**Methods**

*Study Area Management and Fire History*

*Data Sources*

Historical data source. To make historical and modern datasets comparable, we excluded quarter-quarter sections considered by Forest Inventory and Analysis (FIA) to be “non-forested” (total live basal area < 9 m2 ha-1; n = 104), totaling 1,379 individual quarter-quarter sections in our final historical dataset.

The scope of inference for current forest conditions (Figure 1) is approximately 788,000 ha (7880 km2) centered primarily on western Plumas county, a bounding box we generated based on where we believed FIA data would closely resemble the topographic and climatic characteristics of the historical dataset. Due to FIA inventories being conducted on 10-year rotations, we identified FIA plots that were surveyed between 2011 to 2018 and fell within the footprint defined above (n = 274). We then conducted a series of filtering exercises to select plots that would serve as an appropriate comparison to our historical data. This included selecting plots within the elevation range of the historical data (1113 – 1923 m; n = 204) and removing plots where the reported elevation was markedly different than what was extracted from a digital elevation model based on plot coordinates. To account for this public “fuzzing” of coordinates, we only retained plots that were within 150 m of elevation discrepancy (n = 182). We then selected plots recorded with one “condition” only, indicating that the entirety of the plot could be described as having similar ages, species composition, and disturbance history (n = 110). We also removed an additional 29 plots due to a history of fire activity (n = 81) and 1 plot where stand age was recorded as zero (only seedlings present; n = 80). Since we did not expect the historical timber surveys to contain data pertaining to hardwoods, we also excluded 3 plots where the dominant vegetation was categorized as California black oak and 1 plot categorized as tanoak (n = 76). We also excluded hardwoods from all estimates of tree density and basal area. All plots we analyzed had to contain trees at least 30.5 cm (12.0 in) in diameter at breast height (DBH) to match the DBH cut-off for historical inventories and have at least 9 m2 ha-1 of live basal area to be considered “forested.” This filtering approach totaled 71 plots in our final modern dataset for analyses.

For both the 1,379 lots in the historical dataset and the 71 plots in the modern FIA dataset, we extracted underlying climate and topographic data. A digital elevation model was acquired from the US Geological Survey National Elevation Dataset (https://viewer.nationalmap.gov/) at 1/3 arc-second (8 x 10 m) resolution and converted to corresponding slope and aspect layers using QGIS. We converted aspect to a categorical variable with breakpoints at 135° and 315° to correspond to northeast-facing and southwest-facing slopes. Climate data were acquired from the Basin Characterization Model dataset (Flint et al. 2013) for the same geographic footprint at 270 m resolution. Climate variables included 30-year mean values from 1981-2010 for maximum annual temperature, maximum June temperature, annual climatic water deficit, June climatic water deficit, actual evapotranspiration, April 1 snowpack, and annual precipitation. For spatial predictive models, all data were scaled and aligned at 270m resolution.

*Data Analysis*

An initial set of all seven climatic variables, elevation, slope, and aspect was considered to explain the variation in historical forest conditions including trees ha-1 (TPH) , live basal area (m2 ha-1), and pine fraction (calculated as the live basal area of *Pinus* spp. divided by the total live basal area of a given plot). Multicollinearity amongst explanatory variables was reduced by removing variables with a Pearson’s correlation coefficient greater than 0.7 (Appendix A). This threshold resulted in a final candidate set of 5 variables, including maximum annual temperature, slope, aspect, annual climatic water deficit (CWD), and annual precipitation. While we also evaluated higher and lower thresholds of correlation to test the number of parameters included in our models, we did not find any substantial improvements to our final models.

We then input our reduced number of predictor variables into a random forest model using the *randomForest* package in R (Liaw and Wiener 2002; R Core Team 2020) to predict which variables were the most important in explaining historical forest conditions. Random forest is a machine learning algorithm that aggregates bootstrapped estimates of multiple decision trees, which leads to greater accuracy and lower error rates relative to traditional linear regression models (Povak et al. 2014). Similar to methods established by Povak et al. 2014, we started with all five predictor variables in the same random forest model for each forest condition. Based on the percentage increase in mean standard error, we removed the least important variable from each model and re-ran random forest. We repeated this process until only two variables remained in each model. We selected the “best” performing model predicting each forest condition based on the greatest percentage of variation explained and lowest root mean standard error (Appendix B). The variables contained within these models were used as inputs in a regression tree analysis using the *rpart* package in R (Therneau and Atkinson 2019) to identify important thresholds in those variables that are associated with the different mean values of historical forest conditions. We used an ANOVA method for splitting variables and a complexity parameter of 0.02 (the increase in R2 value at each split that must occur for the split to be accepted). For the pine fraction response variable, we used a complexity parameter of 0.03 to avoid an overly complex regression tree.

To compare how TPH, live basal area, and pine fraction may differentiate between historical and modern forests given the same environmental conditions, we used the breakpoints identified by the regression tree analysis of our historical data and aggregated the modern FIA dataset according to those thresholds. We then estimated the mean values of each forest structure variable within that environmental space. Finally, we estimated historical TPH, live basal area, and pine fraction at a landscape scale by applying the best random forest model predicting each structure to a 270 m resolution raster dataset containing each model’s associated climatic and topographic variables. To avoid extrapolating beyond the range of the sampled environmental space, we excluded any topographic or climatic values that were not within the environmental envelope of the historical dataset.

**Results**

For all historical forest structure metrics, maximum annual temperature, slope, CWD, and precipitation were predictor variables in the top random forest models. Regression tree analysis of TPH in 1924 suggest a strong influence of maximum annual temperature, precipitation, and CWD (Figure 2a). Temperature was the primary driver of density, with colder sites reaching an annual maximum < 15 °C having average TPH of 37 trees ha-1. Although sites warmer than 15 °C had higher densities than colder sites, the driest sites (annual precipitation < 1019 mm) limited average tree density to 42 tree ha-1. While we observed greater average TPH in wetter sites (precipitation > 1019 mm), the wettest sites also limited tree density (~44 trees ha-1) when annual precipitation was > 1179 mm and CWD was < 517 mm. The highest densities (~49 – 51 trees ha-1) were observed within intermediate levels of wetness when precipitation was > 1179 mm and CWD was > than 517 mm or precipitation was between 1019 – 1179 mm.

Based on breakpoints established in the regression tree analysis of historical TPH, modern forests had 64 – 78 % higher average TPH than historical forests across all environmental conditions (Figure 2b). Unlike the historical dataset, colder and wetter sites had the highest densities when annual temperature was < 15 °C (~ 170 trees ha-1) or when annual precipitation was > 1179 mm and CWD was < 517 mm (~ 181 trees ha-1). Although the driest sites had lower tree density (~141 trees ha-1) when annual precipitation was < 1019 mm, these sites were dissimilar to the historical dataset in that they did not contain the lowest densities. Rather, sites that had greater precipitation combined with greater CWD showed the lowest densities (~136 trees ha-1). Other sites with intermediate levels of wetness (precipitation between 1019 – 1179 mm) showed intermediate levels of tree density (~166 trees ha-1).

A picture containing table, map, various, bunch

Description automatically generated

**Figure 2** Regression tree analysis of tree density (TPH; trees ha-1) in (a) historical and (b) modern FIA datasets. Inset histograms display distribution of density data, with bin colors in histogram corresponding to partitioned boxes in the regression tree. Parameters for top random forest model predicting TPH displayed at the top of the panel.

Fewer environmental predictors influenced historical live basal area (Figure 3a), with maximum annual temperature and CWD being the only variables present in the regression tree analysis. This may be partly due to the narrow range in average live basal area (Table 1) that we observed in our historical dataset, which was generally between 9 and 42 m2 ha-1. Similar to historical tree density, the warmest and driest sites with maximum temperature > 15 °C and CWD > 436 mm showed the highest levels of live basal rea (~18 m2 ha-1). Conversely, the coldest sites (maximum annual temperature < 15 °C) and sites with lower drought stress (CWD < 436 mm) had lower basal area relative to drier sites (~15 m2 ha-1).

Similar to tree density, the modern FIA dataset showed substantial increases in live basal area (50 – 53 %) relative to historical forests across most environmental conditions (Figure 3b). Consistent with trends found in the historical dataset, colder sites with maximum temperature < 15 °C had the lowest basal area (~33 m2 ha-1), while warmer and drier sites with temperature > 15 °C and CWD > than 446 mm had the highest BA (~36 m2). It is important to note that we did not find any plots within the FIA dataset that existed in environments with temperature > 15 °C and CWD < 436 mm. Therefore, we could not compare changes in live basal area between historical and modern forests subjected to those particular environmental conditions.

A picture containing transport, map, boat, table

Description automatically generated

**Figure 3** Regression tree analysis of live basal area (BA; m2 ha-1) in (a) historical and (b) modern FIA datasets. Inset histograms display distribution of live basal area data, with bin colors in histogram corresponding to partitioned boxes in the regression tree. White box in panel b indicates that no samples within the modern FIA dataset met the criteria for that threshold. Parameters for top random forest model predicting live BA displayed at the top of the panel.

**Table 1** Summary of mean (range) forest structure and environmental data used in random forest modeling and regression tree analysis. Environmental data were averaged using long-term values from 1981 – 2010.

|  |  |  |
| --- | --- | --- |
| Forest structure and data | 1924 | 2011 - 2018 |
| Trees ha-1 | 45 (17 - 123) | 160 (25 – 493) |
| Live basal area (m2 ha-1) | 16.5 (9.1 – 42.9) | 34.7 (9.6 – 92.1) |
| Pine fraction | 0.6 (0.03 – 1.0) | 0.3 (0 – 1) |
| Elevation (m) | 1557 (1113 – 1923) | 1610 (1128 – 1920) |
| Max annual temperature (°C) | 15.4 (13.8 – 18.4) | 15.6 (12.8 – 19.6) |
| Annual climatic water deficit (mm) | 498.8 (82.9 – 688.0) | 526.9 (196.7 – 717.6) |
| Annual precipitation (mm) | 1240.0 (859.0 – 1714.8) | 1381.7 (552.6 – 2447.8) |
| Slope (%) | 11 (1 – 31) | 12 (1 – 35) |

Regression tree analysis showed that historical data were strongly characterized as pine dominated forests (pine fraction > 0.50), regardless of environmental condition (Figure 4a). Unlike the other regression tree analyses, all variables were shown as strong drivers of historical pine fraction. In sites where slope was low (< 10%), drier sites with precipitation < 1497 mm had 23% higher pine fraction (~0.71) than sites where precipitation was > 1497 mm (~0.55). When slope was steeper (> 10%), the driest sites with maximum temperature > 15 °C had the lowest pine fraction (~0.53), whereas colder sites with temperature > 15 °C and less drought-stressed with CWD < 476 mm had similar pine fractions (~0.56). Colder sites with temperature < 15 °C had a higher pine fraction (~0.70) when coupled with higher drought stress (CWD > 476 mm).

Pine dominance shifted dramatically in modern forests, with our regression tree analysis on FIA data indicating that pine fraction diminished even under the same environmental conditions as our historical dataset (Figure 4b). Unlike the historical dataset, low slopes were not pine dominated, with wetter sites (precipitation > 1497 mm) showing the lowest levels of pine fraction (~0.13). Drier sites with precipitation < 1497 mm had higher pine fraction (~0.38), but were still 46% lower than historical pine fraction under the same conditions. This trend was consistent in sites with steeper slopes (> 10%). Although the warmest sites where maximum temperature was > 15°C were similar to historical forests in that they showed lower pine fraction (~0.20) relative to the other environmental conditions, these sites were still 62% lower than historical forests. While colder (maximum temperature < 15 °C) and less drought-stressed (CWD < 476 mm) sites showed higher levels of pine fraction (~0.46), the highest levels of pine fraction (~0.56) were apparent in colder sites with higher drought stress (CWD > 467 mm).

A picture containing table, boat, dog, hanging

Description automatically generated

**Figure 4** Regression tree analysis of pine fraction (basal area of Pinus spp. divided by total live basal area of a given plot) in (a) historical and (b) modern FIA datasets. Inset histograms display distribution of pine fraction data, with bin colors in histogram corresponding to partitioned boxes in the regression tree. Parameters for top random forest model predicting pine fraction displayed at the top of the panel.

Based on maps generated from our best-fit random forest models predicting tree density (Figure 5), live basal area (Figure 6), and pine fraction (Figure 7), there were considerable patterns in forest structure across the historical landscape. Higher tree densities appeared to concentrate at the southern portion of the study area, while lower tree densities were more apparent in the northern portion. While there was less variation in the distribution of live basal area across the landscape, this may be expected considering the lack of variation in basal area within the historical dataset. However, the higher ends of predicted live basal area aligned with higher levels of predicted tree density, while the lower ends of predicted live basal area aligned with lower levels of predicted tree density. Our maps showed that pines dominated the historical landscape, with the highest portions concentrated in the northern region of our study area. The spatial distribution of pine also aligned with the distribution of tree density, with a linear regression detecting a negative relationship between tree density and pine fraction (Appendix C; p < 0.001).

**A close up of a map

Description automatically generated**

**Figure 5** Model prediction of historical landscape variation in tree density (TPH; trees ha-1). Predictions were generated from bet-fit random forest model using the historical dataset. Predictors included means of maximum annual temperature, annual climatic water deficit, annual precipitation, and slope. Predictions were only made within the environmental envelope of the historical dataset (sampled QQs) and scaled to 270 m resolution.

A picture containing flower, tree, animal, coral

Description automatically generated

**Figure 6** Model prediction of historical landscape variation in total live basal area (BA; m2 ha-1). Predictions were generated from bet-fit random forest model using the historical dataset. Predictors included means of maximum annual temperature, annual climatic water deficit, annual precipitation, and slope. Predictions were only made within the environmental envelope of the historical dataset (sampled QQs) and scaled to 270 m resolution.

A picture containing tree, flower

Description automatically generated

**Figure 7** Model prediction of historical landscape variation in pine fraction (calculated as the live basal area of Pinus spp. divided by total live basal area of the plot). Predictions were generated from bet-fit random forest model using the historical dataset. Predictors included means of maximum annual temperature, annual climatic water deficit, annual precipitation, and slope. Predictions were only made within the environmental envelope of the historical dataset (sampled QQs) and scaled to 270 m resolution.

**Appendices**

**Appendix A** Correlations between continuous predictor variables including (from top to bottom) elevation, slope, annual climatic water deficit, June climatic water deficit, actual evapotranspiration, April 1 snowpack, annual precipitation, maximum annual temperature, and maximum June temperature. The diagonal displays the distribution of each variable, the bottom diagonal displays the bivariate scatterplots and trendline (in red), and the top diagonal displays Pearson's product moment correlation coefficient with significant correlations indicated for p < 0.001 (\*\*\*).

A close up of text on a white background

Description automatically generated

**Appendix B** Trends in model performance metrics predicting total live BA, pine fraction, and TPH using various number of predictor variables. The top panel displays the percentage of variation explained for a model with a given number of predictor variables and the bottom panel display the root mean standard error (RMSE) associated with that model. RMSE was estimated using the difference between observed values of total live BA, pine fraction, and TPH and predicted values using the random forest model.

A close up of a map

Description automatically generated

**Appendix C** Relationship between predicted tree density (TPH, trees ha-1) and predicted pine fraction from best-fit random forest models. Black dots represent predictions from random forest while blue line represents the model response curve from linear regression predicting the relationship. Significance of relationship and coefficient of determination (r2) are stated on the top right of the figure.

A close up of a logo

Description automatically generated

**References**

Flint, L. E., A. L. Flint, J. H. Thorne, and R. Boynton. 2013. Fine-scale hydrologic modeling for regional landscape applications: the California Basin Characterization Model development and performance. Ecological Processes **2**:1-21.

Liaw, A. and M. Wiener. 2002. Classification and regression by *randomForest*. R News **2**: 18-22.

Povak, N.A., Hessburg, P.F., McDonnell, T.C., Reynolds, K.M. Sullivan, T.J., Salter, R.B., and B.J. Cosby. 2014. Machine learning and linear regression models to predict catchment-level base cation weathering rates across the southern Appalachian Mountain region, USA. Water Resources Research **50**: 2798-2814.

R Core Team. 2020. R: A language environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org.

Therneau, T. and B. Atkinson. 2019. *rpart*: Recursive partitioning and regression trees. R package version 4.1-15. https://CRAN.R-projection.org/package=rpart.