5 and 6). These effects were most pronounced in the fine-resolution models but still present in the coarse-resolution models. Models with small occurrence datasets tended to have high variability in performance while the performance of models with large occurrence datasets seem to

be less variable (Table 8). The bootstrapping procedure used here illustrated the need for caution when interpreting the accuracy of models with non-random sampling and small sample sizes. In addition, these result show that increased sample sizes may insulate models somewhat from the effect of spatial autocorrelation, and may help to identify areas that should be targeted for increased sampling to decrease the effects of spatial autocorrelation.

While the range of performance differs between the models built at the two scales, the areas predicted as suitable were generally the same. The additional areas predicted by the coarse model can be attributed to the larger number of occurrence data available at that resolution (Table 8; Figure 14). This generally consistency in habitat prediction is encouraging as this study only considered one of the many rare-plant species managed by USFS in Region One. Broad-fruit Mariposa was selected primarily due to its relative abundance of data; however, most of the other rare-plant species have an even smaller number of known occurrences. This comparable performance with only 20 data records suggest that the procedures used here should be applicable with only minor modification to the other rare-plant species within Region One, even with low numbers of occurrence data.

This project aimed to test the influence of a variety of factors on the determination of variable importance for the prediction of suitable habitat for Broad-fruit mariposa. The variables identified by this type of statistical inference may have little or no biophysical meaning and should not be used to infer any biophysical relationships. The procedure used here is most appropriate when applied to relatively small areas where the species of interest is known to exist for the purpose of targeting new areas to survey in

the hope of finding new occurrences that can be used to further refine the model. The output of this procedure should not be used to definitively define the species range or extent. In addition it should not be used to extrapolate the occurrence of the species to new environments or areas. Using HSMs for these types of applications are often problematic and when attempted should utilize randomly sampled presence/ absence occurrence data and the inclusion of environmental variables with strong biophysical connections to the target species.

5.2 **Recommendations for Habitat Suitability Models**

The following lists are specific recommendations aimed at assisting managers to A): select variables to be considered in the modeling effort, and B): some general considerations for modeling rare-plant habitat.

5.2.1 Variable Selection

- Selection of variables should be based on the expert's ecological knowledge of species when possible and basic ecological principles when little species-specific information is known (Guisan and Zimmerman, 2000; Section 2.4). Inclusion of variables related to disturbances (from roads, fire, etc.) should only be done when there is a previous expectation of an ecological relationship to that disturbance.
- The use of a large number of variables is undesirable, particularly when correlated, effort should be given to the statistical determination of which of these correlated variables is most influential, and only those should be utilized for the main model. See earlier discussion of variable selection procedures in Section 2.4 and 3.1.2 for discussion.
- Outputs ranking variable importance can be an effective tool for narrowing down the list of potential variables to be included in modeling efforts (Section 4.3; Dud´ık et al., 2004).
- The specific ranking of these variables is sensitive to numerous factors, is made more difficult when variable are correlated, and should therefore be regarded skeptically (Section 5.1).

- Causal relationships cannot be inferred from the statistical associations identified between occurrence locations and environmental variables (Franklin, 1995)

- While in the field, experts should attempt to assess the validity of environmental variables that could be used in modeling efforts and identify variables that should be incorporated in the future (Section 3.1.2).

5.2.2 General recommendations for developing and implementing HSMs

- Continue to encourage the development of more proximal environmental data at finest resolution possible (Section 3.1.2).

- Do not restrict models or data availability to administrative boundaries as this allows algorithms to take advantage of the largest possible occurrence datasets and more thoroughly describe the species/environment relationship (Section 3.1.2).

- Future data collection should stress locational precision using differentially corrected GPS data to the maximum extent possible (Engler et al., 2004; Section 3.1.1).

- Replicate models can quantify the variability in model performance, particularly when sample size is small and data is spatially correlated (Section 3.4).

- The HSM for a particular species should be iteratively updated as new occurrence and environmental data becomes available (Section 5.1).

- HSMs that provide continuous probability predictions offer more flexibility in the selection of thresholds can be used to focus and improve the efficiency of future field surveys (Phillips et al., 2006).

- Use independent presence/ absence occurrence data for validation of future modeling efforts (Section 3.4.1).

- Conduct field surveys in an effort to ground truth model results (Nock, 2008).

Future research that may improve habitat-suitability modeling of rare-plant species found within Region One of the USFS include a formal examination of the effect of environmental variable resolution on model performance with a larger high-precision occurrence data set. To more accurately assess presence-only models, at least a small set

of presence/ absence data should be developed for those species of rare plants that will likely be included in any future modeling efforts. Conduct field surveys of those areas determined to be the most suitable in an effort to ground truth the results of all HSMs. Compare the result of those efforts to a simple model driven by only biophysically relevant data. Further testing of the procedures used here on a variety different species and environments would help to determine its overall applicability. Further examination of the effects of spatial autocorrelation on model performance would also be useful.

BIBLIOGRAPHY

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Auto. Cont. 19 (6), 716–723.

Austin, M.P., 1980. Searching for a model for use in vegetation analysis. Vegetatio. 42 (1), 11-21.

Austin, M. P., Smith, T. M., 1989. A New Model for the Continuum Concept. Vegetatio. 83 (1), 35-47.

Austin, M P., 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. Ecol. Model.157 (2-3), 101-118.

Berger, A., Della Pietra, S., Della Pietra, V., 1996. A Maximum Entropy Approach to Natural Language Processing. Comp. Linguistics 22 (1), 39-71.

Beers, T.W., Dress, P.E., Wensel, L.C., 1966. Aspect transformation in site productivity research. J. For. 64 (10), 691-692.

Beven, K. J., Kirkby, M. J., 1979. A physically based, variable contributing area model of basin hydrology. Hydrol. Sci. Bull. 24 (1), 43–69.

Brotons, L., Thuiller, W., Araujo, M.B., Hirzel, A.H., 2004. Presence-Absence Versus Presence-Only Modeling Methods for Predicting Bird Habitat Suitability. Ecography 27 (4), 437-448.

Busby, J.R., 1991. Bioclim – A bioclimatic analysis and prediction system. *Nature Conservation: Cost Effective Biological Surveys and data Analysis* (Eds). C.R. Margules and M.P. Austin, 64–68. CSIRO, Melbourne, Australia.

Carpenter, G., Gillison, A.N., Winter, J., 1993. DOMAIN: A Flexible Modeling Procedure for Mapping Potential Distributions of Plants and Animals. Biodivers. Conserv. 2 (6), 667-680.

Chander, G., Markham, B. L., Barsi, J. A., 2007. Revised Landsat-5 Thematic Mapper radiometric calibration. IEEE Geos. Rem. Sens. Lett. 4 (3), 490−494.

Dud´ık, M., Phillips, S.J., Schapire, R.E.., 2004. *Performance guarantees for regularized maximum entropy density estimation.* In: Proceedings of the 17th Annual Conference on Computational Learning Theory. ACM Press, New York, 655–662.

Ebdon, D., 1985. *Statistics in Geography. (2nd Ed.).* Oxford. Blackwell.

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., McC. Overton, J., Peterson, A.T., Phillips, S.J., Richardson, K., Scachetti-Pereira, R., Schapire, R.E., Soberon, J., Williams, S., Wisz, M.S., Zimmermann, N.E., 2006. Novel Methods to Improve Prediction of Species' Distributions from Occurrence Data. Ecography 29 (2), 129-151.

Elith, J., Phillips, S.J., Hastie, T., Dud´ık, M., Yung, E.C., Yates, C.J., 2011. A statistical explanation of MaxEnt for ecologists. Divers. Distrib.17 (2), 43–57.

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. J. Appl. Ecol. 41 (2), 263–274.

Environmental Systems Research Institute, Inc., 2009. ArcMap v9.3.1, Spatial Analyst Extension. Redlands, California.

Feilding, A.H., Bell J.F., 1997. A Review of Methods for the Assessment of Prediction Errors in Conservation Presence/Absence Models. Environ. Conserv. 24 (2), 38-49.

Franklin, J. (1995) Predictive vegetation mapping: geographic modelling of biospatial patterns in relation to environmental gradients. Prog. Phys. Geogr., 19(4), 474–499.

Friedman, J. H., Hastie, T., Tibshirani, R., 2000. Additive logistic regression: a statistical view of boosting. Ann. Stat. 28 (2), 337-407.

Grinnell, J., 1917 Field tests of theories concerning distributional control. Am. Nat. 51 (602), 115–128.

Guisan, A., Zimmermann, N.E., 2000. Predictive Habitat Distribution Models in Ecology. Ecol. Model. 135 (2-3), 147-186.

Guisan, A., Thuiller, W., 2005. Predicting species distribution: offering more than simple habitat models. Ecol. Lett. 8 (9), 993-1009.

Guisan, A., Araujo, M. B., 2006. Five (or so) challenges for species distribution modelling. J. Biogeogr. 33(10), 1677-1688.

Guisan, A., Graham, C.H., Elith, J., Huettmann, F. Group, N.S.D.M., 2007. Sensitivity of predictive species distribution models to change in grain size. Divers. Distrib.13 (3), 332-340.

Hays, M., 2010. (Field notes) from: Idaho Fish and Game, 2010. Plant Conservation Database, accessed Aug 2010.

Hays, M., 2011. (Personal communication).

Hastie, T., Tibshirani, R., 1987. Generalized additive models: some aplications. J. Am. Stat. Assoc. 82 (398), 371–386.

Hastie, T., Tibshirani, R., 1990. Generalized additive models, Chapman & Hall : New York.

Heikkinen, R.K., Luoto, M., Araújo, M.B., Virkkala, R., Thuiller, W., Sykes, M.T., 2006. Methods and uncertainties in bioclimatic envelope modelling under climate change. Progress in Physical Geography 30 (6) 751–777.

Hitchcock, C. L., Cronquist, A., Ownbey, M., Thompson, J. W., 1955-1969.Vascular Plants of the Pacific Northwest. Vols. 1-5. Univ. Washington Press, Seattle.

Hirzel, A.H., Hausser, J., Perrin, N., 2000. BIOMAPPER 1.0 – A new software to compute habitat-suitability maps. Laboratory for Conservation Biology, University of Lausanne, Switzerland.

Hirzel, A., Hausser, J., Chessel, D., Perin, N., 2002. Ecological-Niche Factor Analysis: How to Compute Habitat-Suitability Maps Without Absence Data. Ecology 83 (7), 2027-2036.

Hirzel, A.H., Le Lay, G., 2008. Habitat suitability modelling and niche theory. J. App. Ecol. 45 (5), 1372–1381.

Hughes, J. D., 1975. Ecology in ancient Greece. Inquiry 18 (2), 115-128.

Hutchinson, G.E., 1957. Concluding remarks, Cold Spring Harbor Symp. Quant. Biol. 22, 415–427.

Idaho Fish and Game, 2010. Plant Conservation Database, accessed Aug 2010. Available from: https://fishandgame.idaho.gov/ifwis/portal/page/obtain-information.

Jaynes, E.T., 1957. Information theory and statistical mechanics. Phys. Rev. 106 (4), 620–630.

Jensen, M.E., Haise, H.R., 1963. Estimating evapotranspiration from solar radiation. J. Irrig. Drainage Div. Am. Soc. Civil Eng 89, (4) 15-41.

Johnson, J.B., Omland, K.S., 2004. Model selection in ecology and evolution. Trends Ecol. and Evol. 19 (2), 101–108.

Kauth, R.J., Thomas, G.S., 1976. The tasseled Cap -- A Graphic Description of the Spectral-Temporal Development of Agricultural Crops as Seen by LANDSAT." Proceedings of the Symposium on Machine Processing of Remotely Sensed Data, Purdue University of West Lafayette, Indiana, 4B-41- 4B-51.

Kelly, R. L., 1983. Hunter-Gatherer Mobility Strategies, J. Anthro.Res. 39 (3), 277-306.

Leathwick, J., Rowe, D., Richardson, J., Elith, J., Hastie. T., 2005. Using multivariate adaptive regression splines to predict the distributions of New Zealand's freshwater diadromous fish. Freshwater Biol. 50 (12), 2034–2052.

Legendre, P., 1993. Spatial autocorrelation: trouble or new paradigm? Ecology 74 (6), 1659–1673.

Liu, H., Huete, A.R., 1995. A feedback based modification of the NDVI to minimize canopy background and atmospheric noise. IEEE Trans. Geos. Rem. Sens.33 (2), 457– 465.

Moore, I. D., Grayson, R. B., Ladson, A. R., 1991. Digital terrain modeling – a review of hydrological, geomorphological, and biological applications, Hydrol. Proc. 5 (1), 3–30.

National Environmental Policy Act, 1986. 40 CFR Parts 1500-1508.

Nock, E.E., 2008. A simple GIS approach to predicting rare-plant habitat: North central Rocky Mountains, United States Forest Service, Region One. M.S. Thesis, The University of Montana.

Parolo, G., Rossi, G., Ferrarini, A., 2008. Toward improved species niche modelling: Arnica montana in the Alps as a case study. J. App. Ecol. 45 (5), 1410-1418.

Parviainen, M., Luoto, M., Ryttäti, T., Heikkinen, R.K., 2008. Modelling the occurrence of threatened plant species in taiga landscapes: methodological and ecological perspectives. J. Biogeogr. 35 (10), 1888-1905.

Phillips, S. J., Anderson, R. P., Schapire, R. E., 2006. Maximum entropy modeling of species geographic distributions. Ecol. Model. 190 (3-4), 231-259.

Phillips, S. J., 2008. Transferability, sample selection bias and background data in presence-only modeling: a response to Peterson et al. (2007). Ecography 31 (2), 272-278.

Phillips, S. 2008. A Brief Tutorial on Maxent. Available from: www.cs.princeton.edu/~schapire/maxent/tutorial/tutorial.doc.

Phillips, S.J., Dudı´k, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J., Ferrier, S., 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecol. Appl. 19 (1), 181–197.

PRISM Group, Oregon State University, accessed March, 2010. Available from: http://prism.oregonstate.edu/.

Pulliam, H.R., 2000. On the relationship between niche and distribution. Ecol. Lett. 3 (4), 349-361.

Ramsey, F. L., Shafer, D. W., 2002. *The Statistical Sleuth: A Course in Methods of Data Analysis* (2nd ed.). Pacific Grove, CA: Duxbury.

Reinhart, S., 2008. (Personal communication).

Schwarz, G. E., 1978. Estimating the dimension of a model. Ann. Stat. 6 (2), 461–464

Sergio, C., R. Figueira, Draper, D., Menezes, R., Sousa, A.J., 2007. Modelling bryophyte distribution based on ecological information for extent of occurrence assessment. Biol. Conserv. 135 (3),341–351.

Shelly, S., 2011. (Personal communication).

Stockwell, D., Peters, D., 1999. The GARP Modeling System: Problems and Solutions to Automated Spatial Prediction. Int. J. Geogr. Inf. Sci. 13 (2), 143-158.

Ter Braak, C.J.F., 1986, Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology 67 (5), 1167–1179.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J.R. Stat. Soc. B 58 (1), 267–288.

United States Department of Agriculture, 2005. National Forest System LandManagement Planning 36 CFR Part 219: Final Rule. Federal Register vol.3.United States Department of Agriculture.

United States Department of Agriculture Forest Service Geospatial Group, 2006. VMAP Tree Dominance Type, accessed April 2006. Available from: http://www.fs.fed.us/r1/gis/thematic_data/vmap_treedominancetype.

United States Department of Agriculture Forest Service, 2007. (Draft) Nez Perce National Forest Proposed Land Management Plan. Available from: http://www.fs.fed.us/cnpz/forest/documents/index.shtml.

United States Department of Agriculture Forest Service Geospatial Group, 2007. 30-meter Digital Elevation Model. Unpublished data.

United States Department of Agriculture Forest Service, 2011. Regional Forester's Sensitive Species List. Unpublished document.

United States Department of Agriculture Forest Service Geospatial Group, 2010. accessed May 2010. Available from: http://www.fs.fed.us/r1/gis/thematic_data.

United States Fish and Wildlife Service, 1988. Endangered Species Act of 1973 as amended through the 100th Congress. U.S. Department of the Interior, Fish and Wildlife Service, Washington, D.C., USA.

United States Geological Survey, 2010. Accessed May 2010. Available from: http://www. glovis.usgs.gov.

Van DerWal, J., Shoo, L. P., Graham, C., Williams, S. E., 2009. Selecting pseudo-absence data for presence-only distribution modelling: how far should you stray from what you know? Ecol. Model. 220 (4), 589-594.

Van Manen, F., 2010. Average relative solar insolation using hillshade functions,USGS LSC-SAFB. Accessed January 10, 2010. Available from: http://www.lsc.usgs.gov/gis/data/shen/LandscapeGradients.asp

Wainwright, J., Mulligan, M., 2004. *Environmental Modelling: Finding Simplicity in Complexity*, Wiley, Chichester.

Walker, P.A., Cocks, K.D., 1991. HABITAT: a procedure for modeling a disjoint environmental envelope for a plant or animal species. Glob. Ecol. Biogeogr. Lett. 1 (4), 108–118.

Walter, C.A., In Prep. M.S. Thesis, The University of Montana.

Williams, J. N., Seo, C., Thorne, J., Nelson, J. K., Erwin, S., Schwartz, M. W., 2009. Using species distribution models to predict new occurrences for rare plants. Divers. Distrib.15 (4), 565-576.

Willmott, C.J., 1984. *On the evaluation of model performance in physical geography. Spatial Statistics and Models*, Hingham, Massachusetts.

Wilson, E. H, Sader ,S.A., 2002. Detection of forest harvest type using multiple dates of Landsat TM imagery. Rem. Sens. Environ. 80 (3), 385–396.

Whittaker, R. H., 1967. Gradient Analysis of Vegetation, Biol. Rev. 42 (2), 207-264.

Zimmermann, N. E., Edwards, T. C., Moisen, G. G., Frescino, T. S., Blackard, J. A., 2007. Remote sensing-based predictors improve distribution models of rare, early successional and broadleaf tree species, In Utah. J. App. Ecol. 44 (5), 1057-1067.

Table 1. Species list used as input in DOMAIN habitat predictions (Nock, 2008). Species considered in the current study are highlighted.

Species		State	State	Number of
Common Name	***Scientific Name***	**Listed**	**Rank**	**Occurrences**
Broad-fruit Mariposa	*Calochortus nitidus*	ID	S3	261
Constance's bittercress	*Cardamine constancei*	ID	S3	74
Evergreen kittentail	*Synthyris platycarpa*	ID	S3	83
Idaho Douglasia	*Douglasia idahoensis*	ID	S2	20
Idaho strawberry	*Waldsteinia idahoensis*	ID	S3	45
Pacific dogwood	*Cornus nuttallii*	ID	S1	99
Payson's milkvetch	*Astragalus paysonii*	ID	S3	190
Puzzling halimolobos	*Halimolobos perplexa*	ID	S3	42
Clustered lady's slipper	*Cypripedium fasciculatum*	ID/MT	S2 S3	81
Tapered-root orogenia	*Orogenia fusiformis*	MT	S2	69
Coville Indian paintbrush	*Castilleja covilleana*	MT	S2	86
Hall's rush	*Juncus hallii*	MT	S2	24
Hollyleaf clover	*Trifolium gymnocarpon*	MT	S2	47
Howell's gumweed	*Grindelia howellii*	MT	S2 S3	100
Jove's buttercup	*Ranunculus jovis*	MT	S2	27
Lemhi beardtongue	*Penstemon lemhiensis*	MT	S3	153
Missoula phlox	*Phlox kelseyi missoulensis*	MT	S2	25
Northern rattlesnake-plantain	*Goodyera repens*	MT	S2 S3	133
Sapphire rockcress	*Arabis fecunda*	MT	S2	43
Short-styled colombine	*Aquilegia brevistyla*	MT	S2	47
Small onion	*Allium parvum*	MT	S2 S3	102

Table 2. Percent of known occurrence data contained in DOMAIN's predicted area (Nock, 2008) . (Species considered in the current study highlighted)

Species	New Occurrences	Known Occurrence Accuracy	Acreage (Hectares) of Predicted Habitat in Nez Perce NF
Broad-fruit Mariposa	2	95	304,173 (123,094)
Constance's bittercress	5	100	552,866 (223,736)
Evergreen kittentail	2	83	586,778 (237,460)
Idaho Douglasia	0	95	179,061 (72,463)
Idaho strawberry	11	98	531,617 (215,137)
Pacific dogwood	4	100	552,866 (223,736)
Payson's milkvetch	0	92	596,401 (241,354)
Puzzling halimolobos	0	83	385,235 (155,899)
Clustered lady's slipper	0	88	542,224 (219,430)
Tapered-root orogenia	16	84	67,624 (27,366)
Coville Indian paintbrush	1	67	75,823 (30,684)
Hall's rush	Not surveyed	83	4,964 (2,008)
Hollyleaf clover	1	83	5,539 (2,241)
Howell's gumweed	Not surveyed	90	2,145 (868)
Jove's buttercup	Not surveyed	96	557 (225)
Lemhi beardtongue	1	43	3,391 (1,372)
Missoula phlox	0	52	1,478 (598)
Northern rattlesnake-plantain	0	89	11,287 (4,567)
Sapphire rockcress	Not surveyed	76	196 (79)
Short-styled colombine	Not surveyed	100	100 (40)
Small onion	1	74	26,950 (10,906)

Table 3. Euclidean distance between occurrences.

	Minimum (m)	Mean (m)	Maximum (m)
30 m	34	10383	27705
200 m	112	19381	63862

Table 4. List of environmental variables initially considered (variables highlighted in red were removed as they were mapped at a scale too course for the resolution of the modeling effort). Bold text are those identified by experts as important, underlined variables represent variables with a clear biophysical relationship to plant species, italicized variables represent climactic variables only the most significant of which should be included, Variables with an asterisk represent variables commonly used in other modeling efforts or related to disturbance that may prove important to rare plants.

	Variable	Resolution	Description	Resampling	Source
Climatic	*April Precip*	800 × 800 m	30 yr (71-00) climate data	Bi-linear interpolation	PRISM 2010
	May Precip	800 × 800 m	30 yr (71-00) climate data	Bi-linear interpolation	PRISM 2010
	June Precip	800 × 800 m	30 yr (71-00) climate data	Bi-linear interpolation	PRISM 2010
	July Precip	800 × 800 m	30 yr (71-00) climate data	Bi-linear interpolation	PRISM 2010
	August Precip	800 × 800 m	30 yr (71-00) climate data	Bi-linear interpolation	PRISM 2010
	September Precip	800 × 800 m	30 yr (71-00) climate data	Bi-linear interpolation	PRISM 2010
	April Max Temp	800 × 800 m	30 yr (71-00) climate data	DEM aided interpolation	PRISM 2010
	May Max Temp	800 × 800 m	30 yr (71-00) climate data	DEM aided interpolation	PRISM 2010
	June Max Temp	800 × 800 m	30 yr (71-00) climate data	DEM aided interpolation	PRISM 2010
	July Max Temp	800 × 800 m	30 yr (71-00) climate data	DEM aided interpolation	PRISM 2010
	August Max Temp	800 × 800 m	30 yr (71-00) climate data	DEM aided interpolation	PRISM 2010
	Sept Max Temp	800 × 800 m	30 yr (71-00) climate data	DEM aided interpolation	PRISM 2010
	April Min Temp	800 × 800 m	30 yr (71-00) climate data	DEM aided interpolation	PRISM 2010
	May Min Temp	800 × 800 m	30 yr (71-00) climate data	DEM aided interpolation	PRISM 2010
	June Min Temp	800 × 800 m	30 yr (71-00) climate data	DEM aided interpolation	PRISM 2010
	July Min Temp	800 × 800 m	30 yr (71-00) climate data	DEM aided interpolation	PRISM 2010
	August Min Temp	800 × 800 m	30 yr (71-00) climate data	DEM aided interpolation	PRISM 2010
	Sept Min Temp	800 × 800 m	30 yr (71-00) climate data	DEM aided interpolation	PRISM 2010
	Spring Mean Precip	800 × 800 m	30 yr (71-00) climate data	DEM aided interpolation	PRISM 2010
	Summer Mean Precip	800 × 800 m	30 yr (71-00) climate data	DEM aided interpolation	PRISM 2010
	Spring Max Temp	800 × 800 m	30 yr (71-00) climate data	DEM aided interpolation	PRISM 2010
	Summer Max Temp	800 × 800 m	30 yr (71-00) climate data	DEM aided interpolation	PRISM 2010
	Spring Min Temp	800 × 800 m	30 yr (71-00) climate data	DEM aided interpolation	PRISM 2010
	Summer Min Temp	800 × 800 m	30 yr (71-00) climate data	DEM aided interpolation	PRISM 2010
Topographic	**Elevation**	30 × 30 m	DEM	200m = mean of 30m pix.	USDA 2007
	Slope*	30 × 30 m	DEM derived	200m = mean of 30m pix.	USDA 2007
	Aspect*	30 × 30 m	DEM derived	200m = mean of 30m pix.	USDA 2007
	Beer's Aspect	30 × 30 m	1+cos((45°-aspect)div deg)	200m = mean of 30m pix.	USDA 2007
	Topo. Wetness Index	30 × 30 m	DEM derived	200m = mean of 30m pix.	USDA 2007
	Solar radiation	30 × 30 m	DEM derived	200m = mean of 30m pix.	USDA 2007
	August Pot.ET	30 × 30 m	Solar rad. & Temp derived	200m = mean of 30m pix.	PRISM 2010
Other	**Dominant Veg.Type**	30 × 30 m	Dominant Vegetation Type	200m = reclass 30m pix.	USDA 2010
	NDVI	30 × 30 m	(NIR-RED)/(NIR+RED)	200m = mean of 30m pix.	USGS 2010
	NDMI	30 × 30 m	(NIR-MIR)/(NIR+MIR)	200m = mean of 30m pix.	USGS 2010
	Fire History*	1:24000	1988-2010 Fires	Polygon to 30m raster	USDA 2010
	Timber Harvest Hist*	1:24000	1980-2010 Mech.Treat.	Polygon to 30m raster	USDA 2010
	Road Proximity*	1:24000	Distance to nearest road	Polygon to 30m raster	USDA 2010
	Climate Zones*	1:100000	Bailey's land units	Polygon to 30m raster	USDA 2010
	Geologic Material*	1:100000	Land type associations	Polygon to 30m raster	USDA 2010
	Ecological subregions*	1:100000	Ecological subregions	Polygon to 30m raster	USDA 2010
	Geomorphology	1:100000	Land type associations	Polygon to 30m raster	USDA 2010
	Soil	1:100000	Land type associations	Polygon to 30m raster	USDA 2010

Table 5. Correlation between the top ten variables from the fine and coarse resolution models (not in rank order, and bolded items indicate highly correlated variables).

30m reduced model correlation matrix

	apr_mint	**aug_ppt**	beers_aspt	**elev**	jul_maxt	may_ppt	ndmi	ndvi	slope	veg_type
apr_mint	1.000									
aug_ppt	**-0.865**	1.000								
beers_aspt	-0.023	-0.028	1.000							
elev	**-0.888**	0.754	-0.006	1.000						
jul_maxt	**0.906**	-0.755	-0.031	**-0.961**	1.000					
may_ppt	-0.701	**0.873**	-0.011	0.538	-0.610	1.000				
ndmi	-0.461	0.470	0.164	0.474	-0.511	0.476	1.000			
ndvi	-0.455	0.430	0.209	0.480	-0.532	0.479	**0.859**	1.000		
slope	0.388	-0.363	0.009	-0.315	0.341	-0.336	-0.193	-0.196	1.000	
veg_type	-0.526	0.511	0.156	0.545	-0.565	0.473	0.707	0.709	-0.169	1.000

200m reduced model correlation matrix

	apr_mint	**aug_ppt**	beers_aspt	elev	jul_maxt	may_ppt	**ndmi**	ndvi	slope	veg_type
apr_mint	1.000									
aug_ppt	**-0.864**	1.000								
beers_aspt	-0.026	-0.026	1.000							
elev	**-0.886**	0.754	-0.004	1.000						
jul_maxt	**0.911**	-0.755	-0.030	**-0.962**	1.000					
may_ppt	-0.698	**0.872**	-0.008	0.537	-0.607	1.000				
ndmi	-0.470	0.478	0.174	0.483	-0.522	0.482	1.000			
ndvi	-0.461	0.436	0.223	0.487	-0.543	0.486	**0.865**	1.000		
slope	0.415	-0.390	0.003	-0.340	0.357	-0.361	-0.212	-0.217	1.000	
veg_type	-0.524	0.510	0.164	0.544	-0.568	0.474	0.713	0.713	-0.189	1.000

Table 6. Examples of jackknife variable ranking procedures, gain sorted in ascending order for runs without a particular variable and descending for runs with only a particular variable.

Test gain without variable	Gain	rank	Test gain with only variable	Gain	Rank
elev200	1.7641	1	sep_maxt200	0.6643	1
ndmi200	1.7956	2	elev200	0.6593	2
aug_ppt200	1.8532	3	summer_maxt200	0.6432	3
fire_hist200	1.8536	4	jun_maxt200	0.6334	4
may_ppt200	1.8603	5	aug_maxt200	0.6248	5
road_prox200	1.8623	6	jul_maxt200	0.6212	6
apr_maxt200	1.8642	7	spring_maxt200	0.5893	7
summer_ppt200	1.8664	8	apr_maxt200	0.5777	8
spring_mint200	1.8732	9	may_maxt200	0.572	9
ndvi200	1.8747	10	may_ppt200	0.5565	10
solar_avg200	1.8752	11	apr_ppt200	0.4779	11
veg_type200	1.8755	12	apr_mint200	0.4451	12
aug_ept200	1.8783	13	sep_ppt200	0.4386	13
sep_maxt200	1.8803	14	spring_ppt200	0.4192	14
aug_maxt200	1.8846	15	may_mint200	0.4171	15
jun_ppt200	1.8849	16	jun_ppt200	0.417	16
twi200	1.8855	17	spring_mint200	0.4009	17
jul_mint200	1.8856	18	jul_ppt200	0.3539	18
slope200	1.8859	19	summer_ppt200	0.3515	19
spring_ppt200	1.8869	20	jun_mint200	0.3427	20
may_mint200	1.8879	21	aug_ppt200	0.3261	21
sep_mint200	1.8898	22	aug_ept200	0.2379	22
jul_ppt200	1.8903	23	jul_mint200	0.1839	23
jul_maxt200	1.8919	24	slope200	0.1684	24
aug_mint200	1.8954	25	aug_mint200	0.164	25
jun_maxt200	1.896	26	summer_mint200	0.157	26
apr_mint200	1.897	27	ndmi200	0.1564	27
summer_mint200	1.8982	28	sep_mint200	0.1556	28
beers_aspect200	1.9006	29	solar_avg200	0.1356	29
jun_mint200	1.9053	30	road_prox200	0.1115	30
spring_maxt200	1.9054	31	ndvi200	0.0996	31
may_maxt200	1.9063	32	veg_type200	0.0666	32
aspect200	1.9092	33	aspect200	0.0411	33
apr_ppt200	1.9103	34	beers_aspect200	0.0348	34
sep_ppt200	1.913	35	fire_hist200	0.0339	35
summer_maxt200	1.939	36	twi200	0.0296	36
treatments200	1.9665	37	treatments200	-0.068	37

Table 7. Top ten variables from the fine and coarse resolution model runs as determined by contribution to training gain, the training, test and AUC jackknife procedure, and the combined procedure.

200m top 10 variables									
cont elev200	1	Training elev200	1	Test elev200	1	AUC elev200	1	elev200	1
cont ndmi200	2	Training may_ppt200	2	Test apr_maxt200	2	AUC apr_maxt200	2	apr_maxt200	2
cont may_ppt200	3	Training apr_maxt200	3	Test may_ppt200	3	AUC may_ppt200	3	may_ppt200	3
cont may_mint200	4	Training jul_maxt200	4	Test sep_maxt200	4	AUC sep_maxt200	4	sep_maxt200	4
cont jul_maxt200	5	Training ndmi200	5	Test aug_maxt200	5	AUC aug_ppt200	5	aug_ppt200	5
cont veg_type200	6	Training sep_maxt200	6	Test aug_ppt200	6	AUC spring_mint200	6	aug_maxt200	6
cont road_prox200	7	Training jun_maxt200	7	Test spring_mint200	7	AUC aug_maxt200	7	spring_mint200	7
cont sep_ppt200	8	Training sep_ppt200	8	Test summer_ppt200	8	AUC apr_mint200	8	jul_maxt200	8
cont apr_mint200	9	Training sum._maxt200	9	Test ndmi200	9	AUC jun_ppt200	9	ndmi200	9
cont aug_ppt200	10	Training may_mint200	10	Test jul_maxt200	10	AUC aug_ept200	10	apr_mint200	10
30m top 10 variables									
cont veg_type	1	Training veg_type	1	Test elev	1	AUC apr_mint	1	elev	1
cont elev	2	Training apr_mint	2	Test apr_ppt	2	AUC sep_ppt	2	sep_ppt	2
cont beers_aspect	3	Training ndmi	3	Test veg_type	3	AUC elev	3	veg_type	3
cont ndvi	4	Training ndvi	4	Test sep_ppt	4	AUC jun_mint	4	apr_ppt	4
cont may_ppt	5	Training sep_ppt	5	Test may_ppt	5	AUC apr_ppt	5	jun_mint	5
cont aug_ppt	6	Training apr_ppt	6	Test summer_ppt	6	AUC aug_maxt	6	may_mint	6
cont apr_mint	7	Training beers_aspect	7	Test spring_mint	7	AUC ndmi	7	ndvi	7
cont road_prox	8	Training elev	8	Test ndvi	8	AUC may_mint	8	apr_mint	8
cont ndmi	9	Training may_ppt	9	Test jun_mint	9	AUC ndvi	9	may_ppt	9
cont aug_mint	10	Training jun_mint	10	Test may_mint	10	AUC veg_type	10	ndmi	10

Table 8. Statistical comparison of models developed to test MAXENT's variable selection outputs, variable resolution, and AIC variable selection methods. *Suitable area was calculated by predicting suitable habitat as a percentage of the study area, as determined by thresholding the outputs at approximately a 10% test omission rate. For Scenarios 1-15, the smaller, reduced study area was used. Baseline models are included at the bottom for comparison. **The full extent model was run over the entire Nez Perce National Forest and should not be compared to the others quantitatively.

Intermediate models, variable importance

Scenario	Occ. Data	Resolution	Ranking method	Figure	Mean AUC	F stat/ p-value	t- stat/ p-value	Suitable Area*
1	30 m	30×30 m	%Contribution	9a	0.987	NA	H_0: both equal 1.326/ 0.19	9.97
2	30 m	30×30 m	Combined	9b	0.982	NA		13.25
3	200 m	200×200 m	%Contribution	10a	0.961	NA	H_0: both equal 2.092/ 0.04	21.56
4	200 m	200×200 m	Combined	10b	0.958	NA		23.4

Reduced models, variable importance

Scenario	Occ. Data	Resolution	Ranking method	Figure	Mean AUC	F stat/p-value	t- stat/ p-value	Suitable Area*
5	30 m	30×30 m	%Contribution	11a	0.962	H_0: all equal 24.995/ <.00001	H_0: 7=5, 6 & 8 5.821/ <.00001	27.05
6	30 m	30×30 m	training gain	11b	0.938			23.44
7	30 m	30×30 m	test-combined	11c	0.968			26.49
8	30 m	30×30 m	AUC contribution	11d	0.929			30.15
9	200 m	200×200 m	%Contribution	12a	0.931	H_0: all equal 4.94/.0092	H_0: 9=10 & 11 3.10/ .0026	27.59
10	200 m	200×200 m	training gain	12b	0.924			32.31
11	200 m	200×200 m	test-comb-AUC	12c	0.923			35.01

Resolution, variable importance

Scenario	Occ. Data	Resolution	Ranking method	Figure	Mean AUC	F stat/p-value	t- stat/ p-value	Suitable Area*
12	200 m	30×30 m	%Contribution	14a	0.964	NA	H_0: both equal 1.685/ .0978	20.81
13	200 m	200×200 m	%Contribution	14b	0.961	NA		21.56

MAXENT vs AIC , variable importance

Scenario	Occ. Data	Resolution	Ranking method	Figure	Mean AUC	F stat/p-value	t- stat/ p-value	Suitable Area*
14	200 m	200×200 m	%Contribution	15a	0.973	NA	H_0: both equal 7.398/ <0.00001	20.74
15	200 m	200×200 m	%Contribution	15b	0.963	NA		31.06

All other model comparisons

Description	Occ. Data	Resolution	Study Area Extent	Ranking method	Figure	Mean AUC	Suitable Area*
High Autocorr. 1	30m	30×30 m	Small	%Contribution	5a	0.998	18.59
Low Autocorr. 1	30m	30×30 m	Small	%Contribution	5b	0.827	8.27
High Autocorr. 2	200 m	200×200 m	Small	%Contribution	6a	0.967	20.65
Low Autocorr. 2	200 m	200×200 m	Small	%Contribution	6b	0.869	20.45
Full extent**	200 m	200×200 m	Large	%Contribution	7	0.990	7.71
Red. extent	200 m	200×200 m	Small	%Contribution	8	0.976	19.40

Table 9. Comparison the areas of suitable Broad-fruit Mariposa habitat produced by Domain (Nock, 2008) and MAXENT using thresholds applied for the minimal and maximum predicted area.

Model	Resolution	# of variables	Predicted (acres)	Predicted (hectares)
Domain	60 × 60 m	7	304,173	123,094
MAXENT largest	200 × 200 m	4	313,571	126,898
MAXENT smallest	200 × 200 m	4	58,111	23,517

Figure 1. Broad-fruit Mariposa (*Calochortus nitidus Dougl.*) in flower. Photo by Bob Moseley, Idaho Conservation Data Center.

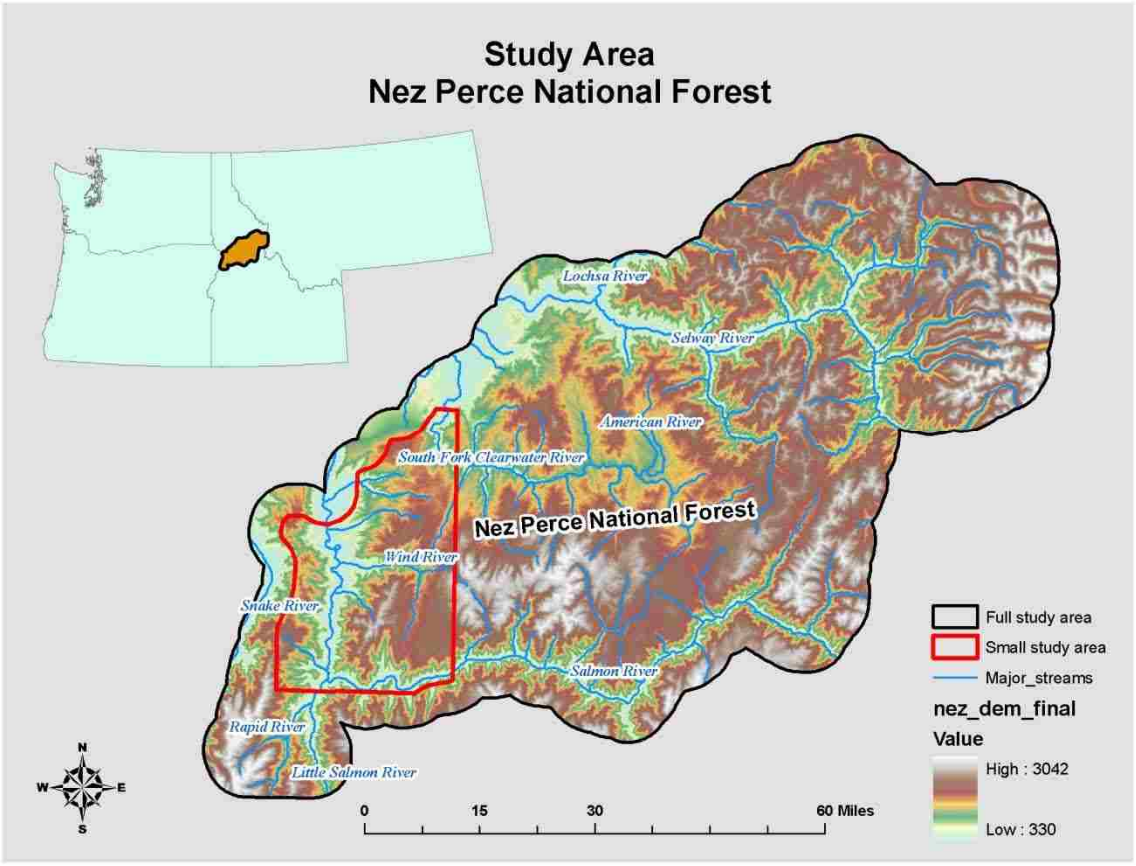


Figure 2. Map of the Nez Perce National Forest and the original full study area and the smaller study area used for most of the research.

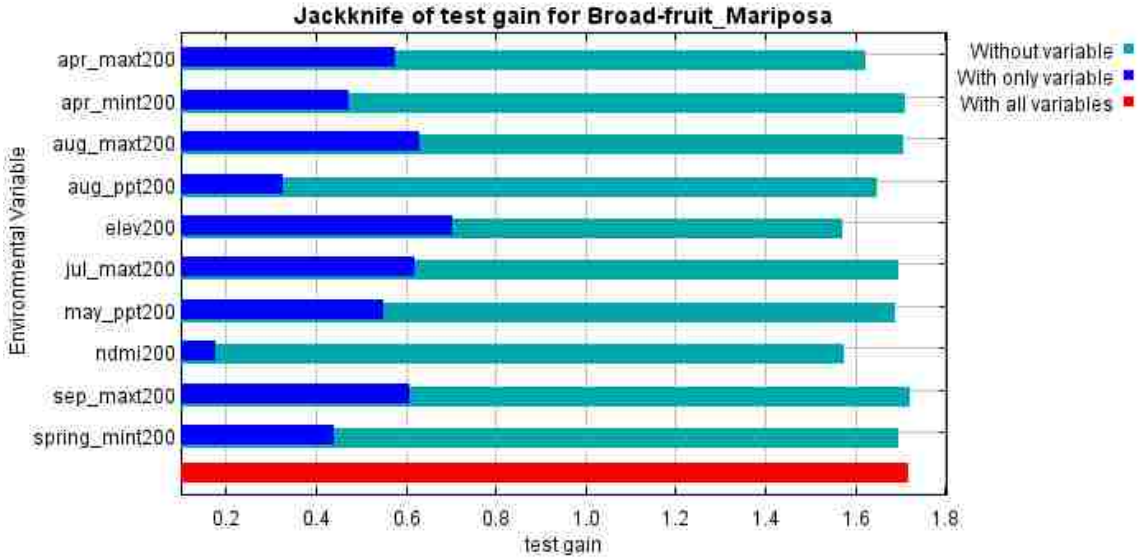


Figure 3. MAXENT's graphical output of jackknife test of variable importance, also available in tabular form. Details are given in Phillips (2008).

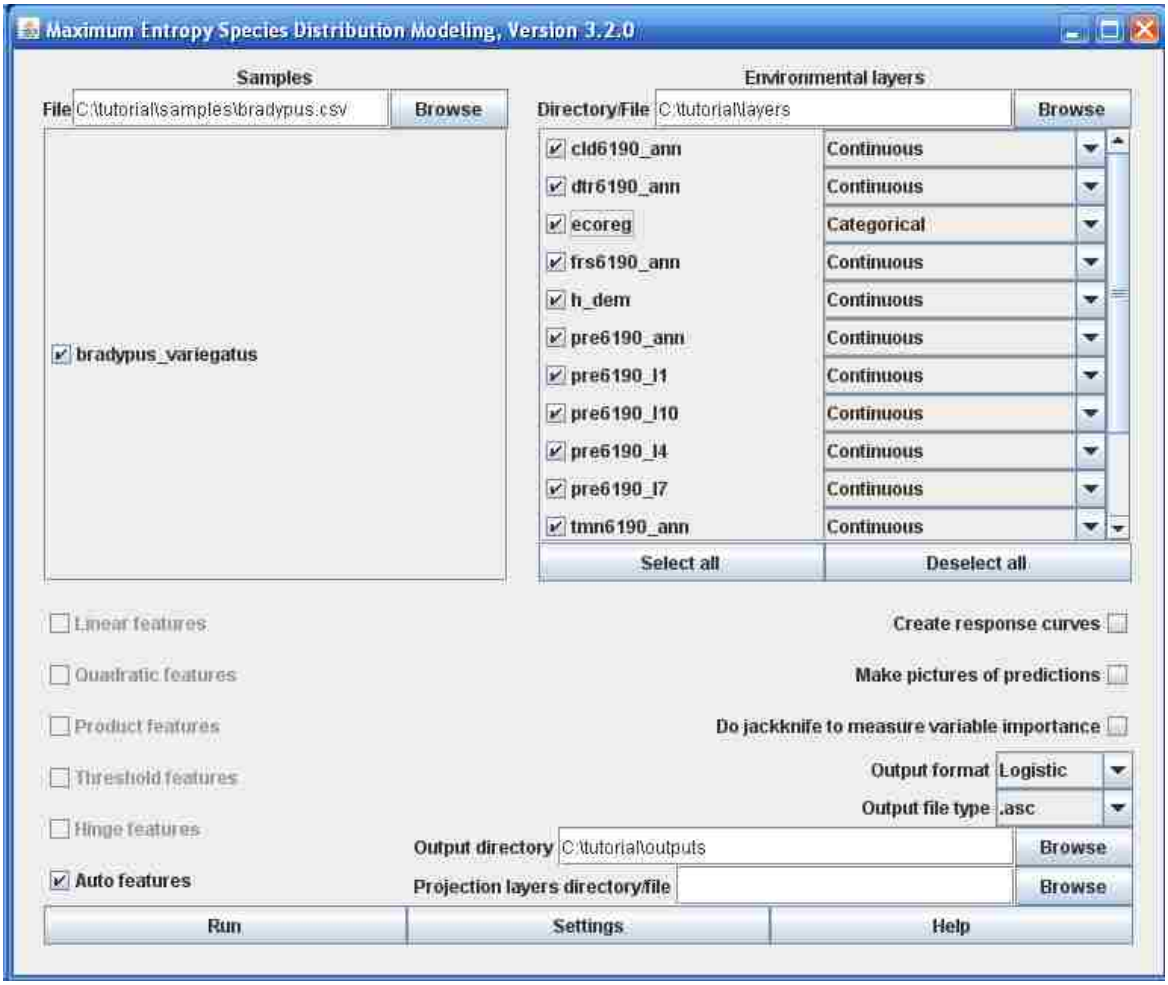


Figure 4. MAXENT's data entry and settings page. Details are given in Phillips (2008).

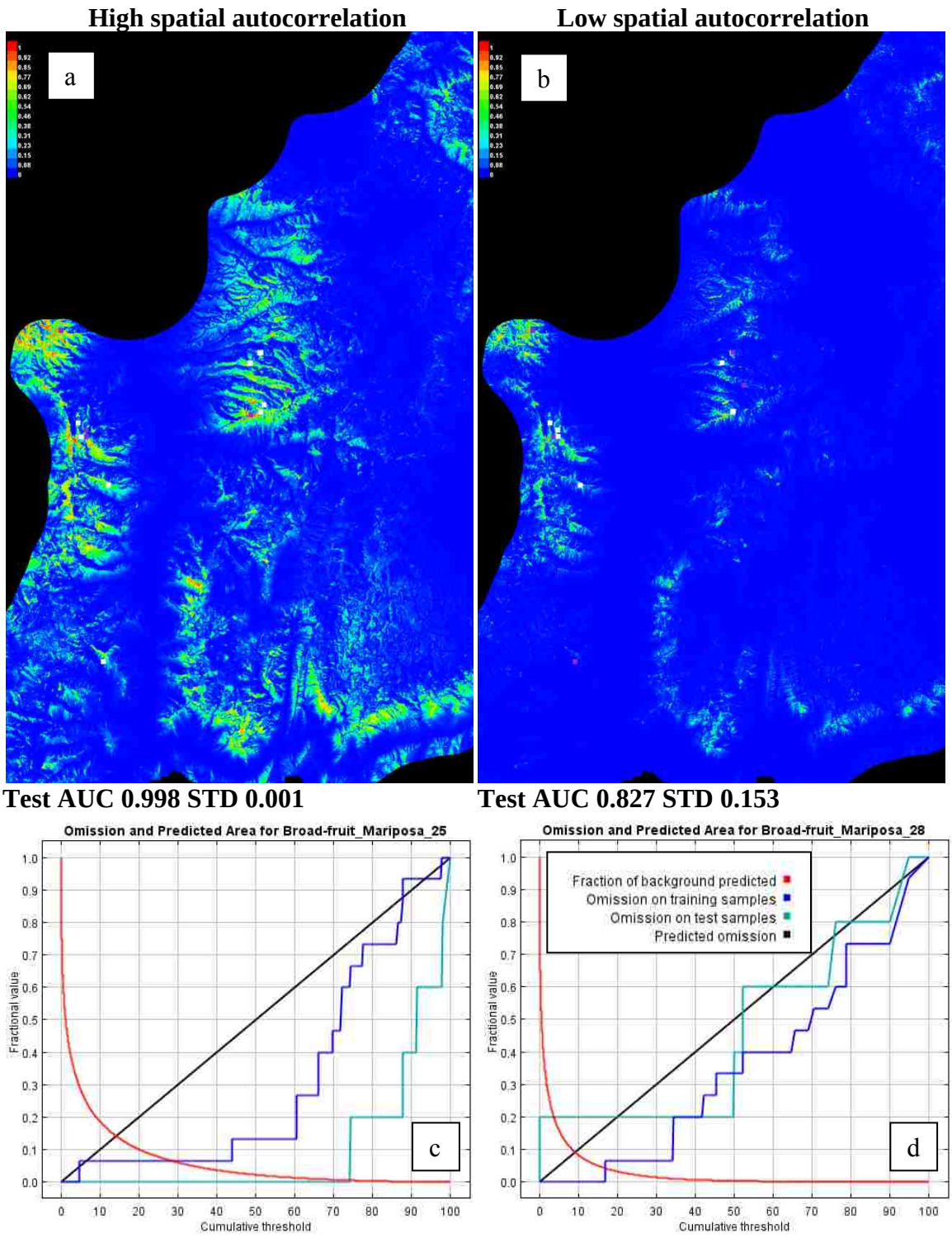


Figure 5. Comparison of 30×30 m resolution models with highest and lowest spatial autocorrelation of training and test datasets. Inset a) and b) depict predicted areas of suitable habitat, while c) and d) show the deviation from MAXENTs predicted omission rates. Additional details are given in Phillips (2008).

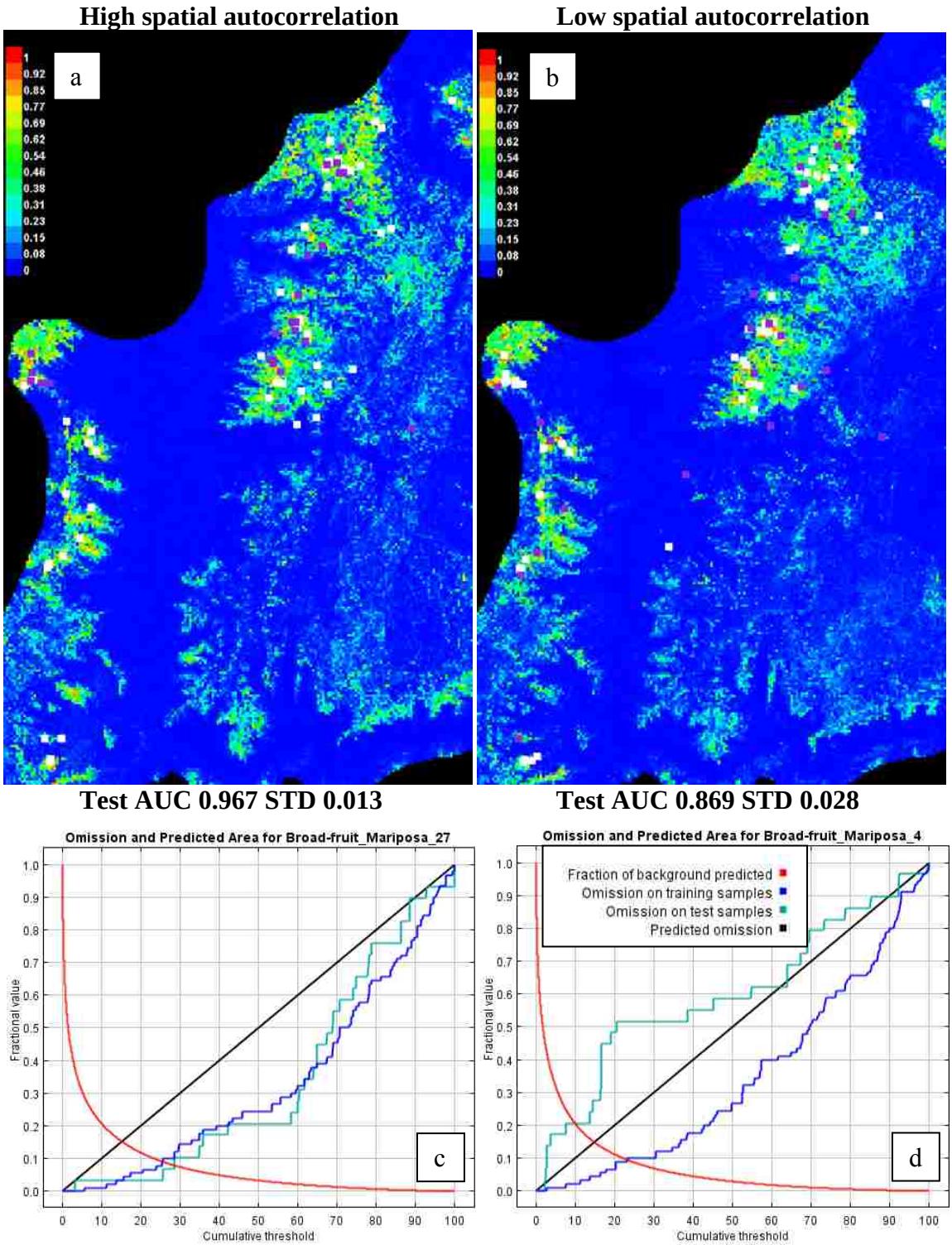
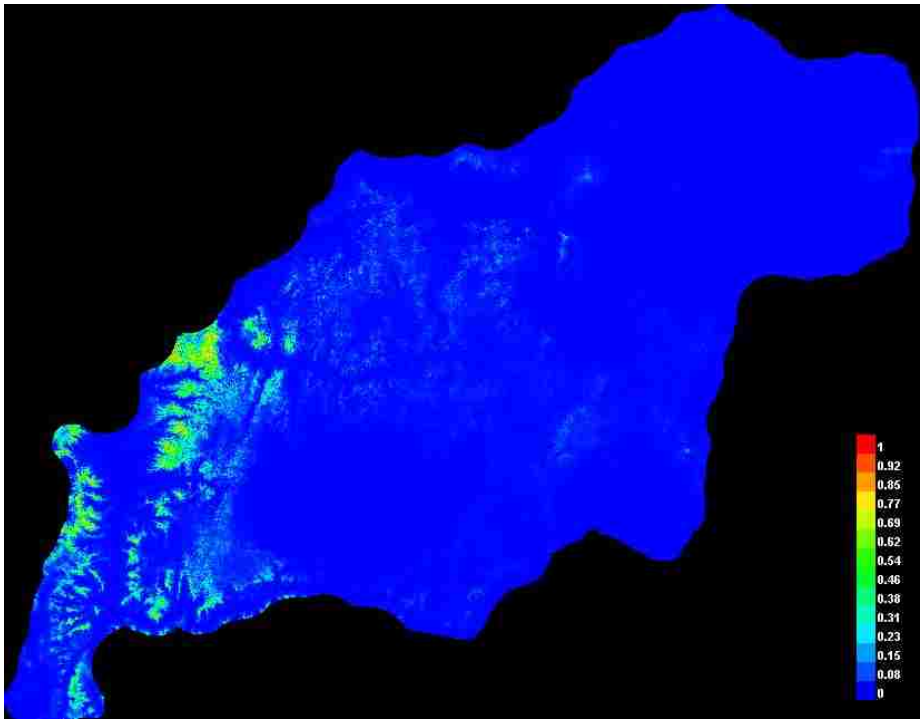
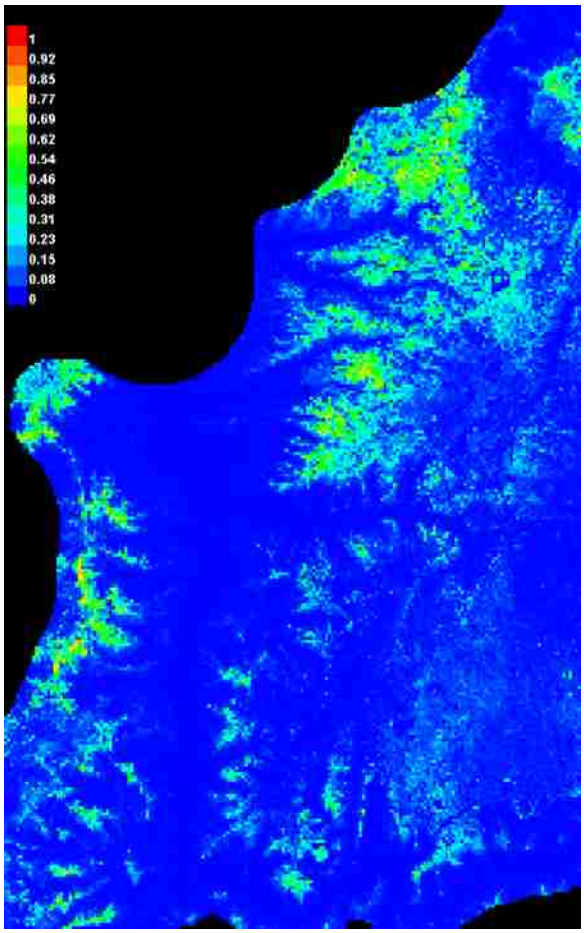


Figure 6. Comparison of 200×200 m resolution models with highest and lowest spatial autocorrelation of training and test datasets. Inset a) and b) depict predicted areas of suitable habitat, while c) and d) show the deviation from MAXENTs predicted omission rates. Additional details are given in Phillips (2008).



Variable	Percent contribution	Permutation importance
road_prox200	19.1	29.3
may_ppt200	11.8	24.3
elevation200	9.8	15
ndmi200	6.7	1
sep_maxt200	5.7	0.1
apr_mint200	5.7	0.5
jul_maxt200	5.3	1.8
jun_maxt200	3.9	1.2
sep_ppt200	3.5	7.5
aug_maxt200	2.7	0
jun_ppt200	2.6	0.3
may_mint200	2.6	0.3
veg_type200	2.5	0.3
aug_etp200	2.4	0.1
sep_mint200	2.4	0.1
ndvi200	2.4	0.4
aug_ppt200	2	3.3
apr_maxt200	1.3	10.3
jun_mint200	1.3	0.1
fire_hist200	1.2	1.7
twi200	1.1	0.1
beers_aspt200	1	0.3
slope200	0.9	0.2
solar_avg200	0.6	0.3
aug_mint200	0.5	0.6
aspect200	0.5	0.2
treatments200	0.4	0.4
jul_mint200	0.2	0.2
jul_ppt200	0.1	0
apr_ppt200	0.1	0
may_maxt200	0	0

Figure 7. Variable ranks of the full model for the full Nez Perce NF study area.

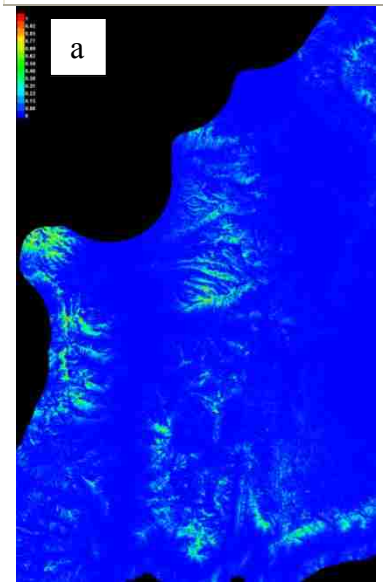


Variable	Percent contribution	Permutation importance
elev200	21.9	23.3
ndmi200	10.5	4.4
may_ppt200	9.1	19.8
may_mint200	8.3	1.9
jul_maxt200	7.4	5.7
veg_type200	4.5	1
apr_mint200	4.1	2.4
sep_ppt200	3.2	4.9
slope200	2.7	1.4
ndvi200	2.7	1
apr_maxt200	2.2	8.8
road_prox200	2.2	1.3
aspect200	2	1.2
aug_ppt200	1.8	5.6
aug_ept200	1.7	1
solar_avg200	1.7	1.2
beers_aspect200	1.6	1
jun_maxt200	1.5	0.3
jun_ppt200	1.4	0.1
fire_hist200	1.3	1.9
twi200	1.3	0.3
treatments200	1.3	0.2
jun_mint200	1	2.7
sep_mint200	1	1
apr_ppt200	0.9	0.1
jul_mint200	0.7	3.9
sep_maxt200	0.6	0.2
jul_ppt200	0.5	0.3
aug_mint200	0.5	3.1
aug_maxt200	0.4	0.1
may_maxt200	0.1	0

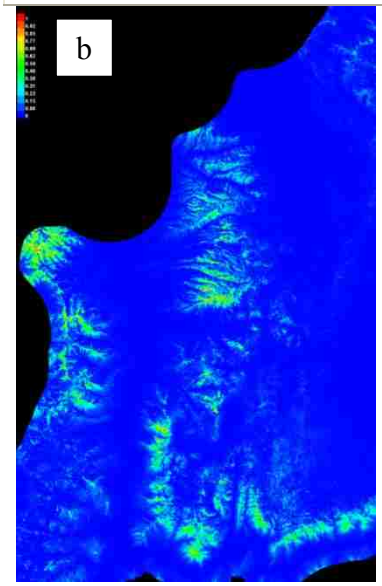
Figure 8. Variable rank of the full model for the reduced study area.

Variable	Percent contribution	Permutation importance
elev	18.1	16.7
veg_type	17.3	2.3
beers_aspect	11.4	5.2
apr_mint	10.1	30.6
ndvi	9.9	12.1
aug_ppt	9.5	8.3
ndmi	8.6	4.7
may_ppt	7.8	8.1
road_prox	6.1	2.6
jul_maxt	1.3	9.3

Variable	Percent contribution	Permutation importance
veg_type	21.4	4
elev	21.2	16
ndvi	15.8	13.1
apr_mint	11.5	24.8
may_ppt	10.1	18
ndmi	7.5	1.1
may_mint	6.4	0.1
sep_ppt	4.6	13
apr_ppt	1	7.4
jun_mint	0.4	2.4

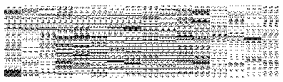


30mcontbest10 AUC 0.987



30mbest10 AUC 0.982

c



d

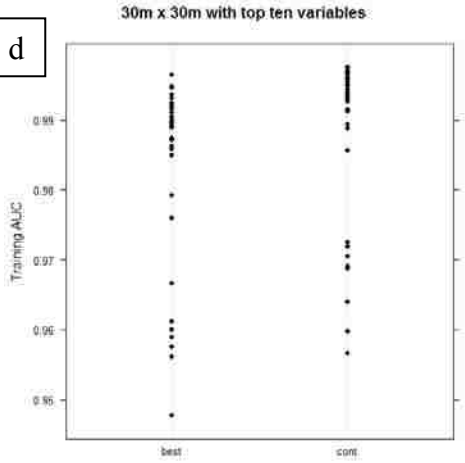
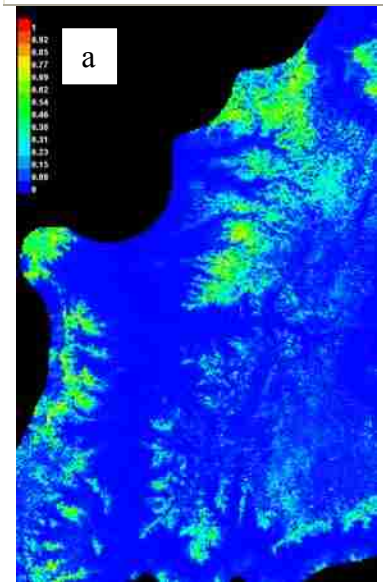


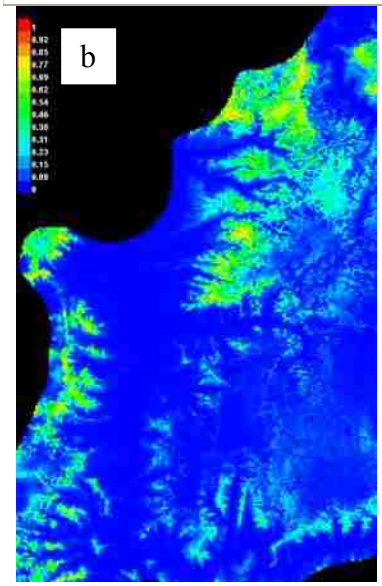
Figure 9. Variable ranks of intermediate models: a) scenario 1, b) scenario 2, while c & d) show plots of bootstrapped models.

Variable	Percent contribution	Permutation importance
elev200	28.1	18.4
may_ppt200	15.1	30.3
apr_mint200	13.9	8.2
ndmi200	12.7	7.3
jul_maxt200	11.7	9
veg_type200	5	2.2
aug_ppt200	4.1	17.5
ndvi200	3.9	2.9
beers_aspt200	3.5	2.9
road_prox200	2	1.3

Variable	Percent contribution	Permutation importance
elev200	28.5	17.7
may_ppt200	15.2	11.2
ndmi200	14.9	3.3
jul_maxt200	12	26.8
spg_mint200	9.3	3.7
apr_mint200	6.2	2.5
aug_ppt200	5.5	10.1
apr_maxt200	4.2	21.4
sep_maxt200	2.9	2.8
aug_maxt200	1.4	0.5



200mcontbest10 AUC 0.961



200mbest10 AUC 0.958

c

d

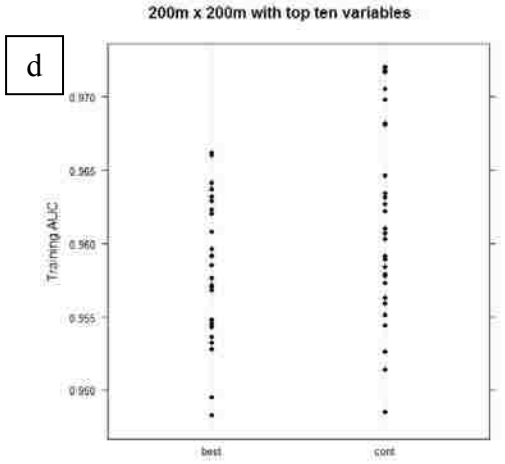
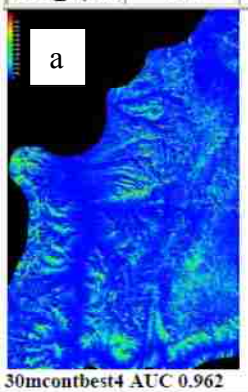


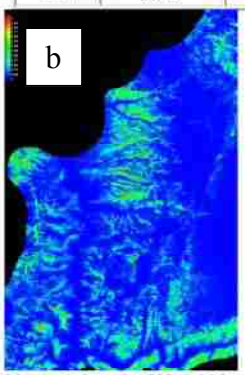
Figure 10. Variable rank of intermediate models: a) scenario 3, b) scenario 4, and c & d) show plots of bootstrapped AUC.

Variable	Percent contribution	Permutation importance
veg_type	33.1	12.9
elev	28.5	61.7
ndvi	20.8	7.2
beers_aspect	17.6	18.3



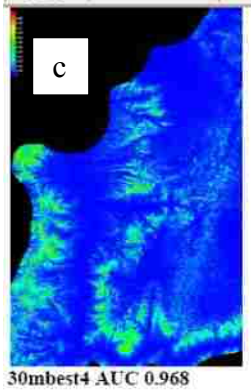
30mcontbest4 AUC 0.962

Variable	Percent contribution	Permutation importance
apr_mint	40.2	63.5
veg_type	32.7	10.4
ndmi	14.6	15.7
ndvi	12.5	10.4



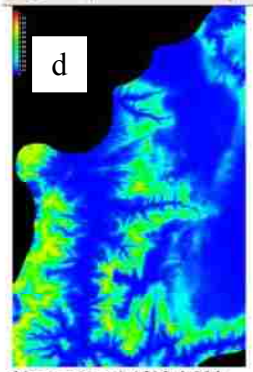
30mtrainbest4 AUC 0.938

Variable	Percent contribution	Permutation importance
veg_type	47.4	23.4
elev	28.3	42.3
sep_ppt	20.1	26
apr_ppt	4.2	8.3

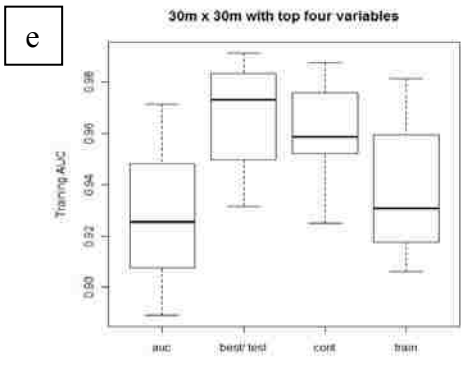


30mbest4 AUC 0.968

Variable	Percent contribution	Permutation importance
elev	39.7	33.1
apr_mint	27	52.6
sep_ppt	22.7	12.6
jul_mint	10.7	1.7



30maucbest4 AUC 0.929



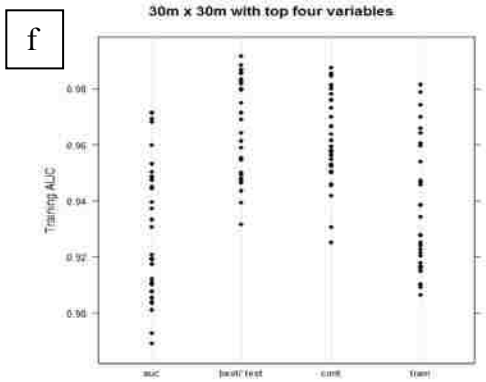
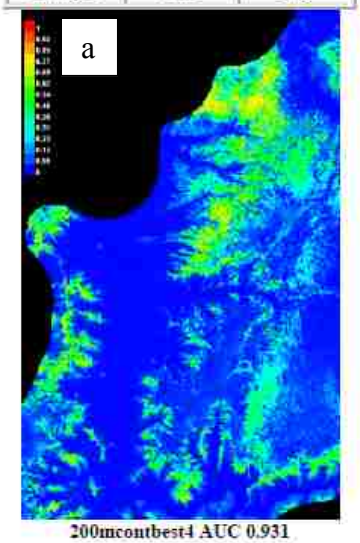


Figure 11. Variable ranks of reduced fine scale output for a) scenario 5, b) scenario 6, c) scenario 7, d) scenario 8, while e and f) show plots of bootstrapped AUC. While b) and d) predicted very different areas of suitability based upon different input variables, their AUC scores are quite close (less than 0.01 difference). This clearly illustrates the limitations of using just this one statistic for evaluation of the performance of SDMs.

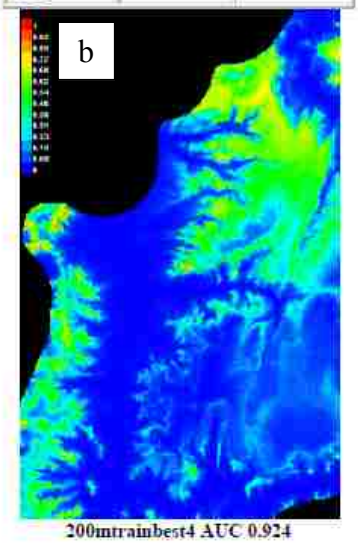
Variable	Percent contribution	Permutation importance
elev200	39.6	35.2
may_mint200	21.1	17.3
may_ppt200	20.2	20.2
ndmi200	19.1	27.3

Variable	Percent contribution	Permutation importance
elev200	45.3	15.3
jul_maxt200	25.5	54.2
may_ppt200	23.1	7.8
apr_maxt200	6.2	22.7

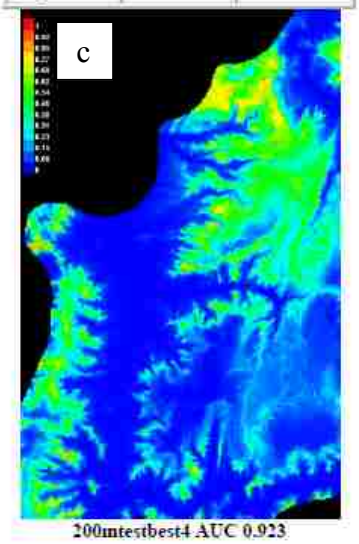
Variable	Percent contribution	Permutation importance
elev200	47.9	14.4
may_ppt200	24.4	6
sep_maxt200	23	52.6
apr_maxt200	4.7	27



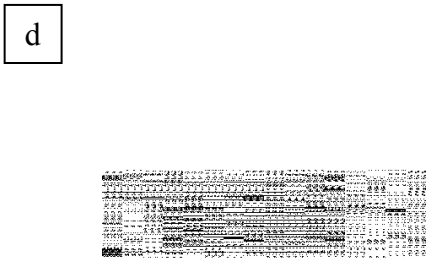
200mcontbest4 AUC 0.931



200mtrainbest4 AUC 0.924



200mtestbest4 AUC 0.923



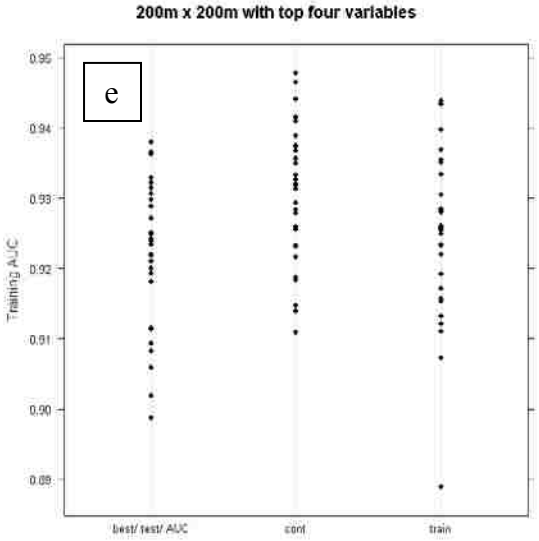
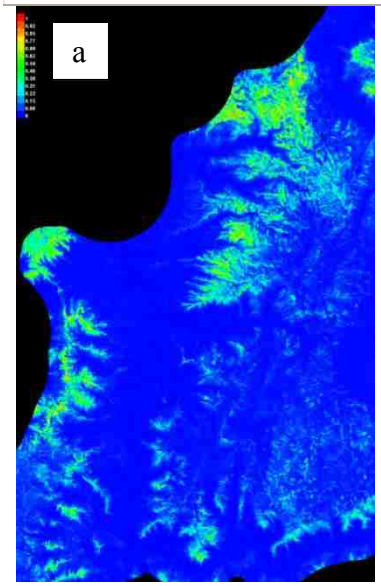


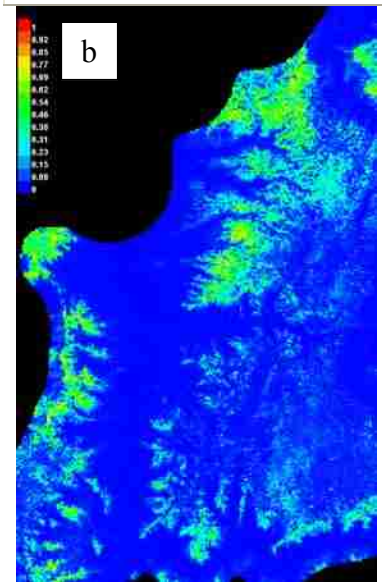
Figure 12 Variable rank of reduced coarse scale output: a) scenario 9, b) scenario 10, c) scenario 11, d and e) plots of bootstrapped models.

Variable	Percent contribution	Permutation importance
elev	29.3	16.9
ndmi	15	7.5
may_ppt	15	31.4
jul_maxt	11.3	7.9
apr_mint	10.2	8.4
ndvi	6.2	4.2
beers_aspect	4.1	3.2
road_prox	3.3	2.7
aug_ppt	2.8	16.8
veg_type	2.7	1.1

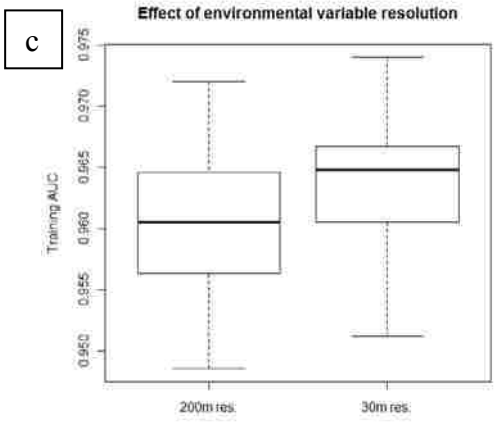
Variable	Percent contribution	Permutation importance
elev200	28.1	18.4
may_ppt200	15.1	30.3
apr_mint200	13.9	8.2
ndmi200	12.7	7.3
jul_maxt200	11.7	9
veg_type200	5	2.2
aug_ppt200	4.1	17.5
ndvi200	3.9	2.9
beers_aspect200	3.5	2.9
road_prox200	2	1.3



30m res occ. n=119 AUC 0.964



200m res occ. n=119 AUC 0.961



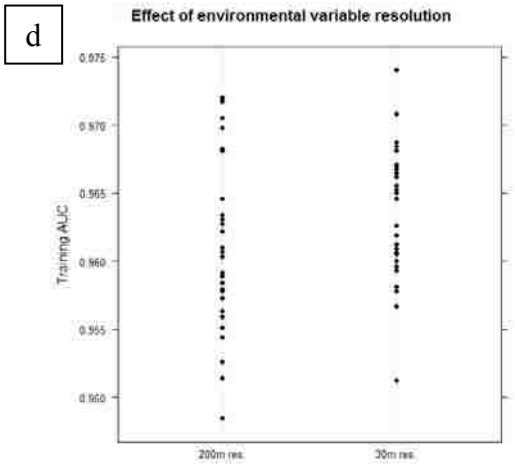


Figure 13. Variable rank of fine and coarse resolution models: a) scenario 12, b) scenario 13, and c & d) plots of bootstrapped AUC.

a

Variable	Percent contribution	Permutation importance
elev200	27.1	21.9
ndmi200	12.2	2.9
jul_maxt200	11.6	13.1
may_ppt200	9.4	19.5
veg_type200	4.7	1
sep_ppt200	4.3	3.6
sep_maxt200	3	0.6
sum_mint200	3	1.9
aug_ppt200	2.9	12.3
road_prox200	2.6	1
apr_maxt200	2.6	12.7
slope200	2.5	1.7
jun_ppt200	2.5	0.6
beers_aspt200	2.2	1.2
solar_avg200	2.2	1.7
twi200	2.2	0.7
treatments200	1.4	0.6
sep_mint200	1.3	1.8
sum_maxt200	1.3	0.7
apr_ppt200	0.5	0.2
aug_maxt200	0.3	0.2

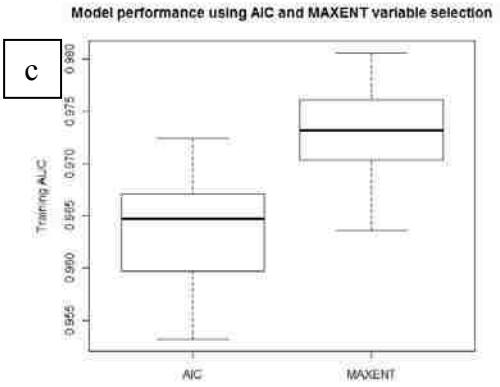
MAXENT selected variables AUC 0.973

b

Variable	Percent contribution	Permutation importance
elev200	28	15.2
ndmi200	14.2	5
may_ppt200	11.3	23
jul_maxt200	10	13.3
may_mint200	9.1	0.7
sep_ppt200	5	7.2
apr_mint200	4.5	2.8
spring_ppt200	2.9	0.3
slope200	2.5	1
apr_maxt200	2.4	10.2
jun_maxt200	2.1	1.2
sum_ppt200	1.7	8.8
jun_mint200	1.3	2.1
sum_maxt200	1.1	0.3
spring_mint200	0.8	0.3
aug_mint200	0.8	3
jul_mint200	0.8	4.7
aug_maxt200	0.7	0.3
sum_mint200	0.5	0.4
spring_maxt200	0.3	0.1
may_maxt200	0.2	0.1

AIC selected variables AUC 0.963

c



d

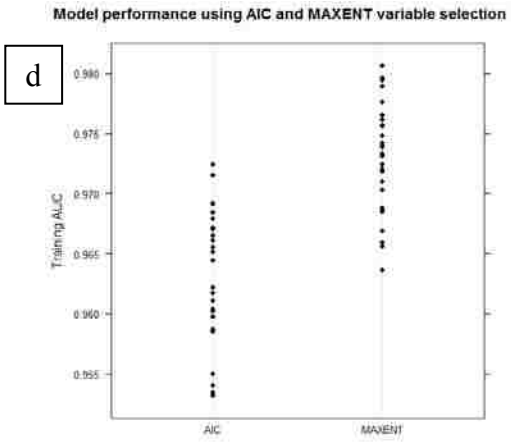


Figure 14. Variable rank of MAXENT vs. AIC model output: a) scenario 14, b) scenario 15, and c & d) plots of bootstrapped AUC.

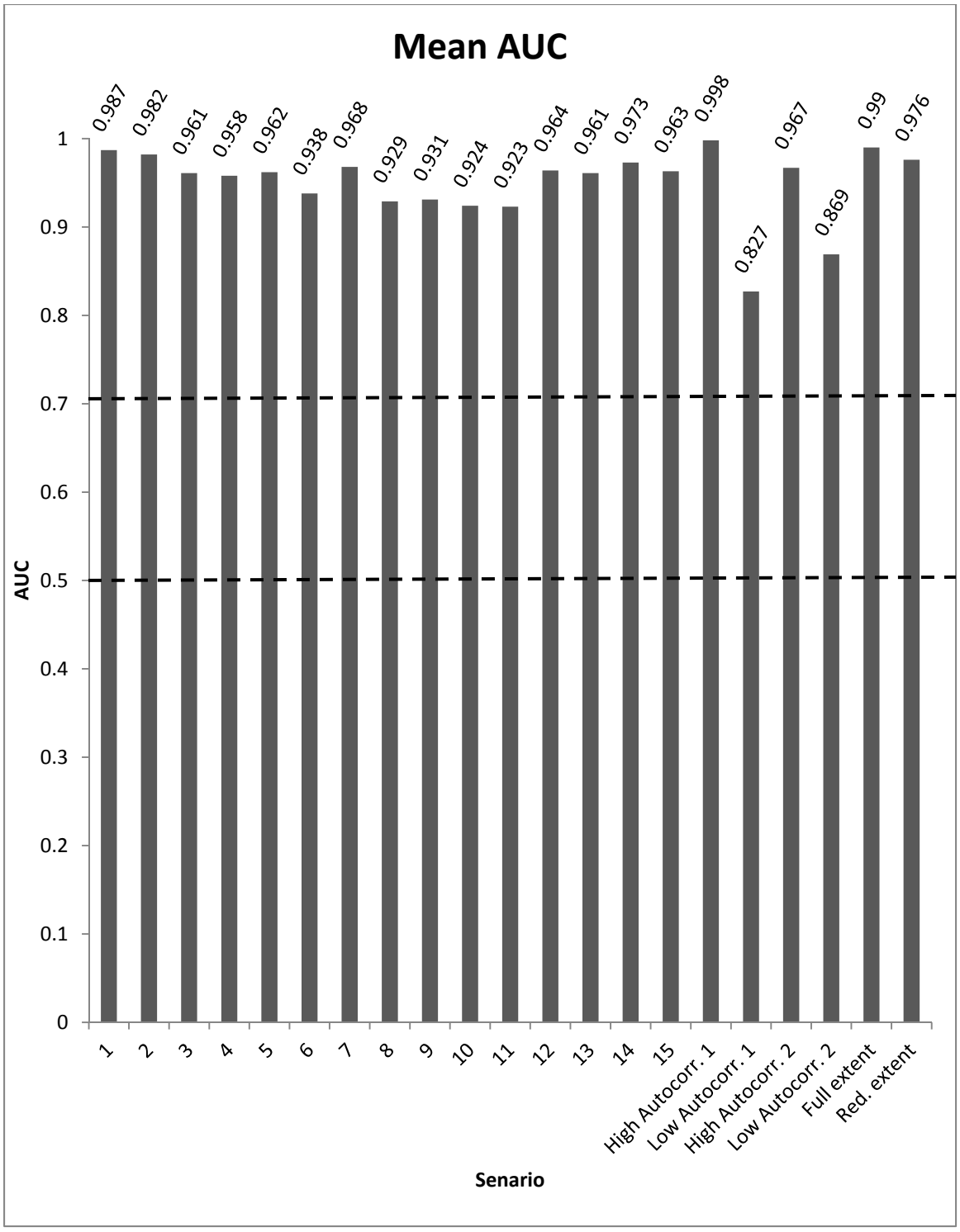


Figure 15. Graph of mean AUC values for all scenarios listed in Table 8. A value of 0.5 indicates predictions no better than random while a value of 0.7 indicates prediction accuracies that have value in conservation planning (Elith et al., 2006). The full extent model was run over the entire Nez Perce National Forest and should not be compared to the others quantitatively.