Predicting Unimpaired Flow in Ungauged Basins: "Random Forests" Applied to California Streams

By

Elaheh White
B.S. (University of Kentucky) 2015

Thesis

Submitted in partial satisfaction of the requirements for the degree of

Master of Science

in

Civil and Environmental Engineering

in the

Office of Graduate Studies

of the

University of California

Davis

Approved:

Chair Jay R. Lund

Jonathan D. Herman

Robert J. Hijmans

Committee in Charge

2017

To Mattye White,...
the world would be a much better place if we were all just a little bit more like you.

Contents

List of Figures

List of Tables

Abstract

Predicting Unimpaired Flow in Ungauged Basins:
"Random Forests" Applied to California Streams

Predicting and forecasting streamflow at ungauged sites is a grand challenge for hydrology (Sivapalan et al., 2003). Such predictions are needed for improved streamflow restoration, flood and drought forecasting, and reservoir release decisions. Traditional hydrologic models are mechanistic; they require a set of system characteristics such as, basin geometry, channel slope, and climate conditions, and they use physics-based governing equations for fluid flow to predict runoff. An alternative approach is to use statistical models to predict water flows from climate and basin characteristics. Such models are easy to construct, run fast, and require little expert intervention in calibrating or tweaking parameters, but they have not been widely used in hydrology. This study used Random Forest (RF) models, a regression-tree based statistical learning algorithm, to model monthly unimpaired flows in 69 California basins. The test set error (Coefficient of Determination, R^2=0.69, Nash-Sutcliffe Efficiency, NSE=0.74) from cross-validation reflects the models ability to capture the variations in flow at a monthly resolution. Next, All predictor variables were ranked based on their relative importance (i.e., contribution to reducing the prediction errors). The most important variables were: precipitation, basin drainage area, precipitation lagged one month, month, and basin relief ratio. The RF model was benchmarked against the Basin Characterization Model (BCM), a mechanistic model, and a Linear Multivariate Regression (LMR) model with the same predictor variables as that of the RF model. The RF model out-performs the LMR, but falls short of the BCM. The RF model quality in predicting unimpaired flow was highly spatially variable. Model improvement strategies are discussed.

Keywords: statistical learning models; Random Forest; hydrologic prediction; ungauged basins; watershed management

Acknowledgments

This thesis owes its completion to many people. My heartfelt thanks to the three members of my committee for all their advice and assistance: Jay Lund and Jon Herman for their comments and encouragement making this a much better study than it otherwise would have been, and Robert Hijmans for teaching me R programming and spatial statistics in his wonderfully difficult *Quantitative Geography* course.

Special thanks to Bonnie Magnuson-Skeels for getting me interested in this topic and paving the way with her masters thesis. Also, many thanks to: Marielle Pinheiro for her patience during all the time we spent programming together; my friends in the Water Resources Graduate Group and the Climate Change Water And Society Trainees for their feedback on my presentations; Professor Stephen Wheeler for his invaluable guidance given in the *Research Design* course; University Writing Program Graduate Writing Fellow Gabi Kirk for her expert advice on writing; Clancy McConnel and Yiwei Huang for their comments on the draft thesis; and the Data Science Initiative and its affiliates for their programming workshops.

And, as always, many thanks to Brad White for his endless encouragement and support on whatever I decide to do in my life.

This material is based upon work partly supported by the NSF GRFP under Grant No. 1650042 and the Climate Change, Water, and Society NSF IGERT, to UC Davis DGE No. 1069333. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Chapter 1

Introduction

> We are drowning in data and starving for knowledge
>
> John Naisbitt, *"Megatrends"*, 1982

1.1 Background

Our ability to extract insights from large diverse datasets has rapidly improved with growing computing power and sophisticated algorithms. The term "statistical learning" has emerged to provide a framework that includes simple linear regression and complex algorithmic methods (James et al., 2013). A main contribution of this field is the development of modeling techniques that allow for the semi-automatic creation of complex models, with many interacting predictor variables, which are not overfit, and predict well.

These developments allow for more accurate and flexible empirical models to manage complex systems. For example, in hydrology, runoff formation processes are highly variable, non-linear, and spatially heterogeneous, which creates a challenge for predicting processes such as streamflow (Dooge, 1986).

Hydrologic models can be classified as *mechanistic* (physical process-based) or *empirical* (statistical) (Guisan & Zimmermann, 2000). Each approach sacrifices some generality, realism, cost, and precision for better understanding, predicting, and managing natural resources (Levins, 1966; Klemes, 1982). Hydrologists often develop mechanistic models to capture complex runoff processes. Such models require considerable effort to collect input field data and calibration to obtain basin-specific parameters (Singh & Frevert, 2005). As mechanistic models increase in complexity, it is unclear if hydrologic predictions improve commensurately (Beven, 2011). Without a unifying approach across these various models, and considering the increasing availability of environmental data, there is merit in the more economical predictive power of empirical models.

Initial empirical studies have explored the prediction of streamflow in data-scarce regions (Shortridge et al., 2016), the prediction of streamflow to fill in the gaps in the gauge record (Petty & Dhingra, 2017), the prediction of natural flow in the headwaters of California basins (Carlisle et al., 2010), and the prediction of natural flow of California streams in dry months (Magnuson-Skeels, 2016). These methods can also be applied to hydrologic prediction with climate changed parameters to assess water supply vulnerability; they can provide

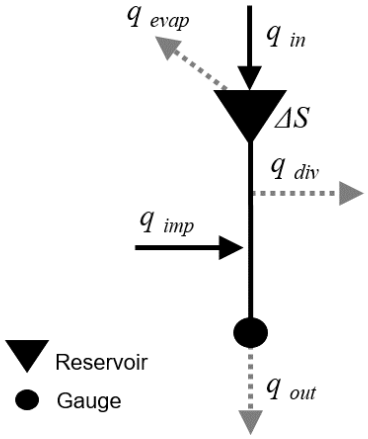


Figure 1.1: Unimpaired flow is calculated by adding back in diversions, subtracting imports, accounting for change in storage and evaporation caused by the reservoir.

opportunities for sensitivity and probabilistic analysis by providing a range of estimates.

1.2 Unimpaired Flow

This study investigates the relationships between the response variable: "unimpaired flow" and the predictor variables: climate and basin characteristics. Before continuing, it is useful to know how unimpaired flow is defined. Unimpaired flow is the flow that is produced by the basin in its current state, but, without dams and diversions (California Department of Water Resources, Bay-Delta Office, 2016). Unimpaired flow calculations are used mostly in California, where dams have created major changes to the natural flow regime. It is often calculated by a simple accounting of water in the system (Figure 1.1 and Equation 1.1).

$$q_{uf} = q_{out} - q_{imp} + q_{div} + \Delta S + q_{evap} \tag{1.1}$$

Where q_{uf} is unimpaired flow, q_{out} is observed guage data, q_{imp} is imported flows, q_{div} is diverted flows, ΔS is the change in storage, and q_{evap} is the evaporation out of the system.

In contrast, "natural flow" is the runoff produced by a basin in its pre-development state prior to any human alterations (Poff et al., 1997). The differences between unimpaired flow and natural flows are usually driven by effects of levees, upland land use, wetlands, and groundwater. This study, however, is only concerned with unimpaired flow; the models are built with unimpaired flow data from the California Data Exchange Center (CDEC), and the predictor variables are taken from various sources discussed in Chapter 3.

1.3 Research Questions

Predicting and forecasting streamflow at ungauged sites is a grand challenge for hydrology (Sivapalan et al., 2003). Such predictions serve as input in many water management models and are needed for improved streamflow restoration, flood and drought forecasting, and reservoir release decisions. In this study, a statistical learning model predicts the unimpaired

flow of California basins on a monthly time step. This report will address the following questions:

1. What is the relative importance of predictor variables that contribute to unimpaired flow predictions?

2. How accurate are the statistical learning model's predictions, and how does the accuracy vary?

3. Can statistical learning models give better streamflow predictions compared to mechanistic models?

1.4 Limitations & Assumptions

Klemes (1982) warns modelers of the general limitations of empirical modeling, the most important being: 1) In search of "better calculus", the modeler may be in danger of overfitting (i.e., regarding part of the noise in the data as information); 2) Empirical models must be regarded as interpolation formulas, and so, they have no justification outside the range of the underlying datasets (Klemes, 1982). Faced with climate change imposing nonstationarity on environmental variables like precipitation and temperature, empirical models for flow should not be used to extrapolate beyond the limits of the variables the model observes or it will run the risk of large errors; 3) Another drawback is the complexity in model structure, especially in ensemble statistical learning methods, which are sometimes referred to as black-box models.

The models in this study were fitted with data on the California Sierra Nevada mountainous basins, as well as some coastal, and southern California basins (Chapter 3.3). These training data sets more or less span the same hydrologic region (United States Geological Survey Region No. 18). As such, the model will not be applicable to basins outside this spatial range.

Another limitation is due to the structural nature of the Random Forest Algorithm. Chapter 3 discusses how regression-based Random Forest models make predictions by averaging predictions made by multiple regression trees. Therefore, the ensemble model limits the predictions it makes to the range seen in the training data; in other words, the predictions do not extrapolate to ranges not seen in the training data. In fact, the averaging dampens the density function when we compare the observed to the predicted data (Figure 4.2).

Furthermore, this method assumes stationarity in the response variable: unimpaired flow. Given the relatively short gauge records in this study, this assumption is not too egregious.

Chapter 2

Literature Review

> Our responsibility is to do what we can, learn what we can, improve the solutions, and pass them on.
>
> Richard P. Feynman, *"What Do You Care What Other People Think?"*, 1988

2.1 Background in Hydrological Modeling

Since the mid-19$^{\text{th}}$ century, with the employment of the "rational method", empirical relationships have been used in rainfall-runoff modeling (Beven, 2011). According to Todini (1988), engineers developed the rational method in response to problems in which the design discharge was of major concern (i.e., urban sewer, land reclamation drainage systems, and reservoir spillway design). This method, based on the concept of concentration time, calculates runoff by simply multiplying a runoff coefficient by rainfall intensity and the basin's drainage area. It proved only applicable to small or mountainous catchments where the rainfall duration does normally exceed the concentration time–the time it takes to reach the maximum discharge of a basin.

To address more complexities in rainfall duration, basin size, and non-uniform characteristics, other methods emerged: in the 1930s, the "unit hydrograph method" materialized; in the 1950s, mathematical techniques such as Z, Laplace or Fourier transforms led to the derivation of the response function from the analysis of input and output data; in the 1960s, grander approaches emerged to model the physical processes of the hydrologic cycle. Models increasing in complexity over time and lacking a one-to-one relationship between model and reality (e.g., unrealistic parameter estimates) have led researcher to other ambitious modeling efforts (Todini, 1988).

In the past two decades, complex statistical learning models, here referred to as "machine learning", have been applied to rainfall-runoff modeling. In juxtaposition with physical or semi-physical models, machine learning models learn from the data itself, with no assumptions as to the underlying process. As Solomatine and Ostfeld (2008) explains, most machine learning techniques that are applied to the rainfall-runoff problem use neural networks (Minns & Hall, 1996; Dawson & Wilby, 1998; Tokar & Johnson, 1999; Hsu et al., 2002; Hu et al., 2007; Abrahart et al., 2007; Govindaraju & Rao, 2013). Other studies use support

vector machines (Asefa et al., 2006; Lin et al., 2006), and tree based algorithms (Iorgulescu & Beven, 2004; Galelli & Castelletti, 2013; Magnuson-Skeels, 2016).

The wide range of models employed may suggest that no one single modeling method is useful across all locations, timescales, and problems. The essential arbitrariness in the selection of the form of an empirical model is one of empirical modeling's drawbacks (Klemes, 1982). Most studies report using one modeling method, which perhaps suggests that researchers are not employing more than one modeling method. Similarly, the application and comparison of different machine learning models to the ungauged basin problem was not considered in this study. Such a study could provide insights into the system by revealing the sensitivity of results to the algorithms employed. Alternatively, this study does benchmark the Random Forest model against a mechanistic, and a linear multivariate regression model.

2.2 Considerations for Empirical Modeling

Cross Validation

Most studies ignore the spatial or temporal structure in the data when devising a cross-validation strategy. When validation data are randomly selected from the entire spatial domain, training and validation data from nearby locations will be dependent (due to spatial autocorrelation). Therefore, if the objective is to project outside the spatial structure of the training data (e.g., to an ungauged basin), error estimates from random cross-validations will be overly optimistic (Roberts et al., 2017). The studies, in which a random test-train split is considered, are most appropriate for predicting flow for a sparsely incomplete gauge record, and the studies, in which holding out blocks of data in time is considered the cross-validation strategy, are most appropriate for predicting streamflow in time for that location. One should not expect to use these cross-validation strategies and get the same predictive accuracy in a purely ungauged basin problem, where blocks are supposed to be designed across geographic space. Adding to the complexity, here, the correlation structure in the gauge data is more complicated than merely proximity of the gauges (i.e., two gauges may be close in proximity but be fed by two different basins and therefore not as correlated as gauges on the same river). A blocking cross validation strategy, like the leave-one-group-out cross-validation (LOGOCV) method explained in Chapter 3, is most appropriate for a study that intends to model the response variable in locations for which no data was observed.

Spatial Autocorrelation

Most studies fail to examine the spatial autocorrelation of the errors produced by the model; as such, inferential results can be biased. Either including additional predictor variables, or choosing a different functional form has to be considered if a strong autocorrelation is detected in a model. The problem of inference cannot be diagnosed without explicitly checking for the spatial variability of the residuals, which are supposed to be independent and not correlated. A simple visual check or a formal test of the significance of Moran's I or Geary's C can help in this regard.

Chapter 3

Methods

> All science is the search for unity in hidden likeness.
>
> Jacob Bronowski, *"The Nature of Scientific Reasoning"*, 1972

3.1 Model Selection

The choice of a suitable model relies on striking the desired balance between three model properties: generality, reality, and precision. Model selection shouldn't solely rely on statistics; some models better reflect physical foundations in hydrology, and conceptual considerations need to include the desired level of trade-off between optimizing accuracy versus optimizing generality (Guisan & Zimmermann, 2000). Unfortunately, no guide to empirical model selection exists in hydrology.

Trees generally do not have the same level of predictive accuracy as some of the other regression and classification approaches (James et al., 2013). However, in recent years, the popularity of tree learning methods (notably, Random Forest) has increased strongly, possibly due to the following advantages:

- Decision trees can handle different variable types (quantitative, ordered, categorical or a mix) (James et al., 2013).
- Decision trees can handle complex relationships between the predictor and response variables, without any a priori assumption; they can have strong nonlinear relationships with high order interactions (James et al., 2013).
- Decision trees intrinsically implement feature selection, making them somewhat robust to irrelevant or noisy variables (Louppe, 2014).
- Decision trees are robust to outliers or errors in the data (Louppe, 2014).
- Due to the *Law of Large Numbers*, which states that as the number of trees in the forest increases, the actual ratio of predictions will converge on the theoretical, or expected, ratio of predictions, Random Forest models are effective in prediction. Injecting the right kind of randomness makes them accurate classifiers and regressors (Breiman, 2001).

- A select few trees in the Random Forest model can be displayed graphically, which make it easier to explain to non-expert users.
- The strength of the individual predictors and their correlations give insight into the ability of the Random Forest model to predict (Breiman, 2001), and "Partial Dependence Plots" can provide further information making Random Forest models a semi-black-box model.

Given these advantages and due to their ease of application and understanding, Random Forest models are developed in this study.

3.2 Model Development

All data processing and model development for this study was done in `R`, a statistical programming language, (R Core Team, 2017), and used the following packages: `sp` (Pebesma & Bivand, 2005; R. S. Bivand et al., 2013), `raster` (Hijmans, 2016), `rgeos` (R. Bivand & Rundel, 2017), `rgdal` (R. Bivand et al., 2017), `dismo` (Hijmans et al., 2017), `geosphere` (Hijmans, 2017), and `randomForest` (Liaw & Wiener, 2002).

Model development followed the steps depicted in Figure 3.1. First, the test and training dataset was split randomly (20/80 split) to get the optimal value of the model parameters and a first cut at the variable importance list. The tuning parameters gleaned from the first step were then applied to the models developed next: the "ungauged basin" problem. Here, a different test-train splitting method was applied: the leave-one-group-out cross-validation (LOGOCV) method, which effectively replicates the ungauged basins problem. In this method, the data for the basin to be modeled was left out of the training data and becomes the test set. The training data is then the data from all the other basins. This process was repeated for all basins in the study. Therefore, for model evaluation purposes, one Random Forest model exists for each basin. With the developed models and the test set, we calculated model measures of fit: the Absolute Error (AE), Relative Error (RE), Root Mean Square Error (RMSE), RMSE standard deviation ratio (RSR), Coefficient of Determination (R^2), Nash-Sutcliffe Efficiency (NSE), and Percent Bias (PBIAS) (Equations 3.1 to 3.7).

$$AE = Y_i^{sim} - Y_i^{obs} \tag{3.1}$$

$$RE = 2 * \frac{Y_i^{sim} - Y_i^{obs}}{Y_i^{sim} + Y_i^{obs}} \tag{3.2}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \left(Y_i^{sim} - Y_i^{obs}\right)^2}{n}} \tag{3.3}$$

$$RSR = \frac{RMSE}{STDEV_{obs}} = \frac{\sqrt{\sum_{i=1}^{n} \left(Y_i^{sim} - Y_i^{obs}\right)^2}}{\sqrt{\sum_{i=1}^{n} \left(Y_i^{obs} - \overline{Y_i^{obs}}\right)^2}} \tag{3.4}$$

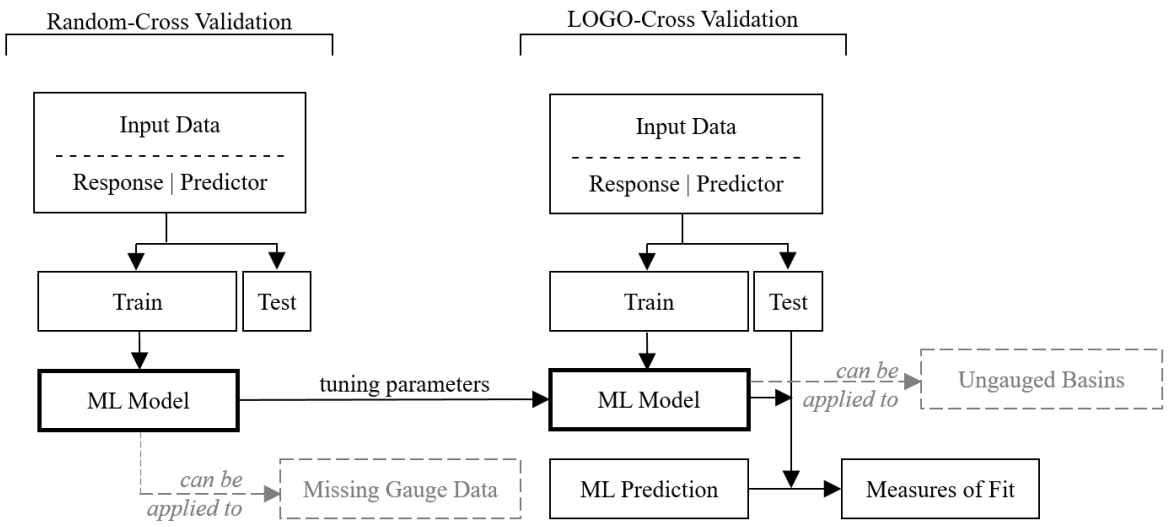


Figure 3.1: Research design.

$$R^2 = \frac{\sum_{i=1}^{n}\left(Y_i^{sim} - \overline{Y_i^{obs}}\right)^2}{\sum_{i=1}^{n}\left(Y_i^{obs} - \overline{Y_i^{obs}}\right)^2} \tag{3.5}$$

$$NSE = 1 - \frac{\sum_{i=1}^{n}\left(Y_i^{sim} - Y_i^{obs}\right)^2}{\sum_{i=1}^{n}\left(Y_i^{obs} - \overline{Y_i^{obs}}\right)^2} \tag{3.6}$$

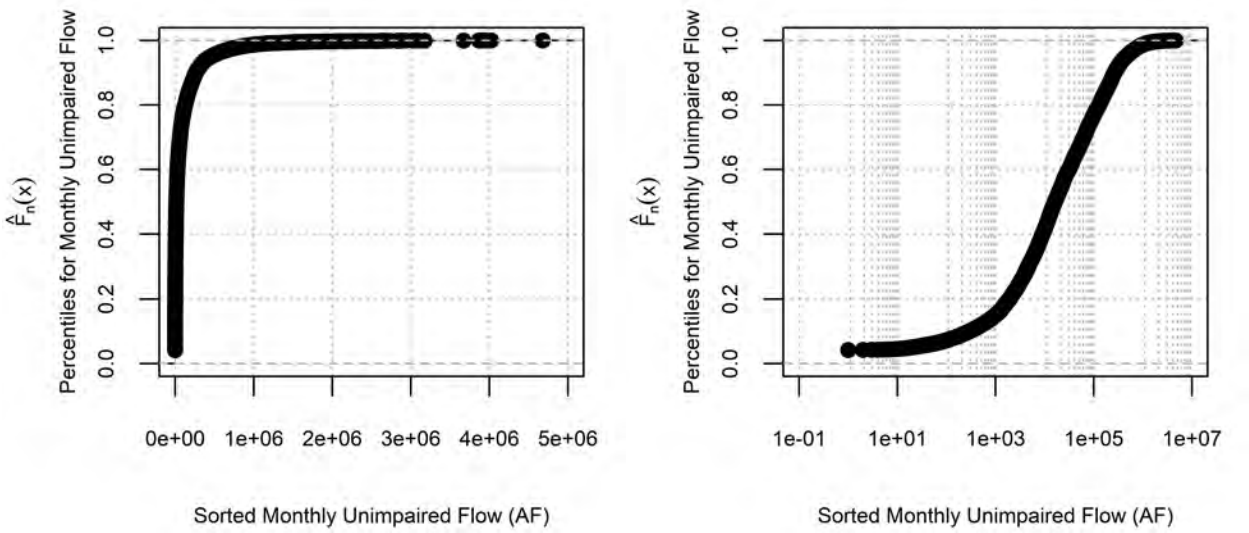
$$PBIAS = \frac{\sum_{i=1}^{n}\left(Y_i^{obs} - Y_i^{sim}\right) * 100}{n\sum_{i=1}^{n} Y_i^{obs}} \tag{3.7}$$

Where Y_i^{obs} are the observed unimpaired flows, and Y_i^{sim} are the predicted or simulated unimpaired flows, and n is the number of observations.

3.3 Study Area & Response Variable

This study used the monthly unimpaired flows dataset developed and maintained by the California Data Exchange Center (CDEC). The data spans 69 California basins (Figure 3.2) from 1982 to 2014. It can be downloaded with the `sharpshootR` package in R (Beaudette, 2016). It has approximately 18,500 monthly streamflow observations in acre-feet (AF) and as a continuous variable can be used for regression type studies (Figure 3.3, and Appendix A).

Figure 3.2: The 69 California basins under study are the CDEC unimpaired flow basins.



(a) The cumulative distribution function.

(b) The cumulative distribution function in log space.

Figure 3.3: Distributions of the response variable: approximately 18,500 unimpaired flows in acre-feet (AF).

3.4 Predictor Variables

Predictor attributes were calculated for each observation point (Table 3.1). A total of 28 predictor variables were selected based on the knowledge of basin characteristics and processes that influence a watershed's response to precipitation: evaporation (temperature); snowfall (cumulative sum of precipitation below 2°C); storage in soil (with soil and land cover parameters); antecedent conditions (with lagged precipitation and temperature parameters); and groundwater processes (with geology and depth to restricted layer).

The climate data were derived from the Parameter elevation Regression on Independent Slopes Model (PRISM) dataset, which contains gridded rasters for the continental United States at $4km$ resolution from 1891 to 2014. The **temperature** variable and its lagged forms are the basin averaged PRISM *tmean* variable, which in turn was calculated by the mean of the monthly minimum temperatures and the monthly maximum temperatures. The **precipitation** variable and its lagged forms are the basin averaged PRISM *ppt* variable, which is a measure of total precipitation (rain and snow).

Low flows in some Sierra Nevada basins exhibit a "memory" effect in which they depend on the current and previous year's snowpack (Godsey et al., 2014). Since we did not want to include 24 lagged precipitation parameters in the Random Forest model, we developed a snow variable. The **snow** variable was the cumulative sum of precipitation, starting in October of each water year, for temperatures under 2°C.

Basin shape can affect the peak discharge; peak discharge for a circular basin arrives sooner than for an elongated basin of the same area. Because of how the tributary network in a circular basin is organized, the flows in a circular basin enter the main stem at roughly the same time, so more runoff is delivered to the outlet together, sooner. In an elongated basin, because of the mismatch in timing, peak runoff is more attenuated, except for some slow

moving streams. The **shape** parameter, calculated by basin length divided by basin width, and the **compactness** parameter, calculated by basin area divided by (basin perimeter)^2, account for this phenomenon. Although, this phenomenon is more pronounced in runoff on a smaller time step, we included these parameters in the final model to see their rankings in the variable importance list.

Basin hypsometric information was derived from the Shuttle Radar Topography Mission (SRTM) $90m$ model, which is a gridded raster of static elevation at a $3arc-second$ resolution. The vertical error of the model is reported to be less than $16m$. The **mean basin elevation** and **basin relief ratio** parameters (Pike & Wilson, 1971) were calculated from this dataset. Basin relief ratio is calculated by the difference in maximum and minimum elevations divided by basin length.

Soil properties were derived from the POLARIS dataset, a Soil Survey Geographic Database (SSURGO) processed dataset at a $3arc - second$ resolution. Percent **clay**, **silt**, and **sand**, **saturated hydraulic conductivity**, **lambda** and **n** pore size, **available water content**, and **depth to restricted layer** information was averaged for each basin.

The land cover property was derived from the California Vegetation (CALVEG) dataset, which includes the following land cover types: urban (URB), barren (BAR), shrub (SHB), conifers (CON), hardwoods (HDW), water (WAT), mix (MIX) and agriculture (AGR). The **percent vegetated** parameter, is the percent of land in a basin that is not covered by URB, BAR, and WAT. The **dominant basin geology** parameter taken from the Natural Resources Conservation Service (NRCS) dataset *rocktype2* variable. Here, the percent of basin area in each rock type category was calculated and the dominant class is preserved.

Table 3.1: Summary of the variables used in the implementation of the Random Forest model.

Type	Variable	Description	Source
Response	Unimpaired Flow	monthly estimated unimpaired flows, in AF	CDEC (Beaudette, 2016)
Time	Month	categorical: Jan, Feb, ..., Dec	-
	Ordinal Month	numerical distance till June: Jan:6, Feb:5, ..., Dec:6	
	Season	categorical: Fall, Winter, Spring, Summer	
	Year	numeric	
Climate	Temperature, Lag 1, 2 and 3 Months	temperature and lagged monthly temperature, in °C	PRISM (Edmund, 2015)
	Precipitation, Lag 1, 2 and 3 Months	precipitation and lagged monthly precipitation, in mm	
	Snow	cumulative precipitation of the same water year for temperatures bellow 2 °C, in mm	
Hypsometric	Relief Ratio	(max(elev) - min(elev))/ basin length in, m/m	SRTM90 (Jarvis et al., 2008)
	Mean Elevation	mean basin elevation, in m	

Type	Variable	Description	Source
Basin Boundaries	Area	basin drainage area, in m^2	NHD2PLUS (McKay et al., 2012)
	Shape	basin length/basin width, in m/m	
	Compactness	basin area/(basin perimeter)^2, in m^2/m^2	
Soil	% Clay	percent clay in surface layer, in %	POLARIS (Chaney et al., 2016)
	% Silt	percent silt in surface layer, in %	
	% Sand	percent sand in surface layer, in %	
	Sat. Hydraulic Conductivity	hydrologic conductivity of surface layer, in cm/hr	
	Lambda	pore size distribution index (brooks-corey)	
	N	measure of the pore size distribution (van genuchten)	
	Available water content	, in m^3/m^3	
Land Cover	Vegetated	Percent of area in the basin vegetated in %	CALVEG (Forest Service, USDA, Pacific Southwest Region, 2006)
Ground Water	Dominant Geology	dominant rock type in basin, categorical	NRCS (NRCS, USDA, 2006)
	Depth to Restricted Layer	in cm	POLARIS (Chaney et al., 2016)

Before developing the machine learning model we inspected the cross correlation of the predictor variables (Figure 3.4) and their correlation with monthly average flow (Figure 3.5). As expected the drainage area of a basin and precipitation are more positively correlated with monthly unimpaired flow. These variables should become important in the Random Forest model. Also, the partial correlations of predictor variables with flow showed that most of the information content lies within drainage area, precipitation, and some measures of infiltration (i.e., lambda pore size, n pore size, and saturated hydraulic conductivity). The correlated variables were not removed from the Random Forest model, because, in such models the dimensionality of a problem is not a concern, and Random Forests intrinsically implement feature selection, making them somewhat robust to multi-collinearity in the variables. As explained in Chapter 3.5, the `mtry` input parameter is designed to have a diversifying effect on the estimates of multiple trees, and it effectively decorrelates the trees.

3.5 Tree Building Algorithms

Regression and classification trees are used for predicting continuous and categorical data, respectively. *Classification and Regression Trees* (CARTs) involve stratifying or segmenting the predictor space, into a number of regions, using a series of if-then statements (Figure 3.6). At each internal node in the tree, a test is made to one of the inputs. Depending

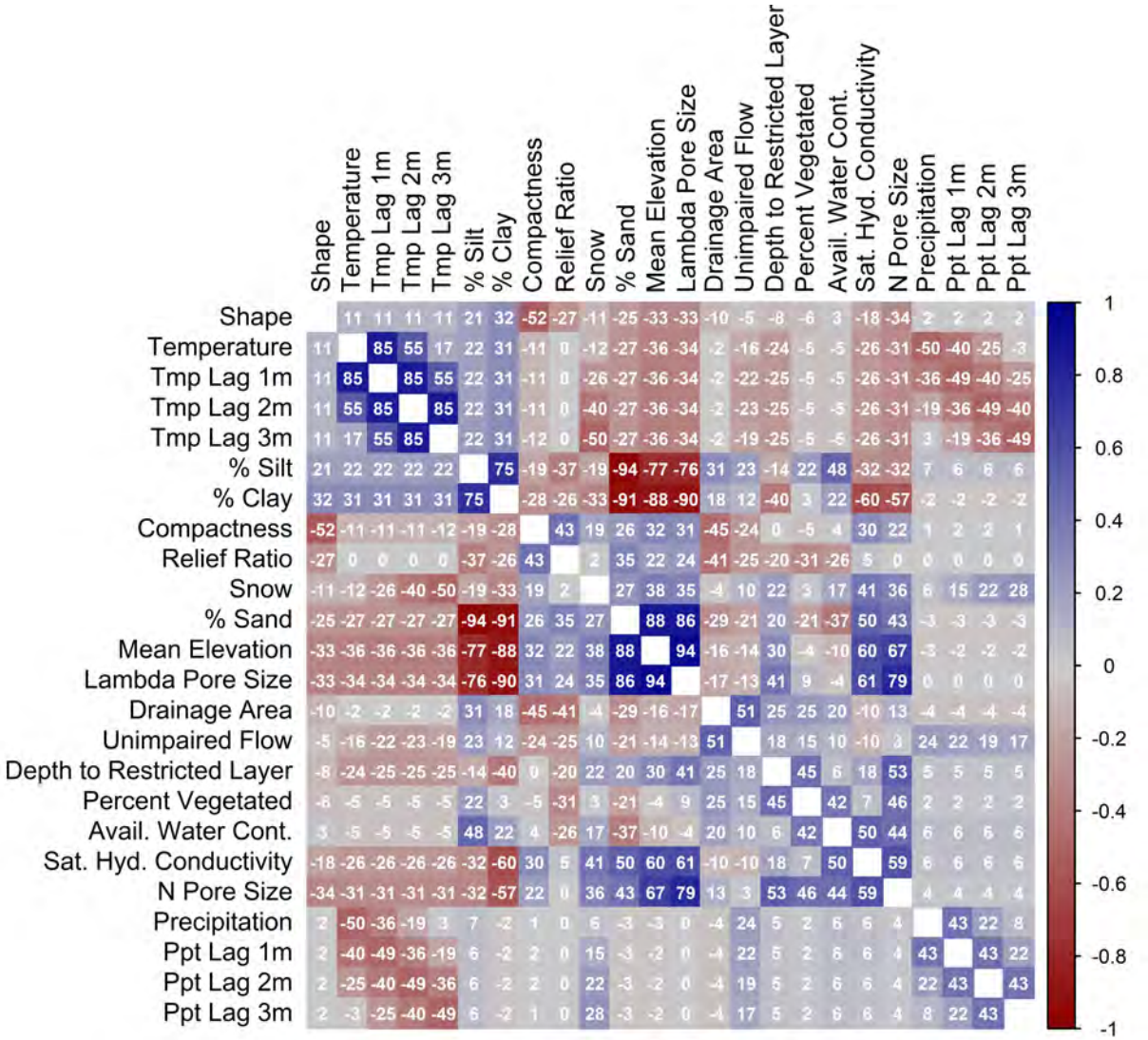
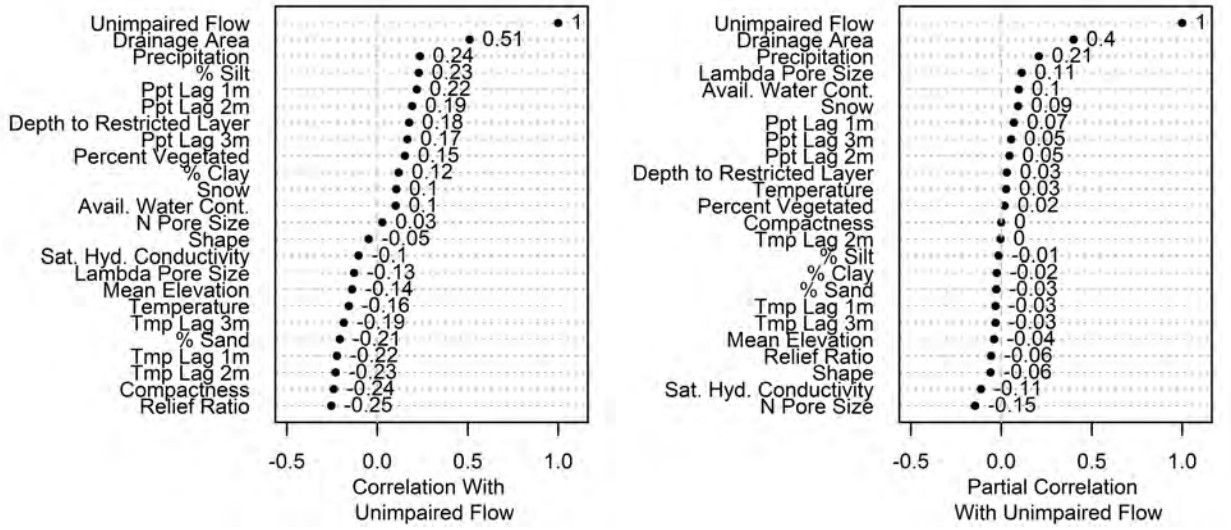


Figure 3.4: Correlation matrix of predictor variables showing groupings in the dataset.

on the outcome of the test (or split rule), the algorithm goes to either the left or the right sub-branch of the tree. Eventually the algorithm arrives at a terminal node, which contains a prediction. The prediction for a given observation is the mean or the mode of the training observations in the region to which it belongs (Breiman et al., 1984).

In essence, tree building algorithms are a series of split rules. The split rule is found using a greedy top-down search for recursively splitting of the data into binary partitions. It is greedy, because, the split rule at each internal node is selected to maximize the homogeneity of its child nodes, without consideration of nodes further down the tree, yielding only locally optimal trees (Grubinger et al., 2011). For regression trees, the mean of all the observation points that fall within a branch is considered the prediction of that branch in the tree. The best tree is one which has the minimum test error rate calculated by the Residual Sum of Squares (RSS).

Since trees have a finite number of terminal nodes (CARTs are pruned based on a complexity parameter, α), the prediction of these methods are discrete, and therefore, not par-



(a) Pearson's Correlation (b) Partial Correlations

Figure 3.5: Correlation of predictor variables with monthly flow volumes. Drainage area and precipitation correlate the most with flow.

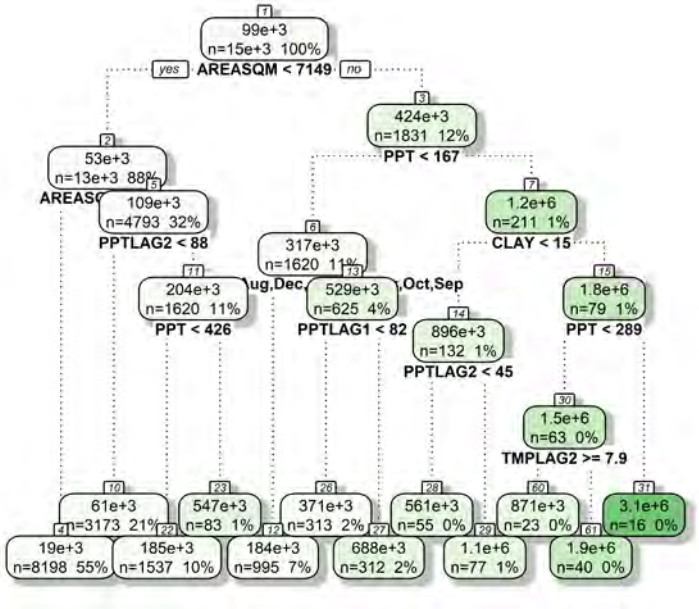


Figure 3.6: An example CART. In Random Forests multiple CARTs are built with randomization applied to the training data and the predictor variables to split on.

ticularly suited to modeling a continuous variable. However, CARTs main problem is the model's high variance; trees grown on different subsets of the training set will produce different predictions. This phenomenon is one of the major drawbacks of CARTs. Methods such as *bagging* (Breiman, 1996), *random forests* (Breiman, 2001), *boosting* (Friedman, 2001) and *bumping* (Grubinger et al., 2010) attempt to improve the prediction accuracy of trees with the idea that combining and averaging trees reduces variance.

A Random Forest consists of an assemblage of unpruned CART models. Each CART model is different because it is grown on: 1) a new training set: in each bootstrap training set, about one-third of the instances are left out; and 2) using random feature selection: each time a split in a tree is considered, a random sample of predictors is chosen as split candidates from the full set of predictors. This process de-correlates the trees. Suppose there is one very strong predictor in the dataset, along with other moderately strong predictors. Then, in the collection of trees, most or all trees will use this strong predictor in the top split. Consequently, all trees will look quite similar. So, the predictions from the trees will be highly correlated. However, by forcing each split to consider only a subset of the predictors makes the resulting trees less variable and more reliable (James et al., 2013). This strategy, using a random selection of features to split each node, introduces some randomness that improves the accuracy of the predictions of the trees as a whole and yields error rates that are robust with respect to noise (Breiman, 2001).

Estimates of Input Parameters for the Random Forest Model

The `randomForest` library, written by Liaw and Wiener (2002), constructs Random Forest models using the `randomForest` function. This function takes in tuning parameters such as mtry, ntree, sampsize, and maxnodes:

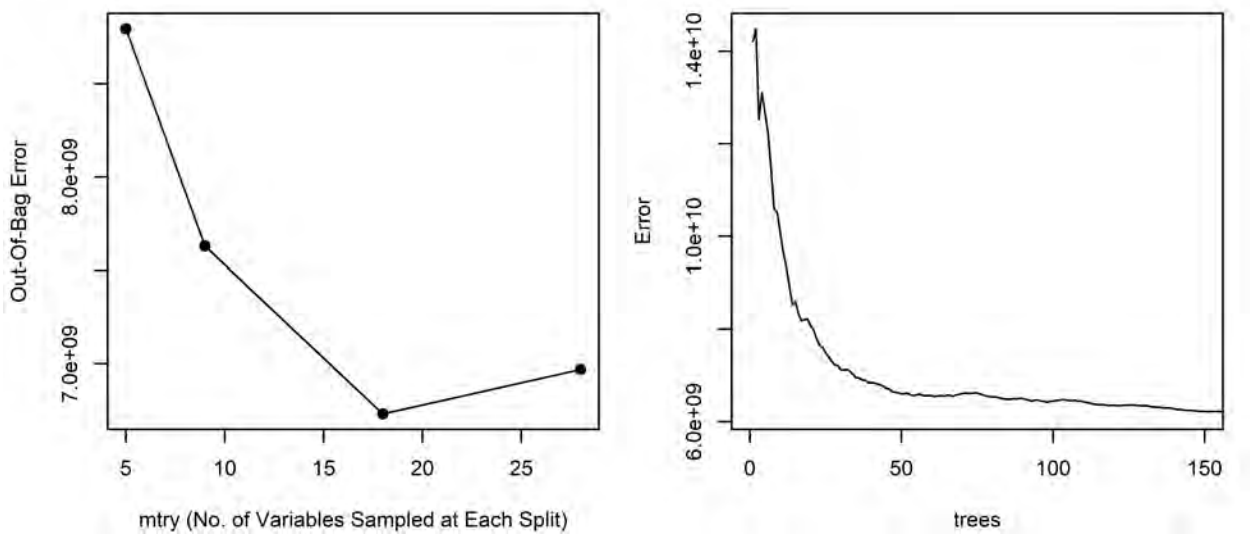
`mtry`: In Random Forest models, internal estimates monitor error, strength and correlation, which are used to show the response to increasing the number of features used in the splitting (Figure 3.7a). Here, this parameter was set to 18 out of the full 28 predictor variables available.

`ntree`: The generalization error of a forest of trees depends on the strength of the individual trees in the forest and the correlation between them (Breiman, 2001). This error converges to a limit as the number of trees in the forest increases (Figure 3.7b). Here, the number of trees was set to 250.

`sampsize`: In Random Forest models, the trees are built on a bootstrap sample of the training data, a sample equal in size to the original dataset, but selected with replacement. Therefore, some rows are not selected, and others are selected more than once. Here, the sample size is set to the default value: length of training set.

`maxnodes`: Using the maximum number of terminal nodes, the user can "prune" the trees back to a smaller version of itself. Here, we used the default value, which is a function of `nodesize` or the allowed minimum number of observations in each node. The default value for nodesize is 5.

These parameters can be fixed by the user or optimized. The benefit of optimizing the parameters become evident when overfitting is concerned (Breiman, 2001). Figure 3.8 shows that, with the exception of mtry, the optimal parameters are the default parameters for this study.



(a) Based on the out-of-bag-error estimate, the optimal value of mtry is 18.

(b) The optimal value of ntree is approximately 150 or more.

Figure 3.7: Tuning parameters to consider before building the Random Forest model: *mtry* and *ntree*. mtry is the number of random subset of parameters to use when building the split rule at each node, and ntree is the number of trees to include in an Random Forest model, in order to achieve stability in the predictions of unimpaired flow.

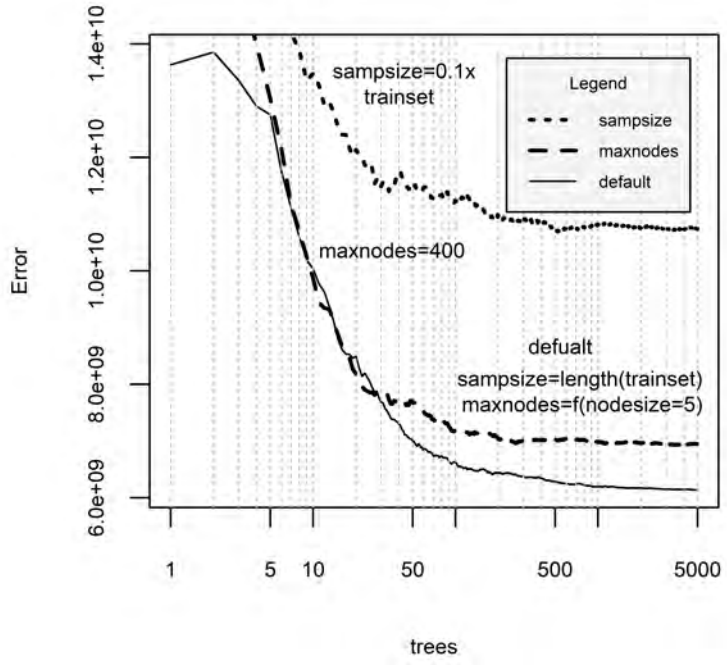


Figure 3.8: Optimal tuning parameters in the Random Forest model predicting unimpaired flow. Reducing *sampsize* and increasing *maxnodes* increases the error of the model. The optimal parameters are default parameters.

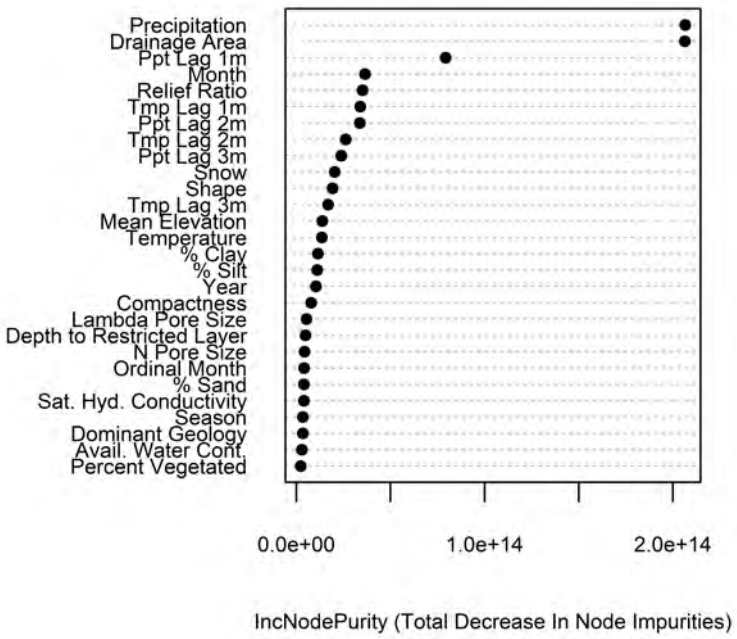


Figure 3.9: Variable importance of the Random Forest model predicting unimpaired flow. The variables that most contributed to the reduction in prediction error are precipitation, basin drainage area, precipitation lagged one month, month, and basin relief ratio.

Variable Importance

In building Random Forest models, internal estimates are used to measure variable importance (Figure 3.9). These estimates answer the first research question: "What is the relative importance of predictor variables that contribute to unimpaired flow predictions?". The variable importance list shows how much each variable that explains unimpaired flow (e.g., basin size, shape, topography, and soil properties) improves the model's predictive capabilities (i.e., node purity or goodness-of-fit). Moreover, the variable importance list can serve as a guide for parameter selection and accuracy for modeling runoff processes in other studies.

Models used to benchmark the Random Forest model are explained in the following sections.

3.6 Linear Multivariate Regression

A regression with more than one explanatory variable is called a multiple regression. In contrast to simple linear regression where the mean is modeled as a straight line, it is now modeled as a function of several predictor variables (Equation 3.8). These predictor variables can be continuous (e.g., precipitation), discrete (e.g., ordinal month) or categorical (e.g., dominant geology).

$$Y_i^{obs} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_i X_{ki} + \epsilon_i \qquad i = 1, \cdots, k \tag{3.8}$$

Where β_i, the partial regression coefficient, is the change in mean for Y_i^{obs} when variable X_i increases by 1 unit, while holding the $k-1$ remaining independent variables constant. This

is also referred to as the slope of Y_i^{obs} with variable X_i holding the other predictors constant.

Given the model, the fitted values can be estimated by Equation 3.9.

$$Y_i^{sim} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_i X_{ki} \tag{3.9}$$

The unknown parameters in Equation 3.8 and 3.9 are: β_0, the overall mean, and β_k, the regression coefficients. To find the best fit, much like simple linear regression, we need to estimate the unknown parameters by minimizing the residual sum of squares (RSS) (Equations 3.10 and 3.11).

$$\epsilon_i = Y_i^{obs} - Y_i^{sim} \tag{3.10}$$

$$\begin{aligned} RSS &= \sum_{i=1}^{n} e_i^2 \\ &= \sum_{i=1}^{n} (Y_i^{obs} - Y_i^{sim})^2 \\ &= \sum_{i=1}^{n} (Y_i^{obs} - \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_i X_{ki}) \end{aligned} \tag{3.11}$$

The `lm()` function in R constructs linear multivariate regression models. This model was built with the same predictor variables explained in Chapter 3.4 and the same method as explained in Chapter 3.2.

3.7 Basin Characterization Model

The California Basin Characterization Model (Flint et al., 2013) applies a monthly regional water-balance model to simulate hydrologic responses of some California basins including those that drain into the state. The model was developed at a $270m$ spatial resolution, using monthly data. It mechanistically models the pathways of precipitation into evapotranspiration, infiltration into soils, runoff, or percolation below the root zone to recharge groundwater. The *evapotranspiration* component is derived through the use of potential evapotranspiration equations that rely on the calculation of solar radiation using slope, aspect, topographic shading, and atmospheric parameters. The *soil storage* component of the model uses soil properties to calculate how much soil moisture is available for plant evapotranspiration. Soil storage is independent from the other hydrologic dynamics, except that *groundwater recharge*, calculated as infiltration below the zone of evapotranspiration, is calculated only from surplus, after soil moisture capacity has been filled. Groundwater recharge is also tied to runoff, and the relationship between the two is driven by the level of permeability of bedrock. Model outputs are calculated for each grid cell, allowing results to be summarized for a variety of planning units including hillslopes, watersheds, ecoregions, or political boundaries (Flint et al., 2013).

Parameters in the model have been calibrated using a total of 159 relatively unimpaired watersheds for the California region. As a result of calibration, predicted basin discharge

closely matches measured data for validation watersheds. The model's recharge and runoff estimates, combined with estimates of snowpack and timing of snowmelt, provide a basis for assessing variations in water availability. This great modeling effort has been supported by numerous federal, state, and local agencies, and international organizations.

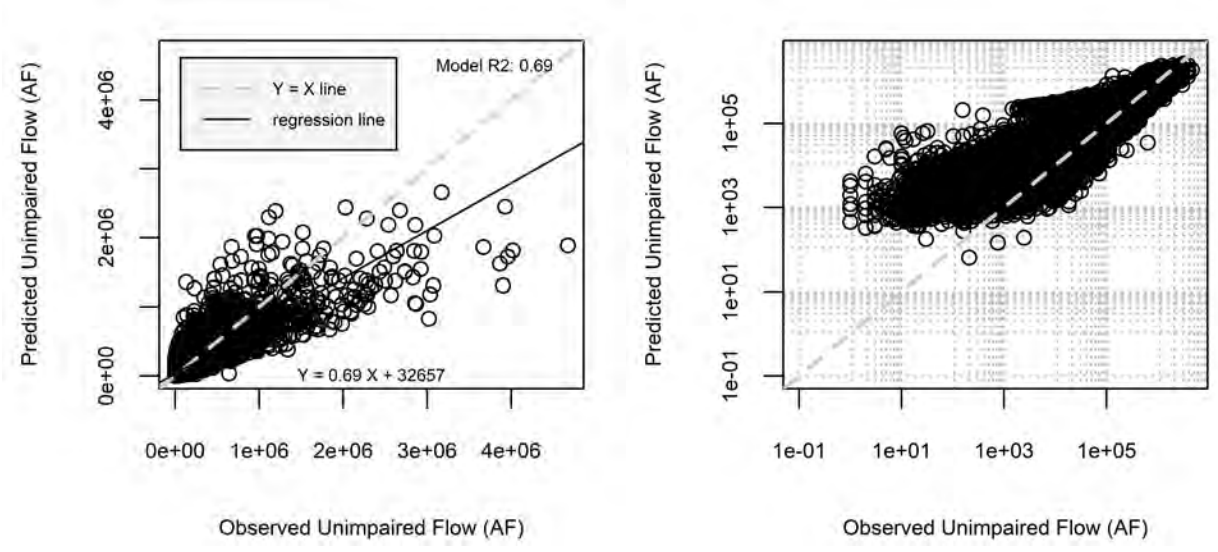
Chapter 4

Results

> The Universe is under no obligation to make sense to you.
>
> Neil deGrasse Tyson, *"Astrophysics for People in a Hurry"*, 2017

The test data is run through the trees in the Random Forest model. Figure 4.1 shows the predicted versus observed values in normal and log space for the combined test sets. Appendix B shows this plot disaggregated by basin. These plots show that the model is over-predicting low flows and under-predicting high flows.

Also, when comparing the probability density's of the observed and predicted unimpaired flows, we see the averaging of predictions made by trees in the Random Forest model dampening the probability density of unimpaired flow (Figure 4.2).



(a) Predicted vs. observed.

(b) Predicted vs. observed in log space.

Figure 4.1: Predicted vs. observed on the test set for all basins combined. Model results show the over-prediction of low flows and the under-prediction of high flows.

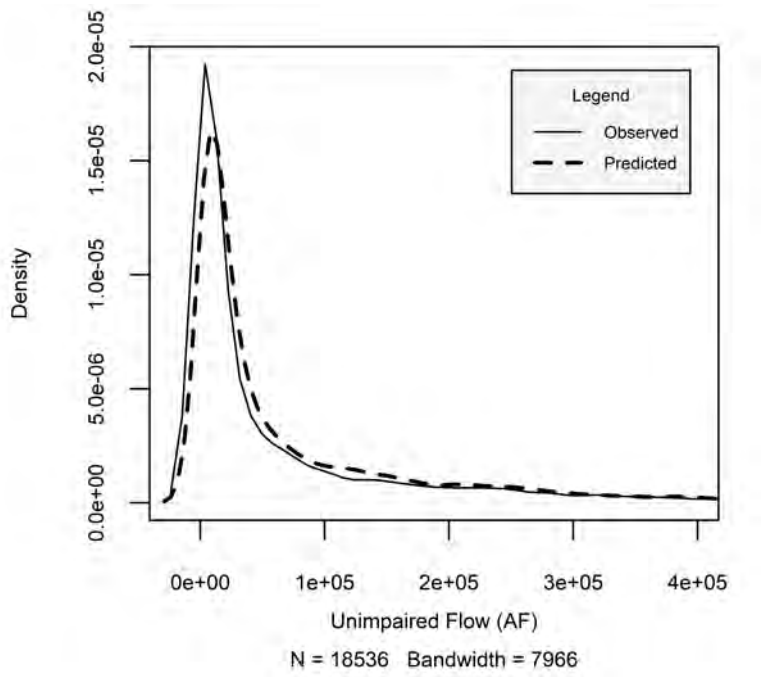


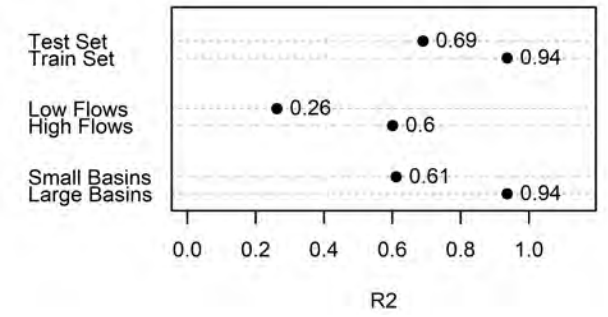
Figure 4.2: The probability density function shows a dampening effect produced by averaging predictions to arrive at an ensemble prediction.

4.1 Model Evaluation

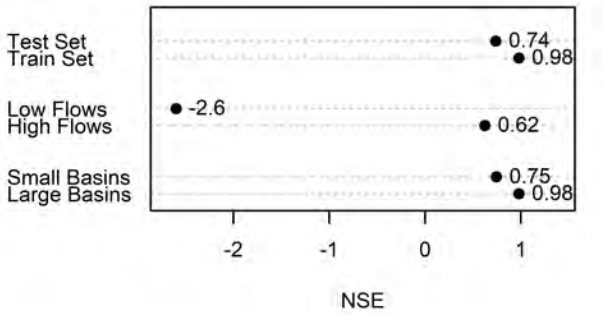
Here, we answer the second research question: "How accurate are the statistical learning model's predictions, and how does the accuracy vary?". To get a numerical estimate of the model's performance, any statistical measure of model fit can be applied. In the combined test set, the Random Forest model performed as follows:

- Mean Absolute Error (AE): 1,700 AF
- Mean Relative Error (RE): 0.52
- Root Mean Square Error (RMSE): 123,000 $AF/month$
- RMSE standard deviation ratio (RSR): 0.51
- Coefficient of Determination (R^2): 0.69
- Nash-Sutcliffe Efficiency (NSE): 0.74
- Mean Percent Bias (PBIAS): 0.05 %.

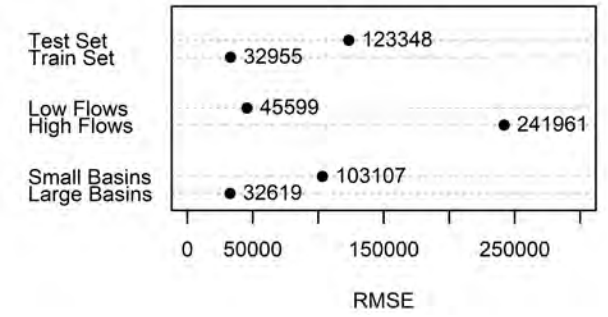
According to Moriasi et al. (2007), the performances in RSR and NSE constitute a "good" model evaluation. Appendix C gives a detailed breakdown of these model performance metrics by basin.



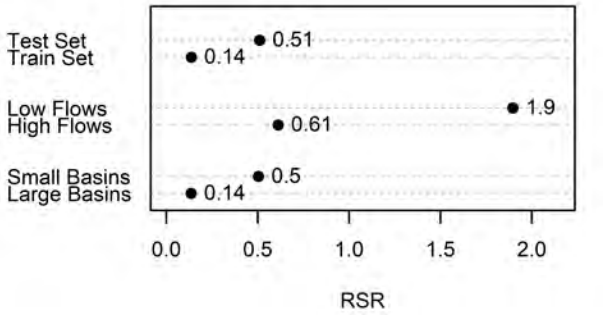
(a) Coefficient of Determination (R^2)



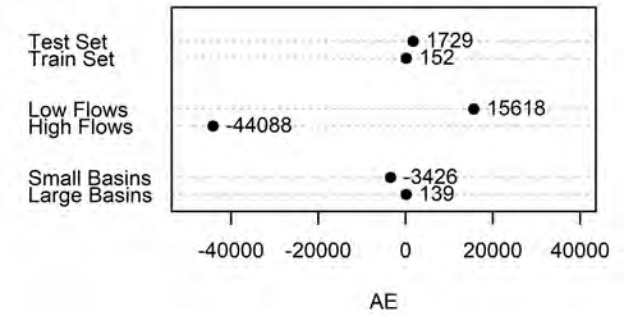
(b) Nash-Sutcliffe Efficiency (NSE)



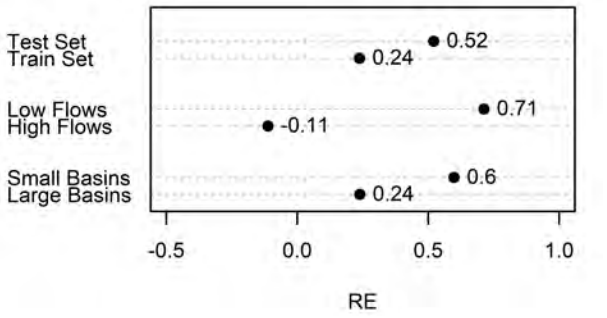
(c) Root Mean Square Error (RMSE)



(d) RMSE standard deviation ratio (RSR)



(e) Absolute Error (AE)

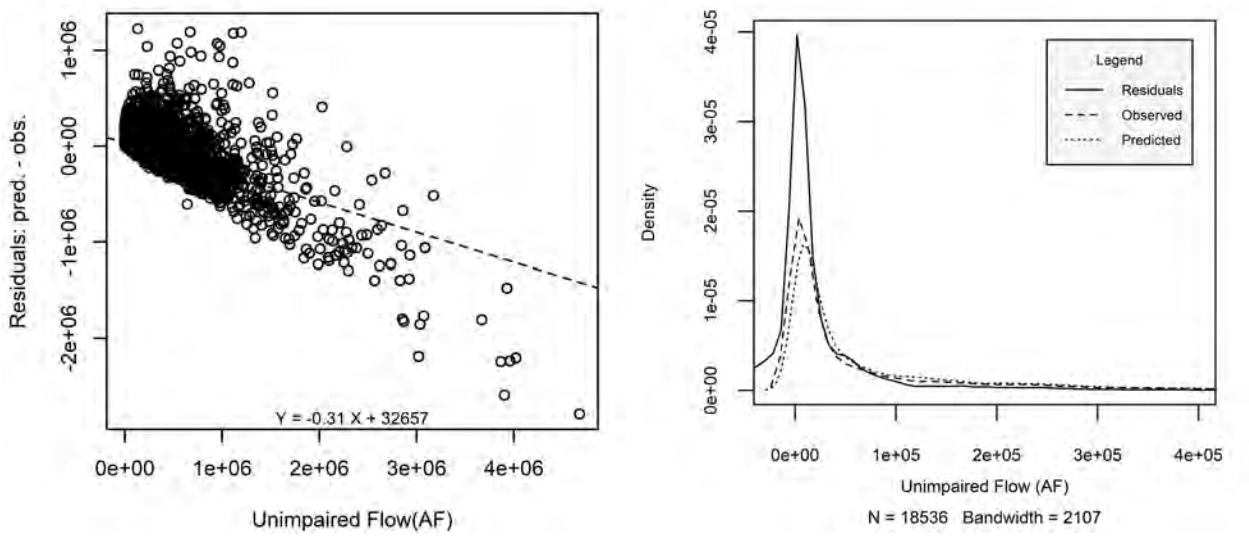


(f) Relative Error (RE)

Figure 4.3: The Random Forest model developed here is most suited to modeling higher flows, which typically occur in larger basins.

4.2 Model Error for Various Categories

Calculating the measures of model fit for various categories shown in Figure 4.3, can help diagnose problem areas in the model. For our purposes, basins with drainage areas smaller than the average, approximately 2,700m^2 or 29,000ft^2, are considered small basins, and the flows less than the average flows, approximately 100,000$AF/month$, are considered low flows. In this model, the higher R^2 values for the training set indicates the model is overfitting (Figures 4.3a, 4.3c, and 4.3e). Also, the model performs worse on low flows and smaller basins (Figures 4.3b, 4.3d, and 4.3f). This phenomenon could be due to the RSS performance measure of the `randomForest` algorithm, which pushes the model to accurately model high flows at the expense of lower flows (Figure 4.4).



(a) Residuals vs. unimpaired flow.

(b) The probability density function of the residuals.

Figure 4.4: Lower absolute error at higher flows show the Random Forest model accurately predicting high flows at the expense of low flows.

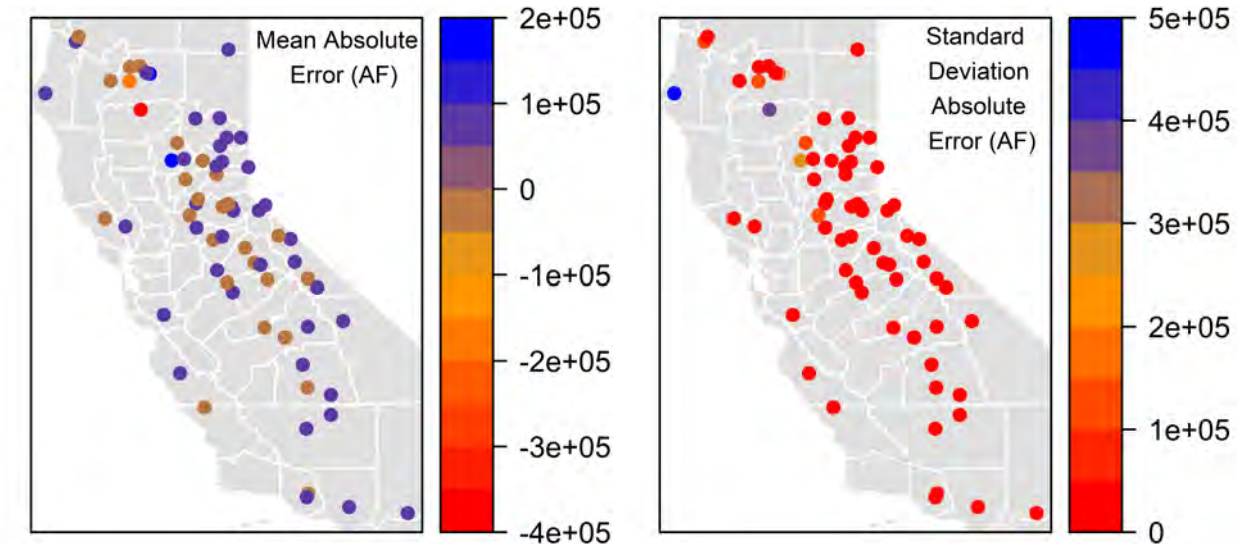
4.3 Spatial Distribution of Model Errors

Figures 4.5 and 4.6 show the spatial distribution of the errors and the models spatially heterogeneous ability for accurately predicting unimpaired flow. These plots show that predicting flow in the headwater basins need to be improved. Figure 4.6b shows a ridge down the middle of California where on the western side the model is accurately predicting flows (same area we typically observe high flows and large basins) and on the eastern side it is performing worse. Figure 4.5b shows the standard deviation of the absolute error, which is in essence a wetness parameter, with a higher standard deviation expected in northern basins.

4.4 Benchmarking

Next, we compared the test set R^2 of the Random Forest model, reflecting the model's ability to capture the variation of flow, with that of the Basin Characterization Model, for basins that overlap the two studies. We also compared the test set R^2 of the Random Forest model with that of a linear multivariate regression model explained in Chapter 3.6. For detailed results see Appendix E.

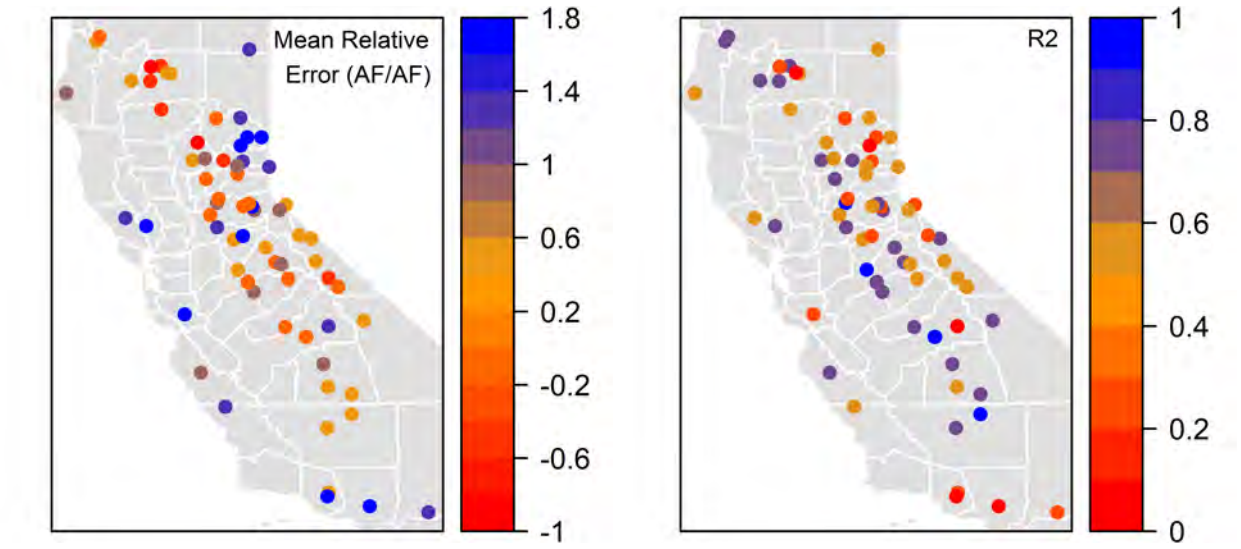
The results of the comparison answer the third research question: "Can statistical learning models give better streamflow predictions compared to mechanistic models?". The Random Forest model out-performs the mechanistic model in two basins (the Stanislaus River at Melones Reservoir and the Feather River near Oroville), and it out-performs the simple linear multivariate regression model in all basins (Figure 4.7). The latter result is expected since we know runoff processes are not linear and are too complex to be modeled using a linear multivariate regression model. However, we expected the Random Forest model to, at the very least, be on par with the mechanistic model, which in most basins it was not.



(a) Mean absolute error (AE) of the test set.

(b) Standard deviation of the absolute error of the test set.

Figure 4.5: The spatial distribution of the absolute error. The absolute error is spatially autocorrelated. Model improvement strategies should consider adding data to the model or switching to another modeling method.



(a) Mean relative error (RE) of the test set.

(b) Coefficient of determination (R^2).

Figure 4.6: The spatial distribution of the relative error (RE) and the coefficient of determination (R^2) statistics show that model modification strategies should consider improving predictions in the headwater basins.

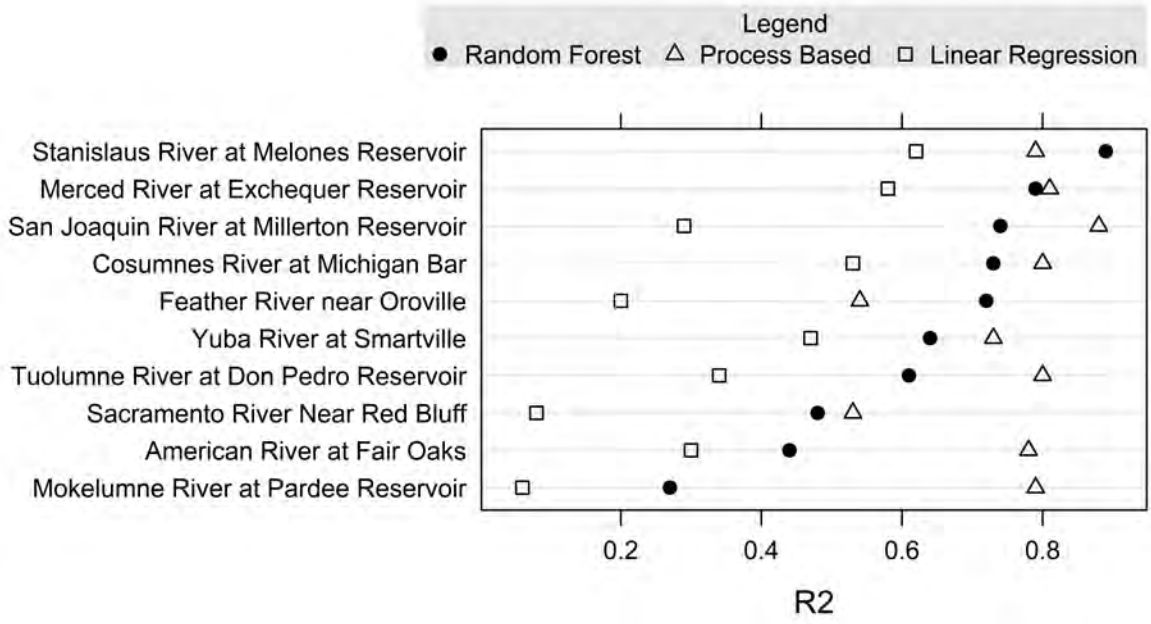


Figure 4.7: Benchmarking. R^2 comparisons on the test set of the three models show the variability in the performance of machine learning models compared to more complex mechanistic and simpler linear regression models.

Chapter 5

Conclusions

> The real purpose of the scientific method is to make sure nature hasn't misled you into thinking you know something you actually don't know.
>
> Robert M. Pirsig, *"Zen and the Art of Motorcycle Maintenance"*, 1974

5.1 Next Steps

This statistical learning application shows a small potential to "learn" from existing environmental data and produce a model that is easy to build and convenient to run. In this case, the statistical learning approach compares favorably to the mechanistic model only in terms of cost of construction, and there is much left to be desired in predictive accuracy. Some strategies, to improve upon this study, are:

1) **Reducing the dimensionality** of the dataset. Adding noise features that are not truly associated with the response will worsen the fitted model, and consequently increase test set error. This is because noise features increase the dimensionality of the problem, exacerbating the risk of having trees in the forest that do not make splits on the variables that better describe flow. Referring back to the variable importance list (Figure 3.9), the prime candidates for eliminating or modifying are those that fall low on the variable importance list. The reason for which may be that the information content of these variables are housed in other variables higher on the list.

2) On the other hand, we can also improve the model by **including more data** that provides information about the hydrologic cycle not quite captured by this dataset. Among possible datasets are: the Moderate Resolution Imaging Spectroradiometer (MODIS) evapotranspiration dataset, the California Irrigation Management Information System (CIMIS) reference evapotranspiration dataset, and the Soil Conservation Survey (SCS) method of reducing hydrologic soil types into one "curve number".

3) Another possible improvement strategy is **adding in a dataset** that covers the problematic areas of the model, the headwater basins. The United States Geological Survey GAUGESII dataset includes reference basins that are generally small, have low flows and are located in the headwaters. The addition of this dataset, although costly in implementation effort, might improve results.

4) Lastly, due to the overfitting problem observed, strategies can be devised for **limiting the depth of the trees** in the forest. Although the `randomForest` function does not have a tree-depth parameter, proxy parameters exist that can somewhat replicate what this parameter is intended to do: i) `nodesize`: increasing the node size increases the minimum number of observations that must fall within a branch. This parameter stops the further splitting of nodes into smaller ones and effectively reduces the depth of the tree. ii) `maxnodes`: decreasing the maximum number of terminal nodes will decrease the number of branches in the tree and effectively builds a smaller or simpler tree. Other Random Forest algorithms that have a tree-depth parameter can remedy this problem as well: max.depth in `xgboost`, nodedepth in `randomForestSRC`, and nLevel in `Rborist`

5) Because the Random Forest algorithm uses the RSS error in evaluating trees, the model is more sensitive to error in high flows. This problem may be remedied by weighted sampling strategies, log-transforming the data, or making multiple overlapping models for different classes.

6) Another option is to **turn to other statistical learning algorithms**: neural networks; support vector machines; and ridge regression.

The importance of understanding the processes that govern runoff formation and the general behavior of statistical learning algorithms becomes evident in model improvement.

5.2 Empirical Modeling for Ungauged Basins

The major challenges in this study were mostly data related: 1) gathering, in formats that are easy to process with languages like python and R; 2) cleaning; and 3) processing. Discipline-specific repositories, curated by university libraries, could help in this regard and prevent "data rot" usually seen in federal and state government agencies. These repositories can assist in discovering, accessing, and acquiring different types of data. They can help researchers understand, develop, and apply strategies for organizing and managing their study's data, and they can help in locating standards for documentation so that the study's output data can be discovered, understood, and reused. Most importantly repositories curated by libraries can aid in the preservation of the input, intermediary, and output data and its scholarly value over time.

On the technical side of empirical modeling, studies in water resources management should carefully examine the structures within the data and their modeling purposes. Researchers should take care when devising test-train splitting strategies for structured data (i.e., data that has a spatial, temporal or hierarchical structure). To remedy this problem, blocking strategies, already employed in ecology, should be applied to empirical modeling with water data.

Also, statistical learning studies should try to employ more than one algorithm. Thus, researchers in the field of water resource management can develop a heuristic, which informs them as to the modeling methods most appropriate with specific data sets. The various algorithms can be tested at different spatial and temporal ranges and resolutions.

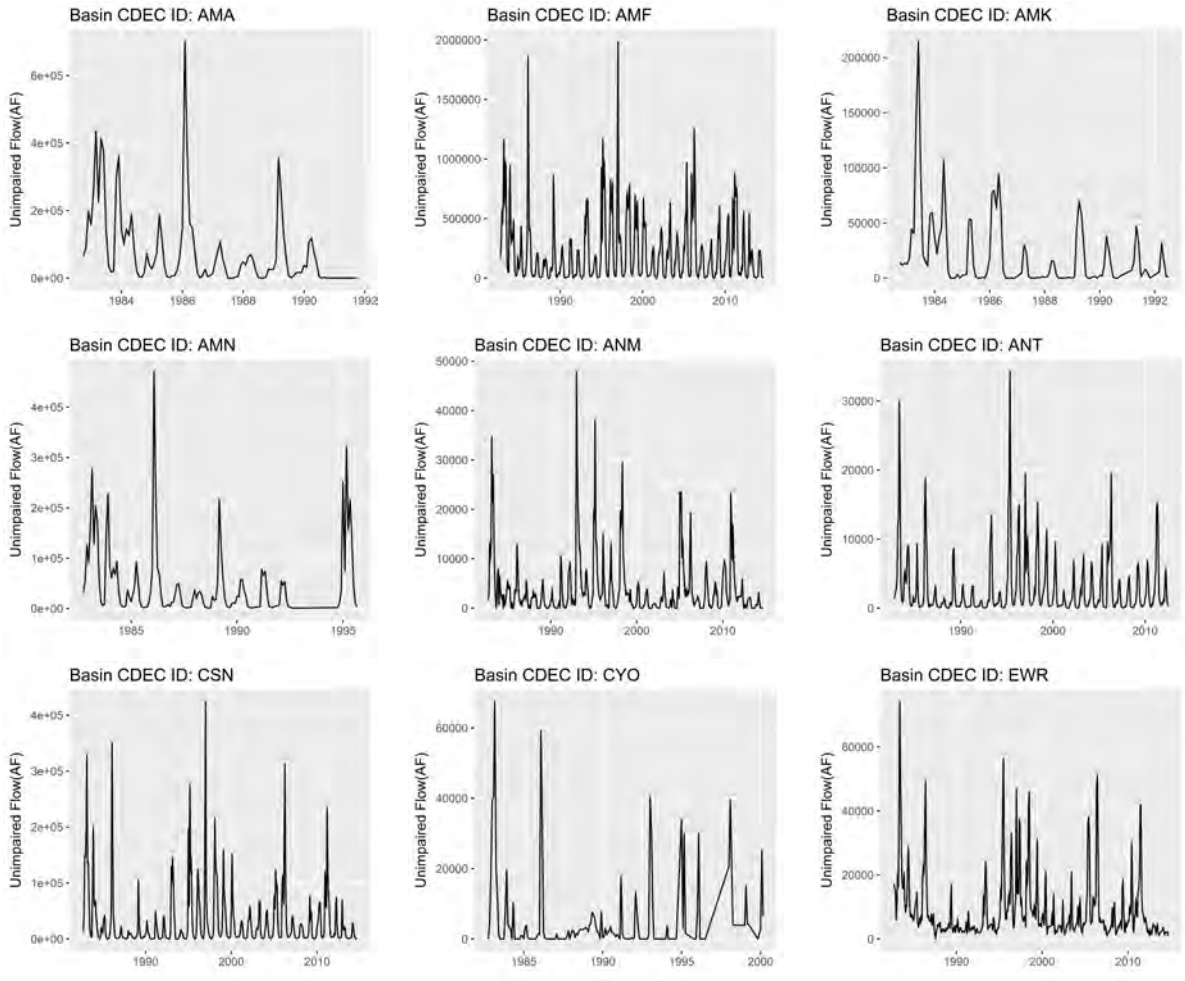
Future endeavors may include ways to employ statistical learning methods in mechanistic models, when accuracy of predictions is of more concern than interpretability. For example, ensemble learning methods can be used to calibrate hydrologic and water quality parameters, a difficult step in developing traditional mechanistic models. In principle, statistical learning

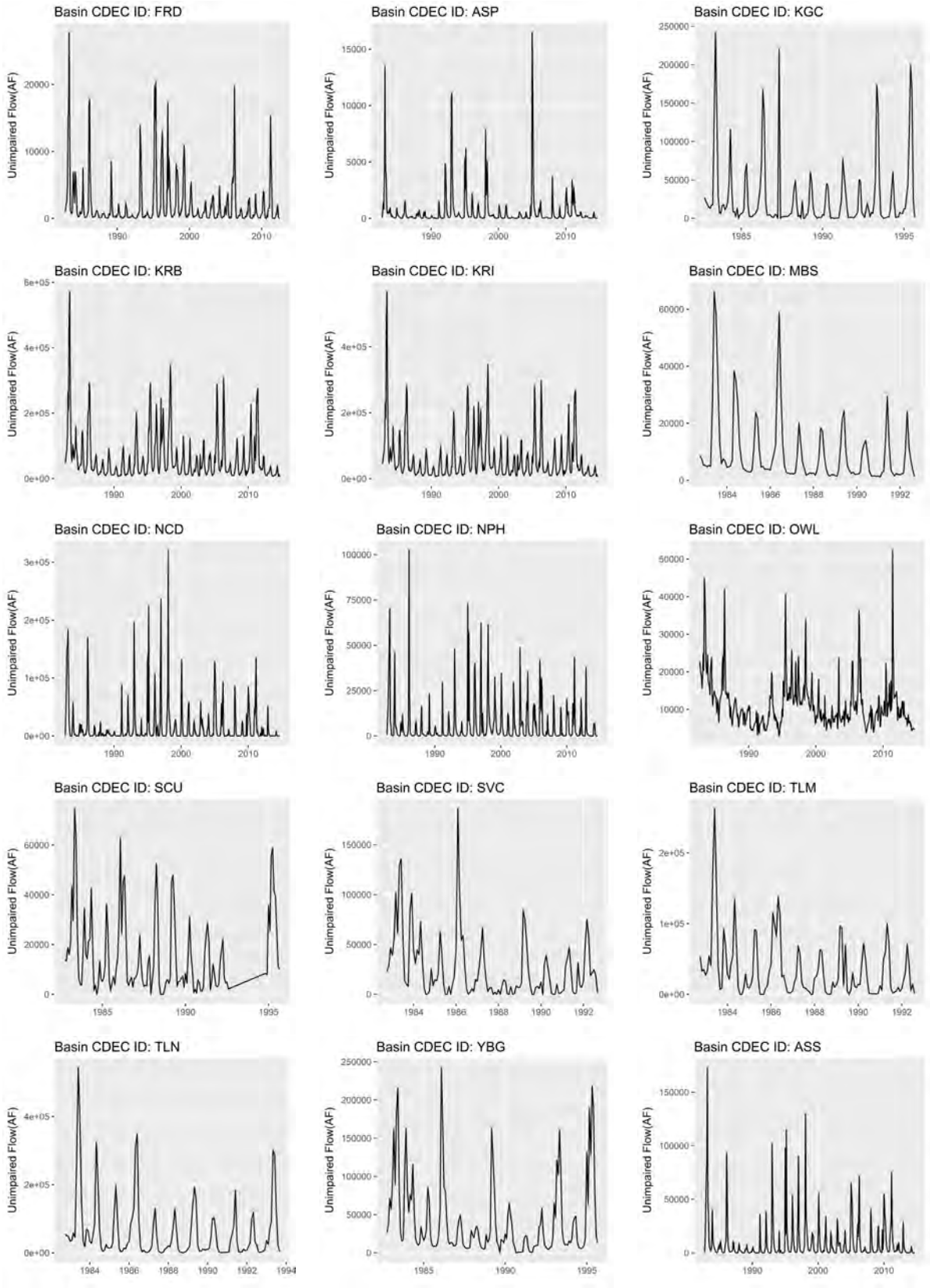
is about embracing noise, variability, and even errors in the data. It is precisely for this reason that they offer the potential for a more "intelligent" model than traditional approaches.

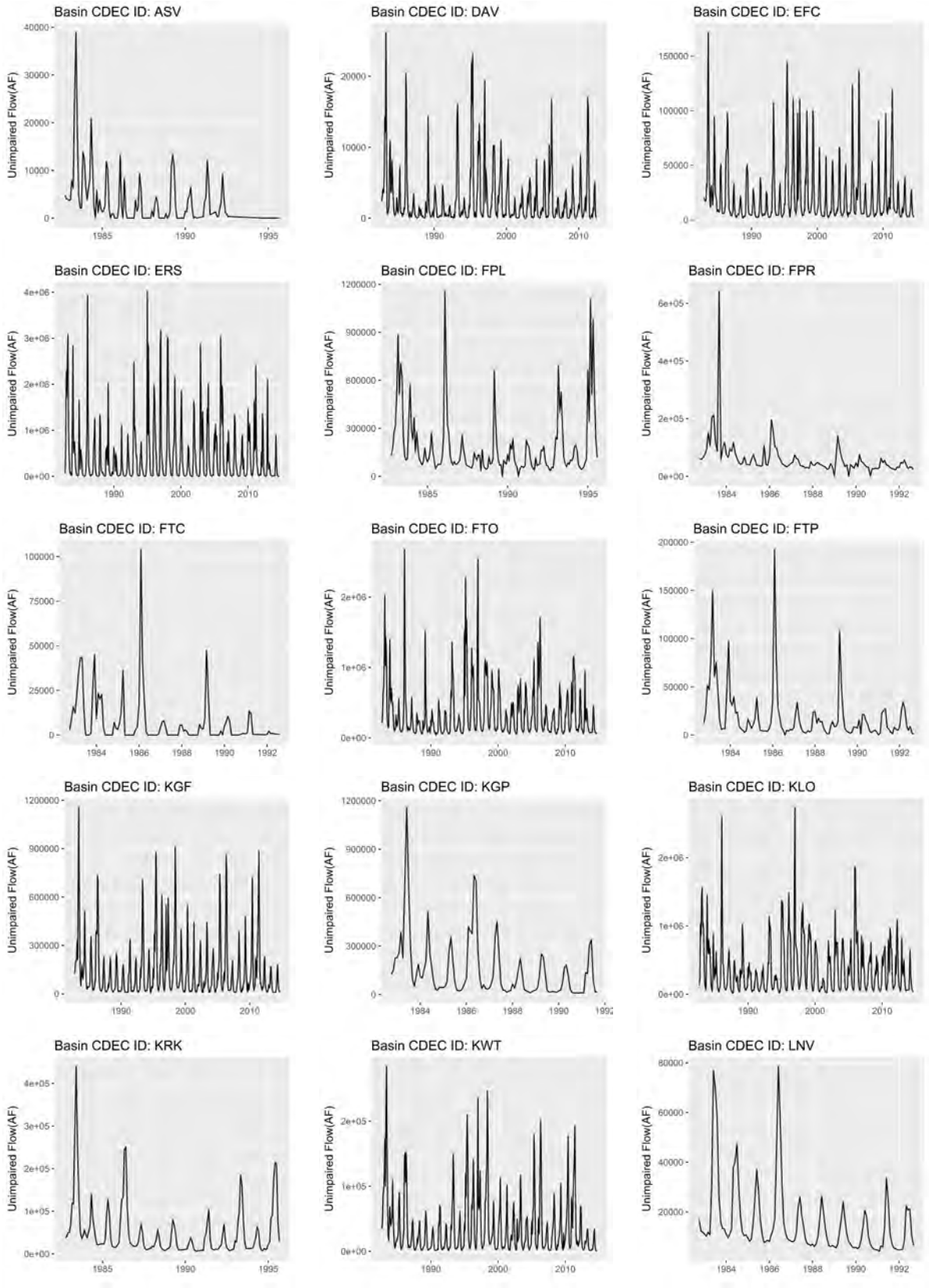
Appendix A

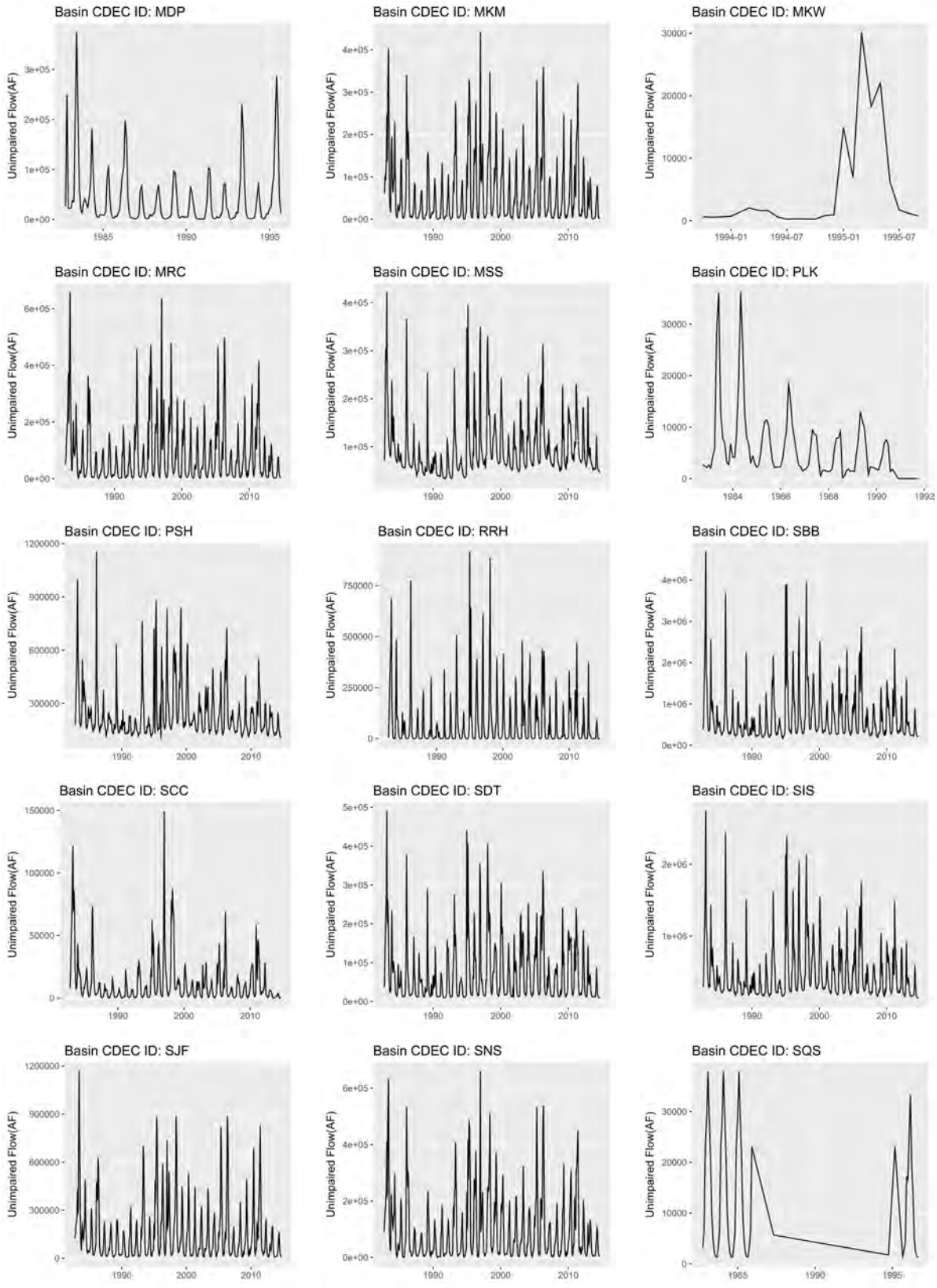
Unimpaired Flow Data Used in Modeling

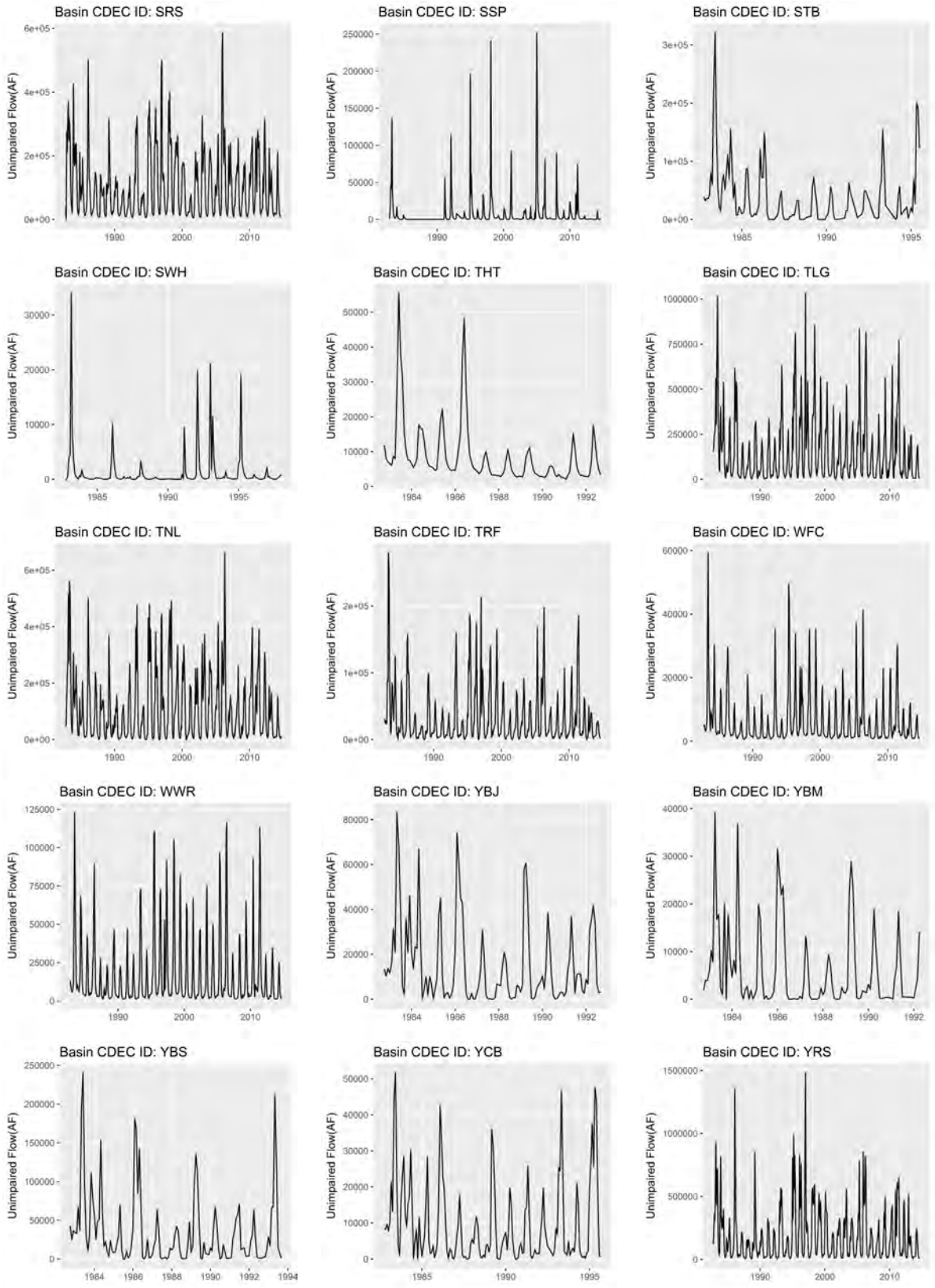
Time series plots of unimpaired flow for all CDEC basins are housed here.




Basin CDEC ID: FRD
Basin CDEC ID: ASP
Basin CDEC ID: KGC
Basin CDEC ID: KRB
Basin CDEC ID: KRI
Basin CDEC ID: MBS
Basin CDEC ID: NCD
Basin CDEC ID: NPH
Basin CDEC ID: OWL
Basin CDEC ID: SCU
Basin CDEC ID: SVC
Basin CDEC ID: TLM
Basin CDEC ID: TLN
Basin CDEC ID: YBG
Basin CDEC ID: ASS
Unimpaired Flow(AF)


Basin CDEC ID: ASV
Basin CDEC ID: DAV
Basin CDEC ID: EFC
Basin CDEC ID: ERS
Basin CDEC ID: FPL
Basin CDEC ID: FPR
Basin CDEC ID: FTC
Basin CDEC ID: FTO
Basin CDEC ID: FTP
Basin CDEC ID: KGF
Basin CDEC ID: KGP
Basin CDEC ID: KLO
Basin CDEC ID: KRK
Basin CDEC ID: KWT
Basin CDEC ID: LNV
Unimpaired Flow(AF)

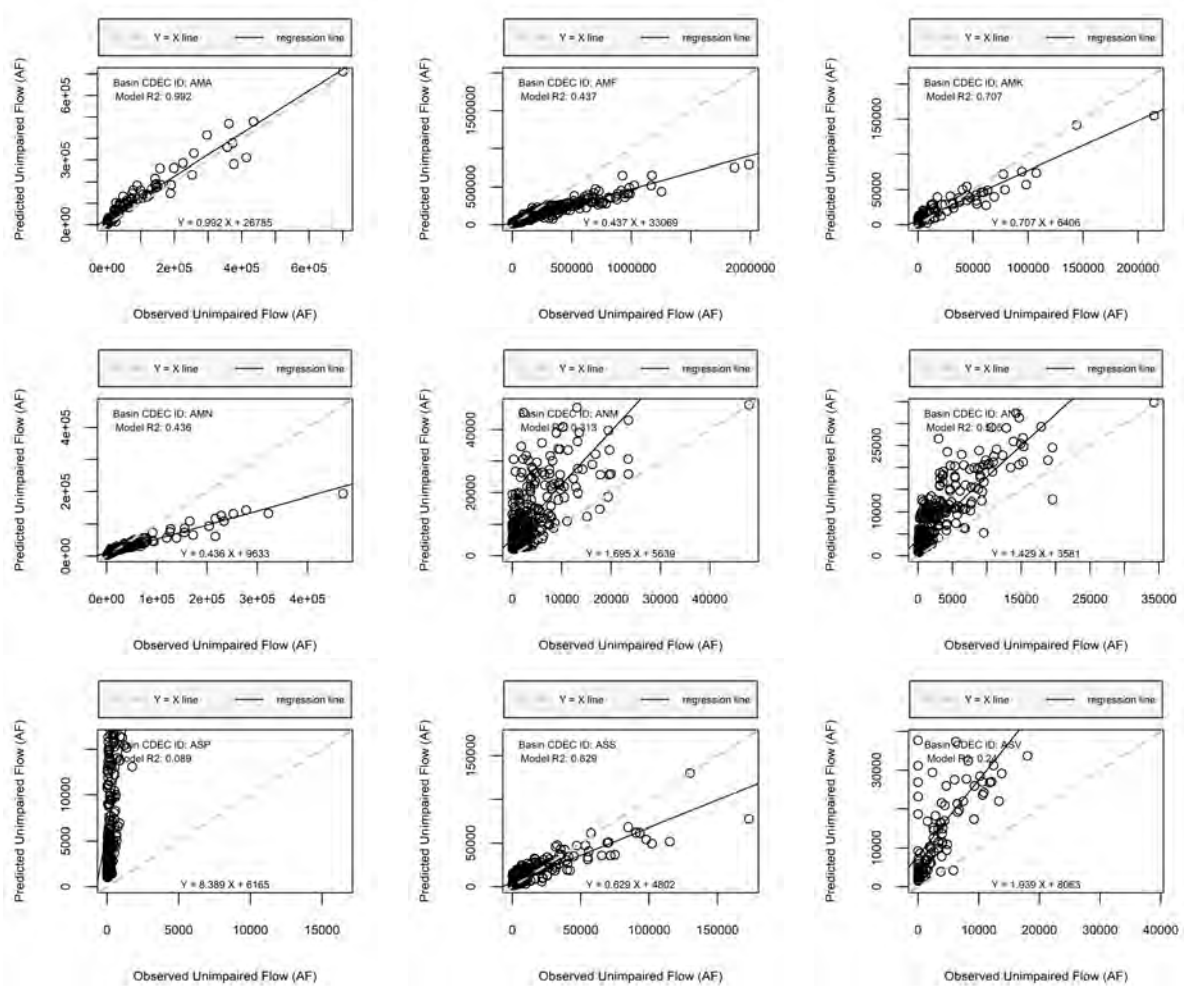

Basin CDEC ID: MDP
Basin CDEC ID: MKM
Basin CDEC ID: MKW
Basin CDEC ID: MRC
Basin CDEC ID: MSS
Basin CDEC ID: PLK
Basin CDEC ID: PSH
Basin CDEC ID: RRH
Basin CDEC ID: SBB
Basin CDEC ID: SCC
Basin CDEC ID: SDT
Basin CDEC ID: SIS
Basin CDEC ID: SJF
Basin CDEC ID: SNS
Basin CDEC ID: SQS
Unimpaired Flow(AF)

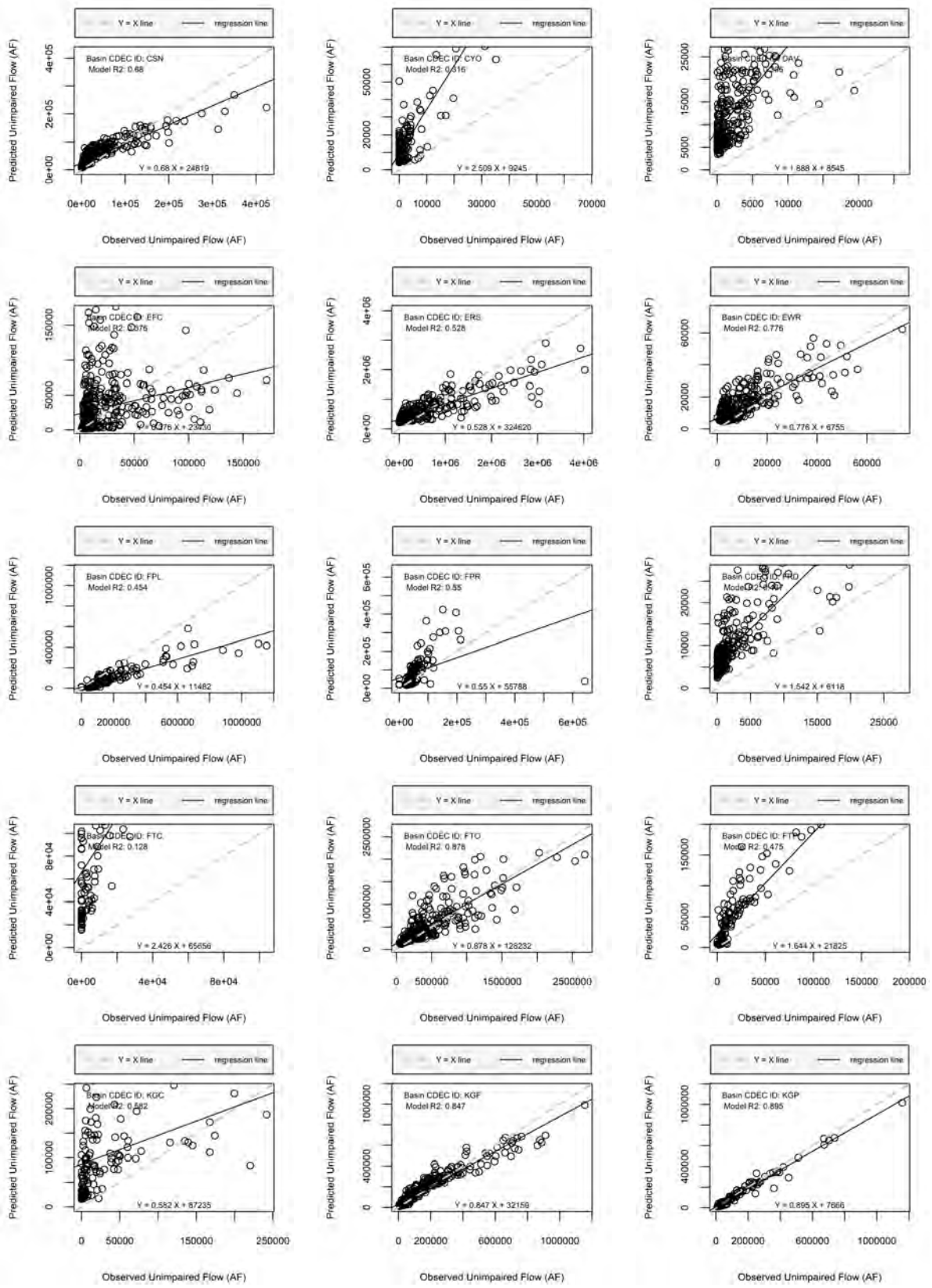

Basin CDEC ID: SRS
Basin CDEC ID: SSP
Basin CDEC ID: STB
Basin CDEC ID: SWH
Basin CDEC ID: THT
Basin CDEC ID: TLG
Basin CDEC ID: TNL
Basin CDEC ID: TRF
Basin CDEC ID: WFC
Basin CDEC ID: WWR
Basin CDEC ID: YBJ
Basin CDEC ID: YBM
Basin CDEC ID: YBS
Basin CDEC ID: YCB
Basin CDEC ID: YRS
Unimpaired Flow(AF)

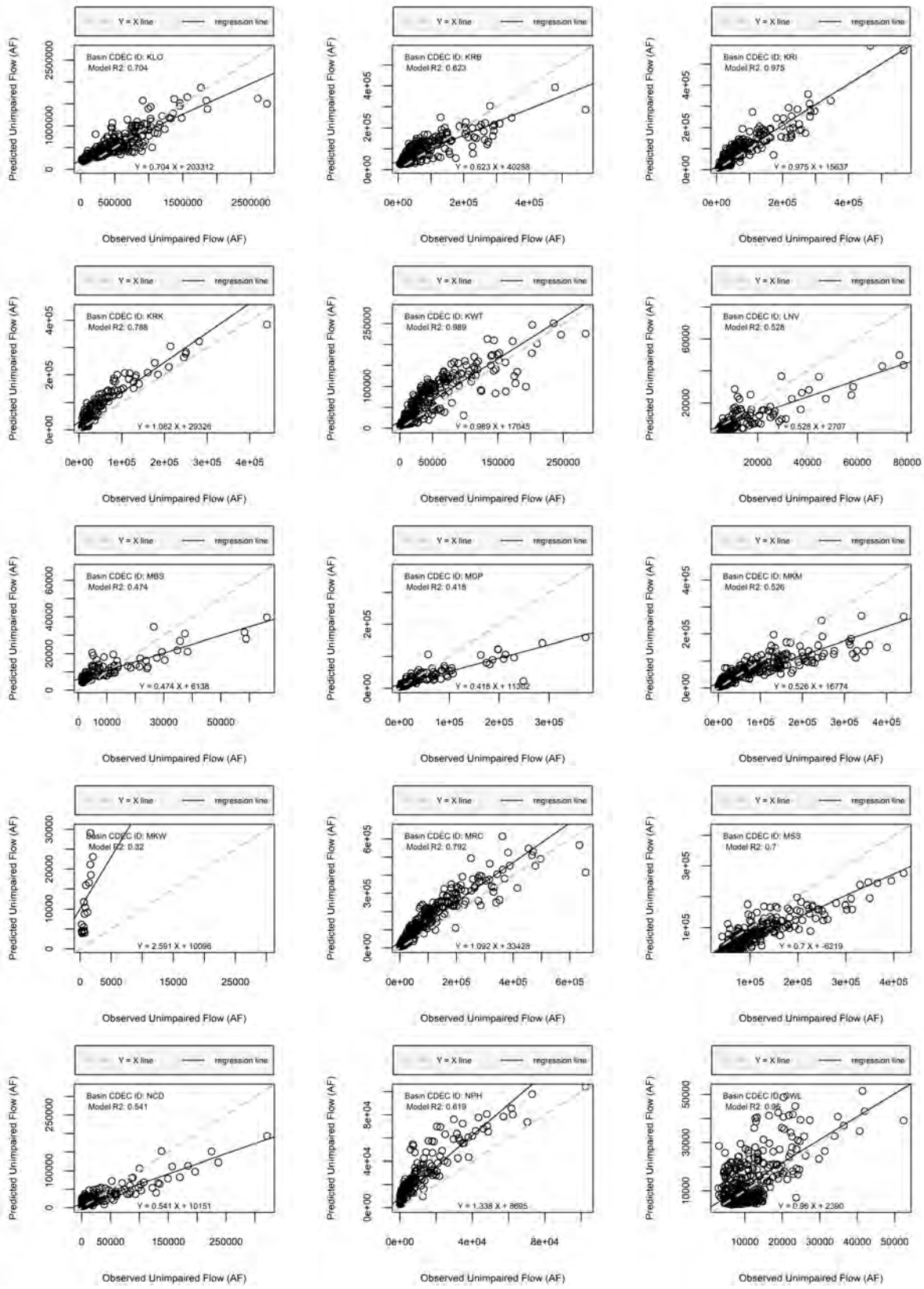
Appendix B

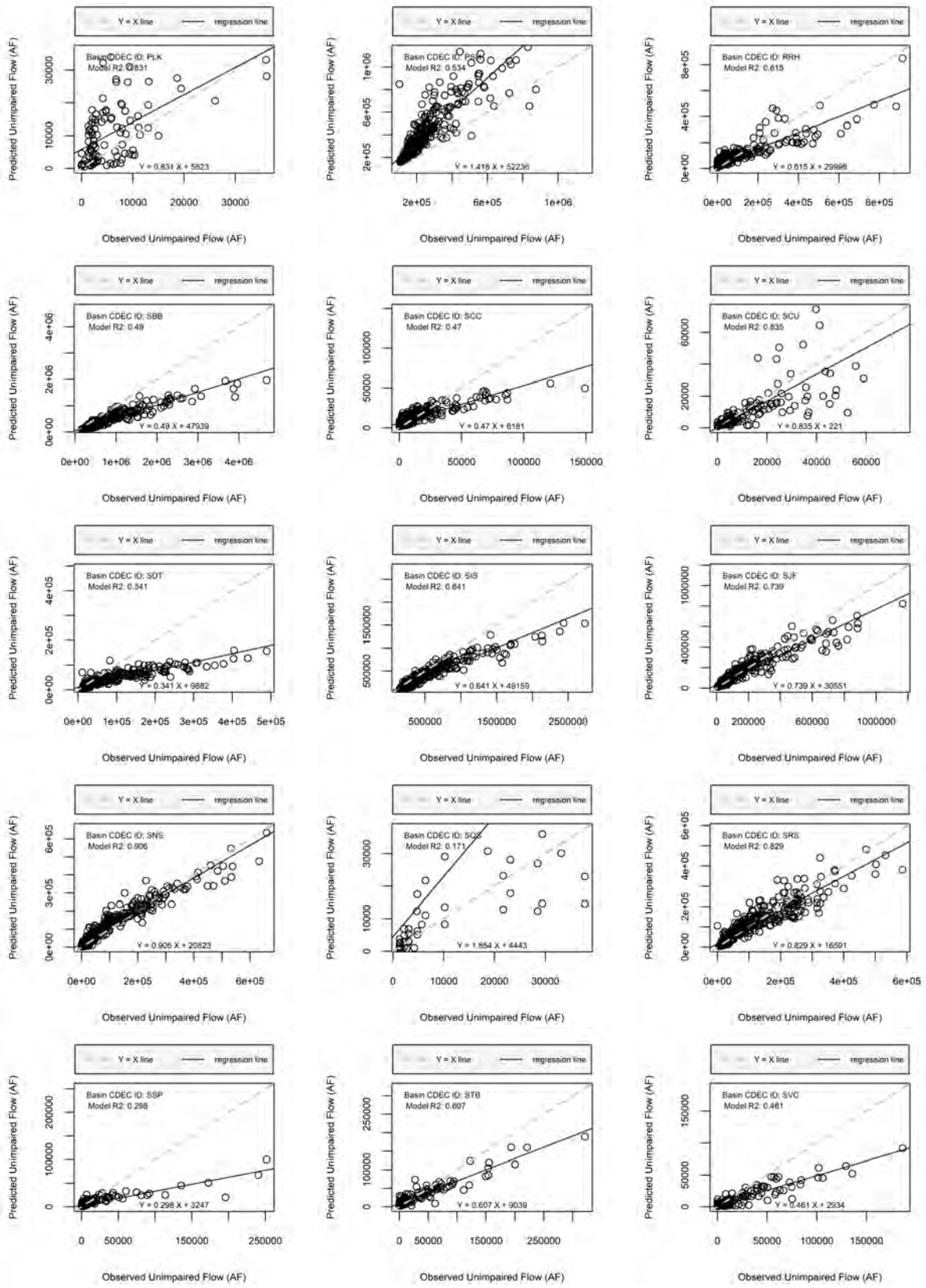
Detailed Random Forest Model Results By Basin

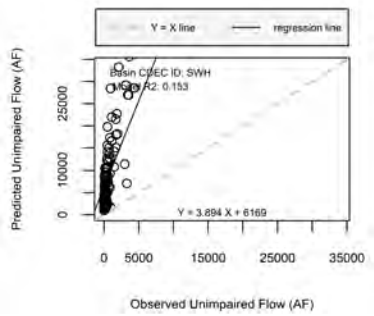
Predicted versus observed plots disaggregated by basins are housed here.

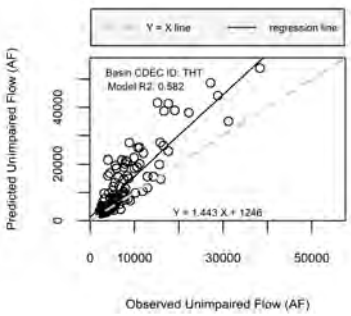


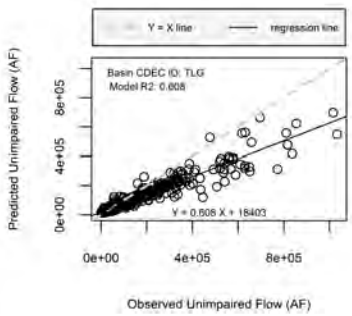

Y = X line
regression line
Basin CDEC ID: CSN
Model R2: 0.68
Y = 0.68 X + 24819
Basin CDEC ID: CYO
Model R2: 0.516
Y = 2.509 X + 9245
Basin CDEC ID: DAV
Model R2: 0.486
Y = 1.888 X + 8545
Basin CDEC ID: EFC
Model R2: 0.376
Y = 0.376 X + 23430
Basin CDEC ID: ERS
Model R2: 0.528
Y = 0.528 X + 324620
Basin CDEC ID: EWR
Model R2: 0.776
Y = 0.776 X + 6755
Basin CDEC ID: FPL
Model R2: 0.454
Y = 0.454 X + 11482
Basin CDEC ID: FPR
Model R2: 0.55
Y = 0.55 X + 55788
Basin CDEC ID: FRD
Model R2: 0.437
Y = 1.542 X + 6118
Basin CDEC ID: FTC
Model R2: 0.128
Y = 2.426 X + 65656
Basin CDEC ID: FTO
Model R2: 0.878
Y = 0.878 X + 126232
Basin CDEC ID: FTO
Model R2: 0.475
Y = 1.644 X + 21825
Basin CDEC ID: KGC
Model R2: 0.582
Y = 0.582 X + 87235
Basin CDEC ID: KGF
Model R2: 0.847
Y = 0.847 X + 32158
Basin CDEC ID: KGP
Model R2: 0.895
Y = 0.895 X + 7666
Predicted Unimpaired Flow (AF)
Observed Unimpaired Flow (AF)

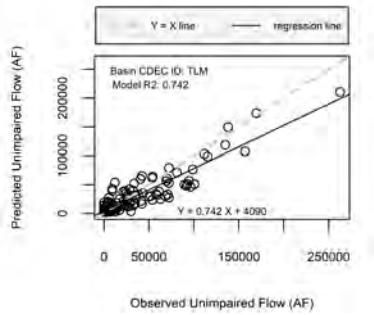

Y = X line
regression line
Basin CDEC ID: KLO
Model R2: 0.704
Y = 0.704 X + 203312
Basin CDEC ID: KRB
Model R2: 0.623
Y = 0.623 X + 40288
Basin CDEC ID: KRI
Model R2: 0.975
Y = 0.975 X + 15637
Basin CDEC ID: KRK
Model R2: 0.788
Y = 1.082 X + 29326
Basin CDEC ID: KWT
Model R2: 0.989
Y = 0.989 X + 17045
Basin CDEC ID: LNV
Model R2: 0.528
Y = 0.528 X + 2707
Basin CDEC ID: MBS
Model R2: 0.474
Y = 0.474 X + 6138
Basin CDEC ID: MGP
Model R2: 0.418
Y = 0.418 X + 11302
Basin CDEC ID: MKM
Model R2: 0.526
Y = 0.526 X + 16774
Basin CDEC ID: MKW
Model R2: 0.32
Y = 2.591 X + 10096
Basin CDEC ID: MRC
Model R2: 0.792
Y = 1.092 X + 33428
Basin CDEC ID: MSS
Model R2: 0.7
Y = 0.7 X + -6219
Basin CDEC ID: NCD
Model R2: 0.541
Y = 0.541 X + 10151
Basin CDEC ID: NPH
Model R2: 0.619
Y = 1.338 X + 8695
Basin CDEC ID: OWL
Model R2: 0.96
Y = 0.96 X + 2390
Predicted Unimpaired Flow (AF)
Observed Unimpaired Flow (AF)

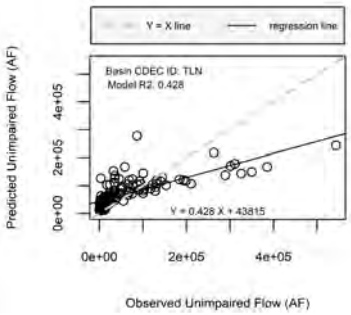

Y = X line
regression line
Basin CDEC ID: PLK
Model R2: 0.831
Y = 0.831 X + 5823
Basin CDEC ID: PSC
Model R2: 0.534
Y = 1.416 X + 52236
Basin CDEC ID: RRH
Model R2: 0.615
Y = 0.615 X + 29998
Basin CDEC ID: SBB
Model R2: 0.49
Y = 0.49 X + 47939
Basin CDEC ID: SCC
Model R2: 0.47
Y = 0.47 X + 6181
Basin CDEC ID: SCU
Model R2: 0.835
Y = 0.835 X + 221
Basin CDEC ID: SDT
Model R2: 0.341
Y = 0.341 X + 9882
Basin CDEC ID: SIS
Model R2: 0.641
Y = 0.641 X + 49159
Basin CDEC ID: SJF
Model R2: 0.739
Y = 0.739 X + 30551
Basin CDEC ID: SNS
Model R2: 0.906
Y = 0.906 X + 20823
Basin CDEC ID: SOS
Model R2: 0.171
Y = 1.854 X + 4443
Basin CDEC ID: SRS
Model R2: 0.829
Y = 0.829 X + 16501
Basin CDEC ID: SSP
Model R2: 0.298
Y = 0.298 X + 3247
Basin CDEC ID: STB
Model R2: 0.607
Y = 0.607 X + 9039
Basin CDEC ID: SVC
Model R2: 0.461
Y = 0.461 X + 2934
Predicted Unimpaired Flow (AF)
Observed Unimpaired Flow (AF)

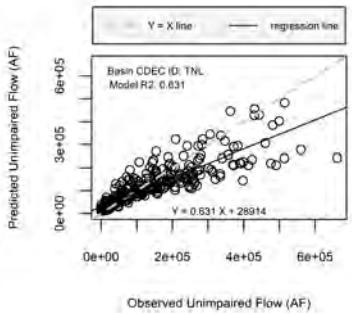

Y = X line
regression line
Basin CDEC ID: SWH
Model R2: 0.153
Y = 3.894 X + 6169
Predicted Unimpaired Flow (AF)
Observed Unimpaired Flow (AF)

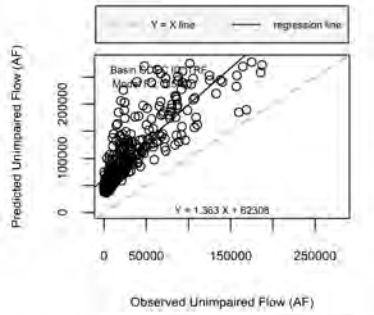

Y = X line
regression line
Basin CDEC ID: THT
Model R2: 0.582
Y = 1.443 X + 1246
Predicted Unimpaired Flow (AF)
Observed Unimpaired Flow (AF)

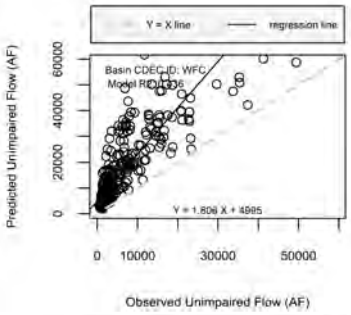

Y = X line
regression line
Basin CDEC ID: TLG
Model R2: 0.808
Y = 0.508 X + 18403
Predicted Unimpaired Flow (AF)
Observed Unimpaired Flow (AF)

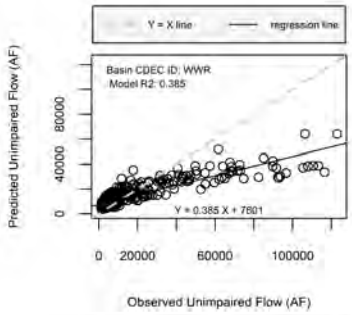

Y = X line
regression line
Basin CDEC ID: TLM
Model R2: 0.742
Y = 0.742 X + 4090
Predicted Unimpaired Flow (AF)
Observed Unimpaired Flow (AF)

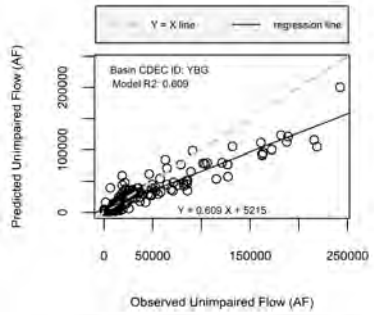

Y = X line
regression line
Basin CDEC ID: TLN
Model R2: 0.428
Y = 0.428 X + 43815
Predicted Unimpaired Flow (AF)
Observed Unimpaired Flow (AF)

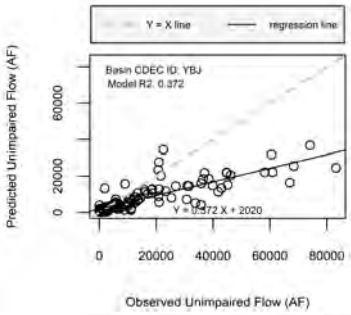

Y = X line
regression line
Basin CDEC ID: TNL
Model R2: 0.631
Y = 0.631 X + 28914
Predicted Unimpaired Flow (AF)
Observed Unimpaired Flow (AF)

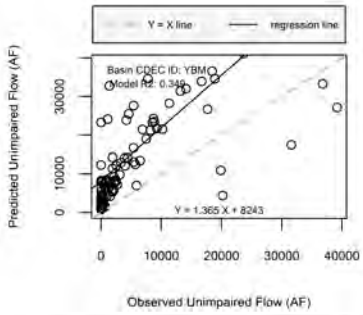

Y = X line
regression line
Y = 1.363 X + 62308
Predicted Unimpaired Flow (AF)
Observed Unimpaired Flow (AF)

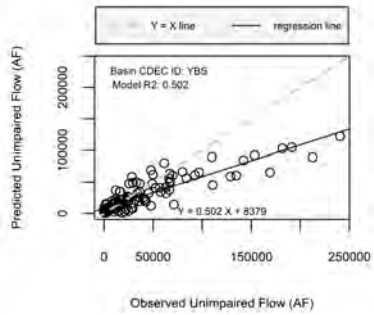

Y = X line
regression line
Basin CDEC ID: WFC
Y = 1.806 X + 4995
Predicted Unimpaired Flow (AF)
Observed Unimpaired Flow (AF)

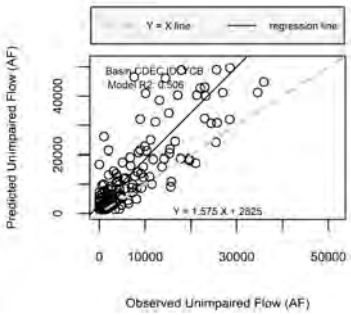

Y = X line
regression line
Basin CDEC ID: WWR
Model R2: 0.385
Y = 0.385 X + 7601
Predicted Unimpaired Flow (AF)
Observed Unimpaired Flow (AF)

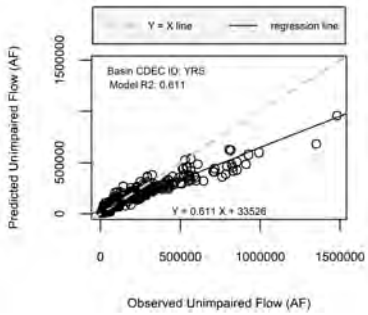

Y = X line
regression line
Basin CDEC ID: YBG
Model R2: 0.609
Y = 0.609 X + 5215
Predicted Unimpaired Flow (AF)
Observed Unimpaired Flow (AF)


Y = X line
regression line
Basin CDEC ID: YBJ
Model R2: 0.372
Y = 0.372 X + 2020
Predicted Unimpaired Flow (AF)
Observed Unimpaired Flow (AF)


Y = X line
regression line
Basin CDEC ID: YBM
Y = 1.365 X + 8243
Predicted Unimpaired Flow (AF)
Observed Unimpaired Flow (AF)


Y = X line
regression line
Basin CDEC ID: YBS
Model R2: 0.502
Y = 0.502 X + 8379
Predicted Unimpaired Flow (AF)
Observed Unimpaired Flow (AF)


Y = X line
regression line
Basin CDEC ID: YCB
Y = 1.575 X + 2825
Predicted Unimpaired Flow (AF)
Observed Unimpaired Flow (AF)


Y = X line
regression line
Basin CDEC ID: YRS
Model R2: 0.611
Y = 0.611 X + 33526
Predicted Unimpaired Flow (AF)
Observed Unimpaired Flow (AF)

Appendix C

Detailed Measures of Random Forest Model Performance

Model performance tables are housed here.

Table C.1: Random forest model fit summary for the combined data sets.

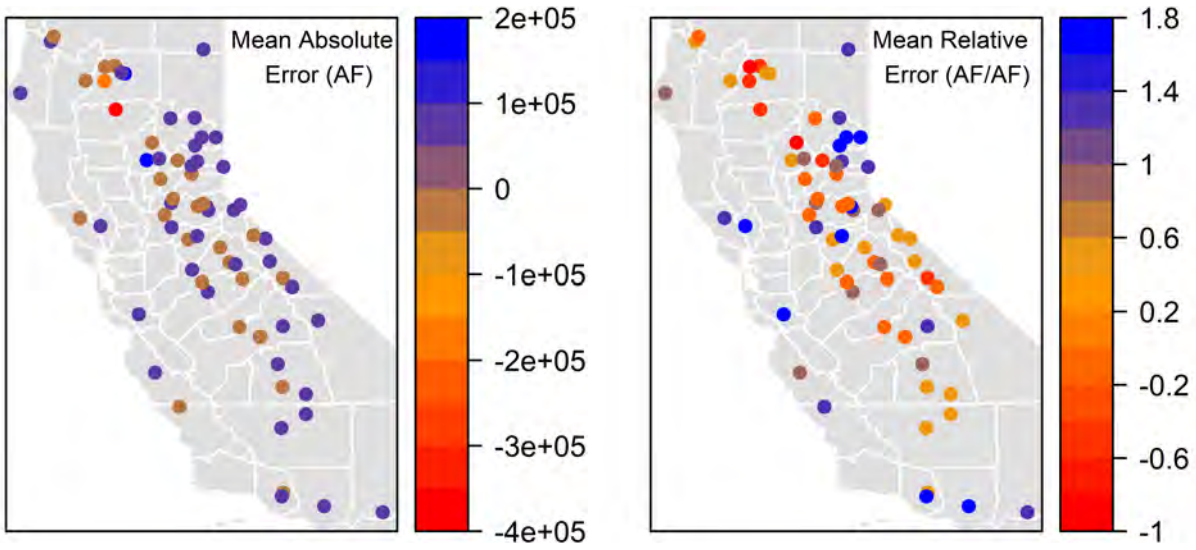
Statistical Measure	Train Set	Test Set	High Flows	Low Flows	Large Basins	Small Basins
No. of Observation	1,260,448	18,536	4,312	14,224	291,563	13,523
Mean AE	152	1729	-44088	15618	139	-3426
Mean RE	0.24	0.52	-0.11	0.71	0.24	0.60
RMSE (AF)	32955	123348	241961	45599	32619	103107
RSR	0.14	0.51	0.61	1.90	0.14	0.50
R2	0.94	0.69	0.60	0.26	0.94	0.61
NSE	0.98	0.74	0.63	-2.60	0.98	0.75
Mean PBIAS	0.00	0.05	0.00	0.09	0.00	0.09

Table C.2: Random forest model fit summary by basin.

CDEC ID	No. Train Set	No. Test Set	Mean AE	Sd AE	Mean RE	R^2	RMSE	NSE	RSR	Mean PBIAS
AMA	18436	100	26369	32353	0.69	0.94	41612	0.88	0.35	174
AMF	18152	384	-90321	163756	-0.03	0.44	186826	0.56	0.66	104
AMK	18426	110	136	13059	0.81	0.68	13001	0.85	0.39	219
AMN	18404	132	-21126	46702	-0.02	0.37	51096	0.52	0.69	18
ANM	18158	378	8814	10244	1.22	0.32	13504	-3.95	2.22	1315
ANT	18180	356	5562	5203	1.24	0.45	7611	-1.92	1.71	827
ASP	18152	384	9036	13192	1.78	0.10	15976	-88.68	9.47	4949
ASS	18152	384	800	10297	0.75	0.64	10315	0.74	0.51	1583
ASV	18412	124	11354	12312	1.44	0.25	16712	-7.96	2.99	42998
CSN	18152	384	16928	25136	1.03	0.73	30277	0.70	0.55	947
CYO	18367	169	19107	22654	1.61	0.31	29584	-6.52	2.74	3741

CDEC ID	No. Train Set	No. Test Set	Mean AE	Sd AE	Mean RE	R^2	RMSE	NSE	RSR	Mean PBIAS
DAV	18178	358	9684	6417	1.49	0.35	11612	-6.90	2.81	3308
EFC	18152	384	11187	38394	0.26	0.22	39943	-1.11	1.45	174
ERS	18152	384	94081	408151	0.91	0.49	418335	0.64	0.60	1639
EWR	18152	384	4106	6274	0.53	0.78	7491	0.47	0.73	118
FPL	18381	155	-84870	118117	-0.65	0.47	145136	0.49	0.71	-47
FPR	18416	120	19472	74470	0.14	0.39	76673	-0.39	1.18	29
FRD	18178	358	7135	4463	1.54	0.44	8412	-3.90	2.21	3474
FTC	18429	107	75840	60095	1.77	0.12	96588	-42.82	6.62	3551
FTO	18152	384	126960	251732	0.41	0.72	281643	0.49	0.71	67
FTP	18418	118	25418	24002	0.74	0.55	34889	-0.51	1.23	561
KGC	18388	148	63025	73211	1.33	0.20	96415	-3.58	2.14	1187
KGF	18152	384	10953	53529	0.20	0.84	54570	0.91	0.30	31
KGP	18428	108	-7505	35285	0.00	0.89	35914	0.96	0.19	4
KLO	18152	384	97237	181884	0.55	0.69	206036	0.73	0.52	141
KRB	18152	384	18294	39972	0.53	0.64	43913	0.64	0.60	102
KRI	18152	384	13766	27022	0.29	0.90	30295	0.82	0.43	44
KRK	18380	156	34140	27611	0.60	0.79	43852	0.50	0.71	110
KWT	18152	384	17630	24790	0.61	0.77	30393	0.57	0.65	126
LNV	18416	120	-5194	8490	-0.37	0.51	9923	0.51	0.70	-25
MBS	18416	120	652	6891	0.41	0.48	6893	0.66	0.58	79
MDP	18380	156	-11731	37209	0.20	0.44	38900	0.59	0.64	66
MKM	18152	384	-12277	43328	0.23	0.54	44980	0.68	0.56	80
MKW	18512	24	21058	20389	1.58	0.27	29015	-12.40	3.66	1065
MRC	18152	384	40665	47491	0.62	0.79	62475	0.66	0.58	121
MSS	18152	384	-33668	30088	-0.57	0.72	45127	0.48	0.72	-39
NCD	18152	384	-678	20035	1.06	0.54	20021	0.72	0.53	375
NPH	18152	384	8143	7655	1.41	0.67	11170	0.19	0.90	4465
OWL	18152	384	2168	7906	0.06	0.42	8188	-0.49	1.22	30
PLK	18437	99	10708	8310	1.03	0.40	13528	-3.13	2.03	363
PSH	18152	384	154495	132149	0.47	0.54	203191	-0.85	1.36	69
RRH	18152	384	-1120	73088	1.05	0.52	73001	0.71	0.54	4235
SBB	18152	384	-304917	366164	-0.56	0.48	476132	0.49	0.71	-43
SCC	18152	384	-373	11287	0.52	0.44	11278	0.62	0.61	412
SCU	18405	131	-1623	10815	-0.06	0.72	10895	0.52	0.69	3
SDT	18152	384	-36091	54947	-0.63	0.34	65680	0.34	0.81	-42
SIS	18152	384	-127369	167921	-0.33	0.64	210587	0.72	0.53	-27
SJF	18152	384	-5275	69269	0.10	0.74	69380	0.86	0.37	18
SNS	18152	384	10195	33311	0.23	0.89	34794	0.91	0.29	201
SQS	18481	55	16296	35337	0.35	0.16	38620	-9.70	3.27	122
SRS	18152	384	-3940	41304	0.01	0.78	41438	0.84	0.40	7
SSP	18221	315	-3303	21186	0.60	0.28	21409	0.44	0.75	397
STB	18408	128	-5480	22885	0.39	0.62	23445	0.79	0.45	92

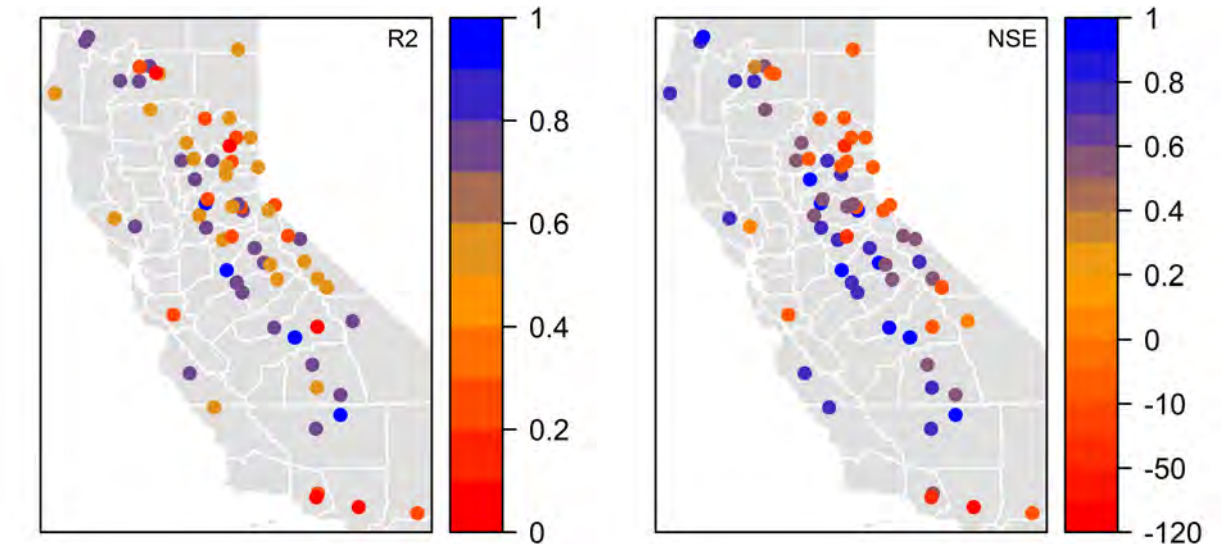
CDEC ID	No. Train Set	No. Test Set	Mean AE	Sd AE	Mean RE	R^2	RMSE	NSE	RSR	Mean PBIAS
SVC	18416	120	-11358	19146	-0.15	0.45	22193	0.52	0.69	-14
SWH	18355	181	9969	16555	1.79	0.15	19285	-23.91	4.99	6605
THT	18416	120	5036	6389	0.37	0.60	8114	0.14	0.93	61
TLG	18152	384	-42470	82911	0.05	0.61	93059	0.76	0.49	204
TLM	18418	118	-5080	17926	0.11	0.77	18559	0.81	0.44	196
TLN	18405	131	11197	61633	0.71	0.42	62410	0.51	0.70	818
TNL	18152	384	-16023	63954	0.22	0.60	65850	0.71	0.54	64
TRF	18152	384	73193	38960	1.32	0.54	82892	-2.71	1.93	1114
WFC	18152	384	9635	10317	0.98	0.43	14107	-1.87	1.69	242
WWR	18152	384	-2093	15211	0.37	0.40	15334	0.58	0.64	82
YBG	18380	156	-10868	23468	-0.23	0.61	25794	0.74	0.51	1
YBJ	18416	120	-7580	12439	-0.32	0.37	14522	0.36	0.80	45
YBM	18427	109	10542	12562	1.26	0.36	16355	-2.62	1.90	998
YBS	18409	127	-8895	26395	0.12	0.50	27755	0.64	0.60	62
YCB	18383	153	8700	11615	0.75	0.50	14482	-0.58	1.26	333
YRS	18152	384	-34307	93642	-0.12	0.64	99615	0.80	0.45	-1



(a) Mean AE of the test set.

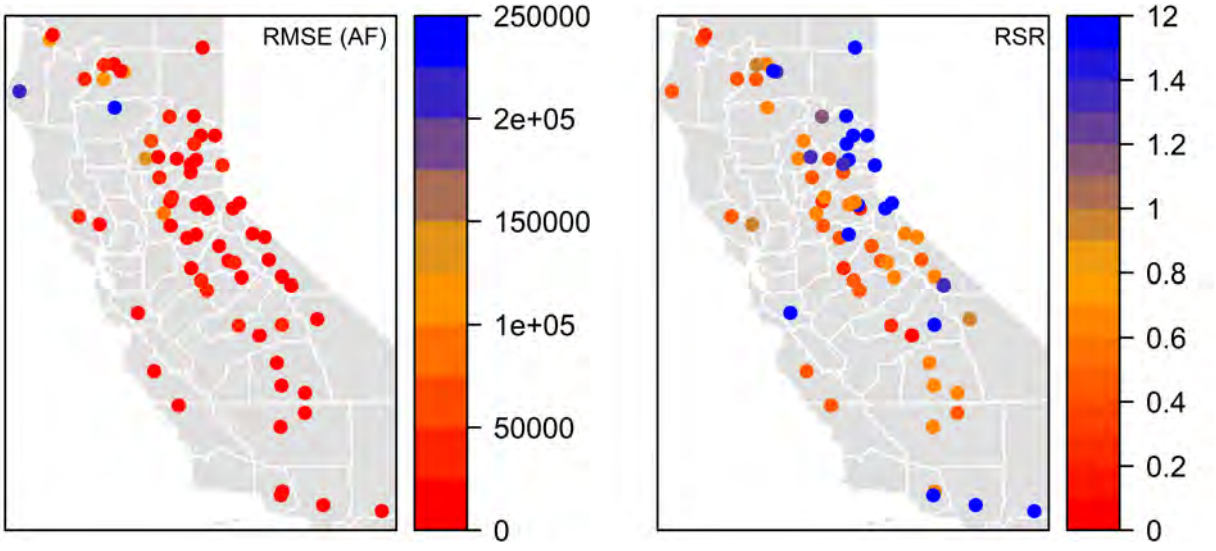
(b) Mean RE of the test set.

Figure C.1: The spatial autocorrelation of the residuals.



(a) R^2 of the test set. (b) NSE of the test set.

Figure C.2: The spatial performance of the model.



(a) RMSE of the test set. (b) RSR of the test set.

Figure C.3: The spatial distribution of the test set errors

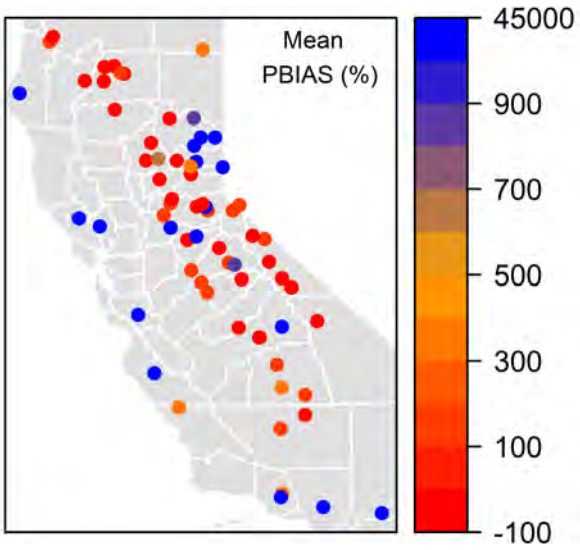
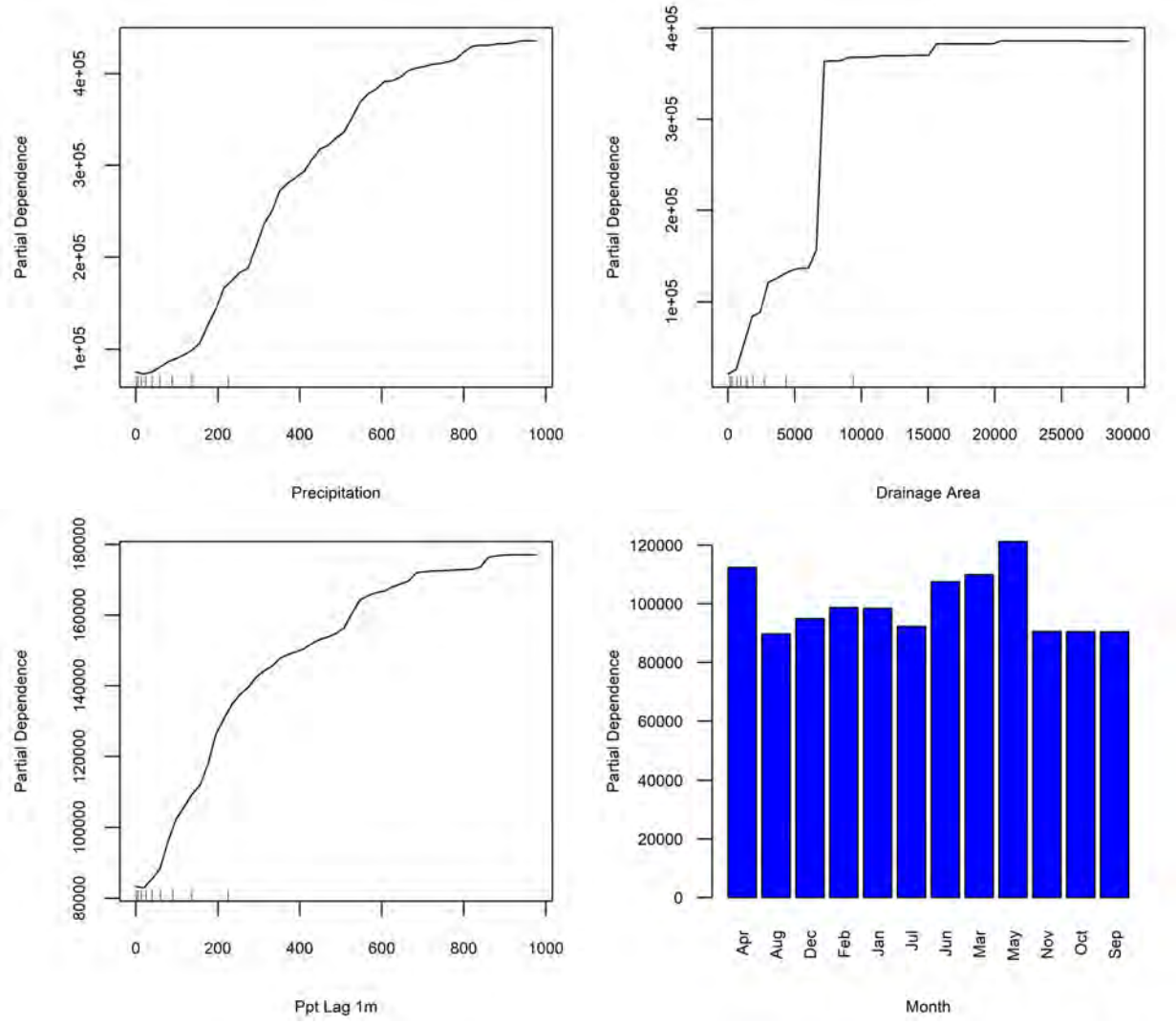


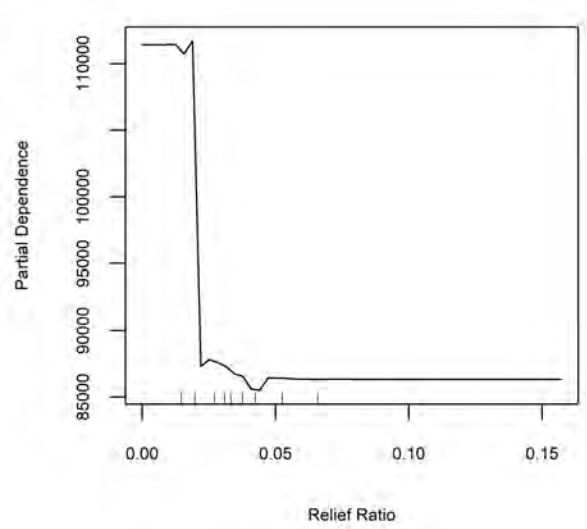
Figure C.4: Mean PBIAS of the test set.

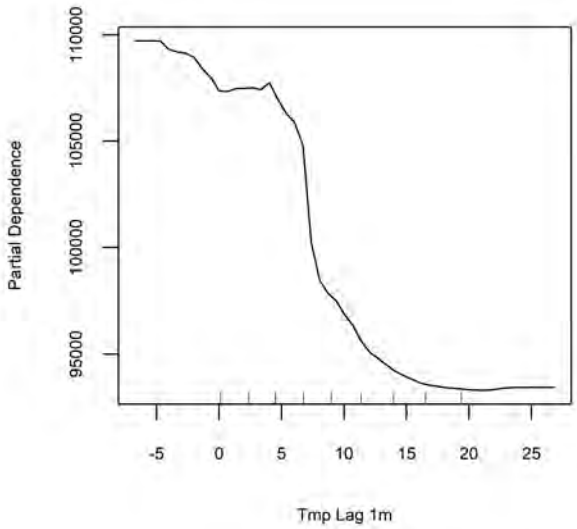
Appendix D

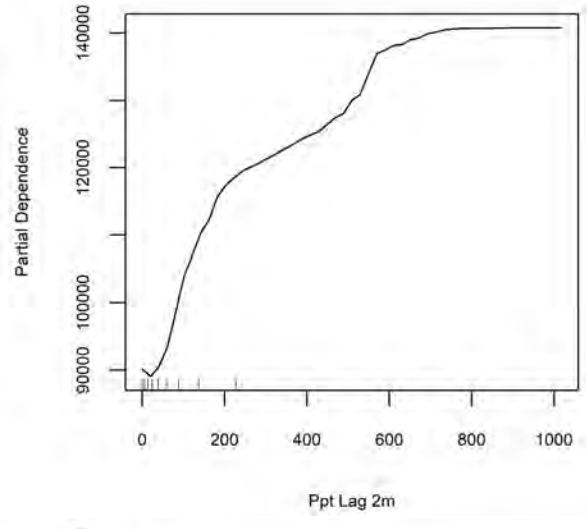
Random Forest Partial Plots

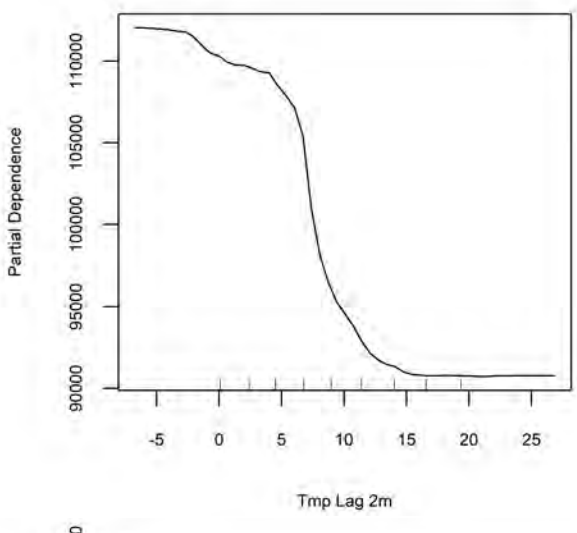
Partial plots, in the order given by the variable importance list, are housed here.

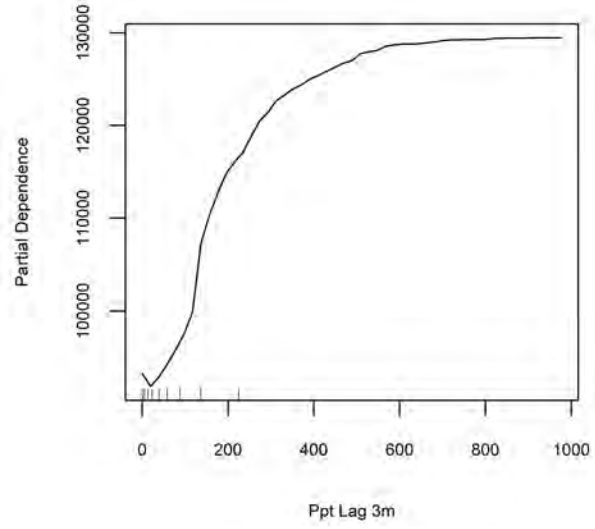


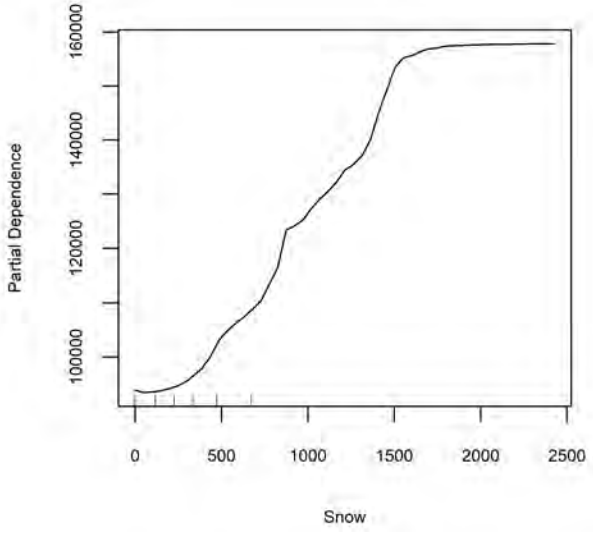

Partial Dependence
85000
90000
95000
100000
110000
0.00
0.05
0.10
0.15
Relief Ratio

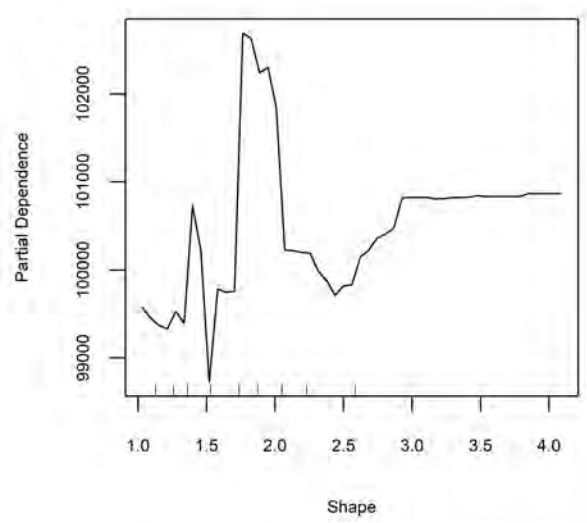

Partial Dependence
95000
100000
105000
110000
-5
0
5
10
15
20
25
Tmp Lag 1m

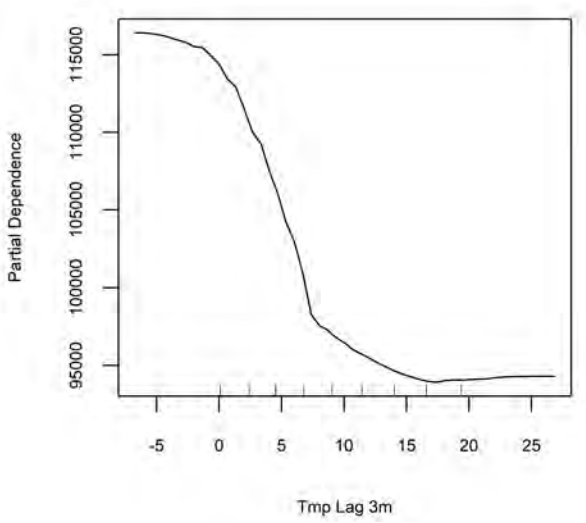

Partial Dependence
90000
100000
120000
140000
0
200
400
600
800
1000
Ppt Lag 2m

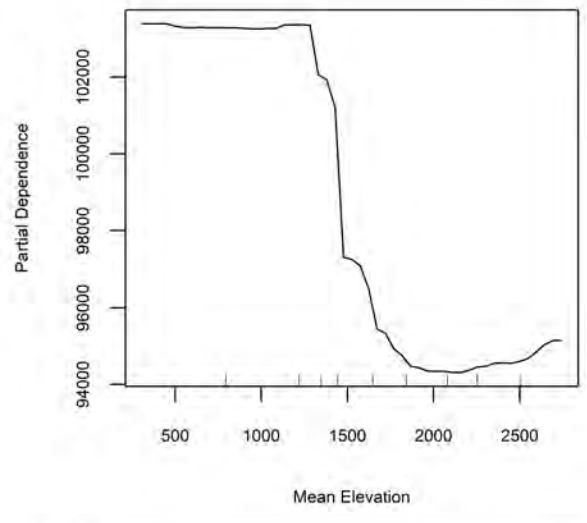

Partial Dependence
90000
95000
100000
105000
110000
-5
0
5
10
15
20
25
Tmp Lag 2m

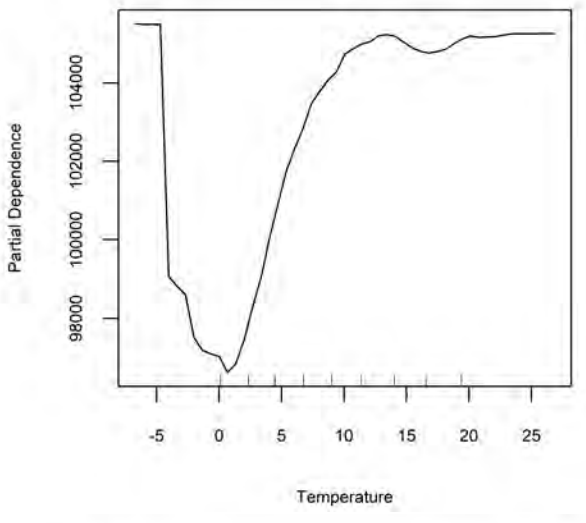

Partial Dependence
100000
110000
120000
130000
0
200
400
600
800
1000
Ppt Lag 3m

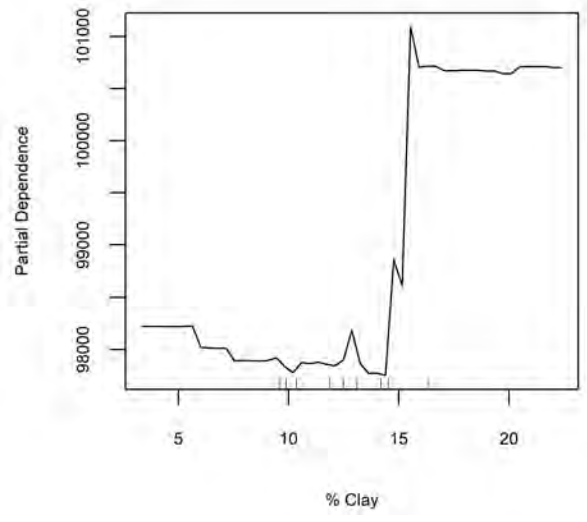

Partial Dependence
100000
120000
140000
160000
0
500
1000
1500
2000
2500
Snow

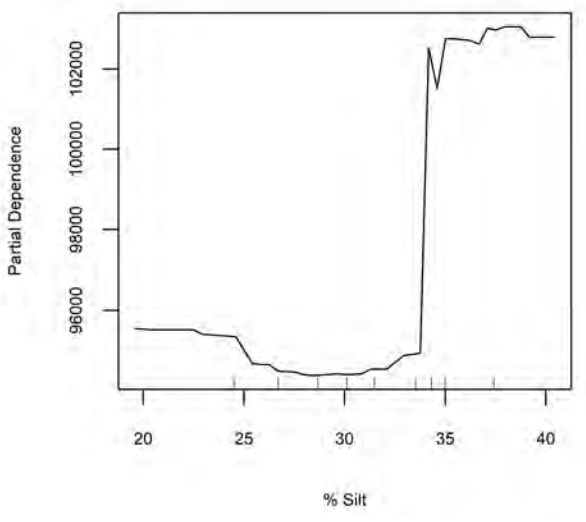

Partial Dependence
99000
100000
101000
102000
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Shape

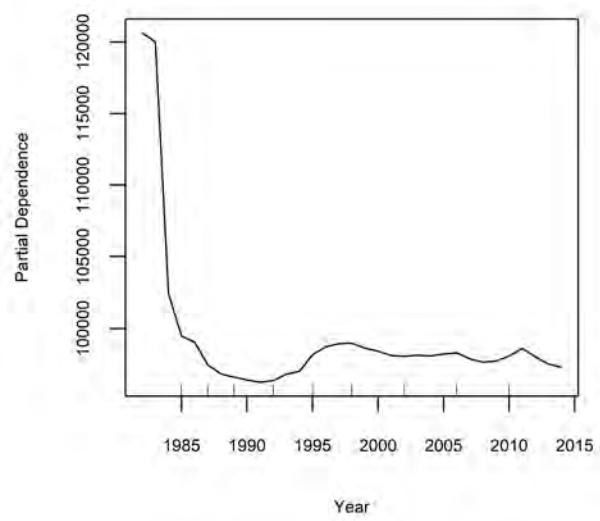

Partial Dependence
95000
100000
105000
110000
115000
-5
0
5
10
15
20
25
Tmp Lag 3m

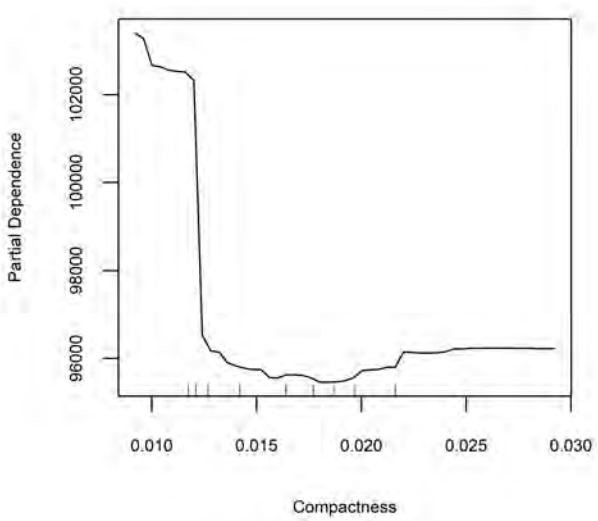

Partial Dependence
94000
96000
98000
100000
102000
500
1000
1500
2000
2500
Mean Elevation

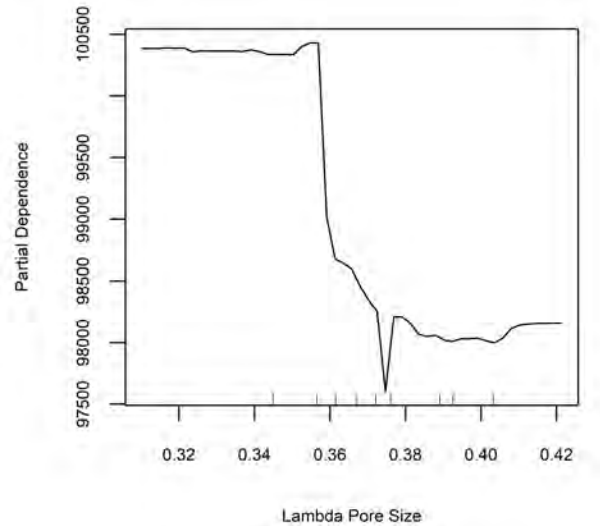

Partial Dependence
98000
100000
102000
104000
-5
0
5
10
15
20
25
Temperature

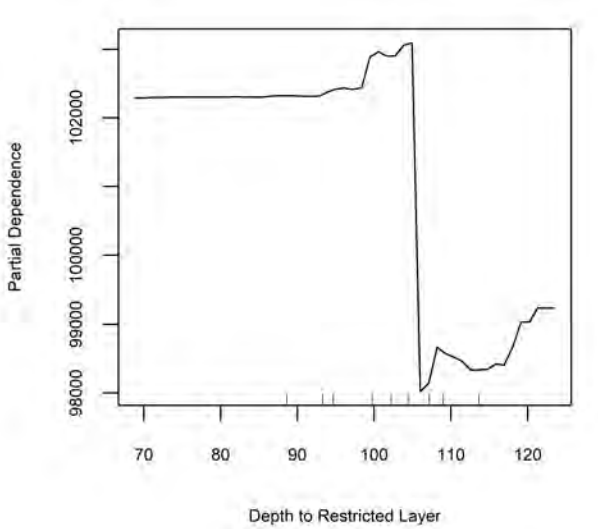

Partial Dependence
98000
99000
100000
101000
5
10
15
20
% Clay

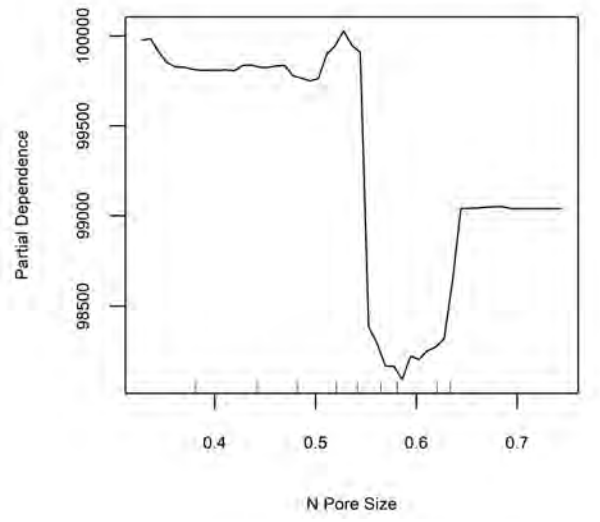

Partial Dependence
96000
98000
100000
102000
20
25
30
35
40
% Silt

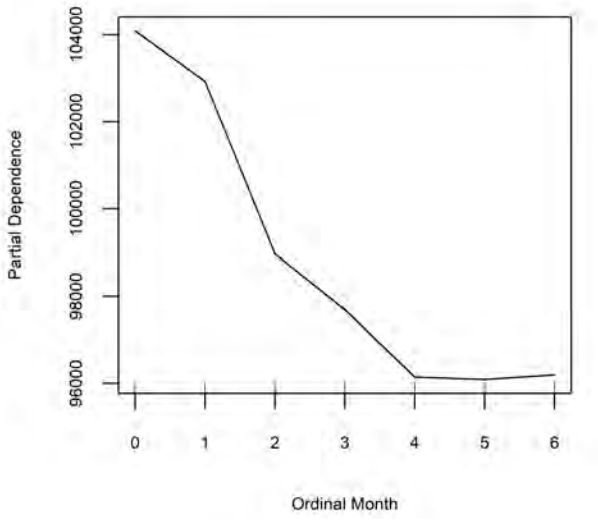

Partial Dependence
100000
105000
110000
115000
120000
1985
1990
1995
2000
2005
2010
2015
Year

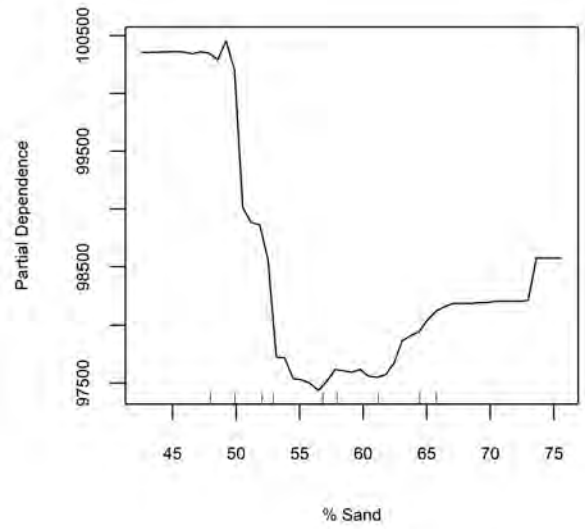

Partial Dependence
96000
98000
100000
102000
0.010
0.015
0.020
0.025
0.030
Compactness

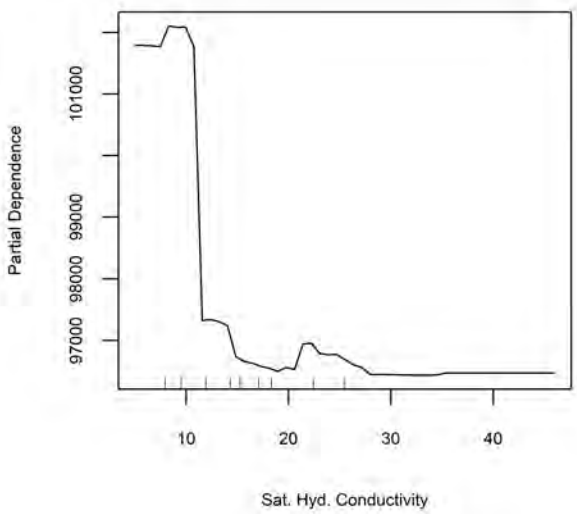

Partial Dependence
97500
98000
98500
99000
99500
100500
0.32
0.34
0.36
0.38
0.40
0.42
Lambda Pore Size

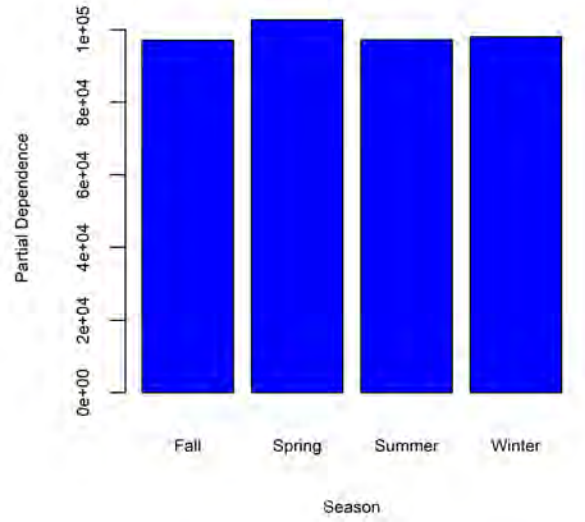

Partial Dependence
98000
99000
100000
102000
70
80
90
100
110
120
Depth to Restricted Layer

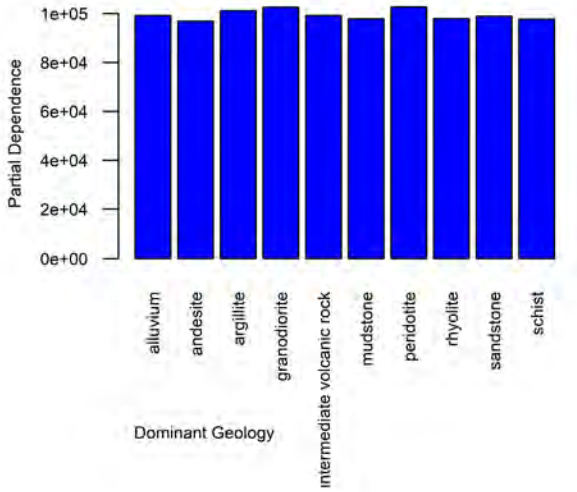

Partial Dependence
98500
99000
99500
100000
0.4
0.5
0.6
0.7
N Pore Size

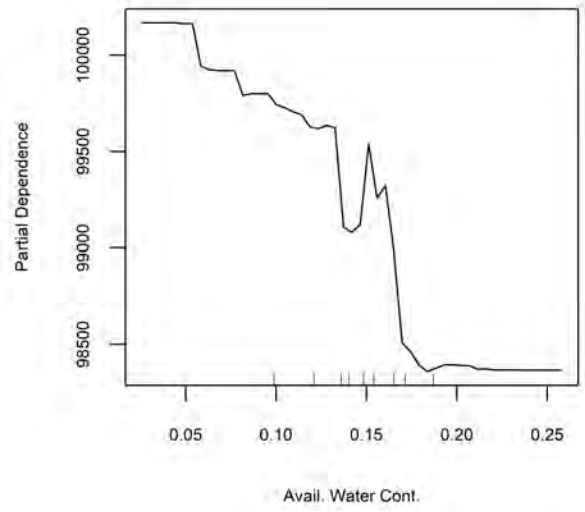

Partial Dependence
96000
98000
100000
102000
104000
0
1
2
3
4
5
6
Ordinal Month

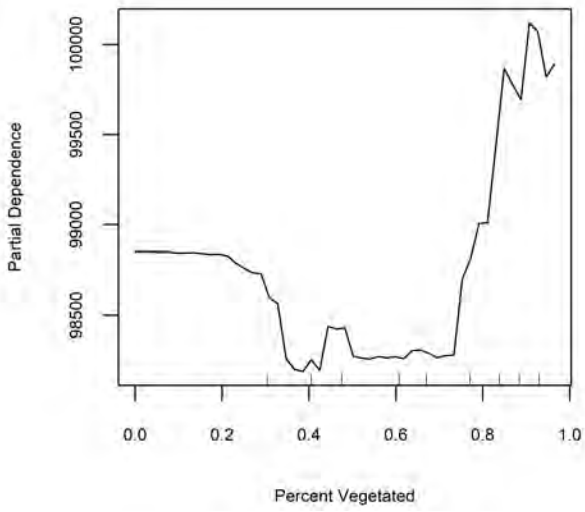

Partial Dependence
100500
99500
98500
97500
45
50
55
60
65
70
75
% Sand


Partial Dependence
101000
99000
98000
97000
10
20
30
40
Sat. Hyd. Conductivity


Partial Dependence
1e+05
8e+04
6e+04
4e+04
2e+04
0e+00
Fall
Spring
Summer
Winter
Season


Partial Dependence
1e+05
8e+04
6e+04
4e+04
2e+04
0e+00
alluvium
andesite
argillite
granodiorite
intermediate volcanic rock
mudstone
peridotite
rhyolite
sandstone
schist
Dominant Geology


Partial Dependence
100000
99500
99000
98500
0.05
0.10
0.15
0.20
0.25
Avail. Water Cont.


Partial Dependence
100000
99500
99000
98500
0.0
0.2
0.4
0.6
0.8
1.0
Percent Vegetated

Appendix E

Detailed Measures of Linear Multivariate Regression Model Performance

Model performance table is housed here.

Table E.1: Linear multivariate regression model fit summary by basin.

CDEC ID	No. Train Set	No. Test Set	R^2	RMSE	NSE	RSR	Mean PBIAS
AMA	18436	100	0.79	67797	0.68	0.57	31
AMF	18152	384	0.30	214659	0.42	0.76	351
AMK	18426	110	0.20	101450	-8.41	3.07	326
AMN	18404	132	0.53	68084	0.14	0.93	-1184
ANM	18158	378	0.07	127537	-440.13	21.00	14354
ANT	18180	356	0.04	138002	-959.48	30.99	16096
ASP	18152	384	0.01	136892	-6583.45	81.14	-493323
ASS	18152	384	0.22	132449	-42.63	6.61	-136761
ASV	18412	124	0.02	117560	-442.57	21.06	-78414
CSN	18152	384	0.53	71562	-0.68	1.30	-453
CYO	18367	169	0.16	64066	-34.26	5.94	-16912
DAV	18178	358	0.04	79212	-366.54	19.17	-46512
EFC	18152	384	0.07	121164	-18.42	4.41	-1126
ERS	18152	384	0.06	699226	0.00	1.00	1020
EWR	18152	384	0.08	71256	-46.65	6.90	552
FPL	18381	155	0.40	133254	0.57	0.65	43
FPR	18416	120	0.31	85009	-0.71	1.31	-8
FRD	18178	358	0.04	68408	-322.73	17.99	-13267
FTC	18429	107	0.17	67979	-20.71	4.66	2728
FTO	18152	384	0.20	336912	0.28	0.85	28
FTP	18418	118	0.18	113174	-14.92	3.99	-6008
KGC	18388	148	0.17	96328	-3.57	2.14	-1940
KGF	18152	384	0.29	152664	0.30	0.84	4

CDEC ID	No. Train Set	No. Test Set	R^2	RMSE	NSE	RSR	Mean PBIAS
KGP	18428	108	0.26	156939	0.29	0.84	43
KLO	18152	384	0.11	604578	-1.35	1.53	579
KRB	18152	384	0.54	82087	-0.27	1.12	247
KRI	18152	384	0.59	63159	0.20	0.89	106
KRK	18380	156	0.48	98852	-1.54	1.59	-430
KWT	18152	384	0.41	69501	-1.23	1.49	-16
LNV	18416	120	0.04	84574	-34.47	5.96	-685
MBS	18416	120	0.06	65652	-29.61	5.53	-1349
MDP	18380	156	0.27	94431	-1.39	1.54	-1283
MKM	18152	384	0.60	72546	0.18	0.91	-1187
MKW	18512	24	0.06	106706	-180.30	13.46	-10206
MRC	18152	384	0.58	95270	0.21	0.89	500
MSS	18152	384	0.45	77048	-0.51	1.23	1
NCD	18152	384	0.57	44949	-0.43	1.20	1778
NPH	18152	384	0.13	111523	-79.25	8.96	49064
OWL	18152	384	0.04	64680	-91.85	9.64	-451
PLK	18437	99	0.06	107592	-260.44	16.17	2835
PSH	18152	384	0.30	146260	0.04	0.98	63
RRH	18152	384	0.51	185711	-0.85	1.36	30289
SBB	18152	384	0.08	662161	0.02	0.99	-1
SCC	18152	384	0.21	89952	-23.01	4.90	2393
SCU	18405	131	0.11	112202	-49.43	7.10	66
SDT	18152	384	0.55	85349	-0.11	1.05	-333
SIS	18152	384	0.14	357235	0.21	0.89	13
SJF	18152	384	0.29	147602	0.37	0.80	141
SNS	18152	384	0.62	78131	0.57	0.66	505
SQS	18481	55	0.05	203412	-295.89	17.23	276
SRS	18152	384	0.86	53013	0.74	0.51	-64
SSP	18221	315	0.23	82325	-7.30	2.88	-38485
STB	18408	128	0.21	133246	-5.66	2.58	1140
SVC	18416	120	0.26	75141	-4.47	2.34	-867
SWH	18355	181	0.03	87966	-517.26	22.77	-243022
THT	18416	120	0.06	73510	-69.82	8.42	-1224
TLG	18152	384	0.34	142496	0.43	0.75	212
TLM	18418	118	0.23	94022	-3.95	2.23	-3527
TLN	18405	131	0.36	111728	-0.57	1.25	-3153
TNL	18152	384	0.77	87590	0.49	0.71	213
TRF	18152	384	0.39	58005	-0.82	1.35	-1173
WFC	18152	384	0.05	79888	-91.08	9.60	-1443
WWR	18152	384	0.15	69167	-7.46	2.91	76
YBG	18380	156	0.28	108706	-3.64	2.16	-770
YBJ	18416	120	0.10	111670	-36.88	6.15	-10361
YBM	18427	109	0.04	119752	-193.22	13.94	-16066

CDEC ID	No. Train Set	No. Test Set	R^2	RMSE	NSE	RSR	Mean PBIAS
YBS	18409	127	0.22	110385	-4.63	2.37	-3084
YCB	18383	153	0.06	137000	-140.76	11.91	-8585
YRS	18152	384	0.47	176506	0.37	0.79	-197

References

Abrahart, R. J., Heppenstall, A. J., & See, L. M. (2007). Timing error correction procedure applied to neural network rainfall–runoff modelling. *Hydrological sciences journal*, *52*(3), 414–431.

Abrahart, R. J., & See, L. (2000). Comparing neural network and autoregressive moving average techniques for the provision of continuous river flow forecasts in two contrasting catchments. *Hydrological processes*, *14*(11-12), 2157–2172.

Asefa, T., Kemblowski, M., McKee, M., & Khalil, A. (2006). Multi-time scale stream flow predictions: the support vector machines approach. *Journal of Hydrology*, *318*(1), 7–16.

Beaudette, M. D. (2016). Package 'sharpshootr'.

Beven, K. J. (2011). *Rainfall-runoff modelling: the primer*. John Wiley & Sons.

Bivand, R., Keitt, T., & Rowlingson, B. (2017). rgdal: Bindings for the 'geospatial' data abstraction library [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=rgdal` (R package version 1.2-15)

Bivand, R., & Rundel, C. (2017). rgeos: Interface to geometry engine - open source ('geos') [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=rgeos` (R package version 0.3-26)

Bivand, R. S., Pebesma, E., & Gomez-Rubio, V. (2013). *Applied spatial data analysis with R, second edition*. Springer, NY. Retrieved from `http://www.asdar-book.org/`

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123–140.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

California Department of Water Resources, Bay-Delta Office. (2016). Estimates of natural and unimpaired flows for the central valley of california: Water years 1922-2014.

Carlisle, D. M., Falcone, J., Wolock, D. M., Meador, M. R., & Norris, R. H. (2010). Predicting the natural flow regime: models for assessing hydrological alteration in streams. *River Research and Applications*, *26*(2), 118–136.

Chaney, N. W., Wood, E. F., McBratney, A. B., Hempel, J. W., Nauman, T. W., Brungard, C. W., & Odgers, N. P. (2016). Polaris: A 30-meter probabilistic soil series map of the contiguous united states. *Geoderma*, *274*, 54–67.

Dawson, C. W., & Wilby, R. (1998). An artificial neural network approach to rainfall-runoff modelling. *Hydrological Sciences Journal*, *43*(1), 47–66.

Dibike, Y. B., Solomatine, D., & Abbott, M. B. (1999). On the encapsulation of numerical-hydraulic models in artificial neural network. *Journal of Hydraulic research*, *37*(2), 147–161.

Dooge, J. C. (1986). Looking for hydrologic laws. *Water Resources Research*, *22*(9S).

Edmund, H. (2015). Package 'prism'.

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*(1), 25.

Forest Service, USDA, Pacific Southwest Region. (2006). Existing vegetation–vegetation classification and mapping for region 5.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Galelli, S., & Castelletti, A. (2013). Tree-based iterative input variable selection for hydrological modeling. *Water Resources Research*, *49*(7), 4295–4310.

Godsey, S. E., Kirchner, J. W., & Tague, C. L. (2014). Effects of changes in winter snowpacks on summer low flows: case studies in the sierra nevada, california, usa. *Hydrological Processes*, *28*(19), 5048–5064.

Govindaraju, R. S., & Rao, A. R. (2013). *Artificial neural networks in hydrology* (Vol. 36). Springer Science & Business Media.

Grubinger, T., Kobel, C., & Pfeiffer, K.-P. (2010). Regression tree construction by bootstrap: Model search for drg-systems applied to austrian health-data. *BMC Medical Informatics and Decision Making*, *10*(1), 1.

Grubinger, T., Zeileis, A., Pfeiffer, K.-P., et al. (2011). *evtree: Evolutionary learning of globally optimal classification and regression trees in r.* Department of Economics (Inst. für Wirtschaftstheorie und Wirtschaftsgeschichte).

Guisan, A., & Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological modelling*, *135*(2), 147–186.

Hijmans, R. J. (2016). raster: Geographic data analysis and modeling [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=raster` (R package version 2.5-8)

Hijmans, R. J. (2017). geosphere: Spherical trigonometry [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=geosphere` (R package version 1.5-7)

Hijmans, R. J., Phillips, S., Leathwick, J., & Elith, J. (2017). dismo: Species distribution modeling [Computer software manual]. Retrieved from

https://CRAN.R-project.org/package=dismo (R package version 1.1-4)

Hsu, K.-l., Gupta, H. V., Gao, X., Sorooshian, S., & Imam, B. (2002). Self-organizing linear output map (solo): An artificial neural network suitable for hydrologic modeling and analysis. *Water Resources Research*, *38*(12).

Hu, T., Wu, F., & Zhang, X. (2007). Rainfall–runoff modeling using principal component analysis and neural network. *Hydrology Research*, *38*(3), 235–248.

Iorgulescu, I., & Beven, K. J. (2004). Nonparametric direct mapping of rainfall-runoff relationships: An alternative approach to data analysis and modeling? *Water Resources Research*, *40*(8).

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 6). Springer.

Jarvis, A., Reuter, H. I., Nelson, A., Guevara, E., et al. (2008). Hole-filled srtm for the globe version 4. *available from the CGIAR-CSI SRTM 90m Database (http://srtm. csi. cgiar. org)*.

Klemes, V. (1982). Empirical and causal models in hydrology.

Levins, R. (1966). The strategy of model building in population biology. *American scientist*, *54*(4), 421–431.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomforest. *R News*, *2*(3), 18-22. Retrieved from http://CRAN.R-project.org/doc/Rnews/

Lin, J.-Y., Cheng, C.-T., & Chau, K.-W. (2006). Using support vector machines for long-term discharge prediction. *Hydrological Sciences Journal*, *51*(4), 599–612.

Louppe, G. (2014). Understanding random forests: From theory to practice. *arXiv preprint arXiv:1407.7502*.

Magnuson-Skeels, B. (2016). *Using machine learning to statistically predict natural flow.* MS Thesis.

McKay, L., Bondelid, T., Dewald, T., Johnston, J., Moore, R., & Rea, A. (2012). Nhdplus version 2: user guide. *National Operational Hydrologic Remote Sensing Center, Washington, DC*.

Minns, A., & Hall, M. (1996). Artificial neural networks as rainfall-runoff models. *Hydrological sciences journal*, *41*(3), 399–417.

Moriasi, D., Arnold, J., Van Liew, M., Bingner, R., Harmel, R., & Veith, T. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE*, *50*(3), 885-900. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-34447500396 (cited By 2311)

NRCS, USDA. (2006). Land resource regions and major land resource areas of the united states, the caribbean, and the pacific basin. *US Department of Agriculture Handbook*, *296*.

Pebesma, E. J., & Bivand, R. S. (2005, November). Classes and methods for spatial data in R. *R News*, *5*(2), 9–13. Retrieved from `https://CRAN.R-project.org/doc/Rnews/`

Petty, T., & Dhingra, P. (2017). Streamflow hydrology estimate using machine learning (shem). *JAWRA Journal of the American Water Resources Association*.

Pike, R. J., & Wilson, S. E. (1971). Elevation-relief ratio, hypsometric integral, and geomorphic area-altitude analysis. *Geological Society of America Bulletin*, *82*(4), 1079–1084.

Poff, N. L., Allan, J. D., Bain, M. B., Karr, J. R., Prestegaard, K. L., Richter, B. D., ... Stromberg, J. C. (1997). The natural flow regime. *BioScience*, *47*(11), 769–784.

R Core Team. (2017). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., ... others (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*.

Shortridge, J. E., Guikema, S. D., & Zaitchik, B. F. (2016). Machine learning methods for empirical streamflow simulation: a comparison of model accuracy, interpretability, and uncertainty in seasonal watersheds. *Hydrology and Earth System Sciences*, *20*(7), 2611.

Singh, V. P., & Frevert, D. K. (2005). *Watershed models*. CRC Press.

Sivapalan, M., Takeuchi, K., Franks, S., Gupta, V., Karambiri, H., Lakshmi, V., ... others (2003). Iahs decade on predictions in ungauged basins (pub), 2003–2012: Shaping an exciting future for the hydrological sciences. *Hydrological sciences journal*, *48*(6), 857–880.

Solomatine, D. P., & Ostfeld, A. (2008). Data-driven modelling: some past experiences and new approaches. *Journal of hydroinformatics*, *10*(1), 3–22.

Todini, E. (1988). Rainfall-runoff modelingpast, present and future. *Journal of Hydrology*, *100*(1), 341–352.

Tokar, A. S., & Johnson, P. A. (1999). Rainfall-runoff modeling using artificial neural networks. *Journal of Hydrologic Engineering*, *4*(3), 232–239.