AOS C204 Final Project
Bowen Wang

Projecting the Distribution of *Pinus monophylla* under Climate Change

Bowen Wang

2023.12.8

1, Introduction

Pinus monophylla, also known as the single-leaf pinyon pine, is a principal species in the pinyon-juniper woodlands where it coexists with the California and Utah juniper and serves important ecological functions in its local ecosystem. It is distributed widely in the southwestern United States, especially in the mountainous areas of California, Nevada, Arizona, and Utah, typically between 1000 and 3000 m in elevation and dry, coarse soil. The distribution of single-leaf pinyon pines, like that of any other species, is influenced by the habitats' local climate (Cole et al., 2008). However, anthropogenic climate change is driving changes in hydroclimate conditions at a rate that exceeds what many species can acclimate or shift to new locations. It thus comes to question whether there remain suitable locations for many tree species to survive in. In the southwestern United States (US) where *P. monophyla* primarily lives, the climate is projected to become warmer and dryer and thus subject to the threat of intensifying drought conditions.

Despite the apparent risk, less is known about how the distribution of *P. monophyla* is going to change. First of all, the distribution of any tree species – including the single-leaf pinyon pines - is determined by many different hydroclimate variables. How changes in each variable will affect an area's suitability as a potential habitat is often unclear and hard to quantify. Such relationships are often highly non-linear, with a certain range of conditions being the most ideal and any condition that deviates from it is less ideal. Changes in different variables may also interact with one another, which can further complicate the predictions of how the species' distribution will change.

For these reasons, machine learning (ML) algorithms are extremely helpful tools to predict a species' range of suitable conditions. ML classification methods are often apt for taking datasets with large numbers of variables and samples and can learn from larger datasets through iterations. They are also able to represent non-linear relationships that are often necessary for understanding how climatic variables affect an area's suitability for a species. In this project, I will first build an ML model that can predict the suitable habitats for *P. monophylla* in the historical climate based on observational presence data of the species; using this model, I will use projected future climate data to project how the distribution of potential habitat for the species may change. In an earlier class project, I demonstrated that a random forest model can serve this purpose for predicting the distribution of *P. monophyla* in California, here I'm interested in expanding the scope of the model to include its entire habitat in the southwestern US and understanding the role each variable plays on deciding where the species is predicted to be able to survive.

2, Methods

2.1, Historical and Projected Climate Data

In this project, I used hydroclimate data from 12 global climate models (GCMs) from the 5th Coupled Model Intercomparison Project (CMIP5). Since the resolution of the raw model output is often very coarse, the models are statistically downscaled to a horizontal resolution of 1/8th degree (i.e., ~13.89 km) to allow for analyses on a regional scale. While neither the models nor the downscaling method represent the most up-to-date methods, I selected these datasets because they are both publicly available and based on easy-to-use rectangular grids defined by longitude and latitude coordinates. The data can be downloaded at https://gdo-dcp.ucllnl.org/downscaled_cmip_projections.

Here, I selected the model simulations that follow the Representative Concentration Pathway 8.5 (RCP8.5), which is the worst-case carbon emissions scenario. While this scenario has been deemed unrealistic based on the carbon emissions in the recent past, it can demonstrate the effects of climate change on the distribution of P. monophyla most clearly due to its large projected changes in the hydroclimate.

For training the model, I used the variables in the historical climate (i.e., 1990-2020) with a total of 5 variables: precipitation, daily maximum temperature, daily minimum temperature, soil moisture content, and evapotranspiration. Snow water equivalent was also evaluated but led to worse model performance and was therefore not included in the model. To project for future species distribution, I tested the model using the climate conditions of mid-century (i.e., 2035-2065) and end-of-century (i.e., 2070-2100).

2.2, Observed species distribution

To construct a model that can predict the distribution of *P. monophylla*, observed presence data of the species is needed to understand where the species is currently distributed. I first used data from iNaturalist (inaturalist.org, 2023), which is a platform that delivers datasets of crowd-sourced plant and animal observations. Users of the iNaturalist application use their phones to take photos of plant samples that they observe, and the application, through its algorithm, will return to the user the species of the sample that they captured in the picture. With the users' permission, the information collected by the users will become a part of the iNaturalist’s dataset where the times and locations at which the species is observed are recorded. For *P. monophylla*, I downloaded from iNaturalist a dataset with a total of 4804 observations, where most of them occurred in California and Nevada, with a decent representation of the species in Arizona and Utah.

Crowd-sourced databases like iNaturalist can be extremely useful and can save much time and resources from the often time- and labor-intensive scientific surveys, they are limited in the sense that they are not as systematic as the more rigorous surveys and often underrepresent the distribution of samples in areas that are more difficult to access. These include less populated and less traveled areas, places where conditions are unfit for human activities, etc. To account for this limitation, I complemented the iNaturalist observations with a peer-reviewed range map of *P. monophyla* in the Great Basin created by the US Geological Survey based on sources like existing digitized maps, tabular data, etc (Cole et al., 2003).

Since both of the datasets described here are in vector data, to incorporate them into the same model as the climate projections, I overlaid these two datasets and gridded them in the same way as the climate data. A grid cell would get a value of 1 as long as there is at least one observation, regardless of the number of observations that occurred in each grid cell. If there are no observations at all in a grid cell, it will get a value of 0 (Figure 1).

It is important to note that although here we have data on where *P. monophyla* is observed to be present, there is no data available on where the species is absent. For this reason, a common practice in species distribution modeling is to create pseudo-absence points, which are randomly generated points/grids in the domain that are assumed to have no samples of that species present. While the generation of such pseudo-absence points can have a wide range of complexities, here I adopted a simple approach that among grid cells with no overlaps with any observed presence data, I randomly selected grid cells so that the number of pseudo-absence grid cells is the same as the grid cells with presence data. This approach ensures that the training dataset is balanced between positive and negative data values.

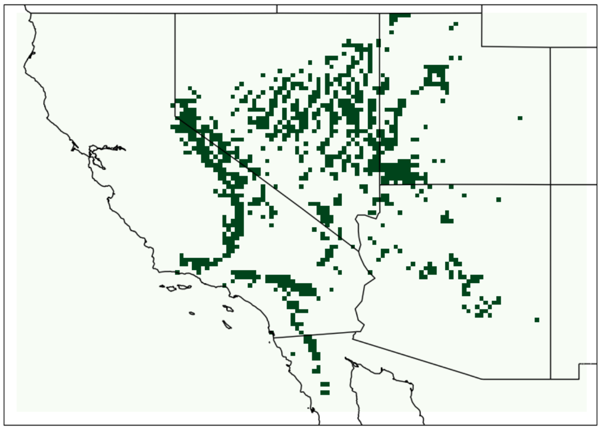


Figure 1. Observed distribution of *P. monophyla* on the grids of GCM climate data. Green grids represent the presence of the species.

2.3, Machine learning (ML) model

Since the question asked in this research project is to predict and project if each grid cell is a suitable habitat for *P. monophyla*, it is a classification problem and needs to be modeled as such. With the observed species distribution data available as training data, it requires supervised learning where the model needs to be trained with a portion of the observed dataset, tested with a subset of the dataset that is not used for training, evaluated with some metric, and applied to make predictions in both historical and future climates.

The earlier version of this project used a random forest and was able to successfully predict the range of suitable habitats for *P. monophyla* in California. However, random forest classifiers struggle with the larger domain used in this project and cannot faithfully represent suitable habitats for the species under the historical climate. For this reason, I instead used a multilayer perceptron (MLP) classifier, which is an artificial neural network with fully connected neurons on different hidden layers. A standard scaler is used to scale the training data before exposing it to the neural network.

To decide the model setup, I used the f1 score to evaluate the model using a cross-validation approach. That is, instead of simply separating the dataset between training and testing and getting a test score just once, here I used a random 75% of the data for training and the remaining 25% for testing, trained the model, and evaluated the test score, and repeated the process for 20 times. In this way, for each model setup, I could get 20 f1 scores and take their mean to average out the potentially large variabilities in model performance. The best-scoring model is then used to predict the suitability of each grid cell in the domain as a potential habitat for *P. monophyla*. It is important to note that it is expected that the predicted range is larger than the observed range, because the predicted term here is all the potentially suitable habitats in terms of climate conditions, but many other non-climatic factors may prevent the species from potentially appearing in these locations.

The resulting model has 2 hidden layers with 20 neurons on each hidden layer of the MLP with a cross-validation f1 score of 0.62, with each score ranging between 0.57 and 0.66, showing that the model performance is stable and not subject to large random variabilities. Adding another one or two hidden layers of neurons may slightly improve the f1 scores (by

~0.02), but the projected results from these different variants are very similar, and the simplicity of the model structure is here favored over the minor improvement in f1 scores.

3, Results

To answer how the potential range of distribution of *P. monophyla* may be affected by climate change, it is first important to understand the current state of the climate in the southwestern US and how the climate is projected to change. Using the ensemble of 12 CMIP5 GCMs, it is clear that the geographical variability is large in the baseline climate in the historical period in the southwestern US (Figure 2). For example, there is a wide range of precipitation, ranging from ~10 mm/month to 500 mm/month. The variabilities in evapotranspiration and soil moisture follow closely with that in precipitation, as they reflect the amount of water available at a certain location. Daily minimum and maximum temperatures look very similar in a relative sense, and both show a very clear elevation effect, where areas of higher elevations have cooler temperatures and vice versa. Snow water equivalent is concentrated mostly in the high-elevation mountain regions.

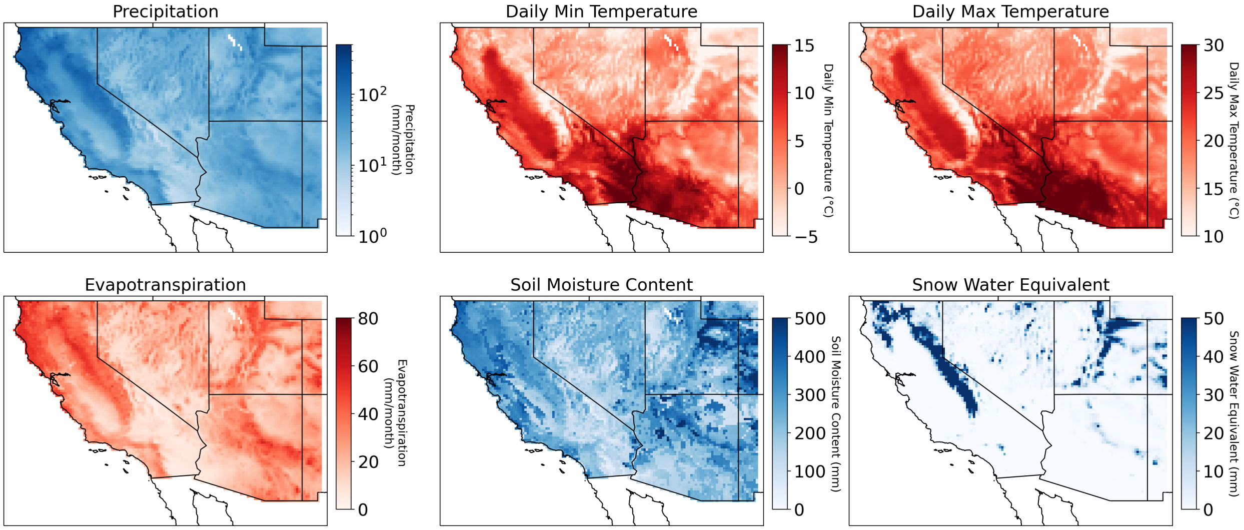


Figure 2. Hydroclimate conditions of the southwestern US in the historical period (1990-2020) as simulated by the GCM ensemble.

The temperature in the southwestern US is projected to increase almost uniformly by 4-5 °C by the end of the 21st century in comparison with the 1990-2020 historical period, which leads to a large decrease in snow water equivalent wherever it was available in the historical period (Figure 3). Although in general, the projected temperature increase tends to be milder near the coast and more intense further inland, the difference is small in comparison with the overall magnitude of the increase. Precipitation is projected to increase in most of California, Nevada, and Utah, while it will slightly decrease in Arizona; evapotranspiration follows a very similar pattern in terms of the regions of increases and decreases. Soil moisture is projected to increase quite significantly in Nevada and Utah, but California and Arizona are mostly expecting decreases in soil moisture.

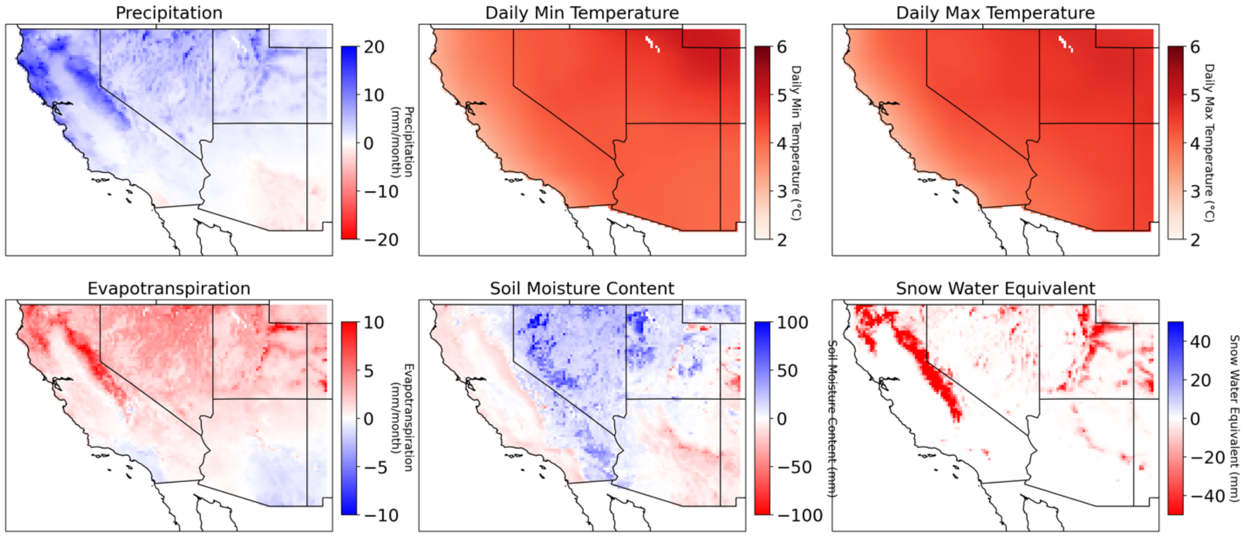


Figure 3. Change in hydroclimate conditions of the southwestern US between the historical period (1990-2020) and the end-of-century (2070-2100) as simulated by the GCM ensemble.

Incorporating the climate conditions into the MLP classifier, we can see that the predicted distribution of potential habitats for *P. monophyla* generally matches the observed records (Figure 4). Notably, the Sierra Nevada in California and much of Nevada, which are the two major clusters of the observed *P. monophyla*, both show up in significant areas as potential habitats for the species. The observed distributions in Utah and Arizona are also both covered within the predicted range. Northern California and Nevada are also projected to be suitable despite having few observed samples in those areas. However, as the climate continues to change, the historically suitable areas will start to become unfavorable conditions for the species. Even in mid-century, both Central Nevada and the Sierra Nevada will start to lose a significant portion of the habitats that used to be available in the historical period. This decline becomes even more intense by end-of-century, when only a very small fraction of the Sierra Nevada and Nevada will remain as suitable habitats for *P. monophyla*. However, it is noticeable that Utah is expected to have some more available climatic conditions for the species by end-of-century in comparison with the historical period.

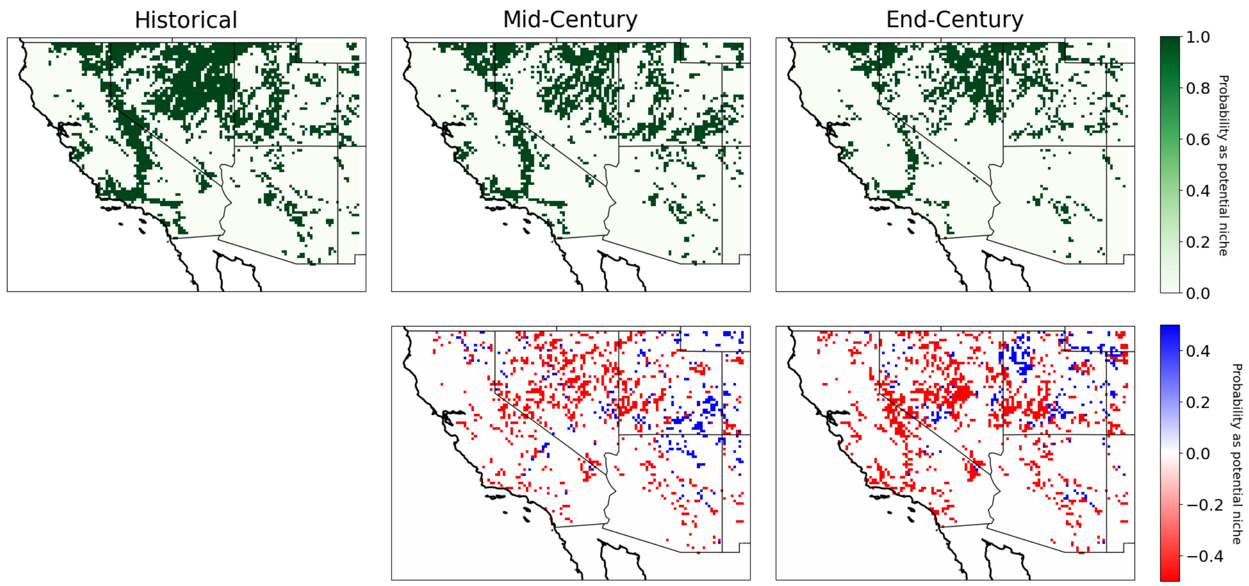


Figure 4. Deterministic estimates of each grid cell as a suitable habitat for P. monophyla using the MLP classifier in the historical period, mid-century, and end-of-century, and the changes from the historical baseline.

In comparison to Figure 4 which shows deterministic estimates, Figure 5 instead shows probabilistic estimates of each grid's chance as a suitable habitat. While both figures show very similar information, Figure 5 provides additional information regarding the whole domain, rather than only the areas that are predicted to be suitable. It is shown here that by the end of the 21st century, a decent portion of Utah and only thin stretches of areas in Nevada and Arizona will have higher probabilities of being a suitable habitat, while the rest of the domain will all become less likely to be able to host *P. monophyla*. Notably, none of the areas that already have a concentration of either observed or predicted range is projected to have a continued increase in probabilities, indicating that all the species will likely need to shift to new habitats in the future climate.

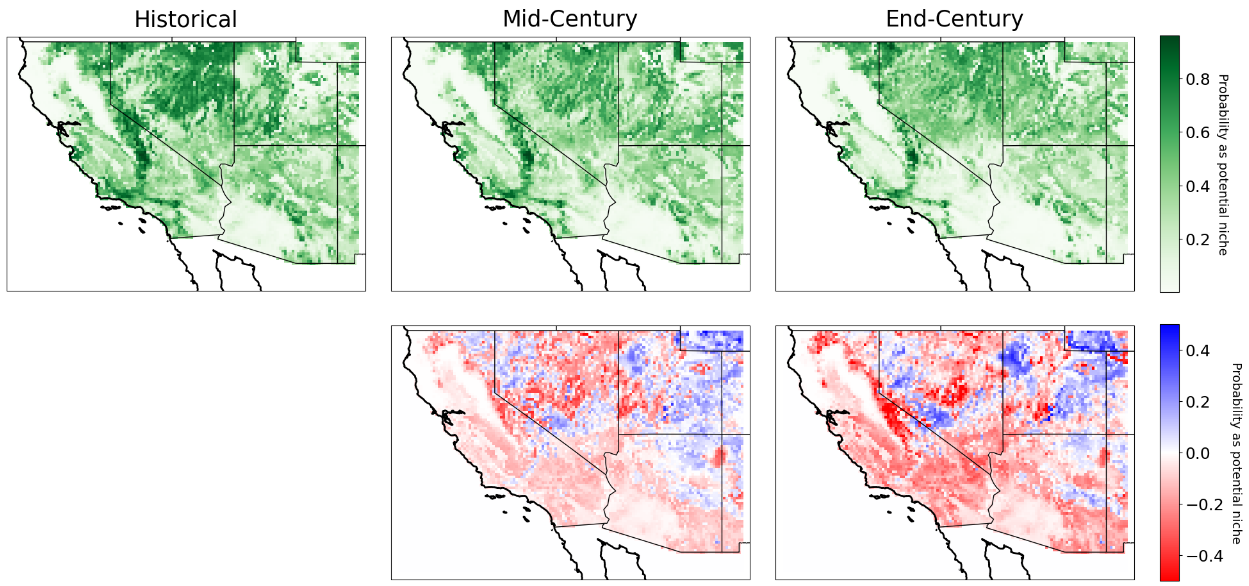
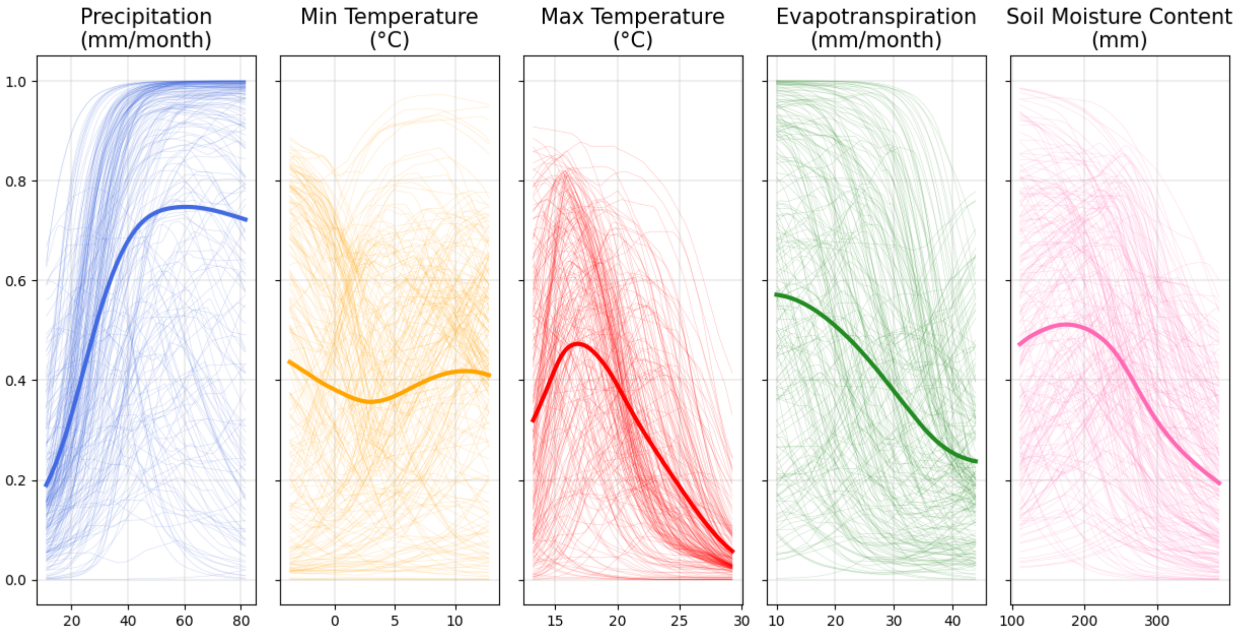


Figure 5. Probabilistic estimates of each grid cell as a suitable habitat for P. monophyla using the MLP classifier in the historical period, mid-century, and end-of-century, and the changes from the historical baseline.

To partition the influence of each variable on the estimates of suitable habitats, I applied a partial dependence approach to detect how changes in each variable's values affect the probabilities. In Figure 6, it is shown that daily minimum temperature tends to have the smallest effect on the predictions than any other variable. The probabilities associated with different minimum temperature values barely fluctuate, in contrast to the daily maximum temperature, whose associated probabilities have a clear peak near 17°C and can quickly decrease above or below that ideal temperature. The partial dependence plot also shows that P. monophyla tend to prefer areas with lower evapotranspiration and soil moisture content values, suggesting that they tend to live in dryer soil conditions, which corresponds well with the existing documentation of the species' climatic preferences. While there is a consistent decreasing trend in probabilities associated with higher evapotranspiration, soil moisture has a peak at around 200 mm beyond which the probabilities can quickly decline. Lastly, Figure 6 shows a preference for higher precipitation values (i.e., >40 mm/month, or 480 mm/year), which tends to be higher than the normal expected range of precipitation for P. monophyla. However, the higher probabilities at higher precipitation values may be due to a smaller number of samples with such precipitation rates in the southwestern US.



4, Discussion and Conclusion

In this project, I demonstrated that using an MLP neural network classifier can help predict the range of potentially suitable habitats for *P. monophyla* in its whole range in the southwestern US and project how climate change might affect its future distribution. It is demonstrated that there is as the climate in this region continues to warm, the suitable habitat for the species will shrink, and current favorable locations may no longer be suitable in the future. It thus clearly suggests that P. monophyla is under the threat of climate change. A partial dependence approach also helps elucidate what changes may be driving the shrinkages in habitat range. The significant warming and increasing maximum temperatures in this region is likely the main reason as the temperature quickly goes beyond the ideal threshold for *P. monophyla*. Increasing soil moisture and evapotranspiration in certain areas may also make it difficult for the species to adapt as it prefers dry soil.

Despite the important conclusions that we can draw, it is however important to note a major limitation of this study and perhaps any other study in the realm of species distribution modeling. The concept of potentially suitable habitat is not observable or verifiable and may only be proven with experiments. For example, there may be two places with the same climate conditions, and one of them has *P. monophyla* while the other one doesn't. Climatologically, the two locations are equally suitable for hosting the species, but there may be many other factors that determine this difference in the observed world, including inter-specific competition, presence/absence of pests, or even random chance. Without research actually trying to plant the species in that place, and without knowing what potentially suitable means in a concrete and quantifiable sense, there is no good way of evaluating how well the model does in validating the observed ranges.

5, References

Cole, K. L., Ferguson, G., Cannella, J., Spellenberg, R., Sanders, A., Arundel, S., & Riser, A. J. (2003). Range Map of Single-Needle Pinyon Pine (Pinus monophylla) [Data set]. https://www.sciencebase.gov/catalog/item/537f6bb6e4b021317a870eae

Cole, K. L., Fisher, J., Arundel, S. T., Cannella, J., & Swift, S. (2008). Geographical and climatic limits of needle types of one- and two-needled pinyon pines. Journal of Biogeography, 35(2), 257–269. https://doi.org/10.1111/j.1365-2699.2007.01786.x

INaturalist. (2023). iNaturalist. http://inaturalist.org