Modeling individual tree mortality in the Sierra Nevada in response to the 2012-2016 California drought

Nicole Hemming-Schroeder[1] (hemmingn@uci.edu), Markelle Kelly[2], Carl Norlen[3], Steven Allison[1,3], and James Randerson[1]
University of California, Irvine, Departments of [1]Earth System Science, [2]Computer Science, and [3]Ecology and Evolutionary Biology

Background

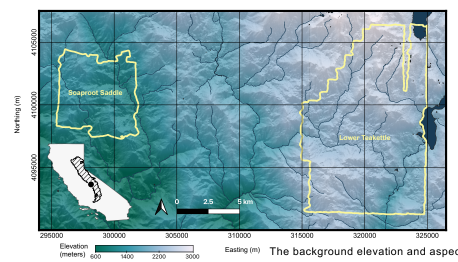
- California experienced a severe drought between 2012 and 2016.
- Tree mortality in some areas of the Sierra Nevada was as high as 50%[1].
- Droughts like this may be more frequent in the future.
- Modeling tree mortality risk may help inform future projections of carbon losses from forests and forest conservation efforts.

Central Question

To what extent can we model individual tree mortality risk in the Sierra Nevada in response to a severe drought using random forests, extreme gradient boosting, and neural networks?

Methods

- We used a set of more than 1 million trees mapped from LiDAR and multispectral data from the National Ecological Observatory Network (NEON) for two sites in the Sierra Nevada for the years 2013, 2017, 2018, 2019, and 2021.



The background elevation and aspect are mapped from NASA SRTM data

- We partitioned the data set into a 60/20/20% split of training, validation, and testing data.
- We chose our target variable to be whether a tree is dead in 2017.
- We tested three machine learning methods: (1) random forests, (2) extreme gradient boosting, and (3) neural networks.
- We resampled the training and validation data to have an even number from each class (live and dead) for each 5th-percentile of tree height to avoid a height-based bias.
- We performed a hyperparameter search for each of the model types.
- For the best model of each type, we shuffled each variable one at a time to quantify its importance.

Feature Variables

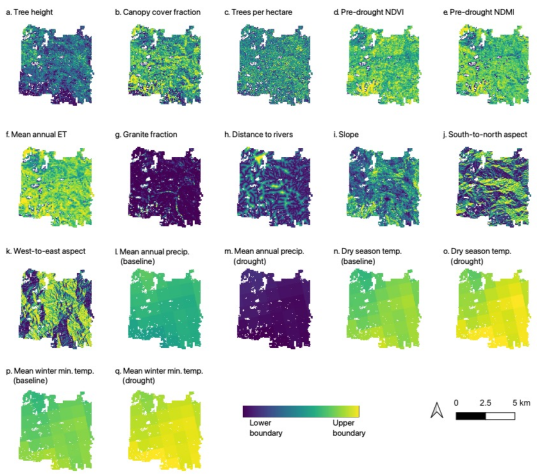


Figure 1. Feature variables for Soaproot Saddle

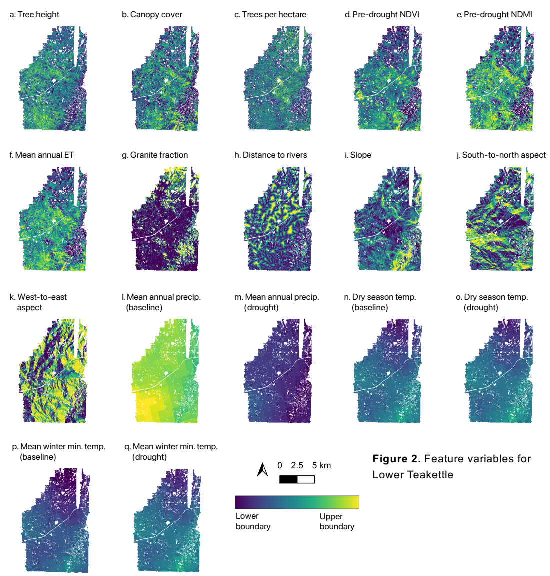


Figure 2. Feature variables for Lower Teakettle

Figures 1 & 2. The feature variables for the tree mortality models are mapped at the two sites to visualize the spatial patterns of each one. Each model is run at the level of individual trees.

- The metrics in **a-c** are derived from 2013 NEON LiDAR point clouds. (We assumed that these metrics have not changed significantly since 2011.)
- Pre-drought normalized difference vegetation and moisture indices (NDVI and NDMI, respectively) in **d-e** are computed from September 2011 Landsat data.
- We computed evapotranspiration (ET) for **f** following the methods of Norlen and Goulden[2] and averaged annual ET from 2009-2011.
- We derived granite fraction for **g** from NEON canopy height model and multispectral reflectance data.
- We calculated distance to rivers between our tree locations and rivers for **h** from the high-resolution National Hydrography Dataset.
- We derived slope and aspect for **i-k** from NEON's digital terrain model.
- Climate variables in **l-q** are derived from 1-km Daymet data.

Results

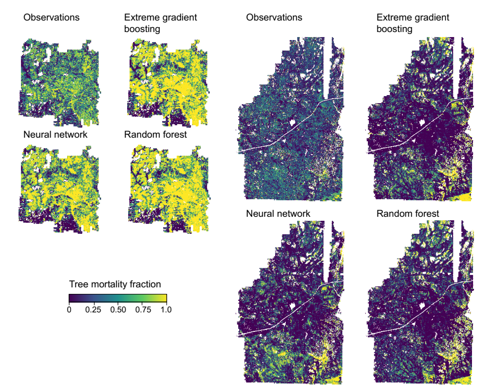


Figure 3. The observed and modeled mortality fraction for Soaproot Saddle (left) and Lower Teakettle (right). Resampling the training and validation data set to avoid a climate-driven bias may help reduce extreme values at both sites.

Training data and validation data	Modeled dead	Modeled live	Class accuracies
Observed dead	228,857	180,810	55.9%
Observed live	83,043	326,632	79.7%

Table 1. Resampled training data confusion matrix. Accuracy on training data: 67.8%

Test data	Modeled dead	Modeled live	Class accuracies
Observed dead	73,998	62,211	54.3%
Observed live	23,132	108,087	82.4%

Table 2. Resampled validation data confusion matrix. Accuracy on validation data: 66.8%

Tables 1 & 2. We trained a model on the training data set and used the validation data set to select the best hyperparameters for each model. We show the training and validation data confusion matrices for the optimum hyperparameters of the best model (extreme gradient boosting) after penalizing for overfitting.

Shuffled Feature	Extreme gradient boosting accuracy (%)	Random forests accuracy (%)	Neural networks accuracy (%)
Tree height	**61.9**	**60.5**	60.1
Mean min. winter temp. during the drought	**62.9**	**52.6**	59.7
Mean annual precipitation during the drought	**65.1**	**63.8**	58.8
Mean baseline dry season temp.	65.4	65.8	**58.0**
Mean baseline annual precipitation	65.8	64.3	**56.8**
NDVI before the drought	66.0	64.3	63.7
Canopy cover	66.5	65.3	64.3
Slope	66.7	64.6	64.7
Distance from rivers	66.7	64.9	65.0
NDMI before the drought	67.2	65.6	63.9
No features shuffled	67.7	65.8	65.8

Table 3. We used our best hyperparameters for each model type and trained a model on the combined resampled training and validation data set. We shuffled the input variables one at a time to obtain the accuracy of the model on the combined resampled training and validation data. We show the top ten feature variables for the best model (extreme gradient boosting). The bold font shows the top three from each model. Not shown is mean dry season temp. during the drought which yielded 57.3% accuracy on the fitted neural network when shuffled.

Key Takeaways

- Extreme gradient boosting performed the best for tree mortality prediction and had an accuracy of 66.8% on the validation data set.
- The most important predictors of individual tree mortality were tree height, mean minimum winter temperature during the drought, and mean annual precipitation during the drought.
- Our next steps include resampling our training and validation data sets to reduce bias among additional feature variables such as mean minimum winter temperature during the drought.

Acknowledgments

This project is supported by Future Investigators in NASA Earth, Space Science and Technology grant 80NSSC20K1626.

References

[1]Hemming-Schroeder et al. (2023), *J. Geophys. Res. Biogeosci.*, 128, e2022JG007234
[2]Norlen & Goulden (2023), *AGU Advances*, 4, e2022AV000810