American chestnut (*Castanea dentata*) habitat modeling: identifying suitable sites for restoration in Shenandoah National Park, Virginia

Jennifer A. Santoro

Final Draft
December 6th, 2013

Masters Project submitted in partial fulfillment of the requirements for the Master of Forestry (MF) and Master of Environmental Management (MEM) degrees at the Nicholas School of the Environment, Duke University

Dr. Jennifer Swenson, Advisor

Abstract

Since 2008, The American Chestnut Foundation's (TACF) Appalachian Trail MEGA-Transect Project has engaged citizen scientists to collect American chestnut occurrence data over the length of the Appalachian Trail. This data helps TACF to locate surviving trees for use in their breeding program and expand their knowledge of chestnuts across the East Coast. However, this dataset is limiting in that it considers only the ridge-top habitat of the trail. To remedy this, we conducted an extensive sampling of side-trails in Shenandoah National Park in order to study more diverse elevation and habitat gradients. Expanding the dataset allows us to draw more informed conclusions about habitat for surviving American chestnuts. To achieve this, I developed a series of species distribution models, including GLM, CART, and Maxent models, based on field observations and spatial data of environmental variables. These predictive distribution models were then combined to generate a comprehensive map of the most likely surviving American chestnut occurrence locations across Shenandoah National Park. Additionally, projections based on future climate were made for the Maxent model to 2050 and 2070 in order to see if habitat for surviving trees might shift in the face of climate warming. Overall, the three species distribution modeling techniques tended to agree on location, but not quantity, of suitable habitat for surviving chestnuts. All models found elevation, sand, and slope to be the most significant habitat predictors in Shenandoah. Climate change models produced only subtle range shifts; as a generalist species, American chestnuts may not face adverse effects of future climate warming. Mapping these results provides valuable information to both Shenandoah National Park and TACF as they continue to search for, study, and restore American chestnuts in the Appalachian forest.

Key words: species distribution modeling, habitat modeling, American chestnut, GLM, CART, Maxent, habitat suitability modeling, GIS, Shenandoah National Park, climate modeling, restoration.

Table of Contents

Introduction

1) Background

Once a fundamental part of Appalachian ecosystem, the American chestnut was devastated by invasive chestnut blight (*Cryphonectria parasitica*), virtually eliminating mature trees from the eastern forest (Anagnostakis 1995, Paillet 2002). The species is not extinct, however; today, natural American chestnut saplings can be found sprouting from the stumps of once-large trees. Unfortunately, since the chestnut blight fungus can lie dormant for long periods of time, it persists in many areas and thus the majority of these saplings succumb to disease before they reach maturity (Paillet 2002). Very little new seed stock is established in the wild. While regenerating without reproducing, the species has effectively come to a genetic halt.

The American Chestnut Foundation (TACF), a non-profit organization whose mission is to restore the chestnut to the forests of the East Coast, has employed a backcross breeding program with the ultimate goal of breeding blight-resistant Chinese chestnut (*Castanea mollissima*) genes into the American trees and restoring forest populations (Sisco n.d.). Their work is supplemented by intensive research and planting trials. After thirty years of breeding to create a tree that contains 95% American chestnut genes, including physical characteristics of American chestnuts and blight resistance inherited from its Chinese chestnut parent, TACF has created some promising stock, but the "perfect tree" is still years away (Freinkel 2007).

In order to continue the pursuit of this "perfect tree," TACF plants every new generation of backcross-bred American chestnut stock in breeding orchards and progeny test plantations (Sisco n.d.). These orchards and plantations are located on public and private property throughout the tree's former range. In order to keep genetic variation and local adaptations in the gene pool, TACF crosses trees by putting backcrossed male pollen from the main Meadowview, Virginia orchard onto the female flowers of naturally-occurring, 100% American chestnut trees (Sisco n.d.). The female flowers are covered with wax paper bags to prevent cross-contamination by other pollen, and the mature burs are harvested in the fall. The resulting nuts are labeled with their parentage and planted in an orchard located in the same general area as the mother tree. For example, nuts from a Pennsylvania mother tree would only be planted in Pennsylvania; they would not be planted in a Virginia orchard. These methods help to keep local adaptations present in the American chestnut gene pool, which could prove necessary in the face of climate warming (Sisco n.d.).

In addition to planting new generations of trees, TACF is also actively on the hunt for surviving American chestnuts; mature trees that could be used for future "mother trees" are always in demand. Since 2008, TACF, in partnership with the Potomac Appalachian Trail Club (PATC), has spearheaded the Appalachian Trail MEGA-Transect Project, a study that trains citizen scientists to hike, find, and collect information on surviving (and naturally-occurring) American chestnut sprouts and trees along the Appalachian Trail (Marmet & Fitzsimmons 2008; Dufour & Crisfield 2008). This kind of field research is directly applicable to their extensive genetic work. When a new generation of American-Chinese hybrid trees is grown, TACF must plant them in progeny tests to see how they fare. Genetics play an obvious role in tree survival, but poor environmental conditions may prevent that good genetic stock from realizing its full potential (Rhoades et al. 2009). TACF's ultimate goal is to reestablish a viable, reproducing population of hybrid trees so natural evolution can once again take its course.

The species is of great historical importance to the eastern forest in many realms. Ecologically, its nuts served as a food source for many animals and the tree a dominant hardwood species from Maine to Florida (Paillet 2002). Culturally, rural Appalachian towns relied on its wood and nuts for building material and food (Lutts 2004). Economically, the chestnut produced lightweight, rot-resistant wood and was an important, high-value timber tree (Lutts 2004). All of these benefits were lost with the introduction of chestnut blight.

2) Species Distribution Modeling

Past studies have shown that species distribution modeling, classically used for rare fauna, is effective in predicting habitat for plant species given presence locations for the species in question (Elith et al. 2006; Hirzel et al. 2006). A study of multiple species of ferns in New Zealand compared a generalized additive model (GAM) to an ecological niche factor analysis (ENFA) in order to identify biodiversity hotspots across the country (Zaneiwski et al. 2002). With numerous data points for many species of ferns, this study emphasized the statistical rigor of GAMs and ENFA models. In a similar vein, a study of multiple tree species in thousands of study plots in Spain developed a Gaussian envelope model to predict tree species distribution with sixteen environmental variables (Montoya et al. 2009). This dataset included species presence and absence locations, which make the models more robust.

Generalized linear models (GLMs) and classification and regression tree (CART) models have also been used in plant distribution studies. One notable example of species distribution modeling for plants explored species-specific models for various types of flora in Nevada (Guisan et al. 1999). When

compared to canonical correspondence analysis (CCA) methods, GLMs were better at fitting individual plant species to a set of environmental variables, but performed best with larger datasets (Guisan et al. 1999). Another study modeled habitat for three oak species in California using GLM and CART models (Vayssieres et al. 2000). With a large dataset, the CART significantly out-performed the GLM in describing oak habitat and was better able to determine interactions between the environmental variables (Vayssieres et al. 2000).

While less statistically rigorous, species distribution modeling can be effective for uncommon plants with small presence-only datasets. A study of the rare New Caledonian tree species *Canacomyrica monticola* used Maxent to predict habitat suitability across the island from 11 species records (Kumar & Stohlgren 2009). Despite the small dataset, maximum entropy modeling was able to isolate the most important environmental variables for *C. monticola* and produce a habitat probability surface for the study area. Although it is more difficult to test a model with very few data points, maximum entropy modeling of small datasets has been successful for rare species predictions relative to other tools (Hernandez et al. 2006; Kumar & Stohlgren 2009).

Few studies have applied species distribution modeling to the once-common American chestnut. A study in Kentucky's Mammoth Cave National Park employed an ENFA model to identify which of seven environmental variables contributed most to habitat affinities for surviving chestnut trees (Fei et al. 2007). Unlike Shenandoah, Mammoth Cave National Park was once a patchwork of farms and forest. Land use history turned out to play an important role in survivability; trees almost never existed on past cultivated land, likely due to elimination of root stock and lack of reestablishment (Fei et al. 2007). Aside from historical land use, geology, slope, and elevation proved to be significant predictors or chestnut habitat in Mammoth Cave National Park (Fei et al. 2007). Such models, based off of actual tree presence data, are useful in identifying sites likely to support restoration efforts. This theme is at the core of species distribution modeling.

3) Shenandoah National Park Study

Grounded in chestnut biology, this project uses Shenandoah National Park in Virginia as a case study to identify habitat characteristics for surviving chestnut trees based on environmental conditions through spatial modeling. Either through genetic advantage, excellent habitat conditions, or a combination of both, some American chestnut trees can survive and reproduce without completely succumbing to the blight. Looking at where these trees occur in historically forested Shenandoah and

modeling to identify areas with similar habitat characteristics can provide valuable information about the location of additional surviving trees.

While models do provide valuable habitat information, it is important to remember that environmental characteristics only answer a piece of the puzzle. American chestnut survival is based on habitat conditions for the trees and their pathogens, genetic makeup, intraspecies competition, and perhaps even more factors. Due to the microscopic nature of chestnut pathogens such as *Cryphonectria parasitica* and *Phytophthora cinnamomi* (root rot fungus), it is difficult to model their distributions. Genetics play a vital role in individual tree survival, but this data is more difficult to obtain and model spatially. More complicated models might be able to show the interactions between multiple tree species, but they are not explored here. It is most straightforward to model the species of concern and combine the spatial results with other information about chestnut growth and survival.

By anchoring our knowledge of chestnut habitat preferences in the results of species distribution modeling, we can gain a more solid understanding of how TACF's backcross-bred saplings will fare in today's Appalachian forest, on what sites they might show the most improvement, and where additional survivors might be found. A habitat suitability model based on many spatially-relevant environmental variables and statistical validation can answer these suitability questions (Fei et al. 2007). Three techniques were used to model the environmental conditions of larger (>4.5 in or 11.4 cm DBH) surviving chestnut trees in Shenandoah National Park: Generalized Linear Models (GLM), Classification and Regression Tree (CART) models, and Maximum Entropy modeling (Maxent) (Phillips et al. 2006, 2008). These models were run independently and combined in an ensemble map of all three. Additionally, one model was used to project habitat under future climate scenarios for 2050 and 2070. This analysis will provide TACF with a comprehensive, statistically-based map displaying current suitable locations to find survivors and future predictions of chestnut habitat in Shenandoah.

Methods

1) Field Data Collection

Data on surviving American chestnuts were collected during the summer of 2013 along the side trails in Shenandoah National Park. Side trails included all named trails and fire roads within the park border, excluding the Appalachian Trail (although many side trails connected with the Appalachian

Trail). Two types of data were obtained: "small trees," a simple count of individual chestnut trees seen along a section of trail, and "large trees," which involved more detailed data on all chestnut trees that were > 4.5 in or 11.4 cm DBH. This cutoff was determined by TACF as a diameter class that typically reaches reproductive maturity (Marmet & Fitzsimmons 2008). Only trees that were within 4.6 m (15 ft.) of either side of the trail edges and at least 1 m (~3 ft.) tall were included in the tallies. Clusters of sprouts within a 0.3 m (1 ft.) radius of each other likely sprouted from the same parent tree and were thus counted as a single occurrence. TACF volunteers were trained in the data collection protocol, which is similar to TACF's Appalachian Trail MEGA-Transect chestnut project protocol (Marmet & Fitzsimmons 2008; Dufour & Crisfield 2008). In the protocol, each Shenandoah trail was given a code related to its name. For example, Buck Ridge trail has the code "BR" and White Oak Canyon trail has the code "WOC."

Shenandoah Chestnut Count Report Table
AT MEGA Transect Chestnut Project: Shenandoah Side Trails

Trail Name: **PAINE RUN TRAIL** Code: **PR** Date of Data Collection: 6/29/13
Start GPS Label: **PRSTART** End GPS Label: **PREND**
GPS Coordinate System: WGS 84 *(example: WGS 84, NAD 1927, NAD 1983 - check your GPS unit)*

Person Completing This Form: Taylor Cochran Phone Number: 555-5555 Other Team Members:

COUNT ALL TREES WHOSE TALLEST LIVE STEM IS 3 FT OR MORE IN HEIGHT AND WHOSE BASE IS WITHIN 15 FT OF TRAIL.
COUNT MULTIPLE SHOOTS FROM A SINGLE ROOT SYSTEM AS ONE TREE.
For each Large Tree found, please fill out a Large Tree Report Form.
For GPS End, please append chestnut count number. Please also indicate where you finish the trail.

GPS Start	GPS End (Indicate)	Starting Point (Coordinates)	Ending Point (Coordinates)	Count	Large Trees Included	Obstructed Visibility			
						% Right	% Left	Ft Right	Ft Left
PRSTART	PRA	38.7639823 -78.5989301	38.7847362 -78.5974763	12	0				
PRA	PRB	38.7847362 -78.5974763	38.7611123 -78.5951002	27	1				
PRB	PRC								
PRC	PRD								

Figure 1: Sample datasheet used by hikers to collect large chestnut tree information for the Shenandoah Side Trails Protocol.

The protocol required that volunteers bring a GPS unit, measuring or DBH tape, clicker counter, timing device, and data forms when collecting data in Shenandoah (Figure 1). First, they would select a trail to hike and print the appropriate data sheets for that trail. At the trailhead, they marked a waypoint on the GPS unit labeled with a pre-defined trail code and the word "start," indicating the start of their hike. The volunteer would then set their timer for 10 minutes and begin hiking. While hiking, the volunteer would count the number of chestnut trees and/or sprouts seen within 15 feet of either side of the trail. For ease, clicker counters were provided to volunteers. After 10 minutes, the volunteer would stop and record another waypoint on the GPS unit, labeled with the trail code, the letter "A," and the number of chestnuts they counted in their ten-minute hike. For example, a hiker on White Oak Canyon trail who saw 23 chestnuts in the first 10 minutes of hiking would record "WOCA23" in their GPS unit.

This data, in addition to the coordinates of the waypoints, was also recorded on paper. The volunteer would reset the timer for another 10 minutes, zero out the clicker counter, and continue to hike, counting chestnut trees along the way, until the timer stopped again. They labeled the next waypoint as "B." This naming convention continued until the volunteer reached the end of the trail, which was marked by a waypoint labeled "end" and the chestnut count on the final section of trail. Unlike the AT MEGA-Transect Project, the Shenandoah Side Trails protocol was based on hiking time instead of hiking distance. Ten minute intervals were selected to provide a finer scale of count data. Some sections of the AT MEGA-Transect Project were over 1 mile long, but by setting the timing interval to 10 minutes in the Shenandoah Side Trails protocol, hiking distances were rarely over half a mile.

Any large tree found was marked on the GPS unit with the trail code and "LT" for "large tree." If multiple large trees were located on a trail, sequential numbers were appended to "LT." Other attributes, such as estimate height, presence of reproductive structures, blight scars, and distance from the trail were noted on a Large Tree Report Form. When collecting data for large trees, volunteers were instructed to pause the timer, and restart it when hiking resumed.

These data were compiled into a database separating the 10-minute waypoints associated with chestnut counts and the coordinates for large trees. This project only used data from the "large tree" dataset, but other concurrent projects are using the "small tree" dataset to measure chestnut density along hiked trails. Fifty-seven large trees were located along the surveyed trails in Shenandoah. At the time of this project, about 162 miles of trail, or 42% of the side trail mileage in Shenandoah National Park were surveyed for surviving American chestnut trees (Figure 2).

2) Spatial Data

a) Variables Used in the Models

When American chestnuts were prevalent across the Appalachian Mountains before they were decimated by the chestnut blight, the trees grew nearly everywhere (Paillet 2002). Today, chestnuts are restricted to certain areas based on intraspecies competition, proximity of hardy genetic stock, and an optimal habitat envelope; these are only three of many potential reasons for the trees to be located at these points. This analysis attempts to explain surviving chestnut habitat in terms of the third restriction: the physical habitat. For that, we must consider a wide variety of environmental variables and determine which ones have the most influence on optimal habitat of survivors.

Ten environmental variables were included in all three models: these included seven digital elevation model (DEM)-derived variables and three soil variables (Table 1). The DEM-derived variables are measures of elevation, slope, insolation, distance to nearest river, topographic convergence index (TCI), topographic relative moisture index (TRMI), and terrain shape index (TSI). Soil variables are percentages of clay and sand in the soil in addition to pH of soils. A temperature variable was added to the Maxent climate projection model.

The main goal in including these 11 environmental variables was diversity in habitat gradients. This can be broken down into measures of temperature and moisture, which are important considerations for plant growth. Variables that can be considered proxies for temperature include elevation (higher elevations tend to be cooler), insolation, TSI, and maximum monthly temperatures. Moisture variables include distance to rivers, slope (steepness can determine water retention), TCI, and soil characteristics. TRMI attempts to approximate the interaction between temperature and moisture, as it considers solar effects through the aspect component (Parker 1982).

Data on average temperature in the warmest month (July) was included only in the Maxent model as a basis for habitat predictions given future climate scenarios. It was not included in the GLM or CART models because they are less often used for extrapolating climate projections and the temperature data is spatially coarse. Elevation data typically captures a proxy for temperature via lapse rates, so only the fine-resolution elevation layer was included in the GLM and CART. Maxent models do not suffer from the inclusion of correlated variables such as elevation and temperature, so both were included to project chestnut habitat onto a future climate scenario (Phillips et al. 2008). Data describing average temperature in the warmest month was chosen over mean annual temperature because the warming during the hottest part of the year was assumed to have a greater effect on chestnut survival. Overall, annual temperature across Shenandoah National Park averaged 45-55^{o}F and ranged from 74-87^{o}F in the warmest month (Hijmans et al. 2005). Lower elevations typically had warmer temperatures than mountain ridges. Annual precipitation ranged from 97-131 cm across the park, with greater precipitation occurring at higher elevations. This dataset was downscaled from climate prediction models to a cell size of 1 km.

Distance to rivers was calculated as a simple Euclidian distance from every raster cell to the nearest water feature. Flowpaths included both ephemeral mountain streams and permanent channels.

Insolation and slope are both simple surface analyses derived from digital elevation model data. The former has a fixed illumination angle that describes annual solar irradiation on the land surface. An azimuth of 225^{o} and altitude of 30^{o} describes hotter southwest-facing slopes and cooler northeast-facing slopes. It is an improvement over aspect because aspect simply describes which direction a slope faces; insolation takes radiation energy into account. Slope is defined as the maximum rate of change in elevation for each raster cell across a land surface. In this analysis, it is calculated as the rise over the run in a three by three neighborhood of raster cells.

TCI and TRMI both model potential moisture conditions based on topographic features. TCI, sometimes referred to as Topographic Wetness Index (TWI), only considers moisture as it relates to physical terrain variables, such as slope steepness or shallowness (Sorensen et al. 2006; Kopecky & Cizkova 2010). It acts as a proxy for soil moisture and groundwater flow. Steeper slopes tend to shed surface water faster than shallower slopes, and thus could have a lower, or drier, TCI value. On the other hand, TRMI is a field-based technique that approximates relative moisture availability stored in the soil based on slope and aspect (Parker 1982). TRMI is the sum of four scalar environmental variables: relative slope position, slope configuration, terrain steepness, and aspect. Including aspect, or solar angle, in this calculation allows the TRMI to provide an estimate of potential evaporative water loss and soil water retention across the landscape. Although the weighting of the four contributing variables is subjective, this analysis uses the variable weights described in Parker (1982). It is important to keep in mind that these moisture indices are not measurements of the landscape, but approximations of environmental characteristics based on factors that could contribute to soil moisture.

TSI approximates the geometric shape of the land surface, from exposed to sheltered, and was originally used to describe the relationship between topography and tree growth (McNab 1989). It is calculated as the mean elevation of a defined circular area divided by the radius of that area and can have a range from negative infinity to positive infinity (McNab 1989). For raster grid cells, this calculation was modified to fit a square. Each cell's elevation was compared to that of its eight neighbors in a three by three neighborhood; if the cell had a value greater than one, it was considered exposed (higher than the average of its neighbors). If the cell value was less than one, it was considered sheltered (lower than the average of its neighbors).

Soil variables were obtained from the NRCS Web Soil Survey and were computed based on averages from the individual soil types. Each soil type is described as an average of at least three complete soil profiles and at least ten smaller samples for each of three transects; soils covering a larger

area involve additional complete profiles and transects (SSURGO 2013). This means that one value was used over the entire surface of a soil type.

b) Eliminated Variables

Other variables were considered but ultimately eliminated due to high (>0.5; Pearson's) correlations with other variables. These included cation exchange capacity (CEC), percentage of silt in the soil, relative slope position, topographic position index, and aspect. The highest correlation between variables remaining in the model was 0.47 between elevation and distance to rivers.

Of the five soil variables considered in this analysis, some were correlated highly with other soil variables. CEC was correlated highly with silt and clay (0.75 and 0.83, respectively), while silt was also correlated with clay (0.82). Silt was eliminated first because it had correlations over 0.8 with both clay and CEC. Next, CEC was eliminated because it depends on pH; two correlated variables describing similar estimates are unnecessary, and pH is assumed to be more reliable. Relative slope position and aspect are both components of the TRMI calculation, and were eliminated due to redundancy with that index. Finally, TPI and TSI had a correlation of 0.52. Both indices could have remained in the model, but due to the similarity of environmental variation each explained, TPI was eliminated.

c) Data Resolution

Data compatible with ESRI's ArcGIS 10.2 were downloaded from the USDA and USGS (Dollison 2010; SSURGO 2013). A National Elevation Dataset at one-third arc second resolution (~10 m resolution) for Shenandoah National Park was used to derive most of the environmental variables used in this study (Dollison 2010). These included measures of slope, insolation, distance to nearest river, topographic convergence index (TCI), topographic relative moisture index (TRMI), and terrain shape index (TSI). Percentages of clay and sand in the soil, in addition to pH of soils, were downloaded from the USDA Web Soil Survey. This data was digitized from 7.5 minute topographic quadrangles (SSURGO 2013). All analyses were performed in the North American Datum of 1983, Universal Transverse Mercator Zone 17 North (NAD 1983 UTM Zone 17N), which encompasses Virginia and much of the East Coast of the United States.

Table 1: Descriptions of 11 environmental variables used in the three species distribution models, their calculations, and sources.

Environmental Variable	Spatial Resolution	Description	Calculation	Primary Source	Download Source
Distance to Rivers	10m	Distance (in meters) to the nearest stream (ephemeral or permanent).	Euclidean distance from flowpaths	USGS National Mapping Services	nationalmap.gov
Elevation	10m	Height above sea level, in meters.	*Pre-prepared from Digital Elevation Model*		
Insolation	10m	Measure of solar radiation energy on Earth's surface, from hot (southwest-facing) to cool (northeast-facing).	Hillshade with azimuth of 225° and altitude of 30°		
Slope	10m	Measure of steepness along terrain.	$Slope = \sqrt{\left(\frac{dz}{dx}\right)^2 + \left(\frac{dz}{dy}\right)^2}$		
Topographic Convergence Index (TCI)	10m	Approximation of the moisture content, or wetness, of an area, taking into account topography as it controls hydrology (Sorensen et al. 2006).	$TCI = \ln \frac{upslope\ area}{\tan(slope)}$		
Topographic Relative Moisture Index (TRMI)	10m	Approximation of the relative soil moisture availability of a site using slope and aspect, from mesic (moist) to xeric (dry) (Parker 1982)	Sum of relative slope position, slope configuration, slope steepness, and slope aspect (scale of 0-60)		
Terrain Shape Index (TSI)	10m	Approximation of the geometric shape of the land surface (McNab 1989).	$TSI = \frac{mean\ elevation}{radius}$		
Percent Clay	30m	Percentage of clay content in particular soil types.	*Pre-prepared from Web Soil Survey*	NRCS SSURGO Database	websoilsurvey.nrcs.usda.gov
Percent Sand	30m	Percentage of sand content in particular soil types.	*Pre-prepared from Web Soil Survey*	NRCS SSURGO Database	websoilsurvey.nrcs.usda.gov
pH	30m	Measure of acidity/basicity of a soil based on hydrogen ion content.	*Pre-prepared from Web Soil Survey*	NRCS SSURGO Database	websoilsurvey.nrcs.usda.gov
Temperature*	1km	Maximum temperature of the warmest month (for 2013, this is July) in degrees Celsius.	*Pre-prepared; methods from WorldClim (Hijmans et al. 2005).*	Hijmans et al. 2005	worldclim.org

* *Used only in the Maxent model.*

3) Modeling

a) Modeling Description

Three models were explored: a generalized linear model (GLM), classification and regression tree (CART) model, and maximum entropy (Maxent) model (Sing et al. 2005; Phillips et al. 2006, 2008; Goslee & Urban 2007; R Development Core Team 2008). Models were mapped individually and then combined to show the intersection of habitat predicted by all three models.

GLM and CART models require true absence or pseudo-absence points with which to compare presence data. One hundred points were randomly generated in ArcGIS within the boundary of Shenandoah National Park to serve as "not habitat" values for the model. Values for the environmental variables were assigned for all presence and pseudo-absence points. Maxent generates its own pseudo-absence points.

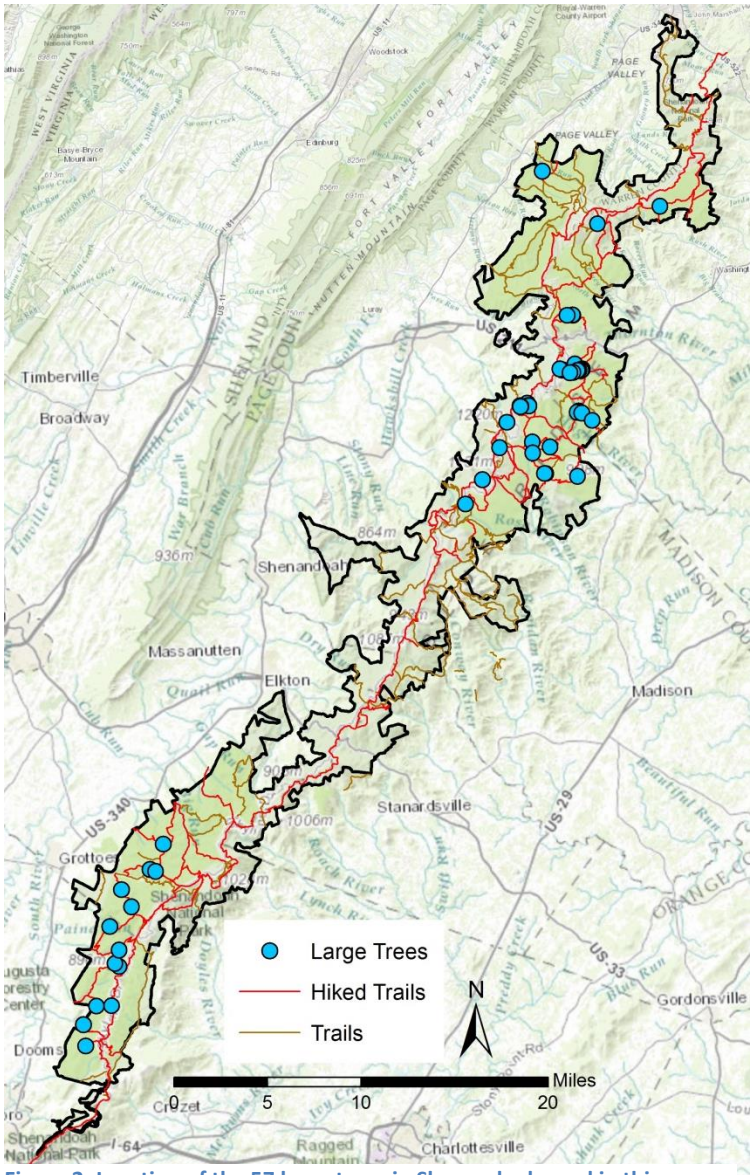


Figure 2: Location of the 57 large trees in Shenandoah used in this analysis and trails hiked to find them.

Maxent also has the ability to consider future climate scenarios and project the species occurrence data onto extrapolated climate variables (Phillips et al. 2006, 2008). After developing a model based on the input environmental variables, Maxent runs the model on the new set of projected variables. Often, all of these variables are the same except for a few key climate warming factors such as increased temperature or precipitation.

In this model, I used maximum monthly temperature for the warmest month of the year (July) and extrapolated maximum monthly temperatures for July

2050 and 2070 using the Community Climate System Model (CCSM4) developed by the National Center for Atmospheric Research in the US (NCAR 2012). This scenario is part of the coupled model intercomparison project, phase five (CMIP5). Data are available in four representative concentration pathways (RCPs) developed by the IPCC, which describe the radiative forcing of greenhouse gas concentration trajectories to the year 2100 (Hijmans et al. 2005). The four RCPs are 2.6 W/m^2, 4.5 W/m^2, 6.0 W/m^2, and 8.5 W/m^2. Here, Maxent was used to model future chestnut habitat predictions using maximum temperature of the warmest month for RCP 8.5 in both 2050 and 2070.

b) Modeling Background

GLMs use maximum likelihood methods and a defined link function to estimate the probability that a given sample belongs to a certain group (Guisan et al. 2002). Here, I use a logistic link function to estimate the probability of chestnut habitat. Since the response variable for logistic regressions is binary, the result is no longer linear, and instead takes the form of an "S" curve between zero and one (Guisan et al. 2002). Tuning the GLM with receiver operating characteristics (ROC) curves allows the user to find the optimal threshold, or break point, between zero (not habitat) and one (habitat); these models have successfully been applied to studies of plants and are known for their ability to fit species-specific functions (Guisan et al. 1999, 2002).

CART models, here used to predict categorical responses, are often used in species distribution modeling. They employ recursive partitioning to develop a series of "and/or" contingencies on the input variables, which is able to cover intricacies that linear models and logistic regressions may miss (De'Ath & Fabricius 2000; Vayssieres et al. 2000). In the case of this analysis, these splits aim to divide the each relevant variable into "habitat" and "not habitat" groups. CART models tend to fit a given dataset as accurately as possible, and often must be pruned down to create a more conservative model able to accept additional datasets (De'Ath & Fabricius 2000; Vayssieres et al. 2000; Loiselle et al. 2003). CART models suffer when used with small datasets because there are fewer points on which to base variable splits (De'Ath & Fabricius 2000). Overall, because GLM and CART models approach species distribution modeling from different angles, it is valuable to explore both in this analysis.

Maxent is a machine-learning program that uses species occurrence data and maximum entropy analysis to predict the probability distribution for the species of concern (Phillips et al. 2006). Predictions are based on user-inputted environmental variables that act as constraints in the study area, and the program generates its own pseudo-absence points (Yackulic et al. 2013). A convenient aspect of Maxent

is its repeatability. Models can be post-validated with another set of species occurrence data, or a portion of data can be withheld in the original model and used as "training" data (Phillips et al. 2006). Maxent also has the ability to bootstrap or cross-validate the data during the initial model run; these options partition the data into random groups, test the groups in turn, and then average the results and performance over the entire model (Elith et al. 2011).

Maxent offers other explanations of variable contribution, including a "jackknife" test and a table of variable importance. In the jackknife test, the overall model is assessed by creating two additional models. The first is a model that includes only one environmental predictor variable; this is repeated for each variable included. The second model is constructed with every variable except one. These jackknife models can help to determine which single variables contribute most to the habitat model. Important variables generate a good model by themselves and a poor model when they are excluded (Pearson et al. 2007). Past studies have shown that compared to other methods of species distribution modeling, Maxent performs well with small datasets (Hernandez et al.. 2006; Pearson et al. 2007; Komar & Stohlgren 2009; Thorn et al. 2009) and has been used to predict habitat for rare or endangered plant species (Engler et al. 2004; Komar & Stohlgren 2009).

GLM and CART models were chosen in this analysis for their ease of construction, implementation, user tuning, and statistical interpretation. Each offers a unique statistical approach to habitat classification, so comparing the two model outputs is helpful in developing a potentially more robust prediction of chestnut habitat in Shenandoah. Maxent has less room for post-processing user tuning, but it balances fitting the supplied data with maintaining statistical reproducibility. Another benefit of Maxent is its ability to generate reliable models with small numbers of species occurrence points (Hernandez et al. 2006; Komar & Stohlgren 2009). Although the Maxent modeling processes is significantly harder to interpret for many users, it provides a robust prediction of habitat that complements GLM and CART models (Phillips et al. 2006). However, modeling has the potential to over-fit the given species occurrences, so it is important to exercise caution when construction, tuning, and interpreting these models for conservation decisions (Loiselle et al. 2003).

Results

During the summer of 2013, over 40% of the side trail mileage in Shenandoah National Park was hiked by TACF volunteers, and 57 large chestnut trees were located (Figure 2). The three modeling approaches, as described above, were run to estimate chestnut habitat based on the 57 large surviving chestnut trees located in the field. All three of these models predicted that elevation, percent sand, and slope have the biggest impact on chestnut presence in the park.

1) Generalized Linear Model

Of the ten variables included in the GLM, none had correlations >0.5. Six proved to be significant habitat predictors in both the GLM and subsequent ANOVA (Table 2). In decreasing order of importance, sand, elevation, and slope had the most significant contributions, while TCI, TRMI, and clay were also significant. Distance to rivers, insolation, pH, and TSI were not significant.

Stepwise logistic regression did not improve the model, so the original GLM including all variables was used. Using a receiver operating characteristics (ROC) calculation, I determined a cutoff value of 0.44 between habitat and non-habitat (Figure 3). This ROC tuning serves to maximize the number of true habitat classifications (true positives) while minimizing incorrect classifications (false positives). The GLM classified 74% of the original large tree points correctly and 26% incorrectly (Table 3). In total, the GLM estimated that 12.4% (9,858 ha) of Shenandoah is suitable American chestnut habitat (Table 4). Overall, the GLM described surviving chestnut habitat as sandy, high-elevation, low- to mid-slopes that were generally dry.

Table 2: Significance of variable predictors in the GLM.

Variable	Coefficient	P-value
(Intercept)	-12.17487	0.000935 ***
Distance to Rivers	-0.000176	0.852537
Elevation	0.005088	0.003261 **
Insolation	0.004399	0.402270
Slope	-0.081305	0.022821 *
TCI	-0.135383	0.025500 *
TRMI	-0.054111	0.033197 *
TSI	0.133374	0.240310
Percent Clay	0.139958	0.091788 ₀
Percent Sand	0.100584	0.000218 ***
pH	0.386365	0.100269

Significance codes: 0.0001: *** 0.001: ** 0.01: * 0.05: ₀

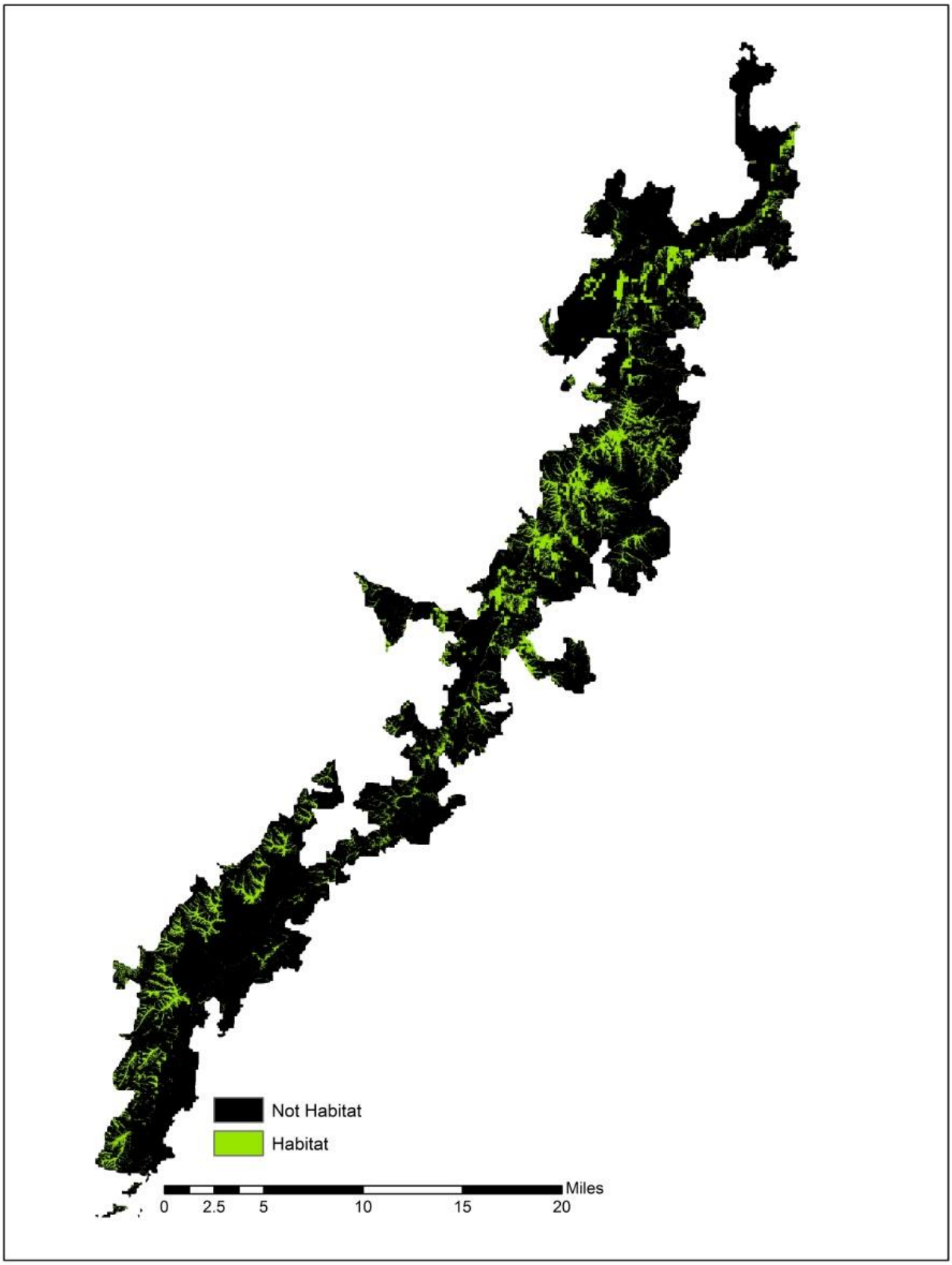


Figure 3: GLM output of chestnut habitat (0.44 threshold).

2) Classification and Regression Tree Model

Of the ten variables included in the CART model, five were significant habitat predictors. Elevation was the most important habitat predictor, followed by slope, sand, TCI, and TSI (Figure 4). Distance to rivers, clay, insolation, pH, and TRMI were not significant.

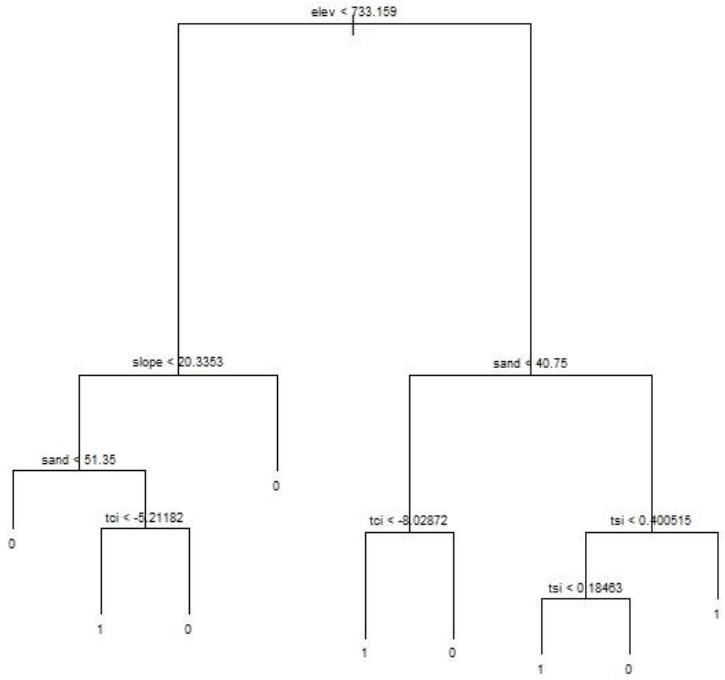


Figure 4: Tree generated for chestnut habitat from the CART model. This tree was pruned to 9 branches.

Table 3: Confusion matrix describing the number of included and excluded present points in each model.

GLM Threshold: 0.44	**Included**	42
	Excluded	15
	% Success	74%
CART Threshold: see tree	**Included**	28
	Excluded	29
	% Success	49%
Maxent Threshold: 0.17	**Included**	52
	Excluded	5
	% Success	91%

Table 4: Area classified as habitat for all four models. The area of Shenandoah in hectares is included for reference.

	Area Classified as Habitat	
	Percent	**Hectares**
GLM	12.4%	9,858
CART	6.76%	5,375
Maxent	19.15%	15,223
Ensemble Model	25.37%	20,168
Shenandoah National Park	100%	79,507

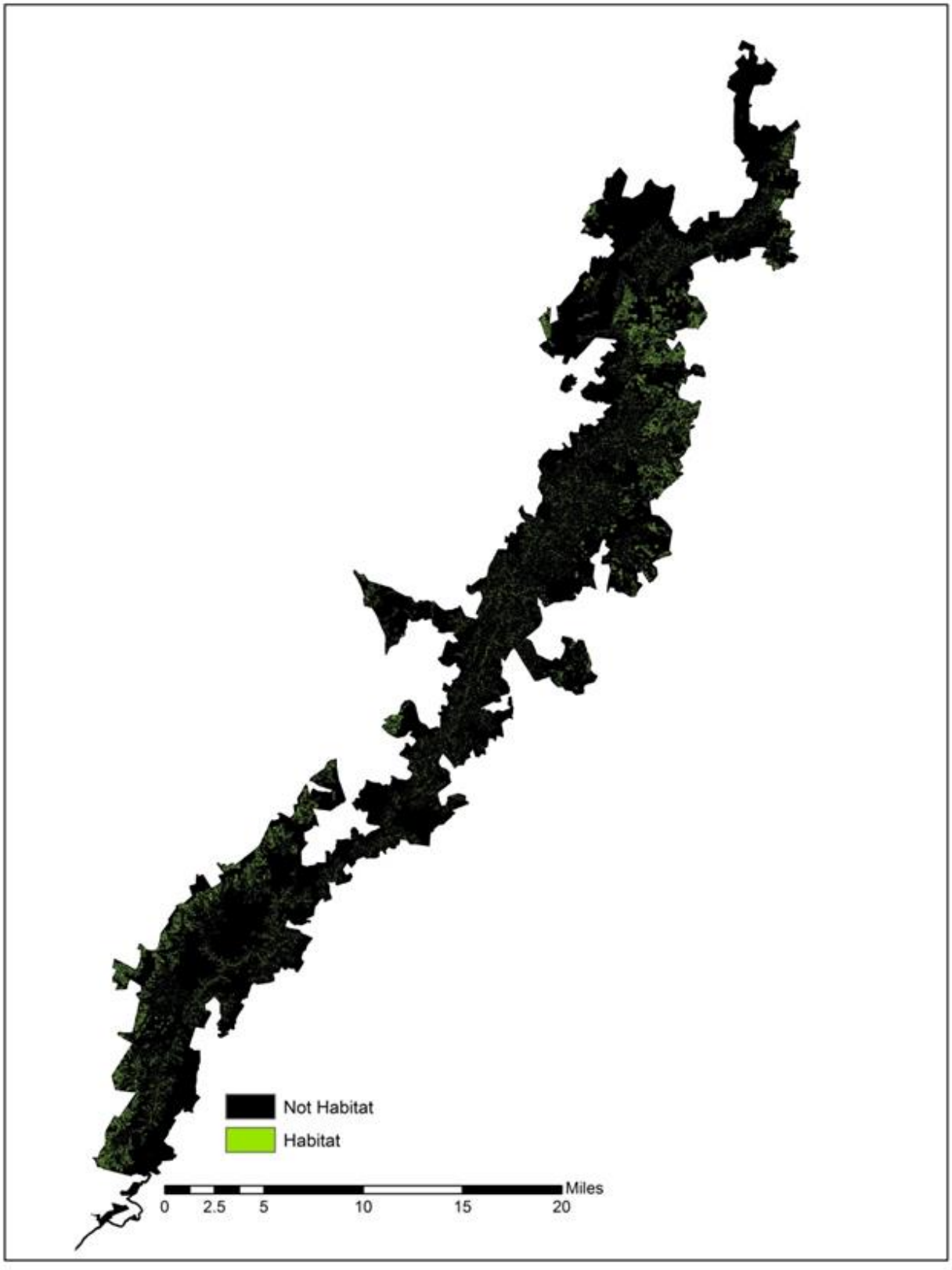


Figure 5: CART output of chestnut habitat.

CART models create a classification tree that determines habitat cutoff values for each variable included in the model. Terminal nodes determine if the series of branches is habitat or non-habitat (Figure 4). To find out which combination of environmental variables determine chestnut habitat, follow the tree from its start at elevation down branches until a terminal node is 1, indicating presence.

Cross-validation suggested that a fitted classification tree would be pruned from sixteen to nine terminal nodes (Figure 4). After pruning the classification tree, the CART model classified 49% of chestnut input points correctly and 51% incorrectly (Table 3; Figure 5). The CART model predicted that 6.76% (5,375 ha) of Shenandoah National Park was suitable American chestnut habitat (Table 4). The CART model described chestnut habitat as high elevation, low- to mid-slopes, mid-sand ranges for soil, dry, and slightly convex terrain.

3) Maximum Entropy Model

The Maxent model based on 11 variables (including temperature) revealed that the variable with the most contribution to the model was sand, which explained 28.3% of the model variation (Table 5). Elevation had the second most important contribution (26.1%), but according to the jackknife analysis was the most important variable when used by itself (Figure 6). In other words, elevation has the most unique explanatory power of all of the environmental variables if used in isolation; the model improves greatly when elevation is included, and suffers when it is removed. Other important variables are slope, pH, and TRMI. Distance to rivers was the least important variable. The Maxent model described chestnut habitat as sandy, high elevation, low-slope terrain with lower soil pH.

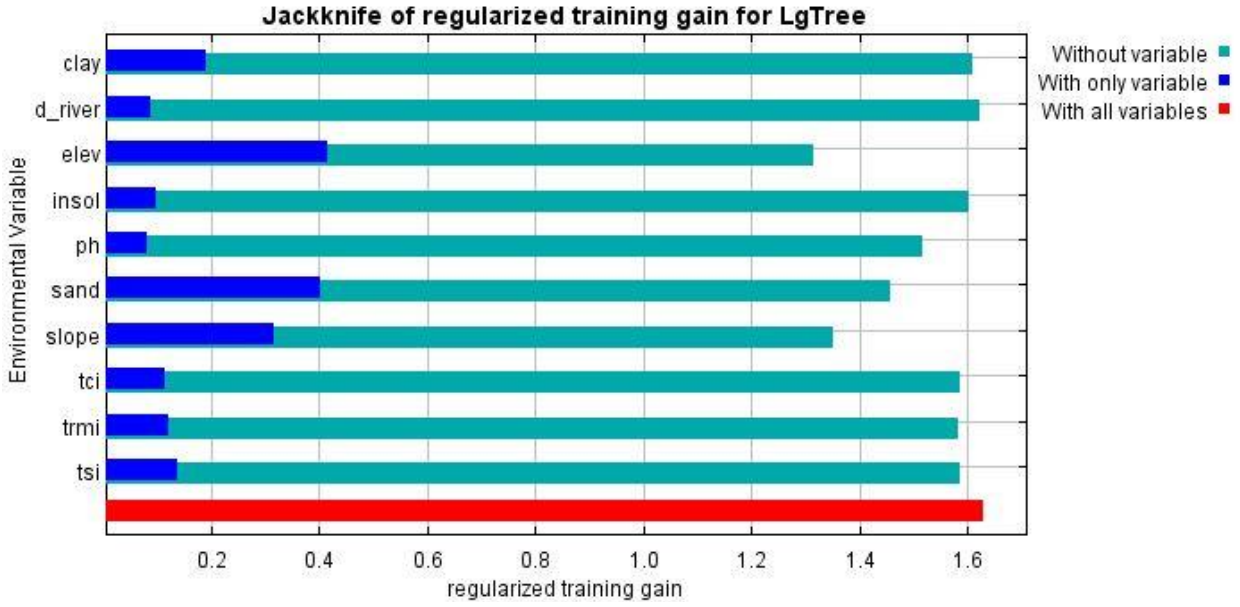


Figure 6: Jackknife diagram from Maxent showing relative variable importance when run by itself (blue) or excluded from a run (teal).

Table 5: Percent contribution and permutation importance for each habitat variable in the Maxent models for 2013, 2050, and 2070. Variable importance rank is shown on the right hand side of each permutation importance column.

Variable	2013			2050			2070		
	Percent contribution	Permutation importance		Percent contribution	Permutation importance		Percent contribution	Permutation importance	
elev	28.3%	25.4	1	31.1%	21.9	1	32.1%	28	1
sand	27.1%	27.3	2	25.3%	30.1	2	26.4%	30.9	2
slope	17.1%	21.3	3	16.5%	21.6	3	17.2%	19.7	3
trmi	7.4%	6.6	4	7%	7	4	7.2%	4.2	4
tci	6%	0.7	5	2.9%	1.2	7	2.8%	0.7	6
ph	4.2%	5.3	6	6.1%	8.2	5	5.3%	8	5
d_river	3.2%	1.5	7	2.7%	0.4	8	2.2%	0.3	8
insol	2.2%	3	8	3.6%	4	6	1.7%	2.9	10
clim	1.8%	1.7	9	0.8%	1.5	11	1.1%	1.6	11
clay	1.6%	5	10	1.9%	2.2	10	2.3%	2.6	7
tsi	1%	2.1	11	2%	2	9	1.7%	1.1	9

The Maxent model had an AUC of 0.939, meaning the model fit the presence data well. The threshold was set at 0.17, which is the average 10th percentile training presence logistic threshold for all ten runs (Figure 7). It was chosen because it displays suitable habitat that includes at least 90% of the input presence data; in case there were any outliers or errors in the data, this threshold is safe because it uses the best 90% to determine potential chestnut habitat in Shenandoah. At this threshold, the model classified 91% of the original chestnut presence points correctly and 9% incorreclty (Table 3). The Maxent model predicted that 19.15% (15,223 ha) of Shenandoah is suitable chestnut habitat (Table 4).

4) Ensemble Model

The ensemble model was formed by overlaying the three non-climate change models (Figure 8). There was very little area where all three models agreed, but a significant amount of predicted habitat that was the same in two of the models. The combined model predicted that 25.37% (20,168 ha) is suitable chestnut habitat (Table 3). Of this, 1.61% (1,282 ha) is agreed upon by all three models, 8.37% (6,651 ha) is agreed upon by two models, and 15.39% (12,234 ha) is considered habitat by only one model. Since the ensemble model is a combination of the GLM, CART, and Maxent models, it will have the same important variables: elevation, sand, and slope (Table 6). To get a better idea of the three most important habitat variables (elevation, sand, and slope), the model outputs, and climate scenarios, refer to Figure 9.

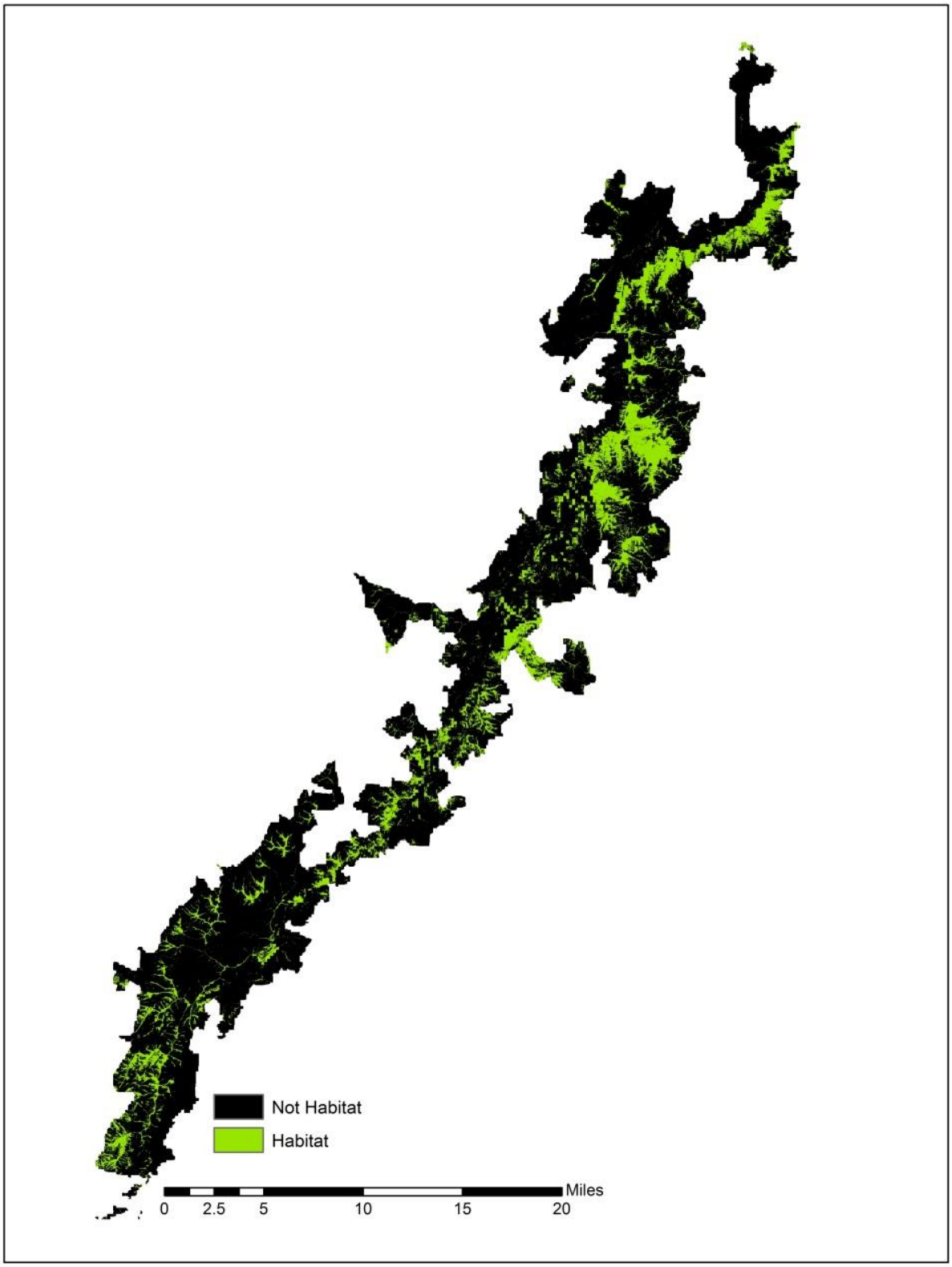


Figure 7: Maxent output of chestnut habitat.

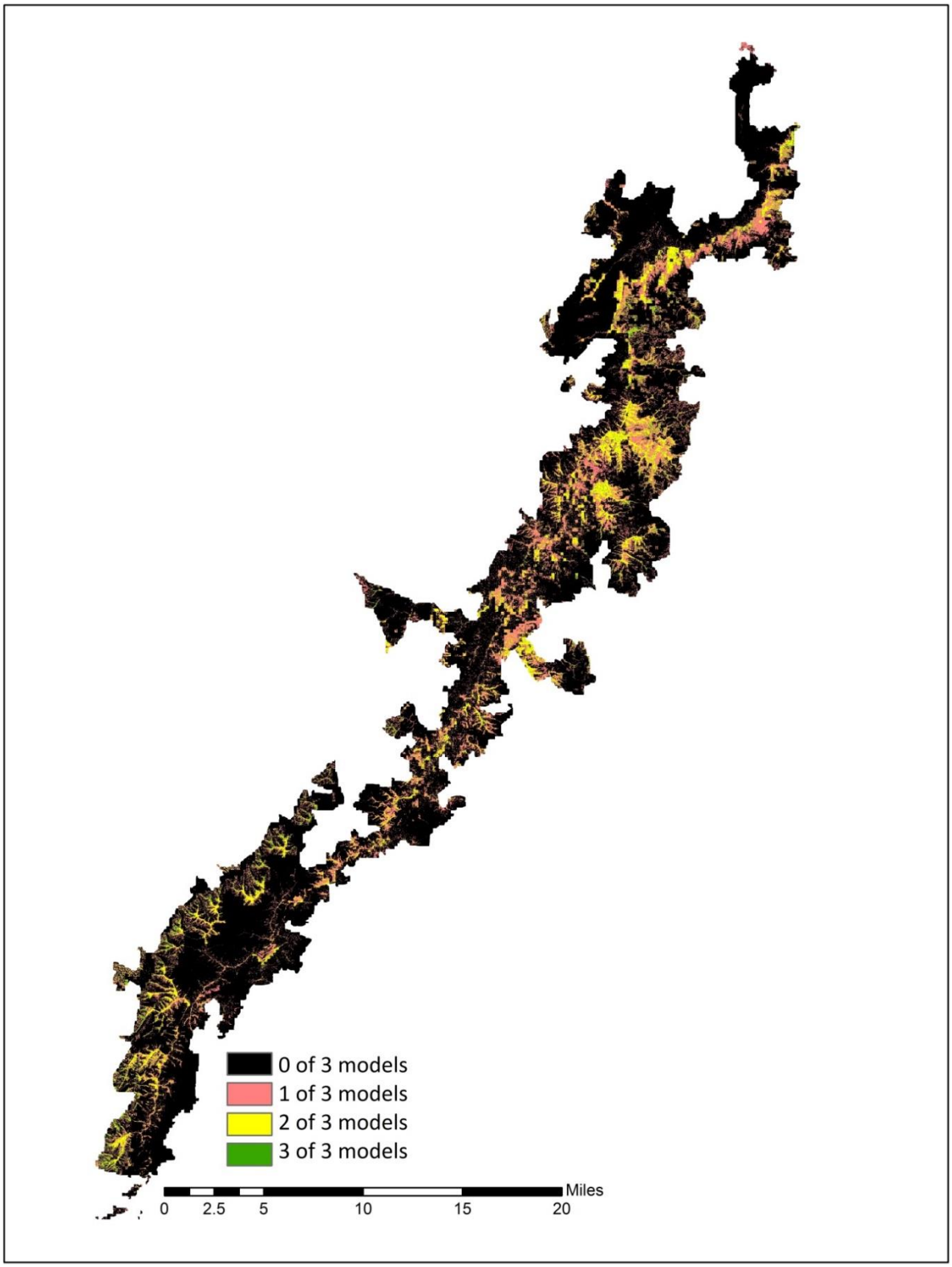


Figure 8: Combined model of GLM, CART, and Maxent outputs showing chestnut habitat.

Table 6: Variable rank from all three models (GLM, CART, and Maxent). More important variables have higher total scores, which are the sums of the scores of the three models.

	Elev	% Sand	Slope	TCI	TRMI	pH	TSI	% Clay	Dist. to Rivers	Insol
GLM	4	5	3	2	1					
CART	5	3	4	2			1			
Maxent	4	5	3		1	2				
Score	**13**	**13**	**10**	**4**	**2**	**2**	**1**	**0**	**0**	**0**
Rank	1	1	2	3	4	4	5	6	6	6

1 = Least important variable in the model
5 = Most important variable in the model

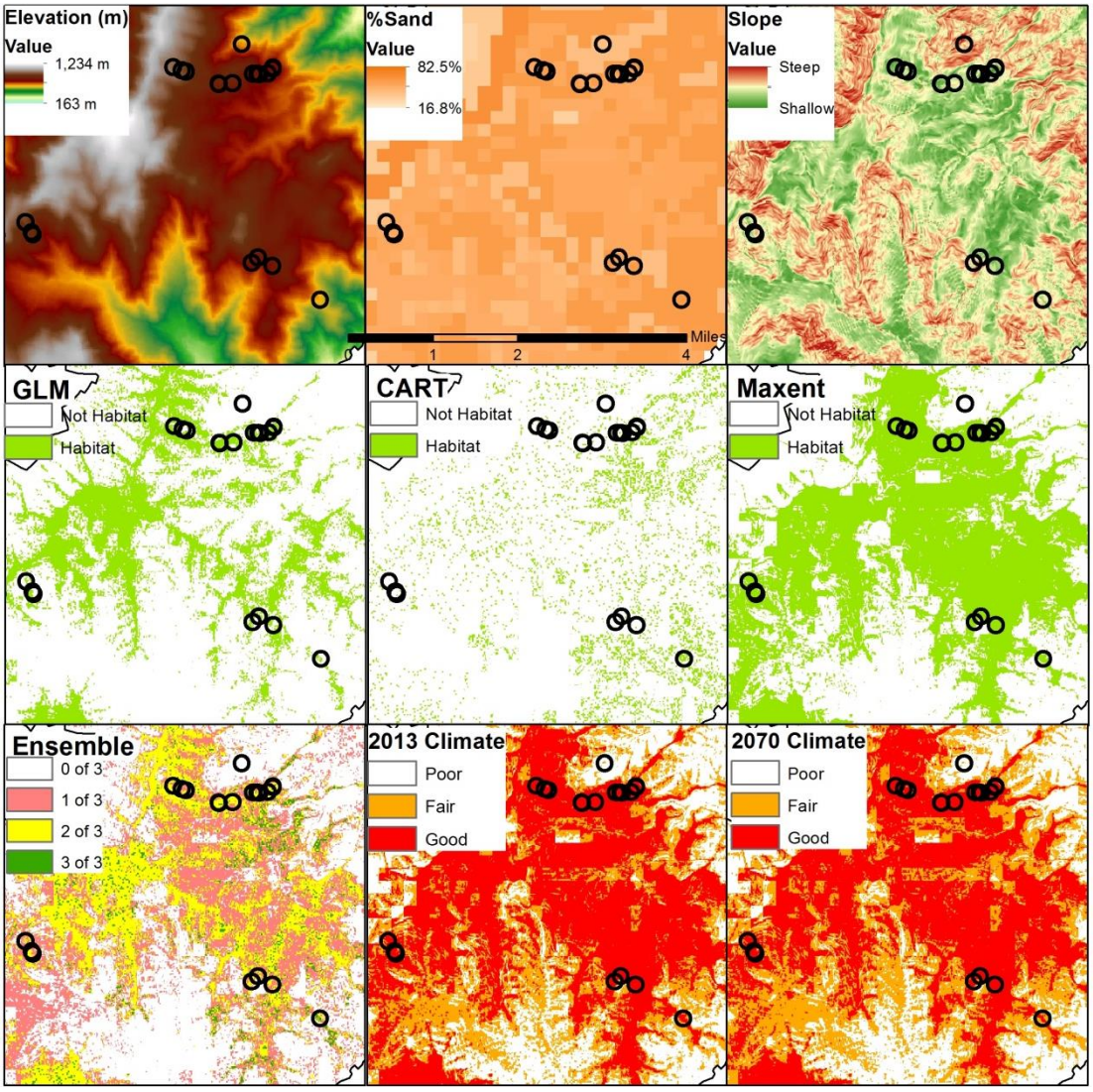


Figure 9: Close-up view of the three most important habitat variables (elevation, sand, and slope; top row), the four models (GLM, CART, Maxent, and Ensemble; middle row and first in bottom row), and climate change habitat projections for 2013 and 2070 (bottom row).

5) Maximum Entropy Climate Projection Model

The Maxent climate prediction model was the run with an eleventh variable (maximum temperature in the warmest month) and was projected to both 2050 and 2070 to predict possible chestnut habitat in the future as the climate warms (Figure 10). Probability of future habitat was divided into classes of "good," "fair," and "poor" habitat based on two thresholds output by Maxent (Table 7) (Phillips et al. 2006). The lower threshold separating "poor" from "fair" habitat is defined at Maxent's minimum training presence logistic threshold for each model. This threshold includes all of the presence points used in the model, and thus show a less conservative prediction of habitat. Any habitat below this threshold was classified as "poor." The second cutoff was defined as the maximum training presence logistic threshold for the Maxent model and caps the habitat considered "fair." This is a conservative threshold because it only includes the species occurrence points that contribute the most to habitat prediction to define areas above, which are considered good habitat for American chestnuts. In Shenandoah, the Maxent climate projection models predicted that 13.5% (10,722 ha) of the terrain was good habitat today, 14.59% (11,596 ha) would be good habitat in 2050, and 11.56% (9,191 ha) would be good habitat in 2070 due to climate warming (Table 7).

Table 7: Total area classified as habitat across all three quality categories.

Year	Threshold	Habitat Quality	Area of habitat (%)	Area of habitat (ha)
2013	0-0.05	poor	55.93%	44,471
	0.05-0.22	fair	30.57%	24,303
	0.22-1	good	13.50%	10,733
2050	0-0.05	poor	58.78%	46,732
	0.05-0.22	fair	26.64%	21,178
	0.22-1	good	14.59%	11,596
2070	0-0.05	poor	58.78%	46,732
	0.05-0.27	fair	29.66%	23,584
	0.27-1	good	11.56%	9,191

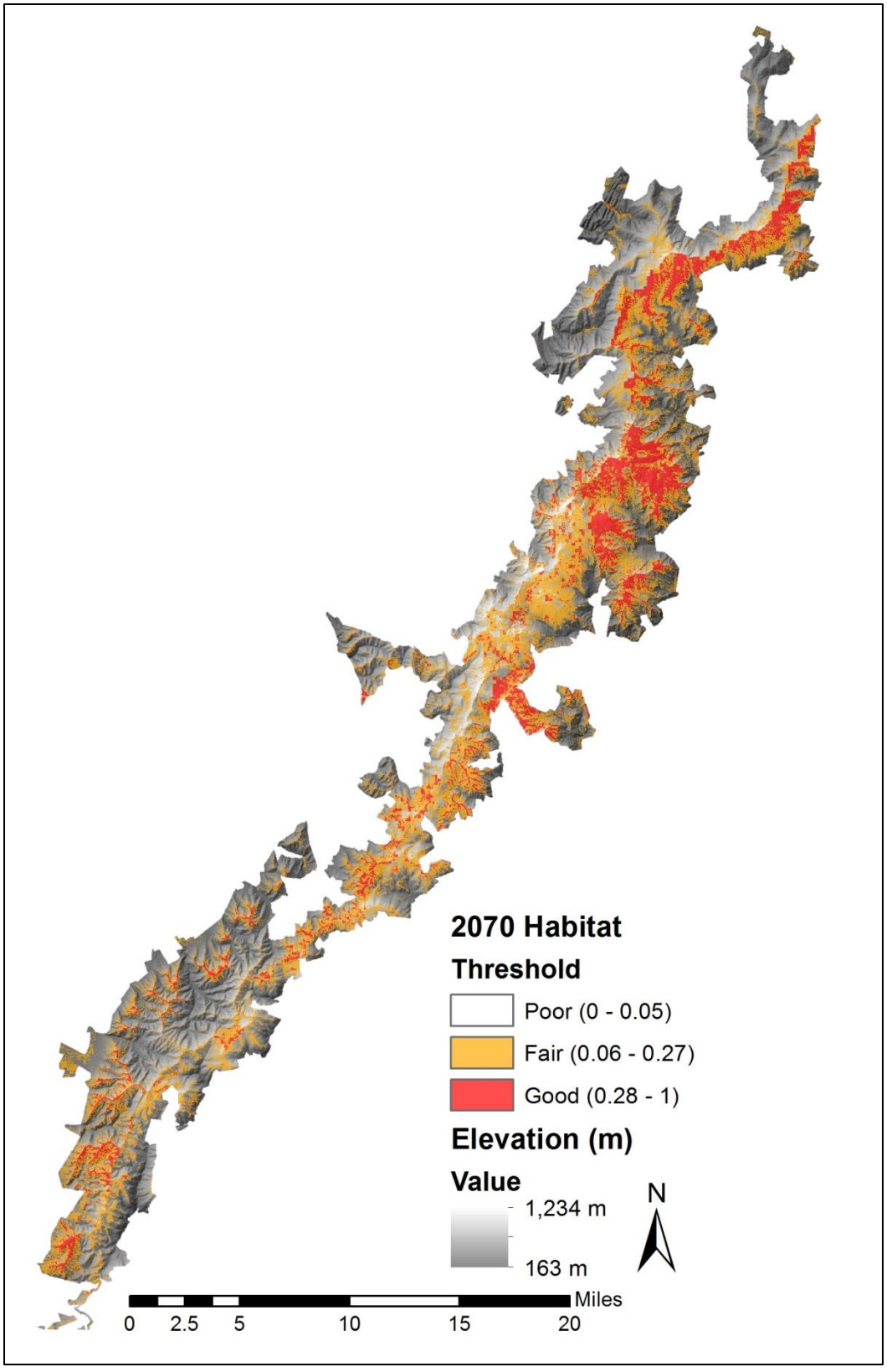


Figure 10: Maxent output for poor, fair, and good chestnut habitat in 2070 using the CCSM4 climate prediction model.

Overall, the Maxent climate prediction model does not show much change in habitat for 2050 and 2070, but there are small areas of habitat change (Figures 10, 11). In general, habitat for surviving chestnut trees stays close to the ridgetops. The probability of good habitat shifts northward between now and 2070; when the most likely habitat is compared, the probability of chestnuts habitat seems to decrease in the southern part of Shenandoah while increasing slightly in the northern section (Figure 12). While, these changes are very slight, they could suggest a subtle habitat migration northwards. However, the probability of chestnut habitat in most areas of Shenandoah remains unchanged between the two time periods.

Because the current and future temperature datasets were extrapolated into the future at a coarse resolution, they contain a high level of uncertainty. Additionally, temperature was not a strong predictor variable for any of the three climate scenarios, but elevation is a temperature proxy and had the strongest influence on the model (Table 5). Before the blight, chestnut was considered a generalist species; combined with the low-resolution temperature data, the models may be too coarse to detect any significant effects of rising temperature.

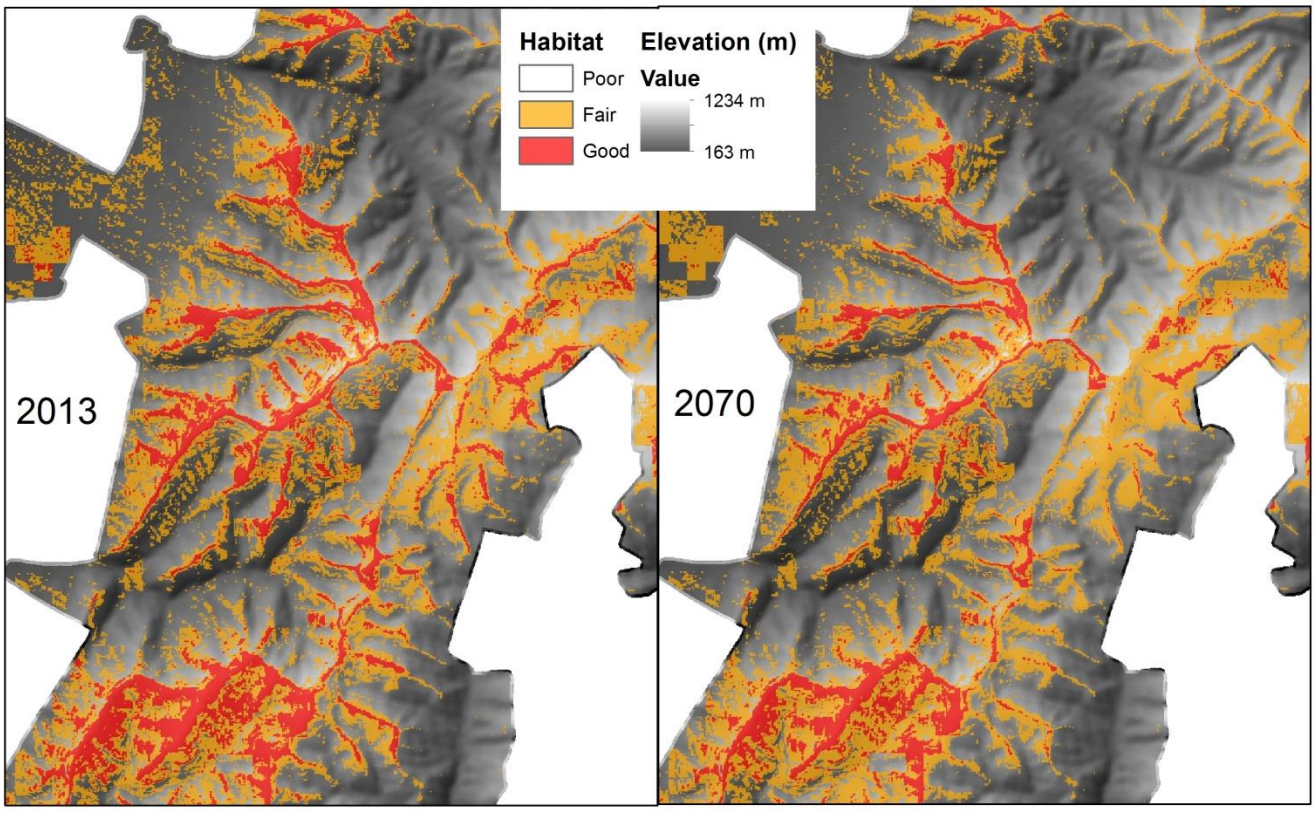


Figure 11: Zoomed-in area of predicted habitat for 2013 (left) and 2070 (right) in the Maxent climate model. There is slightly less habitat along the ridge-tops in 2070, but the change is subtle.

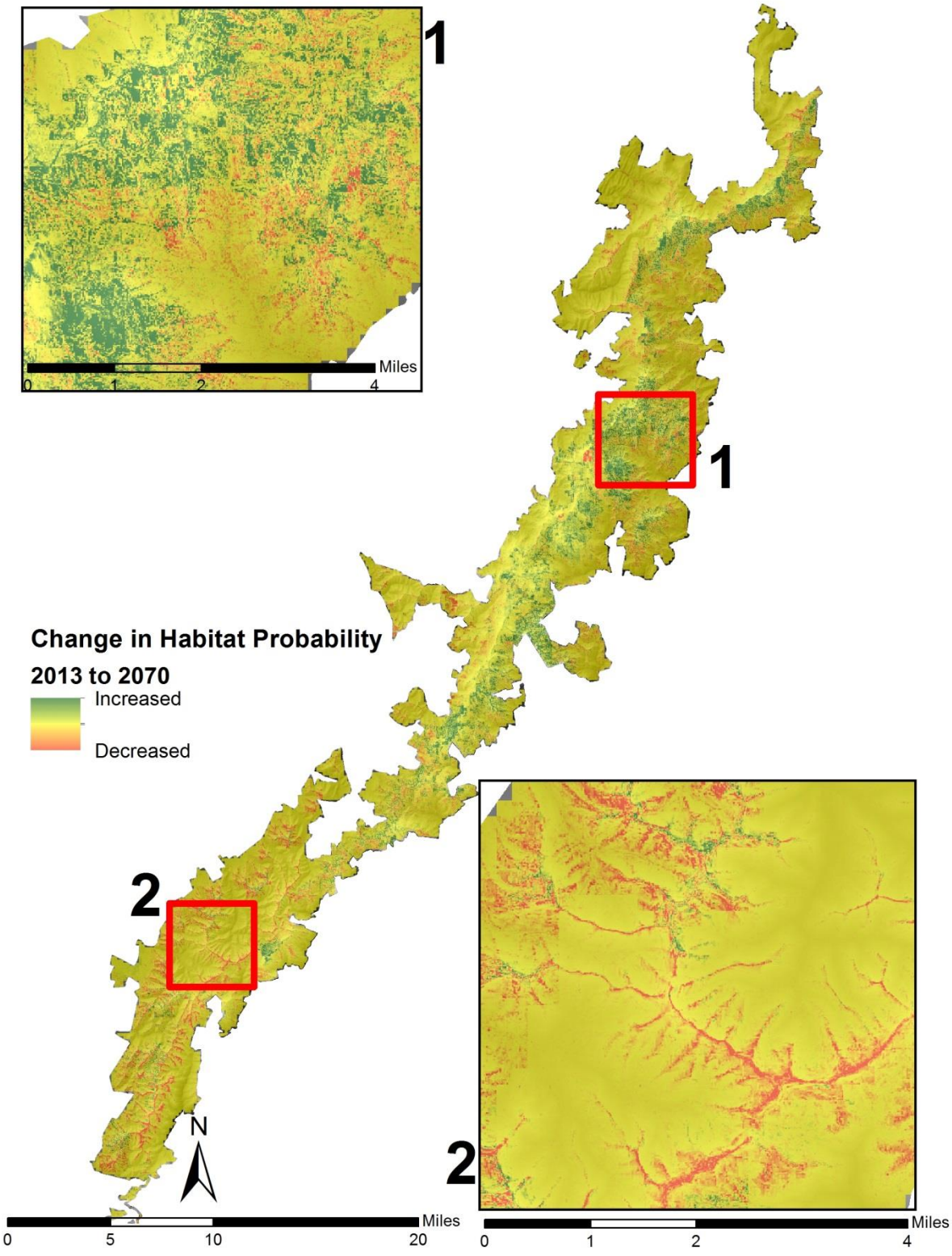


Figure 12: Change in habitat probability from 2013 to 2070. Red indicates a decreased probability of chestnut habitat between 2013 and 2070, while green indicates a slightly increased probability of chestnut habitat by 2070. Yellow indicates no change. All changes were very slight, but show a subtly northward shift.

Discussion

1) Modeling Results

All three models and the Maxent climate models showed that elevation, sand, and slope were the best predictors of chestnut habitat, which have been confirmed in other studies (Fei et al. 2007, 2012; Thomas-Van Gundy & Strager 2011). Overall, the models confirm that chestnuts prefer high, well-drained habitats. While GLM and CART models are easier to interpret, the Maxent model offers the most valuable habitat predictions because it is better able to fit a model to the species occurrence data without sacrificing repeatability (Loiselle et al. 2003; Phillips et al. 2006; Elith et al. 2011).

GLM and CART models require both presence and absence points. Lacking true absences, pseudo-absence points can be substituted, but can degrade the result as the models assume they are true absences. Despite using a random sampling method that avoided the true species occurrence points, there still could have been numerous false absences due to the small size of the dataset. Field sampling was heavier in the north and south sections of Shenandoah; the narrow middle section was not heavily sampled despite having similar habitat characteristics to the north and south sections of the park (Figure 2). Since there was no data collection in this section of the park, random points located there (and thus away from input species occurrence points) may have actually fallen on good chestnut habitat. This leads to the risk of false negatives in the dataset. Regardless of this potential sampling bias, inferior models still add to habitat prediction and robustness of the combined model (Loiselle et al. 2003). These results for larger trees are conservative: sprouts exist over a wider extent, but the data analyzed here consist of limited locations of larger chestnuts that have been able to thrive and reach maturity. Adding more presence points could allow for more robust models.

In the extrapolated climate models to 2050 and 2070, predicted chestnut habitat did not change significantly from the present. The most likely explanation for this result is that maximum temperature in the warmest month did not have a significant effect on chestnut habitat in 2013 or any future year, or that temperature effects were captured by the elevation variable. Reasons for this lack of an effect may have to do with the uncertainty of climate predictions and the large scale of the data. Climate extrapolations are an uncertain science to begin with, and large-scale climate models were downscaled to 1km for use in this model (Hijmans et al. 2005). This may be a good sign for American chestnuts, however: if climate truly does not have much of an effect on habitat suitability, the trees might be able

to thrive as they grow up in Shenandoah. If other tree species do not fare so well in a warming environment, this may open a new ecological niche for chestnuts.

However, American chestnuts were considered a generalist species before the introduction of chestnut blight (Thomas-Van Gundy & Strager 2011). The minimal change in habitat under a conservative warming scenario could echo this tolerance for a wide range of temperatures. Since these models only look at *Castanea dentata*, it is impossible to say how other tree species will react to climate warming. The expansion, reduction, or shift in good habitat for surviving chestnuts may rely heavily on the movement of other competing species (Shugart & West 1977).

2) Data Quality

Sampling procedures are designed to balance data biases and feasible fieldwork, so the computed models contain some biases as a result of data collection methods. The Shenandoah sampling procedure restricted volunteers to sampling along designated trails in the park, so presence points are restricted to those trees visible from a trail. Additionally, trails are more likely to be located on gentler slopes that are easier and less dangerous for hikers to navigate. Large chestnuts that aren't visible from a trail were not located and therefore considered as false absences. Their presence points were not logged in the dataset, which excludes valuable data.

Additionally, the model contained just 57 presence points, which were located over 42% of the side trail mileage in Shenandoah National Park. Trails were picked by volunteers on the basis of hiking desirability and ease of access, so there are many more remote or difficult trails that were not surveyed. Since participants tended to live near the northern or southern ends of Shenandoah, those areas were sampled the most heavily. Consequently, the narrow middle section of the park has no samples, and thus the predictions for that section are far weaker than the habitat surfaces for areas where numerous large trees were recorded (Figure 2).

Since a majority of the environmental variables considered in this analysis are based off a one-third arc second (~10m) DEM, the resolution is suitable to predict relatively fine-scale habitat differences. However, the soil data was based on a lower resolution dataset and thus provides a coarser (30m) scale analysis (SSURGO 2013). It is also important to note that areas covered by a single soil type are classified based on averages of their attributes (SSURGO 2013). This means that the pH, clay content, and sand content are averaged over the extent of a soil type. While this cuts down on the file size, it

diminishes the accuracy of the dataset. For that reason, the soil data is not as precise as elevation-based datasets.

As stated in the beginning, these environmental habitat models only tackle one piece of the puzzle. American chestnut genetics and biology of *Cryphonectria parasitica* are important factors in determining whether a given tree will survive to reproductive maturity or succumb to the blight at a young age. While these models do not consider tree genetics or blight biology, they are still significant factors impacting tree survival, as environmental factors play a large role in habitat suitability.

Only large trees with GPS locations were included in this analysis because I wanted to narrow the focus to trees that had survived to reproductive maturity. Because chestnut was once known as a generalist species, it may not have specific affinities for narrow environmental ranges. Small trees and sprouts would exhibit this generalist character and may not show any meaningful results when modeled. However, habitat quality may assist genetic prowess in helping these "large trees" survive. Even if these habitat models really point to areas that disadvantage chestnut blight and *Phytophthora cinnamomi* root rot or minimize intraspecies competition, this knowledge is still useful to TACF when hunting for wild survivors and planning progeny tests.

Finally, American chestnut are considered a generalist species and thus able to grow in a variety of environmental conditions. Instead of actual chestnut habitat preferences, these models may reflect places where chestnut trees experience the least competition from other plant species. Given their thirst for full-sun conditions yet blight-induced restrictions to the forest understory, large trees may grow better in locations where other tall competing vegetation grows poorly. However, regardless of whether these models show true chestnut preferences or simply locations where trees can survive to maturity, they are still important in pinpointing locations where chestnut reintroduction will be successful.

3) Model Assumptions and Biases

GLM and CART models are easy to construct, simple to interpret, and provide a quick way to glean information about the species in question. However, they are less suited to this type of data collection because they require absence points, where we only have randomly-generated absence (or pseudo-absence) points. Without field verification, it is impossible to know if these sites contain chestnuts or not. Therefore, the models treat the underlying environmental variables of pseudo-absence points as characteristics of poor chestnut habitat. Because of this, GLM and CART models work best with

random plot-based sampling, which allows for a better determination of presence and absence (Elith et al. 2011).

One way to solve this problem would be to determine true absence points in the field, which would involve marking coordinates for locations where chestnuts are not present. However, this raises a question of true habitat unsuitability or simply that chestnuts have not occurred at that location by random chance (Loiselle et al. 2003). Even if certain habitat is suitable for chestnuts, they may not occupy that entire suitable habitat.

Because Maxent is a presence-only model, it is considered more appropriate; however, like GLM and CART models, it too runs the risk of generating false absence points (Phillips et al. 2006; Elith et al. 2011; Yackulic et al. 2013). Maxent models are also subject to higher sample selection biases because more intensively-sampled areas have a stronger effect on the model outputs (Loiselle et al. 2003; Elith et al. 2011). In this case, areas around trails were intensively sampled, while areas in the middle of the forest were not sampled at all. Without confirmed absence points, those un-sampled areas in the middle of the forest may not be classified as ideal habitat because there are no presence records there. This is an unavoidable data collection bias, as bushwhacking to find chestnuts in Shenandoah was highly discouraged.

4) Model Comparison

Overall, all models (but especially GLM and CART models) require numerous data points to make good predictions, so the models suffer with fewer chestnut presences. Because CART models generate decision trees based on computed break points, they suffer more from a small dataset (De'Ath & Fabricius 2000; Hernandez et al. 2006). GLMs are less affected by dataset size than CARTs, but the addition of more data points can result in a much more predictive linear regression function (Vayssieres et al. 2000; Hernandez et al. 2006). Of these three models, Maxent is the most reliable for rare species and small datasets (Hernandez et al. 2006; Kumar & Stohlgren 2009).

The statistics behind Maxent are more difficult to interpret than for GLM and CART models. Maxent is a machine learning program and is often termed as a "black box" that uses difficult to visualize algorithms to reach its habitat prediction output (Phillips et al. 2006, 2008; Elith et al. 2011). Despite this, Maxent is regarded as a competitive and high-performing model despite its opacity.

Conclusion

The GLM, CART, Maxent, and ensemble models described here complement the ENFA model from Fei et al. (2007). Each model identifies similar important habitat variables across Shenandoah National Park and even in other locations in the American chestnut's former range. Such agreement adds confidence to my predictions that elevation, sand content, and slope are important factors for surviving chestnut trees in Shenandoah and even elsewhere in the South.

Surviving chestnut trees are important to TACF's backcross breeding program as they provide new genetics and local adaptations that are crucial to include in a blight-resistant tree. Chestnuts that survive to reproductive maturity in the wild may hold the keys to an American line of partial blight resistance that can augment TACF's efforts with Chinese chestnut genes. Conversely, there may be some characteristics of this combination of habitat variables that is hostile to chestnut enemies such as the blight and *Phytophthora cinnamomi*, a water-borne root fungus. Even though these habitat models do not consider genetics or competition between other species, they are important because they pinpoint several locations that might have surviving chestnuts. For a small non-profit with limited time and funds, narrowing the search area is necessary to take strides in restoring this tree to the Appalachian forest.

Looking toward the future, models like these will become more important and can be refined to create more reliable habitat predictions in the face of climate warming. While these models only describe how chestnut habitat will change, other models for different tree species can help us to understand any possible competition effects (Shugart & West 1977).

While this study was restricted to Shenandoah National Park, it could be useful for chestnut restoration over the length of the East Coast. The Shenandoah side trails protocol was designed to be applicable to other areas throughout the American chestnut's former range. For this reason, it would be straightforward to implement the same data collection methods in other National Parks, National Forests, or hiking trails throughout the Appalachians. Shenandoah was selected for this study because it has enjoyed many years of preservation. Chestnuts typically do not occur on recovered agricultural lands, so historically forested Shenandoah eliminates land use biases (Fei et al. 2007).

As chestnut research and data collection continue in Shenandoah National Park, models will improve given the influx of additional presence points (Loiselle et al. 2003; Yackulic et al. 2013). While venturing off designated trails remains impractical, achieving 100% of trail surveys increases the area

sampled over Shenandoah's regular network of side trails and improves the quality of the dataset. Additional sampling could incorporate plot-based techniques and incorporation of true absence points to further improve the models.

It is important to remember that models do not show the absolute truth; they are predictions based on what we know and may be biased due to what we do not know. However, modeling is important to give us a picture for the future of the American chestnut. These models provide valuable information for finding American chestnuts in Shenandoah National Park and locating possible sites for plantings and restoration.

Acknowledgements

Thanks is extended to numerous Nicholas School faculty, including Dr. Jennifer Swenson, the advisor for this project; John Fay & Pete Harrell, for GIS and GPS support; and Dr. Dean Urban, for assistance with maximum entropy modeling. Many people with TACF made this project possible: Kathy Marmet and Matt Brinckman, for advising my internship; John Scrivani, for GIS support & protocol construction; Taylor Cochran, for hiking Shenandoah side trails and assisting me with trainings; Katy McCune, for protocol support; and numerous volunteers, for data collection in Shenandoah.

Literature Cited

Anagnostakis, Sandra L. 1995. The Pathogens and Pests of Chestnuts. Advances in Botanical Research 21: 125-145.

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

Dollison, R.M. 2010. The National Map: New viewer, services, and data download: U.S. Geological Survey Fact Sheet 2010–3055, 2 p. (Also available at http://pubs.usgs.gov/fs/2010/3055/.)

Dufour, C., and E. Crisfield, eds. 2008 The Appalachian Trail MEGA-Transect. Harpers Ferry, WV: Appalachian Trail Conservancy

Elith, Jane, Catherine H. Graham, Robert P. Anderson, Miroslav Dudı´k, Simon Ferrier, Antoine Guisan, Robert J. Hijmans, Falk Huettmann, John R. Leathwick, Anthony Lehmann, Jin Li, Lucia G. Lohmann, Bette A. Loiselle, Glenn Manion, Craig Moritz, Miguel Nakamura, Yoshinori Nakazawa, Jacob McC. Overton, A. Townsend Peterson, Steven J. Phillips, Karen Richardson, Ricardo Scachetti-Pereira, Robert E. Schapire, Jorge Sobero´n, Stephen Williams, Mary S. Wisz and Niklaus E. Zimmermann. 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29: 129-151.

Elith, Jane, Steven J. Phillips, Trevor Hastie, Miroslav Dudik, Yung En Chee, and Colin J. Yates. 2011. A statistical explanation of MaxEnt for ecologists. Diversity and Distributions 17: 43-57.

ESRI 2013. ArcGIS Desktop: Release 10.2. Redlands, CA: Environmental Systems Research Institute.

Fei, Songlin, Joe Schibig, and Mark Vance. 2007. Spatial habitat modeling of American chestnut at Mammoth Cave National Park. Forest Ecology and Management 252: 201-207.

Fei, Songlin, Liang Liang, Frederick L. Paillet, Kim C. Steiner, Jingyun Fang, Zehao Shen, Zhiheng Wang, and Frederick V. Hebard. 2012. Modeling chestnut biogeography for American chestnut restoration. Diversity and Distributions 18: 754-768.

Freinkel, Susan. 2007. American Chestnut: The Life, Death, and Rebirth of a Perfect Tree. University of California Press.

Goslee, S.C. and D.L. Urban. 2007. The ecodist package for dissimilarity-based analysis of ecological data. Journal of Statistical Software 22(7): 1-19.

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

Guisan, Antione, Stuart B. Weiss, and Andrew D. Weiss. 1999. GLM versus CCA spatial modeling of plant species distribution. Plant Ecology 143: 107-122.

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

Hijmans, R.J., S.E. Cameron, J.L. Parra, P.G. Jones and A. Jarvis, 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25: 1965-1978.

Hirzel, A.H., G. Le Lay, V. Helfer, C. Randin, A. Guisan. 2006. Evaluating the ability of habitat suitability models to predict species presences. Ecological Modeling 199: 1422–2152.

Kopecky, Martin, and Stepanka Cizkova. 2010. Using topographic wetness index in vegetation ecology: does the algorithm matter? Applied Vegetation Sciences 13: 450-459.

Kumar, Sunil, and Thomas J. Stohlgren. 2009. Maxent modeling for predicting suitable habitat for threatened and endangered tree *Canacomyrica monticola* in New Caledonia. Journal of Ecology and natural Environment 1(4): 94-98.

Loiselle, Bette A., Christine A. Howell, Catherine H. Graham, Jaqueline M. Goerck, Thomas Brooks, Kimberly G. Smith, and Paul H. Williams. 2003. Avoiding pitfalls of using species distribution models in conservation planning. Conservation Biology 17(6): 1592-1600.

Marmet, Kathleen, and Sara Fitzsimmons. 2008. The Appalachian Trail MEGA-Transect Project: A preliminary pilot project report. Science and Natural History 22(2): 10-18.

McNab, W. Henry. 1989. Terrain Shape Index: Quantifying effect of minor landforms on tree height. Forest Science 15(1): 91-104.

Montoya, Daniel, Drew W. Purves, Itziar R. Urbieta, and Miguel A. Zavala. 2009. Do species distribution models explain spatial structure within tree species ranges? Global Ecology and Biogeography 18: 662-673.

NCAR Community Climate System Model 4.0 (CCSM4.0). 2012. National Center for Atmospheric Research, USA. Available online at http://www.cesm.ucar.edu/models/ccsm4.0/

Lutts, Ralph H. 2004. Like manna from God: The American chestnut trade in southwestern Virginia. Environmental History 9: 497:525.

Paillet, Frederick L. 2002. Chestnut: History and ecology of a transformed species. Journal of Biogeography 29: 1517-1530.

Parker, Alfred J. 1982. The topographic relative moisture index: an approach to soil-moisture assessment in mountain terrain. Physical Geography 3(2): 160-168.

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

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

Phillips, Steven J. and Miroslav Dudik, 2008. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography 31: 161-175.

R Development Core Team. 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rhoades, Chuck, David Loftis, Jeffrey Lewis, and Stacy Clark. 2009. The influence of silvicultural treatments and site conditions on American chestnut (Castanea dentata) seedling establishment in eastern Kentucky, USA. Forest Ecology and Management 258: 1211-1218.

Shugart, H.H., and D.C. West. 1977. Development of an Appalachian deciduous forest succession model and its application to assessment of the impact of the chestnut blight. Journal of Environmental Management 5: 161-179.

Sing, T., O. Sande, N. Beerenwinkel, and T. Lengauer. 2005. "ROCR: visualizing classifier performance in R." Bioinformatics, 21(20): 7881. http://rocr.bioinf.mpi-sb.mpg.de.

Sisco, Paul H. n.d. Breeding blight-resistant American chestnut trees. The American Chestnut Foundation, Southern Appalachian Regional Office.

SSURGO Soil Survey Staff, Natural Resources Conservation Service, United States Department of Agriculture. 2013. Web Soil Survey. Available online at http://websoilsurvey.nrcs.usda.gov/

Sorensen, R., U. Zinko, and J. Seibert. 2006. On the calculation of the topographic wetness index: evaluation of different methods based on field observations. Hydrology and Earth System Sciences, 10: 101-112.

Thomas-Van Gundy, Melissa, and Michael Strager. 2011. Site characteristics of American chestnut, oak, and hickory witness trees on the Monongahela National Forest, West Virginia. Proceedings of the 17th Central Hardwood Forest Conference GTR-NRS-P-78 (2011): 208-218.

Thorn, J.S., V. Nijman, D. Smith, and K.A.I. Nekaris. 2009. Ecological niche modeling as a technique for assessing threats and setting conservation priorities for Asian slow lorises (Primates: Nycticebus). Diversity and Distributions 15: 289-298.

Vayssieres, M.P., R.E. Plant, and B.A. Allen-Diaz. 2000. Classification trees: an alternative nonparametric approach for predicting species distributions. Journal of Vegetation Science 11: 679-694.

Watershed Boundary Dataset for Virginia. Available URL: http://datagateway.nrcs.usda.gov

Yackulic, Charles B., Richard Chandler, Elise F. Zipkin, J. Andrew Royle, James D. Nichols, Evan H. Campbell Grant, and Sophie Vera. 2013. Presence-only modeling using MAXENT: when can we trust the inferences? Methods in Ecology and Evolution 4: 236-243.

Zaniewski, A. Elizabeth, Anthony Lehmann, and Jacob McC. Overton. 2002. Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. Ecological Modeling 157: 261-280.

Appendix A: GLM Output

```
# GLM

# Quick EDA and read-in
trees.data<-read.csv("trees.csv")
habitat.data<-read.csv("habitat.csv")

names(trees.data)
names(habitat.data)
pairs(habitat.data[,2:11])
summary(habitat.data)

cade<-trees.data$cade
cade

cade.data<-cbind(cade, habitat.data[,-1])
names(cade.data)

# Fit the GLM
cade.glm<-glm(as.factor(cade)~., data=cade.data, family=binomial)
attributes(cade.glm)
print(cade.glm)

# output:
#Call:  glm(formula = as.factor(cade) ~ ., family = binomial, data = cade.data)
#Coefficients:
#(Intercept)     d_river        elev       insol       slope         tci
# -12.174870    -0.000176     0.005088     0.004399    -0.081305    -0.135383
#       trmi         tsi        clay        sand          ph
#  -0.054111     0.133374     0.139958     0.100584     0.386365
#Degrees of Freedom: 156 Total (i.e. Null);  146 Residual
#Null Deviance:      205.7
#Residual Deviance: 127.7        AIC: 149.7

plot(cade.glm) # the weird plots

#run the next 2 lines together
plot(cade.glm$linear.predictor, cade.glm$fitted.values, col="green", ylim=c(0,1), xlab="Linear Predictor",
ylab="P(habitat)")
points(cade.glm$linear.predictor, cade, col="blue")

summary(cade.glm) # null and resid deviance: how much deviance is explained by the model
# SIGNIFICANT VARIABLES:
# elevation (0.003261)
# slope (0.022821)
# tci (0.025500)
# trmi (0.033197)
# clay (0.091788)
# sand (0.000218)
# Null deviance: 205.72  on 156  degrees of freedom
# Residual deviance: 127.70  on 146  degrees of freedom
```

```
# AIC: 149.7

anova(cade.glm, test="Chi")
# SIGNIFICANT VARIABLES
# elevation (2.772e-06)
# slope (0.0008977)
# tci (0.0335491)
# trmi (0.0193044)
# clay (0.0092204)
# sand (7.443e-06)

cade.glm.p<- 1-pchisq(cade.glm$null-cade.glm$deviance,146)
cade.glm.p
# p=0.9999992
cade.glm.r2<-1-(cade.glm$deviance/cade.glm$null)
cade.glm.r2
# r2 = 0.3792417

# step-wise model
cade.glm.step<-step(cade.glm)
summary(cade.glm.step)
# new r2:
1-(cade.glm.step$deviance/cade.glm.step$null)
# r2 = 0.3624921: model got worse, slightly


# threshold to a binary prediction a p(hab)=0.5
cade.glm.p50<-cade.glm$fitted.value
cade.glm.p50[cade.glm.p50<0.50]<-0
cade.glm.p50[cade.glm.p50>=0.50]<-1

# confusion matrix
table(cade.glm.p50,cade)
#           cade
#cade.glm.p50  0  1
#          0 88 15
#          1 12 42

# tuning model using ROC curves
library(ROCR)
cade.pred<-prediction(cade.glm$fitted.values,cade) # prediction object
cade.perf<-performance(cade.pred, "tpr", "fpr")

# plot ROC
plot(cade.perf, colorize=TRUE)
abline(0,1)
performance(cade.pred, "auc")
# 0.8878947

# sensitivity and specificity together
plot(performance(cade.pred, "sens"))
plot(performance(cade.pred, "spec"), add=TRUE)
```

```
# phi correlation (maximized for best total prediction)
plot(performance(cade.pred, "phi"))

source("cutoff.ROCR.R")

# retrieve actual cutoff values
cutoff.ROCR(cade.pred)     # accept the default: "tpr", target=0.95
# 0.4401751
cutoff.ROCR(cade.pred, method="max")       # maximize TPR + TNR
# 0.4401751
cutoff.ROCR(cade.pred, "x")                # intersection
# 0.3907426
cutoff.ROCR(cade.pred, "tpr", target=0.90)   # change the TPR target


# see how i did with tuning
cutoff<-cutoff.ROCR(cade.pred, method="max")
cade.glm.px<-cade.glm$fitted.value
cade.glm.px[cade.glm.px<cutoff]<-0
cade.glm.px[cade.glm.px>=cutoff]<-1
table(cade.glm.px,cade)
#          cade
#cade.glm.px  0  1
#         0 84 11
#         1 16 46
# model improves slightly with tuning, though cutoff of 0.5 was pretty close.



## GLM INPUTTING ALL 9 MILLION POINTS IN SNP
snp.data<-read.csv("glmpts.csv")
all.data<-cbind(cade, snp.data[,-1])
names(snp.data)


# Now ROCR to predict for all of SNP
# tuning model using ROC curves
library(ROCR)
snp.pred<-prediction(cade.glm$fitted.values,cade) # prediction object
snp.perf<-performance(cade.pred, "tpr", "fpr")

# plot ROC
plot(cade.perf, colorize=TRUE)
abline(0,1)
performance(cade.pred, "auc")
# 0.8878947

# sensitivity and specificity together
plot(performance(cade.pred, "sens"))
plot(performance(cade.pred, "spec"), add=TRUE)

# phi correlation (maximized for best total prediction)
plot(performance(cade.pred, "phi"))
```

Appendix B: CART Model Output

```
## CART Model

# Quick EDA and read-in
trees.data<-read.csv("trees.csv")
habitat.data<-read.csv("habitat.csv")

names(trees.data)
names(habitat.data)
pairs(habitat.data[,2:11])
summary(habitat.data)

cade<-trees.data$cade
cade

# Correlations
library(ecodist)
cor2m(as.matrix(cade),habitat.data[,-1])

#Results:
#d_river  0.1477826
#elev     0.3840124
#insol    0.0000000
#slope   -0.3347072
#tci     -0.2435036
#trmi    -0.2404827
#tsi      0.2403402
#clay    -0.1716172
#sand     0.3520950
#ph       0.0000000

par(mfrow=c(2,5))
boxplot(habitat.data$d_river~cade, ylab="Distance to Rivers")
boxplot(habitat.data$elev~cade, ylab="Elevation")
boxplot(habitat.data$clay~cade, ylab="% Clay")
boxplot(habitat.data$insol~cade, ylab="Insolation")
boxplot(habitat.data$ph~cade, ylab="pH")
boxplot(habitat.data$sand~cade, ylab="% Sand")
boxplot(habitat.data$slope~cade, ylab="Slope")
boxplot(habitat.data$tci~cade, ylab="Topographic Convergence")
boxplot(habitat.data$trmi~cade, ylab="TRMI")
boxplot(habitat.data$tsi~cade, ylab="Terrain Shape")

plot(habitat.data$elev, cade, xlab="Elevation", ylab="Chestnut", pch=19)


# TREE library for CART model
cade.data<-cbind(cade, habitat.data[,-1])
names(cade.data)

library(tree)
```

```
cade.tree<-tree(as.factor(cade)~., data=cade.data)

plot(cade.tree)
text(cade.tree,cex=0.6)

summary(cade.tree)
# Resid mean deviance: 0.3676 = 51.83/141
# Misclass error rate: 0.08917 = 14/157

print(cade.tree)


# Prune the tree
cade.tree.prune<-prune.tree(cade.tree, method="misclass")
cade.tree.prune
plot(prune.tree(cade.tree, method="misclass"))
# try 9 terminal nodes

cade.pt9<-prune.tree(cade.tree, method="misclass", best=9)
plot(cade.pt9)
text(cade.pt9, cex=0.6)
print(cade.pt9)


# Cross-validation, using misclass as the rule:
cade.tree.cv<-cv.tree(cade.tree, FUN=prune.misclass)
par(mfrow=c(2,1))
plot(cade.tree.prune)
plot(cade.tree.cv) # low dips = model sweet spot. Have that many nodes
# sweet spot might be 4 nodes
cade.pt4<-prune.tree(cade.tree, method="misclass", best=4)
plot(cade.pt4)
text(cade.pt4, cex=0.6)


# Confusion matrices
# full tree:
cade.pred<-predict(cade.tree, type="class")
table(cade.pred, cade)
#         cade
#cade.pred  0  1
#        0 94  9
#        1  6 48

#pruned to 9:
cade.pred9<-predict(cade.pt9, type="class")
table(cade.pred9, cade)
#          cade
#cade.pred9  0  1
#         0 94  8
#         1  6 49

#pruned to 4:
```

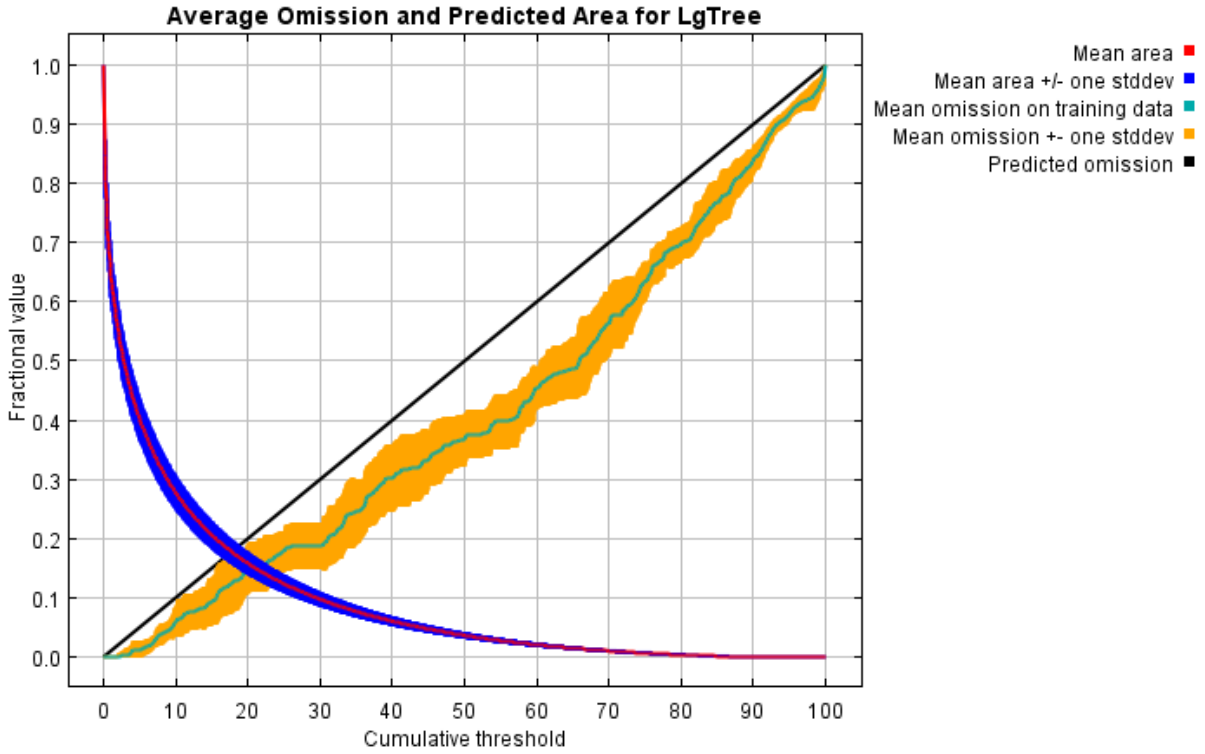
```
cade.pred4<-predict(cade.pt4, type="class")
table(cade.pred4, cade)
#         cade
#cade.pred4  0  1
#         0 89 12
#         1 11 45

# sum diagonals, divided by total, for % correct
```

Appendix C: Maxent Model Output (2013 to 2070)

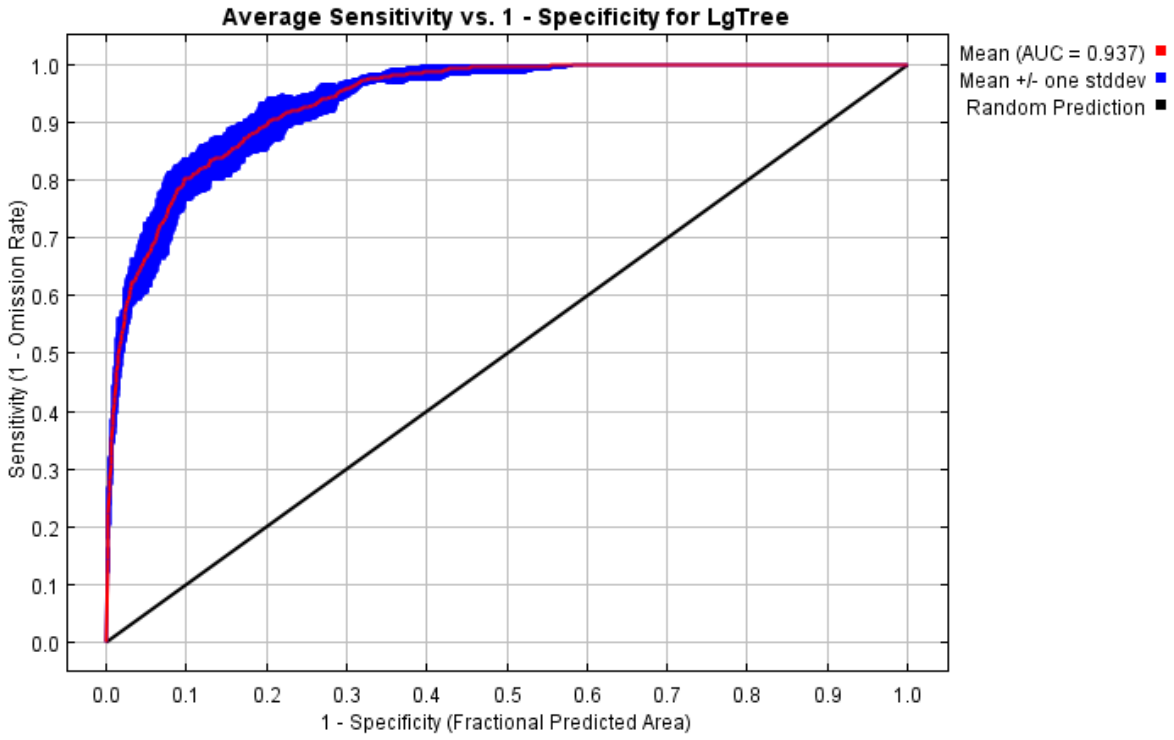
Analysis of omission/commission

The following picture shows the training omission rate and predicted area as a function of the cumulative threshold, averaged over the replicate runs.



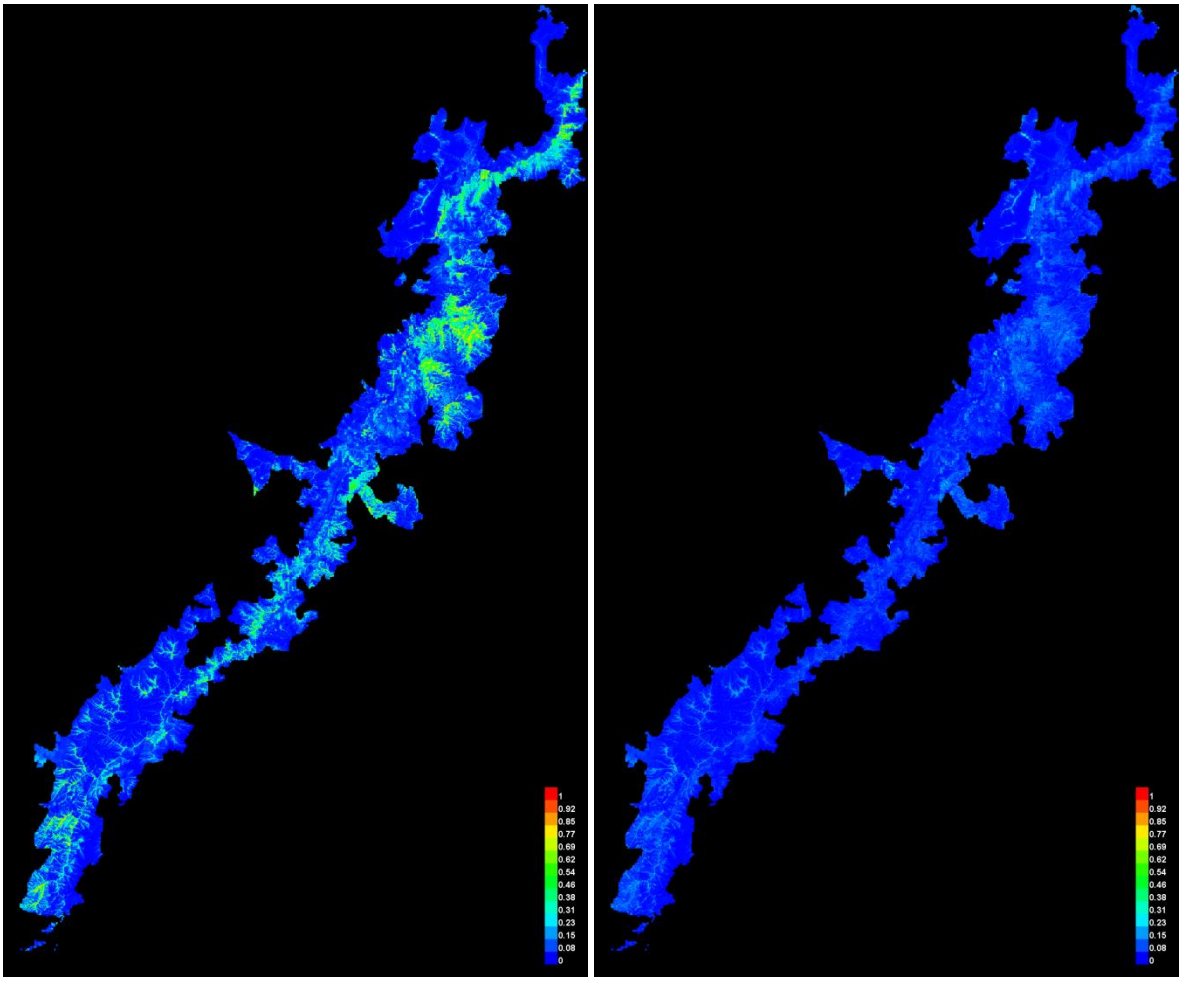
The next picture is the receiver operating characteristic (ROC) curve for the same data, again averaged over the replicate runs. Note that the specificity is defined using predicted area, rather than true commission (see the paper by Phillips, Anderson and Schapire cited on the help page for discussion of what this means). The average training AUC for the replicate runs is 0.937, and the standard deviation is

0.008.

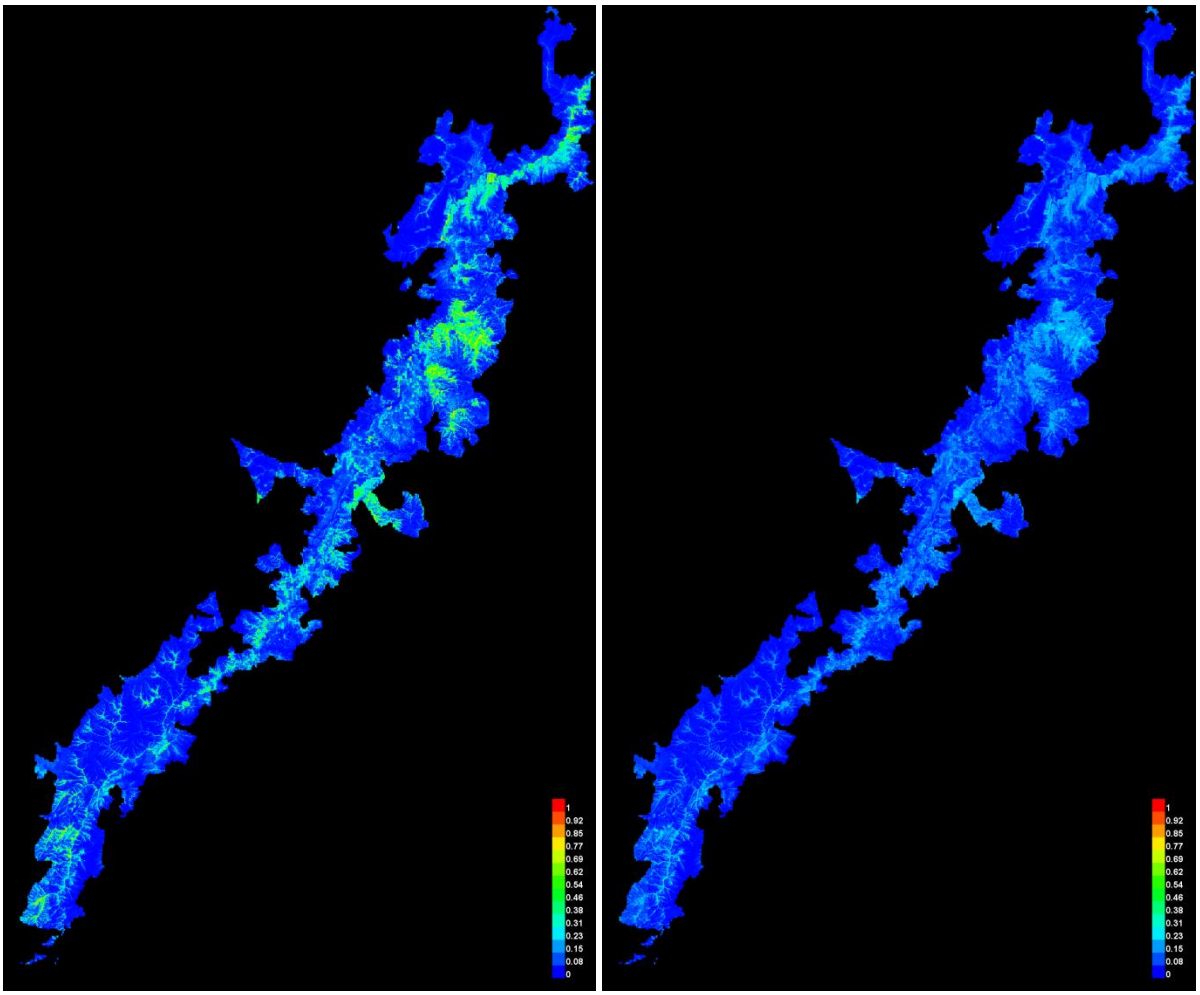


Pictures of the model

The following two pictures show the point-wise mean and standard deviation of the 10 output grids. Other available summary grids are min, max and median.

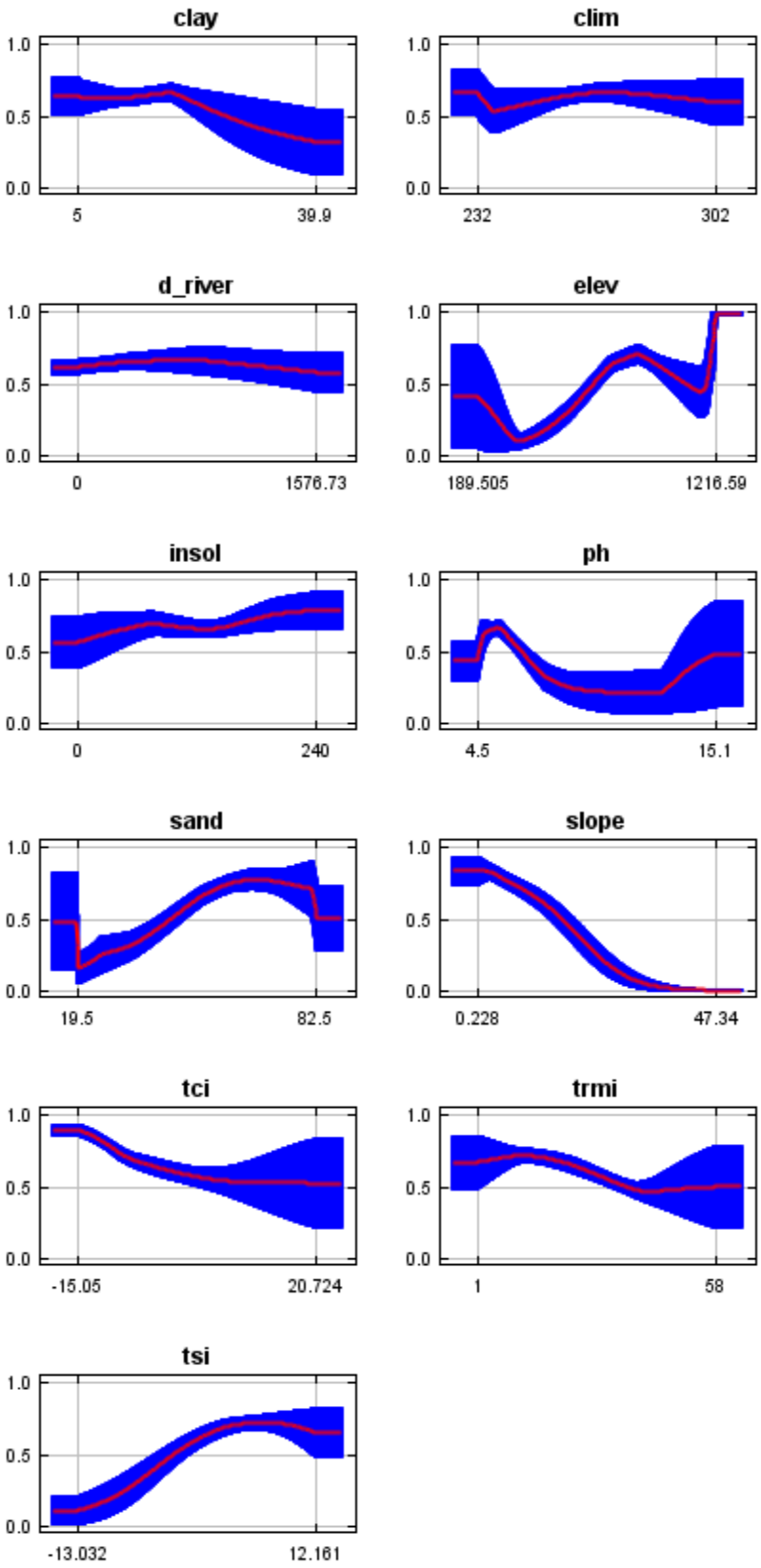


The following two pictures show the point-wise mean and standard deviation of the 10 models applied to the environmental layers in ASCIISNP_2070RCP86. Other available summary grids are min, max and median.

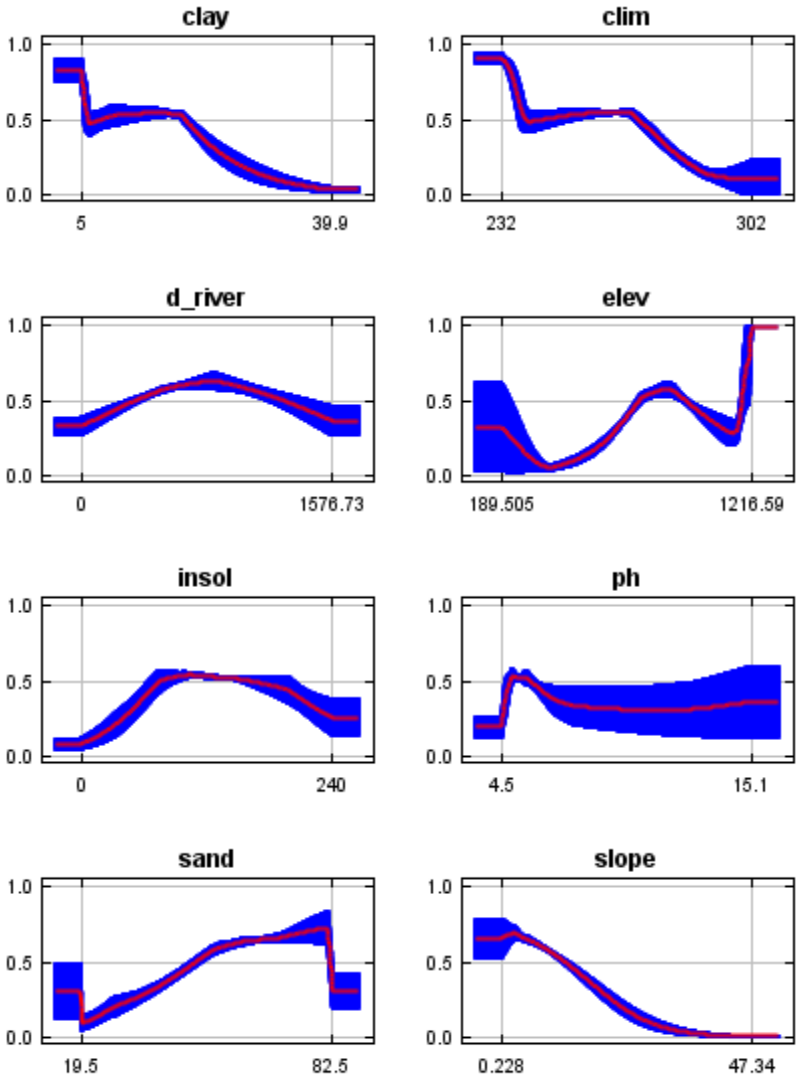


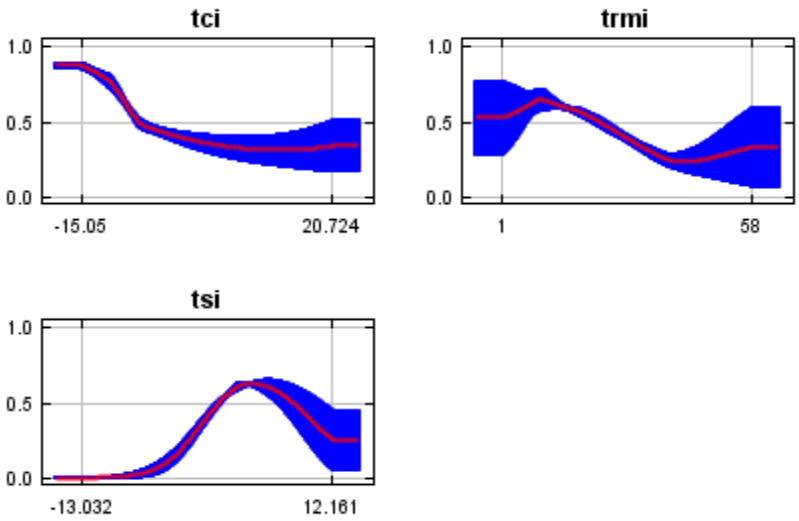
Response curves

These curves show how each environmental variable affects the Maxent prediction. The curves show how the logistic prediction changes as each environmental variable is varied, keeping all other environmental variables at their average sample value. Click on a response curve to see a larger version. Note that the curves can be hard to interpret if you have strongly correlated variables, as the model may depend on the correlations in ways that are not evident in the curves. In other words, the curves show the marginal effect of changing exactly one variable, whereas the model may take advantage of sets of variables changing together. The curves show the mean response of the 10 replicate Maxent runs (red) and and the mean +/- one standard deviation (blue, two shades for categorical variables).


clay
1.0
0.5
0.0
5
39.9
clim
1.0
0.5
0.0
232
302
d_river
1.0
0.5
0.0
0
1576.73
elev
1.0
0.5
0.0
189.505
1216.59
insol
1.0
0.5
0.0
0
240
ph
1.0
0.5
0.0
4.5
15.1
sand
1.0
0.5
0.0
19.5
82.5
slope
1.0
0.5
0.0
0.228
47.34
tci
1.0
0.5
0.0
-15.05
20.724
trmi
1.0
0.5
0.0
1
58
tsi
1.0
0.5
0.0
-13.032
12.161

In contrast to the above marginal response curves, each of the following curves represents a different model, namely, a Maxent model created using only the corresponding variable. These plots reflect the dependence of predicted suitability both on the selected variable and on dependencies induced by correlations between the selected variable and other variables. They may be easier to interpret if there are strong correlations between variables.





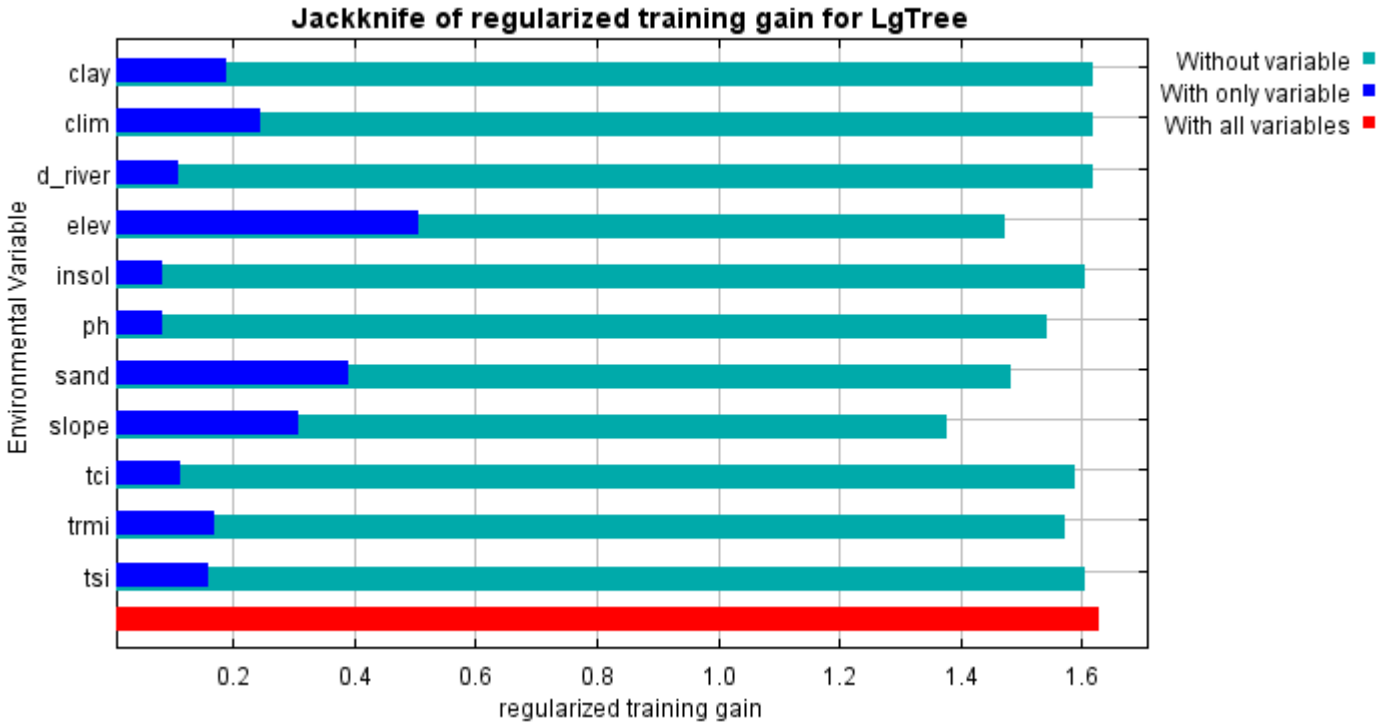
Analysis of variable contributions

The following table gives estimates of relative contributions of the environmental variables to the Maxent model. To determine the first estimate, in each iteration of the training algorithm, the increase in regularized gain is added to the contribution of the corresponding variable, or subtracted from it if the change to the absolute value of lambda is negative. For the second estimate, for each environmental variable in turn, the values of that variable on training presence and background data are randomly permuted. The model is reevaluated on the permuted data, and the resulting drop in training AUC is shown in the table, normalized to percentages. As with the variable jackknife, variable contributions should be interpreted with caution when the predictor variables are correlated. Values shown are averages over replicate runs.

Variable	Percent contribution	Permutation importance
elev	32.1	28
sand	26.4	30.9
slope	17.2	19.7
trmi	7.2	4.2
ph	5.3	8
tci	2.8	0.7

clay	2.3	2.6
d_river	2.2	0.3
tsi	1.7	1.1
insol	1.7	2.9
clim	1.1	1.6

The following picture shows the results of the jackknife test of variable importance. The environmental variable with highest gain when used in isolation is elev, which therefore appears to have the most useful information by itself. The environmental variable that decreases the gain the most when it is omitted is slope, which therefore appears to have the most information that isn't present in the other variables. Values shown are averages over replicate runs.



Command line to repeat this species model: java density.MaxEnt nowarnings noprefixes -E "" -E LgTree responsecurves jackknife outputdirectory=F:\MP_data\Data\MAXENT\outputs2070RCP86 projectionlayers=F:\MP_data\Data\MAXENT\ASCIISNP_2070RCP86 samplesfile=F:\MP_data\Data\MAXENT\LGTREE.csv environmentallayers=F:\MP_data\Data\MAXENT\ASCIISNP randomseed replicates=10 replicatetype=bootstrap nooutputgrids

Appendix D: Environmental Index Calculations