Water Resources Research

RESEARCH ARTICLE
10.1029/2019WR024894

Special Section:
Big Data & Machine Learning in Water Sciences: Recent Progress and Their Use in Advancing Science

A Data-Driven Framework to Characterize State-Level Water Use in the United States

E. Wongso[1], R. Nateghi[1], B. Zaitchik[2], S. Quiring[3], and R. Kumar[4]

[1]School of Industrial Engineering, Purdue University, West Lafayette, IN, USA, [2]Department of Earth and Planatery Sciences, Johns Hopkins University, Baltimore, MD, USA, [3]Department of Geography, Ohio State University, Columbus, OH, USA, [4]UFZ-Helmholtz Centre for Environmental Research, Leipzig, Germany

Abstract Access to credible estimates of water use is critical for making optimal operational decisions and investment plans to ensure reliable and affordable provisioning of water. Furthermore, identifying the key predictors of water use is important for regulators to promote sustainable development policies to reduce water use. In this paper, we propose a data-driven framework, grounded in statistical learning theory, to develop a rigorously evaluated predictive model of state-level, per capita water use in the United States as a function of various geographic, climatic, and socioeconomic variables. Specifically, we compare the accuracy of various statistical methods in predicting the state-level, per capita water use and find that the model based on the random forest algorithm outperforms all other models. We then leverage the random forest model to identify key factors associated with high water-usage intensity among different sectors in the United States. More specifically, irrigated farming, thermoelectric energy generation, and urbanization were identified as the most water-intensive anthropogenic activities, on a per capita basis. Among the climate factors, precipitation was found to be a key predictor of per capita water use, with drier conditions associated with higher water usage. Overall, our study highlights the utility of leveraging data-driven modeling to gain valuable insights related to the water use patterns across expansive geographical areas.

Key Points:
- Drier regions are associated with higher water use
- Irrigated farming, thermoelectric energy generation, and urbanization are the most water-intensive anthropogenic activities
- Random forest outperforms all other tested algorithms in predicting the state-level, per capita water use

Supporting Information:
- Supporting Information S1

Correspondence to:
R. Kumar and R. Nateghi,
rohini.kumar@ufz.de;
rnateghi@purdue.edu

Citation:
Wongso, E., Nateghi, R., Zaitchik, B., Quiring, S., & Kumar, R. (2020). A data-driven framework to characterize state-level water use in the United States. *Water Resources Research*, *56*, e2019WR024894. https://doi.org/10.1029/2019WR024894

Received 31 JAN 2019
Accepted 16 JUN 2020
Accepted article online 8 JUL 2020

1. Introduction

Integrated water resource management has been receiving increasing attention globally (Calder, 2012; Cui et al., 2018; Giordano & Shah, 2014; Loucks & Van Beek, 2017; Rahaman & Varis, 2005). Rapid growth in population and increased rates of economic development and urbanization have resulted in increased demands for fresh water in energy, agriculture, industry, and the commercial and residential sectors, all of which have severely stressed water resources in many regions (Bruss et al., 2019; Obringer & Nateghi, 2018; Worland et al., 2018). Sustainable management of demand for water has been brought into the limelight in the United States following several devastating, multiyear drought episodes in California and the Midwest, which led to adverse impacts on agricultural productivity and energy generation capacity, costing the U.S. economy tens of billions of dollars (Bauer et al., 2014; Chaussee, 2014; Devineni et al., 2015). According to the U.S. Environmental Protection Agency (EPA), 40 out of 50 states will expect water shortages in some portion of their jurisdiction in the next 10 years, even under average conditions (EPA, 2017).

Credible estimates and projections of short-, medium-, and long-term demand for water is valuable for urban planners, regulators, and operators of critical infrastructure systems to ensure reliable and affordable provisioning of many critical services including water (Obringer et al., 2019, 2020). The need for rigorous empirical research related to water management and adaptation has emerged as a critical area (Olmstead, 2014). Optimal investments in the design, operation, modernization, and expansion of water infrastructure systems are largely dependent on access to realistic and credible predictions and projections of the spatiotemporal variability in demand for water (Billings & Jones, 2008). According to (Hall et al., 1989), "the success of any water resource development is critically dependent upon the reliability of the forecasts of future water demands that are employed in its design (and management)."

In this paper, we propose a data-driven framework—grounded in statistical learning theory— to (a) develop rigorously validated models for per capita water uses in various sectors in the United States, (b) identify the key predictors of state-level, per capita water use, (c) understand the relationship between each of the key predictors and per capita water use, and (d) analyze the sensitivity of the water use patterns to changes in

climate variability (e.g., precipitation changes) under changing climate conditions. Our data-driven water use models were developed using state-level, per capita water use data over the past two decades—together with various geographic, climatic, and socioeconomic factors—to identify the key factors that are associated with high water-usage intensity among different sectors in the United States.

We hypothesize that the commonly used parametric empirical models that assume "rigid" functional forms —such as linearity and additivity (e.g., based on the ordinary least squares method)—would not adequately capture the complex dependencies between state-level water use and socioeconomic and geoclimatic conditions and that more robust nonparametric statistical learning algorithms (e.g., ensemble-of-trees) will be more effective in predicting state-level water use. Moreover, given that the largest fraction of water use occurs in the agricultural and thermoelectric generation sectors, we hypothesize that irrigated farming and power generation will be the key predictors of state-level wateruse.

The structure of this paper is as follows. The review of the existing literature in predicting water use is summarized in section 2. Data and methods are introduced in sections 3 and 4, respectively. Results are summarized in section 5, followed by the concluding remarks in section 6.

2. Background

A plethora of research studies have focused on analyzing, predicting, and projecting water demands/uses—with various different spatiotemporal scales and lead time horizons—using a range of methods such as simulation, econometrics, and statistical learning theory. Donkor et al. (2014) reviewed research articles on water demand forecasting—published between 2000 and 2010—to identify useful models for water utility decision making. They concluded that artificial neural networks were more popular for short-term demand forecasts, while econometrics, scenario-based, and simulation models were more likely to be used for making long-term strategic decisions. They also highlighted the value in probabilistic forecasting to capture uncertainties associated with future demand. Sebri (2016) surveyed the empirical literature on urban water forecasting using a meta-analytical approach. Their meta-regression analysis concluded that model accuracy depended on the scale of analysis, the type of approach used, model assumptions, and sample size. Hamoda (1983) examined the impact of socioeconomic factors on the residential water consumption in Kuwait. More specifically, Hamoda (1983) leveraged linear regression to characterize the impacts of income, market value of land, rents of dwellings, and household size on average per capita water consumption. They concluded that the hot climate of Kuwait and its continually improving standards of living were the primary factors contributing to high water consumption rates in the country. More recently, Worland et al. (2018) leveraged Bayesian-hierarchical regression to analyze the variability in public supply water use across the United States. Their study concluded that "the environmental, economic, and social controls on water use are not uniform across the United States." and underscored the importance of "accounting for regional variability to understand the drivers of water use."

Another study by Lutz et al. (1996) leveraged a variation of the Electric Power Research Institute (EPRI) model to study the patterns of residential hot water consumption. Their study shed light on the impacts of efficiency standards for water heaters and other market transformation policies. Jorgensen et al. (2009) analyzed the social factors in residential water use and highlighted the importance of interpersonal and institutional trust for implementation of effective water conservation schemes. Sovacool and Sovacool (2009) implemented a county-level analysis of the water-energy nexus in the United States and concluded that 22 counties will likely face severe water shortages, brought about primarily due to increased capacity expansion in thermoelectric generation. Chandel et al. (2011) leveraged a modified version of the U.S. National Energy Modeling Systems (NEMS) together with thermoelectric water use factors from the Energy Information Administration (EIA) to investigate the impact of various climate change policies on the energy mix. They found that all of the climate policy scenarios that were considered in the study could lead to a reduction in fresh water uses for power generation, compared to the "business-as-usual" scenario and that water use was inversely related to carbon price. Davies et al. (2013) leveraged a Global Change Assessment Model (GCAM)—an integrated assessment modeling of energy, agriculture, and climate change—to assess water intensity associated with electricity generation until 2095. They found that water use would likely decrease with turnover in power plans (i.e., replacing old power plants with new ones with advanced cooling technology).

The majority of the empirical studies to date have focused primarily on either a particular geographical location or a given sector in the United States and leveraged either linear models, in which the assumptions may not be supported by the empirical data or "black-boxes" (e.g., artificial neural network) to project demand. This paper will use state-of-the-art statistical learning techniques to analyze water use data—available from U.S. Geological Survey (USGS) over the past two decades for the entire United States—and develop a *rigorously validated* and *interpretable*predictive water use model as a function of socioeconomic, geographic, and climatic conditions. To address these gaps, we use state-of-the-art statistical learning techniques to analyze water use data—as provided by USGS over the past two decades for the entire United States—and develop a *rigorously validated* and *interpretable* predictive water use model as a function of socioeconomic, geographic, and climatic conditions. It is worth highlighting that in this study, we use the total water use data provided by the USGS and do not distinguish between the consumptive and nonconsumptive water use. The difference between these two basically depends on the degree and form of water use. Specifically, nonconsumptive water use refers to water that can be recycled and reused (e.g., in industrial or domestic applications), while in the consumptive use, water is effectively removed from a system and therefore cannot be recaptured (e.g., water transpired by crops or water lost via evaporation in cooling of thermoelectric power turbines).

It is noteworthy that, though not pursued in this study, there exist another fundamentally different approach to modeling water use—based on complex, mechanistic hydrologic models with integrated elements of human-water interfaces (e.g., Pokhrel et al., 2016; Wada et al., 2017). Models in this category include, for instance, PCR-GLOBWB (Sutanudjaja et al., 2018; Wada et al., 2014), WaterGAP (Alcamo et al., 2003; Flörke et al., 2013), and H08 (Hanasaki et al., 2008a, 2008b). These models have varying ranges of processes, accounting for the coupled human and natural systems. Despite the utility of these models in providing a mechanistic understanding on the functioning of the system, they are inherently complex and difficult to parameterize—partly owing to the limited availability of observational data sets. Different sorts of simplifications and conceptualizations are therefore necessary to model the complex interactions between human and natural systems (e.g., Wada et al., 2008b).

Our proposed data-driven modeling paradigm can be complementary to hydrological modeling efforts by offering key advantages of (a) computational efficiency and (b) requiring a limited set of predictors to reconstruct the continuous space-time evolution of water use, which can then be used to further constrain the parameterization of more complex, mechanistic hydrological models. In summary, our approach can help identify the most water-intensive sectors across various states and inform policy makers, stakeholders, regulators, and researchers on the exiting U.S. water use patterns. It can also help identify sectors and areas where efficiency and conservation mechanisms could yield maximum return in terms of enhanced sustainability of water use in urban areas.

3. Data and Initial Analysis

Data were collected from publicly available sources such as the USGS (2017) website, the EIA (2017), the Bureau of Economic Analysis (BEA, 2017), the U.S. Census Bureau (USCB, 2017), the Climate Prediction Center (CPC), the National Weather Service (NWS) (National Oceanic and Atmospheric Administration [NOAA], 2017a), the U.S. Department of Agriculture (USDA, 2007), the Coastal States Organization (CSO, 2017), the U.S. EPA (2017), and other sources (see Table 1). Below is a brief description of our response variable (i.e., per capita water use) and various socioeconomic, hydroclimatic, and geographic predictors that were used in our analyses. It should be pointed out that since the water use data are only available at 5-year increments, the predictors were processed to match the temporal scale of our response variable.

3.1. Response Variable: Per Capita, State-Level Water Use

State-level water use per capita (in million gallons per day) was selected as our response variable, which contains both consumptive and nonconsumptive water use. Note that we normalized water use by population as opposed to land area, as we did not find an intuitive interpretation for gallons per square feet of water use. However, normalization is an important factor in interpreting the ensuing analyses presented in this paper, and future research on exploring the sensitivity of results to different normalization techniques is needed. Water use data sets were obtained from USGS for the period 1991–2015. Data from 1991–2010 were used to train our data-driven model, while the 2011–2015 range was used as "test" data to assess the accuracy of the developed model. USGS water use data are collected and compiled every 5 years for each of the 50

Table 1
Summary of the Response and Predictor Variables Used for Developing the Water Use Models

Variable type	Variable name	Units	Source
Response	State-level water use	Mgal/d/capita	USGS (2017)
Predictors	Gross state product	US$ M	BEA (2017)
	Household median income	US$	BEA (2017)
	Education level data	% pop	USCB (2017)
	Thermoelectric energy generation	MWh	EIA (2017)
	Cooling degree days	°F	NOAA (2017a)
	Heating degree days	°F	NOAA (2017a)
	Annual precipitation	mm/year	NOAA (2017b)
	Standardized precipitation index	—	NOAA (2017b)
	CPC-modeled soil-water content	m	NOAA (2017a)
	Coastal status indicator	-	CSO (2017)
	Total irrigated farmland area	ha	USDA (2007)

Note. Each variable collected from different sources were spatially and temporally aggregated to match the state-wide, 5-year water use data sets for the period 1990–2015.

states, the District of Columbia, Puerto Rico, and the U.S. Virgin Islands. The data source provides a breakdown of water usage in eight different sectors (Figure 1) including thermoelectric, irrigation, aquaculture, livestock, and mining as well as the consumptive use in public supply, industry, and domestic sectors. Thermoelectric and irrigation are the two dominant sectors that account for almost two thirds of total water use across the United States. We note that there is a large regional variability in water use patterns—the states in the east are dominated by the thermoelectric and industrial water sectors, while irrigation is the main source of water use in the central and western part of the United States. To control for the varying sizes of states, we normalized the state-wide total water use data by the total population of each state. The distribution of state-wise, normalized water use for years of 2006–2010 can be seen in Figure 1 (bottom panel). States highlighted in shades of red represent high per capita water use, while the states in blue represent low per capita water use. Figure 1(bottom panel) reveals that Idaho has the highest per capita water use for the 2006–2010 period.

The distribution of per capita water use (in million gallons per day) for the period 1991–2010 is depicted in Figure 2. The distribution of per capita water use is right skewed and has a heavy-tail distribution. In fact, it can be seen that the power-law distribution provides a reasonable fit to the tail of the data (red line in Figure 2a). Power-law distributions describe phenomena where large events are quite rare but small events are very frequent. Figure 2 suggests that a small fraction of the states in the United States tend to consume disproportionately large volumes of water on a per capita basis. However, these numbers have to be cautiously interpreted as they do not reflect the virtual water use. For instance, the virtual water represented by growing Idaho potatoes is consumed mainly outside of Idaho. Similar statements can be made about other agricultural products as well as the export of thermoelectric generated electricity.

3.2. Socioeconomic Predictors

Gross state product (GSP) data (in millions of USD) were collected from the U.S. BEA for the years of 1991–2010. Household Median Income data (in USD) were collected from the Bureau of Labor Statistics. The values of GSP and income data were converted to a common baseline time period accounting for the deflation of GDP as well the Consumer Price Index Research Series Using Current Methods (CPI-U-RS), respectively.

The education level data—obtained from the USCB—contain the following four levels for each reported year: (a) percentage of population with less than high school diploma, (b) percentage of population with high school diploma only, (c) percentage of population some college (1–3 years), and (d) percentage of population with 4 years of college or higher. We leveraged generalized additive models (GAMs) to impute the missing data and align the temporal scale of the education data with that of water use. The premise for including this variable in the analysis is to test whether educational levels are predictive of the public supply wateruse.

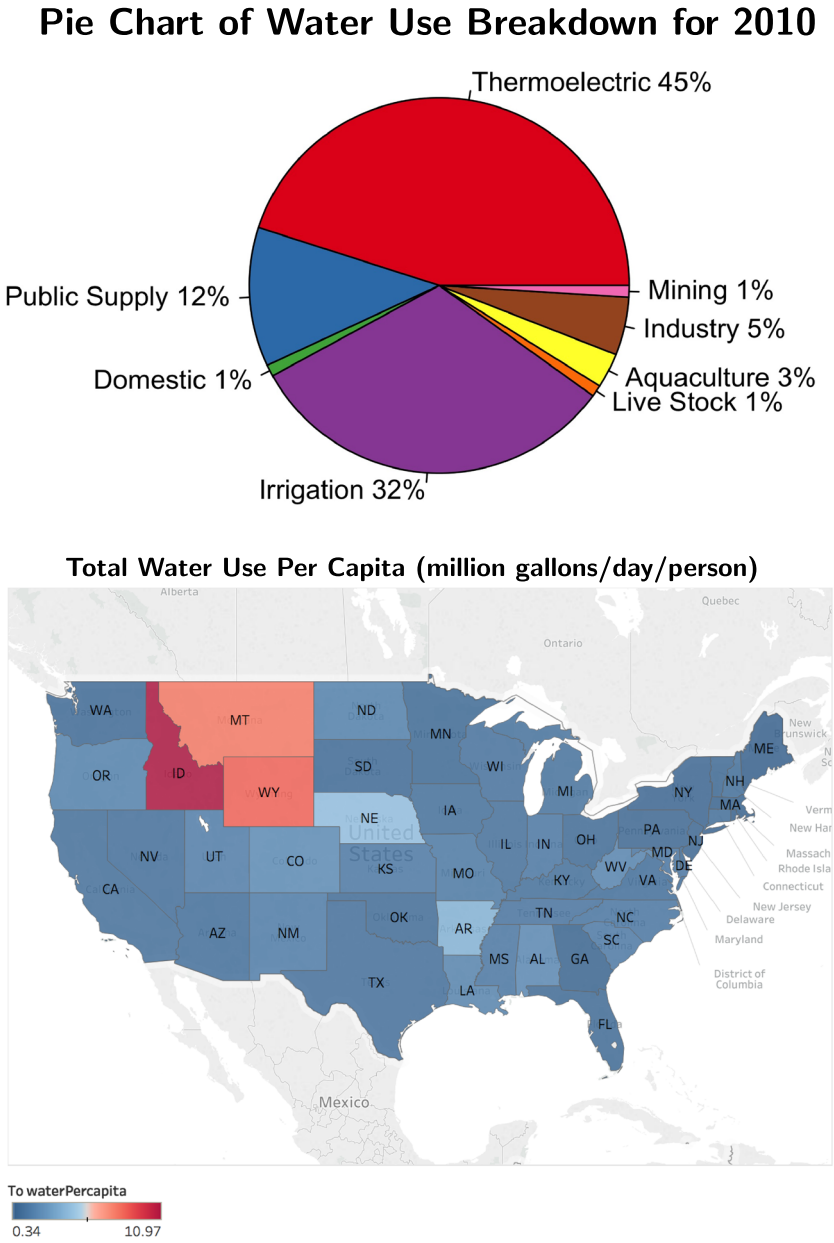

Figure 1. Top: The breakdown of U.S.-wide water uses across the eight major sectors during the period 2006–2010. Bottom: Spatial distribution of the U.S.-wide per capita water use (in million gallons per day).

Data related to thermoelectric energy generation (in mega watt-hours) (e.g., coal, petroleum, and gas-fired plants and nuclear and geothermal technologies) were collected from the EIA. Coal production, available from the EIA, was used as a proxy for mining industry, since coal is the largest profit-generating mining production in the United States. The percentage of urban population data were collected from the U.S. Census. Since the temporal scale of the urban population data was decadal, the years did not match the years in the USGS water data set. Therefore, we imputed the missing years of data using GAMs to match the years across the two datasets.

3.3. Hydroclimatic and Geographic Predictors

Time series data related to cooling degree days (CDDs) and heating degree days (HDDs) are based on variation in air temperature estimates, available from CPC and NWS. Other hydroclimatic variables such as the

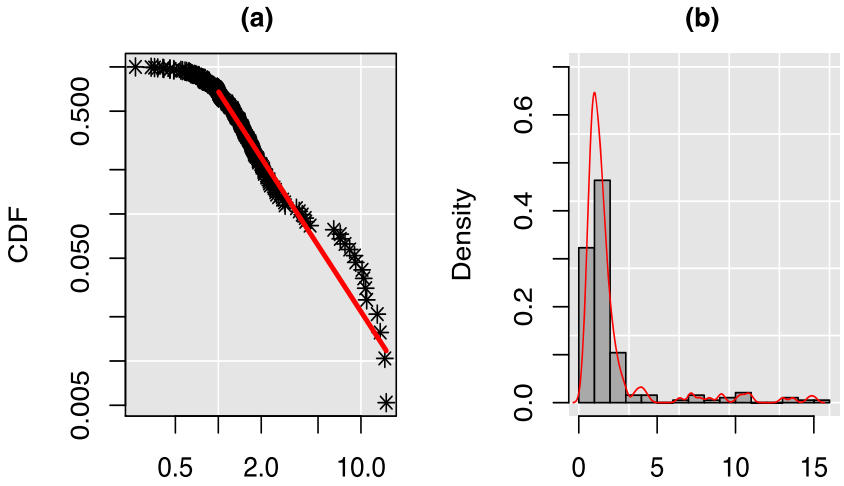

Figure 2. The empirical distribution of per capita water uses (in million gallons per day) for the period 1991–2010: (a) the red line shows that power law fits the tail of the empirical cumulative distribution reasonably well; and (b) the histogram of per capita water demand with overlain kernel density line (in red).

standardized precipitation index (SPI), soil moisture, and annual precipitation data were extracted from the National Centers for Environmental Information. The SPI characterizes the inter- and intra-annual variability of precipitation, with positive values indicating wetter than normal conditions and the negative values indicating drier than normal conditions (Hayes et al., 2010; McKee et al., 1993). Additionally, we

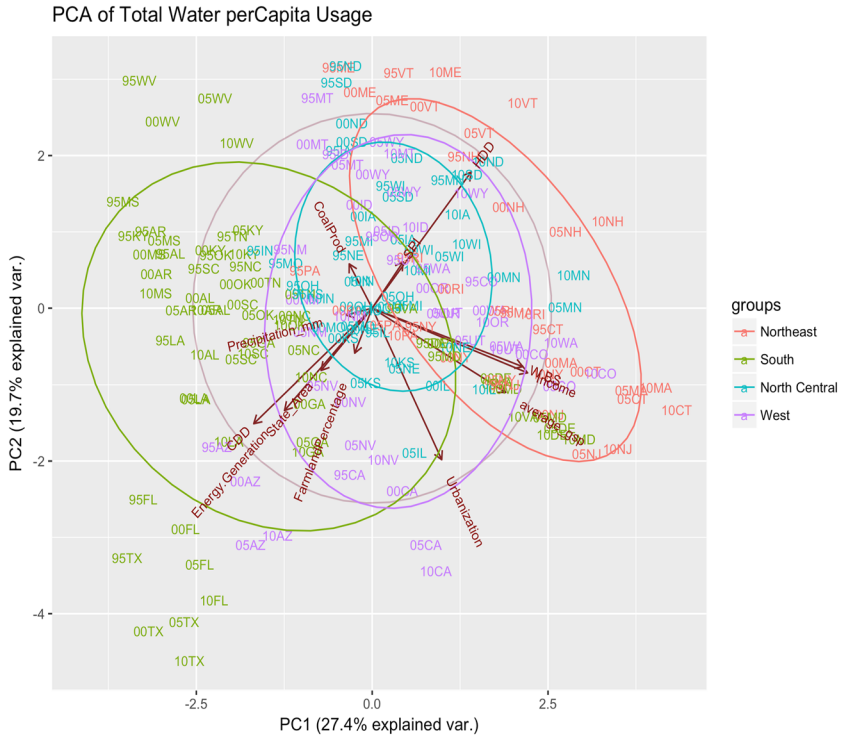

Figure 3. Principal component analysis (PCA) biplot of the per capita water use (in million gallons per day) for the period 1995–2010. The states are color coded based on their proximity to water bodies, and the two digits next to the state codes indicate the year associated with the water use data for the state. On the biplot of leading two principal components (PC1 and PC2), states are identified by year of data collection and standard state abbreviation; e.g., 95TX = Texas, 1995 data.

used the upper 1-m simulated soil-water content (mm), based on the CPC model-based simulations, to represent the near-surface wet and dry conditions (see Fan & van den Dool, 2004, for more details).

Coastal status was calculated for each state by creating dummy variables indicating whether the state borders (a) the Atlantic Ocean, (b) the Pacific Ocean, (c) the Gulf of Mexico, or (d) the Great Lakes. Those states were coded as "1," and otherwise as "0." The estimates of the total irrigated farmland area were collected from the Census of Agriculture Farm and Ranch Irrigation Survey (2008), conducted by the National Agricultural Statistics Service (NASS) in the USDA. The surveys are conducted every five years, starting from year 1992. To align the time steps of the farm data with that of water usage, we used data from 1992 to represent irrigated farmland size between 1991, 1995, and 1997 data were used to represent the value between 1996 and 2000. We normalized the data by the total land size of each state to obtain the percentage of irrigated farmland area per state. Prior to the analysis and the model setup, all predictor variables were aggregated spatially and temporally to match the state-wide, 5-year water use data sets.

3.4. Exploratory Data Visualization and Analysis

A "biplot" is a useful visualization technique that attempts to show multivariate data on the same plot. One of the most commonly used types of a biplot is based on principal component analysis (PCA) (James et al., 2013). A PCA biplot is a two-dimensional representation of multivariate data, using only the first two principle components. In a PCA biplot, vector lengths approximate standard deviations, and the cosines of their angles are proportional to the correlation between the variables.

It can be seen from Figure 3 that over the years of 1995–2010, the state-level water use did not change significantly. For example, on the bottom left corner of the plot, we observe that water use of Arizona, Louisiana, Texas, and Florida are located close to each other across the different years. The energy generation and CDD vectors, extended in the direction of Texas, suggesting that thermoelectric power generation and the warmer climate can explain the variance of water usage in Texas, as opposed to water usage in the states of Colorado or North Dakota, which lie close to the HDD vector. Moreover, Figure 3 reveals that while water usage in the densely populated states in the Northeast can be explained by socioeconomic factors such as income and education and measures of urbanization, water use in the larger Midwestern and Western states of North and South Dakota, Nebraska, Iowa, and New Mexico tends to be dominated by farming and mining practices.

4. Methodology

The existing empirical literature in water resources analysis reveal a unilateral focus on descriptive and explanatory statistical modeling. Predictive modeling of water resources analysis has largely been underexplored. To address this gap, we have proposed a data-driven, predictive framework to analyze state-level water use in the United States. Unlike descriptive or explanatory modeling, which is concerned with best explaining the past variability in the data, predictive modeling is concerned with predicting new or unseen data. The expected prediction error (*EPE*) for a new observation x can be summarized by the equation below

$$\begin{aligned} EPE &= E\left[Y-\hat{f}(x)\right]^2 \\ &= E[Y-f(x)]^2+\left[E\left(\hat{f}(x)\right)-f(x)\right]^2+E\left[\hat{f}(x)-E\left(\hat{f}(x)\right)\right] \\ &= Var(Y)+Bias^2+Var\left(\hat{f}(x)\right) \end{aligned} \tag{1}$$

The first term represents the irreducible error, which is the result of the inherent stochasticity in any process. The second term (the bias) represents how closely the estimated function mimics the process of interest, and the third term (variance) arises due to using (noisy) samples to estimate the response function. Descriptive and explanatory statistical models often focus on reducing the bias of the estimate. However, predictive modeling focuses on minimizing the bias and variance *simultaneously*. With the recent accelerated rate of large and complex data sets becoming available, predictive modeling can be leveraged as a powerful tool to identify complex and nonlinear dependencies that can lead to generating new hypothesis and advance the scientific discovery in the field.

In the next section, we will present a brief discussion on supervised learning theory and predictive modeling. We will then present a detailed discussion of the algorithms that were used to develop predictive models of state-level water uses.

4.1. Supervised Learning Theory (Predictive Modeling)

This paper proposes a data-driven framework, grounded in supervised learning theory, to develop predictive models for state-level water uses and identify their most important predictors of in the United States. The main objective of supervised learning is to estimate a system of interest (e.g., water uses) as a function of various independent predictors (e.g., geographic, climatic, and socioeconomic factors). Mathematically, the prediction process can be summarized by $y = f(X) + \epsilon$, where the stochastic additive Gaussian noise ϵ represents the dependence of y on factors other than X that are not "controllable." The goal of supervised learning is to leverage the observed records and estimate the response $\hat{f}(X)$ (i.e., water use) such that the loss function of interest L is minimized over the entire domain of the input data space

$$L = \int w(X)\Delta\left(\hat{f}(x), f(x)\right)dX \tag{2}$$

where $w(X)$ is a possible weight function and Δ represents the Euclidean distance (or other measures of distance). The value of L in the equation above characterizes the accuracy of the estimate over the entire domain (Hastie et al., 2009).

Since the goal of this study was not only to identify the key predictors of state-level water uses but also to characterize their relationship with water use, model interpretability was a key factor in algorithm selection. Therefore, the methods of deep learning and support vector regression were not considered in this analysis. This is because these learning methods generally involve several rounds of data transformations from the input to the output layer, rendering the final model difficult to interpret. Among the rich library of algorithms available in supervised learning, we selected algorithms that ranged from easily interpretable *parametric models* with rigid model structure (e.g., generalized linear model [GLM]), to more flexible *nonparametric models* (e.g., ensemble-of-tree approaches such as random forest [RF]), as well as *semiparametric models* (e.g., multivariate adaptive regression splines [MARS] and GAMs) that allow for adding "local" complexity over the input parameter space while still allowing for parameterized interpretability. Theoretical details of the algorithms used in our analysis are detailed below.

4.1.1. Generalized Linear Model

GLM is an extension of ordinary linear regression (OLR), relaxing the assumptions of linearity and normality in OLR (Nelder & Wedderburn, 1972). In GLM, the dependent variable is assumed to be generated from a particular distribution in an exponential family. GLM is one of the most widely used methods for function approximation. GLM is mathematically summarized below

$$E(y_i) = g^{-1}(X_i'\beta) \tag{3}$$

where $E(y_i)$ is the expected value of the independent variable, β is the slope parameter, and $g()$ is the link function. GLMs are popular because they can be easily fitted (even with limited data) and they are easily interpretable. However, their rigid structure often fails to approximate the true function, especially when the response is a complex (nonlinear) function of input variables. Their predictive accuracy is therefore often inferior to more flexible models (James et al., 2013).

4.1.2. Generalized Additive Models

GAM is a natural extension from GLM, in order to preserve the additive model while extending to nonlinear relationship between the response and predictors (Hastie et al., 2009). GAM leverages a nonparametric (local parametric) fitting procedure where the conditional expectation of y is related to the input variables space as shown below:

$$y_i = \beta_0 + \sum_{j=1}^{p} f_j(x_{i,j}) + \epsilon_i \tag{4}$$

where $f_j(x_{i,j})$ is a smoothing splines over the p-dimensional input space, with $i = 1,\ldots, n$. GAM relaxes the linearity assumption of multiple linear regression with smoothing functions $f_j(x_{i,j})$. This allows for

capturing any nonlinear relationships between the predictors and the response variable. The flexibility of GAM often results in better approximating the true function and therefore often outperforms GLM in predictive accuracy.

4.1.3. Multivariate Adaptive Regression Splines

MARS is a nonparametric regression techniques developed by Friedman (1991). It extends the use of piecewise linear basis function of form $(x-t)_+$ and $(x-t)_-$, where

$$(x-t)_+ = \begin{cases} x-t & x > t \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

$$(x-t)_- = \begin{cases} t-x & x<t \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

and MARS has the function form of

$$f(X) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(X) \tag{7}$$

where each $h_m(X)$ is a function in form of piecewise linear basis function or the product of two or more such functions. The coefficients β_m are estimated by minimizing the residual sum of squares given the choices of $h_m(X)$(Hastie et al., 2009).

4.1.4. Random Forests

Decision trees are the basic building blocks of the RF algorithm. Trees are "learned" by splitting the data space recursively into subregions (nodes). The split variables and the split values are selected based on maximizing a preselected fitting criterion. In growing a tree, each observation is assigned to a unique terminal node, where the conditional distribution of the univariate response variable y is estimated. To avoid overfitting, the grown tree is "pruned" back based on some cost-complexity criterion. While trees are generally good at capturing the structure of the data and therefore low in bias, they can be quite unstable and suffer from high variance. To minimize the variance, meta-algorithms such as bagging (e.g., bootstrapped aggregating) can be applied to tree-based methods to create tree-ensembles. Tree-ensembles are low in both bias and variance and therefore offer robust predictive accuracy, rendering them a highly popular technique for assessing the risk, sustainability, and resilience of critical infrastructure systems such as water and energy systems (Mukherjee et al., 2018; Mukhopadhyay & Nateghi, 2017; Nateghi, 2018; Obringer & Nateghi, 2018).

RF is an ensemble decision tree-based method developed by Breiman (2001) and can be mathematically represented as

$$F(x) = \frac{1}{m_{\text{tree}}} \sum_{i=1}^{m_{\text{tree}}} T_i(x) \tag{8}$$

where T_i is a single decision tree, trained on bootstrap samples from the original data and x represent a p-dimensional vector of input data predictors (e.g., the geographic, climatic, and socioeconomic factors used in this analysis). The subset of predictors for building each decision tree is randomly selected, and best split values are chosen such that the sum of squared errors (or least absolute deviation) within each node of T_i is minimized. Each decision tree is developed by recursively splitting the data space into terminal nodes, until each terminal node contains no more than a certain predefined minimum number of records. The average is then assigned to the terminal nodes. $F(x)$ estimates the response value, by aggregating m such decision trees. The estimation of prediction error of RF can be obtained by leveraging the out-of-bag (OOB) data (i.e., the test data that were set aside during the development of each tree and not used in building that tree) to compute the mean square error as below

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - y_i')^2 \tag{9}$$

where y_i' is the average OOB predictions data for the ith observation (Liaw & Wiener, 2002).

Table 2
Summary of Models Performance Given as Correlation Coefficient (R^2), Fitted Root Mean Square Error (RMSE; Million GPCD), and Leave-One-Out Cross-Validation (LOOCV) RMSE

Model	R^2	RMSE	LOOCV RMSE
Mean-ONLY	—	2.60	2.62
Multiple Linear Regression (MLR)	0.57	1.71	1.84
Generalized Additive Model (GAM)	0.61	1.62	1.62
Multivariate Adaptive Regression Splines (MARS)	0.85	0.99	1.40
Random Forest (RF)	**0.97**	**0.47**	**0.98**

Note. Each model is trained and tested using all available data records for the period 1991–2010. The model (RF) performing best is shown in bold.

Since the method of RF is nonparametric, partial dependence plots (PDPs) can be used to implement variable inference. PDPs calculate the marginal effects of a given predictor variables x_j in a *ceteris paribus* condition (i.e., controlling for all the other predictors). Mathematically, the estimated PDP is given as (Hastie et al., 2009)

$$(\hat{f}_J)(x_j) = \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_J)(x_j, x_{-j,i}) \tag{10}$$

where $\hat{f}_J$ is the approximation of the true function that generates y, n is the size of the response vector (i.e., the size of the training data set), and x_{-j} represents all input variables except x_j. The estimated PDP of the predictor x_{-j} provides the average value of the function $\hat{f}$ when x_j is fixed and x_{-j} varies over its marginal distribution.

4.1.5. Bias-Variance Trade-Off

Predictive performance of a statistical model depends on its capability to yield accurate predictions for an independent test sample. Generally, simple models are more stable but do not adequately estimate the structure of the true function and therefore are high in bias. Complex models can approximate the shape of the true function, more effectively, but they are prone to overfitting and therefore have high variance. The bias-variance trade-off lies at the heart of developing models with high generalization power (Hastie et al., 2009; Shmueli, 2010). Cross-validation is one of the most widely used methods in balancing bias and variance. We use leave-one-out cross-validation (LOOCV) to estimate predictive accuracy. The LOOCV procedure is defined as holding out 1 data point as the test set and using the rest of data as the training set. The model generated from the training data is then used to predict the test set and the MSE of that point is calculated. LOOCV MSE is defined by

$$MSE_{\text{LOOCV}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{11}$$

where i represents the iteration with 1 data point being left out, y_i represents the true value of the ith iteration, $\hat{y_i}$ represents the predicted value, and n represents the length of data.

5. Results and Discussion

We first compare the accuracy of the range of predictive models developed based on the parametric and nonparametric algorithms discussed in section 4. We hypothesize that the more flexible, nonparametric models will outperform the more rigid parametric models due to their ability to account for complex and nonlinear data dependencies.

Table 2 summarizes the performance of each model. The first column summarizes the goodness of fit for each of the models. MARS and the method of RF fit the data substantially better compared to GLM and GAM. The second and third columns in Table 1 show the in-sample and out-of-sample root mean squared errors for each of the models. Again, it can be observed that MARS and RF are competitive in terms of in-sample fit, but RF significantly outperforms all other models, in terms of out-of-sample accuracy. In fact, the analysis of variance test on the prediction errors of the different models revealed statistically significant differences between the mean errors, with a p value $<2 \times 10^{-16}$.

Table 3 summarizes model fit and predictive accuracy. Based on the results summarized in the table and the plot (Figure 4), it can be inferred that RF outperforms all other models. In fact, RF is able to estimate the water use above 5 million GPCD (gallons per capita per day) accurately, even though there are fewer observation points. While MARS performs well below 5 million GPCD (where there are more observations), it performs poorly for more extreme values, where the observations are more sparse. Figure 4 visualizes the fit of each of the prediction models. As initially hypothesized, the predictions based on the RF algorithm substantially outperform all other models in terms of the goodness of fit. The model developed using the RF algorithm was therefore selected as the best model.

Table 3
Summary of Models Predictive Accuracy to Enable Model Selection Based on Their Out-of-Sample Performance

Model	R^2	RMSE	LOOCV RMSE	Prediction RMSE
Mean-ONLY	—	2.75	2.77	2.11
Multiple Linear Regression (MLR)	0.59	1.76	2.00	1.52
Generalized Additive Model (GAM)	0.65	1.63	1.68	1.31
Multivariate Adaptive Regression Splines (MARS)	0.95	0.60	1.57	1.35
Random Forest (RF)	**0.97**	**0.48**	**1.00**	**0.79**

Note. Given that the data available at the time of model development were for the period of 1991–2010, each model was trained using 1991–2005 data and assessed using the "validation set" spanning 2006–2010. Summary performance is presented here in terms of correlation coefficient (R^2), fitted root mean square error (RMSE; million GPCD), leave-one-out cross-validation (LOOCV) RMSE, and prediction RMSE (for the test data). See the text on section 4.1.5 for more details on LOOCV-RMSE. The model (RF) performing best is shown in bold.

In order to test the performance of the model—in terms of its capability to forecast water use for a future period—we trained the RF algorithm with the data until the end of 2010. In order to predict water use, we selected an independent testing period of 2011–2015, which was released by the USGS at the end of the year 2018 (and was not available at the time when the RF-based model was originally developed). Figure 5 depicts the observed versus predicted values for water use during 2011-2015, revealing the model's remarkable performance across all states with the exception of five, namely, Arkansas (AR), Nebraska (NE), Montana (MT), South Dakota (SD), and Wyoming (WY) for which our model underestimated wateruse.

These results confirm our initial hypothesis that linear-based models (e.g., GLM) and additive structures such as GAM are not able to capture the complex relationships in the data adequately. Moreover, the fact that RF outperformed MARS is not surprising. MARS can be seen as an extension of recursive partitioning algorithms such as tree-based methods (Friedman, 1991), which is very effective at capturing high order interactions and yielding low-bias estimates. However, the model is not as effective as RF in variance reduction and therefore has an inferior predictive power.

We leveraged a data-driven variable selection, based on an algorithm proposed by Genuer et al. (2010), to implement input variable reduction for the RF model. The variable selection algorithm involved developing multiple forests and ranking their input variables based on their importance—by calculating their contribution to out-of-sample predictive accuracy—and their standard deviations. Variables at the bottom of the list (in terms of importance) whose standard deviation was below the minimum calculated threshold were removed. Multiple nested models were then developed in a stepwise forward strategy. The smallest subset of input data that yielded the best predictive accuracy was retained for the final model. The list of the final key variables selected for each sector are shown in Figure 6.

The importance plot shows the ranking of the variables in terms of their contribution to the model's out-of-sample predictive performance, with the variable highest on the *y*-axis contributing the most to model's performance. It can be observed that the percentage of irrigated farmland is the most important predictor of state-level per capita water use, followed by total state-level precipitation, HDDs, urbanization,

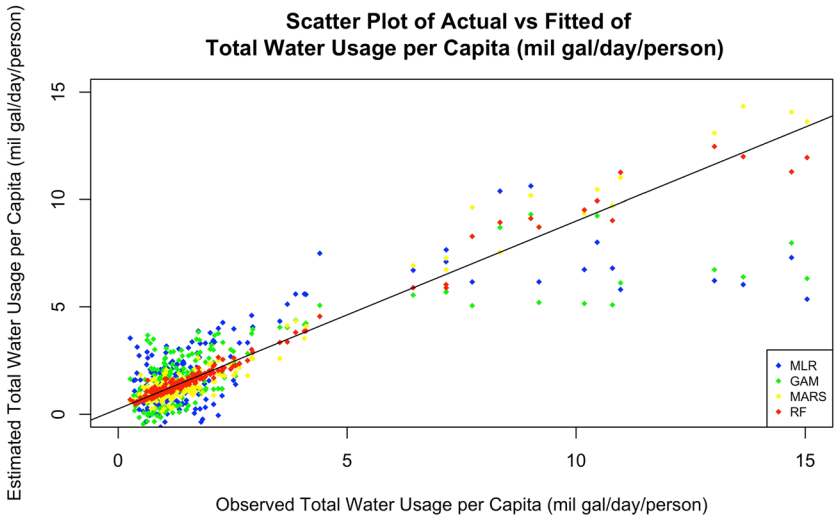

Figure 4. Scatter plot of observed versus estimated values of per capita water use (in million gallons per day) using data of 1995–2010; color-coded on the basis of the statistical models used, namely, Generalized Linear Model (GLM), Generalized Additive Model (GAM), Multivariate Adaptive Regression Splines (MARS), and Random Forest (RF).

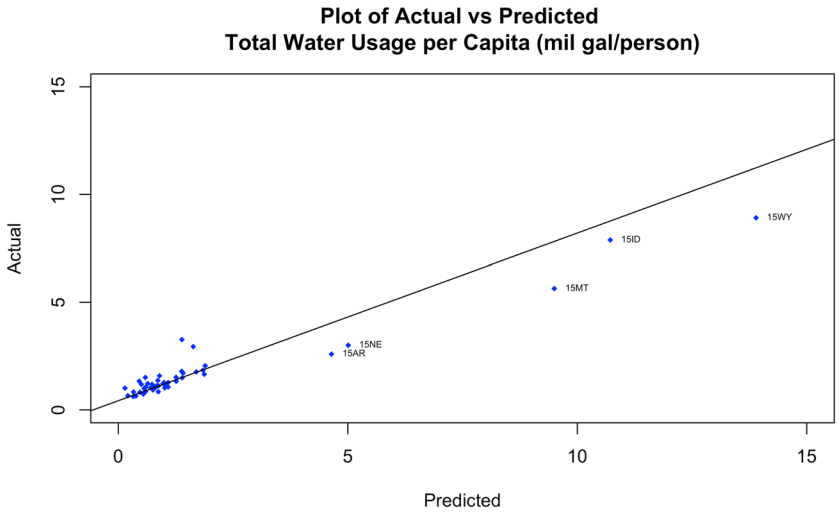

Figure 5. Scatter plot of observed versus predicted values of per capita water use (in million gallons per day) for the period of 2011–2015. States with higher per capita water use identified by year and standard state abbreviation; e.g., 15AR = Arkansas, 2015 data.

thermoelectric energy generation, and state area. This result is intuitive, since irrigation and mining generally comprise a large share of water use in the United States.

In order to understand the association between the topmost important predictors and our response variable (per capita water use), PDPs were examined. Below, we will discuss the partial dependencies for each of the predictors, in order of their importance ranking depicted in Figure 6.

5.1. Effect of Percentage of Irrigated Farmland Areas

The partial dependence between the percentage of irrigated farmland and per capita water use indicates a positive association, with larger irrigated farmlands being associated with higher water use intensity. This is intuitive, as the U.S. agricultural sector accounts for a significant fraction of total water use. Some of the states associated with the different percentiles of water use have been highlighted in Figure 7. As expected, states such as Nebraska and Arkansas lie at the extreme right end of the graph due to their large irrigated agricultural lands. Nebraska (NE) is ranked first in the United States in terms of total irrigated acres

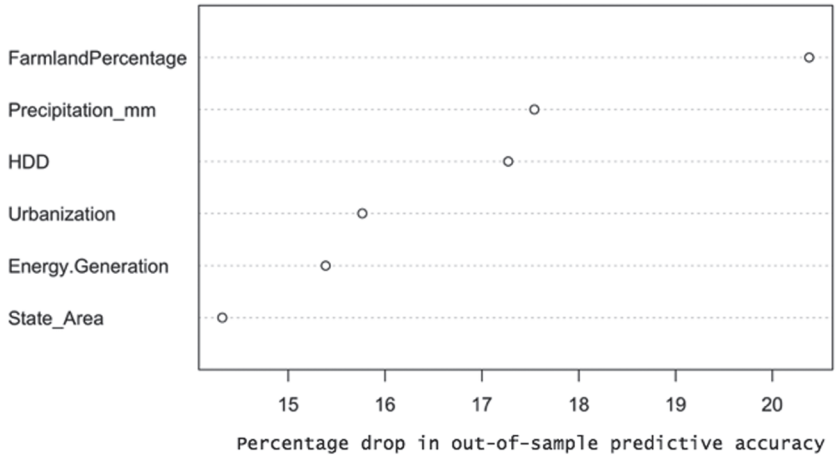

Figure 6. List of the most important predictors identified for the per capita water use predictions, presented here as the percentage drop in the predictive accuracy for the out-of-sample data sets. In other words, the percentages indicate how much the out-of-sample accuracy is decreased if the corresponding variable is removed. The selected predictors are ranked from the most to least influential ones (top to bottom).

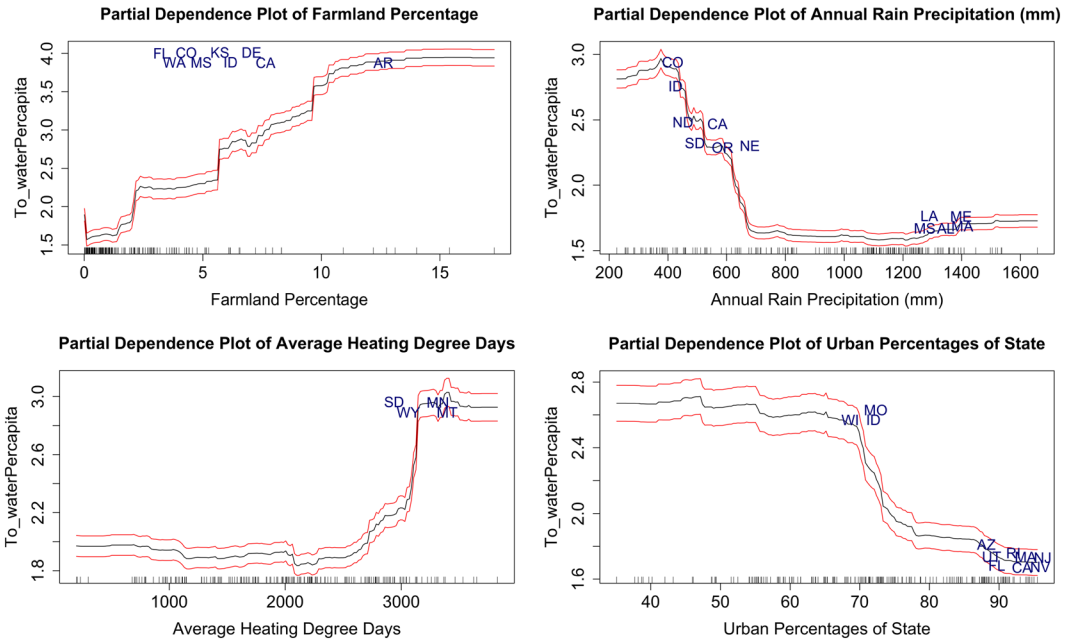

Figure 7. Partial dependency plot (PDP) for the fraction of irrigated farmland, annual precipitation, average heating degree days, and percentage of urban areas, depicting their sensitivity on the per capita water use (in million gallons per-day). The two letters on the plot corresponds to the states, the black line the mean values, and the red lines the 95% confidence intervals. States are identified by the standard state abbreviation; e.g., TX = Texas, FL = Florida, CO = Colorado, KS = Kansas, ND/SD = North/South Dakota.

of land. It has witnessed rapid expansions of irrigated farmlands in recent years, making heavy use of ground water for farming and irrigation. Arkansas (AR)—the number one producer of rice in the United States—also lies at the extreme right end of the plot, which is not surprising since rice is among the most water-intensive crops (Johnson et al., 2011). As mentioned earlier in the paper, however, we make no distinction between consumptive and nonconsumptive water use. It should be noted that a considerable fraction of agricultural water use has return flows, particularly in states such as AR that still harness flood and furrow irrigation. An assessment of what fraction of irrigation is consumptive versus nonconsumptive (i.e., return flows) requires estimates of the volume of water that is withdrawn for irrigation and the fraction of this water that evaporates. In this way, the consumptive fraction of irrigation can be characterized. However, these data are not available in most river basins. (Simons et al., 2020) recently developed a method for estimating the consumed fraction of irrigation water by combining the Budyko method and remote sensing data. However, this type of analysis is beyond the scope of this paper and therefore is a source of uncertainty in our analysis. It is interesting to also note from Figure 7 the step-function jump from states such as Delaware (DE) to California (CA). This may suggest that the crops grown in Delaware (DE) that are mostly corn, soybeans, and wheat based may be less water intensive than the crops grown in California (CA) mainly as nuts and fruits.

5.2. Effect of Precipitation Variability

We hypothesized higher precipitation levels to be associated with decreased water use since precipitation affects a variety of sectors such as thermoelectric power generation, irrigation, public supply, industry, aquaculture, domestic, and life stock. The observed pattern in Figure 7 is consistent with our initial hypothesis, indicating that wetter regions use less water. However, the relationship between precipitation and decreased water use plateaus at about 700 mm of precipitation.

5.3. Effect of HDDs

HDDs measure the difference between average air temperature and an arbitrarily chosen standard baseline temperature (typically 65°F in the United States) to which the built environment would be heated on cold days. Annual HDD measures the time-integrated variation over a year between the average daily

temperature and the baseline "comfort" temperature. Interestingly, there seems to be a subtle, positive association between HDDs and water use, with a sudden jump past 3,000 HDD, which is mostly associated with the states located in the North-Central parts of the United States, such as North Dakota (ND), Minnesota (MN), Wyoming (WY), and Montana (MT) (Figure 7). This pattern may be attributable to the (noncoal) mining and industrial activities such as fracking in these northern states. For instance, in 2005, Minnesota (MN) had the largest share of (sulfide) mining-related fresh water uses in the US. Wyoming (WY) and Montana (MO) also have an active mining sector. Moreover, a significant amount of water is used in North Dakota (ND) in hydraulic fracturing for oil and gas. Due to data limitations and the rapid shifts in these mining and fracking activities, we were not able to test these hypotheses. Nevertheless, there has been indication of generally increased water use intensity in recent years due to intensification of hydraulic fracturing processes across many U.S. shale basins (Kondash et al., 2018).

5.4. Effect of Percentage of the Urbanized Areas

The PDP of the effects of urbanization on water use across the United States clearly shows that the more urbanized states tend to be less water-intensive (Figure 7). This is largely due to the fact that the domestic sector and public supply sector comprise a significantly smaller fraction of total water use as compared to the farmland or energy generation sectors.

6. Conclusions

In this paper, we analyzed the predictive accuracy of various statistical methods in predicting the state-level, per capita water use across the entire United States. The predictive model based on the method of RF was selected as the best model, since it outperformed all other statistical models in terms of both goodness of fit and out-of-sample predictive accuracy.

Our results identified irrigated farming—especially in the states such as Nebraska and Arkansas—and coal mining—especially in states such as Wyoming, West Virginia, and Kentucky—as the most water-intensive anthropogenic activities. Even though mining withdrawals constitute a small fraction of the overall water use in the United States, its share has increased by 40% since 2005 (Maupin et al., 2014).

The water intensity of thermoelectric generation was less than initially hypothesized. According to the USGS, the reduced water use for thermoelectric power generation over the years can be attributed to a reduction in coal-based generation and increased use of natural gas, as well as the newer power plants being equipped with more water-efficient cooling technologies. The USGS also reports declined industrial water use due to higher efficiencies in industrial activities and an emerging emphasis on water reuse and recycling in industrial processes (Maupin et al., 2014).

Climatic conditions such as precipitation and HDDs were also found to be important predictors of per capita water use. Drier conditions (i.e., total annual precipitation less than 600) were intuitively found to be associated with higher water use. However, counterintuitively, we found colder conditions, that is, HDD > 3,000 —which is mostly observed in the North-Central parts of the United States, such as North Dakota, Minnesota, Wyoming, and Montana—to be associated with higher water use. This higher water use might be attributed to hydraulic fracturing for oil and gas and other mining activities beyond coal mining in these states. However, due to data limitations, we were unable to test this hypothesis. While the total per capita water use is lower in more urbanized states, the water use in the public supply is positively associated with urbanization.

In summary, here we present a rigorously validated data-driven framework for characterizing the nationwide water use patterns. Using this framework, we identify and illustrate the key controls of underlying variables affecting the water use rates observed across the United States. The developed data-driven framework has an advantage of being computationally efficient and can compliment the predictions based on mechanistic more detailed hydrologic models. Furthermore, the developed framework can be used as a sensitivity toolbox to analyze the impacts of changing climate and other socioeconomic conditions. To illustrate this utility, we provide one such example analysis in the Supporting Information S1.

Finally, while our study provides a roadmap for efficient and detailed analysis of water usage data to complement mechanistic models, we make no distinction between the consumptive and nonconsumptive use of water. Future studies are encouraged to look further into separating the key factors that influence

consumptive versus nonconsumptive water usage. It is also worth highlighting that the USGS data collection has not remained consistent over time. Specifically, water usage data associated with hydropower and wastewater treatment are not included in the USGS estimates from 2000 onward (USGS, 2017). Given that the main goal in this paper is to provide an efficient framework for large-scale water usage analysis, we have not delved into the implications of the changes in the USGS reporting over time. However, future studies are needed to uncover the implications of such data inconsistencies. Another area for future research is assessing the sensitivity of the results to the choice of "per capita" normalization, as this study did not explore normalizing water use by other factors such as land area.

Acronyms

BAU	business as usual
CDD	cooling degree day (°F)
CPC	Climate Prediction Center
EIA	Energy Information Association
EPA	Environmental Protection Agency
EPRI	Electric Power Research Institute
GAM	generalized additive model
GCM	global circulation model
GDP	gross domestic product
GLM	generalized linear model
GSP	gross state product (millions of USD measured in 2009 real dollars)
HDD	heating degree day (°F)
NEMS	National Energy Modeling Systems
NOAA	National Oceanic and Atmospheric Administration
MARS	multivariate adaptive regression splines
PDP	partial dependence plot
RF	random forest
SPI	standardized prediction index
US	United States
USD	United States Dollar ($)
USGS	United States Geological Survey

Data Availability Statement

All data sets used in this study are collected from publicly available sources and detail references are provided in section 3 and Table 1. Source code and corresponding data sets used in this study can be obtained from https://engineering.purdue.edu/LASCI/research-data/water/index_html.

Acknowledgments
Funding for this project was provided by NSF grant #1826161 & #1832688, and Purdue University C4E Seed Grant. The authors would also like to acknowledge Purdue Climate Change and Research Center (PCCRC) for their initiative in disseminating the research results. We would like to thank the Editor Jim Hall, the Associate Editor, three anonymous reviewers, and Pierre Glynn for their constructive comments, which improved the quality of the manuscript.

References

Alcamo, J., Döll, P., Henrichs, T., Kaspar, F., Lehner, B., Rösch, T., & Siebert, S. (2003). Development and testing of the watergap 2 global model of water use and availability. *Hydrological Sciences Journal*, *48*(3), 317–337.

BEA (2017). BEA regional economic accounts. Retrieved from Bureau of Economic Analysis: Available at: http://www.bea.gov/regional/; Last accessed on 04/04/2017.

Bauer, D., Philbrick, M., Vallario, B., Battey, H., Clement, Z., Fields, F., et al. (2014). The water-energy nexus: Challenges and opportunities. US Department of Energy 2014.

Billings, B., & Jones, C. (2008). *Forecasting urban water demand* (2nd ed.). Denver, CO: American Waterworks Association.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. https://doi.org/10.1023/A:1010933404324

Bruss, C. B., Nateghi, R., & Zaitchik, B. F. (2019). Explaining national trends in terrestrial water storage. *Frontiers in Environmental Science*, *7*, 85.

CSO (2017). Coastal States Organization. Available http://www.coastalstates.org/; Last accessed on 04/04/2017.

Calder, I. (2012). *Blue revolution: Integrated land and water resources management* (2nd ed.). London: Routledge. https://doi.org/10.4324/9781849770613

Chandel, M. K., Pratson, L. F., & Jackson, R. B. (2011). The potential impacts of climate-change policy on freshwater use in thermoelectric power generation. *Energy Policy*, *39*(10), 6234–6242.

Chaussee, J. (2014). California drought expected to cost state 2.2 billion dollars in losses. *Scientific American*. Retrieved from https://www.scientificamerican.com/article/california-drought-expected-to-cost-state-2-2-billion-in-losses/

Cui, R. Y., Calvin, K., Clarke, L., Hejazi, M., Kim, S., Kyle, P., et al. (2018). Regional responses to future, demand-driven water scarcity. *Environmental Research Letters*, *13*(9), 94006.

Davies, E. G. R., Kyle, P., & Edmonds, J. A. (2013). An integrated assessment of global and regional water demands for electricity generation to 2095. *Advances in Water Resources*, *52*, 296–313.
Devineni, N., Lall, U., Etienne, E., Shi, D., & Xi, C. (2015). America's water risk: Current demand and climate variability. *Geophysical Research Letters*, *42*, 2285–2293. https://doi.org/10.1002/2015GL063487
Donkor, E. A., Mazzuchi, T. A., Soyer, R., & Roberson, J. A. (2014). Urban water demand forecasting: Review of methods and models. *Journal of Water Resources Planning and Management*, *140*(2), 146–159.
EIA (2017). U.S. Energy Information Administration (EIA). Detailed State Data. Retrieved from: https://www.eia.gov/electricity/data/state/; Last accessed on 04/04/2017.
EPA (2017). WaterSense EPA. Water Use Today. Retrieved from U.S. Environmental Protection Agency: https://www3.epa.gov/watersense/our_water/water_use_today.html; Last accessed on 04/04/2017.
Fan, Y., & van den Dool, H. (2004). Climate Prediction Center global monthly soil moisture data set at 0.5°C resolution for 1948 to present. *Journal of Geophysical Research*, *109*, D10102. https://doi.org/10.1029/2003JD004345
Flörke, M., Kynast, E., Bärlund, I., Eisner, S., Wimmer, F., & Alcamo, J. (2013). Domestic and industrial water uses of the past 60 years as a mirror of socio-economic development: A global simulation study. *Global Environmental Change*, *23*(1), 144–156.
Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, *19*(1), 1–67.
Genuer, R., Poggi, J.-M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, *31*(14), 2225–2236.
Giordano, M., & Shah, T. (2014). From IWRM back to integrated water resources management. *International Journal of Water Resources Development*, *30*(3), 364–376.
Hall, M. J., Postle, S. M., & Hooper, B. D. (1989). A data management system for demand forecasting. *International Journal of Water Resources Development*, *5*(1), 3–10.
Hamoda, M. F. (1983). Impacts of socio-economic development on residential water demand. *International Journal of Water Resources Development*, *1*(1), 77–84.
Hanasaki, N., Kanae, S., Oki, T., Masuda, K., Motoya, K., Shirakawa, N., et al. (2008a). An integrated model for the assessment of global water resources—Part 1: Model description and input meteorological forcing. *Hydrology and Earth System Sciences*, *12*(4), 1007–1025.
Hanasaki, N., Kanae, S., Oki, T., Masuda, K., Motoya, K., Shirakawa, N., et al. (2008b). An integrated model for the assessment of global water resources—Part 2: Applications and assessments. *Hydrology and Earth System Sciences*, *12*(4), 1027–1037.
Hastie, T., Tibshirani, R., & Friedman, J. (2009). *Overview of supervised learning*. New York: Springer.
Hayes, M., Svoboda, M., Wall, N., & Widhalm, M. (2010). The Lincoln Declaration on drought indices: Universal meteorological drought index recommended. *Bulletin of the American Meteorological Society*, *92*(4), 485–488.
James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). New York: Springer.
Johnson, B., Christopher, T., Anil, G., & NewKirk, S. V. (2011). Nebraska irrigation fact sheet (*Rep. no. 190*): Department of Agricultural Economics, University of Nebraska Lincoln.
Jorgensen, B., Graymore, M., & O'Toole, K. (2009). Household water use behavior: An integrated model. *Journal of Environmental Management*, *91*(1), 227–236.
Kondash, A. J., Lauer, N. E., & Vengosh, A. (2018). The intensification of the water footprint of hydraulic fracturing. *Science Advances*, *4*(8), eaav2110.
Liaw, A., & Wiener, M. (2002). Classification and regression by random forest. *R news*, *2*(3), 18–22.
Loucks, D. P., & Van Beek, E. (2017). *Water resource systems planning and management: An introduction to methods, models, and applications*. Springer. https://link.springer.com/book/10.1007/978-3-319-44234-1
Lutz, J. D., Liu, X., McMahon, J. E., Dunham, C., Shown, L. J., & McCure, Q. T. (1996). *Modeling patterns of hot water use in households*. Berkeley, CA: Office of Scientific and Technical Information (OSTI). Lawrence Berkeley National Laboratory.
Maupin, M. A., Kenny, J., Hutson, S. S., Lovelace, J. K., Barber, N. L., & Linsey, K. S. (2014). *Estimated use of water in the United States in 2010*. Reston, VA: US Geological Survey.
McKee, T. B., Doesken, N. J., & Kleist, J. (1993). The relationship of drought frequency and duration to time scales. In *Eighth Conference on Applied Climatology*, Anaheim, California. 17-22 January 1993.
Mukherjee, S., Nateghi, R., & Hastak, M. (2018). A multi-hazard approach to assess severe weather-induced major power outage risks in the us. *Reliability Engineering & System Safety*, *175*, 283–305.
Mukhopadhyay, S., & Nateghi, R. (2017). Estimating climate-demand nexus to support longterm adequacy planning in the energy sector, *2017 IEEE Power & Energy Society General Meeting* (pp. 1–5). Chicago, IL, USA: IEEE.
NOAA (2017a). National Oceanic and Atmospheric Administration (NOAA), Degree Days Statistics National Weather Service; Center for Weather and Climate Prediction. Retrieved from: http://www.cpc.ncep.noaa.gov/products/analysis_monitoring/cdus/degree_days/; Last accessed on 04/04/2017.
NOAA (2017b). NOAA National Centers for Environmental Information, Local Climatological Data (LCD), Last accessed 04/04/2017.
Nateghi, R. (2018). Multi-dimensional infrastructure resilience modeling: An application to hurricane-prone electric power distribution systems. *IEEE Access*, *6*, 13,478–13,489.
Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, *135*(3), 370–384.
Obringer, R., Kumar, R., & Nateghi, R. (2019). Analyzing the climate sensitivity of the coupled water-electricity demand nexus in the Midwestern United States. *Applied Energy*, *252*, 1–1.
Obringer, R., Kumar, R., & Nateghi, R. (2020). Managing the water–electricity demand nexus in a warming climate. *Climatic Change*, *159*, 233–252.
Obringer, R., & Nateghi, R. (2018). Predicting urban reservoir levels using statistical learning techniques. *Scientific Reports*, *8*(1), 5164.
Olmstead, S. M. (2014). Climate change adaptation and water resource management: A review of the literature. *Energy Economics*, *46*, 500–509.
Pokhrel, Y. N., Hanasaki, N., Wada, Y., & Kim, H. (2016). Recent progresses in incorporating human land–water management into global land surface models toward their integration into earth system models. *Wiley Interdisciplinary Reviews: Water*, *3*(4), 548–574.
Rahaman, M. M., & Varis, O. (2005). Integrated water resources management: Evolution, prospects and future challenges. *Sustainability: Science, Practice and Policy*, *1*(1), 15–21.
Sebri, M. (2016). Forecasting urban water demand: A meta-regression analysis. *Journal of Environmental Management*, *183*, 777–785.
Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310.

Simons, G. W. H., Bastiaanssen, W. G. M., Cheema, M. J. M., Ahmad, B., & Immerzeel, W. W. (2020). A novel method to quantify consumed fractions and non-consumptive use of irrigation water: Application to the Indus Basin Irrigation System of Pakistan. *Agricultural Water Management*, *236*, 106,174.

Sovacool, B. K., & Sovacool, K. E. (2009). Identifying future electricity–water tradeoffs in the United States. *Energy Policy*, *37*(7), 2763–2773.

Sutanudjaja, E. H., Beek, R., Wanders, N., Wada, Y., Bosmans, J. H. C., Drost, N., et al. (2018). PCR-GLOBWB 2: A 5 arcmin global hydrological and water resources model. *Geoscientific Model Development*, *11*(6), 2429–2453. https://doi.org/10.5194/gmd-11-2429-2018

USCB (2017). U.S. Census Bureau, Historical Income Tables: Households. Retrieved from U.S. Census Bureau: https://www.census.gov/hhes/www/income/data/historical/household/USGS. (1995-2010).

USDA (2007). Farm and Ranch Irrigation Survey. Retrieved from USDA, Census of Agriculture. Available from: https://www.agcensus.usda.gov/Publications/2007/Online_Highlights/Farm_and_Ranch_Irrigation_Survey/fris08.pdf; Last accessed on 04/04/2017.

USGS (2017). Water-use data available from USGS. Retrieved from U.S. Geological Survey:http://water.usgs.gov/watuse/data/; Last accessed on 04/04/2017.

Wada, Y., Bierkens, M. F. P., Roo, A., Dirmeyer, P. A., Famiglietti, J. S., Hanasaki, N., et al. (2017). Human–water interface in hydrological modelling: Current status and future directions. *Hydrology and Earth System Sciences*, *21*(8), 4169–4193.

Wada, Y., Wisser, D., & Bierkens, M. F. P. (2014). Global modeling of withdrawal, allocation and consumptive use of surface water and groundwater resources. *Earth System Dynamics*, *5*(1), 15.

Worland, S. C., Steinschneider, S., & Hornberger, G. M. (2018). Drivers of variability in public-supply water use across the contiguous United States. *Water Resources Research*, *54*, 1868–1889. https://doi.org/10.1002/2017WR021268