Modeling Naturally Occurring Wildfires Across the US using Niche Modeling

by

Brandon Polk

B.S., Southern Illinois University, 2014

A Thesis
Submitted in Partial Fulfillment of the Requirements for the
M.S, Geographic Information Science.

Department of Geography and Environmental Resources
in the Graduate School
Southern Illinois University Carbondale
December 2016

TABLE OF CONTENTS

VITA (Optional)

LIST OF FIGURES

LIST OF TABLES

TABLE

PAGE

LIST OF EQUATIONS

TABLE

PAGE

LIST OF ABBREVIATIONS

ROC = Relative operator characteristic

GLM = General Linearized Model

RF = Random Forests

ANN = Artificial Neural Network

MARS = Multiple Adaptive Regression Splines

TSS = True Skill Statistic

LR = Logistic Regression

VIF = Variance Inflation Factor

CSI = Critical Success Index

SR = Success Ratio

FAR = False Alarm Rate

AN ABSTRACT OF THE THESIS/DISSERTATION OF

Brandon Polk, for the Master of Science degree in Geographic Information Science, presented on November 2, 2016, at Southern Illinois University Carbondale.

TITLE: MODELING NATURALLY OCCURRING WILDFIRES ACROSS THE US USING NICHE MODELING

MAJOR PROFESSOR: Dr. Guangxing Wang

Wildfires can cause significant damage to an area by destroying forested and agricultural areas, homes, businesses, and leading to the potential loss of life. Climate change may further increase the frequency of wildfires. Thus, developing a quick, simple, and accurate method for identifying key drivers that cause wildfires and modeling and predicting their occurrence becomes very important and urgent. Various modeling methods have been developed and applied for this purpose. The objective of this study was to identify key drivers and search for an appropriate method for modeling and predicting natural wildfire occurrence for the United States. In this thesis, various vegetation, topographic and climate variables were examined and key drivers were identified based on their spatial distributions and using their correlations with natural wildfire occurrence. Five models including General Linearized Models (GLM) with Binomial and Poisson distribution, MaxEnt, Random Forests, Artificial Neural Networks, and Multiple Adaptive Regression Splines, were compared to predict natural wildfire occurring for seven different climate regions across the United States. The comparisons were conducted using three datasets including LANDFIRE consisting of thirteen variables including characteristics of vegetation, topography and disturbance, BIOCLIM containing climate variables such as temperature and precipitation, and composite data

that combine the most important variables from LANDFIRE and BIOCLIM after the multicollinearity test of the variables done using variance inflation factor (VIF).

This results of this study showed that niche modeling techniques such as MaxEnt, GLM with logistic regression (LR), and binomial distribution were an appropriate choice for modeling natural wildfire occurrence. MaxEnt provided highly accurate predictions of natural wildfire occurrence for most of seven different climate regions across the United States. This implied that MaxEnt offered a powerful solution for modeling natural wildfire occurrence for complex and highly specialized systems. This study also showed that although MaxEnt and GLM were quite similar, both models produced very different spatial distributions of probability for natural wildfire occurrence in some regions. Moreover, it was found that natural wildfire occurrence in the western regions was more influenced by precipitation and drought conditions while in the eastern regions the natural wildfire occurrence was more affected by extreme temperature.

CHAPTER 1

INTRODUCTION

1.1 Statement of Problem

Wildfires can cause significant damage to an area. Wildfires typically destroy forested and agricultural areas but left unchecked they can destroy homes and businesses, and even a potential loss of life. Climate change may lead to an increased frequency of wildfires in the United States (Yonggiqang 2013). As the population continues to grow, wildfires may affect more people in the future, even if climate change does not impact the frequency of wildfires. A quick, simple, and accurate method for modeling wildfire occurrence would aid in land use planning.

Wildfire hazard potential maps are available in the United States from the Forest Service's Rocky Mountain Research Station https://www.firelab.org/project/wildfire-hazard-potential. The LANDFIRE data is paired with wildfire presence points and Fsim, a large fire simulator (Thompson 2013). This approach may not be the most appropriate for modeling wildfires, due to the accuracy of simulators. Additionally, the causes of wildfire are different. Most wildfires are related to human development while some are naturally occurring wildfires (Genton 2006).

A method that can be used to develop probability maps and identify key drivers for any type of raster data would be beneficial. The LANDFIRE data only exists for the United States so a simple method that works with any type of raster data would also be beneficial for international countries, particularly countries without the financial resources to have a significant wildfire monitoring agency. As the causes of wildfires

are different, a method that enables probability maps based on different causes of wildfires could also be beneficial. For example, a map depicting the probability of wildfires due to cigarettes or campfires may be beneficial for more effective management practices. Wildfire probability maps and drivers for naturally occurring wildfires may be beneficial for understanding wildfires in context of climate change.

Perhaps a Bayesian approach that can be paired with any type of raster data would be beneficial. Overly complex models often require highly specialized talent and additional resources that many local agencies do not have. This would simple enough that most agencies could develop their own predictive maps based on conditions that the agency deems important. Using GIS technology and open source programming languages like R, this approach can be made even easier and is cost efficient.

1.2 Purpose Statement

This study will identify key drivers of natural wildfires (grasslands and forest fires) and create naturally occurring wildfire probability maps in the United States for the NOAA Climate Regions. While several techniques and models exist, niche modeling has been successfully applied to mapping potential wildfire occurrence. Niche modeling techniques include both regression based techniques and machine learning techniques. This study will compare two niche modeling techniques, General Linearized Models (GLM) and a machine learning technique called MaxEnt to assess their performance. Additionally, other algorithms will be run to compare how similar they are to Logistic Regression (LR), Poisson distributions, and MaxEnt models.

LR is commonly used for prediction of wildfire occurrence. The reason that LR is popular is that LR can accept continuous/discrete data and produces good results

quickly. This means that LR can be used on a large percentage of real world data, including wildfires. MaxEnt is a machine learning algorithm that can identify key drivers and obtains a specific probability distribution function, namely the one with the maximum amount of Shannon's information entropy. Like LR, MaxEnt is simple to use, able to accept continuous and discrete data, and produces accurate results quickly. In addition, it does not assume normality for the independent variables. Rather, a maximum entropy model produces a normally distributed joint distribution. As *Geman 2015* states, "*The joint distribution g(~x) of asset returns is multivariate Gaussian N (μ, ∑). Assuming normality is equivalent to assuming g(~x) has maximum (Shannon) entropy among all multivariate distributions with the given first- and second-order statistics μ and ∑*".

Recent studies have shown that the MaxEnt software is a very effective tool when compared to other models, including LR. MaxEnt has already shown to perform very well against LR in two wildfire studies (Massada 2012, De Angelis 2015). This study will use the BIOMOD2 package https://cran.rproject.org/web/packages/biomod2/index.html to evaluate the predictive performance of LR and MaxEnt using the same dataset and sampling design.

Another important difference concerns multicollinearity between driver variables. Multicollinearity does not hinder model performance in MaxEnt, but does hinder model interpretation (Estrada-Pena et al 2013). For this study, correlated variables will be removed to reduce errors in model interpretation and should enable a fairer comparison against LR. Even though MaxEnt is a machine learning algorithm and LR is a statistical technique, the above considerations will provide an adequate comparison.

1.3 Research Questions

1.) Does the MaxEnt model predict natural wildfire distributions successfully in each of the NOAA Climate Regions?

2.) How does this model compare with logistic regression and other models?

3.) What are the main drivers (predictor variables) of natural wildfires in each of the NOAA Climate Regions?

1.4 Definition of Terms

There are many definitions of **entropy**, depending on the context in which it is used. In the physical sense, entropy can be described as the logarithm of the amount of combinations of atoms that makes up a macroscopic object. For the purpose of this study, it would be better to think of entropy as the amount of unknown information within a system. **Shannon's entropy** is the entropy of a probability distribution and is the expected value of the information of the probability distribution. The **principle of maximum entropy** states that, subject to precisely stated prior data, the probability distribution which best represents the current state of knowledge is the one with largest entropy. **MaxEnt** refers to the software developed by Steven Phillips which uses Jayne's principle of maximum entropy to calculate the probability distributions of geographic phenomena using occurrence data (a simple csv file with x and y coordinate information) and background data (any kind of continuous or discrete variable that can be put into a raster format). Jayne's principle of maximum entropy has also been called MaxEnt. To avoid a confusion of terms, MaxEnt will refer to the software while the principle of maximum entropy will refer to the principle developed by Jaynes.

CHAPTER 2

LITERATURE REVIEW

2.1 Theoretical Perspective

Wildfires (and other geographic phenomena) are typically modeled using digital information to analyze data in what is called a geographic information system (GIS). To this end, an understanding of how information systems evolve is helpful. Entropy can be used to describe the log of the combination of atoms that describe a macroscopic state within a thermodynamic system, but it can also be used to describe the weight of uncertainty within an information system. Both refer to unknown processes occurring within the system.

Given the constraints, a system will seek the state with the maximum amount of entropy possible. This has due with the efficiency of a system (or machine). The more inefficient the system is (or the less knowledge we have about the information system), the more entropy is generated. Only a machine (system, Laplace's Demon) that is 100% efficient is reversible, because no entropy is generated. Carnot's cycle, equivalent to the 2nd Law of Thermodynamics, shows this. More constraints provide less degrees of freedom for the input of an open system and represents a greater amount of knowledge about the system.

The less the constraints on the system, the greater the amount of entropy (chaos/unknown information) the system can produce. A system with very few constraints would be the most chaotic because, from an information perspective, very little would be known about the system. If nothing is known about a system, then this becomes an example of Laplace's Principle of Indifference. The most uniform

distribution (equal probability to all events) is assigned to the whole distribution because maximum ignorance is assumed.

It is impossible to know everything about a system. If everything is known about the system, then no unknown processes or events can occur. From air mixing around, for example, we can see that the system is heading towards uniformity even though there is increasing disorder in the system. We do not need to know all the processes that are occurring to know that the system is seeking uniformity.

The point is, if it is justifiable to assign the most uniform distribution to unknown processes (Laplace's principle) and it is assumed that these processes seek the most uniform distribution (2nd Law) then it is justifiable to assign the most uniform distribution to unknown information, even when systems have known information. This is exactly what the MaxEnt algorithm does. It finds the most uniform distribution of information, given the constraints or known information (Grendar 2001). In 1957, E.T. Jaynes addressed this in a seminal work for information theory. (Jaynes 1957)

2.2 Fire Studies and Models

Is niche modeling an appropriate approach for modeling wildfire suitability? After all, there are several methods for modeling wildfires throughout the world. One commonly used simulator is the PHOENIX RapidFire software. PHOENIX RapidFire is a wildfire simulator that is designed for large, fast moving fires in Australia that incorporates fuel types, topography, and weather to produce results. The model has also been applied successfully to other areas, such as the southern portion of France, which has a high amount of Wildland-Urban Interface (Pugnet 2013). RapidFire is a

simulator, designed to predict the spread of wildfires and is not useful for predicting the occurrence of wildfires.

Regression models have been used in many different countries. One study compared three types of regression models (multiple linear regression, log-linear regression, and gamma-generalized linear regression) in the Tahe forest region of the Daxing'an Mountains in China, considering nine different meteorological variables (Guo 2015). Another study used LR in the Heilongjiang Province, China and specifically states the reason is because it was "reasonably flexible and accepts a mixture of continuous and categorical variables, as well as non-normally distributed variables (Chang 2013). The study evaluated performance through the Relative Operator Characteristic (ROC) which was 0.906. Regression models can be helpful in prediction of wildfire occurrence, and is considered a niche modeling technique.

More recently, machine learning algorithms have been used for the prediction of wildfire occurrence, and machine learning techniques are also considered a type of niche modeling technique. One study compared three types of machine learning algorithms (Random Forests (RF), Support Vector Machines, and Boosted Regression Trees) against LR in the forested areas of Spain (Rodrigues 2013). The results could show that all three of the machine learning algorithms considered produced higher scores than LR, though RF had the highest scores.

There are many GIS applications useful for modeling different aspects of wildfires. FARSITE is widely used by the U.S. Forest Service, National Park Service, and other federal and state agencies for simulating the spread of wildfires. It can simulate wildfire growth based on topography, fuel, wind, and meteorological variables

(Finney 2004). In addition to other features, FARSITE accounts for spotting. Due to its complexity, generally those who have extensive training in wildfire behavior are the only ones able to utilize its capabilities for making fire and land management decisions. Like PHOENIX RapidFire, it is a simulator and not helpful for the prediction of wildfire occurrence.

FlamMap, described in *An Overview of FlamMap Fire Modeling Capabilities* is also used by several agencies including the U.S. Forest Service and the National Park Service (Finney 2006). FlamMap is an application that is complementary to FARSITE. Because it models fires assuming static environmental conditions, it is unable to perform temporal projections. Even though FlamMap is somewhat less complex than FARSITE, it is still highly complex software. Like FARSITE, only officials with proper fire training and experience are recommended to use FlamMap in order to make fire and land management decisions.

There are other types of wildfire modeling. FOFEM (First Order Fire Effects Model) is designed for predicting tree mortality, fuel consumption, smoke production, and soil heating caused by wildfires (Lutes 2014). The software has many features, including the ability to produce graphs and reports for the above categories, default fuel loadings for different cover types, and batch processing. If modeling the effects of wildfire is important to land use planning, then FOFEM may currently be the most appropriate software available.

FEAT/FIREMON Integrated (FFI): Ecological Monitoring is an ecological monitoring system that is used to fulfill monitoring requirements and is designed to assist officials with the collection, storage, and analysis of ecological information (Lutes

2012). It includes software components for data entry, data storage, summary reports, and analysis tools. FFI also has extensive Microsoft SQL support for database entry. An optional GIS module allows the tools to be incorporated into ArcGIS.

Despite the wealth of wildfire applications available, there is not an application that is specifically designed to produce accurate wildfire distributions quickly and easily. Because of this, many planners have used techniques designed for general inference. Therefore, niche modeling techniques like LR are commonly used. This researcher has drawn the conclusion that niche modeling techniques are a good choice for producing probability maps quick and simple, given the literature review.

2.3 Principle of Maximum Entropy Overview and History

MaxEnt has strong roots in Information Theory and is based on the principle of maximum entropy, a fundamental concept concerning the evolution of information. In the 1940's, Claude Shannon developed a formula for calculating the amount of information or uncertainty within a system, apparently unaware of Boltzmann's formula for entropy. Tribus (1971) accounts Shannon to have stated "*My greatest concern was what to call it. I thought of calling it 'information', but the word was overly used, so I decided to call it 'uncertainty'. When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, 'You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, nobody knows what entropy really is, so in a debate you will always have the advantage*."

Later, Jaynes released an important article that linked Shannon's information entropy to thermodynamic entropy (Jaynes 1957). According to Jaynes, thermodynamic

entropy should be seen as a specific application of Shannon's information entropy. Thermodynamic entropy is interpreted as being proportional to the amount of further Shannon information needed to completely define the physical state. Shannon's information entropy can be used to describe the weight of information in the system.

Jaynes also describes the principle of maximum entropy as an extension to Laplace's principle of insufficient reason, which teaches that there is no justification for assigning a particular probability distribution over another if given the same amount of information. What Jaynes realized was that Laplace was assuming maximum ignorance and therefore was assigning the most uniform distribution, which is the distribution with the highest amount of entropy. He postulated that the justification for assigning a particular probability distribution was that the distribution had the highest amount of entropy.

Laplace's principle describes a system assuming maximum ignorance but what if, like most real world systems, there is some known information about the system but also some unknown information? Jaynes extended Laplace's principle by using the known information to constrain the unknown information. The unknown information is then assigned the most uniform distribution possible, given the constraints. As this distribution, would be the most uniform, it would also be the distribution with the maximum amount of entropy and is therefore justified.

This principle has been utilized for a wide range of scientific fields including image processing and enhancement, biology, and natural language processing. For example, one medical study showed that the principle of maximum entropy can significantly reduce echo-planar correlated spectroscopic imaging, which can help to

reduce scan time and is very important if a patient is in critical condition (Burns 2013). The studies presented here will focus on the contribution the principle has made to geographic information science with a focus on natural hazards.

2.3.1 Applications of MaxEnt to GIS

In 2004, Steven Phillips developed GIS software for prediction of species distributions called MaxEnt from a machine learning perspective (Phillips 2006). As MaxEnt was originally applied to species distribution models and machine learning techniques were often outside the realm of ecologists, he teamed with several authors to describe MaxEnt from a statistical point of view (Elith 2011). The software quickly rose to become one of the most popular species distribution models because of its accuracy and its efficiency. Merow (2013) said that there were over 1000 articles written concerning MaxEnt and species distribution models between 2006 and 2013 and that the reason for its popularity is that it typically outperforms other models and is very easy to use.

Since MaxEnt is a binary classification technique, the software has been used successfully to model other forms of geographic phenomena than species distributions. There are many recent studies utilizing MaxEnt for obtaining probability distributions for different forms of natural hazards including landslides, viral and bacterial outbreaks, and wildfires. MaxEnt has also been applied to land use planning and human settlement, land classification, and streambank erosion potential. One study, for example, used MaxEnt to evaluate optimal transitional areas due to altitude for human settlement in the Qinghai–Tibet Plateau (Cao 2013). The software could find suitable areas along the Huangshui River. Another study used MaxEnt to model streambank erosion potential

with an Area Under the Curve (AUC) of .994 (averaged over 30 iterations) and show that slope, soil type, and bank stress index were the most important variables contributing 32.7, 29.2, and 20.6 respectively (Pitchford 2015). Many studies have used MaxEnt to map bacterial and viral outbreaks. Goka (2013), for example, utilized the MaxEnt software to identify key drivers of avian influenza in Japan. It was determined that the dabbling duck population has a significant role in the spread of influenza in Japan.

MaxEnt has been successfully applied to evaluating landslide risk. Converntino (2013) could show that areas that have medium-high elevation and slope, small-medium hillslope-to-channel distance, or medium erodibility are more susceptible to landslides in the future in the Arno Basin. A similar study evaluated landslide risk in Korea using different variables (Kim 2015). It could show that urban and agricultural areas are at higher risk than other land use types.

2.3.2 Applications of MaxEnt to Wildfires Distributions and A Review of Potential Drivers

The MaxEnt software has been successfully applied to wildfire distributions in many parts of the world. In the study by Arpaci (2014), MaxEnt is compared against another machine learning algorithm, RF, in a section of the Alps mountains using multiple meteorological variables to derive fire weather indices. MaxEnt is shown to perform quite well when compared to RF. The results also showed that both models tended to agree about the most important drivers but diverged when considering drivers of lesser importance.

In the study of De Angelis (2015), MaxEnt was compared against the traditional General Linearized Models (GLM) with binomial distribution and logistic link (same as

LR) in Switzerland. Comparisons between the models were made using meteorological variables, fire weather indices, and a combination of both. The study considered both anthropogenic and naturally occurring wildfires. The top ten models (over several iterations) were MaxEnt. This study gives evidence that MaxEnt outperforms LR, in application.

In the study of Peters (2013), the Wildland Urban Interface (WUI), the Modified Palmer Drought Index (PMDI), and the integrated moisture index were examined for their potential to predict wildfire occurrences in by using a MaxEnt model. Of the variables considered, the WUI and the PMDI were influential variables. It is also revealed from the study that, approximately in August, (Figure 5 of the study) WUI contributes significantly less to wildfire occurrences while PMDI becomes the overwhelming factor, representing roughly 85% of variable contribution. Also, the figure showed that when PMDI variable contribution was high, WUI variable contribution was low.

Batllori (2013) used MaxEnt to model wildfires using BIOCLIM variables, however the study did not compare the model to any other model, such as LR. The article also discusses a key finding, which suggests that areas that show marginal increases in moisture content may show a greater risk. The variables identified in this study will be one of the datasets used.

Massada (2012) compared Random Forests, Generalized Linear Models, and MaxEnt using several predictor variables related to topography, human settlement, and infrastructure. Human settlement and infrastructure were the strongest predictors of wildfire ignitions, though land cover and topography seemed to be strong predictors of

naturally occurring wildfires. MaxEnt was shown to be the most effective, though not by much. The authors also state that there was relatively low variability between the predictor variables in the dataset which may account for similar results among the models. Human structure and infrastructure could be difficult to obtain and would take some time to collect for the conterminous United States. They will not be included in this study because this study is only considering naturally occurring wildfires.

Massada (2012) used MaxEnt to model the variability of wildfires with a set of explanatory variables that characterized ignition sources, flammable vegetation (i.e. fuels), climate and topography. A full model and a model representing non-anthropogenic wildfires were built. The results showed that there was a wide range of responses to the exploratory variables in different areas throughout the western United States. The six most important variables were temperature of the driest month, the remoteness of the area at a 10,000 ha scale, precipitation of the coldest month, ratio of surface to area at the 1 ha scale, gross primary productivity at the 100-ha scale, and percentage land cover of fuels at the 100-ha scale. These variables may be considered in a future study but collecting this dataset currently would be quite time consuming.

Like the previous study, Chen (2015) tried to identify drivers of wildfires in Northeastern China. In this study, weather over the last three days before ignition (temperature, relative humidity, wind speed, rainfall), forest fuel type of ignition, and topography (altitude, slope and aspect) were analyzed. The Variance Inflation Factor (VIF) was also used in this study to determine correlations between the variables and remove highly correlated variables. Among the variables considered, number of

strikes on the day of the ignition, rainfall in the 3 days before, and the intensity of the lightning current seemed to be the most important. That these three variables are the main drivers of lightning caused fires should not be surprising, though they may not be helpful in any long term land use planning. Average wind speed, slope, and altitude did not seem to have much impact in the study area.

LANDFIRE is a collection of several geospatial datasets representing vegetation, wildland fuel, and fire regimes across the United States. The LANDFIRE dataset is a collection of fine scale (30 meters) raster data for several variables related to vegetation and topography. LANDFIRE data can be used in multiple applications, including the applications presented in this study (Opperman 2013). There is also several optional tools that can be downloaded and loaded into ArcMap for LANDFIRE data including the Wildland Fire Assessment Tool, the LANDFIRE Total Fuel Change Tool, the Multi-Raster Classification Tool, and the LANDFIRE Data Access Tool. Thirteen variables from the LANDFIRE dataset will be tested for multicollinearity and used in this study, described in the “Independent Variables” subsection of the “Methodology” section.

One last note considering the literature review. It is important to note that it is not being suggested that MaxEnt is always the most accurate method. For example, Ordóñez (2012) used Generalized Linear Spatial Models (GLSM) and originally considered multiple variables. The AUC was used to determine important factors and were tested for correlation. The dataset was reduced to six factors including percentage of agricultural land, number of dry storm days, mean altitude, number of lightning strikes with a particular cell, percentage of woodland, number of lightning strikes within a broadleaf woodland. A GLM of these variables produced a 73% AUC.

The data was then used on a General Linear Spatial Model (GLSM) and produced a very high AUC of 99%. GLSMs are also a good modeling choice. Of course, it is one study and the ROC may give different results. In any event, MaxEnt typically outperforms logistic regression and several other models but GLSMs may possibly outperform MaxEnt. It is also important to note that this study will evaluate model performance based primarily on the ROC, vs the AUC.

In general, if sufficient information is available on absence then a presences/absence technique is recommended (Elith 2012). Wildfire data, like a lot of real world data, often does not contain information about absence so a presence only technique would be beneficial. A good method that can generate accurate pseudo absence data would also be beneficial. Often, errors are introduced in creating the data, due to sampling techniques and whether the data points represent true absence.

CHAPTER 3

METHODS

3.1 Formulas

The formula for LR is

$$\log\left(\frac{\mu}{1-\mu}\right) = \hat{\beta}_0 + \sum_{j=1}^{p} X_j \beta_j + \varepsilon \quad (1)$$

where μ is the probability of class 1 (meaning occurrence) and 1 - μ is the probability of class 0, β_o is the model intercept, β_j are the regression coefficients, p is the number of independent variables X and ε represents the residuals or errors.

The MaxEnt formula is an extension of Bayes' rule and takes the form:

$$Pr(y=1|z) = f1(z)Pr(y=1) / f(z) \quad (2)$$

Where 1 means presence of natural wildfire, z represents a vector of environmental variables (rasters), and f(z) is the probability density of covariates across the location. Since Pr(y=1) is the term that is lacking, prevalence is not available from presence only data. We need prevalence to calculate conditional probability precisely. A logistic

transformation is done, and calibrates the intercept so that the implied probability of presence at sites with typical conditions, given the parameter τ. If we knew the exact value of τ it would solve the non-identifiability of prevalence.

The MaxEnt formula ends up being an extension of Bayes Rule

$$\Pr(y = 1|z) = \tau * e^{\eta(z) - r} / (1 - \tau + \tau * e^{\eta(z) - r}) \quad (3)$$

where n(z) is a linear score, r is the relative entropy of MaxEnt's estimate of f1(z) from f(z), and τ is the probability of presence at sites with "typical" conditions for the species (Elith 2011).

From equation 1, *we see that a simple approach to estimate Pr(y = 1|z) would be to simply multiply $e^{n(z)}$ by a constant that estimates prevalence; this approach has the disadvantage that $e^{n(z)}$ can be arbitrarily large, which implies that we may get an estimate of Pr(y = 1|z) that exceeds 1. Exponential models can be especially badly behaved when applied to new data, for instance, when extrapolating to new environments. To avoid these problems, and to side-step the non-identifiability of the species prevalence, Pr(y = 1), MaxEnt's logistic output transforms the model from an exponential family model to a logistic model* (Elith 2011).

3.2 Study Area:

The study area will be the United States, divided according to NOAA climate

regions, for the year 2010. This year was chosen because fairly reliable data exists for all of the possible drivers. The climate regions of the landscape were chosen

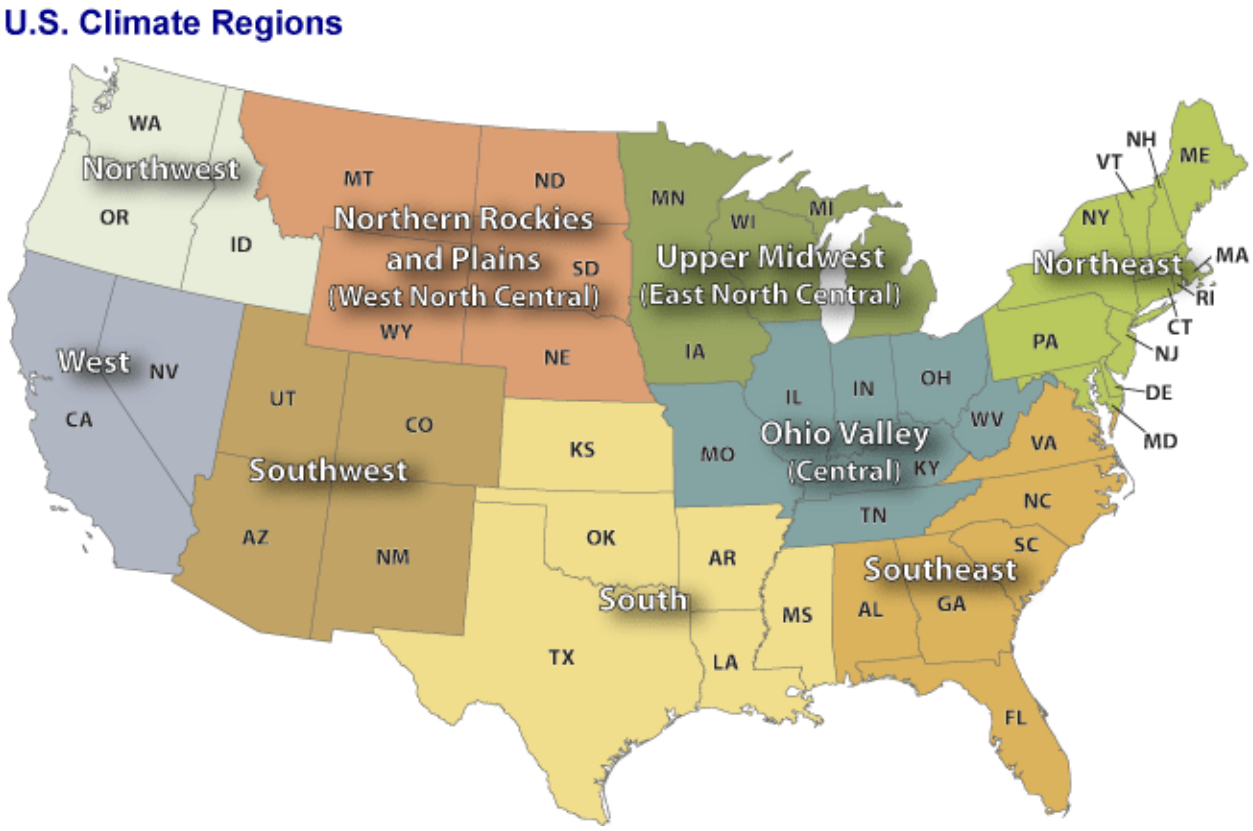

Figure 1:NOAA CLIMATE REGIONS

because it divides regions according to fairly homogenous climates. Figure 1 shows the study area, as broken up according to the NOAA climate regions. The Northeast and East North Central regions are not included due to the limited data and low frequency of occurrence.

The Southeast Region compromises the states of Alabama, Florida, Georgia, North Carolina, South Carolina, and Virginia. Multiple wildfires occur in this area in both forested and non-forested areas. In terms of area burned, this region contains 4 of the top ten wildfire states (North Carolina, Alabama, Georgia, and Florida).

The Central Region compromises the states of Illinois, Indiana, Kentucky, Missouri, Ohio, Tennessee, and West Virginia. While the area has some large forested areas, including Shawnee National Forest, not a lot of forest fires occur in the area. Naturally occurring grassland wildfires, however, happen and most often impact agriculture. Missouri is 5th in the nation in terms of number of reported fires, while other states have significantly less reported fires. In terms of acres burned, Illinois Indiana, Ohio, and West Virginia are almost negligible. Missouri had the most acreage burned by wildfires for the region as well, at about 30,000 acres.

The South Region compromises the states of Arkansas, Kansas, Louisiana, Mississippi, Oklahoma, and Texas. Many wildfires occur in the region. Texas has highest amount of fires in the nation and is seventh highest in the amount of acres burned. Mississippi come in as tenth in the nation in the amount of fires burned. Oklahoma comes in ninth in terms of acres burned. The region saw over 450,000 acres burned in 2015.

The Southwest Region is comprised of the states of Arizona, Colorado, New Mexico, and Utah. This region sees extensive wildfire damage. Every one of these states is in the top ten most wildfire prone states.

The West Region is comprised of the states of California and Nevada. This region sees extensive wildfire damage. California is the most wildfire prone state in the nation, while Nevada is the tenth most wildfire prone. California comes in second in the number of fires, and third in the number of acres burned.

The Northwest Region contains the states of Washington, Oregon, and Idaho. Like the Southwest and West, this region has extensive damage related to wildfires.

All of the states in this region are in the top ten most wildfire prone states. Washington is second in the nation in number of acres burnt by wildfire. Idaho and Oregon are fourth and fifth, respectively.

The West North Central Region is comprised of the states of Montana, Nebraska, North Dakota, South Dakota, and Wyoming. Montana is the 6th in the nation in terms of the amount of acres burned, and 7th in the nation in terms of the frequency of fires. Nebraska does not have many wildfires. All other states in the region have low to moderate amount of fires. **http://www.iii.org/fact-statistic/wildfires.**

3.3 Variables and Datasets

The dependent variable will be the probability for ignition of naturally occurring wildfires, including grassland and forested fires. The data were obtained from the U.S.G.S, available at http://wildfire.cr.usgs.gov/firehistory/data.html and from all available reporting agencies. It is important to note that, like many other sources of data, the accuracy cannot be verified to 100 percent accuracy. This is especially true in species distribution modeling (Graham 2008). The same could be said about the source of ignition for wildfires.

The independent variables are comprised of similar variables identified in literature. The goal is to find variables that are easy to obtain or derive and have been identified as appropriate to use at a regional scale. Five BIOCLIM variables including Temp Seasonality, Mean temp of wettest month (quarter), Mean temp of warmest month (quarter), Total Annual Precipitation and Precipitation of driest

month/quarter, https://www.climond.org/BioclimRegistry.aspx were used in Batllori 2013 will be considered. BIOCLIM has been applied successfully to other forms of niche modeling such as species distribution modeling prior to being applied to forested wildfires, so there is good reason to indicate they may be good drivers for niche modeling of grassland wildfires as well. LANDFIRE data has also been applied successfully to model wildfires. LANDFIRE consists of 4 forest variables (Forest Canopy Bulk Density, Forest Canopy Base Height, Forest Canopy Cover, Forest Canopy Height), 3 general vegetation variables (Existing Vegetation Cover, Existing Vegetation Height, Existing Vegetation Type), 3 topographic variables (Aspect, Elevation, Slope), and 3 disturbance variables (Disturbance, Fuel disturbance and Vegetation Disturbance). It is logical to assume that the vegetation dataset can also help to predict grassland wildfires. It can be obtained at http://www.landfire.gov/version_comparison.php. The BIOCLIM data can be derived through R, using rasters that represent tmax, tmin, tmean, and precipitation data for the given temporal period. Information on creating BIOCLIM in the “climates” package available for R can be seen at https://rforge.net/doc/packages/climates/bioclim.html, or there is already available data at http://www.worldclim.org/download. The already available data was used for this study. The first set of models will be composed of the BIOCLIM data. The second set of models will be composed of the LANDFIRE data.

Studies have shown that moisture content may not be an important consideration. Peters 2013 did not determine the Integrated Moisture Index (IMI) to be an important variable. Vasilko 2009 revealed that 10-h fuel moisture has a

negative effect on wildfire ignition, along with elevation and aspect (LR, the exception, showed positive influence with aspect).

Batllori 2013 looked at moisture content as a driver of wildfire occurrence in relation to climate change extensively. The study considered two future scenarios (warmer/wetter and warmer/ drier). While it was shown that, in general, “fire probabilities were high, and the lowest values were confined to the wettest and driest areas” it was also shown that “the warmer–drier syndrome led to an overall decrease in fire activity across 56% of the Mediterranean biome by the end of the century (2070-2099), whereas warmer–wetter conditions led to an overall increase across 65% of the biome”. It may be that there is a threshold for moisture. Areas that have lower moisture content but receive marginal increases in moisture content in the future could be at greater risk of wildfire occurrence.

3.4 Model framework

Several models will be run in order to provide an assessment of drivers of naturally occurring wildfires in the United States. Differing datasets will be obtained for each climate division. A thirty-meter dataset is very large given the size of the study area. Given hardware restrictions, the dataset will have to be aggregated. This will reduce the file size and allow the rasters to fit into memory within R. While the amount of information will be reduced, it will be sufficient for analysis. The variance inflation factor will be used to test for multicollinearity and remove correlated variables to reduce the potential set of drivers before the models are run. A second correlation test will be done before running the final model.

A third set of models will be developed, after testing for correlation, from a

composite of the other two models. Each region will therefore have at least 3 models for GLM, MaxEnt, and a couple of other niche models to provide an adequate comparison.

3.5 Instruments

The ArcGIS software and the R programming language will be the primary instruments through which modeling is conducted. ArcGIS provides a number of toolboxes that can be used to perform spatial analysis and develop spatial datasets. The software also allows for easy manipulation of spatial data, and for easy development of custom geoprocessing tasks, and has a well-designed interface for cartographic design. It will be used to perform spatial projections, as well as some data conversions. The R programming language is designed for statistical analysis and modeling. It will be used to model both LR (generalized linear model with binomial distribution and logit link) and MaxEnt models.

3.6 Data pre-processing Procedures

The first step is to obtain the data from the various sources and clean it. There were two types of data, occurrence data and background data. The occurrence data will be cleaned first. As only naturally occurring wildfires within the temporal span are being considered, all other types of fire will be removed. The occurrence data is then sectioned according to the climate regions. These spatial points are then projected into an equal area projection for adequate spatial analysis (Albers North American Equal Area Conic was chosen).

The x and y coordinates from the projected occurrence data will need to be obtained and put into a CSV file with three columns delimited by commas. The first two columns contain the x and y coordinates while third column includes the response variable being studied as a heading (in this case wildfires) and presences/absence location as attributes. If (as the case is with this dataset) the response variable is presence only, this column will only have the number 1 in the column. This describes how the CSV file needs to be set up for use in the BIOMOD2 package. The CSV file needs to be set up slightly different if using the MaxEnt software.

A shapefile was created with the correct projection that contains points for all naturally occurring wildfires within the temporal span of the study (2010). A custom Python script was then developed to get the x and y coordinates from the attribute table of the shapefile, create a DBF file for each region and then write the points to a CSV file.

The background data needed to be edited and manipulated in order to be accepted into a MaxEnt model. The LANDFIRE and BIOCLIM rasters will have to be extracted by a masked feature (the extent of the division) for proper analysis. In ArcGIS, this can be accomplished by using the environmental variable “Snap to Raster”, and defining the “Extent” of the raster while running the “Extract by Mask” tool. Like the occurrence data, all of the background rasters need to be projected to the same (equal area) projection. If using the MaxEnt software, it can accept ASCII or BIL raster file format. Using R along with various packages including RGDAL and BIOMOD2, just about any raster format can be used. The TIF file format was used for this study.

After these steps have been taken, both the occurrence points and the background rasters are in the same projection, all of the background have the same extent and resolution. The occurrence data is a csv format and the background data is in the correct format. The data can now input into BIOMOD2 for analysis. Three different sets of models are run, described in the “Model Framework” section.

3.7 Modeling Procedures

To obtain results from the software, all the user must do is input a csv file containing the x and y coordinates of the dependent variable, the folder containing the raster files representing the independent variables, and optionally, set user defined options. The MaxEnt software automatically develops 10,000 pseudo-absence points for the input data. When using a MaxEnt model with popular statistical and machine learning software like R, however, absence data must be developed by the user. Software packages like “BIOMOD2” make creating absence data easy, and can be used to format the data to run with several machine learning techniques including LR, Random Forests (RF), Artificial Neural Networks(ANN), Multiple Adaptive Regression Splines (MARS) MaxEnt, and others.

Since LR will be compared against MaxEnt, the BIOMOD2 package will be used. The BIOMOD2 package allows for the creation of absence data easily, using different sampling strategies. As the BIOMOD2 package includes a function that formats the data to be modeled for GLMs, Artificial Neural Networks, Random Forests, MaxEnt, and other machine learning and statistical techniques, it will be used to compare the LR model to the MaxEnt model. The package includes global default settings and allows

for individual settings for different models. It allows comparison of many different types of models easily. One setting that was changed for GLM's was that the Bayesian Information Criterion was used instead of the default Akaike Information Criterion. This is because MaxEnt is a Bayesian approach and if GLM's are like MaxEnt then they should use the same information criterion for the same dataset. A total of 45 runs are produced, 9 for each of the 5 types of models. Three different models are run, using 3 different permutations of the pseudo-absence data and evaluation metrics (obtained from splitting the data) is reported. Models are run using 10,000 absence points, as this is the standard MaxEnt output and will enable a fairer comparison. An R script was developed to automate the modeling process. Model projections that are done in BIOMOD2 are often converted into a scale between zero and a thousand and is done for memory saving purposes. According to the BIOMOD2 manual, "0 - 1 probabilities are converted into a 0 - 1000 integer scale. This implies a lot of memory saving. User that want to come back on a 0 - 1 scale latter will just have to divide all projections by 1000".

3.8 Variable and Model Evaluation

All the regions were tested for multicollinearity using the Variance Inflation Factor (VIF). Seven variables from the LANDFIRE dataset were used for each region after testing for multicollinearity. If a variable was shown to have collinearity, it was removed. A multicollinearity test was also done on the composite model to ensure no significant correlations existed when the variables from each dataset were combined.

Each of the aforementioned datasets was randomly split into a sub-set (70%) of dataset for model development and a sub-set of dataset for model validation for each of the regions except for the central region in which the data splitting was conducted based on 50% and 50%. When the testing data was used, ROC was primarily discussed for model performance for each region because it is the main metric used for model evaluation for MaxEnt and is also typically used for performance of GLM models. The ROC value illustrates the performance of a binary classification system by plotting the true positive rate (sensitivity) against the false positive rate (specificity) at different thresholds. Other metrics including the True Skill Statistic (TSS) and Kappa were included in the appendices. The LANDFIRE and BIOCLIM datasets were used to identify important drivers to develop the composite model. All the variables were assumed to be continuous unless they were noted.

The coefficient of correlation, R, between a variable and natural wildfire occurring was used to quantify the strength of the variable. In some cases, R^2 values are also given to show the importance of a variable. Moreover, the values of correlation R between the predicted and observed probabilities for natural wildfire occurring were also utilized for evaluate the performance of model runs for each of the regions when the testing data were used. Models were run on a modern hardware (I7 Skylake, 16GB DDR4 RAM) using several statistical packages. A significant level of 0.05 was used.

CHAPTER 4

RESULTS

All the VIF values of predictor variables were less than 10 in the final models, statistically indicating no significant multicollinearity. The used variables included slope, aspect, elevation, existing vegetation type, disturbance, canopy bulk density, and canopy height, temperature seasonality, mean temperature of wettest quarter, mean temperature of warmest quarter, annual precipitation and precipitation of driest quarter. A couple of models showed some collinearity for 5 BIOCLIM variables used. A multicollinearity test was also done for the composite model to ensure no significant correlations existed when the variables from each dataset were combined. In this study, a large number of the probability maps for predicting natural wildfire occurrence were generated from the model runs, but only few of them were provided in the thesis as examples and most of the maps were available in the Appendices.

4.1 Southeast Region

BIOCLIM

The ROC values of probability maps for predicting natural wildfire occurring from BIOCLIM dataset were consistently above 0.75, except for ANN model that produced a score of 0.57. Temperature seasonality was shown to be a most important factor in all the models. Mean temperature of wettest quarter also seemed to be of moderate importance so it was also included. Figure 1 in Appendix A depicts the BIOCLIM model.

LANDFIRE

The ROC values of the probability maps for predicting natural wildfire occurring from LANDFIRE data were similar to those from BIOCLIM data. There was a similar range of 0.75 to 0.9, though one model produced a ROC value of 0.6. According to the importance, the variables were ranked as aspect, slope, elevation, and canopy bulk density, which were used for running the composite model. Figure 2 in Appendix A depicts the LANDFIRE model.

COMPOSITE DATASET

The VIF values of all the variables were smaller than 10 with mean temperature of warmest quarter having the highest value of 8.2. Visually, temperature seasonality appeared to be the most significant variable to predict the probability of natural wildfire occurring. This was also validated by a greatest R value of 0.8 between the predicted and observed probabilities. Most of the probability maps showed that Florida was at higher risk than all other states in this region. Only did a few of the ANN model runs show that temperature seasonality had less importance with R values of about 0.6. This indicated that extreme temperatures were more important for predicting natural wildfire occurrence than other variables. Climate change will lead to increase of temperature and thus probably increase the risk of natural wildfire occurrence in this state in the future. MaxEnt model resulted in the greatest R value of 0.916, while the ANN model led to the smallest R value of 0.733. A couple of the ANN model runs were unable to converge. Figure 3 of

Appendix A depicts the composite model.

4.2 Central Region

BIOCLIM

In central region, mean temperature of wettest quarter and temperature seasonality were most important drivers based on BIOCLIM dataset and were kept in the composite model. MaxEnt produced higher ROC scores (around 0.869) of the probability maps for natural wildfire occurrence in this region than other models, and several RF model runs performed rather poorly with the score values as low as 0.47. This may be due to the relatively small number of data points used, 8 points for calibration and 8 points for testing based on a 50/50 data split, in the region.

LANDFIRE

Based on the results of probability maps for natural wildfire occurring from the runs of MaxEnt, GLM, ANN and RF model, it was found that canopy height, canopy bulk density and existing vegetation type were most important drivers and kept for the composite model. Overall, this dataset produced higher ROC values than the BIOCLIM dataset and MaxEnt led to the highest ROC values of probability maps for natural wildfire occurring. Although some models had extremely low ROC values, most of them were larger than 0.7.

COMPOSITE DATASET

As MaxEnt is similar to the GLM model with a Poisson distribution and logistic

regression (LR), the probability maps of natural wildfire occurrence were created and compared in this region among MaxEnt, GLM – LR with Poisson and binomial distribution respectively. As examples, the probability maps generated using GLM - LR with binomial distribution are shown in Figure 2. Figure 3 presents the probability maps of natural wildfire occurring using GLM – LR with Poisson distribution and Figure 4 depicts the probability maps using MaxEnt model runs. The comment features are that there were higher and lower probabilities in the south and north parts respectively. The differences in the spatial distribution of the probability could also be seen between the models and even within each of the models. For all the model runs, mean temperature of wettest quarter was used as a continual variable. It was noticed that in central region, some of the spatial distributions of probability for natural wildfire occurring (Figures 2, 3 and 4) were dominated by the spatial patterns of mean temperature of wettest quarter used in the model runs in which clear borders existed (Figure 5). In addition, ANN model runs also led to the spatial distribution of the probability that had clear delineations.

Theoretically, both MaxEnt and GLM can utilize both discrete and continuous variables as predictors. This can be regarded as an advantage of MaxEnt and GLM because other models including RF and ANN can use continuous variables only. The existence of distinct borders in the above probability maps may be due to the spatial resolution of the dataset. At this spatial resolution, the raster of mean temperature of wettest quarter may be more appropriately represented as a categorical variable. Figure 6 shows the predicted probability map for natural wildfire occurring using MaxEnt model when mean temp of wettest quarter was represented as a categorical

variable. In Figure 6, it was found that the different spatial patterns of the probability from those in Figures 2, 3, and 3 were created and the distinct borders disappeared. However, the problem is that the categorization led to the loss of variance for this variable.

Moreover, the ROC values of the probability maps for natural wildfire occurring from all the model runs varied from 0.47 to 0.95. Mean temperature of wettest quarter is not as pronounced in the POISSON models and, expectedly, produced much lower R values. ROC scores for the POISSON model were also rather varied, with ranges between .546 and .867. R valued indicated that mean temperature of wettest quarter has stronger influence than the other values, but not significant influence. This suggests that the variable is significant but that the relationship is not linear, given the likeness to mean temperature of wettest quarter visually.

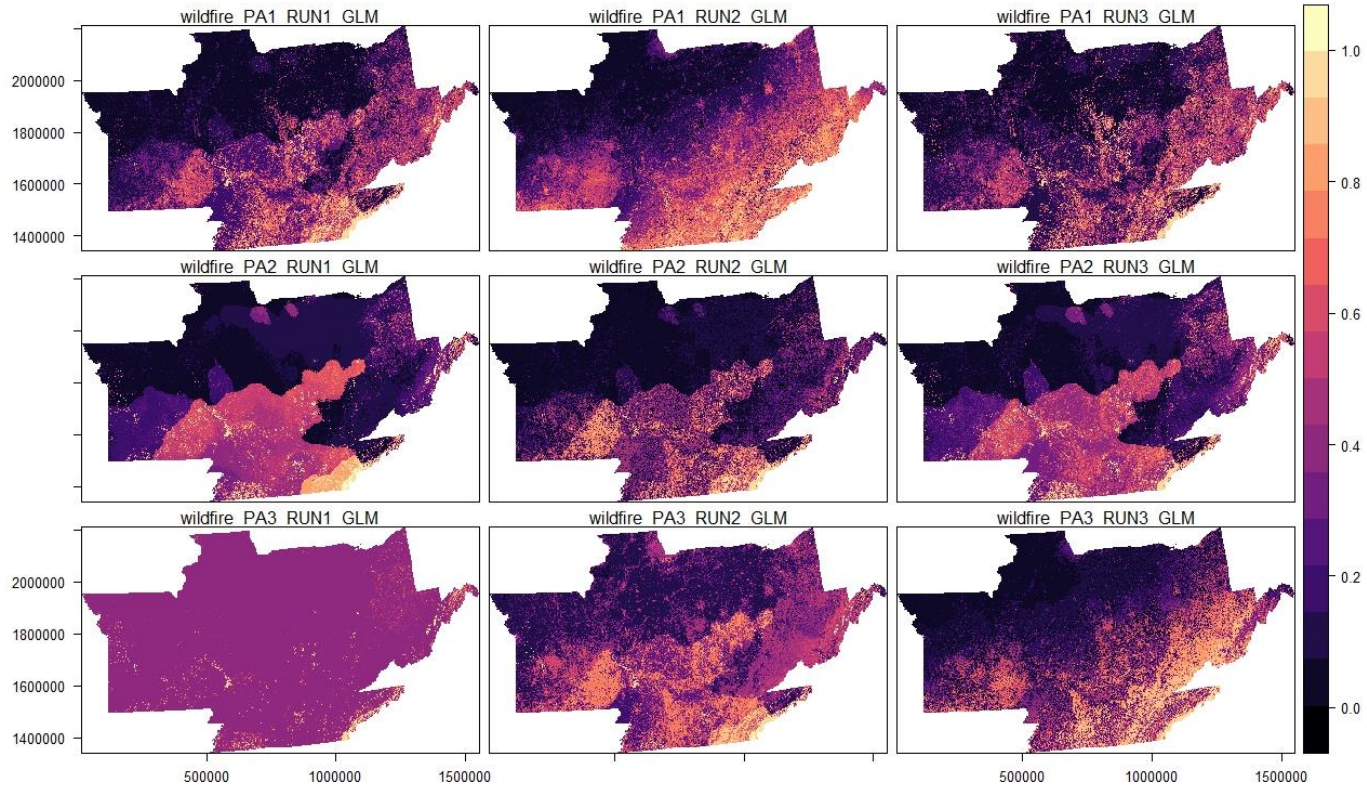

Figure 2: The probability maps of natural wildfire occurring based on General Linearized Model (GLM) with binomial distribution and logistic regression using 50% of data split for Central Region (Note: nine combinations of three model runs: RUN1, RUN2 and RUN3 with three absence data sets: PA1, PA2 and PA3, consisting of 10,000 background points randomly selected).

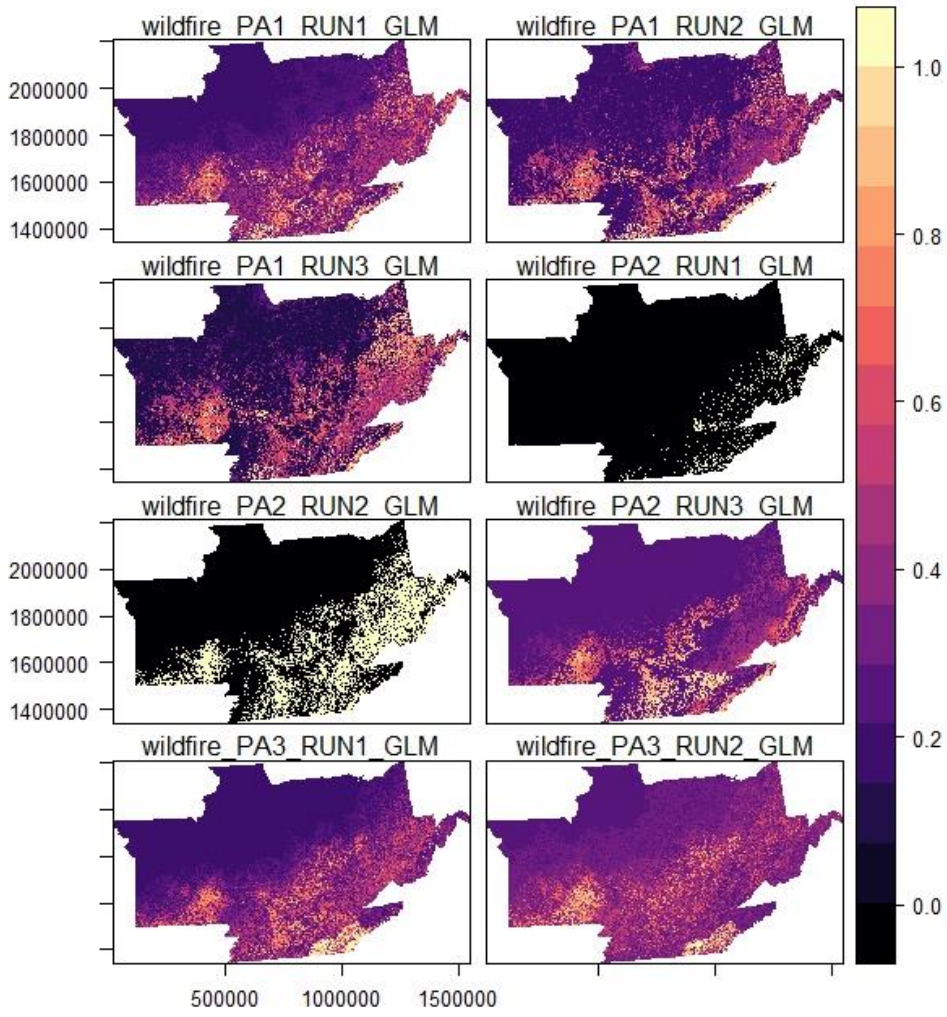

Figure 3: The probability maps of natural wildfire occurring based on General Linearized Model (GLM) with Poisson distribution and logistic regression model using 50% of data split for Central Region (Note: nine combinations of three model runs: RUN1, RUN2 and RUN3 with three absence data sets: PA1, PA2 and PA3, consisting of 10,000 background points randomly selected. The final model did not complete).

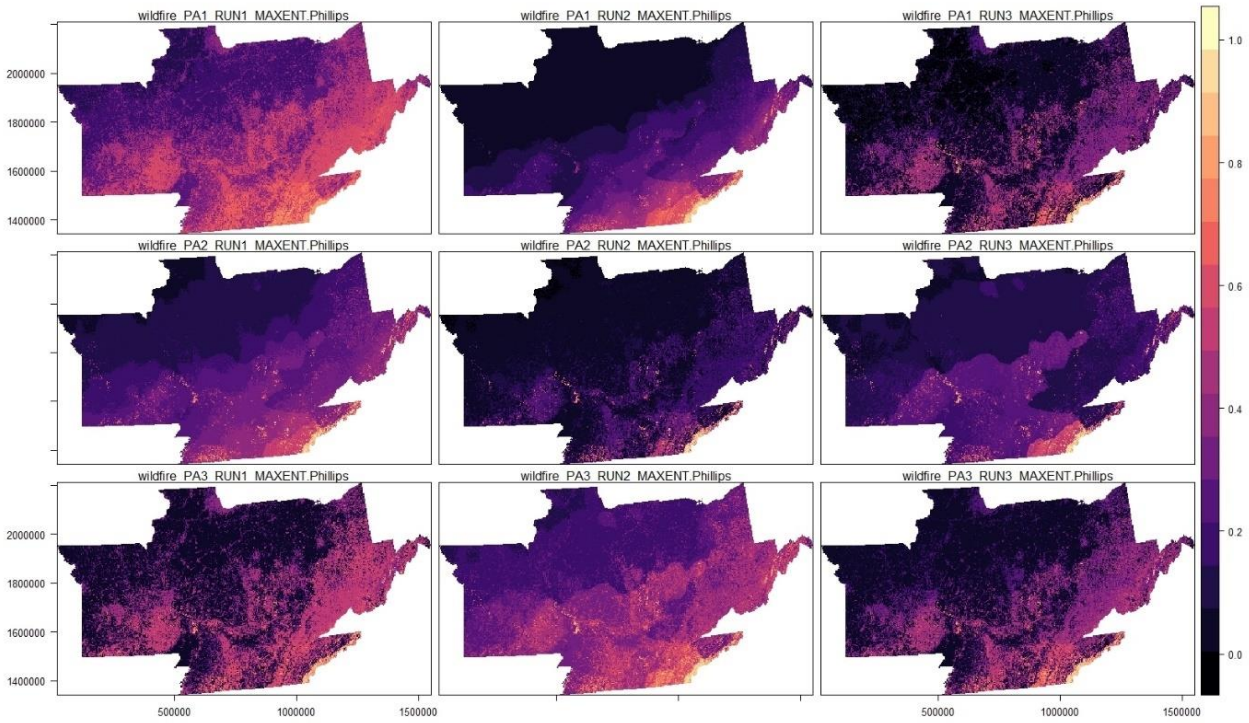

Figure 4: The probability maps of natural wildfire occurring using MaxEnt model with 50% of data split for Central Region (Note: nine combinations of three model runs: RUN1, RUN2 and RUN3 with three absence data sets: PA1, PA2 and PA3, consisting of 10,000 background points randomly selected).

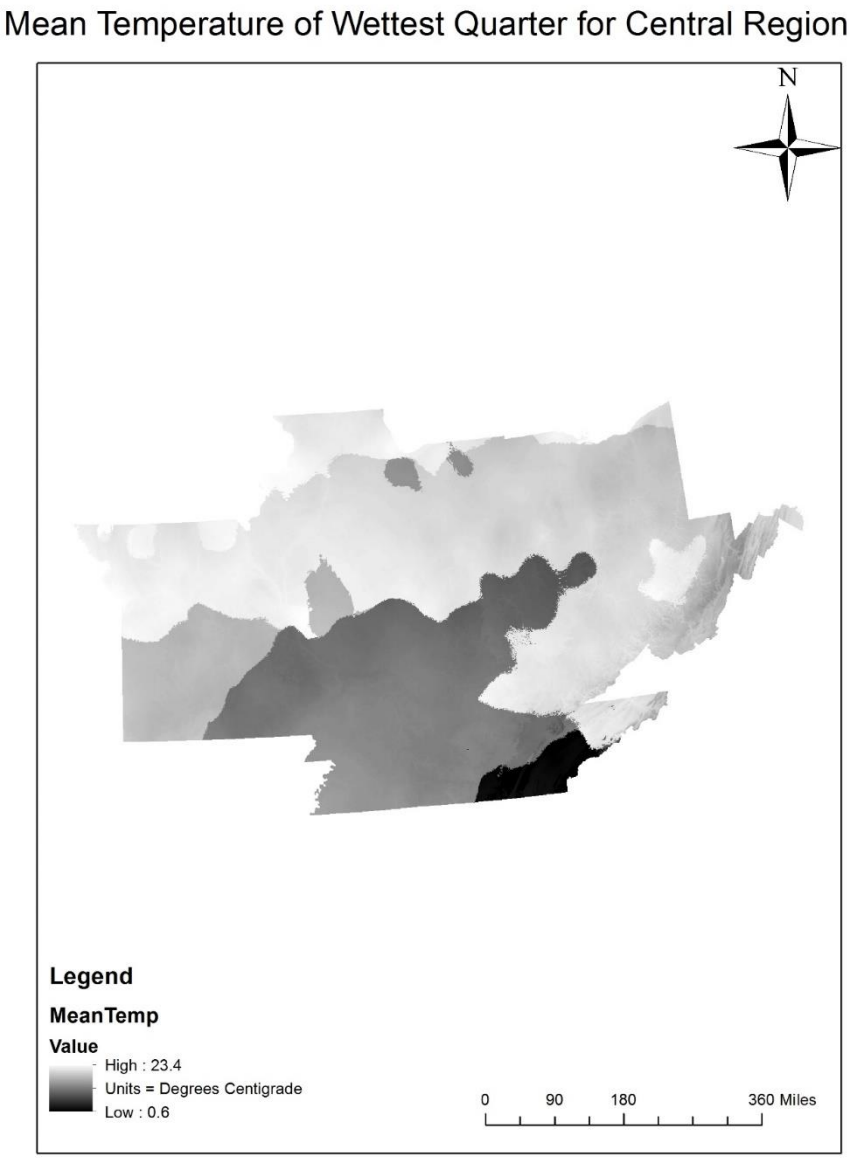

Figure 5: Spatial distribution of mean temperature of wettest quarter for central region. The clear delineations or border lines existed in the raster, which led to the similar spatial patterns of probability for natural wildfire occurrence when the models were run.

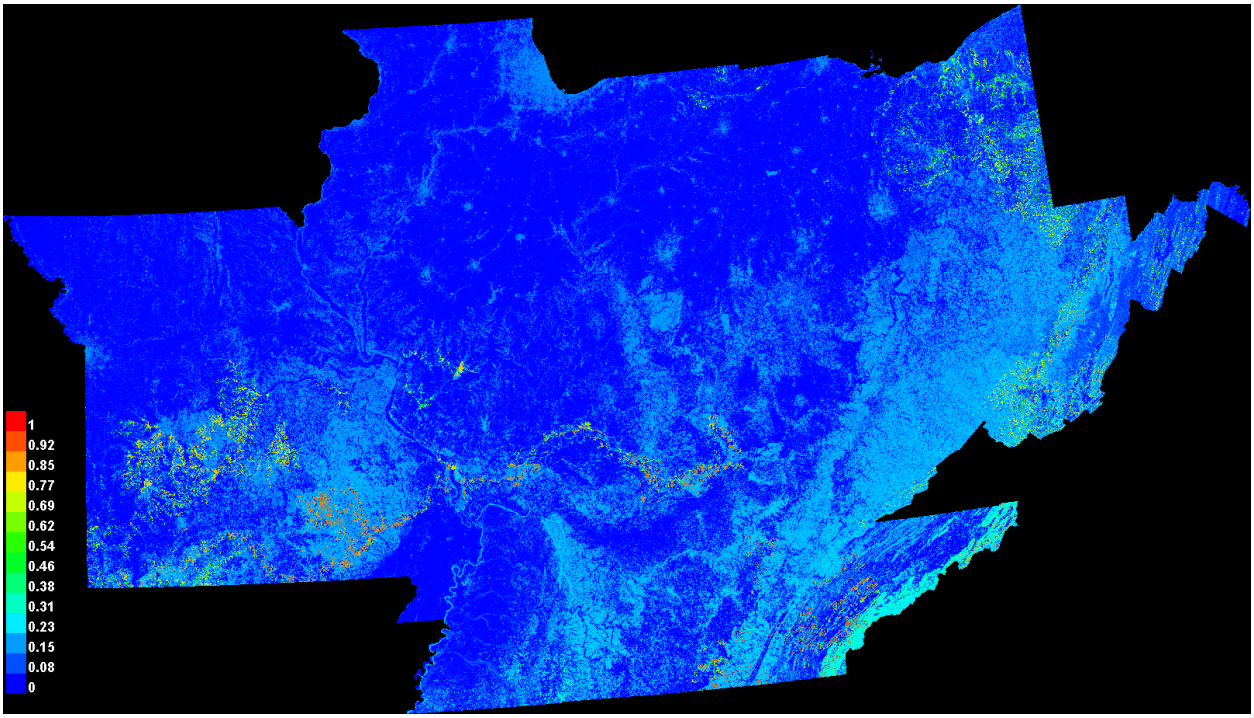

Figure 6: The probability map of natural wildfire occurring using MaxEnt model with 50% of data split for Central Region and representing mean temperature of wettest quarter as a categorical variable with 10,000 background points randomly selected.

4.3 South Region

BIOCLIM

In south region, it was found that there was a strong correlation between annual precipitation and precipitation of driest quarter. Literature review stated that the variables were uncorrelated with each other, but this may be due to different study areas. Moreover, in this region the VIF values ranged from 23 to 26 when annual precipitation was involved. This might indicate that moisture plays a strong role in the region. When annual precipitation was removed and the models were run again, the

VIF values obtained were all less than 2.5. In addition, both the smallest and largest ROC values of 0.769 and 0.930 respectively came from the probability maps of natural wildfire occurring and were created by ANN model runs. MaxEnt and MARS runs consistently resulted in the ROC values of larger than 0.85. All the model runs tended to agree that mean temperature of warmest quarter and precipitation of driest quarter had the strongest influence on the probability of natural wildfire occurring, and were kept for the composite model.

LANDFIRE

The results of LANDFIRE dataset indicated that elevation was a significant driver of natural wildfire occurring for South Region with R values greater than 0.8 between the predicted and observed probabilities for MaxEnt, ANN and MARS model. One MaxEnt model run led to the highest R value of 0.94. The R values for RF and GLM models were lower than those for MaxEnt, ANN, and MARS, but still showed moderate importance with a range of R values between 0.3 and 0.7. Canopy height was also moderate important and retained for the composite model. The ROC values of the probability maps for all of the model runs were above 0.8, except for GLM, which consistently scored from 0.6 to 0.7.

COMPOSITE DATASET

The VIF values for all the variables involved in the composte data were smaller than 4.4. The ROC values of the predicted probability maps for natural wildfire occurring for the composite model ranged from 0.743 created by a GLM model run to

0.940 by a RF model run. MaxEnt and MARS both consistently scored the ROC values of 0.83 or above. All the model runs tended to show that precipitation of driest quarter was the most important predictor and showed significant influence on the probability of natural wildfire occurring. This may indicate that moisture plays a stronger role than other variables in the the region. The R values between the predicted and observed probabilities ranged from 0.628 to 0.940 across the model runs. The examples of the probability maps for natural wildfire occurring from the runs of MaxEnt model, GLM with LR and binormial distribution and GLM with LR and Poisson distribution are shown in Figures 7, 8 and 9. Although it seemed that overall GLM with LR and binormial distribution led to the highest probability, then GLM with LR and Poisson distribution and MaxEnt model, the spatial patterns of the probabilities were similar. That is, the higher and lower probabilities existed in the northeast and south parts respectively.

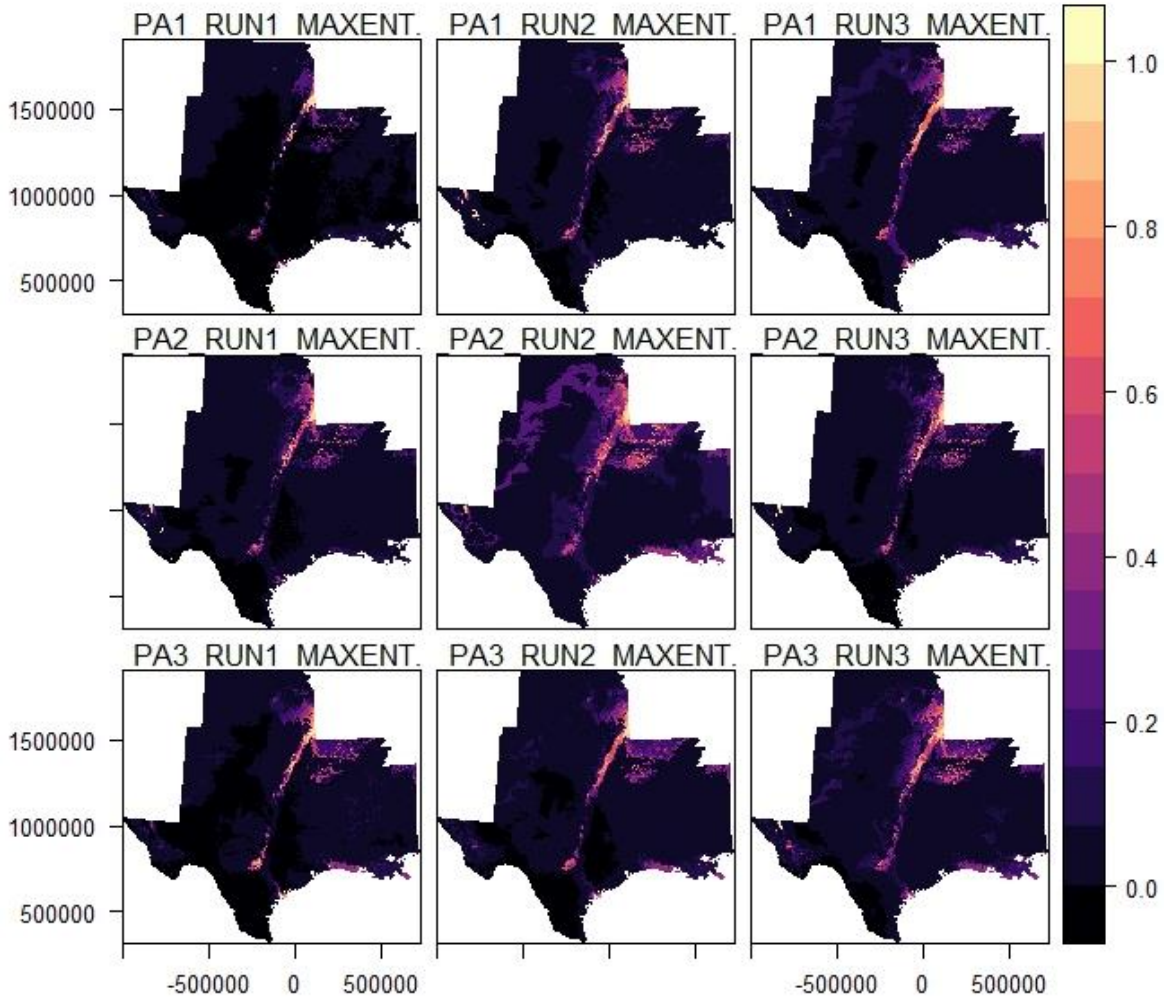

Figure 7: The probability maps of natural wildfire occurring using MaxEnt model for South Region (Note: nine combinations of three model runs: RUN1, RUN2 and RUN3 with three absence data sets: PA1, PA2 and PA3, consisting of 10,000 background points randomly selected).

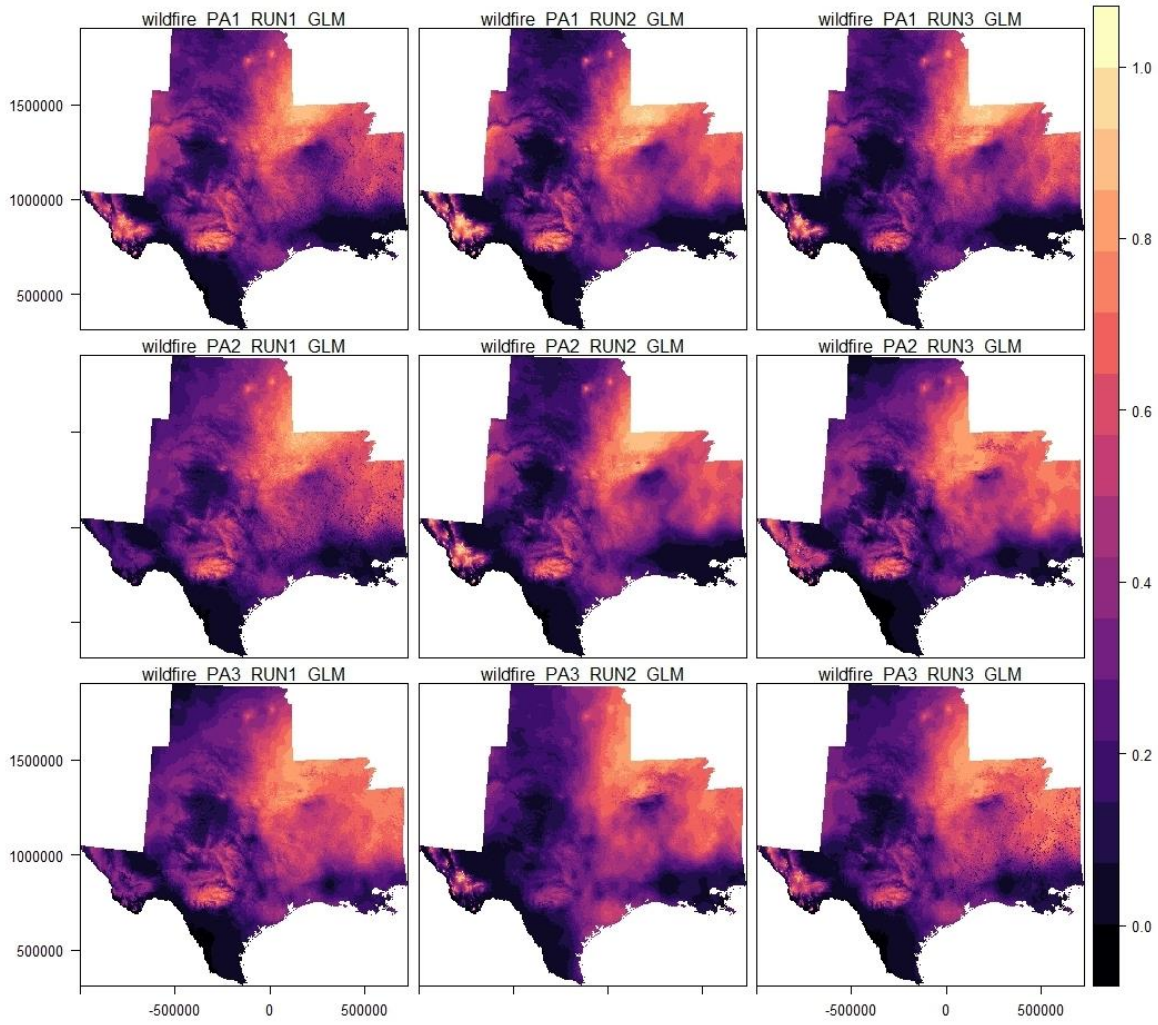

Figure 8: The probability maps of natural wildfire occurring using General Linearized Model (GLM) - with binomial distribution and logistic regression for South Region (Note: nine combinations of three model runs: RUN1, RUN2 and RUN3 with three absence data sets: PA1, PA2 and PA3, consisting of 10,000 background points randomly selected).

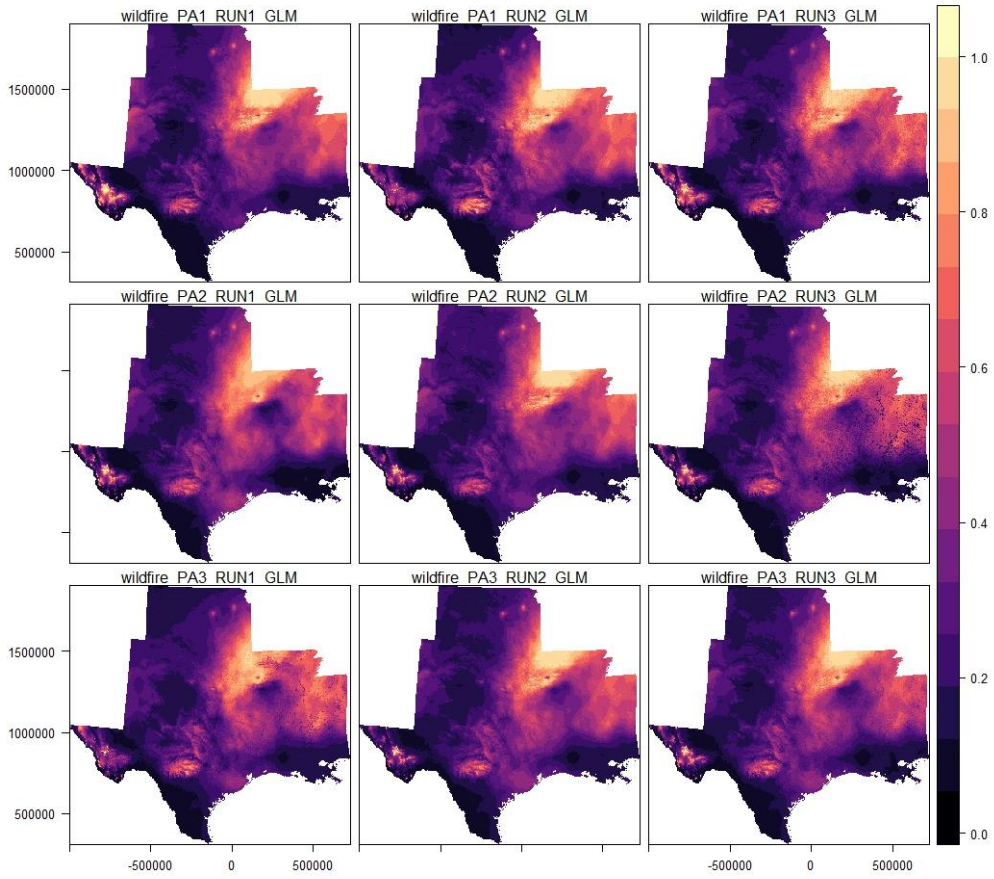

Figure 9: The probability maps of natural wildfire occurring using General Linearized Model (GLM) - with Poisson distribution and logistic regression for South Region (Note: nine combinations of three model runs: RUN1, RUN2 and RUN3 with three absence data sets: PA1, PA2 and PA3, consisting of 10,000 background points randomly selected).

4.4 SOUTHWEST REGION

BIOCLIM

When BIOCLIM dataset was used, the ROC values of probability maps of natural wildfire occurring from all the model runs had a range of 0.747 to 0.872 and most of them 0.8 or above. The only variable that showed significant influence was

precipitation of driest quarter and thus it was kept for the composite model. When precipitation of driest quarter was used, the obtained R values varied from 0.515 to 0.891. Examples of the probability maps from the runs of GLM - with binomial distribution and logistical regression are displayed in Figure 10. Overall, the north and central areas of this region had higher probabilities than other parts.

LANDFIRE

The ROC scores of the probability maps for natural wildfire from LANDFIRE dataset were similar to those from BIOCLIM dataset, with the lowest value of 0.785 by an ANN model run and the highest value of 0.853 by a RF model run. Determining predictor variables was slightly more difficult using this dataset than using BIOCLIM dataset. Most of the ANN model runs tended to show that canopy bulk density was a significant driver with the R values of around 0.8 with a couple of ANN model runs resulting in the R values of around 0.3 for canopy bulk density. Most of the RF model runs led to the R values ranging from 0.45 to 0.55 for canopy bulk density, implying moderate importance. The only other variable with significant R values was elevation. These two variables were thus selected for the composite model. Figure 11 displays the probability maps of natural wildfire occurring using MaxEnt model and LANDFIRE dataset for Southwest Region. The probability maps of natural wildfire occurring looked similar to those in Figure 10 by BIOCLIM dataset, but overall the probability values were greater in Figure 10 than Figure 11.

COMPOSITE DATASET

Three variables used in composite model were precipitation of driest quarter, canopy bulk density, and elevation. Their VIF values were smaller than 2.76, indicating no significant collinearity. The ROC scores of probability maps of natural wildfire occurring were similar to those using LANDFIRE and BIOCLIM datasets, with a range of 0.747 by an ANN model run and the high ROC values and agreement among three datasets provided the great potential to predict the probabilities of natural wildfire occurring using the models in this region. The results of R values indicated that with the highest R value of 0.745, precipitation of driest quarter had a stronger influence on the predictions of probability for natural wildfire occurring than canopy bulk density and elevation. Elevation had influence only for some of the model runs.

The probability maps of natural wildfire occurring from all the model runs show similar spatial patterns for this region (Figures 10 and 11). The spatial patterns were similar to the spatial distribution of precipitation of driest quarter (Figure 12). This may indicate that moisture plays a strong role in the occurrence of natural wildfires in this region. Table 1 shows different evaluation scores for the region including Kappa, True Skill Statistic (TSS), Success Ration (SR), False Alarm Ration (FAR), the Critical Success Index (CSI), and the Relative Operator Characteristic (ROC).

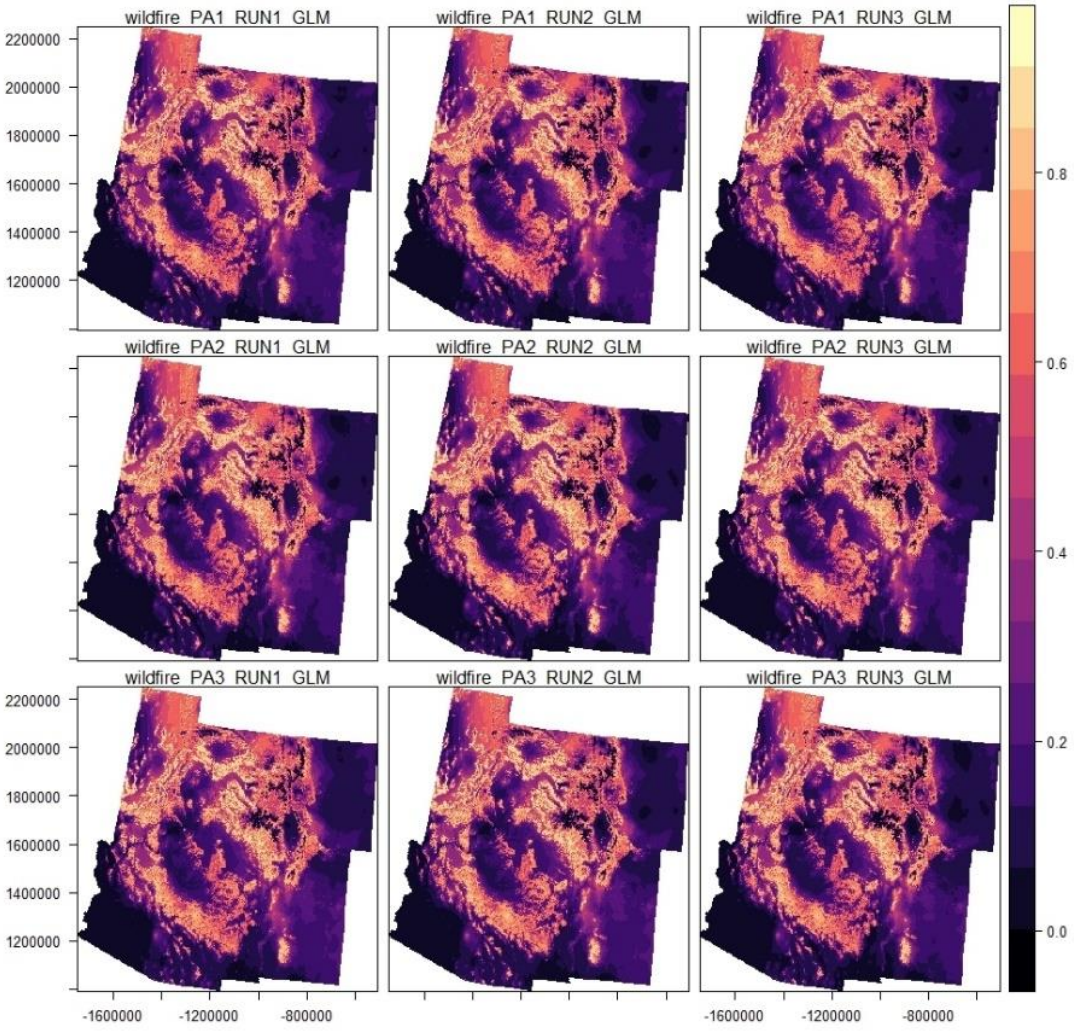

Figure 10: The probability maps of natural wildfire occurring using General Linearized Model (GLM) - with binomial distribution and logistical regression for Southwest Region (Note: nine combinations of three model runs: RUN1, RUN2 and RUN3 with three absence data sets: PA1, PA2 and PA3, consisting of 10,000 background points randomly selected).

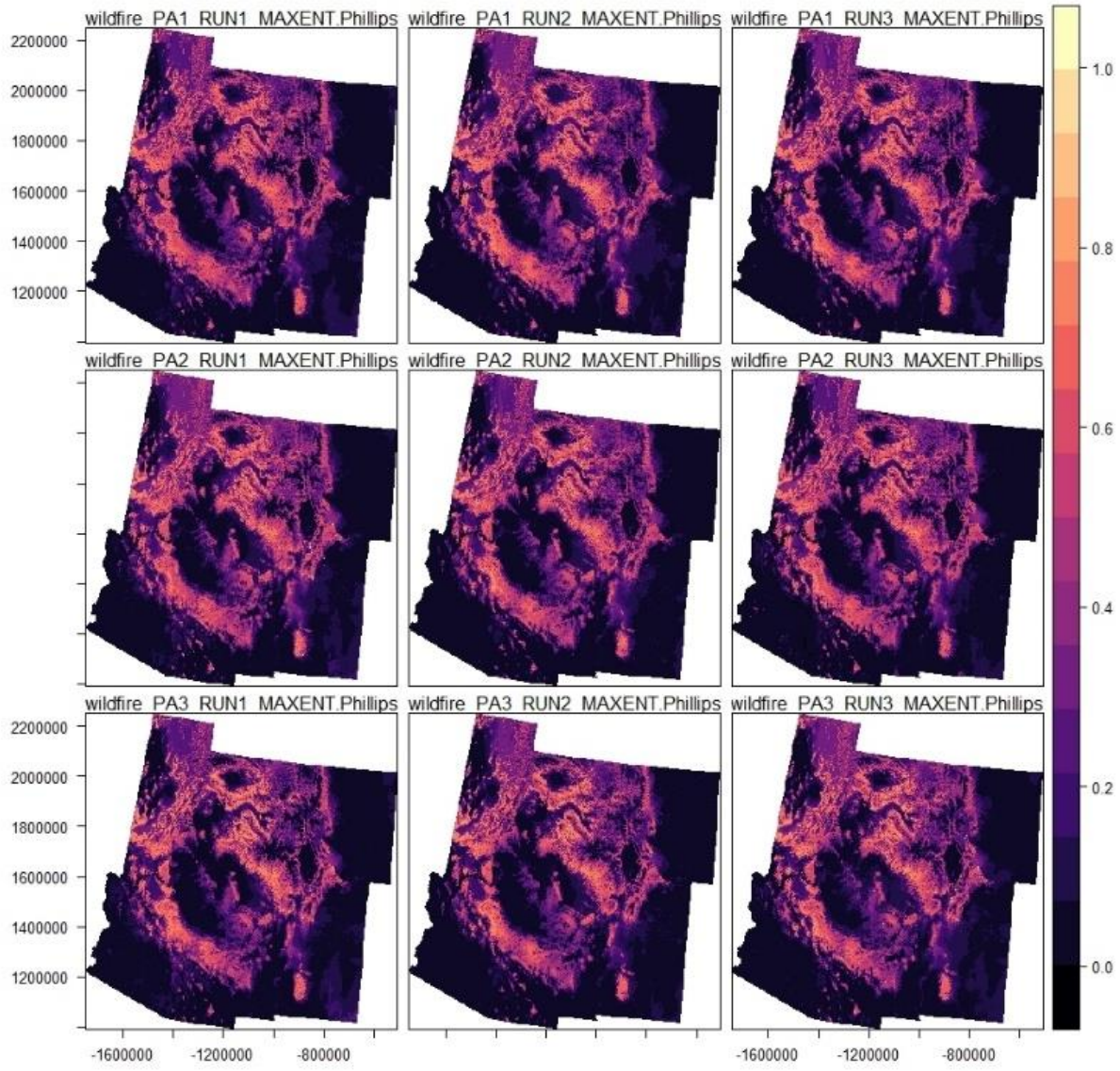

Figure 11: The probability maps of natural wildfire occurring using MaxEnt model for Southwest Region (Note: nine combinations of three model runs: RUN1, RUN2 and RUN3 with three absence data sets: PA1, PA2 and PA3, consisting of 10,000 background points randomly selected).

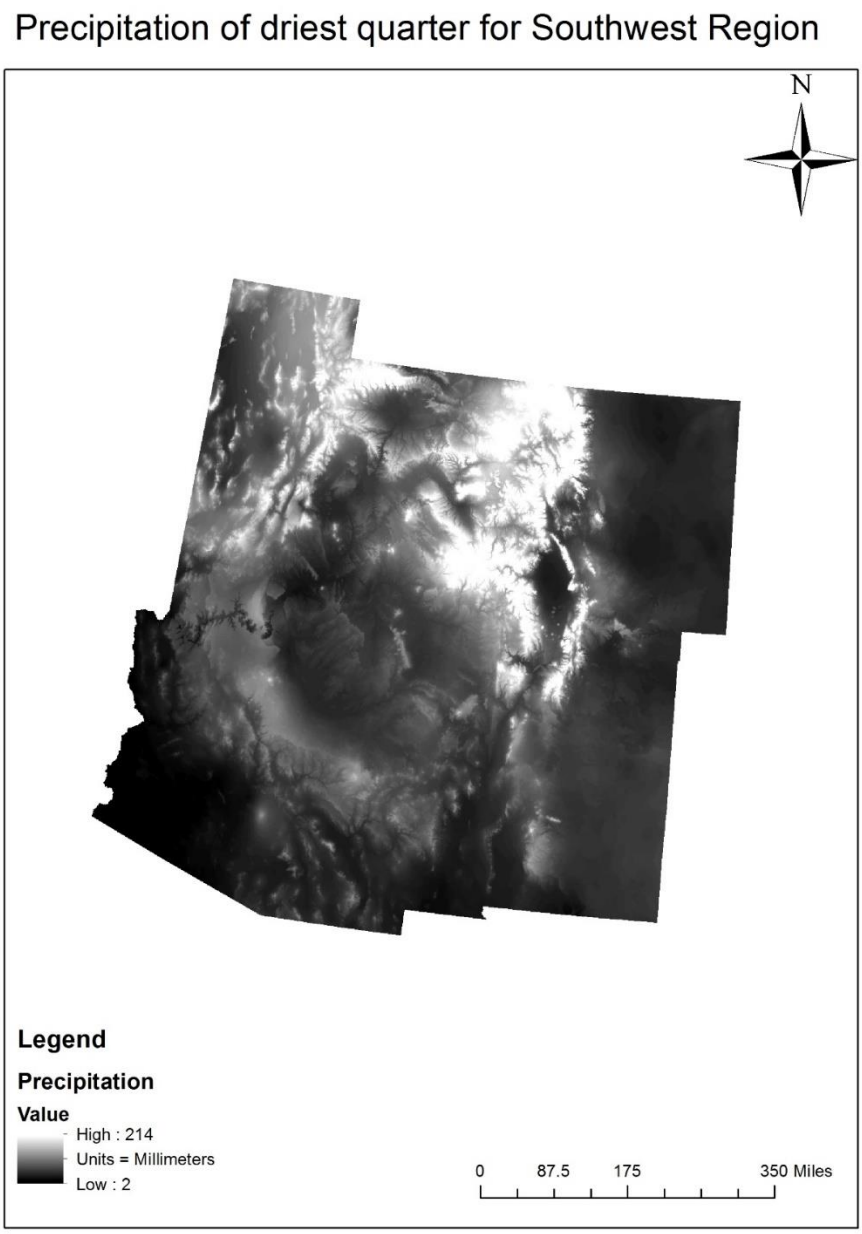

Figure 12: Spatial distributions of precipitation of driest quarter, indicating that precipitation of driest quarter is linked with the probability of naturally wildfire occurring.

		MAXENT	RUN1	PA1
	Testing.data	**Cutoff**	**Sensitivity**	**Specificity**
ROC	**0.832**	**373**	**78.592**	**76.567**
TSS	**0.551**	**369**	**78.592**	**76.467**
SR	**0.6**	**760**	**4.741**	**99.167**
FAR	**0.6**	**760**	**4.741**	**99.167**
KAPPA	**0.451**	**469**	**69.684**	**82.9**
CSI	**0.401**	**469**	**69.684**	**82.9**
		GLM	**RUN1**	**PA1**
	Testing.data	**Cutoff**	**Sensitivity**	**Specificity**
ROC	**0.821**	**532.5**	**78.592**	**72.7**
TSS	**0.512**	**488**	**81.753**	**69.333**
SR	**0.532**	**813**	**25.287**	**94.8**
FAR	**0.532**	**813**	**25.287**	**94.8**
KAPPA	**0.417**	**659**	**61.351**	**84.733**
CSI	**0.371**	**650**	**62.644**	**83.933**
		ANN	**RUN1**	**PA1**
	Testing.data	**Cutoff**	**Sensitivity**	**Specificity**
ROC	**0.829**	**451.5**	**80.747**	**72.667**
TSS	**0.534**	**472**	**79.31**	**74**
SR	**0.55**	**850**	**12.356**	**97.7**
FAR	**0.55**	**850**	**12.356**	**97.7**
KAPPA	**0.431**	**657**	**66.81**	**82.9**
CSI	**0.388**	**608**	**70.402**	**81.2**

Table 1: Binary classification metrics representing evaluative performance for probability of natural wildfire occurrence from the composite dataset for the Southwest Region

		RF	**RUN1**	**PA1**
	Testing.data	**Cutoff**	**Sensitivity**	**Specificity**
ROC	**0.841**	**123**	**84.052**	**72.933**
TSS	**0.569**	**118**	**84.626**	**72.167**
SR	**1**	**941.5**	**0.431**	**100**
FAR	**1**	**941.5**	**0.431**	**100**
KAPPA	**0.442**	**256**	**70.977**	**81.7**
CSI	**0.398**	**217**	**75.287**	**79.367**
		MARS	**RUN1**	**PA1**
	Testing.data	**Cutoff**	**Sensitivity**	**Specificity**
ROC	**0.83**	**545.5**	**78.879**	**75.267**
TSS	**0.54**	**568**	**75.718**	**78.2**
SR	**0.643**	**870**	**2.73**	**99.633**
FAR	**0.643**	**870**	**2.73**	**99.633**
KAPPA	**0.442**	**637**	**68.966**	**82.6**
CSI	**0.395**	**637**	**68.966**	**82.6**
		MAXENT	**RUN2**	**PA1**
	Testing.data	**Cutoff**	**Sensitivity**	**Specificity**
ROC	**0.845**	**364**	**80.316**	**75.833**
TSS	**0.561**	**367**	**80.029**	**75.967**
SR	**1**	**790**	**0.144**	**100**
FAR	**1**	**790**	**0.144**	**100**
KAPPA	**0.472**	**470**	**71.695**	**83.4**
CSI	**0.418**	**462**	**72.557**	**83**
		GLM	**RUN2**	**PA1**
	Testing.data	**Cutoff**	**Sensitivity**	**Specificity**
ROC	**0.828**	**587.5**	**74.713**	**78.033**
TSS	**0.527**	**586**	**75**	**77.7**
SR	**0.692**	**884**	**2.73**	**99.7**
FAR	**0.692**	**884**	**2.73**	**99.7**
KAPPA	**0.449**	**677**	**61.638**	**86.6**
CSI	**0.394**	**640**	**66.81**	**83.733**

Table 1: (Continued) Binary classification metrics representing evaluative performance for probability of natural wildfire occurrence from the composite dataset for the Southwest Region

		ANN	**RUN2**	**PA1**
	Testing.data	**Cutoff**	**Sensitivity**	**Specificity**
ROC	**0.832**	**521**	**75.862**	**78.3**
TSS	**0.541**	**523**	**75.718**	**78.3**
SR	**0.556**	**829**	**31.322**	**94.1**
FAR	**0.556**	**829**	**31.322**	**94.1**
KAPPA	**0.458**	**603**	**68.966**	**83.5**
CSI	**0.405**	**603**	**68.966**	**83.5**
		RF	**RUN2**	**PA1**
	Testing.data	**Cutoff**	**Sensitivity**	**Specificity**
ROC	**0.845**	**95**	**86.925**	**69.1**
TSS	**0.56**	**95**	**86.925**	**69.1**
SR	**1**	**919**	**0.718**	**100**
FAR	**1**	**919**	**0.718**	**100**
KAPPA	**0.446**	**303**	**66.954**	**83.867**
CSI	**0.397**	**256**	**71.839**	**81.2**
		MARS	**RUN2**	**PA1**
	Testing.data	**Cutoff**	**Sensitivity**	**Specificity**
ROC	**0.84**	**579.5**	**77.011**	**78.833**
TSS	**0.557**	**581**	**76.868**	**78.867**
SR	**1**	**874**	**0.144**	**100**
FAR	**1**	**874**	**0.144**	**100**
KAPPA	**0.461**	**651**	**69.684**	**83.567**
CSI	**0.408**	**651**	**69.684**	**83.567**
		MAXENT	**RUN3**	**PA1**
	Testing.data	**Cutoff**	**Sensitivity**	**Specificity**
ROC	**0.833**	**336.5**	**78.161**	**75.7**
TSS	**0.538**	**332**	**78.161**	**75.533**
SR	**0.68**	**768**	**2.73**	**99.733**
FAR	**0.68**	**768**	**2.73**	**99.733**
KAPPA	**0.434**	**418**	**70.546**	**81.367**
CSI	**0.391**	**418**	**70.546**	**81.367**

Table 1 (Continued) Binary classification metrics representing evaluative performance for probability of natural wildfire occurrence from the composite dataset for the Southwest Region

		GLM	**RUN3**	**PA1**
	Testing.data	**Cutoff**	**Sensitivity**	**Specificity**
ROC	**0.821**	**530.5**	**76.58**	**74.467**
TSS	**0.509**	**425**	**83.621**	**67.033**
SR	**0.647**	**886**	**1.58**	**99.8**
FAR	**0.647**	**886**	**1.58**	**99.8**
KAPPA	**0.413**	**660**	**58.908**	**85.6**
CSI	**0.372**	**615**	**66.236**	**81.867**
		ANN	**RUN3**	**PA1**
	Testing.data	**Cutoff**	**Sensitivity**	**Specificity**
ROC	**0.828**	**506.5**	**74.569**	**78.233**
TSS	**0.525**	**509**	**74.138**	**78.4**
SR	**0.552**	**836**	**29.454**	**94.433**
FAR	**0.552**	**836**	**29.454**	**94.433**
KAPPA	**0.437**	**637**	**62.5**	**85.433**
CSI	**0.387**	**581**	**66.092**	**83.533**
		RF	**RUN3**	**PA1**
	Testing.data	**Cutoff**	**Sensitivity**	**Specificity**
ROC	**0.835**	**73**	**83.908**	**69.533**
TSS	**0.531**	**68**	**84.195**	**68.667**
SR	**1**	**929**	**0.862**	**100**
FAR	**1**	**929**	**0.862**	**100**
KAPPA	**0.442**	**284**	**66.379**	**83.933**
CSI	**0.393**	**254**	**68.678**	**82.5**
		MARS	**RUN3**	**PA1**
	Testing.data	**Cutoff**	**Sensitivity**	**Specificity**
ROC	**0.835**	**446.5**	**84.052**	**69.4**
TSS	**0.532**	**480**	**81.609**	**71.5**
SR	**0.706**	**855**	**3.879**	**99.6**
FAR	**0.706**	**855**	**3.879**	**99.6**
KAPPA	**0.436**	**698**	**59.914**	**86.533**
CSI	**0.383**	**576**	**71.121**	**80.033**

Table 1: (Continued) Binary classification metrics representing evaluative performance for probability of natural wildfire occurrence from the composite dataset for the Southwest Region

		MAXENT	**RUN1**	**PA2**
	Testing.data	**Cutoff**	**Sensitivity**	**Specificity**
ROC	**0.838**	**314.5**	**81.609**	**72**
TSS	**0.535**	**320**	**80.891**	**72.6**
SR	**1**	**890**	**0.144**	**100**
FAR	**1**	**890**	**0.144**	**100**
KAPPA	**0.446**	**500**	**64.368**	**85.1**
CSI	**0.392**	**500**	**64.368**	**85.1**
		GLM	**RUN1**	**PA2**
	Testing.data	**Cutoff**	**Sensitivity**	**Specificity**
ROC	**0.825**	**557.5**	**73.707**	**76.9**
TSS	**0.505**	**564**	**72.989**	**77.367**
SR	**0.619**	**877**	**2.155**	**99.6**
FAR	**0.619**	**877**	**2.155**	**99.6**
KAPPA	**0.419**	**653**	**59.914**	**85.533**
CSI	**0.371**	**564**	**72.989**	**77.367**
		ANN	**RUN1**	**PA2**
	Testing.data	**Cutoff**	**Sensitivity**	**Specificity**
ROC	**0.829**	**478.5**	**77.299**	**74.967**
TSS	**0.521**	**454**	**79.598**	**72.433**
SR	**0.566**	**830**	**28.879**	**94.8**
FAR	**0.566**	**830**	**28.879**	**94.8**
KAPPA	**0.426**	**673**	**60.489**	**85.8**
CSI	**0.38**	**540**	**69.971**	**80.333**
		RF	**RUN1**	**PA2**
	Testing.data	**Cutoff**	**Sensitivity**	**Specificity**
ROC	**0.843**	**203**	**73.994**	**79.933**
TSS	**0.539**	**204**	**73.994**	**79.933**
SR	**1**	**930**	**0.431**	**100**
FAR	**1**	**930**	**0.431**	**100**
KAPPA	**0.464**	**282**	**68.534**	**84.433**
CSI	**0.409**	**273**	**69.109**	**84**

Table 1: (Continued) Binary classification metrics representing evaluative performance for probability of natural wildfire occurrence from the composite dataset for the Southwest Region

		MARS	**RUN1**	**PA2**
	Testing.data	**Cutoff**	**Sensitivity**	**Specificity**
ROC	**0.835**	**508.5**	**78.017**	**74.1**
TSS	**0.519**	**509**	**78.017**	**74.1**
SR	**0.633**	**874**	**5.46**	**99.333**
FAR	**0.633**	**874**	**5.46**	**99.333**
KAPPA	**0.438**	**723**	**54.454**	**89.4**
CSI	**0.389**	**607**	**68.678**	**82.2**
		MAXENT	**RUN2**	**PA2**
	Testing.data	**Cutoff**	**Sensitivity**	**Specificity**
ROC	**0.84**	**359.5**	**79.885**	**74.933**
TSS	**0.547**	**360**	**79.885**	**74.933**
SR	**1**	**895**	**0.287**	**100**
FAR	**1**	**895**	**0.287**	**100**
KAPPA	**0.444**	**490**	**67.385**	**83.5**
CSI	**0.396**	**450**	**71.552**	**81.367**
		GLM	**RUN2**	**PA2**
	Testing.data	**Cutoff**	**Sensitivity**	**Specificity**
ROC	**0.822**	**561.5**	**76.006**	**75.2**
TSS	**0.512**	**561**	**76.006**	**75.1**
SR	**0.583**	**886**	**1.293**	**99.767**
FAR	**0.583**	**886**	**1.293**	**99.767**
KAPPA	**0.414**	**660**	**60.776**	**84.9**
CSI	**0.368**	**651**	**61.925**	**84.1**
		ANN	**RUN2**	**PA2**
	Testing.data	**Cutoff**	**Sensitivity**	**Specificity**
ROC	**0.836**	**430.5**	**83.046**	**71.3**
TSS	**0.542**	**434**	**82.615**	**71.567**
SR	**0.75**	**863**	**1.293**	**99.933**
FAR	**0.75**	**863**	**1.293**	**99.933**
KAPPA	**0.447**	**665**	**69.397**	**82.667**
CSI	**0.399**	**640**	**70.977**	**81.867**

Table 1: (Continued) Binary classification metrics representing evaluative performance for probability of natural wildfire occurrence from the composite dataset for the Southwest Region

		RF	RUN2	PA2
	Testing.data	**Cutoff**	**Sensitivity**	**Specificity**
ROC	**0.845**	**151**	**79.023**	**75.4**
TSS	**0.543**	**99**	**83.621**	**70.7**
SR	**1**	**931**	**0.287**	**100**
FAR	**1**	**931**	**0.287**	**100**
KAPPA	**0.46**	**346**	**63.506**	**86.333**
CSI	**0.401**	**327**	**65.086**	**85.567**
		MARS	**RUN2**	**PA2**
	Testing.data	**Cutoff**	**Sensitivity**	**Specificity**
ROC	**0.836**	**558.5**	**76.724**	**76.733**
TSS	**0.534**	**559**	**76.724**	**76.733**
SR	**0.692**	**880**	**2.73**	**99.633**
FAR	**0.692**	**880**	**2.73**	**99.633**
KAPPA	**0.444**	**630**	**69.971**	**82.267**
CSI	**0.397**	**630**	**69.971**	**82.267**
		MAXENT	**RUN3**	**PA2**
	Testing.data	**Cutoff**	**Sensitivity**	**Specificity**
ROC	**0.853**	**343**	**82.902**	**74.733**
TSS	**0.575**	**347**	**82.471**	**75**
SR	**1**	**780**	**0.144**	**100**
FAR	**1**	**780**	**0.144**	**100**
KAPPA	**0.471**	**490**	**68.534**	**84.567**
CSI	**0.413**	**490**	**68.534**	**84.567**
		GLM	**RUN3**	**PA2**
	Testing.data	**Cutoff**	**Sensitivity**	**Specificity**
ROC	**0.838**	**493.5**	**82.615**	**71.267**
TSS	**0.538**	**492**	**82.615**	**71.133**
SR	**0.625**	**876**	**1.724**	**99.767**
FAR	**0.625**	**876**	**1.724**	**99.767**
KAPPA	**0.442**	**644**	**63.218**	**85.367**
CSI	**0.389**	**626**	**66.092**	**83.8**

Table 1: (Continued) Binary classification metrics representing evaluative performance for probability of natural wildfire occurrence from the composite dataset for the Southwest Region

		ANN	**RUN3**	**PA2**
	Testing.data	**Cutoff**	**Sensitivity**	**Specificity**
ROC	**0.848**	**574**	**75.862**	**80.333**
TSS	**0.56**	**569**	**75.862**	**80.067**
SR	**0.62**	**835**	**12.069**	**98.533**
FAR	**0.62**	**835**	**12.069**	**98.533**
KAPPA	**0.458**	**677**	**67.529**	**84.3**
CSI	**0.409**	**577**	**75.575**	**80.367**
		RF	**RUN3**	**PA2**
	Testing.data	**Cutoff**	**Sensitivity**	**Specificity**
ROC	**0.848**	**159**	**78.161**	**77.1**
TSS	**0.551**	**153**	**78.592**	**76.467**
SR	**1**	**893.5**	**0.575**	**100**
FAR	**1**	**893.5**	**0.575**	**100**
KAPPA	**0.478**	**296**	**67.241**	**85.7**
CSI	**0.417**	**296**	**67.241**	**85.7**
		MARS	**RUN3**	**PA2**
	Testing.data	**Cutoff**	**Sensitivity**	**Specificity**
ROC	**0.85**	**535.5**	**80.029**	**76.433**
TSS	**0.564**	**537**	**79.885**	**76.5**
SR	**0.625**	**796**	**29.885**	**95.833**
FAR	**0.625**	**796**	**29.885**	**95.833**
KAPPA	**0.467**	**693**	**63.362**	**86.7**
CSI	**0.41**	**632**	**70.833**	**83.067**
		MAXENT	**RUN1**	**PA3**
	Testing.data	**Cutoff**	**Sensitivity**	**Specificity**
ROC	**0.829**	**414.5**	**73.851**	**79.667**
TSS	**0.532**	**415**	**73.851**	**79.667**
SR	**0.564**	**662**	**33.477**	**93.933**
FAR	**0.564**	**662**	**33.477**	**93.933**
KAPPA	**0.452**	**518**	**64.799**	**85.2**
CSI	**0.397**	**509**	**66.236**	**84.633**
		GLM	**RUN1**	**PA3**

Table 1: (Continued) Binary classification metrics representing evaluative performance for probability of natural wildfire occurrence from the composite dataset for the Southwest Region

	Testing.data	Cutoff	Sensitivity	Specificity
ROC	**0.818**	**551.5**	**76.006**	**75.067**
TSS	**0.509**	**550**	**76.149**	**74.833**
SR	**0.667**	**884**	**2.586**	**99.7**
FAR	**0.667**	**884**	**2.586**	**99.7**
KAPPA	**0.407**	**595**	**70.977**	**79.167**
CSI	**0.375**	**595**	**70.977**	**79.167**
		ANN	**RUN1**	**PA3**
	Testing.data	**Cutoff**	**Sensitivity**	**Specificity**
ROC	**0.78**	**564.5**	**80.46**	**68.067**
TSS	**0.484**	**561**	**80.46**	**67.9**
SR	**0.375**	**707**	**59.195**	**77.133**
FAR	**0.375**	**707**	**59. 195**	**77.133**
KAPPA	**0.333**	**575**	**79.598**	**68.433**
CSI	**0.338**	**568**	**80.316**	**68.1**
		RF	**RUN1**	**PA3**
	Testing.data	**Cutoff**	**Sensitivity**	**Specificity**
ROC	**0.839**	**127**	**81.897**	**72.867**
TSS	**0.546**	**124**	**82.04**	**72.433**
SR	**1**	**927**	**0.718**	**100**
FAR	**1**	**927**	**0.718**	**100**
KAPPA	**0.448**	**296**	**68.678**	**83.033**
CSI	**0.398**	**296**	**68.678**	**83.033**
		MARS	**RUN1**	**PA3**
	Testing.data	**Cutoff**	**Sensitivity**	**Specificity**
ROC	**0.828**	**554.5**	**76.006**	**77**
TSS	**0.529**	**555**	**76.006**	**77**
SR	**0.619**	**856**	**5.891**	**99.133**
FAR	**0.619**	**856**	**5.891**	**99.133**
KAPPA	**0.45**	**667**	**67.385**	**83.833**
CSI	**0.398**	**667**	**67.385**	**83.833**

Table 1: (Continued) Binary classification metrics representing evaluative performance for probability of natural wildfire occurrence from the composite dataset for the Southwest Region

		MAXENT	RUN2	PA3
	Testing.data	**Cutoff**	**Sensitivity**	**Specificity**
ROC	**0.856**	**351.5**	**80.316**	**77.5**
TSS	**0.576**	**347**	**80.316**	**77.267**
SR	**0.671**	**719**	**14.511**	**98.4**
FAR	**0.671**	**719**	**14.511**	**98.4**
KAPPA	**0.484**	**508**	**68.247**	**85.7**
CSI	**0.425**	**458**	**72.126**	**83.733**
		GLM	**RUN2**	**PA3**
	Testing.data	**Cutoff**	**Sensitivity**	**Specificity**
ROC	**0.843**	**495.5**	**80.316**	**73.667**
TSS	**0.538**	**491**	**80.316**	**73.4**
SR	**0.679**	**865**	**5.891**	**99.367**
FAR	**0.679**	**865**	**5.891**	**99.367**
KAPPA	**0.459**	**669**	**59.914**	**88.067**
CSI	**0.398**	**607**	**68.822**	**83.067**
		ANN	**RUN2**	**PA3**
	Testing.data	**Cutoff**	**Sensitivity**	**Specificity**
ROC	**0.81**	**599.5**	**80.747**	**70.133**
TSS	**0.508**	**597**	**80.747**	**70.033**
SR	**0.45**	**715**	**54.598**	**83.5**
FAR	**0.45**	**715**	**54.598**	**83.5**
KAPPA	**0.373**	**684**	**74.569**	**74.7**
CSI	**0.357**	**676**	**75.575**	**74.067**
		RF	**RUN2**	**PA3**
	Testing.data	**Cutoff**	**Sensitivity**	**Specificity**
ROC	**0.859**	**195**	**78.879**	**80.767**
TSS	**0.594**	**171**	**80.46**	**78.9**
SR	**1**	**912**	**0.431**	**100**
FAR	**1**	**912**	**0.431**	**100**
KAPPA	**0.486**	**294**	**67.529**	**85.967**
CSI	**0.43**	**218**	**76.724**	**81.833**

Table 1: (Continued) Binary classification metrics representing evaluative performance for probability of natural wildfire occurrence from the composite dataset for the Southwest Region

		MARS	RUN2	PA3
	Testing.data	**Cutoff**	**Sensitivity**	**Specificity**
ROC	**0.849**	**537.5**	**79.023**	**78.767**
TSS	**0.572**	**531**	**79.167**	**78**
SR	**0.8**	**861**	**2.443**	**99.867**
FAR	**0.8**	**861**	**2.443**	**99.867**
KAPPA	**0.473**	**632**	**68.247**	**84.933**
CSI	**0.419**	**573**	**74.713**	**81.767**

		MAXENT	RUN3	PA3
	Testing.data	**Cutoff**	**Sensitivity**	**Specificity**
ROC	**0.83**	**381.5**	**75.144**	**78.033**
TSS	**0.53**	**376**	**75.431**	**77.467**
SR	**1**	**814**	**0.144**	**100**
FAR	**1**	**814**	**0.144**	**100**
KAPPA	**0.451**	**500**	**66.092**	**84.567**
CSI	**0.398**	**500**	**66.092**	**84.567**

		GLM	RUN3	PA3
	Testing.data	**Cutoff**	**Sensitivity**	**Specificity**
ROC	**0.822**	**585.5**	**74.138**	**77.6**
TSS	**0.514**	**588**	**73.851**	**77.767**
SR	**0.714**	**886**	**2.299**	**99.767**
FAR	**0.714**	**886**	**2.299**	**99.767**
KAPPA	**0.434**	**669**	**61.494**	**85.667**
CSI	**0.385**	**642**	**65.374**	**83.8**

Table 1: (Continued) Binary classification metrics representing evaluative performance for probability of natural wildfire occurrence from the composite dataset for the Southwest Region

		ANN	RUN3	PA3
	Testing.data	**Cutoff**	**Sensitivity**	**Specificity**
ROC	**0.824**	**479.5**	**78.879**	**73.133**
TSS	**0.519**	**478**	**78.879**	**73**
SR	**0.569**	**823**	**32.471**	**94.3**
FAR	**0.569**	**823**	**32.471**	**94.3**
KAPPA	**0.432**	**646**	**63.793**	**84.567**
CSI	**0.385**	**550**	**70.977**	**80.4**

		RF	RUN3	PA3
	Testing.data	**Cutoff**	**Sensitivity**	**Specificity**
ROC	**0.835**	**179**	**77.155**	**76.933**
TSS	**0.54**	**181**	**77.011**	**76.967**
SR	**1**	**923**	**0.862**	**100**
FAR	**1**	**923**	**0.862**	**100**
KAPPA	**0.441**	**285**	**67.241**	**83.433**
CSI	**0.392**	**285**	**67.241**	**83.433**

		MARS	RUN3	PA3
	Testing.data	**Cutoff**	**Sensitivity**	**Specificity**
ROC	**0.829**	**568.5**	**74.569**	**77.967**
TSS	**0.523**	**565**	**74.713**	**77.633**
SR	**0.627**	**843**	**4.885**	**99.267**
FAR	**0.627**	**843**	**4.885**	**99.267**
KAPPA	**0.446**	**686**	**64.943**	**84.833**
CSI	**0.394**	**634**	**69.684**	**81.967**

Table 1: (Continued) Binary classification metrics representing evaluative performance for probability of natural wildfire occurrence from the composite dataset for the Southwest Region

4.5 WEST REGION

BIOCLIM

The ROC values of the probability maps for natural wildfire occurring using BIOCLIM dataset varied from 0.749 by an ANN model run to 0.870 by a RF model run. The R values ranged from 0.18 to 0.62 for annual precipitation and from 0.35 to 0.95 for precipitation of driest quarter, indicating that annual precipitation and precipitation of driest quarter had moderate influence on the predictions of the probabilities for natural wildfire occurring and were retained for the composite model.

LANDFIRE

LANDFIRE dataset did not perform as well as BIOCLIM dataset for predicting the probabilities of natural wildfire occurring in this region, with the lowest ROC value of 0.649 by an ANN model run and the highest ROC value of 0.814 by a RF model run. MaxEnt and RF both consistently scored the ROC values of about 0.8. When LANDFIRE dataset was used, canopy height and elevation indicated moderate influence on the predictions of the probabilities for natural wildfire occurring, and were retained for the composite model.

COMPOSITE DATASET

Compared to LANDFIRE and BIOCLIM datasets, the combination of the datasets, composite model, led to greater ROC values of the probability maps for natural wildfire occurring. The lowest ROC value was 0.795 coming from a GLM

model run, while the highest ROC value was 0.870 obtained by a RF model run. The RF model runs tended to produce the higher ROC values than other model runs in this region.

The R values for four variables greatly varied among different model runs, making it difficult to determine which variables were more strongly linked to natural wildfire occurring in this region. Therefore, the average R values across the model runs for each of five models were calculated and are shown in Table 2. The results indicated that annual precipitation and elevation were more strongly linked to natural wildfire occurring in this study area than other two variables. As found in South region, a further study concerning the role of moisture and elevation may be beneficial in this region. The VIF values with a greatest value of 3.96 indicated that there was little collinearity.

	MAXENT	**GLM**	**RF**	**ANN**	**MARS**
TOTAL ANNUAL PRECIPITATION	**0.4392**	**0.2779**	**0.5410**	**0.5394**	**0.3446**
PRECIPITATION OF DRIEST MONTH/QUARTER	**0.2729**	**0.0766**	**0.4867**	**0.3016**	**0.1381**
CANOPY HEIGHT	**0.0840**	**0.0390**	**0.3233**	**0.1967**	**0.0766**
ELEVATION	**0.2743**	**0.3261**	**0.5894**	**0.5186**	**0.3407**

Table 2. Average values of coefficients of correlation R across all the model runs for West Region

4.6 NORTHWEST REGION

BIOCLIM

BIOCLIM dataset led to the lowest ROC value of 0.671 by an ANN model run and the highest ROC value of 0.817 by a RF model run for the probability maps for natural wildfire occurring. Overall, it appeared that RF model runs produced the highest ROC scores, then MaxEnt and MARS. GLM and ANN model runs tended to lead to the smallest ROC values. The R values indicated that precipitation of driest quarter and temperature seasonality had the strongest influence on the predictions of the probabilities for natural wildfire occurring with the greatest R values of 0.716 and 0.851 respectively, for ANN model runs. Both precipitation of driest quarter and temperature seasonality were retained for the composite model.

LANDFIRE

The ROC values of the probability maps for natural wildfire occurring from LANDFIRE dataset had a range of 0.711 by an ANN model run to 0.796 by a RF model run. Overall, RF, MaxEnt and MARS models runs led similar ROC values that were larger than those from GLM and ANN model runs. The R values indicated that elevation was more important than the other variables to predict the probabilities of natural wildfire occurring. Only the elevation was thus retained for the composite model.

COMPOSITE DATASET

Precipitation of driest quarter, temperature seasonality and elevation were involved in the composite dataset. In this dataset, ANN model runs produced the lowest ROC value of 0.625. GLM model runs consistently led to the ROC values of around 0.7. MaxEnt and MARS model runs resulted in the highest ROC values of around 0.75, while RF model runs consistently had the ROC values of around 0.80.

The R values varied greatly. The average R values were calculated and are listed in Table 3. The overall mean R values were 0.49, 0.43 and 0.27 for elevation, temperature seasonality and precipitation of driest quarter respectively. Thus, elevation contributed the most to the variation of the predicted probabilities of natural wildfire occurring, then temperature seasonality and precipitation of driest quarter. Visually, elevation seemed to be the most significant driver because the spatial distributions of the predicted probabilities for natural wildfire occurring using the composite dataset aere like the spatial patterns of elevation raster.

	MAXENT	**GLM**	**RF**	**ANN**	**MARS**
Precip Driest Qtr	**0.219**	**0.192**	**0.530444**	**0.242**	**0.209111**
Temp seasonality	**0.397333**	**0.235667**	**0.664222**	**0.43875**	**0.453333**
Elevation	**0.398222**	**0.527**	**0.558111**	**0.707**	**0.279778**

Table 3: Average R Values for Northwest Region

4.7 WEST NORTH CENTRAL REGION

BIOCLIM

The values of VIF test for BIOCLIM dataset revealed that precipitation of driest quarter was a significantly correlated variable with annual precipitation and was removed. There were no significant correlations among the remaining four variables. Overall ANN model runs led to the smallest ROC values of the predicted probability maps for natural wildfire occurring, with a range of 0.6 to 0.7. Similar to the results in the previous regions, MARS, RF and MaxEnt model runs resulted in similar ROC values although MaxEnt had the highest ROC values. Most of the R values from all the model runs for mean temperature of wettest quarter were larger than 0.8, indicating that it had strong influence on the predicted probabilities of natural wildfire occurring. Temperature seasonality indicated moderate influence.

LANDFIRE

Most of the ROC values for the predicted probability maps of natural wildfire occurring were greater than 0.8. There was one GLM model run that produced only a ROC value of 0.641. The odd thing about the dataset was that the clear majority of R values indicated that none of the variables were significant and the results from only two GLM model runs showed that canopy height and vegetation type were significant, with the R values of 0.947 and 0.972 (Table 4).

	MAXENT	*GLM*	*RF*	*ANN*	*MARS*
asp_agg.tif	**0.011**	**0.003**	**0.167**	**0.17**	**0.029**
cbd_agg.tif	**0.136**	**0.003**	**0.385**	**0.013**	**0.075**
ch_agg.tif	**0.026**	**0.208**	**0.217**	**0.531**	**0.044**
dem_agg.tif	**0.058**	**0.036**	**0.146**	**0.183**	**0.114**
dist2010.tif	**0.034**	**0.013**	**0.049**	**0.018**	**0.013**
evt_agg.tif	**0.124**	**0.219**	**0.131**	**0.063**	**0.172**
slp.tif	**0.063**	**0.063**	**0.22**	**0.052**	**0.067**
		RUN2,	**PA1**		
	MAXENT	**GLM**	**RF**	**ANN**	**MARS**
asp_agg.tif	**0.022**	**0.009**	**0.157**	**0.188**	**0.036**
cbd_agg.tif	**0.055**	**0.017**	**0.271**	**0.136**	**0.056**
ch_agg.tif	**0.081**	**0.371**	**0.226**	**0.257**	**0.06**
dem_agg.tif	**0.074**	**0.033**	**0.134**	**0.101**	**0.023**
dist2010.tif	**0.047**	**0.017**	**0.066**	**0.028**	**0.022**
evt_agg.tif	**0.129**	**0.18**	**0.111**	**0.06**	**0.176**
slp.tif	**0.078**	**0.079**	**0.219**	**0.082**	**0.105**
		RUN3,	**PA1**		
	MAXENT	**GLM**	**RF**	**ANN**	**MARS**
asp_agg.tif	**0.007**	**0.002**	**0.161**	**0.06**	**0.059**
cbd_agg.tif	**0.168**	**0.008**	**0.311**	**0.267**	**0.072**
ch_agg.tif	**0.017**	**0.196**	**0.214**	**0.213**	**0.041**
dem_agg.tif	**0.076**	**0.04**	**0.156**	**0.06**	**0.027**
dist2010.tif	**0.032**	**0.014**	**0.063**	**0.032**	**0.014**
evt_agg.tif	**0.136**	**0.256**	**0.141**	**0.018**	**0.23**
slp.tif	**0.073**	**0.103**	**0.244**	**0.155**	**0.087**
		RUN1,	**PA2**		
	MAXENT	**GLM**	**RF**	**ANN**	**MARS**
asp_agg.tif	**0.011**	**0**	**0.166**	**0.17**	**0.009**
cbd_agg.tif	**0.16**	**0.007**	**0.373**	**0.009**	
ch_agg.tif	**0.016**	**0.237**	**0.22**	**0.313**	**0.238**
dem_agg.tif	**0.085**	**0.04**	**0.173**	**0.224**	**0.052**
dist2010.tif	**0.061**	**0.015**	**0.109**	**0.01**	**0.02**
evt_agg.tif	**0.15**	**0.209**	**0.128**	**0.056**	**0.147**
slp.tif	**0.086**	**0.121**	**0.244**	**0.125**	**0.108**

Table 4: The coefficients of correlation between the predicted and observed probabilities for natural wildfire occurrence from LANDFIRE dataset for West North Central Region

		RUN2,	PA2		
	MAXENT	**GLM**	**RF**	**ANN**	**MARS**
asp_agg.tif	**0.004**	**0**	**0.113**	**0.057**	**0**
cbd_agg.tif	**0.154**	**0.019**	**0.285**	**0.144**	**0.059**
ch_agg.tif	**0.041**	**0.307**	**0.208**	**0.279**	**0.089**
dem_agg.tif	**0.077**	**0.026**	**0.18**	**0.148**	**0.07**
dist2010.tif	**0.044**	**0.014**	**0.081**	**0.012**	**0.018**
evt_agg.tif	**0.129**	**0.243**	**0.154**	**0.084**	**0.175**
slp.tif	**0.08**	**0.098**	**0.204**	**0.069**	**0.069**
		RUN3,	**PA2**		
	MAXENT	**GLM**	**RF**	**ANN**	**MARS**
asp_agg.tif	**0.017**	**0**	**0.113**	**0.066**	**0.047**
cbd_agg.tif	**0.149**	**0.018**	**0.29**	**0.003**	**0.032**
ch_agg.tif	**0.036**	**0.305**	**0.262**	**0.336**	**0.116**
dem_agg.tif	**0.081**	**0.027**	**0.166**	**0.096**	**0.045**
dist2010.tif	**0.043**	**0.016**	**0.096**	**0.017**	**0.018**
evt_agg.tif	**0.152**	**0.257**	**0.121**	**0.038**	**0.235**
slp.tif	**0.071**	**0.152**	**0.198**	**0.106**	**0.137**
		RUN1,	**PA3**		
	MAXENT	**GLM**	**RF**	**ANN**	**MARS**
asp_agg.tif	**0.008**	**0**	**0.139**	**0.113**	**0.027**
cbd_agg.tif	**0.091**	**0**	**0.278**	**0.094**	**0.113**
ch_agg.tif	**0.03**	**0.947**	**0.222**	**0.258**	**0.038**
dem_agg.tif	**0.086**	**0**	**0.151**	**0.135**	**0.04**
dist2010.tif	**0.021**	**0.034**	**0.067**	**0.021**	**0.014**
evt_agg.tif	**0.159**	**0.326**	**0.133**	**0.044**	**0.217**
slp.tif	**0.086**	**0.544**	**0.236**	**0.06**	**0.084**

Table 4: (Continued) The coefficients of correlation between the predicted and observed probabilities for natural wildfire occurrence from LANDFIRE dataset for West North Central Region

		RUN2,	PA3		
	MAXENT	GLM	RF	ANN	MARS
asp_agg.tif	0.004	0	0.147	0.113	0.044
cbd_agg.tif	0.217	0.029	0.29	0.13	0.033
ch_agg.tif	0.002	0.39	0.235	0.179	0.103
dem_agg.tif	0.048	0.035	0.194	0.088	0.045
dist2010.tif	0.038	0.015	0.078	0.022	0.017
evt_agg.tif	0.111	0.229	0.171	0.102	0.242
slp.tif	0.119	0.163	0.242	0.123	0.157
		RUN3,	PA3		
	MAXENT	GLM	RF	ANN	MARS
asp_agg.tif	0.007	0	0.145	0.214	0.029
cbd_agg.tif	0.07	0	0.285	0.004	0
ch_agg.tif	0.048	0.079	0.247	0.339	0.182
dem_agg.tif	0.082	0	0.138	0.209	0.043
dist2010.tif	0.019	0.001	0.047	0.013	0.017
evt_agg.tif	0.146	0.972	0.136	0.053	0.25
slp.tif	0.08	0.048	0.234	0.068	0.11

Table 4: (Continued) The coefficients of correlation between the predicted and observed probabilities for natural wildfire occurrence from LANDFIRE dataset for West North Central Region

COMPOSITE DATASET

For this region, two composite datasets were built. The first dataset consisted of all the LANDFIRE variables plus two BIOCLIM variables selected. The second dataset consisted of two significant variables in LANDFIRE dataset and two variables selected from BIOCLIM dataset. But, both datasets led to similar probability maps of natural wildfire occurring and ROC values of about 0.8. The first dataset had slight higher ROC values than the second dataset. The obtained R values indicated that these variables had little influence on the probabilities of natural wildfire occurring. This might

be caused by the fact that Montana is in the region and in which there was extremely higher frequency of fires than the other states within this region.

CHAPTER 4

CONCLUSIONS AND DISCUSSIONS

5.1 *Conclusions and Discussions*

In this thesis, one of the research questions was to verify if MaxEnt model can be successfully applied to spatially predict natural wildfire occurring in each of the NOAA Climate Regions in the U.S. The results of this study showed that in most of the regions MaxEnt model led to most accurate probability maps of natural wildfire occurring and in other regions this model resulted in the predictions of natural wildfire occurring close to those from the most accurate models.

The second research question was how different five models including MaxEnt, GLM, ANN, RF, and MARS, are for predicting the probabilities of natural wildfire occurring. The results of this study has shown that the probabilities of natural wildfire occurrence produced by GLM and MaxEnt models spatially greatly varied and their spatial patterns were similar only in some of the regions. The similarity may be because MaxEnt is equivalent to GLM model with a logistical regression and Poisson distribution. More often, there were significant differences of probability maps between GLMs and MaxEnt. In the South Region, for example, the probability maps from MaxEnt looked dissimilar from those by GLM model with logistical regression and Poisson distribution.

In some of the regions, although the algorithms and datasets used differed, the obtained probability maps of natural wildfire occurrence looked very similar to each other. In the South region, for example, ANN, RF, MaxEnt, MARS, and GLM with logistical regression generated very similar probability maps. This might be partly

because the vegetation variables in LANDFIRE dataset and climate variables in BIOCLIM dataset were correlated with each other and partly because the relationships of natural wildfire occurrence with the predictor variables were relatively stable.

The last research question was what the main drivers (predictor variables) of natural wildfire occurring are in each of the NOAA Climate Regions. In the regions located in the western portions of the United States, the predicted probabilities of natural wildfire occurrence seemed to be more influenced by moisture relevant variables. Previous studies have also indicated that slight increases in moisture could lead to the increase of natural wildfire occurrence (Batllori 2013). In the West, South West, and South regions, mean temperature of driest quarter was a most significant driver that affects the probabilities of natural wildfire occurrence, suggesting that drought is the most important consideration for these areas. The roles of precipitation and temperature on natural wildfire occurrence need to be studied in more detail.

Mean temperature of wettest quarter and temperature seasonality significantly affected the probability of natural wildfire occurrence in the central and eastern portions of the United States. This suggests that in the areas natural wildfire occurrence was more influenced by extreme temperature than other variables. Global warming due to anthropogenic activities is leading to the increase of temperature and its variation, thus, taking into account the increase of temperature for modeling natural wildfire occurrence in the future studies would be beneficial for the prediction of natural wildfire occurring.

The objective of this study was to demonstrate a relatively simple method to develop the probability maps of natural wildfire occurring that can be used for land use planning based on MaxEnt and other niche modeling techniques. To accomplish this,

this study introduced niche modeling, provided specific instructions on how to use the MaxEnt software, and sample scripts for using several types of niche modeling techniques through the BIOMOD2 package in R software package. These results have shown that niche modeling is an appropriate and cost efficient alternative to having a dedicated wildfire specialist.

This study was unique and conducted at a broad scale, the continental United States. However, the methods used are also appropriate at finer scales. More detailed analysis should be done for the high risk regions. In this study, only the variables that are involved in LANDFIRE and BIOCLIM datasets were considered. In the future studies, other variables such as soil moisture should be added. In addition, in this study only naturally occurring wildfires were considered. However, most of wildfires are caused by human activities. Thus, in the future studies human-induced wildfires such campfires, cigarette-induced fires, and other recreational activity caused fires should be explored.

One of the surprising findings in this study was the results of the Central, Southwest, and Northwest regions. In each of the regions, the spatial patterns of the obtained probability maps looked similar to those of one BIOCLIM variable. For the Central Region, for example, many probability maps were similar to the map of mean temperature of wettest quarter as a continuous variable in terms of spatial patterns, that is, distinct borders existed. The categorization of this continuous variable mitigated the distinct borders. A future study quantifying the difference between categorical and continual variables in the Southwest, Northwest, and Central regions may be beneficial.

5.2 *Limitations and delimitations*

Like many GIS studies, this study suffered from the boundary problem. Also, the demarcation of the climate regions may not be the most appropriate way to divide the natural landscapes. This study did not describe the interactions between the variables. It also did not make any temporal projections, though it provided the groundwork for a future study of climate drivers and wildfire distributions. The data suffered from the modifiable area unit problem that is common to almost any GIS study. There may also be commission and omission errors in the wildfire occurrence data and the presences of the ignition points cannot be verified. As the cause of some fires was not known to be human or natural, the fires with unknown causes were not considered. LANDFIRE data were aggregated to match the spatial resolution of BIOCLIM data, which might have cause uncertainties that have not been studied.

BIBLIOGRAPHY

Arpaci, A, et al. 2014. “Using multi variate data mining techniques for estimating susceptibility of Tyrolean forests”. *Applied Geography*. 53:258-270

Batllori, Enric, et al. 2013. “Climate change-induced shifts in fire for Mediterranean ecosystems”. *Global Ecology and Biogeography*. 22(10):1118–1129

Burns B, et al. 2013. “Non-uniformly under-sampled multi-dimensional spectroscopic imaging *in vivo*: maximum entropy versus compressed sensing reconstruction” *NMR In Biomedicine* 27(2):191-201

Burrows, A, et al. 2008. “Maximum entropy for gravitational wave data analysis: Inferring the physical parameters of core-collapse supernovae”. *The Astrophysical Journal*. 678(2)1142-1157

Cao, Chunxiang, et al. 2013. “Human settlement evaluation in mountain areas based on remote sensing, GIS and ecological niche modeling”. *Journal of Mountain Science*. 10(3):378-387

Chang Y, et al. 2013. “Predicting fire occurrence patterns with logistic regression in Heilongjiang Province, China”. *Landscape Ecology*. 28(10):1989-2004

Chen F, et al. 2015. “Modeling Forest Lightning Fire Occurrence in the Daxinganling Mountains of Northeastern China with MAXENT” *Forests*. 6(5):1422-1438

Rodrigues, M. and J. Riva. 2014. “An insight into machine-learning algorithms to model human caused wildfire occurrence”. *Environmental Modelling and Software*. 57(2014):192-201

Convertino, M, et al. 2013. “Detecting fingerprints of landslide drivers: A MaxEnt model”. *Journal of Geophysical Research: Earth Surface* 118:1367-1386

De Angelis A, et al. 2015. “Modeling the Meteorological Forest Fire Niche in Heterogeneous Pyrologic Conditions”. *PLoS ONE* 10(2): e0116875. doi:10.1371/journal.pone.0116875

Elith, Jane, et al. 2011. “A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*. 17(1):43- 57

Estrada-Pena, et al. 2013. “Assessing the effects of variables and background selection on the capture of the tick climate niche”. *International Journal of Health Geographics*. 12:43

Finney, M. 2004. “FARSITE: Fire Area Simulator—Model Development and Evaluation” *U.S.D.A. Forest Service Rocky Mountain Research Station* RMRS-RP-4 Revised

Finney, M. 2006 “An Overview of FlamMap Fire Modeling Capabilities” *USDA Forest Service Proceedings RMRS-P-41*

Geman, Donald, Heylette Geman and Nassim Nicholas Taleb. 2015. “Tail Risk Constraints and Maximum Entropy”. *Entropy*. 14:1-14

Genton, Marc, et al. 2006. “Spatio-temporal analysis of wildfire ignitions in the St. Johns River Water Management District, Florida”. *International Journal of Wildland Fire* 15:87-97

Goka, Koichi, Sachiko Moriguchi and Manabu Onuma. 2013. “Potential risk map for avian influenza A virus invading Japan”. *Diversity and Distributions*. 19(1):78-85

Graham, Catherine, et al. 2008. “The influence of spatial errors in species occurrence data used in distribution models”. *Journal of Applied Ecology*. 45:239-247

Grendar, Marian. 2001. “What is the question that MaxEnt answers? A probabilistic interpretation. *AIP Conference Proceedings*. 568(1):83-93

Guo, F, et al. 2015. “Gamma generalized linear model to investigate the effects of climate variables on the area burned by forest fire in northeast China.” *Journal of Forestry Research*. 26(3):545-555

Jaynes, E.T. 1957. “Information Theory and Statistical Mechanics”. *The Physical Review*. 106(4):620-630

Lutes D, et al. 2012. FFI: a software tool for ecological monitoring. *International Journal of Wildland Fire* 18(3) 310–314

Lutes D. 2014. “FOFEM 6.1 First Order Fire Effects Model User Guide”. *Rocky Mountain Research Station Fire Modeling Institute.* http://firelab.org/sites/default/files/images/downloads/FOFEM6_Help_May12014.pdf

Liu, Yougqiang, Scott Goodrick and John Stanturf. 2013. “Future U.S. wildfire potential trends projected using a dynamically downscaled climate change scenario”. *Forest Ecology and Management.* 294(15):120-135

Martyushev, LM. 2005. “Maximum entropy production principle in physics, chemistry and biology” *Physics Reports.* 426(1):1-45

Massada, Avi B, et al. 2012. "Wildfire ignition-distribution modelling: a comparative study in the Huron-Manistee National Forest, Michigan, USA". *International Journal of Wildland Fire*. 22(2) 174-183

Merow C, M. Smith and J. Silander. 2013. “A practical guide to MaxEnt for modeling species’ distributions: what it does, and why inputs and settings matter”. *Ecography*. 36(10):1058-1069

Kim, Ho Gul, et al. 2015. “Evaluating landslide hazards using RCP 4.5 and 8.5 scenarios”. *Environmental Earth Sciences*. 73(3):1385-1400

Opperman, T. and K. Ryan. 2013. LANDFIRE – A national vegetation/fuels data base for use in fuels treatment, restoration, and suppression planning. *Forest Ecology and Management*. 294(April):208-216.

Ordóñez C, et al. 2012. “Using model-based geostatistics to predict lightning-caused wildfires”. *Environmental Modelling & Software*. 29(1):44-50

Parisein, Marc-Andre, et al. 2012. "Spatial Variability in Wildfire probability across the western United States". *International Journal of Wildland Fire*. 21(4):313-327

Peters, Matthew, et al. 2013. “Wildfire hazard mapping: exploring site conditions in the eastern US wildland-urban interfaces”. *International Journal of Wildland Fire*. 22(5):567-578

Phillips, S, R. Anderson and R. Schapire. 2006. “Maximum entropy modeling of species geographic distributions”. *Ecological Modelling*. 190(3-4):231-259

Pitchford, J, et al. 2015. “Modeling streambank erosion potential using maximum entropy in a central Appalachian watershed”. Proceedings of the International Associate of Hydrological Sciences. 367:122-127

Pugnet, L, et al. 2013. “Wildland–urban interface (WUI) fire modelling using PHOENIX Rapidfire: A case study in Cavaillon, France” *20TH INTERNATIONAL CONGRESS ON MODELLING AND SIMULATION (MODSIM2013)*. 228-234

Rodrigues, M. and J. Riva. 2014. “An insight into machine-learning algorithms to model human caused wildfire occurrence”. *Environmental Modelling and Software*. 57(2014):192-201

Thompson, Matthew P, et al. 2013. “A risk-based approach to wildland fire budgetary planning”. *Forest Science*. 59(1):63-77

Tribus M. and E.C. McIrvine. 1971. "Energy and information". *Scientific American*. 224(3):179-188

Vasilako C, et al. 2009. “Identifying wildland fire ignition factors through sensitivity analysis of a neural network”. *Natural Hazards.* 50(1):125-143