ABSTRACT

COMPARISON OF MAXENT AND BOOSTED REGRESSION TREE MODEL PERFORMANCE IN PREDICTING THE SPATIAL DISTRIBUTION OF THREATENED PLANT, TELEPHUS SPURGE (*EUPHORBIA TELEPHIOIDES*)

by Alexa Marie Mainella

A species distribution model (SDM) was developed to predict the presence and suitable habitat of the federally threatened plant, *Euphorbia telephioides*, in northwest Florida using data acquired from the U.S. Fish & Wildlife Service. I used two machine-learning models, MaxEnt and boosted regression trees (BRTs), as previous research has shown them to yield high predictability, especially with presence-only data and different types of predictor variables. Different methods were used to reduce effects of spatial autocorrelation and sampling bias in the model predictions since *E. telephioides* populations are strictly located along the coast. The 29 predictor variables were a combination of categorical, continuous, and distance-based variables. Both the MaxEnt and BRT models had high accuracy as measured by area under the curve (AUC), sensitivity, specificity, and true skill statistic (TSS), but the BRTs had a much lower deviance. The BRT models were also validated with the discovery of a new population in an area predicted as high probability of occurrence. This study demonstrates that machine-learning SDMs can be used by conservation organizations as cost-effective tools to find and protect new populations of threatened or endangered species.

COMPARISON OF MAXENT AND BOOSTED REGRESSION TREE MODEL PERFORMANCE IN PREDICTING THE SPATIAL DISTRIBUTION OF THREATENED PLANT, TELEPHUS SPURGE (*EUPHORBIA TELEPHIOIDES*)

A Thesis

Submitted to the

Faculty of Miami University

in partial fulfillment of

the requirements for the degree of

Master of Environmental Science
Institute for the Environment and Sustainability

by

Alexa Marie Mainella

Miami University

Oxford, Ohio

2016

Advisor: Dr. Mary Henry

Reader: Dr. David Gorchov

Reader: Dr. Thomas Crist

This thesis titled

COMPARISON OF MAXENT AND BOOSTED REGRESSION TREE MODEL PERFORMANCE IN PREDICTING THE SPATIAL DISTRIBUTION OF THREATENED PLANT, TELEPHUS SPURGE (*EUPHORBIA TELEPHIOIDES*)

by

Alexa Marie Mainella

has been approved for publication by

The College of Arts and Science

and

Institute for the Environment and Sustainability

_____

Dr. Mary Henry

_____

Dr. David Gorchov

_____

Dr. Thomas Crist

Table of Contents

List of Tables v
List of Figures vi
Acknowledgements viii
INTRODUCTION: 1
Species Distribution Models (SDMs): 1
Euphorbia telephioides biology & site characteristics 2
METHODS: 3
Study Area 3
Standardizing presence data 4
Generating pseudo-absences for boosted regression tree models 6
Predictor variables 9
Vector data 9
Landsat data 9
Topography 10
Distance-based predictor variables 10
Data Standardization for BRT models 11
Species Distribution Modeling 11
Model Fitting 12
Model Prediction 13
RESULTS: 13
Presence-only MaxEnt models 13
Presence/Pseudo-absence BRT models 18
Comparison of MaxEnt and BRT models 26
Soil types associated with E. telephioides occurrence 27
DISCUSSION: 28
Model Accuracy and Comparison of Machine-learning SDMs 28
Future Research 29
Conclusions 30
Literature Cited: 32
APPENDIX A – LIST OF LAND-USE/LAND COVER TYPES IN STUDY AREA. 37
APPENDIX B – LIST OF SOIL TYPES IN STUDY AREA. 41
APPENDIX C – MAPS OF PREDICTOR VARIABLES. 46

APPENDIX D – LIST OF PREDICTOR VARIABLES..... 55
APPENDIX E – R CODE:..... 57

List of Tables

Table 1. Summary of boosted regression tree (BRT) models..... 6

Table 2. TCT coefficients for Landsat 8 reflectance (Baig et al. 2014). 10

Table 3. Comparison of MaxEnt model evaluation statistics measuring predictive accuracy. 14

Table 4. Percent contribution of top 12 variables in the MaxEnt (1 & 2) models. 17

Table 5. Comparison of BRT model evaluation statistics measuring predictive accuracy 19

Table 6. Percent contribution of top 12 variables in the BRT (Out 1 & 2) models. 19

Table 7. Percent contribution of top 12 variables in the BRT (In 1 & 2) models..... 20

Table 8. Comparing MaxEnt and BRT models using correlation coefficients and AUC. 27

Table 9. List of most important soil types occupied by *E. telephioides*. 27

Table 10. Predicted probabilities for new population at Tyndall Air Force Base. 31

List of Figures

Figure 1. Map of Bay, Gulf, and Franklin counties in Florida where Telephus spurge has historically been found, and the limited study area boundary. 3

Figure 2. Graphic depicting how the presence point standardization was completed for the a) 70.72 m x 70.72 m, and b) 707.11 m x 707.11 m resolution grids; blue points/polygons represent the raw presence data, and red centroids represent the standardized presences. 5

Figure 3. a) Presences (707m) & pseudo-absences (entire study area) for BRT Out 1 model and b) Presences (70m) & pseudo-absences (entire study area) for BRT Out 2 model 7

Figure 4. a) Presences (707m) & pseudo-absences (within 10km buffer) for BRT In 1 model; and b) Presences (70m) & pseudo-absences (within 10km buffer) for BRT In 2 model 8

Figure 5. Receiving Operator Characteristic (ROC) curve for MaxEnt 1 (82 presences) model 14

Figure 6. Receiving Operator Characteristic (ROC) curve for MaxEnt 1 (534 presences) model 15

Figure 7. MaxEnt 1 (82 presences) probability of occurrence map. Circled areas show where the model predicted low probability of occurrence where there are presence points. 16

Figure 8. MaxEnt 2 (534 presences) probability of occurrence map. 17

Figure 9. Partial dependence plots for BRT (Out 1, 82 presences) model showing the top 12 contributing variables 21

Figure 10. Partial dependence plots for BRT (Out 2, 534 presences) model showing the top 12 contributing variables. 21

Figure 11. Partial dependence plots for BRT (In 1, 82 presences) model showing the top 12 contributing variables 22

Figure 12. Partial dependence plots for BRT (In 2, 534 presences) model showing the top 12 contributing variables 22

Figure 13. BRT Out 1 (82 presences) probability of occurrence map. 23

Figure 14. BRT Out 2 (534 presences) probability of occurrence map. 24

Figure 15. BRT IN 1 (82 presences) probability of occurrence map 25

Figure 16. BRT In 2 (534 presences) probability of occurrence map. 26

Figure 17. Map depicting the new population of *E. telephioides* at Tyndall Air Force Base in Panama City, FL. 31

Figure 18. A map displaying the rock types in the study area. 46

Figure 19. Average monthly precipitation (mm) in a) February, b) May, and c) August (worldclim.org). 47

Figure 20. Average monthly temperature (ºCelsius) in a) February, b) May, and c) August (worldclim.org). 48

Figure 21. Maps depicting NDVI in a) February, b) May, and c) August 2014..... 49

Figure 22. Maps depicting EVI in a) February, b) May, and c) August 2014. 50

Figure 23. Maps depicting tasselled cap brightness in a) February, b) May, and c) August 2014..... 51

Figure 24. Maps depicting tasselled cap greenness in a) February, b) May, and c) August 2014..... 52

Figure 25. Maps depicting tasselled cap wetness in a) February, b) May, and c) August 2014..... 53

Figure 26. Digital elevation model for the study area. 54

Figure 27. Map depicting percent slope in study area. 54

Acknowledgements

I would like to thank my adviser, Dr. Mary Henry, for her guidance and time over the past two years leading up to the completion of my thesis. I would also like to thank my committee members, Drs. David Gorchov and Thomas Crist, for their feedback regarding my research and constructive comments on my manuscript. I would like to extend my gratitude to GIS specialist, Robbyn Abbitt, and statistician, Dr. Jing Zhang, for their help with the methodology and technical aspects of the research. I also sincerely thank Dr. Vivian Negrón-Ortiz and the U.S. Fish and Wildlife Service for the opportunity to conduct this research and help protect an important plant species. Lastly, I would like to thank those in the Institute for the Environment and Sustainability at Miami University for the opportunity to pursue my master's degree in a field that I am passionate about.

INTRODUCTION:

The geographical distribution of a species is essential for conservation planning, land management, and ecological understanding (Elith et al. 2006; Williams et al. 2009). Rare species in particular can be difficult to model spatially due to their narrow ranges and specialized habitat requirements (Williams et al. 2009). Species with small sample sizes and small distributions will inherently contain sampling bias that can severely impact model quality (Phillips et al. 2009; Williams et al. 2009). Another issue with modeling species that have narrow ranges is that the occurrence data usually exhibit spatial autocorrelation (Crase et al. 2012). Spatial autocorrelation is a pattern where occurrence points are related to each other based on geographic distance; therefore, locations closer together are more similar than those further apart (Crase et al. 2012). The presence of spatial autocorrelation in a model violates the assumption that all observations are independent, which can lead to incorrectly identifying significant predictor variables, poorly estimating regression coefficients, and overpredicting the species' distribution (Veloz 2009; Crase et al. 2012). Overpredicting a species' distribution makes it difficult to include new data to the model and results in inflated model accuracy. It is therefore essential to consider methods that reduce spatial autocorrelation and sampling bias when using species distribution models (SDMs) to determine areas of suitable habitat and search for new populations.

In this study, I compared the effectiveness of machine-learning species distribution models to predict the geographic range of the threatened plant species, Telephus spurge (*Euphorbia telephioides*). The two models used were MaxEnt and boosted regression trees (BRTs) as previous research has shown them to yield high predictability, especially with presence-only data and different types of predictor variables (Elith et al. 2008; Elith et al. 2011). I chose to test different methods that would address sampling bias and spatial autocorrelation in both models with the main objective of developing an SDM that provided the most accurate prediction of the actual *E. telephioides* distribution.

Species Distribution Models (SDMs):

Species distribution models (SDMs) are cost-effective tools used to predict the distribution of species across a landscape based on their response to environmental factors (Elith and Leathwick 2009; Parviainen et al. 2013). SDMs have been especially helpful for species conservation, management, and recovery, as well as identifying areas containing high biological diversity to be protected (Zaniewski et al. 2009). These models include a combination of species occurrence information with measured environmental variables, such as topo-climatic data and biotic predictors (Elith and Leathwick 2009). SDMs require reliable presence/absence data for the species as well as relevant predictor variables (Elith and Leathwick 2009). There are two categories of SDMs: 1) models that use presence-only data, and 2) models that use presence/absence data (Elith and Leathwick 2009; Zaniewski et al. 2002) If absence data are not available, then computer generated pseudo-absences can be used based on the available presence data (Zaniewski et al. 2009). The best-fit model with the right set of predictors can be determined using a multitude of statistical analyses, including regression-based and machine-learning models (Elith and Leathwick 2009). The statistical analysis chosen depends on the purpose of the

SDM (i.e. predicting current distribution of a species, extrapolating future distribution based on climate change, or identifying biodiversity hotspots) and the type of data available (i.e. presence, absence, and/or pseudo-absence) (Elith and Leathwick 2009). Application and further study of SDMs would be highly beneficial to rare species that face negative effects of a changing landscape and climate so that conservation planning can be implemented now to prevent further population decline.

Euphorbia telephioides biology & site characteristics

Euphorbia telephioides is a perennial plant endemic to the Florida panhandle, specifically to Bay, Gulf, and Franklin counties (Trapnell et al. 2012; Figure 1). The species is restricted to scrubby pine flatwoods with sandy soils within seven kilometers of the Gulf of Mexico (Bridges and Orzell 2002). *E. telephioides* was listed as a threatened species in 1992 and there are currently 41 recorded populations (U.S. Fish and Wildlife Service 2014). *E. telephioides* is an ephemeral plant, meaning that it can remain in a dormant state and still persist because of its large, tuberous root (Trapnell et al. 2012). This suggests a strategy for individuals to persist even in stressful conditions. Some populations recorded years ago have disappeared, but they have the potential to reappear if a fire or mowing disturbance occurs (Negron-Ortiz 2014 pers. obs.).

E. telephioides is subdioecious, composed of male, female, and monoecious plants (Trapnell et al. 2012). Data suggest that the monoecious plants can change sex after a fire disturbance (Negrón-Ortiz 2014 pers. obs.). The plant grows to a maximum 25 cm tall and can be easily shaded out by faster-growing plants, such as palmetto and titi (Bridges and Orzell 2002). Therefore, *E. telephioides* needs frequent disturbance from fire or mowing (historically every 2-3 years) in order to successfully reproduce (Trapnell et al. 2012; Negrón-Ortiz 2014 pers. obs.).

E. telephioides is considered a "spotlight" species because it has high potential to be delisted since the only known, immediate threat is human development (U.S. Fish and Wildlife Service 2014). However, the long-term threat that will potentially harm this species over the next 100 years is climate change and sea level rise (U.S. Fish and Wildlife Service 2014). The IPCC (2013) predicts sea level rise (SLR) to be 0.26 - 0.82 m by year 2100, which will likely cover most of Florida's land mass less than 1m in elevation (Noss 2011). Satellite data shows that average SLR in the Gulf of Mexico, however, is increasing faster than the global average (Bilskie et al. 2014). Since *E. telephioides* grows along the Gulf of Mexico coast in only three Florida counties, SLR will most likely affect the species. The populations of *E. telephioides* in Bay and Franklin counties are approximately 3 m above sea level, and roughly 4 km and 3 km from the coast, respectively. The Gulf County population, however, is only about 2 m above sea level as it is closer to the coast (about 2 km). Telephus spurge is at risk of further decline from SLR because its seeds are not readily dispersed over large distances (Negrón-Ortiz, 2014, pers. obs.), so sea level may rise more quickly than the species can disperse its seeds and establish populations further inland. In addition, *E. telephioides* does not respond well to transplantation (Ecological Resource Consultants 2006; Negrón-Ortiz 2014, pers. obs.). Another major concern is that as more coastline is inundated with water, urban development will expand, decreasing the amount of suitable habitat for *E. telephioides* and impeding the ability of this species to move landward.

METHODS:

Study Area

The study area comprised three counties in the Florida panhandle: Bay, Gulf, and Franklin (Figure 1). In order to reduce the effect of sampling bias, the study area was decreased so that most of the area was about 30 km or less from the Gulf Coast (Young et al. 2011). This distance was chosen so that the new study area encompassed *E. telephioides'* historical range while also including new areas that may contain suitable habitat. The dominant land-use type in the area is coniferous plantations that consist of slash pine (*Pinus elliottii*), which has largely replaced the native longleaf pine (*Pinus palustris*) and wiregrass ecosystem. While *E. telephioides* has been found in coniferous plantations, the species is mainly associated with relatively flat, scrubby pine flatwoods with moderately to poorly-drained sandy soils (Bridges and Orzell 2002; Florida Natural Areas Inventory and Florida Department of Natural Resources 1990). Flatwoods are characterized by an open canopy with widely-spaced pine trees and little to no understory (Florida Natural Areas Inventory and Florida Department of Natural Resources 1990). The densely vegetated groundcover contains a variety of herbs and shrubs, including saw palmetto, swamp titi, and runner oak.

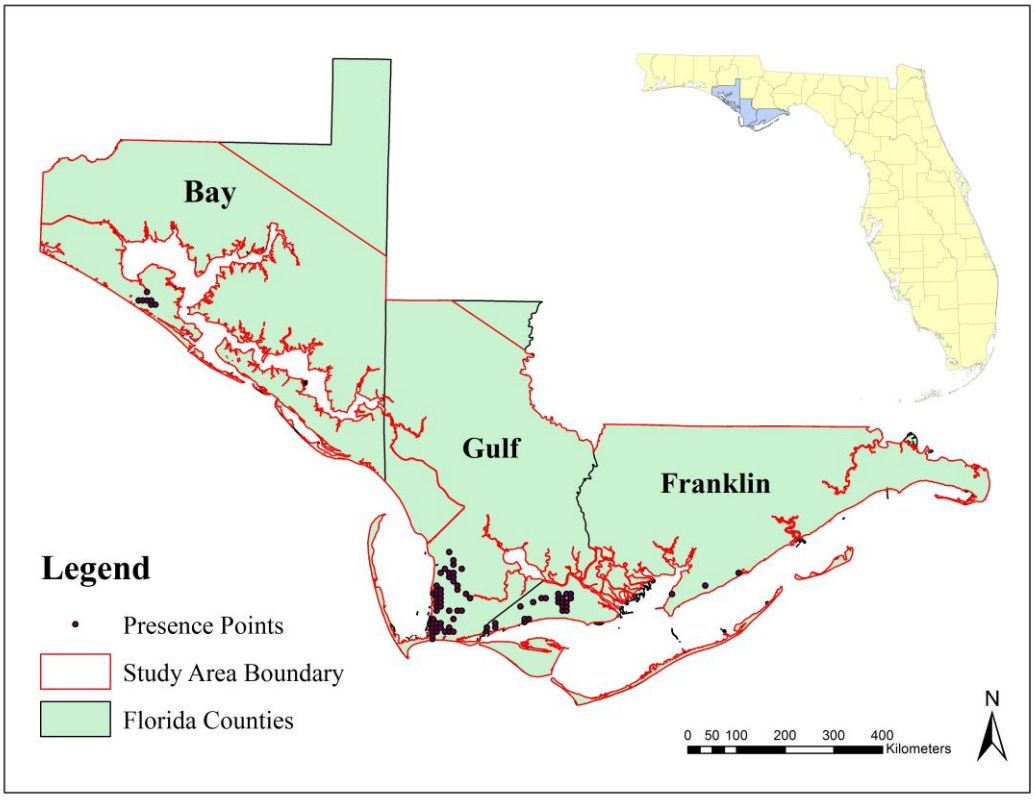


Figure 1. Map of Bay, Gulf, and Franklin counties in Florida where Telephus spurge has historically been found, and the limited study area boundary.

Standardizing presence data

Presence data of *E. telephioides* were gathered by the U.S. FWS and Florida Natural Areas Inventory (FNAI). The data were collected using GPS units as either points (indicating either individual plants or a multitude of plants in the surrounding area) or polygons (used for locations with an abundance of individuals). A method known as spatial filtering was used on the raw presence data to reduce effects of spatial autocorrelation and sampling bias while also converting all presence data to points, thus creating a standardized dataset (Williams et al. 2009; Boria et al. 2014). To standardize the raw presence data, two grids with different resolutions (70.72 m x 70.72 m cells and 707.11 m x 707.11 m cells) were overlaid on the study area and a centroid was added to every grid cell that contained raw presence data (Figures 2a and 2b; Williams et al. 2009). The grid resolutions were calculated so that the centroids would never be more than 50 m (Figure 2a) or 500 m (Figure 2b) away from the raw presence points or polygons. Also, the 50-m and 500-m distances were chosen to compare the effect of spatial autocorrelation at a small scale (50 m) and broad scale (500 m). As a result, the 70.72 m x 70.72 m grid generated 534 standardized presence points, and the 707.11 m x 707.11 m grid generated 82 standardized presences.

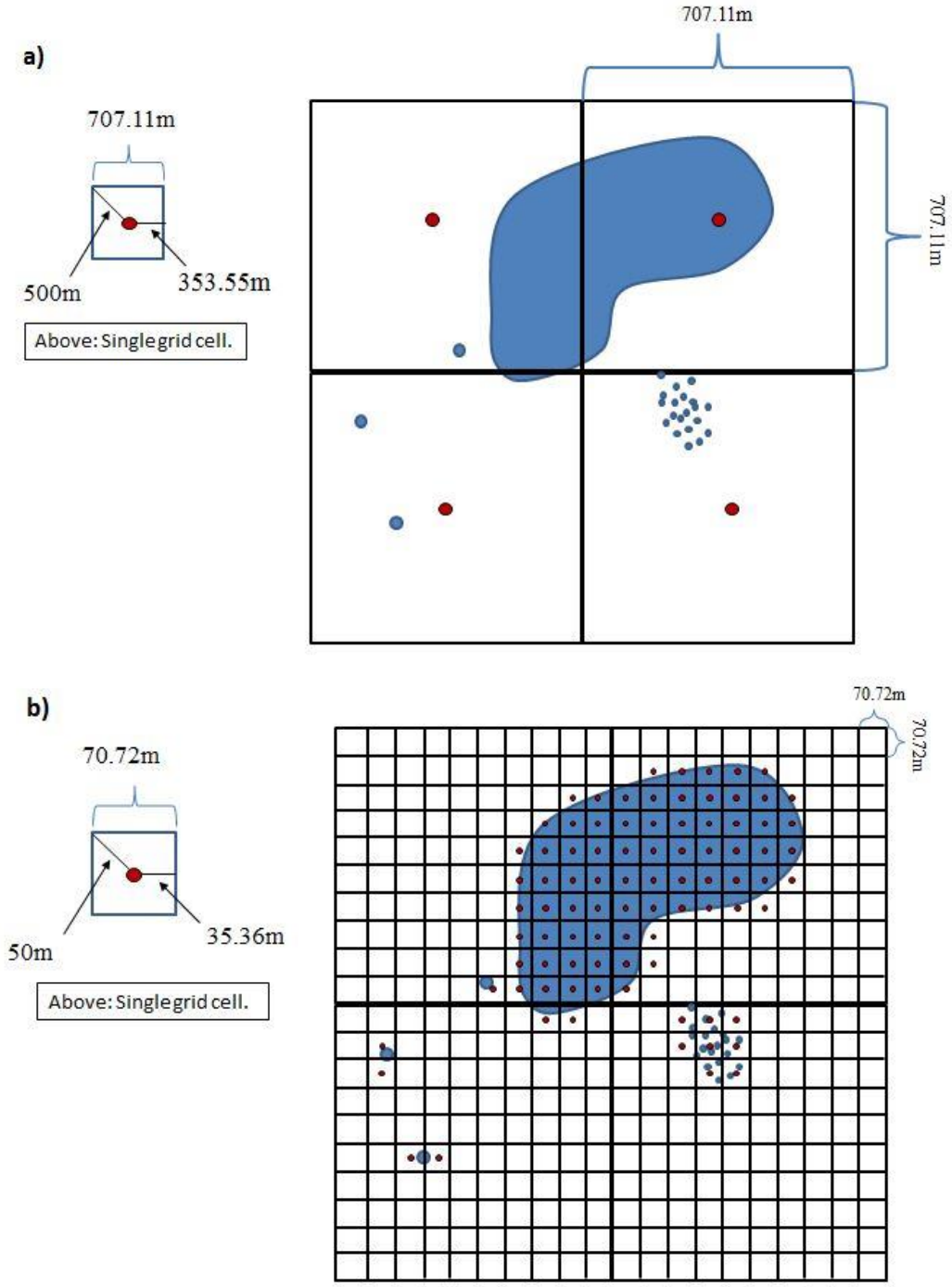


Figure 2. Graphic depicting how the presence point standardization was completed for the a) 70.72 m x 70.72 m, and b) 707.11 m x 707.11 m resolution grids; blue points/polygons represent the raw presence data, and red centroids represent the standardized presences.

Generating pseudo-absences for boosted regression tree models

Pseudo-absences are artificial absences used when true absences are unavailable (Barbet-Massin et al. 2012). Some modelers think that pseudo-absences imply actual absences; however, the purpose of pseudo-absences is to provide the model with background information about the study area (Phillips et al. 2009). The program, MaxEnt, automatically generated background points (i.e. pseudo-absences), so random pseudo-absences were created in ArcMap 10.3 for four boosted regression tree models (Table 1).

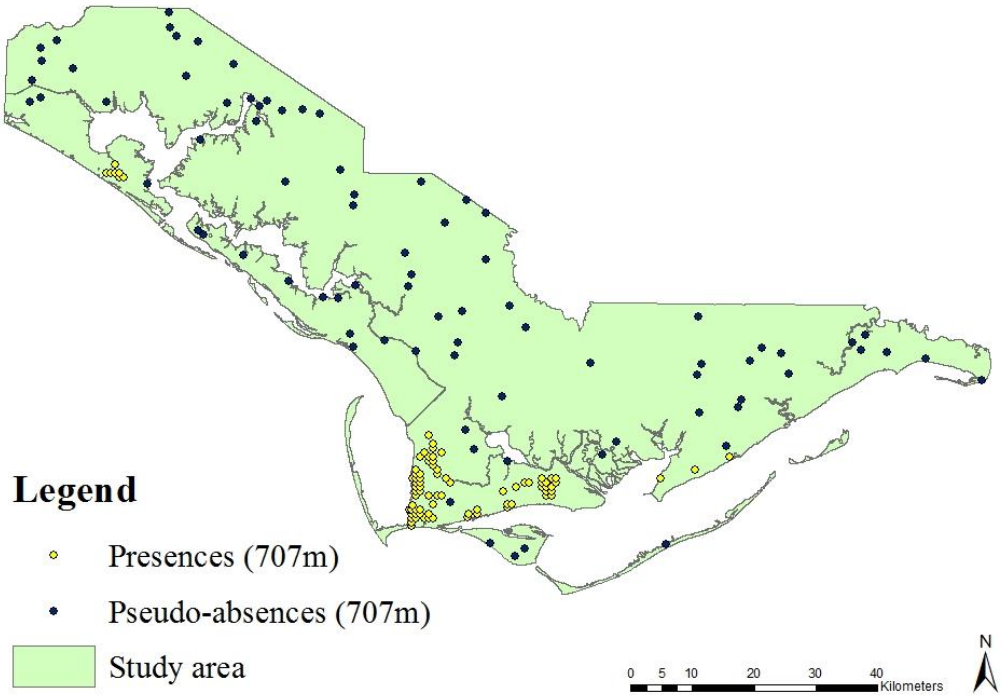
Pseudo-absences were generated for the boosted regression tree (BRT) models because an insufficient number of true absences had been collected in the field. In addition, true absences can introduce unconfirmed assumptions and produce less accurate models compared to models incorporating pseudo-absences (Zaniewski et al. 2002). As a result, true absences may not be reliable for *E. telephioides* since the plant can remain dormant for long periods of time, or because a fire-related disturbance temporarily removed its shoot and leaves (Zaniewski et al. 2002). Thus, the plant may still be present in an area but is not readily visible and any false absences collected can negatively impact model quality (Barbet-Massin et al. 2012). One set of pseudo-absences was used in the BRT models using an equal number of pseudo-absences and presence points. The pseudo-absences and presences were also equally weighted and pseudo-absences were randomly distributed throughout the study area (Table 1; Figures 3a and 3b).

An alternative approach to generating pseudo-absences was also used in order to reduce the effect of sampling bias. The method was based on recommendations from Phillips et al. (2009), and required making pseudo-absences with the same sampling bias as the presence points. This was accomplished by creating a 10-km buffer around the presence points and then generating another set of pseudo-absences within that buffer (Table 1; Figures 4a and 4b). The buffer distance was chosen to encapsulate the entire historical range of *E. telephioides*.

Table 1. Summary of boosted regression tree (BRT) models. Model names reflect how pseudo-absences (PAs) were generated as either within a 10-km buffer around the presence points (BRT In models), or throughout the entire study area (BRT Out models).

BRT model name	Number of presences per model	Number of PAs per model	PAs generated within 10-km buffer or entire study area
BRT Out 1	82	82	Entire study area
BRT Out 2	534	534	Entire study area
BRT In 1	82	82	Within 10-km buffer
BRT In 2	534	534	Within 10-km buffer

a)



b)

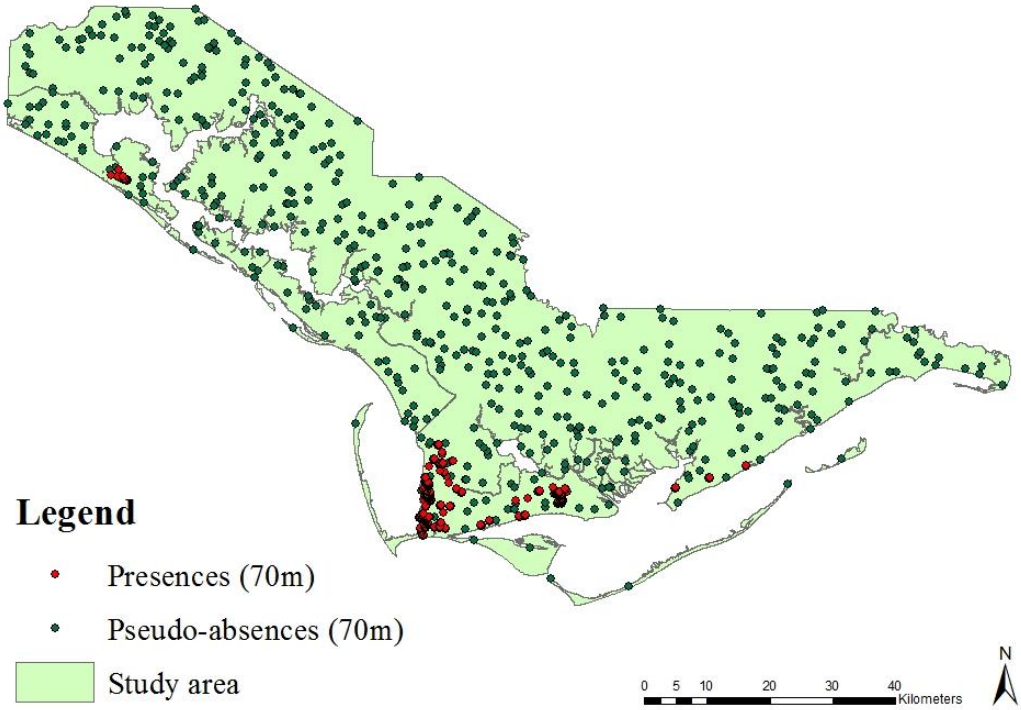
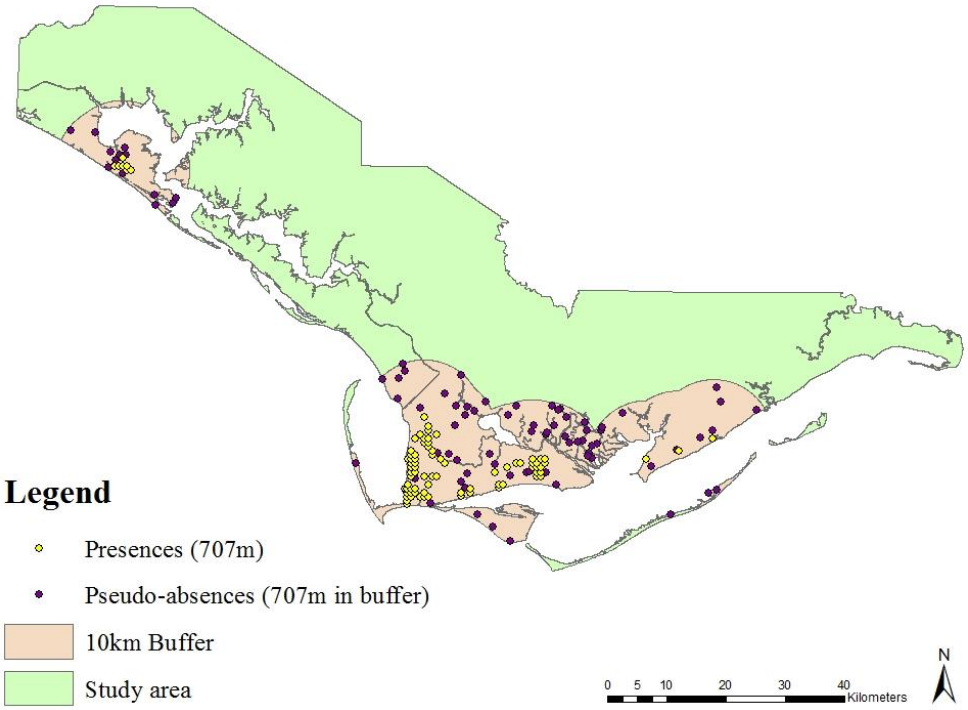


Figure 3. a) Presences (707m) & pseudo-absences (entire study area) for BRT Out 1 model; and b) Presences (70m) & pseudo-absences (entire study area) for BRT Out 2 model.

a)



b)

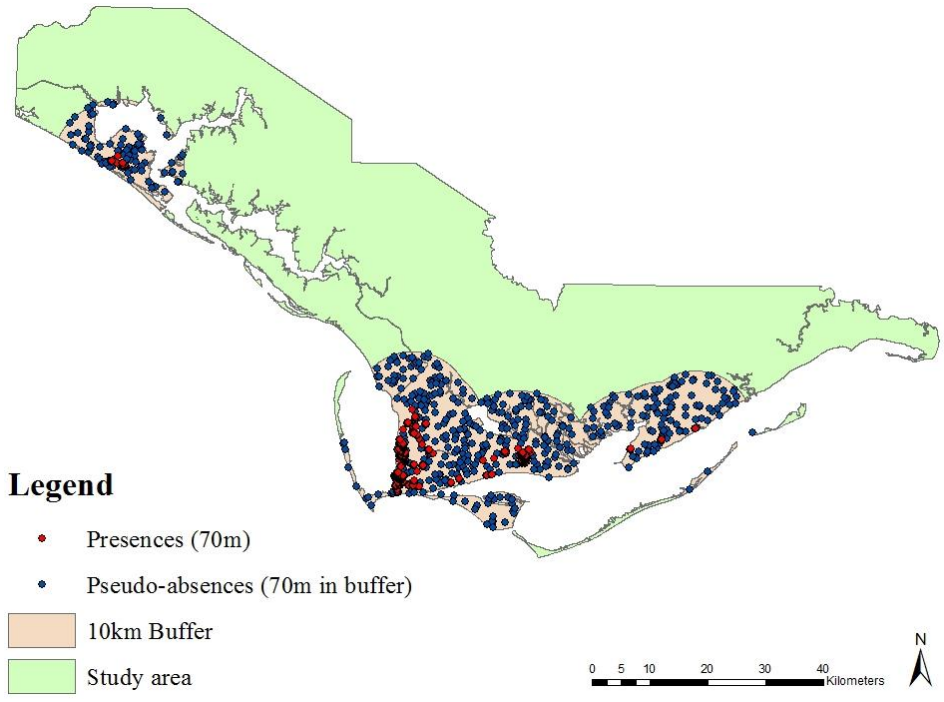


Figure 4. a) Presences (707m) & pseudo-absences (within 10km buffer) for BRT In 1 model; and b) Presences (70m) & pseudo-absences (within 10km buffer) for BRT In 2 model.

Predictor variables

Vector data

Literature on *E. telephioides* site characteristics have stated the importance of land-use/land cover (LULC) and soil type (Bridges and Orzell 2002; Trapnell et al. 2012; U.S. Fish and Wildlife Service 2014). The LULC layer used in the MaxEnt and BRT models is titled the Cooperative Land Cover Map version 3.0 and was created by the Florida Natural Areas Inventory (FNAI) in 2014 using ground-truthed local data sources and high resolution aerial photography (FNAI 2014). This LULC map is updated every 6-12 months, so the latest version should be used if new *E. telephioides* data are added to the models. See Appendix A for a full list of LULC types in the study area. The soils data were downloaded in vector format from the Soil Survey Geographic Database (SSURGO, version 2014) as provided by the Natural Resources Conservation Service (NRCS). The SSURGO database is updated every year on October 1, and modelers should also use the latest version as new presence data are collected (NRCS 2013). See Appendix B for a full list of soil types found in the study area.

Surficial geology was included as a predictor variable to determine if bedrock type influenced *E. telephioides* presence, despite the relative homogeneity in the area (Appendix C). The surficial geology layer was downloaded as vector data from the Florida Department of Environmental Protection (Florida Geological Survey, 2001).

While climatic variables tend to remain relatively homogenous on a small scale, it was important to include precipitation and temperature in this study since coastal conditions in Florida can be temporally variable. Therefore, 1 x 1-km resolution tiles were downloaded from worldclim.org to add climatic data to the model. The average monthly precipitation and temperatures for February, May, and August (Appendix C) were used to determine if seasonal variation in these variables influenced *E. telephioides* distribution. See Hijmans et al. (2005) for a complete description of how the climate data were derived.

Landsat data

Spectral vegetation indices (SVIs) are based on brightness values from satellite images and attempt to measure vegetation biomass (Campbell and Wynne 2011). SVIs are derived from different combinations of spectral bands where the output raster grid indicates the amount of vegetation in each pixel (Campbell and Wynne 2011). Pixels with high SVI values represent areas with more green vegetation than pixels with lower values. Chlorophyll in green vegetation absorbs red light (R) and reflects near infrared (NIR) radiation, thus the NIR/R ratio (also known as the simple ratio) provides an estimate of photosynthetic activity within each pixel (Campbell and Wynne 2011). Consequently, SVIs can discern between different types of vegetation, such as grass, deciduous forest, and coniferous forest.

One of the most widely used SVIs is the normalized difference vegetation index (NDVI), which uses a ratio of near infrared (NIR) and red (R) bands (NDVI = (NIR-R) / (NIR+R)). While NDVI is common to use, it may approach saturation (maximum) before the biomass in a pixel reaches its maximum (Campbell and Wynne 2011). In other words, NDVI should be used with caution

when determining biomass in dense vegetation. Therefore, researchers have proposed using modified indices, such as the enhanced vegetation index (EVI) and tasseled cap transformations (TCTs). EVI was developed by including the blue band to improve sensitivity in dense vegetation by reducing atmospheric effects and decoupling the canopy background signal (Obata et al. 2016). TCTs are categorized as orthogonal SVIs since they combine spectral bands linearly instead of as ratios (Baig et al. 2014). Therefore, the TCTs in this study were calculated using coefficients specific to the operational land imager (OLI) sensor onboard the Landsat 8 satellite to represent vegetation brightness, greenness, or wetness.

Landsat 8 OLI satellite imagery (30-m resolution) was downloaded from United States Geological Service (USGS) EarthExplorer as two images to cover the entire study area. In order to determine if temporal variation in biomass affects *E. telephioides* distribution, images from three different months were used: February 2014 (dormant season), May 2014 (start of growing season), and August 2014 (end of growing season). The August images contained clouds, so pixels underneath and in the immediate vicinity of clouds may not represent true reflectance values. All images were calibrated to at-sensor reflectance in ENVI 5.0 software before each set of two images were mosaicked to form one image. The mosaicked images were then used to calculate the SVIs in ENVI Classic software. NDVI images were created using in-built tools, while EVI and TCT (brightness, greenness, and wetness) images were calculated using manual equations (Appendix C). The coefficients used to calculate EVI were those adopted in the MODIS-EVI algorithm (L=1, C1=6, C2=7.5, and G =2.5) (Appendix C). The coefficients used to generate TCT brightness, greenness, and wetness (Appendix C) images were predetermined by Baig et al. (2014) for Landsat 8 imagery (Table 2).

Table 2. TCT coefficients for Landsat 8 reflectance (Baig et al. 2014).

TCT	(Blue) Band 2	(Green) Band 3	(Red) Band 4	(NIR) Band 5	(SWIR 1) Band 6	(SWIR 2) Band 7
Brightness	0.3029	0.2786	0.4733	0.5599	0.508	0.1872
Greenness	-0.2941	-0.243	-0.5424	0.7276	0.0713	-0.1608
Wetness	0.1511	0.1973	0.3283	0.3407	-0.7117	-0.4559

Topography

A 10-m resolution digital elevation model (DEM) was downloaded from the National Elevation Dataset (NED) provided by USGS. The raster image was resampled in ArcMap 10.2 using bilinear interpolation and thus served as the elevation predictor variable (Appendix C). The DEM was also used to derive a slope predictor variable by using the associated tool in ArcMap 10.2 (Appendix C). Slope is determined as the maximum change in elevation between the original cell in the 10-m DEM and its eight neighboring cells (ESRI 2007).

Distance-based predictor variables

Three distance-based predictor variables used in the BRT models included: distance to roads, distance to wetlands, and distance to the ocean. These predictors were calculated in ArcMap 10.2 using the “Near” tool, which finds the Euclidean distance between each presence and pseudo-

absence point and the nearest feature. These variables were considered in the BRT models based on personal observations of where *E. telephioides* was in relation to these features. The distance-based variables could not be represented as separate data layers because the "Near" tool adds the distances to new columns in existing point shapefiles; therefore, the distance-based variables could not be converted to raster grids and incorporated in the MaxEnt models.

Data Standardization for BRT models

The predictor variable data were extracted for the presence and pseudo-absence points in ArcMap 10.2 and exported to Excel files. Many of the variables with continuous data had significantly different ranges; for instance, the "Distance to Ocean" variable with a range of 5-77,632 would have outweighed the NDVI variables that range from -1 to 1. Therefore, before fitting the BRT models, the data were standardized in Excel using z-score scaling which uses the following formula: (Variable value-Mean of variable) / Standard deviation of variable. As a result, all standardized variables had a mean of 0.0, a variance of 1.0, but different ranges (Milligan and Cooper 1988). Another advantage of z-score scaling is that multicollinearity is minimized with data in standardized form (Marquardt 1980).

Species Distribution Modeling

The first machine-learning method tested was MaxEnt (maximum entropy), which is a program specifically designed to model species' distributions using presence-only data. MaxEnt's predictive performance rivals that of the highest performing models and it has been adopted by government and non-government agencies for real-world mapping applications (Elith et al. 2011). MaxEnt finds the probability distribution of maximum entropy (i.e. closest to uniform) subject to the constraint that the expected values of the environmental predictors match their respective averages (Elith et al. 2006; Phillips et al. 2006). MaxEnt is advantageous because it accommodates both categorical and continuous data, and it handles interactions between predictor variables (Phillips et al. 2006). However, with MaxEnt's simplicity comes disadvantages; the software provides few methods to evaluate model accuracy in the output. In addition, it is easy to over-fit the model in MaxEnt, so care must be taken when choosing a regularization term that prevents the model from becoming too complex (Elith et al. 2011; Shcheglovitova and Anderson 2013). MaxEnt is a beneficial tool because of its predictive power and easy-to-use interface, but not all defaults should be used when fitting a model in MaxEnt.

The second machine-learning method used in this study was boosted regression trees (BRTs), also known as stochastic gradient boosting, which is a presence-absence model that uses two algorithms (boosting algorithm and regression-tree algorithm) in a forward stage-wise fashion in order to make small modifications at each step and produce a better model (Elith et al. 2006). Boosting is advantageous because it incorporates a stochastic component in the model by using a random subset of the data to fit each tree, thereby reducing the overall variance (Elith et al. 2008). Each tree's contribution to the model is controlled by a specified learning rate (lr, also known the shrinkage parameter) while the tree complexity (tc, number of nodes in a tree) term determines the level of interactions between predictor variables that are fitted in the model (Elith et al. 2008). For instance, a tree complexity of 2 fits a model with up to two-way interactions. These two important terms (lr and tc) determine the number of trees (nt) required for optimal prediction (Elith et al. 2008). Finally, the bagging fraction controls stochasticity in the model and

determines what percentage of training data drawn at random are used to fit each tree (Elith et al. 2008).

While some modelers criticize BRTs for their complexity, many agree that BRTs are an excellent tool for modeling complicated data (Elith et al. 2006; De'Ath 2007; Elith et al. 2008; Phillips et al. 2009). BRTs can accommodate many types of response variables (numeric, categorical, censored), predictor variables (continuous, categorical) and loss functions (binomial, Gaussian, Poisson) (De'Ath 2007). In addition, BRTs can accommodate missing data and outliers, and automatically handles interaction effects between predictor variables (Elith et al. 2008). Perhaps the greatest advantage of BRT models is that over-learning (or overfitting) is greatly reduced due to 10-fold cross-validation that withholds portions of data and halts model fitting based on predictive accuracy (Phillips et al. 2009).

Model Fitting

The MaxEnt models were fitted in MaxEnt software (version 3.3.3k). All presence data (n = 82 or n = 534) were included for 10-fold cross-validation, which is advantageous for small datasets. Predictor variables with a variance inflation factor (VIF) greater than 10 were not used for model fitting to avoid multicollinearity effects. The distance-based variables were also excluded since they could not be represented as separate data layers (see "**Distance-based predictor variables**" section for explanation). Predictive performance of the default settings was compared to various models where I specified different settings. The number of replicates was increased from 1 to 10, and the number of iterations was increased from 500 to 1000 to ensure the model converged. The regularization parameter was adjusted multiple times to determine the optimum value that prevented overfitting with each set of presence points (Radosavljevic and Anderson 2014).

The BRT models were fitted in R statistical software (version 3.2.3) using the generalized boosted model (gbm) package (Ridgeway 2007) and BRT functions from Elith et al. (2008). Settings that can be changed for model fitting include loss function, number of trees, learning rate, tree complexity, bagging fraction, and number of cross-validation folds (Ridgeway 2007; Elith et al. 2008). In this study, the BRT settings were varied until an optimum model was fit with more than 1000 trees (Elith et al. 2008). Predictor variables with a variance inflation factor (VIF) greater than 10 were also not used for model fitting to avoid multicollinearity effects.

Cross-validated area under the curve (AUC), deviance, sensitivity, specificity, and the True Skill Statistic (TSS) were used to evaluate both MaxEnt and BRT model accuracy. AUC measures the quality of ranking of each site; in other words, AUC is the probability that a randomly chosen presence point will be ranked above a randomly chosen absence point (Phillips and Dudik 2008). Deviance measures how well calibrated predictions points are in relation to known occurrences (Phillips and Dudik 2008). In other words, deviance is a quality-of-fit measure calculated by multiplying the log of likelihood by negative two (-2*log(likelihood)), where likelihood is comparing a model to the data. Sensitivity is known as the true positive rate and measures the proportion of correctly-identified occurrences. Specificity is known as the true negative rate and measures the proportion of correctly-identified absences in the study area (Crase et al. 2012). TSS is the sum of sensitivity and specificity minus one: (Sensitivity + Specificity) -1, and measures model accuracy without being influenced by prevalence (Allouche et al. 2006). Finally,

the MaxEnt and BRT models were compared using the ArcMap 10.3 "Raster Correlations and Summary Statistics" tool in the SDM Toolbox (version 1.1c). This tool determined how similar the probability maps were.

Model Prediction

The MaxEnt models were predicted in MaxEnt software (version 3.3.3k). The output ASCII files representing the probabilities of occurrence were converted to 30-m resolution raster layers in ArcMap 10.2.

A 100-m resolution grid was created in ArcMap 10.2 that covered the entire study area and a centroid was generated in each cell. The environmental data for every centroid were extracted in the same way as the presence/pseudo-absence points. The grid data were standardized (z-score scaling) using the averages and standard deviations from the presence/pseudo-absence data, resulting in four standardized grid datasets saved as csv files that correspond to their respective presence/pseudo-absence data. The standardized grid data were loaded into R and used to predict the BRT model output using functions in the gbm package. Finally, the predicted probabilities for each dataset were loaded into ArcMap 10.2 and converted to raster layers with a 30-m resolution.

RESULTS:

Presence-only MaxEnt models

The best regularization parameter for both MaxEnt models was 2, which showed less overfitting in the probability maps without compromising accuracy. The first MaxEnt model (82 presences) yielded an AUC of 0.952 but low sensitivity, specificity, and TSS (Table 3). The second model (534 presences) had a slightly higher AUC of 0.961, sensitivity, specificity, and TSS. Finally, MaxEnt 1 had a slightly lower deviance compared to MaxEnt 2 (Table 3).

The receiver operating characteristic (ROC) curve shows that the MaxEnt 1 (82 presences) model has more variation around the mean AUC (Figure 5), while the MaxEnt 2 (534 presences) model follows the curve more closely (Figure 6). This likely reflects greater patchiness in presence at fine scales compared to occurrences that are aggregated in larger grid sizes in greater extents. The MaxEnt 1 map predicted more area away from the presence points (Figure 7) while there are less predicted areas in the MaxEnt 2 map (Figure 8).

The ROC curves also show that Maxent 1 had a lower sensitivity, meaning that it misidentified areas with presence points as having low probability of occurrence (Figure 5). The circled areas on the probability map show where the model did not predict correctly (Figure 7). Zooming in to the same circled areas on the MaxEnt 2 map showed that the model predicted slightly higher probabilities of occurrence, which is likely why MaxEnt 2 had a higher sensitivity (Figure 8). The top five predictor variables are fairly different between the two models; rock type, May precipitation, LULC, soil type, and May NDVI were the highest contributing variables for the

MaxEnt 1 model and May precipitation, February temperature, LULC, soil type, and rock type had the highest percent contributions for the MaxEnt 2 model (Table 4).

Table 3. Comparison of MaxEnt model evaluation statistics measuring predictive accuracy.

Model	#Presences / #Background pts	AUC	Deviance	Sensitivity	Specificity	True Skill Statistic (TSS)
MaxEnt 1	82 / 10,000	0.952 ± 0.029	0.509	0.866 ± 0.038	0.890 ± 0.035	0.756
MaxEnt 2	534 / 10,000	0.961 ± 0.004	0.539	0.948 ± 0.010	0.923 ± 0.012	0.871

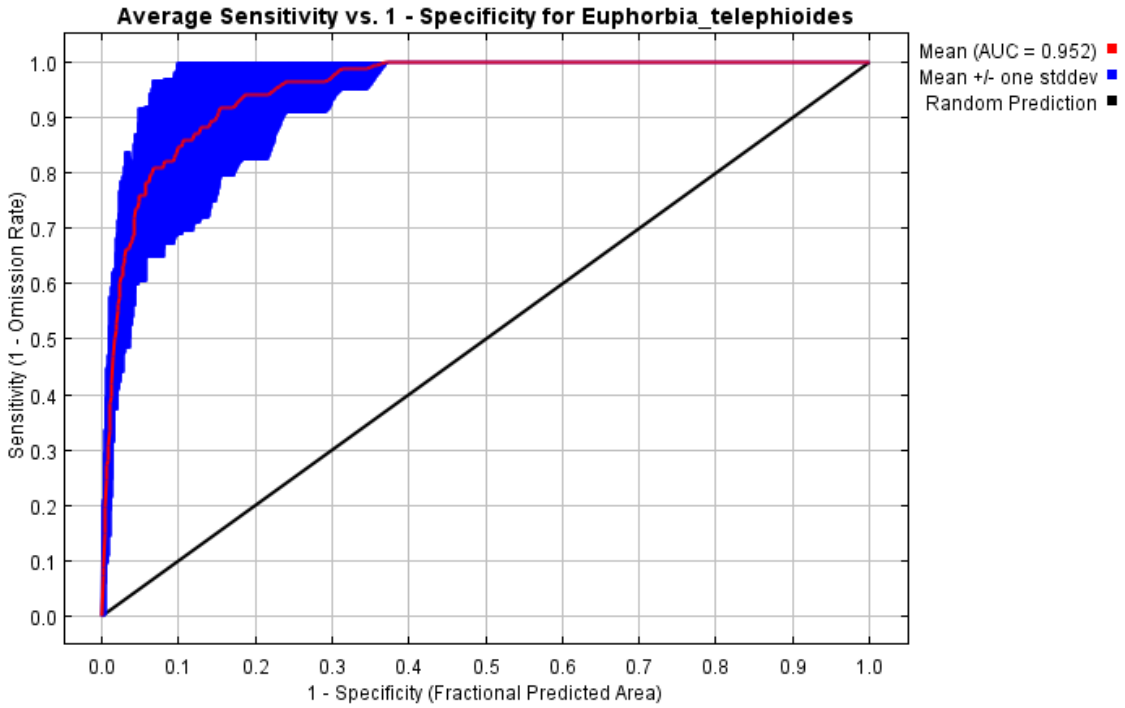


Figure 5. Receiving Operator Characteristic (ROC) curve for MaxEnt 1 (82 presences) model.

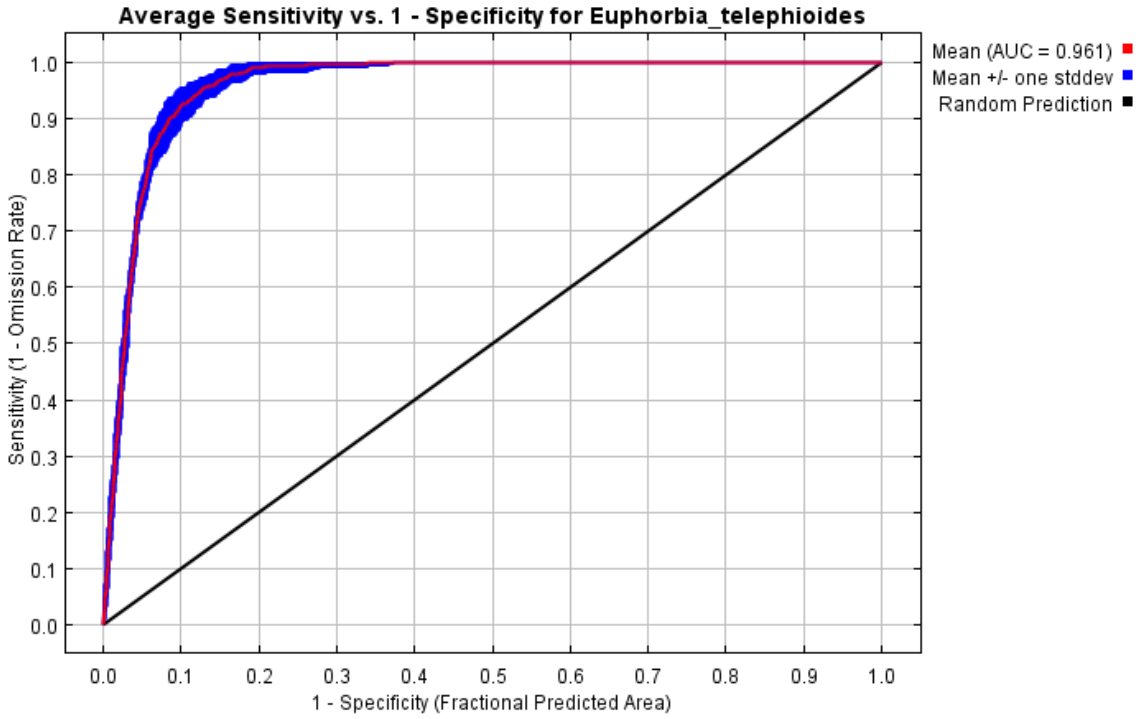


Figure 6. Receiving Operator Characteristic (ROC) curve for MaxEnt 1 (534 presences) model.

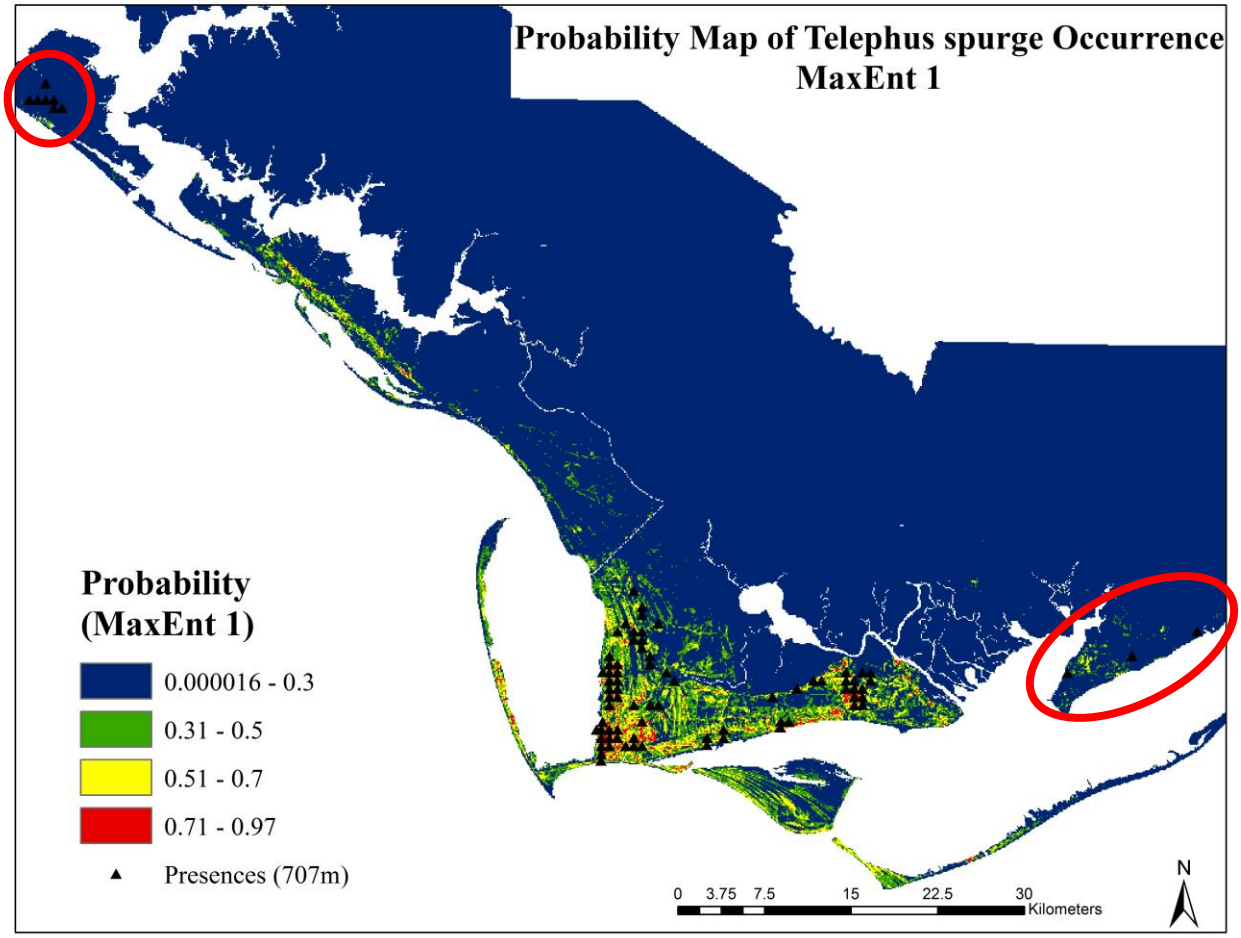


Figure 7. MaxEnt 1 (82 presences) probability of occurrence map. Circled areas show where the model predicted low probability of occurrence where there are presence points.

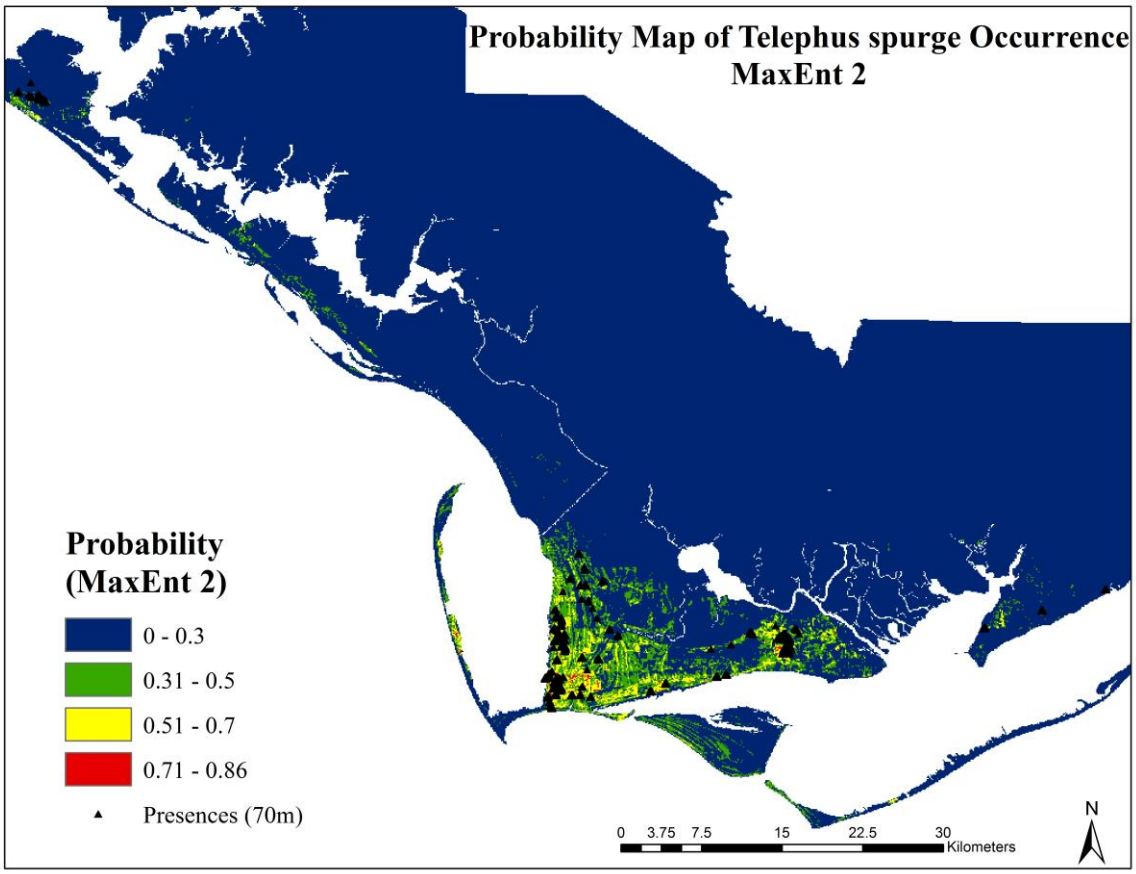


Figure 8. MaxEnt 2 (534 presences) probability of occurrence map.

Table 4. Percent contribution of top 12 variables in the MaxEnt (1 & 2) models. Full predictor variable names listed in Appendix D.

Predictor Variable	Percent Contribution	Predictor Variable	Percent Contribution
	MaxEnt 1 (82 presences)		MaxEnt 2 (534 presences)
Rock type	42.5%	May precip.	63.6%
May precip.	30.2%	Feb. temp.	7.4%
LULC	9.3%	LULC	6.0%
Soil type	9.1%	Soil type	5.6%
May NDVI	3.2%	Rock type	4.6%
Elevation	1.9%	Aug. temp.	3.6%
May TCTB	0.9%	May TCTB	2.4%
Feb. TCTB	0.6%	May TCTW	2.0%
Aug. TCTW	0.6%	Elevation	1.7%
May temp.	0.5%	May EVI	1.4%
Aug. temp.	0.3%	Feb. TCTB	0.8%
Feb. TCTG	0.2%	May temp.	0.6%

Presence/Pseudo-absence BRT models

The settings for all BRT models were as follows: the loss function was binomial, learning rate was 0.001, tree complexity was 5, and bagging fraction was 0.5; the optimal number of trees was determined by the model before fitting, and the default of 10-fold cross-validation was maintained.

The first BRT model (Out 1) was fitted with 1600 trees with an AUC of 0.982 and TSS of 0.866 at an optimal threshold of 0.665 (Table 5). This model also had the highest specificity. The second BRT model (Out 2) was fitted with 5800 trees and produced the highest AUC, sensitivity, and TSS (Table 5). The third BRT model (In 1) was fitted with 1400 trees and had the lowest AUC and TSS among all models. Also, BRT (In 1) had relatively low sensitivity and specificity (Table 5). Finally, the fourth BRT model (In 2) was fitted with 6200 trees and produced a high AUC and sensitivity, low specificity and TSS (Table 5). The deviance was higher for the models with less presence points (BRT Out 1 and BRT In 1) compared to the models with 534 presences (Table 5).

Tables 6 and 7 show the top 12 contributing variables for all models. The top five predictor variables were slightly different between the four models, but soil type and May precipitation were always first or second. In comparison, these variables were less influential on the MaxEnt models (Table 4). LULC and Distance-to-ocean were also in the top five for all BRT models. The partial dependence plots for the BRT models (Figures 9-12) showed most variables had a consistent or slightly positive influence on model performance. However, May precipitation had a negative influence for all models, meaning that lower values (< -0.5) increased probability of occurrence.

The probability of occurrence maps showed high variability between models. The BRT (Out 1) map showed a gradual decrease in probability from the coast to areas more inland, which likely mirrors the progression of soil types and May precipitation (Figure 13); the model also predicted new areas where it would be worth sampling for *E. telephioides*. The BRT (Out 2) map showed more evidence of overfitting since areas of high probability were concentrated around the existing presence points (Figure 14). The BRT (In 1) map was the most different, showing a large area of high probability in Bay County (western-most part of the map) while the other maps showed this area to be low probability of occurrence (Figure 15). Finally, the BRT (In 2) map predicted the largest area of high probability (> 0.70), despite having 534 presence points (Figure 16). This was surprising because the BRT (Out 2) model with the same amount of presence points seemed to have overpredicted.

Table 5. Comparison of BRT model evaluation statistics measuring predictive accuracy; PAs = pseudo-absences.

	#Presences / # PAs	AUC	Optimal Threshold	#Trees	Deviance	Sensitivity	Specificity	True Skill Statistic (TSS)
BRT Out 1	82/82	0.982 ± 0.008	0.665	1600	0.158	0.915 ± 0.035	0.951 ± 0.024	0.866
BRT Out 2	534/534	0.989 ± 0.003	0.813	5800	0.0239	0.998 ± 0.0019	0.934 ± 0.011	0.932
BRT In 1	82/82	0.925 ± 0.020	0.675	1400	0.221	0.915 ± 0.031	0.805 ± 0.044	0.720
BRT In 2	534/534	0.965 ± 0.005	0.810	6200	0.0571	0.953 ± 0.0092	0.873 ± 0.014	0.826

Table 6. Percent contribution of top 12 variables in the BRT (Out 1 & 2) models. Full predictor variable names listed in Appendix D.

Predictor Variable	Percent Contribution	Predictor Variable	Percent Contribution
	BRT Out 1 (82 presences)		BRT Out 2 (534 presences)
Soil type	55.9%	Soil type	32.4%
May precip.	29.8%	May precip.	22.9%
Rock type	4.9%	LULC	22.5%
Distance to Ocean	3.6%	Distance to road	5.5%
LULC	2.4%	Distance to ocean	3.2%
Aug. temp.	1.1%	Feb. TCTB	1.9%
Aug. precipitation	0.5%	May TCTB	1.7%
Distance to road	0.3%	Distance to wetland	1.5%
Aug. TCTW	0.2%	Aug. temp.	1.4%
Aug. TCTB	0.2%	May TCTW	1.3%
Aug. EVI	0.2%	Feb. TCTW	0.9%
Distance to wetland	0.1%	Feb. NDVI	0.8%

Table 7. Percent contribution of top 12 variables in the BRT (In 1 & 2) models. Full predictor variable names listed in Appendix D.

Predictor Variable	Percent Contribution	Predictor Variable	Percent Contribution
	BRT In 1 (82 presences)		BRT In 2 (534 presences)
Soil type	49.1%	May precip.	53.9%
May precip.	22.2%	Soil type	26%
LULC	13.9%	LULC	10.4%
May TCTB	2.7%	Distance to ocean	2.1%
Distance to ocean	2.1%	Distance to wetland	1.2%
Feb. TCTB	2.0%	Distance to road	0.9%
Distance to wetland	1.6%	Aug. TCTW	0.7%
Distance to road	1.5%	May EVI	0.6%
Aug. TCTB	1.5%	Aug. temp.	0.5%
Aug. EVI	0.8%	Aug. precip.	0.5%
Aug. TCTW	0.4%	Feb. temp.	0.5%
Elevation	0.4%	Feb. TCTB	0.5%

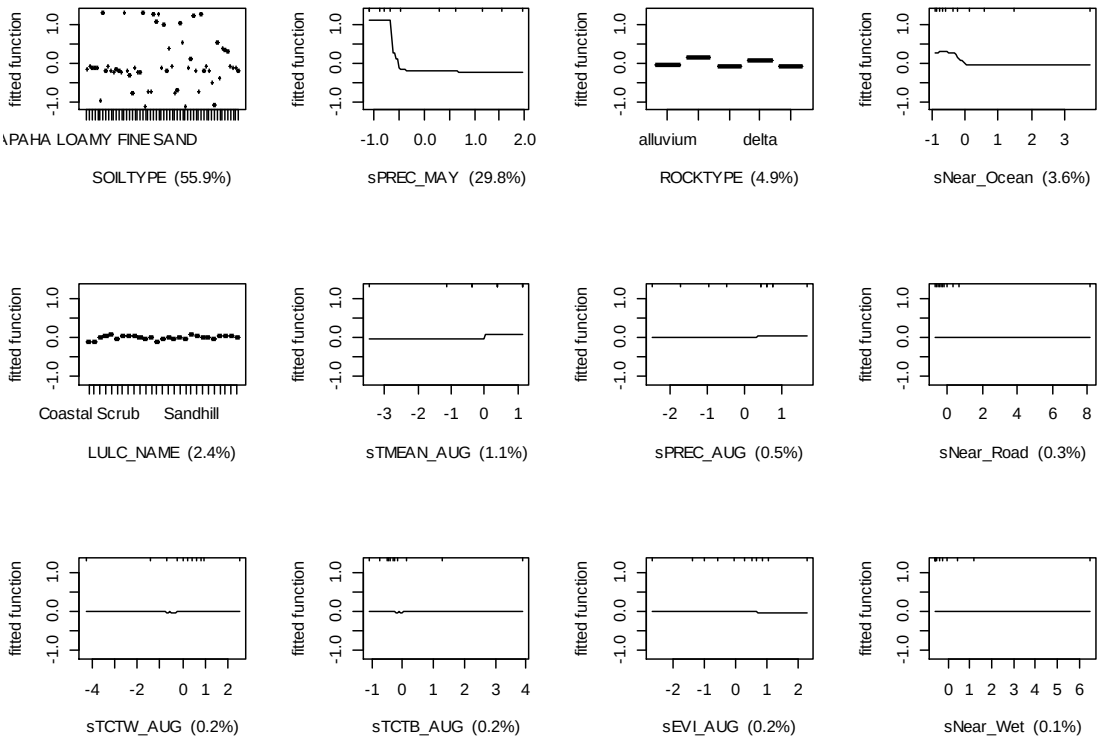


Figure 9. Partial dependence plots for BRT (Out 1, 82 presences) model showing the top 12 contributing variables.

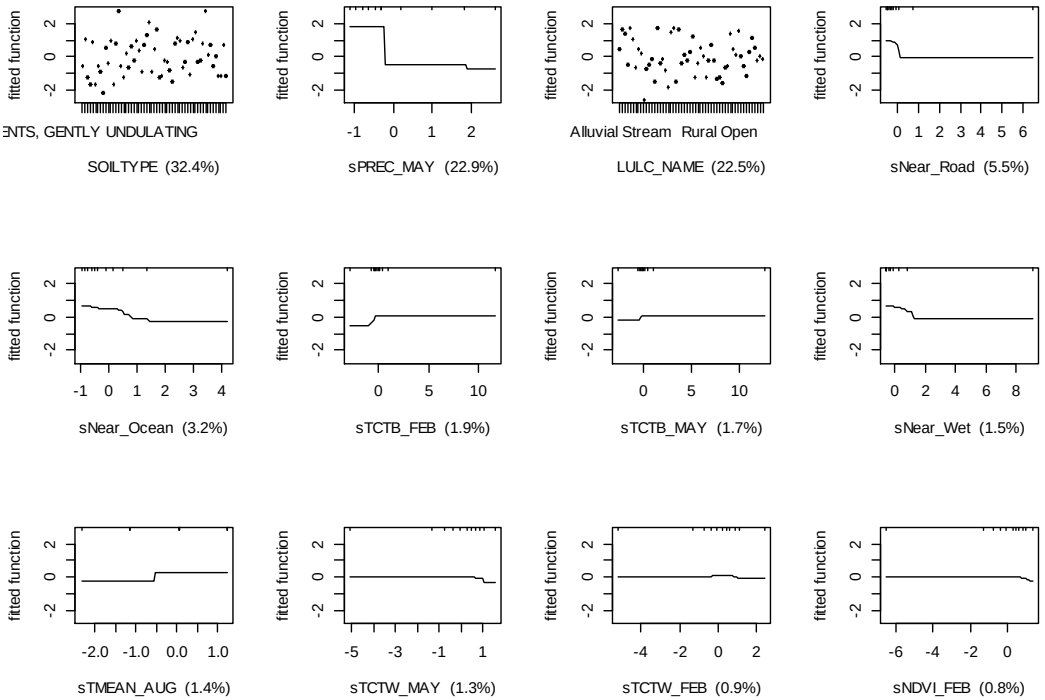


Figure 10. Partial dependence plots for BRT (Out 2, 534 presences) model showing the top 12 contributing variables.

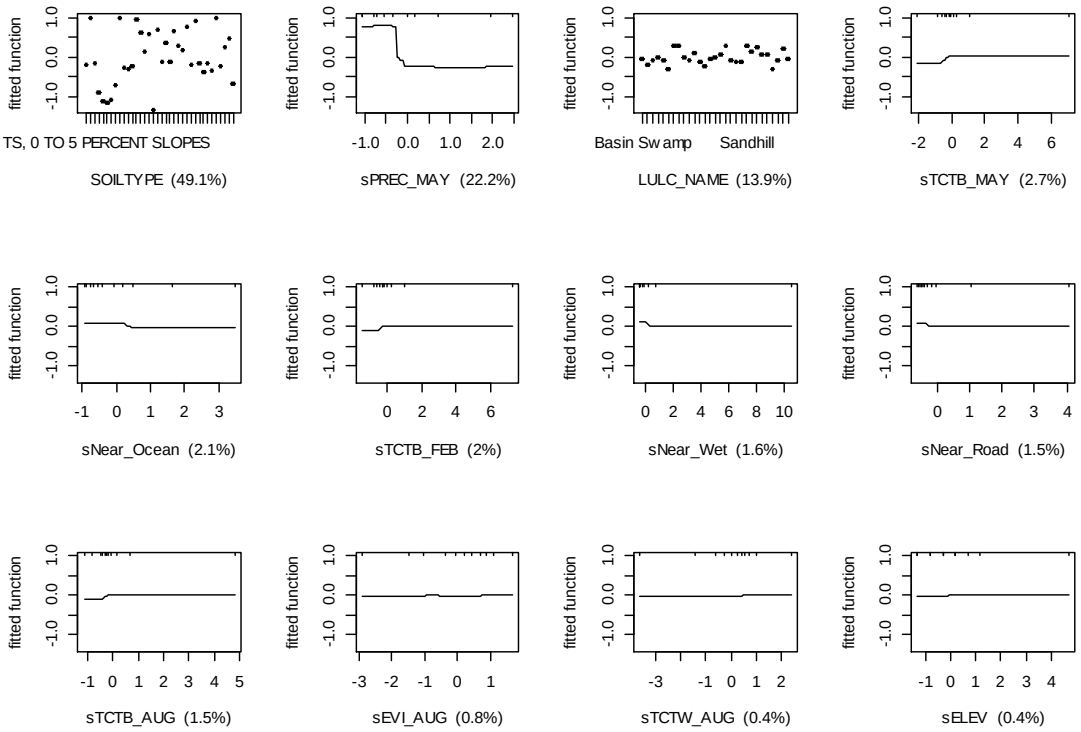


Figure 11. Partial dependence plots for BRT (In 1, 82 presences) model showing the top 12 contributing variables.

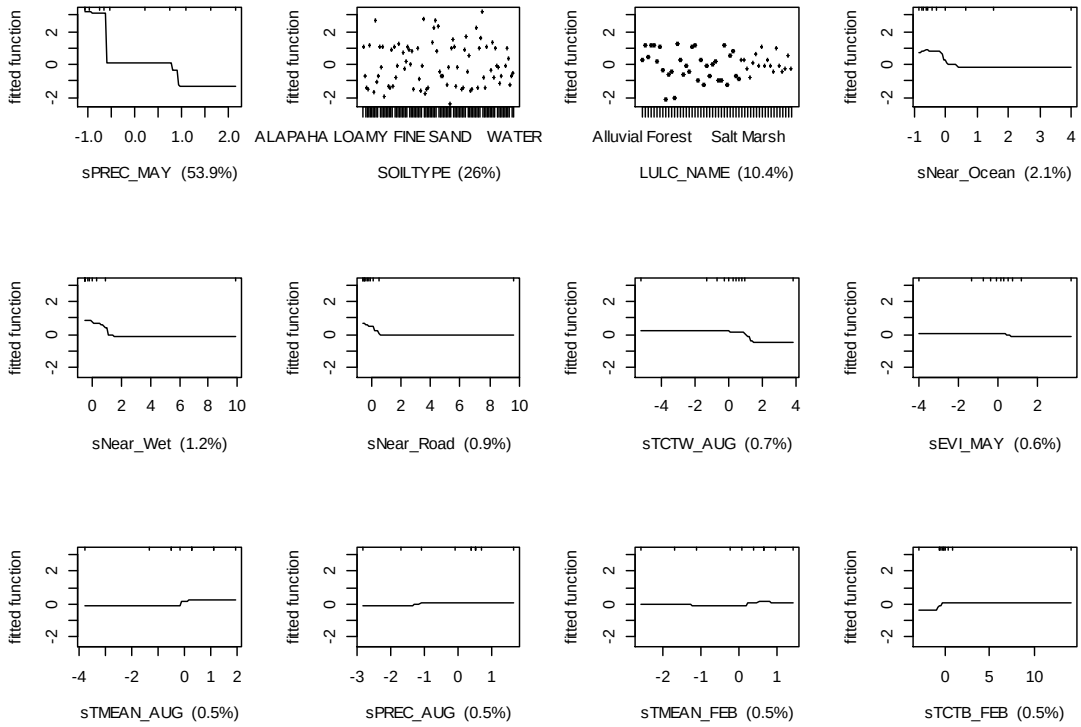


Figure 12. Partial dependence plots for BRT (In 2, 534 presences) model showing the top 12 contributing variables.

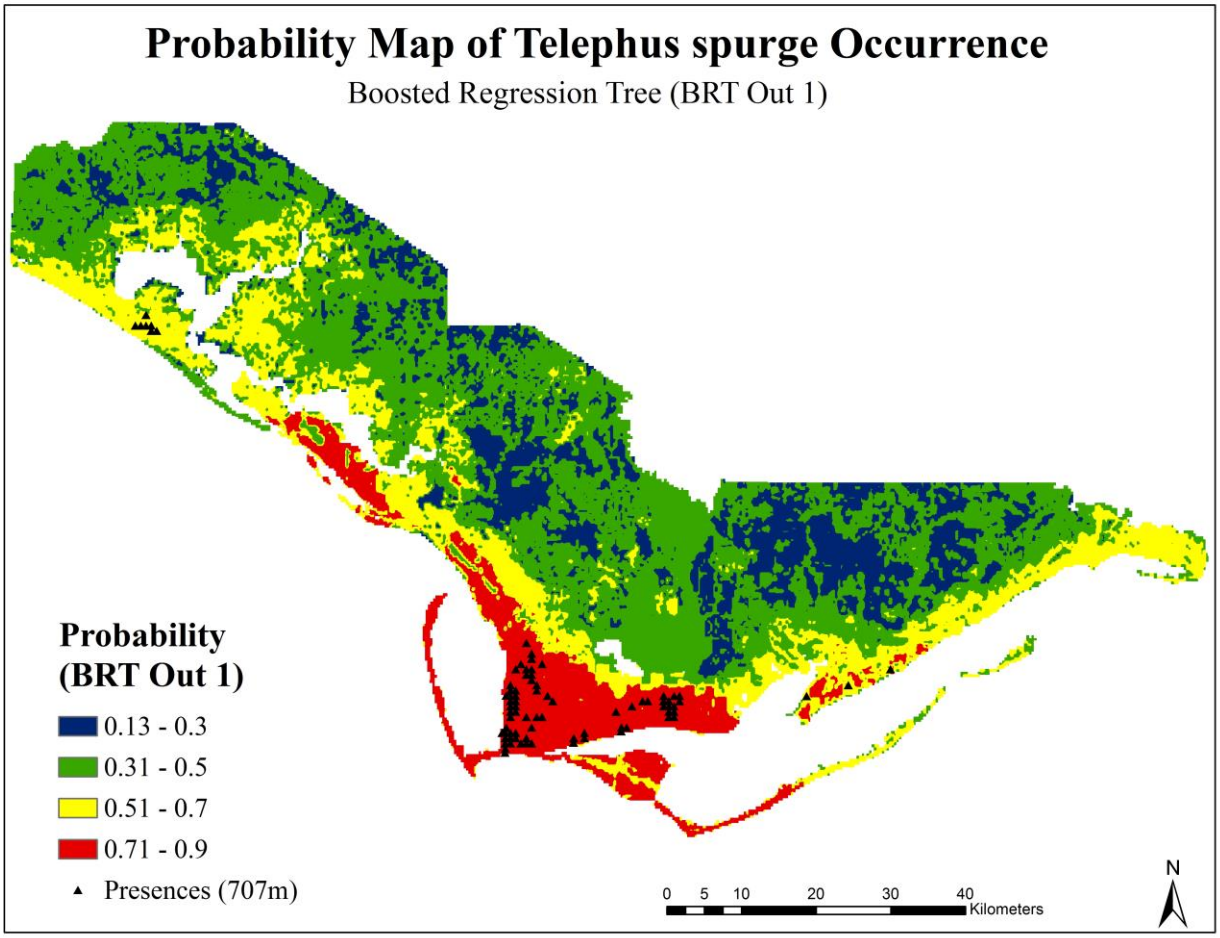


Figure 13. BRT Out 1 (82 presences) probability of occurrence map.

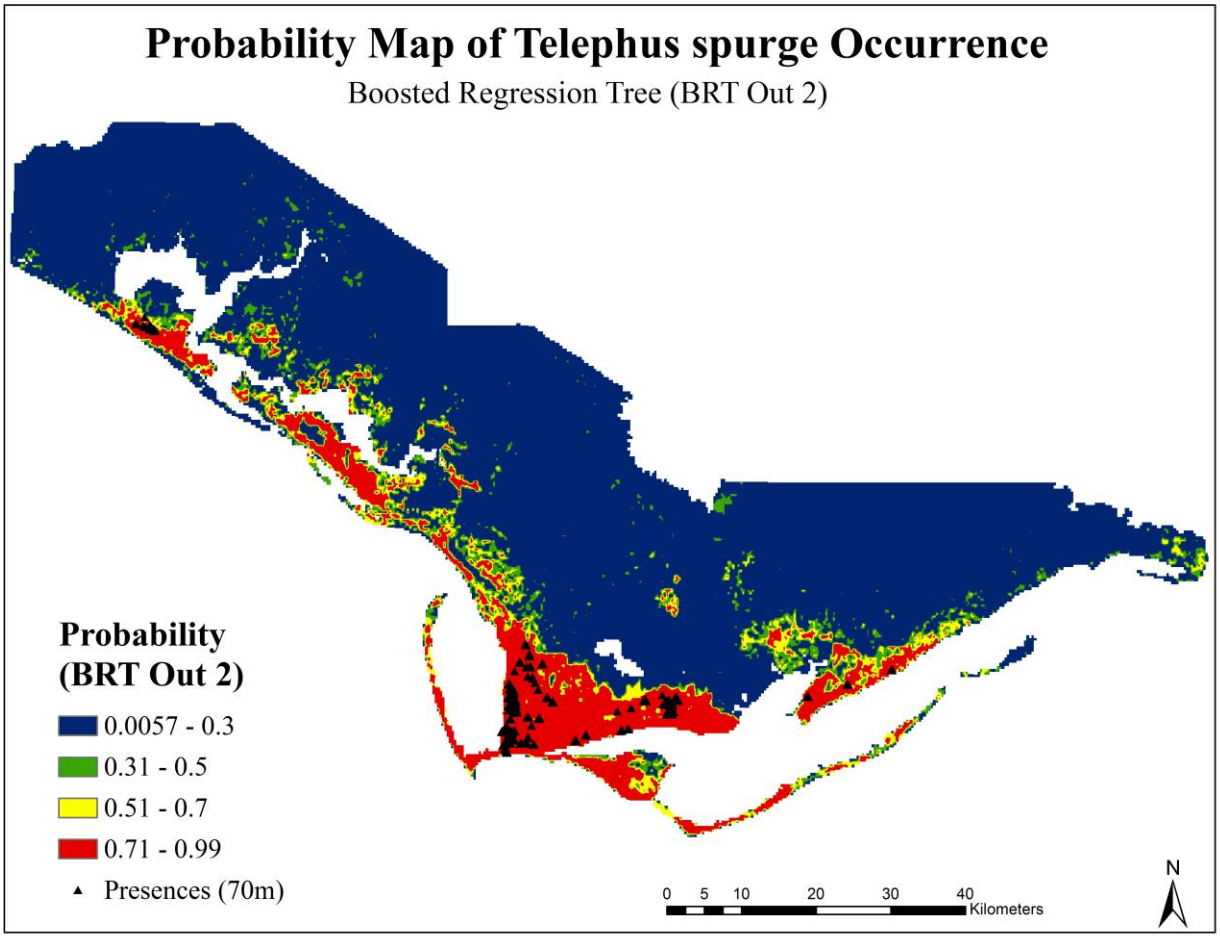


Figure 14. BRT Out 2 (534 presences) probability of occurrence map.

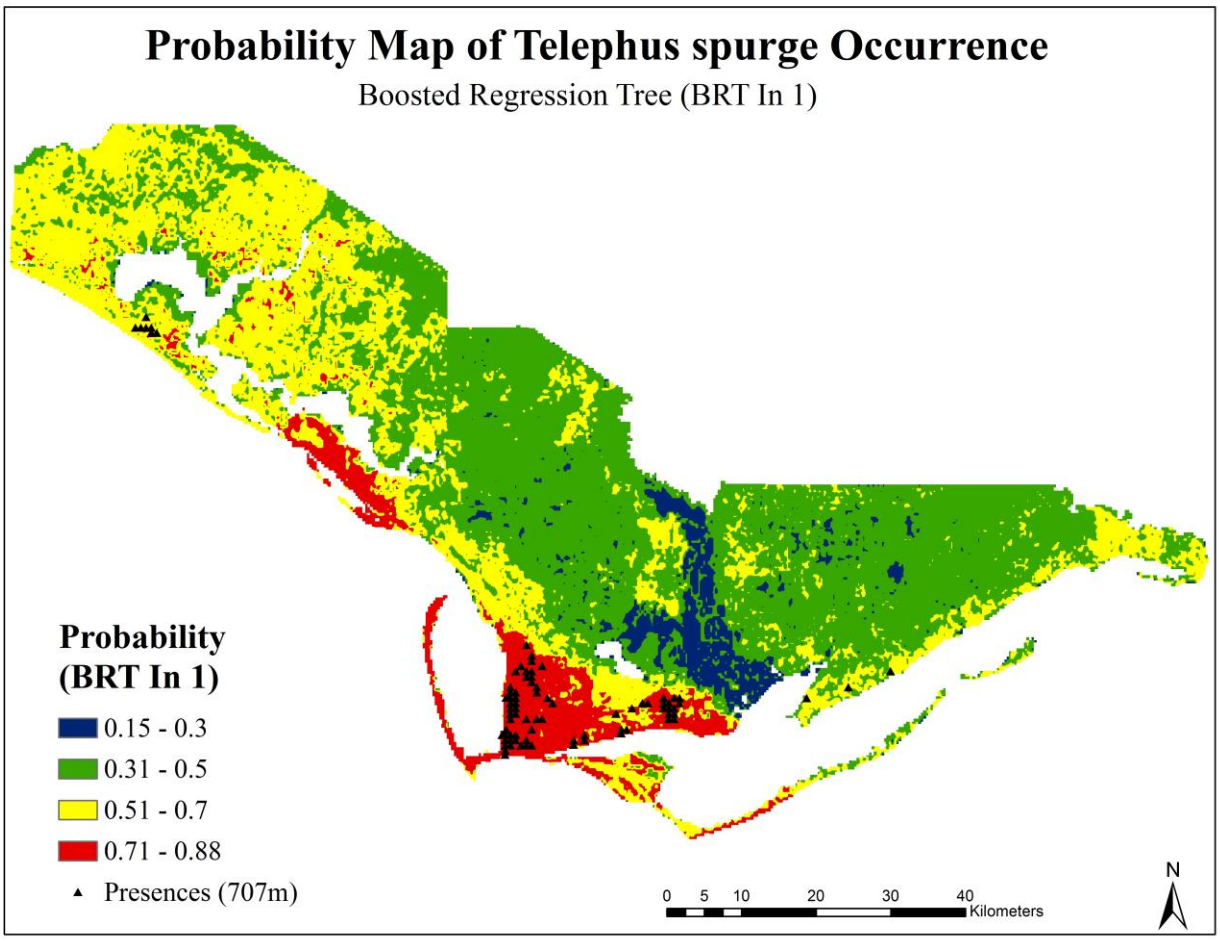


Figure 15. BRT IN 1 (82 presences) probability of occurrence map.

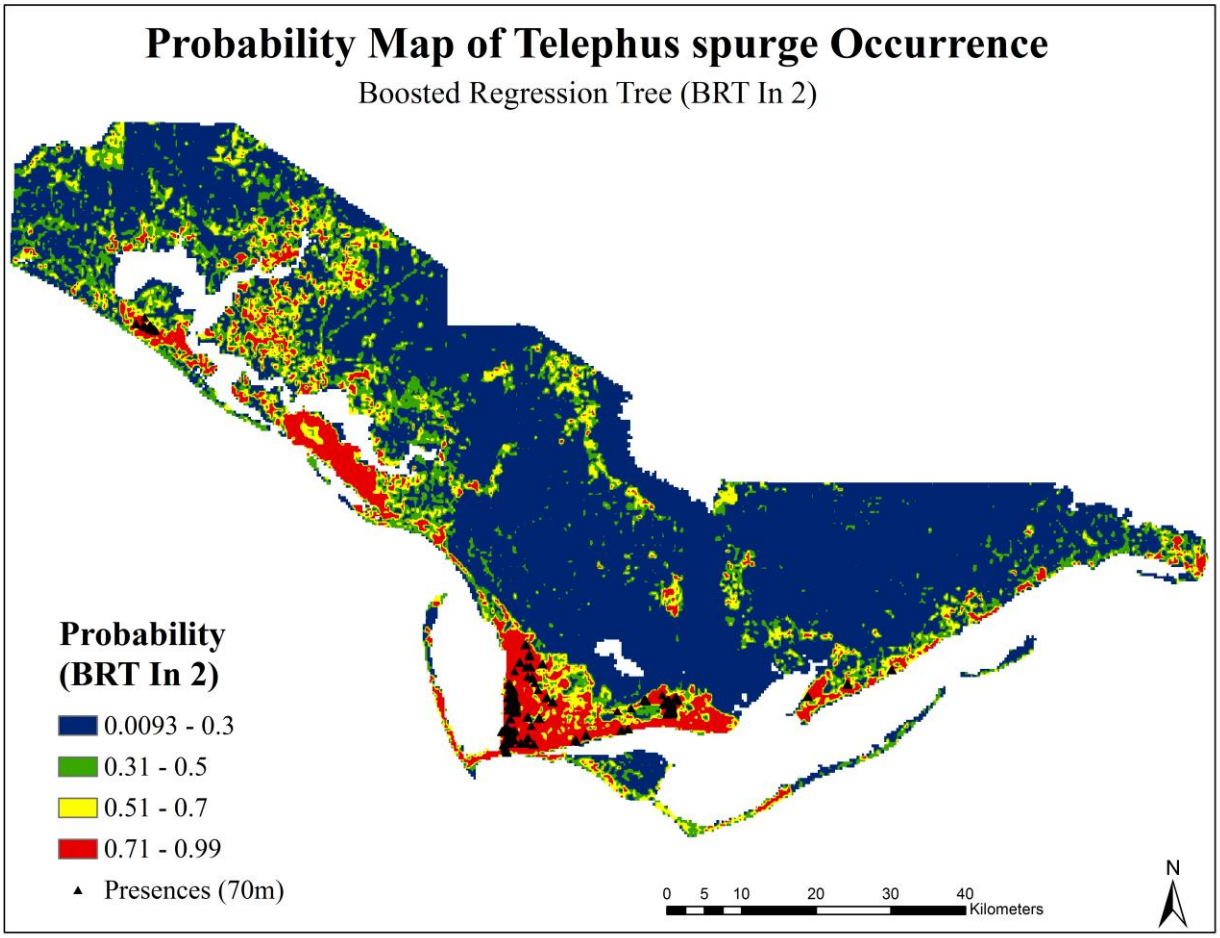


Figure 16. BRT In 2 (534 presences) probability of occurrence map.

Comparison of MaxEnt and BRT models

The MaxEnt models all had lower AUCs when compared to their respective BRT model except for MaxEnt 1 – BRT (In1) (Table 8). Overall, the correlation coefficients showed that predicted values from the MaxEnt models did not correlate with those of the BRT models. The highest correlation coefficient (0.699) resulted when comparing MaxEnt 2 with BRT (Out 2; 534 presences) and the second highest (0.656) was while comparing MaxEnt 1 with BRT (Out 1; 82 presences) (Table 8). These correlations likely resulted because the pseudo-absences in those BRT models were throughout the study area as opposed to within 10 km of the presences.

When comparing correlation coefficients between BRT models, the analysis showed BRT (Out 1) and BRT (Out 2) were more correlated than the other pairs of models. The second highest correlation coefficient (0.826) was while comparing BRT (Out 1) and BRT (In 1) (Table 8).

Table 8. Comparing MaxEnt and BRT models using correlation coefficients and AUC.

Model – Model	Number of presence pts	Correlation coefficient	AUC
MaxEnt 1 – BRT (Out 1)	82	0.656	0.952 - 0.982
MaxEnt 1 – BRT (In 1)	82	0.529	0.952 - 0.925
MaxEnt 2 – BRT (Out 2)	534	0.699	0.961 - 0.989
MaxEnt 2 – BRT (In 2)	534	0.529	0.961 - 0.965
BRT (Out 1) – BRT (In 1)	82	0.826	0.982 – 0.925
BRT (Out 2) – BRT (In 2)	534	0.736	0.989 – 0.965
BRT(Out 1) – BRT (Out 2)	82 - 534	0.838	0.982 – 0.989
BRT (In 1) – BRT (In 2)	82 – 534	0.756	0.925 – 0.965

Soil types associated with E. telephioides occurrence

Soil type was a significant variable for all SDMs, especially the BRTs (Tables 6 and 7). Investigation of the environmental data extracted for both sets of standardized presence points (82 points; 534 points) provided insight on important soil types where *E. telephioides* occurs. The four most important soil types were Leon fine sand, Leon sand, Pickney and Rutlege (depressional), and Mandarin fine sand soils (Table 9), which are all sandy, siliceous, and poorly drained soils that are geographically associated with each other in upland flats and depressions (Soil Survey Staff 2016). Important soil types were defined as those where the percent of presence points occupying a soil type was greater than 2 times (> 2x) the percent of soil type found in the study area. For instance, Leon find sand is only found in 2% of the study area, but 24% of 82 occurrences and 28% of 534 occurrences occupy this soil type (Table 9). The percent occurrences are 12 times (12x) and 14 times (14x) greater, respectively, than the percent soil type found in the study area; therefore, Leon fine sand is likely associated with high probability of *E. telephioides* occurrence.

Table 9. List of most important soil types occupied by *E. telephioides.*

Soil series	Soil type	Percent of soil type found in study area	Percent of 82 occurrences found on soil type	Percent of 534 occurrences found on soil type
Leon	Leon fine sand	2%	24%	28%
	Leon sand	6%	12%	21%
Pickney and Rutlege	Pickney and Rutlege soils, depressional	2%	15%	11%
Mandarin	Mandarin fine sand	0.7%	7%	8%

DISCUSSION:

Model Accuracy and Comparison of Machine-learning SDMs

The results of this study show that machine-learning SDMs are effective at predicting suitable habitat for an endemic rare plant using both broad- and fine-scale environmental variables and methods to reduce overfitting. AUC was high (> 0.9) for all models, and therefore should not be the only measure to consider when choosing the best models. Additionally, both MaxEnt models and the BRT Out 1 and In 1 models may have inflated AUCs since the majority of background and pseudo-absence points were generated far from the presences and should therefore be interpreted with caution for *E. telephioides* and other species with narrow ranges (Phillips et al. 2009; Gogol-Prokurat 2011). The most useful SDM for conservation planning would be one that best discriminates locally between suitable and unsuitable habitat on a continuous scale (Gogol-Prokurat 2011; Lawson et al. 2014). This can be determined based on model deviance, which showed the BRT models to have a better fit and therefore able to more accurately identify areas of suitable habitat compared to MaxEnt. It is highly suggested that future studies consider additional statistical measures to evaluate model performance as AUC can be misleading for species with narrow ranges.

Standardizing presence data was the first priority in this study. The presence data were collected by two agencies over a roughly 40-year period, resulting in a non-standardized way to collect data in the field. Investigating the data in a GIS revealed three different presence data formats: 1) polygons drawn using GPS units in the field to represent an area occupied by numerous plants; 2) a cluster of points where each point represented an individual plant; and 3) single points that indicated multiple plants were in the surrounding area, but the surveyor did not indicate size of area occupied by plants. If similar data are used in other studies, it is recommended that the data be standardized using spatial filtering outlined here in the methods while also testing various grid resolutions to determine an optimal number of presence points. As shown in this study, different numbers of presence points for the same species may produce dissimilar probability maps even though AUC values were relatively the same. Therefore, future studies should continue investigating the optimal number of presence data that would increase model accuracy while decreasing model complexity.

SDMs require thoughtful choices of extent size, model grain size, and environmental variables. The study extent remained constant (three Florida counties covering the entire historical range of *E. telephioides*) when used to generate probability maps, but it is suggested that the extent be limited based on species' prevalence and dispersal capability (Gogol-Prokurat 2011; van Proosdij et al. 2015; Rovzar et al. 2016). This suggestion is worth exploring in future studies using rare species' data to ensure a more accurate model. The model grain size (grid resolution used to map probability of occurrence) for the BRT models was 100 m in order to decrease computer processing time and allow sufficient memory, but was resampled in ArcMap 10.3 to 30 m. MaxEnt predicted the models to the same resolution as the input raster grids (30 m). Accuracy measures (except for deviance) did not differ significantly between the BRT and MaxEnt models despite the difference in model grain sizes, which may be because both grain sizes are spatially considered fine-scale resolutions. Such fine-scale grain sizes were chosen in this study to predict at local spatial scales since the historical range for *E. telephioides* is so narrow. Studies have

confirmed that the use of fine grain sizes actually improved model performance for rare species' SDMs (Gogol-Prokurat 2011; Gottschalk et al. 2011; Song et al. 2013; Rovzar et al. 2016).

The environmental variables included in this study varied in their grain sizes from 10 m (land use-land cover and soil type), to 30 m (Landsat-derived spectral vegetation indices), and finally to 1 km (precipitation and temperature). Since *E. telephioides* is classified as a habitat specialist, I felt it was important for the models to distinguish local differences in soil type and land-use/land cover (LULC) by using highly detailed soils and LULC datasets. Other studies have shown an increase in model performance when incorporating fine-scale environmental predictors, even when comparing results between the use of fine-scale soils data and a simplified soils dataset (Gogol-Prokurat 2011; Rovzar et al. 2016). The machine-learning SDMs in this study were able to handle fine-scale environmental predictors and all resulted in high model performance. In fact, all BRT models had soil type, May precipitation, LULC, and Distance-to-ocean as part of their top five significant variables. Of these variables, only May precipitation is considered a broad-scale predictor (1-km resolution) while the others are fine-scale predictors. MaxEnt, on the other hand, chose two broad-scale predictors as the most significant variables (MaxEnt 1: rock type and May precipitation; MaxEnt 2: May precipitation and February temperature). Broad-scale predictors are more likely to be selected over fine-scale variables if spatial autocorrelation is still affecting model prediction (Franklin 2009); therefore, MaxEnt may not have selected the best predictor variables. However, since the aim of this study was prediction as opposed to explaining the influence of environmental variables, then variable selection in MaxEnt is likely less problematic (Franklin 2009). Therefore, future studies should embrace the use of fine-scale, or high-category, environmental predictors in machine-learning SDMs as well as consider the impact significant variable selection has on interpretation of these models.

Future Research

Disturbance-related environmental predictors are generally absent from SDMs, possibly due to lack of appropriately-scaled data or ecological information about the species (Crimmins et al. 2014). It is well known that *E. telephioides* depends on fire to reproduce and prevent being shaded out by faster-growing plants (Trapnell et al. 2012). However, fire disturbance data was not included in this study because the model output would only map current suitable habitat in the presence of fire as opposed to predicting habitat that may become suitable under future fire regimes (Gogol-Prokurat 2011). By excluding recent fire data in the models, the probability maps can be used to plan areas that would benefit from prescribed fires. Crimmins et al. (2014) showed that fire data did not significantly improve SDMs for fire-dependent plant species; however, this may be because fire occurrences affect plant abundance as opposed to presence/absence (Crimmins et al. 2014). For instance, areas with frequent burning may have higher densities of *E. telephioides* compared to areas where fire is suppressed. Therefore, future SDMs for *E. telephioides* should incorporate demographic patterns of species' abundance when using fire occurrence data.

The majority of SDMs use only abiotic predictor variables to find suitable habitat for the target species. However, these models may be limited because they do not take biotic interactions into account and thus only determine the species' fundamental niche as opposed to its realized niche

(Austin 2002; Meier et al. 2010). Baumberger et al. (2012) showed that using co-occurring species' data along with abiotic predictors produced a model with similar predictive performance as the abiotic-only model, but did not overestimate the target species' presence. Therefore, it could be worth exploring SDMs using presence data for plant species that coexist with *E. telephioides.*

Since *E. telephioides* grows exclusively along the Gulf Coast in northwest Florida, it may be beneficial to develop an SDM predicting suitable habitat under different future climatic conditions and expansion of urban development. Franklin et al. (2014) used a four-step approach to model effects of land-use change and climate change on habitat availability by year 2050 for five plant species in a fire-dominated ecosystem: 1) urban growth scenarios; 2) SDMs based on current conditions; 3) climate change scenarios; and 4) population viability models. Franklin et al. (2014) found that future habitat availability was most affected by land-use change as opposed to climate change in their study area. This study has a well-defined approach to modeling future suitable habitat, but I would suggest modeling land-use change and climate change to year 2100 and maybe even 2150 for *E. telephioides* to better evaluate the full effects of these changes. Also, sea level rise (SLR) would be an important aspect to consider for coastal species, such as *E. telephioides.*

Conclusions

Modeling spatial distributions for rare species with narrow ranges presents a challenge. Predicting the distribution of restricted-range plants may be complicated by dispersal limitations, competition, predation, and stochastic processes (Wiser et al. 1998; Williams et al. 2009). Presence data for rare species contain inherent bias and a deviation from independent observations, which may decrease model accuracy (Phillips et al. 2009; Crase et al. 2012). Therefore, thoughtful preparation of species' presence and absence (or pseudo-absence) data is recommended before fitting each model. This study highlights some methods to improve predictability in machine-learning SDMs while also revealing potential advantages and limitations of using these SDMs with rare species' data. The BRT models were more superior in terms of accuracy and robustness against overfitting the data. MaxEnt did not seem to perform as well in this study, but it should not be ruled out for future models. While exclusively using abiotic predictors did not seem to reduce model performance, future work should investigate biotic interactions and disturbance-related environmental predictors to improve model accuracy.

All models revealed that LULC and soil type were in the top five most significant predictor variables, but three of the four BRTs had soil type as the most significant variable associated with *E. telephioides* occurrence. This may be the most valuable piece of information for conservation planning since *E. telephioides* is known as an edaphic species and thorough soil surveys are available at fine-scale resolutions. Initial data investigation revealed that *E. telephioides* occurrence is associated with Leon soils and other closely related sandy soils. These soils are mostly used for forestry, which may explain why the most dominant LULC type associated with *E. telephioides* presence was coniferous plantations. Conservation of *E. telephioides* and co-occurring species should focus on finding areas where coniferous plantations overlap with Leon, Pickney, Rutlege, or Mandarin sand soils and acquiring these areas for protection. Gagnon and Jokela (2014) investigated uneven-aged restoration and management of Longleaf Pine ecosystems, which may be a way to convince private landowners to convert their

Slash Pine stands to natural Longleaf Pine communities. Restoring and properly managing these areas with prescribed fire, for example, would create more suitable habitat for *E. telephioides.*

As this study was underway, a new population of *E. telephioides* was discovered during vegetation surveys at Tyndall Air Force Base (AFB) in Panama City, FL (Figure 17). This new population was used to initially validate the models after fitting. Overall, the BRT models had higher probabilities of occurrence at the location compared to the MaxEnt models (Table 10). Understandably, more distinct populations need to be found to further validate the models and also help guide future conservation planning for *E. telephioides*.

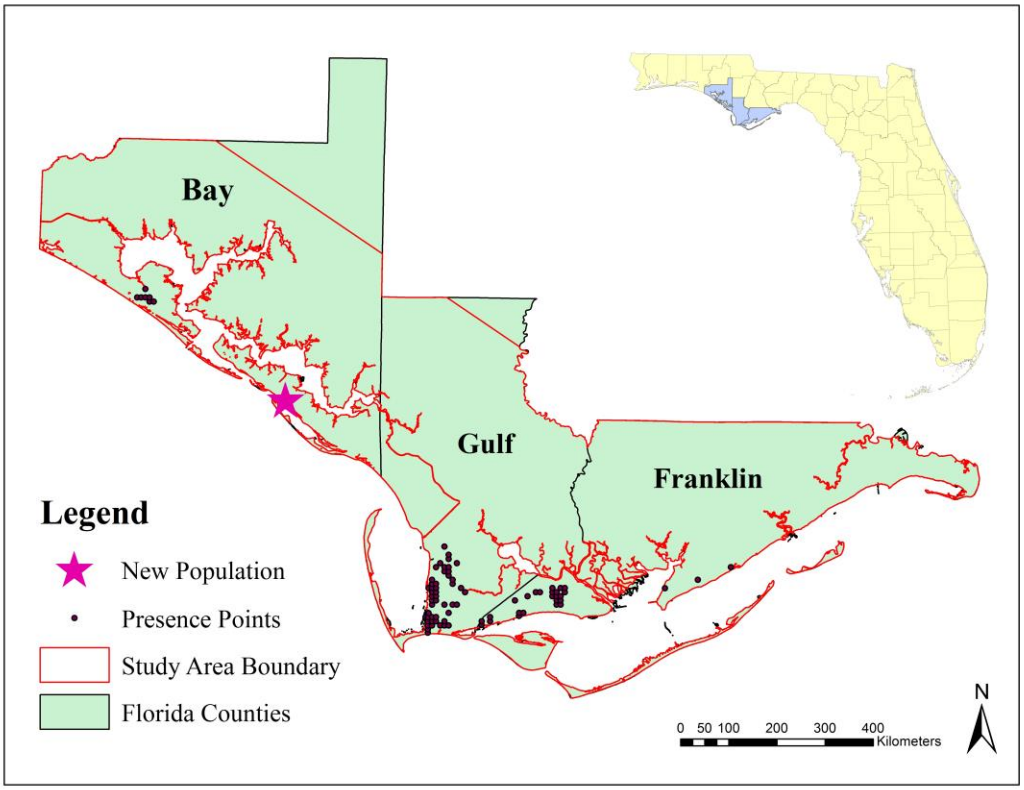


Figure 17. Map depicting the new population of *E. telephioides* at Tyndall Air Force Base in Panama City, FL.

Table 10. Predicted probabilities for new population at Tyndall Air Force Base.

Model	Probability of Occurrence
BRT Out 1	0.859
BRT Out 2	0.942
BRT In 1	0.792
BRT In 2	0.947
MaxEnt 1	0.528
MaxEnt 2	0.298

Literature Cited:

Allouche, Omri, Asaf Tsoar, and Ronen Kadmon. 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology* 43: 1223-1232.

Andriamparany, Jessica N., Katja Brinkmann, Martin Wiehle, Vololoniaina Jeannoda, and Andreas Buerkert. 2015. Modelling the distribution of four *Dioscorea* species on the Mahafaly Plateau of south-western Madagascar using biotic and abiotic variables. *Agriculture, Ecosystems and Environment* 212: 38-48.

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

Baig, Muhammad Hasan Ali, Lifu Zhang, Tong Shuai, and Qingxi Tong. 2014. Derivation of a tasselled cap transformation based on Landsat 8 at-satellite reflectance. *Remote Sensing Letters* 5(5): 423-431.

Barbet-Massin, Morgane, Frédéric Jiguet, Cécile Hélène Albert, and Wilfried Thuiller. 2012. Selecting pseudo-absences for species distribution models: how, where and how many? *Methods in Ecology and Evolution* 3: 327-338.

Baumberger, Teddy, Thomas Croze, Laurence Affre, and François Mesléard. 2012. Co-occurring species indicate habitats of the rare Limonium girardianum. *Plant Ecology and Evolution* 145(1): 31-37.

Bilskie, M.V., S.C. Hagen, S.C. Medeiros, and D.L. Passeri. 2014. Dynamics of sea level rise and coastal flooding on a changing landscape. Geophysical Research Letters, 41(3): 927-934.

Boria, Robert A., Link E. Olson, Steven M. Goodman, and Robert P. Anderson. 2014. Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. *Ecological Modelling* 275: 73-77.

Bridges, E.L., and S.L. Orzell. 2002. *Euphorbia* (Euphorbiaceae) section Tithymalus subsection Inundatae in the southeastern United States. Lundellia. 59-78.

Campbell, James B. and Randolph H. Wynne. Introduction to Remote Sensing, 5th edition. New York: The Guilford Press, 2011.

Crase, Beth, Adam C. Liedloff, and Brendan A. Wintle. 2012. A new method for dealing with residual spatial autocorrelation in species distribution models. *Ecography* 35: 879-888.

Crimmins, Shawn M., Solomon Z. Dobrowski, Alison R. Mynsberge, and Hugh D. Safford. Can fire atlas data improve species distribution model projections? *Ecological Applications* 24(5): 1057-1069.

De'Ath, Glenn. 2007. Boosted Trees for Ecological Modeling and Prediction. *Ecology* 88(1): 243-251.

Ecological Resource Consultants. 2006. Management report for year 2006 for the North Glades Telephus Spurge mitigation. 26 pp.

Elith, Jane, Catherine H. Graham, Robert P. Anderson, Miroslav Dudík, 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, J., J.R. Leathwick, and T. Hastie. 2008. A working guide to boosted regression trees. *Journal of Animal Ecology* 77: 802-813.

Elith, Jane and John R. Leathwick. 2009. Species Distribution Models: Ecological Explanation and Predication Across Space and Time. *The Annual Review of Ecology, Evolution, and Systematics* 40: 677-697.

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

ESRI. 2007. ArcGIS 9.2 Desktop Help. Retrieved from http://webhelp.esri.com/arcgisdesktop/9.2/index.cfm?TopicName=Calculating_slope.

Florida Natural Areas Inventory and Florida Department of Natural Resources. 1990. Guide to the Natural Communities of Florida. Retrieved from http://fnai.org/PDF/Natural_Communities_Guide_1990.pdf.

FNAI. 2014. Cooperative Land Cover Map. Retrieved from http://www.fnai.org/LandCover.cfm.

Franklin, Janet. Mapping Species Distributions: Spatial Inference and Prediction. New York: Cambridge University Press, 2009.

Franklin, Janet, Helen M. Regan, and Alexandra D. Syphard. 2014. Linking spatially explicit species distribution and population models to plan for the persistence of plant species under global change. *Environmental Conservation* 41(2): 97-109.

Gagnon, Jennifer L. and Eric J. Jokela. 2014. Opportunities for Uneven-Aged Management in Second Growth Longleaf Pine Stands in Florida. *University of Florida IFAS Extension.* Retrieved from https://edis.ifas.ufl.edu/fr132.

Gogol-Prokurat, Melanie. 2011. Predicting habitat suitability for rare plants at local spatial scales using a species distribution model. *Ecological Applications* 21(1): 33-47.

Gottschalk, Thomas K., Birgit Aue, Stefan Hotes, and Klemens Ekchmitt. 2011. Influence of grain size on species-habitat models. *Ecological Modelling* 222: 3403-3412.

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.

Lawson, Callum R., Jenny A. Hodgson, Robert J. Wilson, and Shane A. Richards. 2014. Prevalence, thresholds and the performance of presence-absences models. *Methods in Ecology and Evolution* 5:54-64.

Marquardt, Donald W. 1980. A Critique of Some Ridge Regression Methods: Comment. *Journal of the American Statistical Association* 75(369): 87-91.

Meier, Eliane S., Felix Kienast, Peter B. Pearman, Jens-Christian Svenning, Wilfried Thuiller, Miguel B. Araújo, Antoine Guisan, and Niklaus E. Zimmermann. 2010. Biotic and abiotic variables show little redundancy in explaining tree species distributions. *Ecography* 33:1038-1048.

Milligan, Glenn W. and Martha C. Cooper. 1988. A Study of Standardization of Variables in Cluster Analysis. *Journal of Classification* 5: 181-204.

Noss, Reed F. 2011. Between the devil and the deep blue sea: Florida's unenviable position with respect to sea level rise. Climatic Change. 107: 1-16.

NRCS. 2013. Web Soil Survey: Frequently Asked Questions. Retrieved from http://websoilsurvey.nrcs.usda.gov/app/Help/FrequentlyAskedQuestions.htm.

Obata, Kenta, Tomoaki Miura, Hiroki Yoshioka, Alfredo R. Huete, and Marco Vargas. 2016. Spectral Cross-Calibration of VIIRS Enhanced Vegetation Index with MODIS: A Case Study Using Year-Long Global Data. *Remote Sensing* 8(34): 1-17.

Parviainen, Miia, Niklaus E. Zimmermann, Risto K. Heikkinen, and Miska Luoto. 2013. Using unclassified continuous remote sensing data to improve distribution models of red-listed plant species. *Biodiversity Conservation* 22:1731-1754.

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

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

Phillips, Steven J., Miroslav Dudík, Jane Elith, Catherine H. Graham, Anthony Lehmann, John Leathwick, and Simon Ferrier. 2009. Sample Selection Bias and Presence-Only Distribution Models: Implications for Background and Pseudo-Absence Data. *Ecological Applications* 19(1): 181-197.

Radosavljevic, Aleksander and Robert P. Anderson. 2014. Making better MAXENT models of species distributions: complexity, overfitting, and evaluation. *Journal of Biogeography* 41: 629-643.

Ridgeway, Greg. 2007. Generalized Boosted Models: A guide to the gbm package. http://www.saedsayad.com/docs/gbm2.pdf. Accessed March 17, 2016.

Rovzar, Corey, Thomas W. Gillespie, and Kapua Kawelo. 2016. Landscape to site variations in species distribution models for endangered species. *Forest Ecology and Management* 369: 20-28.

Soil Survey Staff, Natural Resources Conservation Service, United States Department of Agriculture. Official Soil Series Descriptions. Available online. Accessed [4/27/2016].

Song, Wonkyong, Eunyoung Kim, Dongkun Lee, Moungjin Lee, and Seong-Woo Jeon. 2013. The sensitivity of species distribution modeling to scale differences. *Ecological Modelling* 248: 113-118.

Shcheglovitova, Mariya and Robert P. Anderson. 2013. Estimating optimal complexity for ecological niche models: A jackknife approach for species with small sample sizes. *Ecological Modelling* 269: 9-17.

Trapnell, Dorset W., J.L. Hamrick, and Vivian Negrón-Ortiz. 2012. Genetic diversity within a threatened, endemic North American species, *Euphorbia telephioides* (Euphorbiaceae). Conservation Genetics. 7: 743-751.

U.S. Fish and Wildlife Service. 2014. 5-year Review: Summary and Evaluation (Telephus spurge). Unpublished report.

van Proosdij, André S. J., Marc S. M. Sosef, Jan J. Wieringa, and Niels Raes. 2015. Minimum required number of specimen records to develop accurate species distribution models. *Ecography* 38: 1-11.

Veloz, Samuel D. 2009. Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. *Journal of Biogeography* 36: 2290-2299.

Williams, John N., Changwan Seo, James Thorne, Julie K. Nelson, Susan Erwin, Joshua M. O'Brien, and Mark W. Schwartz. 2009. Using species distribution models to predict new occurrences for rare plants. *Diversity and Distributions* 15: 565-576.

Wiser, Susan K., Robert K. Peet, and Peter S. White. 1998. Prediction of Rare-Plant Occurrence: A Southern Appalachian Example. *Ecological Applications* 8(4): 909-920.

Young, Nick, Lane Carter, and Paul Evangelista. 2011. A MaxEnt Model v3.3.3e Tutorial (ArcGIS v10). Retrieved from http://ibis.colostate.edu/webcontent/ws/coloradoview/tutorialsdownloads/a_maxent_model_v7.pdf.

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 Modelling* 157: 261-280.

APPENDIX A – LIST OF LAND-USE/LAND COVER TYPES IN STUDY AREA.

Table 10. List of land-use/land cover (LULC) types found in study area in order from highest to lowest area (acres) occupied.

Land-use-Land cover (LULC) type	Area (acres)
Coniferous Plantations	376,796
Hydric Pine Flatwoods	115,895
Other Wetland Forested Mixed	76,121
Floodplain Swamp	76,102
Mixed Scrub-Shrub Wetland	58,672
Tree Plantations	57,995
Transportation	45,500
Wet Prairie	34,085
Rural Open	29,257
Salt Marsh	25,864
Residential, Med.Density -2-5 Dwelling Units/AC	23,341
Mixed Wetland Hardwoods	13,571
Mesic Flatwoods	10,245
Coastal Scrub	10,044
Upland Coniferous	9,713
Cypress	7,924
Residential, High Density > 5 Dwelling Units/AC	7,451
Mixed Hardwood-Coniferous	6,889
Shrub and Brushland	5,527
Marshes	5,393
Commercial and Services	5,380
Wet Flatwoods	5,352
Utilities	4,335
Riverine	4,155
Lacustrine	3,551
Institutional	3,369
Sandhill	3,309
Scrubby Flatwoods	3,220
Pine Flatwoods and Dry Prairie	3,098
Extractive	2,902
Tidal Flat	2,450
Sand Beach (Dry)	2,383
Beach Dune	2,271
Baygall	1,996
Artificial Impoundment/Reservoir	1,951

Titi Swamp	1,754
Wet Coniferous Plantation	1,646
Unimproved/Woodland Pasture	1,554
Marine	1,426
Golf Courses	1,424
Floodplain Marsh	1,409
Estuarine	1,393
Alluvial Stream	1,370
Improved Pasture	1,296
Bare Soil/Clear Cut	1,181
Sod Farms	1,149
Basin Swamp	1,061
Urban Open Forested	1,032
Alluvial Forest	1,029
Field Crops	972
Coastal Interdunal Swale	928
Rural Open Forested	862
Coastal Grassland	843
Industrial	838
Non-vegetated Wetland	831
Urban Open Land	702
Upland Hardwood Forest	661
Community rec. facilities	655
Basin Marsh	576
Shrub Bog	529
Canal	505
Floating/Emergent Aquatic Vegetation	454
Aquacultural Ponds	454
High Intensity Urban	429
Sand Pine Scrub	350
Maritime Hammock	344
Residential, Low Density	303
Rural Structures	299
Flatwoods/Prairie/Marsh Lake	269
Quarry Pond	265
Depression Marsh	241
Rural Open Pine	228
Sand n Gravel Pits	227
Cemeteries	200
Palmetto Prairie	200
Specialty Farms	195
Hydric Hammock	188
Tidally-Influenced Stream	168

Cultural - Terrestrial	165
Hardwood Plantations	114
Communication	112
Strip Mines	97
Sewage Treatment Pond	97
Urban Open Pine	93
Orchards/Groves	88
Mesic Hammock	88
Other Hardwood Wetlands	88
Industrial Cooling Pond	87
Parks and Zoos	86
Scrub	77
Bay Swamp	73
Coastal Uplands	72
Wiregrass Savanna	64
Live Oak	51
Cultural - Lacustrine	46
Slough	37
Trees	37
Blackwater Stream	27
Stormwater Treatment Areas	26
Ornamentals	25
Fallow Cropland	25
Dome Swamp	23
Coastal Dune Lake	21
Unconsolidated Substrate	20
Tree Nurseries	19
Natural Rivers and Streams	17
Successional Hardwood Forest	16
Ballfields	13
Other Open Lands - Rural	8
Dry Flatwoods	7
Other Coniferous Wetlands	6
Oak - Cabbage Palm Forest	6
Roads	6
Coastal Hydric Hammock	6
Vineyard and Nurseries	5
Urban	5
Isolated Freshwater Swamp	4
Mowed Grass	4
Cabbage Palm	3
Rural	3
Oyster Bar	2

Artificial/Farm Pond	1
Cabbage Palm Hammock	1
Spoil Area	1
Natural Lakes and Ponds	-
Bottomland Forest	-

APPENDIX B – LIST OF SOIL TYPES IN STUDY AREA.

Table 11. List of soil types found in study area in order from highest to lowest area (acres) occupied.

Soil type	Area (acres)
LEON SAND	62,261
SCRANTON FINE SAND	61,477
PLUMMER FINE SAND	54,625
POTTSBURG SAND	43,897
RUTLEGE SAND	41,637
RUTLEGE FINE SAND	38,749
PELHAM LOAMY FINE SAND	36,722
BRICKYARD, CHOWAN, AND KENNER SOILS, FREQUENTLY FLOODED	32,181
SCRANTON SAND, SLOUGH	30,208
PICKNEY-PAMLICO COMPLEX, DEPRESSIONAL	29,499
HURRICANE SAND	27,657
ALBANY SAND, 0 TO 2 PERCENT SLOPES	25,897
PAMLICO-DOROVAN COMPLEX	24,689
RUTLEGE-PAMLICO COMPLEX	20,285
PICKNEY AND RUTLEGE SOILS, DEPRESSIONAL	19,817
CHOWAN, BRICKYARD, AND KENNER SOILS, FREQUENTLY FLOODED	19,801
RAINS FINE SANDY LOAM	19,757
LEON FINE SAND	17,448
SURRENCY MUCKY FINE SAND, DEPRESSIONAL	17,440
LEEFIELD SAND	15,776
PLUMMER SAND	15,673
MAUREPAS MUCK, FREQUENTLY FLOODED	15,152
FOXWORTH SAND, 0 TO 5 PERCENT SLOPES	14,937
LAKELAND SAND, 0 TO 5 PERCENT SLOPES	13,084
LEEFIELD LOAMY FINE SAND	13,025
BOHICKET AND TISONIA SOILS, TIDAL	12,579
SURRENCY FINE SAND	12,449
PANTEGO AND BAYBORO SOILS, DEPRESSIONAL	12,199
CROATAN-SURRENCY COMPLEX,	12,071

FREQUENTLY FLOODED	
CHIPLEY SAND, 0 TO 5 PERCENT SLOPES	11,617
WATER	11,590
OSIER FINE SAND	10,766
ALBANY SAND	10,167
RESOTA FINE SAND, 0 TO 5 PERCENT SLOPES	9,848
PAMLICO-PICKNEY COMPLEX, FREQUENTLY FLOODED	9,811
MEADOWBROOK SAND	9,686
BRICKYARD SILTY CLAY, FREQUENTLY FLOODED	9,224
LYNN HAVEN SAND	8,679
ALAPAHA LOAMY FINE SAND	8,552
MEADOWBROOK SAND, SLOUGH	8,159
MEGGETT FINE SANDY LOAM, OCCASIONALLY FLOODED	8,035
MEADOWBROOK FINE SAND, OCCASIONALLY FLOODED	8,026
PELHAM SAND	8,024
MANDARIN FINE SAND	7,855
PICKNEY FINE SAND	6,986
BAYVI LOAMY SAND	6,566
BLANTON FINE SAND, 0 TO 5 PERCENT SLOPES	6,207
DUCKSTON-RUTLEGE-COROLLA COMPLEX	5,731
ARENTS, 0 TO 5 PERCENT SLOPES	5,676
ALAPAHA LOAMY SAND	5,119
KUREB SAND, 0 TO 5 PERCENT SLOPES	4,956
MANDARIN SAND	4,764
RIDGEWOOD SAND, 0 TO 5 PERCENT SLOPES	4,585
ALLANTON SAND	4,405
ORTEGA FINE SAND, 0 TO 5 PERCENT SLOPES	4,333
PELHAM FINE SAND	4,245
HARBESON MUCKY LOAMY SAND, DEPRESSIONAL	4,232
KERSHAW SAND, 2 TO 5 PERCENT SLOPES	4,188
LYNN HAVEN FINE SAND	3,953

SAPELO SAND	3,945
COROLLA SAND, 0 TO 5 PERCENT SLOPES	3,678
PANTEGO SANDY LOAM	3,664
WAHEE-MANTACHIE-OCKLOCKNEE COMPLEX, COMMONLY FLOODED	3,645
DIREGO AND BAYVI SOILS, TIDAL	3,628
DOROVAN-CROATAN COMPLEX, DEPRESSIONAL	3,444
TOOLES-MEADOWBROOK COMPLEX, DEPRESSIONAL	3,318
DOROVAN-PAMLICO COMPLEX, DEPRESSIONAL	3,228
STILSON LOAMY FINE SAND, 0 TO 5 PERCENT SLOPES	3,179
STILSON SAND, 0 TO 5 PERCENT SLOPES	3,077
BLANTON SAND, 0 TO 5 PERCENT SLOPES	2,996
ALBANY FINE SAND	2,865
COROLLA-DUCKSTON COMPLEX, GENTLY UNDULATING, FLOODED	2,799
BLADEN FINE SANDY LOAM	2,599
DUCKSTON SAND, OCCASIONALLY FLOODED	2,553
SAPELO FINE SAND	2,470
AQUENTS, GENTLY UNDULATING	2,371
FRIPP-COROLLA COMPLEX, 2 TO 30 PERCENT SLOPES	2,342
CENTENARY SAND, 0 TO 5 PERCENT SLOPES	2,315
BEACHES	2,227
RAINS SAND	1,957
NEWHAN-COROLLA COMPLEX, ROLLING	1,918
MEADOWBROOK, MEGGETT, AND TOOLES SOILS, FREQUENTLY FLOODED	1,819
BONSAI MUCKY FINE SAND, FREQUENTLY FLOODED	1,800
DIREGO MUCK	1,728
RUTLEGE LOAMY FINE SAND, DEPRESSIONAL	1,653
PANSEY LOAMY SAND	1,623
KUREB-COROLLA COMPLEX, ROLLING	1,563
URBAN LAND	1,533

BAYVI AND DIREGO SOILS, FREQUENTLY FLOODED	1,508
TOOLES SAND	1,483
EBRO-DOROVAN COMPLEX	1,418
FOXWORTH SAND, 5 TO 8 PERCENT SLOPES	1,335
POTTSBURG FINE SAND	1,316
DUCKSTON-BOHICKET-COROLLA COMPLEX	1,265
CHAIRES SAND	1,240
PITS	1,184
WATERS OF THE GULF OF MEXICO	1,155
CLARENDON LOAMY FINE SAND, 2 TO 5 PERCENT SLOPES	1,078
WAHEE FINE SANDY LOAM	1,038
FUQUAY LOAMY FINE SAND	1,004
COROLLA FINE SAND, 1 TO 5 PERCENT SLOPES	874
WEHADKEE-MEGGETT COMPLEX, FREQUENTLY FLOODED	829
OCILLA LOAMY FINE SAND, OVERWASH, OCCASIONALLY FLOODED	809
TROUP SAND, 0 TO 5 PERCENT SLOPES	794
STILSON FINE SAND	724
KERSHAW SAND, 5 TO 12 PERCENT SLOPES	697
RIDGEWOOD FINE SAND	689
ALBANY SAND, 2 TO 5 PERCENT SLOPES	669
DUCKSTON-DUCKSTON DEPRESSIONAL COMPLEX, FREQUENTLY FLOODED	655
LAKELAND SAND, 8 TO 12 PERCENT SLOPES	628
AQUENTS, NEARLY LEVEL	607
DOTHAN-FUQUAY COMPLEX, 5 TO 8 PERCENT SLOPES	529
BONIFAY SAND, 0 TO 5 PERCENT SLOPES	516
QUARTZIPSAMMENTS, UNDULATING	489
LYNCHBURG LOAMY FINE SAND	346
KUREB FINE SAND, 3 TO 8 PERCENT SLOPES	340
DOTHAN LOAMY SAND, 2 TO 5 PERCENT SLOPES	296
LAKELAND SAND, 5 TO 8 PERCENT	266

SLOPES	
BLANTON FINE SAND, 5 TO 8 PERCENT SLOPES	224
CHIPLEY SAND, 5 TO 8 PERCENT SLOPES	179
TROUP SAND, 5 TO 8 PERCENT SLOPES	139
UDORTHENTS, NEARLY LEVEL	115
STILSON SAND, 5 TO 8 PERCENT SLOPES	112
LUCY LOAMY FINE SAND, 0 TO 5 PERCENT SLOPES	109
BONIFAY SAND, 5 TO 8 PERCENT SLOPES	24
KENNANSVILLE-EULONIA COMPLEX, 0 TO 5 PERCENT SLOPES	23
TROUP SAND, 8 TO 12 PERCENT SLOPES	14

APPENDIX C – MAPS OF PREDICTOR VARIABLES.

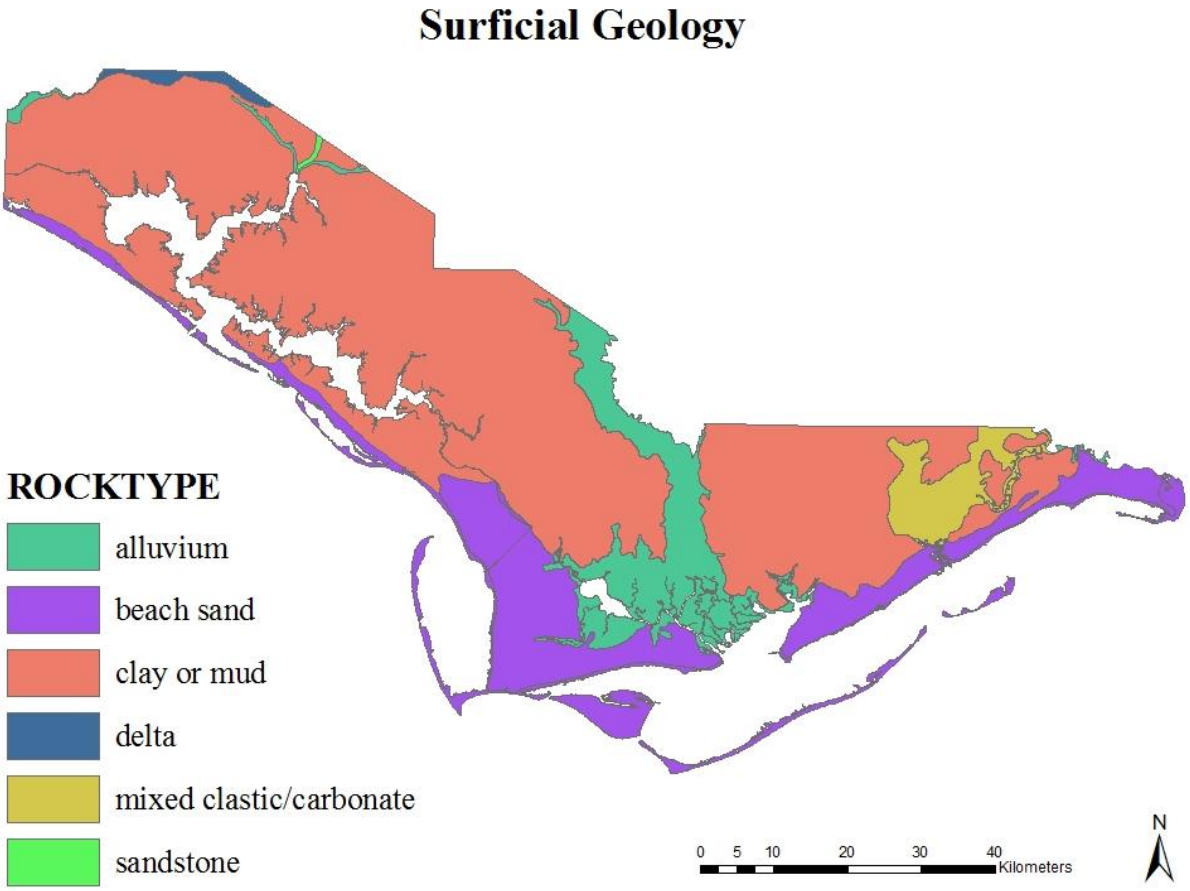
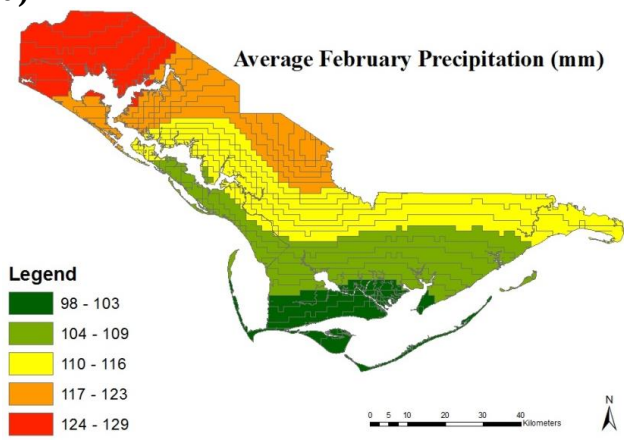
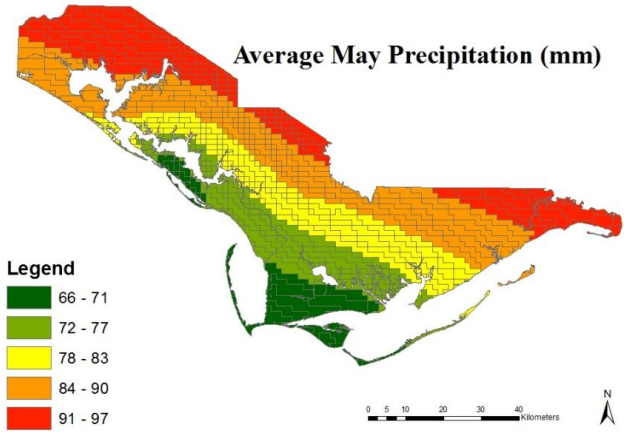


Figure 18. A map displaying the rock types in the study area.

a)



b)



c)

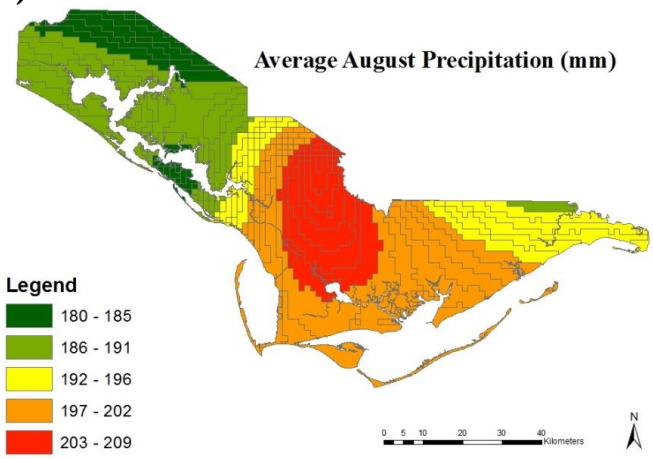
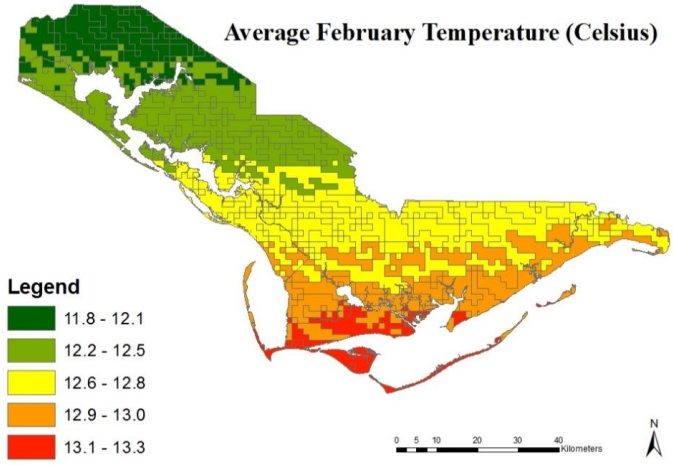
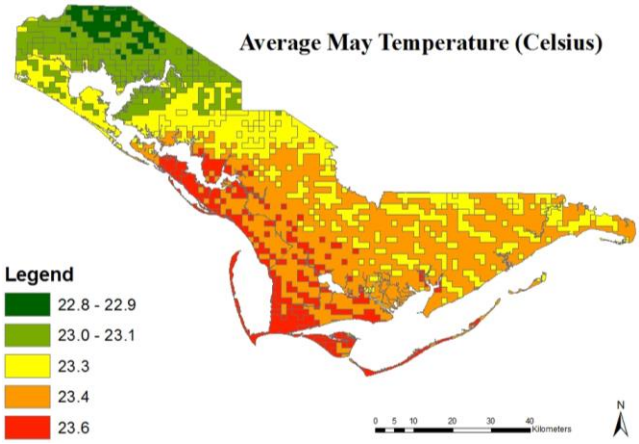


Figure 19. Average monthly precipitation (mm) in a) February, b) May, and c) August (worldclim.org).

a)



b)



c)

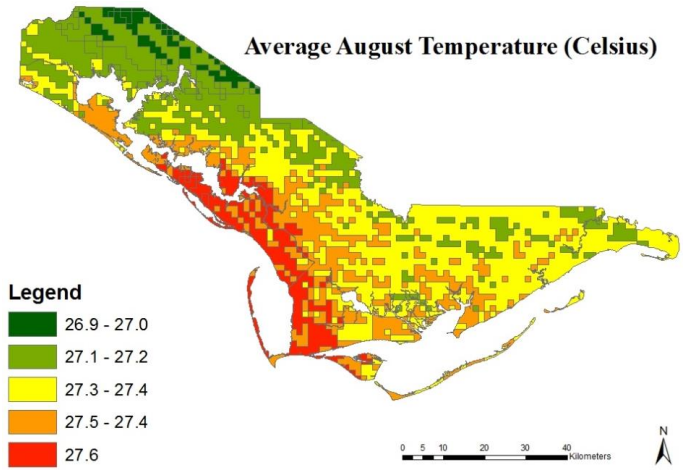


Figure 20. Average monthly temperature (ºCelsius) in a) February, b) May, and c) August (worldclim.org).

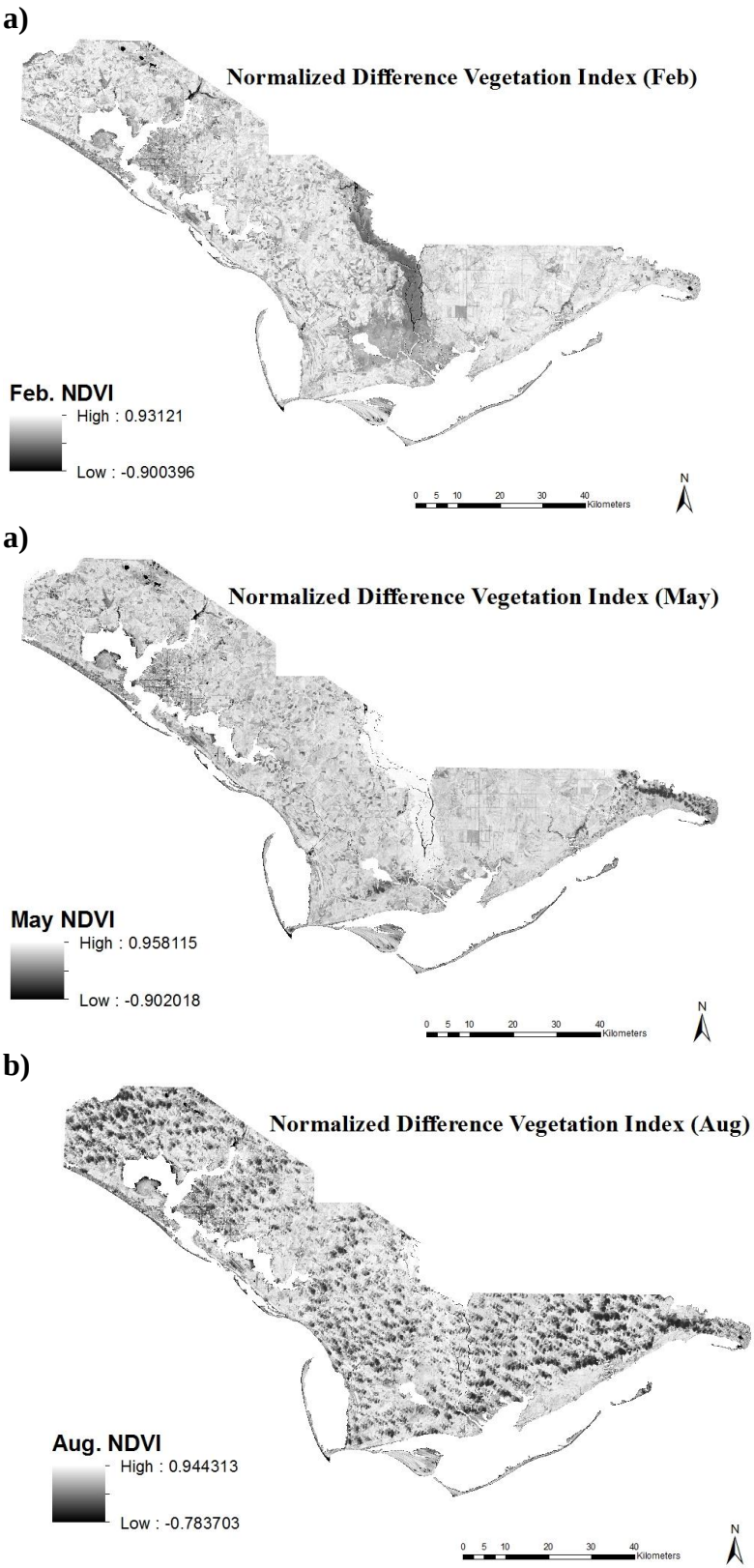


Figure 21. Maps depicting NDVI in a) February, b) May, and c) August 2014.

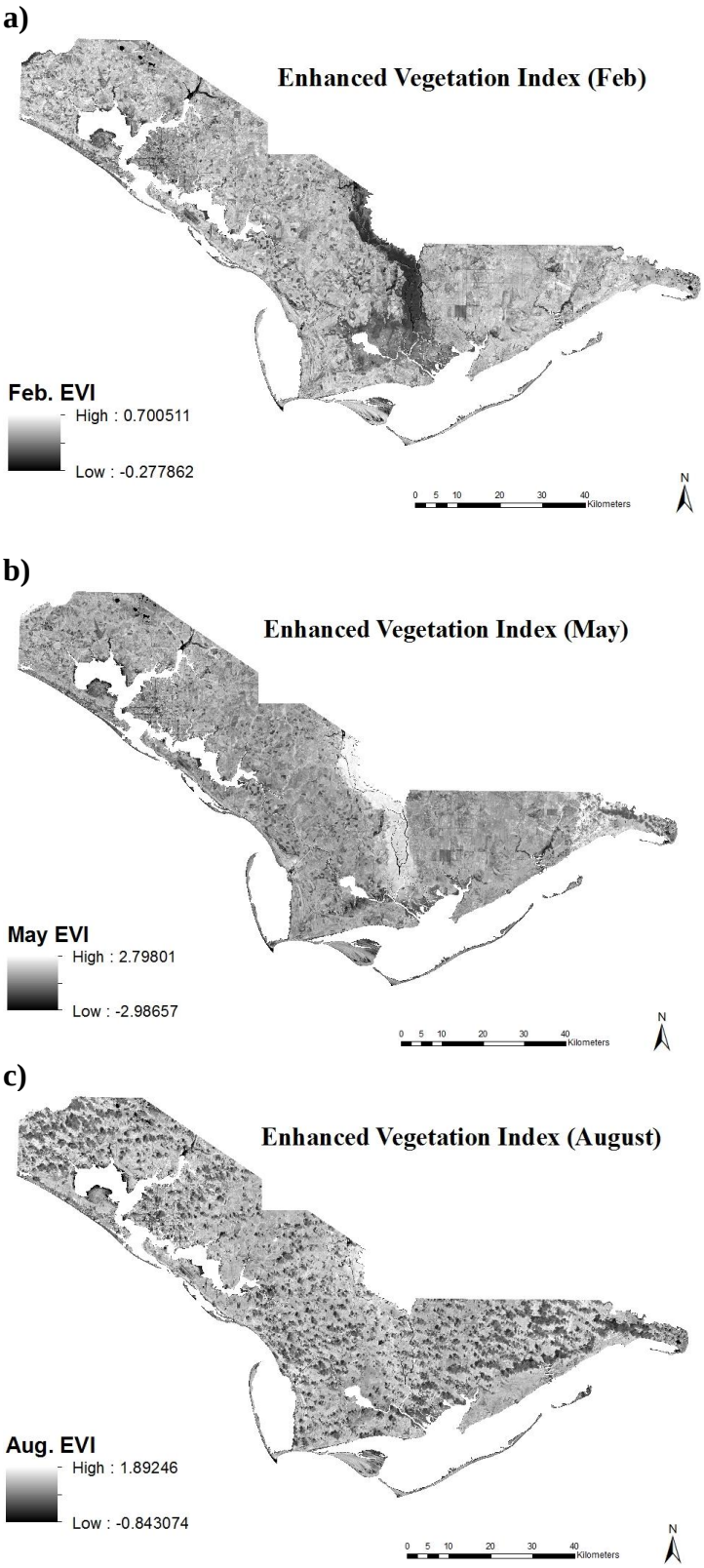


Figure 22. Maps depicting EVI in a) February, b) May, and c) August 2014.

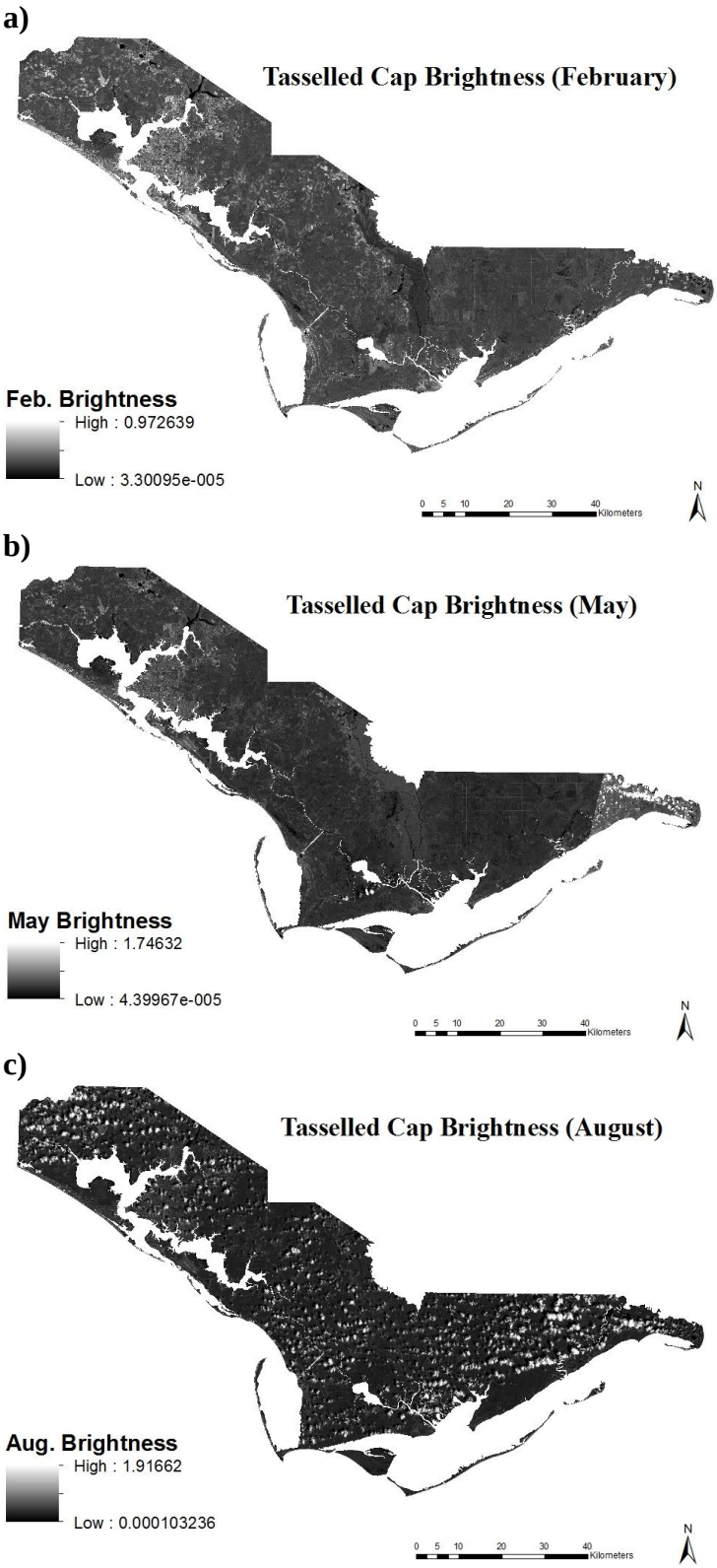


Figure 23. Maps depicting tasselled cap brightness in a) February, b) May, and c) August 2014.

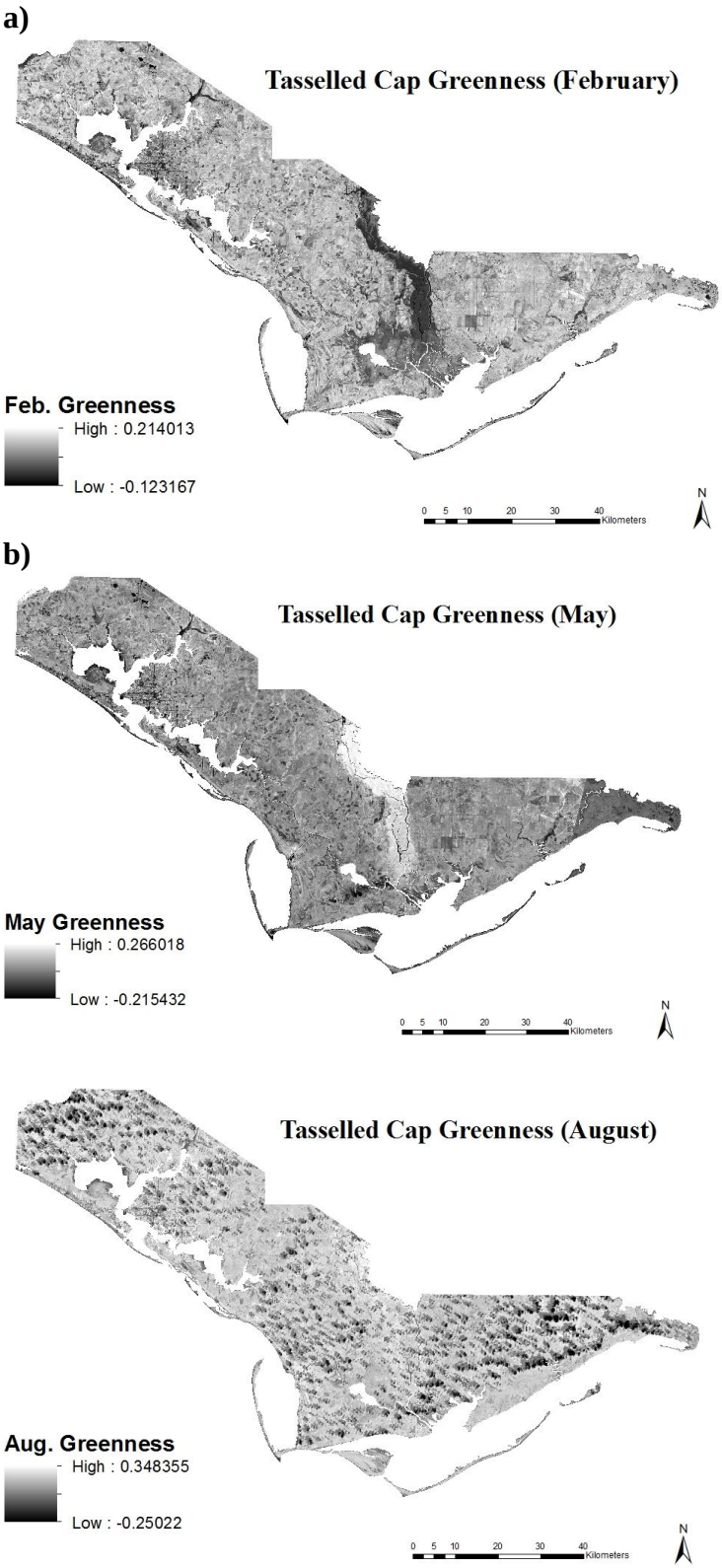


Figure 24. Maps depicting tasselled cap greenness in a) February, b) May, and c) August 2014.

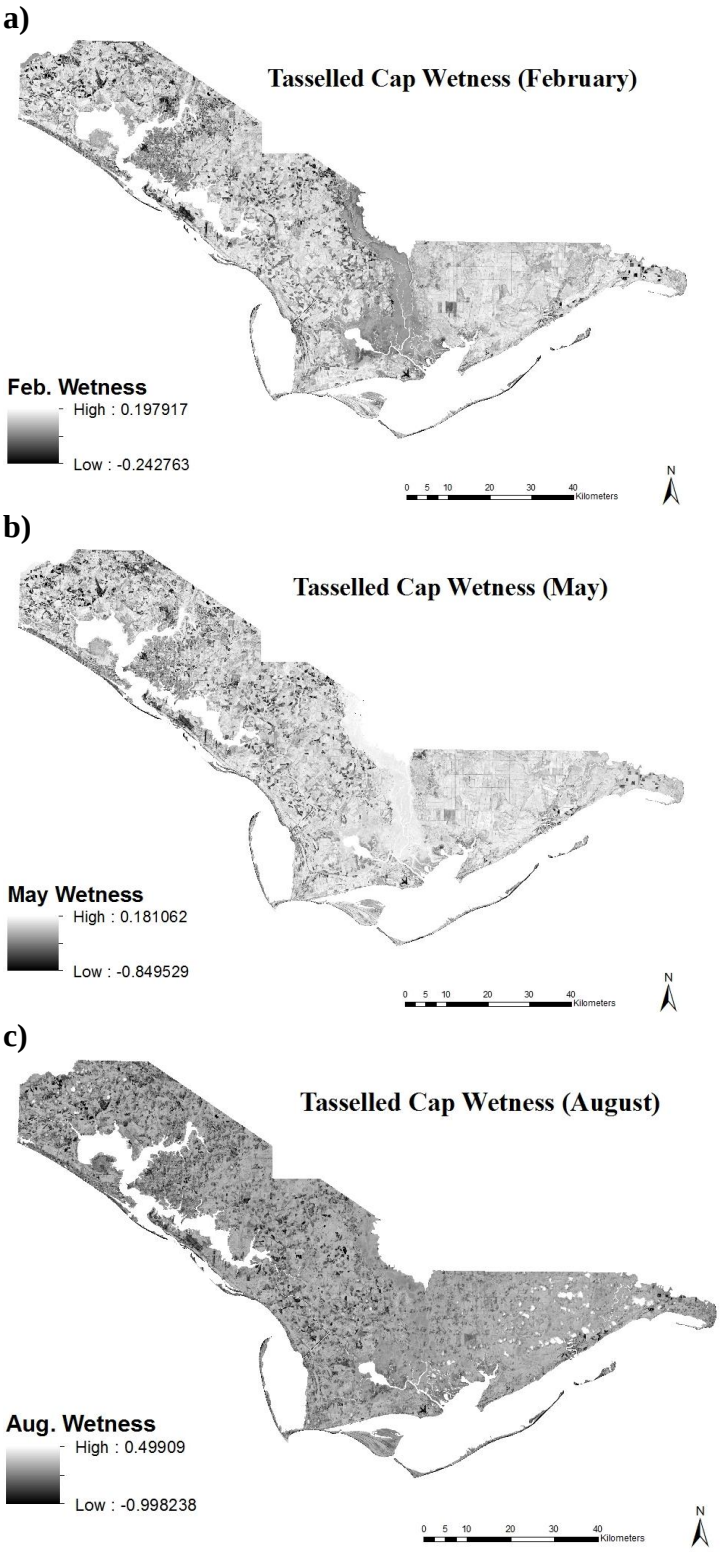


Figure 25. Maps depicting tasselled cap wetness in a) February, b) May, and c) August 2014.

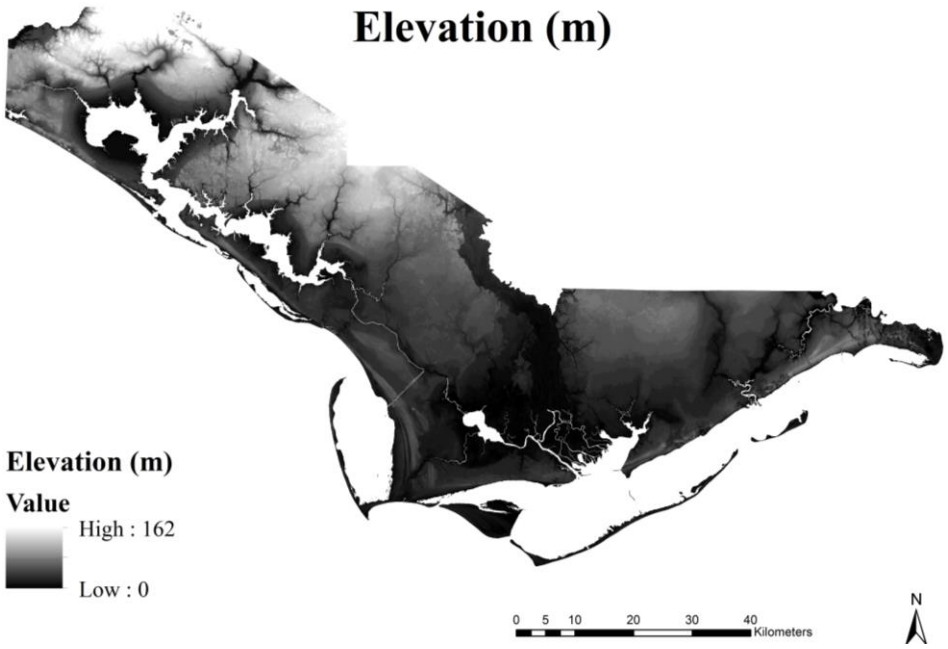


Figure 26. Digital elevation model for the study area.

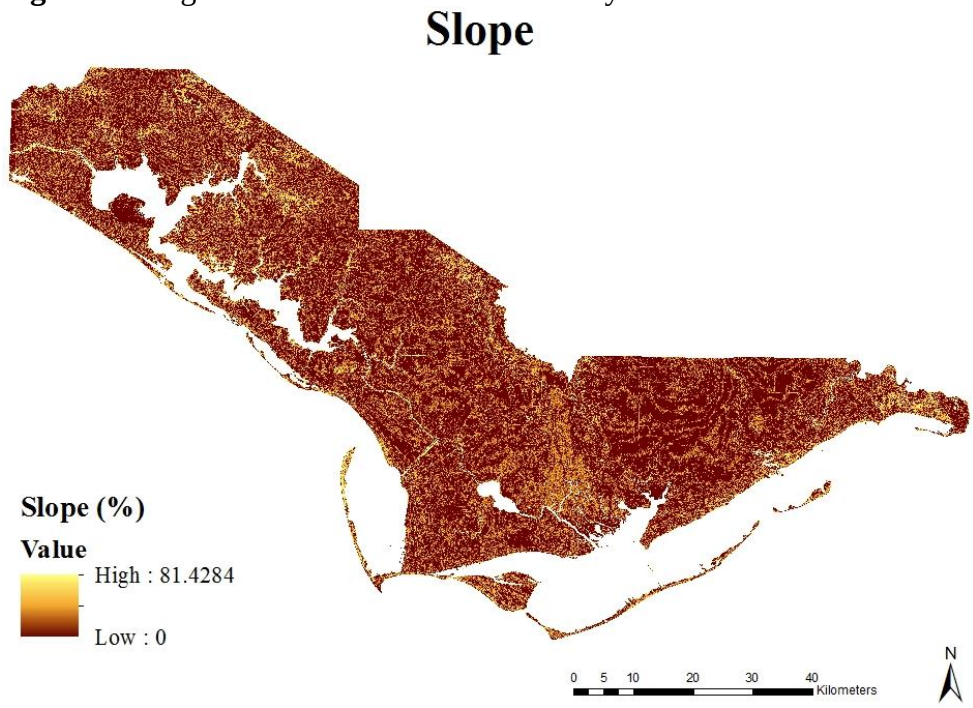


Figure 27. Map depicting percent slope in study area.

APPENDIX D – LIST OF PREDICTOR VARIABLES

Table 12. List of predictor variables used in this study.

Predictor variable name	Predictor variable abbreviation	Scale	Source
Land use/Land cover	LULC	10 m	Florida Natural Areas Inventory (FNAI 2014)
Soil type	Soil type	10 m	Natural Resources Conservation Service (NRCS 2013)
Rock type	Rock type	unknown	(Florida Geological Survey, 2001).
Average February precipitation	Feb. precip.	1 km	Worldclim.org
Average May precipitation	May precip.	1 km	Worldclim.org
Average August precipitation	Aug. precip.	1 km	Worldclim.org
Average February temperature	Feb. temp.	1 km	Worldclim.org
Average May temperature	May temp.	1 km	Worldclim.org
Average August temperature	Aug. temp.	1 km	Worldclim.org
Normalized difference vegetation index (February)	Feb. NDVI	30 m	United States Geological Service (USGS) EarthExplorer (Landsat 8 images, 2015)
Normalized difference vegetation index (May)	May NDVI	30 m	United States Geological Service (USGS) EarthExplorer (Landsat 8 images, 2015)
Normalized difference vegetation index (August)	Aug. NDVI	30 m	United States Geological Service (USGS) EarthExplorer (Landsat 8 images, 2015)
Enhanced vegetation index (February)	Feb. EVI	30 m	United States Geological Service (USGS) EarthExplorer (Landsat 8 images, 2015)
Enhanced vegetation index (May)	May EVI	30 m	United States Geological Service (USGS) EarthExplorer (Landsat 8 images, 2015)
Enhanced vegetation index (August)	Aug. EVI	30 m	United States Geological Service (USGS) EarthExplorer (Landsat 8 images, 2015)
Tasseled cap transformation, brightness (February)	Feb. TCTB	30 m	United States Geological Service (USGS) EarthExplorer (Landsat 8 images, 2015)

Tasseled cap transformation, brightness (May)	May TCTB	30 m	United States Geological Service (USGS) EarthExplorer (Landsat 8 images, 2015)
Tasseled cap transformation, brightness (August)	Aug. TCTB	30 m	United States Geological Service (USGS) EarthExplorer (Landsat 8 images, 2015)
Tasseled cap transformation, greenness (February)	Feb. TCTG	30 m	United States Geological Service (USGS) EarthExplorer (Landsat 8 images, 2015)
Tasseled cap transformation, greenness (May)	May TCTG	30 m	United States Geological Service (USGS) EarthExplorer (Landsat 8 images, 2015)
Tasseled cap transformation, greenness (August)	Aug. TCTG	30 m	United States Geological Service (USGS) EarthExplorer (Landsat 8 images, 2015)
Tasseled cap transformation, wetness (February)	Feb. TCTW	30 m	United States Geological Service (USGS) EarthExplorer (Landsat 8 images, 2015)
Tasseled cap transformation, wetness (May)	May TCTW	30 m	United States Geological Service (USGS) EarthExplorer (Landsat 8 images, 2015)
Tasseled cap transformation, wetness (August)	Aug. TCTW	30 m	United States Geological Service (USGS) EarthExplorer (Landsat 8 images, 2015)
Elevation	Elevation	10 m	National Elevation Dataset (NED) provided by USGS
Slope	Slope	10 m	National Elevation Dataset (NED) provided by USGS
Distance to ocean	Distance to ocean	N/A	Calculated in ArcMap 10.2
Distance to roads	Distance to roads	N/A	Calculated in ArcMap 10.2
Distance to wetlands	Distance to wetlands	N/A	Calculated in ArcMap 10.2

APPENDIX E – R CODE:

```
#########################################################

#####Variance Inflation Factor (VIF)#####

## installs usdm package ##
install.packages("usdm")
library(usdm)

##load model data into R##
model.data<-read.csv("Directory_where_data_are_located")

##systematically calculates VIFs and removes variables with a threshold greater than 10##
v1<-vifstep(model.data,th=10)

##view VIFstep output##
v1

#####Telephus spurge Presence/Pseudo-Absence Boosted Regression Tree Model#####

## installs gbm package ##
Install.packages("gbm")
library(gbm)

## load the gbm and brt.functions.R code into R ##
setwd("C:/Directory_containing_functions_from_Elith_et_al.")
source("brt.functions.R")

## tells R where your data are on disk ##
model.data <- read.csv("C:/directory containing data spreadsheet.csv")

spurge.tc5.lr001 <- gbm.step(data=model.data,
   gbm.x = 6:29,
   gbm.y = 4,
   family = "bernoulli",
   tree.complexity = 5,
   learning.rate = 0.001,
## identifies the optimal number of trees or iterations, then fits a model ##
   bag.fraction = 0.5)


##### Investigating Model #####

## plots partial dependence plots of predictor variables ##
par(mfrow=c(6,6))
```

```
gbm.plot(spurge.tc5.lr001, n.plots=23, write.title = F)

## investigates interactions between predictor variables ##
find.int <- gbm.interactions(spurge.tc5.lr001)
find.int$rank.list
find.int$interactions

## returns most relevant model evaluation statistics (AUC=discrimination.mean) ##
spurge.tc5.lr001$cv.statistics


##### Predicting the Model to a Map #####

## tells R where the fishnet grid data are on disk ##
OutFishnet <- read.csv("C:/Directory_containing_fishnet_grid_csv", header=TRUE, sep=",")

##tells R to predict BRT model to the fishnet grid##
Predict <- predict.gbm(spurge.tc5.lr001, OutFishnet, n.trees =
Insert_optimal_number_of_trees_from_model_output, type="response")

## executes fitted BRT model creating a new Excel file with probabilities of occurrence##
write.table(Predict, "C:/Directory_where_response_file_should_be_saved", sep = " ",
row.names=TRUE, col.names=TRUE)


#####Investigate BRT and MaxEnt model accuracy######

##tells R which directory the observation (presence/absence) and prediction data are in##
ObsProb_70m<-read.csv("C:/Directory_where_Excel_file_is_located", header=TRUE)

##calculates the optimal threshold that maximizes sensitivity + specificity##
optimal.thresholds(ObsProb_70m,threshold=101,which.model=1,opt.methods=3)

##calculates a confusion matrix for a single model using the optimal threshold##
CMX1<-cmx(ObsProb_70m,threshold=0.8125,which.model=1,na.rm=FALSE)

##calculates sensitivity using the confusion matrix##
sensitivity(CMX1,st.dev=TRUE)

##calculates sensitivity using the confusion matrix##
specificity(CMX1,st.dev=TRUE)
```