Examining the impact of climate models on the distribution of tree species in Colorado forests

John D. Drumheller

1 Dept. of Statistics/Master of Applied Statistics Program, Colorado State University, Ft. Collins, CO, USA. john.drumheller@colostate.edu

Abstract

This report describes and experiment where ERA-Intim Climate data is combined with USFS Forest Inventory and Analysis data to examine the relationship between climate and distribution of tree species in Colorado forests. An interpolation scheme is developed to relate the position of plot locations with estimates of climate variables. The XGBoost algorithm, and its associated open source library, is used to predict the distribution of tree species by modeling the presence or absence of a species. The distribution of lodgepole pine has the strongest relationship with climate related variables, in particular the difference between summer high and low temperatures. A method for abundance estimation using a Tweedie loss function is also discussed in addition to the shortcomings of the modeling and suggestions for further investigation.

Author summary

I am a masters of applied statistics distance student at Colorado State University and currently reside outside of Chicago, Il. My anticipated graduation data is Summer 2021. I am very exited to have the opportunity to participate in this competition and make an effort to not only utilize my statistical and computation skills, but also learn more about climate science and forest ecology.

Introduction

The objective of this report is to describe and perform an experiment where climate reanalysis data is combined with observed forest data to predict tree species distributions in Colorado forests. The report addresses whether or not reanalysis climate model data describe the spatial distribution of tree species, and which species are most impacted by climate related variables. The impetus for the experiment arises from research described in the review article by Tinkham et al. (2018). The authors provide background on the United States Forest Service Forest Inventory and Analysis (FIA) and the application of FIA data to a wide range of topics and research. Their section on climate applications discusses how the spatial and temporal ranges of the program allows for monitoring climate related forest issues. Topics include describing shifts in species distribution relating to climate change, climate change modeling applications, and utilizing FIA data to validate outputs of models [18].

The FIA database contains one of the word's largest ecological data sets in terms of spatial and temporal extent lending itself to monitoring climate-related forest issues. Briefly, the FIA program conducts multiple phase inventories and stratified estimation method to estimate population parameters. Phase I remotely senses to stratify the population area by determining land use, and in Phase II permanent grouped plots are randomly distributed. Ecological integrity of plot locations is preserved by "fuzzing" and "swapping" the coordinates of each plot, thus the publicly available coordinates are generally within 1-km of the actual plot location [18]. This report takes inspiration from a series of articles describing climate change modeling applications. These articles generally use, at a minimum, the following three components when generating a database: tree data (either plot level or aggregated), soil data (describing physical and chemical properties) and climate data (typically historical observations). Modeling of distribution or abundance is typically performed with a regression/classification tree or less often an ensemble method like random forests. [1] [6] [7] [9].

Jupyter Intelligence has provided competitors in the 2021 ASA ENVR competition access to three different climate model data-sets. The ERA-Interim reanalysis data produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) is used in this experiment. The ERA-Interim reanalysis utilizes a data assimilation technique, where observations are combined with prior information from a forecast model estimating the changing state of the atmosphere [4]. The reanalysis product produces a consistent and convenient *maps without gaps* atmospheric record for the entire world [5]. Competitors are provided with a spatial snapshot of the ERA-Interim data over the continental United States with daily values of maximum temperature, minimum temperature, and precipitation from 1979 to 2017. The *maps without gaps* concept motivates incorporation of reanalysis data for ecological modeling described in this experiment.

Materials and methods

Examining the impact of reanalysis climate variables on tree distribution requires wrangling data from two disparate databases: the USFS FIA database and the provided ERA-Interim data in a `netCDF` file format. R software is used for the data-wrangling, in addition to the `tidyverse` package for data table manipulation and plotting, and `tidync`, `RNetCDF`, and `ncmeta` packages are used for wrangling climate data [13], [14], [16], [17], [20].

Obtaining Colorado forest data

Plot level data were obtained using the `rFIA` package to download and load forest data for the stat of Colorado from the FIA data-mart [15]. Tree data was obtained by combining the `PLOT` and `TREE` tables for observations since 2002 and with a status code indicating a measurement was taken on a live tree. Counts of species were tabulated to determine the most abundant species in Colorado forests shown in Table 1. The most frequently occurring tree species are selected as candidates for modeling distribution and abundance. The count of species is dichotomized into a $\{0, 1\}$ binary variable indicating the absence or presence of a tree species at a plot location. The count response is retained for abundance estimation.

The FIA database provides soil data measured at various plot locations including physical and chemical properties. Data were compiled at plot locations with complete observations. Physical measurements include forest floor thickness, litter thickness, density, and depth to a restricted layer. Chemical properties include measures of amount of inorganic and organic carbon, and amount of nitrogen. The soil data contains many missing observations and does not exist at every tree plot location. Plots without direct soil measurements are imputed to their nearest neighbor with a full record based on the distance given by the coordinates for each plot.

Table 1. Ranking the top five most frequently surveyed trees in Colorado

Rank	Common Name	Scientific Name	Percent Occurance
1	Engelmann Spruce	*Picea engelmannii*	16.7
2	Quaking Aspen	*Populus tremuloides*	16.6
3	LodgepolePpine	*Pinus contorta*	12.4
4	Gambel Oak	*Quercus gambelii*	10.2
5	Subalpine Fir	*Abies lasiocarpa*	8.9

Interpolation of ERA-Interim data

The ERA-Interim data is utilized to understand the relationship with reanalysis climate variables on forest ecology. The spatial resolution of of ERA-Interim is approximately 80 kilometers necessitating development of an interpolation scheme relating the tree plot locations with climate data. A set of 13 climate related features was developed using the ERA-Interim temperature and precipitation data. A majority of the features were developed based on similar prior studies from Iverson et al. (1998), Jiang et al. (2014). For example both articles considered mean annual temperature, mean January temperature, and annual precipitation, to name a few, when modeling abundance in the eastern United States. Jiang et. al. considered over a dozen climate related variables. The entire time span of the provided climate data is summarized into estimates of mean and variance at each coordinate. Extra features were added to model the variability due to observed non-constant variability over Colorado seen in residual diagnostics of the models developed to smooth/interpolate climate data over the state. Table 2 describes the climate variables used and aggregation method.

Table 2. Description of climate features used for modeling

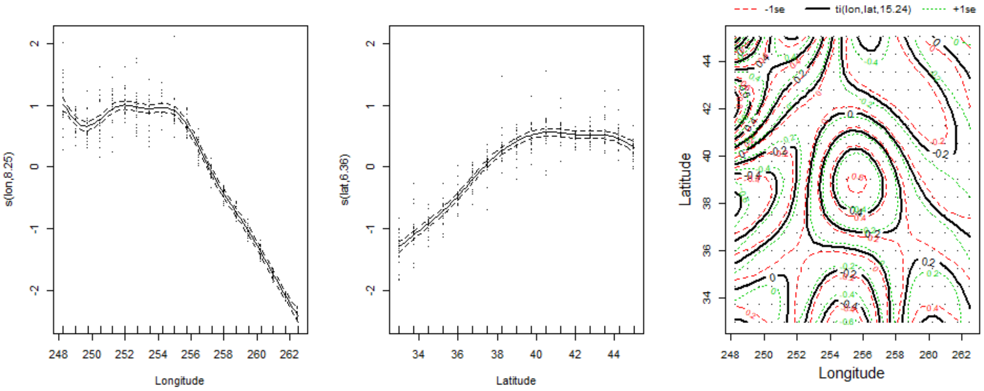
Variable Description	Abbreviation	Time Window	Aggregation Method
Average Temperature	AVGT	Jan - Dec	mean of max and min temps
Winter Temperature	JANT	Jan, Feb	mean of max and min temps
Summer Temperature	JULT	July, Aug	mean of max and min temps
Temperature Variation	VARA	Jan- Dec	mean of max and min temsp
Summer Temperature Variation	VARS	June, July, Aug	variance of daily temps
Winter Temperature Variation	VARW	Dec, Jan, Feb	variance of daily temperature
Difference in Summer High and Low Temp.	DIFS	June, July, Aug	mean difference in high an low temps
Difference in Winter High and Low Temp.	DIFW	Dec, Jan, Feb	mean difference in high and low temps
Cumulative Precipitation	PPT	Jan - Dec	sum of precipitation
Summer Cumulative Precipitation	PPTS	June, July, Aug	sum of precipitation
Winter Cumulative Precipitation	PPTW	Dec, Jan, Feb	sum of precipitation
Precipitation Variation	VARP	Jan - Dec	variance of precipitation

A wide variety of geo-spaital smoothing tools exist, however an approach using generalized additive models (GAMs) is developed in this experiment. Interpolation of a climate related variable f is obtained by spline smooths s of of the spatial variables shown in Eq (1), including main effects terms and an interaction term.

$$f(\text{lon}, \text{lat}) = s_1(\text{lon}) + s_2(\text{lat}) + s_3(\text{lon}, \text{lat}) + \epsilon \quad (1)$$

Models are fit using the `mgcv` package of Wood et al. (2011). The `mgcv` package offers a rich ecosystem of model specifications for fitting GAMs including a wide range of response families and spline types. The main effects, s_1 and s_2, are fit with a thin plate regression spline, the interaction term, s_3, is fit with a tensor product interaction and cubic regression spline, and error term $\epsilon \sim N(0, \sigma^2)$. GAMs provide the analyst with the ability to observe the marginal effects for each smooth and perform residual diagnostics assessing model fit and behavior [21]. Figure 1 displays the marginal effects with residuals modeling the difference in summer high and temperature, which has the strongest association with lodgeploe pine distribution.

Fig 1. GAM model of difference in summer temperature. Marginal effect smooths for the average difference in summer high and low temperatures.



While generally satisfactory marginal effects and residual plots are observed for the climate variables, most models show some degree of heteroskedastic residuals. Inspecting plots of the marginal effect helps to characterize the climate variables behavior across Colorado. The first panel of figure 1 suggests non-constant variance with larger variability in the northwestern part of Colorado. The variability variables were introduced to examine if spatial variability has an effect on tree species distribution, or if the mean estimate provided by the GAM model is sufficient. See the supporting information for a document containing the GAM model results for all climate variables.

Modeling Methods

Regression tree models are a popular modeling technique among the previously discussed articles producing predictions of species distribution or abundance [7] [9] [1]. Regression trees are a flexible modeling technique that produce easily visualized and interpretable models. Regression trees are constructed by recursively splitting the feature space into different regions relative to the mean of the response in that region. The main disadvantage of single tree modes is their lack of robustness or high variance, however ensemble methods aim to ameliorate such problems, but sacrifice model interpretation [8]. Random forests is an ensemble method used by Weiskittlel et al. (2011) to model forest productivity in the western United States [19].

Bagging, random forests, and boosting are popular ensemble methods often applied to regression tress. Bagging repeatedly samples from the training data to train an instance of the model with a sub-sample, then averages over all the models/predictions to get an aggregated estimate. Random forests improves on this procedure by decorrelating the trees by considering a random subset of predictors to grow each tree in the ensemble. Boosting uses an ensemble of trees but goes further by incorporating information from previous trees to grow subsequent trees. This additional information is obtained by fitting a tree to the residuals, thus the construction of subsequent trees are fit with a modified version of the original data set [8]. The increased accuracy with boosting, especially gradient boosting, comes at a cost. Gradient boosting can be slow;

however improvements have been made in recent years with XGBoost.

This experiment uses a popular machine learning modeling framework called XGBoost. Developed by Chen et al., XGBoost was built to be a highly scalable framework, improving computations associated with boosting methods. This algorithm, and its associated open-source package, has become a popular modeling tool among researchers and even among 'data-science enthusiasts' on the Kaggle competition platform. In this experiment gradient boosting was initially considered but the slow computation times made progress almost intractable when considering grid search optimization for hyperparameters. The XGBoost framework provides comparable results roughly 10 to 20 times faster than gradient boosting and allows for a wider specification of response families and tuning/optimization hyperparameters [2].

Both gradient boosting and XGBoosting minimize a regularized objective function, $\mathcal{L}$ to the data. The objective function is $\mathcal{L} = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k)$, where $l(\hat{y}_i, y_i)$ is the loss function and $\Omega(f_k)$ is the regularization term, i.e. a penalization on the complexity of of the model. The `xgboost` library allows users to choose from a wide rage of loss functions depending on the structure of the data and sought after predictions. Classification type problems, like predicting the absence/presence of tree species, use the binary cross-entropy function, i.e.the log-likelihood of a Bernoulli random variable: $l(\hat{y}_i, y_i) = y_i \log(\hat{y}_i + (y_i - 1)(\log(1 - \hat{y}_i)$. Given the proportion of tree species in Colorado, described in Table 1, an estimation of abundance for a particular species results in data with a high degree of zero-inflation. While most examples arise from the insurance industry, the Tweedie distribution accommodates data with zero-inflation. The log-likelihood (up to a constant) of a Tweedie model is $l(\hat{y}_i, y_i) \propto y_i \frac{\hat{y}_i^{1-p}}{1-p} + \frac{\hat{y}_i^{2-p}}{2-p}$, where p is the Tweedie variance power with $p \in (1, 2)$ represents a compound Possion with non negative mass at zero [22] [3].

Results

Boosted models were fit using the `xgboost` package in R (Cite XGboost). The data was standardized (scaled and centered) then split into an 80% training set and 20% testing set. Five different data-sets were generated for the five tree species. The depth of the boosted tree d and learning rate η model parameters were optimized via a grid search on the training set. A routine was written utilizing the `xgb.cv` function to perform 10-fold cross-validation for each pair of candidate parameters for each species of tree. Table 3 summarizes the results of the models with the top five most important variables, the best hyperparameters, and predictive metrics. The train/test accuracy and F1-Scores (to account for imbalance in the data) were computed on the held-out test set. Final models were fit using all of the data and the optimum hyperparameters.

Likely due to the somewhat small sample size ($n = 6922$) and the imbalance of the data over-fitting is an issue. Table **??** shows a common theme where training accuracy slightly exceeds the testing accuracy, The degree of over-fitting was managed by favoring models with small tree depths d and by manually changing parameters in the XGBoost fitting routine. Additional parameters can be included to control over-fitting. The `xgboost` documentation states increasing the maximum delta step parameter helps to make the update step more conservative and can help in classification problems where the classes are extremely imbalanced. Increasing the scaling on the positive weights is also helpful for unbalanced cases. In this experiment these parameters are set to 2 and 1.5 respectively during the cross-validation and final model fitting steps to regulate over-fitting issues.

Ensemble methods do permit some degree of model interpretation by assessing variable importance. Table 3 describes the top five variables with the highest gain indicating their degree of importance. It is perhaps no surprise that elevation plays a role among all five species, in particular englmann Spruce are almost entirely depended on elevation and quaking aspen to a lesser extent. Soil factors play a very marginal role and do not have a high gain/impact on the predictive models. Lodgepole Pine is the most dependent on the climate variables: difference in summer high and low temperatures and difference in winter high and low temperatures have high gain. Both elevation and annual variation in temperature are associated with subalpine fir distributions.

Table 3. Summary of the model fits. ELEV is an abbreviation for Elevation obtained from the FIA database. The climate abbreviations can be found in Table 2

	Quaking Aspen		Engelmann Spruce		Lodgepole Pine		Gambel Oak		Subalpine Fir	
Variable Rank	Variable	Gain	Variable	Gain	Variable	Gain	Variable	Gain	Variable	Gain
1	ELEV	0.739	ELEV	0.920	DIFS	0.412	ELEV	0.656	ELEV	0.600
2	JANT	0.050	JULT	0.017	DIFW	0.200	DIFS	0.085	VARA	0.175
3	VARA	0.028	VARA	0.016	ELEV	0.198	VARS	0.077	JANT	0.066
4	PPTS	0.028	VARW	0.015	JANT	0.031	PPTW	0.050	DIFW	0.046
5	VARS	0.019	JANT	0.015	VARW	0.031	VARA	0.024	AVGT	0.017
Depth (d)	4		3		3		3		3	
Learning Rate (η)	0.3		0.3		0.5		0.3		0.3	
Prob. Threshold	0.5		0.35		0.25		0.35		0.35	
CV Train AUC	0.904		0.959		0.964		0.921		0.937	
CV Test AUC	0.880		0.954		0.950		0.908		0.925	
Train Accuracy	0.812		0.870		0.880		0.869		0.828	
Test Accuracy	0.796		0.853		0.874		0.855		0.835	
Train F1-Score	0.855		0.899		0.924		0.912		0.882	
Test F1-Score	0.840		0.885		0.920		0.903		0.885	

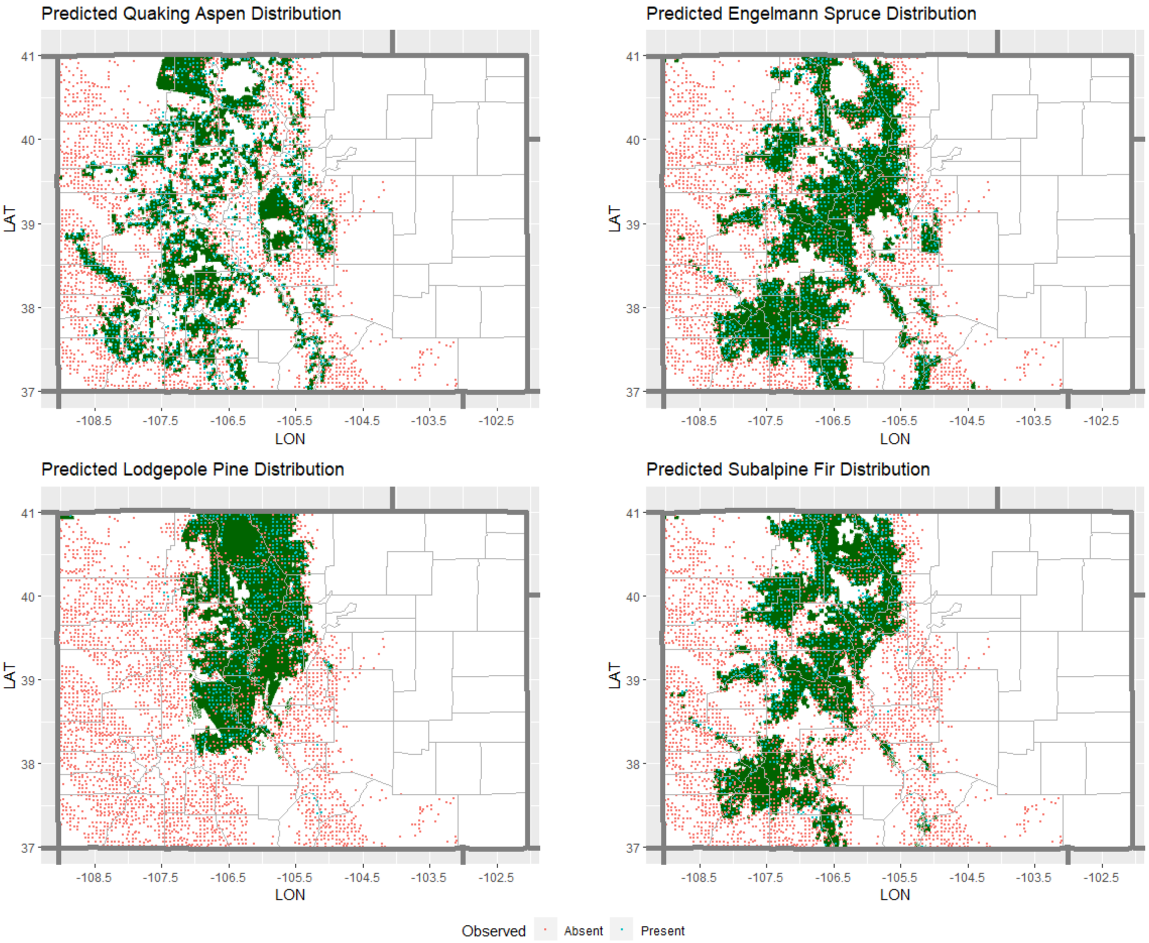
Figure 2 provides a visualization of for four of the five tree distributions across the state of Colorado. The prediction grid was generated by applying the same soil interpolation scheme and the values of the climate related variables used the GAMs to predict the value of each feature on a uniform grid given by latitude and longitude. The prediction grid was applied to each of the tree distribution models to generate a raster image of the distributions. Gambel oak predictions are not included due to a high degree of extrapolation issues, notably in the San Luis Valley oak trees are predicted there when no observations are present. The plot is included with the supplementary material; see the supporting information.

Distribution of Lodgepole Pine and Subalpine Fir under different climate scenarios

The results from Table 3 show Lodgepole Pine has the strongest dependence on climate related variables, notably the difference in summer high and low temperatures DIFS. Using the trained model, a hypothetical scenario is presented where the difference in summer high/low temperatures is varied while holding all other variables fixed. For example, a decrease in DIFS could arise from less variable summer high an low temperatures, and an increase could be caused from higher high temperatures and lower low temperatures.

Figure 3 reflects a one-standard deviation increase and decrease in DIFS across all plot locations reflecting a 0.32 degree Kelvin change across the entire prediction grid for DIFS. A decrease in DIFS is associated with a retreat in pine species from mostly the

Fig 2. Predicted Tree distributions for four species. Raster plot image of predicted distributions in green shading for quaking aspen, engelmann spruce, lodgepole pine, and subalpine fir species.


Predicted Quaking Aspen Distribution
Predicted Engelmann Spruce Distribution
Predicted Lodgepole Pine Distribution
Predicted Subalpine Fir Distribution
LAT
41
40
39
38
37
LON
-108.5
-107.5
-106.5
-105.5
-104.5
-103.5
-102.5
Observed
Absent
Present

northern part of the state, and an increase in DIFS is associated with an overall westward and southern expansion of pine.

Figure 4 is produced in the same way as 3. Subalpine fir behaves similarly to lodgepole pine: a decrease in VARA (annual variation in temperature) produces a retreat in the distribution and an increase produces and expansion. The scenario where VARA increases produces a fir distribution that resembles the predicted Engelmann Spruce distribution, which is almost solely dependent on elevation.

Estimating Abundance

Recall the number of live trees on each plot location is retained in the original data-set. We initially dichotomized the counts into a variable indicating presence and absence to estimate the distribution, but given the live tree count an estimate of abundance can be produced, namely the number of trees per plot area. For example, approximately 16% of the observations for pine are non-zero, so a model ought to account for this high degree of zero-inflation. A model of this form would predict both the abundance and the spatial distribution. The XGBoost package allows for fitting data with a Tweedie response. Table 4 lists the top five most important variables in a model with a Tweedie

Fig 3. Predicted Lodgepole Pine distribution under change in difference in summer high/low temperature. A decrease in DIFS reflect a retreat of pine and increases reflects and expansion.

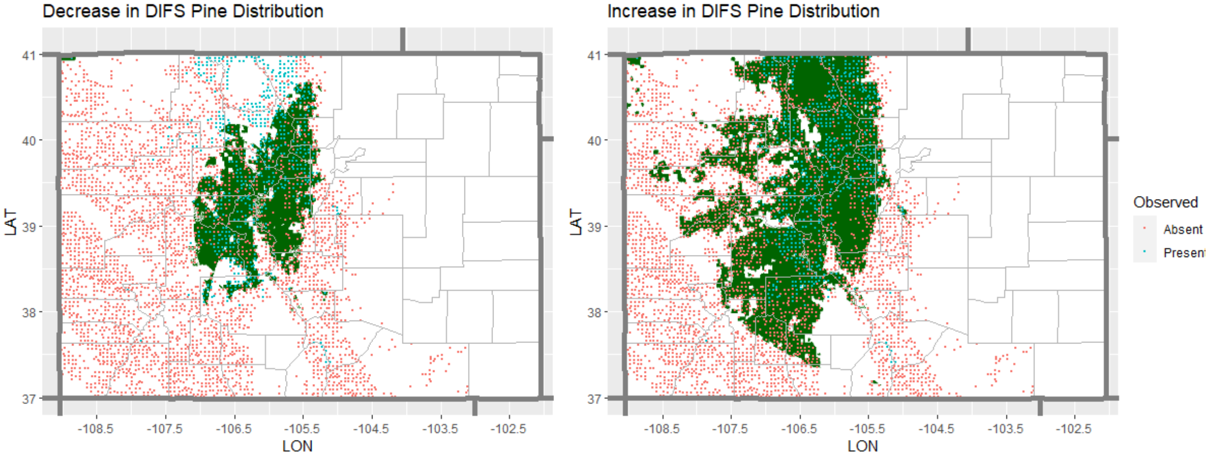
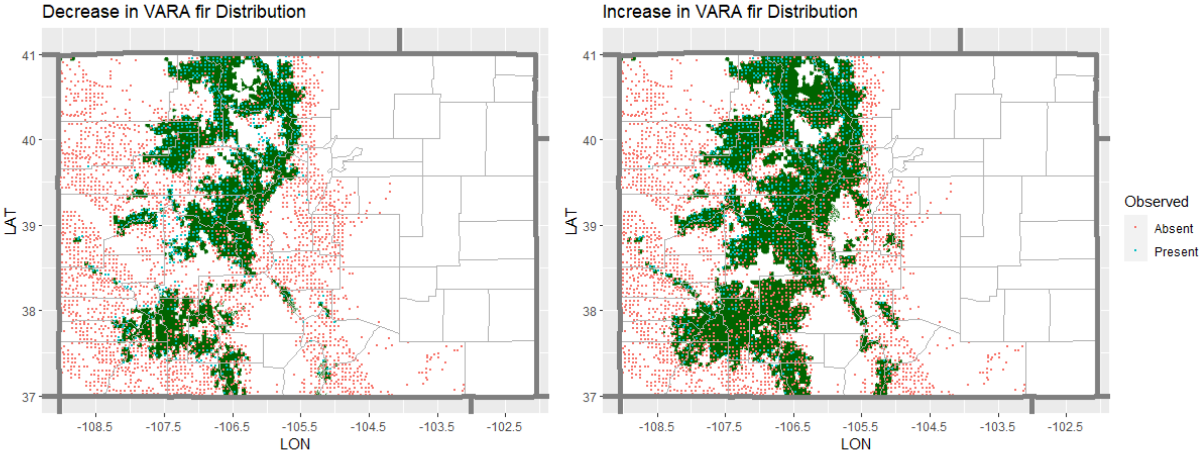


Fig 4. Predicted Subalpine Fir distribution under change in annual temperature variation. A decrease in VARA reflect a retreat of fir and increases reflects and expansion.



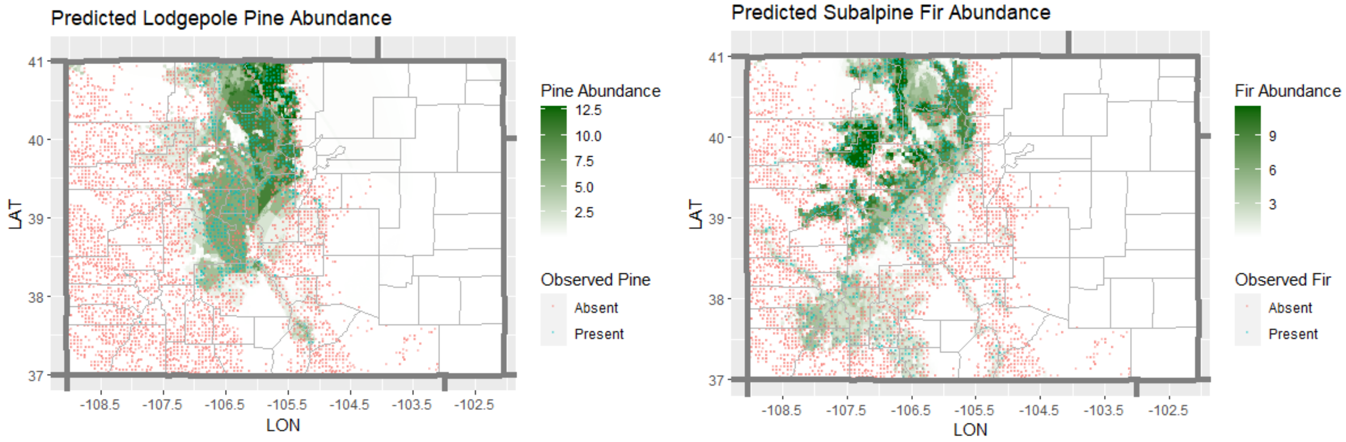
loss function.

Table 4. Variable importance for abundance estimation

	Lodgepole Pine		Subalpine Fir	
Variable Rank	Variable	Gain	Variable	Gain
1	DIFS	0.468	ELEV	0.638
2	ELEV	0.220	VARA	0.075
3	PPTS	0.073	AVGT	0.073
4	DIFW	0.057	JANT	0.063
5	PPTW	0.049	DIFW	0.054

For lodgepole pine the DIFS variable is still the most important and its associated gain increased slightly, elevation plays a more important role, and precipitation becomes slightly more important. Mean square training error for pine abundance is 153.1 and mean square test error is 179.5, still indicating an overfitting problem. Subalpine fir has a stronger dependence on elevation, and a decreased dependence on the annual variation in temperature, and an increased dependence on the average annual temperature. The mean square training error for fir abundance is 36.7, and mean square test error is 47.1. Figure 5 shows raster image plots of the predicted abundance of lodgepole pine and subalpine fir in Colorado forests. The predictions depict both the overall spatial distribution of each species and the number of trees.

Fig 5. Predicted Abundance of Lodgeple Pine and Subalpine Fir. In addition to examination of the distribution, the abundance for the tree species can be predicted.



Discussion

The preceding analysis demonstrates that combining ERA-Interim FIA data is not unreasonable. An ensemble method like `xgboost` and its associated library provides reasonable predictions for estimating spatial distribution and abundance. As discussed previously, a common problem of over-fitting models is present, despite changing some of the parameters in the `xgboost` fitting routine to regularize the degree of over-fitting. Some species depend almost entirely on elevation and their distributions are not associated the climate and soil related variables. It is not clear if this is a result of certain species being highly dependent on elevation or a shortcoming with the data and model. Over-fitting could be resolved by either performing a grid-search on the auxiliary parameters that control over-fitting, or considering a larger spatial distribution and/or aggregating plot observations. Iverson et al. (1998) consider the entire East-Coast of the United States and aggregate plot level information from the FIA database to the county level [6] [7].

Certain species may not be suitable for modeling as described in this report. Both the observed and predicted quaking aspen distribution, shown in Figure 2, have a very discontinuous shape, high degree of overfitting, and poor model performance. This is likely due to the aspen growing singly or in multi-stemmed clones with large clones located in the Rocky Mountains. Moreover, in Colorado and Wyoming, quaking aspen grow in a narrow range of elevations with precipitation ranging from 410-1020mm [12]. The model examined here has very weak association with summer precipitation. Instead of modeling with plot-level data, plots could be aggregated and a proportion of aspen trees determined at the county level, for example. More domain knowledge likely needs to be integrated into the model for Gamble Oak distribution to obtain reasonable predictions.

Lodgepole pine is the only species that does not have elevation ranked first in variable importance. The actual presence/absence observation for pine, shown in Figure 2, are located in the central part of the state, north of 38° N. The grouping of pines is associated with DIFS conditions, illustrated by studying Figure 1. Given the mountainous topography of Colorado, the amount of snowfall may play a role. Both species tend to favor the central and norther parts of the state with mountains and snow. Lodgepole pine are found on the interior of North America where seasonal distribution of precipitation is important; namely locations where snowfall supplies water for rapid growth in early summer and when temperatures are favorable for germination after snow-melt [11]. This also suggests some degree of temporal dependence on lodepole pine behavior.

It may be useful to develop proxy variables for elevation. Coops et al. (2009), consider a set of modifiers modeling the geographic variation in the maximum effect on photosynthesis relative to Douglas fir stands, which include a soil/water modifier, temperature modifier, frost modifier and vapor pressure modifier. They find spatial distribution of tree species depends on these modifiers using regression tree analysis [1]. A similar approach could be taken in the state of Colorado calculating modifiers to engelmann spruce as the reference species due the extent of it's distribution throughout the state.

Conclusion

This report provides a demonstration of combining ERA-Interim climate data with FIA data to model distribution and abundance of tree species in Colorado forests. Generalized additive model were used to smooth and interpolate climate data to FIA plot locations. Lodgepole pine has the strongest relationship with the average difference in summer temperature. A method for abundance estimation (number of trees per plot) was developed using a Tweedie distribution. Although the current experiment shows symptoms of over-fitting, expanding the area of interest to a broader region may further elucidate the relationship between climate and forests.

In addition to expanding the area of interest, different climate models could be examined. One of the other data-sets provided by Jupyter Intelligence is a climate model developed by Liu et al. (2016). Their climate model grew out of research examining the water cycle over the headwaters of the Colorado River to obtain a more precise determination of snowfall and snow pack in areas with varied topography [10]. This may be important for understating the distribution and abundance of species that are known to be dependent on precipitation or spring/early summer snow-melt, such as lodgepole pine. Unfortunately, due to computational limitations, this data could not be examined. Future work could move toward generating a database and modeling strategy to incorporate temporal information. Deep learning methodology, such as recurrent neural networks, could forecast tree distributions and abundance in American forests with climate models as input features.

Supporting information

Supplement. Lorem ipsum. Supplementary plots are avilable here: https://github.com/jddrumheller/Colorado-Forests/blob/main/Forest_Supplement.pdf and a GitHub repository with code and and data set is avilable here: https://github.com/jddrumheller/Colorado-Forests

Acknowledgments

I'd like to thank Phong Nguyen, who had to drop out of the competition, for his help with managing the netCDF files early on in the project.

References

1. Coops, Nicholas C., Waring, Richard H., Schroder, Todd A. (2009). Combining a generic process-based productivity model and a statistical classification method to predict the presence and absence of tree species in the Pacific Northwest, U.S.A., Ecological Modelling, Volume 220, Issue 15,2009,Pages 1787-1796, https://doi.org/10.1016/j.ecolmodel.2009.04.029.

2. Chen, Tanqui and Guestrin, Carlos (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 785–794. DOI:https://doi.org/10.1145/2939672.2939785

3. Chen, Tanqui, and He, Tong, and Benesty, Michael and, et al. (2021). xgboost: Extreme Gradient Boosting. R package version 1.3.2.1. https://CRAN.R-project.org/package=xgboost

4. Dee, D.P., Uppala, S.M., Simmons, A.J., Berrisford, P., et al. (2011). The ERA-Interim reanalysis: configuration and performance of the data assimilation system. Q.J.R. Meteorol. Soc., 137: 553-597. https://doi.org/10.1002/qj.828

5. European Center for Medium-Range Weather Forecasts. (n.d.). Climate Reanalysis. https://www.ecmwf.int/en/research/climate-reanalysis

6. Iverson, Louis R. and Prasad, A.M. (1998). Predicting abundance of 80 tree species following climate change in the eastern United States. Ecological Monographs, 68: 465-485. https://doi.org/10.1890/0012-9615(1998)068[0465:PAOTSF]2.0.CO;2

7. Iverson, Louis R., and Parasad, Anatha, and Schwartz, Mark W. (1998). Modeling potential future individual tree-species distributions in the eastern United States under a climate change scenario: a case study with Pinus virginiana, Ecological Modelling, Volume 115, Issue 1, 1999, Pages 77-93, https://doi.org/10.1016/S0304-3800(98)00200-2.

8. James, Gareth, and Witten, Daniela, et al. (2017). An Introduction to Statistical Learning: with Applications in R. New York, Springer.

9. Jiang, Huiquan and Radtke, Philip J. and Weiskittel, Aaron R., et al. (2014). Climate- and soil-based models of site productivity in eastern US tree species. Canadian Journal of Forest Research. 45(3): 325-342. https://doi.org/10.1139/cjfr-2014-0054

10. Liu, C., Ikeda, K., Rasmussen, R. et al. Continental-scale convection-permitting modeling of the current and future climate of North America (2016). Clim Dyn 49, 71–95 (2017). https://doi.org/10.1007/s00382-016-3327-9

11. Lotan, James E. and Critchfield, Willimam B. (n.d.) Lodgepole Pine. USDA-FS. https://www.srs.fs.usda.gov/pubs/misc/ag_654/volume_1/pinus/contorta.htm

12. Parala, D.A. (n.d.). Quaking Aspen. USDA-FS. https://www.srs.fs.usda.gov/pubs/misc/ag_654/volume_2/populus/tremuloides.htm

13. Michna, Pavel and Woods, Milton (2020). RNetCDF: Interface to 'NetCDF' Datasets. R package version 2.4-2. https://CRAN.R-project.org/package=RNetCDF

14. R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

15. Stanke, H., Finley, A. O., Weed, A. S., et al. (2020). rFIA: An R package for estimation of forest attributes with the US Forest Inventory and Analysis database. Environmental Modelling & Software, 127, 104664.

16. Sumner, Michael (2020). ncmeta: Straightforward 'NetCDF' Metadata. R package version 0.3.0. https://CRAN.R-project.org/package=ncmeta

17. Sumner Michael (2020). tidync: A Tidy Approach to 'NetCDF' Data Exploration and Extraction. R package version 0.2.4. https://CRAN.R-project.org/package=tidync

18. Tinkham, Wade T. and Mahoney, Patrick R. and Hudak, Andrew T., et al. (2018). Applications of the United States Forest Inventory and Analysis dataset: a review and future directions. Canadian Journal of Forest Research. 48(11): 1251-1268. https://doi.org/10.1139/cjfr-2018-0196

19. Weiskittel Aaron R., Crookston Nicholas L., and Radtke Philip J. (2011). Linking climate, gross primary productivity, and site index across forests of the western United States. Canadian Journal of Forest Research. 41(8): 1710-1721. https://doi.org/10.1139/x11-086

20. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

21. Wood, S.N. (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. Journal of the Royal Statistical Society (B) 73(1):3-36

22. Yang, Yi and Quian, Wei and Zou Hui (2016). Insurance Premium Prediction via Gradient Tree-Boosted Tweedie Compound Poisson Models. Preprint on arXiv.org. https://arxiv.org/abs/1508.06378