Data-driven Methods for Hydrologic Inference and Discovery

By

Scott Campbell Worland

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Environmental Engineering

May 11, 2018

Nashville, Tennessee

Approved:

George M. Hornberger, PhD

Jonathan M. Gilligan, PhD

Robert M. Hirsch, PhD

Hiba Baroud, PhD

Ralf Bennartz, PhD

ACKNOWLEDGMENTS

Special thanks to my advisor and mentor, George Hornberger, for making the "existence, origin, movement, and course of waters and the causes which govern and direct their movement" less "secret, occult, and concealed" (quotes derived from an 1861 Ohio-Supreme Court case). Jonathan Gilligan, David Furbish, Tyler Doane, Chris Tasich, and Chelsea Peters were also influential in shaping my time and thoughts at Vanderbilt. I am very grateful to the many hydrologists at the U.S. Geological Survey (USGS) that have supported and encouraged me throughout graduate school. Specifically Bob Hirsch, Timothy Cohn, Bill Wolfe, Scott Gain, Mike Bradley, Rodney Knight, Jenny Murphy, Pierre Glynn, Julie Kiang, Will Farmer, and Stacey Archfield. Tim Cohn (1957-2017) was particularly kind and supportive. In 2014 I gave a talk at the USGS National Center in Reston, VA during the first year of my PhD. My talk was not very good. Possibly sensing my uncertainty, Tim–who is a renowned statistical hydrologist–approached me afterwards and invited me to join him and several others for lunch. From that lunch in 2014 until his death in 2017 Tim was a thoughtful mentor to me. Tim will have a long-lasting legacy both in the hydrologic community and in my own life. For an excellent tribute to Tim's life and work see https://eos.org/articles/timothy-a-cohn-1957-2017.

I am thankful that my siblings and I were able grow up on a piece of land in a rural community. The property has a karst spring that serves as the headwaters of the Chattanooga Creek watershed, Hydrologic Accounting Unit 10 code 0602000112. My parents encouraged us to spend most of our days outside. This constant exposure to the outdoors is largely responsible for my curiosity and love for the natural world–especially hydrology. Finally, I am thankful for my immediate family: Bonnie (spouse), NoraJean (5 yrs) and Harper (3 yrs). They are the best friends I have ever had.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Overview

Data-driven models (DDMs) map connections within a system without making assumptions about the physical behavior of the system. Hydrologists have expressed concerns about the usefulness of models that mostly disregard hydrologic theory [2, 3], yet DDMs have been widely used in hydrology due to their "unreasonable effectiveness" when applied to real-world problems [4, 5]. DDMs work because they approximate functions (e.g., a single-layer neural network is considered a "universal approximator" for a given range [6]). This suggests that a DDM can recreate the functional relationships that hydrologists have discovered through empirical research, physics, and mathematics.

For example, the specific energy (E) of frictionless channel flow of known specific discharge (q_w) and depth (h) can be obtained using the following equation (See Hornberger et al., page 94, for more details [7]),

$$E = \frac{q_w^2}{2gh^2} + h, \tag{1.1}$$

where g is the acceleration due to gravity ($9.8ms^{-1}$). This relationship is an example of hydrologic theory. Can a DDM discover the same relationship? Equation 1.1 was used to generate 90 realizations of E based on 90 values of h ranging from 0.1-1 with q_w set to $0.5m^2s^{-1}$. The values for E and h were then used to train a simple neural network with one hidden layer and 3 neurons (Figure 1.1, A). The model was able to learn the relationship (Figure 1.1, B) without "knowing anything" about the physical processes described in Equation 1.1. The calculated weights (written beside the lines connecting the neurons in Figure 1.1, A) do not help us understand the dynamics of frictionless channel flow, but the

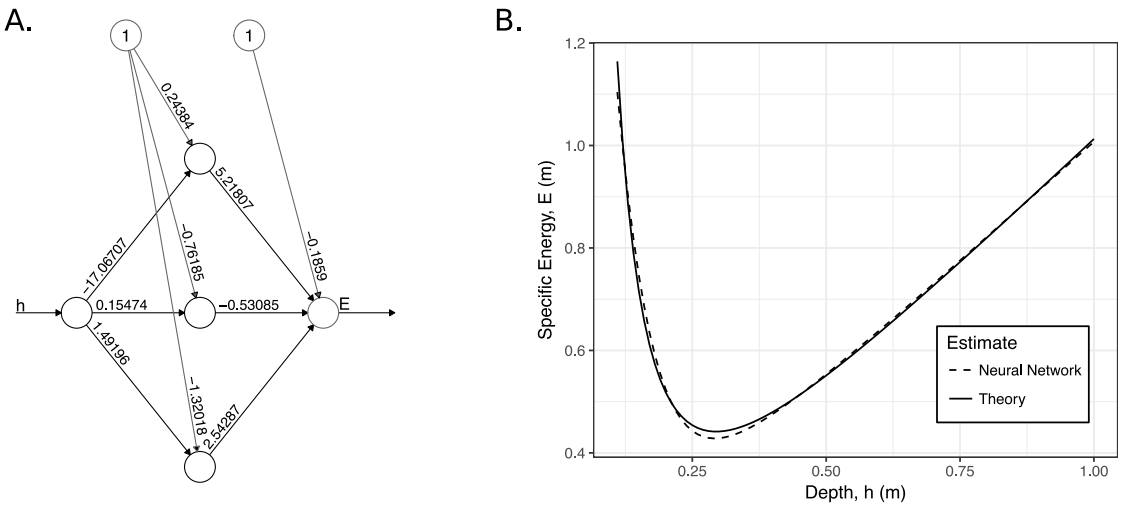

Figure 1.1: [A] Learned weights of simple neural network model using a sigmoid activation function with one hidden layer with 3 nodes. [B] Relationship between depth of frictionless channel and specific energy estimated by neural network model and a known theoretical relationship. Note, q_w was set to $0.5m^2s^{-1}$ across all depths.

DDM does provide a tool to predict E given h.

Hydrologic systems are not easily characterized by deterministic relationships between variables. Outputs of interest are often the result of complex-hydrologic processes occurring at various temporal and spatial scales. With sufficient data a DDM can learn abstract representations of the connections between hydrologic inputs and outputs. Due to this abstraction, DDMs are often considered useful only for prediction. A recent synopsis article by Vogel et al., 2015, however, lists "analyzing 'Big Data' using advances in the fields of statistics, machine learning, signal processing, and data mining" as one of the most important advances "needed to develop an in-depth understanding of the dynamics of the connectedness between human and natural systems and to determine effective solutions to resolve the complex water problems that the world faces today" [8]. How can we extend the capacity of DDMs to include "in-depth understanding?"

Theory-guided data science is an emerging paradigm that integrates scientific understanding and data-driven methods [9]. This integration can occur at multiple steps within

the DDM process and the resulting models are more generalizable and have greater physical interpretability than DDM models alone. Examples range from simple transformations of the response variable, to constrained optimization and learning hybrid theory and DDMs [9]. Leveraging the strength of both approaches can restrict the domain of possible models to ones that are (1) physically consistent and (2) produce accurate predictions. The DDM approach to predicting specific energy from the example above fails when applied to values outside of the range of the training data (Figure 1.2). Theory could improve the model by constraining predictions of E for depths greater than 1 meter to simply be 1:1[1].

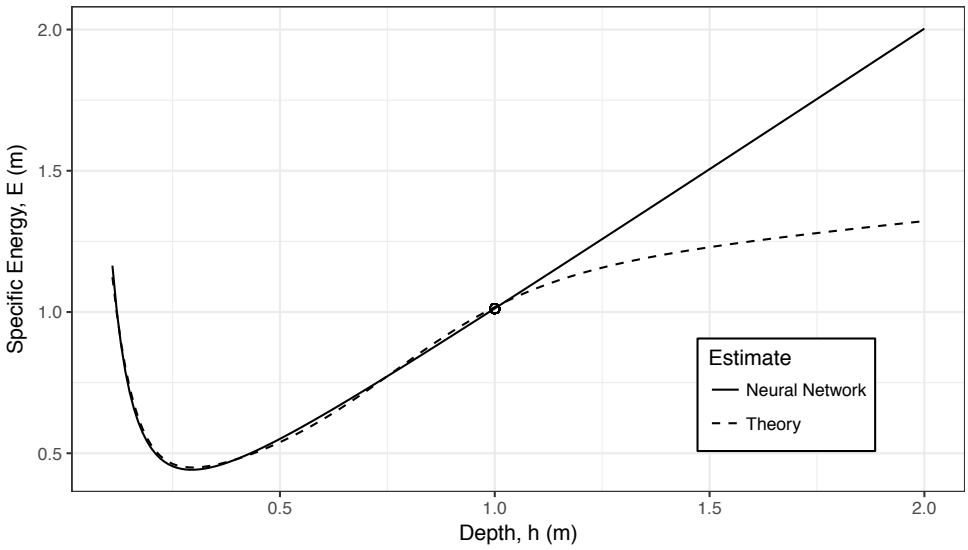

Figure 1.2: Extrapolation of neural network model outside of the range of data used to train the weights results in poor predictions.

1.2 Organization

Improving the predictions of streamflow in ungaged basins has been an multi-decadal objective for the international hydrologic community [10, 11, 12], but despite the rapid advances from this effort, there are still many unexplored questions. Chapters two and three of my dissertation are focused on improving hydrologic predictions in ungaged basins.

[1]This is obviously a toy example, because in reality we would just use Equation 1.1 directly across the entire range of depths.

Chapter four explores questions in sociohydrolgy–an emerging sub-discipline within the hydrologic sciences that seeks to integrate the physical and social aspects of hydrologic systems [13]. The broad research questions for these three chapters are the following,

1. Can machine-learning models improve predictions of a low streamflow statistic (7Q10) in ungaged catchments compared to "baseline" methods?

2. Can multi-output neural networks estimate physically consistent flow duration curves in ungaged catchments?

3. How do the drivers of municipal water use vary across the contiguous United States?

These research questions are explored using various data-driven methods. The second chapter compares the ability of 8 machine-learning models and 4 baseline models to predict a lowflow statistic in ungaged catchments. The mean squared error is decomposed to highlight the major-error components of each model. Additionally, weighted variable importance and partial-dependence plots are used to analyze further the relationship between variables and the hydrologic response. This work is published in Environmental Modeling and Software [14]. The third chapter uses theory-driven multi-output neural networks to estimate flow duration curves and daily streamflow in ungaged basins. The multi-output architecture leverages covariance between multiple quantiles to generate flow duration curves that ensure monotonicity for almost 100% of the estimates. This work is being prepared for submission. The fourth chapter uses hierarchical Bayesian models to explore the drivers of municipal water use across the U.S. The model-design is derived from socio-economic theory and results in an increased understanding of a coupled human-hydrologic system. This work is published in Water Resources Research [15].

CHAPTER 2

IMPROVING PREDICTIONS OF HYDROLOGICAL LOW-FLOW INDICES IN UNGAGED BASINS USING MACHINE LEARNING

2.1 Abstract

We compare the ability of eight machine-learning models (elastic net, gradient boosting, kernel-K-nearest neighbors, two variants of support vector machines, M5-cubist, random forest, and a meta-learning ensemble M5-cubist model) and four baseline models (ordinary kriging, a unit area discharge model, and two variants of censored regression) to generate estimates of the annual minimum 7-day mean streamflow with an annual exceedance probability of 90% (7Q10) at 224 unregulated sites in South Carolina, Georgia, and Alabama, USA. The machine-learning models produced substantially lower cross validation errors compared to the baseline models. The meta-learning M5-cubist model had the lowest root-mean-squared-error of 26.72 cubic feet per second. Partial dependence plots show that 7Q10s are likely moderated by late summer and early fall precipitation and the infiltration capacity of basin soils.

2.2 Introduction

Water managers rely on streamflow data to allocate water resources, define the dilution potential of catchments, set ecological streamflow limits, and ensure sustainable watershed planning [16, 17, 18]. However, many streams do not have observed streamflow data and water managers must depend on the streamflow estimates from various prediction models [19, 16, 20]. Improving the predictions of streamflow in ungaged basins has been a primary objective for hydrologists for decades and international initiatives have resulted in rapid advances in this field [10, 11, 12]. The two primary modeling strategies for predicting streamflow response in ungaged basins are: (1) deterministic physically based models–

i.e. calculating streamflow based on distributed hydrologic parameters, and (2) statistical regionalization–i.e. using regression models to transfer hydrologic information from gaged to ungaged basins [16, 21]. This current paper focuses on the statistical regionalization of a low streamflow statistic: the annual minimum 7-day mean streamflow with an annual exceedance probability of 90% (7Q10).

A stream's "low flow" refers to the amount of water flowing in a stream during prolonged periods of little to no rainfall during an average non-drought year. The low-flow regime for a particular stream is controlled by the physical characteristics of its basin and the local climate [22]. The 7Q10 statistic describes a basin's expected low-flow and provides a way to compare directly the low-flow regimes of different basins. This statistic is commonly used to determine permitted point-source pollutant levels in streams [23]. There are a number of other important low-flow metrics not discussed in this paper; several examples are the 7Q10 for a particular season or month, the annual minimum 7-day mean streamflow with an annual exceedance probability of 50% (7Q2), mean annual minimum, median September streamflow, and ecologically derived values [17, 24, 25, 26]. The contribution of this research is the comparison of statistical estimation techniques; the choice of the specific response variable would not change the structure of the analysis but we cannot conjecture how specific models would perform for a different target variable.

Low-flow regionalization methods attempt to predict low-flow metrics in ungaged basins by leveraging the correlation between basin characteristics and streamflow at gaged basins [16]. The primary goal of 7Q10 regionalization is accurate predictions and not mechanistic explanations of what controls the 7Q10, and this distinction between prediction and explanation should guide the statistical analysis [27]. Regardless of outcome goal or the type of model used, all hydrologic models require assumptions. Deterministic models, for example, assume that the physical relationships between parts of a hydrologic system are adequately captured by a set of static functions and decision rules, while stochastic models may depend on assumptions about the probabilistic constraints on parameters (i.e.

"priors"), the choice of the likelihood and cost functions, the numerical methods used for parameter estimation (e.g., gradient descent, maximum likelihood, numerical integration, etc.), and choices about data preprocessing and transformation. Furthermore, hydrologic models often assume some level of stationarity [28]. These assumptions can have significant effects on the applicability of model results, and researchers must acknowledge how their model design choices propagate into conclusions drawn from the model.

This paper evaluates the predictive performance of various association-based models (e.g., linear regression models) that leverage the covariance structure between variables to make inferences and predictions. Association-based models have proved to be a useful engineering tool for predicting 7Q10s, and have become increasingly sophisticated in the last 30 years [11]. Regression methods have evolved from simple ordinary least squares [29, 30, 31] to time series weighted least squares [32], generalized least squares (GLS) [33], censored regression [34], two step GLS-logistic regression [35], truncated models, and catchment clustering methods [36]. There has also been an increased application of geostatistical low-flow regionalization methods–primarily ordinary kriging, top kriging, and physiographical space-based interpolation [37, 38].

Despite the recent methodological advances mentioned above, few studies have explored machine-learning methods to predict low-flow metrics in ungaged basins. [39] used an ensemble of artificial neural networks for predicting various low-flow metrics in Canada, [40] used regression trees to predict Q95s in Austria, [41] used model tree ensembles to predict a complete flow-duration curve (FDC) for streams in Illinois and Texas, and [42] used random forest models to predict several components of a FDC in New Zealand. These studies contributed valuable baseline assessments of the applicability of machine learning to streamflow-statistic estimation. Yet, however, they compare only 2-3 estimation techniques, each using a unique data set–a practice that confounds direct comparison of model performance between individual studies.

In this paper, twelve different modeling methods were applied to a publicly available

data set [43], and the multi-model comparison approach presented by [44, 45] and [46] was used to determine the predictive performance of the models using multiple assessment criteria. Several machine-learning techniques were introduced–gradient boosting machines, kernel-K-nearest-neighbors, and elastic net-that, to our knowledge, have not yet been used to predict low-flow statistics. A meta-learning M5-Cubist model was also introduced that minimizes the overall generalization error by combining the cross-validated predictions of each machine-learning model. Finally, hydrologic insights to the physical controls of low streamflow were explored through a discussion of the relative importance of predictor variables and their corresponding partial-dependence functions for each model. The novelty of this contribution is the use of multiple machine-learning models, the introduction of meta-modeling approaches for the regionalization of low-streamflow statistics, the comparison with models historically used to estimate 7Q10s, and the large gains in predictive accuracy over historical methods.

Research Objectives and Major Findings

This paper provides the 7Q10 prediction performance estimates of twelve statistical estimation techniques–four "baseline" methods (type I Tobit regression, region of influence type I Tobit regression, ordinary kriging, and an average unit-area discharge null model) and eight machine-learning models: (1) M5-cubist regression trees, (2) gradient boosting machines, (3) kernel-K-nearest neighbors, (4) random forests, (5) elastic net, support vector machines with a (6) polynomial kernel and a (7) radial basis function kernel and an (8) ensemble meta-learning M5-cubist model is also explored. The specific research objectives are,

1. Use leave-one-out cross validation (LOO-CV) to simulate the prediction of 7Q10s at ungaged sites in three states in the southeast U.S. using eleven estimation techniques.

2. Compare the predictive accuracy of each model using root mean squared error (RMSE),

unit area root mean squared area (UA-RMSE), median percentage error (MPE), and the Nash-Sutcliffe efficiency coefficient (NSE), and decompose the RMSE to examine what is controlling the error for each model.

3. Discuss the relative importance and partial dependence functions of predictor variables for each model.

We found that machine-learning methods can produce more accurate predictions of 7Q10s in ungaged basins than baseline models. Variable importance measures and partial dependence plots suggest that 7Q10s are partially driven by landcover, late summer and early fall precipitation, the infiltration rate of soils, and the variability of minimum and maximum monthly temperatures.

Background of Machine Learning in Hydrology

Machine learning–also referred to as statistical learning, data-driven modeling, and computational intelligence–refers to a set of statistical methods that are optimized for predictive performance through a cross-validated parameter tuning process [47, 48]. These methods have been called black-box approaches and criticized for having little connection to the underlying physical processes being modeled (See references in [44] and [3] for examples of these critiques in hydrology). Regardless, machine-learning techniques have become prevalent in the hydrology literature. Artificial neural networks have been used for predictions in hundreds of water-resource studies [49, 50, 51, 52]. Random forest models have been used to predict natural and altered streamflow regimes in ungaged basins [53, 54, 55]; support vector machines have been used to forecast monthly streamflow [56, 57] and to downscale low-flow indices [58]; genetic algorithms have been used to calibrate rainfall-runoff models [59]; Instance-based methods (e.g., K-nearest neighbors) have been used to forecast daily streamflow [60]; and M5-cubist models have been used for low-flow forecasting [61], flood forecasting [62], and monthly streamflow forecasting

[46, 63, 64]. Various ensemble methods have also been used for prediction in hydrology. Two examples are artificial neural network ensembles applied in flood-frequency analysis [65], and the prediction of monthly streamflow using ensemble methods for support vector machine and regression trees [66]. Historically, machine-learning models have been best suited for modeling tasks that are concerned with accurate predictions and not physical interpretability [67]. However applied researchers are exploring new methods to extract knowledge and gain domain-specific insights from data-driven models [68].

2.3 Material and methods

2.3.1 Study Site and Data

Models were developed using 7Q10 values from a total of 224 basins: 45 basins in South Carolina [69, 70, 71, 72, 73], 68 basins in Georgia [74], and 111 basins in Alabama [75] (Figure 2.1, Table 2.1). Predictions in this paper were based on 7Q10s calculated as of 2015 and may not reflect updated or forecasted 7Q10 values.

Table 2.1: Summary statistics for the 7Q10 and several basin descriptors.

Variable [units]	min	mean	max	σ
7Q10 [ft^3s^{-1}]	0.0	42.7	807	95.3
basin area [km^2]	10.3	1023.0	15400.0	1847.0
mean annual precip [cm]	119.0	140.7	194.4	15.7
mean annual temp [°C]	12.15	16.3	19.65	1.57
mean elevation ASL [m]	14.4	195.90	885.7	147.8

The basins were selected from studies that estimated 7Q10 values for near-reference conditions (i.e. basins unregulated and unlikely to be altered given associated measures of development). Over 230 independent variables from the Gages II data set [43] were originally considered as predictor variables. The full Gages II data set consists of basin characteristics for 9,322 streams within the U.S. that have at least 20 years of complete streamflow record since 1950 or streams that have been active since water year 2009. The

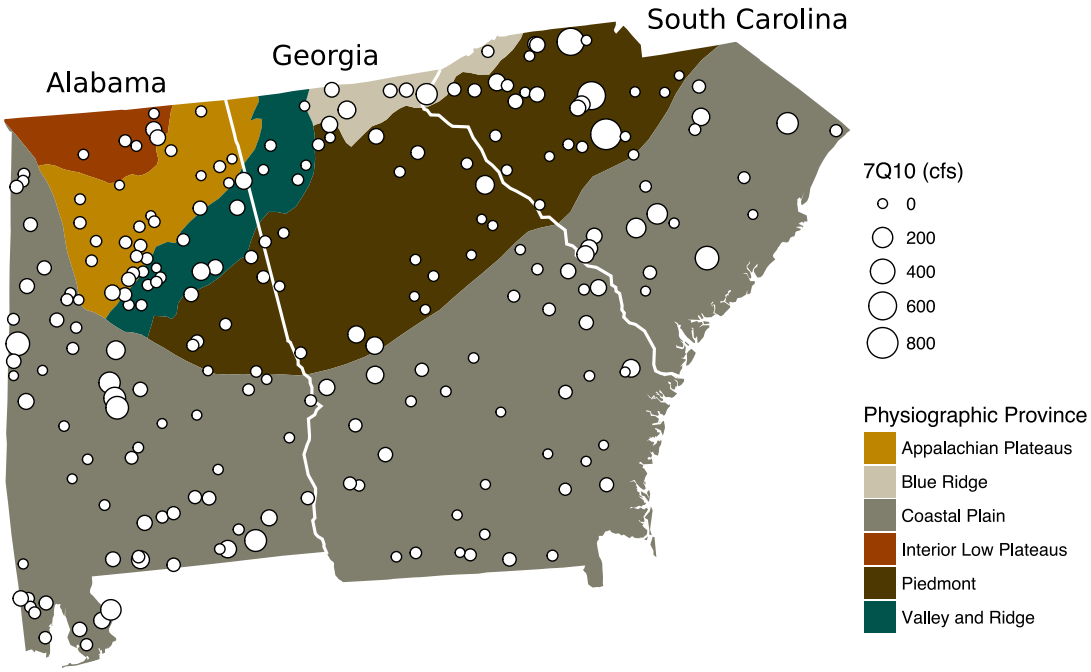

Figure 2.1: Map of streamgages within study area where the size of the symbol represents the 7Q10 value calculated for each gage.

U.S. Geological Survey (USGS) defines a water year as the 12-month period October 1, for any given year through September 30, of the following year. Several of the variables represented measures of regulation in the upstream basins. The focus in this current analysis was unregulated streamflow, so these variables were removed. The method to explore and eliminate variables in this study closely follows that of [76], which left 125 variables for our analysis. The selected variables are listed in [77], and are defined in the supplemental material. Each of the independent variables were standardized by subtracting the mean of the variable and dividing by the standard deviation of the variable prior to model development. Values of 7Q10s can span several orders of magnitude and so the response variable was also transformed:

$$y_i = ln\left(\frac{7Q10_i + 0.001}{DA_i}\right), \tag{2.1}$$

where y_i is the transformed response variable for site i, $7Q10_i$ is the 7Q10 for site i, and DA_i is the drainage area for site i. Because 13 of the sites had 7Q10 values of zero, 0.001 was added to each 7Q10 value to avoid having infinite values. The predicted 7Q10 values were converted back to natural space (cfs, cubic feet per second) by simple algebraic manipulation of Equation 2.1. The response variable transformation for the Tobit models was slightly different than what is presented in equation 1, for reasons particular to censored regression, and details are presented in Section 2.3.4.

2.3.2 Software/Data Availability

All of the analysis was done in the R Language and Environment for Statistical Computing [78] and the required packages for each model are listed within the individual model description section. The input data and R model archive can be accessed in [77].

2.3.3 Tuning Parameter Selection for Machine-Learning Models

A tuning parameter is any free parameter in a model that is provided by the user, and tuning parameters are indicated by italics in the model descriptions below. Model tuning, also referred to as hyperparameter optimization or model training, is the process of searching for values of model parameters that optimize a predefined loss function (e.g., the RMSE). The cross-validated RMSE was used for all tuning parameters for each model and is an arbitrary design choice that ensured consistency between models. We took a two-step model tuning approach for the machine-learning models: (1) 30 initial points were generated in hyperparameter and RMSE space using a simple random search across possible values of hyperparameters, and (2) a Bayesian optimization of the hyperparameters was conducted using a Gaussian process prior and initial points from the random search.

Bayesian optimization attempts to select optimal hyperparameters by treating the relationship between hyperparameter values and the RMSE as an unknown function to be minimized (i.e. the negative RMSE is used for maximization). A Gaussian process model

describes this function by constructing a posterior distribution of functions. The posterior distribution improves as the number of samples from the hyperparameter space grows, and the algorithm becomes more certain of the regions in hyperparameter space that are worth further exploration [79]. In this study, the Gaussian process model was updated for 15 steps and the final model was selected based on the combination of parameters that produced the smallest leave-one-out cross-validation (LOO-CV) RMSE value (Table 2.2). The hyperparameters were tuned using the rBayesianOptimization R package [80]. The optimization function was parameterized with the Matérn 5/2 kernel and the Expected Improvement or Upper Confidence Bound acquisition functions based on the recommendations in [79]. The only exception to this tuning process for the machine-learning models was for the kernel-K-nearest neighbor (KKNN) model, where grid search was used instead of Bayesian optimization because the latter required numeric hyperparameters and the kernel hyperparameter of the KKNN model is a text string. Grid search involves an exhaustive search through a user-defined subset of hyperparameter space [47].

The number of predictor variables used to build the models can also be considered a free parameter. We did not explicitly tune the number of predictor variables, but allowed the machine-learning models to potentially use all 125 variables. The hyper-parameter tuning scheme will naturally avoid overfitting; i.e, the final model architecture is chosen by selecting the combinations of hyperparameters that produce the lowest cross-validated prediction error, thus rejecting model architectures that overfit to the training data. Tuning of the baseline models is described separately in each section.

2.3.4 Baseline Models

We classified multivariate regression and geostatistical techniques as "baseline" models. This classification scheme was used as a way to compare groups of models and does not reflect the complexity, accuracy, or robustness of the method. Four baseline models and the tuning associated with each of those models is described below.

Table 2.2: Tuning parameters for each model. GBM: gradient boosting machine, KKNN: kernel-K-nearest neighbors, RF: random forest, SVMP: support vector machine with a polynomial kernel, SVMPG: support vector machine with a Gaussian kernel, OK: ordinary kriging, ROI-Tobit: region-of-influence Tobit.

Model Family	Model Name	Tuning Params.	Param. Values
Regularized	elastic net	α	0.0
		λ	1.374
Tree-based/ boosting	GBM	shrinkage	0.067
		interaction depth	15
		min obs in node	14
		ntree	439
Instance based	KKNN	neighbors	5
		kernel	triangular
		distance	0.25
Tree-based	M5-cubist	committees	50
		neighbors	8
Tree-based/ bagging	RF	mtry	116
		ntree	500
Support Vectors	SVMP	cost	1
		kernel	polynomial
		degree	2
		scale	0.0025
Support Vectors	SVMG	cost	1.51
		kernel	gaussian
		sigma	0.0044
Tree based	meta-M5 cubist	committees	50
		neighbors	6
Gaussian process	OK	variogram	spherical
Censored	ROI-Tobit	predictor number	8
		number of sites	205
Censored	full-Tobit	predictor number	8

Type I Tobit Model

A left-censored Tobit regression model (full Tobit) was used as the baseline 7Q10 prediction [81, 34]. Left-censored regression is useful for situations where the response variable cannot be observed below a certain value (possibly due to measurement sensitivity), but the predictor variables are known for every observation. Tobit models are frequently used to develop regionalization equations for low flow statistics when the streamflow statistic can be equal to zero [82, 83, 84, 85]. The model can be written as,

$$\hat{y}_i = \left\{ \begin{array}{ll} x_i^T \beta + \varepsilon_i & if \;\; y_i > y* \\ y* & if \;\; y_i =< y* \end{array} \right\} \tag{2.2}$$

where y_i is the response value for observation i, x_i are the values of predictor variables for observation i, β is a vector of regression parameters, ε_i is the unexplained variance for a observation i, and $y*$ is the censoring value. Basins with 7Q10 values equal to zero (n=13) were set to 0.001, the response was transformed using the natural log, and ln(0.001) was used as the censoring value. The natural log of the drainage area was then included as a candidate predictor. We use this response transformation rather than the unit area 7Q10 (Equation 2.1) because it (1) provides a unique censoring value of ln(0.001), and (2) produces better predictions for the Tobit models than Equation 2.1. A Tobit model has the potential to overfit when a large number of predictor variables are included in the model. To mitigate this, forward stepwise selection was used to select explanatory variables within LOO-CV. The final model with the lowest RMSE included 8 predictor variables.

Region-of-Influence Tobit (ROI-Tobit)

The region-of-influence method builds a regression model for a particular site using only a subset of the full data set [86, 87, 88, 36, 85]. In this study, the sites included in the subset were selected based on their similarity to the site of interest, where similarity was measured by Euclidean distances in predictor space. The optimal *number of sites* and

predictor number designated as the region of influence was found using LOO-CV (Figure 2.2).

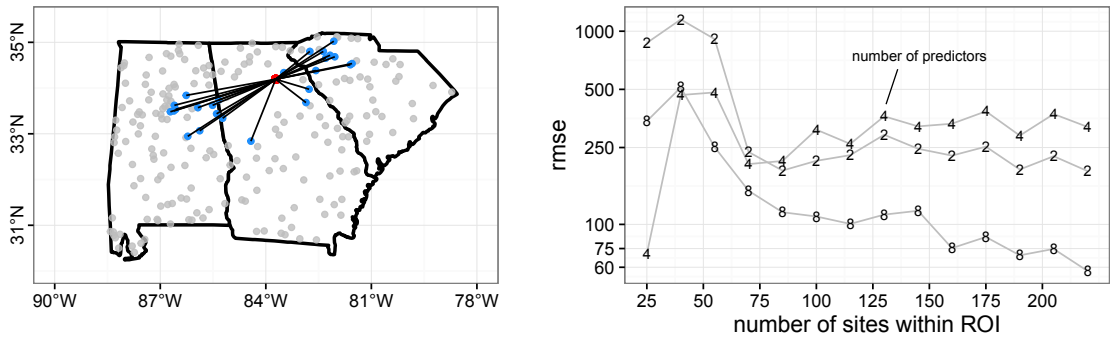

Figure 2.2: *Left*: Example of the region of influence for a basin (red) where the lines connect the basin to its 25 nearest neighbors (blue) calculated by the Euclidean distances in predictor space. *Right*: Example of tuning the number of sites and predictor variables using leave-one-out cross validation for the region-of-influence Type I Tobit model.

Ordinary Kriging

Ordinary kriging is a geostatistical tool that uses the distance between two points to predict the semivariance of a dependent variable [89]. The inter-site semivariances of data from a measured network can be used to create a system of linear equations predicting the semivariance at an unmeasured site to be a weighted, linear sum of the semivariance between all observed sites. For an unmonitored site, these same weights can be used to estimate the unknown quantity on which the semivariances were based. If all the assumptions of ordinary kriging are valid, this tool provides the best linear unbiased estimate. In this study, a spherical model was used to represent the semivariance between 7Q10s. Other hydrologic applications [90, 91] have found success with spherical models, and we used cross validation to confirm that the choice of a model form did not have a substantial impact on the prediction, which is consistent with previous research [91].

Null Model

A null prediction model was created where the 7Q10 prediction for a site was calculated as the left-out mean of value of the unit 7Q10s multiplied by the drainage area for the site:

$$\hat{y}_i = \left[\frac{1}{N} \sum_{j=1}^{N} \frac{7Q10_j}{A_j} \right] \times A_i \tag{2.3}$$

where $\hat{y}_i$ is the prediction for site i, $\frac{7Q10_j}{A_j}$ is the unit 7Q10 value for every site but the site of interest, and A_i is the drainage area for the site of interest. Equation 2.3 can be rewritten as a one parameter single variate regression model where drainage area is the only predictor.

$$\hat{y}_i = 0 + \beta A_i \tag{2.4}$$

2.3.5 Machine-Learning Models

It is rarely possible to make meaningful a priori distinctions between learning algorithms for a given data set [92]. Therefore, it is desirable to select a range of initial models with distinct functional differences (ie. models from a range of "model families") to increase the likelihood of discovering a well-performing model. Each model in this study was fit to the data, and the most promising models were further fine-tuned to achieve optimal performance. We include a brief section describing each model below. We begin each section with a general description of how the model relates to the hydrologic task of predicting 7Q10 values in ungaged basins. We then provide further details describing the specific mechanics of each model.

Elastic Net

General overview: Elastic-net models address overfitting by preventing parameters from inflating in response to a basin with an anomalously large 7Q10. From a hydrologic perspective, this results in out-of-sample predictions with reduced variance but potentially

higher bias than a non-regularized regression model. This can lead to better predictions for sites with large 7Q10 values.

Further details: Elastic-net models are produced by a regularized regression method that combines the two penalties from least absolute shrinkage and selection operator (LASSO) regression and ridge regression [93]. Regularized regression methods (also referred to as shrinkage or penalized regression) provide a less complex model with better fit by including a penalization parameter in the loss function of least squares that shrinks the slope coefficients towards zero [47]. Ridge regression uses a squared penalty in the loss function, which shrinks the parameter estimates towards zero, whereas LASSO regression uses an absolute value penalty in the loss function, which results in some coefficients being set to exactly zero. From a Bayesian perspective, ridge regression is equivalent to assigning a zero-mean normally distributed prior distribution on the parameter vector, and LASSO regression is equivalent to assigning a zero-mean Laplace prior distribution on the parameter vector. Elastic-net is a blend between the two. The loss function for an elastic net model can be written as,

$$\hat{\beta} = \underset{\beta}{\textbf{argmin}} \parallel y - X\beta \parallel^2 + \lambda \left[(1-\alpha) \parallel \beta \parallel^2 /2 + \alpha \parallel \beta \parallel \right], \tag{2.5}$$

where $\hat{\beta}$ is the vector of regression coefficients, y is the response vector, X is the predictor matrix, α is a hyperparameter that serves to "bridge the gap" between LASSO regression and ridge regression, where $\alpha = 1$ results in LASSO and $\alpha = 0$ results in ridge regression, and λ controls the overall strength of the penalty [47]. The elastic net model was fit using the glmnet R package [94].

Gradient Boosting Machine (GBM)

General overview: The gradient boosting algorithm implemented here uses a regression tree as a base learner. A regression tree is a simple rule-based method–basically a flow

chart generated analytically to locate sites with similar basin characteristics. The model then generates predictions by taking the average 7Q10 values of sites that fall within the same nodes of a tree. Gradient boosting takes this a step further by using the model error from a single regression tree to iteratively build models that make better predictions. For example, if a site that falls within a node receives a poor prediction, the algorithm generates a secondary model that tries to predict the residual of the base regression tree. From a physical perspective, this is a way to capture non-linear relationships between 7Q10s and basin characteristics.

Further details: A gradient boosting algorithm uses the residuals (i.e. the gradient) from a base model to subsequently fit new models that are then added to the base model [95]. A regression tree is often used as the base model. For example, a tree with a specified number of terminal nodes (*interaction depth*) is fit to data, its residuals are calculated, and a second tree is built using the residuals from the first tree as the response variable, and the predictions from the second tree are added to the predictions from the first tree resulting in a new model. This process is repeated a certain number of times specified by the user (the *number of trees*). A *shrinkage* parameter (ranging between zero and one) can be used to control the fraction of the new prediction added to the previous model. For regression trees, there is also an additional parameter that restricts the minimum number of observations that must be within each node (*min obs in node*), which can reduce the overall variance in the model. In this study, the GBM model was fit using the gbm R package [96].

K-Nearest-Neighbors (KNN)

General overview: KNN models leverage the proximity of basins in predictor space to predict 7Q10 values. That is, basins with more similar basin characteristics (predictor variables) are considered to be "near" each other. To predict a 7Q10 for a new basin, the algorithm determines the K most-similar basins to the one of interest, and assigns the mean

7Q10 for the K most-similar basins as the prediction at the site of interest [87]. A variant of KNN was used in this study that involved transforming the predictor variables (via a kernel function) to allow the discovery of non-linear relationships.

Further details: The distance between samples can be measured using the Minkowski distances, which is calculated by,

$$\left(\sum_{j=1}^{p} |x_{aj} - x_{bj}|^q \right)^{\frac{1}{q}} \tag{2.6}$$

where p is the number of predictors, x_{aj} and x_{bj} are observations in predictor space, and q is passed to the model as a *distance* parameter. When $q = 2$, the Mikowski distance is simply the Euclidean distance. The predicted value is the average value of the response for a given number of nearest *neighbors* in predictor space. A *kernel* (referred to here as kernel-K-nearest-neighbors, KKNN) can be used to transform the predictors prior to calculating the distances, and has been shown to increase the prediction accuracy of the model [97]. The KKNN model was fit using the kknn R package [98].

M5-Cubist

General overview: Similar to GBM, KNN, and region of influence regression models, cubist models subset groups of basins (via a regression tree method) that have similar basin characteristics and makes predictions based off the subset. Cubist models, however, have two features that improve predictions: (1) they weight the out-of-sample predicted 7Q10 value for a particular basin by the in-sample model performance on basins that are close to the basin of interest (i.e. close in a nearest-neighbor sense), and (2) they use a linear regression in the terminal nodes of a tree rather than the mean to make predictions.

Further details: A M5-cubist model is a type of regression tree [99, 100]. The predictor space is partitioned through a set of recursive binary splits and prediction of the target vari-

able is based on values of the features contained within the partitions. The individual splits are chosen based on a greedy algorithm that seeks to minimize prediction error of possible subsets using only the branch of the subtree where the model is making a split. The major difference between a simple regression tree and a M5-cubist model is how the models make predictions within the nodes. A regression tree produces a single-value prediction for each node, whereas a M5-cubist model produces a prediction using a linear regression model [101]. The regression model in each node is built using only the predictor from the split directly above the node. The final prediction from a single tree is based on the regression model in the terminal node, but can be smoothed using a weighting scheme based on predictions from an arbitrary number of nodes within the subtree. The number of nodes used in the smoothing process is referred to as *neighbors*. The predictions from single-tree M5-cubist model can be improved using a boosting-like ensemble method where subsequent trees are built using the residuals from the single tree. The number of trees used in the ensemble is referred to as *committees*. In this study, the M5-cubist models were fit using the Cubist R package [102].

Random Forest (RF)

General overview: RF models combine the results of multiple regression trees to predict 7Q10 values. RF models differ from other regression tree-based methods because they rely on random sampling to describe persistent relationships between 7Q10s and basin characteristics. For example, if a particular site's 7Q10 is highly correlated with a certain basin characteristic, but most of the other sites do not show the same level of correlation, then an RF model will "sacrifice" a good prediction for that particular site to avoid overfitting. This is accomplished by the repeated random sampling of 7Q10s and basin characteristics.

Further details: RF models aggregate individual regression trees to reduce variance and improve prediction accuracy [103]. Observations are randomly sampled from the training

set, a individual regression tree is built using the random sample, predictions are made for the remaining observations (i.e. out-of-bag samples), and this process is repeated a certain number of times (*ntree*). Randomness is further added by forcing each tree to consider different randomly selected sets of predictor variables (*mtry*) at each split in order to reduce overall variance by lessening the strength of correlation between trees. This results in a bootstrapped aggregation of models (referred to as "bagging") that is almost always more accurate than its constituent models [47]. In this study, the random forest models were fit using the randomForest R package [104].

Support Vector Machines (SVM)

General overview: The physical interpretation of an SVM is similar to the interpretation for the elastic-net model, however, using a kernel function for the predictor variables allows the SVM to discover non-linear relationships between 7Q10s and basin characteristics. The "support vectors" are observations that have the most influence on the regression and are given weight over other observations.

Further details: Support vector machine regression models fit a regression line using only the data points (i.e. the "support vectors") that fall outside of a user-defined threshold (denoted as ε). The residuals outside of the threshold contribute a linear-scaled amount to the model fit and residuals within the threshold do not contribute to the model fit. Hence, SVM regression is considered an ε-insensitive regression. The effect that the large residuals have on the regression is controlled by a *cost* parameter, which can be shown to have a regularizing effect much like ridge regression. A kernel function is often used to extend an SVM to nonlinear regression [105]. Different *kernels* have different effects on the model predictions. In this study, two SVMs were built–one with a Gaussian kernel and one with a polynomial kernel. For the polynomial kernel, the degree of the polynomial (*degree*) was provided. An additional *scale* parameter can be provided that controls how close observa-

tions are in kernel space. In this study, the support vector machine models were fit using the kernlab R package [106].

2.3.6 Stack Generalization Model (meta-M5 cubist)

General overview: Meta models–also referred to as ensemble models–generate predictions by combining the output of multiple-first-order models [107, 108]. First-order models are referred to as level-0 and the meta-model is referred to as level-1.The models do not need to have a similar structure. For example, imagine a linear regression model, a regression tree, and a nearest neighbor model were used to independently predict the value of some observation. A simple "meta-model prediction" could be calculated by taking the mean prediction of the three level-0 models. We expand this idea by using a stacked-regression model that uses regression to combine the predictions from the level-0 models.

Further details: The stacked-regression model employed here combines the LOO-CV predictions from all the level-0 models using a level-1 M5-cubist model (Figure 2.3), a technique similar to bagging used in RF models and boosting used in GBMs. Combining the LOO-CV predictions from a suite of different models can reduce variance and increase prediction accuracy [107, 108]. The level-0 models are the 7 machine learning and 4 baseline models described above, and the level-1 model is the M5-cubist that uses only the predictions from the level-0 models as predictor variables. Prior to building the level-1 M5-cubist model, the mean and median predictions for each model were added as predictor variables for the level-1 model. The stacked model used only the LOO-CV predictions from each level-0 model and the unweighted mean and median predictions across each all models (i.e. it does not use the basin characteristics).

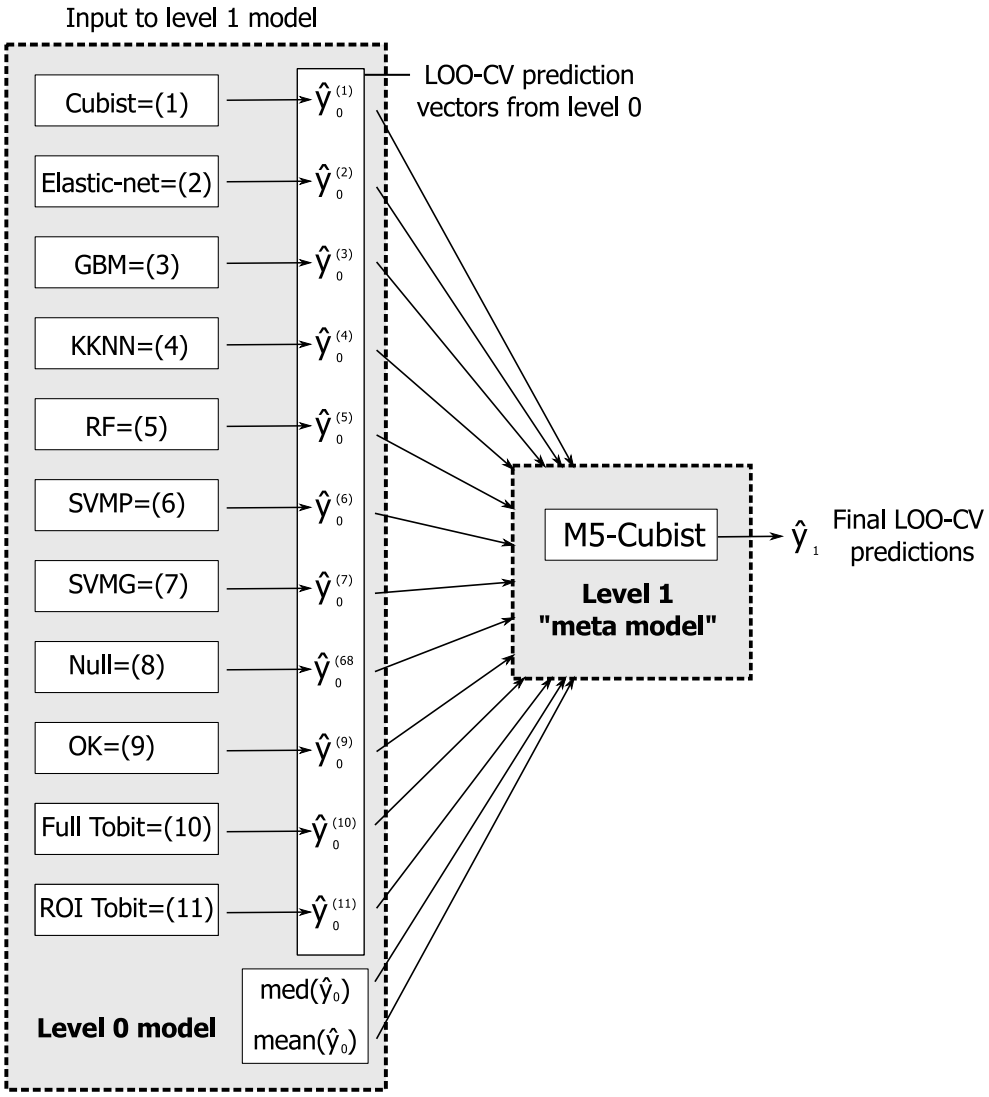

Figure 2.3: Schematic of the stacked generalization ensemble model used in this study. GBM: gradient boosting machine, KKNN: kernel-K-nearest neighbors, RF: random forest, SVMP: support vector machine with a polynomial kernel, SVMG: support vector machine with a Gaussian kernel, OK: ordinary kriging, ROI Tobit: region-of-influence Tobit

2.3.7 Error Metrics and Error Decomposition

The models were all evaluated using LOO-CV–an observation was removed from the data set (i.e. left out), the model was built on $n-1$ observations, and the left-out observation was predicted for each model. This was iteratively done for each observation. The LOO-CV predictions from each model were evaluated using four error metrics: The root mean squared area (RMSE),

$$RMSE = \sqrt{\frac{1}{N}\sum(y-\hat{y})^2}, \tag{2.7}$$

the unit-area RMSE (UA-RMSE),

$$UA-RMSE = \sqrt{\frac{1}{N}\sum(\frac{y}{DA}-\frac{\hat{y}}{DA})^2}, \tag{2.8}$$

the median percent error,

$$MPE = median(|(\hat{y}-y)/y|)*100, \tag{2.9}$$

and the Nash-Sutcliffe efficiency coefficient (NSE),

$$NSE = 1-\frac{\sum(\hat{y}-y)^2}{\sum(y-\bar{y})^2}. \tag{2.10}$$

Where y are the observed values, $\hat{y}$ are the estimated values, N is the total number of sites, DA is the drainage area, and $\bar{y}$ is the grand mean. For the MPE, y and $\hat{y}$ are only for sites where $y > 0$. The MPE, unit area RMSE, and NSE provide a measures of model performance relative to the size of the observed value (MPE), the drainage area (unit area RMSE), or the mean of the observed data (NSE). The RMSE can be decomposed following derivations presented in [109] and [110]. If p is a subscript for the predicted 7Q10 values, and o is a subscript for the observed 7Q10 values, then the RMSE can be decomposed into three parts,

$$RMSE^2 = MSE = (\mu_p - \mu_o)^2 + (\sigma_p - \sigma_o)^2 + 2\sigma_p\sigma_o(1-r), \tag{2.11}$$

where r is the linear correlation coefficient between x_0 and x_p, μ_p and σ_p are the mean and standard deviation of the predicted values, and μ_o and σ_o are the mean and standard deviation of the observed values. If we designate $A = (\mu_p - \mu_o)^2$, $B = (\sigma_p - \sigma_o)^2$, and $C = 2\sigma_p\sigma_o(1-r)$ from Equation 2.11, then we can visualize the absolute contribution of each component, A, B, and C. The first term (A) is a measure of model bias (i.e. the difference in the means of the observed and predicted 7Q10s), the second term (B) is a measure of how well the model matches the variance of the observed values and the third term (C) is the remaining error and is largely controlled by the covariance or correlation of the predicted and observed 7Q10s. Components A and B represent how well the model is able to recreate the location and shape of target distribution while component C accounts for the pairwise relationship between the predicted and observed values.

2.3.8 Variable Importance and Partial Dependence Plots

The relative importance values from the top three predictor variables for each machine-learning model were combined and rescaled. The SVMG model was omitted, but the SVMP model was retained to keep a representative from each model family. The relative variable importance for each model was calculated using the varImp function from the R caret package [111]. For the RF model, the mean squared error was computed on the out-of-bag sample for each tree that was built using a random subset of the predictor variables. The differences in mean squared errors with and without certain predictor variables was then used to determine the relative importance of each predictor. A similar approach was taken for GBM but the relative importance was the sum of the importances from each boosting iteration. The M5-cubist model returned the percentage of times a variable was used for a condition or was used in a linear regression model. The elastic-net and SVM

models returned the absolute value of the non-zero coefficients from the regression. For the KKNN model–a technique without an obvious way to calculate the variable importance–a filter method was used to compare R^2 values for models fit with a certain predictor variable compared to the R^2 from a null model. The predictor variable that showed the largest improvement in R^2 values over the null was considered the most important predictor.

The top three most important predictor variables from each model were then selected and the importance values were rescaled by dividing the sum of the importance value of each predictor variable by six, the maximum possible sum. This provided a combined measure of variable importance from all of the machine-learning models. Partial dependence plots (PDP) were created for the most important predictor variables from the RF, GBM, SVMP, and M5-cubist models. These four models were chosen to show the effect of varying the most important covariates for models from several families. A PDP showed the effect of varying a predictor variable of interest while accounting for the average effects of all other variables [47], providing mechanistic insights from "black-box" algorithms. The PDPs were created using the ICEbox R package [112].

2.4 Results

2.4.1 Prediction Errors and Error Decomposition

The machine-learning models outperformed the baseline models across almost all error metrics (Figure 2.4, Table 2.3). The exception was the ordinary kriging and the null models both had a lower UA-RMSE than the elastic-net model. The full-Tobit model had a lower RMSE than the ROI-Tobit model, but both had similar median percent errors. The level-1 meta-M5 cubist model resulted in the lowest RMSE, highest NSE, lowest MPE, and lowest unit-area RMSE. The machine-learning models generated predictions that more closely matched the observed values than the baseline models (Figure 2.5). The error decomposition revealed that although the Tobit models have the smallest bias errors, they had

relatively large variance and covariance errors (Figure 2.6). The top performing (i.e., lowers RMSE) machine-learning models generally showed smaller overall covariance errors.

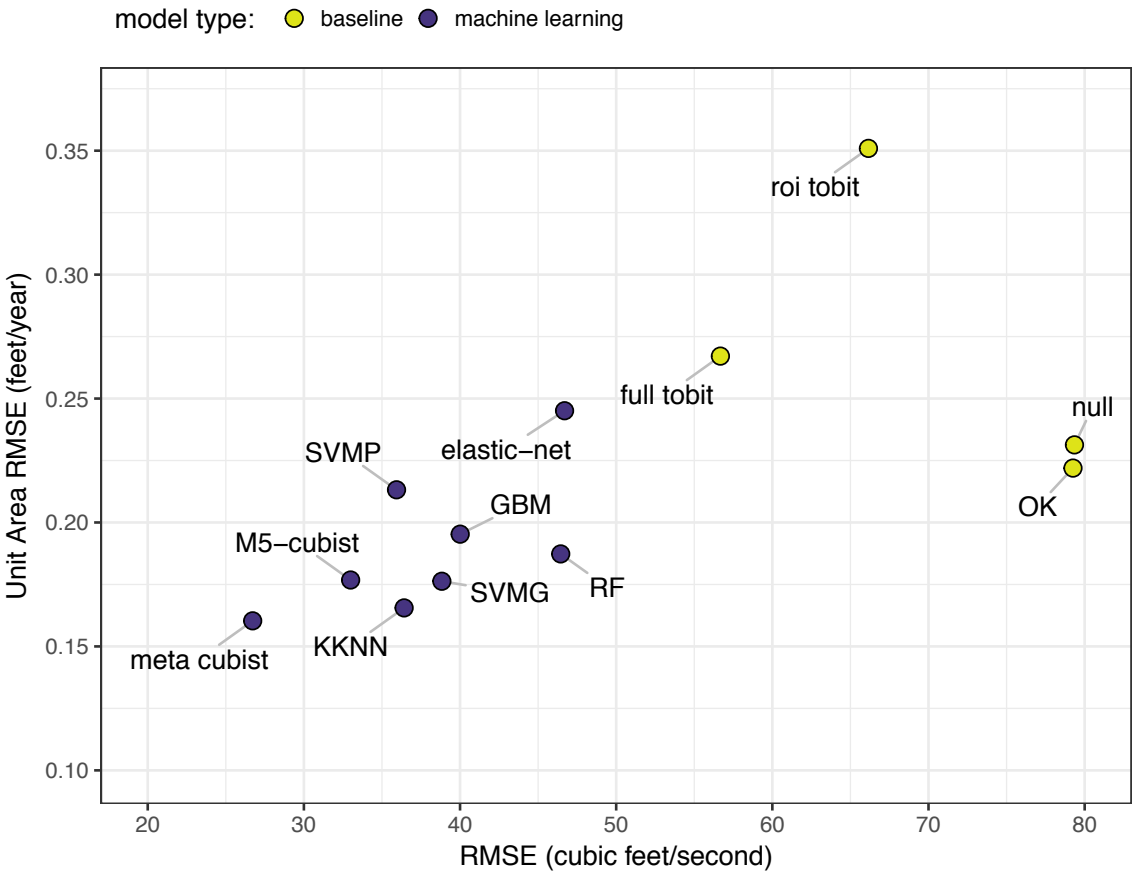

Figure 2.4: Leave-one-out Root Mean Squared Error (RMSE) and Unit Area RMSE for each model (residuals of each model were divided by the drainage area for the basin before calculating the RMSE). The RMSE provides a measure of overall goodness of fit for each model, and the unit area RMSE shows the error normalized by the size of the basin. From right to left, (1) null: null model, (2) OK: ordinary kriging, (3) ROI Tobit: Type I Tobit regression model using the region-of-influence method, (4) full Tobit: Type I Tobit regression model using the full data set, (5) elastic net: elastic-net regularized regression, (6) RF: random forest, (7) GBM: gradient boosting machine, (8) SVMG: support vector machine with Gaussian kernel, (9) KKNN: kernel-K-nearest neighbors, (10) SVMP: support vector machine with a polynomial kernel, (11) M5-cubist, (12) meta-M5-cubist: stacked ensemble M5-cubist model trained on the LOO-CV predictions from the other machine-learning models.

Table 2.3: Error metrics for each model sorted by RMSE value. The rank for each metric and model is listed in square brackets to the right of the metric value. The values were rounded to two decimal places for the table but the ranks were derived from seven decimal places (which is why, for example, the RF model was given a better rank for the NSE over the elastic net when it appears to be a tie in the table). Baseline models are indicated by italics. KKNN: kernel-K-nearest neighbors, RF: random forest, SVMP: support vector machine with a polynomial kernel, SVMPG: support vector machine with a Gaussian kernel, OK: ordinary kriging, ROI-Tobit: region-of-influence Type I Tobit.

Model	RMSE	Med % error	NSE	unit area RMSE
meta-M5 cubist	26.72 [1]	45.45 [1]	0.92 [1]	0.16 [1]
M5-cubist	33.00 [2]	55.00 [5]	0.88 [2]	0.18 [4]
SVMP	35.93 [3]	53.86 [4]	0.86 [3]	0.21 [7]
KKNN	36.42 [4]	52.08 [2]	0.85 [4]	0.17 [2]
SVMG	38.83 [5]	53.67 [3]	0.83 [5]	0.18 [3]
GBM	40.01 [6]	65.22 [7]	0.82 [6]	0.20 [6]
RF	46.45 [7]	60.30 [6]	0.76 [7]	0.19 [5]
elastic net	46.69 [8]	69.25 [8]	0.76 [8]	0.25 [10]
full-tobit	56.67 [9]	70.32 [9]	0.64 [9]	0.27 [11]
ROI-tobit	66.15 [10]	74.80 [11]	0.52 [10]	0.35 [12]
ordinary kriging	79.25 [11]	74.93 [10]	0.30 [11]	0.22 [8]
null	79.35 [12]	85.55 [12]	0.30 [12]	0.23 [9]

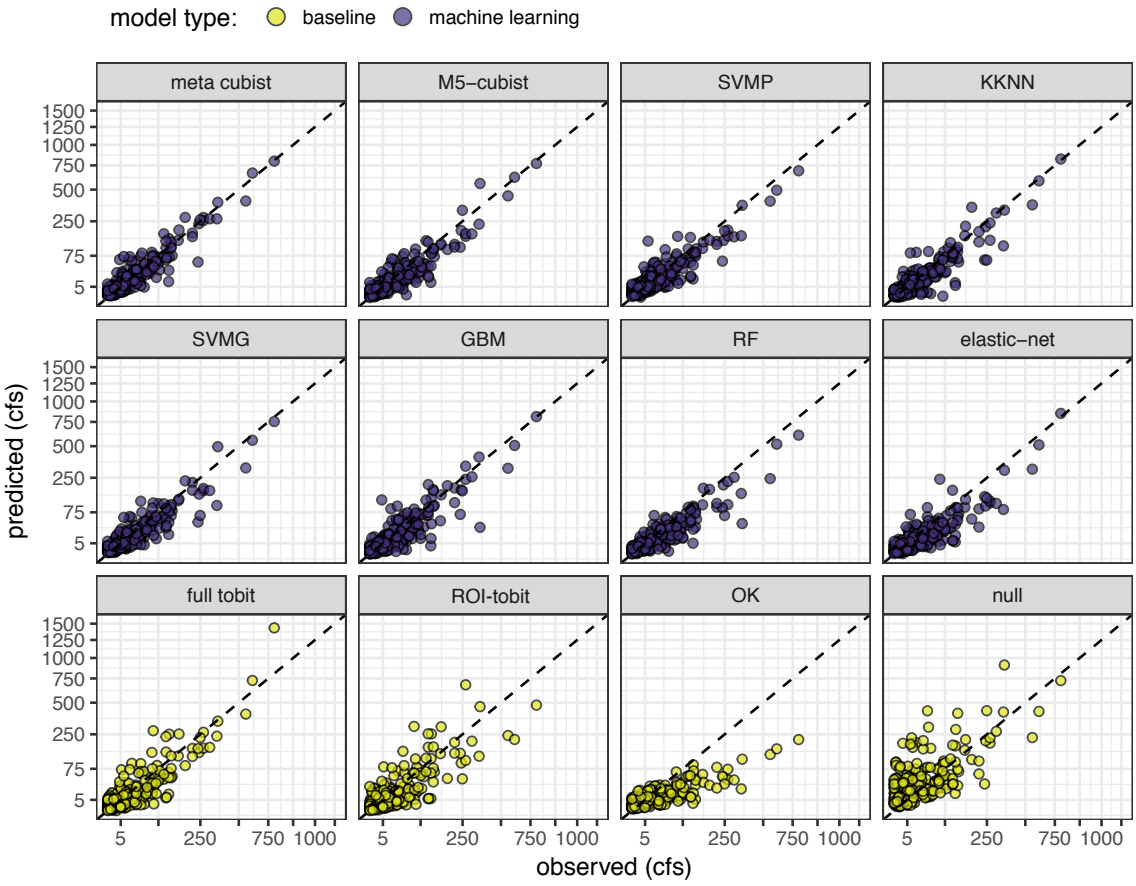

Figure 2.5: Scatter plot of predicted vs observed 7Q10s for all of the models. The panels are sorted by RMSE values from top left to bottom right. SVMP: support vector machine with a polynomial kernel, KKNN: kernel-K-nearest neighbors, SVMG: support vector machine with a Gaussian kernel, GBM: gradient boosting machine, RF: random forest, ROI-Tobit: region-of-influence Type I Tobit, OK: ordinary kriging.

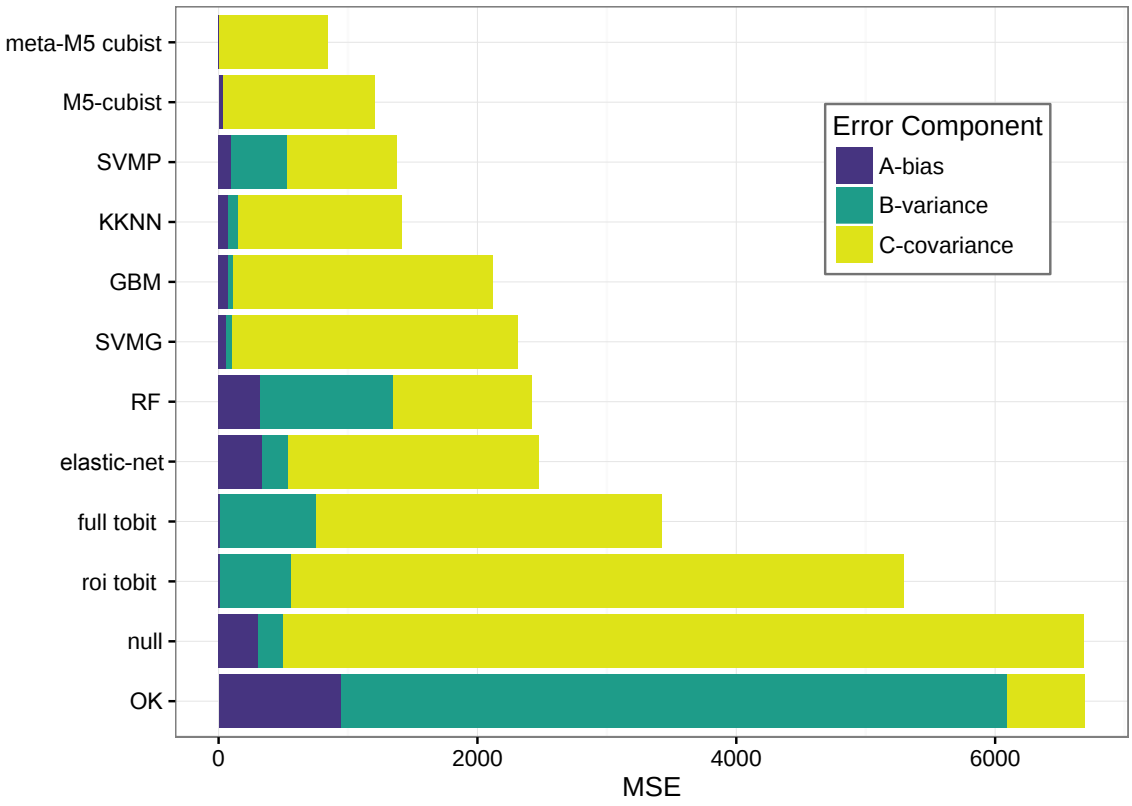

Figure 2.6: Absolute contributions from the decomposed mean squared error (MSE) shown in Equation 2.11. SVMP: support vector machine with a polynomial kernel, KKNN: kernel-K-nearest neighbors, SVMG: support vector machine with a Gaussian kernel, GBM: gradient boosting machine, RF: random forest, ROI-Tobit: region-of-influence Type I Tobit, OK: ordinary kriging.

2.4.2 Variable Importance

The overall variable importance value for percent wetlands was equal to one, which indicated that percent wetlands was the most important predictor variable for each individual machine-learning model (Figure 2.7). Unit 7Q10 values generally decrease with an increase of percentage wetlands in a basin. The second most important predictor variable was the percentage of soils in hydrologic soil group B (HGB, see [113] for more details about the soil classes), which has moderate infiltration rates and is moderately coarse in texture, and was in the top three most important predictor variables for all of the models except elastic net and M5-cubist. Unit 7Q10 values tended to increase with an increase of moderately well-drained soils. Other important predictor variables were the standard deviations of the minimum and maximum temperatures for the basins, depth to the seasonally high water table, mean August and November precipitation, and the percentage of well-drained soils with a high gravel and sand content. Unit 7Q10s increased with a greater amount of August and November precipitation, a higher amount of well-drained soils, and greater variability in the minimum and maximum temperatures for a basin. The results were mixed for the average depth to the water table (Figure 2.7). Further descriptions of the variables can be found in the supplementary material. The most important variables for the machine-learning models can be compared to the variables identified by forward stepwise selection used for the Tobit models. LOO-CV resulted in 8 predictors for both models. The 8 predictors in order of the absolute value of their coefficients are (1) drainage area of the basin, (2) percent well-drained soils, (3) mean estimated March runoff from 1951-2000, (4) percent forest in the basin, (5) percent wetlands in the basin, (6) percent pasture in the basin, (7) percent moderately well-drained soils, and the (8) standard deviation of the minimum temperature.

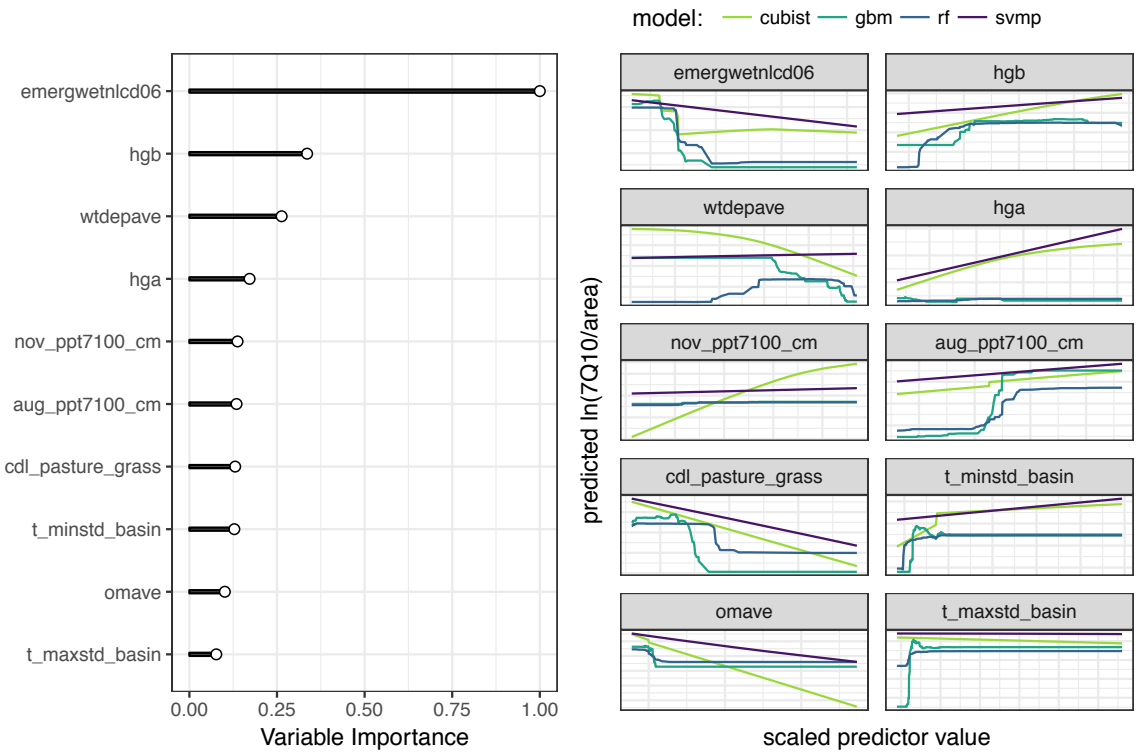

Figure 2.7: [*Left*]: Relative importance of predictor variables for machine-learning models. [*Right*]: Partial dependence plots for the predictor variables from the left panel. SVMP: support vector machine with a polynomial kernel, RF: random forest, GBM: gradient boosting machine. For the plot in the right panel, the numerical values on the axes are omitted as the various predictors have unique ranges of values (for both the scaled predictor value and the predicted value), which would require individual axis values for each the 10 separate facets. Including the individual values clutters the plot and considerably shrinks the size of the individual plots. Removing the axis values does detract from the purpose of the plot–to observe how predictions change based of different values of different predictor variables.

2.5 Discussion

2.5.1 Predictive performance and applications

Our results showed that machine-learning models can produce more accurate 7Q10 estimates for ungaged basins than traditional-statistical methods. The range of RMSEs also indicated that exploring multiple methods is paramount for discovering well-performing models. The relative accuracy of different models can provide additional insights into the nature of the data. For example, the three most accurate level-0 models (M5-cubist, KKNN, and SVMP) learn from data using different approaches. SVMP fits a robust regression line in a kernel feature space, KKNN takes the average response value of sites that are close together in kernel space, and M5-cubist extends tree-based methods. Although these models belong to different families, they all performed well on the largest 7Q10 values and did not exhibit any systematic bias (Figure 2.5). In contrast, the elastic-net model accurately estimated the larger 7Q10 values but under predicted a majority of sites with 7Q10 values between 5-500 cfs. This suggests that the best performing models were able to handle high leverage observations in the training set. The meta-M5 cubist model learned from the LOO-CV errors produced by the level-0 models (Figure 2.3) and further increased the accuracy of the predictions. However, the meta-M5 cubist model was unable to reduce greatly the UA-RMSE, which suggests that the error relative to the size of the basin may be near a threshold and is insensitive to small changes to absolute error (as represented by, e.g., RMSE).

In practice, watershed management goals should guide the choice of the error metric and the "best model" for a given task [46]. For example, if an aquatic ecological habitat depends on a specific streamflow characteristic, then the model that results in the lowest RMSE for that characteristic may provide the best information for managing the ecological streamflow. Conversely, if the management goal is to determine basins with anomalous flows relative to the size of the basin (e.g. predict where a large basin might have

a small 7Q10), then the UA-RMSE performance of models is more relevant to that task than the RMSE. Many states use the 7Q10 for design flows while other states have selected other low-flow statistics that are more suited for a given objective. Our study-design is dependent-variable agnostic–i.e., the data processing and choice of models are not specific to 7Q10s. The same analysis could be rerun for a different statistic and a water manager could select the model that performs best in their region, for their statistic, and their purpose. The model could then be used to make predictions in ungaged locations. The results presented in this paper demonstrate that machine learning methods, including meta-modeling, are viable approaches for future low-streamflow statistic regionalization studies. Comparable results in truly ungaged-basins can only be expected if the same set of explanatory variables are available for ungaged locations, and to the authors knowledge, this type of dataset does not yet exist. However, the 10 variables in Figure 7 (and drainage area) provide a tractable starting place. Additionally, the models in this paper are built using 7Q10 data from 2015 and future models would benefit from updated 7Q10 estimates prior to regionalization.

2.5.2 Error decomposition

The machine-learning models are optimized to minimize the overall prediction error (RMSE), whereas Tobit and ordinary kriging are optimized to produce an unbiased estimate with the smallest variance. Decomposing the overall error can lead to insights about model behavior and allows us to examine what is driving the error for each model (Figure 2.6). For example, the Tobit models were unbiased with low variance, but had a large amount of unsystematic error, whereas the elastic-net model accepted higher bias to reduce variance and decrease the overall error, which is a property of regularized regression [47]. The ordinary kriging and null models had the same MSE, but the ordinary kriging error was almost entirely composed of variance, which is likely the result of transformation bias on a linearly weighted summation [114]. The meta-M5 cubist model was unbiased with

minimal variance. This indicated that the mean and variance of meta-M5 cubist predictions was equal to the mean and variance of the actual 7Q10s. The remaining error results are from imperfect linear dependence between the predicted and the observed 7Q10 values.

Although we recognize the value of incorporating uncertainty into the model predictions, we do not explicitly include uncertainty intervals here because real-world applications of 7Q10 predictions require a single 7Q10 prediction for an ungaged basin. We partially explore uncertainty by a decomposition of the error terms for each model (Section 2.5.2), but a full exploration of the implications of predictive uncertainty are beyond the scope of this work.

2.5.3 Physical controls of 7Q10s

The primary goal of estimating 7Q10 values in ungaged basins is predictive accuracy, and simple interpretable functions rarely produce the most accurate predictions [67]. However, even when the objective is prediction, valuable mechanistic insights can be gained by examining the effect of specific predictor variables [112]. This is often accomplished by calculating the importance of each covariate used to fit a particular model, which although useful, only accounts for the magnitude of the effect and not the direction (Figure 2.7, left panel). In addition to variable importance, we explored the effect of each predictor in more detail using partial dependence plots (Figure 2.7, right panel). In a review of low-flow hydrology, [22] lists several factors that influence the low-flow regime of a basin: the distribution and infiltration characteristics of soils, the hydraulic characteristics and extent of the aquifers, the rate, frequency and amount of recharge, the evapotranspiration rates from the basin, distribution of vegetation types, topography and climate. Five of the seven factors mentioned by [22], including soils, aquifer characteristics, recharge, vegetation type, and climate are reflected in the most important predictor variables identified here for the machine-learning models (Figure 2.7).

The percentage of emergent wetlands was an important covariate for the machine-

learning models, and wetlands have been shown to modulate streamflow [115]. As the percent of emergent wetlands increases, the 7Q10/drainage area decreases 2.7. The distribution of emergent wetlands is heavily left-skewed where only a few basins have large percentages of wetlands, and almost all of those sites are clustered in the Coastal Plain of Georgia and South Carolina. Percent wetlands is negatively correlated with a shallow depth to the water table (ρ=-0.82), November precipitation (ρ=-0.53), and moderately well-drained soils (ρ=-0.45). This correlation suggests that wetlands, depth to water table, and 7Q10s are controlled by precipitation and local surface geology, and that the importance of percent wetlands for predicting unit 7Q10s may simply reflect that they are both influenced by similar processes. The PDP suggested that basins with very low standard deviations of maximum and minimum monthly air temperature (min-max-temp) also tend to have low unit 7Q10s. However, this trend quickly disappears after an increase in unit 7Q10. One possible explanation is that most basins with high standard deviations of min-max-temps are located in the higher elevations of the Piedmont and Blue Ridge physiographic provinces, which consist of a different geology than the Coastal Plain. As with the percent wetlands predictor, it is difficult to infer whether there exists a mechanistic relationship between the standard deviation of min-max-temps and unit 7Q10s, or if the trend is simply an artifact of this particular data set.

2.5.4 Conclusions

Machine-learning methods can produce more accurate predictions of 7Q10s in ungaged basins than historically relevant baseline models. M5-cubist models, kernel-K-nearest-neighbor models, and polynomial kernel support vector machines show the greatest improvements in prediction accuracy over Type I Tobit models and ordinary kriging. The improved prediction accuracy of the machine-learning models can be explained by how each model treats the bias variance tradeoff. Multivariate regression and ordinary kriging are both optimized to produce the best linear unbiased estimator, whereas machine-learning

models minimize the overall prediction error by tuning hyperparameters using Bayesian optimization or grid search. This tuning process commonly accepts bias to reduce variance but generates optimal predictions.

Variable importance measures and partial dependence plots show that percent emergent wetlands in the basin is the most important predictor variable for the machine-learning models. We interpret this correlation as simply an indication that 7Q10s and percent wetlands are likely controlled by similar factors–late summer and early fall precipitation, the infiltration rate of soils, and the variability of minimum and maximum monthly temperatures–which also emerge as some of the most important variables for predicting 7Q10s.

Machine-learning approaches show much promise for improving predictions of low streamflow in ungaged catchments. Additionally, combining the predictions of multiple first order machine-learning models via a global meta-model is a novel yet practical advancement for hydrologic-regionalization studies.

CHAPTER 3

PREDICTING FLOW DURATION CURVES IN UNGAGED BASINS USING L-MOMENTS AND THEORY-INFORMED NEURAL NETWORKS

3.1 Abstract

We develop theory-driven neural network models to predict flow duration curves (FDCs) in ungaged locations. The model architecture contains multiple response variables in the output layer that correspond to either individual quantiles or L-moments. During training, predictions are made for each response variable and a combined loss function is used for back propagation and parameter updating. The combined loss function accounts for the covariance between the response variables and generates physically-consistent outputs (e.g., monotonically increasing quantiles with increasing nonexceedance probabilities). We compare a model that predicts directly 27 quantiles simultaneously (referred to as *direct multi-output neural network*, DMNN), a model that predicts 27 quantiles independently (referred to as *direct single-output neural network*, DSNN), and a model that simultaneously predicts four L-moments that are used to parameterize a 4-parameter Asymmetric Exponential Power Distribution (referred to as *L-moment multi-output neural network*, LMNN). All the predictions are made using a 10-fold cross validation framework. The multi-output neural network model results in realistic flow duration curves. We show that the DMNN model produces more physically-consistent and accurate predictions than the DSNN model. The DMNN model also outperforms the LMNN model over the 27 quantiles. We also show how neural network dropout can be used to generate posterior predictive distributions for FDCs. Finally, we demonstrate how estimating FDCs can be used to estimate daily streamflow in ungaged catchments.

3.2 Introduction

Streamflow at a catchment outlet can be statistically summarized by a flow duration curve (FDC). An FDC maps streamflow values to their corresponding exceedance probabilities [116]. The unique shape of an FDC for a given basin and time period is the result of multiple-hydrologic processes interacting at various temporal and spatial scales [117]. The physical characteristics of a basin (e.g., mean temperature, soil type, potential evapotranspiration, elevation, landuse, etc) filter precipitation into the different components of streamflow [118, 119, 120, 121, 122, 123]. The shape of an FDC is therefore controlled by both basin characteristics and precipitation.

Estimating FDCs is relevant to predicting streamflow in ungaged catchments [124, 76], quantifying streamflow alteration [125, 126], exploring spatial and temporal trends in streamflow [127, 128], and for predicting the frequency of floods or low streamflow [14]. A physically-based approach involves using a rainfall-runoff model to generate streamflow and a corresponding FDC based of the estimated values [118]. A benefit of the physically-based method is that an FDC can be estimated in any location provided the data needed to force the model and calibrate the parameters are available. It also insures that the resulting FDC will have flows that always increase with increasing quantiles (i.e., monotonicity). Physically-based models, however, often do a poor job of recreating the distributional properties of streamflow and statistical models are preferred for generating FDCs [21]. Common statistical methods include (1) regression-based methods that estimate quantiles independently using basin characteristics [76, 129, 130], (2) distributional methods that estimate statistical moments using basin characteristics and fit an analytical distribution to the moments [131, 132], (3) streamflow index-based methods [117], and (4) geostatistical methods [133, 134]. Recent research has improved on each of these methods.

Even though statistical methods tend to produce better estimates, they often require significant post-processing to account for physical implausibility in the results. For example, direct estimation of quantiles is one of the most accurate methods but produces

non-monotonic FDCs (Figure 3.1). This is usually mitigated by interpolating a value to replace the violating quantile. Distributional methods ensure monotonicity but can lead to poor fits in the tails of the distribution. Can we improve on these methods by incorporating what we know about the physical system into the statistical framework? Recent work by Poncelet et el., [135] introduced a method that leverages the physical relationship between quantiles and basin characteristics to ensure quantile solidarity. Their method improves prediction accuracy and results in hydrologically-consistent models, however, it requires significant preprocessing of the explanatory variables and involves multiple models that are independently fit and combined.

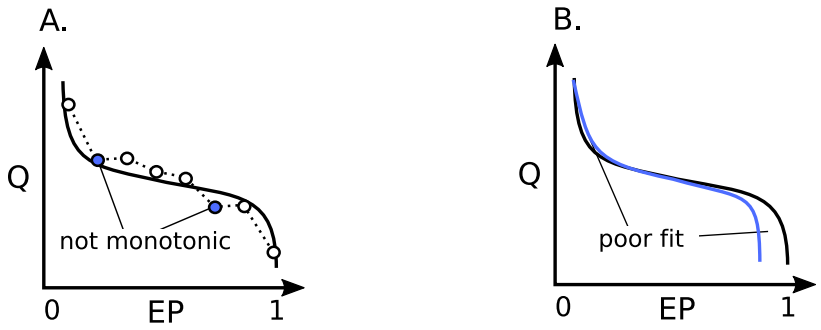

Figure 3.1: Two common problems with (A) regression-based methods and (B) distributional methods.

We propose a multi-output neural-network (NN) model to predict each quantile simultaneously (referred to as *direct multi-output neural network*, DMNN). The network parameters are determined by a weighted-loss function that accounts for the prediction accuracy of each output, which greatly reduces the number of non-monotonic predictions. Additionally, the model requires minimal preprocessing of the data. We also use a multi-output NN to predict 4 statistical moments and parameterize a 4-parameter Asymmetric Exponential Power (AEP) distribution (referred to as *L-moment multi-output neural network*, LMNN). The DMNN and LMNN models are compared to single-output NNs that predict each quantile independently (referred to as *direct single-output neural network*, DSNN). We also use

a variant of the QPPQ [136, 137, 129] ("Q" refers to discharge and "p" refers to exceedance (or non-exceedance) probability. More details can be found in Section 3.3.6) to estimate daily streamflow for an example site using the estimated FDC.

3.3 Methods

3.3.1 Streamflow Data

The streamflow data was downloaded from the USGS National Water Information System (NWIS). Daily streamflow data was downloaded for 1,379 sites and 423 sites were removed that contained greater than 1% negative flows, consisted of provisional data, were located outside of our study area, had less than 10 years of data, or did not have data between 1950-2010 (Figure 3.3). The remaining 956 sites were used for further analysis (Figure 3.2). The daily streamflow was grouped into 6 decades, 1950-1959, 1960-1969,..., 2000-2009, and flow statistics were calculated per decade (Table 3.1). This was done to provide a greater number of observations for regionalization and to partially addressed the non-stationarity in FDCs. The final streamflow dataset contained 2,807 site-decade combinations. The flow statistics include the following 27 quantiles, 0.0002, 0.0005, 0.0010, 0.0020, 0.0050, 0.0100, 0.0200, 0.0500, 0.1000, 0.2000, 0.2500, 0.3000, 0.4000, 0.5000, 0.6000, 0.7000, 0.7500, 0.8000, 0.9000, 0.9500, 0.9800, 0.9900, 0.9950, 0.9980, 0.9990, 0.9995, 0.9998, and L-moments, L1, T2, T3, and T4 (for more information on L-moments see Section 3.3.3).

Table 3.1: Number of sites per decade.

Decade	**Number of sites**
1950	362
1960	487
1970	484
1980	455
1990	455
2000	564

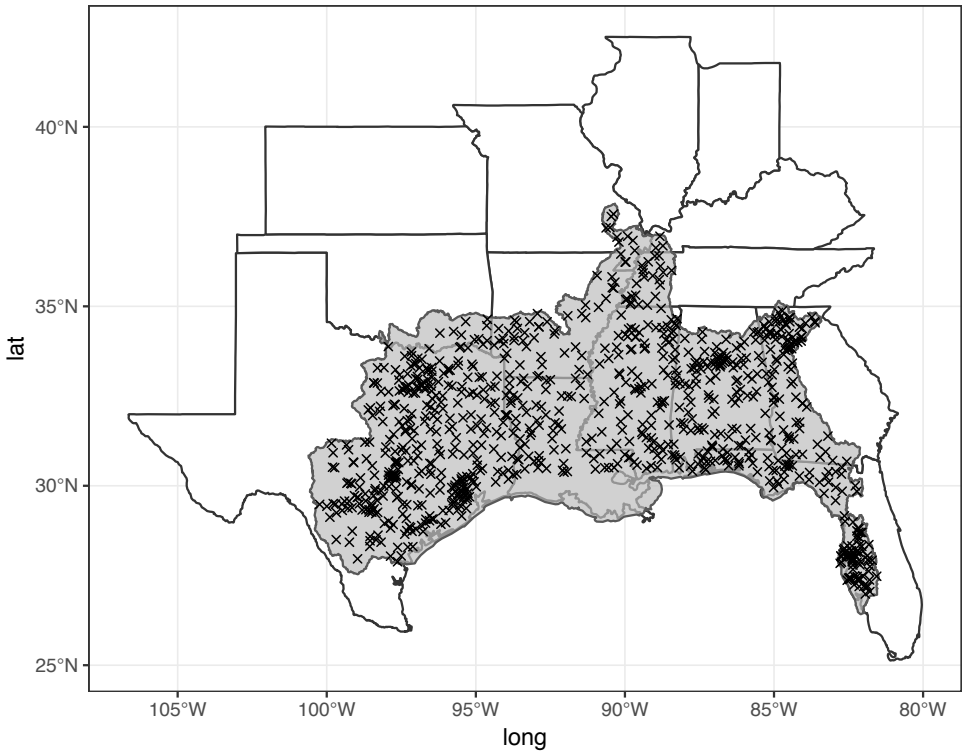

Figure 3.2: Study area location and streamgages.

3.3.2 Basin Characteristics

All basin characteristics were obtained from the USGS data release [138] and are publicly available (https://doi.org/10.5066/F7765D7V) (Table 3.2). Basin characteristics that can potentially evolve through time, e.g., land use, precipitation, reservoir storage, etc, were aggregated by decade and are referred to here as "mutable" variables. Basin characteristics that are generally considered to be fixed through time, e.g., elevation, soil type, physiographic province, etc, did not change by decade and are referred to here as "immutable" variables. The basin characteristics are spatially aggregated by the NHDPlus version 2 catchments and linked to NWIS stream gages and 12-digit Hydrologic Unit Codes (HUC12s) by a unique identifier [138]. Two spatial components were used for aggregation, (1) by reach catchment, which characterizes data at the local scale, or (2) accumulated through the river networks, which characterizes cumulative upstream conditions. The ac-

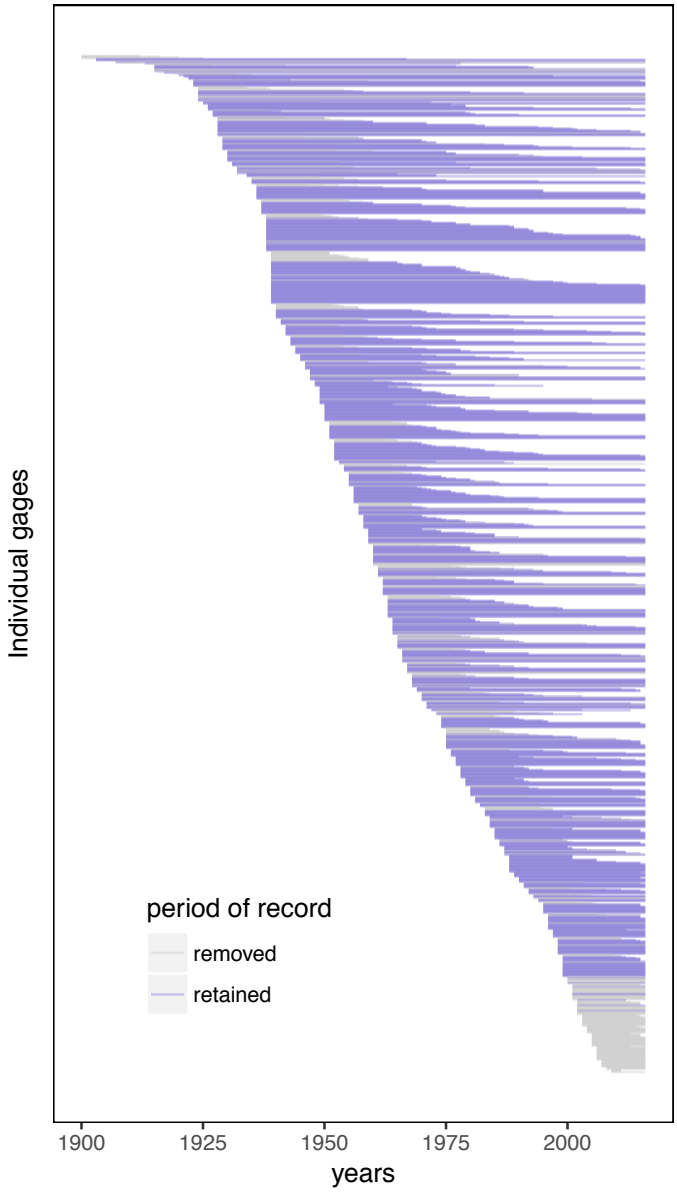

Figure 3.3: Period of record for original 1,379 sites sorted by start year and length of record. Sites that were removed are colored in orange and those that were retained are colored in blue. Note that data prior to 1950 was not used due to the lack of reliable basin descriptor data pre-1950.

cumulation method followed NHDPlus version 2 flowlines upstream until it reached the first divergence.

Table 3.2: The 44 Basin characteristics used for regression models. Two spatial components were used for spatial aggregation, (1) reach catchments (CAT), which characterizes data at the local scale, or (2) through the river network (NET), which characterizes cumulative upstream conditions.

Mutable	**Description**	**Unit**
precip	decadal mean and stdev of mean annual precip	NET
temp	decadal mean and stdev of mean annual temp	NET
runoff (Q)	decadal mean and stdev of mean annual Q from WBM[a]	NET
housing density	decadal housing units per km^2	NET
dams	decadal total number of dams	NET
major dams[b]	decadal total number of major dams	NET
dam storage	decadal dam normal and total storage volume	NET
LULC	decadal percent landcover land use for 12 classes	NET
Immutable	**Description**	**Scale**
lon	longitude of gage	CAT
lat	latitude of gage	CAT
elevation	min, max, and mean elevation	NET
basin area	upstream basin area	NET
baseflow index	ratio of long-term baseflow to total stream flow	NET
TWI	Topographic wetness index	NET
basin slope	average slope	NET
road crossings	number of roads crossing stream	NET
sinuosity	stream length/straightline distance of s	CAT
stream density	stream density in basin	CAT
permeability	permeability of streambed material	CAT
aquifer	primary aquifer code	CAT
HLR	hydrologic landscape region	CAT
ecol3	level III ecoregion	CAT
physio	physiographic province	CAT
soil type	primary soil type	CAT

[a] Water balance model

[b] dams 50 feet or more in height, dams with a normal storage capacity $\geq$ 5,000 acre-feet, and dams with a maximum storage capacity $\geq$ 25,000 acre-feet

3.3.3 L-moments

L-moments are calculated as linear combinations of order statistics and are widely used in regional frequency analysis in favor of product moments (i.e., standard deviation, skew,

etc) [139, 140, 141]. L-moments have both theoretical and practical advantages compared to conventional moments; they are more robust on small datasets and in the presence of outliers, they are less subject to bias, and can characterize a wider range of distributions. Many sources in the literature aptly describe the history and theory of L-moments in hydrology [139, 141, 142, 143] and we only briefly describe the calculations here. Given the product weighted moments [144],

$$\beta_0 = \frac{1}{n}\sum_{j=1}^{n} x_j \tag{3.1}$$

$$\beta_1 = \frac{1}{n}\sum_{j=2}^{n} x_j[(j-1)/(n-1)] \tag{3.2}$$

$$\beta_2 = \frac{1}{n}\sum_{j=3}^{n} x_j[(j-1)(j-2)]/[(n-1)/(n-2)] \tag{3.3}$$

$$\beta_3 = \frac{1}{n}\sum_{j=3}^{n} x_j[(j-1)(j-2)(j-3)]/[(n-1)/(n-2)(n-3)], \tag{3.4}$$

and if r is the order of the L-moment, the first 4 L-moments (λ_r) and 3 L-moments ratios (τ_r), along with their product moment analogs, are calculated as,

$$\lambda_1 = \beta_0 = \mu \tag{3.5}$$

$$\lambda_2 = 2\beta_1 - \beta_0 = \sigma \tag{3.6}$$

$$\lambda_3 = 6\beta_2 - 6\beta_1 + \beta_0 \tag{3.7}$$

$$\lambda_4 = 20\beta_3 - 30\beta_2 + 12\beta_1 - \beta_0 \tag{3.8}$$

$$\tau_2 = \lambda_2/\lambda_1 = coefficent\ of\ variation \tag{3.9}$$

$$\tau_3 = \lambda_3/\lambda_2 = skew \tag{3.10}$$

$$\tau_4 = \lambda_4/\lambda_2 = kurtosis. \tag{3.11}$$

For our purposes, we only need λ_1, τ_2, τ_3, and τ_4 to estimate the parameters of the 4-parameter Asymmetric Exponential Power Distribution (see Section 3.3.4) using the "method of L-moments": a parameter estimation technique analogous to the well-known method of moments. The method provides parameter estimates by choosing (analytically or through numerical methods) the parameters of a probability distribution so as to equate the theoretical L-moments of the distribution to the sample L-moments. The conceptual basis is that the sample L-moments succinctly quantify the distributional geometry of the sample. Every distribution with finite mean is defined by its theoretical L-moments, which can be computed for a given set of parameters. The method of L-moments simply sets the sample and theoretical L-moments as equalities [145].

3.3.4 The 4-parameter Asymmetric Exponential Power Distribution

The regionalized L-moments were used to generate 4-parameter Asymmetric Exponential Power (AEP) distributions that were used to approximate FDCs. The AEP distribution was introduced by [146] and the theoretical L-moments were first derived by [147]. A detailed description of the AEP and it's L-moments can be found in [145]. The distribution functions of the AEP having parameters ξ (location), α (scale, $\alpha > 0$), κ (shape1, $\kappa > 0$), h (shape2, $h > 0$) for probability density f, nonexceedance probability F, and quantile x ($-\infty < x < \infty$) are

$$f(x) = \frac{\kappa h}{\alpha(1+\kappa^2)\Gamma(1/h)} \exp[-(\kappa^{\text{sign}(x-\xi)}(|x-\xi|/\alpha))^h], \tag{3.12}$$

$$F(x) = \begin{cases} [\kappa^2/(1+\kappa^2)]\ \gamma([(\xi-x)/(\alpha\kappa)]^h,\ 1/h) & \text{for } x < \xi, \\ 1-[1/(1+\kappa^2)]\ \gamma([\kappa(x-\xi)/\alpha]^h,\ 1/h) & \text{for } x \geq \xi\ \text{, and} \end{cases} \tag{3.13}$$

$$x(F) = \begin{cases} \xi - \alpha\kappa[\gamma^{(-1)}([1+\kappa^2]F/\kappa^2,\ 1/h)]^{1/h} & \text{for } F < F(\xi), \\ \xi + (\alpha/\kappa)[\gamma^{(-1)}([1+\kappa^2](1-F),\ 1/h)]^{1/h} & \text{for } F \geq F(\xi)\,, \end{cases} \tag{3.14}$$

where $\Gamma(a)$ is the complete gamma function, and is defined as,

$$\Gamma(a) = \int_0^\infty y^{a-1}\exp(-y)\ \mathrm{d}y, \tag{3.15}$$

and where $\gamma(Z,a)$ is the upper tail of the incomplete gamma function,

$$\gamma(Z,a) = \frac{\int_Z^\infty y^{a-1}\exp(-y)\ \mathrm{d}y}{\Gamma(a)}, \tag{3.16}$$

and where $\gamma^{(-1)}(Z,a)$ is the inverse of the upper tail of the incomplete gamma function. The AEP subsumes the Normal ($\kappa = 1, h = 2$) as well as the Laplace or Double Exponential ($\kappa = h = 1$) distributions. The theoretical quantile function is analogous to a FDC (Figure 3.4).

3.3.5 Neural network model

Neural networks (NNs) rely on multiple layers ("hidden layers") of nonlinear processing units (i.e., "neurons", "nodes") to learn associations from data [47, 48]. "Deep learning" refers to neural networks with many hidden layers [148, 149, 150]. NNs have been used extensively in hydrology for predicting water resource variables [49, 50, 51, 52] while deep learning has only recently been proposed for hydrologic inference [151]. Among

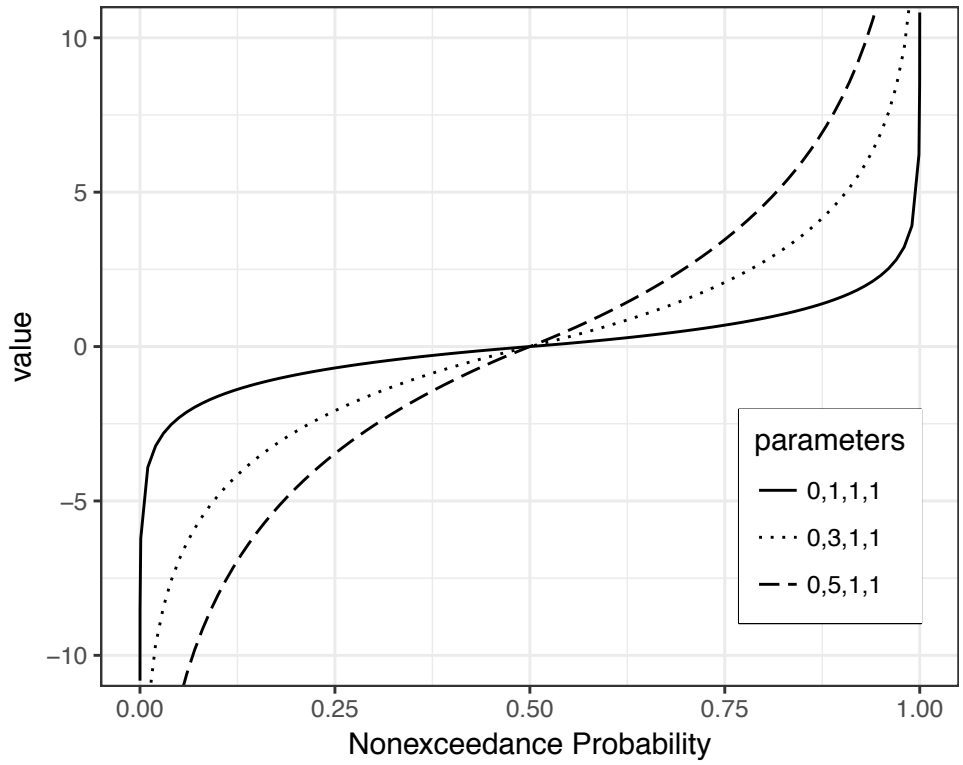

Figure 3.4: AEP quantile function with varying scale parameters.

other advantages, deep NNs learn features at varying levels of abstraction and tend to generalize better to new data than do shallower NNs [152]. Despite the recent success of deep learning in the fields of natural language processing [153, 154], image recognition [155], and beating humans at games [156], there is not consensus that deep learning is necessary or more accurate for most learning tasks [157, 158, 159, 160]. There is also not consensus about how many hidden layers are needed before a NN is considered "deep", but [161] suggest that two or more hidden layers constitutes a deep NN (DNN).

3.3.5.1 Description of basic Neural Network

Nodes in a NN receive inputs and compute outputs. Each input has an associated weight (w) that describes its relative importance compared to other inputs in the same layer (Figure 3.5). The node executes a nonlinear function to the weighted sum of its inputs. This nonlinear function is referred to as the *activation function* and allows neural network models to

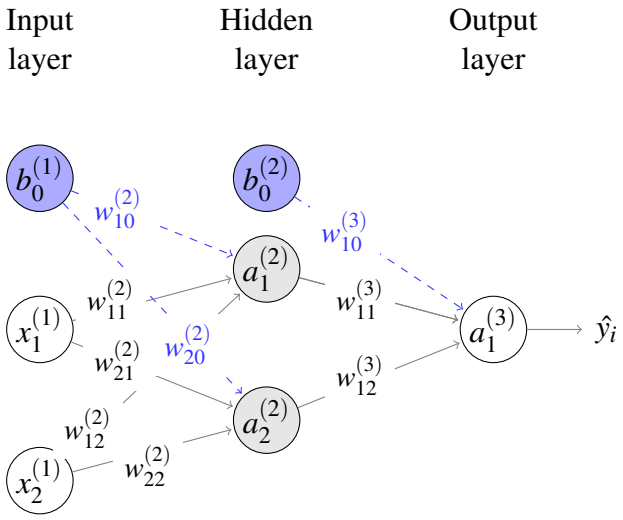

Figure 3.5: Basic neural network structure for the prediction of a single observation, i. The blue nodes are the biases, x_1 and x_2 are the inputs, $w_{10}^{(2)}$, $w_{20}^{(2)}$, $w_{11}^{(2)}$, $w_{12}^{(2)}$, $w_{21}^{(2)}$, and $w_{22}^{(2)}$ are the weights from the first layer, $w_{10}^{(3)}$, $w_{11}^{(3)}$ and $w_{12}^{(3)}$ are the weights from the hidden layer, $a_1^{(2)}$, $a_2^{(2)}$, and $a_1^{(3)}$ are the activation functions, and $\hat{y}_i$ is the predicted output for observation i.

be considered universal approximators [6]. The activation function takes a single number input and performs a predefined mathematical operation to it. Several common activation functions are sigmoid: $\sigma(x) = 1/(1+exp(-x)$, tanh: $tanh(x) = 2\sigma(2x) - 1$, and rectified linear unit (ReLU): $f(x) = max(0,x)$ [161] (Figure 3.6). Each node also has an associated bias (Figure 3.5) that shifts the activation function left and right (Figure 3.7).

Initial training of the network involves assigning random weights and completing a *forward pass* to compute a prediction. A forward pass simply involves multiplying the weights by their inputs, summing them up, passing the result through an activation function, and pushing the value forward to the nodes in the following layer. Using a sigmoid activation function and the parameter notation in [162]. the activation for each hidden node j in layer l can be written as,

$$a_j^{(l)} = \sigma\left(\sum_k (w_{jk}^{(l)} a_k^{(l-1)}) + b_j^{(l)}\right) \tag{3.17}$$

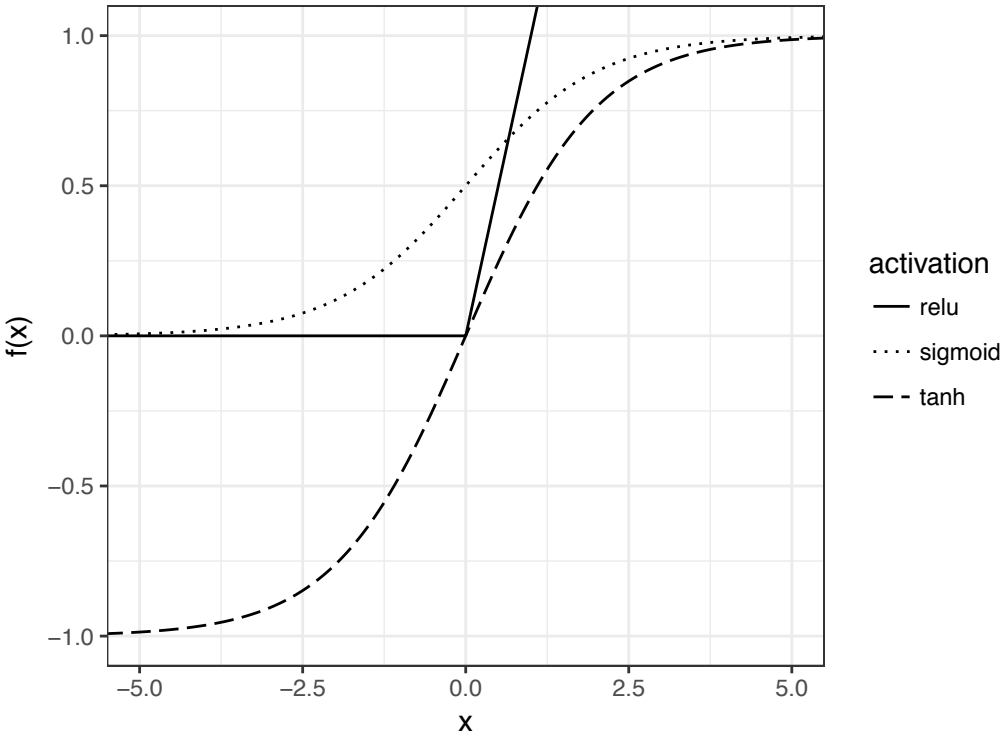

Figure 3.6: Common activation functions computed over the domain x=seq(3,-3,by=0.01).

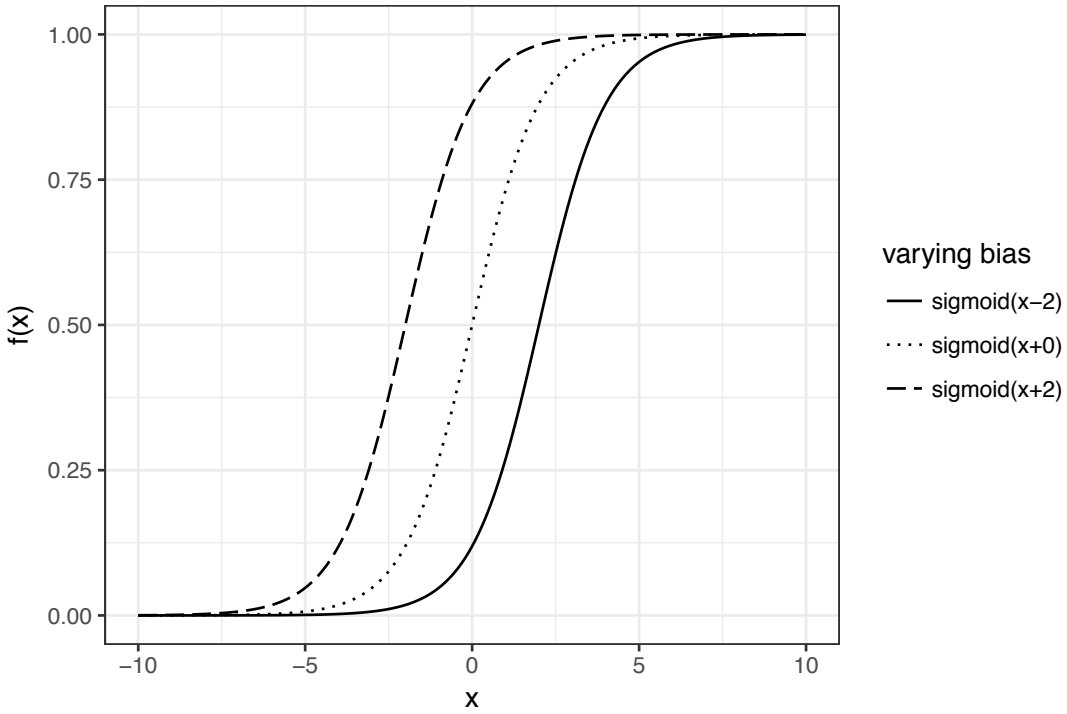

Figure 3.7: Varying bias for sigmoid function computed over the domain x=seq(10,-10,by=0.01).

where σ is the sigmoid activation function, $w_{jk}^{(l)}$ is the weight from the k^{th} node in the $(l-1)^{th}$ layer to the j^{th} node in the l^{th} layer. The sum in Equation 3.17 is the dot product between the outgoing weights from layer l and the activation of layer l,

$$\sum_k (w_{jk}^{(l)} a_k^{(l-1)}) = \left[w_{jk}^{(l)}\right]^T \left[a_k^{(l-1)}\right] \tag{3.18}$$

For example, using a sigmoid activation function, the calculation for the output node in Figure 3.5 is,

$$a_1^{(3)} = \sigma\left(w_{11}^{(3)} a_1^{(2)} + w_{12}^{(3)} a_2^{(2)} + w_{10}^{(3)} b_0^{(2)}\right) \tag{3.19}$$

The output of the activation function for the hidden layer becomes the the input to the activation function of the output layer. The output layer provides a predicted value for one observation $\hat{y}_i$. The squared error of the prediction,$\hat{y}_i$, can be calculated for as,

$$\delta_i^{(3)} = (y_i - \hat{y}_i)^2. \tag{3.20}$$

where y_i is the actual value of observation i. The error is referred to as the *cost*, and is represented as C. The error can then be *back-propagated* through the network using the chain rule [163]. Using the notation above the partial derivative (i.e, the gradient) can be written as,

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = a_k^{(l)} \delta_j^{(l+1)}. \tag{3.21}$$

The partial derivative in Equation 3.21 describes the change in weight $w_{jk}^{(l)}$ associated with a change in the cost function. Training a network using back propagation involves finding values of the weights that minimize C via a gradient descent optimization algorithm, such as RMSprop [164].

The network shown in Figure 3.5 is only for one observation. When doing regression

with many observations, a cost function is chosen that averages the costs for each observation. A common choice is the mean squared error (MSE),

$$C = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \tag{3.22}$$

The weights can be rewritten in matrix form to simplify the weight-updating process. A weight matrix $\Theta^{(l)}$ is defined for each layer l where the entries correspond to the indexing described in Equation 3.17, where the entry in $\Theta^{(l)}$ for j^{th} row and k^{th} column is $w_{jk}^{(l)}$. So in matrix form, the activation for layer l can be written as,

$$a^{(l)} = \sigma(w^{(l)} a^{(l-1)} + b^{(l)}) \tag{3.23}$$

and the associated error for each node j of layer l is,

$$\delta_j^{(l)} = \frac{\partial C}{\partial \sum_k (w_{jk}^{(l)} a_k^{(l-1)}) + b_j^{(l)}}. \tag{3.24}$$

The weights are updated for a set number of iterations, referred to as an *epoch*. The number of epochs is chosen based on when the MSE on the validation set stops improving.

3.3.5.2 Multiple Outputs and Dropout

Multiple-output regression involves predicting several response variables using the same set of explanatory variables [165]. The goal is to leverage correlations between the response variables to improve the predictive accuracy rather than relying on separate regressions that ignore the relationships between individual response variables [166]. This is straightforward to implement in a neural network model. For each epoch, the MSE is calculated for each desired output, and the weighted-average MSE is used to update the weights [164]. The weights assigned to the average MSE are defined based on the range of the output variables and which output should have the most control on the weights.

Dropout is a regularization method where a fraction of randomly selected nodes, along

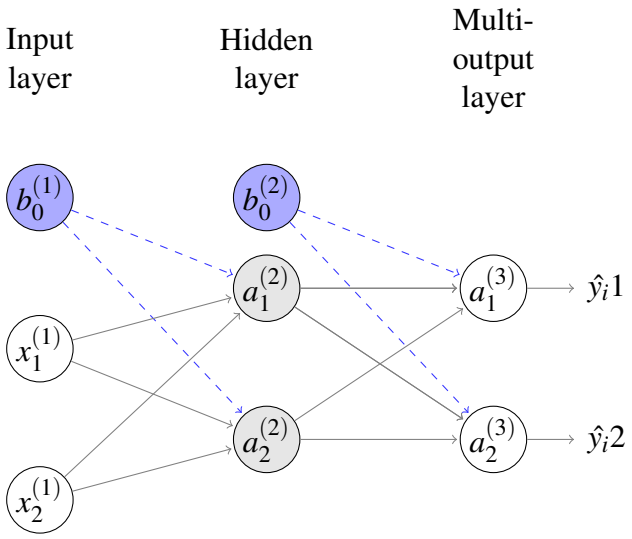

Figure 3.8: Basic neural network structure for predicting multiple outputs. Note, the weights and are not shown for clarity.

with their connections, are removed from a network during training [167]. The nodes are removed at a pre-specified rate for each layer given by probability $p^{(l)}$ (Figure 3.9). Removing nodes forces the network to learn more robust weights that are less prone to overfit by preventing hidden nodes from learning complicated features of the training data that are not present in the test set [168]. Predictions on the test set are made by scaling the weights of retained units in layer l by $p^{(l)}$, which ensures that the expected output from the "thinned network" is maintained at test time. This procedure of scaling the weights has been shown to be a computationally cheap form of model averaging and is similar to bagging used in tree-based methods and naive Bayes used in classification [168].

Dropout can also be used to capture model uncertainty and [169] showed that dropout is mathematically equivalent to a Bayesian approximation (i.e., probabilistic deep Gaussian process marginalized over its covariance function parameters). A predictive-posterior distribution can be obtained by using running multiple forward passes through a network with dropout. This process is referred to as *Monte-Carlo dropout* [167, 169]. Yet recent work by [170] argues that dropout results in ill-posed Bayesian inference. Regardless of the mathematical equivalence to Bayesian approaches, dropout can be used to estimate uncertainty

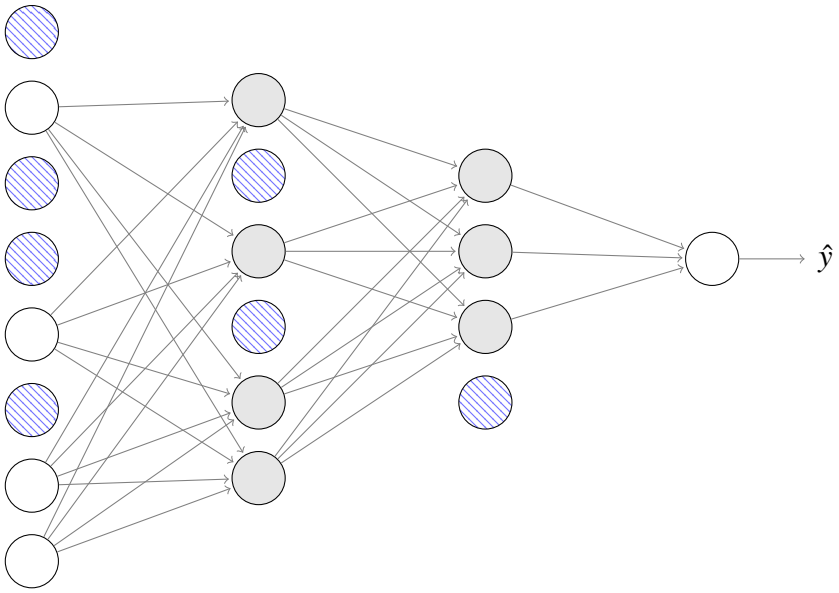

Figure 3.9: Neural network with dropout where the blue hatched nodes illustrate randomly removed nodes and connections during training. p is 0.50 for the first layer, 0.66 for the second layer, and 0.75 for the third layer. Note: the bias nodes and weights are excluded for clarity.

from NNs.

3.3.5.3 Neural networks used in this study

Three different NN architectures were used in this study. Each NN had an input later with 44 units, one unit for each basin characteristic. The DMNN and DSNN models had two hidden layers where the first hidden layer had 40 units with a 10% dropout rate and the second layer had 30 units with a 10% dropout rate. The DMNN had 27 units in the output layer corresponding to the 27 quantiles. The DSNN has one unit in the output layer. The LMNN had one hidden later with 40 units with no dropout and 4 units in the output layer corresponding to the 4 L-moments. ReLU activation functions were used for each of the hidden layers and the output layer. ReLU was used for the output later instead of a linear activation because the outputs have to be ≥ 0. All models were fit using Keras 2.1.3 [171] with a tensorflow backend. Predicting L-moments and quantiles for ungaged locations was simulated by a 10-fold cross-validation approach [164].

3.3.6 Selecting reference sites

We use a variant of the $Q_r P_r P_x Q_x$ (referred to as simply "QPPQ") [136, 137, 129] to estimate daily streamflow for an example site using the estimated FDC. QPPQ uses streamflow at gaged-reference locations (Q_r) to calculate a time series of exceedance probabilities (P_r) that are used to map probabilities (P_x) back to streamflow (Q_x) at ungaged locations (Figure 3.10). Where Q refers to streamflow, P refers to exceedance probabilities, subscript r is reference sites and subscript x is the site of interest.

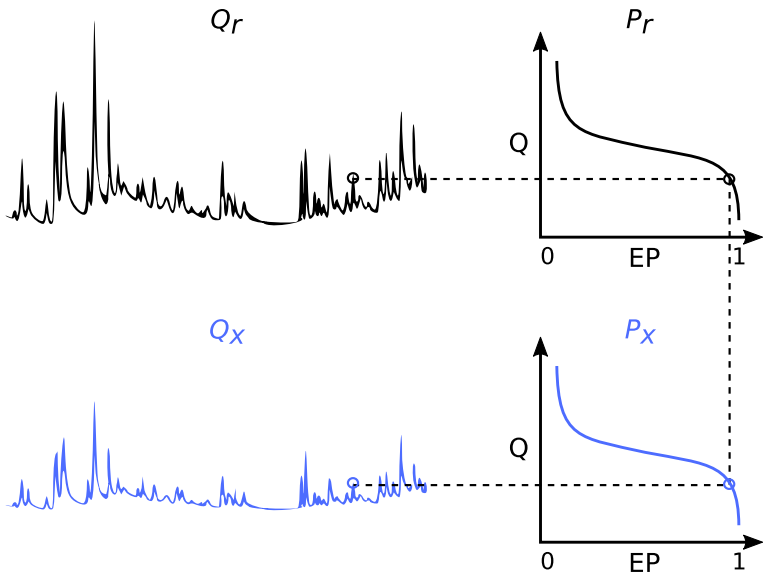

Figure 3.10: Schematic of the QPPQ method. The black hydrograph (Q_r) and FDC (P_r) represent the reference gage and the blue FDC (P_x) and hydrograph (Q_x) represent the estimated site.

The final step in QPPQ is selecting a reference site to donate daily exceedance probabilities (EPs) to the estimated FDC at the ungaged location. The FDC at the ungaged location can then be used to convert EPs to streamflow. The assumption is that the EP for the reference site, P_r, on a given date is equal to the EP for the ungaged site, P_x, on the same date. It is important to note that the streamflow for the reference site, Q_r, is only used to select EPs and is not directly used to calculate the streamflow at the ungaged site Q_x. The

Desiderata for choosing reference gages include the following,

1. The two sites must be close enough geographically to capture the same storm events. This is particularly important for estimating sub-weekly streamflow.

2. The basin characteristics of the reference gage must be similar to the basin characteristics for the ungaged location. This should ensure similar hydrologic responses.

3. The method used to select reference gages must not directly depend on measured streamflow as this data will not be available when estimating flow for ungaged locations.

Multiple methods have been proposed to choose a reference site [136, 137, 129]. We propose a simple two-step method. (1) Subset candidate reference sites for each decade by selecting all sites within 150 km radius of the site of interest (i.e, the ungaged site), and then (2) use Euclidean distances in the 40 dimensions of predictor space to select a final reference site from among those subsetted sites in step 1. These steps are executed per decade to account for differences in active streamgages and changes in mutable basin characteristics (Figure 3.11). This method does not require streamflow information (i.e., it is unsupervised) and can be used to locate reference sites for ungaged locations. The estimated FDC at the target site maps the donated probability from the reference site to a discharge value. A local regression model (LOESS) is used to predict discharge values for quantiles other than the 27 that were directly predicted.

3.4 Results

3.4.1 Monotonic Violations

The number of monotonic violations were calculated for each decade by,

$$\varepsilon_{mon} = \sum_{i=1}^{n} \left[\sum_{k=1}^{m} (cummax(\hat{y_k}) \neq \hat{y_k}) \right]_i \tag{3.25}$$

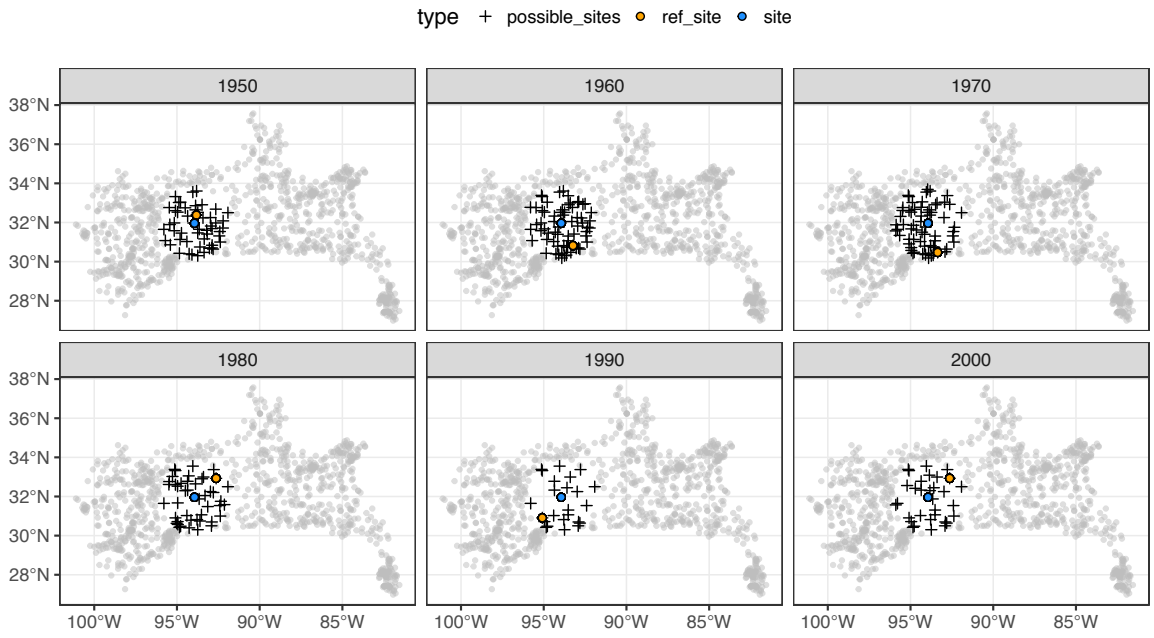

Figure 3.11: Example of two-step method to select reference sites per decade for site number USGS-08023080.

where i is the number of site, k is the number of quantiles, and $\hat{y_k}$ is the predicted quantiles. This definition of monotonic violations accounts for each quantile for site and decade combination rather than only adjacent quantiles. The predictions are first sorted and then each is checked against the cumulative max, where each quantile should be equal to the cumulative max. This means that if the first 5 quantiles were 1.5, 3.0, 1.4, 1.8, 2.5 ft^3s^{-1}, there would be a total of 3 monotonic violations because 1.4, 1.8, and 2.5 are not the cumulative max (because 3 ft^3s^{-1} comes previously) although they are increasing monotonically. The DMNN model produces substantially lower monotonic violations than the DSNN model (Figure 3.12). Most of the violations for the DMNN model are due to very small differences. For example, for site number 07343500 in 1970 the DMNN predicted 0.18 ft^3s^{-1} for the 0.02 quantile and 0.17 ft^3s^{-1} for the 0.05 quantile. The largest absolute difference is for site number 07348500 in 1950, where the DMNN predicted 507,974 ft^3s^{-1} for the 99.95 quantile and 483,460 ft^3s^{-1} for the 99.98 quantile. By definition, the LMNN-AEP model never violates monotonicity. The DSNN model is not used for further analysis.

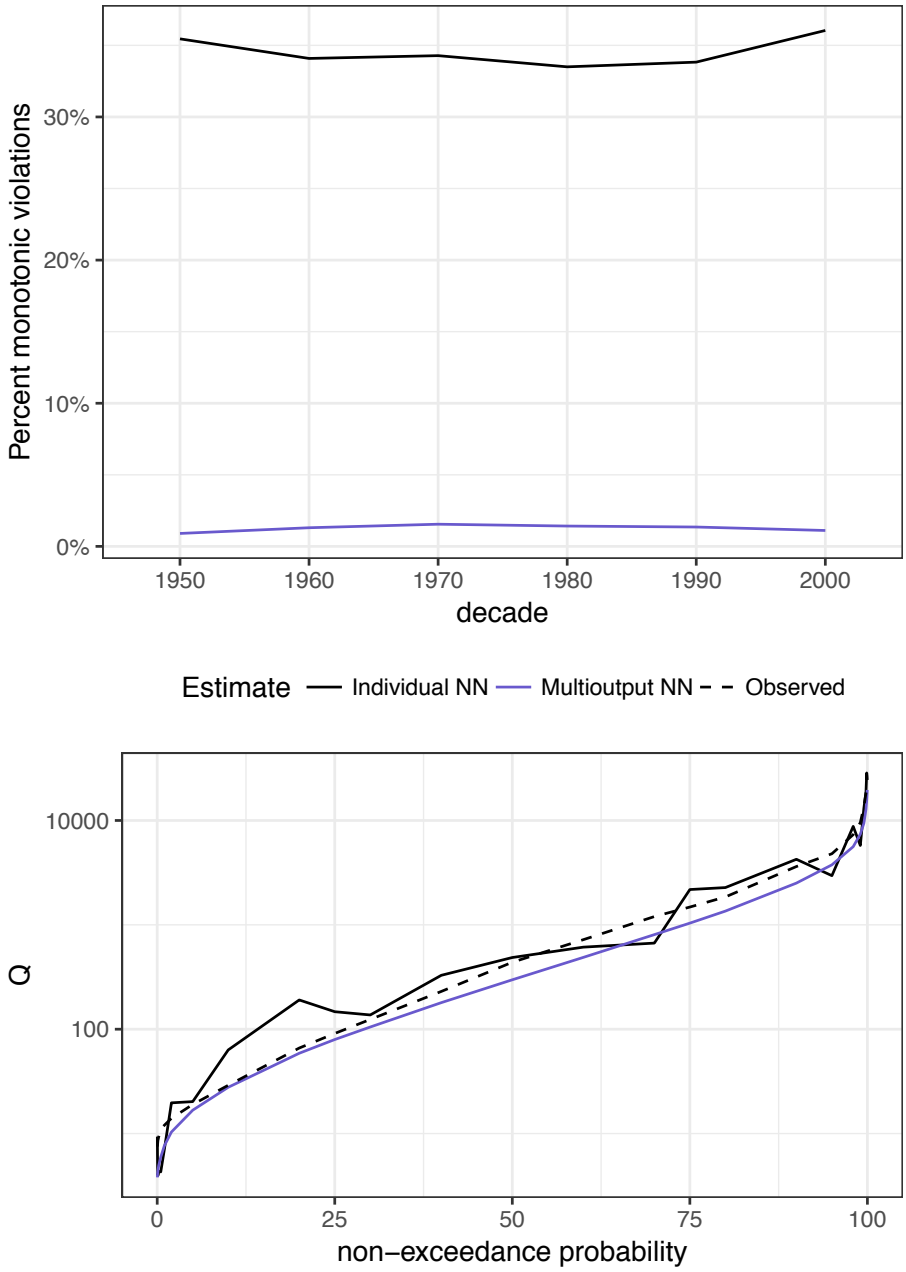

Figure 3.12: [*Top*] Percent of monotonic violations per decade for the DMNN and the DSNN. The number of sites change per decade (Table 3.1) but the average number is $\sim$ 470. There are 26 possible violations (number of quantiles -1) per site, meaning there are around 26*470 = 12,220 possible violations each decade. [*Bottom*] Example quantiles predicted by both the DMNN an the DSNN.

3.4.2 Comparing FDC predictions

The LMNN model is used to predict the L-moments (Figure 3.13) that parameterize an AEP distribution, and the non-exceedance AEP probability function is used to predict the 27 quantiles. The AEP predictions are compared to the DMNN model. On average, The DMNN model tends to perform better towards the tail of the distribution while the AEP model performs better in the middle of the distribution (Figures 3.14 and 3.15). The AEP predicted a large number of negative values that were set to zero, and the effect of this can be seen in the lower tail of the predictions sharply dropping to zero. This can also be seen the in the 100% median percent errors for the the smaller nonexceedance probabilities for the AEP (middle row, Figure 3.15). The DMNN model describes over 70% of the variance for all quantiles and decades, often describing over 90% of the variance. The AEP model describes over 50% of the variation for all quantiles and decades, and also often describes over 90%, but shows a greater range in it's correlations over nonexceedance probabilities (top row, Figure 3.15). The normalized RMSE (n-RMSE) shows a similar trend to the median percent error. The predicted lower tail for AEP model is over 3x smaller than the actual values (the exponent in the RMSE calculation causes only positive RMSE values while the actual difference can be positive or negative). Both models have smaller n-RMSE values moving towards larger non-exceedance probabilities. The results for the non-zero estimates show that for small nonexceedance probabilities, the AEP overpredicts and the DMNN underpredicts (Figure 3.16). The AEP model also overpredicts the larger nonexceedance probabilities while there is less bias in the DMNN model. There is not an obvious spatial trend to the error for the 50th percentile estimates (Figure 3.17).

3.4.3 Dropout uncertainty intervals

Multiple iterations of dropout can be used to generate uncertainty intervals for predictions (Figure 3.18). The width of the interval is a measure of how certain the model was

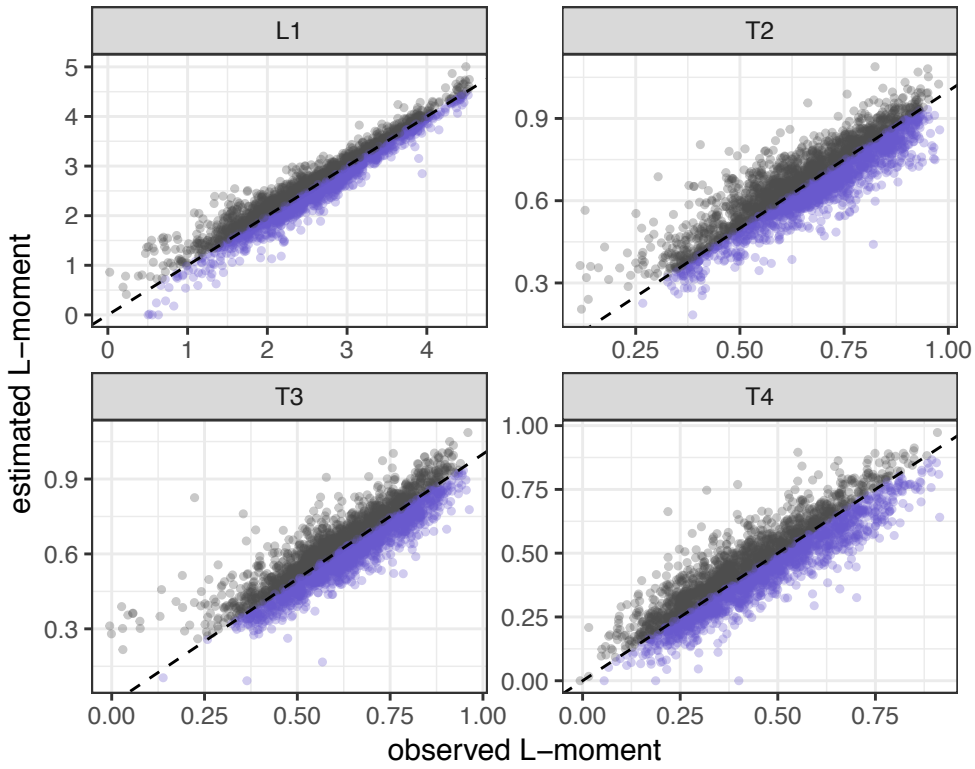

Figure 3.13: Predicted vs observed L-moments from the LMNN model.

in it's predictions. If the interval is small, e.g., site numbers 08047000 and 08091000, then the predictions are robust to stochastic perturbations in the model architecture. If the interval is large, e.g., site numbers 0233100 and 0243800, the model learns very different weights, and therefore makes very different predictions, depending on which hidden units are present.

3.4.4 Streamflow predictions

Although predicting streamflow is not the focus of this study, we provide one example estimation for a full period of record (Figure 3.19). The estimated streamflow overpredicts the observed streamflow on average (mean Q estimated: 1,331 ft^3s^{-1} and mean Q observed:1,040 ft^3s^{-1}), has a higher standard deviation (standard deviation Q estimated: 2,083 ft^3s^{-1}, standard deviation Q observed: 1,780 ft^3s^{-1}), but similar ranges (range Q

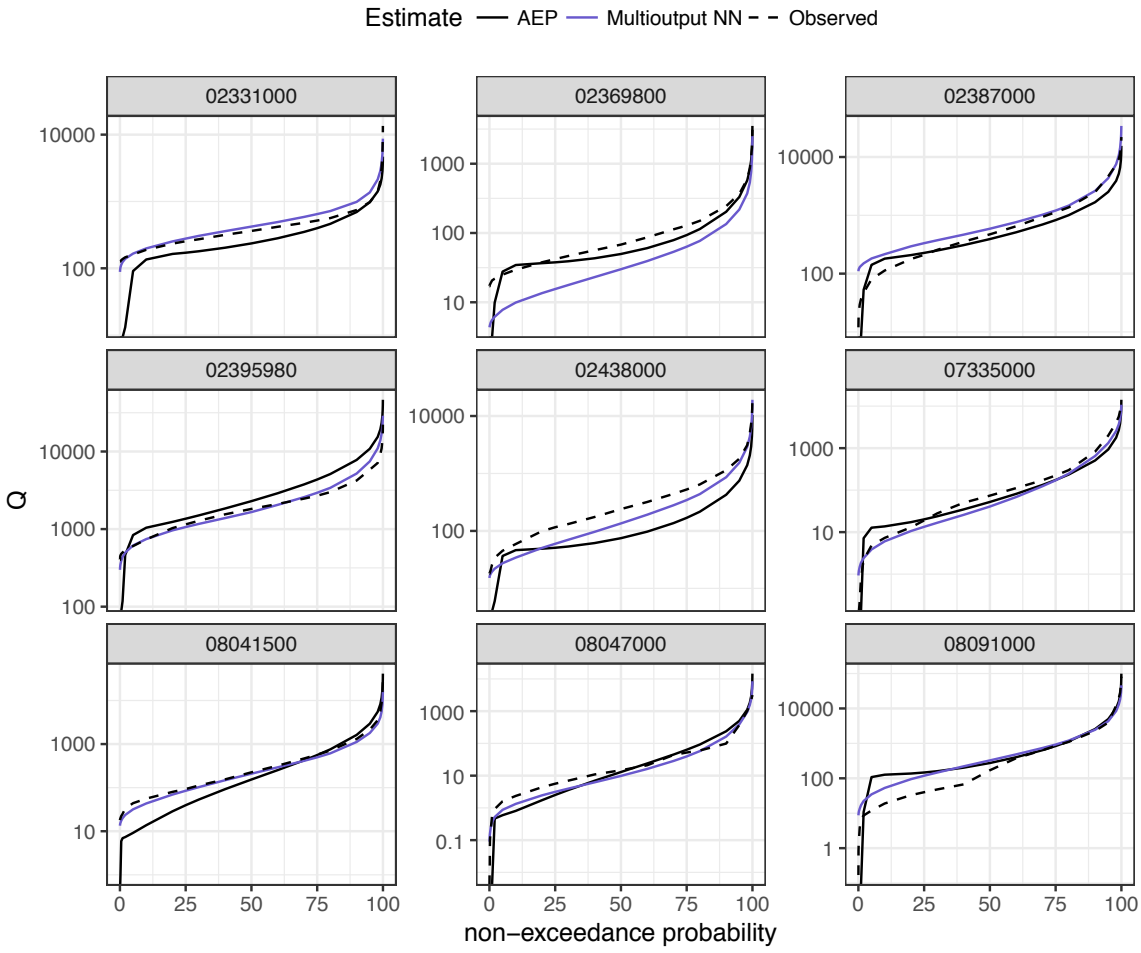

Figure 3.14: Observed quantiles vs predicted quantiles for both the direct multi-output NN the AEP distribution parameterized by the multi-output NN estimated L-moments. The sites were randomly selected. Each panel is for a specific site and decade.

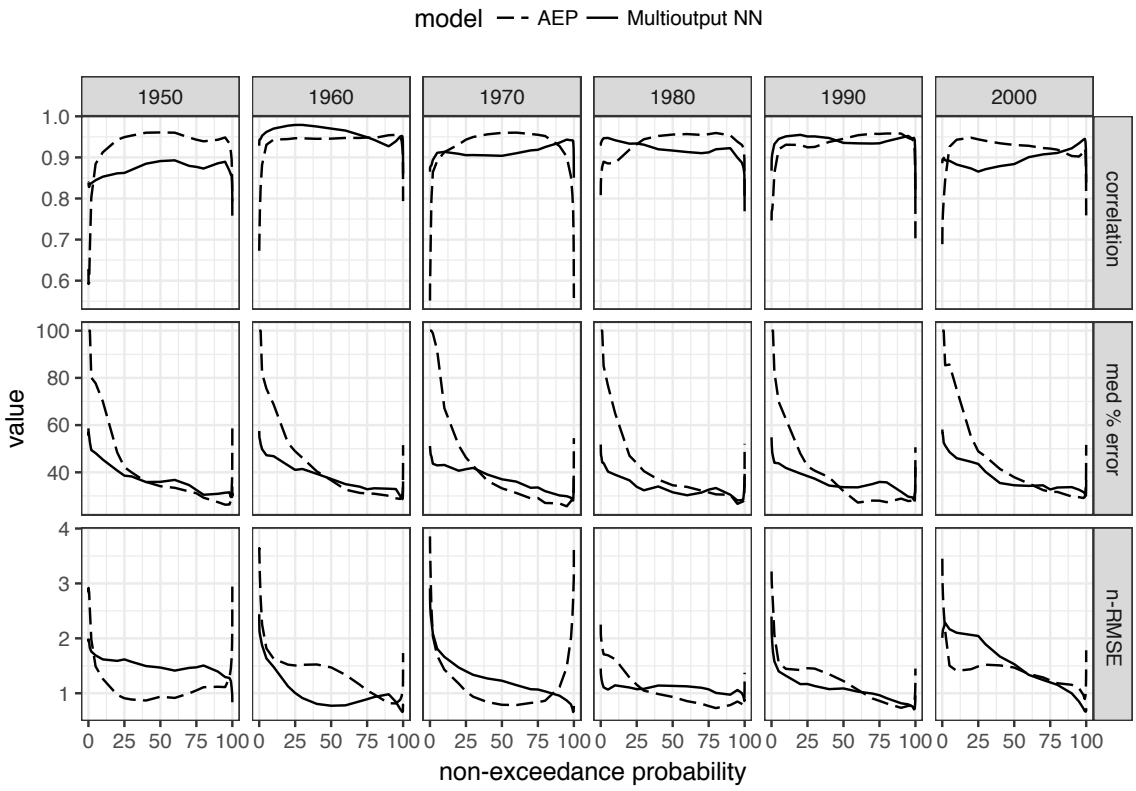

Figure 3.15: Comparison of correlation (*correlation*), median percent error (*med % error*), and the normalized root mean squared error (*n-RMSE*), where the RMSE is divided by the mean observed value for each non-exceedance probability.

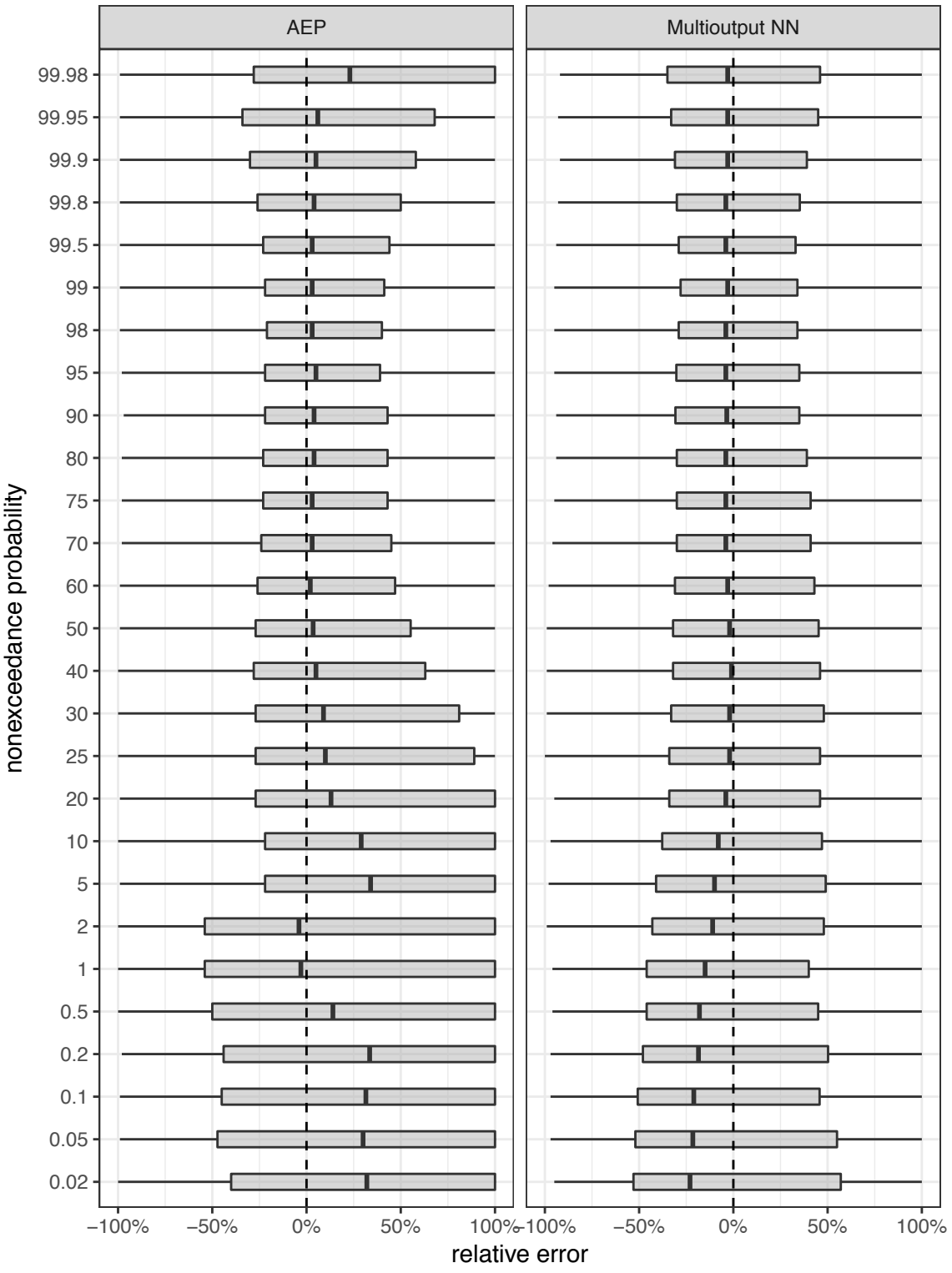

Figure 3.16: Relative percent errors for each non-exceedance probability and model. Zero values were removed for the plot. Positive percent errors greater than 100% were set to 100%.

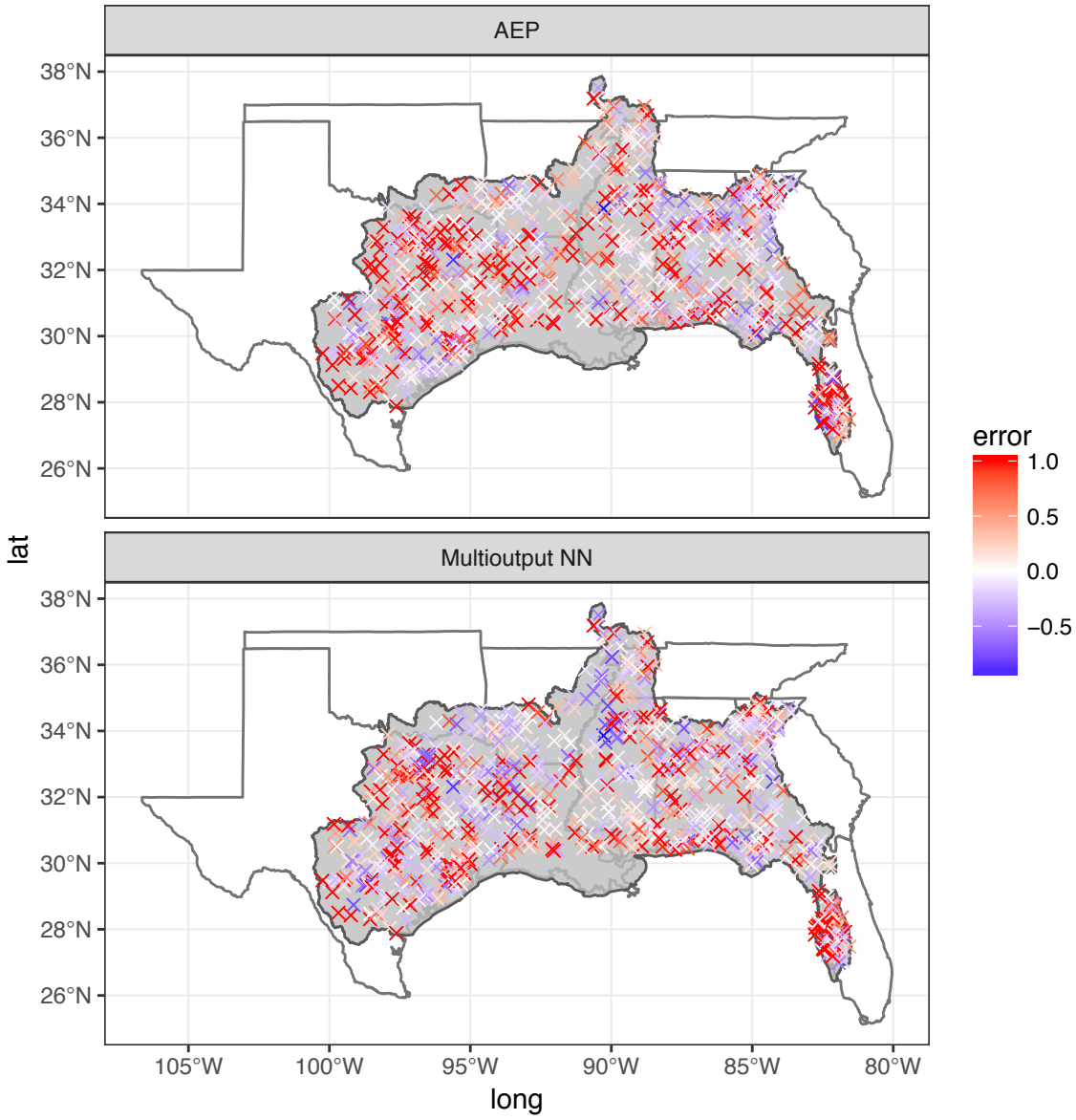

Figure 3.17: Map of relative percent errors the 50th non-exceedance probability and each model. Zero values were removed for the plot. Positive percent errors greater than 100% were set to 100%.

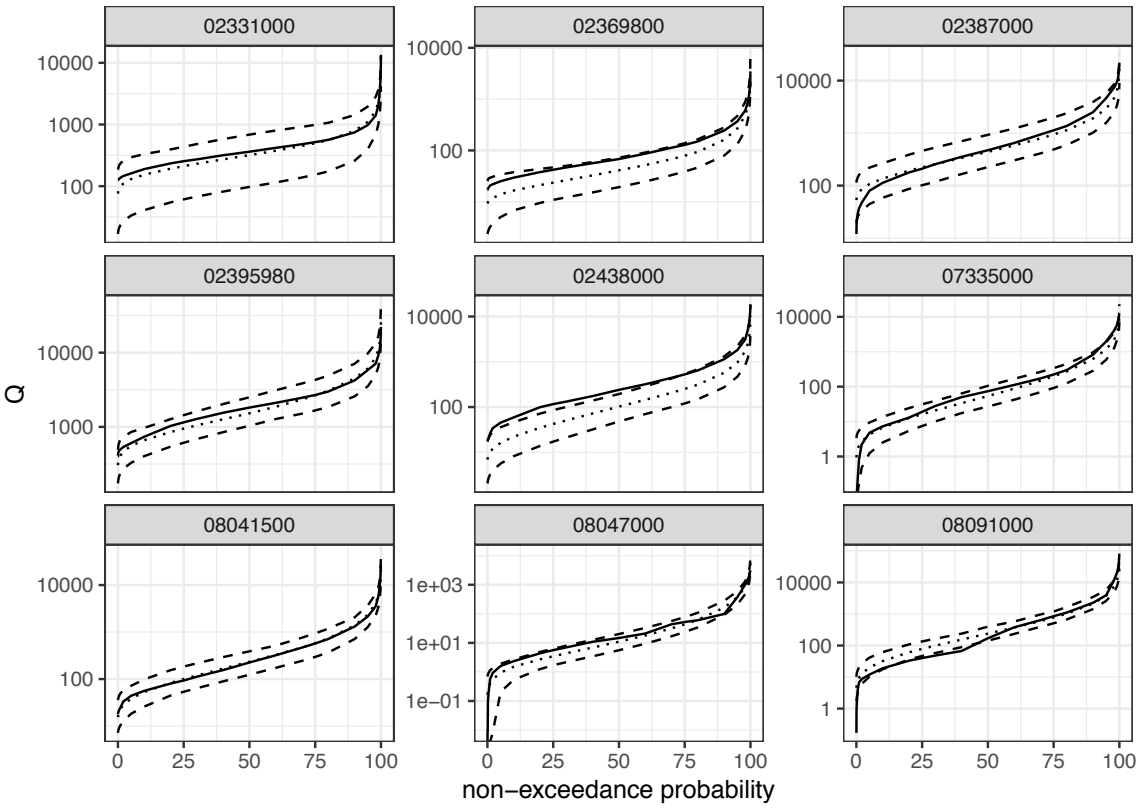

Figure 3.18: Using 100 iterations of neural network dropout to obtain uncertainty estimates for the DMNN model. The dashed lines are the minimum and maximum estimates and the dotted line is the mean. The sites shown are the same sites shown in Figure 3.14.

estimated: 6-32,776 ft^3s^{-1}, range Q observed: 17-34,300 ft^3s^{-1}).

3.5 Discussion

Multioutput neural networks produce improved predictions when there are consistent relationships between multiple response variables for each observation [164]. This can be seen when comparing the output of the DSNN, which predicts each quantile independently, and the DMNN model, which predicts all 27 quantiles simultaneously (Figure 3.12). The quantiles for a FDC for given stream both covary and are monotonic. Knowing something about one quantile provides significant information about other quantiles. A single-output model ignores this relationship and learns model weights and biases (i.e, parameters) that minimize a loss function for each quantile independently. Whereas the multi-output model shares information between all the inputs and outputs to find parameters that minimize a combined loss function. For the DMNN model used in this study, each quantile in the output layer shares parameters with every other quantile in the output layer. The only non-shared parameters are the last connections between each unit in the last hidden layer and the output layer. Parameter-sharing between the outputs helps to estimate smooth and internally consistent FDCs.

Theoretical probability functions, such as the 4-parameter Asymmetric Exponential Power distribution (AEP) used in this study [145], also produce realistic FDCs and only require regionalizing 3-4 statistical moments rather than a large number of individual quantiles. Additionally, after the distribution is parameterized it can generate streamflow values for any nonexceedance probability without the need of an interpolation technique (like the LOESS model for the DMNN). It can also extrapolate into the tails of the distribution, whereas a direct quantile-estimation approach is constrained within the range of quantiles used in the regression (although it is possible to extrapolate beyond the range, it is not advised [135]). However, theoretical distributions also have their drawbacks. There are theoretical constraints on their L-moments that can be violated when the L-moments are

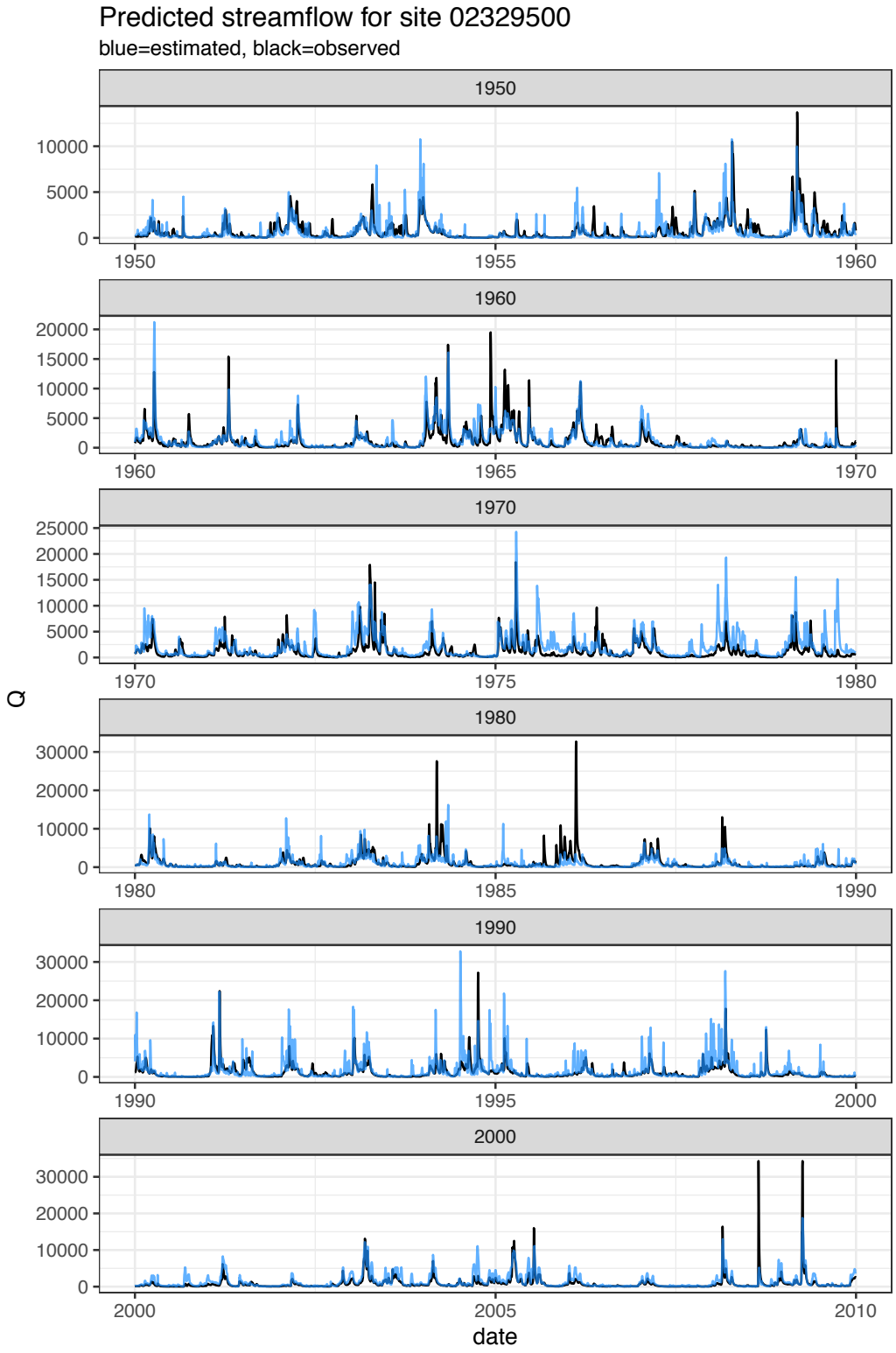

Figure 3.19: Example of cross-validated streamflow estimation using the DMNN model and QPPQ.

regionalized. A distribution cannot be fit to the moments when this occurs [131]. Due to this type of violation we were unable to produce AEP estimates for 3.8% (n=107) of the 2,807 site-decade combinations. Theoretical distributions tend to result in less accurate representations of FDCs at ungaged sites than Direct-quantile estimation methods [134] (Figures 3.16 and 3.14).

Monte-Carlo neural network dropout offers a simple and rigorous method for calculating a posterior-predictive distribution for each quantile [169]. Back propagating the error after randomly dropping a number of hidden units prevents the active units from co-learning relationships unique to the training set. This resistance to overfitting leads to more generalizable models. The range of predicted values provide additional information about the confidence of the model for a given observation. A large range indicates that small changes to the internal architecture of the network leads to large changes in the predictions. The range may also change along the FDC, suggesting that the model is robust to dropout for some quantiles and is more sensitive for others. The width of the prediction intervals indicate how certain the model is about the predictions, yet this does not ensure that the predictions are actually closer to the observed value. It may also be that predicted FDCs near the tails of the posterior distribution are closer to the observed FDC (e.g., site number 02438000 and 02369800 in Figure 3.18).

Predicting streamflow was not an objective of this study but we included an example of how to complete the QPPQ process. Nonexceedance probabilities from reference sites can be used to generate streamflow predictions using the estimated FDCs at ungaged sites (10-fold cross validation was used to simulate ungaged sites in this study). The reference sites can be chosen using many different methods [172, 90, 76], and the best method is dependent on the goal of the study. The method used in this study leveraged both spatial and covariate proximity to select a reference site each decade (Figure 3.11). This method could be improved in several ways. First, sites with small Euclidean distances in predictor space do not necessarily have similar daily nonexceedance probabilities. A method that

accounts for the relationship between predictors and FDC similarity would likely provide a better mapping between the reference FDC and the target FDC. For example, Ganora et al., [173] used the pairwise similarity between dimensionless FDCs and related the distances to basin descriptors. Second, our method uses the raw probability from the reference site directly without accounting for the decaying relationship between sites due to distance (both geographical and covariate distance). Copula-based methods would offer a rigorous treatment of this relationship [174].

Data-driven models can be improved by integrating domain knowledge into model development [9]. In this study we show that multi-output neural networks that directly predict a number of quantiles can estimate FDCs that obey monotonicity between increasing quantiles nonexceedance probabilities. Physically-consistent FDCs can also be generated using distributional methods. The simplicity and flexibility of the distributional methods make them useful models, although the directly estimated FDCs are more accurate across quantiles than distributional methods. Regardless, multi-output neural networks offer significant advantages over other regression modes when regionalizing L-moments or individual quantiles.

CHAPTER 4

EXPLORING THE DRIVERS OF PUBLIC-SUPPLY WATER USE USING HIERARCHICAL-BAYESIAN MODELS

4.1 Abstract

This study explores the relationship between municipal water-use and an array of climate, economic, behavioral, and policy variables across the contiguous U.S. The relationship is explored using Bayesian-hierarchical regression models for over 2,500 counties, 18 covariates, and three higher-level grouping variables. Additionally, a second analysis is included for 83 cities where water price and water conservation policy information is available. A hierarchical model using the nine climate regions (product of National Oceanic and Atmospheric Administration) as the higher level groups results in the best out-of-sample performance, as estimated by the Widely Available Information Criterion, compared to counties grouped by urban continuum classification or primary economic activity. The regression coefficients indicate that the controls on water use are not uniform across the nation: e.g., counties in the Northeast and Northwest climate regions are more sensitive to social variables, whereas counties in the Southwest and East North Central climate regions are more sensitive to environmental variables. For the national city-level model, it appears that arid cities with a high cost of living and relatively low water bills sell more water per customer, but as with the county-level model, the effect of each variable depends heavily on where a city is located.

4.2 Introduction

There are growing concerns that population growth in many regions is leading to unsustainable demands on water supply systems [175, 176], particularly in locations likely to face reduced availability due to changing climate [177, 178, 179, 180]. In the United States,

withdrawal projections suggest that 70% of counties could experience increased water supply risks in the next 50-100 years [177, 178]. Others argue that attempts to reduce CO_2 emissions and stem climate change could result in greater stress on U.S. water systems than the stress caused by climate change itself, e.g., increased water requirements to grow biofuels [181]. Regardless, the impacts of future climate and population growth on water supply security can only be determined if examined in conjunction with other controls of domestic water use. This paper explores the spatial variability in national public-supply water use by assessing the relationship between water use and an array of climate, economic, behavioral, and policy variables and how these relationships vary across the contiguous U.S (CONUS).

More than 280,000,000 people in the U.S. are served by some 150,000 utilities, 80% of which are classified as small (population served $<$ 3,500 people [182]). The 2010 US Geological Survey's national water-use compilation indicates that total municipal water-supply withdrawals in the United States have decreased by 5% between 2005 and 2010 despite a growing population [183], which seems to reflect at least a partial decoupling of withdrawals and population.

The national picture is not uniform for all regions, with domestic withdrawals varying from 55 gpcd to 167 gpcd (gallons per capita per day) across states [183]. Despite this wide variation, per capita water consumption in the U.S. is high relative to absolute needs. The minimum drinking water requirement for human survival is less than 1 gpcd [184]. The requirement is closer to 15 gpcd if the water needed for sanitation, bathing, and food preparation is included [185]. Actual indoor water use in the U.S. is a factor of four higher than the minimum requirements (about 60 gpcd [186]), while current per capita domestic water withdrawals are estimated to be 89 gpcd [183] and has remained greater than 60 gpcd throughout twentieth century [187, 188]. Clearly the per capita use in the U.S. reflects an aggregate of complicated environmental, social, financial, and motivational influences on decisions of end users [189].

A substantial amount of research has been dedicated to exploring the controls on water use. Increased water use is often associated with higher temperatures and lower precipitation [190, 191]. This is mainly attributed to the increased summertime lawn and garden irrigation and the need to replace evaporated water in pools [192, 193, 194]. Household age is generally thought to be associated with increased water use due to older fixtures and a higher propensity for older pipes to leak [195, 194], but other studies suggest that neither efficient water use appliances nor the age of households have a substantive effect on overall use [196, 197]. The effect of household income is also unclear [198]. Households with more disposable income often have larger lawns and are less sensitive to water price, which can lead to greater water use [199, 200]. Alternatively, income and education are often positively correlated, and education level tends to be associated with environmental awareness [201], which in turn, can lead to conservative water-use habits [202, 203, 204, 205]. Low density housing (e.g., suburban and rural) tends to be associated with higher per-capita withdrawals than high density housing (e.g., urban apartments), primarily due to outdoor water use [206]. Water-use restrictions, rebates, water conservation education, and behavioral factors–beliefs about the environment, institutional trust, etc–are generally thought to reduce water use [207, 208].

The price of water also affects use, although water prices are set well below the long term marginal cost of water supply [209]. Price elasticities for domestic water use have been estimated to be in the range of -0.45 and -0.14. [210, 211, 194]. These elasticities can vary widely depending on both pricing structure and household preferences. For example, increasing block tariffs leads to nonlinear price elasticities as households are asked to pay more for the marginal price of water as the volume purchased increases [212], and end user preferences, such as the percent of outdoor water use, result in different elasticities for different homes [209].

Recent research by [202] examined the change in county-level water use efficiency across the U.S. by climate region between 1985 and 2010. This was one of the first studies

to examine trends in efficiency and their underlying drivers on a national scale. They attribute much of the temporal variability in efficiency to state-level North-South gradients and county-level difference in rural vs. urban areas, education levels, and income. They use climate regions as the main spatial unit of aggregation for analysis. Despite the important contribution of [202], there remain several open questions: Are climate regions the most logical unit of analysis for water-use efficiency? What other grouping variables might be important? Do the negative effects of income and education persist after controlling for other explanatory variables? Do urban areas still show greater water-use efficiency after controlling for inter-county water transfers (i.e., counties buying and selling water across county borders [213])? Are drivers of water use consistent if the data are aggregated for cities rather than counties?

This paper builds on the work in [202] to explore the controls on municipal water withdrawals in the CONUS by considering how they are related to climate, economic measures, demographics, policies, and behavior variables. We address the specific questions above by exploring two methods of controlling for inter-county water transfers, leveraging a hierarchical model that allows information sharing between higher-level groups, exploring three grouping variables, including multiple predictor variables in addition to income and education, and considering two separate modeling exercises using county-level withdrawals and city-level demand for the CONUS. We consider different sets of proxies for climatic, economic, social, and policy conditions as covariates for the county and city models, subject to data availability. The county-level and city-level models are built on independent but related datasets, which provides a unique opportunity to test the robustness and scalability of conclusions regarding the primary controls on public water use. We find that precipitation and temperature, partisan voting behavior, the number of people per household, the median age of the structures in a county, and the average water price are variables that explain the most variability in water withdrawals and water deliveries.

4.3 Methods

4.3.1 County-level data

Models of 2010 county-level water withdrawals were constructed using 18 explanatory variables (Table 4.1) for over 2,500 counties. To represent local climate, we used county precipitation, temperature, and overall water yield. Water yield accounts for soils, vegetation cover, wind speed, and other landscape factors in addition to precipitation and temperature [214]. As proxies for economic and behavioral information we included demographic measures of income, the Gini index (measure of income disparity), college attainment, Cook Partisan Voting Index (measures local Republican vs Democratic voting proportions for presidential elections relative to national average), median age, rural to urban class, and rent vs own proportions. The county-level water use data were taken from the 2010 US Geological Survey (USGS) water use compilation [183] (description of dependent variables are provided below). The covariate data was primarily taken from publicly available datasets (Table 4.1). If two predictors were highly correlated ($\rho > 0.4$), then one of the predictors was removed from the dataset and is not shown in (Table 4.1). For example, the correlation coefficient between the percentage of a county below the poverty line (*ppoverty*) and the *Gini index* is 0.57, and so *ppoverty* was removed from the analysis. The full water-use dataset for the contiguous U.S. included 3,109 counties and county equivalents (e.g., Louisiana parishes). When the water-use data were merged with the explanatory variable dataset, 110 counties were dropped because explanatory variables were unavailable. The full dataset for our analyses consisted of data for 2,999 counties and county equivalents. For the national county-level model, direct information on water price, household behavior, fraction of outdoor use, inter-county water transfers, and water conservation policy was not available. Therefore, we conducted an extended analysis for 83 cities where policy, transfers, and price variables were available (Table 4.1). Descriptions of the city-level covariates are in Section 4.3.4.

The full datasets and the R scripts used to process that data are available in a USGS data release [1]. Further information can be obtained by contacting the corresponding author.

4.3.2 Grouping variables and hierarchical-Bayesian models

Fixed-effects models were compared to hierarchical models (also commonly referred to as multilevel models, partially-pooled models, random-effects models, varying-effects models, or adaptively-regularizing models [217, 218, 219, 220]), both of which were estimated in a Bayesian framework. The primary difference between these models relates to how they handle possible clustering in the observations that define unique water-use settings. Fixed-effects models address possible clustering in one of two ways. In a "fully pooled" fixed-effects model, any clustering by group is ignored, and a single, fixed estimate of the coefficient for each covariate is developed using all of the observations. Conversely, in an unpooled fixed-effects model, separate coefficient estimates are developed only using the observations in each group. A hierarchical model provides a compromise between these two extremes. Hierarchical models extend single-level regression to data with a nested structure, whereby the model parameters vary at different levels in the model, including a lower level that describes the actual data and an upper level that influences the values taken by parameters in the lower level. Here, observations indexed by $i = 1, ..., n$ are clustered within two or more groups, $j = 2, ..., J$, that define unique water use settings and the variation in lower-level parameters by group is of interest. The structural dependency of lower-level parameters across groups is defined by prior distributions that characterize their joint parameter space, and these prior distributions have their own parameters (called hyperparameters) that exist in the upper level of the model. Both upper and lower-level parameters are estimated jointly, thereby sharing information between levels of the model and partially pooling parameter estimates across groups [221]. Hierarchical-Bayesian models are very similar to random-effects models (multi-level models fit by using some form of maximum likelihood estimation) but are more flexible in terms of their ability to propagate

Table 4.1: Explanatory and grouping variables considered in this study. More information detailing the data sources for each variable can be found in [1]. Abbreviations used in the table: max=maximum, T=temperature, P = precipitation, pop=population, cons=conservation, °C=degrees Celsius, and mm = millimeters.

County covariate	**Description**	**Data source**
pgrowth	proportion population growth from 2000-2010	census[a]
prop_sw	fraction of withdrawals from surface water	USGS[b]
Qmm	daily water yield	[214]
tmax_40	mean annual max T, 1970-2010	PRISM[c]
tmax_diff	2005-2010 mean annual max T - tmax_40	PRISM
ppt_40	mean annual P, 1970-2010	PRISM
ppt_diff	2005-2010 mean annual P - ppt_40	PRISM
med_income	median household income	ACS[d]
Gini	Gini index	ACS
pcollege	proportion of county with some college education	ACS
house_dens	houses per square mile	ACS
ppl_house	average number of people per household	ACS
med_age_struc	median age of household structures	ACS
prent	fraction of pop that rents residence	ACS
psfh	fraction of single family homes	ACS
papt	fraction of apartments (10+ units)	ACS
cook_pvi	Cook Partisan Voting Index	this paper[e]
rur2urbi	rural to urban (1-9) index	USDA[f]
City covariate	**Description**	**Source**
water_price	price of water	AWWA[g]
bill_type	bill structure (decreasing, uniform, or increasing)	AWWA
reb	number of rebate oriented water cons. policies	[215]
req	number of requirement oriented water cons. policies	[215]
aridity	$P/(T+33)$ with P in mm and T in °C	this paper[h]
rpp	regional price parity	BEA[i]
Grouping variable	**Description**[j]	**Source**
climate_region	nine climate regions that groups states by climate	NOAA[k]
rur2urb	nine urban continuum codes	USDA
econdep	six non-overlapping categories of economic dependence	USDA

[a]Census data: https://www.census.gov/data/developers/data-sets.html
[b]USGS data: http://water.usgs.gov/watuse/data/2010/usco2010.txt
[c]Parameter-elevation Relationships on Independent Slopes Model (PRISM) data:
http://www.prism.oregonstate.edu/documents/PRISM_datasets.pdf
[d]American Community Survey (ACS) data: https://www.census.gov/data/developers/data-sets.html
[e]Positive % = more Democratic than national average for 2004 and 2008 presidential elections.
[f]2003 U.S. Department of Agriculture (USDA) data:
https://www.ers.usda.gov/data-products/rural-urban-continuum-codes/
[g]American Water Works Association (AWWA) Water and Wastewater Rate Survey data:
https://www.awwa.org/store/productdetail.aspx?productid=61841567
[h]Köppen aridity index, more details can be found here [216]
[i]Bureau of Economic Analysis (BEA): https://www.bea.gov/regional/downloadzip.cfm
[j] See Figures 4.12, 4.13, and 4.14
[k]National Oceanic and Atmospheric Administration (NOAA) data:
https://www.ncdc.noaa.gov/monitoring-references/maps/us-climate-regions.php

uncertainty through complex model structures [222].

Importantly, it is not clear *a priori* whether to partially pool, and if so, across what grouping variables. Therefore several grouping variables were explored for the hierarchical models in this study (Table 4.1). Climate regions were included because it allows for the comparison of parameters after partially controlling for long term climate of an area (Figure 4.12). The rural-to-urban gradient was included because it was reported as a major driver of water use in [202] (Figure 4.13). The primary economic dependency of a county was included to explore possible effects of end users purchasing water in each county (Figure 4.14). There are multiple structures that can be used for the varying effects determined by these grouping variables: the intercepts can be partially pooled with fixed slopes, the slopes can be partially pooled with fixed intercepts, or both the intercepts and slopes can be partially pooled. We compare each of these structures. To avoid confusion between the various model names, for the remainder of this paper we refer to fixed effects models as ”fully pooled” (ignores higher level groups) or ”unpooled” (separate models for each higher-level group) and hierarchical models as ”partially pooled” (estimates coefficients for each higher-level group while sharing information between groups via a joint prior).

4.3.3 County-level analysis

When developing the county-level regression, it is essential to control for undocumented inter-county water transfers–an important confounding factor when data is aggregated to scales larger than the footprint of the supply system [213]. For example, if a rural county with a small population is exporting water to an urban county with a large population, the per-capita water withdrawal will suggest high usage in the rural county and low usage in the urban county, when in fact the values are just artifacts of the water transfers. Furthermore, dividing by population in counties where water is being either exported or imported can greatly over-estimate the actual per-capita water use in the county with a small population and only slightly under-estimate the per-capita water use for a county

with a large population. Additionally, the covariates associated with a county selling water are partially exported to a county buying water (i.e., an increase in water price for a county importing water may decrease withdrawals in the county exporting it), and for regression modeling, this results in having the rows of the design matrix not always corresponding with the correct rows of the response variable. This will hinder the interpretation of regression coefficients. Lacking an up-to-date database of inter-county water transfers for the U.S., we address this issue using two different modeling strategies: (A) drop counties with values outside of a threshold taken from the literature and model only the remaining counties using a linear regression model, and (B) use the expected national per-capita withdrawal value to classify counties with a binary response variable (greater or less than expected) and model each county outcome using a logistic regression model.

4.3.3.1 Linear regression model for county-level water withdrawals

The response variable, *wh*, for the county-level linear regression model is the annual gallons of public supply freshwater withdrawn (column "PS-WFrTo" in [183]) per unit household in a given county (Figure 4.1). The number of households served in each county was calculated as: $(pop_{served}/pop_{total}) * houses_{total}$ where population served (pop_{served}) and total population (pop_{total}) for a county was obtained from the USGS water use dataset, and total number of houses per county ($houses_{total}$) was obtained from the American Community Survey. Due to undocumented inter-county water trading, only counties with *wh* values of 50,000 $<$ wh $<$ 300,000 were retained for the analysis (dropped 453 counties) using the linear regression model. These thresholds are consistent with a range of values for eleven municipalities across the U.S. [223, 224]. A manual inspection of specific counties suggests that this method accurately removes counties with known transfers. For example, the counties in the Catskills in NY provide water to New York City and are removed, counties in Western MA that provide water to Boston are removed, and counties surrounding Atlanta GA and large metro areas in Texas are removed. However, this approach does not

remove every county with known water transfers, such as Los Angeles county in CA.

The regression coefficients were estimated using Bayesian regression models with a Gaussian likelihood function:

$$wh_i \sim Normal(\mu_i, \sigma) \tag{4.1}$$

$$\mu_i = X_i \beta_{j[i]} \; for \; i = 1..., N \; \& \; j = 1..., J \tag{4.2}$$

$$\sigma \sim HalfCauchy(0, 1e6) \tag{4.3}$$

where i is the county index, j is the group index, N is the number of counties, J is the number of groups, X is the design matrix, β is the vector of coefficients, μ_i is the within group means, and σ is the within group prior standard deviation. A normal distribution was assumed for the response variable *wh*. A Half-Cauchy distribution was chosen as a prior for the standard deviation because it is constrained to only positive values (standard deviations are only positive) and offers weak regularization for large standard deviation values [225]. The scale parameter for the Half-Cauchy prior was selected by cross validation. This model is developed for each grouping variable (climate region, urban gradient, economic dependency) to compare how model estimation and prediction change based on the grouping structure. We also developed a fully-pooled and unpooled models, where $J = 1$ and β for each covariate was assigned an independent $N(0, 1e5)$ prior distribution and α was assigned a $N(0, 1e6)$ prior distribution (where α is the intercept and is estimated by assigning a vector of ones in the first column of the design matrix). The choice for these priors was based on cross validation and these priors are weakly regularizing. The coefficients that describe μ_i in Equation 4.2 are drawn from a Multivariate-Normal (MVN) distribution,

$$\beta_j \sim MVN(\nu, \Sigma) \; for \; j = 1..., J, \tag{4.4}$$

where ν is the vector of parameter means, and Σ is the parameter covariance matrix. The

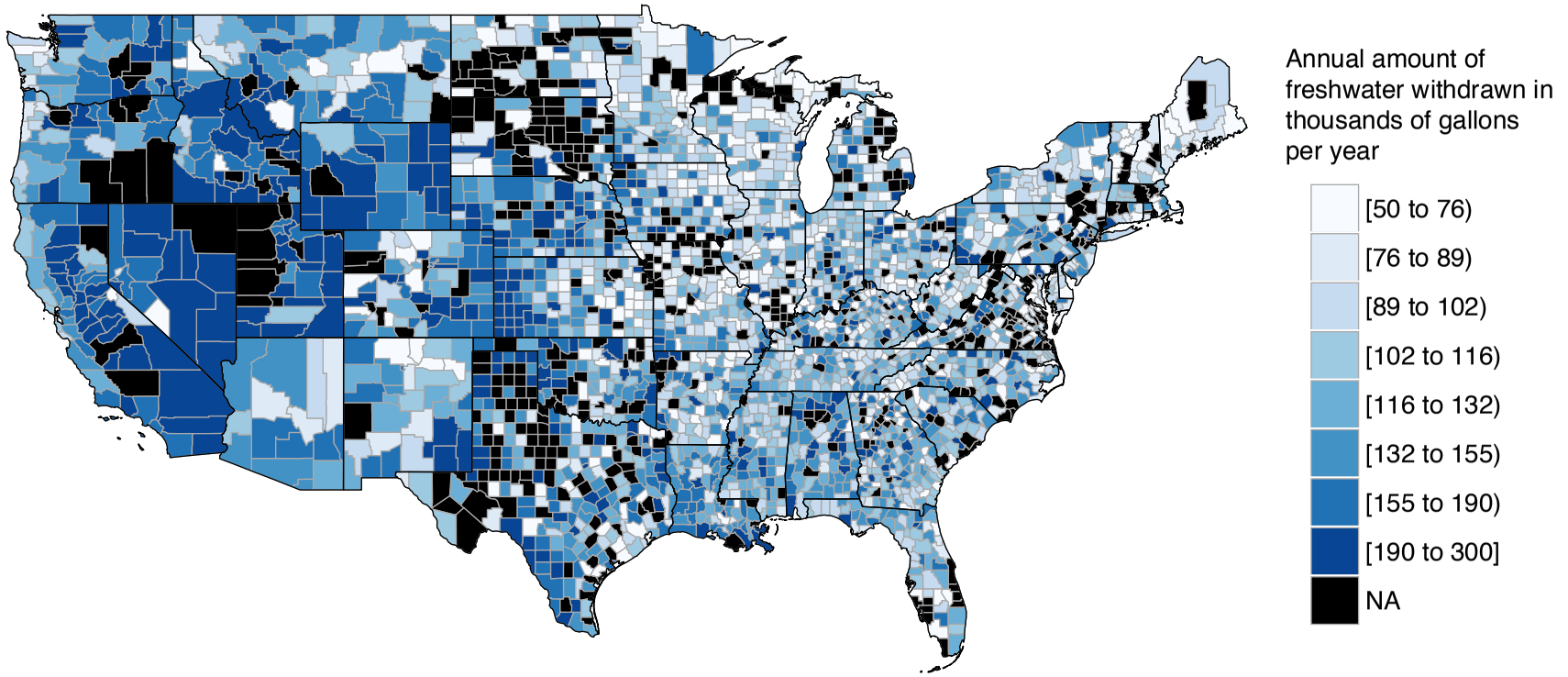

Figure 4.1: Map of response variable, *wh*, for county-level linear regression model for 2546 counties. *wh*, is the annual gallons delivered to a household for a given county. Only counties with *wh* values of 50,000 ¡ wh ¡ 300,000 were retained for the analysis, which resulted in dropping 453 counties (NAs in the figure). See the text for more details.

multivariate normal prior in Equation 4.4 was vectorized using non-centered parameterization to allow for more efficient Markov chain Monte Carlo (MCMC) sampling. The covariance matrix was defined as a quadratic function of a correlation matrix, Ω, and a vector of coefficient standard deviations, τ, each with their own priors:

$$\tau \sim HalfCauchy(0, 1e6) \tag{4.5}$$

$$\Omega \sim LKJ(2) \tag{4.6}$$

The prior for τ is the same prior that was used for σ and is described above. The correlation matrix Ω was assigned a weakly regularizing *LKJ* prior over the correlations [226]. All covariates were converted to z-scores before building the models.

4.3.3.2 Logistic regression model for water-withdrawal classification

The response variable used in the county-level logistic regression model, dw, is calculated as the difference between a county's actual withdrawal and the national population normalized withdrawal expectation:

$$dw_i = w_i - \left[\frac{\sum w_i}{\sum pop_i} * pop_i\right],\ for\ i = 1..., N, \tag{4.7}$$

where $\sum w_i / \sum pop_i$ is the national population normalized withdrawal expectation, and dw_i, pop_i, and w_i are the withdrawal departures, population, and raw withdrawals for the ith county, respectively. dw was then converted to a binomial response variable (Figure 4.2) by:

$$dw_{i\ class} = \left\{ \begin{array}{ll} 0 & if\ \ dw_i \leq 0\ \ (county\ \ withdrew\ \ less\ \ than\ \ expected) \\ 1 & if\ \ dw_i > 0\ \ (county\ \ withdrew\ \ more\ \ than\ \ expected) \end{array} \right\}. \tag{4.8}$$

The use of a binary response variable provides another way to reduce the effects of inter-basin transfers on model inference and provides some advantages over the screening approach taken for the linear regression model. Representing water use as a departure from the national average allows population to still be used for normalization while removing the non-linear effect of normalizing by counties with different population sizes. The resulting classification better reflects water transfers than the raw *wh* calculation, which can be seen in counties such as Davidson (pop=626,681, withdrawals=136.38 mgd) and Williamson (pop=183,182, withdrawals=1.81 mgd) Counties TN. Davidson County sells water to several surrounding counties, including 14 mgd to Williamson County [*personal communication*], and this is reflected in dw_{class} variable, whereas Williamson County is dropped from the *wh* analysis. Water transfers are better represented in *dw* because even small transfers from a county with a large population will result in a departure from the national average that will be on the same scale as the departure for a county with a smaller population. The downside of the binomial approach is a loss of information due to discretization of counties with highly anomalous water withdrawals that might be caused by large anomalies in covariates.

The regression coefficients were estimated using hierarchical Bayesian regression models with a binomial likelihood function and a logit link. The model for each county's *dw* is:

$$dw_i \sim Bernoulli(p_i) \tag{4.9}$$

$$p_i = logit(X_i \beta_{j[i]}),\ for\ i = 1...,N \tag{4.10}$$

$$\beta_j \sim MVN(\nu, \Sigma)\ for\ j = 1...,J \tag{4.11}$$

where p_i is the probability that *dw* for the ith county is equal to one, ν is the vector of parameter means, and Σ is the parameter covariance matrix. The multivariate varying effects

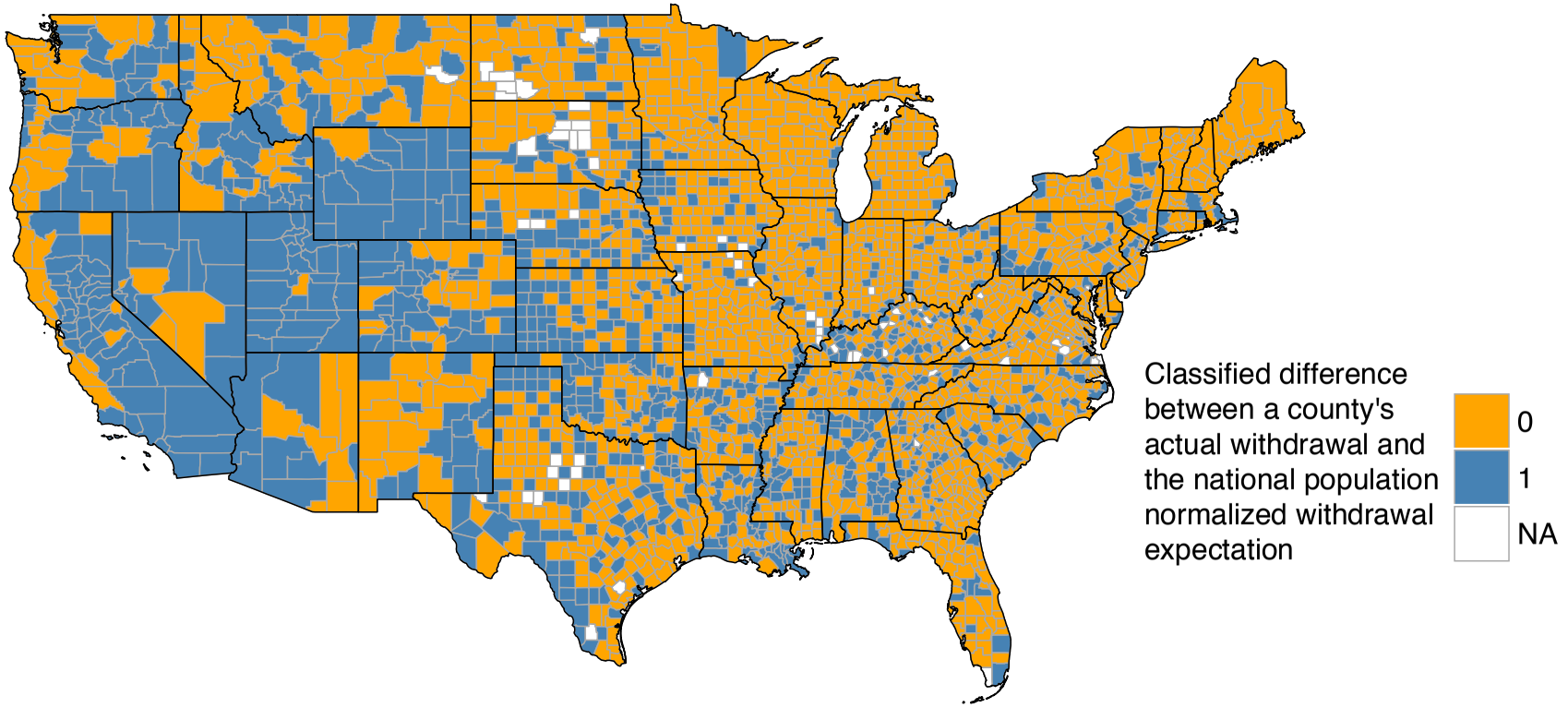

Figure 4.2: Map of county response variable for the logistic regression model (*dw* described in Equation 4.7) for 2999 counties. Orange counties withdrew less than expected and blue counties withdrew more. There were 110 counties where the full covariate dataset could not be obtained and were dropped from the analysis (indicated by NA).

prior in Equation 4.11 was re-parameterized to use a non-centered parameterization. All covariates were converted to z-scores before building the models.

4.3.4 City-level data and analysis

A second dataset was compiled that included water price and conservation measures for 83 cities (Figure 4.3). Transfers are not present in the city dataset as the dependent variable, *gal*, is the water demand rather than the withdrawals for a given city. The demand is calculated as the annual gallons sold (domestic, commercial, and industrial) by a municipality per account for a given city. The covariates included in the city analysis are the average residential monthly water bill (i.e., "water price") across all volumes for a 5/8 inch meter, the bill type, the number of requirement oriented water conservation policies, the number of rebate oriented conservation policies, the Köppen aridity index [216], and the regional price parity. The bill types were converted into a dummy variable, where decreasing block rates were coded as a one, and uniform and increasing block rates were coded as a zero. This choice was made prior to the analysis based on inspection of the response variable distribution per bill type. The city-level water use and price related variables (water price and bill type) were taken from the 2010 AWWA water and wastewater survey dataset [227]. Explanations of the rebate and requirement oriented water-conservation policies are detailed in the appendices of [215]. The regional price parity, an index which compares the cost of living for an area relative to the national average, was taken from the Bureau of Economic Analysis (www.bea.gov). The city-level data was also analyzed using a Bayesian hierarchical model with a Gaussian likelihood. Only climate regions were explored as a grouping variable for the city-level analysis, because the other grouping variables could not be applied to the city-level data (all cities are classified as urban and most of the economic dependence measures would not apply to cities, e.g., farming, mining...etc). Other potentially interesting ways to group the city-level data (by states, demographic indices, etc) could be explored in future work. The model structure mirrored Section 4.3.3.1 and

the range of values for *gal* and *wh* were similar.

4.3.5 Parameter estimation and fit metrics

We performed all regressions using Hamiltonian Monte-Carlo sampling with the Stan modeling language and functions from the *rstan* and *rethinking* R packages [228, 229, 230]. For each model, we ran 2 chains for 4,000 iterations where the first 1000 iterations of each chain were used for warm-up and parameter tuning and the remaining 3000 were sampled to calculate posterior distributions of the parameters. Sampling convergence was confirmed by calculating the Gelman-Rubin statistic and the visual inspection of trace plots [231]. The county-level models were compared using the widely applicable information criterion (WAIC) which is derived from the log pointwise predictive density of the models and can be shown to approximate out-of-sample predictive performance [232, 233]. The WAIC additionally provides a measure of model performance while accounting for the effective number of parameters used in each model.

4.3.6 Interpreting β parameters

We assessed the significance and magnitude of each regression coefficient (β) to characterize how each covariate affects water use. However, the interpretation of these coefficients is different between the linear and logistic regressions. For the linear regression models with mean centered and scaled covariates, a unit change in the response variable is associated with a one standard deviation change in the explanatory variable (Table 4.2). For example, if a β parameter is equal to -5000 for water price with a standard deviation of \$20, then the interpretation is that a \$20 increase in water price corresponds to decrease of 5000 in the units of the response variable (e.g., gal/household/year). For the logistic regression models, the upper bound of the predictive difference in probability of the response equaling one is approximated by dividing the β coefficient by four. This approximation is equal to the maximum of the first derivative of the logistic curve [219]. For example, if the

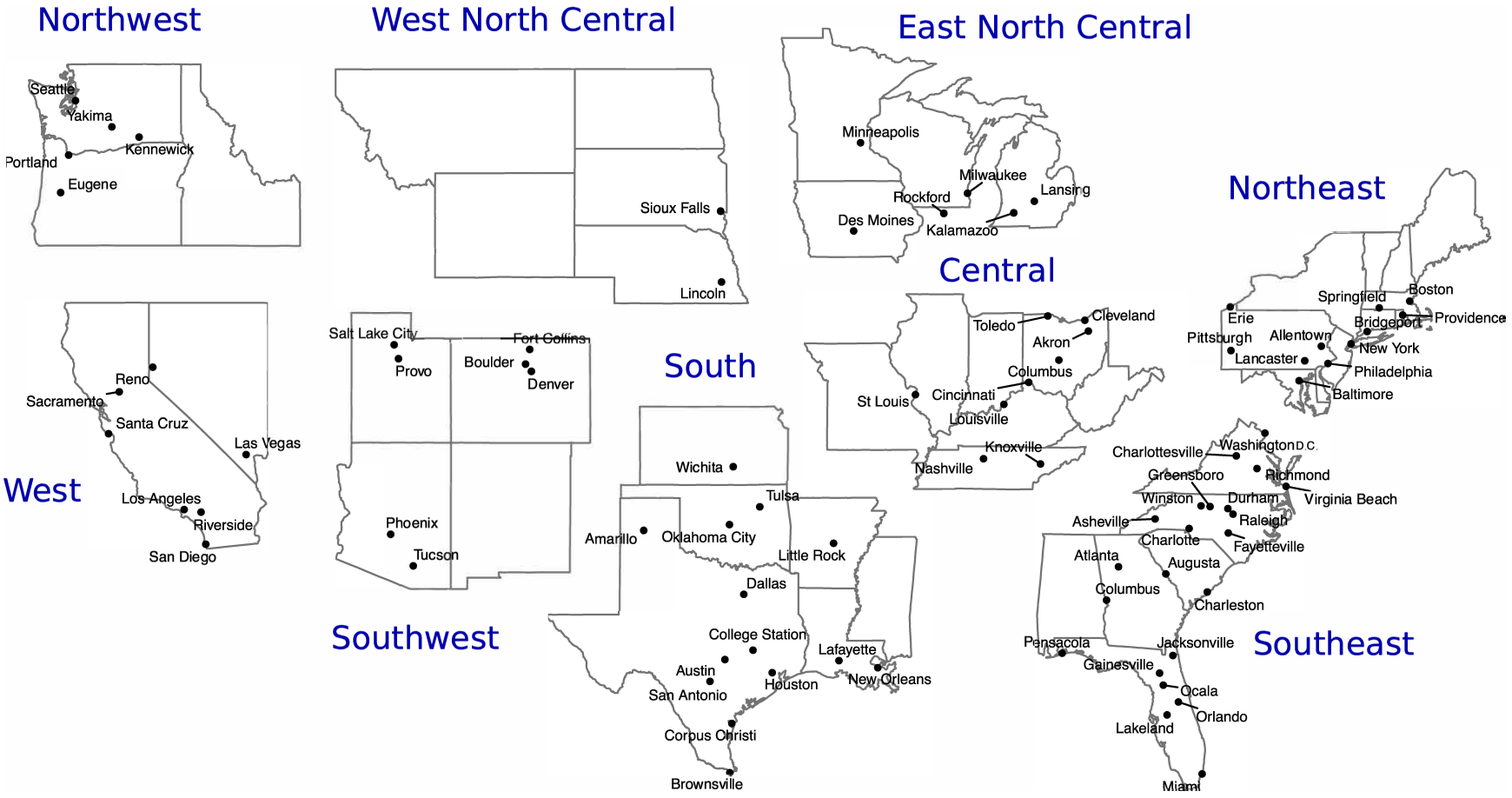

Figure 4.3: Map showing the 83 cities (by NOAA Climate region) used in the city-level analysis regression model.

β parameter is equal to -0.6 for water price, then a standard-deviation change in water price is associated with a $\sim$ -0.6/4 = 15% decrease in the probability that the response variable is equal to one.

The statistical significance of a parameter is determined by examining whether its posterior distribution overlaps zero. It is also helpful to observe the predictive accuracy of the model in a region when interpreting coefficient estimates. For the county-level models we can additionally check consistency between the linear and logistic regression models to gain additional insights. Consistent signs of parameters between the models indicate that the effect of the predictor variable is insensitive to the absolute magnitude of withdrawals. The continuous *wh* variable allows for a large range of values while the binary *dw* variable just reveals if a county's withdrawals are greater or less than the national average based on its population; if the signs are consistent, then we have added confidence in the directional effect of the predictor. If the signs are inconsistent we conclude that the actual effect of the predictor is too nuanced to be fully understood by this analysis.

4.3.6.1 Calculating variable-type importance for county-level models

In order to summarize the parameter estimates by climate region from both county-level models the parameters were grouped by climate region and model type (linear or logistic), each variable was ranked by the absolute value of the mean of posterior parameter estimate and then divided by the total sum for each model and group to create a weight for each parameter. The top 5 ranked variables were retained for each climate region and model, and the weights were then summed by climate region and variable type (environmental or social) to create an overall weight for each region and variable type. The weights for the environmental variables were divided by the total weight for each climate region.

Table 4.2: Mean and standard deviation of the response (wh) and explanatory variables used to construct models of county-level and city-level water use in the CONUS

County variable	**Mean**	**Standard deviation**
wh	126,770.80 gal/year	51,999.20 gal/year
pgrowth	0.05	0.13
prop_sw	0.37	0.42
Qmm	339.57 mm	245.73 mm
tmax_40	29.63 °C	3.07 °C
tmax_diff	0.21 °C	0.33 °C
ppt_40	92.78 cm	32.73 cm
ppt_diff	0.74 cm	10.10 cm
med_income	$22,434.83	$4,760.81
Gini	0.43	0.04
pcollege	0.19	0.04
house_dens	68.86 house/mile^2	244.58 house/mile^2
ppl_house	2.17 people/house	0.33 people/house
med_age_struc	37.93 years	11.14 years
prent	0.22	0.07
psfh	0.74	0.09
papt	0.04	0.05
cook_pvi	-10.39	12.70
rur2urbi	4.90	2.59
City variable	**Mean**	**Standard deviation**
gal	187,055.78 gal/account	87,603.94 gal/account
water_price	$32.74	$11.57
bill_type	categorical	categorical
reb	3.24 policies	3.42 policies
req	6.40 policies	5.95 policies
aridity	19.99	7.20
rpp	97.68	8.03

4.4 Results

4.4.1 County-level analysis

4.4.1.1 Model comparison

Nine different linear regression models were built for *wh* and compared using WAIC (Table 4.3). Climate regions explained more variation in *wh* when used as a grouping variable than either Economic dependency or Urban gradient. Therefore, partial pooling by climate region was employed for both the linear and logistic models to allow comparison of coefficients between the models. Partially pooled varying-intercept and slope models, where the parameters varied by climate region but shared information via a shared prior distribution, resulted in the best out-of-sample performance as estimated by WAIC. The partial pooling greatly increased the performance of the models compared to the unpooled climate region model (i.e., no information shared between regions) and the fully-pooled model (i.e., ignoring differences between regions).

Table 4.3: Comparing performance for the household normalized water withdrawal models (*wh* linear regression models). Based on the results from the *wh* model, climate region was the only grouping variable considered for the logistic regression and city-level model so the results could be compared directly.

Grouping variable	**Model**	**ΔWAIC**[a]	**Parameters**[b]	**Rank**
Urban gradient	varying α^c only, partially pooled	751.1	8.3	9
Econ dependency	varying α only, partially pooled	706.2	8.0	8
Climate region	varying α only, partially pooled	452.1	11.4	7
No group	fixed α and β, fully pooled	195.5	23.4	6
Econ dependency	varying α and β, partially pooled	166.3	79.5	5
Urban gradient	varying α and β, partially pooled	152.4	92.9	4
Climate region	varying α and fixed β, partially pooled	114.4	32.2	3
Climate region	varying α and β, unpooled	60.3	163.8	2
Climate region	varying α and β, partially pooled	0	106.2	1

[a]Difference in WAIC value from top ranked model
[b]Effective parameters = sum of variance of log-likelihood of y_i for each sample from posterior
[c]α is the intercept.

4.4.1.2 Environmental variables

The sign of the β parameters for environmental variables (i.e., sign of the mean of posteriors) switches zero times between the linear and logistic fully pooled models (Figure 4.4), and 7 times out of 45 pairs for the partially pooled models (Figures 4.5, 4.6, and 4.7), indicating fairly good agreement on the effect of each environmental parameter for each climate region. The 40-year precipitation is the largest fixed effect predictor for each model and indicates that an increase of 32 cm of average annual rainfall is associated with a decrease of 13,000 gallons of water withdrawn per year for the linear model and a 14% decrease in the probability that a county's population normalized withdrawals is greater than the national average for the logistic model, after controlling for the effect of all other parameters. However, the effect of average precipitation is not uniform for all regions. The relationship is three times stronger for the Southwest than the national average as represented by the fully-pooled model but is nearly zero for the South, Southeast, Northeast, Northwest, and the Central regions. The 40-year average maximum temperature is associated with an increase in withdrawals for the fully pooled models (stronger for the logistic model) but has a much smaller effect for most of the regions in the partially pooled models. For the Southeast and East North Central regions, an increase of 3 °C in maximum temperature is associated with a water-use increase of 10,000-15,000 gal/household and a 25-35% increase in the probability that dw is greater than the national average for a particular county. Departures from the 40-year average precipitation (ppt_diff) and temperature (tmax_diff) have smaller effects than estimates of average conditions, suggesting that interannual climate fluctuations are less important than the average conditions, at least when comparing water withdrawals across regions.

4.4.1.3 Social variables

The sign of the social β parameters switches 2 times out of 5 pairs between the linear and logistic fully pooled models (Figure 4.4) and 33 times out of 45 pairs for the partially

pooled models (Figures 4.5, 4.6, and 4.7), indicating fairly poor agreement on the effect of each social parameter for each climate region. For the fully pooled linear model, an increase of one person per household is associated with an increase of 4,000 gal/household but is zero for the logistic regression model. The relationship holds for the West North Central, West, Southwest, South, Northeast, and Central climate regions. For the partially pooled logistic regression model, a 42% increase in surface-water fraction of the overall supply portfolio corresponds to a 12.9% increase in the probability that a county withdraws more than the national expected value. A 13% increase in the Cook PVI (more Democratic votes in presidential elections) is associated with a 7,500 gal/household decrease in water use for the national model and a decrease of more than 10,000 gal/household in the West North Central, West, Southwest, and Northwest climate regions. Cook PVI has a smaller but mostly negative effect for the binomial model. An increase of 7% in the proportion of renters in the South, Southwest, and Northwest corresponds to an increase of over 5,000 gal/household and 5% in the probability that a county withdraws more than the national expected value for the binomial model.

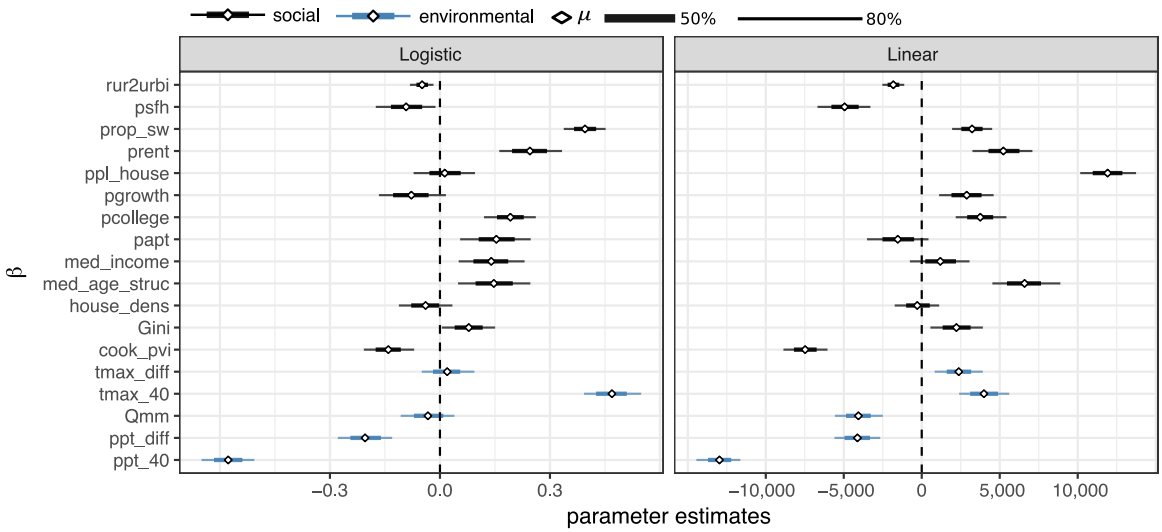

Figure 4.4: 1) Fully-pooled fixed β-parameter estimates for the [left] logistic regression (dw) and [right] linear regression (wh) models. The thick lines represent the 50th percentile interval of the posterior distribution. The thin lines represent the 80th percentile interval of the posterior distribution, and the white diamonds represent the mean (μ) of the posterior distribution.

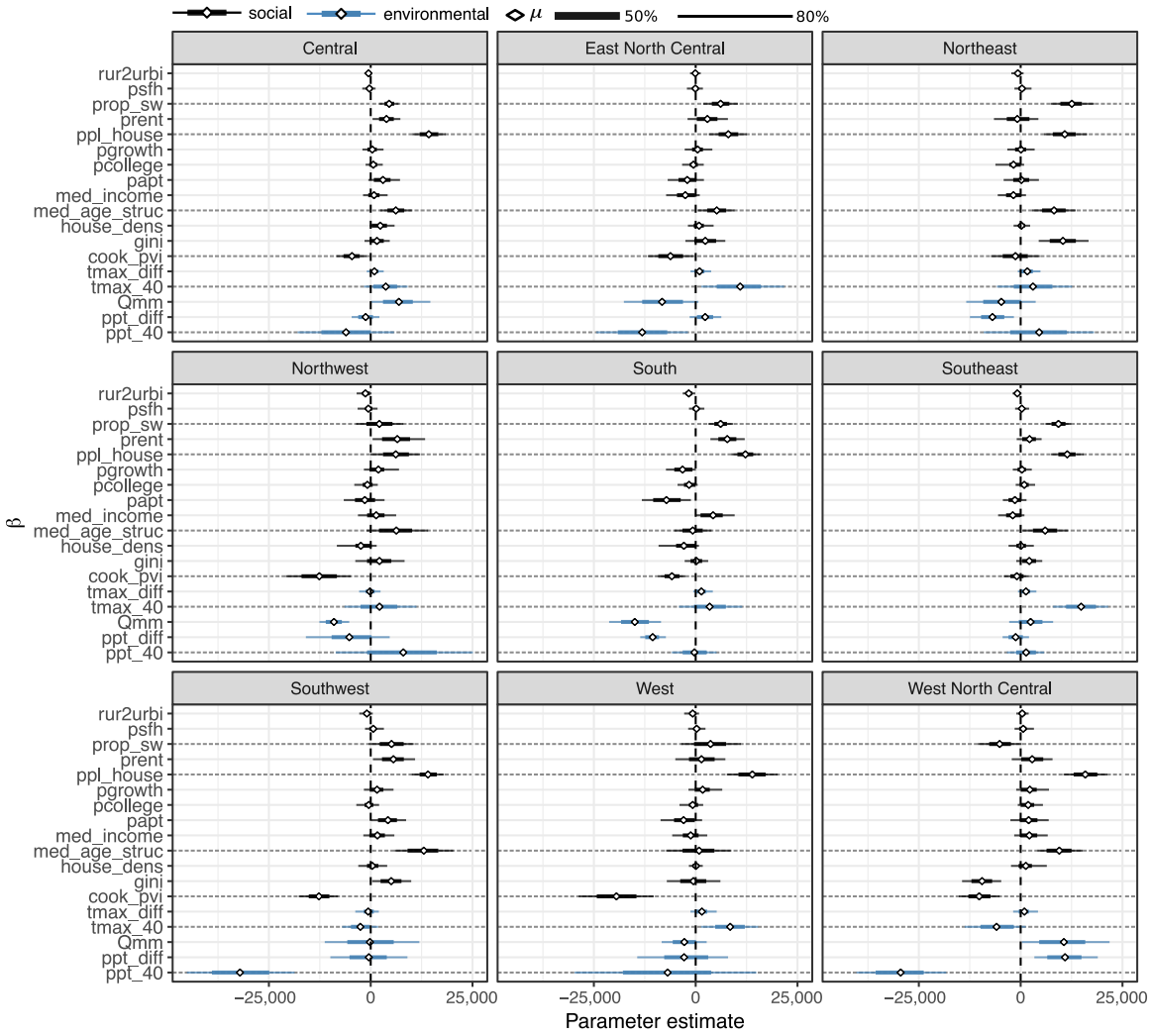

Figure 4.5: Partially pooled β parameter estimates for the linear regression (*wh*) model grouped by climate regions. The thick lines represent the 50th percentile interval of the posterior distribution. The thin lines represent the 80th percentile interval of the posterior distribution, and the white diamonds represent the mean (μ) of the posterior distribution. The most significant covariates are indicated by dashed horizontal lines.

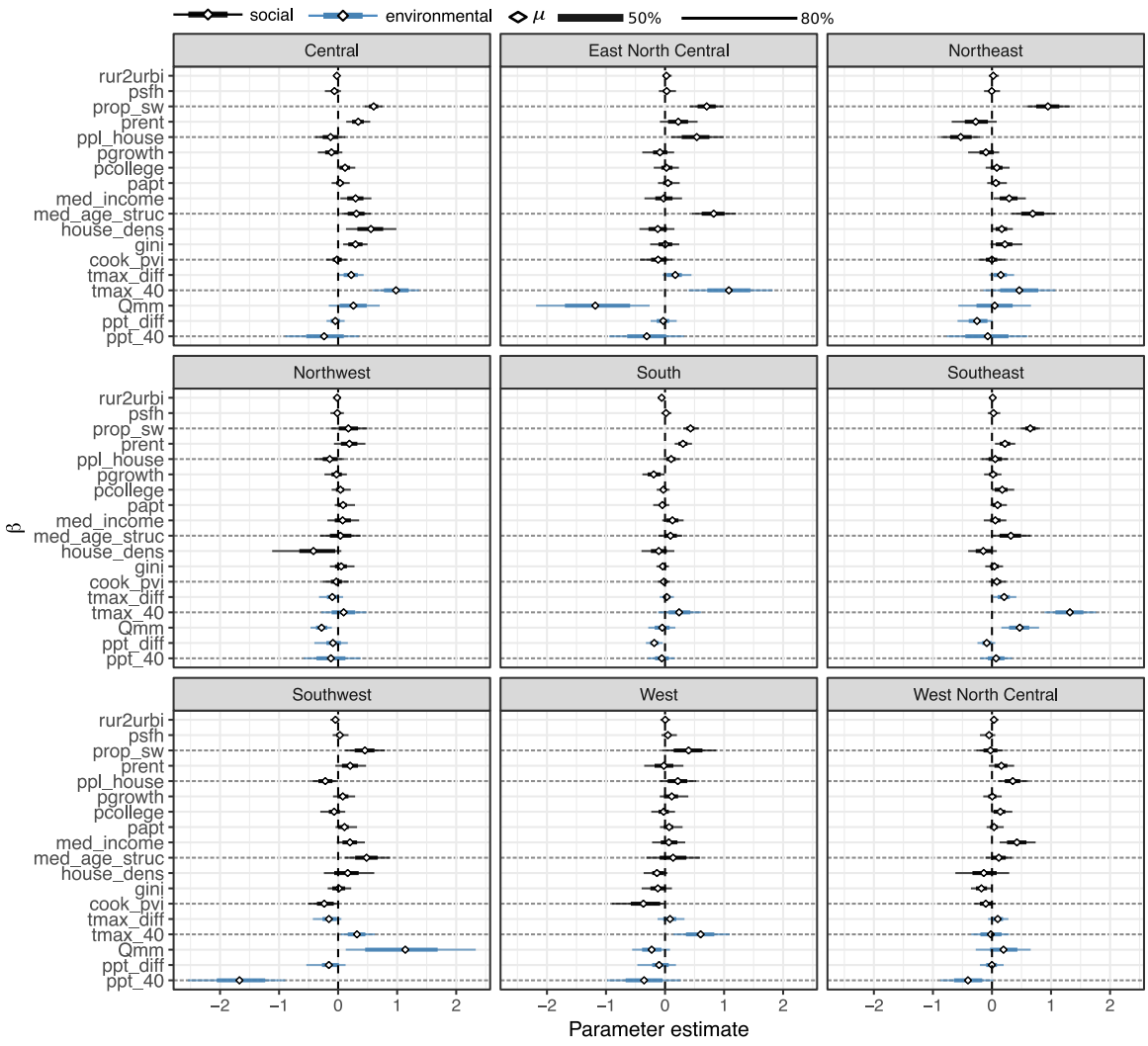

Figure 4.6: Partially pooled β parameter estimates for the logistic regression (dw) model. The thick lines represent the 50th percentile interval of the posterior distribution. The thin lines represent the 80th percentile interval of the posterior distribution, and the white diamonds represent the mean (μ) of the posterior distribution. The most significant covariates are indicated by dashed horizontal lines.

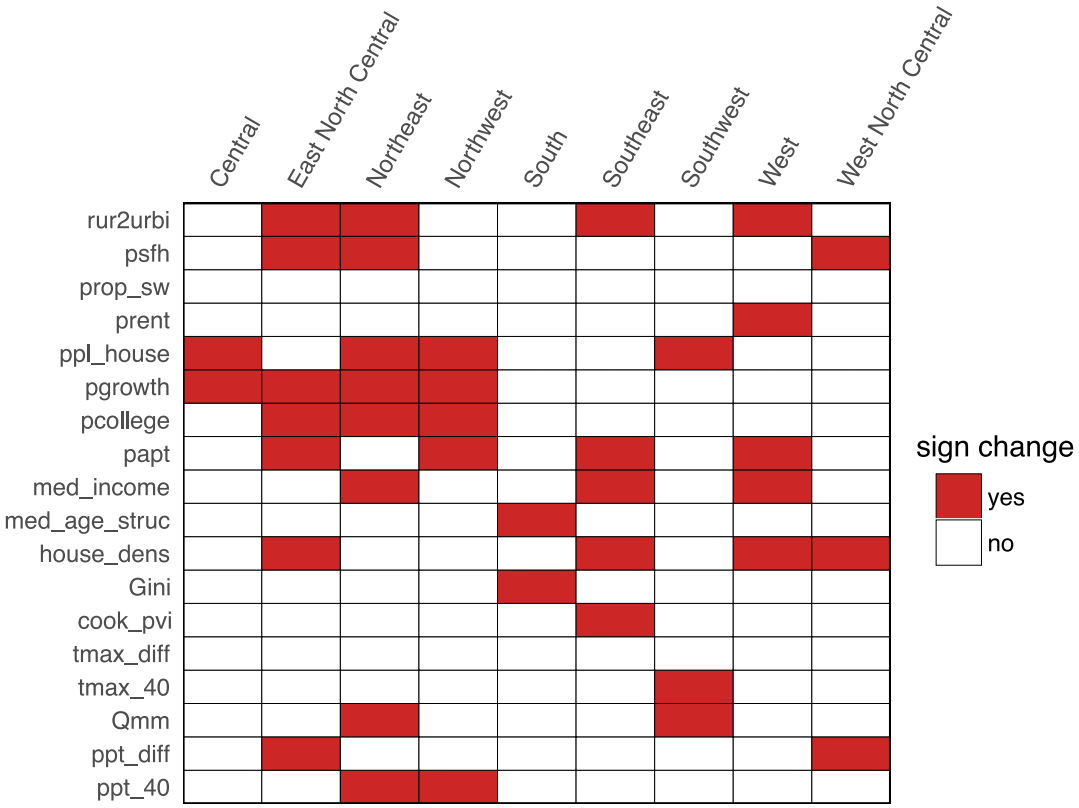

Figure 4.7: Change in signs of the mean of posterior distribution in the pair-wise comparison between the county-level linear and logistic regression models.

4.4.1.4 Ranked and scaled variable-type importance

The Northwest, Northeast, and Central regions show the strongest correlation with social variables. The Southwest, East North Central, West North Central, and the Southeast show the strongest correlation with environmental variables. The West and South regions are influenced equally by both environmental and social variables (Figure 4.8).

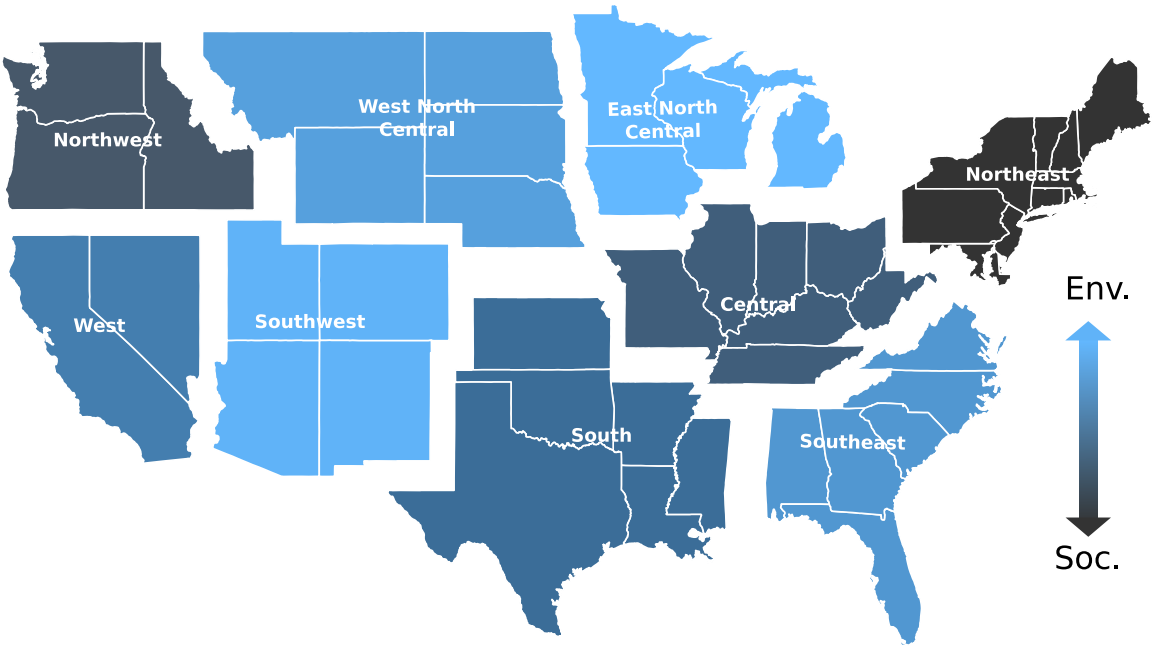

Figure 4.8: Relative influence of environmental and social variables from both of the county-level models.

4.4.1.5 County-level predictions

The most accurate predictions for the linear model were for counties in the West region, where the model explained 66% of the variation in water withdrawals (Figure 4.9). The county-level linear model produced the worst predictions of household water use in the Central region. The logistic model produced the most accurate predictions for counties in the Southwest Region, where the difference in average probability that $dw = 1$ was 0.25. The logistic model produced the worst predictions for the counties located in the South region.

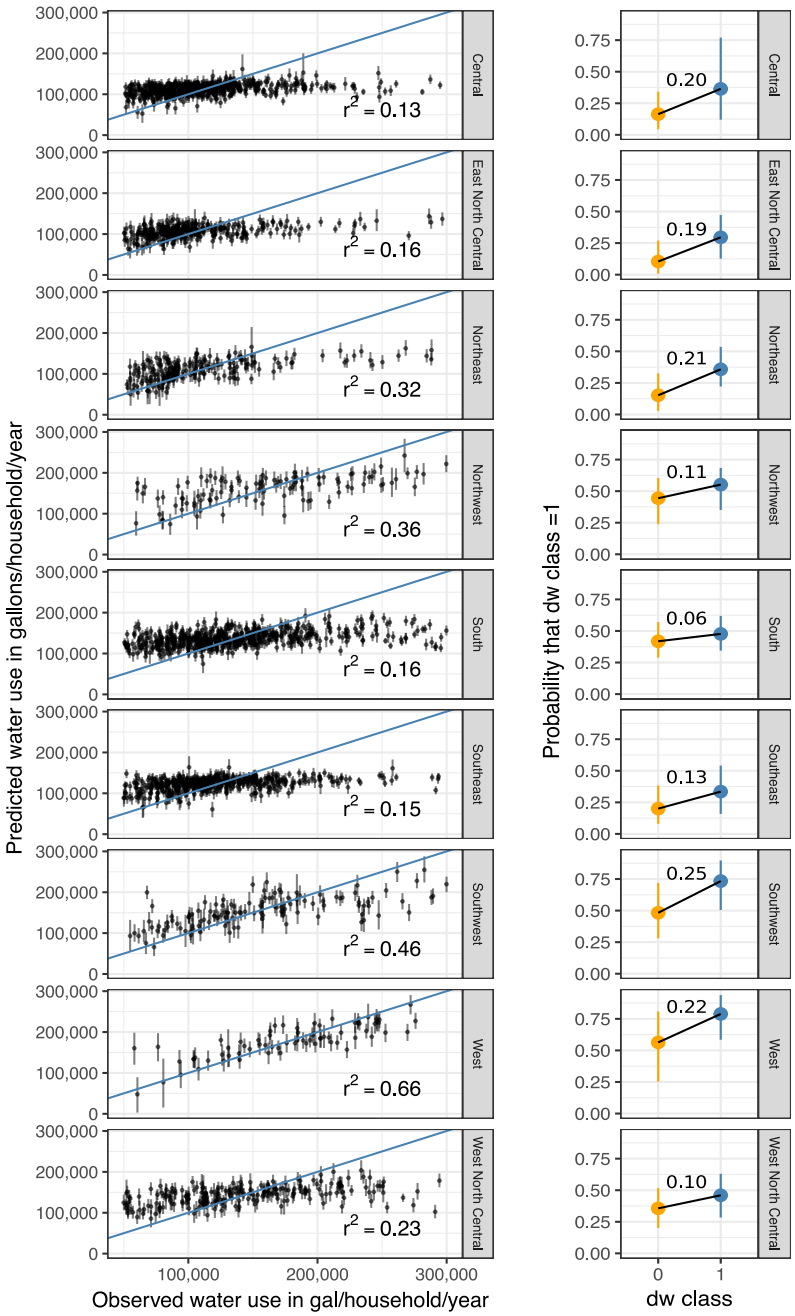

Figure 4.9: [left] Predicted and observed household water use and r^2 values calculated using the mean of the posterior predictive distribution for the partially-pooled linear regression model (*wh*). [right] Distributions of predicted probabilities from the partially-pooled logistic regression model (*dw*). The y axis of the right plot is the distribution of predicted probabilities (μ of posterior distribution) that the difference between a county's actual withdrawal and the national population normalized withdrawal expectation is greater than zero (Equation 4.8) for each climate region. The lower line of the point range extends to the 10% quantile, the point is the mean, and the upper line of the point range extends to the 90% quantile. The number above each line connecting the pointranges is the difference in average probability that $dw = 1$ for each climate region and *dw* class.

4.4.2 City-level analysis

The city-level analysis resulted in better overall predictions than the county-level model. Two exceptions were Houston, TX and St. Louis, MO, both of which sold significantly more water per customer than anticipated by the model (Figure 4.10). The fully pooled and partially-pooled models had similar WAIC values, although the partially pooled model resulted in a better r^2 (Figure 4.10). The largest differences between the models relate to the coefficient estimates for regional price parity and water price. For the fully-pooled city model, an increase of $8 in the regional price parity was associated with an annual increase of 39,000 gal/customer, whereas an $11 increase of the average water price was associated with an annual decrease of 28,000 gal/customer (Figure 4.11). Overall, the fully pooled model suggested that cities located in humid climates (i.e., higher aridity index) that have a high water price and a low price parity tend to, on average, use less water per account than other cities. For the partially-pooled models, an $8 increase in regional price parity was associated with an annual increase in 88,500 gal/customer in the South region. For the West and Southwest regions, an $11 increase of the average water price was associated with an annual decrease of 35,000 gal/customer (Figure 4.11). The bill type had the largest effect in the Southeast region. Water conservation policies had the largest effect in the West, South, and Central regions.

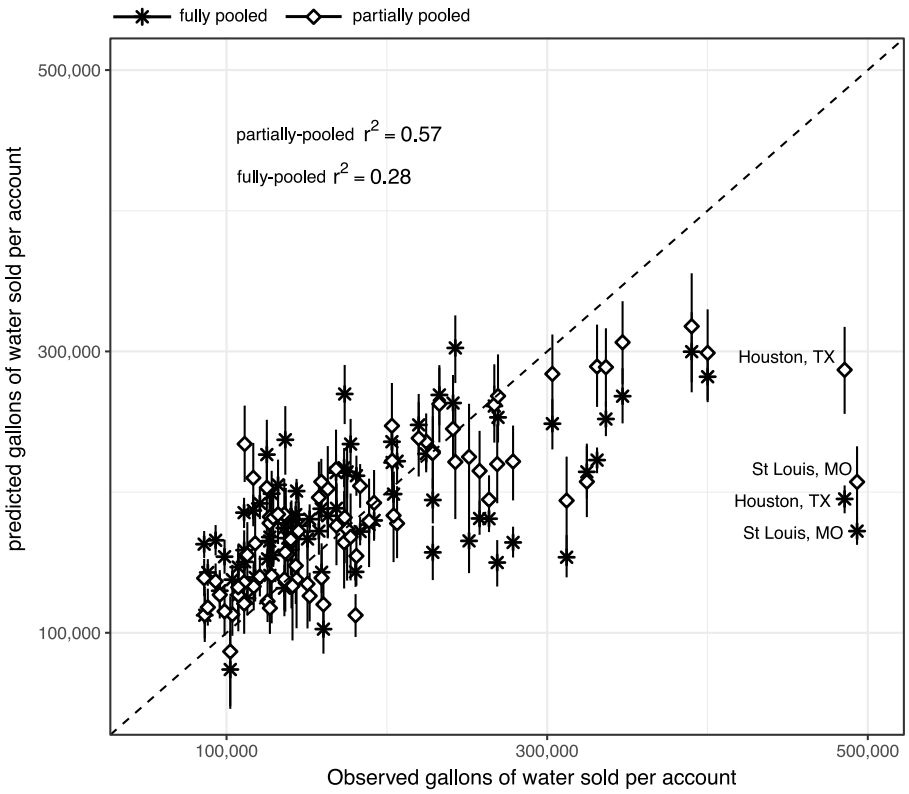

Figure 4.10: Predictions from the partially-pooled and fully-pooled city-level models. Posterior-predictive distributions were generated using 1,000 combinations of parameter values sampled from the posterior parameter distribution. The points and lines are the means and 50th percentile of the posterior-predictive distribution, respectively. The point ranges with the black asterisks were generated using the fully-pooled model and the point ranges with the white diamonds using the partially-pooled model. The partially-pooled model explains significantly more variation in water withdrawals than the fully-pooled model.

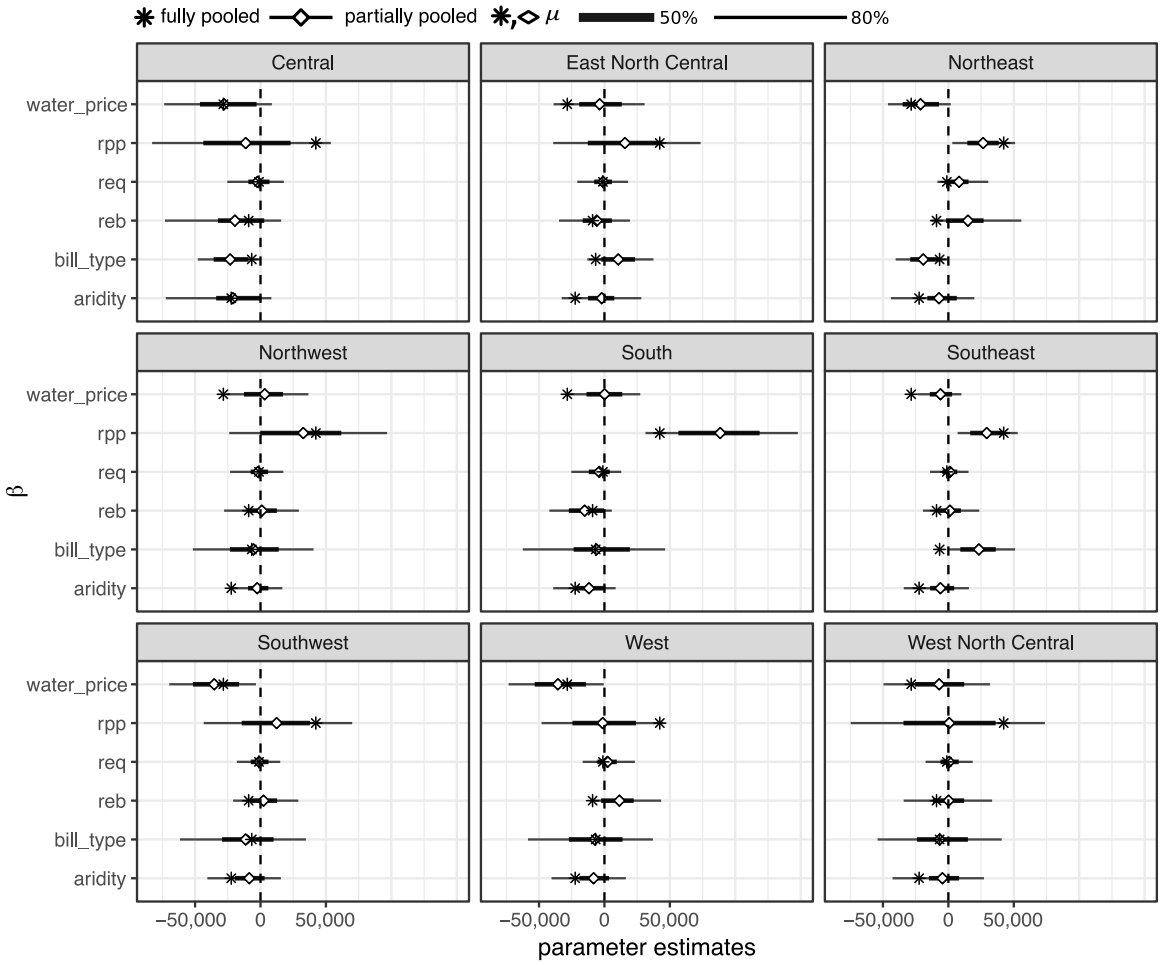

Figure 4.11: β parameter estimates for the city-level models. The partially-pooled estimates are represented by the thick black lines (the 50th percentile interval of the partially-pooled posterior distribution), the thin black lines (the 80th percentile interval of the partially-pooled posterior distribution), and the white diamonds (the mean of the partially-pooled posterior distribution). Fully-pooled estimates are represented by the black asterisks (the mean of the fully-pooled posterior distribution). To avoid cluttering the graph we do not show the percentile intervals for the fully-pooled model. By definition, the fully-pooled estimate is the same for every climate region and is shown to highlight how the climate-region-level estimates vary from the national estimates.

4.5 Discussion

Here we synthesize our results to address the questions laid out at the end of the Introduction.

4.5.1 Climate Regions as a Grouping Variable

Our results demonstrate that the regional "water use context" of a municipal water supplier largely modulates both the volume of water withdrawn and the level of correlation between water use and associated explanatory variables. The regional "water-use context" is defined here by a suite of 18 environmental and social variables grouped into the nine climate regions specified by NOAA (Figures 4.8 and 4.12). The climate regions are identified based on similarities in long term climate histories [234] and are found to explain more variation in water withdrawals compared to several other potential grouping variables (Table 4.3). Climate regions are useful for county-level analyses as the regions align with political boundaries and therefore each county is assigned to a unique climate region (necessary for aggregating social variables). The out-of-sample performance of the models–as estimated by the WAIC–demonstrates that accounting for regional variation is crucial for understanding the relationship between water use and specific covariates. Models that allow partial sharing of information between regions outperform models that estimate parameters that ignore nationally consistent traits. This suggests, for example, that although there is a real difference between the water use context in the Southeastern and the Southwestern U.S., there is also sufficient similarity to warrant allowing information learned in the Southeast to at least partially inform our understanding of water use in the Southwest and vice versa.

The first order effect of climate is accounted for by the varying intercepts in the hierarchal model. The βs associated with precipitation and temperature covariates in each climate region indicate how additional variations in climate across the counties in each region affect water use. For example, the intercept for the Northeast region is smaller than

the intercepts in other regions, (e.g., the West and Southwest), indicating that the Northeast region has lower water use than other regions of the country, possibly linked to its wet and cool climate. However, in the Northeast region, the posterior distribution of the β coefficient for precipitation contains zero. This insignificant β value indicates that additional precipitation fluctuations within the region do not seem to have much impact on water use. This is not the case in the West, for example, where the regression coefficient for precipitation is positive, which likely is the result of outdoor water use [192, 193], and even slight variations in rainfall can greatly affect lawn irrigation needs. These results may indicate a non-linear response of water use to climate, where conditions need to surpass a threshold of aridity before smaller climate variations affect water use. Thus, if locations in the Northeast became significantly more arid, water use may respond to intra-regional variations in climate more strongly.

The regression coefficients indicate that regions vary in their sensitivity to social and environmental explanatory variables. This is most evident when comparing the weighted influence of the top 5 ranked variables from each model for each region (Figure 4.8). The Northeast region is the most sensitive to social variables such as the people per household, proportion of surface-water supply, income disparity, and the age of structures in a county, while the Southwest region is most sensitive to the 40-year precipitation (Figures 4.5 and 4.6). Additionally, the prediction accuracy of the models should be considered when interpreting the regression coefficients. For example, both models for the Southwest region produce relatively good predictions and indicate that, on average, counties in the Southwest that receive more rainfall, vote more Democrat in the presidential election, have younger median age of structures, and have fewer people per household tend to use less water. This region is expected to experience a decreased rainfall in the 21st century [179], which could lead to water shortages if the current relationship between rainfall and water use holds; however, a higher Cook PVI is associated with an increase in water conservation policies in the Southwest region [235], and an increase in water conservation policies could

help to offset increased water withdrawals under a changing climate. The predictions for the logistic model in the Central and East North Central regions are much better than the predictions for the linear model in the same regions. This indicates that the explanatory variables included in this analysis are useful for predicting a binary estimate of high or low withdrawals in these regions but cannot accurately detect a signal in a continuous measure of withdrawals. Finally, we also note that per-capita withdrawals can be predicted with high accuracy in each county based upon water-use values from the 2005 USGS compilation [236], indicating a high degree of memory in water use (not shown). However, a regression using previous water-use values provides little insight regarding the drivers of water use.

4.5.2 Effects of Income and Education

We do not see a strong negative association between water use and either income or education. The fully pooled model actually suggests a positive association (Figure 4.4) between both income and water use and education and water use. For the partially pooled model, the mean posterior parameter value for income is negative for four climate regions for the linear model and one climate region for the logistic model. The mean posterior parameter value for education (measured by percent of population in county with any college education) is negative for six climate regions for the linear model and three climate regions for the logistic model (Figures 4.5 and 4.6). The 80th percentile interval overlaps zero for both variables and models for each climate region. The difference between our findings and those in [202] is likely due to the added information from other variables in this study. For example, the association between income, education, and water use is dampened or reversed after accounting for the effect of the number of people per household and the age of the structures in a county. The partial pooling will pull parameter estimates closer to zero as it accounts for not only the information provided within each climate region but also the joint distribution shared among climate regions.

4.5.3 Effect of Urban to Rural Gradients

The results presented here do not suggest that urban counties use water more efficiently than rural counties, another major conclusion of [202]. Our models indicate there may be a small effect for the fully pooled model (Figure 4.4) but the effect basically disappears for the partially pooled model (Figures 4.5 and 4.6). Why does this discrepancy exist? One explanation is that the rural to urban continuum was treated differently for each analysis. In our study, we treated the discrete codes as a continuous variable that progresses from most rural (code=1) to most urban (code=9). The analysis in [202] classified codes 1-4 as rural and 5-9 as urban. Another possible explanation is undocumented inter-county transfers. Without controlling for transfers, situations where rural counties with small populations are exporting water to urban counties with large populations would be incorrectly interpreted as a rural to urban gradient in water use efficiency. As with education and income, the effect of rural to urban gradients is likely influenced by the addition of the other variables used in this analysis.

4.5.4 City-level Analysis

The city level analysis accounts for water price, conservation policies, and water transfers– variables that were omitted in the county-level analysis due to a lack of data. We also included an environmental variable (aridity) and a socio-economic variable (regional price parity, RPP) to further add to the context for each city. For the fully pooled national model, it appears that arid cities with a high cost of living sell more water per customer. This roughly corresponds to the county-level analysis, where decreased precipitation and a higher median household income are, on average, associated with higher withdrawals per household (Figure 4.11). The fully pooled city-level results reveal a strong relationship between price and water use, where an $11 increase of the average water price is associated with an annual decrease of 28,000 gal/customer (Figure 4.11). The price elasticity

($\beta_{water\ bill}/\alpha$, when *water bill* is mean centered) of water is roughly calculated to be between -0.2 and -0.15 based on the city-level analysis, which is consistent with the lower-end of values from previous estimates [194]. The effect of conservation policies is smaller than the effect of the water price, with much of the posterior distribution overlapping zero, but it does appear that an increase in rebate-related policies is associated with a greater reduction in water use than a change in requirement-related policies. The effect of bill type suggests areas with a decreasing block rate tend to have greater use per customer (although like conservation policy parameters much of the posterior is overlapping zero). A decreasing block rate is effectively a volumetric rebate for using *more water*, so it is sensible that it is associated with a higher water use.

The partially pooled city-level model generates more accurate predictions (Figure 4.10) and indicates that the effects of each predictor strongly depend on where a city is located. The effect of RPP is four times larger than the effect of other variables in the South climate region, which is primarily driven by the large RPP and high water use in Houston and Dallas relative to other cities in the South (Figure 4.11). The South climate region also shows the strongest negative correlation between rebate oriented water conservation policies and water use. This is due to cities like Austin and San Antonio that have a large number of conservation policies and relatively low water use. The parameter estimates are similar for each variable in the Southeast and Northeast, with the exception of the bill type, where a decreasing block rate is associated with an increase in water use in the Southeast and a decrease in water use for the Northeast (Figure 4.11). The Northeast also shows a positive association between water conservation policies and water use, as New York City has the highest relative water use and over 20 rebate and requirement policies, whereas other cities in the Northeast have less than 10 combined policies. One possible explanation for this relates to how a municipality accounts for water provided to apartment buildings. If an apartment building is considered a single account, then locations like New York City with a relatively large number of apartment buildings would seem to have high per capita water

use and the parameter estimates for conservation policies would be difficult to interpret. Regardless, it is important to note that we cannot determine the efficacy of water conservation policies from this study as we do not have temporal data describing when certain policies were implemented and the subsequent change in water use. For example, places like New York City may have had exceptionally high water use and began implementing policies to reduce end user demand, and this analysis simply reflects that positive association but does not indicate that increasing conservation policies increases water use or vice versa.

4.6 Conclusion

The social and environmental controls on water use are not uniform across the CONUS, and national-scale water-use assessments must account for regional variability in order to understand the present drivers of water use and project likely changes into the future. Hierarchical Bayesian regression models offer one way to explicitly account for spatial variability by using higher-level grouping variables. It is also important for large-scale water-use studies to account for a large number of social and environmental variables to avoid over-interpreting the effect of spurious correlations. Finally, our analysis demonstrates the importance of water transfers among public suppliers and the potential of such transfers to confound analyses of per capita water use and its drivers. Absent a national quality-assured database of water transfers and pricing information, large-scale analyses of per-capita water use will continue to be plagued by avoidable but unaccounted error.

4.7 Appendix

4.7.1 Groups for hierarchical models

Table 4.4: Description of urban continuum codes. More information: https://www.ers.usda.gov/data-products/rural-urban-continuum-codes/documentation/

Urban code	**Description**
metro_large	metro with county population >1,000,000
metro_med	metro with county population 250,000-1,000,000
metro_small	metro with county population < 250,000
urb_large_adj	urban with county population ≥ 20,000, adjacent to metro area
urb_large_det	urban with county population ≥ 20,000, detached from metro area
urb_small_adj	urban with county population 2,500-19,999, adjacent to metro area
urb_large_det	urban with county population 2,500-19,999, detached from metro area
rural_adj	rural with county population < 2,500, adjacent to metro area
rural_det	rural with county population < 2,500, detached from metro area

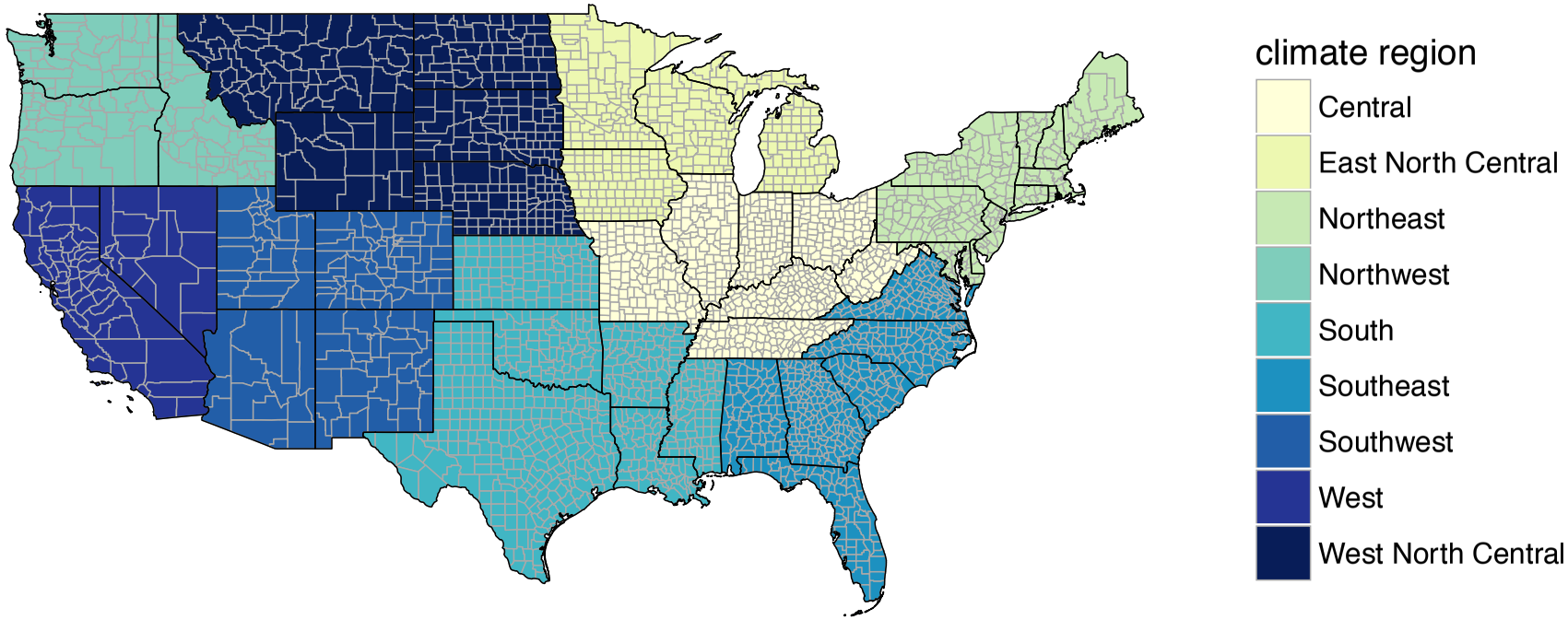

Figure 4.12: Map of NOAA climate regions (variable is labeled ”climate_region” in Table 4.1).

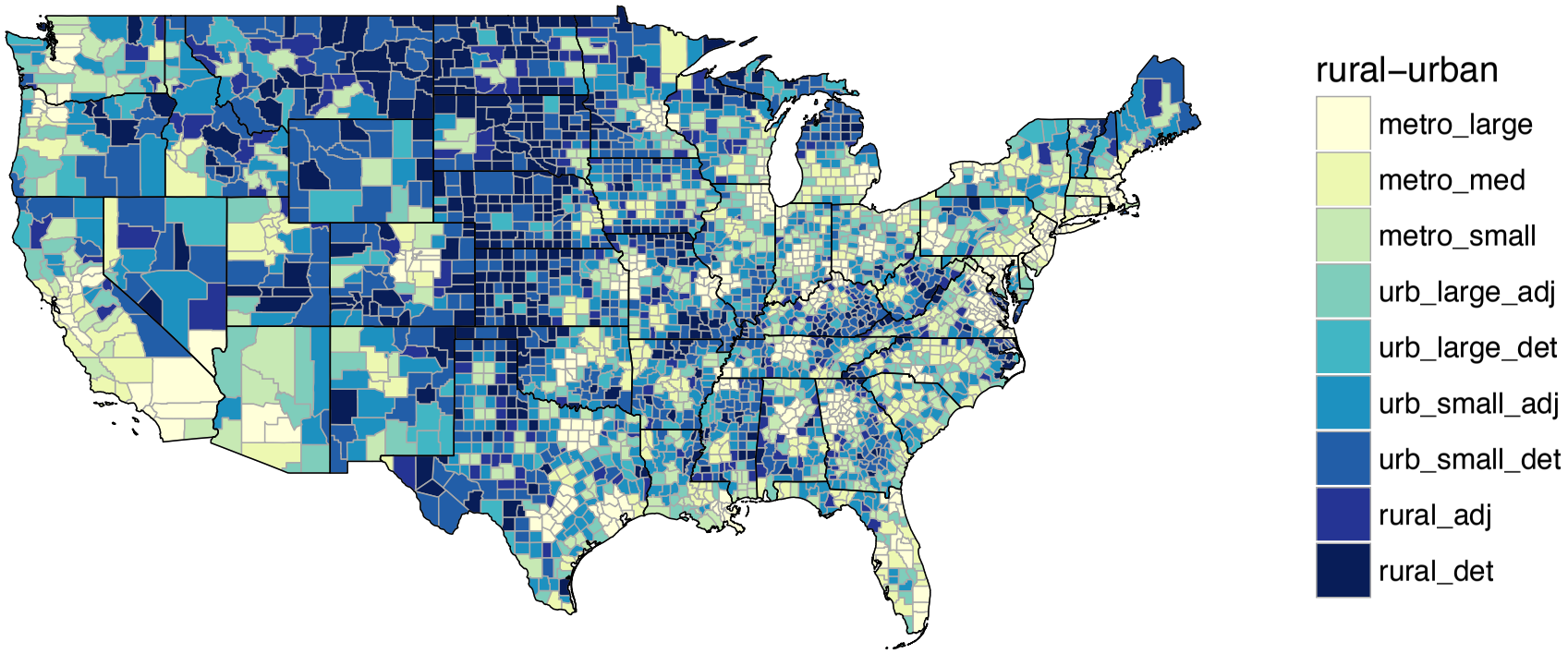

Figure 4.13: Map of the rural-to-urban gradient (variable is labeled ”rur2urb” in Table 4.1) for each county. Descriptions of codes are in Table 4.4

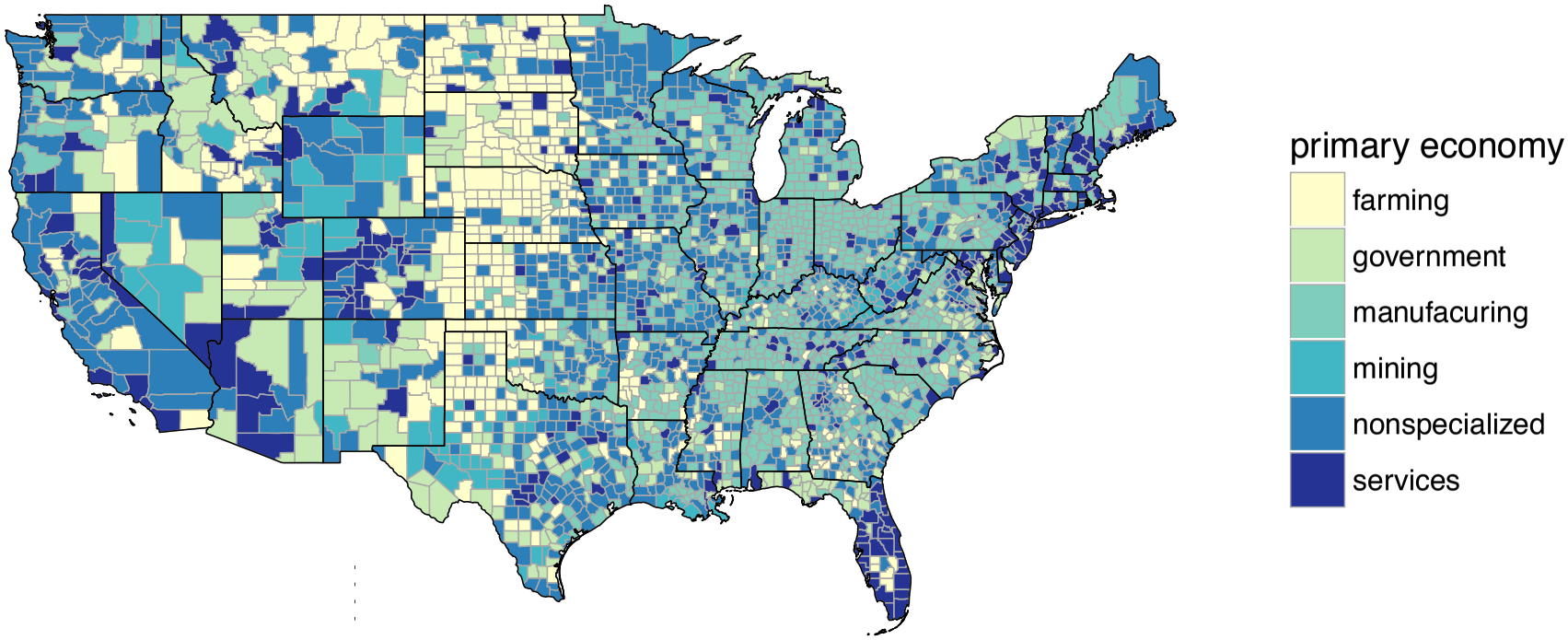

Figure 4.14: Map of the primary economic activity (variable is labeled “econdep” in Table 4.1) for each county.

CHAPTER 5

SYNTHESIS

Data-driven methods are often extolled as alternatives to theory-driven methods. In hydrology, this can be seen in the juxtaposition of stochastic and conceptual hydrologic models [21]. Some argue that the scientific method is becoming obsolete with the rise of bigger data and better data-mining algorithms [237]. The argument is that with sufficient data we don't need hypotheses or conceptual models. We don't need the "old ways" of doing science when "...huge amounts of data, along with the statistical tools to crunch these numbers, offers a whole new way of understanding the world. Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all" [237][1]. Advocates for a data-driven future often cite examples like Google's ability to match ads with content without relying on conceptual models that describe how humans react to advertisements. Google simply lets an algorithm explore petabytes of data to decide what ad should be associated with what content based of previous click-rates. How does this relate to the "old ways" of doing science?

The *The Fourth Paradigm: Data-Intensive Scientific Discovery* [238], a book that offers one perspective on evolution of science, suggests that science has gone through three major paradigm transitions and is entering a fourth. (1) Science was originally *empirical*. For example, Aristotle developed biological theories based off observations he made while dissecting fish. (2) Science became *theoretical*. For example, Newton wrote mathematical equations that describe physical laws. (3) Science then became *computational*. For example, hydrologists modeled groundwater systems by solving flow equations using finite-element methods. (4) Science is becoming *data driven*. For example, earth scientist can mine massive amounts of satellite data to generate new hypotheses about land-use tran-

[1]The article was written to be provocative [http://norvig.com/fact-check.html]. It presents the caricature of an idea with the hopes of stimulating conversation.

sitions. The idea is that data-driven methods will continue to outpace mental conceptual models, basic theories, and simulations.

These "transitions" describe the evolution of the tools that scientists use rather than the scientific method itself. It is not clear how having access to large amounts of data moves the conversation past the traditional framework of deduction, induction, and abduction–the pillars of logical inference (Figure 5.1). For example, the inductive process describes both Aristotle's attempt to understand fish biology and a modern-day researcher mining massive datasets looking for associations between inputs and outputs. Both start with observations, examine possible connections, and generalize rules. Each of the four paradigms described above can be placed somewhere on the gradient of logical inference. The introduction of "huge amounts of data, along with the statistical tools to crunch these numbers" offers a new way to discover relationships and make inferences, while the "ways of understanding the world", remain unchanged. Thus I argue that integrating data-driven methods and domain theory is not a change in paradigm (at least not in the Khunian sense [239]), but rather an extension and improvement in the process of inference and discovery. The remainder of this synthesis involves tracing the path of each dissertation chapter through the various stages of logical inference and describing how data-driven models augment that path.

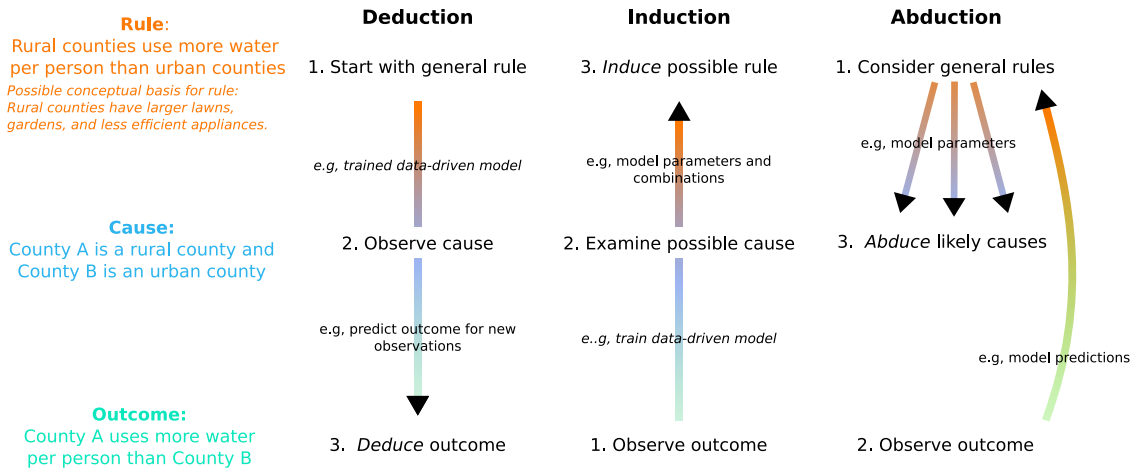

Figure 5.1: Schematic of logical inference using an example from Chapter 4.

Training machine-learning models is an inductive task. Starting with specific training observations, associations are discovered and are encoded as general rules that can be used to make predictions for new observations. The predictive step is a form of deduction as the general rules learned in the training step are used to deduce outcomes for new observations. Chapter 2, *Improving predictions of hydrological low-flow indices in ungaged basins using machine learning*, is an example of applying induction to build rules (models) followed by deduction to make predictions. The variable importance and partial dependence analysis are examples of abduction (Figure 2.7), and it is in this step where theory guides how we interpret model results. We examine the changes in 7Q10s associated with different values of particular explanatory variables (conditioned on the mean of all the other explanatory variables). We see that 7Q10's generally decrease as the percent of emergent wetlands increases. If we observe a small 7Q10 for a new basin, it is possible that the 7Q10 is small because the basin has a large amount of wetlands. Can we improve our understanding of the cause given additional information (e.g., knowledge of region, hydrologic theory, other associations in the dataset)? We know that the basins with the most wetlands are clustered in the Coastal plain of Georgia and South Carolina, a region with shallow groundwater tables that can support wetlands. We also know that evapotranspiration from wetlands during periods of low streamflow can lead to less water being available for the stream. Therefore, we can abduce that the wetlands are using the water that would otherwise be available to support streamflow during dry periods.

The above example illustrates how theory might inform the interpretation of model predictions. Chapter 3, *Predicting flow duration curves in ungaged basins using L-moments and theory-informed neural networks*, involves integrating of theory into a different level of the model. We know that quantiles always increase with increasing nonexceedance probabilities for a given observation. A model that learns each quantile independently often violates this "rule of monotonicity". We can greatly reduce the number of violations by constructing a data-driven model that learns each quantile simultaneously, thus leveraging

the covariance between quantiles. All data-driven models are a form of constrained induction, i.e., the general rules (parameters and how they are combined) of the model are conditional on explanatory variables. A multiple-output model is another layer of constraint where the outputs are conditioned on both the explanatory variables and the values of the other outputs. In this sense we are just narrowing the subset of possible models to ones that are more consistent with the system we are attempting to model [240].

Chapter 4, *Exploring the drivers of public-supply water use using hierarchical-Bayesian models*, is an explicit example of how data-driven methods can be used to aid abductive reasoning[2]. The hierarchical model design is based on the premise that water use can be described on two levels; a lower level that describes the actual data and an upper level that influences the values taken by parameters in the lower level. The grouping variable, climate regions, and final model design, a two-level hierarchical model, were selected using approximate out-of-sample performance of the posterior-predictive distribution. This type of model selection uses the empirical content in the posterior distributions (i.e., the inductive content) to reject alternative hypotheses (i.e., other models) using "deduction within a model", a concept closely related to the idea of Popperian falsification [242, 243]. After a model has been selected, the conditional posterior distributions of each parameter in the model provides rigorous means to abduce likely casual links between water use and explanatory variables. For example, the intercept for the Northeast region is smaller than the intercepts in other regions, (e.g., the West and Southwest), indicating that the Northeast region has lower water use than other regions of the country, possibly linked to its wet and cool climate (Figure 4.5). However, in the Northeast region, the posterior distribution of the β coefficient for precipitation contains zero. This insignificant β value indicates that additional precipitation fluctuations within the region do not seem to have much impact on water use. This is not the case in the West, for example, where the regression coefficient for precipitation is positive, which likely is the result of outdoor water use [192, 193], and

[2]For an excellent discussion of abduction and Bayesian inference see Romeijn 2013 [241]

even slight variations in rainfall can greatly affect lawn irrigation needs. These results may indicate a non-linear response of water use to climate, where conditions need to surpass a threshold of aridity before smaller climate variations affect water use.

Data-driven methods are changing *how* science is done rather than *what* science does. Surprising associations found in large datasets can generate novel hypotheses that might not be imaginable to the researcher at the time of discovery. These new hypotheses can be further refined through traditional deductive methods and data exploration. Data-driven methods can lead to new scientific understandings, which in turn can be used to augment new inductive models. The feedback between data analysis and establishing theory is connate to scientific progress.

BIBLIOGRAPHY

[1] Scott C Worland. 2010 county and city-level water-use data and associated explanatory variables. *U.S. Geological Survey data release*, 2017. https://doi.org/10.5066/F72Z14FR.

[2] Dimitri P Solomatine and Avi Ostfeld. Data-driven modelling: some past experiences and new approaches. *Journal of hydroinformatics*, 10(1):3–22, 2008.

[3] Linda See, Dimitri Solomatine, Robert Abrahart, and Elena Toth. Hydroinformatics: computational intelligence and technological developments in water science applicationseditorial. *Hydrological Sciences Journal*, 52(3):391–396, 2007.

[4] Alon Halevy, Peter Norvig, and Fernando Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2):8–12, 2009.

[5] Eugene P Wigner. The unreasonable effectiveness of mathematics in the natural sciences. richard courant lecture in mathematical sciences delivered at new york university, may 11, 1959. *Communications on pure and applied mathematics*, 13(1):1–14, 1960.

[6] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[7] George M Hornberger, Patricia L Wiberg, Jeffrey P Raffensperger, and Paolo D'Odorico. *Elements of physical hydrology*. JHU Press, 2014.

[8] Richard M Vogel, Upmanu Lall, Ximing Cai, Balaji Rajagopalan, Peter K Weiskel, Richard P Hooper, and Nicholas C Matalas. Hydrology: The interdisciplinary science of water. *Water Resources Research*, 51(6):4409–4430, 2015.

[9] Anuj Karpatne, Gowtham Atluri, James H Faghmous, Michael Steinbach, Arindam Banerjee, Auroop Ganguly, Shashi Shekhar, Nagiza Samatova, and Vipin Kumar. Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering*, 29(10):2318–2331, 2017.

[10] Murugesu Sivapalan, K Takeuchi, SW Franks, VK Gupta, H Karambiri, V Lakshmi, X Liang, JJ McDonnell, EM Mendiondo, PE O'connell, et al. 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, 2003.

[11] M Hrachowitz, HHG Savenije, G Blöschl, JJ McDonnell, M Sivapalan, JW Pomeroy, Berit Arheimer, T Blume, MP Clark, U Ehret, et al. A decade of predictions in ungauged basins (pub)a review. *Hydrological sciences journal*, 58(6):1198–1255, 2013.

[12] Günter Blöschl. Predictions in ungauged basins–where do we stand? *Proc. IAHS*, 373:57–60, 2016.

[13] Murugesu Sivapalan, Hubert HG Savenije, and Günter Blöschl. Socio-hydrology: A new science of people and water. *Hydrological Processes*, 26(8):1270–1276, 2012.

[14] Scott C Worland, William H Farmer, and Julie E Kiang. Improving predictions of hydrological low-flow indices in ungaged basins using machine learning. *Environmental Modelling & Software*, 101:169–182, 2018.

[15] Scott. C. Worland, Scott Steinschneider, and George M. Hornberger. Drivers of variability in public-supply water use across the contiguous united states. *Water Resources Research*.

[16] Tara Razavi and Paulin Coulibaly. Streamflow prediction in ungauged basins: review of regionalization methods. *Journal of Hydrologic Engineering*, 18(8):958–975, 2012.

[17] Rodney R Knight, Jennifer C Murphy, William J Wolfe, Charles F Saylor, and Amy K Wales. Ecological limit functions relating fish community response to hydrologic departures of the ecological flow regime in the tennessee river basin, united states. *Ecohydrology*, 7(5):1262–1280, 2014.

[18] Katherine E Kapo, Kathleen McDonough, Thomas Federle, Scott Dyer, and Raghu Vamshi. Mixing zone and drinking water intake dilution factor and wastewater generation distributions to enable probabilistic assessment of down-the-drain consumer product chemicals in the us. *Science of The Total Environment*, 518:302–309, 2015.

[19] Ashok K Mishra and Paulin Coulibaly. Developments in hydrometric network design: A review. *Reviews of Geophysics*, 47(2), 2009.

[20] Charles Luce. Runoff prediction in ungauged basins: synthesis across processes, places and scales. *Eos, Transactions American Geophysical Union*, 95(2):22–22, 2014.

[21] William H Farmer and Richard M Vogel. On the deterministic and stochastic use of hydrologic models. *Water Resources Research*, 2016.

[22] VU Smakhtin. Low flow hydrology: a review. *Journal of hydrology*, 240(3):147–186, 2001.

[23] Daniel P Ames. Estimating 7q10 confidence limits from data: a bootstrap approach. *Journal of water resources planning and management*, 132(3):204–208, 2006.

[24] Patrick R Kormos, Charles H Luce, Seth J Wenger, and Wouter R Berghuijs. Trends and sensitivities of low streamflow extremes to discharge timing and magnitude in pacific northwest mountain streams. *Water Resources Research*, 2016.

[25] Jennifer C Murphy, Rodney R Knight, William J Wolfe, and W S Gain. Predicting

ecological flow regime at ungaged sites: a comparison of methods. *River Research and Applications*, 29(5):660–669, 2013.

[26] Timothy H Raines and William H Asquith. Analysis of minimum 7-day discharges and estimation of minimum 7-day, 2-year discharges for streamflow-gaging stations in the brazos river basin, texas. Technical report, 1997.

[27] Galit Shmueli. To explain or to predict? *Statistical Science*, 25(3):289–310, 2010.

[28] Harry F Lins and Timothy A Cohn. Stationarity: Wanted dead or alive? *Journal of the American Water Resources Association*, 47(3):475–480, 2011.

[29] HC Riggs. Regional analyses of streamflow characteristics: Us geological survey techniques of water-resources investigations, book 4, chap. Technical report, 1973.

[30] Donald M Thomas and Manuel A Benson. *Generalization of streamflow characteristics from drainage-basin characteristics*. US Government Printing Office, 1970.

[31] Clayton H Hardison. Prediction error of regression estimates of streamflow characteristics at ungaged sites. *US Geological Survey Professional Paper*, 750:C228–C236, 1971.

[32] Gary D Tasker. Hydrologic regression with weighted least squares. *Water Resources Research*, 16(6):1107–1113, 1980.

[33] Jery R Stedinger and Gary D Tasker. Regional hydrologic analysis: 1. ordinary, weighted and generalized least squares compared. *Water Resources Research*, 21(9):1421–1432, 1985.

[34] Charles N Kroll and Jery R Stedinger. Development of regional regression relationships with censored data. *Water Resources Research*, 35(3):775–784, 1999.

[35] Jaysson E Funkhouser, Ken Eng, and Matthew W Moix. Low-flow characteristics and regionalization of low-flow characteristics for selected streams in arkansas. Technical report, Geological Survey (US), 2008.

[36] George S Law, Gary D Tasker, and David E Ladd. *Streamflow-Characteristic Estimation Methods for Unregulated Streams of Tennessee*. US Geological Survey, 2009.

[37] S Castiglioni, A Castellarin, and A Montanari. Prediction of low-flow indices in ungauged basins through physiographical space-based interpolation. *Journal of hydrology*, 378(3):272–280, 2009.

[38] S Castiglioni, A Castellarin, A Montanari, JO Skøien, G Laaha, and G Blöschl. Smooth regional estimation of low-flow indices: physiographical space based interpolation and top-kriging. *Hydrology and Earth System Sciences*, 15(3):715–727, 2011.

[39] TBMJ Ouarda and C Shu. Regional low-flow frequency analysis using single and ensemble artificial neural networks. *Water resources research*, 45(11), 2009.

[40] Gregor Laaha and Günter Blöschl. A comparison of low flow regionalisation methodscatchment grouping. *Journal of Hydrology*, 323(1):193–214, 2006.

[41] Spencer Schnier and Ximing Cai. Prediction of regional streamflow frequency using model tree ensembles. *Journal of Hydrology*, 517:298–309, 2014.

[42] DJ Booker and RA Woods. Comparing and combining physically-based and empirically-based approaches for estimating the hydrology of ungauged catchments. *Journal of Hydrology*, 508:227–239, 2014.

[43] James Falcone. Gages-ii: Geospatial attributes of gages for evaluating streamflow:

U.s. geological survey: Reston, virginia. http://water.usgs.gov/lookup/getspatial?gagesII_Sept2011, 2011.

[44] A Elshorbagy, G Corzo, S Srinivasulu, and DP Solomatine. Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology-part 1: Concepts and methodology. *Hydrology and Earth System Sciences*, 14(10):1931–1941, 2010.

[45] A Elshorbagy, G Corzo, S Srinivasulu, and DP Solomatine. Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology-part 2: Application. *Hydrology and Earth System Sciences*, 14(10):1943–1961, 2010.

[46] Julie E Shortridge, Seth D Guikema, and Benjamin F Zaitchik. 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, 2016.

[47] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*, volume 2. Springer series in statistics Springer, Berlin, 2013.

[48] Max Kuhn and Kjell Johnson. *Applied predictive modeling*. Springer, 2013.

[49] Holger R Maier, Ashu Jain, Graeme C Dandy, and K PKPS Sudheer. Methods used for the development of neural networks for the prediction of water resource variables in river systems: current status and future directions. *Environmental Modelling & Software*, 25(8):891–909, 2010.

[50] KS Kasiviswanathan, Jianxun He, KP Sudheer, and Joo-Hwa Tay. Potential application of wavelet neural network ensemble to forecast streamflow for flood management. *Journal of Hydrology*, 536:161–173, 2016.

[51] Greer B Humphrey, Matthew S Gibbs, Graeme C Dandy, and Holger R Maier. A hybrid approach to monthly streamflow forecasting: Integrating hydrological model outputs into a bayesian artificial neural network. *Journal of Hydrology*, 540:623–640, 2016.

[52] Ioannis N Daliakopoulos and Ioannis K Tsanis. Comparison of an artificial neural network and a conceptual rainfall–runoff model in the simulation of ephemeral streamflow. *Hydrological Sciences Journal*, 61(15):2763–2774, 2016.

[53] Daren M Carlisle, James Falcone, David M Wolock, Michael R Meador, and Richard H Norris. Predicting the natural flow regime: models for assessing hydrological alteration in streams. *River Research and Applications*, 26(2):118–136, 2010.

[54] Kenny Eng, Daren M Carlisle, David M Wolock, and James A Falcone. Predicting the likelihood of altered streamflows at ungauged rivers across the conterminous united states. *River Research and Applications*, 29(6):781–791, 2013.

[55] Bing Li, Guishan Yang, Rongrong Wan, Xue Dai, and Yanhui Zhang. Comparison of random forests and other statistical methods for the prediction of lake water level: a case study of the poyang lake in china. *Hydrology Research*, 47(S1):69–83, 2016.

[56] Aman Mohammad Kalteh. Improving forecasting accuracy of streamflow time series using least squares support vector machine coupled with data-preprocessing techniques. *Water Resources Management*, 30(2):747–766, 2016.

[57] Jun Guo, Jianzhong Zhou, Hui Qin, Qiang Zou, and Qingqing Li. Monthly streamflow forecasting based on improved support vector machine model. *Expert Systems with Applications*, 38(10):13073–13081, 2011.

[58] Deepti Joshi, André St-Hilaire, Anik Daigle, and Taha BMJ Ouarda. Databased

comparison of sparse bayesian learning and multiple linear regression for statistical downscaling of low flow indices. *Journal of hydrology*, 488:136–149, 2013.

[59] Monomoy Goswami and KIERAN MICHAEL O'Connor. Comparative assessment of six automatic optimization techniques for calibration of a conceptual rainfall-runoff model. *Hydrological sciences journal*, 52(3):432–449, 2007.

[60] Dimitri P Solomatine, Mahesh Maskey, and Durga Lal Shrestha. Instance-based learning compared to other data-driven methods in hydrological forecasting. *Hydrological Processes*, 22(2):275–287, 2008.

[61] Luka Štravs and Mitja Brilly. Development of a low-flow forecasting model using the m5 machine learning method. *Hydrological sciences journal*, 52(3):466–477, 2007.

[62] Dimitri P Solomatine and Yunpeng Xue. M5 model trees and neural networks: application to flood forecasting in the upper reach of the huai river in china. *Journal of Hydrologic Engineering*, 9(6):491–501, 2004.

[63] Zaher Mundher Yaseen, Ozgur Kisi, and Vahdettin Demir. Enhancing long-term streamflow forecasting and predicting using periodicity data component: application of artificial intelligence. *Water Resources Management*, 30(12):4125–4151, 2016.

[64] Huma Zia, Nick Harris, Geoff Merrett, and Mark Rivers. Predicting discharge using a low complexity machine learning model. *Computers and Electronics in Agriculture*, 118:350–360, 2015.

[65] Chang Shu and Donald H Burn. Artificial neural network ensembles and their application in pooled flood frequency analysis. *Water Resources Research*, 40(9), 2004.

[66] Halil Ibrahim Erdal and Onur Karakurt. Advancing monthly streamflow prediction

accuracy of cart models using ensemble learning paradigms. *Journal of Hydrology*, 477:119–128, 2013.

[67] Leo Breiman et al. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231, 2001.

[68] Alfredo Vellido, José David Martín-Guerrero, and Paulo JG Lisboa. Making machine learning models interpretable. In *ESANN*, volume 12, pages 163–172. Citeseer, 2012.

[69] T. D. Feaster and W. B. Guimaraes. Low-flow frequency and flow duration of selected South Carolina streams in the Pee Dee River Basin through March 2007. *U.S. Geological Survey Open-File Report*, 2009–1171:1–39, 2009.

[70] W. B. Guimaraes and T. D. Feaster. Low-flow frequency and flow duration of selected South Carolina streams in the Broad River Basin through March 2008. *U.S. Geological Survey Open-File Report*, 2010–1305:1–47, 2010.

[71] T. D. Feaster and W. B. Guimaraes. Low-flow frequency and flow duration of selected South Carolina streams in the Saluda, Congaree, and Edisto River Basins through March 2009. *U.S. Geological Survey Open-File Report*, 2012–1253:1–53, 2012.

[72] T. D. Feaster and W. B. Guimaraes. Low-flow frequency and flow duration of selected South Carolina streams in the Catawba-Wateree and Santee River Basins through March 2012. *U.S. Geological Survey Open-File Report*, 2014–1113:1–34, 2014.

[73] T. D. Feaster and W. B. Guimaraes. Low-flow frequency and flow duration of selected South Carolina streams in the Savannah and Salkehatchie River Basins through March 2014. *U.S. Geological Survey Open-File Report*, 2016–1101, 2016 pages = 1–62.

[74] A.J. Gotvald. Selected low-flow frequency statistics for continuous-record streamgages in georgia, 2013. *U.S. Geological Survey Open-File Report*, 2016–5037:1–20, 2016.

[75] T. D. Feaster and K.G. Lee. Low-flow frequency and flow duration characteristics of selected alabama streams through march 2014. *U.S. Geological Survey Open-File Report*, in press.

[76] William H Farmer, Stacey A Archfield, Thomas M Over, Lauren E Hay, Jacob H LaFontaine, and Julie E Kiang. A comparison of methods to predict historical daily streamflow time series in the southeastern united states. *U.S. Geological Survey Scientific Investigations Report*, 2015.

[77] Scott C Worland, William H Farmer, and Kiang Julie. Data release: 7q10 records and basin characteristics for 224 basins in south carolina, georgia, and alabama (2015), 2017.

[78] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

[79] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012.

[80] Yachen Yan. *rBayesianOptimization : Bayesian Optimization of Hyperparameters*, 2016. R package version 1.1.0.

[81] James Tobin. Estimation of relationships for limited dependent variables. *Econometrica: journal of the Econometric Society*, pages 24–36, 1958.

[82] Rachel A Esralew and Stephen Jerrod Smith. Methods for estimating flow-duration

and annual mean-flow statistics for ungaged streams in oklahoma. Technical report, 2009.

[83] John C Risley, Adam Stonewall, and Tana L Haluska. Estimating flow-duration and low-flow frequency statistics for unregulated streams in oregon. Technical report, US Department of the Interior, US Geological Survey, 2008.

[84] C Kroll and J Luz. The application of censored regression models in low streamflow analyses. In *AGU Fall Meeting Abstracts*, volume 1, page 0967, 2003.

[85] David A Eash and Kimberlee K Barnes. Methods for estimating selected low-flow frequency statistics and harmonic mean flows for streams in iowa. Technical report, 2012.

[86] Ken Eng, Gary D Tasker, and PCD Milly. An analysis of region-of-influence methods for flood regionalization in the gulf-atlantic rolling plains1. *JAWRA Journal of the American Water Resources Association*, 41(1):135–143, 2005.

[87] Naomi S Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992.

[88] Donald H Burn. Evaluation of regional flood frequency analysis with a region of influence approach. *Water Resources Research*, 26(10):2257–2265, 1990.

[89] Edward H Isaaks and R Mohan Srivastava. An introduction to applied geostatistics. 1989. *New York, USA: Oxford University Press. Jones DR, A taxonomy of global optimization methods based on response surfaces. Journal of Global Optimization*, 23:345–383, 2001.

[90] Stacey A Archfield and Richard M Vogel. Map correlation method: Selection of a reference streamgage to estimate daily streamflow at ungaged catchments. *Water Resources Research*, 46(10), 2010.

[91] W. H. Farmer. Ordinary kriging as a tool to estimate historical daily streamflow records. *Hydrology and Earth System Sciences*, 20(7):2721–2735, 2016.

[92] David H Wolpert. The lack of a priori distinctions between learning algorithms. *Neural computation*, 8(7):1341–1390, 1996.

[93] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[94] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

[95] Jerome H Friedman. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378, 2002.

[96] James Hickey, Paul Metcalfe, Greg Ridgeway, Stefan Schroedl, Harry Southworth, and Terry Therneau. *gbm: Generalized Boosted Regression Models*, 2016. R package version 2.1.06.9000.

[97] Kai Yu, Liang Ji, and Xuegong Zhang. Kernel nearest-neighbor algorithm. *Neural Processing Letters*, 15(2):147–156, 2002.

[98] Klaus Schliep and Klaus Hechenbichler. *kknn: Weighted k-Nearest Neighbors*, 2016. R package version 1.3.1.

[99] Max Kuhn, Steve Weston, Chris Keefer, and Nathan Coulter. Cubist models for regression. *R package Vignette R package version 0.0*, 18, 2012.

[100] Wei-Yin Loh. Classification and regression tree methods. *Encyclopedia of statistics in quality and reliability*, 2008.

[101] J Ross Quinlan. Combining instance-based and model-based learning. In *Proceedings of the Tenth International Conference on Machine Learning*, pages 236–243, 1993.

[102] Max Kuhn, Steve Weston, Chris Keefer, and Nathan Coulter. C code for Cubist by Ross Quinlan. *Cubist: Rule- and Instance-Based Regression Modeling*, 2014. R package version 0.0.18.

[103] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[104] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.

[105] Alex J Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004.

[106] Alexandros Karatzoglou, Alex Smola, Kurt Hornik, and Achim Zeileis. kernlab – an S4 package for kernel methods in R. *Journal of Statistical Software*, 11(9):1–20, 2004.

[107] David H Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.

[108] Leo Breiman. Stacked regressions. *Machine learning*, 24(1):49–64, 1996.

[109] Hoshin V Gupta, Harald Kling, Koray K Yilmaz, and Guillermo F Martinez. Decomposition of the mean squared error and nse performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, 377(1):80–91, 2009.

[110] Allan H Murphy. Skill scores based on the mean square error and their relationships to the correlation coefficient. *Monthly weather review*, 116(12):2417–2424, 1988.

[111] Max Kuhn. Variable importance using the caret package. 2012.

[112] Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1):44–65, 2015.

[113] David M Wolock. Statsgo soil characteristics for the conterminous united states. (97-656), 1997.

[114] Dennis R Helsel and Robert M Hirsch. *Statistical methods in water resources*, volume 49. Elsevier, 1992.

[115] Carol A Johnston, Naomi E Detenbeck, and Gerald J Niemi. The cumulative effect of wetlands on stream water quality and quantity. a landscape approach. *Biogeochemistry*, 10(2):105–141, 1990.

[116] James K Searcy. Flow-duration curves. *USGS water supply paper*, 1542-A, 1959.

[117] A Castellarin, G Botter, DA Hughes, S Liu, TBMJ Ouarda, J Parajka, DA Post, M Sivapalan, C Spence, A Viglione, et al. Prediction of flow duration curves in ungauged basins. *Runoff Prediction in Ungauged Basins: Synthesis Across Processes, Places and Scales. University Press, Cambridge*, pages 135–162, 2013.

[118] Yoshiyuki Yokoo and Murugesu Sivapalan. Towards reconstruction of the flow duration curve: development of a conceptual framework with a physical basis. *Hydrology and Earth System Sciences*, 15(9):2805–2819, 2011.

[119] Joshua S Rice and Ryan E Emanuel. How are streamflow responses to the el nino southern oscillation affected by watershed characteristics? *Water Resources Research*, 2017.

[120] Lei Cheng, Mary Yaeger, A Viglione, Evan Coopersmith, Sheng Ye, and Murugusu Sivapalan. Exploring the physical controls of regional patterns of flow duration

curves–part 1: Insights from statistical analyses. *Hydrology and Earth System Sciences*, 16(11):4435–4446, 2012.

[121] Sheng Ye, M Yaeger, E Coopersmith, Lei Cheng, and M Sivapalan. Exploring the physical controls of regional patterns of flow duration curves-part 2: Role of seasonality, the regime curve, and associated process controls. *Hydrology and Earth System Sciences*, 16(11):4447, 2012.

[122] Evan Coopersmith, MA Yaeger, Sheng Ye, Lei Cheng, and Murugusu Sivapalan. Exploring the physical controls of regional patterns of flow duration curves–part 3: A catchment classification system based on regime curve indicators. *Hydrology and Earth System Sciences*, 16(11):4467–4482, 2012.

[123] Mary Yaeger, Evan Coopersmith, Sheng Ye, Lei Cheng, A Viglione, and Murugusu Sivapalan. Exploring the physical controls of regional patterns of flow duration curves–part 4: A synthesis of empirical analysis, process modeling and catchment classification. *Hydrology and Earth System Sciences*, 16(11):4483–4498, 2012.

[124] K Nruthya and VV Srinivas. Evaluating methods to predict streamflow at ungauged sites using regional flow duration curves: A case study. *Aquatic Procedia*, 4:641–648, 2015.

[125] Elizabeth S Homa, Casey Brown, Kevin McGarigal, Bradley W Compton, and Scott D Jackson. Estimating hydrologic alteration from basin characteristics in massachusetts. *Journal of hydrology*, 503:196–208, 2013.

[126] Charles N Kroll, Kelly E Croteau, and Richard M Vogel. Hypothesis tests for hydrologic alteration. *Journal of Hydrology*, 530:117–126, 2015.

[127] Quanxi Shao, Lu Zhang, Yongqin D Chen, and Vijay P Singh. A new method for modelling flow duration curves and predicting streamflow regimes under altered land-use conditions. *Hydrological Sciences Journal*, 54(3):606–622, 2009.

[128] Tao Yang, Xi Chen, Chong-Yu Xu, and Zhi-Cai Zhang. Spatio-temporal changes of hydrological processes and underlying driving forces in guizhou region, southwest china. *Stochastic Environmental Research and Risk Assessment*, 23(8):1071, 2009.

[129] S Archfield, R Vogel, P Steeves, S Brandt, P Weiskel, and S Garabedian. The massachusetts sustainable-yield estimator: A decision-support tool to assess water availability at ungaged sites in massachusetts. *US Geological Survey Scientific Investigations Report*, 5227:2010, 2009.

[130] Neil Fennessey and Richard M Vogel. Regional flow-duration curves for ungauged sites in massachusetts. *Journal of Water Resources Planning and Management*, 116(4):530–549, 1990.

[131] Annalise G Blum, Stacey A Archfield, and Richard M Vogel. On the probability distribution of daily streamflow in the united states. *Hydrology and Earth System Sciences*, 21(6):3093, 2017.

[132] DJ Booker and TH Snelder. Comparing methods for estimating flow duration curves at ungauged sites. *Journal of Hydrology*, 434:78–94, 2012.

[133] Jon Olav Skøien and Günter Blöschl. Spatiotemporal topological kriging of runoff time series. *Water Resources Research*, 43(9), 2007.

[134] Alessio Pugliese, William H Farmer, Attilio Castellarin, Stacey A Archfield, and Richard M Vogel. Regional flow duration curves: Geostatistical techniques versus multivariate regression. *Advances in water resources*, 96:11–22, 2016.

[135] Carine Poncelet, Vazken Andréassian, Ludovic Oudin, and Charles Perrin. The quantile solidarity approach for the parsimonious regionalization of flow duration curves. *Hydrological Sciences Journal*, 62(9):1364–1380, 2017.

[136] DA Hughes and V Smakhtin. Daily flow time series patching or extension: a spatial interpolation approach based on flow duration curves. *Hydrological Sciences Journal*, 41(6):851–871, 1996.

[137] NM Fennessey. *A hydro-climatological model of daily streamflow for the northeast United States: Medford, MA, Tufts University*. PhD thesis, Ph. D. dissertation, variously paged, 1994.

[138] Michael E Wieczorek, Shannon E Jackson, and Gregory E Schwarz. Data release: Select attributes for nhdplus version 2.1 reach catchments and modified network routed upstream watersheds for the conterminous united states, 2017.

[139] Jonathan RM Hosking. L-moments: analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 105–124, 1990.

[140] Jonathan Richard Morley Hosking and James R Wallis. *Regional frequency analysis: an approach based on L-moments*. Cambridge University Press, 2005.

[141] William H Asquith. *Distributional analysis with L-moment statistics using the R environment for statistical computing*. CreateSpace, 2011.

[142] QJ Wang. Direct sample estimators of l moments. *Water resources research*, 32(12):3617–3619, 1996.

[143] Richard M Vogel and Neil M Fennessey. L moment diagrams should replace product moment diagrams. *Water Resources Research*, 29(6):1745–1752, 1993.

[144] J Arthur Greenwood, J Maciunas Landwehr, Nicolas C Matalas, and James R Wallis. Probability weighted moments: definition and relation to parameters of several distributions expressable in inverse form. *Water Resources Research*, 15(5):1049–1054, 1979.

[145] William H Asquith. Parameter estimation for the 4-parameter asymmetric exponential power distribution by the method of l-moments using r. *Computational Statistics & Data Analysis*, 71:955–970, 2014.

[146] Abraham Ayebo and Tomasz J Kozubowski. An asymmetric generalization of gaussian and laplace laws. *Journal of Probability and Statistical Science*, 1(2):187–210, 2003.

[147] Pedro Delicado and MN Goria. A small sample comparison of maximum likelihood, moments and l-moments methods for the asymmetric exponential power distribution. *Computational Statistics & Data Analysis*, 52(3):1661–1673, 2008.

[148] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[149] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

[150] Li Deng, Dong Yu, et al. Deep learning: methods and applications. *Foundations and Trends® in Signal Processing*, 7(3–4):197–387, 2014.

[151] Jean Marçais and Jean-Raynald de Dreuzy. Prospective interest of deep learning for hydrological inference. *Groundwater*, 55(5):688–692, 2017.

[152] Ronen Eldan and Ohad Shamir. The power of depth for feedforward neural networks. In *Conference on Learning Theory*, pages 907–940, 2016.

[153] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent trends in deep learning based natural language processing. *arXiv preprint arXiv:1708.02709*, 2017.

[154] Sercan O Arik, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Jonathan Raiman, Shubho Sengupta,

et al. Deep voice: Real-time neural text-to-speech. *arXiv preprint arXiv:1702.07825*, 2017.

[155] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[156] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[157] Gary Marcus. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*, 2018.

[158] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 427–436, 2015.

[159] Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? In *Advances in neural information processing systems*, pages 2654–2662, 2014.

[160] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[161] Aurélien Géron. Hands-on machine learning with scikit-learn and tensorflow: concepts, tools, and techniques to build intelligent systems. *Sebastopol, CA: O*, 2017.

[162] Michael A Nielsen. Neural networks and deep learning, 2015.

[163] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533, 1986.

[164] Franois Chollet and J J Allaire. *Deep Learning with R*. Manning Press, 2017.

[165] Hanen Borchani, Gherardo Varando, Concha Bielza, and Pedro Larrañaga. A survey on multi-output regression. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5(5):216–233, 2015.

[166] Leo Breiman and Jerome H Friedman. Predicting multivariate responses in multiple linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(1):3–54, 1997.

[167] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research*, 15(1):1929–1958, 2014.

[168] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

[169] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059, 2016.

[170] Jiri Hron, Alexander G de G Matthews, and Zoubin Ghahramani. Variational gaussian dropout is not bayesian. *arXiv preprint arXiv:1711.02989*, 2017.

[171] François Chollet et al. Keras. https://github.com/keras-team/keras, 2015.

[172] William H Farmer and Sara Levin. Characterizing uncertainty in daily streamflow estimates at ungauged locations for the massachusetts sustainable yield estimator. *JAWRA Journal of the American Water Resources Association*, 2017.

[173] An approach to estimate nonparametric flow duration curves in ungauged basins. *Water Resources Research*, 45(10), oct 2009.

[174] C Bracken, KD Holman, B Rajagopalan, and H Moradkhani. A bayesian hierarchical approach to multivariate nonstationary hydrologic frequency analysis. *Water Resources Research*, 2018.

[175] Pete Campana, John Knox, Andrew Grundstein, and John Dowd. The 2007-2009 drought in athens, georgia, united states: A climatological analysis and an assessment of future water availability1. *JAWRA Journal of the American Water Resources Association*, 48(2):379–390, 2012.

[176] Richard Seager, Alexandrina Tzanova, and Jennifer Nakamura. Drought in the southeastern united states: causes, variability over the last millennium, and the potential for future hydroclimate change. *Journal of Climate*, 22(19):5021–5045, 2009.

[177] Thomas C Brown, Romano Foti, and Jorge A Ramirez. Projected freshwater withdrawals in the united states under a changing climate. *Water Resources Research*, 49(3):1259–1276, 2013.

[178] Sujoy B Roy, Limin Chen, Evan H Girvetz, Edwin P Maurer, William B Mills, and Thomas M Grieb. Projecting water withdrawal and supply for future decades in the us under climate change scenarios. *Environmental science & technology*, 46(5):2545–2556, 2012.

[179] Daniel R Cayan, Tapash Das, David W Pierce, Tim P Barnett, Mary Tyree, and Alexander Gershunov. Future dryness in the southwest us and the hydrology of the early 21st century drought. *Proceedings of the National Academy of Sciences*, 107(50):21271–21276, 2010.

[180] Thomas C Brown et al. Projecting us freshwater withdrawals. *Water Resources Research*, 36(3):769–780, 2000.

[181] Mohamad I Hejazi, Nathalie Voisin, Lu Liu, Lisa M Bramer, Daniel C Fortin, John E Hathaway, Maoyi Huang, Page Kyle, L Ruby Leung, Hong-Yi Li, et al. 21st century united states emissions mitigation could increase water stress more than the climate change it is mitigating. *Proceedings of the National Academy of Sciences*, 112(34):10635–10640, 2015.

[182] USEPA. USEPA safe drinking water information system database, 2017. https://www3.epa.gov/enviro/facts/sdwis/search.html.

[183] Molly A Maupin, Joan F Kenny, Susan S Hutson, John K Lovelace, Nancy L Barber, and Kristin S Linsey. Estimated use of water in the united states in 2010. Technical report, US Geological Survey, 2014.

[184] MN Sawka. Dietary reference intakes for water, potassium, sodium, chloride, and sulfate. *The National Academies Press: Washington, DC, USA*, 2005.

[185] Peter H Gleick. Basic water requirements for human activities: Meeting basic needs. *Water international*, 21(2):83–92, 1996.

[186] William B DeOreo, Peter W Mayer, Benedykt Dziegielewski, and Jack Kiefer. *Residential end uses of water, version 2*. Water Research Foundation, 2016. ISBN 978-1-60573-236-7.

[187] Kristina Donnelly and Heather Cooley. Water use trends in the united states. *Pacific Institute*, 2015.

[188] Ezra B Whitman. Per capita water consumption. *American Water Works Association*, 24(4):515–528, 1932.

[189] Bradley Jorgensen, Michelle Graymore, and Kevin O'Toole. Household water use behavior: An integrated model. *Journal of environmental management*, 91(1):227–236, 2009.

[190] Joachim Schleich and Thomas Hillenbrand. Determinants of residential water demand in germany. *Ecological economics*, 68(6):1756–1769, 2009.

[191] Robert C Balling, Patricia Gober, and Nancy Jones. Sensitivity of residential water consumption to variations in climate: an intraurban analysis of phoenix, arizona. *Water Resources Research*, 44(10), 2008.

[192] Geoffrey J Syme, Quanxi Shao, Murni Po, and Eddy Campbell. Predicting and understanding home garden water use. *Landscape and Urban Planning*, 68(1):121–128, 2004.

[193] Gary D Gregory and Michael Di Leo. Repeated behavior and environmental psychology: the role of personal involvement and habit formation in explaining water consumption1. *Journal of Applied Social Psychology*, 33(6):1261–1296, 2003.

[194] Austin S Polebitski and Richard N Palmer. Seasonal residential water demand forecasting for census tracts. *Journal of Water Resources Planning and Management*, 136(1):27–36, 2009.

[195] Subhrajit Guhathakurta and Patricia Gober. The impact of the phoenix urban heat island on residential water use. *Journal of the American Planning Association*, 73(3):317–329, 2007.

[196] Mary E Renwick and Richard D Green. Do residential water demand side management policies measure up? an analysis of eight california water agencies. *Journal of Environmental Economics and Management*, 40(1):37–55, 2000.

[197] Lily House-Peters, Bethany Pratt, and Heejun Chang. Effects of urban spatial structure, sociodemographics, and climate on residential water consumption in hillsboro, oregon1. 2010.

[198] Mark Hoffmann, Andrew Worthington, and Helen Higgs. Urban water demand with fixed volumetric charging in a large municipality: the case of brisbane, australia. *Australian Journal of Agricultural and Resource Economics*, 50(3):347–359, 2006.

[199] Randolph C Martin and Ronald P Wilder. Residential demand for water and the pricing of municipal water services. *Public Finance Quarterly*, 20(1):93–102, 1992.

[200] Weiqi Zhou, Austin Troy, J Morgan Grove, and Jennifer C Jenkins. Can money buy green? demographic and socioeconomic predictors of lawn-care expenditures and lawn greenness in urban residential areas. *Society and Natural Resources*, 22(8):744–760, 2009.

[201] E Clery and R Rhead. Education and attitudes towards the environment. *Background paper prepared for the education for all global monitoring report 2013*, 4, 2013.

[202] A. Sankarasubramanian, J. L. Sabo, K. L. Larson, S. B. Seo, T. Sinha, R. Bhowmik, A. Ruhi Vidal, K. Kunkel, G. Mahinthakumar, E. Z. Berglund, and J. Kominoski. Synthesis of public water supply use in the u.s.: Spatio-temporal patterns and socio-economic controls. *Earth's Future*, 2017.

[203] Víctor Corral-Verdugo, Giuseppe Carrus, Mirilia Bonnes, Gabriel Moser, and Jai BP Sinha. Environmental beliefs and endorsement of sustainable development principles in water conservation: Toward a new human interdependence paradigm scale. *Environment and Behavior*, 40(5):703–725, 2008.

[204] Rachelle M Willis, Rodney A Stewart, Kriengsak Panuwatwanich, Philip R Williams, and Anna L Hollingsworth. Quantifying the influence of environmental and water conservation attitudes on household end use water consumption. *Journal of environmental management*, 92(8):1996–2009, 2011.

[205] Richard A Berk, Daniel Schulman, Matthew McKeever, and Howard E Freeman.

Measuring the impact of water conservation campaigns in california. *Climatic Change*, 24(3):233–248, 1993.

[206] Elena Domene and David Saurí. Urbanisation and water consumption: Influencing factors in the metropolitan region of barcelona. *Urban Studies*, 43(9):1605–1623, 2006.

[207] Heather E Campbell, Ryan M Johnson, and Elizabeth Hunt Larson. Prices, devices, people, or rules: the relative effectiveness of policy instruments in water conservation1. *Review of Policy Research*, 21(5):637–662, 2004.

[208] Jody M Hines, Harold R Hungerford, and Audrey N Tomera. Analysis and synthesis of research on responsible environmental behavior: A meta-analysis. *The Journal of environmental education*, 18(2):1–8, 1987.

[209] Sheila M Olmstead and Robert N Stavins. Comparing price and nonprice approaches to urban water conservation. *Water Resources Research*, 45(4), 2009.

[210] Jasper M Dalhuisen, Raymond JGM Florax, Henri LF De Groot, and Peter Nijkamp. Price and income elasticities of residential water demand: a meta-analysis. *Land economics*, 79(2):292–308, 2003.

[211] R Quentin Grafton, Michael B Ward, Hang To, and Tom Kompas. Determinants of residential water consumption: Evidence and analysis from a 10-country household survey. *Water Resources Research*, 47(8), 2011.

[212] Sheila M Olmstead, W Michael Hanemann, and Robert N Stavins. Water demand under alternative price structures. *Journal of Environmental Economics and Management*, 54(2):181–198, 2007.

[213] Joyeeta Gupta and Pieter van der Zaag. Interbasin water transfers and integrated

water resources management: Where engineering, science and politics interlock. *Physics and Chemistry of the Earth, Parts A/B/C*, 33(1):28–40, 2008.

[214] Vinod Mahat, Thomas C Brown, and J. A Ramirez. Twenty-first-century climate in cmip5 simulations: Implications for snow and water yield across the contiguous united states. *Journal of Hydrometeorology*, 18(8):2079–2099, 2017.

[215] David J Hess, Christopher A Wold, Scott C Worland, and George M Hornberger. Measuring urban water conservation policies: toward a comprehensive index. *JAWRA Journal of the American Water Resources Association*, 53(2):442–455, 2017.

[216] Cheng Quan, Shuang Han, Torsten Utescher, Chunhua Zhang, and Yu-Sheng Christopher Liu. Validation of temperature–precipitation based aridity index: Paleoclimatic implications. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 386:86–95, 2013.

[217] Stephen Raudenbush and Anthony S Bryk. A hierarchical model for studying school effects. *Sociology of education*, pages 1–17, 1986.

[218] Nicholas T Longford. *Random coefficient models*. Springer, 1995.

[219] Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, 2006.

[220] Richard McElreath. *Statistical rethinking: A Bayesian course with examples in R and Stan*, volume 122. CRC Press, 2016.

[221] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*, volume 2. Taylor & Francis, 2014.

[222] William J Browne, David Draper, et al. A comparison of bayesian and likelihood-based methods for fitting multilevel models. *Bayesian analysis*, 1(3):473–514, 2006.

[223] Paul Coomes, Tom Rockaway, Josh Rivard, and Barry Kornstein. *North America residential water usage trends since 1992*. Water Research Foundation, 2010.

[224] Thomas D Rockaway, Paul A Coomes, Joshua Rivard, and Barry Kornstein. Residential water use trends in north america. *Journal AWWA*, 103(2):76–89, 2011.

[225] Nicholas G Polson, James G Scott, et al. On the half-cauchy prior for a global scale parameter. *Bayesian Analysis*, 7(4):887–902, 2012.

[226] Daniel Lewandowski, Dorota Kurowicka, and Harry Joe. Generating random correlation matrices based on vines and extended onion method. *Journal of multivariate analysis*, 100(9):1989–2001, 2009.

[227] AWWA 2010. 2010 water and wastewater rate survey, raftelis finan. consult., denver, colo, 2010. Available at https://www.awwa.org/store/productdetail.aspx?productid=61841567.

[228] Stan Development Team. Stan modeling language users guide and reference manual, 2016. Version 2.15.0. http://mc-stan.org.

[229] Stan Development Team. Rstan: the r interface to stan, 2016. R package version 2.14.1 http://mc-stan.org.

[230] Richard McElreath. rethinking: Statistical rethinking book package, 2017. R package version 1.62, https://github.com/rmcelreath/rethinking.

[231] Andrew Gelman and Donald B Rubin. Inference from iterative simulation using multiple sequences. *Statistical science*, pages 457–472, 1992.

[232] Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, pages 1–20, 2016.

[233] Sumio Watanabe. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(Dec):3571–3594, 2010.

[234] Thomas Karl and Walter James Koss. *Regional and national monthly, seasonal, and annual temperature weighted by area, 1895-1983*. National Climatic Data Center, 1984.

[235] David J Hess, Christopher A Wold, Elise Hunter, John Nay, Scott Worland, Jonathan Gilligan, and George M Hornberger. Drought, risk, and institutional politics in the american southwest. In *Sociological Forum*, volume 31, pages 807–827. Wiley Online Library, 2016.

[236] Joan F Kenny, Nancy L Barber, Susan S Huston, Kristin S Linsey, John K Lovelace, and Molly A Maupin. Estimated use of water in the united states in 2005. 2009. https://water.usgs.gov/watuse/data/2005/usco2005.txt.

[237] Chris Anderson. The end of theory: The data deluge makes the scientific method obsolete. *Wired magazine*, 16(7):16–07, 2008.

[238] Tony Hey, Stewart Tansley, Kristin M Tolle, et al. *The fourth paradigm: data-intensive scientific discovery*, volume 1. Microsoft research Redmond, WA, 2009.

[239] Thomas S Kuhn. *The structure of scientific revolutions*. University of Chicago press, 2012.

[240] F Bergadano, V Cutello, and D Gunetti. Abduction in machine learning. In *Abductive Reasoning and Learning*, pages 197–229. Springer, 2000.

[241] Jan-Willem Romeijn. Abducted by bayesians? *Journal of Applied Logic*, 11(4):430–439, 2013.

[242] Karl Popper. *The logic of scientific discovery*. New York: Basic Books, 1959.

[243] Andrew Gelman et al. Induction and deduction in bayesian data analysis. *Rationality, Markets and Morals*, 2(67-78):1999, 2011.