Using Machine Learning to Statistically Predict Natural Flow: The Sacramento Watershed under Dry Conditions

By

BONNIE ROSE MAGNUSON-SKEELS
B.S. (University of Idaho) 2011
B.A. (University of Idaho) 2011

THESIS

Submitted in partial satisfaction of the requirements for the degree of

MASTER OF ARTS

in

Geography

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

_____

Jay R. Lund, Chair

_____

Robert J. Hijmans

_____

Theodore E. Grantham

Committee in Charge

2016

ABSTRACT

Machine learning techniques were applied to climatic, geologic, and geographic data to statistically model natural river flows in dry years in California's Intermountain and Xeric ecoregions. The model is tailored to predict flows during dry years for use as inputs for the Sacramento River version of the Drought Water Rights Allocation Model (DWRAT), a water rights curtailment model developed at the University of California, Davis. The modeling approach builds on a general-purpose statistical model developed by the US Geological Survey designed to predict natural flows at national and regional scales. Multiple machine learning algorithms were applied, using different techniques to select variables and reduce dimensionality and restricting training data to drier years. The ability of the resulting models to predict flows in dry water years was evaluated with multiple test metrics. The new models consistently tested as well as or superior to the corresponding general-purpose models when used to predict dry year flows, and in some cases, they performed far better. This improvement in predicting natural flows in dry years allows for more accurate estimation of available water and should help make DWRAT more useful for informing water rights curtailment decisions. This research also provides a high-level Python package for easily exploring and evaluating various combinations of machine learning techniques.

TABLE OF CONTENTS

ABBREVIATIONS

BCM – Basin Characterization Model

C2VSIM – California Central Valley Groundwater-Surface Water Simulation Model

CFS – Cubic feet per second

CWS – Center for Watershed Sciences

DWR/CA DWR – California Department of Water Resources

DWRAT – Drought Water Rights Allocation Tool

GAGES – Geospatial Attributes of Gages for Evaluating Streamflow

HUC – Hydrologic Unit Code

ICA – Independent Component Analysis

MSE – Mean Squared error

NOAA – National Oceanic and Atmospheric Administration

NWIS – National Water Information System

NWS – National Weather Service

O/E – observed/expected

PCA – Principle Component Analysis

PRMS – Precipitation-Runoff Modeling System

RMSE – root mean squared error

SWRCB – State Water Resources Control Board

USDA – United States Department of Agriculture

USEPA – United States Environmental Protection Agency

USGS – United States Geological Survey

CHAPTER 1: INTRODUCTION

Background

For many rivers around the world, actual natural flow rates are unknown because extensive development of hydrologic resources through dams, levees, and other water infrastructure has substantially altered natural flow regimes, changing the magnitude, timing, and variability of river flows. Additionally, the network of monitoring stations is sparse, so natural flow rates are often unknown even where the natural flow regime is unaltered (Poff *et al*., 1997; Pringle *et al*., 2000; Nilsson *et al*., 2005). However, it is often useful to have estimates of a river's natural flow. For example, approximations of natural flow can be used to estimate impacts of development on a watershed (Carlisle *et al*., 2010) and to assess which water rights can be fulfilled during a drought based on their legal structure (Lund *et al*., 2014; Lord, 2015).

Several modeling approaches exist for estimating a river's natural or unimpaired flow, including mechanistic and statistical models (Arthington *et al*., 2006). Mechanistic models are detailed simulations of watersheds based on either physical hydrologic principles and historic, "pre-disturbance" records of flow or using observed streamflow from current gage measurements and "unimpairing" them by adding known diversions back in. Although detailed mechanistic models are useful for understanding a watershed's processes and are the standard approach in hydrology, they also have disadvantages. We usually lack pre-disturbance records of flows for most streams (Carlisle *et al*., 2010), and watershed models are often very complex and depend on variables that are often difficult or expensive to obtain at sufficient temporal and spatial resolution (Eng *et al*., 2012). Statistical models are constructed with available data about the variable of interest (in this case, streamflow) and other variables thought to be relevant. The model can then predict the value for the variable of interest across a range of conditions. Statistical models can be simpler and faster to develop than mechanistic models, particularly when predictor variables are readily available, as is the case for elevation and rainfall (Eng *et al.*, 2012). Statistical modeling (generally called machine learning in this thesis) has often been used to predict observed hydrologic phenomena, but its application to prediction of expected natural flows is more recent. Researchers at the US Geological Survey (USGS) have shown it to be effective for predicting natural flow at ungaged locations based on available geospatial data. The USGS statistical natural flow model is based on reference gages in the region of interest with at least 20 years of daily flow data and basin characteristics at each gage, including climatic, geographic, and geologic attributes. Using a random forest algorithm, models are trained to predict monthly observed flow and other flow variables of interest based on the basin characteristics. The trained models can then be used to predict monthly flows at ungaged locations using the same set of basin predictor variables (Carlisle *et al*., 2010).

The ability of the USGS model to predict natural flow at any point in a watershed is useful to the Drought Water Rights Allocation Tool (DWRAT), a water rights curtailment model developed by the Center for Watershed Sciences (CWS) at University of California, Davis as a research project for the California State Water Resources Control Board (SWRCB). DWRAT takes estimates of full natural flow for various points throughout a watershed and, given data on local water rights, uses optimization methods to suggest water users to curtail if available water is insufficient (Lord, 2015; Lund *et al*., 2014). USGS researchers modified the original model in two key ways before it was initially applied in DWRAT. First, they found that a set of monthly regional models (defined by aggregated Level-II ecoregions) generally produced more accurate

predictions than statewide models, so the models train on region-specific data from around California and portions of Oregon and Nevada (Grantham *et al*., 2014; Grantham, 2014; Lund *et al*., 2014; Grantham, 2015). Second, they developed a method to estimate natural flow at a daily time step (rather than the monthly natural flow predicted by the model) for every sub-basin in the watershed. The regionally calibrated predictions of monthly natural flows for each sub-basin were used to extrapolate hydrologic estimates of daily natural flow obtained from the National Weather Service (NWS) at select gage locations to ungaged sub-basins. This was done by calculating the ratio of predicted natural flow values between the gaged and ungaged sub-basins and applying it to the NWS daily flow predictions as a scaling factor (Grantham, 2014). While the intention of this approach was to obtain natural flow estimates at a daily time step, which the current version of DWRAT requires, it also facilitated DWRAT's use of the natural flow model by allowing a user to generate estimates of current natural flow as well as historical natural flow. Since as of 2016 the climate data needed for the model (from Oregon State University's PRISM Climate Group) was only available through 2011, the model could only be used to predict natural flow up to 2011. Before the flow model is used as input to DWRAT, the scaling factors from a year assumed to be representative (the 1977 water year) are used to extrapolate from current NWS estimates to all sub-basins around the watershed (Grantham, 2014).

Natural Flow

Before continuing, it is useful to understand exactly what is meant by "natural flow." Poff *et al*. (1997) define the "natural flow regime" as "the characteristic pattern of a river's flow quantity, timing, and variability" without human alteration. (Others have since divided this concept more precisely into "unimpaired flow" and "full natural flow".) Poff *et al*. explained that flow regimes naturally vary over time (from hour to hour, era to era, and every time step in between) and across regions (due to geographic variation in climate, geology, topography, and vegetation). A natural flow regime can be described by five components: magnitude of discharge, frequency of occurrence, duration of a given flow condition, regularity or predictability, and rate of change between magnitudes, also known as flashiness. This thesis is concerned with the first component: magnitude of discharge, meaning rate of flow.

A key distinction should be drawn between unimpaired flow and full natural flow, which describe slightly different natural flows. Unimpaired flow is an estimate of flow that would occur without dams and diversions. It assumes the current river channel configuration with levees, current upstream and instream vegetation, and current groundwater accretion/depletion rates. It is often the result of hydrologic models that adjust measured flows by adding known agricultural and urban consumptive water use back in to "unimpair" them. In contrast, full natural flow is the theoretical flow of a river in its pre-development state, prior to any human influences, including loss of evaporation from drained wetlands and groundwater flows. It makes different assumptions than unimpaired flow about river channels, vegetation, evaporation, and surface-groundwater interactions (Chung & Ejeta, 2011; CA DWR, 2007; Kadir & Huang, 2015).

To simplify a complicated legal matter to a single sentence, California's riparian water right holders (those whose properties are adjacent to water bodies) are entitled to the full natural flow of the water body, while appropriative water right holders (those who divert water to storage or a non-adjacent property and have seniority based on their date of first diversion) have a right to the remaining unimpaired flow of a water body (Lord, 2015). Table 1.1 compares each of these terms, their applications, and associated sources.

Table 1.1. Natural Flow Definitions and Applications

Term	Definition	Uses	References
Unimpaired flow	Estimate of a river's flow assuming its current channel configuration, vegetation, and surface-groundwater interaction	Water management planning, flood modeling, appropriative water rights assessments	Chung & Ejeta, 2011; CA DWR, 2007; Kadir & Huang, 2015
Full natural flow	Theoretical flow of a river in pre-development state	Environmental impact assessments, riparian water rights assessments	Chung & Ejeta, 2011; CA DWR, 2007; Kadir & Huang, 2015; Carlisle *et al*., 2010

"Natural flow" as used in this thesis is full natural flow as defined above. The models used are built with data on flow at USGS reference gages, which are either undisturbed or "least-disturbed" by human development. This means that anthropogenic impacts on current river channels, vegetation, and surface-groundwater impacts are nonexistent or insignificant upstream of and at those gage locations. Also, this work predicts flow purely based on natural geographic, geologic, and climatic variables, with minimal effects from human development (see Appendix A for a full list of variables).

Hypothesis

DWRAT's curtailment suggestions reflect the underlying natural flow models used to estimate water availability. Previous work has made the natural flow model more accurate, but further improvements in accuracy, particularly in predicting dry year flows, could be helpful for guiding curtailment decisions during droughts. it is worthwhile to evaluate and to improve its accuracy for curtailing water rights during droughts. USGS researchers previously used random forests (Breiman, 2001) as the modeling method because it has proved relatively robust and accurate (Carlisle *et al*., 2010), but other machine learning techniques may be more effective. Additionally, in the same way that the USGS found improved predictive performance by creating separate models for different regions, using a dataset restricted to only drier years or to a more localized geographic area could further improve predictive performance. This thesis applies several machine learning algorithms as well as additional variable selection and dimensionality reduction approaches to model dry year flows in the Sacramento River basin. By evaluating and characterizing the performance of these various approaches, a more accurate model geared toward predicting natural flows during droughts can be created and used as input to DWRAT. The Sacramento River was chosen as a test case because it is a basin currently being evaluated for application of the DWRAT model.

Significance of Research

With better information about available flows (i.e., the natural water supply), water rights curtailments can be suggested more effectively and reliably. Development of multiple alternative flow models also provides a range of estimates of natural flow, which would allow for sensitivity

testing of DWRAT's robustness to different flow model inputs. It could also support probabilistic analysis of drought curtailments. More generally, models for predicting natural flow for dry years are likely more useful for informing curtailment decisions during droughts than a general-purpose natural flow model. Also, this research provides a higher-level Python package built on top of the existing *scikit-learn* package for more easily exploring and evaluating various combinations of machine learning techniques.

Preview of Results

Restricting the training data to drier years and following the USGS approach of creating monthly models for relevant aggregated ecoregions resulted in a set of 24 models (12 monthly models for each of the two regions containing the Sacramento watershed). When evaluated on their prediction of known dry water year flows, these models consistently tested as better than or equivalent to their corresponding general-purpose USGS models on multiple test metrics, and in some cases, they performed far better. This is a significant improvement toward predicting natural flows in dry years and could help DWRAT suggest improved curtailment decisions. However, both the new models and the USGS models tended to underpredict natural flow for the main stem of the Sacramento River by a large margin when compared against hydrologically calculated 1977 natural flow estimates for several points around the watershed. Since both statistical models are trained on data from smaller streams and upper reaches of watersheds, this is an experimental extension of the model, and it was not surprising that both models predicted very different values than the hydrologic model when extended to new territory. While these hydrologic estimates have their own set of modeling simplifications and errors and should not treated as a precise performance benchmark, it is interesting to note the difference in predicted natural flows.

Restricting the training data geographically rather than by water year type created a single model that only uses data from the Sacramento basin, rather than the set of monthly dry-year models created for each region. While this model performed well on all test metrics, it tended to predict very low flows because of the limited nature of and lack of variability in its training and test data, which consist of above-rim flows only. It would be less useful in predicting natural flows for the main stem of the Sacramento River.

Limitations and Assumptions

Natural flow is particularly difficult to assess because many locations lack flows recorded prior to human development and disruption, including the Sacramento River's main stem. The USGS classifies 18 of its gages in the Sacramento watershed as "reference" gages, meaning their historical records represent "hydrologic conditions which are least disturbed by human influences" (Falcone, 2011a; Falcone, 2011b). These can be used as an estimate of natural flow for those locations. However, as one might expect from their definition, none of these gages lie on the main stem of the Sacramento River, so while they can be left out of training data and used to test a model's predictions, they do not represent a natural flow ground truth for the main stem of the Sacramento River. This means that the model can be tested for some areas of the Sacramento watershed but not all. The model could be tested against predictions generated by a mechanistic Sacramento River model based on physical hydrologic principles, but that would

still not represent testing against pre-development ground truth. This limitation should be kept in mind when applying the model.

A second limitation in applying the natural flow model is that it can only predict for time periods for which predictor variable data are available. Most predictor variables are relatively static, unchanging values such as elevation and soil composition, but some predictor variables are climate-related values that change each month. This introduces complications when applying the flow model to the present day. For example, to predict flow for a month at a point, one needs to know that month's precipitation for the drainage area above that point. This makes it difficult to use the model to predict flow for the present day, since the current month's average precipitation is not yet known (although one could use a forecasted value predicted by a weather model). Alternatively, a spatial disaggregation process could be used. If a present-day estimate of natural flow is known for some point in the watershed, such as those modeled by the National Weather Service (NWS) for some locations, then the natural flow model's predicted historical flow between those locations and other points of interest can be used to create flow ratios. These flow ratios can then be used to spatially disaggregate an estimate of present-day natural flow at one location to other locations. This approach is currently used in DWRAT to get natural flow estimates for the current day throughout a watershed (Grantham, 2014). The second approach assumes that upstream-downstream ratios of natural flow are consistent across dry years, which is difficult to test because of the lack of data on natural flow. Both approaches would depend on another model's estimates, limitations, and assumptions, either from weather forecasting or natural flow modeling. This should be kept in mind when applying any historical natural flow model to the present day.

CHAPTER 2: LITERATURE REVIEW

Natural Flow Models for the Sacramento River

The Sacramento is California's largest river, carrying 31% of California's surface water runoff and supporting an intensely productive agricultural region with about 2 million acres of irrigated farmland (Sacramento River Watershed Program, 2010). It has undergone significant human-induced change, making prediction of its natural flows an interesting and valuable endeavor (Buer *et al*., 1989). Previous research on predicting unimpaired and full natural flows in the Sacramento watershed has focused on mechanistic hydrologic modeling, rather than statistical modeling, and prediction of full natural flow has entered the scene only recently. Table 2.1 lists other recent research efforts involving the Sacramento watershed that modeled unimpaired or full natural flow. Several of the models are subsequently highlighted in greater detail.

Table 2.1. Previous Sacramento River Full Natural and Unimpaired Flow Modeling Efforts

Source	Description	Area	Model Type
CA DWR 1980, 1987, 1994, & 2007	Unimpaired flow estimates dating back to 1921	Central Valley and Sacramento-San Joaquin Delta	Mechanistic
Fox *et al*., 2015	Reconstructed natural landscape to estimate long-term annual average outflow of the Delta	Central Valley, Sacramento-San Joaquin Delta, and San Francisco Bay	Mechanistic
Gleick 1987	Water-balance model of unimpaired runoff	Sacramento River	Mechanistic
Grantham, 2014; Carlisle *et al*., 2010	Used USGS reference gages to model natural flow of various rivers	Sacramento River, among others	Statistical
Hay *et al*., 2011	Used the USGS Precipitation-Runoff Modeling System (PRMS) to simulate streamflow under various climate change scenarios	Feather River, among others	Mechanistic
Huang *et al*., 2014; Kadir & Huang, 2015	Used C2VSIM and rainfall-runoff models to estimate both full natural flow and unimpaired flow	Central Valley and Sacramento-San Joaquin Delta	Mechanistic
Koczot *et al*., 2005	Used precipitation-runoff simulation models to predict unimpaired streamflow	Feather River	Mechanistic
MacDonald *et al*., 2008	Predict unimpaired flow based on tree rings	Sacramento River, among others	Statistical
Meko *et al*., 2001	Predict unimpaired	Sacramento River	Statistical

	Sacramento flow based on tree ring widths		
NOAA, 1972	NWS River Forecast System (which includes the Sacramento Soil Moisture Accounting Model and other models)	Sacramento River, among others	Mechanistic
Flint *et al*., 2012	California Basin Characterization Model (BCM)	California	Mechanistic

Starting with observed gage flows and adjusting them for upstream operations, the California Department of Water Resources (CA DWR) has released estimates of Central Valley unimpaired flow obtained regularly over the past 36 years (CA DWR, 1980; CA DWR, 1987; CA DWR, 1994; CA DWR, 2007). Recently, DWR's Bay-Delta Office has put significant effort into modeling both unimpaired and full natural flows for the Central Valley for the first time. Flows on the valley floor are modeled using the California Central Valley Groundwater-Surface Water Simulation Model (C2VSIM), an integrated surface-groundwater hydrologic model, and the "above rim" flows (meaning flows in the upper watersheds, above the rim of the valley) are estimated using daily precipitation-runoff models. By removing anthropogenic influences such as reservoirs and urban development upstream of a location, replacing the landscape with native vegetation, and making allowances for streams overtopping and other mechanistic hydrologic processes, full natural flow can be simulated using the same models (Huang *et al*., 2014; Kadir & Huang, 2015).

Various other mechanistically-based models have been used recently to predict natural streamflow in areas of the Sacramento watershed (Fox *et al*., 2015; Gleick, 1987; Koczot *et al*., 2005; Flint *et al*., 2012; Hay *et al*., 2011). In contrast, Meko *et al*. (2001) used a regression model to predict historical Sacramento River flows from the year 869 to 1977 based on chronologies of tree ring width. Since the training data used to build the model was based on unimpaired flow estimated by CA DWR, this model could be more precisely said to predict unimpaired flow rather than full natural flow as the terms are now understood.

USGS Natural Flow Model

Researchers at the USGS have effectively used machine learning to predict natural flow at ungaged locations based on available geospatial data (Carlisle *et al*., 2010). This section discusses their model and its evolution in more detail, as it directly inspired this thesis.

USGS researchers first used random forest models (Breiman, 2001) to predict biological condition of streams—meaning they classified whether or not the streams had been altered by human development—using widely available geospatial data as predictor variables such as land cover, topography, climate, soils, societal infrastructure, and potential hydrologic modification. Random forests were chosen because they efficiently handle messy data and higher-order interactions and often seem to be more accurate than many other machine learning algorithms because of the robustness and strength derived from tree-based regression and model averaging (Carlisle *et al*., 2009).

Building on this work, USGS researchers later used random forest models and publicly available geospatial data to predict various aspects of the natural flow regime, including the base flow index, number of flood-free days, daily variability, low and high flow pulses, and—most relevant to this thesis and for application in DWRAT—annual runoff of a stream (Carlisle *et al*., 2010). Training data came from the GAGES (Geospatial Attributes of Gages for Evaluating Streamflow) database (Falcone *et al*., 2010), which was developed as part of this effort. GAGES classifies 6785 USGS stream gages in the conterminous United States with at least 20 complete years of streamflow record between 1950 and 2007 and with reliably delineated watershed boundaries into "reference" and "non-reference" gages. This split gages into those that are least disturbed by human influences upstream and those where flow has been altered by human activities. Several hundred watershed and site characteristics, including climate, geology, soils, and topography, were then calculated for each gage based on national data sources (Falcone *et al*., 2010). Choosing 1272 of the reference gages from across the country, the USGS researchers developed a set of predictive random forest models, each of which contained 2000 individual regression trees. Each model used a flow metric calculated based on the gage as the dependent variable and 80 of the gage characteristics as the predictor variables. They calculated several model evaluation metrics: root mean square error (RMSE) as well as mean and standard deviation of the distribution of observed/expected (O/E) ratio values. They then applied the models to predict natural flow metrics in three ungaged watersheds as a demonstration (Carlisle *et al*., 2010).

Since its initial creation, the USGS natural flow model has been refined, expanded and applied in various ways. The underlying GAGES database was updated to GAGES-II in 2011. Notable changes included increasing the number of gages in the database from 6785 to 9322 by adding gages active in water year 2009 and gages in Alaska, Hawaii, and Puerto Rico; correcting basin boundaries and updating their characteristics; and adding mean annual precipitation and air temperature for 1950-2009 based on 4-km climate data from Oregon State University's PRISM Climate Group (Falcone, 2011a; Falcone, 2011b). In 2012, Eng *et al*. used a form of the USGS natural flow model that included human disturbance variables (e.g., number of dams and road density upstream of a gage) to predict likelihood of streamflow alteration.

Grantham applied this model to California for multiple purposes. The model was trained on a subset of 180 reference gages and then applied to generate estimates of mean annual flows in Californian rivers. These estimates were compared with the total annual face value of water right volumes to assess the degree of over-allocation of California's water (Grantham & Viers, 2014). Another paper also used the model to systematically assess which dams warrant environmental flows due to hydrologic alteration and improved predictive performance by creating separate monthly regional models for the first time. It divided California into three subregions following US Environmental Protection Agency (USEPA) Level-II ecoregion boundaries: interior mountains, coastal mountains, and the xeric (i.e., desert) region (Grantham *et al*., 2014; Omernik, 1987). The USGS natural flow model's use as input to DWRAT follows the same division into monthly regional models (Grantham, 2014; Lund *et al*., 2014).

Sub-Basin Extrapolation Process for DWRAT

As of 2016, the climate data needed for the USGS natural flow model was only available through 2011, so the method used to scale monthly natural flow estimates to a daily time step was also used to generate current daily natural flow estimates before the model's estimates are

used in DWRAT. As applied for use in DWRAT, the method extrapolates from current daily point estimates of natural flow from government sources to sub-basins throughout the watershed based on ratios of historical estimated natural flow. To use the USGS natural flow model in the Sacramento DWRAT model, average monthly flow in cubic feet per second (cfs) is predicted for each sub-basin in the watershed for 1977, a year chosen because it was California's worst recorded drought until the current one and because data required by the flow model is available for that year. Next, ratios of estimated 1977 flow are calculated between sub-basins containing a current daily natural flow estimate from the NWS and the California Data Exchange Center and other sub-basins without a current estimate. These ratios are then used to extrapolate from sub-basins with current daily natural flow estimates to sub-basin outlets around the entire watershed. For example, take two hypothetical sub-basins, A and B. A is upstream and had an estimated monthly natural flow of 50 cfs in June 1977. B is downstream, has a current daily natural flow estimate location near its outlet, and had an estimated monthly natural flow of 100 cfs in June 1977. Following the sub-basin extrapolation process, the flow ratio between A and B is ½, so if the NWS estimate of current daily flow in B is 150 cfs, we extrapolate that A has a current daily flow of 75 cfs. (Grantham, 2014). Figure 2.1 shows the location of the six daily natural flow estimates for the Sacramento River and their corresponding spheres of influence in the sub-basin extrapolation process.

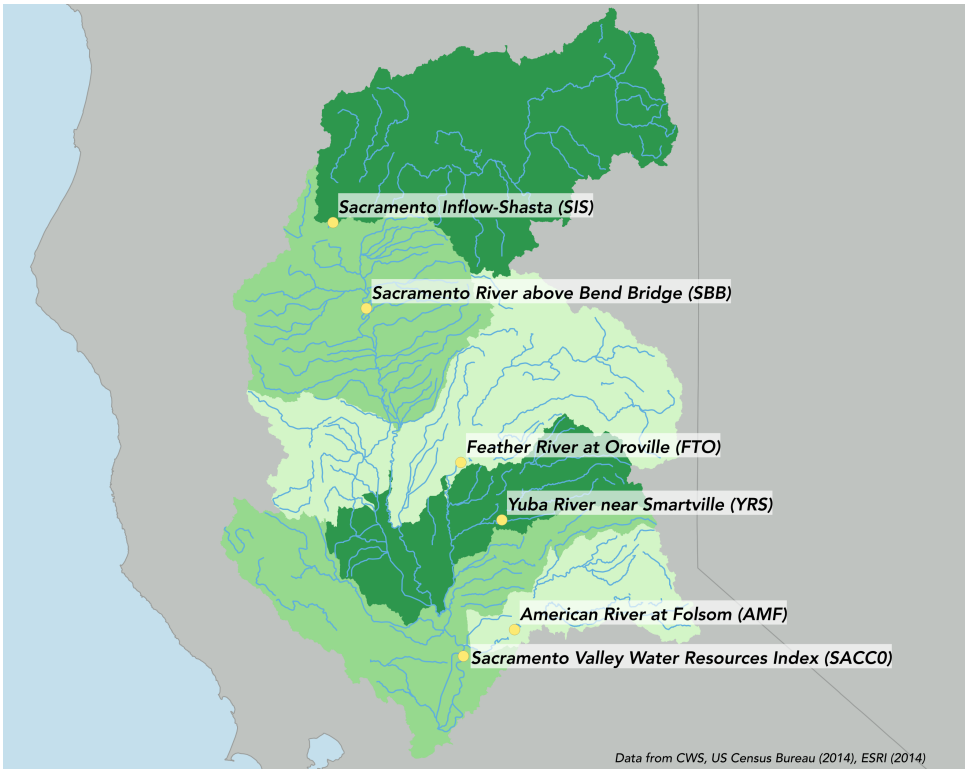


Figure 2.1. Sacramento daily natural flow estimates and sub-basin extrapolation regions[*]

This spatial disaggregation process made it possible to run DWRAT at a daily time step and to use DWRAT for the present day. However, this approach assumes that upstream-

[*] All maps produced for this thesis were made in QGIS (QGIS Development Team, 2015).

downstream ratios of natural flow are consistent across years, which is difficult to test because of the lack of data on natural flow. It also assumes that 1977 is the most appropriate year to use, although that drought may have affected areas differently than the current drought. Finally, it depends on the NWS model's estimates, limitations, and assumptions. Testing these assumptions is beyond the scope of this thesis.

Overview of DWRAT

DWRAT is a water rights curtailment model developed by the Center for Watershed Sciences at University of California, Davis as a research project for California's SWRCB. DWRAT uses estimates of full natural flow extrapolated to each sub-basin in the watershed as described above as input into an integrated set of two linear programming models. Given the predictions of available water in each sub-basin, water right holders' locations and recorded demand for water (data provided by the SWRCB for 2010 through 2013), and the legal priority system of riparian and appropriative water rights in California, DWRAT uses optimization methods to allocate available water given water rights law and available water supplies, suggesting ideal curtailments when available water is insufficient. The goal is to support a more transparent, precise approach for water rights curtailments when California experiences drought conditions. For a more detailed explanation, see Lund *et al*. (2014) and Lord (2015).

Machine learning and model aggregation

Machine learning is a set of techniques for predicting (i.e., estimating) an output based on one or more inputs (if performing supervised machine learning; unsupervised machine learning uses inputs without supervising output to examine the structure of the data). The inputs are commonly called features, predictors, explanatory variables, independent variables, or x-variables, and the outputs are referred to as labels, response variables, dependent variables, or y-variables. These terms can be used interchangeably, although this thesis uses the terms "predictor variables" and "dependent variables" for consistency and clarity for a non-machine learning audience. Machine learning can be considered a subfield of statistical learning. Statistical learning traces back to the earliest form of linear regression—first implemented in the early 1800s—and has grown to include many more flexible techniques for learning from data. Machine learning focuses on prediction from large and complex datasets, not on inference of the relations between variables. The algorithms fit a model from sets of "training data." Their accuracy can be assessed with "test data" withheld from the training set by comparing the observed values to the model predictions. If the model performs well, it can then be used to predict values for other cases for which predictors are available but the response is unknown.

Since its goal is prediction of a rate of flow, this thesis focuses on machine learning techniques that output a continuous quantity rather than a binary, categorical value, so it uses regression techniques rather than classification techniques. It uses a variety of general-purpose machine learning algorithms: ridge regression, regression trees, random forests, k-nearest neighbors, support vector machines, and AdaBoost. This thesis also combines the output of all the above algorithms through model averaging and model stacking. A detailed discussion of the reasoning behind each of these well-known machine learning methods is beyond the scope of this thesis. For more complete discussion of these methods, consult Mitchell (1997), Hastie *et al*. (2009), James *et al*. (2013), and Wolpert (1992).

CHAPTER 3: METHODS

Overview

This chapter describes methods used to test different machine learning algorithms, variable selection and dimensionality reduction techniques, and base datasets to discover which combinations were most effective in predicting natural flow. It first describes the data used to predict natural flow in more detail. This is followed by descriptions of three alternative model concepts: a dry-year model to test whether a more curated dataset was helpful in predicting dry year natural flows, a wet-year model to see if the pattern held true for predicting wet year natural flows, and a Sacramento basin model to examine the effect of geographically restricting the data. It concludes with a general discussion of *mlutilities*, a Python package written for this thesis research that facilitates exploring and evaluating the combinations of models and datasets, and then details of how *mlutilities* was applied for natural flow prediction.

Natural Flow Data

The dataset used to develop the natural flow model is the one used in the USGS model of natural flow used in DWRAT (Grantham, 2014): flow data from region-specific reference gages and basin characteristics from the GAGES-II database spanning the years 1950 to 2011. The Sacramento watershed is contained within the Intermountain and Xeric aggregated ecoregions, so only those datasets were used in this research when creating the monthly region models. (A set of models developed for the coastal mountain region is not considered in this research, being outside the Sacramento watershed.) The Intermountain set contains 38 gages (15,780 observations of average monthly flow across 62 years). The Xeric dataset contains 60 gages (26,665 observations across the same years). One gage in the northwestern Sacramento watershed was classified as belonging to the Coastal Mountain region for the USGS model used in DWRAT. It was left out of the Intermountain and Xeric monthly regional models, but its 513 observations were added to the single Sacramento basin model. Figure 3.1 shows the locations of these gages around California.

Figure 3.1. California reference gages used in developing natural flow models

Each reference gage's recorded average monthly flow rate in cfs is used as the dependent variable to be predicted. Following the example set by Grantham (2014), these are treated as independent observations, a simplifying assumption which is probably less false when using the monthly regional model approach rather than the geographic basin approach. (July's flow rate in year 1 is likely related to July's flow rate in year 2, but July's flow rate in year 1 is likely much *more* related to August's flow rate in year 1.) The flow rates come from the USGS surface water database, the National Water Information System (NWIS). Mean monthly flows were calculated from the daily flow records for each gage (Grantham, 2016). The predictor variables used to predict natural flow rate come from the GAGES-II database (Falcone 2011a) and consist of 210 different watershed and site characteristics calculated for each gage based on national data sources, including information on upstream topography, climate, geology, and soils. Not all of these variables were used in every model, but they were the starting set from which the included variables were chosen. For a complete list of these variables and their sources, see Appendix A.

To apply the USGS natural flow model for use in DWRAT, values for the set of predictor variables were calculated for locations of interest in the Sacramento River watershed (Grantham, 2015). These same data were used to apply the natural flow models developed for this thesis. More specifically, the USGS has systematically subdivided the United States into hydrologic sub-basins, the smallest of which is represented by a hydrologic unit code (HUC) with 12 digits, known as a HUC 12. These smallest subcatchments are the unit at which DWRAT currently operates, so to supply DWRAT with an estimate of water availability for each subcatchment, values for the predictor variables were calculated at the outlet of every HUC 12 sub-basin in the Sacramento watershed. The Sacramento sub-basins were specified by designating an outlet HUC 12 sub-basin at the bottom of the Sacramento River and then using an ArcGIS Python tool that finds all sub-basins upstream of the chosen outlet based on their systematic numbering scheme (Santos, 2015). Using Threemile Slough (whose 12-digit HUC is 180201630703) as the outlet resulted in a total of 768 sub-basins, shown in the map below.

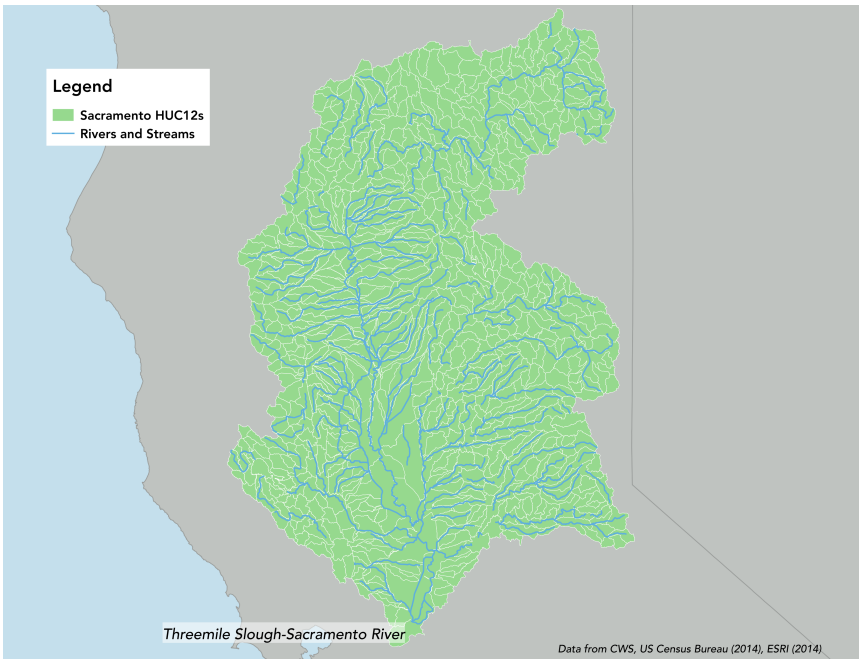


Figure 3.2. Sacramento Watershed Sub-basins

Proposed Natural Flow Models

Three sets of models were developed: dry-year monthly regional models to test whether a more curated dataset was helpful in predicting dry year natural flows, wet-year monthly regional models to see if the pattern held true for predicting wet year natural flows, and a Sacramento basin model to examine the effect of geographically restricting the data. The first two resulted in sets of 24 monthly models for the two aggregated ecoregions of interest (Xeric and Intermountain), while the last one was a single model that would hypothetically predict natural flow for any month or area of the Sacramento basin. Behind each of these proposed models is the idea that a natural flow model for the Sacramento River during drought can serve a narrower purpose. The existing USGS flow model is a general-purpose model, built on a dataset of all observed flows. Rather than using all the data, using only data from drier time periods could make the flow model more accurate for low-flow years. Whatever power is lost by the decreased sample size and variation might be more than compensated for by a "higher-quality" dataset that includes only the conditions of interest for prediction.

To test if dry-year models predicted dry year flows better than models using all years, the first step was to select relatively dry years. Annual precipitation data for the Sacramento drainage from 1950 to 2011 was used to rank water years from driest to wettest (NOAA, 2015). This assumed that years when the Sacramento watershed was relatively dry would also be relatively dry in the rest of California. All water years on the drier half of the ranking list were designated as dry years, while all water years on the wetter half of the ranking list were designated as wet years. The full dataset could then be reduced to a dry dataset of just those observations in dry years.

Model performance was evaluated using five-fold cross-validation. In this process, each observation is randomly assigned to one of five bins. Models are built with the data from four bins (80% of the data) and then evaluated using the data from the remaining bin (20%). This is repeated five times such that the data in each of the five bins is used once for evaluation. The test performance metrics calculated for each test bin can be averaged to give a more stable estimate of each statistic to estimate how the model performs on previously unseen data.

To correctly compare performance of a dry-year model to an all-year model, both models had to be evaluated based on their predictions for the *same* test dataset, even though their training datasets differed. Five-fold cross-validation had to be adapted slightly to handle comparing two datasets at once and ensure model comparability. For each fold, a 20%/80% test/train random split was performed on the dry dataset as in the standard five-fold cross-validation process described above. The 80% bin was the dry-year model's training set. The 20% bin was considered a "universal" test set. All observations in the full, all-year dataset not in the universal test set became the all-year model's training set for that fold. Figure 3.3's Venn diagram depicts how the all-year data, dry-year data, and universal test set datasets overlap.

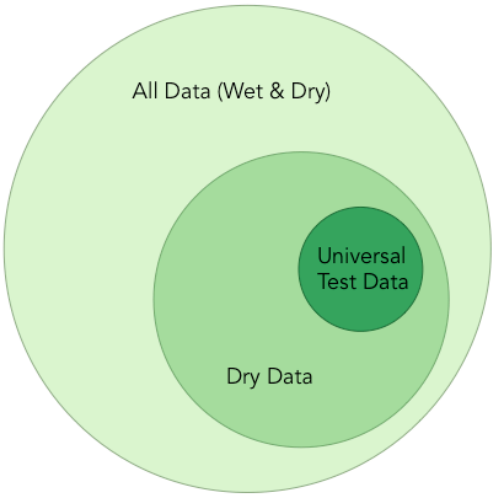


Figure 3.3. Venn Diagram of Dataset Partition in One Fold of Five-Fold Cross Validation

This process guaranteed that both modeling approaches were tested on how well they predicted the same dry year flows. Repeating this process for each bin of the dry dataset, sending the data through the sequence of chosen data transformations and machine learning algorithms (detailed in the final section of this chapter), and calculating performance metrics for each modeling combination based on its success or failure in predicting for the test set performed a five-fold cross-validation. The averaged performance metrics from each fold provided a stable test estimate of model performance. Redoing five-fold cross-validation for the wet years and creating a wet year universal test set then tested whether or not using a more curated dataset also helped more accurately predict wet year natural flows.

The process for the Sacramento basin model was simpler, performing 5-fold cross-validation for a single dataset rather than for two. The first few steps—performed using QGIS (2015)—were to determine which gages are in the Sacramento watershed and which aggregated ecoregions they belong to. For the latter, USEPA Level II ecoregions were aggregated following the guidelines from Falcone *et al*. (2010), since that paper's original shapefiles were unavailable (USEPA, 2008). There are 18 reference gages belonging to two aggregated ecoregions in the Sacramento watershed, with 11 of these in the full dataset made available for this thesis. Three of these 11 belong to the Xeric aggregated ecoregion; the remaining eight belong to the Intermountain aggregated ecoregion. Their distribution appears in the below map.

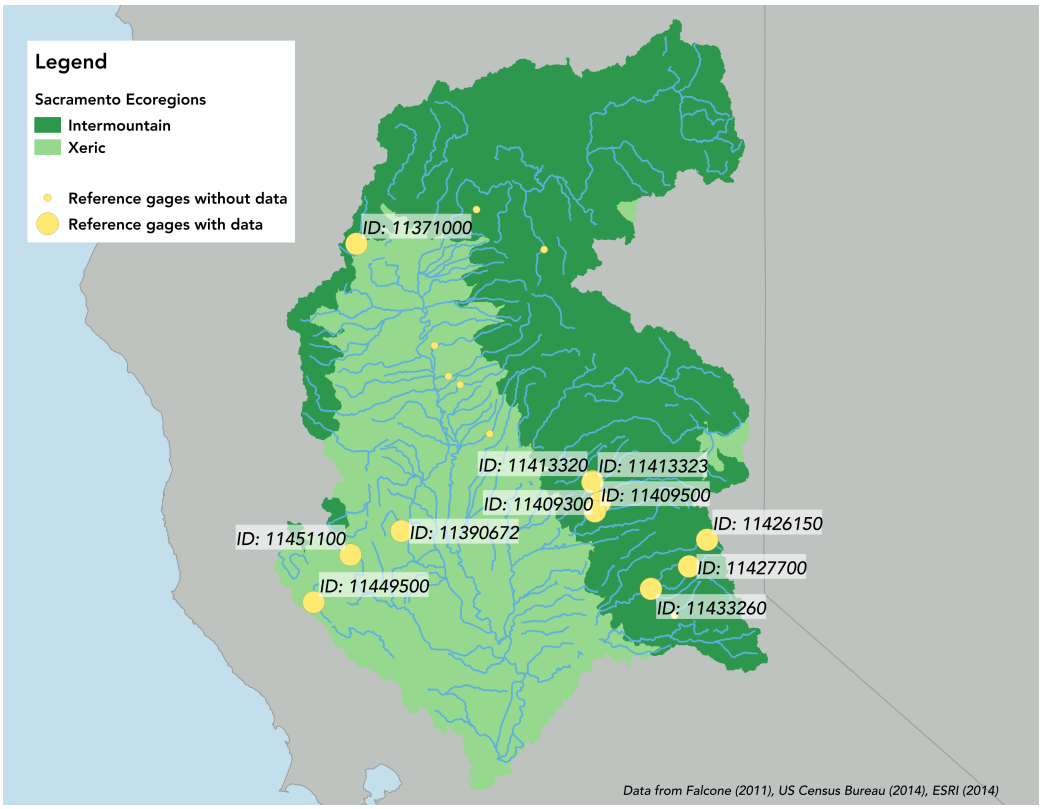


Figure 3.4. Sacramento Watershed Reference Gages in their Aggregated Ecoregions

Next, monthly regional datasets from the original USGS flow model were combined into a single dataset, adding columns containing binary variables that designated whether or not a given observation occurred in a given month or a given region. The dataset was subset to only observations whose station ID numbers matched the gages found to be in the Sacramento watershed in the process detailed above. As mentioned in the previous section, one gage in the northwestern Sacramento watershed (Clear Creek at French Gulch, ID #11371000) was classified as belonging to the Coastal Mountain region for the USGS model used in DWRAT, but it fell into the Intermountain region for this analysis based on the Falcone *et al*. (2010) aggregated ecoregions diagram. This Sacramento-specific dataset (with 4124 independent observations) was then randomly split into 5 folds, each fold was sent through the sequence of chosen data transformations and machine learning algorithms (detailed in the final section of this chapter), and performance metrics were calculated for each modeling combination based on that fold's test set predictions. Once again, the 5-fold cross-validation meant that the test performance metrics for each fold could be averaged to give a more stable estimate of each statistic to estimate how the model performs on previously unseen data.

Using *mlutilities* to Experiment with Machine Learning Approaches

This research provides a higher-level Python 3.x package called *mlutilities* built on top of the existing *scikit-learn* machine learning package for more easily experimenting with and evaluating various combinations of machine learning techniques. *Scikit-learn* is a Python package that integrates a wide range of machine learning algorithms for medium-scale

supervised and unsupervised problems (Pedregosa *et al*., 2011). *Mlutilities* uses the *scikit-learn* structure but adds functionality that facilitates manipulating datasets in multiple ways and applying various machine learning algorithms to evaluate their performance. It also makes extensive use of the *pandas* package for data manipulation. This section provides some brief examples of how *mlutilities* can be used in general to facilitate understanding how it was applied to predict natural flow. The full code for *mlutilities* can be seen at https://github.com/brmagnuson/ (as well as the code applying the package to this research).

Most *mlutilities* operations are built around a DataSet object, which is instantiated by giving it a description, pointing toward a comma-separated values file, and specifying information about the location of the dependent variable/predictor variables columns if necessary. The code below shows a simple example of reading in a dataset and splitting it into a training set and a testing set.

```
# Read data sets
myData = mlutilities.types.DataSet('My Training Data',
                                   pathToData)
splitData = mlutilities.dataTransformation.splitDataSet(myData,
                                                         testProportion=0.3,
                                                         randomSeed=89271)
trainingData = splitData.trainDataSet
testingData = splitData.testDataSet
```

Once a DataSet has been created, the tuneModels function is used to tune (e.g., calibrate) one or more models for one or more training DataSets. It expects a list of Datasets and a list of TuneModelConfigurations, which contain the necessary information to tune each individual model. For example, the below code sets up two models, a random forest model with possible tree numbers of 50, 75, and 100 and a k-nearest neighbor model that uses either 2 or 5 nearest neighbors. (These code snippets are purely for demonstration and were not used in the research for this thesis, so these potential parameter values are just arbitrary examples.) The tuneModels function performs a default 5-fold cross-validation grid search for each TuneModelConfiguration across all the possible combinations of supplied parameters (using *scikit-learn*'s default values for unspecified parameters). By default, the internal cross-validation grid search is scored based on R^2. It then returns a list of TuneModelResults, which contain the best set of model parameters for each model.

```
# Tune models for training data set
tuneScoringMethod = 'r2'

rfParameters = [{'n_estimators': [50, 75, 100]}]
rfMethod = mlutilities.types.ModellingMethod('Random Forest',
                                             sklearn.ensemble.RandomForestRegressor)
rfConfig = mlutilities.types.TuneModelConfiguration('Tune Random Forest',
                                                    rfMethod,
                                                    rfParameters,
                                                    tuneScoringMethod)
```

```
knnParameters = [{'n_neighbors': [2, 5]}]
knnMethod = mlutilities.types.ModellingMethod('K Nearest Neighbors',
                                              sklearn.neighbors.KNeighborsRegressor)
knnConfig = mlutilities.types.TuneModelConfiguration('Tune KNN',
                                                     knnMethod,
                                                     knnParameters,
                                                     tuneScoringMethod)

predictorConfigs = [rfConfig, knnConfig]
tunedModelResults = mlutilities.modeling.tuneModels([trainingData],
                                                    predictorConfigs)
```

Once the models have been tuned using training data, they can be applied to testing data using the applyModels function and their results scored on one or more performance metrics using the scoreModels function. For this demonstration, R^2 and the mean observed/expected ratio value were chosen as the performance metrics. Functions to evaluate the mean and standard deviation of the observed/expected ratio were created in *mlutilities* to supplement the existing *scikit-learn* performance metrics such as R^2 and mean squared error.

```
# Apply tuned models to some test data
applyModelConfigs = []
for tunedModelResult in tunedModelResults:
    applyModelConfig = mlutilities.types.ApplyModelConfiguration(
                            tunedModelResult.description,
                            tunedModelResult.modellingMethod,
                            tunedModelResult.parameters,
                            trainingData,
                            testingData)
    applyModelConfigs.append(applyModelConfig)
applyModelResults = mlutilities.modeling.applyModels(applyModelConfigs)

# Score test results
r2Method = mlutilities.types.ModelScoreMethod('R Squared',
                                              sklearn.metrics.r2_score)
meanOEMethod = mlutilities.types.ModelScoreMethod('Mean O/E',
                                  mlutilities.modeling.meanObservedExpectedScore)
testScoringMethods = [r2Method, meanOEMethod]
testScoreModelResults = mlutilities.modeling.scoreModels(applyModelResults,
                                                         testScoringMethods)
```

To easily display results, scoreModelResults can be converted to a *pandas* DataFrame, which can be printed out, written to a CSV file, or visualized.

```
scoreModelResultsDF = mlutilities.utilities.createScoreDataFrame(
                                                              testScoreModelResults)
mlutilities.utilities.barChart(scoreModelResultsDF, 'R Squared',
                'R Squared for Each Model', 'ExampleData/rSquared.png', '#2d974d')
```

The output of the above barChart function generates a simple chart with a bar for each model, numbered by its index in the results DataFrame.

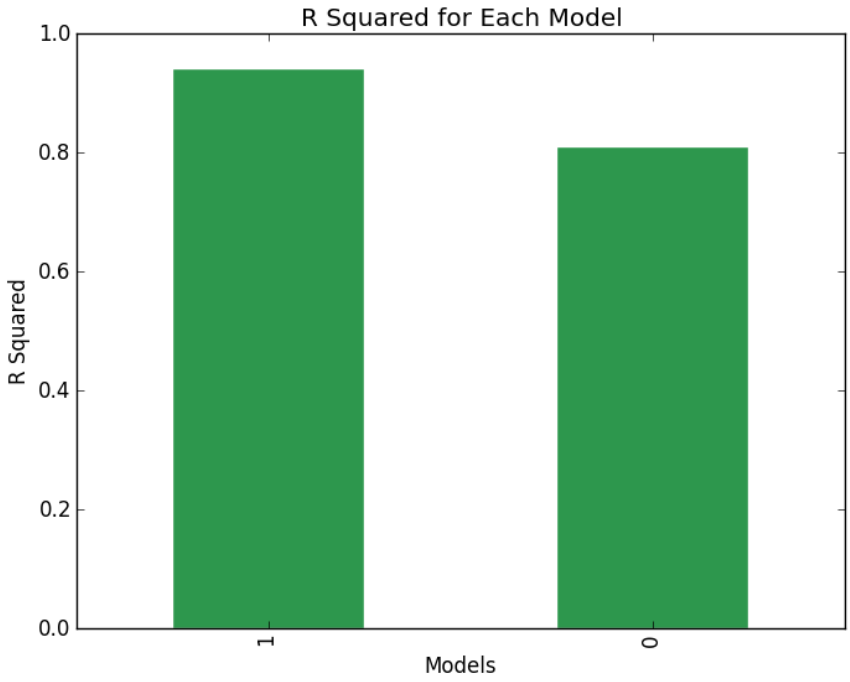


Figure 3.5. Example Random Forest vs. K-Nearest Neighbor R^2

Mlutilities also allows for DataSets to be scaled from 0 to 1 before any models are used, as some machine learning algorithms perform better with numbers reduced to this range, and for other DataSets to be scaled using the same scaling process. Use of the scaleDataSet and scaleDataSetByScaler functions is demonstrated below.

```
# Scale data
scaledTrainingData, scaler = mlutilities.dataTransformation.scaleDataSet(
                                                                trainingData)
scaledTestingData = mlutilities.dataTransformation.scaleDataSetByScaler(
                                                            testingData, scaler)
```

Likewise, *mlutilities* can be used to perform feature engineering (a term coined for this thesis to refer to the set of possible variable selection and dimensionality reduction techniques) on a DataSet before a model is applied. This consists of either variable selection approaches, such as using a variance threshold technique to only keep predictor variables that have relatively high variance or specifying a list of predictor variables to keep, or variable extraction approaches, such as principle component analysis (PCA) or independent component analysis (ICA) to generate a reduced set of predictor variables through dimensionality reduction. The same feature engineering process can then be used to transform another DataSet as well. The

below code demonstrates this process, using PCA to create new DataSets with 5 principle components.

```
# Perform feature engineering
pcaConfig = mlutilities.types.FeatureEngineeringConfiguration('PCA n5',
                    'extraction', sklearn.decomposition.PCA, {'n_components': 5})

pcaTrainingData, transformer = mlutilities.dataTransformation.\
    engineerFeaturesForDataSet(trainingData, pcaConfig)
pcaTestingData = mlutilities.dataTransformation.engineerFeaturesByTransformer(
                                                        testingData, transformer)
```

Mlutilities also provides a way to average or stack models to test improvement due to model aggregation. An example of building a StackingEnsemble is shown below. The tunedModelResults for the random forest and k-nearest neighbors models are processed to extract the tuned model configurations. These are then used as the base predictors for the stacking ApplyModelConfiguration, with the random forest model arbitrarily chosen (by specifying `predictorConfigs[0]`) as the second-level model that predicts based on the base predictors' results. (An AveragingEnsemble can also be created in a similar way. For that approach, rather than using a second-level model, the base predictors' results are all averaged together for a final prediction. By default, a regular arithmetic mean is calculated, but if desired, the user can specify weights for each base predictor, which results in a weighted average instead.)

```
# Create stacking ensemble
predictorConfigs = []
for tunedModelResult in tunedModelResults:
    predictorConfig = mlutilities.types.PredictorConfiguration(
                          tunedModelResult.modellingMethod.description,
                          tunedModelResult.modellingMethod.function,
                          tunedModelResult.parameters)
    predictorConfigs.append(predictorConfig)

stackMethod = mlutilities.types.ModellingMethod('Stacking Ensemble',
                                                mlutilities.types.StackingEnsemble)
stackParameters = {'basePredictorConfigurations': predictorConfigs,
                   'stackingPredictorConfiguration': predictorConfigs[0]}
stackApplyModelConfig = mlutilities.types.ApplyModelConfiguration(
    'Stacking Ensemble', stackMethod, stackParameters, trainingData, testingData)

stackResult = mlutilities.modeling.applyModel(stackApplyModelConfig)
```

Using *mlutilities* to Predict Natural Flow

Combining all the above techniques together in sequence allows for exploration of a great variety of dataset-algorithm combinations and evaluating how they perform relative to each other. The dataset for each proposed model was randomly split into testing/training sets for 5-

fold cross validation as detailed above. (Throughout the research, a constant random seed was used whenever randomness came into play to ensure reproducible results.) Each fold was sent through the same sequence of scaling the data, using multiple variable selection and dimensionality reduction techniques, tuning various machine learning algorithms to the resulting datasets, applying the same combinations to the matching test sets, and then scoring the combination's predictive performance. The resulting test performance metrics for each fold were averaged to give a more stable estimate of each statistic to estimate how the model performs on previously unseen data. This high-level sequence can be seen in the below flow diagram. The rest of this section details each step in the sequence.

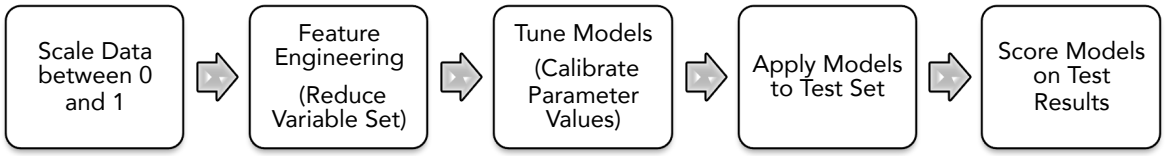


Figure 3.6. Sequence of Applying Dataset-Model Combinations to a Given Dataset Fold

Each predictor variable column of the original dataset (or datasets, in the case of a dry-year/wet-year model that starts with both the full dataset and a reduced dataset) is linearly scaled from 0 to 1, and the same scaling math is applied to a matching copy of the test dataset. (The test dataset's values may go beyond 0 and 1, since its scaling is still based on the range of the training dataset.) The original and scaled training datasets, all of which have more than 200 predictor variables, are then sent through four different feature engineering approaches to reduce to a smaller subset of predictor variables. The same transformation is then applied to the matching copy of the test dataset. The four chosen feature engineering approaches (methods to select variables or reduce dimensionality) were:

1. PCA for 20 components.
2. PCA for 50 components.
3. Removing all predictor variables with a variance below a 0.08 threshold.
4. Retaining a list of predictor variables that, based on their definitions, were thought to be most relevant to predicting natural flows. (This method was referred to as "Expert Selection" in the code, although the author professes only rudimentary expertise in hydrology.) These consisted of various predictor variables recording information about precipitation, temperature, soil, and topography. Appendix B contains a full list of variables retained in this feature selection approach.

These scaling and feature engineering operations rapidly multiplied the number of datasets to be sent through the next step in the sequence, model tuning. Figure 3.7 shows how a training dataset multiplied for a given dry-year monthly regional model because of creating the reduced dataset and then applying data transformation methods to each dataset. After starting with a single dataset, the scaling and feature engineering steps result in a total of 20 different datasets.

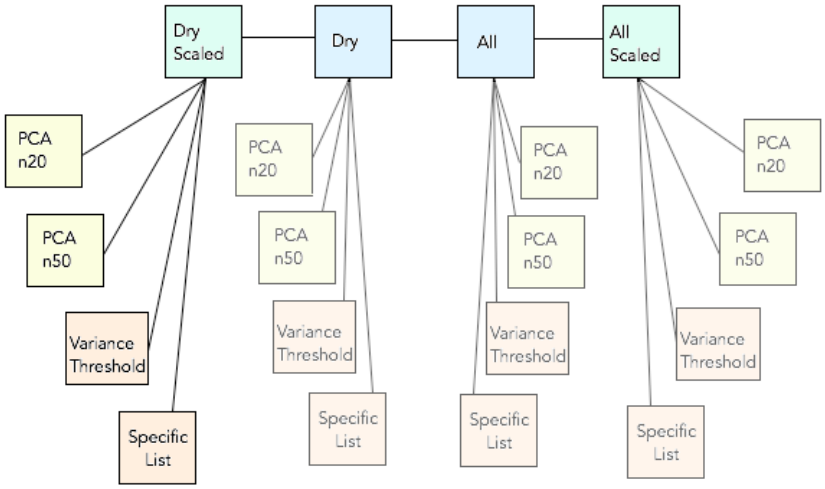


Figure 3.7. Dataset Transformation for a Given Monthly Regional Model Dataset

Six different machine learning algorithms are then tuned for each training dataset using 5-fold cross-validation and a grid search across the given parameter space. These algorithms were chosen as the full set of available *scikit-learn* regression algorithms that could process the dataset quickly, although ridge regression was chosen as the only linear model. Table 3.1 shows the algorithms and their possible parameter sets; unspecified parameters used default values. Algorithms employing randomness were passed a random seed of 47392 (a randomly chosen number) to ensure repeatable results.

Table 3.1. Chosen Scikit-learn Algorithms and Their Possible Parameter Sets

Algorithm	Possible Parameters
AdaBoost	n_estimators: 50, 100 learning_rate: 0.5, 1.0 random_state: 47392
Decision Tree	max_features: ‘sqrt’, ‘auto’ random_state: 47392
K-Nearest Neighbors	n_neighbors: 2, 5, 10 metric: ‘minkowski’ weights: ‘uniform’, ‘distance’
Random Forest	n_estimators: 50, 75, 100 max_features: 10, ‘sqrt’ random_state: 47392
Ridge Regression	alpha: 0.0, 0.1, 0.5, 1.0 normalize: True, False
Support Vector Machine	C: 1.0, 10.0 epsilon: 0.1, 0.2 kernel: ‘rbf’, ‘sigmoid’

Once the six base machine learning algorithms are tuned, they are used to build three more ensemble models. For more details on ensemble methods, see Wolpert (1992), Breiman (1996), Clarke (2003), and Sill *et al*. (2009). The three chosen ensemble models consist of:

1. An averaging ensemble that averages the predictions of each base algorithm weighted by that algorithm's estimated R^2 from the tuning process (as long as $R^2 > 0$).
2. A stacking ensemble that uses the base algorithm with the highest estimated R^2 from the tuning process to predict natural flow based on the predictions of all base algorithms, meaning that for the second-level model, the predictor variables consisted of the flow predictions from each of the 6 models in Table 3.1.
3. A stacking ensemble identical to the above except that it also includes all the original predictor variables that the base algorithms were training on in the first level as predictor variables in the second-level model.

This results in 9 machine learning algorithms for each training dataset. Each algorithm is trained on the full training dataset with the tuned parameters and then applied to the test dataset to make predictions for known but previously unseen data. These predictions, or expected values, are then compared to the actual observed values and used to calculate five different model performance metrics (which are later averaged across folds for each dataset-model combination to get a more stable estimate of each metric):

1. R^2
2. Mean O/E value
3. Standard deviation of O/E values
4. Mean squared error (MSE)
5. Root mean squared error (RMSE)

R^2 measures the proportion of variation in the observed data explained by the model. There are multiple definitions of R^2, but throughout this thesis, the formula used for R^2 is:

$$R^2 = 1 - \frac{Sum\ of\ squared\ residuals}{Total\ sum\ of\ squares}$$

This means that negative R^2 values are possible when a model fits the data worse than simply predicting the dependent variable's observed mean. The O/E metrics give an idea of if the model tends to over- or under-predict flow and by how much of a relative margin. The RMSE gives the magnitude of the average error in flow prediction in cfs.

This set of scores could then be averaged for each dataset-algorithm combination with the corresponding scores from the other folds, resulting in a list of estimated performance metrics for each dataset-algorithm combination. R^2 was treated as the primary metric, meaning the dataset-algorithm combination with the highest R^2 was then chosen as the "best" model, although it was worthwhile to examine each best model's performance on the other performance statistics to ensure it performed well on all dimensions. Performing this process for each monthly regional scenario resulted in a set of 48 best models (12 monthly models for 2 regions for both the dry-year approach and the wet-year approach), while performing it for the Sacramento River basin, a

second method of restricting the data added to this research near the end of the process, resulted in a single best model.

These best models could then be evaluated relative to the USGS natural flow models' 5-fold cross-validated estimated test performance on the same datasets. If model performance had increased by using the best model derived from trying all the combinations, then the new models could be treated as more accurate predictors of dry year natural flows than the existing USGS models.

A Note on Combinatorics

Employing more feature engineering and machine learning algorithms in the modeling sequence expands the number of combinations to try. During initial development, only two feature engineering methods and three machine learning algorithms were used to streamline code testing, and a single fold took only 1-2 minutes to run on a personal computer thanks to parallelization. However, once all combinations were added in, a single fold took approximately 9 minutes to run on a personal computer, so the code to try out all the combinations was run using an Amazon Web Services server with 36 cores. This decreased a fold's run time by almost an order of magnitude to only 1 minute. The results were the downloaded from the cloud for further analysis.

Considering that running the above sequence for the Sacramento basin as well as for dry and wet years in the Intermountain and Xeric regions required training models an estimated 52,920 times,[*] making full use of parallelization techniques and cloud computing was essential for timely performance.

[*] (2 wet/dry scenarios * 12 months * 2 regions + 1 Sacramento basin approach) * (20 datasets * 9 machine learning algorithms * (5 parameter tuning trainings + 1 model application training)) = 52,920 model trainings

CHAPTER 4: RESULTS

Overview

This chapter describes the results of developing the various proposed natural flow models and how they compared to the original USGS natural flow model used in DWRAT. Some detailed results output by *mlutilities* are presented first, using the dry-year July Intermountain scenario as a representative example. Next, overall model performance is discussed, showing improvement relative to the USGS natural flow model when predicting for dry years and that the monthly regional model set approach is better than the geographic basin approach. The chapter concludes with an experimental application of the new models to the main stem of the Sacramento River.

Example Case: Dry-Year Intermountain July Output

The *mlutilities* package produces a large amount of data for any given model run. This section discusses only the dry-year Intermountain July results as a representative example. Examining detailed output for a dry July in the Intermountain ecoregion is an interesting example because July tends to be the start of the dry season, after the snowmelt period is over and low baseflow conditions start, and because most of the Sacramento River's water comes from the Sierra Nevada Mountains. If curtailments were to happen in a year, they would usually be underway by July. Appendix C has charts of cross-validation results for each scenario (the dry-year regional monthly models, the wet-year regional monthly models, and the geographically-restricted Sacramento basin model).

First, we examine results for a single fold to see the low-level results before averaging for cross-validation estimates. Table 4.1 shows the top output of *mlutilities* for one of the five folds to show *mlutilities*' most detailed outputs; the averaged, cross-validated results are in Table 4.2. Because 180 model-dataset combinations were tried, only those with the highest 10 R^2 values are shown here. The first column describes the base data and any transformations applied to it. ("Features selected via…" means that the predictor variables were selected via the named feature engineering method.) The second column names the modeling method used. ("Stacking OF Ensemble" refers to a stacking model that, in addition to the base algorithms' predictions, also includes as predictor variables all the original predictor variables that the base algorithms were training on in the first level as predictor variables in the second-level model.) The remaining columns display the models' performance metric results; RMSE gives the average magnitude of the error in cubic feet per second. These results are specific to this fold and will look somewhat different once the results for each model-dataset combination have been averaged across the five folds.

Table 4.1. Top 10 Dry-year July Intermountain Models from Fold 2

Base DataSet	Model Method	R^2	Mean O/E	Std. Dev. O/E	MSE	RMSE (cfs)
Dry Years Scaled	Stacking Ensemble	0.889	1.137	0.842	5729.818	75.696
Dry Years	Random Forest	0.887	0.851	0.471	5800.291	76.160
Dry Years Scaled	Random Forest	0.887	0.851	0.471	5800.711	76.162

Dry Years Scaled	Averaging Ensemble	0.872	-0.046	12.633	6588.165	81.168
Dry Years Scaled	Stacking OF Ensemble	0.870	0.893	0.472	6700.249	81.855
Dry Years features selected via Variance Threshold .08	Random Forest	0.868	0.815	0.469	6812.465	82.538
All Years Scaled features selected via PCA n50	K Nearest Neighbors	0.867	0.803	0.409	6863.459	82.846
Dry Years	Averaging Ensemble	0.862	0.838	2.750	7098.036	84.250
Dry Years	Stacking Ensemble	0.854	1.022	0.748	7540.411	86.836
All Years Scaled	K Nearest Neighbors	0.849	0.798	0.408	7759.998	88.091

For this fold, most of the best-performing models are dry-year models, and scaling appeared to increase predictive accuracy, although PCA and the variance threshold method each make an appearance. The stacked ensemble model had the highest R^2 value of 0.889, which seems to be a respectable result. However, based on the mean O/E results, it tends to under-predict flow, since a ratio greater than 1 means that the observed value is larger than the predicted value. Other ensembles, random forests, and k-nearest neighbors are the other top performers.

To understand several general trends for the full spread of results for the same fold from the dry-year July Intermountain model, the below figures graph R^2 versus mean O/E and show how different dataset-model combinations performed relative to each other. The left portion of each figure represents the full results set, while the right portion is zoomed in to show detail for the main cluster of results. Negative R^2 values are possible when a model fits the data worse than simply predicting the dependent variable's observed mean. Negative mean O/E values occur when using linear models such as ridge regression that extrapolate beyond the range of observed values for the dependent variable in the training set. Figure 4.1 shows that for this fold, dry-year datasets as a group tended to have much better R^2 values than all-year datasets, although their mean O/E values range more widely from the ideal value of 1. In fact, many of the all-year datasets result in a negative R^2 value for the test data.

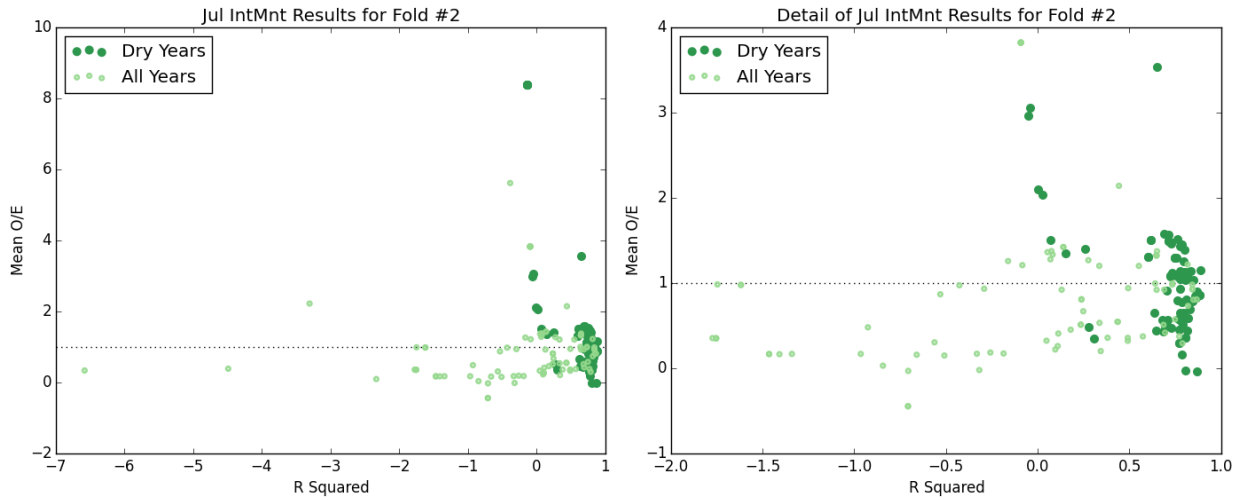


Figure 4.1. Dry vs. All Approach, Dry-year July Intermountain Fold 2 Test Performance[*]

Figure 4.2 shows that for this fold, an ensemble model performs best. Ensemble model methods[*] generally perform well in terms of R^2, although there are a few poor performers. Using ensemble methods seems to stabilize the mean O/E values. In contrast, non-ensemble methods have a broad range of performance metric values.

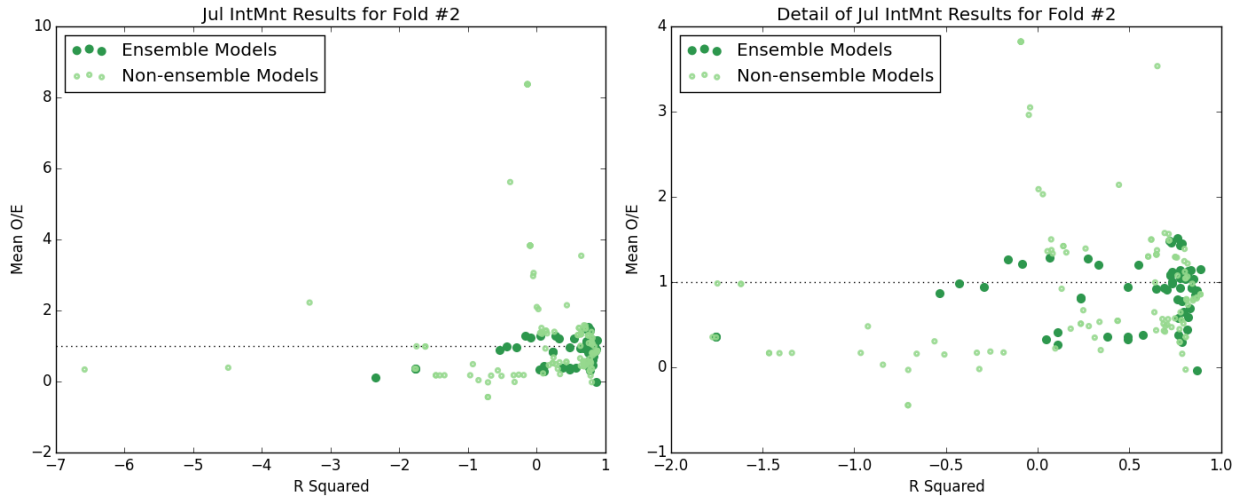


Figure 4.2. Ensemble vs. Base Models, Dry-year July Intermountain Fold 2 Test Performance

Now, moving on to the cross-validation results for the dry-year Intermountain July scenario, we can see how the final results changed from those of an individual fold. The top 10 models ranked by R^2 value are shown in Table 4.2.

[*] All graphs in this thesis were made using Python's *matplotlib* library (Hunter, 2007).

[*] Averaging ensembles and the two versions of stacked ensembles. Technically, random forests are an ensemble of multiple decision trees, but "ensemble" is used here to mean the modeling methods employed to combine the results of six *different* base modeling methods.

Table 4.2. Top 10 Cross-validated Dry-year July Intermountain Models

Base DataSet	Model Method	R^2	Mean O/E	Std. Dev. O/E	MSE	RMSE (cfs)
Dry Years features selected via PCA n50	Stacking Ensemble	0.784	0.891	0.687	8006.155	85.999
Dry Years features selected via PCA n50	Stacking OF Ensemble	0.767	1.246	1.855	8506.871	89.280
Dry Years	Stacking Ensemble	0.766	0.875	0.694	8271.001	88.602
Dry Years	Stacking OF Ensemble	0.763	1.250	1.832	8654.498	90.711
Dry Years features selected via Variance Threshold .08	Stacking OF Ensemble	0.762	1.253	1.833	8929.398	90.528
Dry Years features selected via Variance Threshold .08	Stacking Ensemble	0.760	0.885	0.658	8962.852	90.347
Dry Years features selected via PCA n50	Decision Tree	0.752	4.354	23.863	8307.093	90.063
All Years features selected via Expert Selection	K Nearest Neighbors	0.749	1.026	1.643	9588.479	94.562
Dry Years Scaled	Stacking OF Ensemble	0.747	0.968	0.863	8642.635	90.340
Dry Years features selected via PCA n50	Averaging Ensemble	0.747	0.542	0.530	9277.459	93.269

In general, R^2 values have decreased somewhat (from a top value of 0.889 to 0.784) with proportionate increases in MSE and RMSE. This shows the stabilizing effect of using multiple test sets, rather than relying on a single test set that a model is better-suited to by chance than it would be to other random test sets. The O/E metrics seem to have remained fairly consistent. Once again, dry-year datasets predominate. PCA and variable selection using a variance threshold have a much stronger presence than they did in Fold 2, as do the various ensemble modeling methods. A stacking ensemble remains the top performer.

Graphs of the full cross-validation estimates show the same broad trends. Dry-year datasets tend to perform much better than all-year datasets, many of which have negative test R^2 values (showing that these results were not a fluke of a single test set). Ensemble methods seem to have moved up in the ranks, with fewer poor performers as in Fold 2. (As one might suspect, the remaining poor performers are trained on an all-year dataset.) See Figures 4.3 and 4.4. The left portion of each figure represents the full results set, while the right portion is zoomed in to show detail for the main cluster of results.

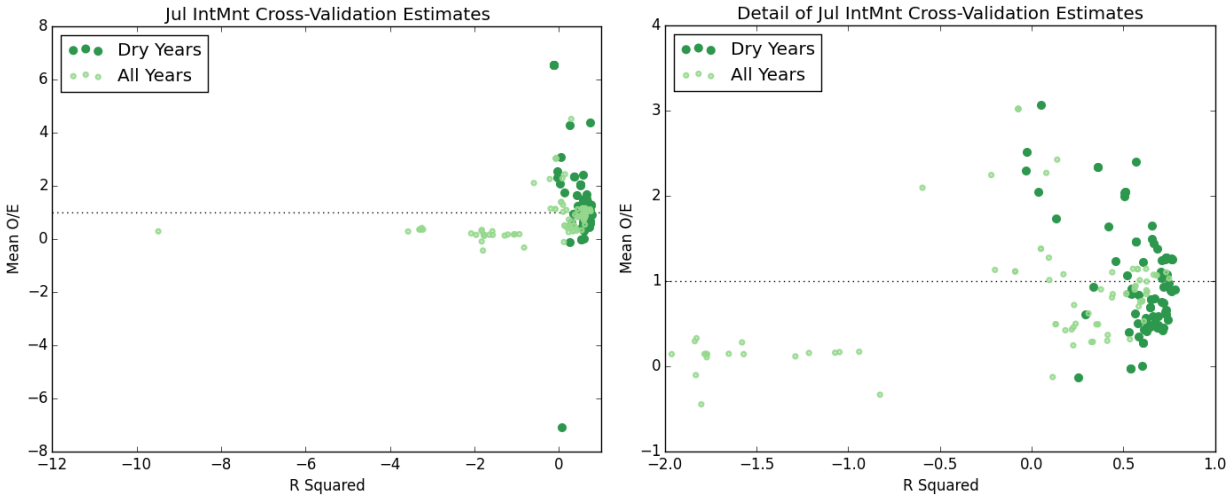


Figure 4.3. Dry vs. All Approach, Dry-year July Intermountain Cross-Validation Estimates

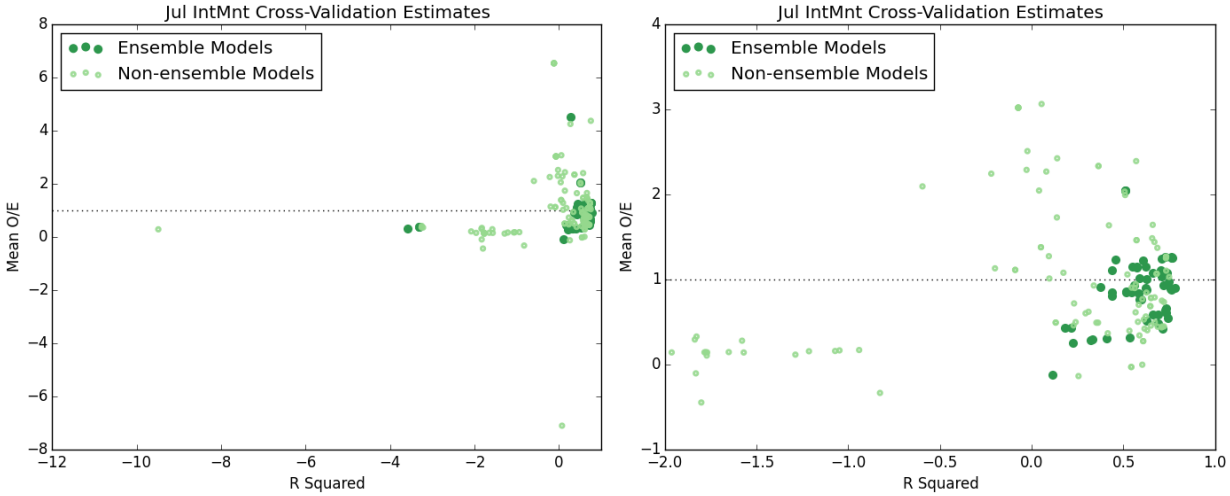


Figure 4.4. Ensemble vs. Base Models, Dry-year July Intermountain Cross-Validation Estimates

It is also worth looking at an example of how scaling and feature engineering affected model performance. Figure 4.5 demonstrates how scaling and feature engineering methods affected performance of the best model for the July Intermountain scenario, a stacking ensemble trained on a dry-year dataset and with predictor variables selected via PCA with 50 components. The points represent what one could consider the best model's "relatives"—all the combinations of different versions of the dry-year dataset with a stacking ensemble—and show their performance in terms of R^2 and mean O/E value.

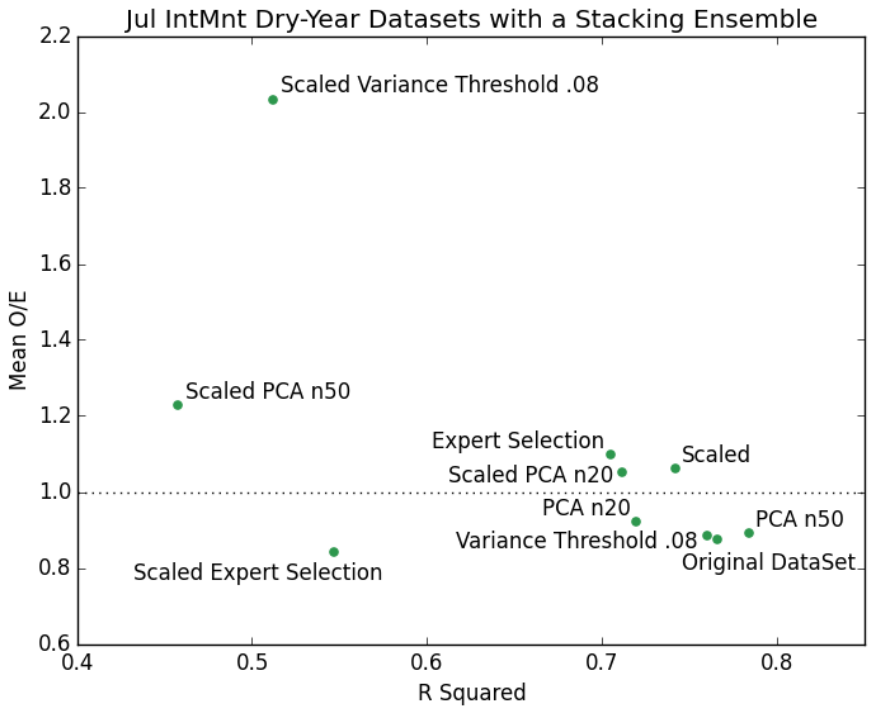


Figure 4.5. Effect of Scaling and Feature Engineering on Dry-year July Intermountain Models

At least for this scenario, there seem to be few easily distinguishable trends in how scaling, dimensionality reduction, and variable selection affected performance. Although PCA with 50 components performed best based on R^2 for the July Intermountain scenario, not every dataset that employed PCA performed well, and the original dry-year dataset with no scaling, dimensionality reduction, or variable selection was the second-best R^2 performer among the relatives. However, in this scenario, all base datasets that combined scaling with some form of feature engineering—either dimensionality reduction or variable selection—are in the bottom five of the relatives ranked by R^2. While the "expert selection" method (choosing specific predictor variables that, based on their definitions, were thought to be most relevant to predicting natural flows) resulted in generally respectable performance metrics, validating the list of chosen predictor variables as an acceptable one, it was far from the top performer among the best model's relatives.

Best Monthly Regional Models

This section presents overall results and general trends in model performance for all best monthly regional models (meaning the model with the top R^2 value for each month in each region).

For the dry-year Intermountain scenario, the top-ranked model for two-thirds of the months was trained on a dry-year dataset, with the notable exception of three spring months: April, May, and June (as well as November). This difference might be explained by these months representing the peak season of Sierra runoff flow, when the snowpack melts and releases torrents of water to rivers. High runoff months likely see more variable flows, so training the model on the complete dataset would give the model more of these variable examples to learn

from. However, for the dry-year Xeric scenario, nearly all the top-ranked models were trained on the full all-year dataset. Likewise, for the wet-year scenario, most top-ranked models were also trained on the full all-year dataset rather than the more restricted wet-year dataset. See Figure 4.6. In these cases, then, the greater variation and power of additional data must have been more important than the type of that data.

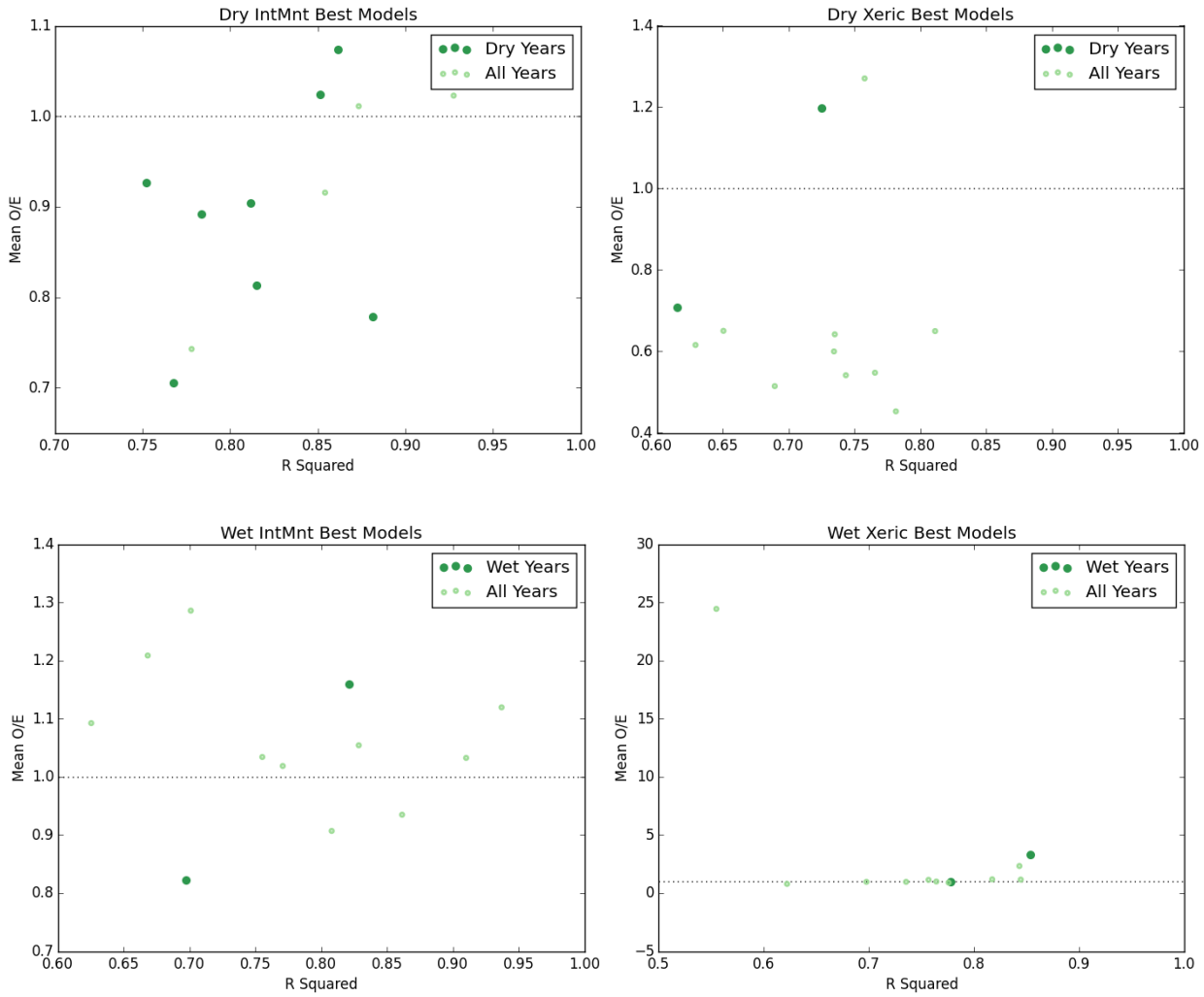


Figure 4.6. Best Monthly Regional Models: Restricted vs. Complete Datasets

Across all scenarios, the stacking ensemble approach—most often including the original predictor variables as part of the second-level model, although not always—was by far the most frequent top performer. (See Figure 4.7.) It was the top performer for three-quarters of the dry scenarios and five-sixths of the wet scenarios. This demonstrates the power of stacking models together and allowing them to correct for each other's biases. Random forests were the next-most frequent top performer, reinforcing that the USGS chose well when using it for their natural flow model (Carlisle *et al.*, 2010), followed by averaging ensembles and k-nearest neighbors. Unsurprisingly, the single linear model, ridge regression, was never a top performer. This makes

sense, as streamflow is hardly a linear phenomenon. Support vector machines, AdaBoost, and single decision trees also never made an appearance as top-performing algorithms.

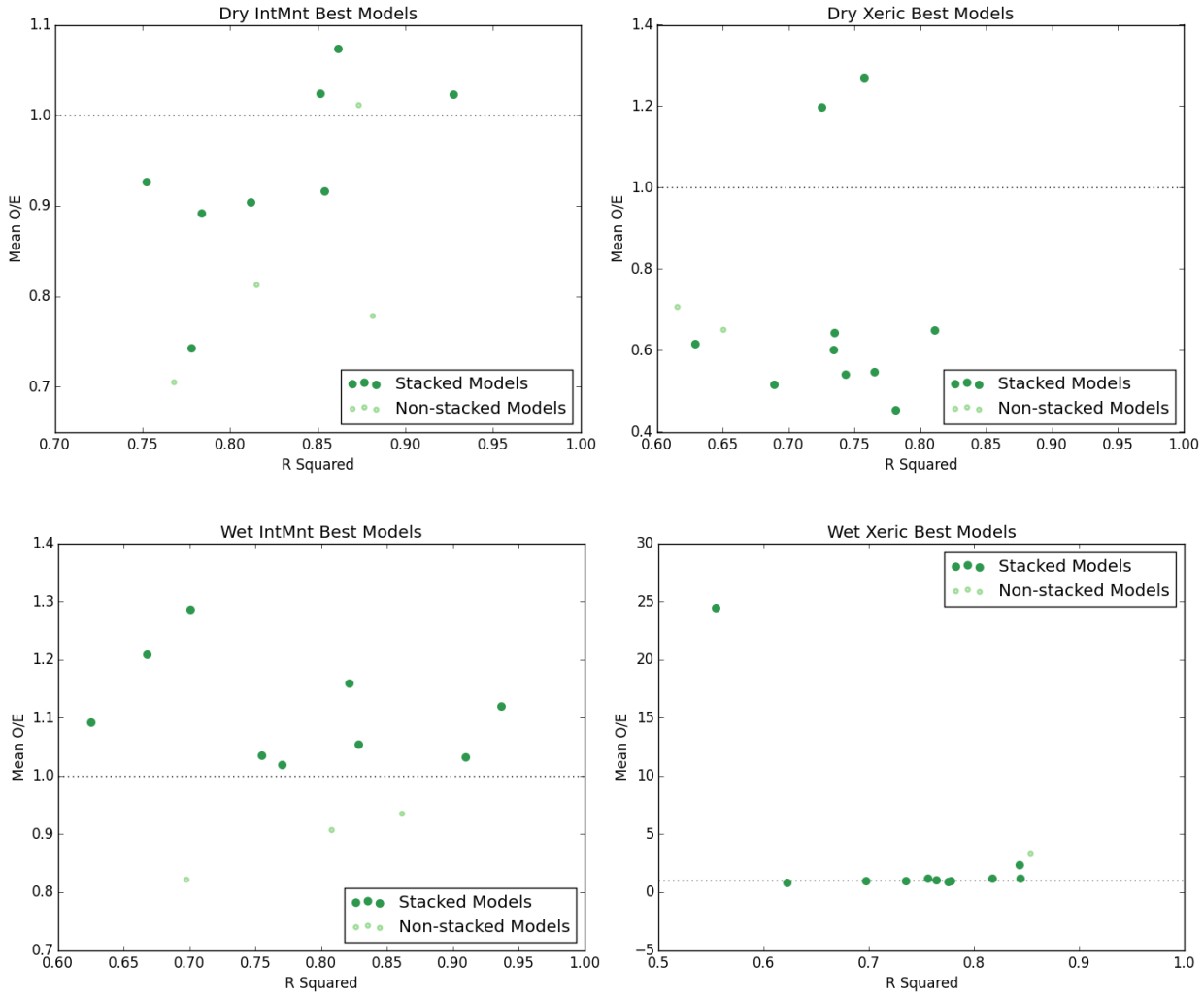


Figure 4.7.Best Monthly Regional Models: Stacked vs. Non-stacked Models

Of the 48 best monthly regional models, 23 of their training datasets used scaling, although that was combined with a feature engineering method in only 3 of the best models. This continued the trend seen in the detailed results for the dry-year July Intermountain scenario that it was usually better to either scale or engineer predictor variables rather than to do both. The "expert selection" variable selection method topped the list 6 times, reinforcing that the list of predictor variables chosen based on their definitions and likelihood to affect natural flow was a respectable one. The variance threshold method of variable selection proved far more popular than dimensionality reduction, as it showed up four times more often than PCA's two top model showings. Approximately one quarter of the time, it was best to simply do nothing and use all the predictor variables, at least according to R^2 rankings.

Appendix D contains complete tables of best monthly regional models. Each scenario's performance metrics were averaged by region and are briefly summarized in Table 4.3. The

highest R^2 values were obtained for the dry-year Intermountain models on average, and the lowest for the dry-year Xeric models, but the average R^2 values are quite respectable across the board. The magnitude of the average error seems reasonable, especially when predicting for larger streams. The dry-year models tend to have mean O/E values below 0, meaning they over-predict flow slightly, while the wet-year models tend to under-predict flow. The wet-year Xeric scenario has an egregiously high average standard deviation of the O/E values, but this is due to the October model's results; the average of all other months' values for that column is 4.409.

Table 4.3. Best Monthly Regional Model Average Performance

Scenario	Region	Mean Performance Metrics				
		R^2	Mean O/E	Std. Dev. O/E	MSE	RMSE (cfs)
Dry	Intermountain	0.830	0.900	0.722	9,259	71
	Xeric	0.720	0.699	2.038	656	17
Wet	Intermountain	0.782	1.056	0.864	41,160	165
	Xeric	0.754	3.260	28.957	4,600	46

Comparison to USGS Models for Dry-year Prediction

Given all these models, are they any more accurate in predicting for dry years than the models created by the USGS? This question was answered by recreating the original model in Python using the variables and random forest parameters from Grantham (2014) and running it through the exact same cross-validation process. Because the random shuffling and splitting of the data into folds had been done based on a known seed value to initialize the randomization and ensure repeatability, the USGS models saw the same training and testing data, so its estimated test performance is directly comparable to that of the above models. Figure 4.8 compares the dry-year performance of the new models to those of the USGS models for the Intermountain region.

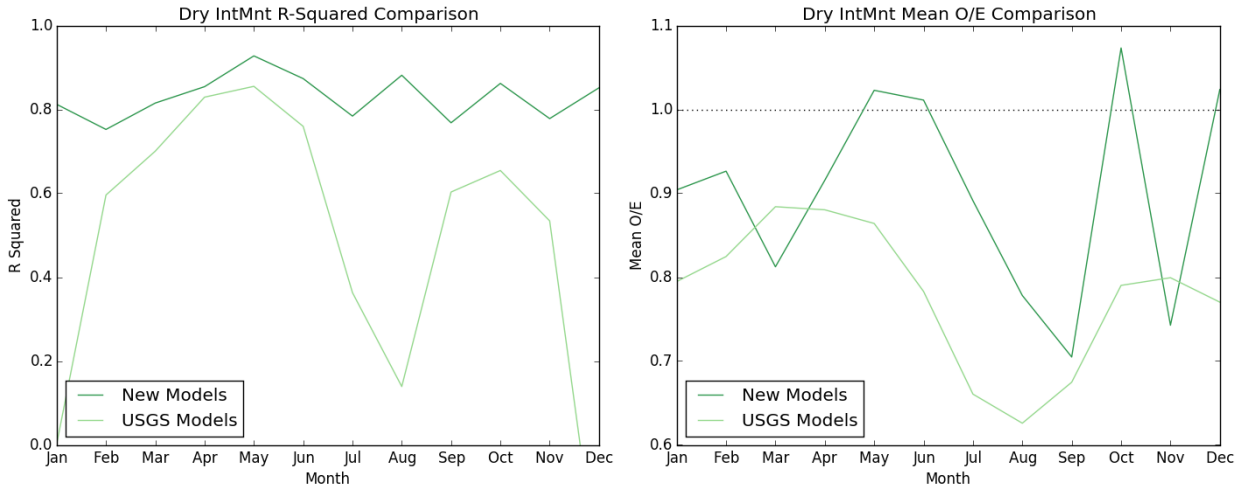


Figure 4.8. Performance Comparison with USGS Models for Dry Years: Intermountain Region

The new Intermountain models have higher R^2 values than the USGS models in every month. In some months, particularly July, August, December, and January, the improvement in model fit is quite large. The USGS model actually gets a negative R^2 value for December. For all except March and November, the mean O/E value is closer to the ideal of 1. This all demonstrates that the new models appear to be more accurate when predicting for dry years than the USGS models. Seen below, Figure 4.9 demonstrates that the new models make fewer improvements for the Xeric region over the USGS models. In this case, eight months show an improved R^2 for the new models, while November through February show a slight decrease in R^2. Mean O/E values show some improvements for the summer and early winter months but not in the spring.

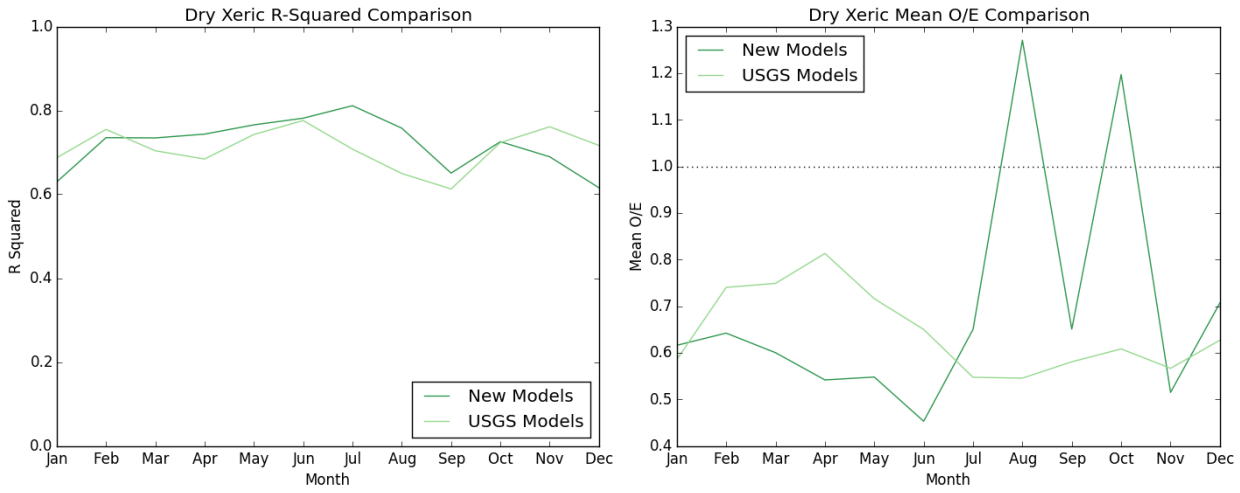


Figure 4.9. Performance Comparison with USGS Models for Dry Years: Xeric Region

These comparisons demonstrate the value of the improvements added to the modeling process for dry-year prediction. While the gains are not dramatic in every case, they generally point to fairly equivalent or improved accuracy, and predictions for some months should now be much improved. Given that water right curtailments are likely to occur in July and August if they occur for any month in a year, using the new flow models would likely improve DWRAT's suggested curtailment decisions compared to using the USGS model as input to DWRAT.

This process was not repeated for the wet-year models, since creating a more accurate dry-year model was the focus of this research. Therefore, the wet-year models cannot be said to be more accurate than the USGS models at this time.

Best Sacramento Basin Model

Restricting the data geographically to the Sacramento basin rather than by month and region resulted in a single best model that uses only the 11 Sacramento watershed reference gages for which predictor variable information was available as training data. While the monthly regional models developed for this research had shown improved accuracy over the USGS natural flow model, it seemed worthwhile to see if even further accuracy gains could be eked out by restricting the training data geographically. Its performance and that of the four next-best models are presented in Table 4.4.

Table 4.4. Top 5 Sacramento Basin Models

Base DataSet	Model Method	R^2	Mean O/E	Std. Dev. O/E	MSE	RMSE (cfs)
Sacramento Basin Scaled	Stacking OF Ensemble	0.841	0.849	0.520	5648.109	74.707
Sacramento Basin Scaled	Stacking Ensemble	0.820	1.018	0.806	6376.173	79.531
Sacramento Basin, Variance Threshold .08	Stacking OF Ensemble	0.816	0.802	0.612	6514.261	80.067
Sacramento Basin Scaled	Averaging Ensemble	0.815	0.443	5.103	6522.098	80.425
Sacramento Basin	Averaging Ensemble	0.814	0.365	4.620	6564.478	80.675

These are very encouraging estimated test performance metrics—one of the higher test R^2 value for any best model and a mean O/E near 1. As was seen with the monthly regional models, scaling or variable selection via the variance threshold method tended to do better most often, and some form of stacking ensemble predominated.

However, concerns about the generalizability of this model to the entire watershed arose due to the small number of and lack of diversity in Sacramento basin reference gages. When the model was applied to predict for sub-basin outlets across the entire Sacramento basin as detailed in the next section, the predicted flow rates were often extremely and unrealistically low, so the approach was not pursued further. (However, following the USGS approach and using runoff per unit of drainage area as the dependent variable rather than measured flow rate might improve the model's application, as discussed in the next section.) As shown in Figure 3.4, most reference gages were concentrated in two general areas of the watershed and were often on upper reaches of streams. Restricting the training data to only those reference gages in the Sacramento basin appears to have taken too much variation out of the data and resulted in a heavily biased model. Apparently, a natural flow model does not benefit from Sacramento-specific idiosyncrasies captured in the available data. While the results of the dry-year modeling process showed that in some cases, using a more curated dataset is helpful, a too-limited dataset with insufficient variation can make a worse model. In this case, using the traditional, "more data is better" approach is superior, so basing the models on the larger ecoregions is a better approach.

Case Study: Application of New Best Models for Use in the Sacramento DWRAT

Can the models developed based on reference gages be generalized to the entire Sacramento watershed? To answer this question, the new dry-year monthly regional models were applied to predict natural flow for the outlet of every sub-basin in the Sacramento basin for the 1977 calendar year. The year 1977 was chosen because the Sacramento DWRAT uses ratios of 1977 natural flow estimates to spatially disaggregate current natural flow estimates at six locations to every sub-basin outlet in the watershed, as mentioned in Chapter 2. The Intermountain models were used to predict for any sub-basin in the Intermountain region as well

as any Xeric sub-basin downstream of an Intermountain sub-basin (found using Santos, 2015). The Xeric models were used to predict for the remaining Xeric sub-basins. Since the Intermountain sub-basins tended to have more flow, this prevented a sudden drop in predicted natural flow rates at the Intermountain-Xeric border. This process accounted for a Xeric sub-basin being downstream of Sierra runoff and tried to maintain hydrologic logic in a statistical model. In total, 90 Xeric sub-basins were included in the Intermountain region in this way. They are highlighted in white in Figure 4.10.

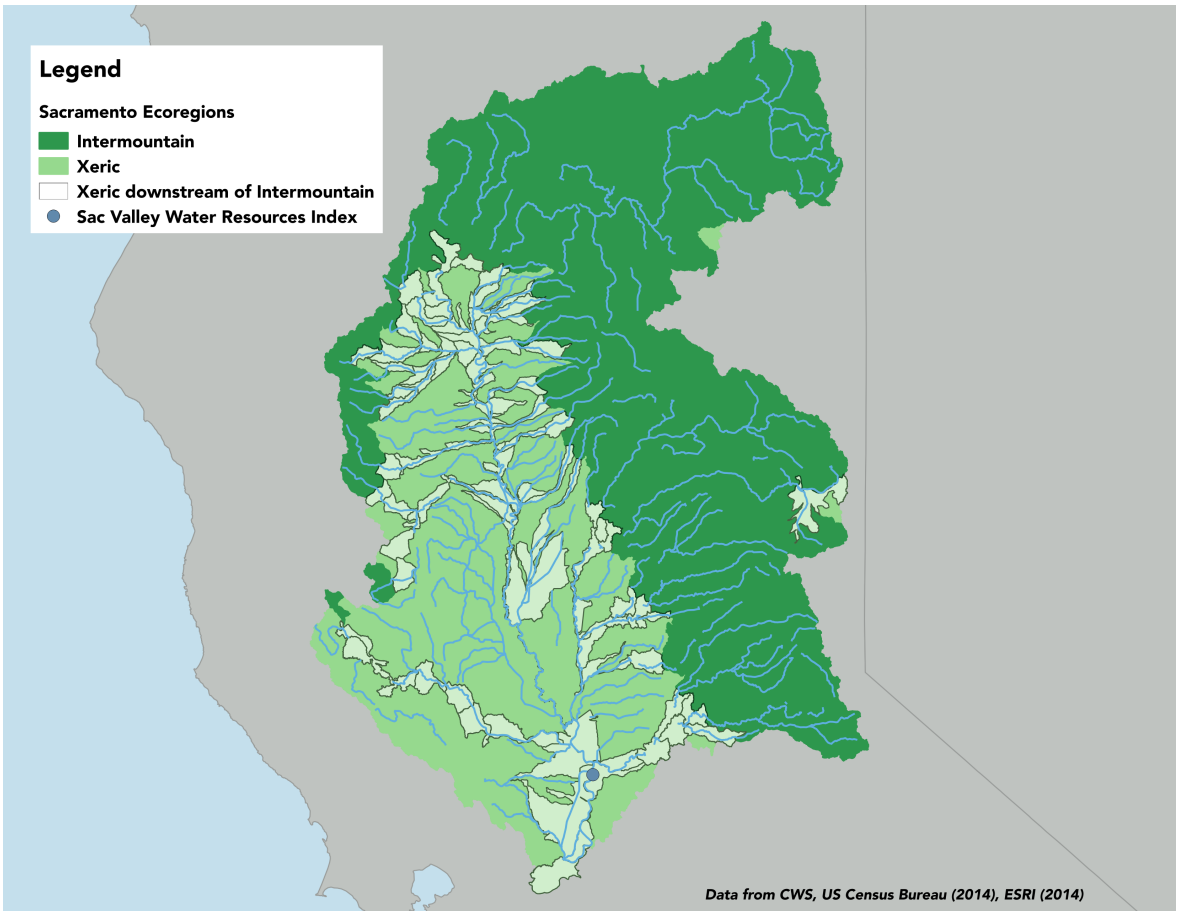


Figure 4.10. Map of Xeric Sub-basins Downstream of Intermountain Sub-basins

As Carlisle *et al*. (2010) noted when developing the original USGS model, in cases where the predictor variables were not within the scope of the reference sites used in training the model, the model should be applied with caution. As they stated, "we believe the models can be applied to all stream segments…that occur within the experience of the reference sites we identified." Since this case study applies the model to locations that have no parallel in the set of reference gages—for example, there are no reference gages on the main stem of a river as large as the Sacramento—this case study has moved into purely experimental territory. The performance metrics estimated based on the test data are at best a guess at how well the model will perform when applied to new situations.

Keeping this warning in mind, the flows generated by the model were compared to those estimated by traditional hydrologic modeling to get a sense of how well the model generalizes

for the entire Sacramento watershed. Historical monthly full natural flow estimates were obtained for the six gage locations used in the sub-basin extrapolation process (mapped in Figure 2.1). Three of the locations lie on the Sacramento River (SACC0, SBB, and SIS), and the other three are on major tributaries. See Table 4.5.

Table 4.5. Natural flow estimation locations for the Sacramento River

Gage	Name	Basin Location	Source
SIS	Sacramento-Inflow Shasta	Upper reaches of the Sacramento River	(CA DWR, 2016b)
SBB	Sacramento River above Bend Bridge	Sacramento River	(CA DWR, 2016b)
FTO	Feather River at Oroville	Eastern tributary	(CA DWR, 2016b)
YRS	Yuba River near Smartville	Eastern tributary	(CA DWR, 2016b)
AMF	American River at Folsom	Eastern tributary	(CA DWR, 2016b)
SACC0	Sacramento Valley Water Resources Index	Near the Sacramento River's outlet	(CA DWR, 2016a)

The first five locations represent hydrologically-modeled estimates at actual gage locations. The last location, SACC0, is an index that aggregates four upstream natural flow estimates: the Sacramento River at Bend Bridge (SBB), the Feather River at Oroville (FTO), the Yuba River at Englebright, and the American River at Folsom (AMF) (CA DWR, 2016). The historical estimates are available in monthly acre-feet. By converting these to average daily natural flow estimates in cfs, they could be directly compared to the statistical models' predictions for those locations.

At first, the results of generalizing the model were not encouraging. Other than a spike in predicted natural flow for May and a smaller spike in July, results from the best models developed in the previous sections of this chapter were much lower than the hydrologic estimates. While the hydrologic estimates have their own errors and should not be treated as a precise performance benchmark, such a large difference in estimates seemed concerning. The USGS models' predictions were also generally lower than the hydrologic estimates for locations on the Sacramento River, although they sometimes overpredicted for tributary locations. Neither set of models seemed to follow the traditional hydrologic model's estimates and general shape. Figure 4.11 shows each model's estimates for each comparison location with Sacramento River locations on the left and tributary locations on the right.

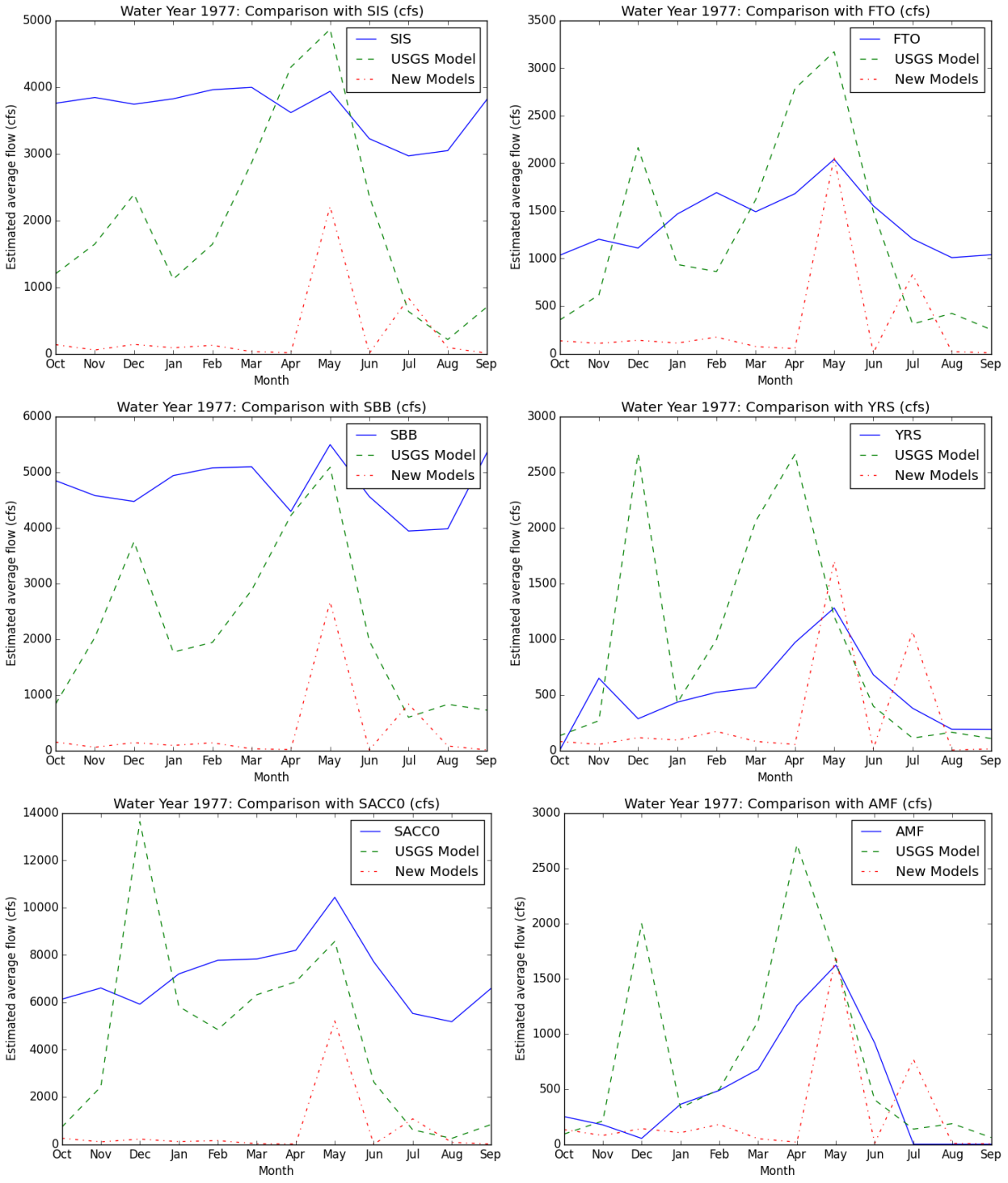


Figure 4.11. Comparison of 1977 Flow Estimates at Six Sacramento River Gages

This means that the "best models" as conceptualized thus far in this thesis performed better than the USGS models for the test data, but they did not generalize well to the main stem of the Sacramento River or its major tributaries. However, neither model seems to match the hydrologic model reliably. This is a concern for the Sacramento DWRAT, although since both

statistical models were developed based on data for smaller streams and upper reaches of larger rivers and cannot be guaranteed to generalize well to larger rivers, these findings are not unexpected.

The USGS had found improved performance by using runoff per unit of drainage area as the dependent variable, rather than simply flow rate (Grantham, 2015). One of the available predictor variables was the total drainage area that feeds to a gage in square kilometers (*drain_sqkm*). Dividing the measured flow (*qmeas*) by *drain_sqkm* resulted in a new variable that represented the amount of runoff generated per square kilometer draining to a location. After removing *drain_sqkm* from the list of available predictor variables and training the model to predict this scaled runoff unit, the predictions were multiplied by *drain_sqkm* to convert them back to an estimate of flow in cfs. Early on in this research, this process had not improved estimated test performance when experimented with for the month of July as part of the *mlutilities* sequence, so the models generated for this thesis had been developed to predict *qmeas*, not *qmeas/drain_sqkm*.

When, as an experiment for this case study, this process was left out of the USGS model, meaning it was trained to predict *qmeas* instead of *qmeas/drain_sqkm*, the resulting 1977 natural flow predictions for the Sacramento watershed were very similar to those generated by the new best models in magnitude. Continuing the experiment, the new best models were retrained to predict *qmeas/drain_sqkm* rather than *qmeas*. The 1977 predictions were much improved but did not lead to a very smooth curve, jumping between high and low flows across months. Rather than using different models for each month, a consistent model for each monthly region scenario was chosen based on general performance characteristics found in this chapter: a dry-year, scaled dataset combined with a stacking ensemble. Using this new process, the predictions for 1977 were dramatically improved. See Figure 4.12 for details.

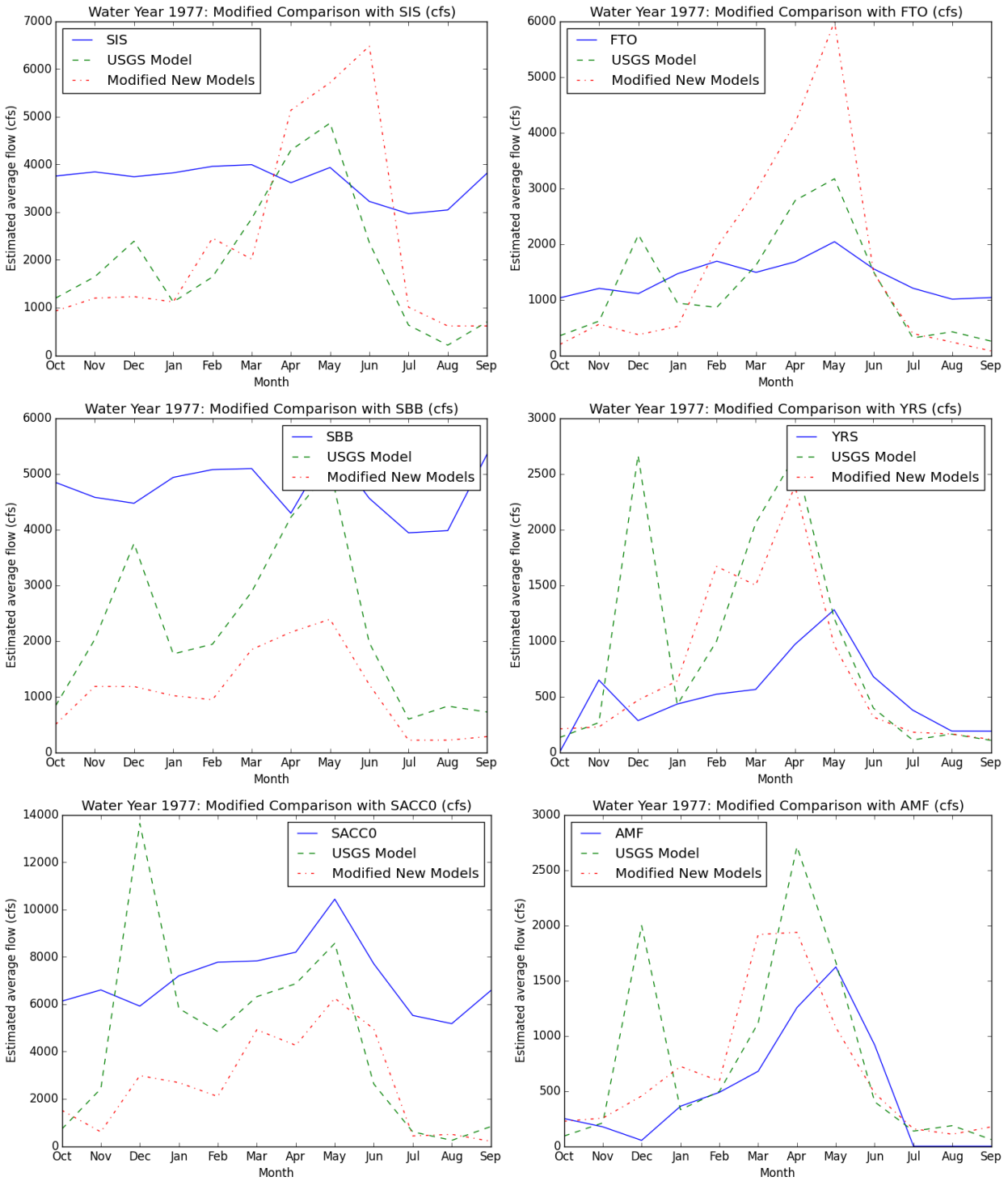


Figure 4.12. Comparison of 1977 Flow Estimates at Six Sacramento River Gages with Modified Dependent Variable & Consistent Model

The new models' 1977 predictions now seem to mimic the temporal pattern of the hydrologic model's estimates at SBB, SACC0, and AMF, although the predictions are still lower for the first of those two locations and higher at AMF. The new models tend to overpredict in the

spring when compared with the hydrologic model's estimates at SIS, FTO, and YRS, but the magnitudes are much improved relative to those shown in Figure 4.11. Using *qmeas/drain_sqkm* as the dependent variable and a consistent model seems to help consciously enforce hydrologic logic and make sure that locations with large drainage areas have larger flows. However, on average, both statistical models still predict much lower natural flows than the hydrologic model does on the main stem of the Sacramento River, as Table 4.6 shows. While their predictions tend to be closer (although too high) for tributary locations, as would be expected for statistical models developed for small streams and upper reaches of large rivers, neither model consistently predicts natural flows close to the hydrologic model's predictions.

Table 4.6. Statistical Models' Difference from Hydrologic Models (cfs)

Location	Mean of Modified Difference	Standard Deviation of Modified Difference	Mean of USGS Difference	Standard Deviation of USGS Difference
SIS	-1267	2072	-1649	1283
SBB	-3622	698	-2499	1364
SACC0	-4469	1152	-2628	3517
FTO	198	1536	-128	761
YRS	226	590	421	875
AMF	191	449	300	668

What do these results mean for DWRAT? This case study only compares flow at one location in the basin, so no conclusive recommendation can be made. Since neither the USGS models nor the new models seem to predict well for the main stem of the river, switching the Sacramento DWRAT to the new flow models does not appear to be worthwhile. Modifying the models to predict *qmeas/drain_sqkm* seems helpful, but it would require further study and a rigorous estimation of the model's performance on unseen test data to understand its accuracy. More importantly, if we can treat the SACC0 estimate as a guideline, both the USGS and the new statistical models seem to drastically underpredict summer flows on the main stem of the Sacramento River. This highlights a common difficulty in machine learning. Models can only be applied with confidence to situations for which they have been trained and on which they have been tested. By definition, random forest predictions can only be combinations of observed dependent variable values from the training data, so it will never predict a flow value higher than a flow it has seen before. The same is true for the k-nearest neighbors and decision tree algorithms. Since decision trees were used as the base estimator for AdaBoost in this research, this is true for AdaBoost as well. In contrast, parametric statistical models like ridge regression can extrapolate, but prediction of natural flow does not seem to be a linear phenomenon, making those extrapolations relatively unhelpful. Therefore, without new training data that better covers natural flow of large rivers, neither model will likely predict well for the main stem of the Sacramento River.

DWRAT's spatial disaggregation process to update the estimated historical flows to the current day corrects for these problems to some extent, although if flow estimates for the main stem of the Sacramento River are too low, the ratios of estimated historical flows between sub-basins are likely also too low. This would mean that natural flows for sub-basins near a hydrologically estimated natural flow gage are in the right neighborhood, but natural flows for

sub-basins at the upper limits of a natural flow gage's spatial disaggregation region (see Figure 2.1) are likely too high. By using these scaling ratios, DWRAT may be over-predicting tributary flows using either set of natural flow models. Smaller spatial disaggregation regions likely experience less of this problem. Because the Sacramento DWRAT uses six estimated natural flow gages, those spatial disaggregation regions are kept from growing too large. Including additional estimated natural flow gages would further mitigate the issue. Until a detailed mechanistic model can be used to estimate natural flow for each sub-basin outlet in the Sacramento basin, DWRAT is limited to the statistical models and their limitations.

CHAPTER 5: CONCLUSIONS

Discussion of results

After exploring many combinations of datasets and machine learning approaches and testing their performance, some general conclusions can be drawn. The utility of stacking ensemble models to correct for biases of their underlying base predictors and to generate more accurate predictions for unseen test data was demonstrated for multiple scenarios. Using a more restricted, curated dataset was helpful in some cases, and is an approach worth considering when trying to create a targeted model such as dry-year natural flow prediction rather than a general-purpose model for predicting flows across all water year types. The new models generally were more accurate than the original USGS models for dry years, particularly in the Intermountain region. However, because both are trained on reference gage flows generally located in upper reaches of streams, neither the USGS models nor the new models seem to predict natural flow of the main stem of the Sacramento River well when this was attempted experimentally and compared to the natural flow estimates of a traditional hydrologic natural flow model (although the hydrologic model has its own set of simplifications and errors). This is not unexpected, however, considering the premise of a statistical model. When the USGS models' outputs are used in DWRAT, the spatial disaggregation process likely corrects some of this problem. Using a dependent variable that enforces some degree of hydrologic logic—such as runoff generated per unit of drainage area (*qmeas/drain_sqkm*)—also appears to help in generating usable natural flow estimates for DWRAT.

Restricting the training data to drier years and following the USGS approach of creating monthly models for relevant aggregated ecoregions resulted in a set of 24 models (12 monthly models for each of the two aggregated ecoregions containing the Sacramento watershed). In all scenarios, stacking ensemble models tended to perform best on withheld test data. Random forests were the next-most frequent top performer, indicating that the original models are still well-performing models in many cases.

The monthly Intermountain scenarios often benefitted from training on a dry-year dataset, demonstrating that using a more targeted dataset to train the model can be helpful. This contrasts with the usual expectation in statistics and machine learning that "more data is better." The main exceptions were the high-runoff spring months, which likely see more variable flows. Perhaps this is because training the model on the complete dataset gave the model more variable examples from which to learn. In contrast, the dry-year monthly Xeric scenarios did not benefit from using a more restricted, curated dataset in most months. The wet-year/dry-year disparity in the Intermountain region is probably much greater than in the Xeric region. Given this, including wet years in the dataset when training a model to predict for dry years in the Xeric region probably does not bias model predictions as strongly toward wet years as it seems to in the Intermountain region. When examined to test the effect of restricting the data in the other direction, the wet-year scenarios also followed the traditional "more data is better" axiom. Most of their top-ranked models for both regions were also trained on the full all-year dataset. In predicting for wet years, then, the greater variation and power of additional data must have been more important than the type of that data. Restricting the data geographically was also unhelpful. This approach resulted in models that performed well on geographically-restricted test data but that could only predict very low flows when applied because of the lack of variation in the training data.

When evaluated on their prediction of known dry water year flows, the monthly regional models consistently tested as better than or equivalent to their corresponding general-purpose USGS models on multiple test metrics, and in some cases, they performed far better. This is a significant improvement toward predicting natural flows in dry years and could make DWRAT a more useful tool for suggesting curtailment decisions. As was explored in the experimental case study at the end of Chapter 4, however, applications of a statistical natural flow model beyond the situation it was trained for can lead to untrustworthy results and should be treated as suspect. Measures to correct for likely errors are important, such as the spatial disaggregation process that spreads current, validated, multiple hydrologic estimates of natural flow around a basin based on ratios of the statistical model's historical predictions.

Limitations

The caveats from Chapter 1 bear repeating and are true for both the USGS model and the new models developed in this research. An expected value for natural flow is particularly difficult to assess because many locations lack flows recorded prior to human development and disruption, including the main stem of the Sacramento River. The USGS reference gages used to train the model can be left out of training data and used to test a model's predictions, but they only represent a particular type of location. Training the model to predict for those locations does not mean that it will predict well for all locations, as Carlisle *et al*. (2010) also indicated. Predictions can be evaluated against the result of more traditional mechanistic hydrologic models of natural flow, as was done for the main stem of the Sacramento River, but those models have their own uncertainties and errors and should not be treated as ground truth.

Recommendations for further research

This thesis has built on the work of Carlisle *et al*. (2010) and Grantham (2014) and lays a path for further inquiries. One clear next step would be to rerun much of the analysis with the dependent variable discovered to perform more logically in Chapter 4's case study: runoff generated per unit of drainage area (*qmeas/drain_sqkm*). At this point, a clear understanding of how it would affect test performance during the model development process is unknown but would be helpful for those who wish to apply the model. The idea of restricting data could be explored further. Rather than a simple wet-dry categorization, a three-category continuum of dry, average, and wet years might be more helpful. This research examined multiple performance metrics but chose the "best" model for each scenario based only on its estimated R^2 value, but creating a process that choses the best model based on multiple metrics might select better models, since they would have been shown to perform across multiple dimensions. Also, this research created a wealth of models for every scenario, and although some dataset-algorithm combinations proved to be truly terrible, there were many good models that performed well, even though one was chosen as best. Some of the well-performing models could perhaps be used to generate a probability distribution or an envelope of possible numbers rather than a single flow value for a given point.

Using the same basic framework to create monthly regional models but finding new sources to supplement the training data with information on natural flow of larger rivers would help make the models more accurate for larger rivers like the Sacramento. Perhaps estimates of natural flow from a mechanistic hydrologic model like the Sacramento Valley Water Resources

Index could be included as training observations—although they would not represent a ground truth, they would be an expertly developed estimate that could enlarge the applicable scope of the model. Alternatively, the model could be made to predict unimpaired flow rather than natural flow if trained on calculations of unimpaired flow rather than only on reference gages.

On the subject of applying the model, an evaluation of the spatial disaggregation model that uses ratios of the historical natural flow estimates as a basis for disaggregating current natural flow estimates from mechanistic hydrologic models from a single point to the rest of the watershed could be interesting. It rests on the premise that ratios of downstream and upstream flows are fairly consistent across years and that the ratios for any drought year are probably similar to those of 1977, but this may not be the case. This analysis was beyond the scope of this thesis and would have been difficult to do with the existing dataset, but it would be an important part of evaluating the trustworthiness of the natural flow estimates fed into DWRAT.

Conclusions

This thesis demonstrates the power of stacking ensemble models to correct for the biases of their underlying base predictors and to generate more accurate predictions for unseen test data in multiple scenarios. Although statistics and machine learning usually expect more data to be better, the monthly Intermountain scenarios often benefitted from training on a dry-year dataset, demonstrating that using a more targeted dataset to train the model can be helpful at times. The new models generally tested as more accurate than the original USGS models for dry years, particularly in the Intermountain region, although because of the scope of their training data and the nature of statistical modeling, both models can only be applied to streams' upper reaches with confidence.

A river's natural flow is an important quantity that we often can only do our best to estimate. Even with their limitations, further research into statistical models of natural flows to supplement more traditional hydrologic models can lend additional insights and foster more rapid development of usable models. Better information on natural flows can improve decision-making on both environmental problems and water rights curtailment, issues which will likely only grow in importance in California over time.

ACKNOWLEDGEMENTS

This thesis owes its completion to the advice and assistance of many people. My heartfelt thanks to the three members of my committee for all their support and guidance: Jay Lund for involving me in this project and his vision for improving administration of water rights in California, Robert Hijmans for teaching me the glories of statistical learning and all his time spent working out the details with me, and Ted Grantham for patiently explaining the logic of the existing models and answering one thousand questions about the many pieces of data. Thanks to Ben Lord for paving the way for the rest of us with the Eel River model, to Andy Tweet and Chad Whittington for our many companionable working hours and mutual assistance these past two years, and to Wes Walker and Brad Arnold for being willing to pick up the torch. And Christopher Skeels, thank you for all the brainstorming sessions, expert advice, and endless encouragement.

BIBLIOGRAPHY

Arthington, A. H., Bunn, S. E., Poff, N. L., & Naiman, R. J. (2006). The Challenge of Providing Environmental Flow Rules to Sustain River Ecosystems. *Ecological Applications, 16*(4), 1311-1318.

Breiman, L. (1996). Stacked Regressions. *Machine Learning, 24(*1), 49-64.

Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5-32.

Buer, K., Forwalter, D., Kissel, M., & Stohler, B. (1989). The Middle Sacramento River: Human Impacts on Physical and Ecological Processes along a Meandering River: USDA Forest Service.

California Department of Water Resources. (1980). California Central Valley Natural Flow Data. (1st ed.).

California Department of Water Resources. (2016a). *Sacramento Valley Water Resources Index - Monthly Full Natural Flow*. [Spreadsheet]. Provided via email by Brett Whitin, National Oceanic and Atmospheric Association, 8 March 2016.

California Department of Water Resources. (2016b). *Monthly Data by Water Year* [Spreadsheet]. Flow, Full Natural (AF). Retrieved from http://cdec.water.ca.gov/cgi-progs/queryWY

California Department of Water Resources, Bay-Delta Office. (2007). California Central Valley Unimpaired Flow Data. (4th ed., pp. 52).

California Department of Water Resources, Division of Planning. (1987). California Central Valley Unimpaired Flow Data. (2nd ed., pp. 38).

California Department of Water Resources, Division of Planning. (1994). California Central Valley Unimpaired Flow Data. (3rd ed.).

Carlisle, D. M., Falcone, J., & Meador, M. R. (2009). Predicting the biological condition of streams: use of geospatial indicators of natural and anthropogenic characteristics of watersheds. *Environmental Monitoring and Assessment, 151*(1-4), 143-160.

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

Clarke, B. (2003). Comparing Bayes Model Averaging and Stacking When Model Approximation Error Cannot be Ignored. *Journal of Machine Learning Research*, *4*, 683-712.

Di Luzio, M., Johnson, G. L., Daly, C., Eischeid, J. K., & Arnold, J. G. (2008). Constructing Retrospective Gridded Daily Precipitation and Temperature Datasets for the Conterminous United States. *Journal of Applied Meteorology and Climatology, 47*(2), 475-497.

Eng, K., Carlisle, D. M., Wolock, D. M., & Falcone, J. A. (2012). Predicting the Likelihood of Altered Streamflows at Ungauged Rivers across the Conterminous United States. *River Research and Applications, 29*(6), 781-791.

ESRI. (2014). *USA Rivers and Streams* [Digital spatial dataset]. Retrieved from: http://beta.esri.opendata.arcgis.com/datasets/0baca6c9ffd6499fb8e5fad50174c4e0_0

Falcone, J. A. (2011a). *GAGES-II (Geospatial Attributes of Gages for Evaluating Streamflow) Summary Report* (pp. 25). US Geological Survey.

Falcone, J. A. (2011b). *GAGES-II: Geospatial Attributes of Gages for Evaluating Streamflow* [Digital spatial dataset]. US Geological Survey. Retrieved from http://water.usgs.gov/GIS/metadata/usgswrd/XML/gagesII_Sept2011.xml

Falcone, J. A., Carlisle, D. M., Wolock, D. M., & Meador, M. R. (2010). GAGES: A stream gage database for evaluating natural and altered flow conditions in the conterminous United States. *Ecology, 91*(2), 621-621.

Flint, L. E., Flint, A. L., Thorne, J. H., & Boynton, R. (2013). Fine-scale Hydrologic Modeling for Regional Landscape Applications: The California Basin Characterization Model Development and Performance. *Ecological Processes, 2*(25), 1-21.

Fox, P., Hutton, P. H., Howes, D. J., Draper, A. J., & Sears, L. (2015). Reconstructing the natural hydrology of the San Francisco Bay–Delta watershed. *Hydrology and Earth System Sciences, 19*(10), 4257-4274.

Gleick, P. H. (1987). The Development and Testing of a Water Balance Model for Climate Impact Assessment: Modeling hte Sacramento Basin. *Water Resources Research, 23*(6), 1049-1061.

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.

Grantham, T. E. (2015, August 5). [Flow Model Discussion].

Grantham, T. E. (2016, January 22). [Flow Model Data Discussion].

Grantham, T. E., & Viers, J. H. (2014). 100 years of California's water rights system: patterns, trends and uncertainty. *Environmental Research Letters, 9*(8), 084012.

Grantham, T. E., Viers, J. H., & Moyle, P. B. (2014). Systematic Screening of Dams for Environmental Flow Assessment and Implementation. *BioScience, 64*(11), 1006-1018.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). New York: Springer.

Hay, L. E., Markstrom, S. L., & Ward-Garrison, C. (2011). Watershed-Scale Response to Climate Change through the Twenty-First Century for Selected Basins across the United States. *Earth Interactions, 15*(17), 1-37.

Huang, G., Kadir, T., & Chung, F. (2014). *Estimating Natural Flows into California's Sacramento-San Joaquin Delta*. Paper presented at the American Geophysical Union Fall Meeting, San Francisco, California. Retrieved from https://www.researchgate.net/publication/270589558_Estimating_Natural_Flows_into_the_California's_Sacramento_-_San_Joaquin_Delta

Hunt, C. D. (1979). *National Atlas of the United States of America - Surficial Geology* [Map]. US Geological Survey.

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering, 9*(3), 90-95.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning, with Applications in R*. New York: Springer.

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

Koczot, K. M., Jeton, A. E., McGurk, B. J., & Dettinger, M. D. (2005). Precipitation-Runoff Processes in the Feather River Basin, Northeastern California, with Prospects for Streamflow Predictability, Water Years 1971-97 (pp. 92). Reston, Virginia: US Geological Survey. Scientific Investigations Report 2004-5202.

Lord, B. (2015). Water rights curtailments for drought in California: Method and Eel River Application. (Master of Science Thesis), University of California, Davis.

Lund, J., Lord, B., Fleenor, W., & Willis, A. (2014). *Drought Curtailment of Water Rights: Problems and Technical Solutions* (pp. 23). Center for Watershed Sciences: University of California, Davis.

MacDonald, G. M., Kremenetski, K. V., & Hidalgo, H. G. (2008). Southern California and the perfect drought: Simultaneous prolonged drought in southern California and the Sacramento and Colorado River systems. *Quaternary International, 188*(1), 11-23.

Meko, D. M., Therrell, M. D., Baisan, C. H., & Hughes, M. K. (2001). Sacramento River Flow Reconstructed to A.D. 869 from Tree Rings. *Journal of the American Water Resources Association, 37*(4), 1029-1039.

Mitchell, T. M. (1997). *Machine Learning*. Boston: McGraw-Hill.

National Oceanic and Atmospheric Administration. (1972). National Weather Service River Forecast System, Forecast Procedures. Silver Spring, Maryland: US Department of Commerce. NOAA Technical Memo NWS HYDRO-14.

National Oceanic and Atmospheric Administration. (2015). *Climate At a Glance: Time Series* [Table]. Retrieved from: https://www.ncdc.noaa.gov/cag/time-series/us/4/2/pcp/12/9/1950-2011?base_prd=true&firstbaseyear=1901&lastbaseyear=2000

National Oceanic and Atmospheric Administration. (2016). *Sac and SJ 4 River FNF* [Excel workbook]. Provided via email by Brett Whitin, NOAA.

Nilsson, C., Reidy, C. A., Dynesius, M., & Revenga, C. (2005). Fragmentation and Flow Regulation of the World's Large River Systems. *Science, 308*(5720), 405-408.

Olson, J. R., & Hawkins, C. P. (2012). Predicting natural base-flow stream water chemistry in the western United States. *Water Resources Research, 48*(2), 19.

Omernik, J. M. (1987). Ecoregions of the Conterminous United States. *Annals of the Association of American Geographers, 77*(1), 118-125.

Oshiro, T. M., Perez, P. S., & Baranauskas, J. B. (2012). *How Many Trees in a Random Forest?* Paper presented at the International Conference on Machine Learning and Data Mining, Berlin, Germany.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2825-2830.

Poff, N. L., Allan, J. D., Bain, M. B., Karr, J. R., Prestegaard, K. L., Richter, B. D., Sparks, R. E., & Stromberg, J. C. (1997). The Natural Flow Regime: A paradigm for river conservation and restoration. *BioScience, 47*(11), 769-784.

Pringle, C. M., Freeman, M. C., & Freeman, B. J. (2000). Regional Effects of Hydrologic Alterations on Riverine Macrobiota in the New World: Tropical-Temperate Comparisons. *BioScience, 50*(9), 807-823.

PRISM Climate Group. (2011). *PRISM Climate Data* [Digital spatial data].

QGIS Development Team. (2015). Quantum GIS.

Reed, J. C., & Bush, C. A. (2005). *Generalized geologic map of the United States, Puerto Rico, and the U.S. Virgin Islands, Version 2.0* [Digital maps]. US Geological Survey.

Sacramento River Watershed Program. (2010). The Sacramento River Basin: A Roadmap to Watershed Management (pp. 18). Retrieved from: http://www.sacriver.org/aboutwatershed/roadmap/sacramento-river-basin

Santos, N. R. (2015). select_upstream_hucs. Retrieved from https://bitbucket.org/nickrsan/sierra-code-library/

Shao, J. (1993). Linear Model Selection by Cross-Validation. *Journal of the American Statistical Association, 88*(422), 486-494.

Sill, J., Takacs, G., Mackey, L., & Lin, D. (2009). Feature-Weighted Linear Stacking. *arXiv preprint arXiv:0911.0460*, pp. 17.

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

United States Department of Agriculture. (2008). *U.S. General Soil Map (STATSGO) NRCS NCGC* [Digital spatial dataset].

United States Environmental Protection Agency. (2008). *National Hydrography Dataset Plus (NHDPlus)* [Digital spatial dataset].

United States Geological Survey. (2008). *USGS Water Data for the Nation (National Water Information System)* [Digital spatial dataset].

Wolock, D. M. (2003). *Base-flow index grid for the conterminous United States* [Digital dataset]. US Geological Survey.

Wolock, D. M., Hornberger, G. M., Beven, K. J., & Campbell, W. G. (1989). The relationship of catchment topographic and soil hydraulic characteristics to lake alkalinity in the northeastern United States. *Water Resources Research, 25*, 829-838.

Wolock, D. M., & McCabe, G. J. (1999). Estimates of runoff using water-balance and atmospheric general circulation models. *Journal of the American Water Resources Association, 35*(6), 1341-1350.

Wolock, D. M., Winter, T. C., & McMahon, G. (2004). Delineation and Evaluation of

Hydrologic-Landscape Regions in the United States Using Geographic Information System Tools and Multivariate Statistical Analyses. *Environmental Management, 34*(S1), S71-S88.

Wolpert, D. H. (1992). Stacked Generalization. *Neural Networks, 5*(2), 241-259.

APPENDIX A: DATA USED IN MODELING

All variables were calculated using publicly available datasets. For further details, see Falcone (2011a). These variables were mapped to their sources in consultation with Grantham (2016); any errors are my own. Most variables describe the total, average, or percentage value for the area draining to that point (for example, "DRAIN_SQKM" is the total drainage area that feeds to a gage). Some variables are simple ID keys and were not used in prediction. "Year" and "STAID" were never used as predictor variables. "Month" was ignored as superfluous in the monthly regional models but was transformed into 11 binary variables indicating month (January-November) to be used as a predictor variable in the Sacramento basin model. "IntMnt" was ignored as superfluous in the monthly regional models but was used as a binary variable indicating if a location was within the Intermountain aggregated ecoregion for the Sacramento basin model.

Variable	Definition	Source
anorthositic	Rock type	Reed & Bush, 2001
APR_PPT7100_CM	Long term average precipitation for that month (1971-2000)	Falcone, 2011b
APR_TMP7100_DEGC	Monthly average temperature by water year	Falcone, 2011b
AUG_PPT7100_CM	Long term average precipitation for that month (1971-2000)	Falcone, 2011b
AUG_TMP7100_DEGC	Monthly average temperature by water year	Falcone, 2011b
AWCAVE	Soil water capacity	USDA, 2008
BDAVE	Soil average bulk density	USDA, 2008
BFI_AVE	Average base flow index	Wolock, 2003
CaO_pct	Mean percent of rock's CaO content	Olson & Hawkins, 2012
CLAYAVE	Soil average clay content	USDA, 2008
CONTACT	Subsurface flow contact time	Wolock *et al*., 1989
DEC_PPT7100_CM	Long term average precipitation for that month (1971-2000)	Falcone, 2011b
DEC_TMP7100_DEGC	Monthly average temperature by water year	Falcone, 2011b
DRAIN_SQKM	Drainage area	USGS, 2008
ELEV_MEAN_M_BASIN_30M	Mean watershed elevation	USEPA, 2008
ET	Evapotranspiration	Falcone, 2011b
FEB_PPT7100_CM	Long term average precipitation for that month (1971-2000)	Falcone, 2011b
FEB_TMP7100_DEGC	Monthly average temperature by water year	Falcone, 2011b
gneiss	Rock type	Reed & Bush, 2001
granitic	Rock type	Reed & Bush, 2001

HGA	Soil hydrologic group A	USDA, 2008
HGAC	Soil hydrologic groups A & C	USDA, 2008
HGAD	Soil hydrologic groups A & D	USDA, 2008
HGB	Soil hydrologic group B	USDA, 2008
HGBC	Soil hydrologic groups B & C	USDA, 2008
HGBD	Soil hydrologic groups B & D	USDA, 2008
HGC	Soil hydrologic group C	USDA, 2008
HGCD	Soil hydrologic groups C & D	USDA, 2008
HGD	Soil hydrologic group D	USDA, 2008
HGVAR	Soil hydrologic group VAR	USDA, 2008
HLR1	Hydrologic landscape region 1	Wolock *et al*., 2004
HLR10	Hydrologic landscape region 10	Wolock *et al*., 2004
HLR11	Hydrologic landscape region 11	Wolock *et al*., 2004
HLR12	Hydrologic landscape region 12	Wolock *et al*., 2004
HLR13	Hydrologic landscape region 13	Wolock *et al*., 2004
HLR14	Hydrologic landscape region 14	Wolock *et al*., 2004
HLR15	Hydrologic landscape region 15	Wolock *et al*., 2004
HLR16	Hydrologic landscape region 16	Wolock *et al*., 2004
HLR17	Hydrologic landscape region 17	Wolock *et al*., 2004
HLR18	Hydrologic landscape region 18	Wolock *et al*., 2004
HLR19	Hydrologic landscape region 19	Wolock *et al*., 2004
HLR2	Hydrologic landscape region 2	Wolock *et al*., 2004
HLR20	Hydrologic landscape region 20	Wolock *et al*., 2004
HLR3	Hydrologic landscape region 3	Wolock *et al*., 2004
HLR4	Hydrologic landscape region 4	Wolock *et al*., 2004
HLR5	Hydrologic landscape region 5	Wolock *et al*., 2004
HLR6	Hydrologic landscape region 6	Wolock *et al*., 2004
HLR7	Hydrologic landscape region 7	Wolock *et al*., 2004
HLR8	Hydrologic landscape region 8	Wolock *et al*., 2004
HLR9	Hydrologic landscape region 9	Wolock *et al*., 2004
intermediate	Rock type	Reed & Bush, 2001
IntMnt	Presence in Intermountain aggregated ecoregion	Falcone, 2011a
JAN_PPT7100_CM	Long term average precipitation for that month (1971-2000)	Falcone, 2011b
JAN_TMP7100_DEGC	Monthly average temperature by water year	Falcone, 2011b
JUL_PPT7100_CM	Long term average precipitation for that month (1971-2000)	Falcone, 2011b
JUL_TMP7100_DEGC	Monthly average temperature by water year	Falcone, 2011b
JUN_PPT7100_CM	Long term average precipitation for that month (1971-2000)	Falcone, 2011b
JUN_TMP7100_DEGC	Monthly average temperature by	Falcone, 2011b

	water year	
KFACT_UP	Soil average K-factor	USDA, 2008
LPerm	Mean hydraulic conductivity	Olson & Hawkins, 2012
MAR_PPT7100_CM	Long term average precipitation for that month (1971-2000)	Falcone, 2011b
MAR_TMP7100_DEGC	Monthly average temperature by water year	Falcone, 2011b
MAY_PPT7100_CM	Long term average precipitation for that month (1971-2000)	Falcone, 2011b
MAY_TMP7100_DEGC	Monthly average temperature by water year	Falcone, 2011b
MgO_pct	Mean percent of rock's MgO content	Olson & Hawkins, 2012
Month	Month	Falcone, 2011b
NO10AVE	Soil material < 0.07 mm	USDA, 2008
NO200AVE	Soil material < 2 mm	USDA, 2008
NO4AVE	Soil material < 5 mm	USDA, 2008
NOV_PPT7100_CM	Long term average precipitation for that month (1971-2000)	Falcone, 2011b
NOV_TMP7100_DEGC	Monthly average temperature by water year	Falcone, 2011b
OCT_PPT7100_CM	Long term average precipitation for that month (1971-2000)	Falcone, 2011b
OCT_TMP7100_DEGC	Monthly average temperature by water year	Falcone, 2011b
OMAVE	Soil average organic matter	USDA, 2008
p0	Monthly average precipitation, current month	PRISM Climate Group, 2011
p1	Monthly average precipitation, previous month	PRISM Climate Group, 2011
p10	Monthly average precipitation, 10 months ago	PRISM Climate Group, 2011
p11	Monthly average precipitation, 11 months ago	PRISM Climate Group, 2011
p12	Monthly average precipitation, 12 months ago	PRISM Climate Group, 2011
p2	Monthly average precipitation, 2 months ago	PRISM Climate Group, 2011
p2sum	Sum of monthly average precipitation from previous 2 months	PRISM Climate Group, 2011
p3	Monthly average precipitation, 3 months ago	PRISM Climate Group, 2011
p3sum	Sum of monthly average precipitation from previous 3 months	PRISM Climate Group, 2011

p4	Monthly average precipitation, 4 months ago	PRISM Climate Group, 2011
p5	Monthly average precipitation, 5 months ago	PRISM Climate Group, 2011
p6	Monthly average precipitation, 6 months ago	PRISM Climate Group, 2011
p6sum	Sum of monthly average precipitation from previous 6 months	PRISM Climate Group, 2011
p7	Monthly average precipitation, 7 months ago	PRISM Climate Group, 2011
p8	Monthly average precipitation, 8 months ago	PRISM Climate Group, 2011
p9	Monthly average precipitation, 9 months ago	PRISM Climate Group, 2011
PERDUN	Dunne overland flow	Wolock, 2003
PERHOR	Horton overland flow	Wolock, 2003
PERMAVE	Soil permeability	USDA, 2008
PPTAVG_BASIN	Annual average precipitation	Falcone, 2011b
PRECIP_SEAS_IND	Monthly preciptiation variability	Falcone, 2011b
qmeas	Average monthly flow rate	Falcone, 2011b
quarternary	Rock type	Reed & Bush, 2001
RFACT	Rainfall/runoff factor	USDA, 2008
ROCKDEPAVE	Average soil thickness	USDA, 2008
RUNAVE7100	Average annual runoff (1970-2000)	Falcone, 2011b
S_pct	Mean percent of rock's S content	Olson & Hawkins, 2012
SANDAVE	Soil average sand content	USDA, 2008
sedimentary	Rock type	Reed & Bush, 2001
SEP_PPT7100_CM	Long term average precipitation for that month (1971-2000)	Falcone, 2011b
SEP_TMP7100_DEGC	Monthly average temperature by water year	Falcone, 2011b
SGEO1	Surficial geology class 1	Hunt, 1979
SGEO10	Surficial geology class 10	Hunt, 1979
SGEO11	Surficial geology class 11	Hunt, 1979
SGEO12	Surficial geology class 12	Hunt, 1979
SGEO13	Surficial geology class 13	Hunt, 1979
SGEO14	Surficial geology class 14	Hunt, 1979
SGEO15	Surficial geology class 15	Hunt, 1979
SGEO16	Surficial geology class 16	Hunt, 1979
SGEO17	Surficial geology class 17	Hunt, 1979
SGEO18	Surficial geology class 18	Hunt, 1979
SGEO19	Surficial geology class 19	Hunt, 1979
SGEO2	Surficial geology class 2	Hunt, 1979

SGEO20	Surficial geology class 20	Hunt, 1979
SGEO21	Surficial geology class 21	Hunt, 1979
SGEO22	Surficial geology class 22	Hunt, 1979
SGEO23	Surficial geology class 23	Hunt, 1979
SGEO24	Surficial geology class 24	Hunt, 1979
SGEO25	Surficial geology class 25	Hunt, 1979
SGEO26	Surficial geology class 26	Hunt, 1979
SGEO27	Surficial geology class 27	Hunt, 1979
SGEO28	Surficial geology class 28	Hunt, 1979
SGEO29	Surficial geology class 29	Hunt, 1979
SGEO3	Surficial geology class 3	Hunt, 1979
SGEO30	Surficial geology class 30	Hunt, 1979
SGEO31	Surficial geology class 31	Hunt, 1979
SGEO32	Surficial geology class 32	Hunt, 1979
SGEO33	Surficial geology class 33	Hunt, 1979
SGEO34	Surficial geology class 34	Hunt, 1979
SGEO35	Surficial geology class 35	Hunt, 1979
SGEO36	Surficial geology class 36	Hunt, 1979
SGEO37	Surficial geology class 37	Hunt, 1979
SGEO38	Surficial geology class 38	Hunt, 1979
SGEO39	Surficial geology class 39	Hunt, 1979
SGEO4	Surficial geology class 4	Hunt, 1979
SGEO40	Surficial geology class 40	Hunt, 1979
SGEO41	Surficial geology class 41	Hunt, 1979
SGEO42	Surficial geology class 42	Hunt, 1979
SGEO43	Surficial geology class 43	Hunt, 1979
SGEO44	Surficial geology class 44	Hunt, 1979
SGEO45	Surficial geology class 45	Hunt, 1979
SGEO5	Surficial geology class 5	Hunt, 1979
SGEO6	Surficial geology class 6	Hunt, 1979
SGEO7	Surficial geology class 7	Hunt, 1979
SGEO8	Surficial geology class 8	Hunt, 1979
SGEO9	Surficial geology class 9	Hunt, 1979
SILTAVE	Soil average silt content	USDA, 2008
SLOPE_PCT_30M	Mean watershed slope	USEPA, 2008
STAID	Gage ID	Falcone, 2011b
T_AVG_BASIN	Long-term average max annual temperature (1970-2000)	Falcone, 2011b
T_MAX_BASIN	Long-term average annual temperature (1970-2000)	Falcone, 2011b
T_MIN_BASIN	Long-term average min annual temperature (1970-2000)	Falcone, 2011b
t0	Monthly average air temperature,	PRISM Climate Group,

	current month	2011
t1	Monthly average air temperature, previous month	PRISM Climate Group, 2011
t10	Monthly average air temperature, 10 months ago	PRISM Climate Group, 2011
t11	Monthly average air temperature, 11 months ago	PRISM Climate Group, 2011
t12	Monthly average air temperature, 12 months ago	PRISM Climate Group, 2011
t2	Monthly average air temperature, 2 months ago	PRISM Climate Group, 2011
t3	Monthly average air temperature, 3 months ago	PRISM Climate Group, 2011
t4	Monthly average air temperature, 4 months ago	PRISM Climate Group, 2011
t5	Monthly average air temperature, 5 months ago	PRISM Climate Group, 2011
t6	Monthly average air temperature, 6 months ago	PRISM Climate Group, 2011
t7	Monthly average air temperature, 7 months ago	PRISM Climate Group, 2011
t8	Monthly average air temperature, 8 months ago	PRISM Climate Group, 2011
t9	Monthly average air temperature, 9 months ago	PRISM Climate Group, 2011
UCS	Mean uniaxial compressive strength	Olson & Hawkins, 2012
ultramafic	Rock type	Reed & Bush, 2001
volcanic	Rock type	Reed & Bush, 2001
wb0	Monthly runoff estimate from water balance model, current month	Wolock & McCabe, 1999
wb1	Monthly runoff estimate from water balance model, previous month	Wolock & McCabe, 1999
wb10	Monthly runoff estimate from water balance model, 10 months ago	Wolock & McCabe, 1999
wb11	Monthly runoff estimate from water balance model, 11 months ago	Wolock & McCabe, 1999
wb12	Monthly runoff estimate from water balance model, 12 months ago	Wolock & McCabe, 1999
wb2	Monthly runoff estimate from water balance model, 2 months ago	Wolock & McCabe, 1999
wb3	Monthly runoff estimate from water balance model, 3 months ago	Wolock & McCabe, 1999
wb4	Monthly runoff estimate from water balance model, 4 months ago	Wolock & McCabe, 1999
wb5	Monthly runoff estimate from water balance model, 5 months ago	Wolock & McCabe, 1999

WB5100_ANN_MM	Average monthly runoff (1950-2010)	Falcone, 2011b
WB5100_APR_MM	Average monthly runoff (1950-2010)	Falcone, 2011b
WB5100_AUG_MM	Average monthly runoff (1950-2010)	Falcone, 2011b
WB5100_DEC_MM	Average monthly runoff (1950-2010)	Falcone, 2011b
WB5100_FEB_MM	Average monthly runoff (1950-2010)	Falcone, 2011b
WB5100_JAN_MM	Average monthly runoff (1950-2010)	Falcone, 2011b
WB5100_JUL_MM	Average monthly runoff (1950-2010)	Falcone, 2011b
WB5100_JUN_MM	Average monthly runoff (1950-2010)	Falcone, 2011b
WB5100_MAR_MM	Average monthly runoff (1950-2010)	Falcone, 2011b
WB5100_MAY_MM	Average monthly runoff (1950-2010)	Falcone, 2011b
WB5100_NOV_MM	Average monthly runoff (1950-2010)	Falcone, 2011b
WB5100_OCT_MM	Average monthly runoff (1950-2010)	Falcone, 2011b
WB5100_SEP_MM	Average monthly runoff (1950-2010)	Falcone, 2011b
wb6	Monthly runoff estimate from water balance model, 6 months ago	Wolock & McCabe, 1999
wb7	Monthly runoff estimate from water balance model, 7 months ago	Wolock & McCabe, 1999
wb8	Monthly runoff estimate from water balance model, 8 months ago	Wolock & McCabe, 1999
wb9	Monthly runoff estimate from water balance model, 9 months ago	Wolock & McCabe, 1999
WD_APR_BASIN	Number of wet days	Falcone, 2011b
WD_AUG_BASIN	Number of wet days	Falcone, 2011b
WD_BASIN	Average annual wet days	Falcone, 2011b
WD_DEC_BASIN	Number of wet days	Falcone, 2011b
WD_FEB_BASIN	Number of wet days	Falcone, 2011b
WD_JAN_BASIN	Number of wet days	Falcone, 2011b
WD_JUL_BASIN	Number of wet days	Falcone, 2011b
WD_JUN_BASIN	Number of wet days	Falcone, 2011b
WD_MAR_BASIN	Number of wet days	Falcone, 2011b
WD_MAY_BASIN	Number of wet days	Falcone, 2011b
WD_NOV_BASIN	Number of wet days	Falcone, 2011b

WD_OCT_BASIN	Number of wet days	Falcone, 2011b
WD_SEP_BASIN	Number of wet days	Falcone, 2011b
WDMAX_BASIN	Max annual wet days	Falcone, 2011b
WDMIN_BASIN	Min annual wet days	Falcone, 2011b
WTDEPAVE	Average depth to water table	USDA, 2008
Year	Year	Falcone, 2011b

APPENDIX B: LIST OF PREDICTOR VARIABLES RETAINED BASED ON DEFINITIONS

As part of the feature engineering step explained in Chapter 3, a list of specific predictor variables (from the comprehensive list in Appendix A) was used to perform an informed variable selection for each model, in addition to using PCA with 20 components, PCA with 50 components, and a variance threshold of 0.8. These tables show the contents of each list of chosen predictor variables for each proposed model.

For both the dry-year/wet-year models, the same set of 34 predictor variables was selected:

Variable	Definition
DRAIN_SQKM	Drainage area
ELEV_MEAN_M_BASIN_30M	Mean watershed elevation
p0	Monthly average precipitation, current month
p1	Monthly average precipitation, previous month
p2	Monthly average precipitation, 2 months ago
p3	Monthly average precipitation, 3 months ago
p4	Monthly average precipitation, 4 months ago
p5	Monthly average precipitation, 5 months ago
p6	Monthly average precipitation, 6 months ago
p7	Monthly average precipitation, 7 months ago
p8	Monthly average precipitation, 8 months ago
p9	Monthly average precipitation, 9 months ago
p10	Monthly average precipitation, 10 months ago
p11	Monthly average precipitation, 11 months ago
p12	Monthly average precipitation, 12 months ago
p2sum	Sum of monthly average precipitation from previous 2 months
p3sum	Sum of monthly average precipitation from previous 3 months
p6sum	Sum of monthly average precipitation from previous 6 months
PERMAVE	Soil permeability
RFACT	Rainfall/runoff factor
t0	Monthly average air temperature, current month
t1	Monthly average air temperature, previous month
t2	Monthly average air temperature, 2 months ago
t3	Monthly average air temperature, 3 months ago
t4	Monthly average air temperature, 4 months ago
t5	Monthly average air temperature, 5 months ago
t6	Monthly average air temperature, 6 months ago
t7	Monthly average air temperature, 7 months ago
t8	Monthly average air temperature, 8 months ago

t9	Monthly average air temperature, 9 months ago
t10	Monthly average air temperature, 10 months ago
t11	Monthly average air temperature, 11 months ago
t12	Monthly average air temperature, 12 months ago
WD_BASIN	Average annual wet days

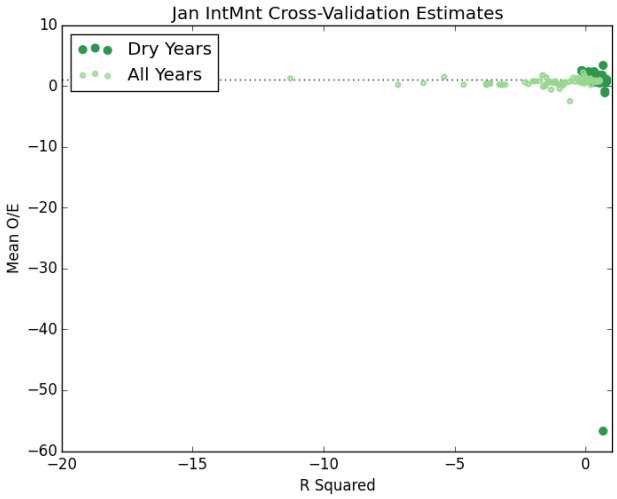
For the Sacramento basin model, the above list of 34 predictor variables was retained, plus the below 12 predictor variables:

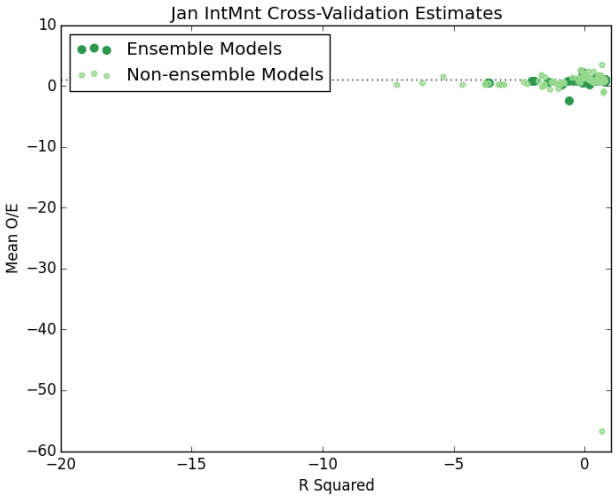
Variable	Definition
IntMnt	Binary indicating presence in Intermountain ecoregion
jan	Binary indicating occurrence of flow in January
feb	Binary indicating occurrence of flow in February
mar	Binary indicating occurrence of flow in March
apr	Binary indicating occurrence of flow in April
may	Binary indicating occurrence of flow in May
jun	Binary indicating occurrence of flow in June
jul	Binary indicating occurrence of flow in July
aug	Binary indicating occurrence of flow in August
sep	Binary indicating occurrence of flow in September
oct	Binary indicating occurrence of flow in October
nov	Binary indicating occurrence of flow in November

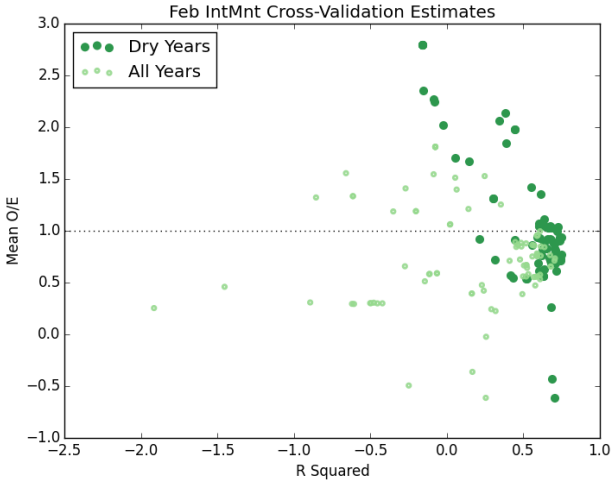
APPENDIX C: CROSS-VALIDATION RESULTS FOR EACH SCENARIO

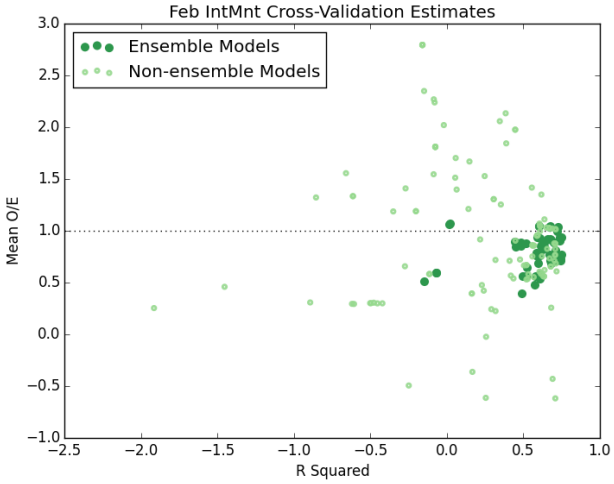
Dry-Year Regional Monthly Models

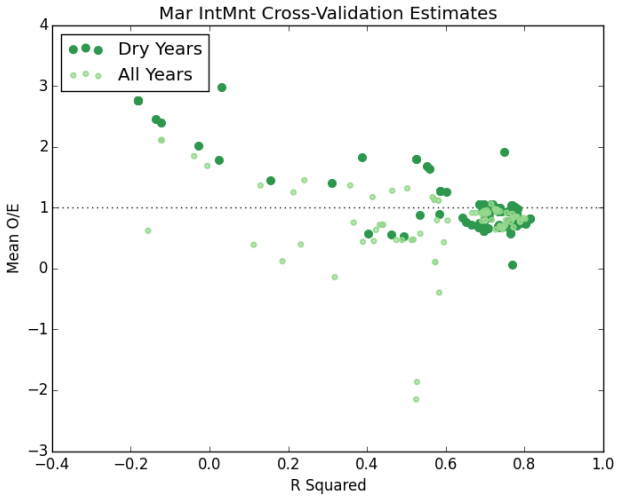
Intermountain Monthly Models

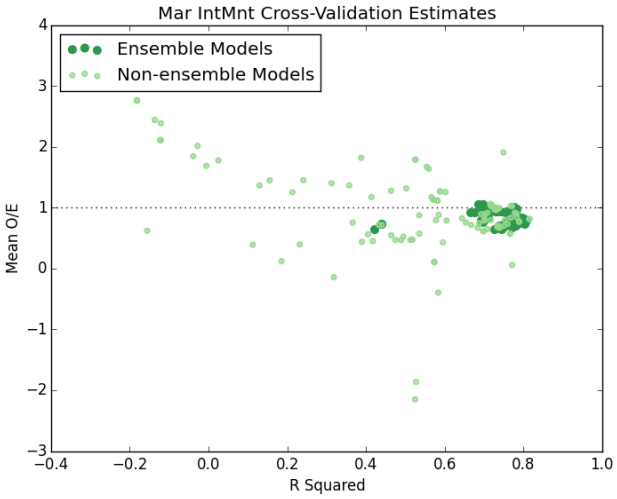


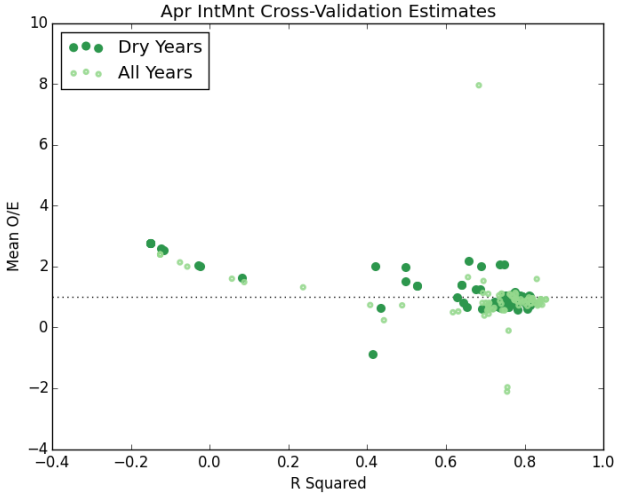


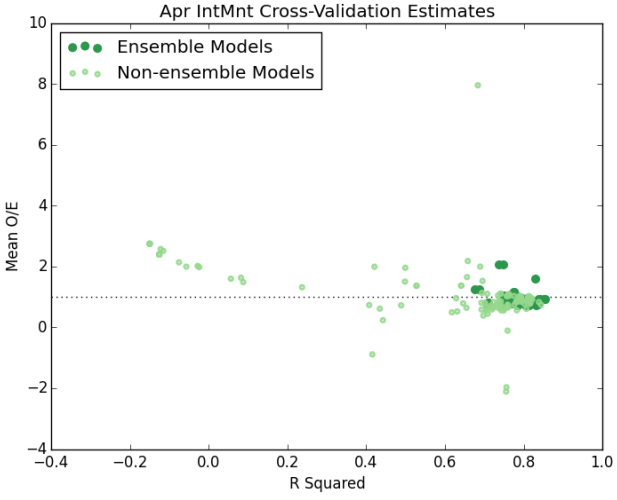


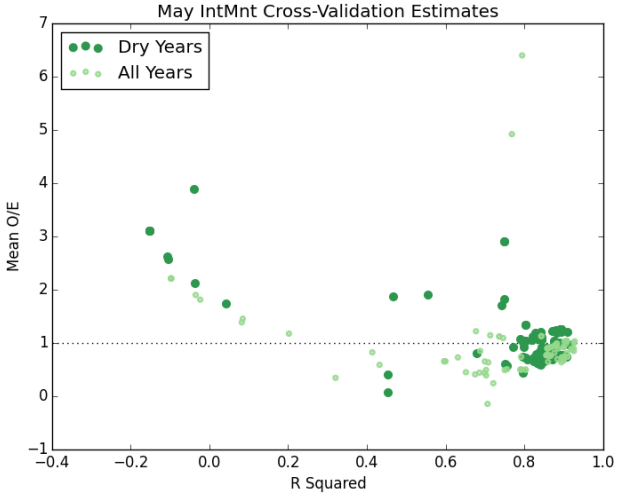


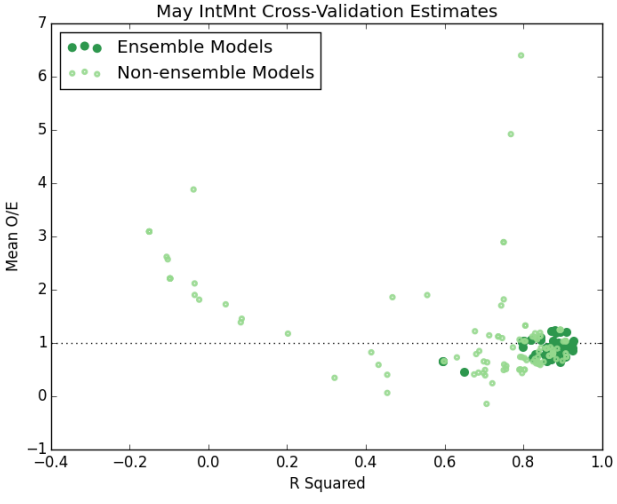

Mar IntMnt Cross-Validation Estimates
Dry Years
All Years
Mean O/E
R Squared

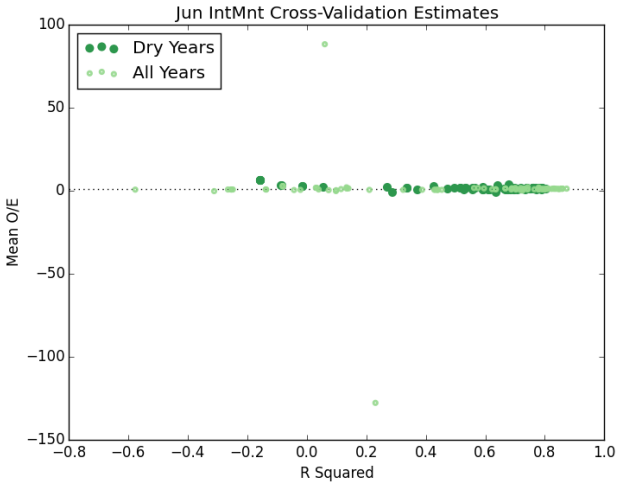

Mar IntMnt Cross-Validation Estimates
Ensemble Models
Non-ensemble Models
Mean O/E
R Squared

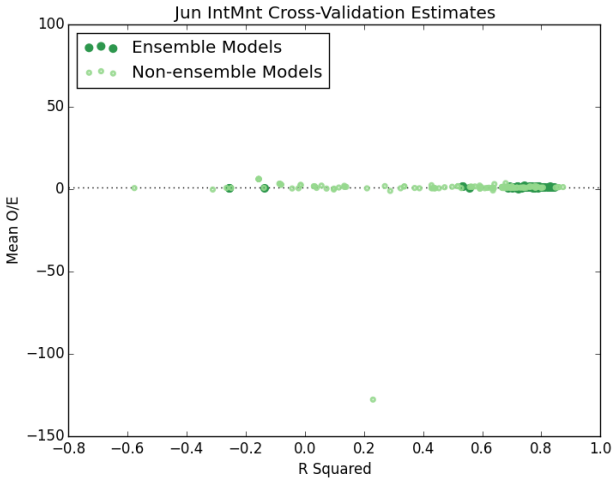

Apr IntMnt Cross-Validation Estimates
Dry Years
All Years
Mean O/E
R Squared

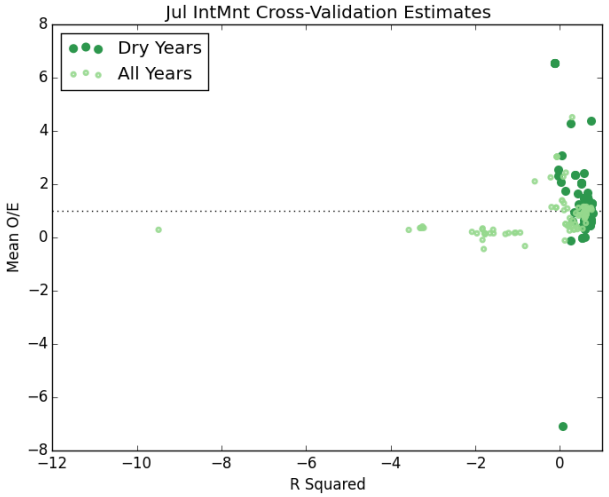

Apr IntMnt Cross-Validation Estimates
Ensemble Models
Non-ensemble Models
Mean O/E
R Squared

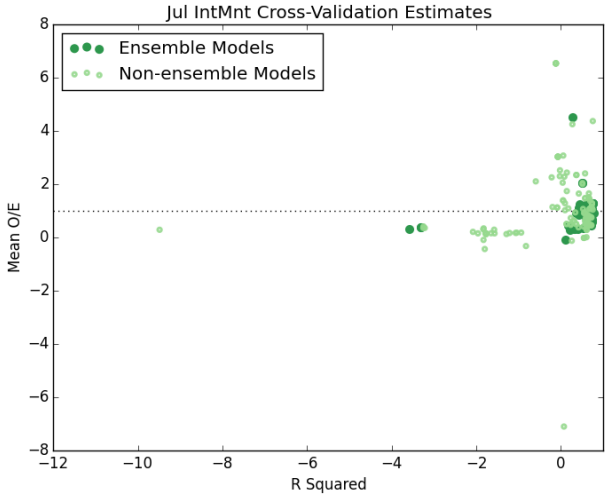

May IntMnt Cross-Validation Estimates
Dry Years
All Years
Mean O/E
R Squared

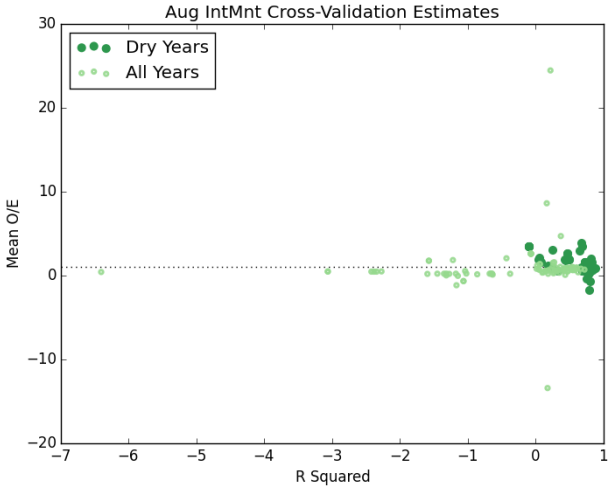

May IntMnt Cross-Validation Estimates
Ensemble Models
Non-ensemble Models
Mean O/E
R Squared

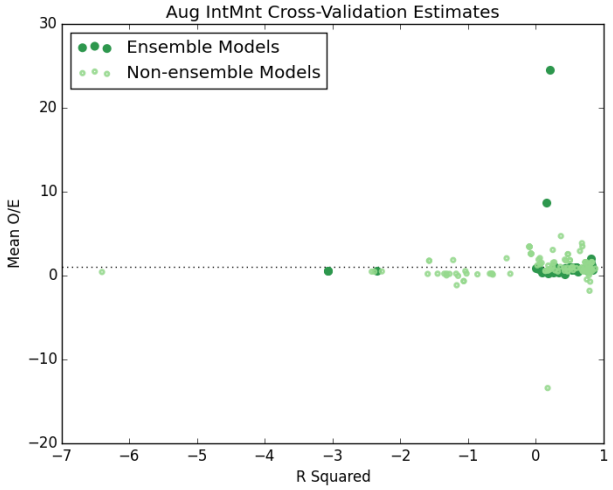

Jun IntMnt Cross-Validation Estimates
Dry Years
All Years
Mean O/E
R Squared

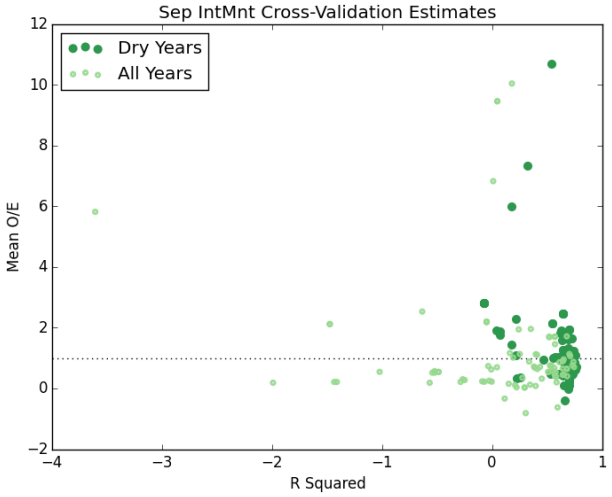

Jun IntMnt Cross-Validation Estimates
Ensemble Models
Non-ensemble Models
Mean O/E
R Squared

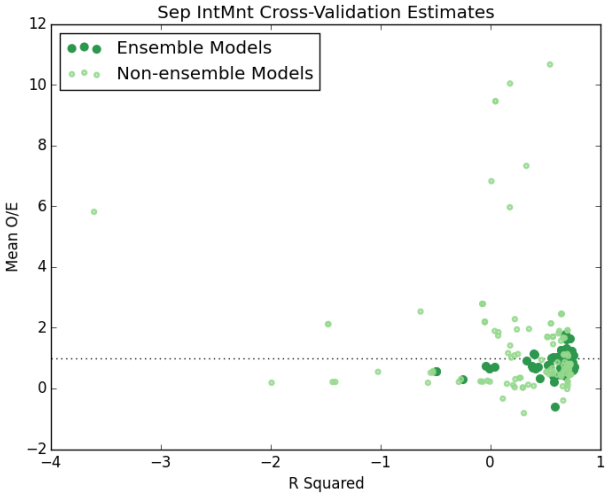

Jul IntMnt Cross-Validation Estimates
Dry Years
All Years
Mean O/E
R Squared

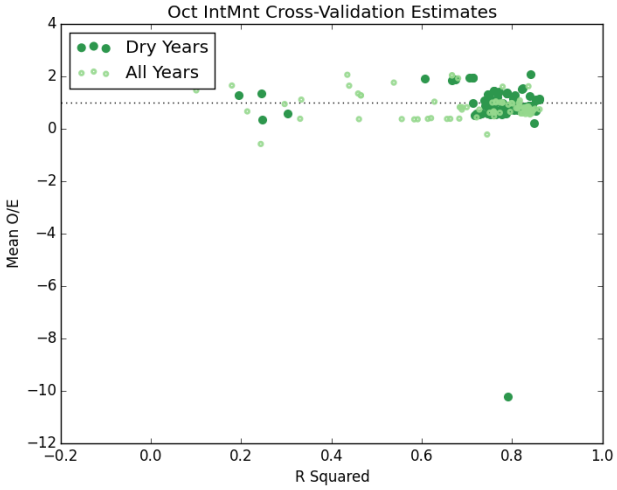

Jul IntMnt Cross-Validation Estimates
Ensemble Models
Non-ensemble Models
Mean O/E
R Squared

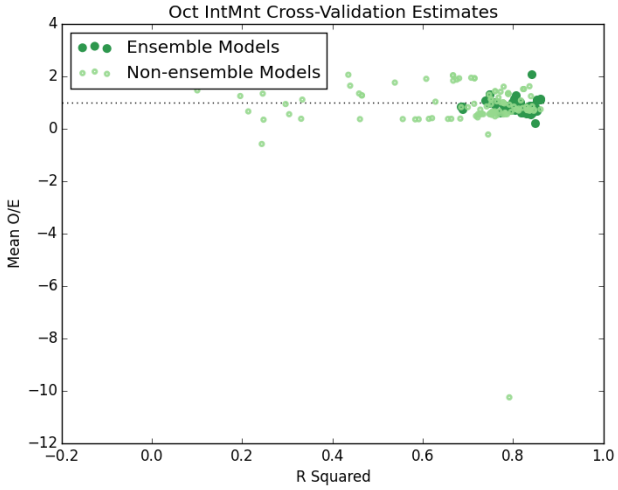

Aug IntMnt Cross-Validation Estimates
Dry Years
All Years
Mean O/E
R Squared

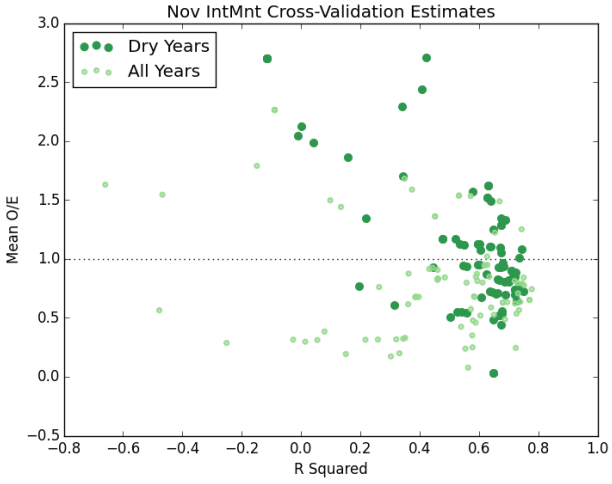

Aug IntMnt Cross-Validation Estimates
Ensemble Models
Non-ensemble Models
Mean O/E
R Squared

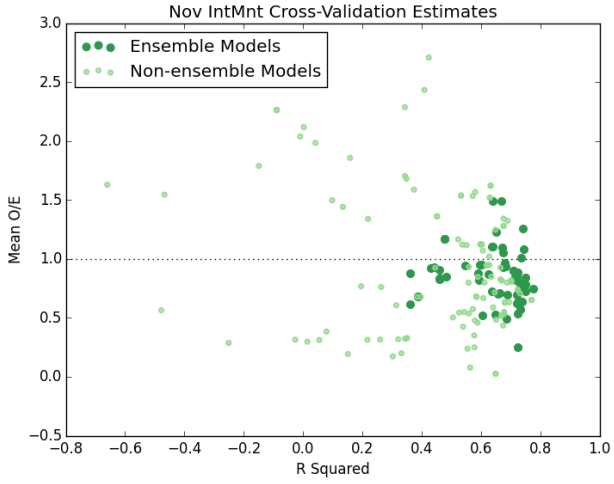

Sep IntMnt Cross-Validation Estimates
Dry Years
All Years
Mean O/E
R Squared

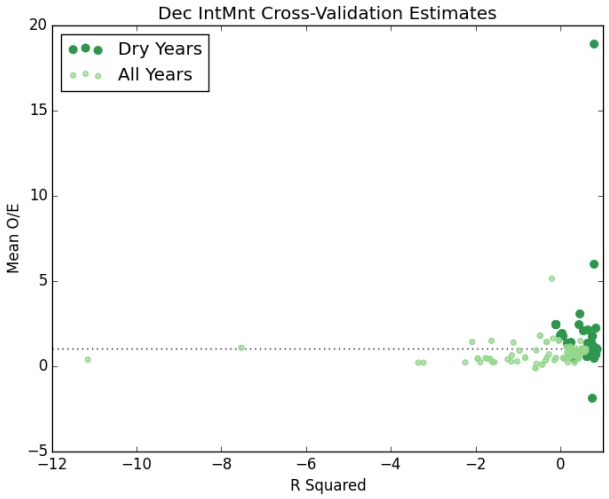

Sep IntMnt Cross-Validation Estimates
Ensemble Models
Non-ensemble Models
Mean O/E
R Squared

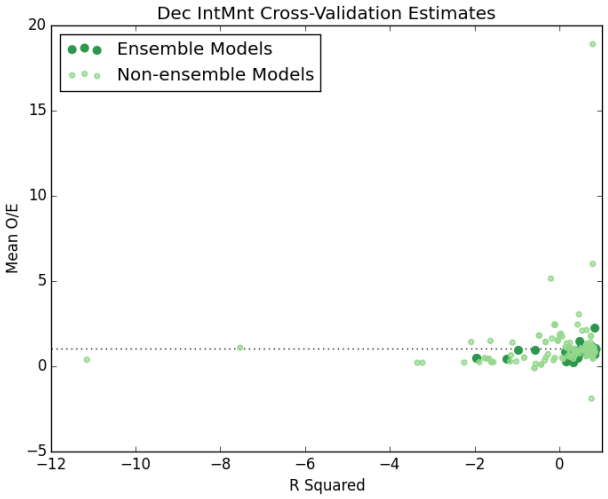

Oct IntMnt Cross-Validation Estimates
Dry Years
All Years
Mean O/E
R Squared


Oct IntMnt Cross-Validation Estimates
Ensemble Models
Non-ensemble Models
Mean O/E
R Squared

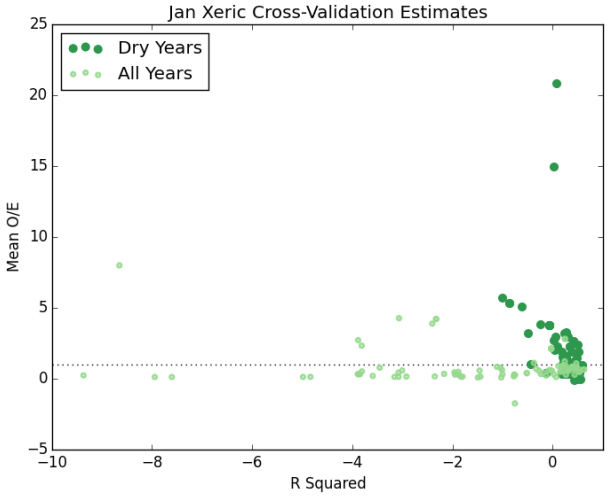

Nov IntMnt Cross-Validation Estimates
Dry Years
All Years
Mean O/E
R Squared

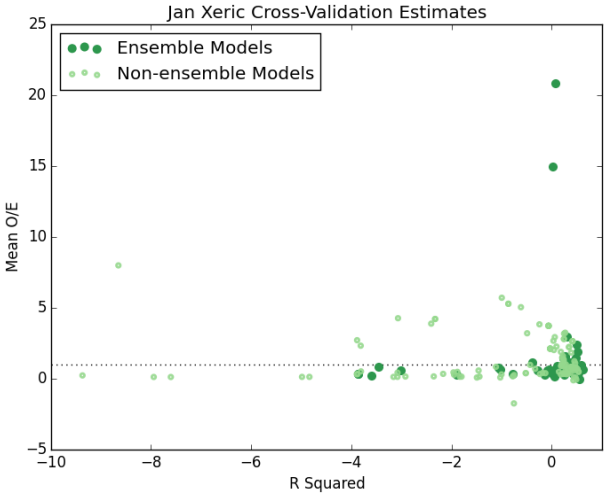

Nov IntMnt Cross-Validation Estimates
Ensemble Models
Non-ensemble Models
Mean O/E
R Squared

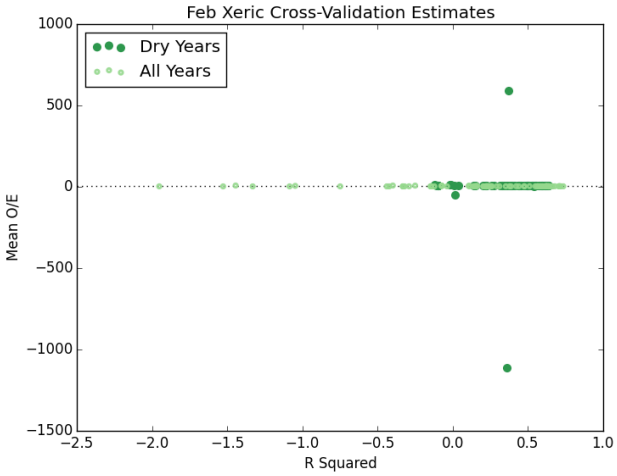


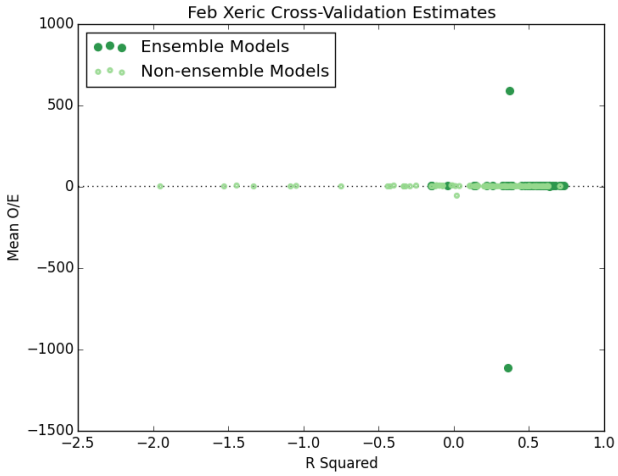


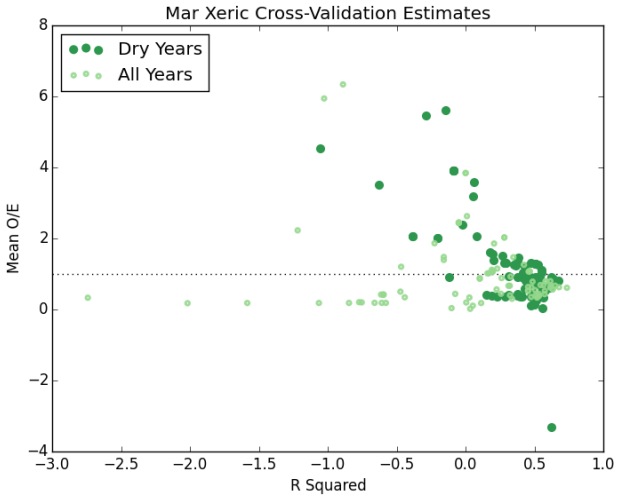
Xeric Monthly Models

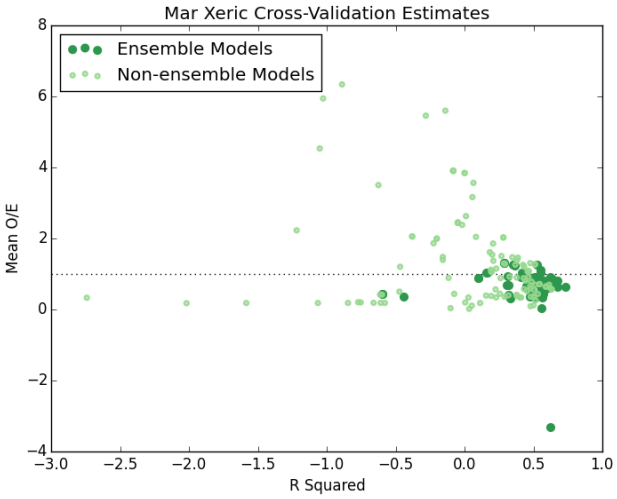


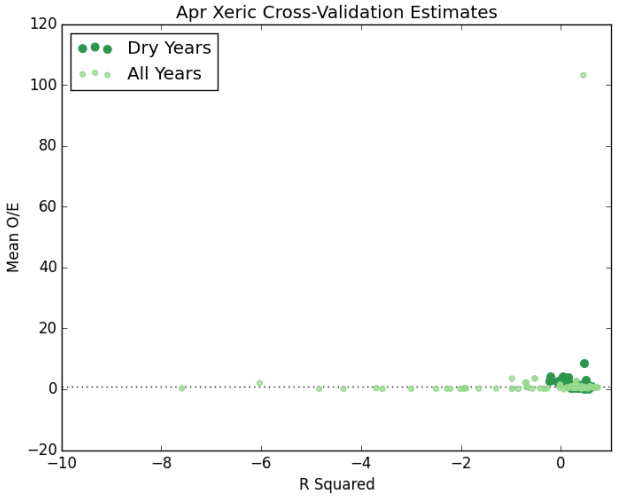


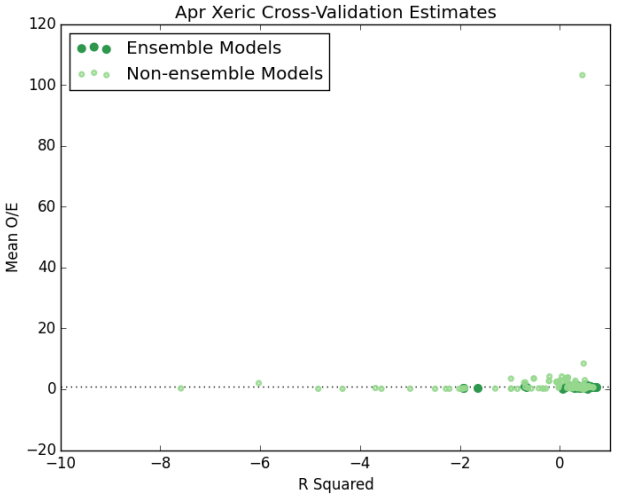


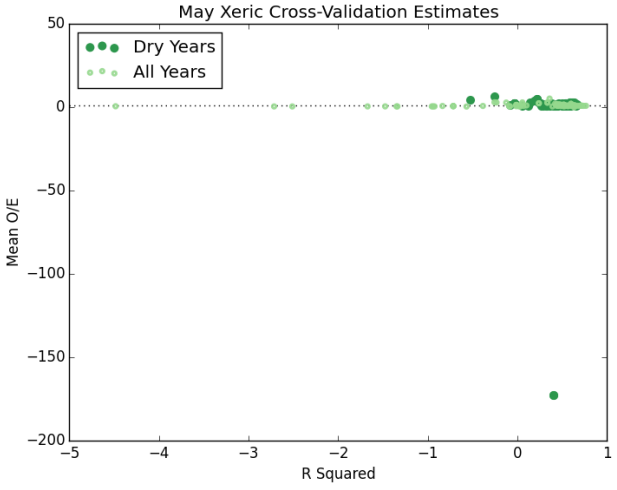


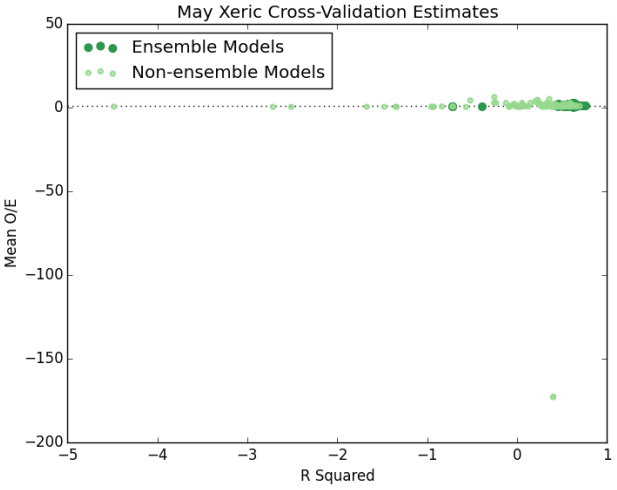

Mar Xeric Cross-Validation Estimates
Dry Years
All Years
Mean O/E
R Squared

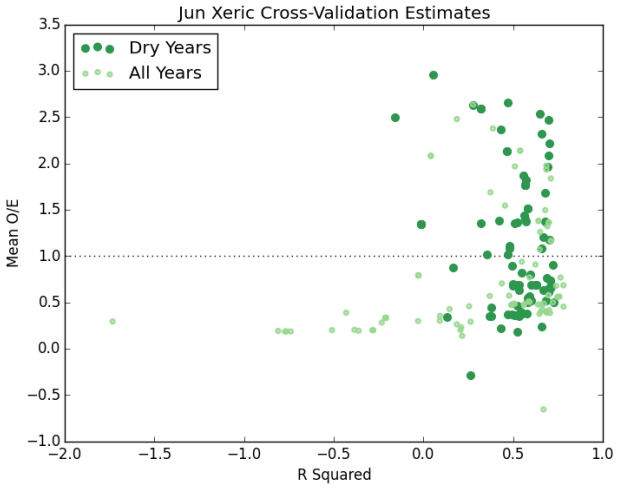

Mar Xeric Cross-Validation Estimates
Ensemble Models
Non-ensemble Models
Mean O/E
R Squared

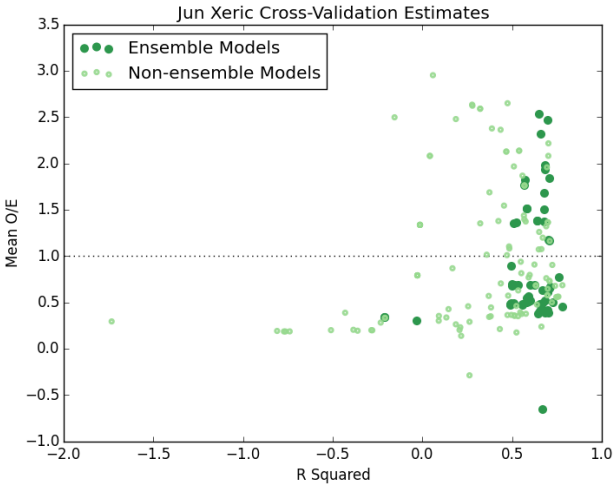

Apr Xeric Cross-Validation Estimates
Dry Years
All Years
Mean O/E
R Squared

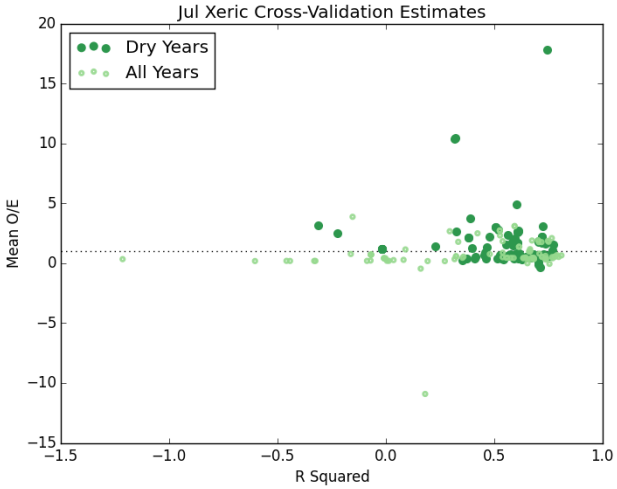

Apr Xeric Cross-Validation Estimates
Ensemble Models
Non-ensemble Models
Mean O/E
R Squared

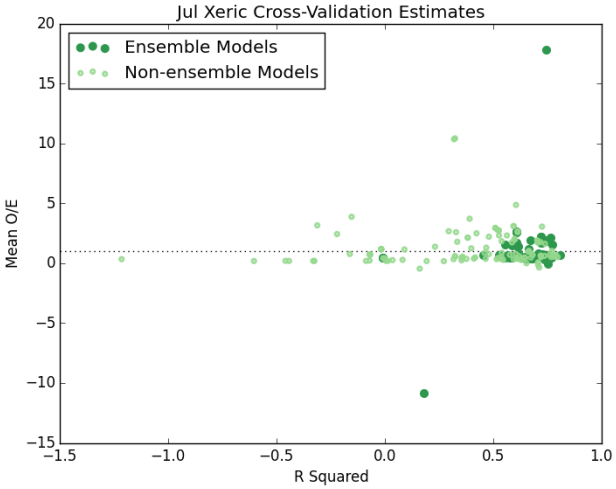

May Xeric Cross-Validation Estimates
Dry Years
All Years
Mean O/E
R Squared

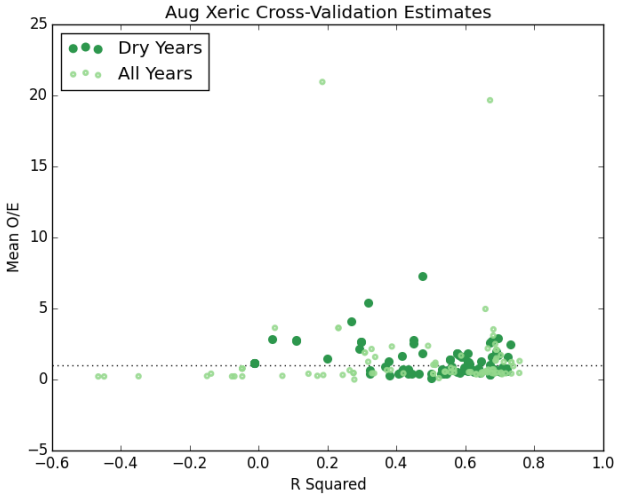

May Xeric Cross-Validation Estimates
Ensemble Models
Non-ensemble Models
Mean O/E
R Squared

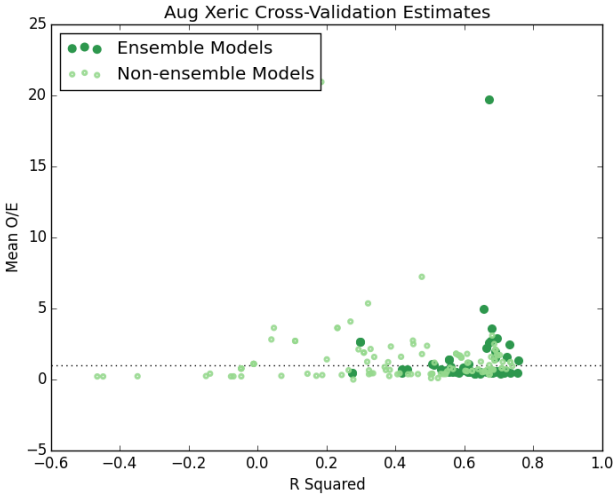

Jun Xeric Cross-Validation Estimates
Dry Years
All Years
Mean O/E
R Squared

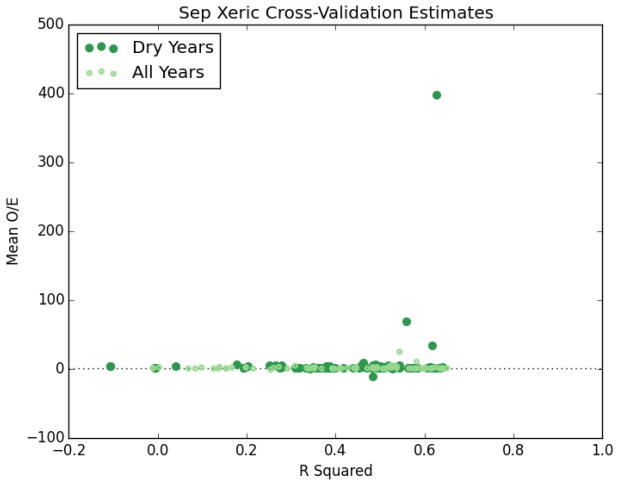

Jun Xeric Cross-Validation Estimates
Ensemble Models
Non-ensemble Models
Mean O/E
R Squared

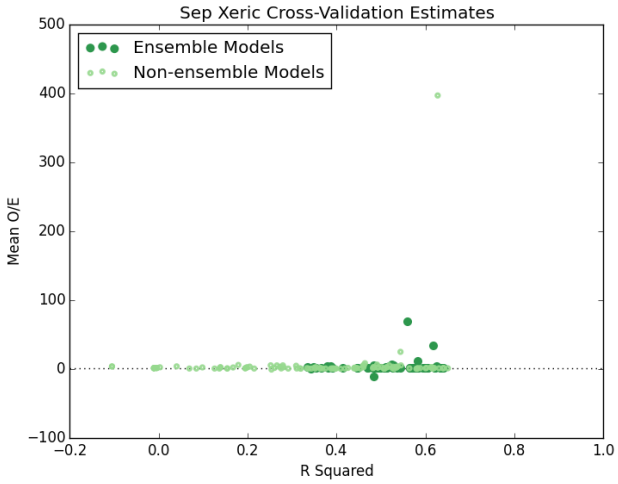

Jul Xeric Cross-Validation Estimates
Dry Years
All Years
Mean O/E
R Squared

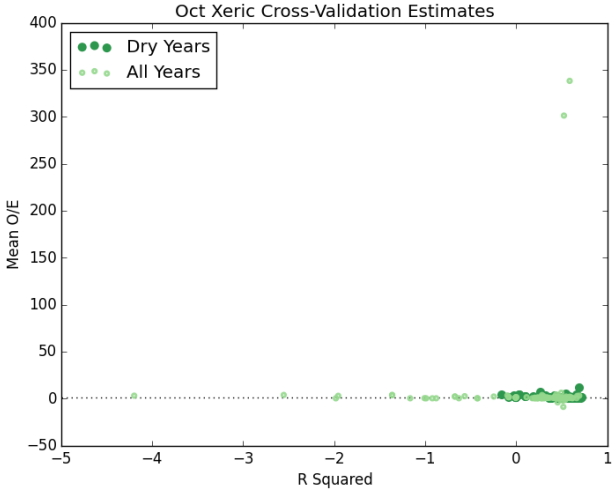

Jul Xeric Cross-Validation Estimates
Ensemble Models
Non-ensemble Models
Mean O/E
R Squared

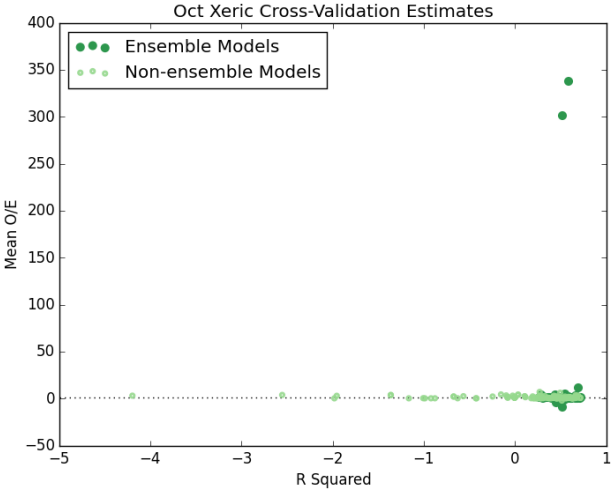

Aug Xeric Cross-Validation Estimates
Dry Years
All Years
Mean O/E
R Squared

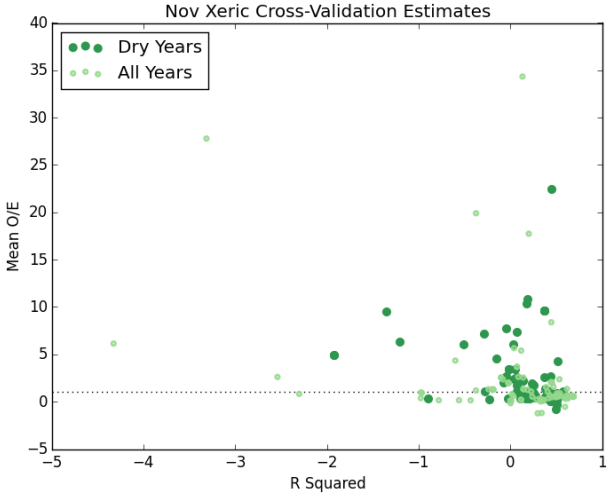

Aug Xeric Cross-Validation Estimates
Ensemble Models
Non-ensemble Models
Mean O/E
R Squared

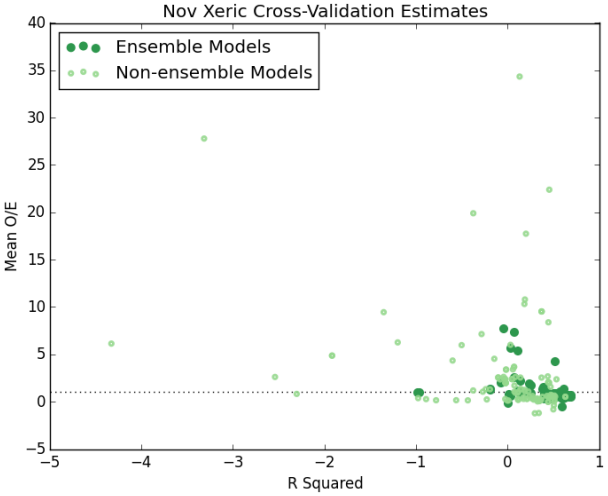

Sep Xeric Cross-Validation Estimates
Dry Years
All Years
Mean O/E
R Squared

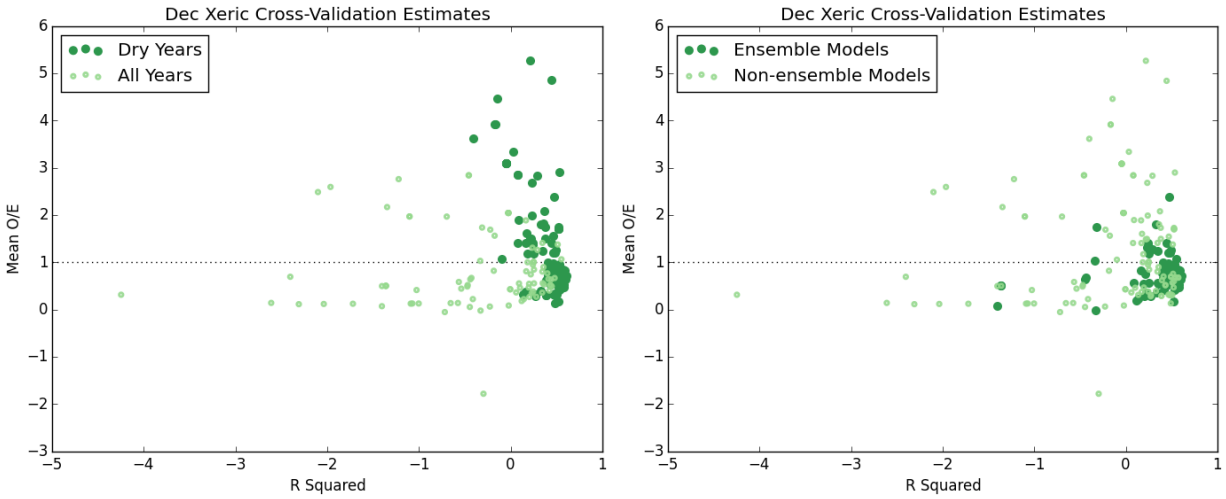

Sep Xeric Cross-Validation Estimates
Ensemble Models
Non-ensemble Models
Mean O/E
R Squared


Oct Xeric Cross-Validation Estimates
Dry Years
All Years
Mean O/E
R Squared


Oct Xeric Cross-Validation Estimates
Ensemble Models
Non-ensemble Models
Mean O/E
R Squared

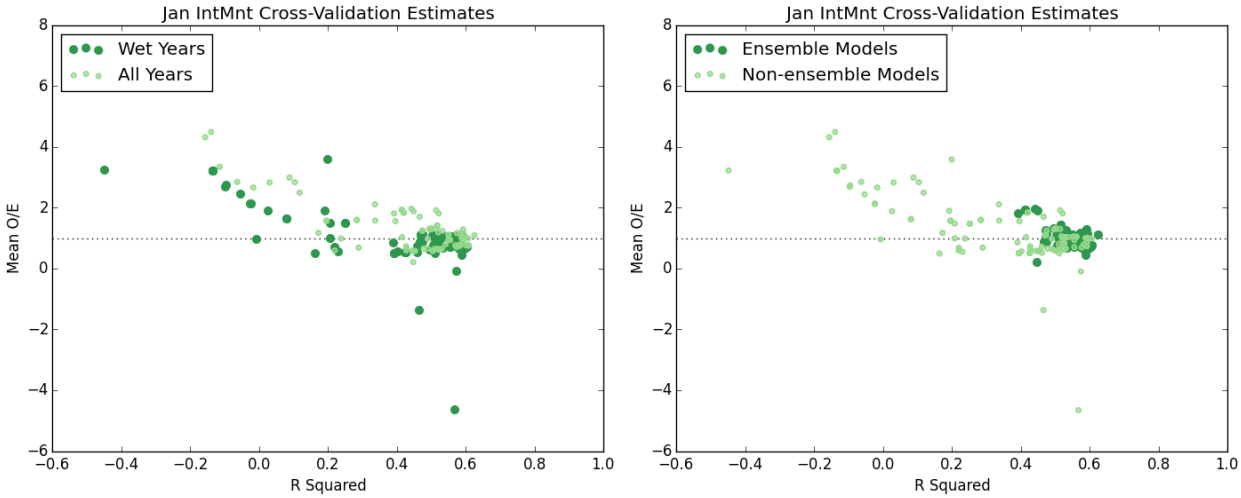

Nov Xeric Cross-Validation Estimates
Dry Years
All Years
Mean O/E
R Squared

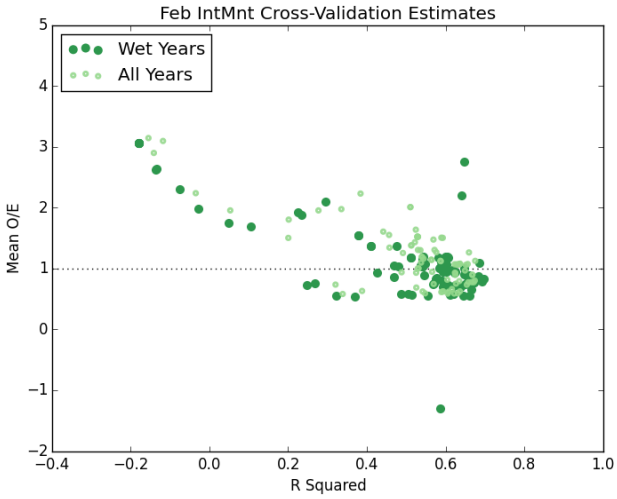

Nov Xeric Cross-Validation Estimates
Ensemble Models
Non-ensemble Models
Mean O/E
R Squared

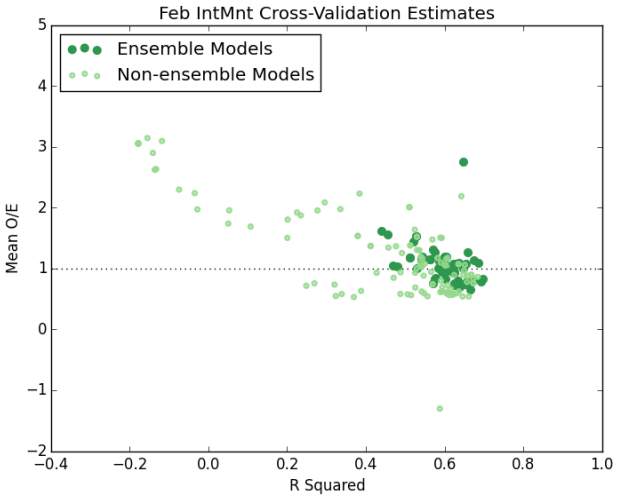


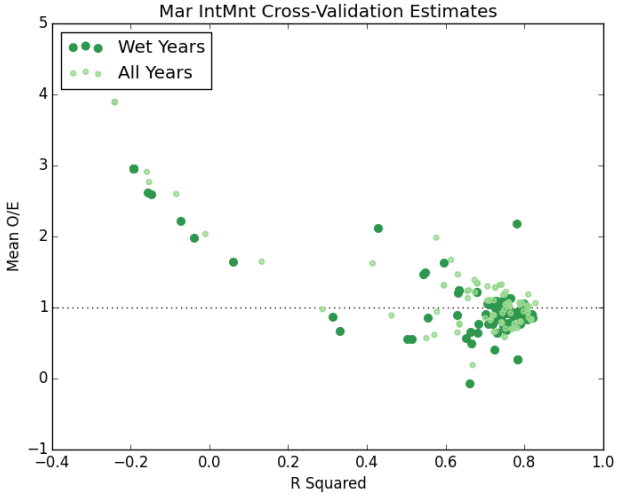
Wet-Year Regional Monthly Models

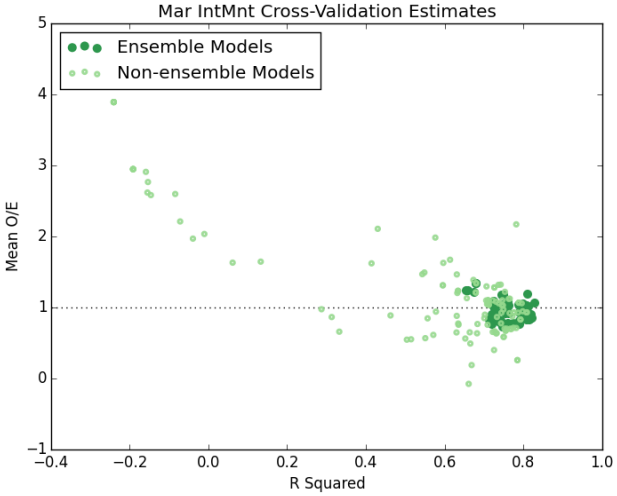
Intermountain Monthly Models

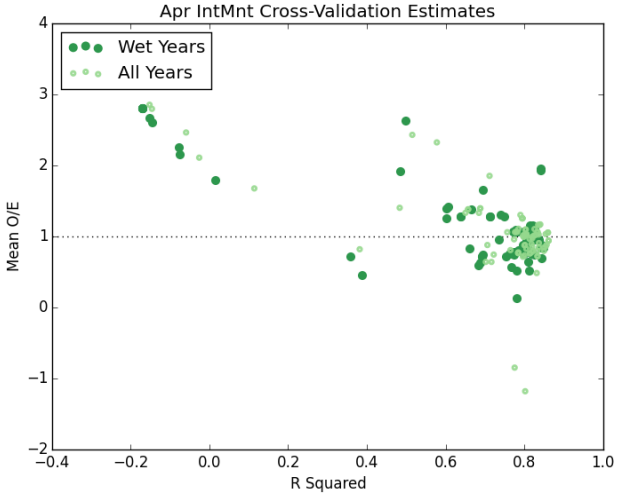


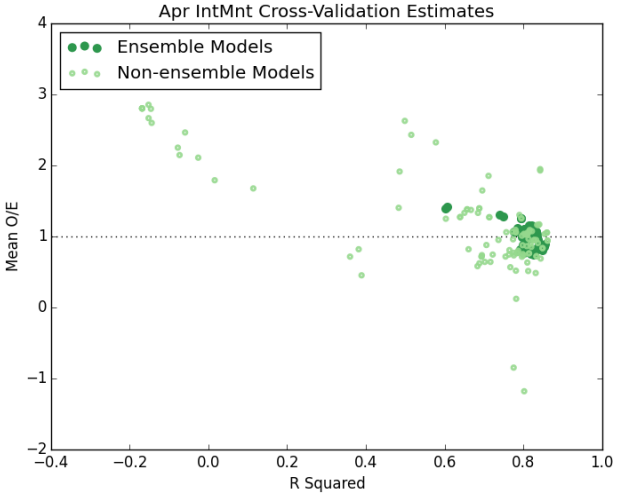

Feb IntMnt Cross-Validation Estimates
Wet Years
All Years
Mean O/E
R Squared

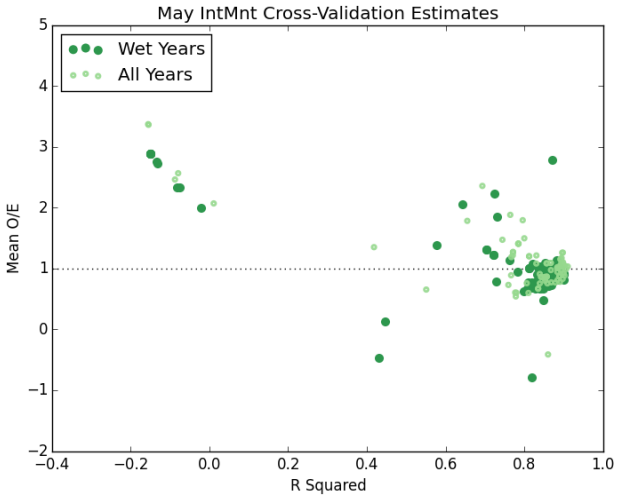

Feb IntMnt Cross-Validation Estimates
Ensemble Models
Non-ensemble Models
Mean O/E
R Squared

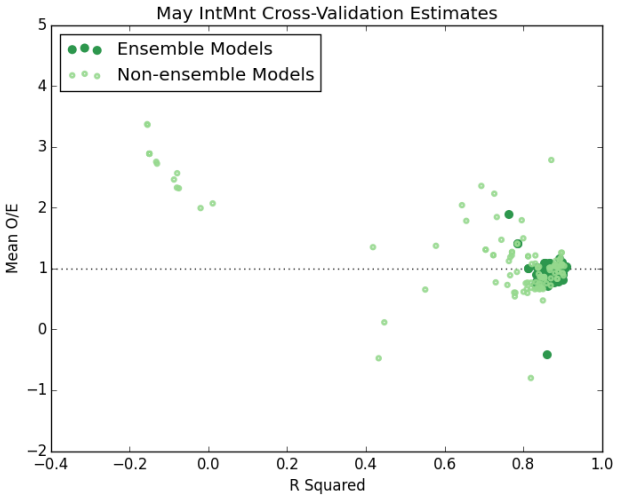

Mar IntMnt Cross-Validation Estimates
Wet Years
All Years
Mean O/E
R Squared

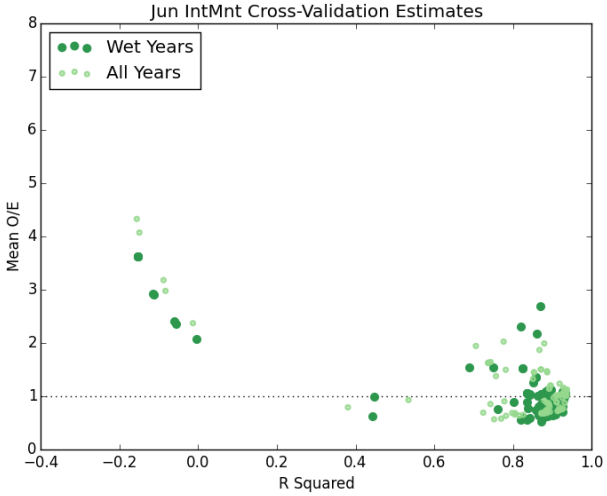

Mar IntMnt Cross-Validation Estimates
Ensemble Models
Non-ensemble Models
Mean O/E
R Squared

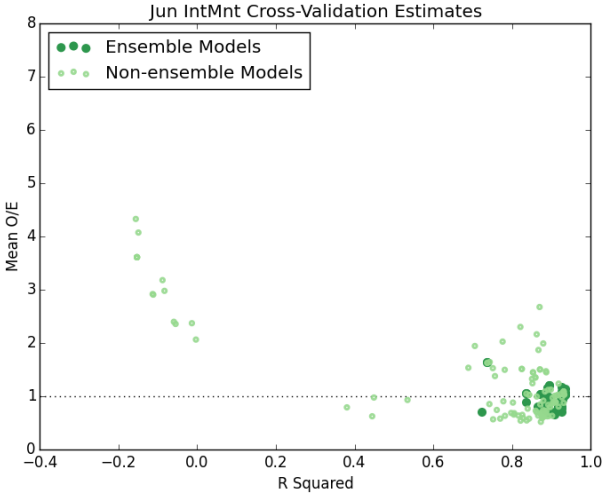

Apr IntMnt Cross-Validation Estimates
Wet Years
All Years
Mean O/E
R Squared

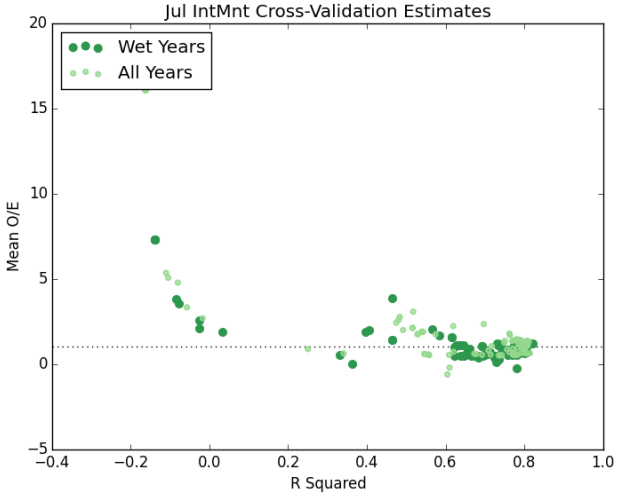

Apr IntMnt Cross-Validation Estimates
Ensemble Models
Non-ensemble Models
Mean O/E
R Squared

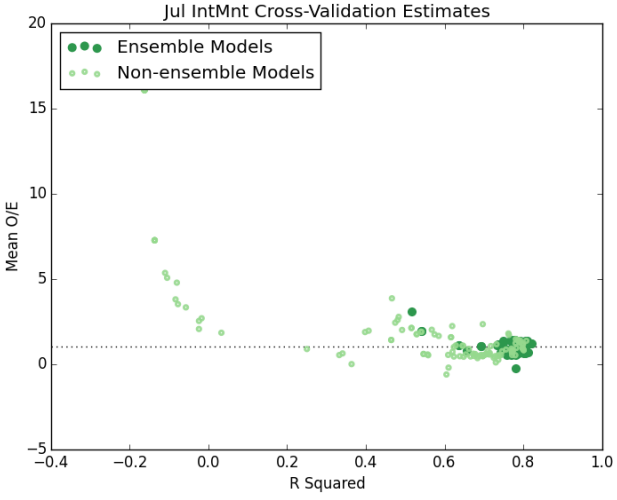

May IntMnt Cross-Validation Estimates
Wet Years
All Years
Mean O/E
R Squared

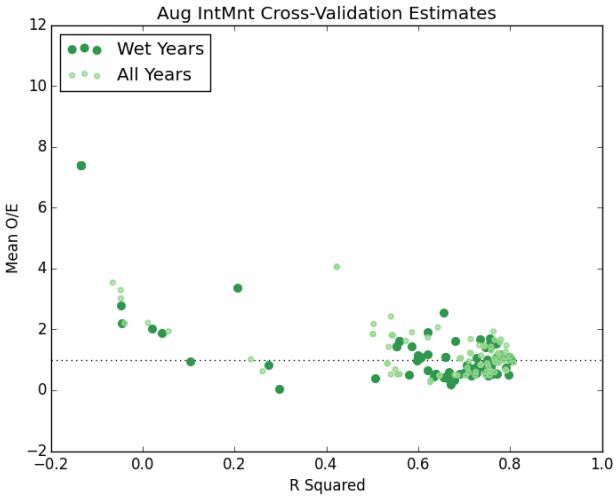

May IntMnt Cross-Validation Estimates
Ensemble Models
Non-ensemble Models
Mean O/E
R Squared

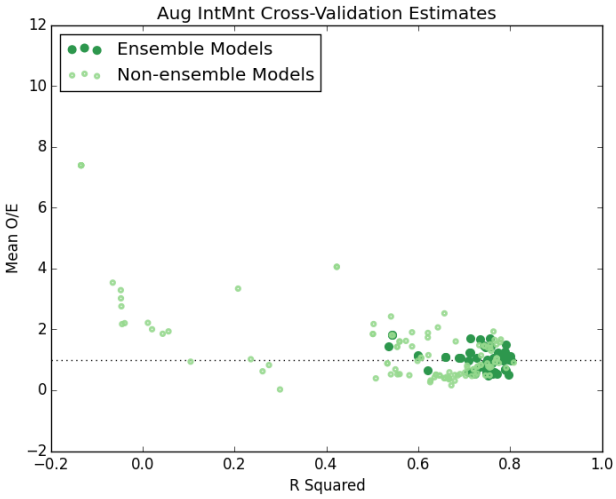

Jun IntMnt Cross-Validation Estimates
Wet Years
All Years
Mean O/E
R Squared

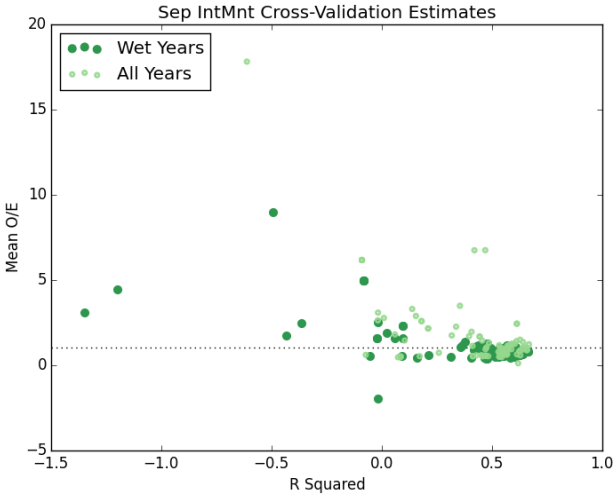

Jun IntMnt Cross-Validation Estimates
Ensemble Models
Non-ensemble Models
Mean O/E
R Squared

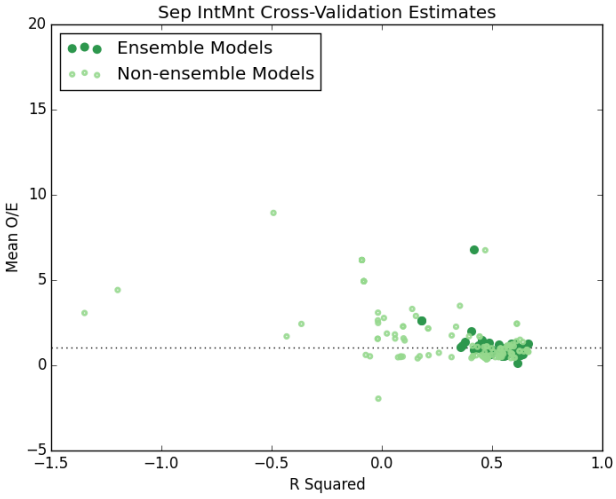

Jul IntMnt Cross-Validation Estimates
Wet Years
All Years
Mean O/E
R Squared

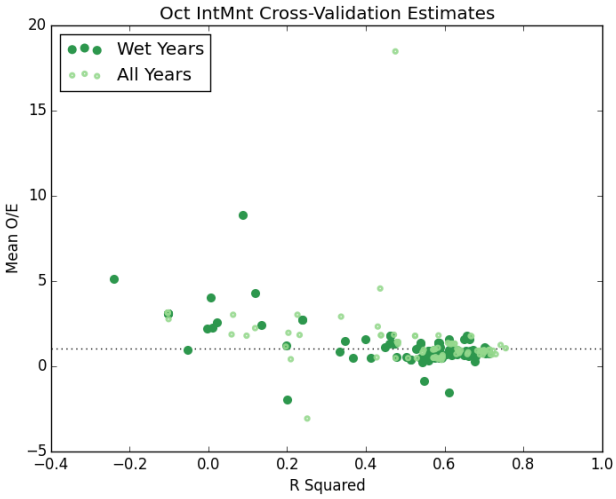

Jul IntMnt Cross-Validation Estimates
Ensemble Models
Non-ensemble Models
Mean O/E
R Squared

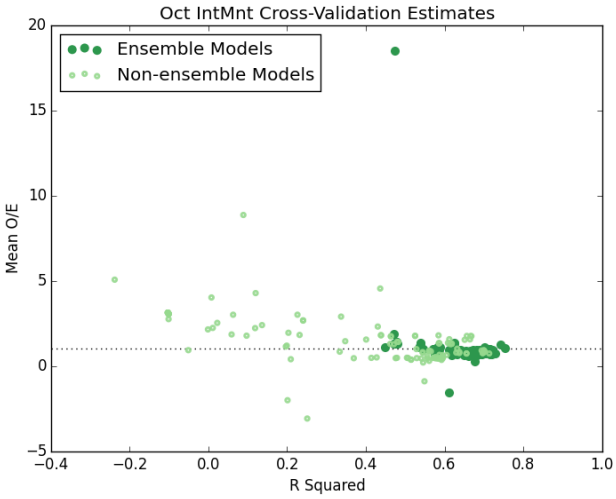

Aug IntMnt Cross-Validation Estimates
Wet Years
All Years
Mean O/E
R Squared

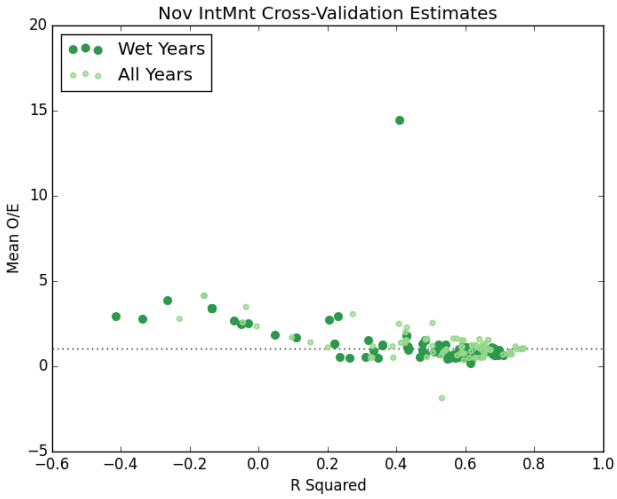

Aug IntMnt Cross-Validation Estimates
Ensemble Models
Non-ensemble Models
Mean O/E
R Squared

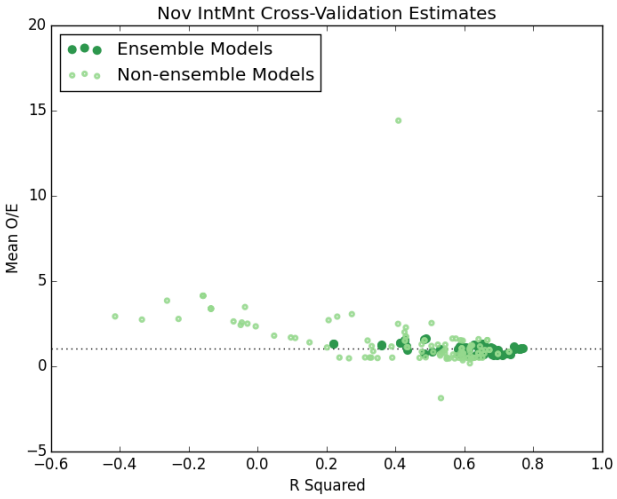

Sep IntMnt Cross-Validation Estimates
Wet Years
All Years
Mean O/E
R Squared

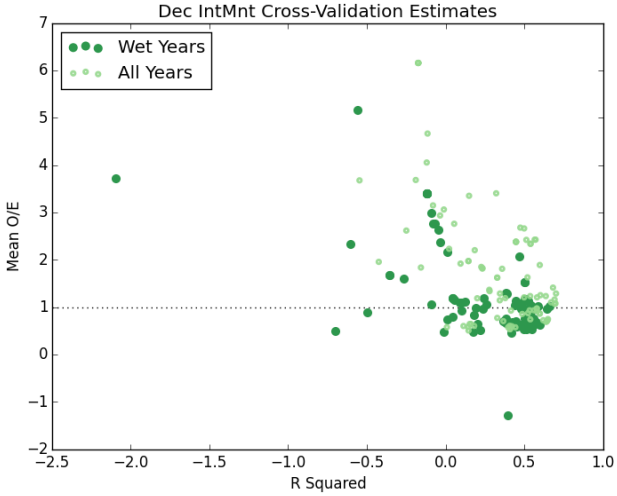

Sep IntMnt Cross-Validation Estimates
Ensemble Models
Non-ensemble Models
Mean O/E
R Squared

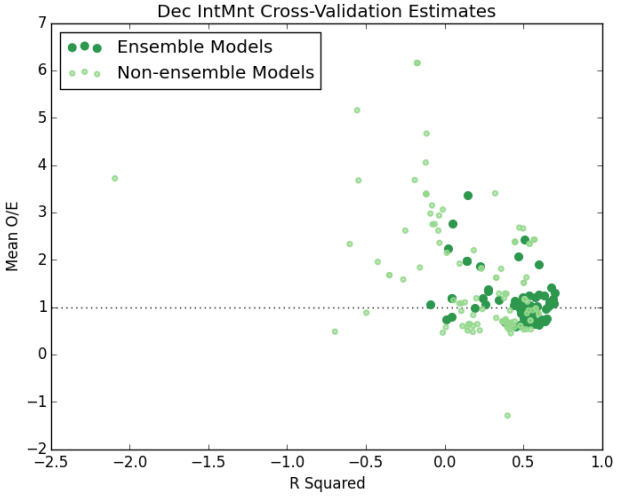

Oct IntMnt Cross-Validation Estimates
Wet Years
All Years
Mean O/E
R Squared


Oct IntMnt Cross-Validation Estimates
Ensemble Models
Non-ensemble Models
Mean O/E
R Squared

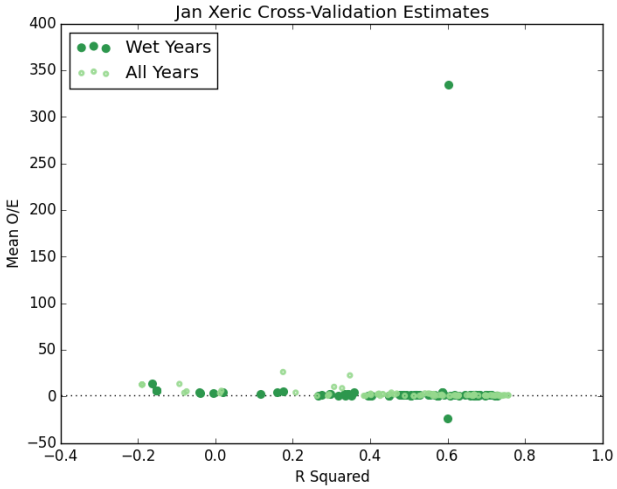


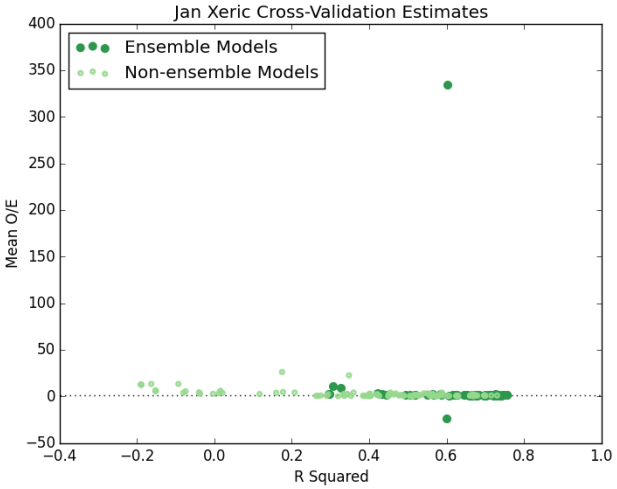


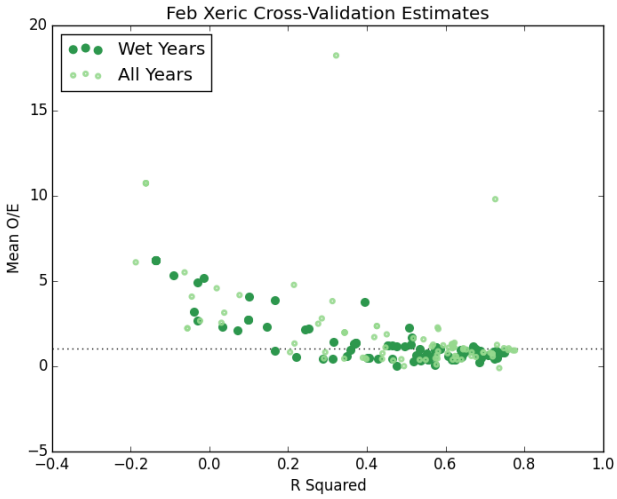


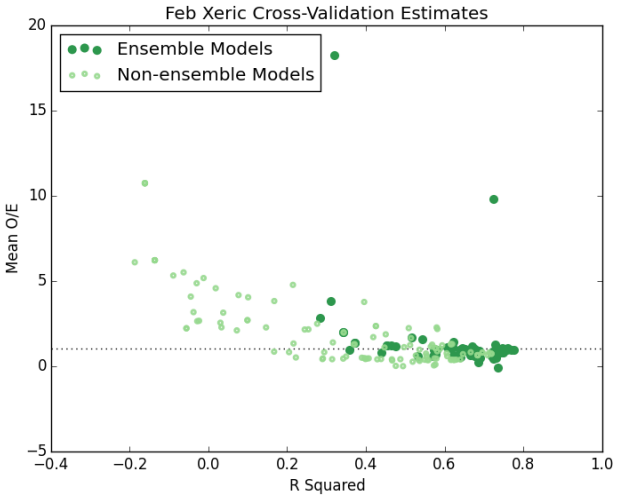


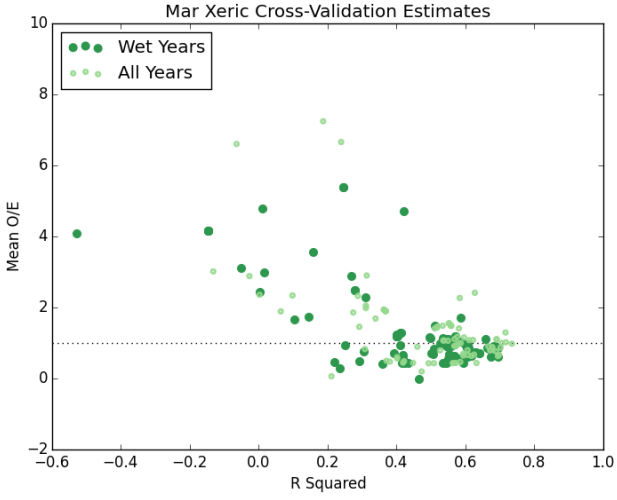
Xeric Monthly Models

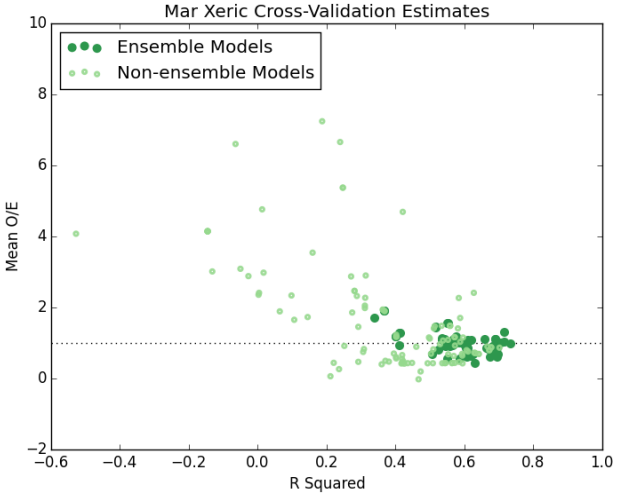


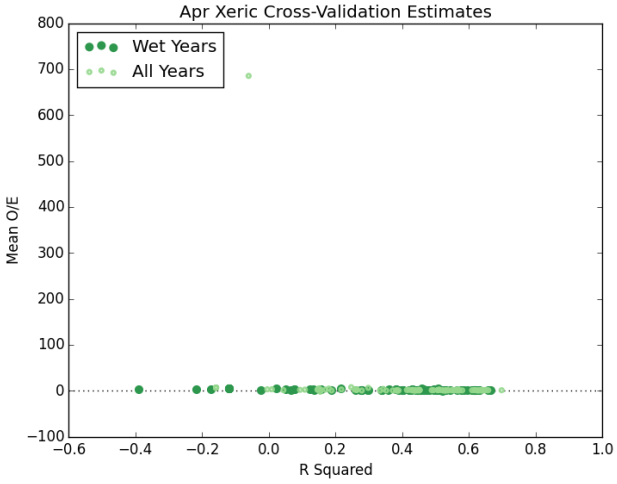


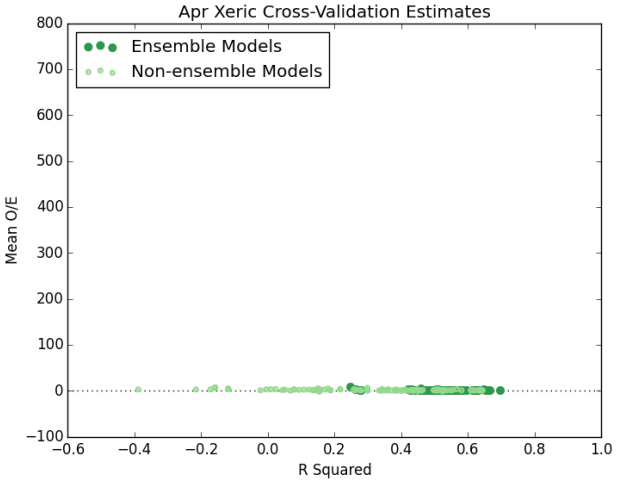

Feb Xeric Cross-Validation Estimates
Wet Years
All Years
Mean O/E
R Squared

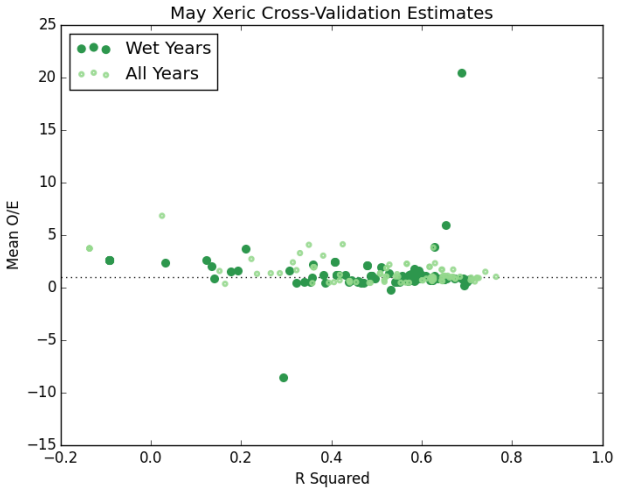

Feb Xeric Cross-Validation Estimates
Ensemble Models
Non-ensemble Models
Mean O/E
R Squared

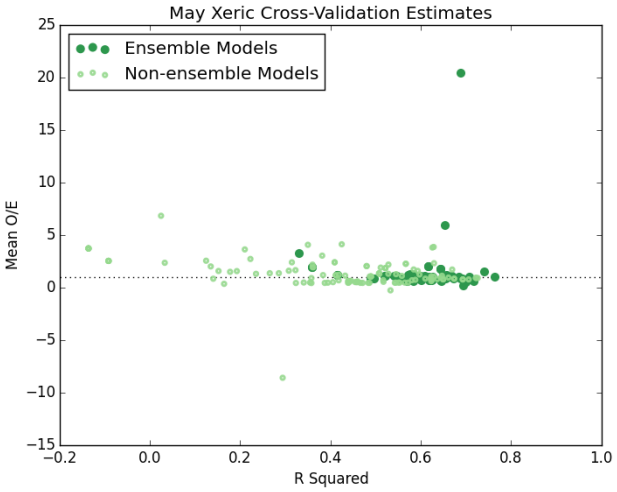

Mar Xeric Cross-Validation Estimates
Wet Years
All Years
Mean O/E
R Squared

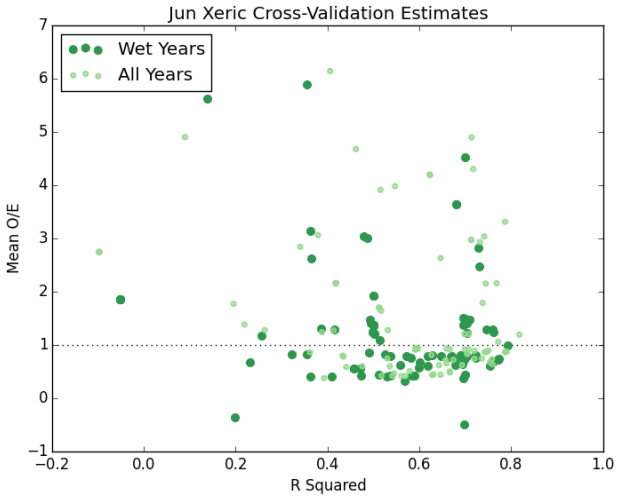

Mar Xeric Cross-Validation Estimates
Ensemble Models
Non-ensemble Models
Mean O/E
R Squared

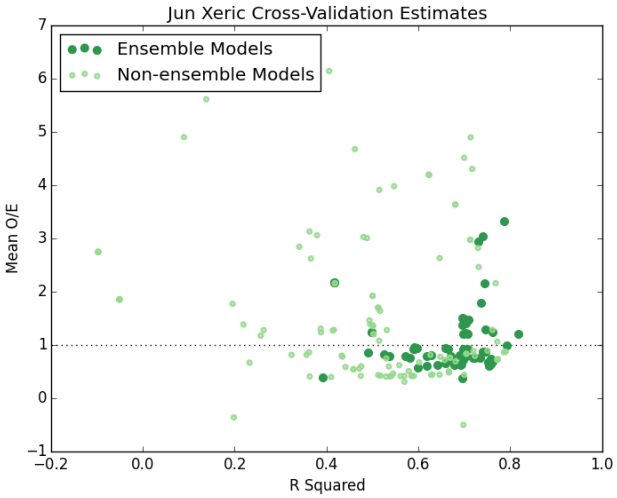

Apr Xeric Cross-Validation Estimates
Wet Years
All Years
Mean O/E
R Squared

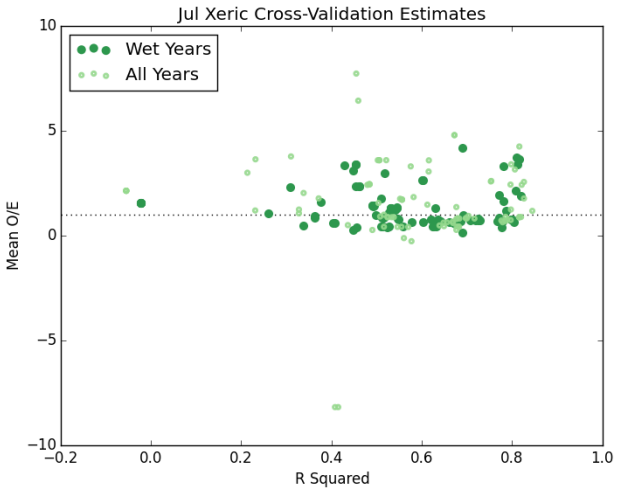

Apr Xeric Cross-Validation Estimates
Ensemble Models
Non-ensemble Models
Mean O/E
R Squared

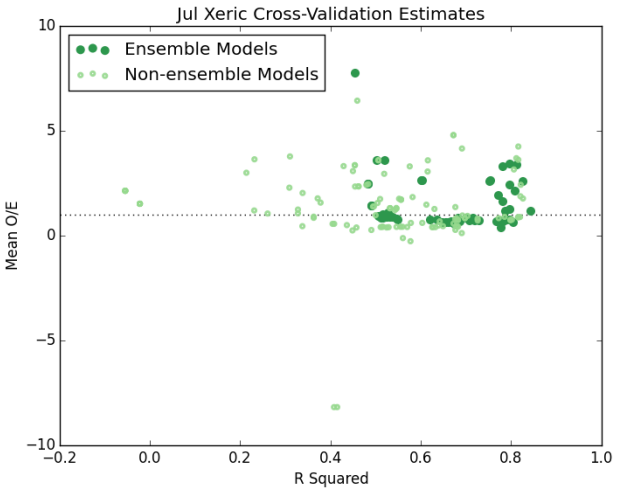

May Xeric Cross-Validation Estimates
Wet Years
All Years
Mean O/E
R Squared

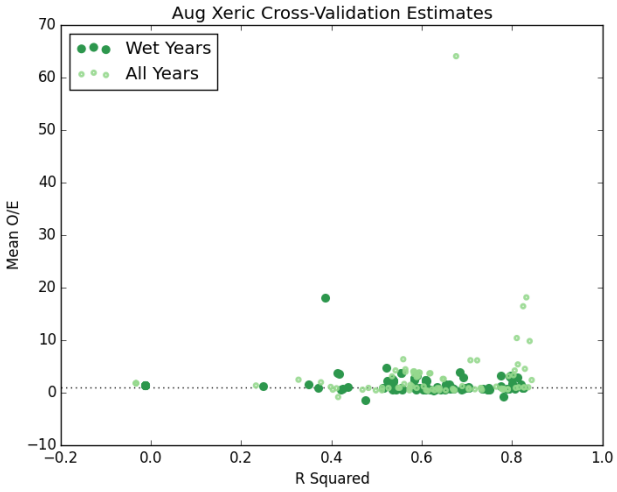

May Xeric Cross-Validation Estimates
Ensemble Models
Non-ensemble Models
Mean O/E
R Squared

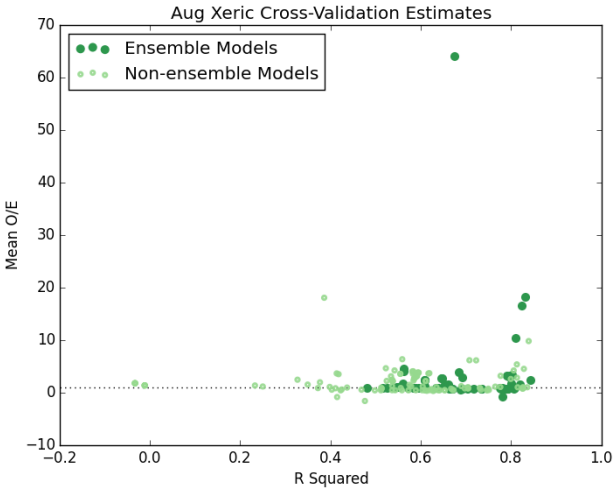

Jun Xeric Cross-Validation Estimates
Wet Years
All Years
Mean O/E
R Squared

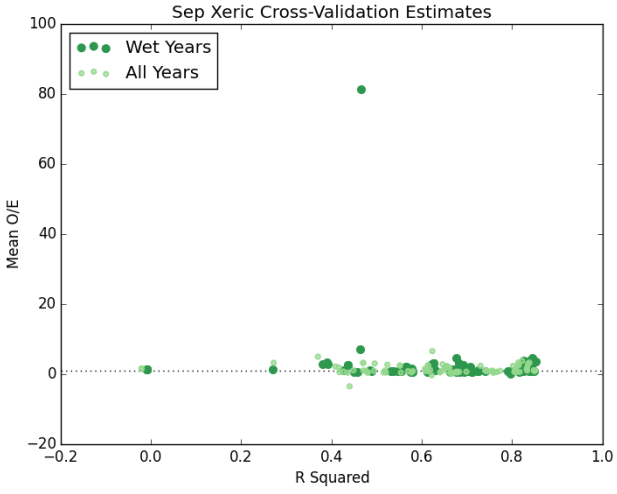

Jun Xeric Cross-Validation Estimates
Ensemble Models
Non-ensemble Models
Mean O/E
R Squared

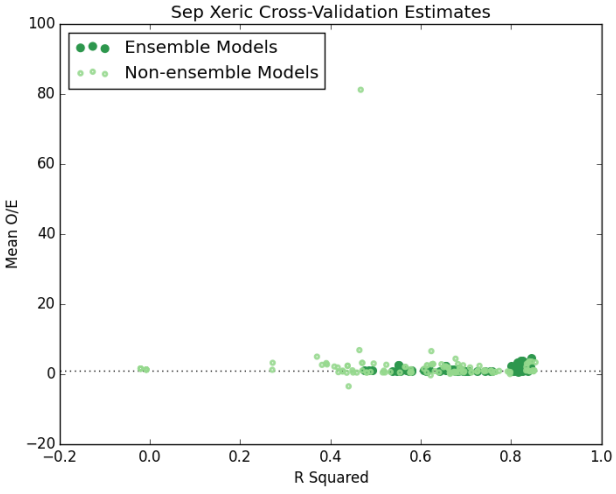

Jul Xeric Cross-Validation Estimates
Wet Years
All Years
Mean O/E
R Squared

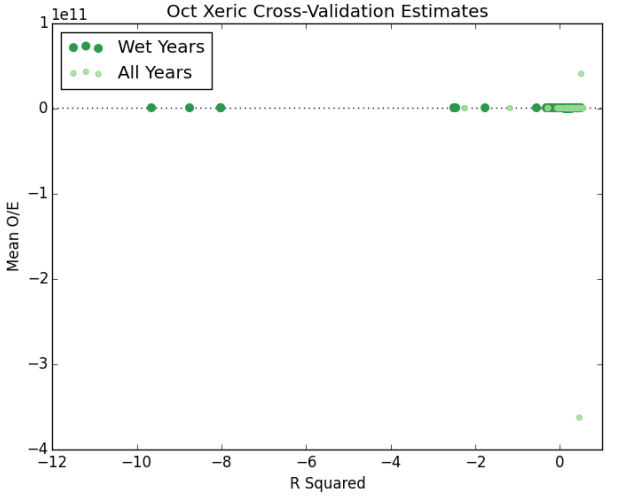

Jul Xeric Cross-Validation Estimates
Ensemble Models
Non-ensemble Models
Mean O/E
R Squared

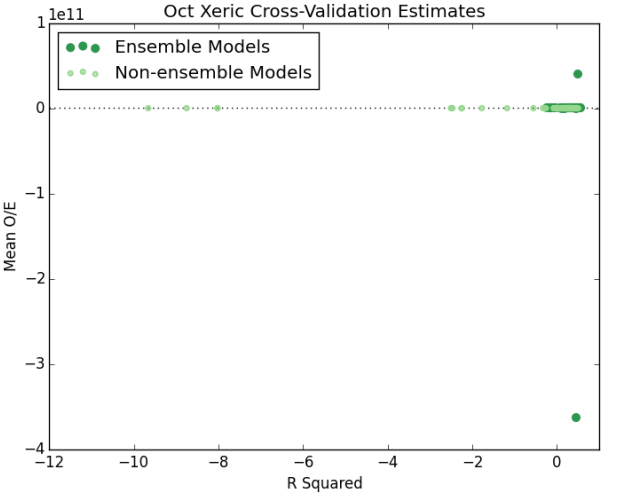

Aug Xeric Cross-Validation Estimates
Wet Years
All Years
Mean O/E
R Squared

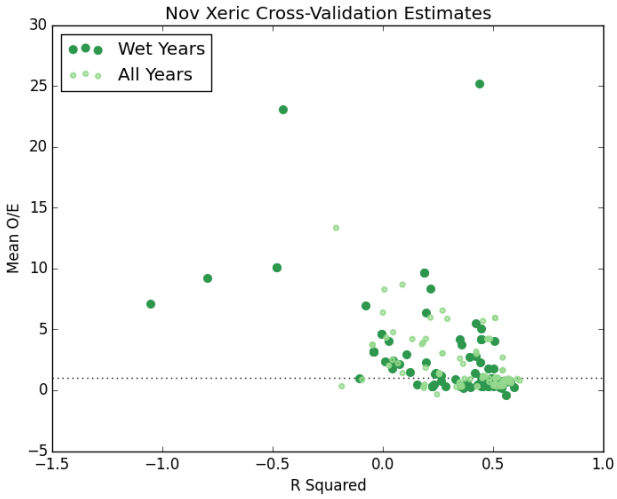

Aug Xeric Cross-Validation Estimates
Ensemble Models
Non-ensemble Models
Mean O/E
R Squared

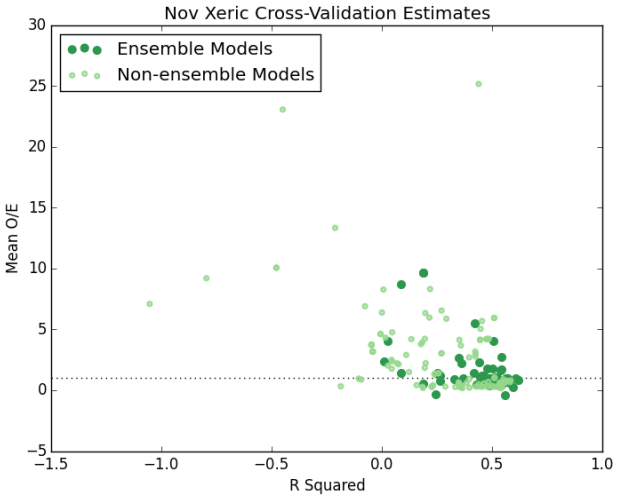

Sep Xeric Cross-Validation Estimates
Wet Years
All Years
Mean O/E
R Squared

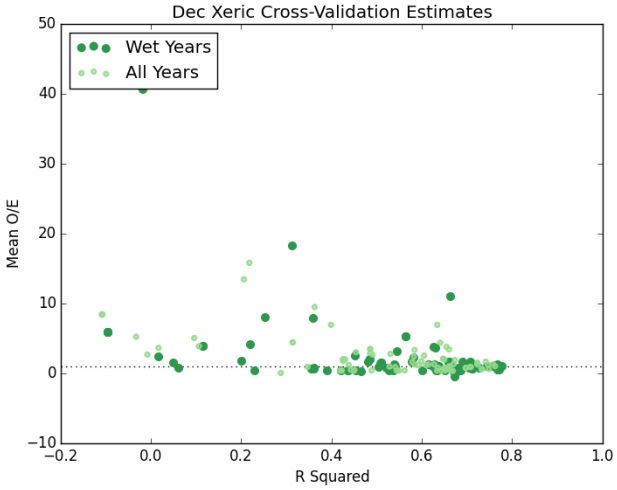

Sep Xeric Cross-Validation Estimates
Ensemble Models
Non-ensemble Models
Mean O/E
R Squared

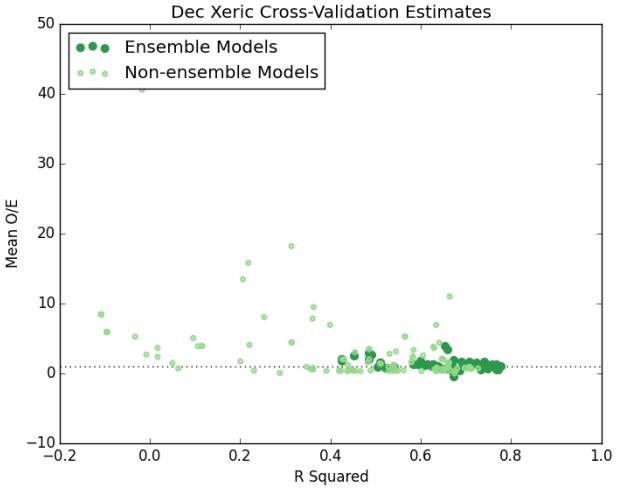

1e11
Oct Xeric Cross-Validation Estimates
Wet Years
All Years
Mean O/E
R Squared


1e11
Oct Xeric Cross-Validation Estimates
Ensemble Models
Non-ensemble Models
Mean O/E
R Squared

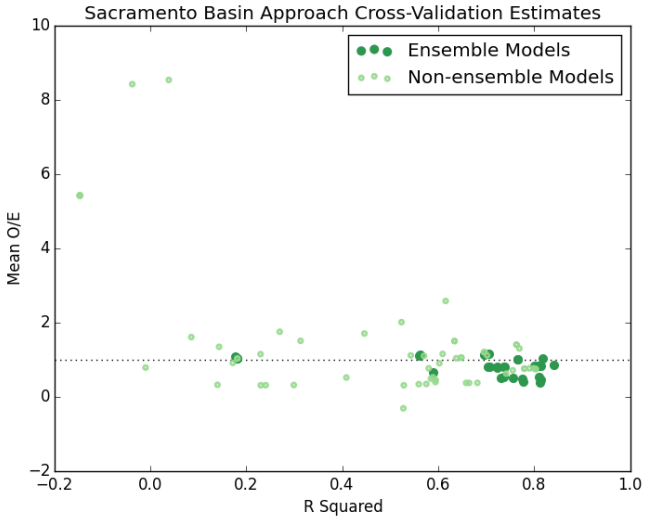








Sacramento Model



APPENDIX D: DETAILS OF BEST MONTHLY REGIONAL MODELS

These tables detail the base dataset and modeling method combination chosen as "best" (based on its R^2 value) for each monthly regional scenario as well as its estimated test performance metrics.

Best Dry-Year Intermountain Models for Each Month

Month	Base DataSet	Model Method	R^2	Mean O/E	Std. Dev. O/E	MSE	RMSE (cfs)
Jan	Dry Years Scaled	Stacking OF Ensemble	0.812	0.904	0.564	690.790	25.902
Feb	Dry Years	Stacking OF Ensemble	0.752	0.926	0.686	2901.184	50.814
Mar	Dry Years, Expert Selection	Random Forest	0.815	0.812	0.430	4616.699	65.426
Apr	All Years, Expert Selection	Stacking Ensemble	0.854	0.916	0.483	19688.253	138.999
May	All Years	Stacking OF Ensemble	0.928	1.023	0.778	34197.542	182.207
Jun	All Years, Expert Selection	K Nearest Neighbors	0.873	1.011	1.011	38603.548	195.324
Jul	Dry Years, PCA n50	Stacking Ensemble	0.784	0.891	0.687	8006.155	85.999
Aug	Dry Years Scaled, PCA n20	Random Forest	0.881	0.778	0.550	545.025	21.277
Sep	Dry Years Scaled, Expert Selection	Averaging Ensemble	0.768	0.705	1.341	508.578	21.159
Oct	Dry Years Scaled	Stacking OF Ensemble	0.862	1.073	1.061	328.793	17.567
Nov	All Years Scaled	Stacking OF Ensemble	0.778	0.743	0.455	610.925	24.365
Dec	Dry Years Scaled	Stacking Ensemble	0.851	1.024	0.619	409.535	19.694

Best Dry-Year Xeric Models for Each Month

Month	Base DataSet	Model Method	R^2	Mean O/E	Std. Dev. O/E	MSE	RMSE (cfs)
Jan	All Years, Expert Selection	Stacking Ensemble	0.629	0.616	0.880	1246.150	34.475
Feb	All Years, Variance Threshold .08	Stacking OF Ensemble	0.735	0.642	0.585	3501.843	57.141

Mar	All Years	Stacking OF Ensemble	0.734	0.600	0.534	1890.997	40.557
Apr	All Years	Stacking OF Ensemble	0.743	0.542	0.515	236.831	15.064
May	All Years, Variance Threshold .08	Stacking OF Ensemble	0.766	0.548	0.542	50.886	6.873
Jun	All Years	Stacking OF Ensemble	0.781	0.453	0.493	12.615	3.385
Jul	All Years Scaled	Stacking OF Ensemble	0.811	0.650	1.361	2.899	1.684
Aug	All Years Scaled	Stacking OF Ensemble	0.758	1.271	9.502	2.535	1.555
Sep	All Years, Variance Threshold .08	Random Forest	0.650	0.651	2.037	4.633	2.118
Oct	Dry Years Scaled	Stacking OF Ensemble	0.725	1.197	4.613	5.332	2.162
Nov	All Years	Stacking OF Ensemble	0.689	0.515	0.667	199.921	12.307
Dec	Dry Years, Variance Threshold .08	Averaging Ensemble	0.616	0.707	2.728	721.460	25.115

Best Wet-Year Intermountain Models for Each Month

Month	Base DataSet	Model Method	R^2	Mean O/E	Std. Dev. O/E	MSE	RMSE (cfs)
Jan	All Years Scaled	Stacking OF Ensemble	0.625	1.092	0.717	35513.983	177.415
Feb	Wet Years Scaled	Averaging Ensemble	0.698	0.821	0.826	18041.230	131.488
Mar	All Years Scaled	Stacking OF Ensemble	0.829	1.054	0.452	12316.922	110.517
Apr	All Years Scaled, via Expert Selection	Random Forest	0.861	0.935	0.403	33528.940	173.316
May	All Years, Variance Threshold .08	Stacking OF Ensemble	0.910	1.032	0.505	119626.081	336.743
Jun	All Years	Stacking OF Ensemble	0.937	1.119	0.881	114379.293	323.183
Jul	Wet Years	Stacking Ensemble	0.821	1.159	2.402	120571.644	334.724
Aug	All Years, Variance Threshold .08	Random Forest	0.808	0.907	0.589	12426.334	108.453
Sep	All Years Scaled	Stacking Ensemble	0.668	1.209	0.930	8649.976	86.035
Oct	All Years Scaled	Stacking OF	0.755	1.034	0.873	885.746	28.110

		Ensemble					
Nov	All Years	Stacking Ensemble	0.771	1.019	0.817	2313.822	45.884
Dec	All Years Scaled	Stacking Ensemble	0.701	1.286	0.982	15670.291	119.312

Best Wet-Year Xeric Models for Each Month

Month	**Base DataSet**	**Model Method**	**R^2**	**Mean O/E**	**Std. Dev. O/E**	**MSE**	**RMSE (cfs)**
Jan	All Years, Variance Threshold .08	Stacking OF Ensemble	0.757	1.140	1.009	15818.753	120.102
Feb	All Years, Variance Threshold .08	Stacking OF Ensemble	0.776	0.902	0.797	19421.435	132.405
Mar	All Years Scaled	Stacking OF Ensemble	0.736	0.968	0.680	9960.594	97.347
Apr	All Years Scaled	Stacking OF Ensemble	0.698	0.968	0.817	4703.123	65.993
May	All Years Scaled	Stacking OF Ensemble	0.764	0.995	0.725	449.808	20.367
Jun	All Years Scaled	Stacking OF Ensemble	0.818	1.185	1.639	66.212	8.095
Jul	All Years Scaled	Stacking OF Ensemble	0.845	1.161	2.585	16.638	3.853
Aug	All Years Scaled	Stacking OF Ensemble	0.843	2.330	16.663	6.522	2.497
Sep	Wet Years Scaled	K Nearest Neighbors	0.854	3.279	21.087	3.669	1.881
Oct	All Years	Stacking OF Ensemble	0.555	24.434	298.987	134.464	9.337
Nov	All Years, Variance Threshold .08	Stacking OF Ensemble	0.623	0.792	1.144	801.180	26.431
Dec	Wet Years	Stacking Ensemble	0.779	0.971	1.350	3819.094	58.608