Boosted regression tree modeling of selected plant species distributions in California at 22ka, 6ka and the Present

Scott Farley and Roger Byrne

Draft 1/27/2015

## Abstract

We report here on the results of boosted regression tree modeling of plant species distributions in California at 22kya, 6kya, and the 1981-2010 modern baseline. For six indicator species, we generated estimates of species presence at the three time periods. These time periods were chosen because high-resolution paleo-climate data for these periods has recently been made available. Using these data, we assessed a variety of metrics, including shifts in elevation and latitude of distribution, to determine large-scale migrational history. We also compared the results to fossil pollen and other paleoproxy records published in the literature to assess model accuracy. The comparisons show general agreement in response to climatic trends; however, there are several discrepancies. We hypothesize that these mismatches between fossil evidence and modeled species distributions are due to uncertainties in the precipitation estimates of the underlying climate data for the two historical periods.

## 

# Introduction

## Species Distribution Modeling

As anthropogenic warming threatens to disrupt vegetation and wildlife regimes, species distribution modeling is becoming a popular method of predicting future species ranges given climatic changes. So-called climate envelope models estimate the relationship between species records at sites and the environmental and/or spatial characteristics of those sites (Elith et al. 2011). Using these response profiles to predict species ranges under future climate scenarios is speculative, insofar as they can not be currently tested. One way to test modeling techniques is to use them to model species distributions for time periods in the past for which there is fossil pollen and other paleo proxy records against which to validate model output. We use this method here. More specifically, we model the distributions of six species in the state of California at 22ka (last glacial maximum) and 6ka (Mid-Holocene) along with the 0ka modern baseline (1981AD - 2010AD averaged climate). Numerous published studies present contain fossil pollen and packrat (*Neotoma spp.)*  midden reconstructions of vegetation in California extending back at least to the last glacial maximum (LGM). These studies suggest significant large-scale climate change in California between the LGM, Mid-Holocene (M-H), and modern period. The LGM is generally understood to have been wetter and colder, while many studies suggest warmer and drier conditions during the M-H. Given these changes in climate, we expect that many species shifted southward and downslope during the LGM when compared to the modern baseline, while the inverse is expected for the M-H.

## 1.2. Regression Trees and Boosting

A variety of species modeling algorithms are available to predict or reconstruct changes in plant and animal species distributions. The general method is to assess present climate at observed locations and then project this climatic profile onto novel climate grids. In this paper, we chose to use the Boosted Regression Tree (BRT) method for its predictive power and its flexibility. A regression tree is a statistical modeling structure designed to predict a real number outcome of some variable Y, given a set of predictor variables. In concept, it is similar to a traditional regression, except that it involves binary rules, rather than continuous functions. Regression trees are thus typically superior at fitting local variations in the data. In our case, the outcome variable is the probability that the species in question will grow in a given location, and the predictors are the climatic variables given as input.

The generation of regression trees is combined in this method with boosting, a statistical technique used to improve accuracy by combining many weak-learning trees into a highly predictive ensemble rather than building a single large tree. Single tree models tend to be less accurate than other potential modeling algorithms, such as generalized additive models or generalized linear models (Elith et al. 2008). However, with the addition of boosting, predictive power is greatly increased by adding trees that minimize the deviance in the model (Schonlau 2005). The first tree in the model attempts to maximally reduce the deviance of the model within a given level of complexity. Each additional tree attempts to explain variation in the response that is not so far explained by the model. The process continues stagewise, so that existing trees are left unchanged, until the addition of an another tree would act to increase the deviance, thus decreasing the model’s predictive performance. The final model contains a linear combination of hundreds of small trees that can be seen as akin to a linear regression model with each tree as an individual term.

Elith et al (2008) discusses several important advantages of this type of model. First, the model generation process is stochastic, using only a random subset of the input observations to generate each additional tree. In our case, each tree was developed with a ¾ of the input data. This inherent randomness ensures that outlying or misclassified observations do not skew the model generation. Therefore, unlike many other predictive models, such as MaxEnt, reclassification and outlier elimination is generally not necessary. Secondly, because the process is stagewise, the bulk of the model attempts to focus on the variations in the responses that are the most difficult to explain. Initial trees aim to reduce the deviance of the model as much as possible, and subsequent trees are added solely to reduce the residuals. Finally, because of the hierarchical nature of the tree structure, the model is capable of resolving interactions between the predictor variables, which is useful when considering many variables, as we do here.

## Indicator Species

For this paper, we chose six species that are considered indicators of larger change in California for the time periods in consideration: Bis Sagebrush (*Artemisia tridentata),* Ponderosa Pine *(Pinus ponderosa),* Gray Pine *(P. sabiniana),* Greasewood *(Sarcobatus vermiculatus),* Western Hemlock *(Tsuga heterophylla),* and Joshua Tree (*Yucca brevifolia). A. tridentata, S. vermiculatus,*  and *Y. brevifolia* are key species in the cold-desert ecosystem commonly found today in the Great Basin of the intermountain West. Bartlein et al. included *A. tridentata* as a taxon that is broadly distributed in the western United States (Bartlein et al. 1998). *P. ponderosa*  and *P. sabiniana*  are both common trees native to California, typically found at hot, dry sites with minimal summer rain. Finally, *T. heterophylla* is a large evergreen tree, included here to be indicative of relatively wet climates. According to a study Gavin and Hu (2006), the modern *T. heterophylla* distribution is most strongly controlled by actual evapotranspiration rates. Today *T. heterophylla* is distributed in the northern reaches of the California Coast Ranges, where it is supported by frequent fog and precipitation throughout the growing season.

# 2. Methods

## 2.1. Model Inputs

### 2.1.1. High Resolution Climate Grids

The climate at each of the three time periods in this study is represented by an ensemble of 19 bioclimatic variables. These variables are listed in Table 1, and encompass a wide range of potentially limiting environmental factors. These environmental conditions were obtained in gridded format for each of the time periods. These grids are composed of grid cells sized 800 meters by 800 meters. Modern data was obtained from the PRISM Project [Daly and Bryant 2013], representing the 1981-2010 thirty-year climatological average. Bias-corrected downscaled Community Climate System Model 4 (CCSM4) output was used to represent the LGM and the M-H (Mondal. et al. 2015).

Modeling using these 800m by 800m climate grids allows us to improve the reslolution of many prior species distribution modeling studies (Thompson et al. 1998, Bartlein et al. 1998). Prior studies typically used lower resolution climate grids and covered larger spatial extents in their analysis. By using higher resolution grids, we are able to resolve nuances in species distributions that may not be possible using lower resolution data. Furthermore, we are able to limit our study region to the state scale, rather than the national or continental scale, which allows for more detailed comparison with paleoproxy data. However, it is important to note that even the best downscaled climate data is derived from coarse GCM output. Therefore, climatic details, such as the effect of the coastal margin, may not be properly resolved.

Using standard GIS techniques, the climate grids were loaded into the R programming environment, and manipulated for spatial analysis [R Core Team 2014]. The grids were kept in unprojected latitude and longitude format, with a WGS1984 datum, to facilitate point computations necessary in subsequent steps. The grids were clipped from their original continental extent to the spatial extent of California (California Department of Forestry and Fire Protection, 2010).

### 2.1.2. Presence Points

Presence points are locations where a particular species has been recorded growing in the wild or where a live specimen was collected. These points can be used to determine the environmental factors that make up a species’ realized niche. Databases of presence locations are often an integral part of herbarium and museum collections, and typically include metadata on the specimen and its collection location.

For this project, presence points were collected from the Berkeley Ecoinformatics Engine (Ecoengine) [Berkeley Initiative In Global Change Biology, 2015]. These points were loaded directly into R using the Ecoengine’s R Package (Ram 2014). The number of presence points for the species included in this paper ranged from 75 to 361. Each point’s geographic location was then used as the basis for a point extraction function, using the R package dismo’s extract function (Hijmans et al 2014), which yielded the value of each of the 19 bioclimatic variables at that point.

### 2.1.3. Absence Points

Absence points are generated using a pseudorandom technique that aimed to create a number of points that represent climatic conditions outside of the species’ present distribution. All of these points are at least 10 kilometers from any known occurrence, and thus, it is a reasonable assumption that conditions are not suitable for its growth.

We generated absence points as follows. Presence points were first buffered by 10 kilometers. They were then buffered again by 1000 kilometers, yielding a ring of potential sample points. From this ring, 10,000 random points were selected. As with the presence points, the modern bioclimatic variable assemblage was extracted at each of the points. Due to the way the points are generated, many of the potential absences are outside of the boundary of the state. These points are disregarded, and the points within the spatial extent of the state are used to describe the outside-of-niche environmental conditions. Because of the stochastic nature of the sampling, every run of the model yields different numbers of absence points, even for the same species. One important metric that we assessed was the ratio of presence to absence points. Because we wanted to accurately model the full range of conditions not suitable for species presence, we aimed for a 1:10 ratio of presence to absence points. The model inputs are presented in Table 2. Upon completion of this step, we were left with two large matrices – one with the environmental conditions at known presence locations, and one with the environmental conditions at suspected absence locations.

## 2.2. Model Generation, Projection, and Validation

### 2.2.1. Model Generation

The two matrices were combined and used as the input data for model generation. The boosted regression tree technique described in Elith et al. (2008) was used, implemented by the gbm.step function in the R package dismo (Hijmans et al. 2014). This function assesses the climatic data given as input, and generates a response profile -- an estimate of how the species responds to changes in each of the 19 variables. We ran the model three times for each species, using the same settings, to ensure that the technique provides a robust foundation for range predictions.

### 2.2.2. Model Projection

Once each model was generated, it was projected onto climate grids to estimate of the range of the species. It was first projected onto the modern PRISM grid, and subsequently, onto the downscaled M-H and LGM grids. This projection was done with the dismo function project (Hijmans et al. 2014).

The results were returned as probability of presence. Fractional values were transformed into a binary description of presence/absence using a threshold of 0.4, following a standard convention (Thompson et al. 1998, Bartlein et al. 1998). Below this value, the species was considered to be absent.

The binary rasters were then projected onto an elevation surface and filtered to exclude the top and bottom 10% of the predicted range of elevation. This step was done to eliminate any outlying points that would negatively impact the following analyses. The filtering was done by first converting the binary presence-absence rasters to a set of points. These points were then used to sample the elevation raster. The resulting matrix of latitude, longitude, and elevation was then filtered to include only the 10th through 90th percentile. These data were then used for analysis presented below, and will be referred to as the outlier-eliminated dataset. Using the R package raster function `rasterize` (Hijmans 2014), these points were converted back into raster format and saved for future analysis.

### 2.2.3. Model Evaluation

We used two techniques to evaluate the effectiveness and predictive power of the models. The first was the model output statistics provided by the gbm.step function for each model generation. For each species, we generated three predictive models using the same variables. Because of the stochastic nature of both the absence point generation process and the inherent randomness in the boosting technique, generating three models yielded a better understanding of the modeling paradigm’s predictive performance. Each model generation process uses a 10-fold cross validation to determine the optimum number of trees needed to minimize the model’s deviance. To ensure sufficient predictive power, we aimed for no less than 1,000 trees per model, although *A. tridentata* was included with a mean of 983 trees. The cross validation results were used to assess model predictive power, including 10 accuracy assessments for each model. Discrimination, reported as the area under the receiver operator curve (ROC), describes the ability of the model to correctly separate presence and absence when compared to the observation set. We also show correlation between the cross verification folds, and the deviance of the model. Table 3 shows the model evaluation statistics, averaged across all three model runs.

The second method of validation was a comparison to the estimated biophysical elevation constraints of the species. The known elevation range was obtained from the Jepson Manual, the standard floristic treatment of California Plant Species (Hickman 1996). Using the outlier-eliminated elevation ranges, the maximum elevation of the modeled distribution was compared with the Jepson maximum elevation. The results are presented in table 4, showing that the modeled results come within a several meters of the known elevation limits. The similarities between the known range constraints and the modeled results indicate that the modeling process tends to do a good job of estimating modern the modern climatic response profiles of the species we are investigating.

## 2.3. Analysis

Modeled results were analyzed by assessing changes in distributions at the M-H and LGM by comparing them to the modern baseline distribution. For each species, we extracted the elevations and latitudes of each of points in the outlier-eliminated dataset. These points were then used to compute frequency distributions of the two variables. In addition, spatial analysis of the distribution of the species at each point in time was done using R (R Core Team, 2014). Together, these techniques led to an understanding of the uphill or downhill shift and the north-south shift that the species would have undergone between the three time periods. These trends and the modeled distributions were then compared against well-known paleoproxy evidence, including *Neotoma* midden macrofossils and lacustrine and marine pollen records. These proxies provide evidence of each species’ distribution in the past and are a key to many vegetation and climate reconstructions. Figure 1 shows the approximate geographic location of several of the most important of these sites.

# 3. Discussion

### 3.1 *Sarcobatus vermiculatus*

The modern range of *S. vermiculatus* extends throughout much of Western North America. While its primary distribution is in the Great Basin, presences have been recorded in Alberta and Saskatchewan, as well as some sites east of the Rocky Mountains. In California, its distribution is confined to the far eastern reaches of the state: primarily concentrated in the Owens Valley and in basin and range country north of Lake Tahoe. As a desert plant, it tends to grow in ephemeral stream channels and alluvial salt flats between 600 and 2400 meters. We included this species in this paper because of its distinctive pollen and easily identifiable pollen type.

Modeled distributions suggest that during the M-H, this species significantly expanded its range, primarily to the north. During this expansion, the distribution’s average latitude increased 0.64 degrees (approximately 71 kilometers), while its mean elevation increased by 5.05 meters. During the LGM, the species appears to have undergone a dramatic southward shift, moving all the way into the Mojave lowlands. The mean latitude of the distribution during this period was 3.1 degrees latitude south and 864 meters below the modern distribution.

Davis (1999) reports significant percentages of *Sarcobatus* pollen at Tulare Lake, in the southern San Joaquin Valley at throughout the glacial period until about 8,200 BP. While we do not explicitly see this signal in the modeled results, *S. vermiculatus* shows an LGM distribution significantly west of its present distribution, which does not extend west of the Sierra Crest. During the M-H, modeled distributions of *Sarcobatus* show expansion in the northern part of the state. Conditions at this time on the Modoc Plateau are documented by Anderson et al. (2008) to be significantly more arid than at the present. They interpret a core taken from Flycatcher Basin that shows extreme oxidation and decomposition as evidence that the area was likely a dry depression for the majority of the period 8,400 to 4,500 BP. This type of arid flat may have indeed been suitable for *S. vermiculatus.*

### 3.1 *Artemisia tridentata*

Like *S. vermiculatus,* *Artemisia tridentata* grows throughout the Western United States, with its modern range extending throughout the intermountain West north into Canada. It is fairly common in California, occupying much of eastern California, as well as covering the mountains of southern California. We included this species for two reasons. First, it is included in many prominent vegetation change modeling studies (Bartlein et al. 1998, Thompson et al. 1998), and second, it is found in many pollen records in California and the Western U.S., and has a distinctive and easily identifiable pollen type.

*A. tridentata* is displays a more discontinuous than the many of the other species reported on in this paper. Therefore, the latitudinal shifts may be less meaningful. In fact, we see the opposite of the hypothesized latitudinal and elevation shifts in the M-H, as it shifts to the south of 0.41 degrees latitude (~45 km) and downslope about 300 meters. This modeled distribution shows the species’ to be concentrated in the lower slopes of the Transverse ranges of southern California with patchy presences throughout the Tehachapis and into the Owens Valley. The LGM distribution shows a mean shift of 0.25 degrees latitude to the south (~28 kilometers), and a drastic 1038-meter downslope shift in elevation compared to the modern baseline. It continued to dominate the lower slopes of the mountains of Southern California, and spread out into the Mojave lowlands during this time. It also appears to expand into the southern end of the San Joaquin Valley.

*A. tridentata* pollen is very commonly encountered in fossil pollen assemblages and is often used in vegetation reconstructions to support the presence of cold steppe conditions. During the LGM, the Owens Valley is thought to have been dominated by this species (Woolfenden 1996, Koehler et al. 2005). Moreover, over the last 180,000 years, numerous cold and wet glacial periods have been show to be accompanied by a downslope expansion of juniper woodland and sagebrush, and a displacement of warm desert scrub. These events are departures of between 2°C and 6°C from today’s climate and an additional 100 to 350 mm of annual precipitation. Conversely, interglacial periods have been routinely accompanied by the expansion of warm desert scrub and an upslope movement of the sagebrush steppe (Woolfenden 2003). While the model indicates the downslope expansion of *Artemisia* during the last glacial period, the expansion does not include the Owens Valley drainage. Davis contends that steppe conditions existed in much of Southern California, extending as far north as the San Joaquin Valley. Fossil pollen evidence from Tulare Lake confirms the presence of both *S. vermiculatus* and *A. tridentata* together in this region during the LGM. Indeed, the model suggests that *A. tridentata* was likely present in the San Joaquin Valley during the LGM.

*Artemisia tridentata* was also important during the M-H, especially on the western side of the Sierra Nevada, where it is not found today. The M-H is widely recognized as a warm and arid period in much of California (Davis et al 1985, Lindstrom 1990, Benson et al. 1989), and is often associated with general upslope movement of species distributions (Koehler 2005, Smith and Anderson 1992). Increased percentages of low- to mid- altitude montane shrubs, such as *A. tridentata* are present during the Early and Middle Holocene at several high altitude lakes in the Sierra Nevada (Anderson 1990). Studies also indicate that nonarboreal pollen types were more important in the southern Sierra Nevada between 11,000 and 7,000 years BP than at present*.* At lower elevations, packrat midden evidence from Kings Canyon suggests that an *A. tridentata-*dominated steppe was present, similar to contemporary vegetation found 500-1000 meters higher on the eastern side of the Sierra (Cole 1983). The Tulare Lake record (Davis 1999) also notes the M-H expansion of *Artemisia* steppe conditions between 7,000 and 4,000 BP, and a lake lowstand between 6,900 and 6,600 BP. While modeled distributions show some presence of *A. tridentata* in the southern Sierra Nevada and Central Valley, they are not indicative of the widespread sagebrush steppe shown by these paleoenvironmental studies to dominate west of the Sierra Crest.

### 3.1 *Yucca brevifolia*

While occupying a similar niche to *A. tridentata* and *S. vermiculatus,* *Y. brevifolia*  shows a much more confined range than these species. Its modern distribution is limited to eastern California, southern Nevada, and Northern Arizona. In California, its range is primarily concentrated in the Tehachapi and Transverse Ranges, as well as the mountains near Death Valley. The modeled distributions of *Y. brevifolia* shifted north approximately 0.2 degrees latitude in the M-H (~13 km). This northward shift is accompanied by a 68-meter upslope shift. This relatively small shift suggests that the M-H distribution was similar to its modern range. During the LGM, the species displayed a downslope shift of 600 meters when compared to the present, and a latitudinal shift of 0.12 degrees to the south (~20km). This shift indicates the presence of small, isolated stands throughout the Mojave region during the LGM.

Packrat evidence shows that *Yucca brevifolia* was absent from the Mojave by c. 10,000 BP (Koehler et al. 2005), and that during the M-H, the *Y. brevifolia* distribution displayed a similar pattern to that of the present. While little evidence is readily available about the distribution of *Y. brevofolia* during the LGM*,* packrat middens (Koehler et al 2005) suggest that it would likely have been have been found in the relatively mild-mesic region south of 36N, which coincides well with the modeled distribution. Additional midden studies have found *Y. brevifolia* macrofossils at elevations as low as 425 meters in Death Valley, and 258 meters in the Sonoran Desert (Wells and Woodcock 1985). These environments are no longer capable of sustaining ‘high-desert’ juniper-*Y. brevifolia* woodlands, but would be within the elevation range of our modeled LGM distribution.

### 3.2. Dry Climate Conifers

### 3.1 *Pinus ponderosa*

Ponderosa pine is another taxon that is common throughout the Western United States, extending into Canada. Capable of withstanding very cold winters as well as hot and dry summers, its potential sites are widespread. In California, it occupies the much of the Sierra Nevada and the North Coast Ranges, extending into the Tehachapi and Transverse Ranges as well. It is included in the Thompson et al. (1998) modeling effort. It is included here to because of its modern importance in California. During the M-H, its distribution shifted north 0.8 degrees latitude (~89 km) and downslope of around 36 meters. This shift pushed the species fully into the Coast Range, as well as the expanded its presence in the northern Sierra. During the LGM, the distribution was 1.2-degree (~130km) south and a 800 meters below the modern mean. This suggests the distribution was widespread in the Central Valley and the southern Coast Ranges, where it is not present today.

Packrat midden evidence at Kings Canyon includes macrofossils of *P. ponderosa* at 20,000 years BP (Cole 1983). *P. ponderosa* is not present this area today, and this downslope shift coincides well with our modeled results. Anderson (1990) lends support for this theory, citing that modern mixed conifer forests grew at elevations 300-700 meters lower than today during the late Pleistocene. Beginning around 11, 000BP, there is a transition away from an *A. tridentata* dominated steppe towards the emergence of *P. ponderosa* woodland (Smith and Anderson, 1992). Moreover, macrofossil evidence of the species was recovered from Nichols Meadow from between 8800 and 6000BP (Koehler and Anderson 1994). This site’s elevation of 1509 meters fits well within our modeled M-H distribution.

### 3.1 *Pinus sabiniana*

Another conifer, *P. sabiniana,* is also important in the California Foothills. This species occupies sites at lower elevations than its counterpart, and is widespread at the edges of the Sacramento and San Joaquin Valleys, as well as the Southern Coast Ranges. Outside of California, its presence is very limited, with just a few stands in Southern Oregon. *P. sabiniana* was included in this study because of an unusual gap in its distribution between the Kings and Tulare Rivers in Central California. This gap has been thoroughly investigated (Whitaker et al. 2014). Our model is successful in reflecting this phenomenon, lending support to our model’s ability to accurately predict species ranges.

During the M-H, *P. sabiniana* shifts 0.93 degrees (~100 km) to the north. Unlike *P. ponderosa,* however, *P. sabiniana* shows an upward shift during this time period of about 100 meters, suggesting that the two species may have shared the same niche during this time period. Indeed, the modeled distributions appear to overlap significantly in the North Coast Ranges and the northern Sierra. During this time, *P. sabiniana* also included a significant distribution in the Tehachapi Range of southern California.During the LGM, however, *P. sabiniana* appears to have been isolated to only a few stands in the Mojave. This drastic shift is characterized by a 74-meter upslope shift, and a 2.3-degree (~252 kilometer) southward shift in latitude. This distribution suggests that the species was primarily distributed in the lower San Joaquin Valley and the Eastern Mojave.

Pollen and macrofossil evidence for this species is extremely limited. The species was primarily included to investigate the break in distribution in Central California. Because we are able to successfully resolve this gap, we suggest that this discontinuity is due entirely to climate. Whitaker et al. indicate that the gap is older than 6ka, while our model shows widespread presence in the region during the M-H. This implies an inaccuracy in the climate layers used as input for our model. Limited macrofossil evidence from Swamp Lake confirms an upward shift of the species during the Holocene, indicating lower winter precipitation and higher summer temperatures than at present (Smith and Anderson, 1992).

### 3.3 *Tsuga heterophylla*

*T. heterophylla* is a large tree that typically requires conditions significantly wetter and cooler than the other species discussed here. It is widespread up the pacific coast into Alaska and Western Canada, with inland distributions in Oregon and Washington. In California, it is limited to sites close to the coast, where it is supported by the fog common to this region. Its pollen type is clearly identifiable, and it is reported on in many paleoecological studies.

During the M-H, the *T. heterophylla* distribution was 0.3 degrees (~33km) north and 15 meters below the modern distribution, while during the LGM, the average latitude was 0.91 degrees (~102km) south and 130 meters above the modern distribution. Its modeled M-H distribution was further confined to the northernmost coastal areas of the state. During the LGM, however, the model suggests that it expanded both inland and southward, reaching as far south as the San Francisco Bay Area. Figure 4 displays these changes in distribution.

The coastal margin is more difficult to model and validate than the interior of the continent. Major regional differences exist along the coast of California, and these differences are difficult to resolve with a GCM, even with advanced downscaling techniques (Mondal et al 2015). Ocean circulation in the northern California region is thought to have reached its current configuration at about 13,000 BP (Barron et al. 2003). Several estimates show sea surface temperatures changes between 2°C and 5°C between the LGM and the modern period off the coast of California (Benson et al. 2002, Ortiz 1997, Mix et al. 1999). Cooler sea surface temperatures could have promoted the growth of *T. heterophylla*, however, it is not clear that the fog and moisture, needed to support these trees was present during the LGM. *Sequoia sempervirens*, which is well understood to need fog and mild temperatures, does not appear on the Northern California coast until approximately 5000 BP (???). Benson suggests that the Middle and Early Holocene were warm and dry, with especially warm winters, and that the maritime influence did not reach its current pattern around 5,000 BP.

A core taken at Clear Lake is shows the transition to the modern inland Coast Range’s chaparral vegetation around 7,000 BP, as evidenced by the growing importance of the Rosaceae, Amaranthacae, and Rhamanacae pollen type, a chaparral indicator. This is reflected in our modeling results, as *T. heterophylla* shows significant contraction from its LGM distribution during the M-H. The upward shift in the LGM is a result of the evapotranspiration control on the species distribution, and indicates an expansion of the coastal influence further in the interior of the state. Conversely, during the M-H, the maritime influence is smaller than at present, which would align with studies suggesting a warm and arid Early to Middle Holocene.

# 4. Conclusion

When comparing the modeled estimates of species presence to records of species derived from the fossil record, it is clear that the model seems to lack the ability to precisely determine the location of the species at the particular time periods. While the summary statistics of the distributions and the general trends in latitude and elevation between the modern, M-H, and LGM tend to agree with the expected changes, few species are present at the locations known to have supported the species in the past. For example, Davis’ study 1999 study indicates the presence of *S. vermiculatus* at Tulare Lake during the LGM, while our modeled results show that while its distribution at this time was significantly west of the modern distribution, it was far from the Southern San Joaquin Valley. Similarly, Davis’ 150kya study of the Owen’s valley suggests the presence of *A. tridentata* in the Owens Valley throughout the LGM. However, our results show the near complete absence of this species from this entire region. Bartlein et al. refered to these discrepancies as mismatches, and identified three potential causes for these, to which we will add a fourth (1998).

The two potential causes of mismatches between observed and modeled distributions is the inaccuracy of using the fossil pollen record as an estimate of past distribution and the inaccuracy of using the fossil record as a standard for comparison. Complex topography in California can make assessing the precise locations of wind-pollenated plant distributions difficult from pollen records alone. Packrat middens, however, typically reflect the environments within 30 meters of the site (Finley 1990). The combination of the two is generally accepted as a good indicator of environmental change and species ranges.

The third cause of observed-modeled range mismatches identified by Bartlein is that the observed paleoecological is correct, while the models are inaccurate. This is a tempting explanation, however, our models actually preform quite well when predicting the range of the species for the modern time period. For example, the modeled distribution of *P. sabinaiana* clearly reflects the unusual gap in distribution between the Kings and Tulare Rivers in central California. The other species in the study also conformed to known ranges and theoretical maximum elevations. Over all runs in the study, the avergage departure from known range constraints was only -55.3 ± 116.8 meters, indicating great predictive power for current ranges. Thus, this explanation for the discrepancy between modeled distributions and fossil evidence seems unlikely, instead, we propose a fourth explanation for the mismatches.

While the modern climate used as input is based on the 1981-2010 average of observed climate, the two historical time periods are based on general circulation model output, and thus are subject to the assumptions and errors inherent to these models. These models are widely known to accurately estimate temperatures, but are far more error prone when modeling precipitation. The signal to noise ratios for GCM precipitation projections are typically far lower than those for temperature (Hawkins and Sutton, 2011). Our modeled distributions tend to show the correct trends in shifting southward and downslope in the LGM and northward in the M-H. These trends are typically temperature controlled, especially the changes in the elevation of the distribution. It follows then that the mismatches in precise observed and modeled ranges are due to errors in the precipitation variables used as input for the models. Further research should be pointed at characterizing the magnitude of mismatches. Such a study would allow forward-projected modeling results to more accurately quantify their error under global warming scenarios.

# 5. References

Anderson, R. S. (1990). Holocene Forest Development and Paleoclimates Within the Central Sierra Nevada, California. *Journal of Ecology*, *78*(2), 470–489. doi:10.2307/2261125

Anderson, R. S., Smith, S. J., Jass, R. B., & Geoffrey Spaulding, W. (2008). A Late Holocene Record of Vegetation and Climate from a Small Wetland in Shasta County, California. *Madroño*, *55*(1), 15–25. doi:10.3120/0024-9637(2008)55[15:ALHROV]2.0.CO;2

Barron, J. A., Heusser, L., Herbert, T., & Lyle, M. (2003). High-resolution climatic evolution of coastal northern California during the past 16,000 years. *Paleoceanography*, *18*(1), 1020. doi:10.1029/2002PA000768

Benson, L. V., Currey, D. R., Dorn, R. I., Lajoie, K. R., Oviatt, C. G., Robinson, S. W., … Stine, S. (1990). Chronology of expansion and contraction of four great Basin lake systems during the past 35,000 years. *Palaeogeography, Palaeoclimatology, Palaeoecology*, *78*(3–4), 241–286. doi:10.1016/0031-0182(90)90217-U

Berkeley Initiative In Global Change Biology. (n.d.). Berkeley Econinformatics Engine. Retrieved January 19, 2015, from https://ecoengine.berkeley.edu/

California Department of Forestry and Fire Protection. (2009, February). 1:24,000 County Boundaries. Shapefile, CalAtlas.

Cole, K. (1983). Late Pleistocene vegetation of Kings Canyon, Sierra Nevada, California. *Quaternary Research*, *19*(1), 117–129. doi:10.1016/0033-5894(83)90031-5

Daly, C., & Bryant, K. (2013, June). The PRISM Climate and Weather System - An Introduction. Oregon State University. Retrieved from http://prism.oregonstate.edu/documents/PRISM\_history\_jun2013.pdf

Daniels, M. L., Anderson, S., & Whitlock, C. (2005). Vegetation and fire history since the Late Pleistocene from the Trinity Mountains, northwestern California, USA. *The Holocene*, *15*(7), 1062–1071. doi:10.1191/0959683605hl878ra

Davis, O. K. (1999). Pollen analysis of Tulare Lake, California: Great Basin-like vegetation in Central California during the full-glacial and early Holocene. *Review of Palaeobotany and Palynology*, *107*(3–4), 249–257. doi:10.1016/S0034-6667(99)00020-2

Davis, O. K., Scott Anderson, R., Fall, P. L., O’Rourke, M. K., & Thompson, R. S. (1985). Palynological evidence for early Holocene aridity in the southern Sierra Nevada, California. *Quaternary Research*, *24*(3), 322–332. doi:10.1016/0033-5894(85)90054-7

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, *77*(4), 802–813. doi:10.1111/j.1365-2656.2008.01390.x

Elith, J., Phillips, S., Hastie, T., Dudik, M., Chee, Y. E., & Yates, C. L. (2011). A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, *17*, 43–57.

Gavin, D. G., & Hu, F. S. (2006). Spatial variation of climatic and non-climatic controls on species distribution: the range limit of Tsuga heterophylla. *Journal of Biogeography*, *33*(8), 1384–1396. doi:10.1111/j.1365-2699.2006.01509.x

Hawkins, E., & Sutton, R. (2011). The potential to narrow uncertainty in projections of regional precipitation change. *Climate Dynamics*, *37*(1-2), 407–418. doi:10.1007/s00382-010-0810-6

Hickman, J. C. (Ed.). (1996). *The Jepson Manual: Higher Plants of California*. University of California Press.

Hijmans, R. J., Phillips, S., & Elith, J. L. and J. (2014). dismo: Species distribution modeling (Version 1.0-5). Retrieved from http://cran.r-project.org/web/packages/dismo/index.html

Koehler, P. A., Anderson, R. S., & Spaulding, W. G. (2005). Development of vegetation in the Central Mojave Desert of California during the late Quaternary. *Palaeogeography, Palaeoclimatology, Palaeoecology*, *215*(3–4), 297–311. doi:10.1016/j.palaeo.2004.09.010

Koehler, P. A., & Scott Anderson, R. (1994). The paleoecology and stratigraphy of Nichols Meadow, Sierra National Forest, California, USA. *Palaeogeography, Palaeoclimatology, Palaeoecology*, *112*(1–2), 1–17. doi:10.1016/0031-0182(94)90132-5

Lindstrom, S. (1990). Submerged Tree Stumps as Indicators of Mid-Holocene Aridity in the Lake Tahoe Basin. *Journal of California and Great Basin Anthropology*, *12*(2), 146–157.

Mondal, Y., Farley, S., Chiang, J., & Koo, M. S. (n.d.). Reconstructions of the Last Glacial Maximum and Mid-Holocene using CCSM4 simulations over the Continental United States for Ecological Modeling Applications.

Mozingo, H. N. (1987). *Shrubs of the Great Basin: A natural history*. Reno, NV.: University of Nevada Press.

Ram, K. (2014). Ecoengine: Programmatic interface to the API serving UC Berkeley’s Natural History Data. (Version R package version 1.6). Retrieved from http://CRAN.R-project.org/package=ecoengine

R Core Team. (2014). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Robert S. Thompson, Steven W. Hostetler, Patrick J. Bartlein, & Katherine H. Anderson. (1998). A Strategy for Assessing Potential Future Changes in Climate, Hydrology, and Vegetation in the Western United States. United State Geological Survey.

Smith, S. J., & Anderson, R. S. (1992). Late Wisconsin paleoecologic record from Swamp Lake, Yosemite National Park, California. *Quaternary Research*, *38*(1), 91–102. doi:10.1016/0033-5894(92)90032-E

Spaulding, W. G. (1990). Vegetation Dynamics during the Last Deglaciation, Southeastern Great Basin, U.S.A. *Quaternary Research*, *33*, 188–203.

Spaulding, W. G. (1991). A middle holocene vegetation record from the Mojave Desert of North America and its paleoclimatic significance. *Quaternary Research*, *35*(3, Part 1), 427–437. doi:10.1016/0033-5894(91)90055-A

Whitaker, A. R., Rosenthal, J. S., & Wohlgemuth, E. (2014). The Holocene biogeography of Pinus sabiniana Douglas ex D. Don (gray pine) in California. *Vegetation History and Archaeobotany*, *23*(6), 683–692. doi:10.1007/s00334-014-0449-z

Woolfenden, W. B. (1996). Late Quaternary vegetation history of the southern Owens Valley region, Inyo County, California. Retrieved from http://arizona.openrepository.com/arizona/handle/10150/282184

Woolfenden, W. B. (2003). A 180,000-year pollen record from Owens Lake, CA: terrestrial vegetation change on orbital scales. *Quaternary Research*, *59*(3), 430–444. doi:10.1016/S0033-5894(03)00033-4

Wright, H. E., & Porter, S. C. (1983). *Late Quaternary Environments of the United States: The Late Pleistocene*. U of Minnesota Press.

# Tables