Building Better Species Distribution Models with Machine Learning:
Assessing the Role of Covariate Scale and Tuning in Maxent Models

by

Cass E. Kalinski

A Thesis Presented to the
Faculty of the USC Graduate School
University of Southern California
In Partial Fulfillment of the
Requirements for the Degree
Master of Science
(Geographic Information Science and Technology)

May 2019

Dedicated to Annette Wagner, my wife, my love, my ever-patient partner in this endeavor. Without her support and encouragement, the journey to this waypoint would never even have started.

"All models are wrong, but some are useful"
George Box

Table of Contents

List of Figures

List of Tables

Acknowledgements

My most profound gratitude to Dr. Karen Kemp for her support leading up to and including the work on this paper. The analysis and modeling skills I learned in her courses set me on the path to the topics covered in this thesis. Her guidance has been invaluable as the thesis unfolded. I will miss the routine of the lively dialogs we had while exploring the twists and turns of this research.

The time and effort provided by my thesis committee members, Dr. Laura Loyola and Dr. Travis Longcore, has been appreciated. Dr. Longcore's experience with Maxent proved helpful at several junctures and nudged the research in better directions accordingly. Dr. Loyola, in addition to the work on this thesis effort, took extra time to advise and guide a somewhat bewildered graduate student back in the first class I took in this program. The seeds planted there have borne fruit. Thank you both.

A thank you to Dr. Chris Daly and Dr. Matt Doggett for the gracious help provided in answering my questions on the PRISM data. They provided background information and insights that I could not have obtained without their willingness to respond to my queries. Their patience with some random GIS student muddling through their climate data is much appreciated.

I would like to thank Dr. Andreas Hamann for the dialogs regarding the ClimateWNA program outputs. His explanation of the limitations of the data were a significant contribution to this paper. Dr. Hamann also provided guidance and feedback on the direction of that part of this work, helping set context and affirming the validity of the questions being explored.

Without the hours of proofreading and editing provided by Annette Wagner, my wife, this paper would have been painful to read. She patiently red-lined my endless battles with tense

and word choice, provided alternatives on structure, and somehow managed to include these distractions into her own, already busy schedule.

These fine people all had a hand in the final product that follows. If there is found to be anything lacking or in error in this paper, it is not for their lack of effort and best intentions. The ownership of any shortcomings is, no doubt, mine.

List of Abbreviations

3DEP	3D Elevation Program
AIC	Akaike Information Criterion
AICc	Akaike Information Criterion, corrected
AUC	Area Under the Curve
CSV	Comma-Separated Values
cWNA	ClimateWNA
DD0	Degree Days Less Than 0°C
DEM	Digital Elevation Model
DLG	Digital Line Graph
ENM	Environmental (or Ecological) Niche Model
FC	Feature Class
GBIF	Global Biodiversity Information Facility
GIST	Geographical Information Science and Technology
HSA	Hot Spot Analysis
HSM	Habitat Suitability Model
MANIS	Mammal Networked Information System
MAT	Mean Annual Temperature
MCMT	Mean Coldest Month Temperature
MWMT	Mean Warmest Month Temperature
or10pct	10% Omission Rate
orMTP	Omission Rate Minimum Training Presence
PCS	Projected Coordinate System

PRISM   Parameter-elevation Regressions on Independent Slopes Model

REMIB   La Red Mundial de Información sobre Biodiversidad

RM   Regularization Multiplier

ROC   Receiver Operating Characteristic

SDM   Species Distribution Model

SSI   Spatial Sciences Institute

TSS   True Skill Statistic

USC   University of Southern California

USGS   United States Geological Survey

VGI   Volunteered Geographic Information

Abstract

Machine learning has emerged as a growing area of interest in species distribution modeling. Maxent is one machine learning tool that has gained wide use in such modeling. Maxent has shown good to superior performance compared to other SDM methods in studies using presence-only species data when the tool is used properly. Often, however, due diligence with the selection of input data and model parameters is neglected, resulting in models of questionable quality. A range of factors need to be considered when setting up Maxent modeling. This study explored two of these. The performance impact of covariate scaling and the results of model tuning on Maxent species distribution models were examined, evaluating two questions related to these factors. Do higher resolution covariates yield a better performing Maxent model of potential habitat extent? Does a tuned Maxent model yield a better performing model of potential habitat than a model using the default Maxent settings? Two approaches to Maxent modeling, default parameters and tuned parameters, were used at two different covariate resolutions, yielding four evaluation models. Presence data for bristlecone pines (*Pinus longaeva*) provided the species example for the evaluation. Covariates were selected that are relevant to the species. These were scaled to match the two study resolutions. Model tuning was performed using the ENMeval R package. Quantitative and qualitative evaluations of the resulting models demonstrated improvements in the model performance in the tuned models. Results from the resolution aspects of the study were less conclusive. Issues with the quality of certain aspects of the climate and elevation data raised questions about the certainty of results at either resolution.

Chapter 1 Introduction

Environmental niche models (ENMs), habitat suitability models (HSMs), and species distribution models (SDMs) collectively are variations on models seeking to answer questions such as where a species might be found, where might its range be extended to, what is its potential range, etc. These models use data on species presence, climate, and environment to estimate habitat ranges for a selected species of interest to determine areas of past, present, or potential species distributions (Franklin and Miller 2014, 1-12). A review of scientific research in the field of SDMs includes a wide variety of methodologies for their application (Elith et al. 2006; Franklin and Miller 2014; Guisan, Thuiller, and Zimmermann 2017).

Among these methodologies is a class of modeling tools derived from the artificial intelligence (AI) and statistics fields called machine learning. Machine learning methods create data classification rules and map functions to fit data patterns by inductively examining the data itself to create algorithms that best describe the data (Elith et al. 2006; Franklin and Miller 2014, 154-155). The machine learns from the data, fits an initial algorithm, then methodically iterates through variations of the algorithm until it is optimized.

Most notably among the machine learning tools available to the SDM field, Maxent (Phillips, Anderson, and Schapire 2006) has emerged as a leading modeling tool for experimentation with SDMs (Elith et al. 2011; Fitzpatrick, Gotelli, and Ellison 2013; Merow, Smith, and Silander 2013). However, use of Maxent default parameters and poor understanding of the data requirements has led to misleading or scientifically unsupportable results (Fourcade et al. 2014; Merow, Smith, and Silander 2013; Morales, Fernández, and Baca-González 2017; Warren and Seifert 2011; Yackulic et al. 2013).

When developing a SDM in Maxent, users must critically evaluate each of the model's required inputs. Decisions include selection of environmental variables (covariates) to use and at what spatial scale; determining the pattern for training data partitioning; and confirming the proper regularization multiplier (RM) and feature class (FC) parameters. Covariates include criteria such as climate, elevation, aspect, slope, landforms, etc. that are pertinent to the species habitat. Climate data at various spatial scales are available, but obtaining resolution relevant to the study area can limit results (Bedia, Herrera, and Gutiérrez 2013). Training data, a subset of the species presence and covariate data used to tune the model, can often contain spatial bias that impacts the model accuracy if not handled properly (Fourcade et al. 2014; Kramer-Schadt et al. 2013). An incorrect RM setting can over or under-fit the model (Warren and Seifert 2011). FC selection, a mathematical transformation of the different covariates, needs to be applied with an understanding of the impact these have on the model complexity and results (Merow, Smith, and Silander 2013).

While Maxent is much lauded for its ease of use, the data preparation and the determination of appropriate model parameters can be challenging (Radosavljevic and Anderson 2014). Tools such as ENMeval (Muscarella et al. 2014), MIAT (Mazzoni, Halvorsen, and Bakkestuen 2015), MIAmaxent (Vollering 2017), and Wallace (Kass et al. 2018) have been developed to assist in decision making for model parameters and to assist with reducing spatial bias in the training data. However, tools and techniques for evaluating the impact of spatial scale of covariates are not well covered in the current literature. Spatial scale, when it is mentioned, is typically in the context of the extent of the study area for the species (Anderson and Gonzalez 2011; Elith and Leathwick 2009; Peterson and Soberón 2012). Bedia, Herrera, and Gutiérrez

(2013) evaluated regional use of different resolution climate models in Maxent, finding that scale in this context has significant impacts on the model validity.

Given all of these challenges related to the proper use of Maxent, the attractiveness of it as an easily implemented modeling tool, and evidence of its widespread misuse, a careful assessment of the impact of model input choices was warranted.

1.1. Thesis Scope

Focusing on key challenges identified in the literature, this study explored the performance impact of covariate spatial scale and the results of model tuning on species distribution models created with a machine learning method. Using Maxent as a representative example of machine learning, two questions related to these factors were evaluated: do higher resolution covariates yield a better performing Maxent model of potential habitat extent and does a tuned Maxent model yield a better performing model of potential habitat than a model using the default Maxent settings. Two approaches to setting up a Maxent model, using default parameters and using tuned parameters, were examined at two different covariate resolutions, yielding four models to evaluate. Specifically, four models of bristlecone pine (*Pinus longaeva*) potential habitat were built using default parameters and tuned parameters at model resolutions of 30-meters and 800-meters.

Presence data for the bristlecones provided the species example for the evaluation. The species habitat is a high elevation, tree line environment. This mountainous terrain includes abrupt terrain changes that challenge the modeling of the covariates and species presence locations. The concise range of the species, a limited set of key environmental factors for its habitat, and the high relief terrain made it a suitable species to model for the study.

Elevation related covariates derived from 30-meter and 800-meter digital elevation models and climate related covariates of the same resolution derived from the PRISM 800-meter climate model were used as inputs for Maxent. For Maxent parameter optimization, ENMeval (Muscarella et al. 2014) was selected to tune the models. ENMeval is a R programming language package that has seen some acceptance in Maxent user circles (Maxent Forum, https://groups.google.com/group/Maxent).

Performance, quantitatively and qualitatively, of the four model versions were compared using a combination of quantitative model metrics and qualitative spatial analysis. A portion of the spatial analysis involved a novel use of hot spot analysis to explore tuning changes between the models.

By using both default Maxent models and models tuned for bristlecone pine related data in the study extent, variations due to parameter setting were isolated and demonstrated alongside the spatial scaling impacts. The hypotheses of the research were that the optimized model would present a more accurate model than the one using default Maxent parameters and that the models based on the 30-meter resolution covariates would perform better than 800-meter resolution models.

1.2. Structure of This Document

This thesis report unfolds through a series of four additional chapters. The next chapter provides context information on SDMs and related data challenges. The Maxent program is discussed as well as model evaluation approaches. The ENMeval utility used for model tuning is introduced. Chapter 3 describes in detail the data used in the study and elaborates on the methods used to create and evaluate the resulting models. Chapter 4 discusses the tuning results and then reviews the results of running the default and tuned models at both the 30-meter and 800-meter

resolution. Quantitative metrics are provided as well as subjective reviews of the Maxent

prediction surfaces. Hot spot analysis, a novel method of spatial analysis of the Maxent outputs,

is introduced. Chapter 5 discusses the conclusions of the study, highlighting the results of the

questions on tuning and resolution core to the study.

Chapter 2 Background and Context

This chapter sets the context of the study by providing background information on species distribution models (SDMs), data challenges for these models, and, more specifically, the use of Maxent for SDM development. Key Maxent parameters are explained, with a focus on their influence on the modeling. SDM model evaluation is covered, providing background for the approach chosen in this study. The chapter closes with the introduction of the ENMeval tool used to tune the study's Maxent models.

2.1. Species Distribution Models

As mentioned previously, SDMs, ENMs, and HSMs, among other names, are variations on models seeking to answer questions such as where a species might be found, where might its range be extended to, what is its potential range, etc. (Franklin and Miller 2014, 1-12). There is debate surrounding the nomenclature and taxonomy of these various models, with advocates arguing for various nuances in definition (Araújo and Peterson 2012; Franklin and Miller 2014, 5-7; McInerny and Etienne 2013; Warren 2012, 2013). The questions of spatial scale and model tuning examined in this study are relevant across this spectrum of definitions. For the purposes of this study, therefore, the term *species distribution model* is used generically.

The field of species distribution modeling has expanded tremendously in the last several decades. Advances in techniques, technical capabilities, and a massive increase in the volume of data available has contributed to this expansion. Even more so, the need for the models and the information they represent has grown exponentially. A subject that was once primarily focused on the understanding of species geography now addresses a wide range of purposes (Franklin and Miller 2014, 7-20). SDMs assist in not only the understanding of species distributions, they have become critical tools for environmentalists, conservationists, biogeographers, governments,

agencies, and climatologists for a broad range of endeavors. SDMs assist in managing

endangered species, mitigating invasive species, guiding habitat restorations, setting policies for

urban development and resource extraction, evaluating the impacts of climate change, revealing

interactions between ecosystems, and a host of other applications across a range of fields and

interests. They have become foundational in many of these areas.

SDMs are powerful and prevalent tools. They are, however, mere approximations of

reality. As George Box, a pioneer in modeling theory once stated, "All models are wrong but

some are useful" (Box and Draper 1987). Modeling species and their interactions with the

environment is complex beyond measure. Scientists readily admit they do not fully understand

all the possible facets of these interactions. SDMs, therefore, should be considered as guidance

for exploration and understanding, not as absolute truths. The goal of modeling practitioners,

therefore, is to strive for Box's "some are useful".

2.2. Maxent

A wide range of tools and techniques have been developed for SDMs (Elith et al. 2006;

Franklin and Miller 2014, 113-205; Guisan, Thuiller, and Zimmermann 2017, 151-236; Peterson

et al. 2011, 101-111), with Maxent (Phillips, Anderson, and Schapire 2006) emerging as a

leading choice in this area (Elith et al. 2011; Fitzpatrick, Gotelli, and Ellison 2013; Merow,

Smith, and Silander 2013). While sometimes criticized for its misuse and for being too "black

box" (e.g. Halvorsen et al. 2015; Morales, Fernández, and Baca-González 2017; Phillips et al.

2017), Maxent's popularity is due in great part to its ease of use and to its proven performance

potential for many SDM scenarios (Elith and Graham 2009; Elith et al. 2006).

The tool borrows from the field of machine learning, applying the principles of maximum

entropy to species distribution models. Maximum entropy seeks to define a probability

distribution that is closest to uniform given a set of constraints in the form of covariates and species presence locations. In other words, Maxent examines the covariate values and their relationships at each of the species presence locations, then creates a set of transformations (i.e. formulas) that best represents this data across the extent of the study area (Phillips, Anderson, and Schapire 2006). The number of transformations can often exceed the number of covariates as different combinations and coefficient values are joined into new transformations in an attempt to find the optimum model fit. The output is a relative probability surface for the species, indicating areas more and less likely to align with the species preferences given the input covariates.

Maxent's power, ease of use, and the complexity of the underlying model have led to issues with its use. Use of Maxent's default parameters without due diligence in assessing their applicability, combined with poor understanding of the data requirements, has often led to misleading or scientifically questionable results (Fourcade et al. 2014; Merow, Smith, and Silander 2013; Morales, Fernández, and Baca-González 2017; Warren and Seifert 2011; Yackulic et al. 2013). When developing a SDM in Maxent, users must critically evaluate each of the model's required inputs. Decisions include selection of environmental variables (covariates); determining what spatial scale is needed; assessing the pattern for model training partitioning; confirming the proper regularization multiplier (RM) and feature class (FC) parameters.

Covariates can include criteria such as climate, elevation, aspect, slope, landcover, geology, soil composition, etc. that are pertinent to the species' habitat. Covariate data at various spatial scales is available and widely used. However, this data, in some sense, is too readily available. Studies often take a "shotgun" approach to the covariates in Maxent studies, applying a wide range of often correlated covariates just because they are available. Maxent does employ

mechanisms to disregard superfluous covariates, but as Fourcade, Besnard, and Secondi (2017) demonstrated using a random painting as a source of covariate information, these mechanisms do not insure the model will yield biologically relevant results. A more robust approach is to use covariates selected for relevance to the species under study (Fourcade, Besnard, and Secondi 2017; Guevara et al. 2018).

Two concerns come into play when the question of scale is approached for an SDM. What is the appropriate scale for the species under study? And, what is required to get data at that resolution? For the species scale, is its domain regional? Continental? Gentle terrain or mountainous? For the resolution, is there high-quality data available at the necessary scale? If not, what processing is required to bring the data to the resolution? What quality and uncertainty issues does this processing bring to the model? All too often, studies simply take datasets that are available without understanding or accounting for the quality lineage of the data (Connor et al. 2017; Fournier et al. 2017; Franklin et al. 2012; Manzoor, Griffiths, and Lukac 2018). All data aggregation, disaggregation, and interpolation methods, by their nature, alter the source data as it is transformed from one resolution to another. Care must be taken in the choice of methods used to minimize distortions and to understand what potential impacts the transformation may have on the model.

The process of building a Maxent model involves segregating the species locations and the background covariate data into training and testing partitions in some pattern. The training data is used by Maxent to create the model. The withheld test data, as the name implies, is used to validate and evaluate the fitness of the model on data the model has not yet seen. Both the species presence location distribution and the covariate data often contain spatial bias that impacts the model accuracy if not handled properly (Anderson and Gonzalez 2011; Boria et al.

2014; Fourcade et al. 2014; Kramer-Schadt et al. 2013). The test/train process must account for this bias and either reflect it in the partitioning or use partitioning methods to mitigate undesired impacts.

The sophistication of the underlying maximum entropy approach in Maxent can lead to overly complex and overfit models (Elith et al. 2011; Merow, Smith, and Silander 2013; Warren and Seifert 2011). Maxent uses a regularization multiplier (RM) parameter to apply smoothing to the model, allowing the user to tune the level of fitting in the model. Higher regularization values loosen the maximum entropy requirements for each of the transformation variables, increasing the acceptable range of the coefficients. It also applies penalties to the large magnitude coefficients, driving down the complexity of the model as necessary. Experimenting with multiple RM settings and evaluating the model performance to determine the proper setting given the species data and study goals is an often-overlooked aspect of using Maxent (Radosavljevic and Anderson 2014; Warren and Seifert 2011).

Feature classes (FCs) in Maxent are the mathematical transformations used on the different covariates to create the model's distribution surface. Maxent currently has five FCs available: linear, quadratic, product, hinge, and threshold. These are used in isolation or, more commonly, in combination to create the set of transformations required for a well fit model. By default, Maxent uses the species location count to select the FCs set (Phillips, Anderson, and Schapire 2006). Linear is always used. Quadratic requires a minimum of 10 species presence locations and hinge requires 15. Threshold and product only come into play when the location count exceeds 80. Allowing Maxent to choose the FC combination to use can be beneficial. The tool will consider these in combination with the covariates provided to optimize the fit of each covariate to the model. Indeed, Maxent can be used to explore covariate responses and

interactions this way, leading to an iterative approach in further culling of covariates to simplify the model. There is some risk the FC may not reflect the actual biological response dynamics of the covariate (Merow, Smith, and Silander 2013). Covariate response curves should be scrutinized accordingly.

While the creators of Maxent set the default parameters to what they considered an optimal mix, they also caution that each species dataset is unique (Phillips 2017; Phillips, Anderson, and Schapire 2006). At a minimum, testing of the RM settings is required to insure proper fitting of the model. Covariates and FC settings should be selected based on their relevance to the species under study (Fourcade, Besnard, and Secondi 2017; Guevara et al. 2018). Maxent is a powerful and well performing tool for the SDM modeler. Like any powerful tool, it must be approached with an understanding on how it can be used properly. High performance results can be expected with the correct amount of due diligence.

2.3. Presence Data

Presence data are, obviously, the most needed element of an SDM. Presenting relevant known location data for the species under study to the model and analyzing the covariates associated with those presence locations enables the model to predict potential habitat within the study extent. Without reasonably accurate, representative presence data for the species, the models would be meaningless.

Acquiring presence data specifically for a study can be time and resource intensive. A common approach to mitigate these costs is for researchers to use existing datasets from prior fieldwork. One type of data source that has grown in acceptance in recent years are collaborative biodiversity data networks such as the Global Biodiversity Information Facility (GBIF), the Mammal Networked Information System (MANIS), The World Information Network on

Biodiversity (REMIB; La Red Mundial de Información sobre Biodiversidad), and the European

Natural History Specimen Information Network, to name a few (Graham et al. 2004). These

sources are typically not databases themselves. Instead, they are portals accessing a network of

species data provided by a diverse range of sources: prior studies, volunteered geographic

information (VGI), universities, and natural history collections are common. With caveats, these

sources have significantly enhanced the abilities of researchers to carry out species studies that

would have been difficult or logistically impossible without the data provided (Edwards 2004;

Graham et al. 2004; Sobern and Peterson 2004).

The sources do have limitations and issues that must be considered. The range of taxa

covered, while expanding as more institutions come online, is limited. Many networks have

strong regional bias in the sample sets that can skew studies of broader extent (Beck et al. 2014;

Boakes et al. 2010). The most pressing issue is data quality. Geographic inaccuracy or ambiguity

of the location data, misidentification or outdated taxonomy of the specimens, and sampling bias

are common issues (Beck et al. 2013; Beck et al. 2014; Boakes et al. 2010; Graham et al. 2004;

Sobern and Peterson 2004). The data received from these networks must be carefully analyzed

for fitness-of-use for the species and area being studied.

GBIF was the source of presence data for the bristlecone pine (*Pinus longaeva)* examined

in this study. GBIF is currently the largest single biodiversity portal for occurrence data (Beck et

al. 2013; Yesson et al. 2007) and provides a plentiful set of presence locations for *P. longaeva*.

While the overall GBIF dataset has all the issues noted above, regional bias and taxonomy issues

are not significant factors in the GBIF data for this species and the study area. With proper

adjustments and detailed data cleansing, GBIF has been shown to be an excellent source of

quality data (Yesson et al. 2007). Chapter 3 elaborates on the quality of and the cleansing

performed on the GBIF data for *P. longaeva* to yield data within the well-defined study region.

2.4. Climate Data

Climate data used in SDMs such as precipitation and temperature are derived from

climate models. Climate models project climate patterns spatially based on general circulation

models and additional physiographic features such as elevation, nearness to coastal features, and

rain shadows, among others. These models are typically global in nature, though continental

scale models are common and regional models have been developed. A full explanation of this

highly complex subject is beyond the scope of this current work. Creating accurate climate

models relies on a wide variety of factors, not all of which are as richly represented in the

underlying data as would be desired. Even the best climate models contain gaps and limitations.

(Araújo and Peterson 2012; Daly 2006; Raäisaänen 2007)

This presents issues for SDMs as they rely heavily on climate model data for covariates.

Climate models are often broad scale given their global and continental scope. This may suffice

for species studies at similar scale, but often SDMs are more narrowly focused in spatial extent.

The climate models must, therefore, be disaggregated or interpolated to the scale needed for the

species study. Disaggregation risks missing possible local effects. Interpolation introduces yet

further processing of data that may already be lacking key component details. Both introduce

uncertainty into the SDMs that is extremely difficult to quantify (Daly 2006).

Two commonly used datasets for SDMs are WorldClim and PRISM. Both offer high

resolution (for climate models) datasets in roughly the $1km^2$ range. WorldClim is often preferred

in SDM studies as it has a rich set of climate variables that can be examined in the modeling.

However, WorldClim has issues in certain topographical situations, including mountainous

terrain such as the study extent of this project. Hijmans et al. (2005) highlights this deficiency and warns against its use in extents that include abrupt topographical transitions. PRISM, on the other hand, attempts to recognize and compensate for mountainous terrain in those regions where it is warranted. The PRISM model uses a combination of approaches to build its climate projections. Besides the base climate algorithm, expert guided adjustments and compensating algorithms for areas of temperature inversion, rain shadows, and coastal effects are applied. The climate data provided are generally more accurate for difficult terrain areas (Daly and Bryant 2013; Daly et al. 2000).

All climate datasets are approximations and inherently have significant unquantifiable uncertainties for their use in SDMs. Modeling results should be viewed accordingly. While introducing uncertainty in the model, the climate covariate data are held as a constant in this study. The elevation and aspect covariates drive the study's scale variability. Any uncertainty introduced from the climate covariates from a species habitat perspective should be consistent between the model evaluations in this study, allowing for isolation of the spatial scale questions.

2.5. Evaluation Metrics

Comparing metrics across SDMs is an area of study that continues to elude specific solutions. Controversy surrounds many of the widely used SDM metrics such as area under the curve (AUC; aka receiver operating characteristic or ROC), kappa, true skill statistic (TSS), and Akaike information criteria (AIC). There is no one metric that indicates a "right model" (Allouche, Tsoar, and Kadmon 2006; Galante et al. 2018; Jiménez-Valverde 2011; Leroy et al. 2018; Lobo, Jiménez-Valverde, and Real 2007; Peterson, Papeş, and Soberón 2008; Peterson et al. 2011; Warren and Seifert 2011). Each of the metrics must be evaluated in context, with its merits and limitations considered. When assessing SDMs, the model suitability is best

determined by a mix of evaluation metrics (Jarnevich et al. 2015; Muscarella et al. 2014; Radosavljevic and Anderson 2014).

AUC, the most widely used SDM metric, quantifies whether the model correctly orders random presence data versus random background data. In other words, the AUC score indicates how well a model can discriminate presence locations from random background locations. A higher AUC number indicates better discrimination performance of the model. AUC values below 0.7 indicate poor model performance, between 0.7 – 0.9 is considered moderate performance, and values above 0.9 are consider high performance (Franklin and Miller 2014, 223).

AUC is not without its criticisms. As it relies heavily on prevalence, comparing AUC scores across models of different species with different prevalence rates is not recommended as the measures will be skewed for comparison purposes. The mapping of sensitivity and threshold inherent in the measure is not always applicable to biological studies, especially on the tail ends of the curve. The metric treats omission and commission errors with equal weight, an approach that introduces some bias into measures of presence-only models such as MaxEnt. AUC is also a poor indicator of model fit. (Jiménez-Valverde 2011; Lobo, Jiménez-Valverde, and Real 2007; Peterson, Papeş, and Soberón 2008). While useful, AUC should always be used in concert with other measures of model performance.

Kappa is another widely used metric for SDMs. Kappa measures agreement between predictions and observations, both presence and absence observations in the case of SDMs, with a correction applied for the amount of agreement expected to occur by chance. It thus incorporates omission and commission error into one statistic. The value range for a kappa score is from -1 to +1, where +1 indicates perfect agreement and values zero or less indicate no better

than random chance. The measure does require establishing a threshold value for its measurements, typically set at 0.5.

Kappa has been shown to be impacted by the level of species prevalence and range size of the study area, resulting in bias in the results (Allouche, Tsoar, and Kadmon 2006; McPherson, Jetz, and Rogers 2004). The metric does consider both presence and absence occurrences in its calculations, with its primary use being presence-absence SDMs. While work arounds exist, its appropriateness for presence-only models such as Maxent is marginal.

TSS was introduced by Allouche, Tsoar, and Kadmon (2006) as an alternative to kappa that addresses the prevalence issues. In the authors' own words, "TSS is a special case of kappa", retaining all of the benefits of that metric while mitigating the prevalence issue. Some minor criticism of TSS exists, but the issues seem to be corner cases rather than systematic (Leroy et al. 2018).

AICc (Akaike information criterion with correction) is a comparison of model complexity and fit. It is a variation of the full AIC metric, with adjustments to the formula to account for small sample sizes (Burnham and Anderson 2003, 66). Under the statistical principle of parsimony, AICc seeks to strike a balance between too little and too much model complexity, while at the same time considering the overall fit of the model. Models with lower AICc scores are desirable, all other factors being the same, as they indicate lower complexity and better fit. The limitation of AICc is that comparisons of the metric can only be used on models of the same dataset. AICc values across model resolutions, as done in this study, are not comparable. AICc is, however, one of the stronger, independent measures of model fit and performance (Galante et al. 2018; Muscarella et al. 2014; Warren and Seifert 2011).

2.6. ENMeval

      While Maxent has been shown to be effective in estimating potential species habitat (Elith et al. 2006; Phillips, Anderson, and Schapire 2006), determining appropriate Maxent program parameters to avoid overfitting, determining optimum model complexity, and dealing with sampling bias of the species locations can be problematic (Merow, Smith, and Silander 2013; Radosavljevic and Anderson 2014).

      Several methods are available for building and tuning Maxent models. Besides the Maxent native user interface and batch mode processing, tools are available in the R programming environment (R Core Team 2018). ENMeval (Muscarella et al. 2014), MIAMaxent (Vollering 2017), MIAT (Mazzoni, Halvorsen, and Bakkestuen 2015), and Wallace (Kass et al. 2018) are a few examples. High quality packages such as dismo (Hijmans et al. 2017) and biomod2 (Thuiller et al. 2016) are widely used in the SMD field. However, these packages often require crafting together multiple pieces of code to create a model tuning and evaluation workflow.

      ENMeval incapsulates much of this workflow, leveraging the dismo and biomod2 packages as dependencies. ENMeval streamlines the data processing and metric generation of the tuning exercise tremendously, saving time and ensuring quality metric data for decision making. ENMeval incorporates the findings of Radosavljevic and Anderson (2014), providing a variety of the test/training partitions (Table 1 and Figure 1) and tuning options. The methods mitigate fitting and sample bias issues. The partitioning schemes in combination with the range of RM and FC settings available allow for multiple views of the ensuing models. The suite of metrics generated (Table 2) provides a quantitative basis for model tuning selections.

Table 1 - ENMeval partitioning schemes

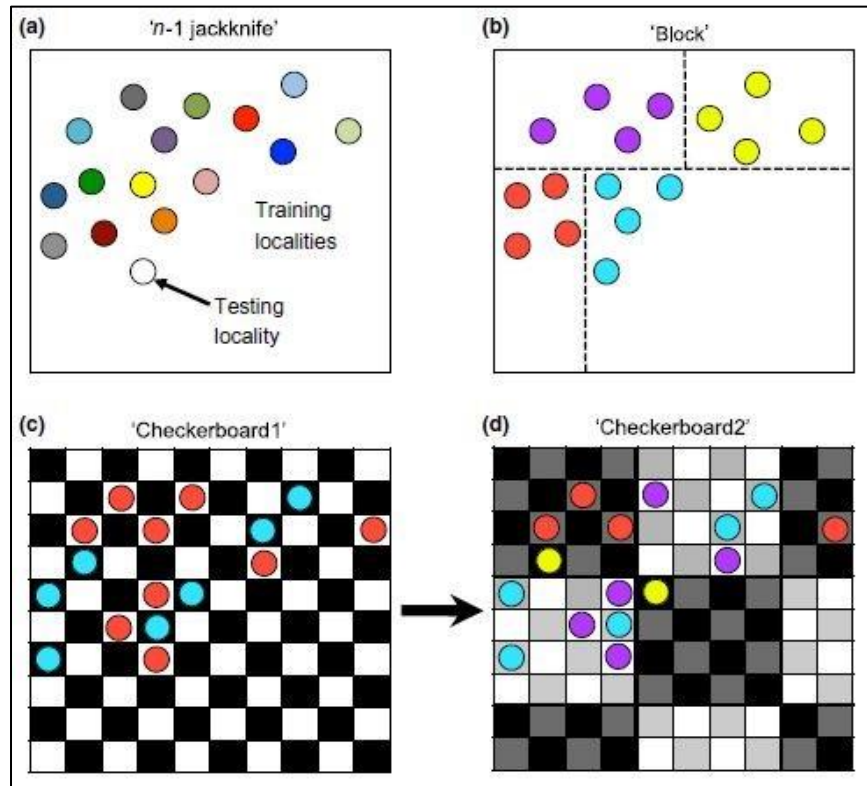| Partitioning Scheme	Description
Jackknife	Each of n presence locations used once for test. Results averaged across all *n*.
Random k-fold	Locations randomly placed into *k* number of bins (*k* set by user). Equivalent to Maxent's cross-validate partitioning.
Block	4 spatial bins with equal number of locations in each bin.
Checkerboard1	2 bins, but in a checkerboard/chessboard pattern. Uneven location distribution.
Checkerboard2	4 bins, with and additional layer of aggregation from checkerboard1.
User defined	User creates spatial and location partitioning based on study requirements.

Figure 1 - ENMeval partitioning examples (Source: Muscarella et al. 2014)

Table 2 - Evaluation metrics from ENMeval

Metric	Description	References
AUC_{TEST}	The threshold-independent metric AUC based on predicted values for the test localities (i.e. localities withheld during model training), averaged over k iterations. Higher values reflect a better ability for a model to discriminate between conditions at withheld (testing) occurrence localities and those of background localities (by ranking the former higher than the latter based on their predicted suitability values). The rank-based AUC does not indicate model fit	Hanley & McNeil (1982), Peterson et al. (2011)
AUC_{DIFF}	The difference between the AUC value based on training localities (i.e. AUC_{TRAIN}) and AUC_{TEST} ($AUC_{TRAIN} - AUC_{TEST}$). If $AUC_{TRAIN} < AUC_{TEST}$, the returned value is zero. Value of AUC_{DIFF} is expected to be positively associated with the degree of model overfitting	Warren & Seifert (2011)
OR_{MTP} ('Minimum Training Presence' omission rate)	A threshold-dependent metric that indicates the proportion of test localities with suitability values (MAXENT relative occurrence rates) lower than that associated with the lowest-ranking training locality. Omission rates greater than the expectation of zero typically indicate model overfitting	Fielding & Bell (1997), Peterson et al. (2011), Radosavljevic & Anderson (2014)
OR_{10} (10% training omission rate)	A threshold-dependent metric that indicates the proportion of test localities with suitability values (MAXENT relative occurrence rates) lower than that excluding the 10% of training localities with the lowest predicted suitability. Omission rates greater than the expectation of 10% typically indicate model overfitting	Fielding & Bell (1997), Peterson et al. (2011)
AICc	The Akaike Information Criterion corrected for small samples sizes reflects both model goodness-of-fit and complexity. The model with the lowest AICc value (i.e. $\Delta AICc = 0$) is considered the best model out of the current suite of models; all models with $\Delta AICc < 2$ are generally considered to have substantial support	Burnham & Anderson (2004), Warren & Seifert (2011)

(Source: Muscarella et al. 2014)

ENMeval outputs includes additional metrics besides the traditional AUC and AICc metrics (Table 2) used to evaluate SDMs. Omission rate minimum training presence (orMTP), 10% training omission rate (or10pct), and the AUC difference between training and test data (diff.AUC) are provided as indicators of potential overfitting. Both the orMTP and or10pct metrics are threshold measures, suggested by Radosavljevic and Anderson (2014) as indicators of overfitting, with orMTP indicating the proportion of species presence locations in the test data falling below the lowest ranking training locations. Likewise, the or10pct sets the test threshold at the 10% level of the training data. The diff.AUC measure is simply the difference between the AUC of the training data less the AUC of the test data (Warren and Seifert 2011). High diff.AUC values indicate the model is overfit.

The ENMeval package, through the dependent dismo R package (Hijmans et al. 2017), uses the same dataset as the final Maxent model and creates the models using the Maxent Java application. Therefore, there is no difference in the models created by ENMeval versus those

created with the native Maxent user interface or Maxent's batch mode. The package facilities

running multiple parameter and partitioning combinations in parallel and provides consolidated

performance metrics to guide parameter selection for a tuned model. ENMeval is well accepted

within the Maxent user community (Maxent Forum, https://groups.google.com/group/Maxent), is

well supported, and is a relatively mature implementation compared to the other options.

2.7. Summary

This chapter briefly reviewed the role of SDMs in species habitat research and Maxent's

place in the collection of tools available for that research. The data required to create an SDM

were explored, along with the challenges of choosing performance measures of the resulting

models. The Maxent program was explained, followed by an introduction of ENMeval for

Maxent tuning. In the next chapter, the application of the information provided in this chapter

will provide the context as the data choices and methodologies specific to this study are

explained.

Chapter 3 Data and Methods

This chapter describes the data and methods used for this study. The chapter begins with an overview section to provide context for the chapter before moving into the detailed sections on data, modeling, and evaluations. The datasets are then reviewed, including discussion of the data preparation steps. Next, the methodologies for building the default and tuned Maxent models are described. Finally, the model evaluation procedures and metrics are discussed before transitioning to Chapter 4 Results.

3.1. Data and Methods Overview

The project was comprised of four major phases (Figure 2). Data were gathered and prepared, the model tuning requirements analyzed, the models processed, and then the models were evaluated.
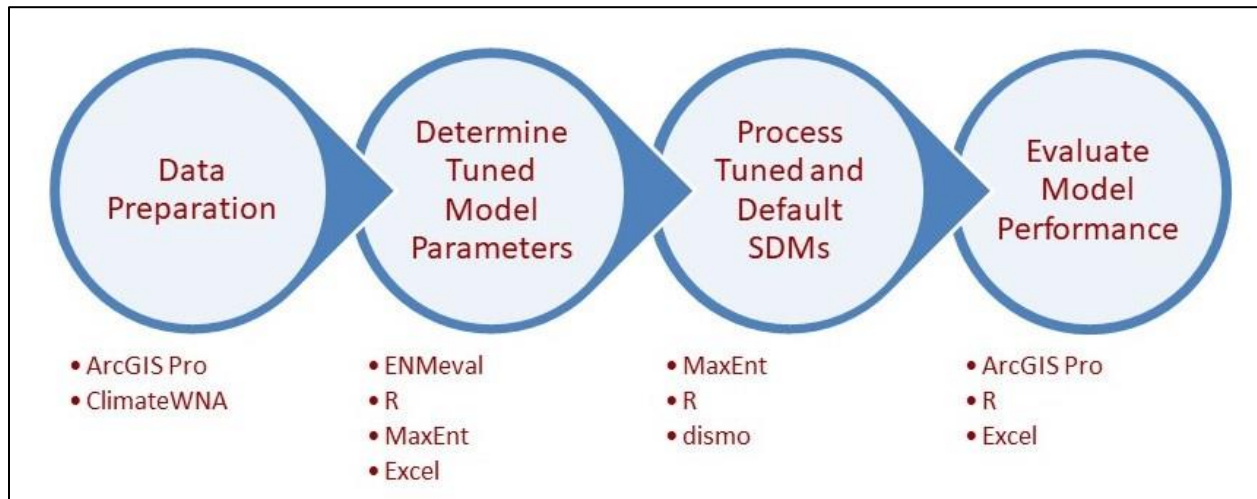
Figure 2 - Project phases and major tools

Two types of models were built in Maxent: one using the Maxent defaults; the other using tuned Maxent performance parameters. Of these types, models were generated at two different resolutions (30-meter and 800-meter), for a total of four models (Figure 3).
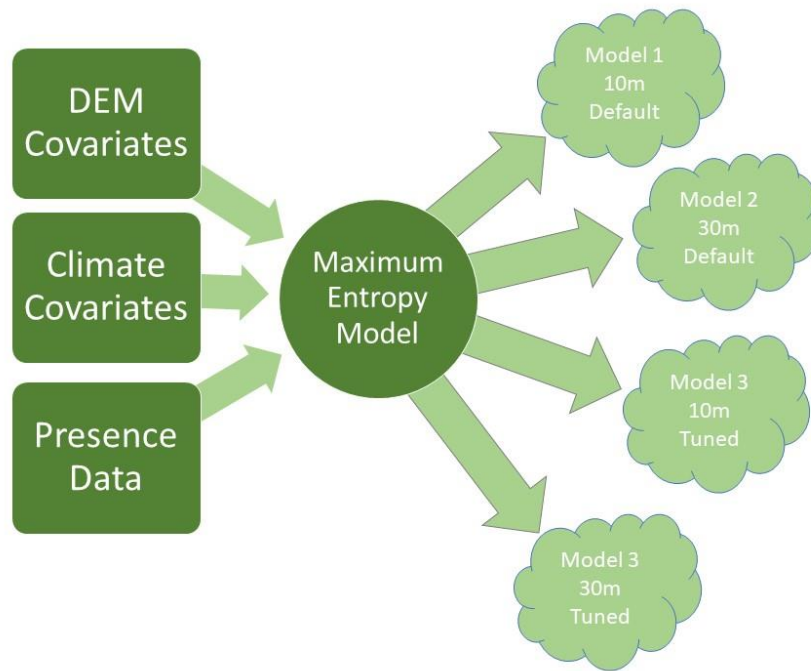
Figure 3 - High level modeling process

Presence data for the target species (bristlecone pine, *Pinus longaeva*) were acquired from the GBIF (GBIF 2018). A study area was established, with a corresponding subset of the presence data used to build the models. The models used two different digital elevation models (DEMs) to generate elevation, curvature, and aspect covariates at 30-meter and 800-meter resolutions. Aspect was expressed as northness (cosine of aspect degrees) and eastness (sine of aspect degrees). Climate covariates at matching resolutions were generated by ClimateWNA software (cWNA; Wang et al. 2016). ClimateWNA interpolates PRISM 800-meter climate data to finer resolutions and calculates several useful climate variables. Climate covariates chosen for this study were mean annual temperature (MAT), mean warmest month temperature (MWMT), mean coldest month temperature (MCMT), and degree days less than 0°C (DD0 or DD<0). Covariate and presence data were prepared for Maxent input using ArcGIS Pro.

Figure 4 shows the high-level flow of the data preparation and model methodology. The Maxent models were built and evaluated using various R programming tools (see Appendix A Software Used in The Study). The R package ENMeval (Muscarella et al. 2014) was used to determine the tuned model parameters. The R environment interfaces with the Java version of the Maxent program. The programming environment streamlines the tuning, model creation, and data collection processes. The models and output rasters are, however, from the Maxent program.
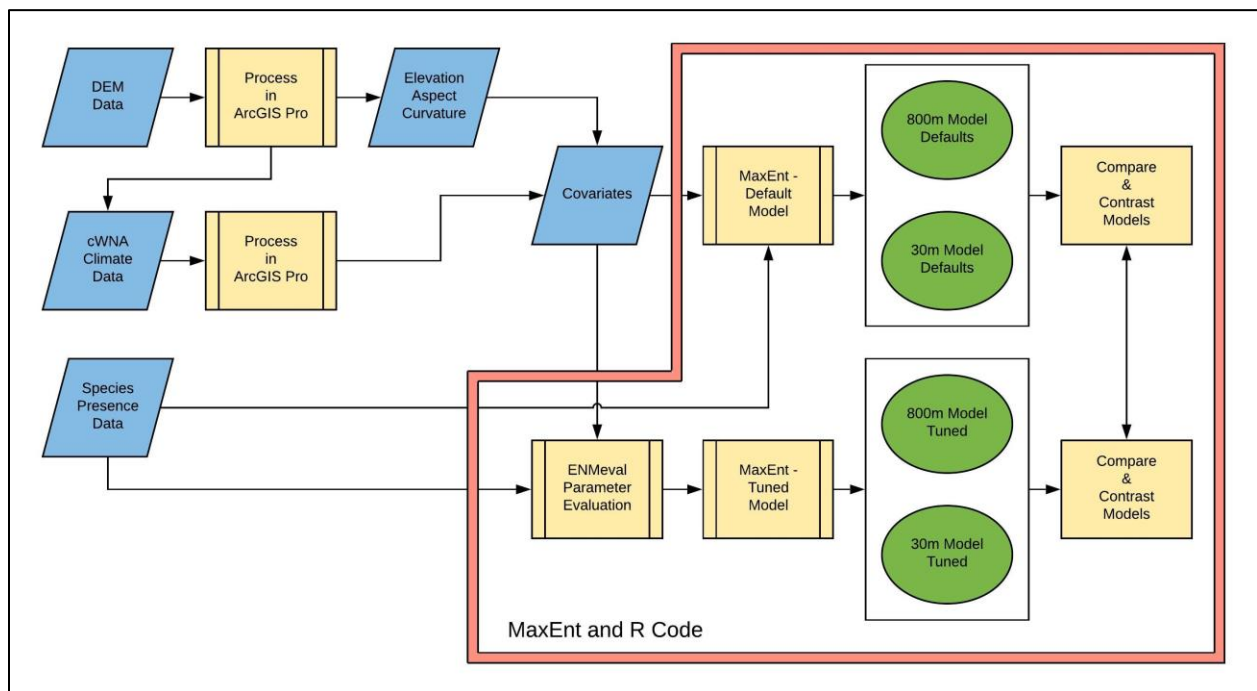
Figure 4 - High level flow of data preparation and methodology

Three types of evaluations were performed on the models. First, metrics were generated from the Maxent models. These included area under the curve (AUC), Akaike information criterion with correction (AICc), omission rate at minimum training presence (orMTP), and omission rate at 10 percentile training presence (or10pct). As an additional measure of performance, the models were used to predict species distributions in a Utah study area and the

AUC measured. This provided an evaluation of how well the model can be applied to other habitat areas for the species. Finally, the prediction surfaces from each of the models were compared for a subjective spatial evaluation. A hot spot analysis was performed to study the changes between the tuned and default model surfaces.

Elaborations on each of these areas of data, modeling methodology, and evaluation are provided in the following sections.

3.2. Data

In a typical use of a Maxent model, species presence data are combined with key environmental factors (i.e. covariates) to predict potential species distribution. While Maxent can be used to algorithmically recommend a set of covariates from a large set input into the model, it is recommended to select covariates based on known ecology for the species under study (Fourcade, Besnard, and Secondi 2017; Guevara et al. 2018).

In the case of bristlecone pines used in this study, key covariates include elevation, aspect, curvature, growing season, and temperature (Bruening et al. 2017; Bruening 2016). Other factors are likely involved besides these named covariants. For example, more sophisticated factors include microclimate/topoclimatic variances or inter-species dependencies (Peterson and Soberón 2012). However, these more sophisticated factors are beyond the scope of this study, as is determining the full range of possible factors. Studies of Bristlecones indicate the covariates outlined above are the main drivers for bristlecone pines and similar high elevation species (Bruening et al. 2017; Coops and Waring 2011; Körner 2012).

Four primary datasets were used in the study. Each is described in detail in subsequent subsections.

- GBIF species presence data for bristlecone pines (GBIF 2018)

- ClimateWNA climate covariates at 30-meter and 800-meter resolutions (Hamann et al. 2018)

- USGS 3D Elevation Program (3DEP) 1arc-sec DEM (USGS 2017)

- PRISM 800-meter elevation data (PRISM Climate Group 2018)

All datasets were projected to NAD 1983 UTM Zone 11N (California locations) or 12N (Utah locations) for the modeling. The projection preserves areas and has minimal distortion of shape within the bounding parallels of the projection (Esri 2000). While Maxent accepts a variety of projections, the use of this projected coordinate system (PCS) allowed for an evenly spaced grid network, simplifying portions of the processing. Cell sizes, for example, in the Utah and California study areas were consistent, with limited distortions due to differences in latitude.

ArcGIS Pro software (Esri 2018) was used for the preparation of the species data and the covariate rasters used as input for the models. For specific methods used for the data preparation see Appendix B Data Preparation Procedures.

3.2.1. Species Presence Data

The example species used for this study was bristlecone pine (*Pinus longaeva*). Bristlecones are a tree line species inhabiting mountainous terrain within the interior of the western United States.

Figure 5 - Bristlecone Pine, White Mountains, CA (Photo credit: C. Kalinski)

The species was chosen for four reasons: 1) its habitat consists of high relief terrain; 2) the species has a fairly concise range; 3) the species has a small set of key climate variables; 4) sufficient occurrence data for the species were available. The high relief terrain served to highlight the potential impacts of covariate resolution at 800-meter versus 30-meter. The concise range reduced processing time of the models and allowed attention to a relatively small study area. With the focus of the study being on the impacts of resolution rather than finding a definitive range for the species, the reduced number of climate variables needed simplified the study approach.

The species is especially long lived, with lifespans up to five thousand years (Currey 1965). Experts speculate the changes in habitat are measured at the century scale for these species and there is a significant lag time in range changes given the life cycle of the trees. Prior studies indicate the current habitat extent likely represents stabilization circa 1380 CE per tree ring data and has not fully adjusted to current climate conditions (Bruening et al. 2017; Lloyd and Graumlich 1997; Scuderi 1987). While some adjustments were made in the modeling to

partially account for this, the lag should be kept in mind when viewing habitat extents in this study. The lag is applicable across the study data, so the relative results of the model comparisons are still valid.

Presence only data were acquired from the Global Biodiversity Information Facility (GBIF). GBIF is a collection of pooled datasets. Most of the historical data in the GBIF dataset used in this study are from field observations by scientists, university studies, and collection information from museum specimens. More recent data include a large proportion of iNaturalist/citizen data. For example, data from 2011-2017 shows 30 of the 33 observations are from iNaturalist. The iNaturalist data are from geotagged photos submitted through an application on volunteer's cell phones. There is some general risk of misidentification with iNaturalist data as expertise is more limited than trained scientists doing field work. However, the risk in the context of this study is minimal. The locations presented appear to be within the bounds of other trusted data sources.

3.2.1.1. Presence Data Cleansing and Preparation

As advocated by Yesson et al. (2007), a conservative approach was taken with the presence data selected for the study. A balance was struck between the need to have an adequate sample count against the need to avoid introducing redundant or questionable quality data. Cleansing criteria are outlined in Table 3. The full download of raw GBIF data for the species was 362 records. Cleansing of the data netted 180 presence-only locations for consideration. This pool was assembled into a comma-separated values (CSV) file. ArcGIS Pro was used to create point features from the CSV file for further processing.

Table 3 - Presence data cleansing

Reason for Removal	Count
Location outside of study area (i.e. New Zealand, Berkeley)	5
No observation date	23
Coordinate uncertainty >1km	87
Collection comments question observation validity	6
Duplicate records (i.e. same date and coordinates)	61
Total Records Removed	182

The pool of cleansed records fell into roughly four major habitat areas in the western contiguous United States (Figure 6). The species occupy high elevation niches in the mountains of eastern California (Sierra Nevada, White Mountains, Inyo Mountains, and Panamint Range); the Great Basin region of Nevada; the Mount Charleston area near Las Vegas, Nevada; and in the mountain ranges of Utah. Counts for the areas are listed in Table 4.

Table 4 - Bristlecone pine major habitat areas' presence location counts

Area	Count
Easthern California	109
Great Basin, NV	23
Mount Charleston, NV	19
Utah	29
TOTAL	**180**
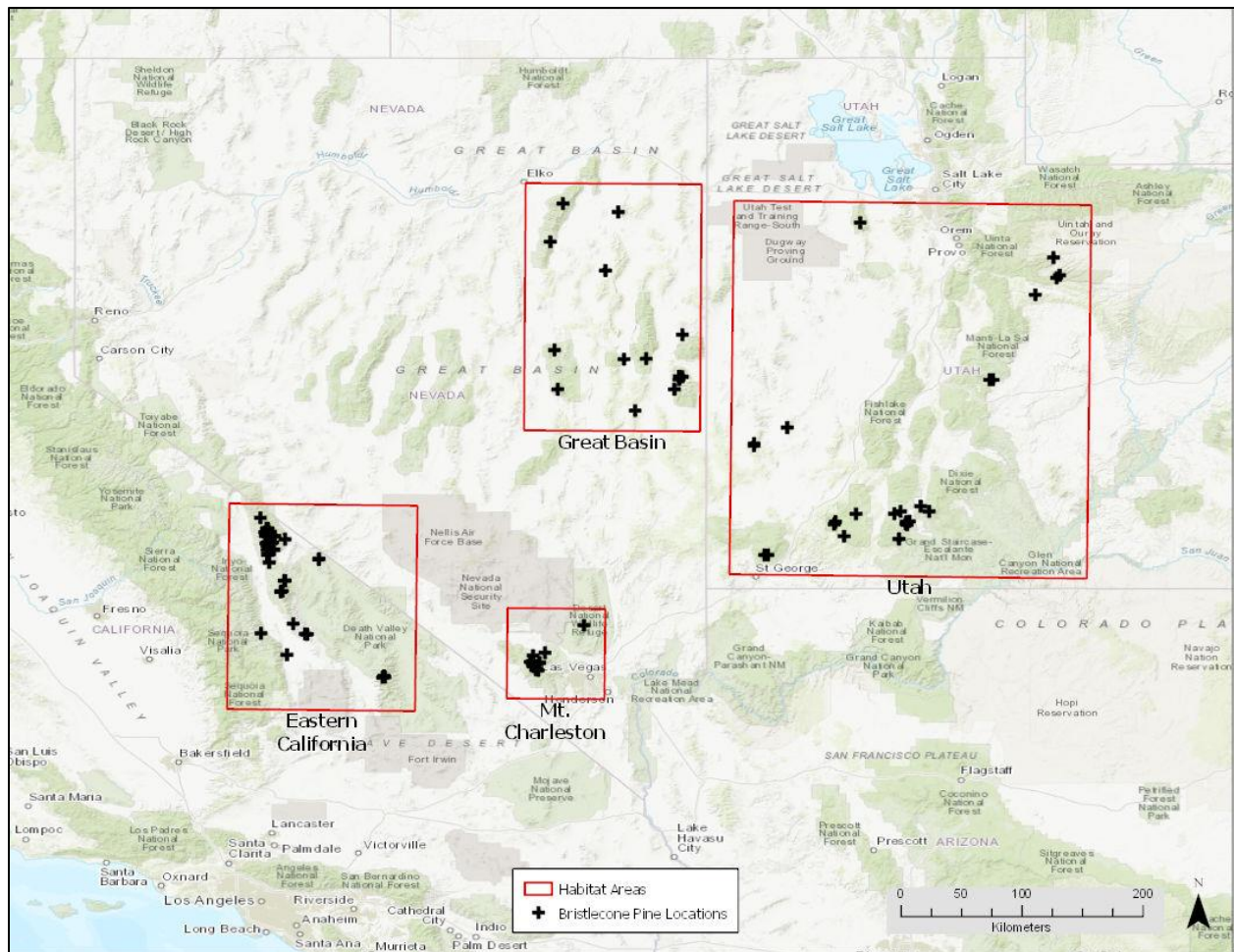
Figure 6 - Bristlecone pine major habitat areas in the Western U.S.

3.2.1.2. Subset Presence Locations for Study Area

With the dataset cleansed, the next step was to determine the extent of the study area. The goal of this study was to examine model performance at different resolutions and with different modeling approaches rather than making a definitive study of the bristlecone habitat. Modeling the full range or potential species distribution was not necessary. Given the resolution of the covariate datasets, modeling the full range of the species would have required substantial computer processing. The study area selected needed to be valid and sufficient for the species modeling but did not need to be all inclusive. To that end, a multistep process of selection and reduction was followed to determine the study area.

While Maxent has been shown to perform well with limited presence data (Baldwin 2009; Elith et al. 2006; Hernandez et al. 2006; Wisz et al. 2008), sample sizes greater than 50 are generally recommended for SDMs (Franklin and Miller 2014, 62-63; Guisan, Thuiller, and Zimmermann 2017, 116; Hernandez et al. 2006). From the pool of 180 locations remaining after cleansing the original GBIF data, the eastern California region was the first pass selection as the species presence location count exceeds this sample size recommendation. The region was then broken down into four subregions (Table 5).

Table 5 - Eastern California subregions' species presence location counts

Subregion	Count
White Mountain/Inyo Mountain Range	76
Last Chance Range	3
Sierra Nevada	2
Panamint Range	28

The locations in the region (Figure 7) lie primarily within the White Mountain/Inyo Mountain Range subregion with a second population grouping in the Panamint Range near the Death Valley area. Two presence locations in the Sierra Nevada and three locations in the Last Chance Range were excluded given their small record count and distance from other populations. While likely valid presence locations, including them would have significantly expanded the evaluation area and thus the processing time for the model while adding little analysis value. Of the two remaining subregions, the White Mountain/Inyo Mountain range provided a greater presence location count and was more dispersed than the Panamint population.
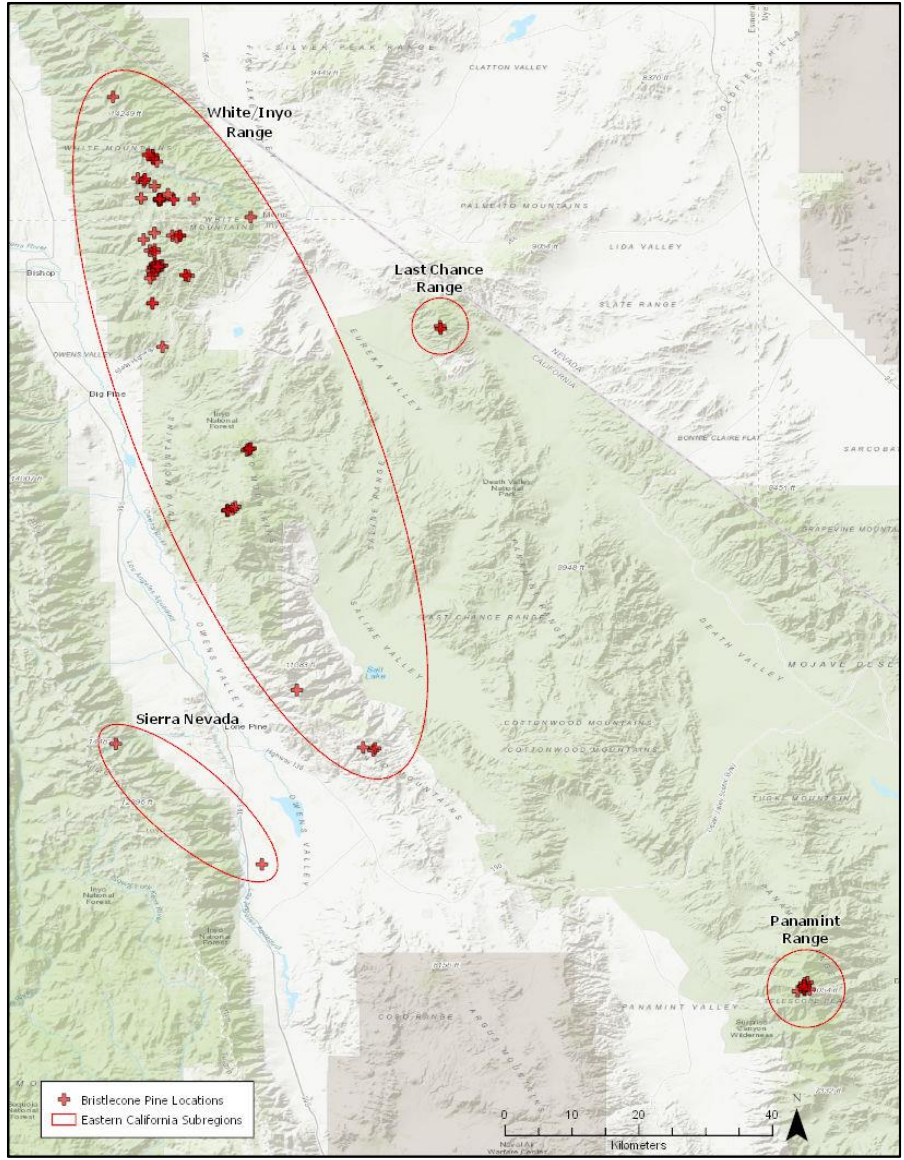
Figure 7 - Eastern California bristlecone pine subregions

The study area was set to a rectangle bounding the White Mountain/Inyo Mountain crest (Figure 8). A convex hull bounding the bristlecone pine locations was considered but deemed too limited. The choice of the rectangular area allowed for a capture of the full mountain range and nearby terrain. The area selected represented the species habitat well, had sufficient species presence locations for accurate modeling, had the desired variations in terrain, and was concise enough for reasonable processing times.
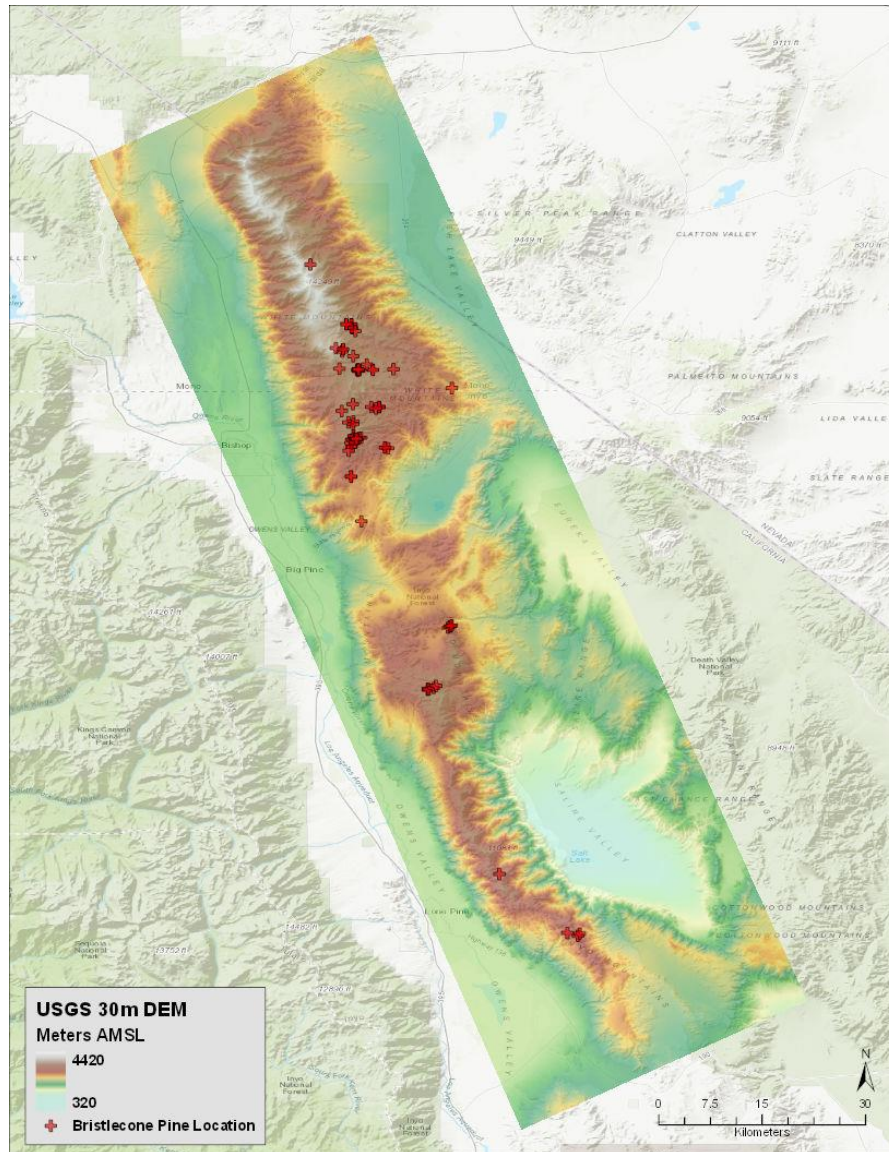
Figure 8 - Study area: White Mountain/Inyo Mountain Range

Using similar logic, the major populations groups (Figure 6) were reviewed to select a

prediction area for model evaluation. The Great Basin group was first considered as it was the

next largest number of species presence locations (23). The range of this group, however, was

dispersed across several ridges in a basin and range terrain. Coherent modeling would have been

difficult. The southern portion of the Utah population of bristlecone pines included 19 locations

within a concise extent (Figure 9). This area was used in the evaluation of the Maxent models created from the California location data.
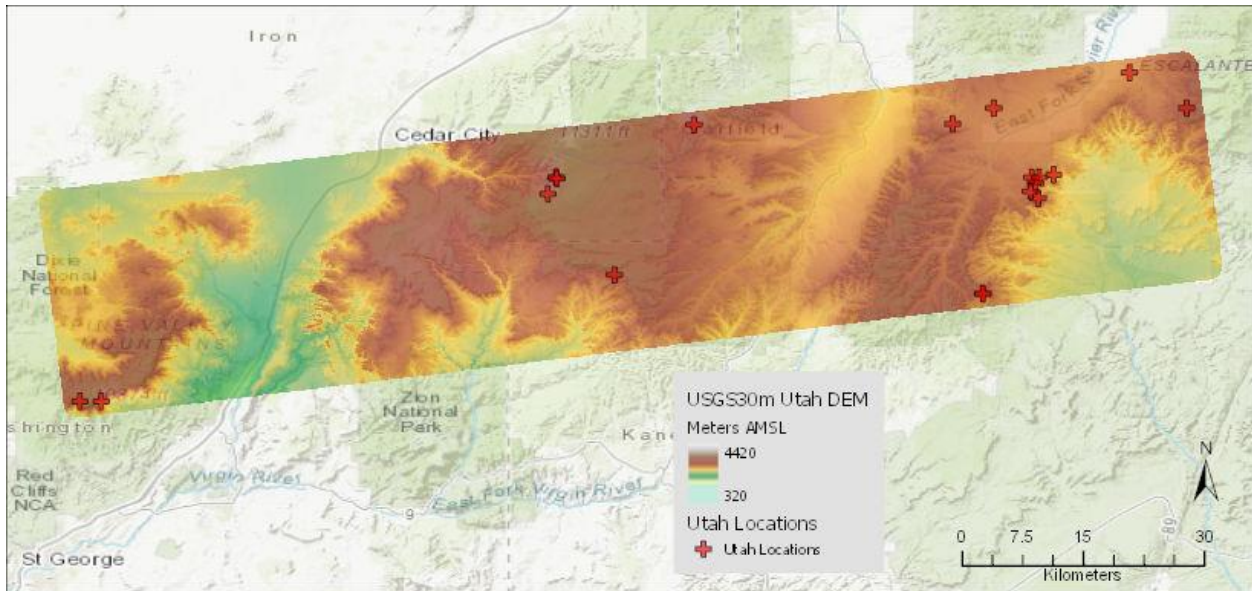
Figure 9 - Utah prediction area

3.2.2. Climate Data

The PRISM 30-Year Normals were selected as the source for the climate covariates. PRISM provides datasets at 800-meter and at 4km resolution (PRISM Climate Group 2018). The 1981-2010 Normals at 800-meter resolution were used for the study. PRISM is a quality, high resolution dataset comparing favorably with other datasets in climate models (Daly 2006; Daly et al. 2000). While not without limitations, PRISM performs reasonably well in mountainous terrain (Radcliffe and Mukundan 2016; Strachan and Daly 2017), a needed element for this study area. PRISM source weather stations are well represented in the region, including several at elevations relevant for the species (Figure 10). This increases the likelihood the PRISM climate interpolations reflect study area conditions.
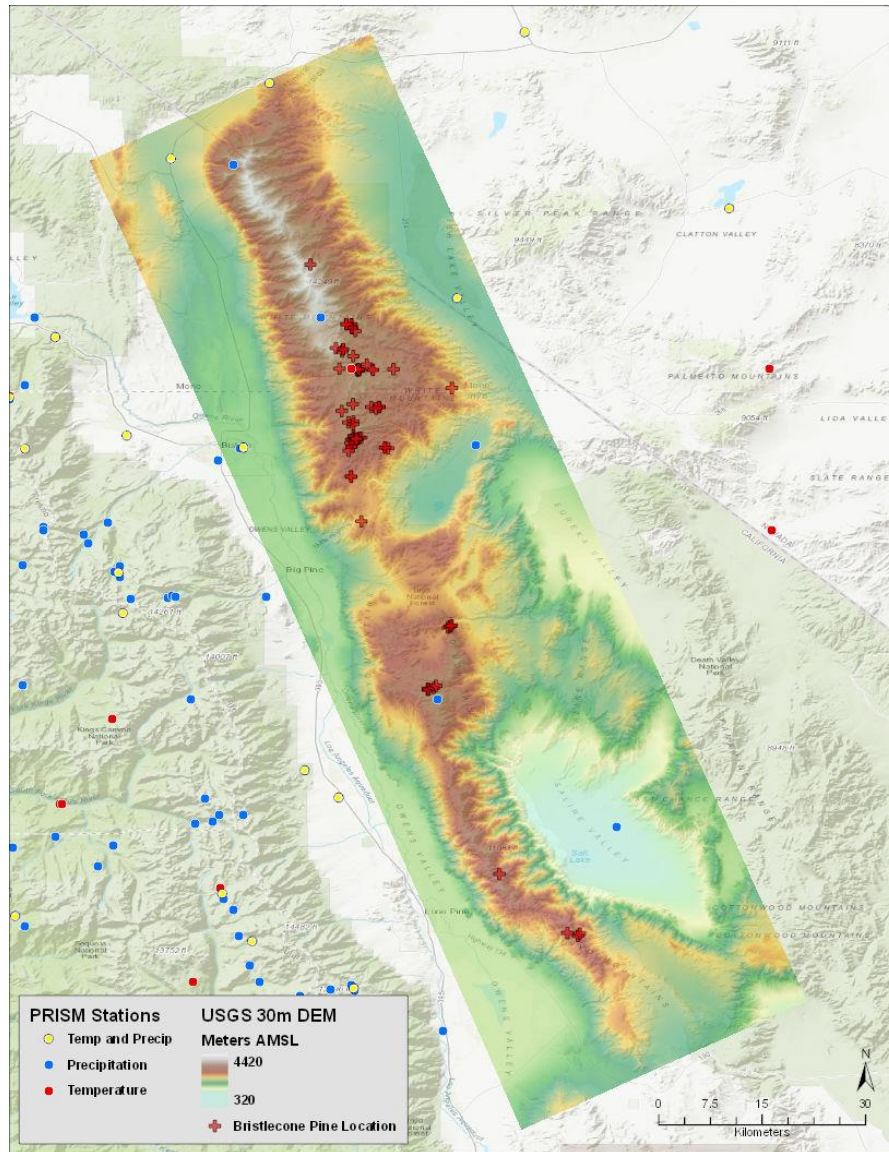
Figure 10 - PRISM station locations

The climate variables provided in the PRISM 30-year Normal data include mean temperature, minimum temperature, and maximum temperature in both annual and monthly aggregations. Growing days, a desired climate attribute for the bristlecone pines, was lacking. While the monthly 30-year Normal data could have been used to assemble seasonal temperature covariates, calculating the growing days from the provided data was an issue. The detail in the data provided by PRISM was not sufficient to calculate this covariate. Either an alternative was

needed, or the growing season covariate needed to be removed from the model. Removal would have allowed for an accurate assessment of the resolution question of the study, further simplifying the study, but would have removed one of the key climate variables associated with the bristlecone pines. Fortunately, an alternative climate data source was found, ClimateWNA (cWNA; Hamann et al. 2018), that offered a proxy for the growing days covariate, provided the preferred seasonal aggregations of the temperature covariates, and also addressed concerns around the data disaggregation.

Disaggregating the climate covariates to the 30-meter resolution without compromising the data integrity was a concern in the study design. Techniques such as resampling the PRISM 800-meter data were explored. Using cWNA, however, provided a peer reviewed approach, developed by climate experts, that reduced the risk of inadvertent errors being introduced using other approaches.

ClimateWNA is a software package developed by climatologists to generate custom climate surfaces at any resolution (Wang et al. 2011). The base data for cWNA is the PRISM 800-meter dataset. One of several similar programs developed by the authors (http://cfcg.forestry.ubc.ca/projects/climate-data/climatebcwna/), cWNA covers western North America from 100°W longitude to the Pacific and from 20° to 80°N latitude. While the source PRISM data use a combination of climate model interpolations from weather station data and expert-based algorithms to account for inversion areas, rain shadows, and coastal effect, cWNA extends the PRISM data to finer resolutions using bilinear interpolation and lapse rate adjustments. While the authors acknowledge the approach used is a relatively simple statistical method, testing indicates the model performs well in the mountainous regions it was designed for (Spittlehouse et al. 2012; Wang et al. 2011).

A further advantage of using cWNA is the seasonal aggregation of the climate covariates it provides. As with the spatial aggregation, having climate experts provide the temporal aggregations reduced the risk of data processing error. For this study, mean annual temperature (MAT), mean warmest month temp (MWMT), mean coldest month temp (MCMT), and degree-days $< 0°C$ (DD0) were selected from the available climate covariates. The three temperature data points provide a good characterization of the temperature ranges in the study area.

To reflect the climate lag mentioned in Section 3.2.1, all the temperature variables were adjusted downward by $1.5°C$. This is aligned with the recommendations in the bristlecone studies by Bruening et al. (2017). The caveats mentioned in above still apply. Properly identifying the appropriate temperature factors given the long-lived nature of the bristlecones in this study is problematic and model results from a species study perspective should be viewed accordingly. However, the focus of this study is on the impacts of resolution and tuning on the modeling. Holding these species variables constant with adjustments still provides a basis for evaluation within that scope.

The DD0 was selected as an inverse proxy for growing season. Growing season is typically defined by a measure called degree-days (University of California Agriculture and Natural Resources 2016). One degree-day is one day with the temperature above a determined threshold. It is a cumulative measure that, in this context, provides the potential growing season for a species. Tree line species such as bristlecone pines typically cease growing activity around $0.9°C$ (Bruening et al. 2017; Bruening 2016; Körner 2012; Salzer et al. 2014). The choice of DD0 (days below $0°C$) as a climate covariate in this study approximates degree-days outside the growing season for the species, thus providing an inverse proxy for the Maxent modeling. The higher the DD0 number, the shorter the growing season for the species.

In order to generate the data needed from cWNA, preparation was required in ArcGIS Pro. Figure 11 shows an overview of the process. (Details can be found in Appendix B Data Preparation Procedures.) Point grids were created from the rasters, with points centered on each cell in the raster. The point grid coordinate and elevation information were exported to a CSV file for input into cWNA. cWNA then interpolated the baseline PRISM climate data for each point in the grid. A CSV file was returned by cWNA with the climate data appended to each point. The process of CSV file to point to raster was then reversed, with a raster created for each of the covariates (MAT, MWMT, MCMT, DD0, elevation). The ArcGIS rasters were exported as ASCII files for Maxent input.
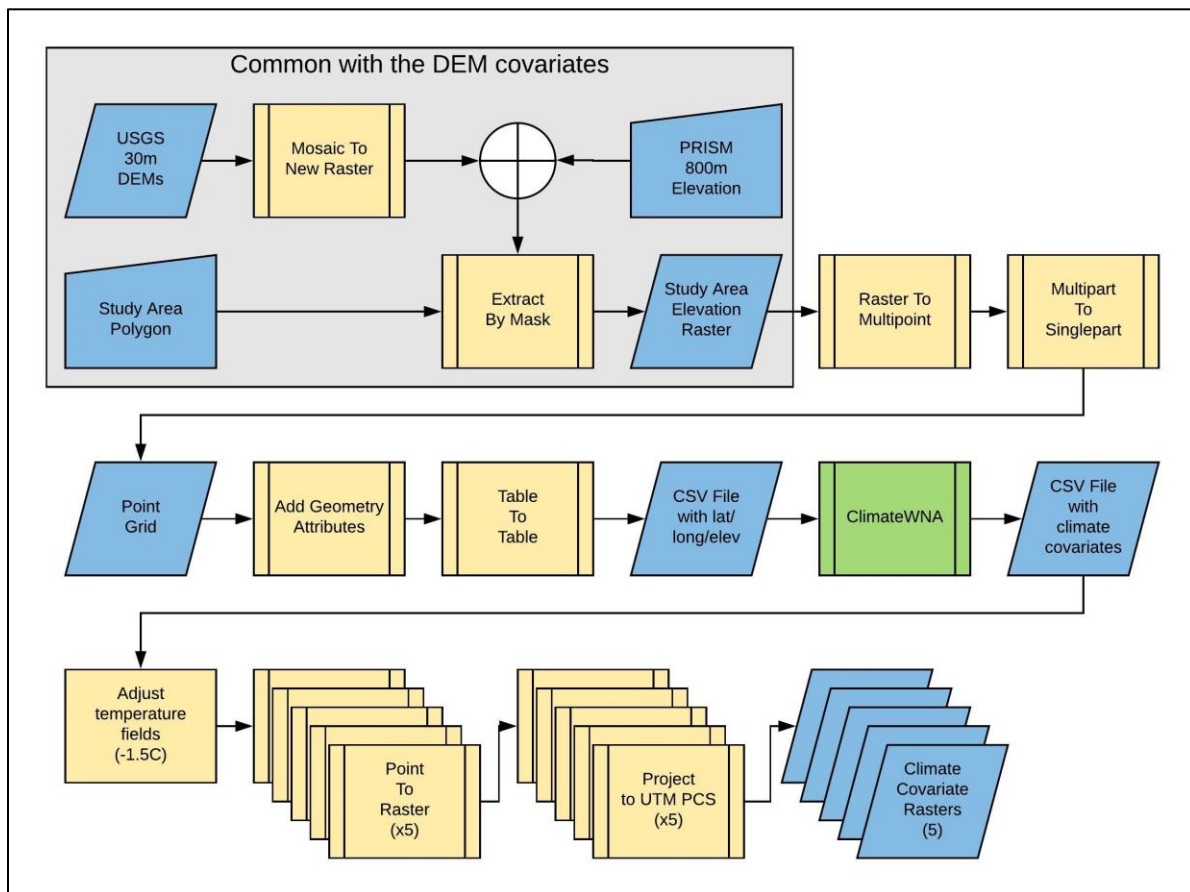
Figure 11 - Climate covariate preparation

An issue with the climate data interpolation to 30-meter was discovered during the processing. Grid blocks with large value differences across the margins appeared in portions of the rasters (Figure 12). The image on the left displays a close up of one such set of grid blocks, showing a distinct edge down the left side and fainter edges elsewhere. This closeup was located near the northern end of the study area (upper arrow in the image on the right). As can be seen in the broader view on the right, the grids are only apparent in certain locations, ranging from obvious (northern arrow), to subtle (southern arrows), to not apparent across most of the extent.
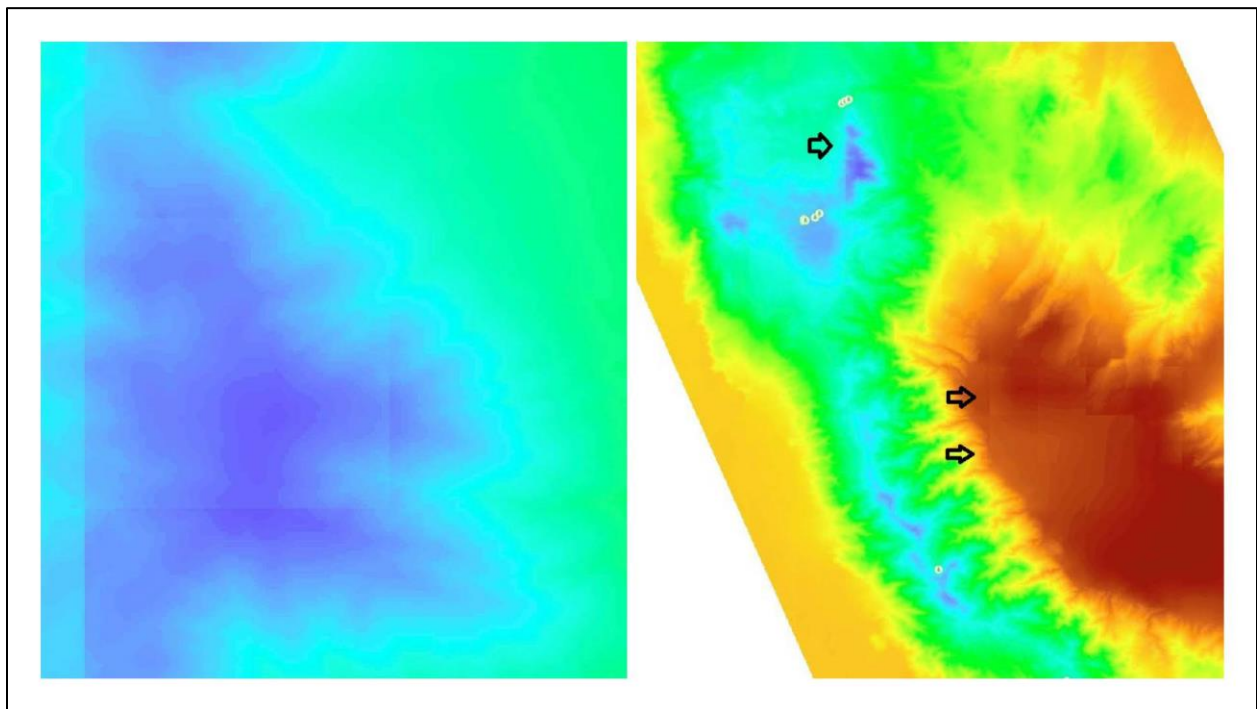
Figure 12 - cWNA gridding issue (MCMT example)

The climate rasters were examined manually and found to have the anomaly apparently randomly scattered across the extent. The edge delta varied from 0C to 0.5C for the temperature rasters in the samples that were examined (~1-2% of the temperature range). Similar results were found in the DD0 raster. No specific pattern to these anomalies could be discerned. The original

DEM and the elevation data output from cWNA were also examined. The anomaly was limited to the four climate rasters and did not appear in the elevation rasters.

The authors of cWNA (Wang et al. 2011) were queried about the appearance of the grids in the 30-meter interpolation. Dr. Hamann indicated cWNA currently can only interpolate PRISM data to 200m resolution (Andreas Hamann, email exchanges, December 2018). The grid blocks are artifacts of the interpolation. Program updates to allow finer resolution interpolations are planned but were not available during the time of this study. Options suggested included changing the study resolution to 200m or using additional interpolation methods to smooth the anomalies down to the 30-meter.

The decision was made to continue with the 30-meter resolution with the existing anomalies. The anomalies, when used in concert with the other elevation derived covariates, had minor impact on the study given the context of the research questions. Adding an additional interpolation to the climate data would have added unquantifiable uncertainty to the models. The 200m resolution is also not one seen in any of the literature reviewed for this study so its relevance would be questionable. Using the data at the 30-meter resolution, especially if the anomalies do indeed introduce error, served as input into one of the goals of the study: does higher resolution provide a better model? When presented the context of the study, Dr. Hamann concurred.

3.2.3. Digital Elevation Models

Two different DEMs were used in this study. The 30-meter resolution models used USGS 3DEP 1 arc-second DEMs (USGS 2015). The 800-meter models used the PRISM 800-meter elevation data (PRISM Climate Group 2018). The PRISM 800-meter elevation model is based on the USGS NED 3 arc-second DEM (Daly et al. 2008). Besides the elevation covariate, curvature

and aspect (expressed as northness and eastness) covariates were generated from the DEMs using ArcGIS Pro. Data were acquired to cover both the California study area and the prediction area in Utah.

The USGS 3DEP 1 arc-second product (here after simply "30-meter DEM") is provided in a series of tiles, each representing 1° x 1° extent. Five tiles were required to cover the California study area (Figure 13) and an additional three tiles for the Utah prediction area (Figure 14). The California tiles spanned 36°N to 39°N latitude and 117°W to 119°W longitude. The Utah tiles spanned 37°N to 38°N latitude and 111°W to 114°W longitude.
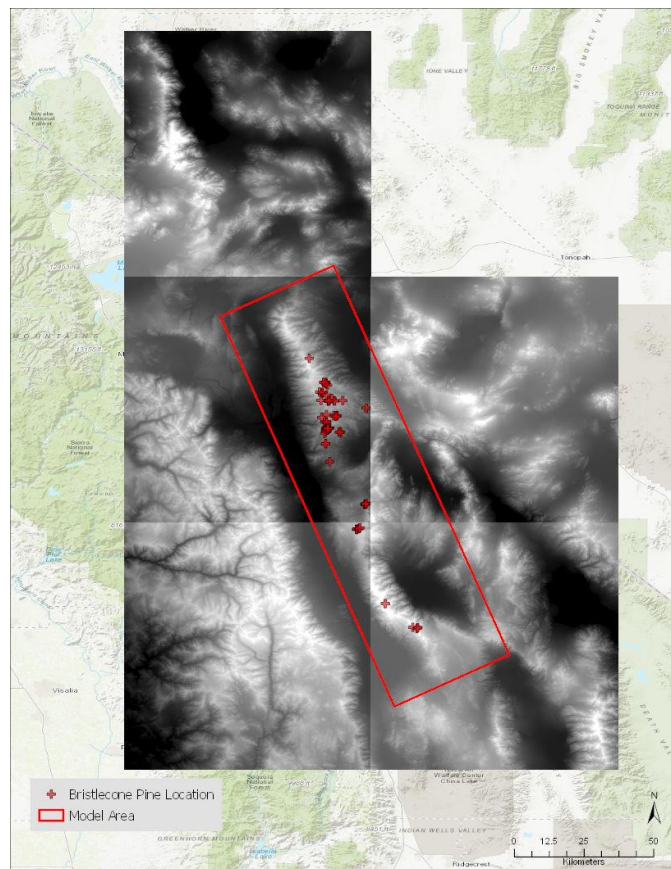
Figure 13 - USGS 1x1 degree tiles for the California study area
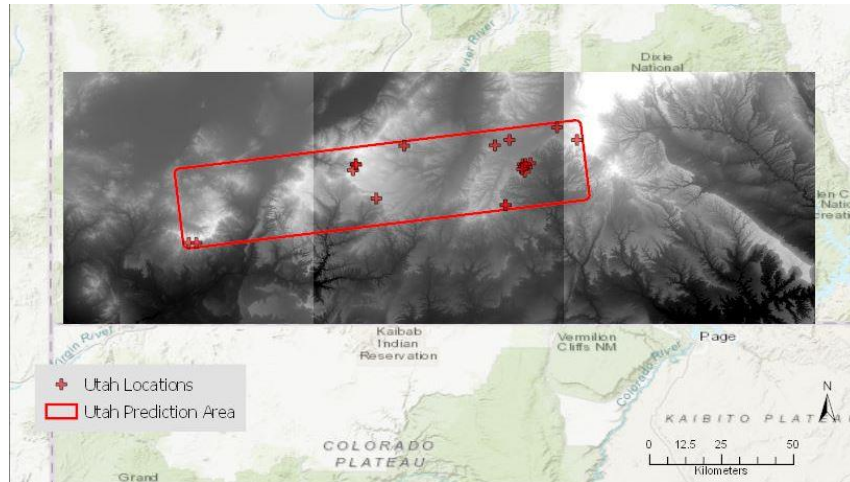
Figure 14 - USGS 1x1 degree tiles for the Utah prediction area

While the goal of the USGS 3DEP program is have all of the elevation data for the National Map based on lidar sources, many of the DEMs are still derived from other sources (Gesch 2007; USGS 2018). The 30-meter DEM metadata indicates the study area is derived from LT4X production methods (Figure 15). The LT4X software used by the USGS produces DEMs by interpolating elevations from contour lines of photogrammetrically derived digital line graph (DLG) sources. These DEMs have been shown to be of high quality (Hatfield 2000; Osborn et al. 2001; Wilson 2018, 35), with a vertical accuracy of 1-2m (Gesch, Oimoen, and Evans 2014; Wilson 2018, 35). Fitness for use in this study was good and no issues were encountered with these 30-meter DEMs.
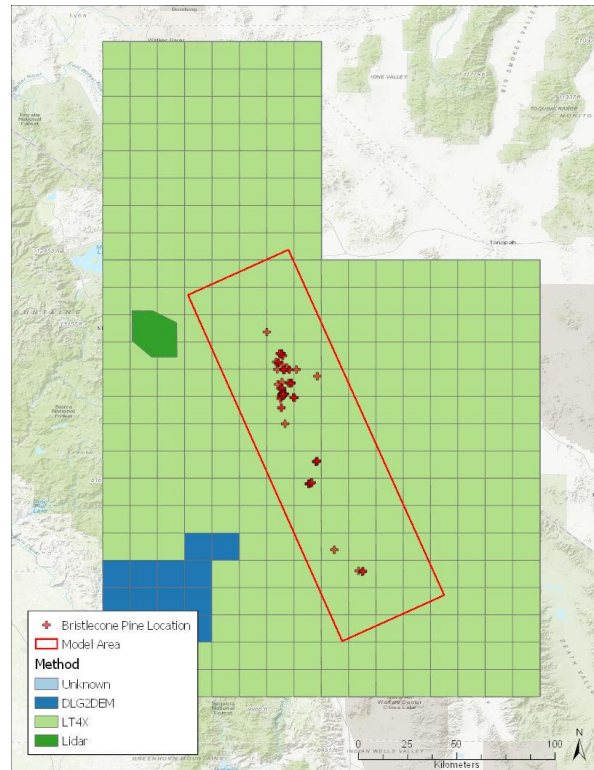
Figure 15 - USGS metadata for study area

The PRISM elevation model has a more complex lineage. Clarification is first needed on

nomenclature. PRISM uses the term "800m" for its fine resolution, 30 arc-second DEM. For

USGS data and other climate models, 30 arc-second is typically referred to as "1km" resolution.

The PRISM Group confirmed that the naming is a result of approximations of the cell size as the

product transitioned from 4km resolution to the new "800m" resolution in the 1990's. The new

resolution was approximately a 5x increase in cell count and 800-meters was roughly 1/5 of 4km.

The naming stuck and propagated through the subsequent literature. (Matt Doggett, email

exchanges, December 2018.) Similarly, references in the literature to the "80m" source data for

the 800-meter DEM refers to the USGS NED 3 arc-second products (aka 90m). For consistency,

the PRISM data is referred to as 800-meter DEM throughout this study. The underlying USGS

NED 3 arc-second data, however, is referred to as the 90m DEM.

The PRISM elevation model is derived from the USGS 90m DEM. The 90m DEM was aggregated to the 800-meter resolution using a modified Gaussian filter technique (Daly et al. 2008). The technique uses a circular pattern based on distance from cell centers to aggregate values to the new cell resolutions. There is mention of filtering of the DEM data for the climate model in the literature, removing terrain features with <~7km extent from the precipitation modeling. It was confirmed that this does not apply to the elevation data provided on the PRISM site (Chris Daly, email exchange, November 2018).

The amount of uncertainty introduced by the aggregation method for the elevation values is unstated in the literature. Likely it varies across the extent, depending on the complexity of the source terrain for the specific cell-to-cell interpolation. Further, the USGS Fact Sheet (USGS 2000) indicates the vertical accuracy of the 90m DEM is ±30 meters. Taken together, the covariates based on the 800m DEMs contain an undetermined amount of uncertainty, likely significant. As the PRISM climate covariates are modeled, in part, from the elevation information, this uncertainty extends to all covariates in the 800-meter models.

The source DEMs were processed in ArcGIS Pro as outlined in Figure 16. The 30-meter USGS DEMs were combined into one raster mosaic. Both the 30-meter and the 800-meter DEMs were clipped to match the study and prediction areas. Curvature and aspect were generated using the ArcGIS Pro tools. Aspect is typically expressed in degrees ranging from 0° to 360°, with north being both 0° and 360°. Statistical studies remove this circular issue and provide a linear value by decomposing aspect to "northness" and "eastness" using the cosine of the aspect and the sine of the aspect respectively. This provides a measure more suitable for analysis (Olaya 2009, 147; Wilson 2018, 63). The resulting northness and eastness rasters, as well as the curvature raster, were then exported as ASCII files for input into Maxent.
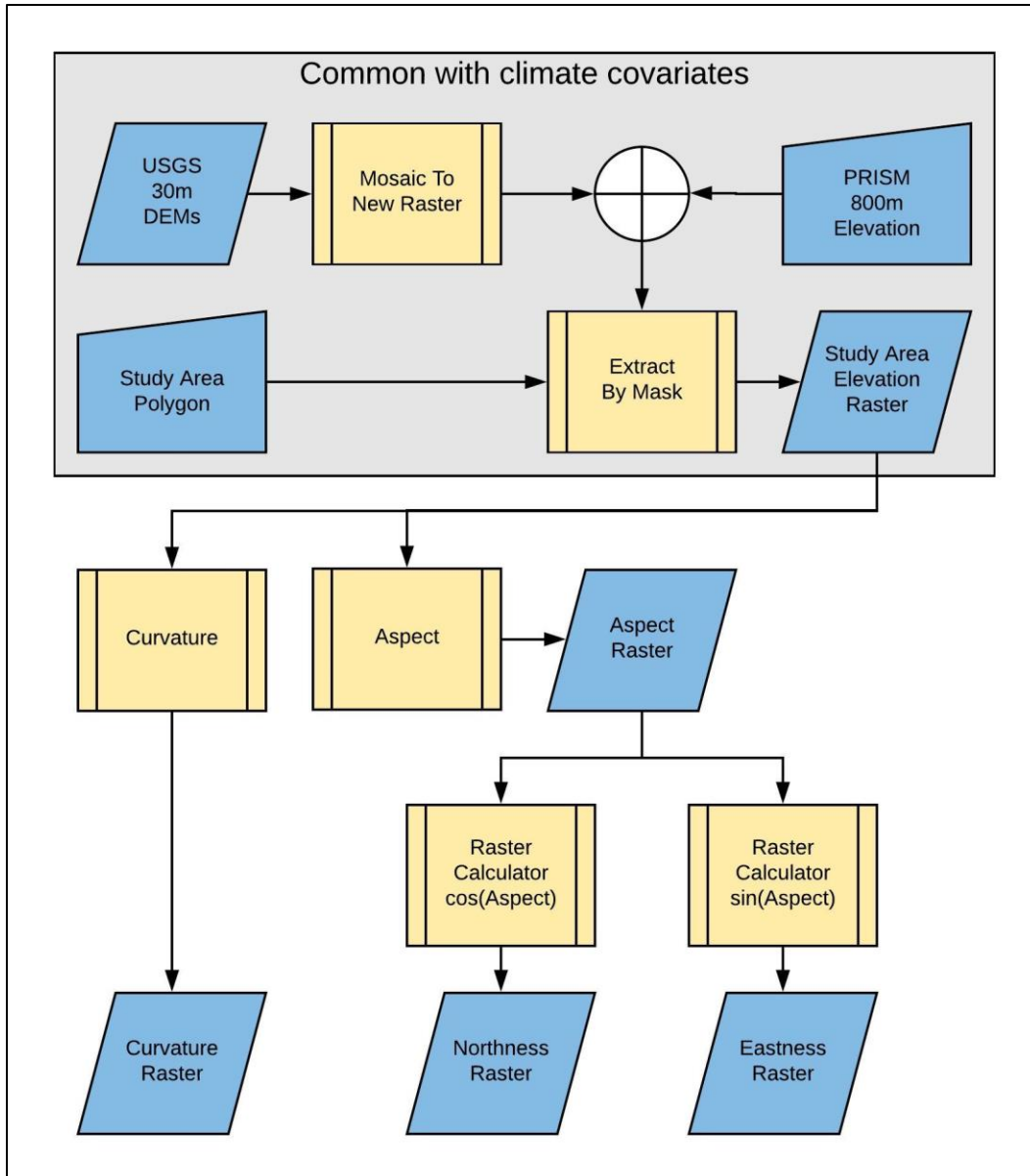
Figure 16 - DEM covariate preparation

3.3. Model Tuning

The study generated models using the default Maxent settings as well as models using tuned Maxent parameters. The ENMeval R package (Muscarella et al. 2014) was used to determine the tuned parameters for the bristlecone dataset.

3.3.1. Model Tuning Setup

Three choices were required for the setup of the ENMeval tuning exercise. The partitioning schemes to be used were determined. A range of RM values to test were selected. Finally, a set of FCs to be tested were chosen.

Four partitioning schemes were tested with the species data: jackknife, random k-fold, block, and checkerboard2. Checkerboard1 and user-defined were not used. Checkerboard1 is somewhat redundant to checkerboard2, the latter being adequate for the scope of this study. User-defined partitioning is only useful where specialized knowledge of the species distribution is available, and a specific sampling scheme is indicated.

For the random k-fold, k was set to 5 folds. Folds of 5, 10, or 20 are recommended as statistically stable, with the choice of fold count determined by the population size of the study (Kohavi 1995; Salzberg 1997). With the population in this study being 76 locations, a k-fold of 5 was deemed appropriate. This resulted in a bin size of ~15 locations per fold for testing and training partitions.

The checkerboard pattern was set to 5km strips across the study area (Figure 17). The choice of strip size was somewhat arbitrary. Given the extent is approximately 50km by 185km, the 5km grid size seemed reasonable given the distribution of the species locations. Groupings within the checkerboard pattern were neither too sparse nor too concentrated. See Appendix C Partitioning Scheme Patterns for the block, random k-fold, and checkerboard2 partitions specific to this study.
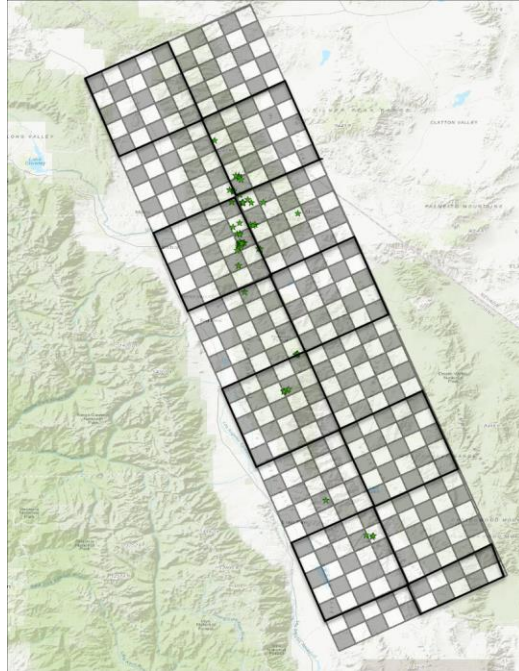
Figure 17 - Checkerboard partitions at 5km grids (approximate)

For the RM list, values of 1, 2, 3, 4, and 5 were initially selected, with the option to expand the list and rerun the test if tuning results suggested higher values would be useful. They were not necessary given the results of the initial test runs with the 1 to 5 values.

For FCs, values of L, Q, P, LQ, and LQP (where L=linear, Q=quadratic, and P=product) were chosen for testing. Simplifying the models by excluding the hinge (H) and threshold (T) FCs can lead to improved model performance (Merow, Smith, and Silander 2013; Syfert, Smith, and Coomes 2013), so these values were not considered for the tuned models in this study.

With the partitioning schemes and parameter sets selected, testing was performed on the full set of combinations. Two resolutions, four partitioning schemes, five RM values, and five FC values resulted in an array of 200 model test configurations. Using R programming, primarily the ENMeval package, the model arrays were processed in Maxent, and the evaluation metrics gathered. Evaluation metrics were consolidated for comparisons in Microsoft Excel. Details of the programming are provided in Appendix D Model Tuning R Script.

3.3.2. Tuning Results Evaluation

To evaluate the results from the tuning tests, the output metrics across each partitioning scheme were compared. The top two models from each partitioning scheme were compiled and compared across schemes. From these eight test models, the best model per the evaluation was used for the subsequent model processing. This process was repeated for both the 30-meter and 800-meter resolution models. Details of the tuning results for the 200 test models can be found in Appendix E Tuning Test Results. A summary of the results and the tuning conclusions for the 30-meter and 800-meter resolutions are discussed in Chapter 4.

As discussed in Chapter 2, there is no consensus on what the most important metric is for comparing Maxent models. An evaluation of an array of metrics, with emphasis given to those that best fit the study objectives is recommended. For this study, priority was given to the AICc value of each model, then the AUC value. The AICc value allowed comparisons of complexity and goodness of fit between the models of the same resolution. The AUC score measures the model's ability to discriminate between background points and presence locations. While not without controversy as noted in Chapter 2, AUC score is useful in this context as the study area and associated data are fixed across the models being compared. Lastly, the orMTP, or10pct, and diff.AUC scores were also used in the evaluation, being indicators of potential overfitting.

Use of the kappa and TSS metrics was considered. The prevalence issues with kappa make it questionable with the sample size and extents involved in this study. The lack of true absence data in the study's models limited the usefulness of both TSS and kappa in this context. Given the scope of the study, the AICc, AUC, and fitting measures were deemed sufficient.

The naming of each test model follows the pattern of "FC_RM". For example, LQP_2 would indicate the model using the combined linear, quadratic, and product FC set in combination with a RM value of 2.

3.4. Model Processing

Maxent models were built for the comparisons of default and tuned settings at both 30-meter and 800-meter resolutions; a total of four models. The models were then used to create the prediction surfaces for the California study area and for the Utah prediction area. Species and covariates from the California study area provide the inputs for model creation. The prediction surfaces use local covariate data from either the California or Utah extents.

The R programming environment was used to facilitate building of the models, creation of the prediction surfaces, and for gathering the evaluation metrics. Twenty-five iterations of each model were executed to insure stability in the resulting metric output (Morales, Fernández, and Baca-González 2017). The primary R packages used were dismo (Hijmans et al. 2017) and raster (Hijmans 2018). Details of the code are provided in Appendix F Model Comparison R Script. The dismo package runs the Maxent Java application in Maxent's batch mode for model and prediction surface creation. There is no difference in Maxent functionality or outputs using this approach versus running Maxent with its desktop user interface.

Table 6 lists the key Maxent parameters for the default and tuned models that impact model performance for this study. The FC and RM values for the tuned model presented here were derived from the tuning tests in ENMeval. The results of the ENMeval tests are discussed in Chapter 4. Maxent chose the LQH FC for the default model. This was validated by examining the standard Maxent outputs. Prediction rasters were output in both the Maxent raw and logistics formats using the dismo::predict function. The raw output was required for the metric calculations (Merow et al. 2014; Muscarella et al. 2014). The logistics output was used for the spatial analysis as this provided a more intuitive value while retaining the same scaling as the raw output.

Table 6 - Maxent model parameters

Maxent Parameter	Choices	Maxent Default	Comments	Default Model	Tuned Model
Feature Class (FC)	Linear (L) Quadratic (Q) Product (P) Threshold (T) Hinge (H) Auto	Auto	Auto FC set per sample size: Linear is always used; quadratic with at least 10 samples; hinge with at least 15; threshold and product with at least 80. Ranges can be overridden.	Auto: LQH for this dataset	LQ per ENMeval analysis
Regularization multiplier (RM)	Number >=0.5, in steps of 0.5	1.0		1.0	2.0 per ENMeval analysis
Output format	Cloglog Logistic Cumulative Raw	Cloglog	Raw recommended for comparing models of the same species. (Merow et al 2013, others) Raw required for ENMeval AICc metric.	Raw & Logistic	Raw & Logistic
Max number of background points	Number	10,000	Random points selected using dismo::randomPoints function rather than the Maxent function	CA: 10,000 UT: 5,000	CA: 10,000 UT: 5,000
Maximum iterations	Number	500	Insure convergence of the covariate evaluation within Maxent	500	5000

The number of background points was reduced to 5,000 for the Utah models. The size of the 800-meter Utah raster was <10,000. For consistency, the 5,000 points choice was used for the 30-meter Utah raster as well. In all cases, new random background points were selected for each of the twenty-five model iterations using the dismo::randomPoints function. Using the dismo function, rather than the native Maxent random selection function, allowed the same set of random background points to be applied to both the default and tuned models within each iteration. Each model type, therefore, executed off the same dataset of background points, presence locations, and covariates within each iteration.

The maximum iterations parameter determines how many iterations Maxent processes internally as it is evaluating the covariates and determining the coefficients for the model being created. (Note that these iterations are internal to Maxent and not to be confused with the twenty-five iterations performed for the model evaluations.) Maxent continues to iterate until this

iteration count is reached or until convergence of the coefficient value to a minimum threshold (0.00001) occurs. Testing during the tuning experiments showed that convergence for this dataset often did not occur until 700+ iterations. The default model was left at the default value of 500 to reflect the intent of the study, but the tuned model was increased to an arbitrarily high value of 5000 to insure convergence of the covariate parameters in the model.

3.5. Model Evaluation Methods

Model evaluation comprised three approaches. First, model metrics measuring the quality and degree of fit were reviewed. Second, the models were used to predict species distribution in the Utah area. This allowed assessment on how well the models perform when applied to other species habitat areas. Lastly, spatial analysis of differences between the default and tuned models was performed on the Maxent prediction rasters. Maxent output rasters were compared across resolutions and between default and tuned models. A hot spot analysis was also performed on the difference between the default and tuned models.

3.5.1. Model Evaluation Metrics

For each of the twenty-five model iterations, metrics were calculated for the models. As with the tuning evaluations, TSS and kappa were not considered. Metrics for AICc, AUC, AUC difference (diff.AUC), and two omission rates (orMTP and or10pct) were collected and averaged across the iterations. AICc is an independent measure of model fit and complexity. Lower AICc scores indicate a parsimonious balance between those two factors. AUC, while not without controversy, is a measure of the model's ability to discriminate between background points and presence locations. Higher AUC scores indicate better models. The AUC difference, orMTP, and or10pct metrics are measures of model fit. Higher values on these scores indicate overfitting of the model. As with the tuning evaluations, priority was given to the AICc score of each model,

then the AUC score, with the diff.AUC, orMTP, and or10pct scores considered as indicators of overfitting.

The AICc metric was calculated by the ENMeval::calc.aic function. The AUC metric was calculated by the dismo::evaluate function. The orMTP and or10pct metrics were extracted from Maxent's maxentResults.csv file. Omission rate metrics were only applicable to the California study area and were not available for the Utah predictions. Maxent generates these metrics for the model area where training and test data are available, but not for areas predicted using the model. All metrics were output to CSV files and then imported into Excel for consolidation and presentation purposes. Details of the code involved with the metric generation and collection are included in Appendix F Model Comparison R Script.

3.5.2. Model Prediction to Utah Area

Another view of the model quality was obtained by using the models generated from the California training data to predict a species distribution in another region. Transferring a model (a "projection" in Maxent terms) to a new area or to a different climate scenario is a contentious topic for Maxent modeling (Phillips 2008; Townsend Peterson, Papeş, and Eaton 2007). Given the complexities of species interactions with their environments, not all the relevant factors may translate well from one region to another or from one climate scenario to another. Key factors for the new area might be missing. Approaches to mitigate this when doing species distribution modeling for ecological studies are available (Sequeira et al. 2018), but they are not included here due to the focus of this study. However, since transferring the models to another area was included as a means of comparing models, and taking a cue from Sequeira et al. (2018), a selection of covariates were compared between the California and Utah extents. The ranges of

the covariates in both regions were compared to example response curves from the models built on the California data.

In addition, performance metrics were generated for the Utah area. Prediction surfaces were created for the Utah southern region using the default and tuned models. AUC and AICc metrics were calculated on the models using the species presence data and covariates from that new region. Maxent does not generate omission rates for predictions to new regions, so these metrics were not available for the Utah area. Likewise, diff.AUC is not available as there is no test/train performed on the prediction.

3.5.3. Spatial Analysis of Model Tuning

Two approaches were taken for spatial analysis of the models. Maxent output rasters were compared across the resolutions as well as between the default and tuned models. A hot spot analysis (HSA) was also performed on the differences between the default and tuned models to visualize the changes in the models from tuning.

Prediction rasters using the Maxent logistics output were used as a basis for the spatial analysis rather than the raw output format. The logistics output provides values of relative likelihood of species occurrence across the extent of the study area. The values are scaled directly from the raw data outputs used for the metric generation but provide values that are more intuitive for interpretation.

Prediction rasters were gathered from each of the twenty-five iterations of the models. The rasters were assembled into raster stacks and calculations of mean values performed using the raster::mean function in R. The resulting rasters provided the average values across the twenty-five iterations for each cell in the extent. Four prediction rasters were created; one raster for each resolution/model type combination.

An additional set of difference (delta) rasters was created for the HSA. Differences between the prediction rasters of the default model and tuned model were calculated by subtracting each mean tuned prediction cell value from the corresponding mean default prediction cell value. A set of four rasters was created from this: the "raw" delta values (positive and negative values); the absolute value of the delta; just the negative values; and just the positive values. These are summarized in Table 7. These rasters provided inputs for the subsequent hot spot analysis of the tuning changes.

Maxent's logistic output was used for these rasters. The logistic output maintains the same relationships between cell values as presented in the Maxent raw output that was required for the metric calculations. The relative probability values in the logistic output are more intuitive than the Maxent raw output values.

The mean rasters were created in the R environment as part of the model processing, then imported into ArcGIS Pro for subsequent processing. The ArcGIS "Raster Calculator" tool was used to create the delta rasters of the differences between the default and tuned model outputs. These were created at both the 30-meter and 800-meter resolution for the California study area.

Table 7 - Rasters used for the spatial analysis

Raster	Description
LogMean_Default	Mean values from the default model prediction rasters in logistic format
LogMean_Tuned	Mean values from the tuned model prediction rasters in logistic format
LogDelta	LogMean_Default minus LogMean_Tuned. Positive values indicate the default model had higher predictions at that cell location. Negative values indicate the tuned model had higher values at that cell location.
LogDelta_abs	The absolute value of LogDelta. Shows the variance between the tuned and default models.
LogDelta_pos	LogDelta filtered for just the positive values. Negative values were set to null for those cell locations.
LogDelta_neg	LogDelta filtered for just the negative values. Positive values were set to null for those cell locations.

For use in the hot spot analysis, point arrays were created at 30-meter and 800-meter resolutions to match the output rasters. Values from the rasters listed in Table 7 were spatially joined to their corresponding resolution point arrays as attributes of the points using the ArcGIS "Multi Values To Points" tool. These point arrays were used as inputs into the ArcGIS Pro "Optimized Hot Spot Analysis" tool. The hot spot analysis provided indications of spatial patterns in the change between the default and tuned model outputs. "Cold spots" would indicate areas of small change between the models. "Hot spots" would indicate greater deltas in prediction values. The data and the hot spot analysis were parsed into absolute, negative, and positive values to provide insights into different aspects of the model change patterns.

Chapter 4 Results

This chapter reviews the results of the tuning and modeling performed for the study. First, the data from the tuning tests are reviewed and model parameter conclusions presented. Next, results of the model evaluation methods discussed in Chapter 3 are reviewed in turn. Performance metrics were derived from the models and the prediction surfaces of the California study area. The models were used to create predictions in the species' southern Utah range to determine how well the models perform when applied to new areas. Finally, the prediction surfaces of the default and tuned models were compared spatially to assess how the tuning changed the prediction surfaces. Each of these evaluations are discussed in the following sections.

4.1. Model Tuning Results

As discussed in Chapter 3, ENMeval provided data to determine the parameters for the tuned Maxent models. This section discusses the results of the tuning tests for each model resolution/partitioning scheme combination. The RM and FC parameter selections for the 30-meter and 800-meter tuned models are listed at the end of their respective results sections. Detailed metrics for each of the 200 test models can be found in Appendix E Tuning Test Results. The appendix also includes graphs of the five metrics (AICc, AUC, AUC.diff, orMTP, or10pct) for each combination of resolution/partitioning scheme as these were too large to include inline here (20 graphs over four pages).

4.1.1. Tuning Results: 30-meter Resolution

In the block partitioning tests, the overall orMTP scores look better for this partition scheme compared to the others. The or10pct scores, however, were much higher than the other schemes. This indicates an overfitting of the model. The range on the AUC scores was small,

indicating little variation in discrimination across the test models. Interestingly, the L_1 model had good scores across the metrics except for the AICc score. This reinforces the literature contention that an independent AICc score needs to be considered outside the model generated AUC scores. LQ_1 had the best AICc and the second-best AUC scores, making it the top candidate in this scheme. LQ_2, while having a slightly lower AUC score than L_1, scored better on the AICc score. The other metrics were not significantly different.

Overall results were scattered on the checkerboard2 partition scheme, with the AICc and AUC scores not aligning on the models and the other metrics mixed for top ranked AICc candidates. The checkerboard2 scheme, like the block model, scored high on the or10pct metric across the model. The orMTP scores were high as well. These indicators of overfitting question the appropriateness of this scheme for the species model. As the overall AUC range was very tight, as was the diff.AUC measures, LQ_1 and LQ_2 were chosen as the top two candidates given their AICc scores. Their or10pct scores were of concern, however, as they were higher than many of the other models in the scheme.

The jackknife partition results were less ambiguous. Metrics overall were fairly consistent, ranges small, and the top contenders easy to assess. The LQ_2 candidate had the top AICc and AUC scores, making it the top pick. LQ_1 had the second highest AICc and a AUC score only 0.0002 off from the second highest score; virtually identical, making it the second candidate from this scheme. The orMTP scores were flat across all the models, holding values of either 0.0132 or 0.0263. As the jackknife uses all the locations for test in turn, averaging the results, the speculation is this flattens the metric results. Overall, the orMTP and or10pct scores tended to be high.

Random k-fold results were similarly straight forward in regards to AICc and AUC scores, but mixed with the other metrics. While the orMTP scores were high across the board, the range and the scores on the diff.AUC were very tight. The or10pct was very mixed, with some good scores and many that were high. Overall, a mixed message in the metrics as far as overfitting in the model. LQ_1 and LQ_2 were selected as the top two models out of this scheme given their good AICc and AUC scores. LQ_1 had a better or10pct number, with the other metrics being comparable.

Consolidating the top scores from each of the partitioning scheme (Table 8) shows that all the partitioning schemes resolved to the LQ_1 and LQ_2 models. The range on the AICc scores was only 3.2 points, making the models essentially ties in this metric. This is as would be expected given the FC and RM parameters are the same. AICc is calculated across the full extent of the data, regardless of the partition scheme and weighs parameterization in the scoring. The other metrics in the mix become more important in this stage of the evaluation.

Table 8 - Top two models from each partition scheme for 30-meter resolution

Partition	Settings	FC	RM	AUC	diff.AUC	orMTP	or10pct	AICc	par.
Block	LQ_1	LQ	1	0.9408	0.0167	0.000	0.145	2179.5	8
Block	LQ_2	LQ	2	0.9405	0.0193	0.000	0.145	2181.7	8
Checkerboard2	LQ_1	LQ	1	0.9342	0.0259	0.029	0.165	2179.0	8
Checkerboard2	LQ_2	LQ	2	0.9345	0.0259	0.048	0.145	2180.2	8
Jackknife	LQ_1	LQ	1	0.9456	0.0275	0.026	0.118	2181.5	9
Jackknife	LQ_2	LQ	2	0.9461	0.0270	0.026	0.105	2180.6	8
Random k-fold	LQ_1	LQ	1	0.9421	0.0091	0.053	0.106	2182.1	9
Random k-fold	LQ_2	LQ	2	0.9415	0.0091	0.040	0.118	2181.0	8
	RANGE			0.0119	0.0184	0.053	0.059	3.2	1
Ranking	*Best*	*Second*							

The high or10pct scores for the block and checkerboard2 schemes eliminated them immediately. While the other measures of overfitting, diff.AUC and orMTP, had good measures

for the block scheme, the relatively low AUC and high or10pct made it questionable. The checkerboard2 scheme had high overfitting metrics and even lower AUC scores.

The random k-fold and the jackknife yield similar metrics. Both had mixed, but good overfitting numbers, with random k-fold showing significantly better diff.AUC values. The slightly better AUC and orMTP scores tipped the decision in favor of the jackknife scheme, with LQ_2 being the configuration selected for the tuned 30-meter models. Either of the random k-fold models, however, would have been viable candidates to use.

4.1.2. Tuning Results: 800-meter Resolution

The results from the block partition scheme models were mixed. The or10pct and the orMTP scores were high, making this partitioning scheme an unlikely candidate for the final tuned model. Scores on the or10pct metric ranged as high as 0.1974. The orMTP scores were as high as 0.526. The LQ_1 and the LQ_3 models had the lowest AICc scores, but LQ_3 score poorly on the fitting metrics and the AUC. L_1 had better AUC and fitting numbers, with an AICc well towards the lower end of the range. It was chosen as the second candidate from this scheme.

The checkerboard2 results were very mixed. Good AICc metrics did not correspond with good AUC numbers. The fitting metrics were mixed. While not as muddled as the random k-fold, choosing the candidate models entailed compromises. The L_1 and L_2 models were chosen as they had the best overall metrics. The AUC scores were the highest. The AICc scores were near the low end of the range, not far off the top performers. Their fitting scores were better than the other models in the scheme.

Overall, the jackknife partition results were the most consistent of the schemes. The range on the AUC scores was tight, as was the AICc range. The orMTP and or10pct scores were

slightly high, with a couple high outliers in both. The diff.AUC scores also slightly high, though consistent across the models. LQ_1 was the top candidate given its top AICc score and second-best AUC score. LQ_2 and LQP_3 were the next candidates considered. LQ_2 prevailed as the AUC score was better and the LQP_3 or10pct and orMTP scores high.

The random k-fold results were very mixed. The omission rates were high across all the models, though the diff.AUC values were the best across all the 800-meter partition schemes. LQ_2 and LQ_1 were chosen as the candidate models given their low AICc scores and high AUC values. Both had the same high orMTP and or10pct scores, a cause for concern.

Consolidating the top partition scheme candidates (Table 9) yielded a slightly more diverse mix of FC/RM models compared to the 30-meter schemes. As with the 30-meter, the AICc scores had a tight range (5.2 points). Evaluation again came down to the other metrics. The block schemes were immediately removed given the very high or10pct numbers. The jackknife models were nearly identical. Both had strong AUC scores, though the orMTP and or10pct were slightly on the high side. The omission rates were the second best in the list, however. The random k-fold candidates had good AUC scores and excellent diff.AUC numbers but were very high on the orMTP and or10pct scores. The choice of candidate for the final 800-meter tuning model was LQ_2. The LQ_1 model would also have made good candidate. Using LQ_2 had an edge as it would be consistent with the 30-meter candidate, albeit via a different decision path.

Table 9 - Top two models from each partition scheme for 800-meter resolution

Partition	Settings	FC	RM	AUC	diff.AUC	orMTP	or10pct	AICc	par.
Block	L_1	L	1	0.9472	0.0136	0.013	0.184	1137.4	6
Block	LQ_1	LQ	1	0.9470	0.0174	0.000	0.184	1132.2	8
Checkerboard2	L_1	L	1	0.9411	0.0127	0.031	0.098	1137.3	6
Checkerboard2	L_2	L	2	0.9405	0.0125	0.031	0.098	1136.8	5
Jackknife	LQ_1	LQ	1	0.9474	0.0260	0.013	0.118	1132.2	8
Jackknife	LQ_2	LQ	2	0.9471	0.0261	0.013	0.118	1133.2	8
Random k-fold	LQ_1	LQ	1	0.9472	0.0093	0.040	0.120	1134.4	9
Random k-fold	LQ_2	LQ	2	0.9467	0.0093	0.040	0.120	1132.9	8
	RANGE			0.0069	0.0168	0.040	0.086	5.2	4
Ranking	*Best*	*Second*							

4.2. Model Performance Metrics

Four performance metrics were derived from the models: AICc, AUC, orMTP, and or10pct. Direct comparisons of AICc and AUC between the 30-meter and 800-meter models is not possible. While the models used the same species presence locations and overall extent, the difference in cell count between the two resolutions results in differences in the scale of the values for AICc and AUC metrics. The metrics are, however, useful for comparing the default with the tuned models. Summaries of the results follow. Details of each of the 25 iterations for each of the four models can be found in Appendix G Evaluation Metrics.

Table 10 summarizes the AICc metrics averaged from 25 runs of the models using the California study area data. Tuning of the models demonstrated a marked increase in the AICc performance at both resolutions. The mean value on the 800-meter model increased by 9%. For the 30-meter model, the performance increased by ~7%.

Table 10 - AICc summary

Resolution		AICc Default	AICc Tuned	AICc Delta	parameters Default	parameters Tuned
800m	Mean	1235.3	1133.1	9.0%	36.8	7.9
	Median	1229.9	1133.2	8.5%	37	8
	Range	197.9	2.9		24	1
30m	Mean	2339.0	2184.2	7.1%	42.7	8.4
	Median	2312.2	2184.5	5.8%	41	8
	Range	277.9	6.9		22	3

The AICc scores were more volatile on the default models compared to the tuned models. The box-and-whisker graphs (or simply "box charts") in Figure 18 illustrates this. A box chart provides a sense of both the range and the grouping of the data. The box represents the interquartile range between Q1 and Q3, containing 50% of the values. The line within the box is the median value. The "X" within the box is the mean. The "whisker" lines extending above and below the boxes represent either the upper and lower limits of the data range or 1.5 times the interquartile range if outliers are present (as is the case in the 800-meter graph here). Finally, outliers are represented by dots on the whisker lines. The tuned models are essentially flat. The default model interquartile range in the box chart is much broader, with a range on the default 800-meter model ~198 and the default 30-meter model ~278. with some extension of the range in the direction of the minimum values and more extension of the range in the direction of the maximum values.
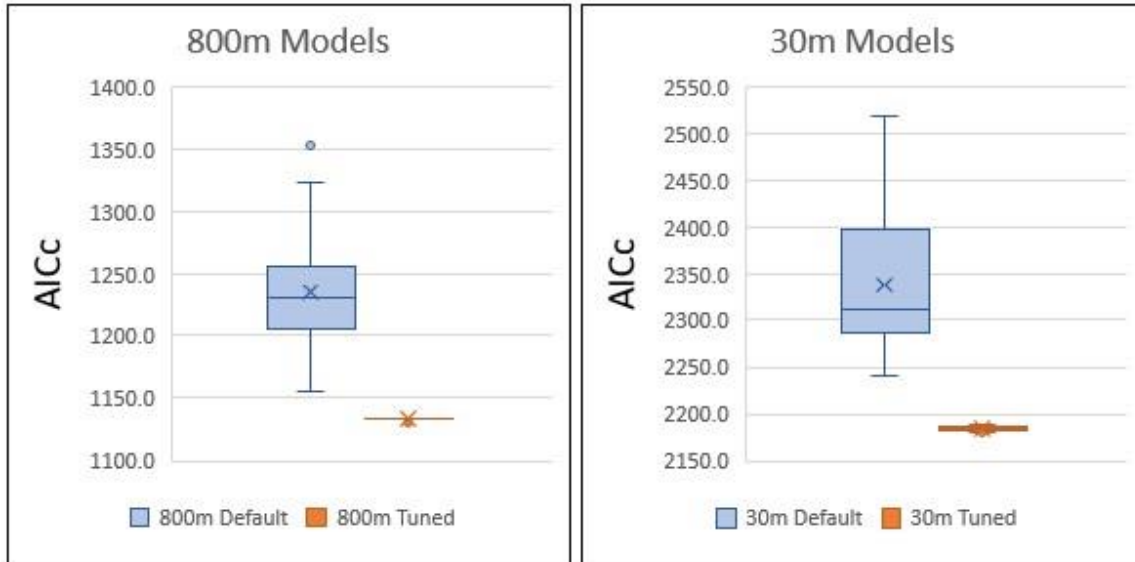
Figure 18 - AICc comparison

Reflecting the AICc values, the parameter counts on the default models varied widely (Table 10), with a range of 24 on the 800-meter model (24 to 48) and 22 on the 30-meter model (32 to 54). The tuned models showed much less volatility, with ranges of 1 for the 800-meter and 3 for the 30-meter models.

The summary of the AUC metrics (Table 11) showed little variation in the AUC scores across the iterations. Ranges were <0.3% on the 800-meter model and ~0.6% on the 30-meter model. While the AICc scores showed marked improvement with the tuned models, the AUC values for the tuned models were lower, indicating a slight decrease in discretionary ability compared to the default model. The decrease was 1.0% or less, however. All AUC scores were approximately 0.95 to 0.96, indicating very good discretionary ability of the models.

Table 11 - AUC summary

Resolution		AUC Default	AUC Tuned	AUC Delta
800m	Mean	0.9577	0.9498	-0.8%
	Median	0.9577	0.9499	-0.8%
	Range	0.0033	0.0023	
30m	Mean	0.9569	0.9475	-1.0%
	Median	0.9566	0.9472	-1.0%
	Range	0.0059	0.0059	

Omission rates for the tuned 800-meter model showed some slight variance with the or10pct metric, but all other omission rate measures were flat across the model iterations. The orMTP values were at zero. The or10pct rates were at ~9% for all the models except for the 800-meter tuned models. The 800-meter tuned models were slightly lower at ~8%. Overall, overfitting was not indicated on the models given these metrics.

Table 12 - Omission rate summary

Resolution		orMTP Default	orMTP Tuned	or10pct Default	or10pct Tuned
800m	Mean	0.0	0.0	0.091	0.085
	Median	0	0	0.091	0.079
	Range	0.0	0.0	0.000	0.013
30m	Mean	0.0	0.0	0.094	0.092
	Median	0	0	0.094	0.092
	Range	0.0	0.0	0.000	0.000

4.3. Utah Prediction Surfaces and Metrics

The models developed with the California study area data were used to create prediction surfaces for the southern Utah species presence area. AUC metrics were derived from these surfaces. AICc metrics were generated on the tuned models but the NA values on the default

model relegate this metric to information-only as there is no comparative value available. The models presented poor performance metrics at both resolutions and for both the default and tuned models (Table 13). The prediction surfaces generated (Figure 19 and Figure 20) did not align with known species locations in the extent.

The AUC values in the 0.6xxx range for both the models indicate poor discrimination performance. When compared to the California region (AUC Δ Default and AUC Δ Tuned columns in Table 13), the AUC values dropped ~27% for the default model and ~36%. While comparing AUC values between regions has known issues, changes of this magnitude are sufficient to indicate inadequate model fit.

Table 13 - Metrics for the Utah prediction surfaces

Resolution		UT AUC Default	UT AUC Tuned	AUC Delta	UT AICc Default	UT AICc Tuned	AUC Δ Default	AUC Δ Tuned
800m	Mean	0.6976	0.6056	-13.2%	NA	376.0	-27.2%	-36.2%
	Median	0.6979	0.6064	-13.1%	NA	376.5	-27.1%	-36.2%
	Range	0.0264	0.0145		NA	7.3		
30m	Mean	0.6434	0.6040	-6.1%	NA	637.2	-32.8%	-36.3%
	Median	0.6442	0.6018	-6.6%	NA	637.8	-32.7%	-36.5%
	Range	0.0516	0.0576		NA	31.2		

The tuned models demonstrated even worse performance than the default models in this scenario. The 800-meter results showed the AUC of tuned models ~13% below that of the default models. Likewise, the 30-meter AUC scores were off by ~6%.

Looking at the prediction rasters for the 800-meter models (Figure 19) and the 30-meter models (Figure 20) reinforces the findings of the metrics. A small population in the southwest corner of the extent and another in the north central area aligned with high probability areas in

the rasters, but most of the presence locations did not. Note in particular the locations in the

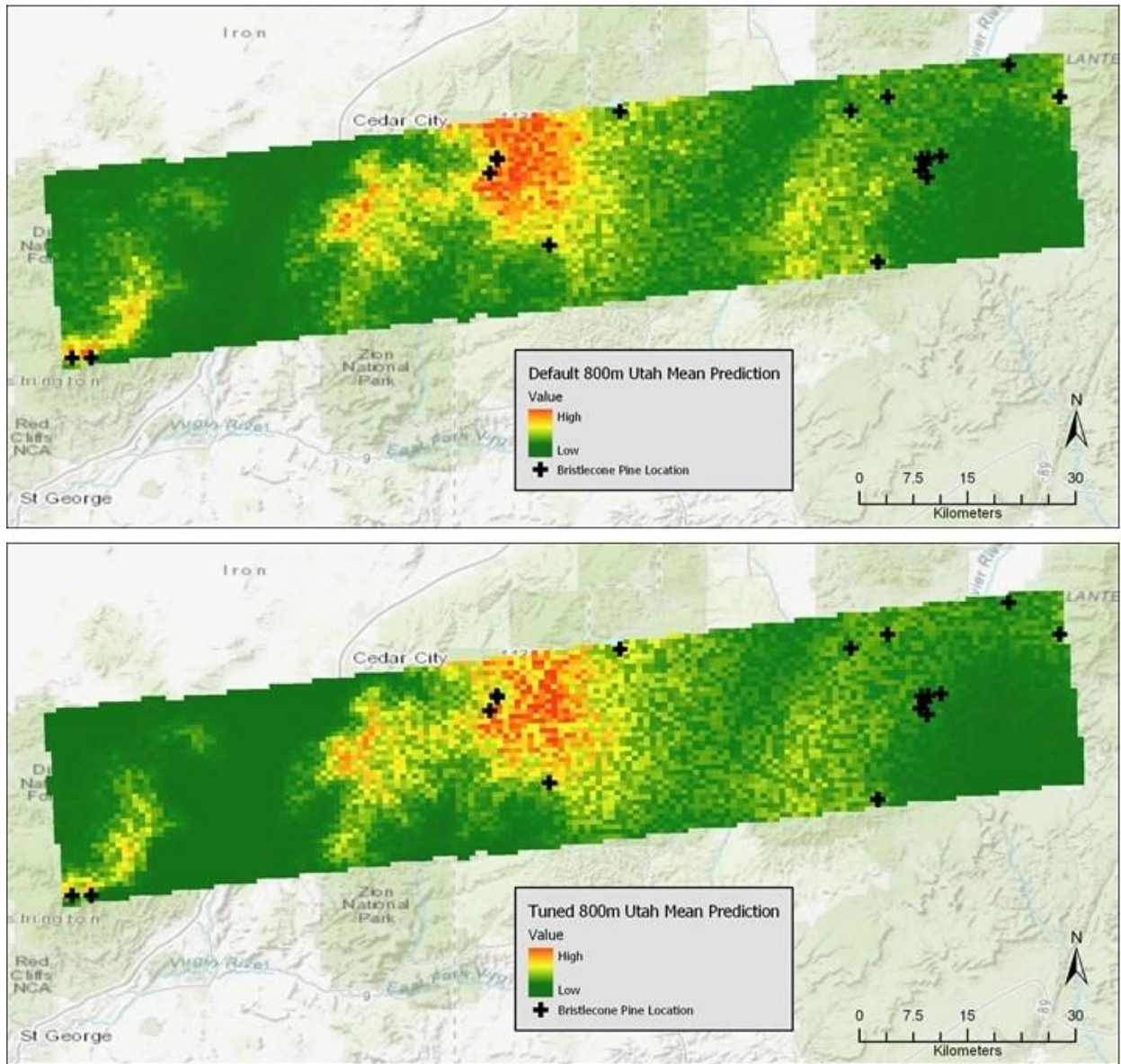eastern area of the extent that lie within low probability predictions.

Figure 19 - Utah 800-meter prediction rasters (Maxent logistic output)
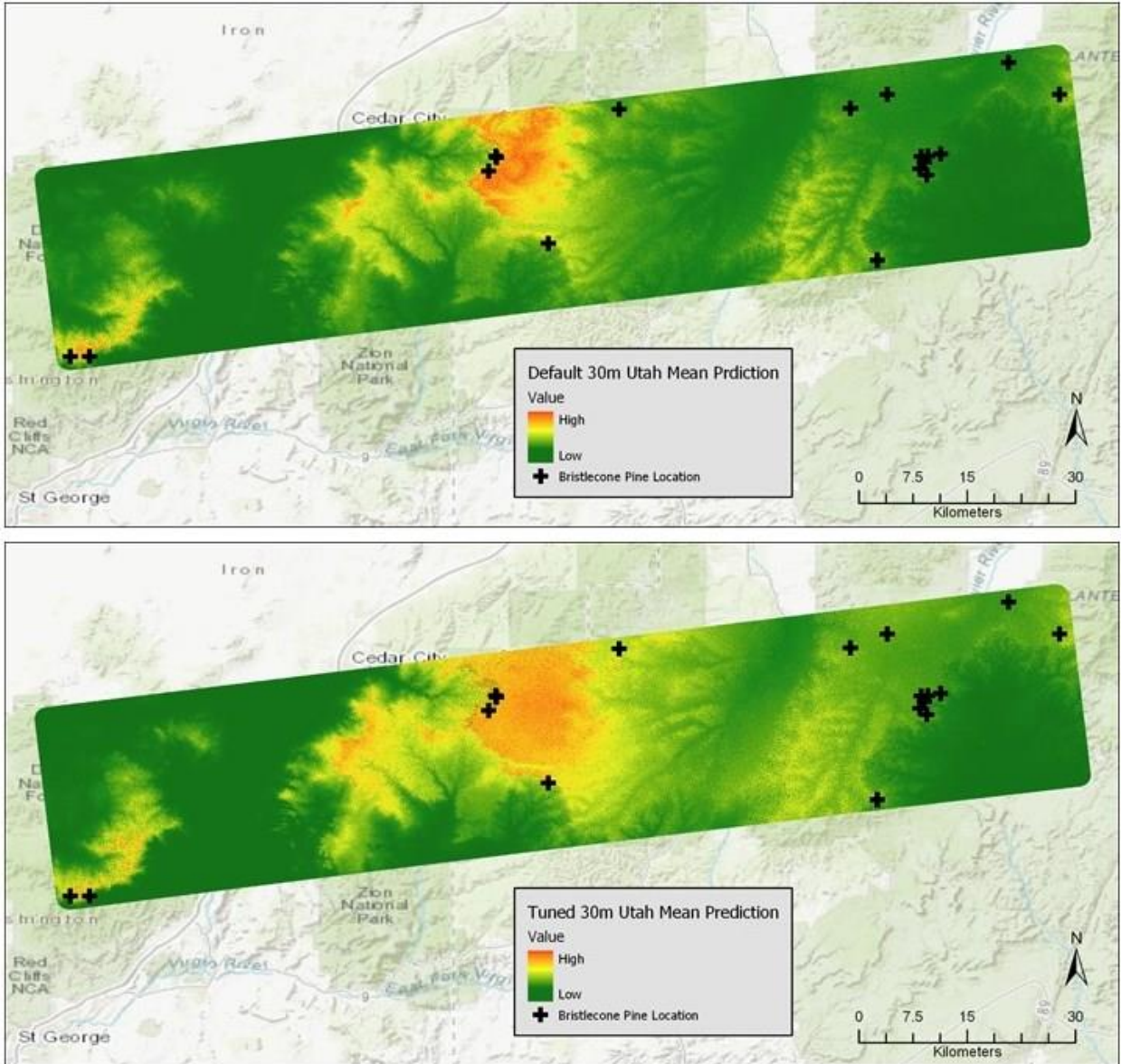
Figure 20 - Utah 30-meter prediction rasters (Maxent logistic output)

While a full evaluation of the Maxent covariate metrics, response curves, and jack-knife results is outside the scope of this study, examination of the covariate values at the California species presence locations versus the values at the Utah presence locations provided insights on why the model failed to transfer to the Utah extent. Two Maxent measures are briefly touched on here, then the impacts of four key covariates are examined in context of the model transfer to the Utah extent.

In the standard Maxent outputs, two measures of the covariate impact are provided in the "Analysis of variable contributions" section. Percent contribution represents the net contribution the covariate provided to the regularization gain as the model was iterating through the algorithm. To obtain the second measure, permutation importance, each variable is permutated in turn and the drop in the model's AUC evaluated. Large decreases, normalized to percentages, denotes the reliance of the model on that covariate. There are nuances and cautions around each of these measures (see Phillips, Anderson, and Schapire 2006), but they serve here as guides to picking a selection of covariates to examine for the Utah transfer evaluation.

Figure 21 shows the average percent contribution for the 25 model iterations. The DD<0 covariate provided the highest percent contribution for the default models at 30-meters and 800-meters. For the two tuned models, the MAT (mean annual temperature) covariate had the highest percent contribution.
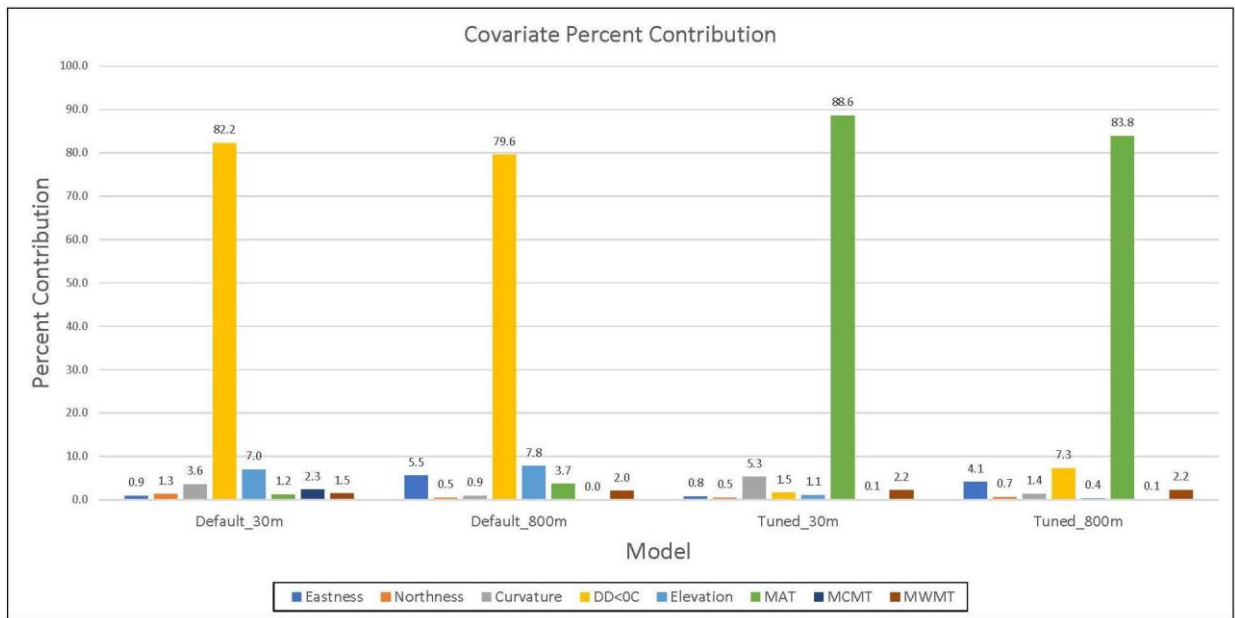

Figure 21 – Variable percent contribution for the four models

Figure 22 similarly shows the average permutation importance for the 25 model iterations. For this measure, the elevation covariate showed the most importance for the default models and the MWMT (mean warmest month temperature) was most important for the tuned models.
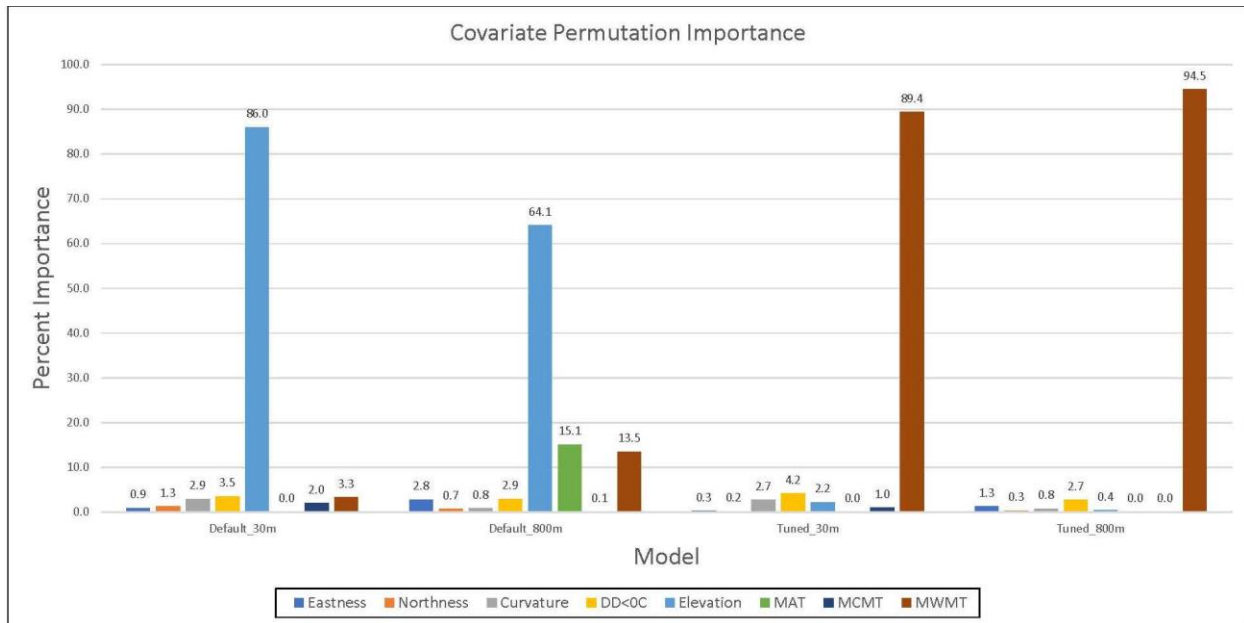

Figure 22 – Variable permutation importance for the four models

These four main covariates (DD<0, MAT, elevation, and MWMT) were the focus of the next evaluation step as they had the most impact on the models. Representative samples of the covariate response curves for the California models were compared to the covariate values at the Utah presence locations. Covariate value ranges at the presence locations in California and Utah were also compared.

The response curves plot Maxent's response to the variable's values. Keep in mind that the response curves shown provide the model gain with just the variable being used. No interaction with the other variables is yet implied in these curves. The response curves do shift and change as Maxent combines the variables to build the final model. The peaks of the response

curves will not always align with the groupings of covariate values as the Maxent modeling must adjust its internal algorithms as it attempts to fit all the values. However, these variable-only curves are good indicators of what covariate value ranges are influencing the final model.

In addition to the response curves for the four covariates, box charts were assembled for the covariate values at each of the presence locations in the California and Utah at both the 30-meter and 800-meter resolution. Composite graphs were assembled for each of the four covariates to facilitate examination of the data. Each composite consists of a representative response curve from a 30-meter model, one from an 800-meter model, and a box chart of the values of the presence location covariates in California and Utah. Note that the vertical green line in the response charts provided below is a visual clue added to the Maxent graphs to provide a common reference value between the chart pairs within each composite. This is needed as Maxent is inconsistent in how it scales the graphics of the response curves.

The graphs for DD<0 (degree days less than zero) are shown in Figure 23. DD<0 was the variable with the strongest contribution to the default models. The peak response to DD<0 is approximately 900, with a broad taper to either side. Looking at the variable values in the box chart, California encompasses a broader range of these DD<0 values than does the much narrower range of the Utah data. Utah had several outlier values, further complicating the fit of the California derived model on the Utah data.
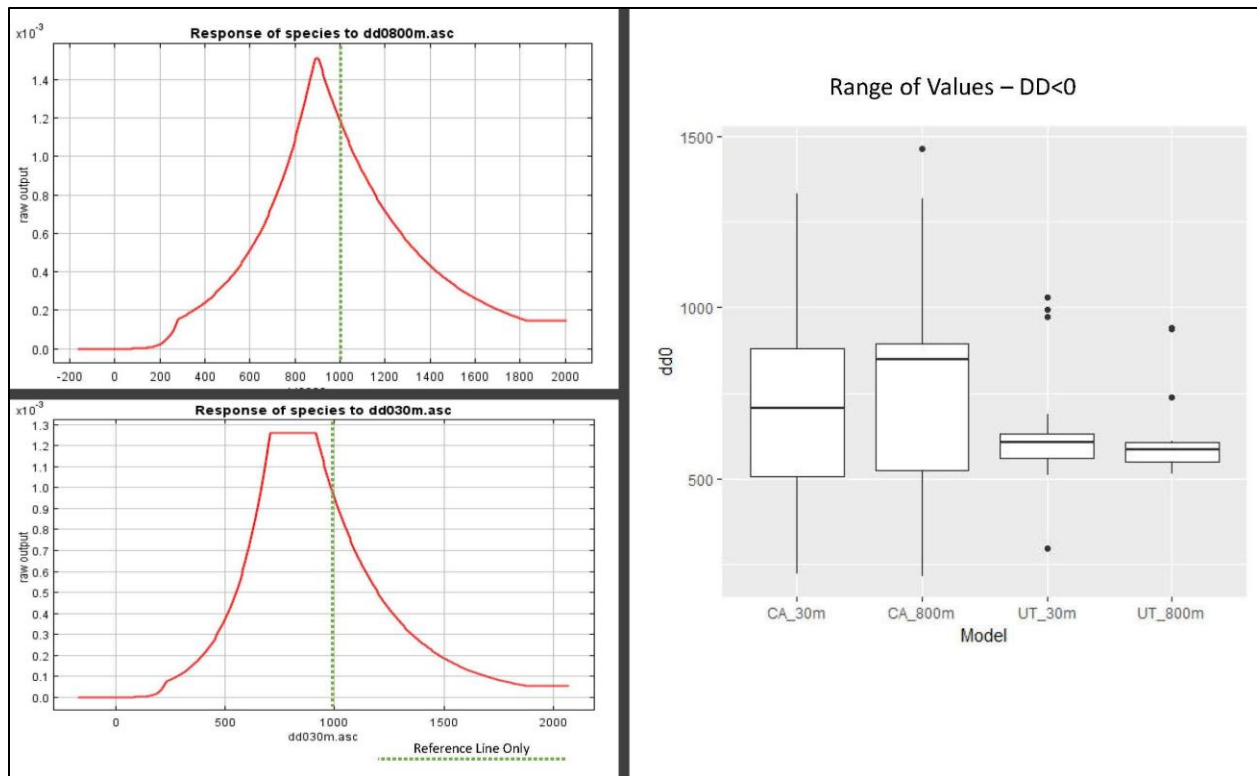
Figure 23 - DD<0 variable response curves and box chart of variable values

MAT (mean annual temperature), illustrated in Figure 24, was the variable with the biggest contribution to the tuned models. The peak of MAT values is approximately 0°C, with a steep falloff to either side of that value. Again, the California value range shown in the box chart encompass more of that range than the Utah values. The Utah values are also shifted upward significantly compared to the California values. The Utah presence locations are in a much warmer climate compared to the California presence locations, skewing the model for Utah.
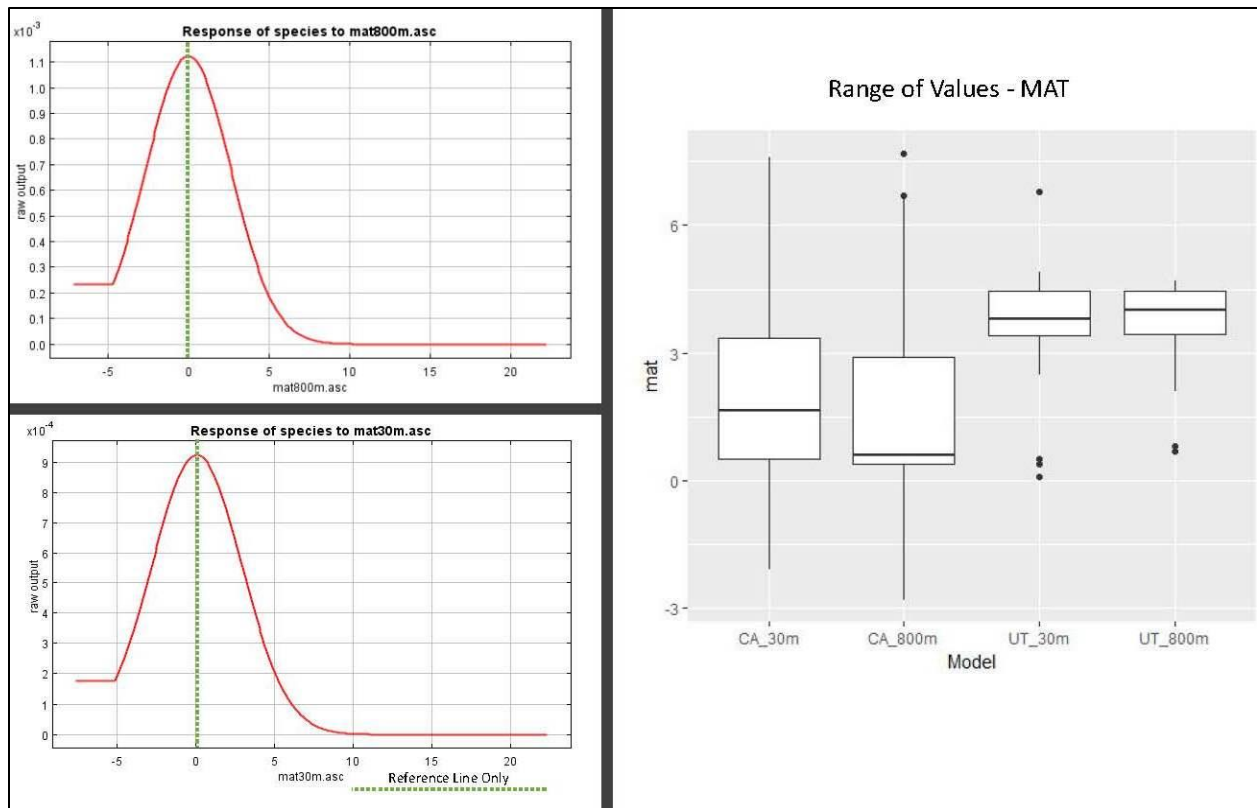
Figure 24 – MAT variable response curves and box chart of variable values

Elevation was the variable with the greatest importance for the default models. The response curves for elevation (Figure 25) had the most pronounced spike of all four variables. This is as expected given the bristlecones are a tree line species, located within a narrow elevation band. In this case, the curve aligns well with the range of values shown in the box chart for the California data. Notably, the Utah box charts show the elevation for the presence locations in the region are considerably lower than the California extent. The Utah ranges fall well down the left shoulder of the response curve for the models.
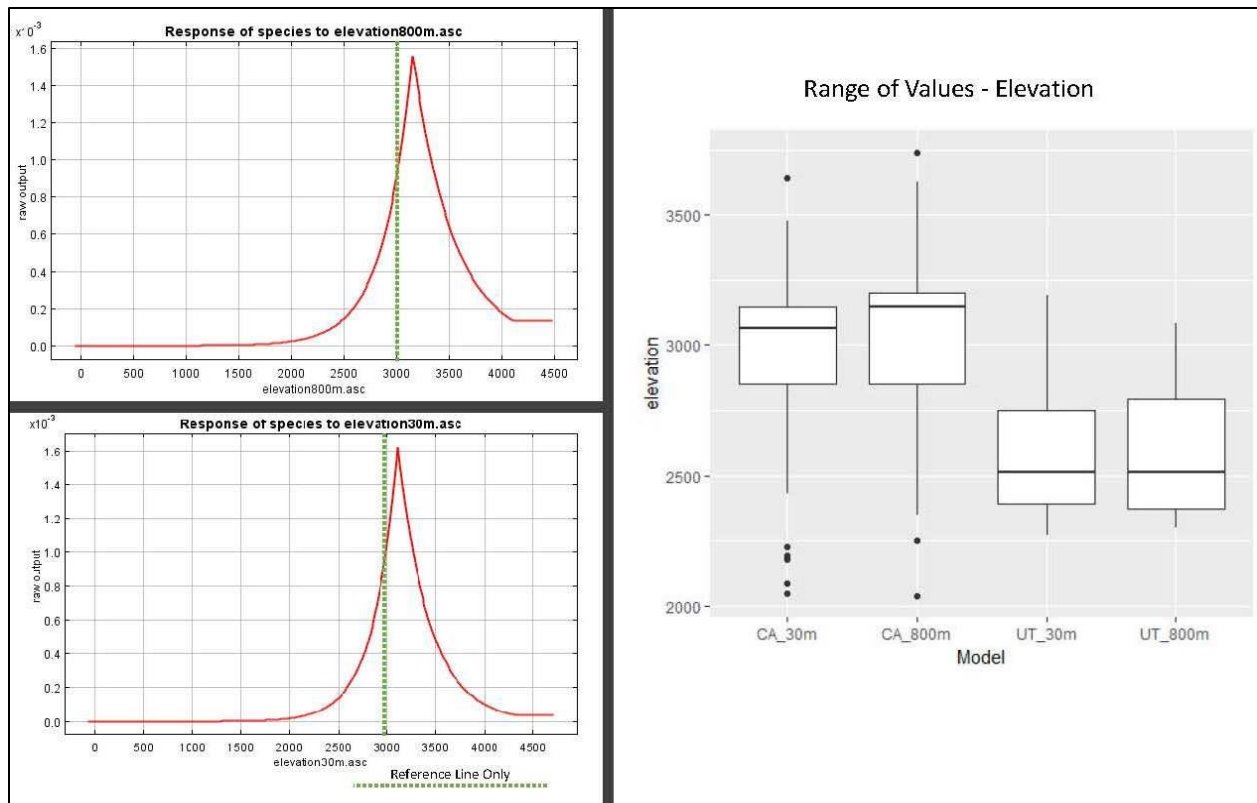
Figure 25 - Elevation variable response curves and box chart of variable values

The last variable to look at is MWMT (mean warmest month temperature), illustrated in Figure 26. This was the variable with the highest importance value for the tuned models. The response curves for MWMT were slightly more complex than the others. There is a plateau of response values up to the 5°C point, likely reflecting the colder, high elevation portions of the study area. Then a short rise to maximum followed by a long right shoulder on the curve. The values above 15°C in the curve likely represent the warmer, lowland areas around the edges of the study area. The California values shown in the box chart fall primarily in the upper portions of the right shoulder of the response curve. The covariate values for the Utah presence locations fall significantly higher than the California values, landing well down the response curve for MWMT, indicating the Utah presence locations are generally in much warmer terrain than their California counterparts.
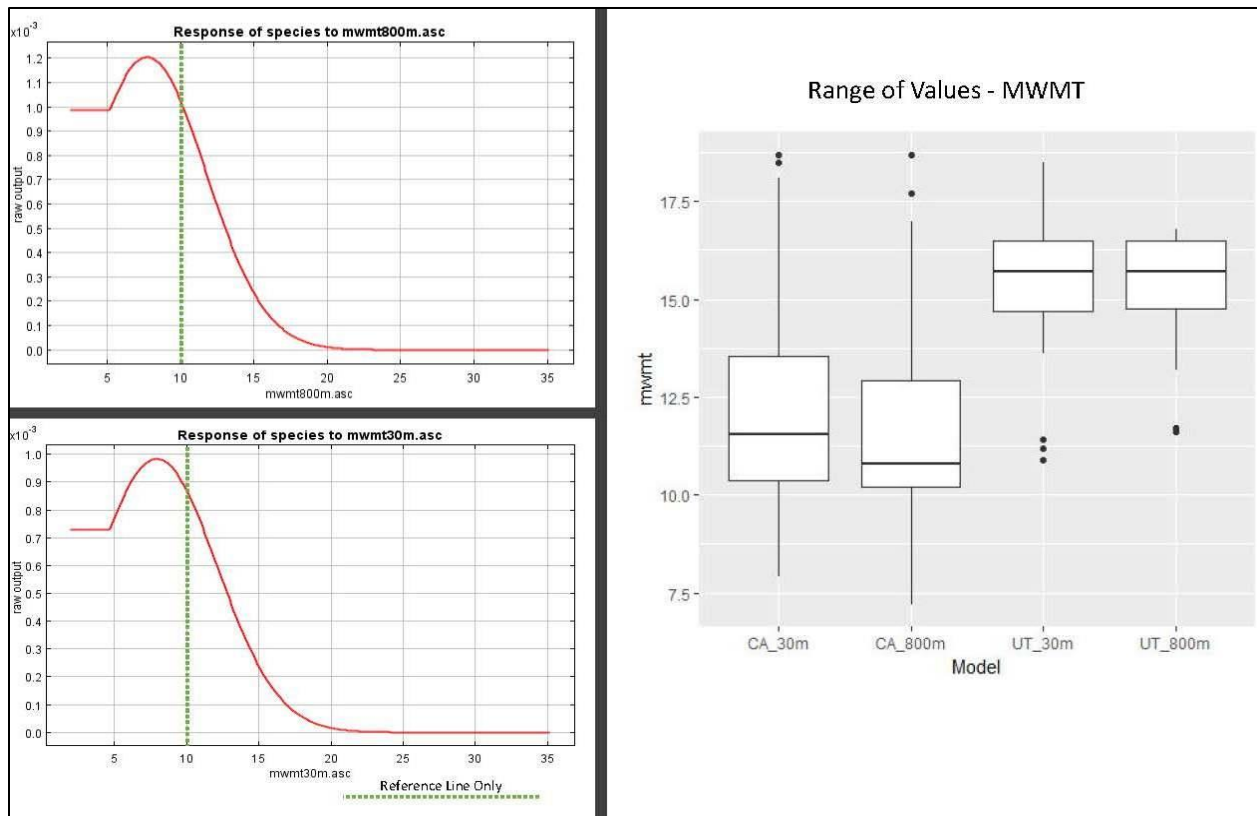
Figure 26 - MWMT variable response curves and box chart of variable values

In summary, using these four top predictive variables, it is evident that the ranges of the covariates for Utah presence locations were significantly different than the California presence locations the model was built on. Considered together with the metrics and the maps of the presence locations on the prediction surfaces, the Maxent models performed poorly when transported to the new region. Neither resolution exhibited a substantially better model in this case either.

4.4. Spatial Analysis of Model Tuning

The spatial analysis of the models was more subjective than the evaluations of the metrics and the Utah prediction surface. The analysis consisted of a visual inspection of the outputs from the Maxent models and from a hot spot analysis of the tuning changes. Each of the four model

prediction surfaces for California were compared to the species presence locations in the study area and evaluated for fit to the expected species distribution of the bristlecone pine. The HSA of the 800-meter delta surface was examined to see how the tuning of the model changed the prediction surfaces. As will be explained further, the HSA was not performed on the 30-meter delta data due to processing issues.

4.4.1. Comparing the Prediction Surfaces

As expected, the 800-meter prediction surfaces (Figure 27) showed less granularity than the 30-meter prediction surfaces (Figure 28). Owing to the resolution, the overall pattern of the species distribution predictions, for both the default and tuned models, were more sharply defined with the 30-meter surfaces than with the 800-meter surfaces. The 30-meter predictions showed higher predictions on the narrow ridgeline "tail" in the southern portion of the extent, aligning with the species locations in the area. The small central population showed similar results. The northern White Mountain grouping, the bulk of the observations, did not show as distinct a difference in predictions between resolutions.

Differences between the 800-meter default and tuned models (Figure 27) were mixed. The central and southern groupings showed some upward shift in the prediction values, but the northern area had little noticeable differences in the pattern. Given the tuned model had a slightly higher RM setting, more dispersal of the predictions was expected compared to the default model than what is seen in these results.

The differences between the 30-meter default and tuned models (Figure 28) exhibited a different pattern shift than the 800-meter. Overall, the prediction results were more dispersed on the tuned model, as would be expected given the loosening of the RM setting. This is most apparent with the northern area, somewhat less in the central area, but is limited in the southern

grouping. The pattern differences between north, central, and south areas were not as apparent in the 30-meter default versus tuned comparison as they were in the 800-meter models. The prediction results for the central and southern areas were very similar between the default and tuned models.
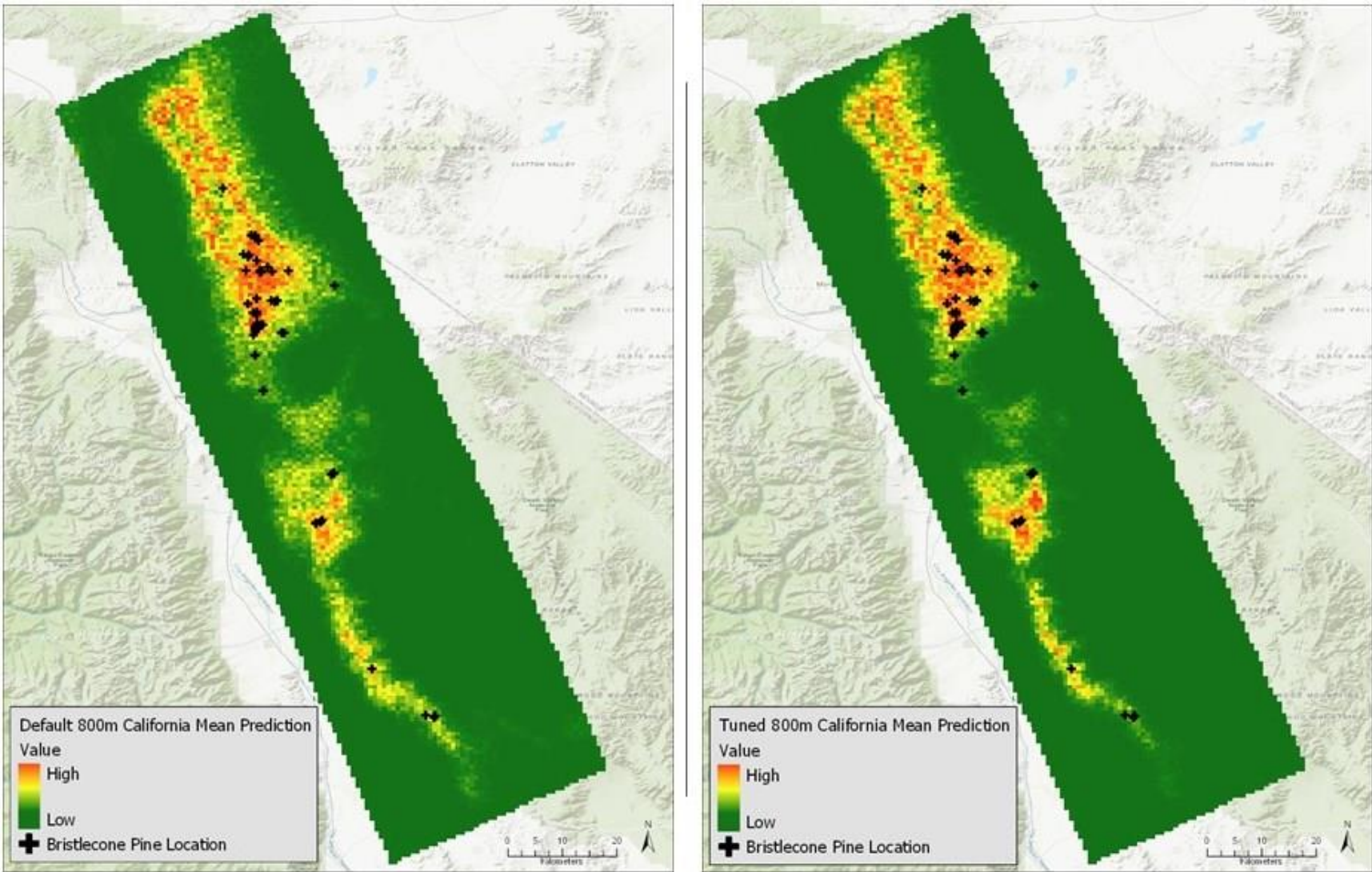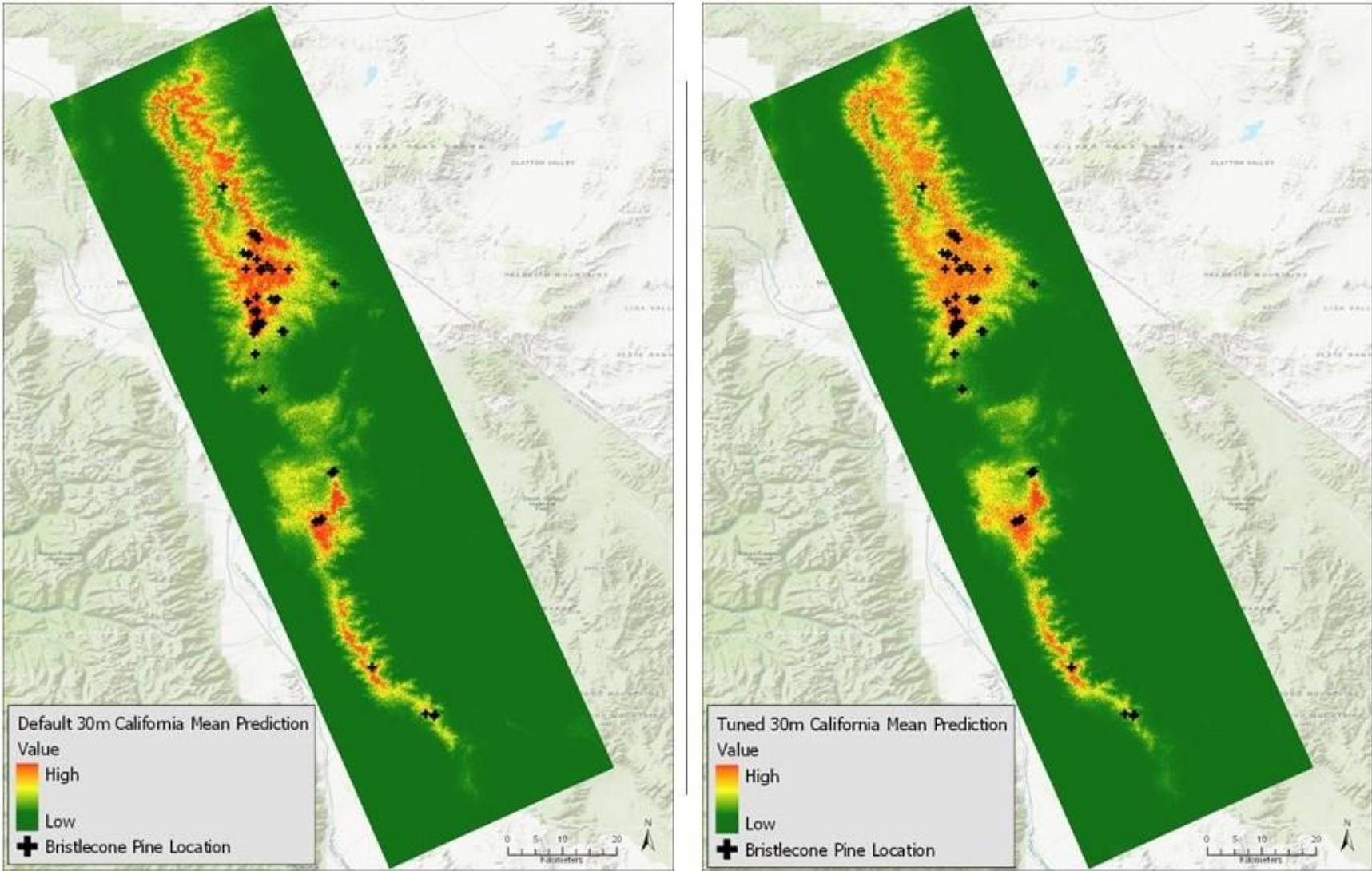
Figure 27 - 800-meter prediction rasters

Figure 28 - 30-meter prediction rasters

4.4.2. Hot Spot Analysis

Hot spot analysis was performed on the delta raster of the 800-meter default versus tuned prediction surfaces. The intent of the HSA is to evaluate where the tuned model shifted relative to the default model. The analysis does not provide quantitative data about the model quality but does provide insights into how Maxent is adjusting the models given the tuning process. This approach is somewhat unique and experimental.

The data presentation requires some explanation for proper interpretation. The delta raster was created by subtracting the tuned model mean cell values from the default model mean cell values. Therefore, a positive value in the delta surface indicates the default model presented a higher probability at that cell location than the tuned model. A negative value in the delta surface indicates the tuned model presented a higher probability than the default model. The other dynamic to keep in mind is that larger negative values show up as cold spots in the analysis as they are linearly smaller numbers, even though they represent a larger variance from the default in this study. The importance of this point becomes apparent when the negative deltas are explained.

Finally, keep in mind the maximum entropy approach used by Maxent to create the prediction surfaces sums to unity across the surface in the raw output format. A value increased or decreased in one cell is offset by a corresponding decrease or increase in another cell on the surface. This offset is reflected in the logistics output format that is the basis of the spatial analysis comparisons here.

Four different views of the delta information were examined. The raw delta values (left panel, Figure 29) and the absolute values (right panel, Figure 29) provided some information on the tuning impacts, but the delta values filtered to just include the positive values (left panel,

Figure 30Figure 29) or the negative values (right panel, Figure 30) were more informative about the tuning shifts.

The hot spots in the raw values (red areas) in Figure 29's left panel shows the areas where the default model assigned higher probabilities than the tuned model, with the cold spots (blue areas) showing where the tuned model presented higher than the default model. The overall pattern suggests the dispersion of the tuned model probabilities given the loosening of the model fit with the higher Maxent RM value. The hot spots align roughly with the ridgeline of the terrain in the study area. This suggests the tuned model shifted probabilities from the highlands to the lower elevation areas, but the pattern is general. The absolute value data (Figure 29, right panel), representing the variance of the delta data, can be interpreted similarly.
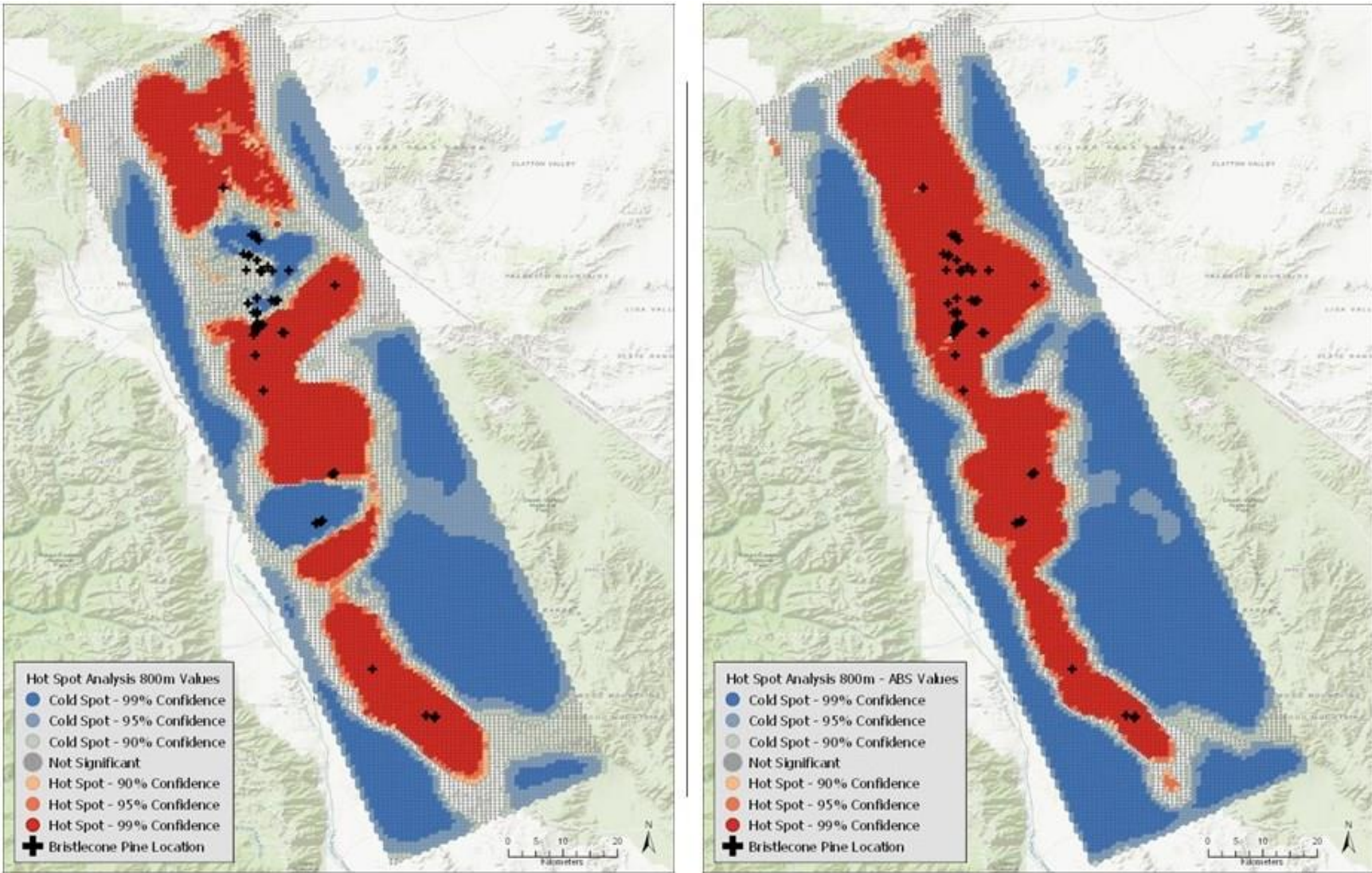
Figure 29 - 800-meter hot spot analysis for delta values and delta absolute values
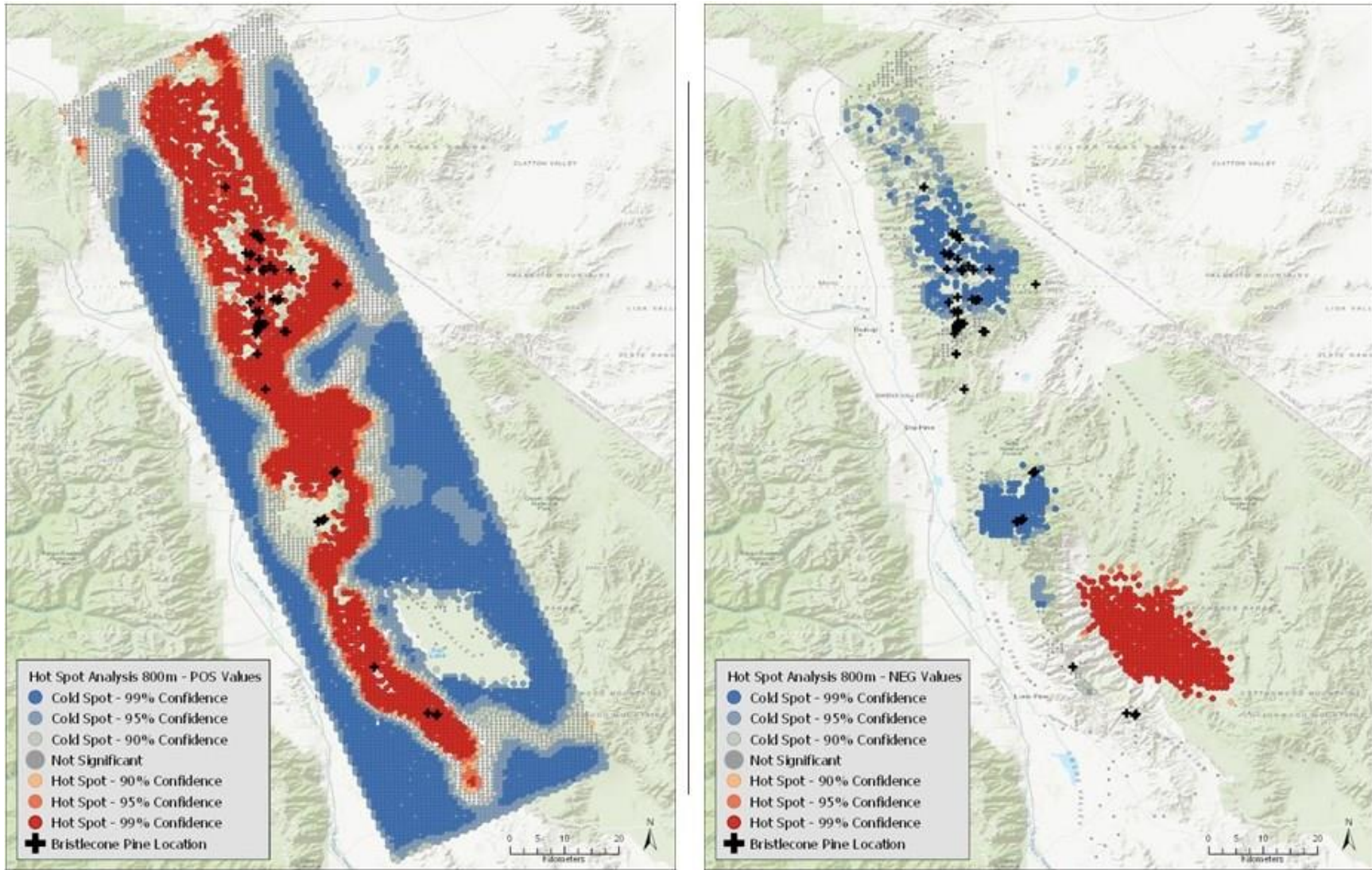
Figure 30 - 800-meter hot spot analysis for delta positive values and delta negative values

Looking at just the positive values (left panel, Figure 30) is somewhat more helpful. The pattern is slightly more concise than the raw and absolute values. The hot spots represent areas where the tuned model reduced the probability significantly, the cold spots where the tuned model reduced values less. In both cases, the data represents where the default model had larger probabilities than the tuned model. Here we can see the default model had a high bias toward the highland areas as would be expected given the tighter fitting of the model to the species presence locations. The hot spots, with a couple exceptions in the central area of the extent, align closely with the bristlecone locations used to create the model.

The negative delta data (right panel, Figure 30) further explained the tuning changes, with the most concise data of the four HSAs. Given the values are all negative in this delta data, cold spots represent the largest variance (larger probabilities in the tuned model), hot spots the least variance between the default and tuned models. The large hot spot area in the southern portion of the study area corresponds to Saline Valley in Death Valley National Park. As the name suggests, this is a salt flat, very far from ideal bristlecone habitat. The hot spot, in this case, indicates that the tuned model changed the prediction values very little compared to the default model. This indicates that both models correctly identified this region as unfit for the species. The tuned model had no need to shift prediction values from the area as they were already low.

The cold spots provide insight on the areas where the tuned model dispersed probabilities with the loosening of the model fit. When comparing the default and prediction surfaces, the dispersion of the probabilities with the loosening of the tuned model fit was not apparent. In this negative delta view, the probability dispersion in the northern population group is more defined, especially when contrasted with the positive delta values. The tuned model values were drawn

down in the cells nearest the species probabilities (the hot spots in the positive deltas) and dispersed slightly to adjacent areas (the cold spots in the negative deltas).

This pattern, however, is less distinct in the central population group and non-existent in the HSA of the southern population. There is a strong cold spot present in the negative deltas for the central population, but it does not align well with the presence locations there nor exhibit as strong a pattern compared to the positive deltas. There is no strong cold spot in the negative deltas for the southern group, though there is a strong hot spot present in the positive deltas.

HSA was not completed on the 30-meter surfaces given the size of the dataset and unresolved issues with the ArcGIS tool handling the data complexities. The Esri "Optimized Hot Spot Analysis" tool was unable to complete the HSA, giving warnings regarding the size of the dataset, the large number of outliers, and the large number of invalid (i.e. null) records. It was unable to establish a distance band for the analysis with the data, causing the tool to fail. Deeper examination of the tool functionality and the data issues was beyond the capabilities or scope of this study. The decision was to proceed with just the 800-meter analysis as the HSA is a unique and experimental approach to evaluating the Maxent models in any case.

Chapter 5 Conclusions

This study explored two important aspects of species distribution modeling using the Maxent modeling tool. Do higher resolution covariates yield a more accurate Maxent model of possible habitat extent than Maxent models with lower resolution covariates? Does a tuned Maxent model yield a more accurate model of possible habitat extent than a model using default settings? The study examined these questions by creating tuned and default models at 800-meter and 30-meter resolutions, then comparing the model performance metrics and predicted habitat extents.

The tuned models did demonstrate improved AICc metrics compared to the default models. The AUC and omission rate measures of the tuned models were comparable or better than the default models across the California study area. As reflected in the AICc score and the number of model parameters, the tuned models were much less complex and demonstrated more stable outputs across multiple iterations.

The subjective evaluation of the prediction surfaces of the tuned versus default was less conclusive. Qualitatively, the 800-meter tuned prediction surface did not appear to vary significantly from the default model. The 30-meter tuned prediction surface was not significantly different in the central and southern population areas of the study extent. The northern population area did show a wider dispersion of potential habitat in the tuned model, as would be expected given the loosening of the fit of the tuned model. Whether this is more accurate or not would require additional ground-truth data for validation of the species distribution.

The lower AUC scores of the models used in the Utah prediction area are a concern. Both the tuned and the default models performed poorly when transported to the new extent, with the tuned model showing even more degradation of performance than the default model. Further study is required to determine the cause of the degradation. While the same species, the covariate

mix for the Utah bristlecone population may be different enough from the California range as to throw off the modeling results. Alternatively, the covariate mix for the study area may have been too narrowly defined in relation to the broader use of the species model. Covariate selection for this study was based on species specific recommendations from expert studies. Broadening the selection of covariates and testing the impacts on the models would be required to validate these suggestions.

Overall, the HSA provided some insights into the tuning behavior of the Maxent models. The expected dispersion of the species probabilities from the loosening of the model fit was more apparent in parts of the study extent, but not conclusive across the full range. The lack of additional HSA data from the 30-meter models limited insights on the tuning behavior across the models in this study. While not definitive, the use of HSA for spatial analysis demonstrated potential areas of further study for quantifying Maxent model comparisons.

The findings on the question of resolution were inconclusive. Comparing metrics across models that have such a wide range of differences in extent (cell count in this case) is an area of study that continues to be elusive in offering solutions (Lobo, Jiménez-Valverde, and Real 2007; Peterson et al. 2011; Warren and Seifert 2011). This was apparent in this study. The AICc metric was of no value in comparing across resolutions, though it did prove very useful in the tuning exercise and in comparing the default to tuned models within a resolution. The AUC and omission rate measures did provide some quantitative value, though with due caution (Lobo, Jiménez-Valverde, and Real 2007). The visual evaluation of the prediction surfaces allowed for limited qualitative comparisons.

In the case of the metrics, no clear distinction was obvious between the 30-meter and 800-meter models. Both scored very well on the AUC. Omission rate measures were strong on

both as well. In fact, the measures for AUC and omission rates were nearly identical between the resolutions. Based on just these measures, no model resolution was clearly superior to the other.

Subjectively, the 30-meter models did present a more defined species distribution than the 800-meter models in the graphics. This was, no doubt, a factor of the granularity of the covariate grids, but this observation was not supported by the metrics discussed above. Whether the finer resolution yielded a quantitatively superior model would require substantial ground-truthing of the suggested habitat extent and further analysis. From a practical standpoint, it was not clear the 30-meter models presented a better estimate of the potential habitat of the species across the study extent. The suggested potential habitat of both model resolutions appears to roughly align.

Compounding all the analysis are some serious questions around data uncertainty. As seen in the data section above, both the climate data and the DEM data have defects. For the 800-meter models, the climate variables and the elevation data are created from USGS NED products that have a vertical margin of error of ± 30 meters. That is a significant number given the terrain modeled in this study. Further, aggregation techniques were used by the PRISM team to scale the USGS NED data from the 90m source to the 800-meter resolution. This aggregated data was used in the climate modeling. It was also the source elevation data used to create the elevation 800-meter covariates for the study. The uncertainty of the underlying 800-meter elevation data propagated to all the covariates directly and indirectly, with the uncertainty possibly being magnified throughout the processing chain.

The 30-meter data had substantial issues as well. While the USGS 3DEP DEM data are of high quality, the climate covariates, being derived from the PRISM model, inherit all the same issues as the 800-meter models described above. While ClimateWNA was used to scale the data

down to 30-meter resolution, it is using the same PRISM source data as the 800-meter models. Adding to the issues, cWNA was shown to have limits on its ability to disaggregate the climate data below a 200-meter resolution. Between the uncertainty in the PRISM data, the uncertainty introduced by the cWNA disaggregation of that data, and the known limits of cWNA to create data surfaces below a resolution of 200-meter, considerable questions exist on the data basis for the climate covariates of the 30-meter models.

In summary, there is demonstrated value in applying tuning exercises before executing models in Maxent. This is already well documented in the literature and confirmed with this further study. Tuned models quantitatively perform better than the default models, are less complex, and the tuning exercise allows the models to be tailored to the specifics of the species and covariate data provided.

On the question of the benefits of using higher resolution data, the author's opinion is that the climate data does not support moving beyond the 800-meter resolution currently offered by the major climate models such as PRISM and WorldClim. While advances have been made in providing high quality DEM data at finer resolutions, the climate models are not at the same level of quality at comparable resolutions. Perhaps they never will be. The question to be explored is whether it even makes sense to consider climate variables below a certain spatial resolution given the dynamics of what is being measured. For example, are temperature or precipitation measures at 30-meter resolution even meaningful, even if methods were devised to interpolate data to that level? Another key question is what is the meaningful resolution for the species under study? While 30-meter resolution may be useful (if appropriate data can be found) for a plant, how meaningful would it be for large vagile species such as mountain sheep or wolves?

This study highlighted several areas worthy of further research. Several have been mentioned in passing above. To start, the impact of the climate data resolution on SDMs offers several avenues to explore. What is the impact to SDMs of the various aggregation, disaggregation, and interpolation techniques used to align climate data with the other covariates? Which of these techniques are better to employ than others and why? What is the smallest resolution supported by the current state of the climate models? Is the 200m limit seen in the cWNA data that limit? Overall, the spatial aspects of climate data used in SDMs has not received the same level of scrutiny as seen with DEMs and modeling of soil movement and waterflow. It is possible there are parallels in research and methods in those areas of modeling that can be applied to the climate models and their use in SDMs.

There was some evidence of spatial bias shown in the tuning tests in this study, highlighted by the high omission rates in several of the partitioning schemes. The use of the partitioning schemes in the tuning tests, in particular the block and checkerboard methods, was meant to take that bias into account when determining the tuning parameters. The results of the testing done were sufficient for the scope of the study. However, further evaluation of this potential bias is warranted if a more thorough species habitat distribution for bristlecone pines was needed.

The biggest gap highlighted in the study was the need for more conclusive, consistent measures of SDM performance across models, across resolutions, and across extents. It proved to be the most difficult and frustrating part of the research. While the metrics chosen were valid for the scope of study, they were less than satisfying given the various limitations and caveats for their use. The picture they paint of the models, even when viewed collectively, leaves large gaps in understanding of the model effectiveness. The inconsistency of the AUC rankings versus the

AICc rankings versus the omission rates in the tuning tests is an example. Choices of what is "best" often resolve to subjective decisions. As was discussed in Chapter 2, finding these ideal measures is already a contentious and active area of study within the modeling community. Progress is being made, though slowly.

The use of the hot spot analysis was an intriguing experiment, attempting to understand the spatial dynamics of the model tuning. It offered some value in understanding the changes Maxent was applying to the results but yielded little in the way of why. The HSA technique was developed with other uses in mind. Exploration of the tool, perhaps using derivatives more aligned with the need to measure model results, would be an interesting and potentially valuable contribution to the body of knowledge surrounding SDMs.

Returning to George Box's quote, this study has highlighted both where "all models are wrong", but hopefully also where "some are useful". The climate models, despite the issues mentioned, are no doubt invaluable to SDM performance. Tuning also has repeatedly been shown to produce better models, both in this study and in a wide range of other literature. If there is one message to be taken from this thesis, it is to know where your models are wrong so that you may fully appreciate where they are most useful.

References

Allouche, Omri, Asaf Tsoar, and Ronen Kadmon. 2006. "Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS)." *Journal of Applied Ecology* 43 (6):1223-1232. doi: 10.1111/j.1365-2664.2006.01214.x.

Anderson, Robert P., and Israel Gonzalez. 2011. "Species-specific tuning increases robustness to sampling bias in models of species distributions: An implementation with Maxent." *Ecological Modelling* 222 (15):2796-2811. doi: 10.1016/j.ecolmodel.2011.04.011.

Araújo, Miguel B., and A. Townsend Peterson. 2012. "Uses and misuses of bioclimatic envelope modeling." *Ecology* 93 (7):1527-1539. doi: 10.1890/11-1930.1.

Arnold, Jeffrey B. 2018. "ggthemes: Extra Themes, Scales and Geoms for 'ggplot2'", ver. 4.0.1. https://CRAN.R-project.org/package=ggthemes.

Baldwin, A. Roger. 2009. "Use of Maximum Entropy Modeling in Wildlife Research." *Entropy* 11 (4). doi: 10.3390/e11040854.

Beck, Jan, Liliana Ballesteros-Mejia, Peter Nagel, and Ian J. Kitching. 2013. "Online solutions and the 'Wallacean shortfall': what does GBIF contribute to our knowledge of species' ranges?" *Diversity and Distributions* 19 (8):1043-1050. doi: 10.1111/ddi.12083.

Beck, Jan, Marianne Böller, Andreas Erhardt, and Wolfgang Schwanghart. 2014. "Spatial bias in the GBIF database and its effect on modeling species' geographic distributions." *Ecological Informatics* 19:10-15. doi: 10.1016/j.ecoinf.2013.11.002.

Bedia, Joaquín, Sixto Herrera, and José Manuel Gutiérrez. 2013. "Dangers of using global bioclimatic datasets for ecological niche modeling. Limitations for future climate projections." *Global and Planetary Change* 107:1-12. doi: 10.1016/j.gloplacha.2013.04.005.

Bivand, Roger, Tim Keitt, and Barry Rowlingson. 2018. "rgdal: Bindings for the 'Geospatial' Data Abstraction Library", ver. 1.3-6. https://CRAN.R-project.org/package=rgdal.

Boakes, Elizabeth H, Philip JK McGowan, Richard A Fuller, Ding Chang-qing, Natalie E Clark, Kim O'Connor, and Georgina M Mace. 2010. "Distorted views of biodiversity: spatial and temporal bias in species occurrence data." *PLoS Biology* 8 (6):e1000385. doi: 10.1371/journal.pbio.1000385.

Boria, Robert A., Link E. Olson, Steven M. Goodman, and Robert P. Anderson. 2014. "Spatial filtering to reduce sampling bias can improve the performance of ecological niche models." *Ecological Modelling* 275:73-77. doi: 10.1016/j.ecolmodel.2013.12.012.

Box, George E. P., and Norman R. Draper. 1987. *Empirical model-building and response surfaces*, *Empirical model-building and response surfaces*. Oxford, England: John Wiley & Sons.

Bruening, Jamis M, Tyler J Tran, Andrew G Bunn, Stuart B Weiss, and Matthew W Salzer. 2017. "Fine-scale modeling of bristlecone pine treeline position in the Great Basin, USA." *Environmental Research Letters* 12 (1):014008. doi: 10.1088/1748-9326/aa5432.

Bruening, Jamis M. 2016. "Fine-scale Topoclimate Modeling and Climatic Treeline Prediction of Great Basin Bristlecone Pine (*Pinus longaeva*) in the American Southwest." Master of Science Masters Thesis, Environmental Sciences, Western Washington University.

Burnham, Kenneth P, and David R Anderson. 2003. *Model selection and multimodel inference: a practical information-theoretic approach*. New York, NY: Springer Science & Business Media.

Connor, Thomas, Vanessa Hull, Andrés Viña, Ashton Shortridge, Ying Tang, Jindong Zhang, Fang Wang, and Jianguo Liu. 2017. "Effects of grain size and niche breadth on species distribution modeling." *Ecography* 41 (8):1270-1282. doi: 10.1111/ecog.03416.

Coops, Nicholas C., and Richard H. Waring. 2011. "Estimating the vulnerability of fifteen tree species under changing climate in Northwest North America." *Ecological Modelling* 222 (13):2119-2129. doi: 10.1016/j.ecolmodel.2011.03.033.

Currey, Donald R. 1965. "An Ancient Bristlecone Pine Stand in Eastern Nevada." *Ecology* 46 (4):564-566. doi: 10.2307/1934900.

Daly, Christopher. 2006. "Guidelines for assessing the suitability of spatial climate data sets." *International Journal of Climatology* 26 (6):707-721. doi: 10.1002/joc.1322.

Daly, Christopher, and Kirk Bryant. 2013. "The PRISM climate and weather system—an introduction." Accessed November 12, 2018. http://www.prism.oregonstate.edu/documents/PRISM_history_jun2013.pdf.

Daly, Christopher, Michael Halbleib, Joseph I. Smith, Wayne P. Gibson, Matthew K. Doggett, George H. Taylor, Jan Curtis, and Phillip P. Pasteris. 2008. "Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States." *International Journal of Climatology* 28 (15):2031-2064. doi: 10.1002/joc.1688.

Daly, Christopher, GH Taylor, WP Gibson, TW Parzybok, GL Johnson, and PA Pasteris. 2000. "High-quality spatial climate data sets for the United States and beyond." *Transactions of the ASAE* 43 (6):1957-1962. doi: 10.13031/2013.3101.

Edwards, James L. 2004. "Research and Societal Benefits of the Global Biodiversity Information Facility." *BioScience* 54 (6):485-486. doi: 10.1641/0006-3568(2004)054[0486:RASBOT]2.0.CO;2.

Elith, Jane, and Catherine H. Graham. 2009. "Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models." *Ecography* 32 (1):66-77. doi: 10.1111/j.1600-0587.2008.05505.x.

Elith, Jane, Catherine H. Graham, Robert P. Anderson, Miroslav Dudík, Simon Ferrier, Antoine Guisan, Robert J. Hijmans, Falk Huettmann, John R. Leathwick, Anthony Lehmann, Jin Li, Lucia G. Lohmann, Bette A. Loiselle, Glenn Manion, Craig Moritz, Miguel Nakamura, Yoshinori Nakazawa, Jacob McC. M. Overton, A. Townsend Peterson, Steven J. Phillips, Karen Richardson, Ricardo Scachetti-Pereira, Robert E. Schapire, Jorge Soberón, Stephen Williams, Mary S. Wisz, and Niklaus E. Zimmermann. 2006. "Novel methods improve prediction of species' distributions from occurrence data." *Ecography* 29 (2):129-151. doi: 10.1111/j.2006.0906-7590.04596.x.

Elith, Jane, and John R. Leathwick. 2009. "Species Distribution Models: Ecological Explanation and Prediction Across Space and Time." *Annual Review of Ecology, Evolution, and Systematics* 40 (1):677-697. doi: 10.1146/annurev.ecolsys.110308.120159.

Elith, Jane, Steven J. Phillips, Trevor Hastie, Miroslav Dudík, Yung En Chee, and Colin J. Yates. 2011. "A statistical explanation of MaxEnt for ecologists." *Diversity and Distributions* 17 (1):43-57. doi: 10.1111/j.1472-4642.2010.00725.x.

Esri. 2000. "Understanding Map Projections." Accessed March 25, 2018. http://downloads2.esri.com/support/documentation/ao_/710Understanding_Map_Projecti ons.pdf.

Esri. 2018. "ArcGIS Pro", ver. 2.2.4. Esri Inc., Redlands, CA.

Fitzpatrick, Matthew C., Nicholas J. Gotelli, and Aaron M. Ellison. 2013. "MaxEnt versus MaxLike: empirical comparisons with ant species distributions." *Ecosphere* 4 (5):art55. doi: doi:10.1890/ES13-00066.1.

Fourcade, Yoan, Aurélien G. Besnard, and Jean Secondi. 2017. "Paintings predict the distribution of species, or the challenge of selecting environmental predictors and evaluation statistics." *Global Ecology and Biogeography* 27 (2):245-256. doi: 10.1111/geb.12684.

Fourcade, Yoan, Jan O. Engler, Dennis Rödder, and Jean Secondi. 2014. "Mapping Species Distributions with MAXENT Using a Geographically Biased Sample of Presence Data: A Performance Assessment of Methods for Correcting Sampling Bias." *PLOS ONE* 9 (5):e97122. doi: 10.1371/journal.pone.0097122.

Fournier, Alice, Morgane Barbet-Massin, Quentin Rome, and Franck Courchamp. 2017. "Predicting species distribution combining multi-scale drivers." *Global Ecology and Conservation* 12:215-226. doi: 10.1016/j.gecco.2017.11.002.

Franklin, Janet, Frank W. Davis, Makihiko Ikegami, Alexandra D. Syphard, Lorraine E. Flint, Alan L. Flint, and Lee Hannah. 2012. "Modeling plant species distributions under future climates: how fine scale do climate projections need to be?" *Global Change Biology* 19 (2):473-483. doi: 10.1111/gcb.12051.

Franklin, Janet, and Jennifer A. Miller. 2014. *Mapping species distributions: Spatial inference and prediction*. Cambridge: Cambridge University Press.

Galante, Peter J., Babatunde Alade, Robert Muscarella, Sharon A. Jansa, Steven M. Goodman, and Robert P. Anderson. 2018. "The challenge of modeling niches and distributions for data-poor species: a comprehensive approach to model complexity." *Ecography* 41 (5):726-736. doi: 10.1111/ecog.02909.

GBIF. 2018. "GBIF Occurrence Download, *Pinus longaeva*." Accessed March 3, 2018. https://doi.org/10.15468/dl.2yfsa3.

Gesch, Dean. 2007. "Chapter 4: The National Elevation Dataset." In *Digital elevation model technologies and applications: the DEM users manual*, edited by David F. Maune, 99-118. Bethesda, Maryland: American Society for Photogrammetry and Remote Sensing.

Gesch, Dean, Michael J. Oimoen, and Gayla A. Evans. 2014. *Accuracy assessment of the U.S. Geological Survey National Elevation Dataset, and comparison with other large-area elevation datasets-SRTM and ASTER: U.S. Geological Survey Open-File Report 2014–1008*. Reston, VA: U. S. Geological Survey. doi: 10.3133/ofr20141008.

Graham, Catherine H., Simon Ferrier, Falk Huettman, Craig Moritz, and A. Townsend Peterson. 2004. "New developments in museum-based informatics and applications in biodiversity analysis." *Trends in Ecology & Evolution* 19 (9):497-503. doi: 10.1016/j.tree.2004.07.006.

Guevara, Lázaro, Beth E. Gerstner, Jamie M. Kass, and Robert P. Anderson. 2018. "Toward ecologically realistic predictions of species distributions: A cross-time example from tropical montane cloud forests." *Global Change Biology* 24 (4):1511-1522. doi: 10.1111/gcb.13992.

Guisan, Antoine, Wilfried Thuiller, and Niklaus E. Zimmermann. 2017. *Habitat suitability and distribution models: with applications in R*. Cambridge: University of Cambridge Press.

Halvorsen, Rune, Sabrina Mazzoni, Anders Bryn, and Vegar Bakkestuen. 2015. "Opportunities for improved distribution modelling practice via a strict maximum likelihood interpretation of MaxEnt." *Ecography* 38 (2):172-183. doi: 10.1111/ecog.00565.

Hamann, A., T. Wang, D.L. Spittlehouse, and T.Q Murdock. 2018. "ClimateWNA", ver. 5.60. University of British Columbia. http://www.climatewna.com/.

Hatfield, David C. . 2000. "TopoTools - A Collection of Topographic Modeling Tools for ArcINFO." 2000 Esri International User Conference, San Diego, CA. http://proceedings.esri.com/library/userconf/proc00/professional/papers/PAP560/p560.htm.

Hernandez, Pilar A., Catherine H. Graham, Lawrence L. Master, and Deborah L. Albert. 2006. "The effect of sample size and species characteristics on performance of different species distribution modeling methods." *Ecography* 29 (5):773-785. doi: 10.1111/j.0906-7590.2006.04700.x.

Hijmans, Robert J. 2018. "raster: Geographic Data Analysis and Modeling", ver. 2.8-4. https://CRAN.R-project.org/package=raster.

Hijmans, Robert J., Susan E. Cameron, Juan L. Parra, Peter G. Jones, and Andy Jarvis. 2005. "Very high resolution interpolated climate surfaces for global land areas." *International Journal of Climatology* 25 (15):1965-1978. doi: 10.1002/joc.1276.

Hijmans, Robert J., Steven Phillips, John Leathwick, and Jane Elith. 2017. "dismo: Species Distribution Modeling", ver. 1.1-4. https://CRAN.R-project.org/package=dismo.

Jarnevich, Catherine S., Thomas J. Stohlgren, Sunil Kumar, Jeffery T. Morisette, and Tracy R. Holcombe. 2015. "Caveats for correlative species distribution modeling." *Ecological Informatics* 29:6-15. doi: 10.1016/j.ecoinf.2015.06.007.

Jiménez-Valverde, Alberto. 2011. "Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling." *Global Ecology and Biogeography* 21 (4):498-507. doi: 10.1111/j.1466-8238.2011.00683.x.

Kass, Jamie M., Bruno Vilela, Matthew E. Aiello-Lammens, Robert Muscarella, Cory Merow, and Robert P. Anderson. 2018. "Wallace: A flexible platform for reproducible modeling of species niches and distributions built for community expansion." *Methods in Ecology and Evolution* 9 (4):1151-1156. doi: 10.1111/2041-210X.12945.

Kohavi, Ron. 1995. "A study of cross-validation and bootstrap for accuracy estimation and model selection." *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)* 14 (2):1137-1145.

Körner, Christian. 2012. *Alpine treelines: functional ecology of the global high elevation tree limits*. Basel: Springer Science & Business Media.

Kramer-Schadt, Stephanie, Jürgen Niedballa, John D. Pilgrim, Boris Schröder, Jana Lindenborn, Vanessa Reinfelder, Milena Stillfried, Ilja Heckmann, Anne K. Scharf, Dave M. Augeri, Susan M. Cheyne, Andrew J. Hearn, Joanna Ross, David W. Macdonald, John Mathai, James Eaton, Andrew J. Marshall, Gono Semiadi, Rustam Rustam, Henry Bernard, Raymond Alfred, Hiromitsu Samejima, J. W. Duckworth, Christine Breitenmoser-Wuersten, Jerrold L. Belant, Heribert Hofer, and Andreas Wilting. 2013. "The importance of correcting for sampling bias in MaxEnt species distribution models." *Diversity and Distributions* 19 (11):1366-1379. doi: 10.1111/ddi.12096.

Leroy, Boris, Robin Delsol, Bernard Hugueny, Christine N. Meynard, Chéïma Barhoumi, Morgane Barbet-Massin, and Céline Bellard. 2018. "Without quality presence–absence data, discrimination metrics such as TSS can be misleading measures of model performance." *Journal of Biogeography* 45 (9):1994-2002. doi: 10.1111/jbi.13402.

Lloyd, Andrea H., and Lisa J. Graumlich. 1997. "Holocene Dynamics of Treeline Forests In The Sierra Nevada." *Ecology* 78 (4):1199-1210. doi: 10.1890/0012-9658(1997)078[1199:HDOTFI]2.0.CO;2.

Lobo, Jorge M., Alberto Jiménez-Valverde, and Raimundo Real. 2007. "AUC: a misleading measure of the performance of predictive distribution models." *Global Ecology and Biogeography* 17 (2):145-151. doi: 10.1111/j.1466-8238.2007.00358.x.

Manzoor, Syed Amir, Geoffrey Griffiths, and Martin Lukac. 2018. "Species distribution model transferability and model grain size – finer may not always be better." *Scientific Reports* 8 (1):7168. doi: 10.1038/s41598-018-25437-1.

Mazzoni, Sabrina, Rune Halvorsen, and Vegar Bakkestuen. 2015. "MIAT: Modular R-wrappers for flexible implementation of MaxEnt distribution modelling." *Ecological Informatics* 30:215-221. doi: 10.1016/j.ecoinf.2015.07.001.

McInerny, Greg J., and Rampal S. Etienne. 2013. "'Niche' or 'distribution' modelling? A response to Warren." *Trends in Ecology & Evolution* 28 (4):191-192. doi: 10.1016/j.tree.2013.01.007.

McPherson, Jana M., Walter Jetz, and David J. Rogers. 2004. "The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact?" *Journal of Applied Ecology* 41 (5):811-823. doi: 10.1111/j.0021-8901.2004.00943.x.

Merow, Cory, Mathew J. Smith, Thomas C. Edwards, Antoine Guisan, Sean M. McMahon, Signe Normand, Wilfried Thuiller, Rafael O. Wüest, Niklaus E. Zimmermann, and Jane Elith. 2014. "What do we gain from simplicity versus complexity in species distribution models?" *Ecography* 37 (12):1267-1281. doi: 10.1111/ecog.00845.

Merow, Cory, Matthew J. Smith, and John A. 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. doi: 10.1111/j.1600-0587.2013.07872.x.

Microsoft and R Core Team. 2017. "Microsoft R Open", ver. 3.5.1. Microsoft, Redmond, Washington. https://mran.microsoft.com/.

Morales, Narkis S., Ignacio C. Fernández, and Victoria Baca-González. 2017. "MaxEnt's parameter configuration and small samples: are we paying attention to recommendations? A systematic review." *PeerJ* 5:e3093. doi: 10.7717/peerj.3093.

Muscarella, Robert, Peter J. Galante, Mariano Soley-Guardia, Robert A. Boria, Jamie M. Kass, María Uriarte, and Robert P. Anderson. 2014. "ENMeval: An R package for conducting spatially independent evaluations and estimating optimal model complexity for Maxent ecological niche models." *Methods in Ecology and Evolution* 5 (11):1198-1205. doi: 10.1111/2041-210X.12261.

Olaya, V. 2009. "Chapter 6 Basic Land-Surface Parameters." In *Developments in Soil Science*, edited by Tomislav Hengl and Hannes I. Reuter, 141-169. Amsterdam: Elsevier.

Oliveira, Brunno F. . 2016. "Building and comparing the performance of Ecological Niche Models (ENMs)." Last Modified December 4, 2016, Accessed December 17, 2018.

https://oliveirabrunno.wordpress.com/2016/12/04/compare-the-performance-of-ecological-niche-models-enms/.

Osborn, Kenneth, John List, Dean Gesch, John Crowe, Gary Merrill, Eric Constance, James Mauck, Christine Lund, Vincent Caruso, and John Kosovich. 2001. "Chapter 4: National digital elevation program (NDEP)." In *Digital elevation model technologies and applications The DEM users manual*, edited by David F. Maune, 83-120. Bethesda, Maryland: American Society for Phtogrammetry and Remote Sensing.

Peterson, A. Townsend, Monica Papeş, and Jorge Soberón. 2008. "Rethinking receiver operating characteristic analysis applications in ecological niche modeling." *Ecological Modelling* 213 (1):63-72. doi: 10.1016/j.ecolmodel.2007.11.008.

Peterson, A. Townsend, and Jorge Soberón. 2012. "Integrating fundamental concepts of ecology, biogeography, and sampling into effective ecological niche modeling and species distribution modeling." *Plant Biosystems* 146 (4):789-796. doi: 10.1080/11263504.2012.740083.

Peterson, A. Townsend, Jorge Soberón, Richard G Pearson, Robert P Anderson, Enrique Martínez-Meyer, Miguel Nakamura, and Miguel B Araújo. 2011. *Ecological niches and geographic distributions (MPB-49)*. Princeton, NJ: Princeton University Press.

Phillips, Steven J. 2008. "Transferability, sample selection bias and background data in presence-only modelling: a response to Peterson et al. (2007)." *Ecography* 31 (2):272-278. doi: 10.1111/j.0906-7590.2008.5378.x.

Phillips, Steven J. . 2017. "A Brief Tutorial on Maxent." Accessed November 3, 2018. http://biodiversityinformatics.amnh.org/open_source/maxent/.

Phillips, Steven J., Robert P. Anderson, Miroslav Dudík, Robert E. Schapire, and Mary E. Blair. 2017. "Opening the black box: an open-source release of Maxent." *Ecography* 40 (7):887-893. doi: 10.1111/ecog.03049.

Phillips, Steven J., Robert P. Anderson, and Robert E. Schapire. 2006. "Maximum entropy modeling of species geographic distributions." *Ecological Modelling* 190 (3):231-259. doi: 10.1016/j.ecolmodel.2005.03.026.

Phillips, Steven J., Miroslav Dudík, and Robert E. Schapire. 2017. "Maxent software for modeling species niches and distributions", ver. 3.4.1. http://biodiversityinformatics.amnh.org/open_source/maxent/.

PRISM Climate Group. 2018. "PRISM Datasets". Oregon State University, Corvallis, Oregon. http://prism.oregonstate.edu/normals/.

R Core Team. 2018. "R: A language and environment for statistical computing.". R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Raäisaänen, Jouni. 2007. "How reliable are climate models?" *Tellus A: Dynamic Meteorology and Oceanography* 59 (1):2-29. doi: 10.1111/j.1600-0870.2006.00211.x.

Radcliffe, D. E., and R. Mukundan. 2016. "PRISM vs. CFSR Precipitation Data Effects on Calibration and Validation of SWAT Models." *Journal of the American Water Resources Association* 53 (1):89-100. doi: 10.1111/1752-1688.12484.

Radosavljevic, Aleksandar, and Robert P. Anderson. 2014. "Making better Maxent models of species distributions: complexity, overfitting and evaluation." *Journal of biogeography* 41 (4):629-643. doi: 10.1111/jbi.12227.

RStudio Team. 2016. "RStudio: Integrated Development Environment for R", ver. 1.1.463. RStudio, Inc., Boston, MA. http://www.rstudio.com/.

Salzberg, Steven L. 1997. "On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach." *Data Mining and Knowledge Discovery* 1 (3):317-328. doi: 10.1023/A:1009752403260.

Salzer, Matthew W., Evan R. Larson, Andrew G. Bunn, and Malcolm K. Hughes. 2014. "Changing climate response in near-treeline bristlecone pine with elevation and aspect." *Environmental Research Letters* 9 (11):114007.

Scuderi, Louis A. 1987. "Late-Holocene upper timberline variation in the southern Sierra Nevada." *Nature* 325:242. doi: 10.1038/325242a0.

Sequeira, Ana M. M., Phil J. Bouchet, Katherine L. Yates, Kerrie Mengersen, and M. Julian Caley. 2018. "Transferring biodiversity models for conservation: Opportunities and challenges." *Methods in Ecology and Evolution* 9 (5):1250-1264. doi: 10.1111/2041-210X.12998.

Smith, Adam B. , Danielle Christianson, Camilo Sanín, and Danielle Svehla. 2017. "A Hands-on Short Course in Species Distribution Modeling Using R: From Start to Finish." The Global Change Conservation Lab at the Missouri Botanical Garden, Last Modified June 4, 2018, Accessed November 5, 2018. http://www.earthskysea.org/workshops-classes/.

Sobern, Jorge, and Townsend Peterson. 2004. "Biodiversity informatics: managing and applying primary biodiversity data." *Philosophical Transactions of the Royal Society B: Biological Sciences* 359 (1444):689-698. doi: 10.1098/rstb.2003.1439.

Spittlehouse, DL, Trevor Murdock, Gerd Berger, and Tongli Wang. 2012. "Final Report, FFESC project 014: High Resolution Spatial Climate Data for Climate Change Research in BC." British Columbia Ministry Forest, Lands, and Natural Resource Operations. https://www2.gov.bc.ca/assets/gov/environment/natural-resource-stewardship/nrs-climate-change/applied-science/spittlehousefinalreportrevised.pdf.

Strachan, Scotty, and Christopher Daly. 2017. "Testing the daily PRISM air temperature model on semiarid mountain slopes." *Journal of Geophysical Research: Atmospheres* 122 (11):5697-5715. doi: 10.1002/2016JD025920.

Syfert, Mindy M., Matthew J. Smith, and David A. Coomes. 2013. "The Effects of Sampling Bias and Model Complexity on the Predictive Performance of MaxEnt Species Distribution Models." *PLOS ONE* 8 (2):e55158. doi: 10.1371/journal.pone.0055158.

Thuiller, Wilfried , Damien Georges, Robin Engler, and Frank Breiner. 2016. "biomod2: Ensemble Platform for Species Distribution Modeling", ver. 3.3-7. https://cran.r-project.org/web/packages/biomod2/biomod2.pdf.

Townsend Peterson, A., Monica Papeş, and Muir Eaton. 2007. "Transferability and model evaluation in ecological niche modeling: a comparison of GARP and Maxent." *Ecography* 30 (4):550-560. doi: 10.1111/j.0906-7590.2007.05102.x.

University of California Agriculture and Natural Resources. 2016. "About degree-days." Accessed January 27, 2019. http://ipm.ucanr.edu/WEATHER/ddconcepts.html.

Urbanek, Simon. 2018. "rJava: Low-Level R to Java Interface", ver. 0.9-10. https://CRAN.R-project.org/package=rJava.

USGS. 2000. "US GeoData Digital Elevation Models - Fact Sheet 040-00 (April 2000)." Accessed November 29, 2018. https://egsc.usgs.gov/isb//pubs/factsheets/fs04000.html.

USGS. 2015. "Elevation Products." Accessed November 14, 2018. https://eros.usgs.gov/products-data-available/elevation-products.

USGS. 2017. "3DEP products and services: The National Map, 3D Elevation Program." Accessed April 2, 2018. https://nationalmap.gov/3DEP/3dep_prodserv.html.

USGS. 2018. "Seamless DEMs - Production Methods." Accessed November 5, 2018. https://www.usgs.gov/media/images/seamless-dems-production-method.

Vollering, Julien. 2017. "MIAmaxent ", ver. 0.4.0. https://cran.r-project.org/package=MIAmaxent.

Wang, Tongli, Andreas Hamann, Dave Spittlehouse, and Carlos Carroll. 2016. "Locally Downscaled and Spatially Customizable Climate Data for Historical and Future Periods for North America." *PLOS ONE* 11 (6):e0156720. doi: 10.1371/journal.pone.0156720.

Wang, Tongli, Andreas Hamann, David L. Spittlehouse, and Trevor Q. Murdock. 2011. "ClimateWNA—High-Resolution Spatial Climate Data for Western North America." *Journal of Applied Meteorology and Climatology* 51 (1):16-29. doi: 10.1175/JAMC-D-11-043.1.

Warren, Dan L. 2012. "In defense of 'niche modeling'." *Trends in Ecology & Evolution* 27 (9):497-500. doi: 10.1016/j.tree.2012.03.010.

Warren, Dan L. 2013. "'Niche modeling': that uncomfortable sensation means it's working. A reply to McInerny and Etienne." *Trends in Ecology & Evolution* 28 (4):193-194. doi: 10.1016/j.tree.2013.02.003.

Warren, Dan L., and Stephanie N. Seifert. 2011. "Ecological niche modeling in Maxent: the importance of model complexity and the performance of model selection criteria." *Ecological Applications* 21 (2):335-342. doi: 10.1890/10-1171.1.

Wickham, Hadley. 2016. "ggplot2: Elegant Graphics for Data Analysis", ver. 3.1.0. Springer-Verlag, New York. https://cran.r-project.org/web/packages/ggplot2/index.html.

Wilson, John P. 2018. *Environmental applications of digital terrain modelling*. Hoboken, NJ: Wiley-Blackwell.

Wisz, M. S., R. J. Hijmans, J. Li, A. T. Peterson, C. H. Graham, and A. Guisan. 2008. "Effects of sample size on the performance of species distribution models." *Diversity and Distributions* 14 (5):763-773. doi: 10.1111/j.1472-4642.2008.00482.x.

Yackulic, Charles B, Richard Chandler, Elise F Zipkin, J Andrew Royle, James D Nichols, Evan H Campbell Grant, and Sophie Veran. 2013. "Presence‑only modelling using MAXENT: when can we trust the inferences?" *Methods in Ecology and Evolution* 4 (3):236-243. doi: 10.1111/2041-210x.12004.

Yesson, Chris, Peter W Brewer, Tim Sutton, Neil Caithness, Jaspreet S Pahwa, Mikhaila Burgess, W Alec Gray, Richard J White, Andrew C Jones, and Frank A Bisby. 2007. "How global is the global biodiversity information facility?" *PLoS One* 2 (11):e1124. doi: 10.1371/journal.pone.0001124.

Appendix A Software Used in The Study

Name	Version	Type	Source
ArcGIS Pro	2.2.4	MS Windows	(Esri 2018)
ClimateWNA	5.60	MS Windows	(Hamann et al. 2018)
dismo	1.1-4	R package	(Hijmans et al. 2017)
ENMeval	0.3.0	R package	(Muscarella et al. 2014)
gglot2	3.1.0	R package	(Wickham 2016)
ggthemes	4.0.1	R package	(Arnold 2018)
Maxent	3.4.1	Java jar file	(Phillips, Dudík, and Schapire 2017)
Microsoft R Open	3.5.1	Programming language	(Microsoft and R Core Team 2017)
R	3.4.4	Programming language	(R Core Team 2018)
raster	2.8-4	R package	(Hijmans 2018)
rgdal	1.3-6	R package	(Bivand, Keitt, and Rowlingson 2018)
rJava	0.9-10	R package	(Urbanek 2018)
RStudio	1.1.463	MS Windows	(RStudio Team 2016)

Additional software snippets and instructional materials:

- "Building and comparing the performance of Ecological Niche Models (ENMs)" (Oliveira 2016)
- "A Hands-on Short Course in Species Distribution Modeling Using R: From Start to Finish" (Smith et al. 2017)

Appendix B Data Preparation Procedures

The detailed steps used in ArcGIS Pro and ClimateWNA (cWNA) for the preliminary data processing of the Maxent inputs can be found at the thesis' GitHub site: https://github.com/CassKal/GIST_Thesis. The same processing was used for both the 30-meter and 800-meter resolutions and for both the main California study area as well as the Utah prediction area. All data in GCS North American 1983 unless otherwise noted.

GitHub File: Detailed Data Preparation Procedures.pdf

Appendix C Partitioning Scheme Patterns

Colors indicate members of the partition in the scheme.
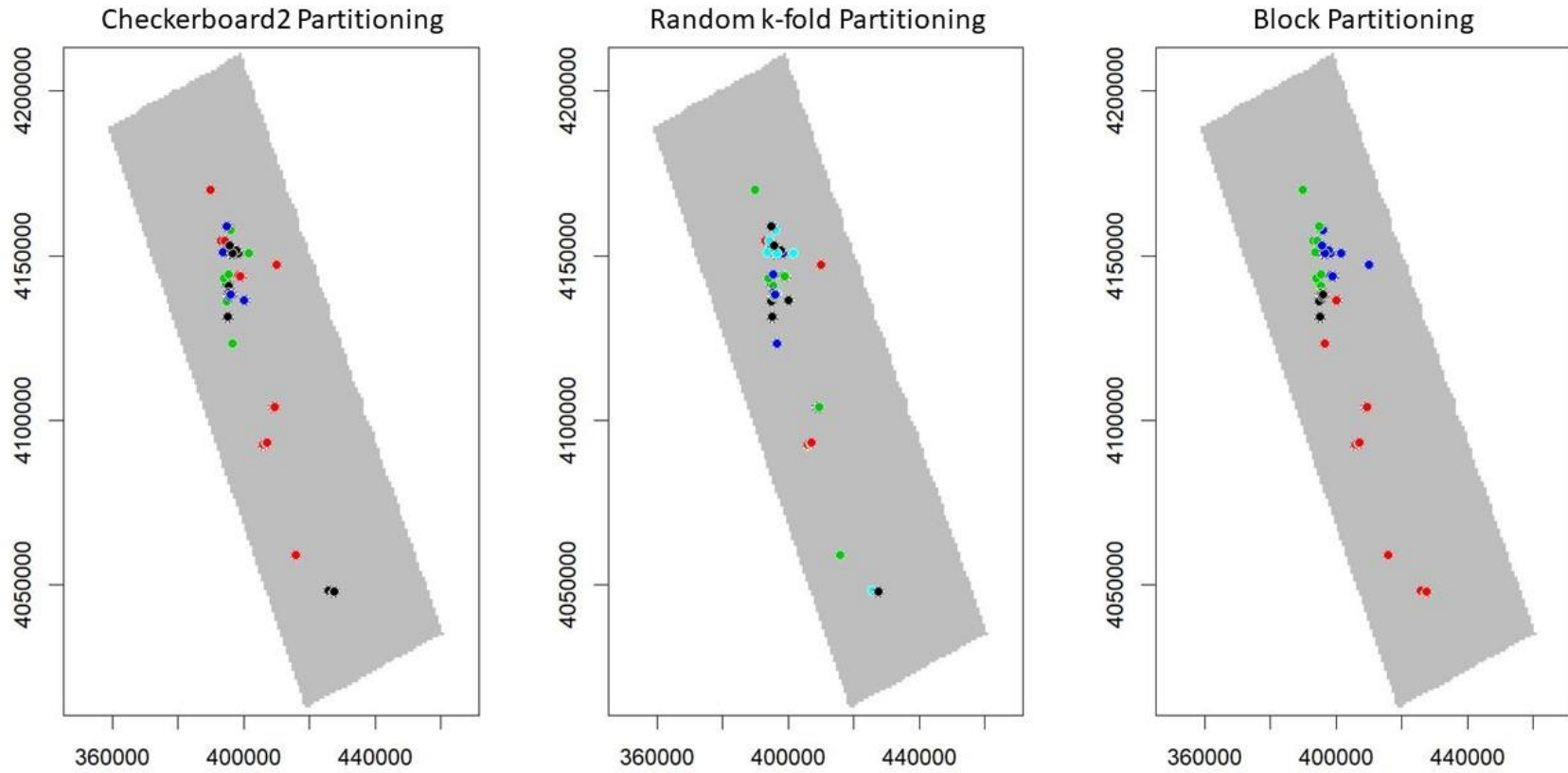
Figure 31 - Partitioning scheme patterns for the 800-meter tests
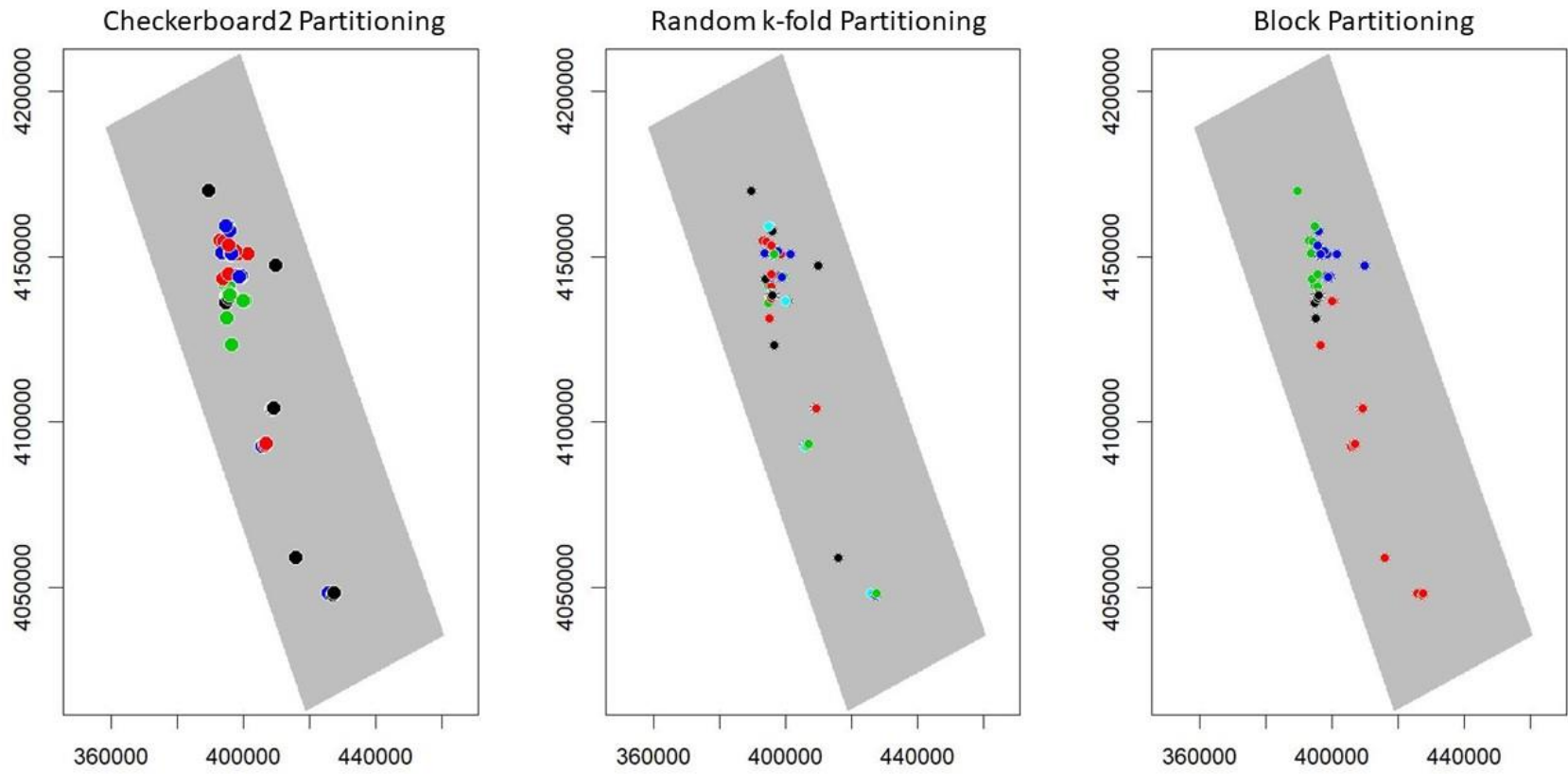
Figure 32 - Partitioning scheme patterns for the 30-meter tests

Appendix D Model Tuning R Script

The ENMeval package (Muscarella et al. 2014) and other R code was used to facilitate the evaluation and determination of the tuned model parameter selection. ENMeval uses either the MaxNet R package or the Maxent java jar file to create the Maxent models. The latter was used in this study. Code within this module was copied from or modeled after code provided by Muscarella et al. (2014) and from the ENMeval R Vignette (https://cran.r-project.org/web/packages/ENMeval/vignettes/ENMeval-vignette.html). The R script for this module can be found at the thesis' GitHub site: https://github.com/CassKal/GIST_Thesis.

GitHub File: Thesis_Kalinski_R_Script_ModelTuning_v06.R

Appendix E Tuning Test Results

Graphs of the tuning test results follow in Figure 33 to Figure 40. Each figure is a composite of graphs for a specific resolution/partition combination. The five graphs in each composite show the AICc, AUC, AUC.diff, orMTP, and or10pct metrics for the test. A total of 200 models were tested.

Microsoft Excel files containing the full results of the evaluation model testing can be found at the thesis' GitHub site: https://github.com/CassKal/GIST_Thesis. The test results list the AICc, AUC, diff.AUC, orMTP, and or10pct results for each of the 200 model configurations along with the evaluation notes for each.

GitHub Files:

ResultsOf30mTuningRuns_v2.xlsx

ResultsOf800mTuningRuns.xlsx
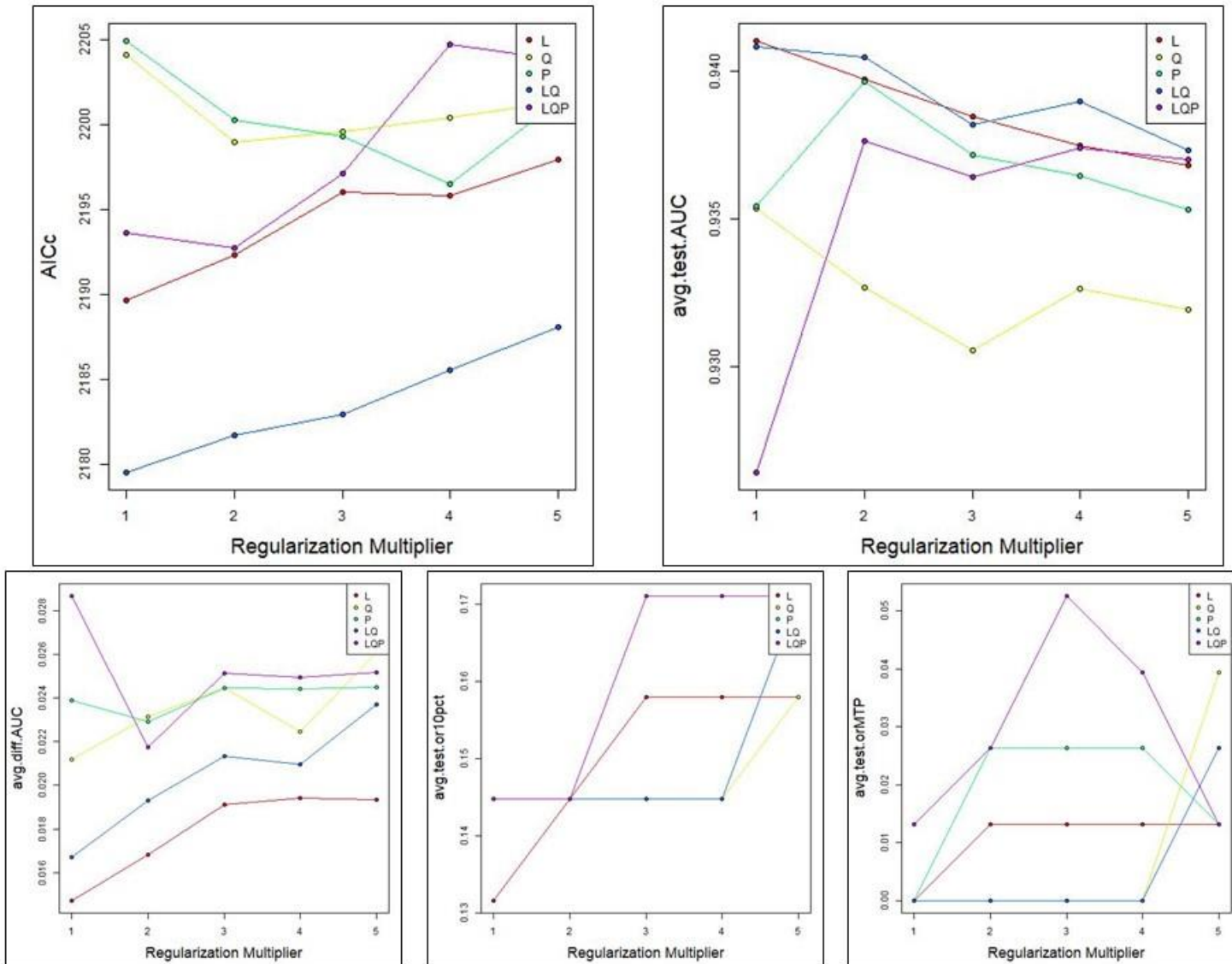
Figure 33 - Tuning test results: 30-meter block partition
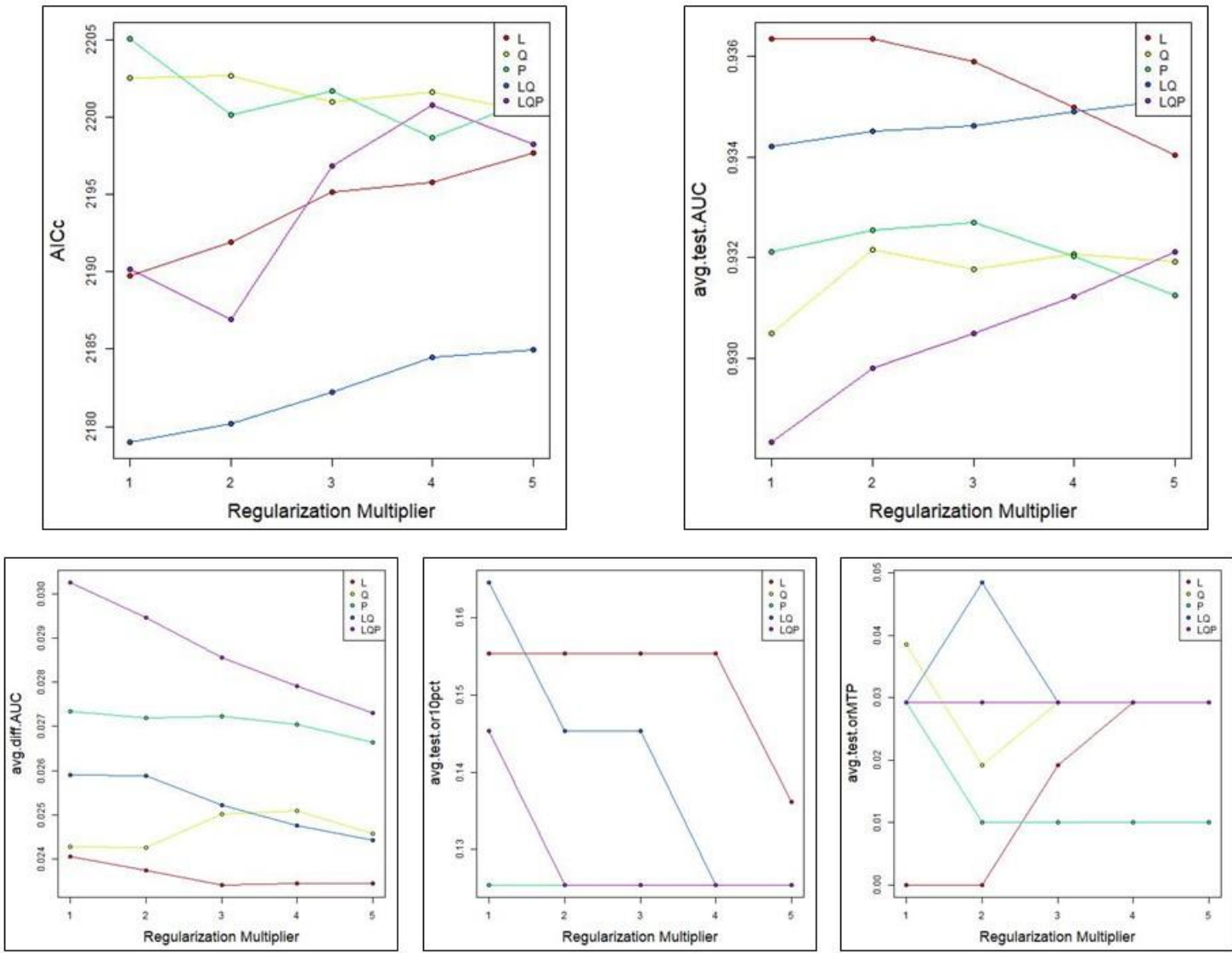
Figure 34 - Tuning test results: 30-meter checkerboard2 partition
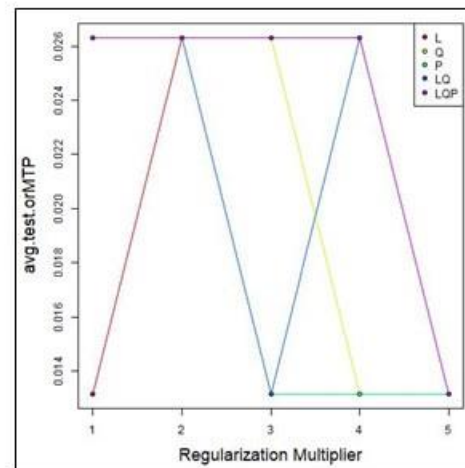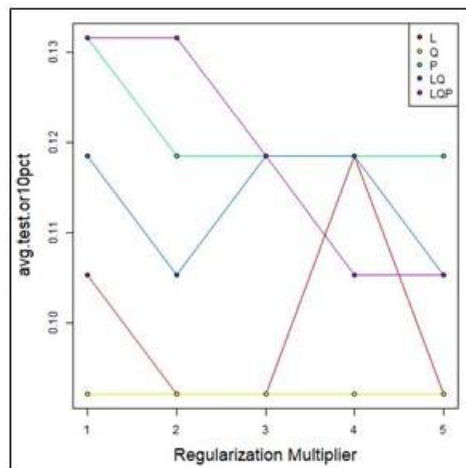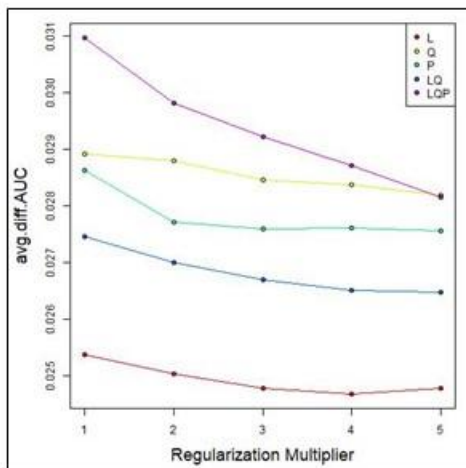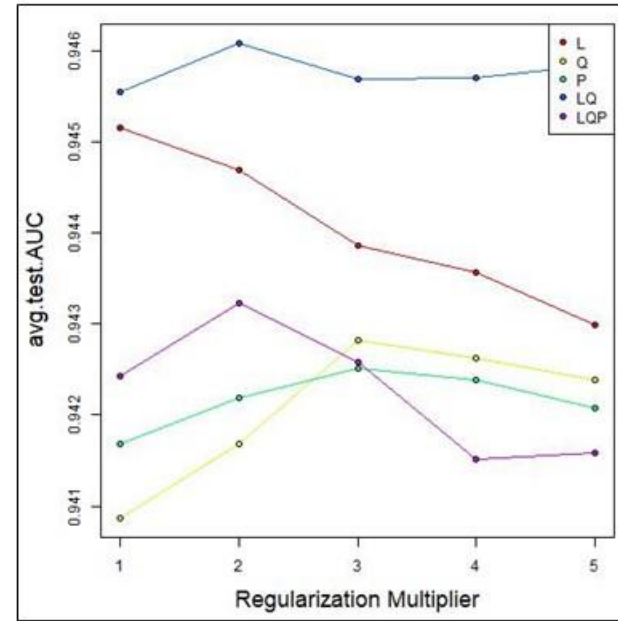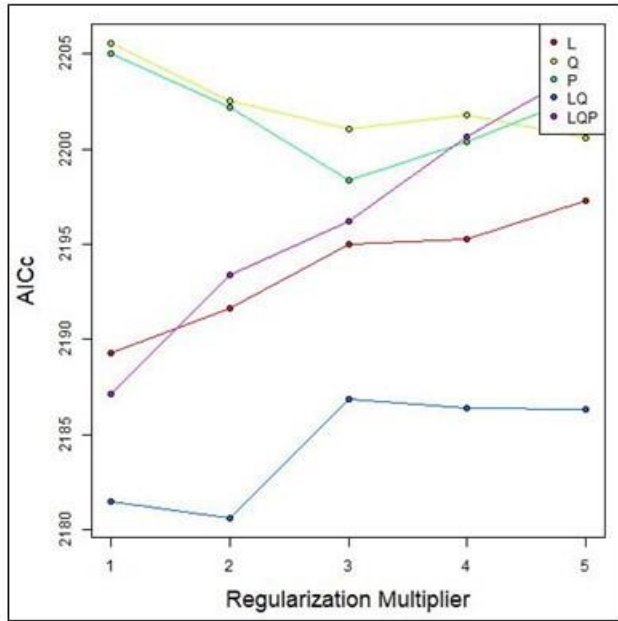
Figure 35 - Tuning test results: 30-meter jackknife partition
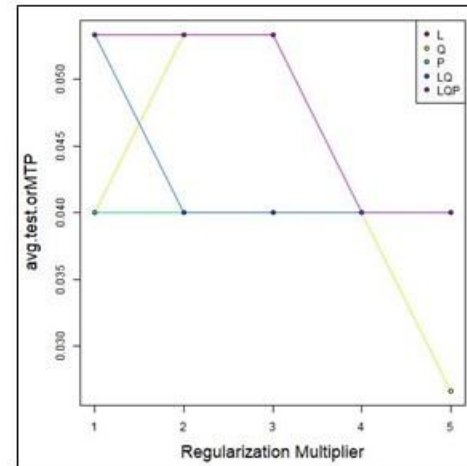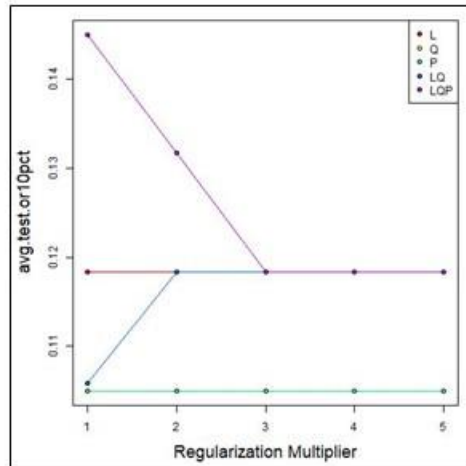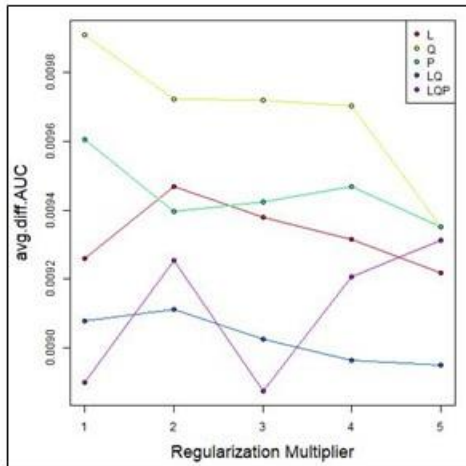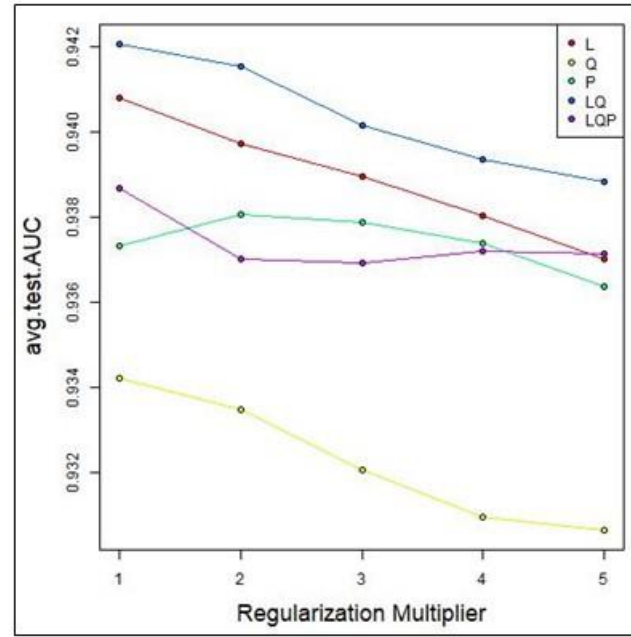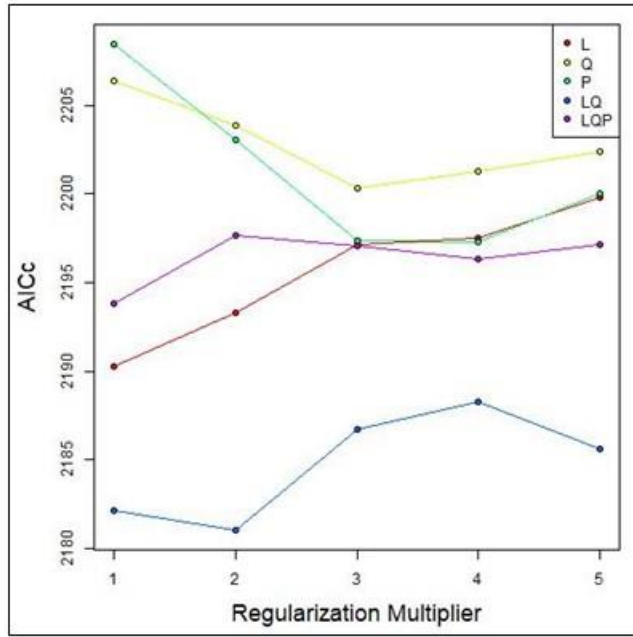
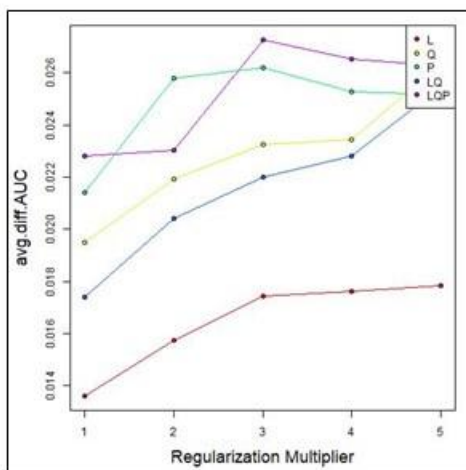Figure 36 - Tuning test results: 30-meter random k-fold partition
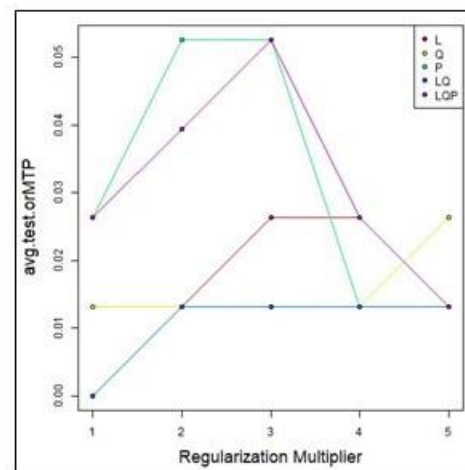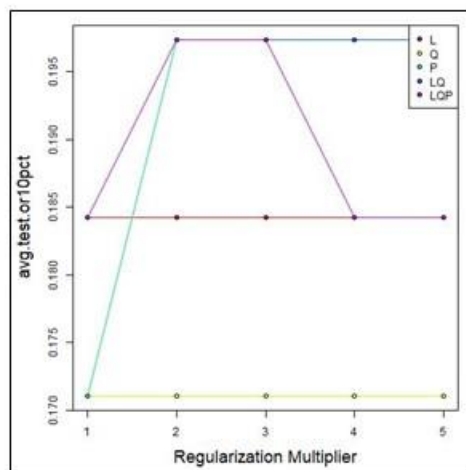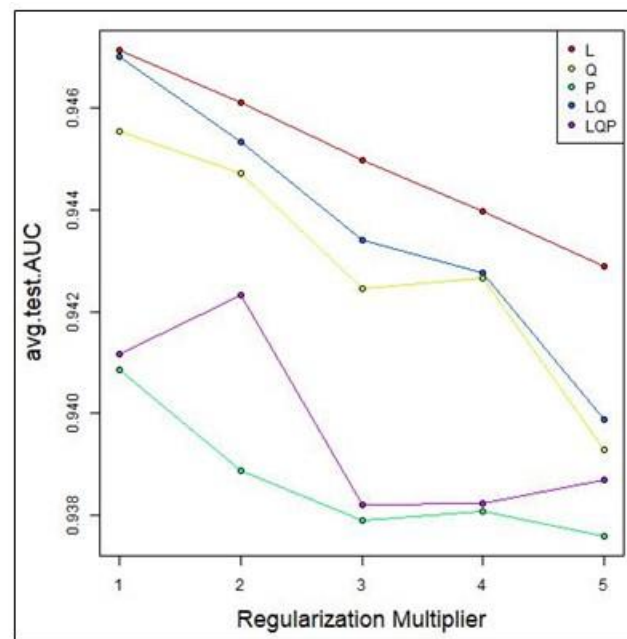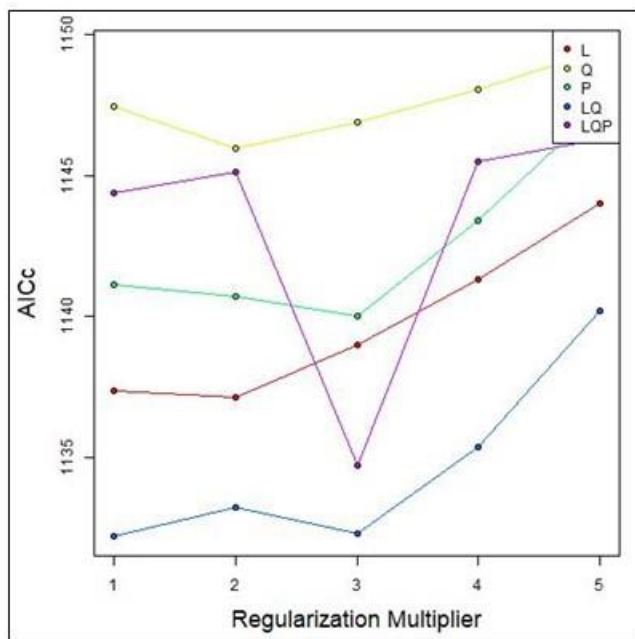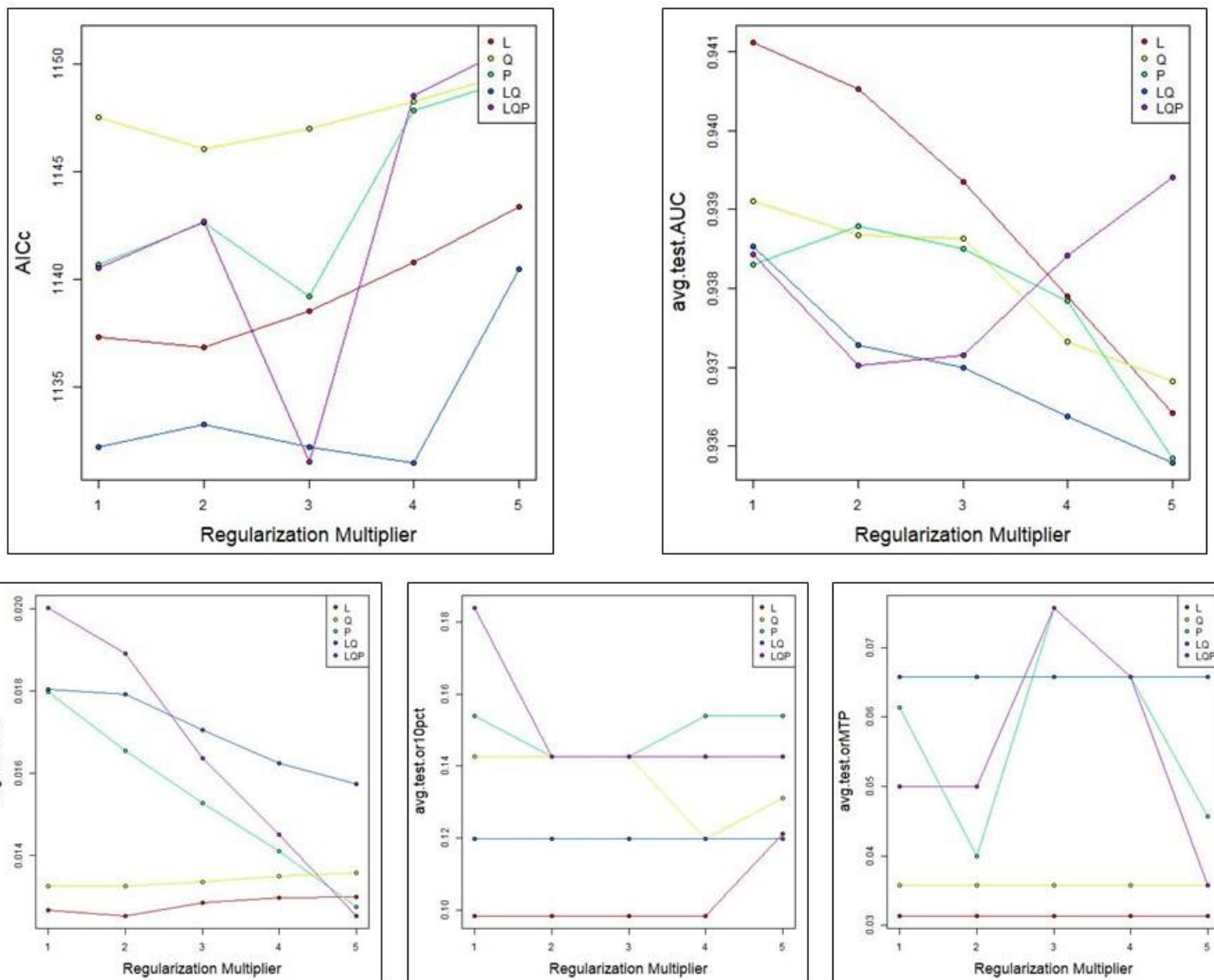
Figure 37 - Tuning test results: 800-meter block partition

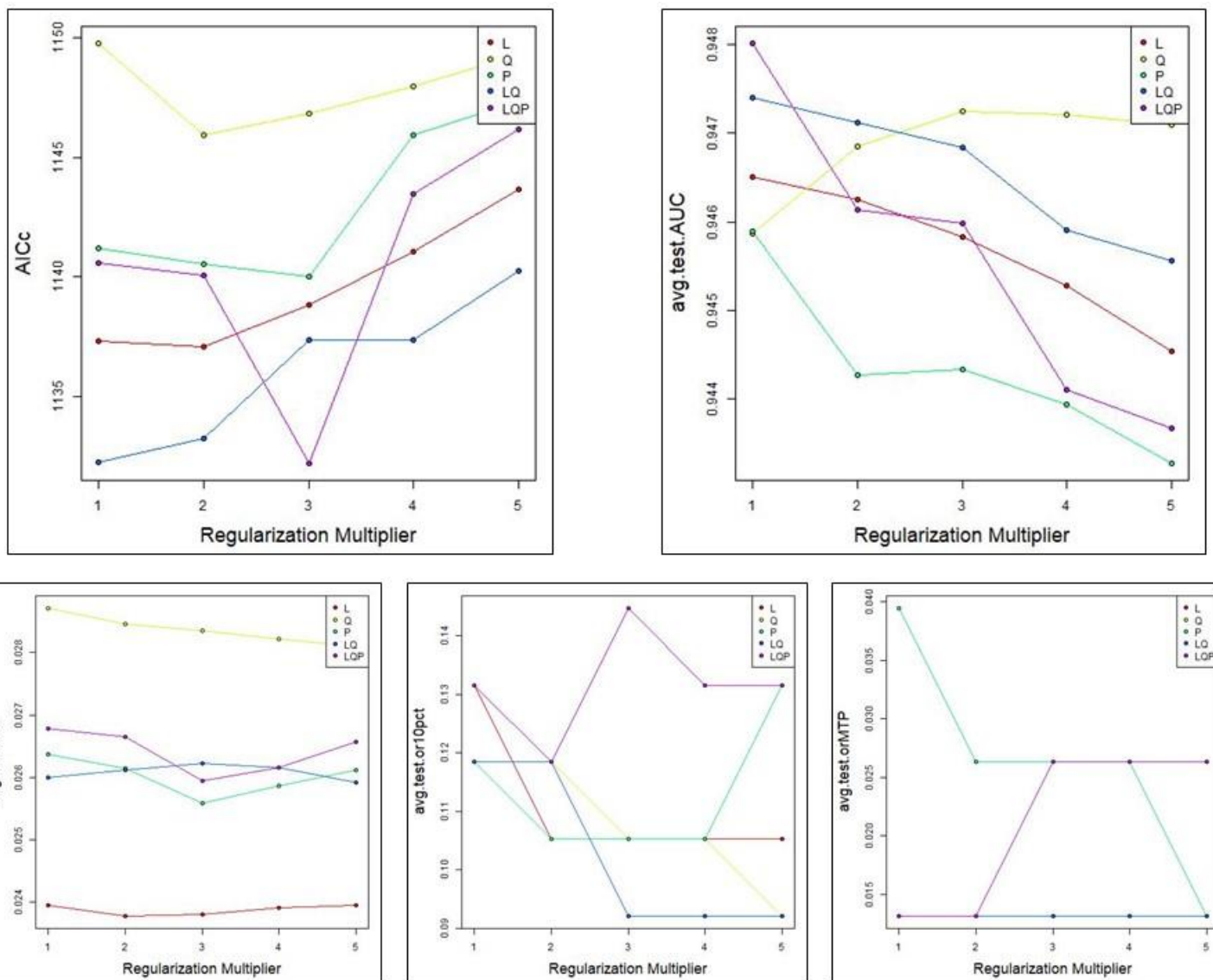Figure 38 - Tuning test results: 800-meter checkerboard2 partition

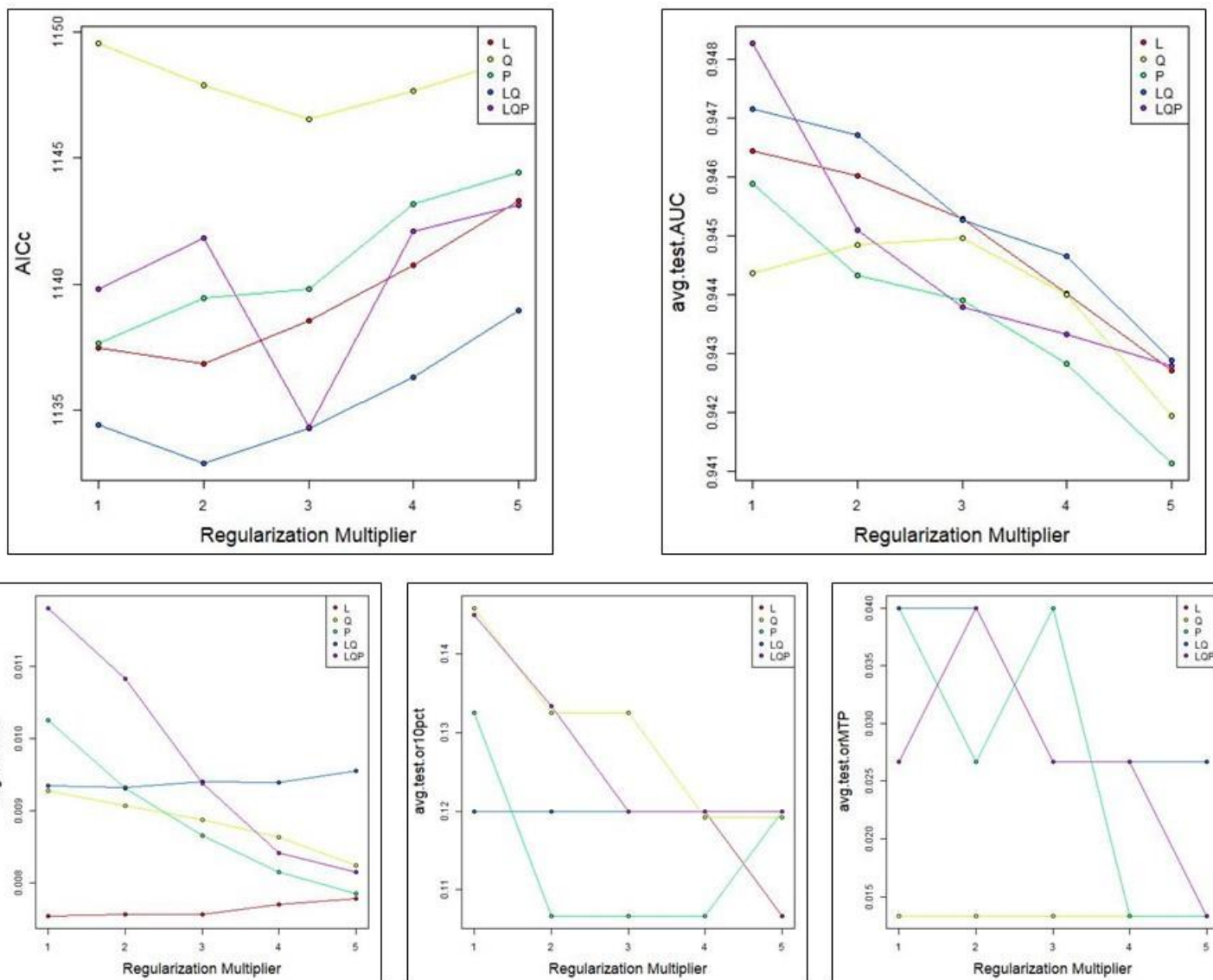Figure 39 - Tuning test results: 800-meter jackknife partition

Figure 40 - Tuning test results: 800-meter random k-fold partition

Appendix F Model Comparison R Script

The code in this R module builds the tuned and default Maxent models, primarily using the dismo package (Hijmans et al. 2017). Metrics are generated for the models and saved to storage for offline evaluation. Twenty-five iterations of the models are created, the prediction rasters averaged, and saved to storage. Code within this module was copied from or modeled after code provided by Muscarella et al. (2014) and from the ENMeval R Vignette (https://cran.r-project.org/web/packages/ENMeval/vignettes/ENMeval-vignette.html). Code from the Oliveira (2016) blog was used throughout this module. The R script for this module can be found at the thesis' GitHub site: https://github.com/CassKal/GIST_Thesis.

GitHub File: Thesis_Kalinski_R_Script_BuildModels_v09.R

Appendix G Evaluation Metrics

A Microsoft Excel file containing detailed metrics for the four study models can be found at the thesis' GitHub site: https://github.com/CassKal/GIST_Thesis. The file contains the data for each of the 25 iterations for each of the four models as well as a summary sheet comparing the data across models.

GitHub File: Evaluation Metrics v01.xlsx