Using Machine Learning to Statistically Model Natural Flow

The Sacramento Watershed under Dry Conditions

Bonnie Magnuson-Skeels, Jay Lund, Robert Hijmans, and Theodore Grantham

April 12, 2016

UC DAVIS
CENTER FOR WATERSHED SCIENCES

Background

Estimating natural flow

- Research goal: create an improved natural flow model for dry years in the Sacramento watershed for use in DWRAT
  - DWRAT is a water rights curtailment model developed at UC Davis and funded by the SWRCB that suggests ideal curtailments for a basin
  - Currently uses a USGS statistical natural flow model as input
- Two general approaches for natural flow modeling
  - Mechanistic hydrologic modeling
  - Statistical models
- Tricky to evaluate because of limited ground-truth data

Definition of Natural Flow

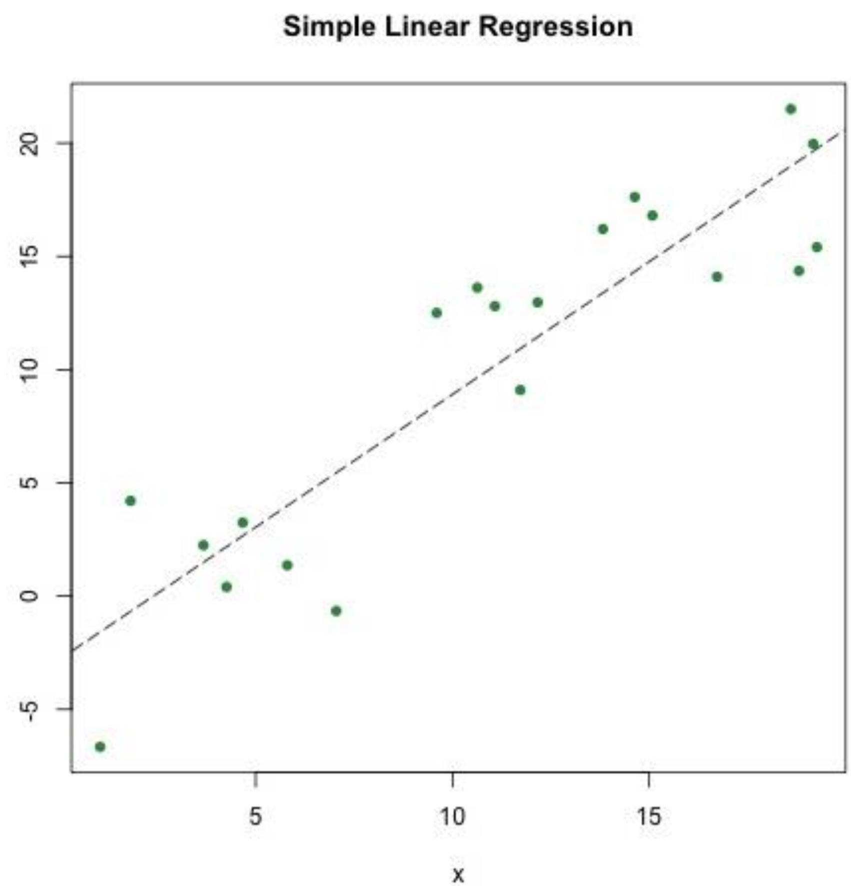
- Unimpaired flow
  - Assumptions about the current river channel configuration, vegetation, groundwater accretion/depletion rates, etc.

- Full natural flow
  - Theoretical flow of a river in its pre-development state

Chung & Ejeta, 2011; CA DWR, 2007; Kadir & Huang, 2015

Definition of Machine Learning

- A set of techniques for predicting an output based on one or more inputs
  - Mostly the same thing as statistical learning, although more focused on accurate prediction than inference
  - Regression, K Nearest Neighbors, Random Forests, Support Vector Machines…



USGS Natural Flow Model

- Uses random forests to predict average flow rate (cfs) based on publicly available geospatial data
  - Label (y variable): Flow from GAGES II reference gages
  - Features (x variables): precipitation, temperature, elevation, soil characteristics, etc.
  - Data covers 1950-2011

- Set of 36 (3 x 12) monthly regional models:
  - California's 3 ecoregions (Coastal, Intermountain, and Xeric)
  - 12 months

Carlisle *et al.*, 2010; Grantham, 2014

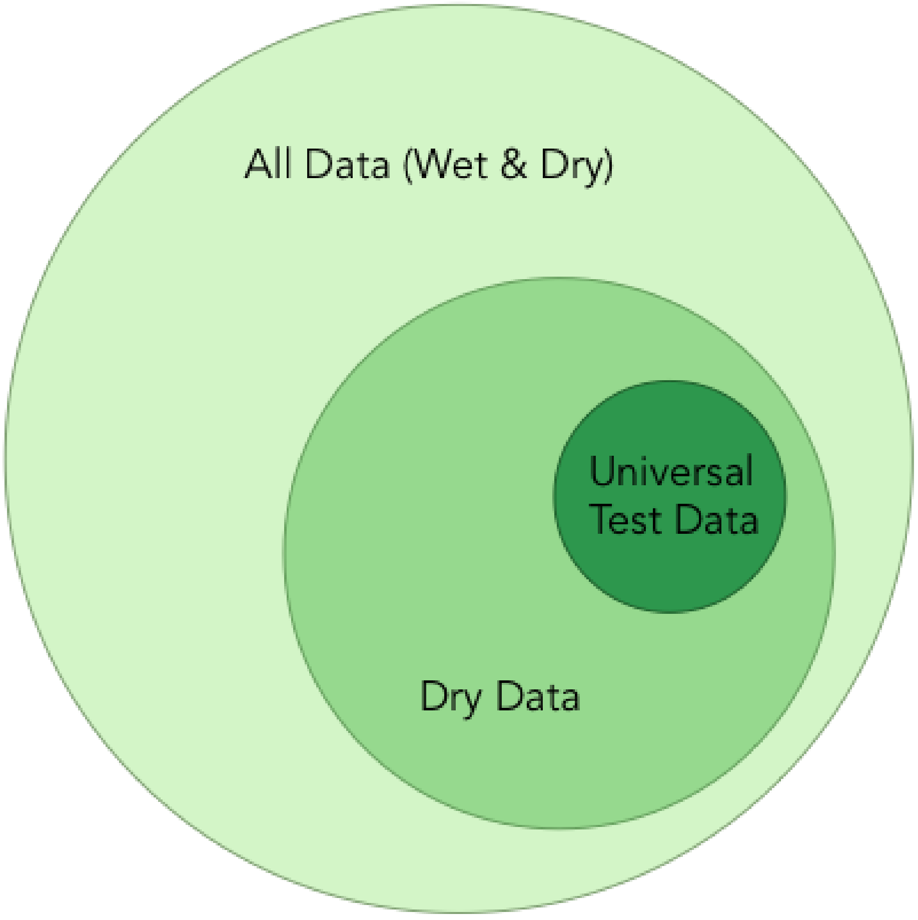
Proposed Improvements

- Additional machine learning algorithms
- New feature selection methods & dimensionality reduction
- Training model on more applicable datasets
  - Dry-year datasets for monthly regional models
  - Sacramento basin dataset
- All this means trying out a LOT of different combinations of models and datasets to see how they compare.

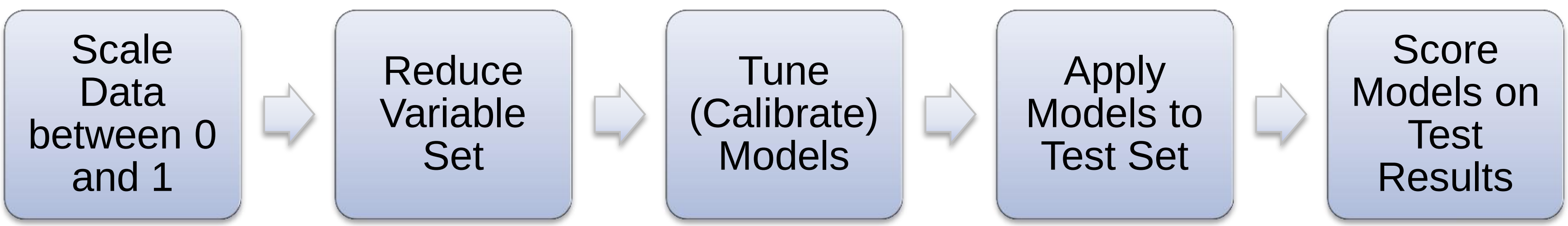
Method

Evaluation

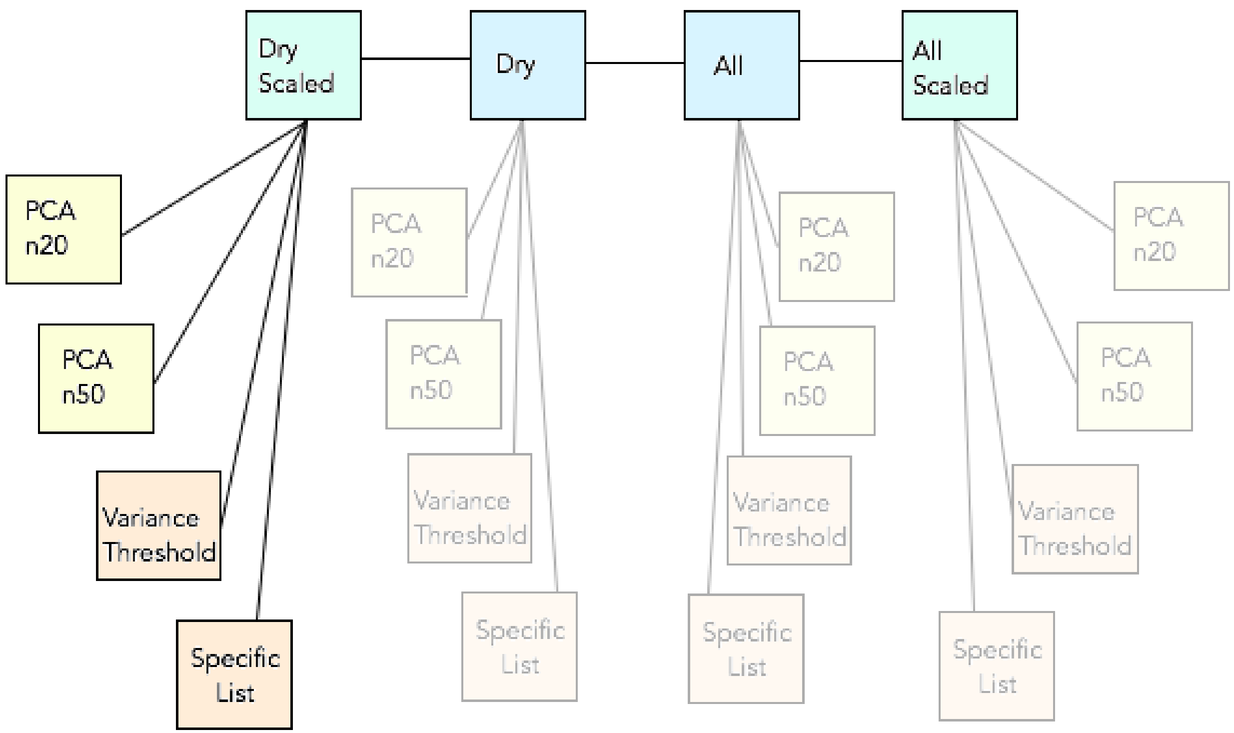
- Five-fold cross-validation
  - Randomly splitting data from drier years into 5 different 80/20 train/test sets
  - Dry-year test sets were used as a "universal test set"
  - Average results from each fold to get stable estimates of performance on previously unseen test data



Sequence for Each Fold



Dataset Transformations



Calibrating Machine Learning Models

- Machine learning algorithms:
  - Ridge regression
  - Random forest
  - K nearest neighbors
  - Support vector machine
  - Decision tree
  - AdaBoost
  - Averaging Ensemble
  - Stacking Ensemble
  - Stacking Ensemble with original features
- The first six are tuned (e.g., calibrated) on the training data using a grid search and 5-fold cross-validation
- The latter three are tuned based on these tuning test scores

Model Evaluation

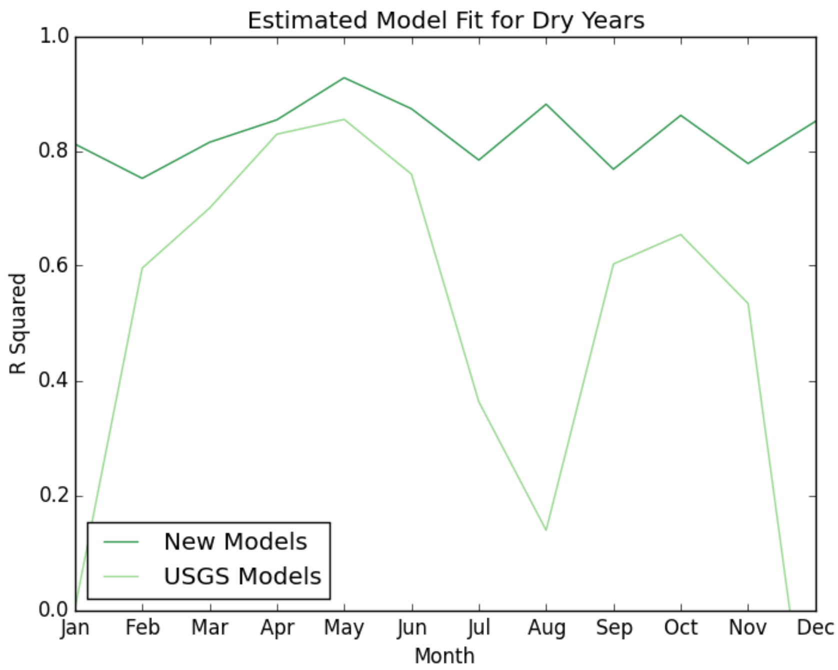
- Each trained algorithm is then applied to the testing dataset to find the best approach for predicting natural flow
  - 9 algorithms * 20 datasets = 180 algorithm-dataset combos
- Evaluation metrics:
  - R^2
  - Observed/expected ratio (mean and standard deviation)
  - Mean squared error and root mean squared error

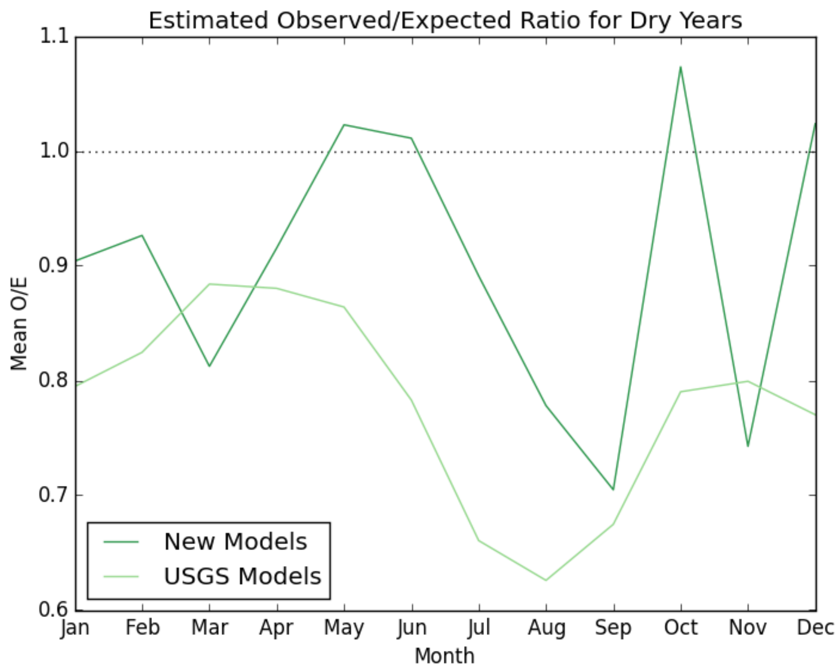
Results

General Dry-Year Results

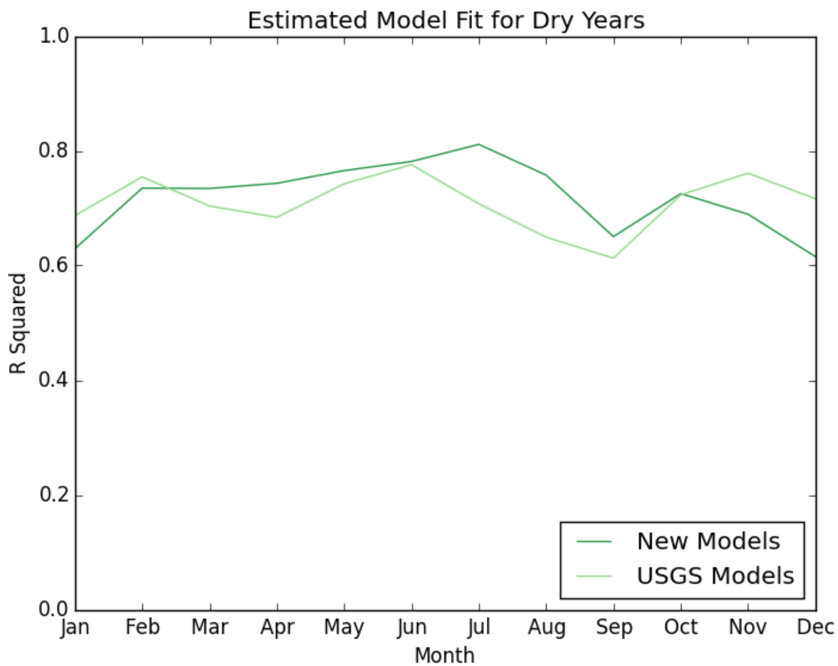
- Running the sequence for every month for both the Intermountain and Xeric regions resulted in 24 (2 regions x 12 months) best models.
- Stacking models are most often the best algorithm.
- Reducing training data to dry years often helped in the Intermountain region, but not very much in the Xeric region.

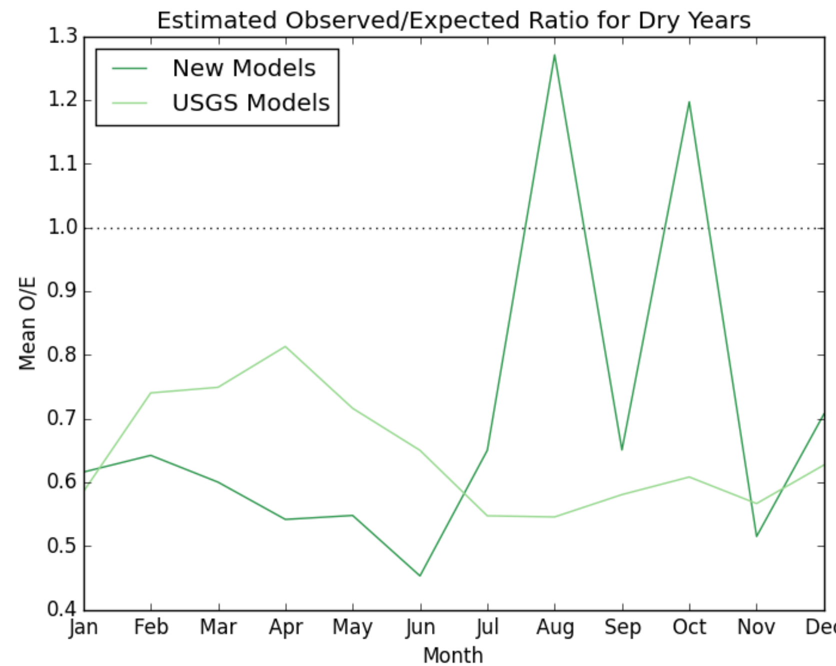
Comparison to USGS: Intermountain



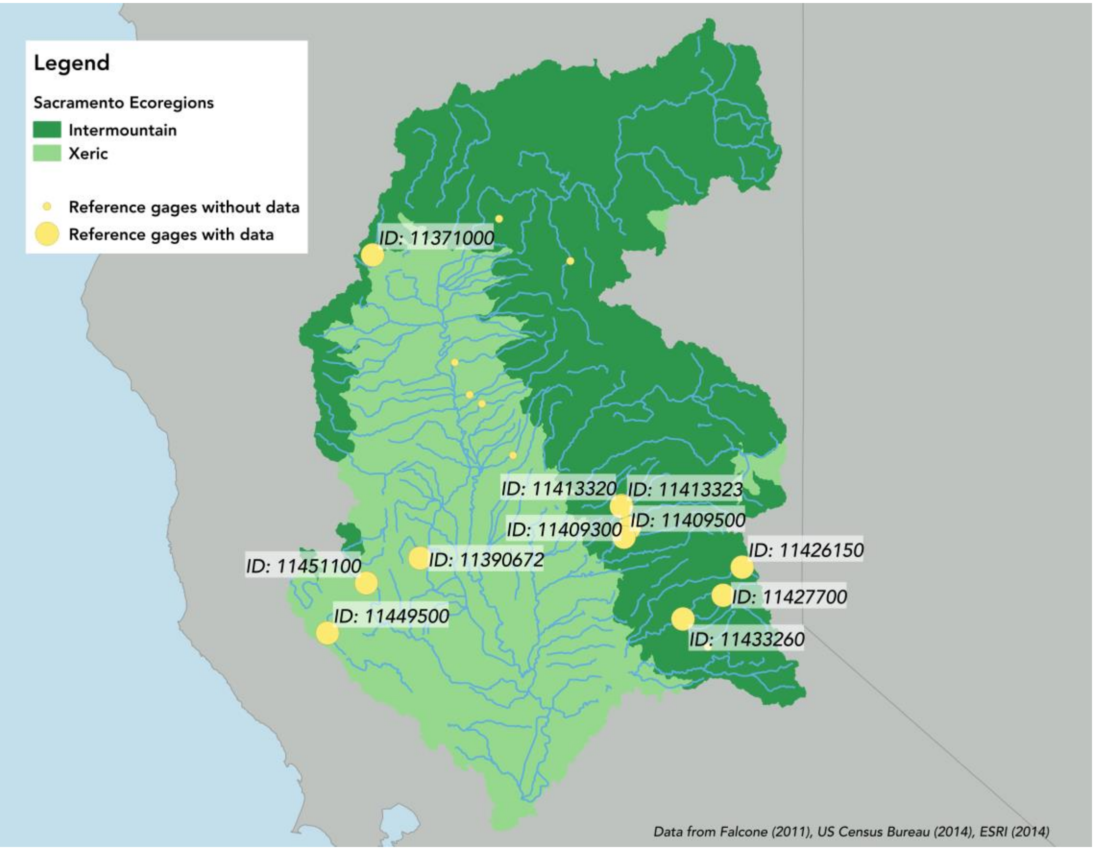


Comparison to USGS: Xeric





Restricting the Data Geographically?



- Not enough variation in the dataset.
- Models scored well on test data, but they tend to predict very low flows, probably because the dataset is made up of only a few above-rim gages.

Technical Details

- Written in Python
  - Wrote *mlutilities* package to facilitate experimenting with different combinations of datasets and machine learning techniques
  - *mlutilities* uses *pandas* and *sklearn* packages
- Parallelized and ran full process on Amazon Web Services
  - Running sequences for all scenarios required training models over 50,000 times
  - Reduced run time from ~36 hours to ~3 hours

Conclusions

- Training the Intermountain models on a dry-year dataset improved performance.

- Stacking ensemble modeling increases model performance.

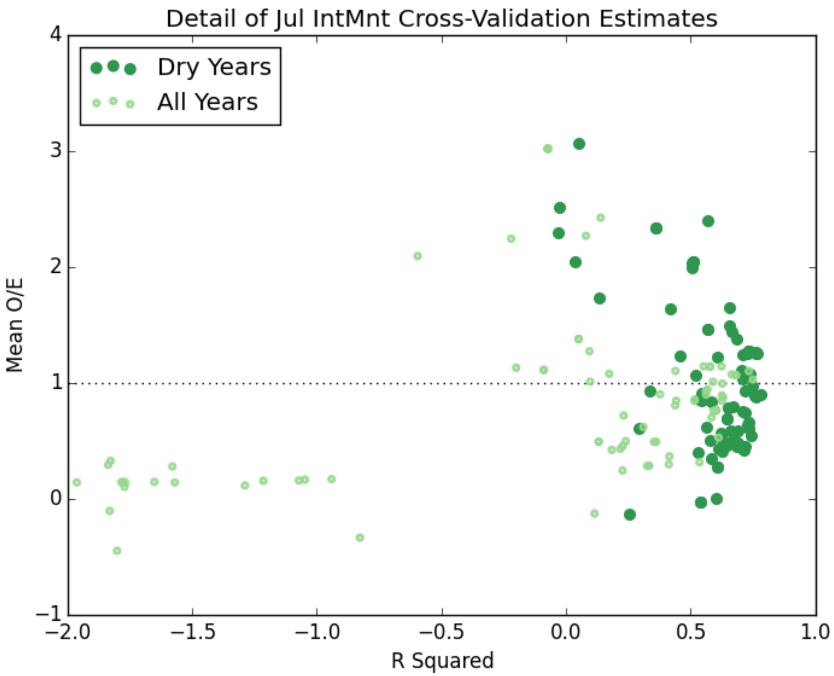
Questions?

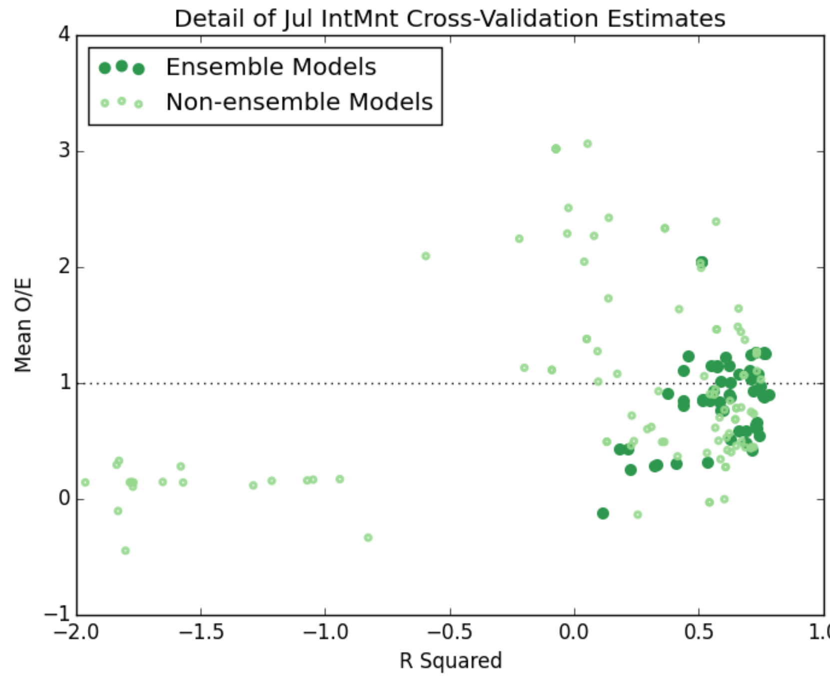
Sources

- California Department of Water Resources, Bay-Delta Office. (2007). California Central Valley Unimpaired Flow Data. (4th ed., pp. 52).
- Carlisle, D. M., Falcone, J., Wolock, D. M., Meador, M. R., & Norris, R. H. (2010). Predicting the Natural Flow Regime: Models for Assessing Hydrological Alternation in Streams. *River Research and Applications, 26*(2), 118-136.
- Chung, F., & Ejeta, M. (2011). *Estimating California Central Valley Unimpaired Flows*. Presentation to the California State Water Resources Control Board. Retrieved from http://www.waterboards.ca.gov/waterrights/water_issues/programs/bay_delta/sds_srjf/sjr/docs/dwr_uf010611.pdf.
- ESRI. (2014). *USA Rivers and Streams* [Digital spatial dataset]. Retrieved from: http://beta.esri.opendata.arcgis.com/datasets/0baca6c9ffd6499fb8e5fad50174c4e0_0
- Falcone, J. A. (2011b). *GAGES-II: Geospatial Attributes of Gages for Evaluating Streamflow* [Digital spatial dataset]. Retrieved from http://water.usgs.gov/GIS/metadata/usgswrd/XML/gagesII_Sept2011.xml
- Grantham, T. E. (2014). *Appendix B Section 2 (Drought Water Rights Allocation Tool Supply Estimation) of Drought Curtailment of Water Rights: Problems and Technical Solutions* (pp. 6). Center for Watershed Sciences: University of California, Davis.
- Kadir, T., & Huang, G. (2015). *Unimpaired Flows vs. Natural Flows to the Sacramento-San Joaquin Delta: What's the Difference?* Paper presented at the California Water and Environmental Modeling Forum, Folsom, California. Retrieved from http://www.cwemf.org/AMPresentations/2015/Kadir_NaturalFlow.pdf
- United States Census Bureau. (2014). *Cartographic Boundary Shapefiles - States (500k)* [Digital spatial dataset]. Retrieved from: https://www.census.gov/geo/maps-data/data/cbf/cbf_state.html

- All code used for this research is located at: https://github.com/brmagnuson/MachineLearningPipeline

Example: July Intermountain Model





Best model: Stacked ensemble based on dry-year dataset reduced to 50 components using PCA.

Restricting the Data for Wet Years?

- Repeated the same process to test using a wet-year data set to predict for wet years.

- The full dataset of all water years tends to do better.
  - This might be because a more varied dataset helps predict the greater variation in wet years.