Machine Learning for Hydrologic Prediction in Water Resources Management Models

Ellie White [1*], Jon D. Herman [1], Marielle C. Pinheiro [2], Jay R. Lund [1]

[1]Civil and Environmental Engineering, University of California, Davis
[2]Atmospheric Science, , University of California, Davis
*Corresponding author: elawhite@ucdavis.edu

1 "We are drowning in data and starving for knowledge" –Rutherford D. Roger

Research Questions:

Q1) Can machine-learning models give better and more cost effective streamflow predictions for drought, flood and water management applications compared to mechanistic models?

Q2) How will including other variables that contribute to flow (e.g., basin topography, location, and soil properties) improve the model's predictive capabilities?

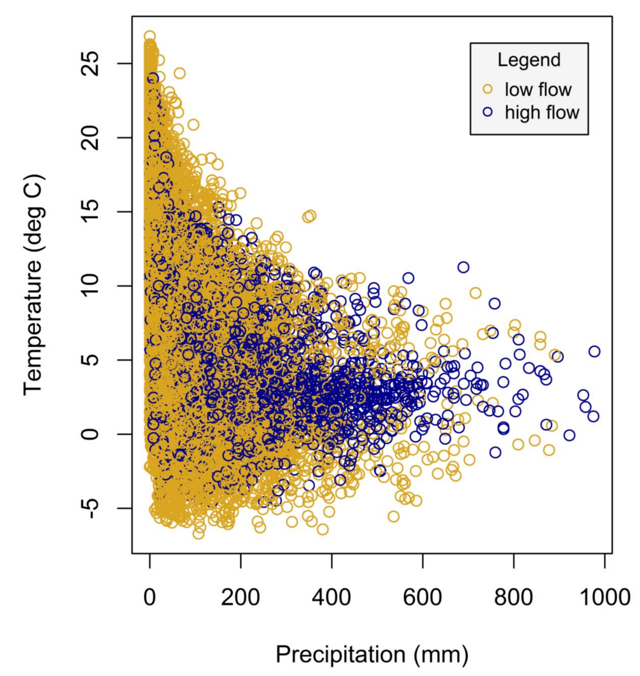


Figure 1. Temperature vs. Precipitation "Cloud". In machine learning, computers apply statistical learning techniques to automatically identify patterns and boundaries in the data.

2 Data sources for predictor and response variables: 69 California catchments

Figure 2. The Response Variable: *Monthly Unimpaired Flow*. Depicted is the data from the California Data Exchange Center (CDEC). It spans 69 California basins from 1982 to 2014, for a total of 18,500 instances.

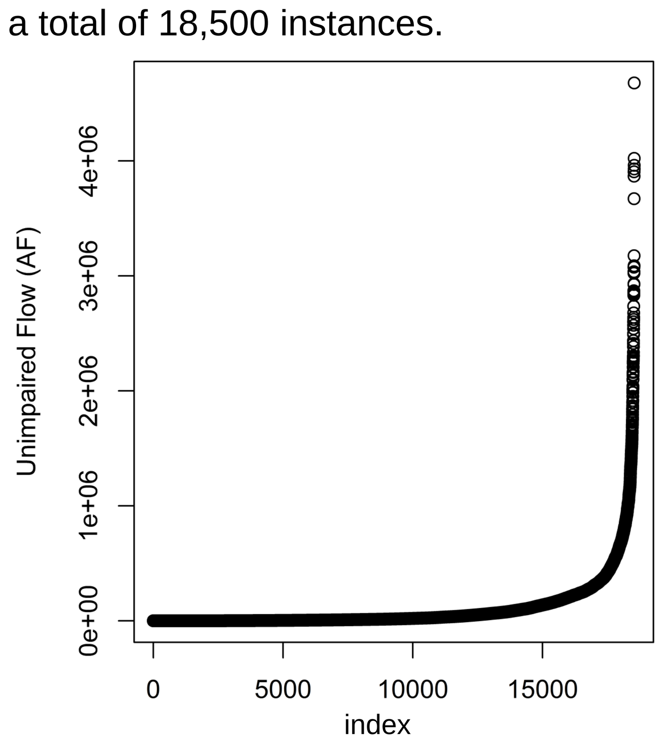
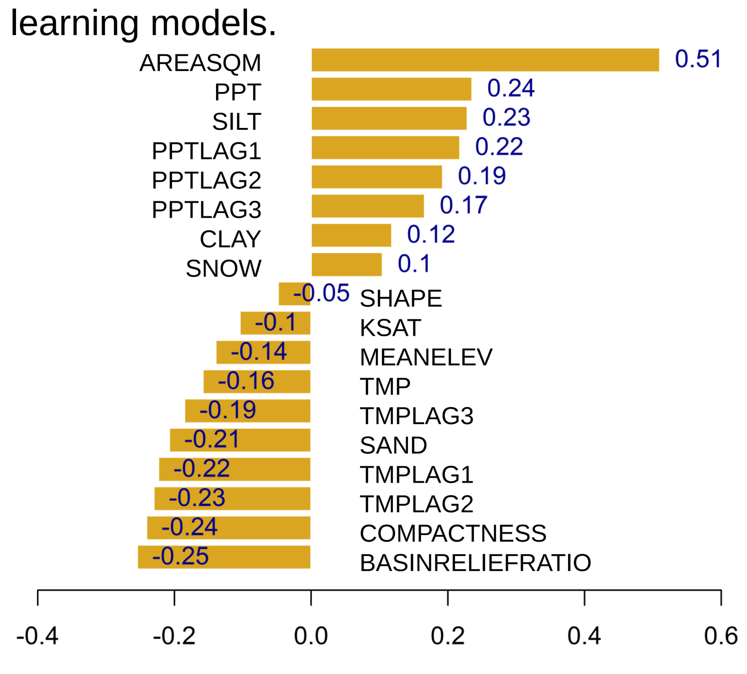


Figure 3. 69 Unimpaired Flow Catchments

Table 1. The Predictor Variables

Variable	Source
Climate PPT and lags TMP and lags SNOW	PRISM
Hypsometric BASIN RELIEF RATIO MEAN ELEV	SRTM90
Time MONTH	---
Basin Boundaries DRAINAGE AREA SHAPE COMPACTNESS	NHD2PLUS
Soil CLAY SILT SAND KSAT	POLARIS: (modified SURRGO)
Geology DOMINANT GEOLOGY	NRCS

Figure 4. Correlation of Predictor Variables Considered with the Response Variable (Unimpaired Flow). We generally expect to see the top variables show up in the machine learning models.



3 Decision trees (e.g. CART and RF) make binary splits on the predictor variables.

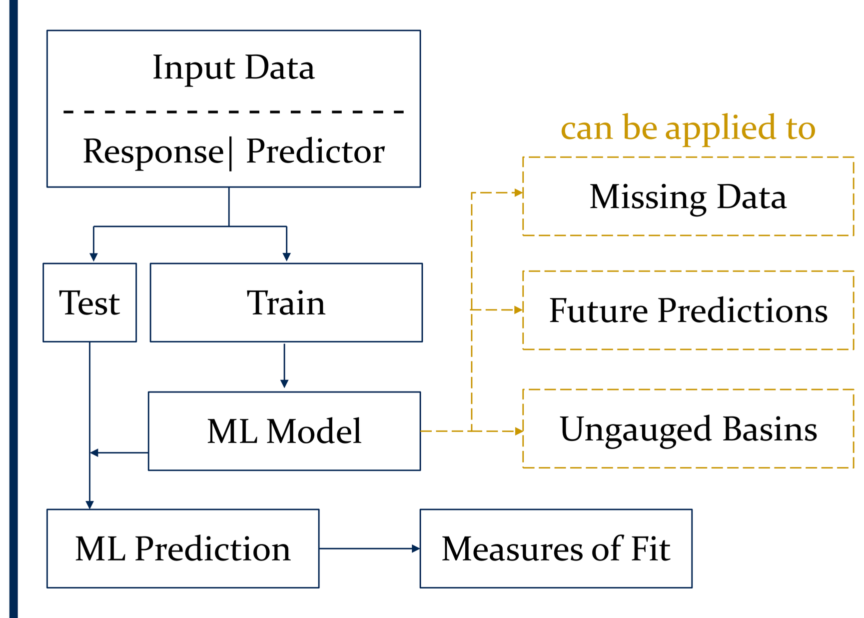


Figure 5. Study Flowchart

In this study, the choice of a suitable model shouldn't solely rely on statistics; some models better reflect theoretical foundations in hydrology. Some applicable supervised machine learning models to consider are: support vector machines, neural networks, ensemble methods, random forests (RF), lasso and ridge regression. In recent years, the popularity of tree learning methods like *Classification and Regression Trees (CART)* and *Random Forests* has exploded mainly due to their ease of understanding.

4 A CART model is one tree and therefore can't predict a continuous variable very well.

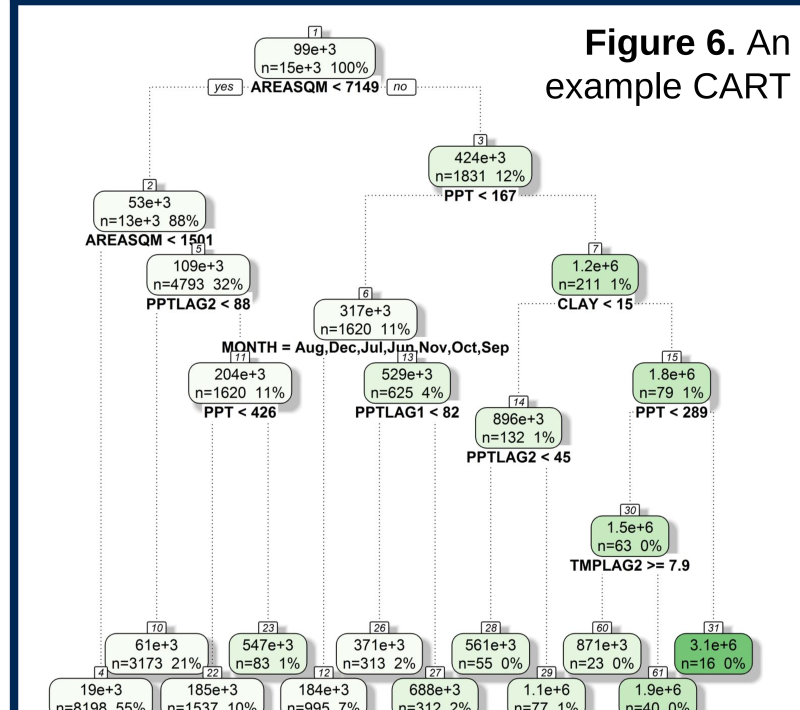
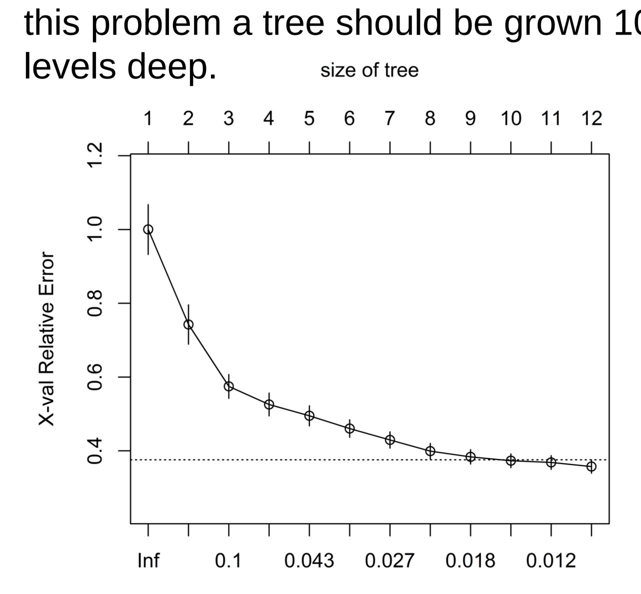


Figure 6. An example CART

Figure 7. The Optimal Size of a CART Based on a Complexity Parameter. For this problem a tree should be grown 10 levels deep.



5 A RF model grows many trees and averages the predictions making a better predictor of continuous variables compared to the CART.

Figure 8. At each node in a RF, the model can only make a split on a random subset of the variables. In this problem, the optimal number of random parameters is 12 out of the full 20 predictor variables. This introduces randomness and decorrelates the trees in the forest.

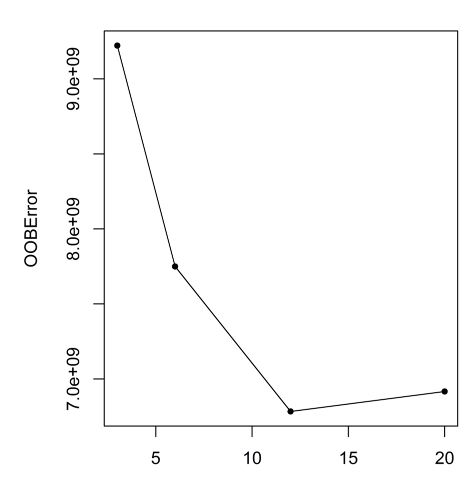


Figure 9. Number of Trees Needed for Stability of Model Predictions. In this problem it is around 100 trees.

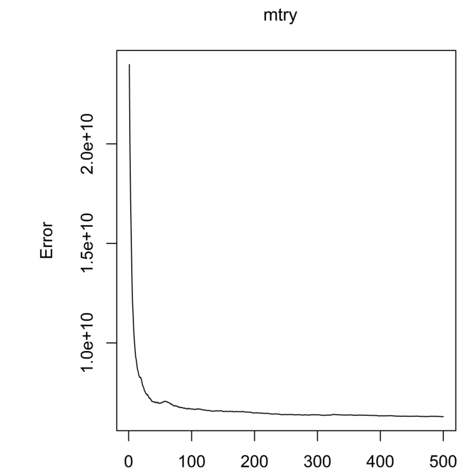
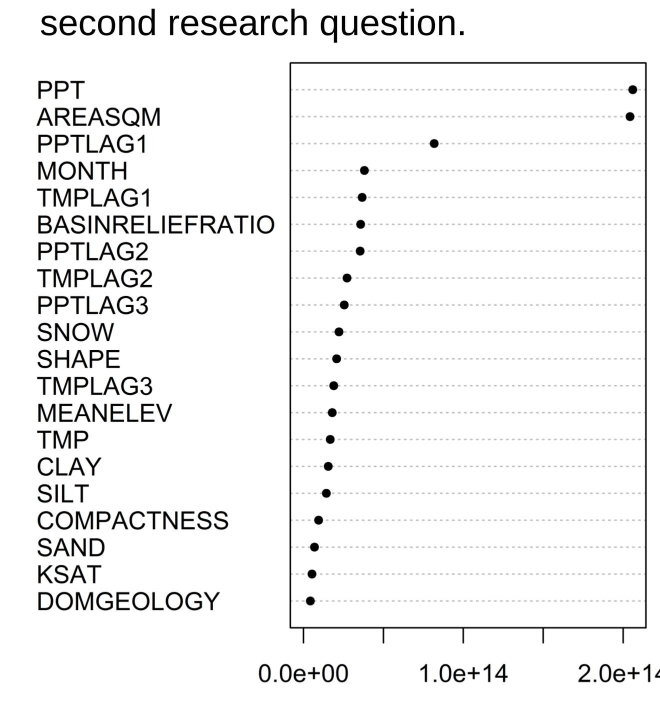


Figure 10. Variable Importance List. In this problem the precipitation, catchment area and month can explain most of the variability in unimpaired flow. This answers the second research question.



6 The random forest model outperforms a process-based model with NSE=0.88.

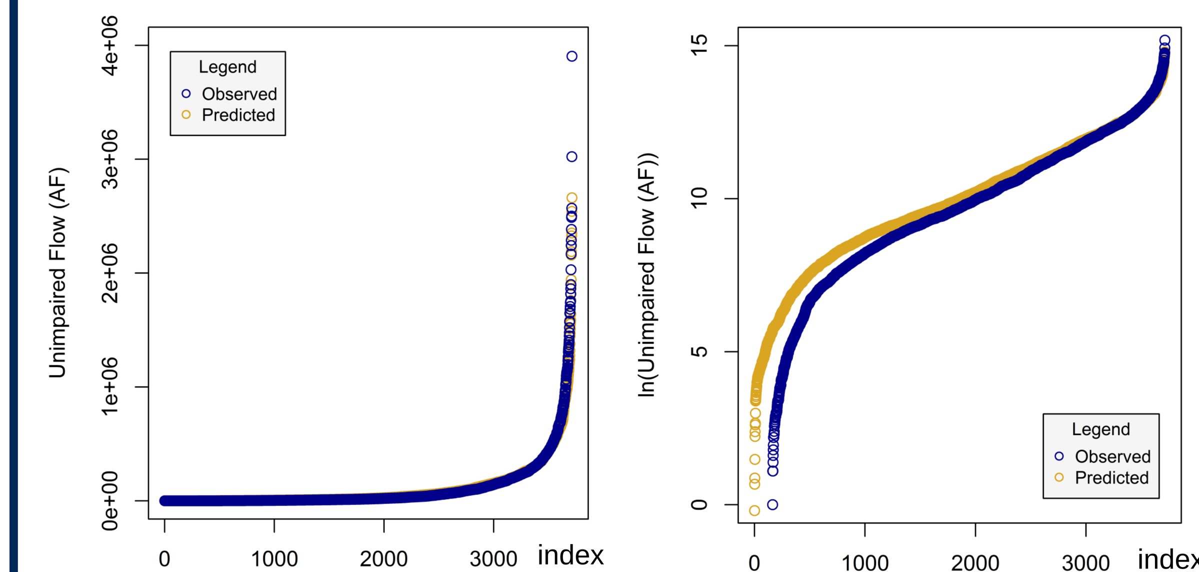


Figure 11. Observed vs. Predicted Unimpaired Flows. The model is slightly over predicting the low flows.

Model Fit Results: R^2 (Coefficient of Determination): **0.883**, RMSE (Root Mean Square Error): **85134 AF**, RSR (RMSE standard deviation ratio): **0.346**, NSE (Nash-Sutcliffe Efficiency): **0.880**, PBIAS (Percent Bias): **0.061**%. According to Moriasi 2007 this constitutes a "very good" performance.

Figure 12. Model Fit Diagnostics. Calculating the R^2 for different categories can help diagnose problem areas in the model.

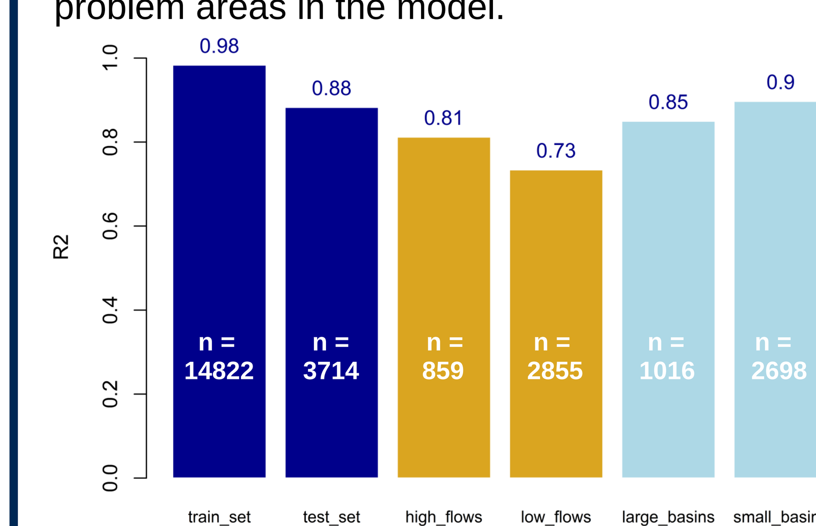


Figure 13. R^2 For Each Basin

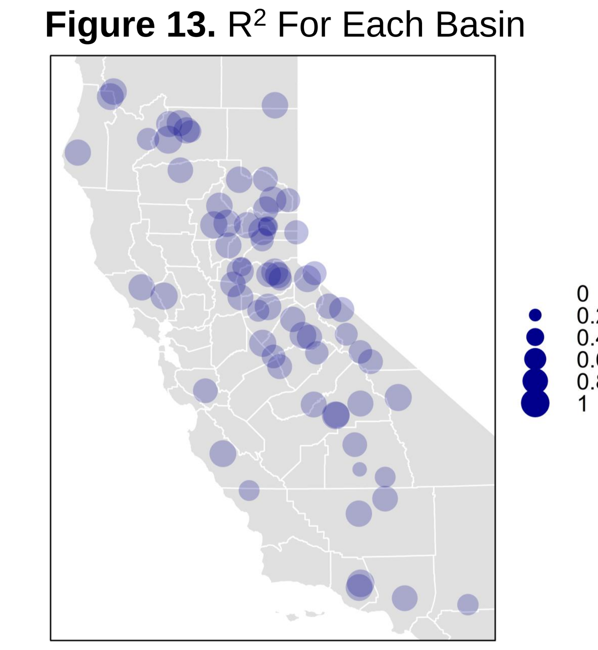


Figure 14. Benchmarking Results. The random forest model does better in all 10 basins compared to a process based model. These basins were selected because they overlap between the two studies. The random forest model also does better than the linear model in all but one basin. This is expected since we know runoff processes are not linear. These results answer the first research question.

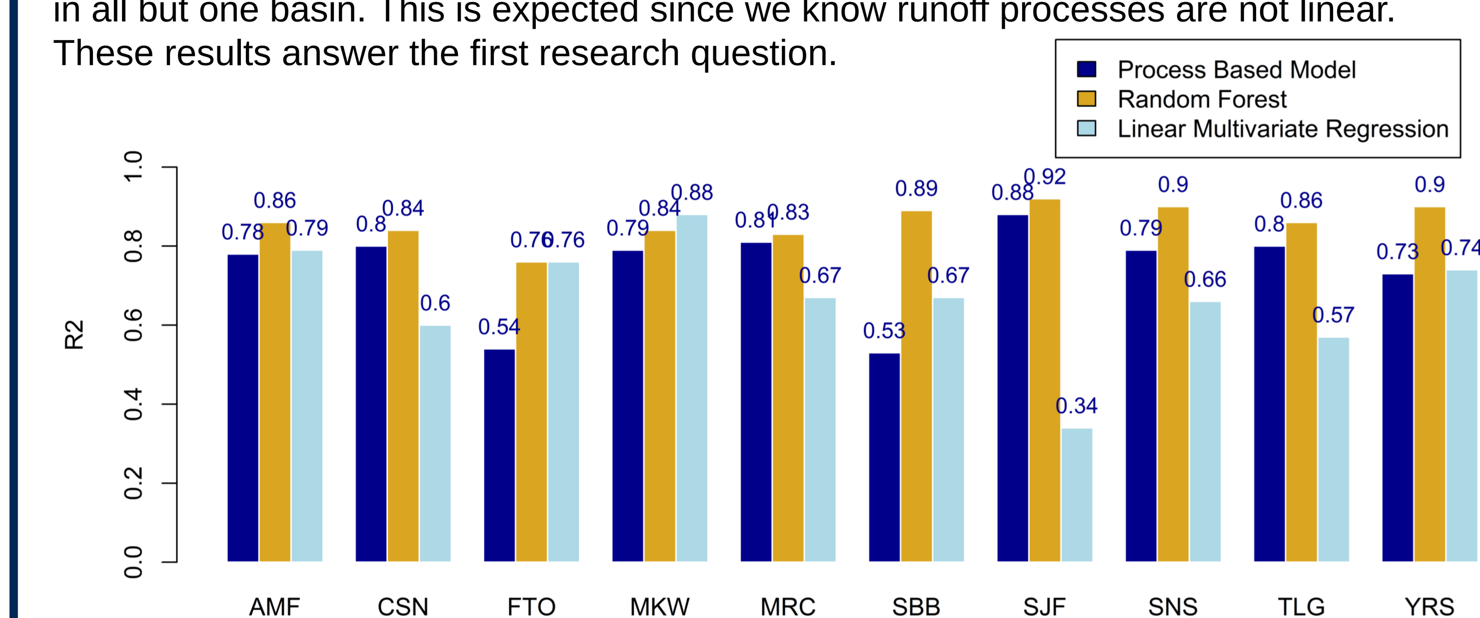
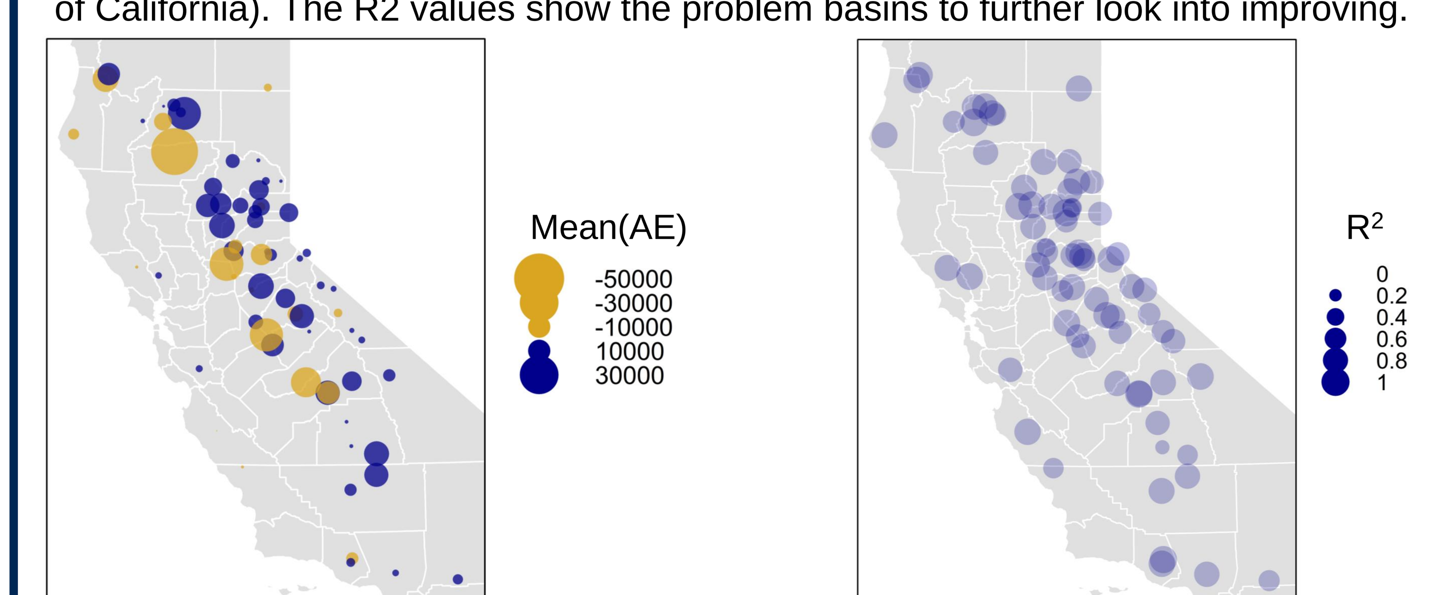


Figure 15. Model Improvement. There may be a variable that could eliminate the spatial trend in the mean of the absolute error (see the ridge of under-prediction down the middle of California). The R2 values show the problem basins to further look into improving.



References

[1] R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/ . R version 3.4.0 (2017-04-21) -- "You Stupid Darkness"
[2] Taiyun Wei and Viliam Simko (2016). corrplot: Visualization of a Correlation Matrix. R package version 0.77. https://CRAN.R-project.org/package=corrplot
[3] Pebesma, E.J., R.S. Bivand, 2005. Classes and methods for spatial data in R. R News 5 (2), https://cran.r-project.org/doc/Rnews/ .
[4] Roger S. Bivand, Edzer Pebesma, Virgilio Gomez-Rubio, 2013. Applied spatial data analysis with R, Second edition. Springer, NY. http://www.asdar-book.org/
[5] Terry Therneau, Beth Atkinson and Brian Ripley (2017). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-11. https://CRAN.R-project.org/package=rpart
[6] A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18--22.
[7] Robert J. Hijmans (2016). raster: Geographic Data Analysis and Modeling. R package version 2.5-8. https://CRAN.R-project.org/package=raster
[8] Roger Bivand and Colin Rundel (2017). rgeos: Interface to Geometry Engine - Open Source (GEOS). R package version 0.3-23. https://CRAN.R-project.org/package=rgeos
[9] Roger Bivand, Tim Keitt and Barry Rowlingson (2017). rgdal: Bindings for the Geospatial Data Abstraction Library. R package version 1.2-7. https://CRAN.R-project.org/package=rgdal
[10] Edmund M. Hart and Kendon Bell (2015) prism: Download data from the Oregon prism project. R package version 0.0.6 http://github.com/ropensci/prism DOI: 10.5281/zenodo.33663
[11] USDA-NRCS Soil Survey Staff (2016). sharpshootR: A Soil Survey Toolkit. R package version 1.0. https://CRAN.R-project.org/package=sharpshootR
[12] Hadley Wickham (2007). Reshaping Data with the reshape Package. Journal of Statistical Software, 21(12), 1-20. URL http://www.jstatsoft.org/v21/i12/ .
[13] Nathaniel W. Chaney, Eric F. Wood, Alexander B. McBratney, Jonathan W. Hempel, Travis W. Nauman, Colby W. Brungard, Nathan P. Odgers, POLARIS: A 30-meter probabilistic soil series map of the contiguous United States, Geoderma, Volume 274, 15 July 2016, Pages 54-67, ISSN 0016-7061, https://doi.org/10.1016/j.geoderma.2016.03.025.

Supported by:

NSF DGE # 1069333, the Climate Change, Water, and Society IGERT, to UC Davis

Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors(s) and do not necessarily reflect the views of the National Science Foundation.