Leveraging statistical learning theory to characterize the U.S. water consumption

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

Key Points:

- Statistical inference of the nation wide water withdrawal patterns in the U.S.
- Irrigated farming, thermoelectric energy generation and urbanization are the most water-intensive anthropogenic activities
- Water withdrawal patterns across U.S. show varying sensitivity (between ±10%) to future changes in precipitation changes under the RCP8.5 scenario.

Corresponding author: Rohini Kumar; Roshanak Nateghi, `rohini.kumar@ufz.de;rnateghi@purdue.edu`

Abstract

Access to accurate estimates of water withdrawal is requisite for urban planners as well as operators of critical infrastructure systems to make optimal operational decisions and investment plans to ensure reliable and affordable provisioning of water. Furthermore, identifying the key predictors of water withdrawal is important to regulators for promoting sustainable development policies to reduce water use. In this paper, we developed a rigorously evaluated predictive model, using statistical learning theory, to estimate state-level, per-capita water withdrawal as a function of various geographic, climatic and socio-economic variables. We then harnessed the data-driven predictive model to identify the key factors associated with high water-usage intensity among different sectors in the U.S. We analyzed the predictive accuracy of a range of parametric models (e.g., generalized linear models) and non-parametric, flexible learning algorithms (e.g., generalized additive models, multivariate adaptive regression splines and random forest). Our results identified irrigated farming, thermo-electric energy generation and urbanization as the most water-intensive anthropogenic activities, on a per-capita basis. Among the climate factors, precipitation was also found to be a key predictor of per-capita water withdrawal, with drier conditions associated with higher water withdrawals. Results of the first-order sensitivity analysis indicated changes between ±10% in the future water withdrawal across the U.S., in response to precipitation changes, by the end of the 21st Century under the business-as-usual scenario. Overall, our study highlights the utility of leveraging statistical learning theory in developing data-driven models that can yield valuable insights related to the water withdrawal patterns across expansive geographical areas.

1 Introduction

Integrated water resource management has been receiving increasing attention globally (Giordano & Shah, 2014; 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. Sustainable management of demand for water has been brought into the limelight in the United States following several devastating, multi-year 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. According to the U.S. Environmental Protection Agency, 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).

Accurate estimates 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. 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 spatio-temporal variability in demand for water (Billings & Jones, 2008). According to Hall, Postle, and Hooper (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 leverage statistical learning theory to: a) develop accurate predictive models for per-capita water use in various sectors in the U.S., b) identify the key predictors of state-level, per-capita water withdrawal, c) understand the relationship between each of the key predictors and per-capita water use, and d) analyze the sensitivity of the water withdrawal patterns to changes in climate variability (e..g, precipitation changes) under changing climate conditions. Our predictive water withdrawal models were developed using state-level, per-capita water withdrawal data over the past two decades

– together with various geographic, climatic, and socio-economic factors – to identify the key factors that are associated with high water-usage intensity among different sectors in the U.S.

We hypothesized that statistical models that assume 'rigid' functional forms – such as linearity and additivity (e.g., multiple linear regression) – would not adequately capture the complex dependencies between state-level water withdrawals and socio-economic and geoclimatic conditions; and that more robust statistical learning algorithms (e.g., ensemble-of-trees), would be more effective in predicting state-level, water withdrawals. Moreover, given that the largest fraction of water-withdrawals occur in the agricultural and thermoelectric generation sectors, we hypothesize irrigated farming and power generation to be the key predictors of state-level water withdrawals.

The structure of this paper is as follows. The review of the existing literature in predicting water withdrawal 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 demand – with various different spatio-temporal scales and lead time-horizons – using a range of methods such as simulation, econometrics and statistical learning theory. Donkor, Mazzuchi, Soyer, and Roberson (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. More recently, 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 socio-economic 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 together with its continually improving standards of living were the primary factors contributing to high water consumption rates in the country.

In an another study by Lutz et al. (1996) leveraged a variation of the EPRI (Electric Power Research Institute) 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, Graymore, and O'Toole (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 energy-water nexus in the U.S., and concluded that twenty-two counties will likely face sever water shortages, brought about primarily due to increased capacity expansion in thermoelectric generation. Chandel, Pratson, and Jackson (2011) leveraged a modified version of the U.S. National Energy Modeling Systems (NEMS) together with thermoelectric water-use factors from the EIA to investigate the impact of various climate change policy 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 withdrawal for power generation, compared to the business as usual scenarios. Moreover, they found that water-use decreased as the policy's carbon price increased. Davies, Kyle, and Edmonds (2013) leveraged GCAM – an integrated assessment modeling of energy, agriculture, and climate change – to assess the water intensity associated with electricity generation until 2095. They found that water use would likely decrease with capital stock turnover.

The majority of the empirical studies to date have focused primarily on either a particular geographical location, or a given sector in the U.S., and leveraged either linear models (the assumptions of which 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 withdrawal data – available from USGS over the past two decades for the entire U.S. – and develop an accurate and interpretable predictive water withdrawal model as a function of socio-economic, geographic, climatic conditions.

It is noteworthy that, though not pursued in this study, there exist another fundamentally different approach to modeling water withdrawal, based on complex, mechanistic hydrologic models with integrated elements of human-water interfaces (e.g., Pokhrel, Hanasaki, Wada, & Kim, 2016; Wada et al., 2017). Models in this category include, for instance, PCR-GLOBWB (Sutanudjaja et al., 2018; Wada, Wisser, & Bierkens, 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 datasets. Different sorts of simplifications and conceptualizations are therefore necessary to model the complex interactions between human and natural systems (e.g., Wada et al., 2017). Our proposed modeling paradigm – based on statistical learning theory – can be complementary to hydrological modeling efforts. Our approach offers key advantages of a) being computationally efficient, and b) requiring a limited set of predictors to re-construct the continuous space-time evolution of water withdrawal; which can the be used to further constrain the parameterization of more complex, mechanistic hydrologic models. In summary, our approach can help identify the most water-intensive sectors across various states, inform policy makers, regulators and researchers on the exiting U.S. water use patterns and identify sectors and areas where efficiency and conservation mechanisms could yield maximum return, in-terms of enhanced sustainability of our urban ecology.

3 Data and Initial Analysis

Data were collected from various publicly available sources such as the Geological Survey website (USGS, 2017), the Energy Information Administration (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 (NOAA, 2017), the U.S. Department of Agriculture (USDA, 2007), the Coastal States Organization (CSO, 2017), the U.S. Environmental Protection Agency (EPA, 2017) and other sources (IOWA, 2017). Below, we will provide a brief description of our response variable (i.e., per-capita water usage) and various socio-economic, hydro-climatic and geographic predictors that were used in our analyses. It should be pointed out that since the water withdrawal data is only available at five-year increments, the predictors were processed to match the temporal scale of our response variable.

3.1 Response Variable: Per-Capita, State-Level Water Withdrawal

State-level water withdrawal data (in million gallons per day) were selected as our response variable, and were obtained from U.S. Geological Survey website (USGS) for the period of 1991-2010. USGS water usage data are collected and compiled every five years for each of the 50 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 (depicted in Fig. 1) such as thermoelectric, irrigation, public supply, industry, aquaculture, domestic, livestock and mining. Thermoelectric and irrigation are the two dominant sectors that account for almost two-third of the total water withdrawal across the U.S. We, however, note that there is a large regional variability in water withdrawal patterns – the States in the east is more dominated by the thermoelectric and industrial water sectors, while the irrigation is the main water usages in the central and western part of the U.S. To control for the varying sizes of states, we normalized the state-wide total water withdrawal data by the total population of each state. The distribution of state-wise, normalized water withdrawal for years of 2006–2010 can be seen in Fig. 1(bottom panel). States highlighted in shades of red represents high per-capita water usage, while the states in blue represent low per-capita water usage. Fig. 1(bottom panel) reveals that Idaho has the highest per-capita water usage for the year 2006–2010.

The distribution of the per-capita water withdrawal (in million gallons per day) for the period 1991-2010 is depicted in Fig. 2. The distribution of per-capita water withdrawal 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 Fig. 2a). Power-law distributions describe phenomena where large events are quite rare, but small events are very frequent. Fig. 2 suggests that a small fraction of the states in the U.S. tend to consume disproportionately large volumes of water per capita.

3.2 Socio-Economic Predictors

Gross State Product (GSP) data were collected from the U.S. Bureau of Economic Analysis for the years of 19912010 in current value. The GSP data (in millions of USD) were then converted to time value of 2010, using the GDP deflator. Household Median Income (in USD) was collected from the Bureau of Labor Statistics. The value of income data was converted to 2013 CPI-U-RS (Consumer Price Index Research Series Using Current Methods) USD.

The education level data obtained from the U.S. Census Bureau contains 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 four years of college or higher. We leveraged generalized additive models to impute the missing data and align the temporal scale of the education data with that of water withdrawal. The premise for including this variable in the analysis is to test whether educational levels are predictive of the public supply water withdrawal.

Datasets related to thermoelectric energy generation – e.g., coal, petroleum, and gas fired plants, nuclear and geothermal technologies – in mega watt-hours were collected from the Energy Information Administration (EIA). Coal production, available from the EIA, was used as a proxy for mining industry, since coal is the biggest profit generating mining production in the U.S. The percentage of urban population data were collected from the U.S. Census. Since the temporal scale of the urban population data were decadal, the years did not match the years in the USGS water dataset. We therefore imputed the missing years of the percentage of urban population data a using generalized additive model to match the years across the two datasets.

3.3 Hydro-climatic and Geographic Predictors

Time-series of datasets related to Cooling Degree Days (CDD) and Heating Degree Days (HDD) are based on variation in air temperature estimates which were made available from Climate Prediction Center (CPC) and National Weather Service (NWS). Other hydro-climatic variables as predictor variables include Standardized Precipitation Index (SPI), soil moisture, and annual precipitation data were provided by the National

Centers for Environmental Information. The SPI characterizes the inter/intra-annual variability of precipitation with positive values indicating wetter than normal conditions and the negative values being indicative of drier than normal conditions(Hayes, Svoboda, Wall, & Widhalm, 2010; McKee, Doesken, & Kleist, 1993). Additionally, we used the upper 1 m simulated soil-water content (mm) based on the CPC model based simualtions 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 is in the borders of (a) the Atlantic Ocean, (b) the Pacific Ocean, (c) the Gulf of Mexico, and (d) the Great Lakes. The states in proximity of any of the above-mentioned water-sheds, 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 U.S. Department of Agriculture (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 and 1995, and 1997 data was used to represent the value between 1996–2000. 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 set-up, all predictor variables were aggregated spatially and temporally to match the state-wide, five-yearly available water withdrawal datasets.

3.4 Exploratory Data Visualization and Analysis

A 'biplot' is a useful visualization tool for multivariate data. One of the most commonly used types of a biplot is based on principle component analysis. A PCA-biplot is a low-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 Fig. 3 that over the years of 1995–2010, the state-level water usage did not change significantly. For example, on the bottom left corner of the plot, we observe that water usage of Arizona, Louisiana, Texas, and Florida are located close to each other across the different years. The energy generation and cooling-degree-days (CDD) vectors extended in the direction of Texas suggest that the state's thermoelectric power generation and its hot climate can help explain the variance of water usage in Texas, as opposed to states of Colorado or North Dakota which lie close to the heating-degree-day (HDD) vector. Moreover, the Fig. 3 reveals that while water usage in the densely populated states of the Northeast can be explained by socio-economic factors such as income and education and measures of urbanization, the water usage in the larger Midwestern and Western states of North and South Dakota, Nebraska, Iowa and New Mexico tend to be dominated by farming and mining practices.

4 Methodology

The existing empirical literature in field of water analysis has almost exclusively focused on descriptive and explanatory statistical modeling, while predictive modeling of water analysis has largely been under-explored. Unlike descriptive or explanatory modeling which is concerned with best explaining the past variability in the data, predictive modeling is concerned with predicting 'new/unseen' data. The expected prediction error (EPE) for a new observation x can be summarized by the equation below [11]:

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

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.* The central thesis in this paper is that, with the recent accelerated pace of large complex datasets becoming available, predictive modeling can be leveraged as a powerful tool to identify complex and non-linear 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 algorithm that was used to develop the final best predictive model of the state-level, water withdrawal data.

4.1 Supervised Learning Theory (Predictive Modeling)

Supervised learning theory was leveraged to develop accurate predictive models for state-level water withdrawals, and identify their most important predictors of in the U.S. The main objective of supervised learning is to approximate a process of interest (e.g., water withdrawals) as a function of various independent predictors (e.g., geographic, climatic and socio-economic 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 approximate the response $hatf(X)$ (i.e., water withdrawal) such that the loss function L is minimized over the entire domain of the input data space:

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

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, Tibshirani, & Friedman, 2009).

We trained our data with various parametric (e.g., generalized linear models) and non-parametric (e.g., generalized additive models (GAM), multivariate adaptive regression splines (MARS) and random forests (RF)) methods – description of which can be found in the Appendix. Given that the ensemble tree-based algorithm (the method of random forest) outperformed all other algorithms in terms of out-of-sample predictive accuracy (see Section 5), we selected it as our final best model. A brief description of the random forest (RF) algorithm is provided below.

4.2 Random Forests (RF)

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

$$F(x) = \frac{1}{m_{\text{tree}}} \sum_{i=1}^{m_{\text{tree}}} T_i(x) \tag{3}$$

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 socio-economic factors used in this analysis). The subset of predictors for building each decision tree is randomly selected, and best splits values are chosen such that the sum of squared errors (or least absolute deviation) within each node t within 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 (or mode value as for the case of classification) is then assigned to the terminal nodes. $F(x)$ estimates the response value, by aggregating m such decision trees.

Regression trees are low in bias, particularly if they are grown sufficiently deep, since the tree structure follows the structure of the data well so that the estimated target mean is close to the true mean (Hastie et al., 2009). They are, however, notoriously noisy, and generally have high variance. They are unstable and not particularly robust to outliers, and this makes the procedure non-ideal for datasets that contain many outliers. The issue of high variance is solved by leveraging the ensemble methodology as a variance reduction technique. The ensemble-of-trees methods such as random forest are generally very robust to outliers and offer strong predictive power. The estimation of prediction error of random forest can be obtained by leveraging the out-of-bag (OOB) data (i.e., the test data that was set aside during the development of each tree and not used in building that tree) to compute the mean square error as below:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \overline{y}_i^{'})^2 \tag{4}$$

where $\overline{y}_i^{'}$ is the average OOB predictions data for the i^{th} observation (Liaw & Wiener, 2002). Since the method of random forest is non-parametric, 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):

$$(\hat{f}_J)(x_j) = \frac{1}{n} \sum_{i=1}^{n} (\hat{f}_J)(x_j, x_{-j,i}) \tag{5}$$

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 dataset); 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.

5 Results and Discussion

Table 1 summarizes the performance of each of the models. The first column summarizes the goodness-of-fit for each of the models. Multivariate adaptive regression splines (MARS) and the method of random forest (RF) fit the data substantially better compared to multiple linear regression (MLR) and generalized additive (GAM) model. 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 significance differences between the mean errors, with a p-value $< 2 \times 10^{16}$.

Fig. 4 (top panel) visualizes the fit of each of the prediction models. The prediction model based on the random forest algorithm substantially outperforms all other models in terms of the goodness-of fit. The model developed using the random forest algorithm was therefore selected as the final best model.

In order to further demonstrate the predictive capability of the model, we trained the random forest algorithm with the data until the end of 2005 in order to predict water withdrawals in an independent testing period of 2006–2010. Table 2 summarizes model fit and predictive accuracy, and Fig. 4 (bottom panel) provides a graphical representation of the predicted and observed values of per-capita water withdrawals. Based on the results summarized in the table and the plot, it can be inferred that RF outperforms all other models. In fact, RF is able to estimate the water usage above 5 million gal/day/person accurately, even though there are less observation points. While MARS performs well below 5 million gal/day/person (where there is more observations) it performs poorly where the data is sparse.

These results confirms our hypothesis that simple linear-based models (e.g., MLR) and additive structures such as GAMs 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 in variance reduction and therefore has an inferior predictive power.

We leveraged a data-driven variable selection, based on an algorithm proposed by Genuer, Poggi, and Tuleau-Malot (2010), to implement input variable reduction for the RF model. The variable selection algorithm first 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 step-wise forward strategy. The smallest subset of input data that yielded the best predictive accuracy were retained for the final model. The list of the final key variables selected for each sector are shown in Fig. 5.

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 withdrawal, followed by total state-level precipitation, heating degree days (HDD), urbanization, thermoelectric energy generation and state-area. This result is intuitive, since irrigation and mining generally comprise a large share of water withdrawal in the U.S.

In order to understand the association between the top most important predictors and our response variable (per-capita water withdrawal), partial dependence plots were examined. Below, we will discuss the partial dependencies for each of the predictors, in order of their importance ranking depicted in Fig. 5.

5.1 Effect of Percentage of Irrigated Farmland Areas

The partial dependence between the percentage of irrigated farmland and per-capita water withdrawal indicates a positive association, with larger irrigated farmlands being associated with higher water withdrawal intensity. This is intuitive, as the U.S. agricultural sector accounts for a significant fraction of total water consumption. Some of the states associated with the different percentiles of water withdrawal have been highlighted in Fig. 6. 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 is ranked first in the U.S. in terms of total irrigated acres of land, and has seen rapid expansions of irrigated farmlands in recent years. It is located on the Ogallala Aquifer which is among the largest in the world, and makes heavy use of ground water for farming and irrigation. In fact, most of the irrigation in Nebraska (and effectively all of the more recent expansion in irrigated farming) is pumped from the High Plains (aka Ogallala) Aquifer. Arkansas, the number one producer of rice in the U.S., also lies at the extreme right end of the table, which is not surprising since rice is among the most water-intensive crops (Johnson, Christopher, Anil, & NewKirk, 2011). It is interesting to note the step-function jump from the states such as Delaware to the state of California. This could suggest that the crops grown in Delaware that are mostly corn, soybeans and wheat-based may be less water intensive than the crops grown in CA (mainly nuts, and fruits).

5.2 Effect of Precipitation Variability

We hypothesized higher precipitation levels to be associated with decreased water usage 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 Fig. 6 is consistent with our initial hypothesis, indicating that wetter regions use less water. However, the decreased water-use plateaus at the threshold of 700 mm of precipitation

5.3 Effect of Heating Degree Days

Heating degree days (HDD) measure the difference between average air temperature and an arbitrarily chosen standard baseline temperature (typically 65°F in the US) 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 heating degree days and water withdrawal, with a sudden jump past HDD of 3000 which is mostly associated with the states located in the North-Central parts of the U.S., such as North Dakota, Minnesota, Wyoming and Montana (Fig. 6). This might be attributable to the (non-coal) mining and industrial activities such as fracking in these northern states. For instance, in 2005, Minnesota had the largest share of (sulfide) mining-related fresh water withdrawals in the U.S. Wyoming and Montana also have an active mining sector. Moreover, a significant amount of water is used in North Dakota in hydraulic fracturing for oil and gas. Unfortunately, data limitation as well as the diversity and rapid shifts in these mining and fracking activities make it difficult to test these hypotheses.

5.4 Effect of Percentage of the Urbanized Areas

The partial dependency plot for the urbanization effects on water withdrawal patterns across U.S. clearly shows that the more urbanized states tend to be less water-intensive (Fig. 6). Again, this is largely due to the fact that the domestic sector and public supply sector comprises a significantly smaller fraction of total water withdrawal as compared to the farmland or energy generation sectors.

5.5 Sensitivity of Water Withdrawal to Future Climate Variability

In this section, We demonstrate the utility of leveraging the predictive model, based on the random forest algorithm, in assessing the sensitivity of changes in water withdrawal patterns across U.S. in response to changing climate conditions. To this end, we used the precipitation datasets from the five CMIP5 Global Circulation Models (GCMs: HadGEM2-ES, IPSL-CM5A-LR, MIROC- ESM-CHEM, GFDL-ESM2 and NorESM1-M), available in a bias-corrected form by the Inter-Sectoral Impact Model Intercomparison Project (ISI-MIP; Warszawski et al., 2014, see also www.isimip.org for more details). For this demonstration purpose, we aggregated the daily precipitation dataset to create state-wide, mean annual estimates for the two time periods indicating the contemporary condition (1995-2010) and the future one (2070-2085), which are taken from the runs corresponding to the RCP8.5 future pathways under the narration of a "business-as-usual" scenario. For these periods, we run the established RF model to predict state-wide water withdrawal using their respective precipitation data-sets while keeping other variables at nominal values following a "ceteris paribus" condition. We estimate the ensemble mean of the state-wise, projected changes in the water withdrawal rates based on the RF model outputs driven by five GCM based precipitation data-sets.

We observed a clear north-south gradient in the relative changes of the water withdrawal patterns across U.S. between future and contemporary period estimates (Fig. 7). Our simulation results indicated increased water withdrawal rates in the southern States, while the declined rates are expected in the Northern states – in response to future precipitation changes. The southern states such as Texas (TX), Florida (FL), Louisiana (LA), and Arizona (AZ) show a projected increase of more than 5% in their water withdrawal rates relative to the contemporary condition. The changes in the future water withdrawal rates across the majority of States is in-between ± 10% with the driving precipitation changes being projected ± 15%. Results of this analysis also indicate a varying level of sensitivity in the projected water withdrawal rates to changes in precipitation estimates (Fig. 7; bottom scatter plot). For example, in states such as Texas (TX) and Arizona (AZ), a small change in mean annual precipitation (around 2%) creates a relatively larger change in water withdrawal (6-8%). Notably, all of the above presented estimates corresponds to ensemble mean of the modeled water withdrawal (based on the RF model run with five GCMs outputs); analysis based on the individual model estimates revealed a substantial uncertainty owing to the differences in projected precipitation from different GCMs.

6 Conclusions

In this paper, we analyzed the predictive accuracy of various statistical methods in predicting the state-level, per-capita water withdrawal across the entire U.S. The predictive model based on the method of random forest was selected as the best model, since it out-performed 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 Kentuky as the most water-intensive anthropogenic activities. Even though mining withdrawals constitute a small fraction of the overall water use in the U.S., 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 withdrawals for thermoelectric power generation over the years can be attributed to a reduction in coal consumption 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 withdrawals 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 heating-degree days were also found to be important predictors of per-capita water withdrawal. Drier conditions (i.e., total annual precipitation less than 600) were intuitively found to be associated with higher water withdrawals. However, counter-intuitively, we found colder conditions i.e., HDD > 3000 which is mostly observed in the North-Central parts of the U.S., 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. While the total, per-capita water withdrawals are lower in more urbanized states, the water withdrawal in the public supply is positively associated with urbanization.

Using the developed predictive model, we were able to infer the first-order sensitivity of the projected changes in the water withdrawal to changing climate conditions such as precipitation. Our analysis results revealed a distinct north-south gradient in the projected changes of the water withdrawal pattern across U.S. (mostly between ± 10%), with the southern (northern) states showing projected increase (decrease) in future water usages in response to the projected changes in mean annual precipitation by the end of Century under the RCP8.5 scenario. In a similar fashion, our data-driven modeling framework allows for analyzing and documenting the sensitivity of future changes in water withdrawal in response to other climatic (e.g., HDD changes) and socioeconomic factors (e.g., changes in farmland expansion, urbanization, energy generation); either individually (considering one at a time) or in combination.

Table 1. Summary of models performance given as correlation coefficient (R^2), fitted Root Mean Square Error (RMSE; million gal/day/person), and Leave one out cross validation (LOOCV) RMSE. Each model is trained and tested using all available data records for the period 1991-2010.

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**

Table 2. Summary of models predictive accuracy. Each Model is trained using 1991-2005 data and tested using 2006-2010 data. Summary performance is presented here in terms of correlation coefficient (R^2), fitted Root Mean Square Error (RMSE; million gal/day/person), Leave one out cross validation (LOOCV) RMSE, and prediction RMSE (for the test data). See Appendix D for more details on LOOCV-RMSE.

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**

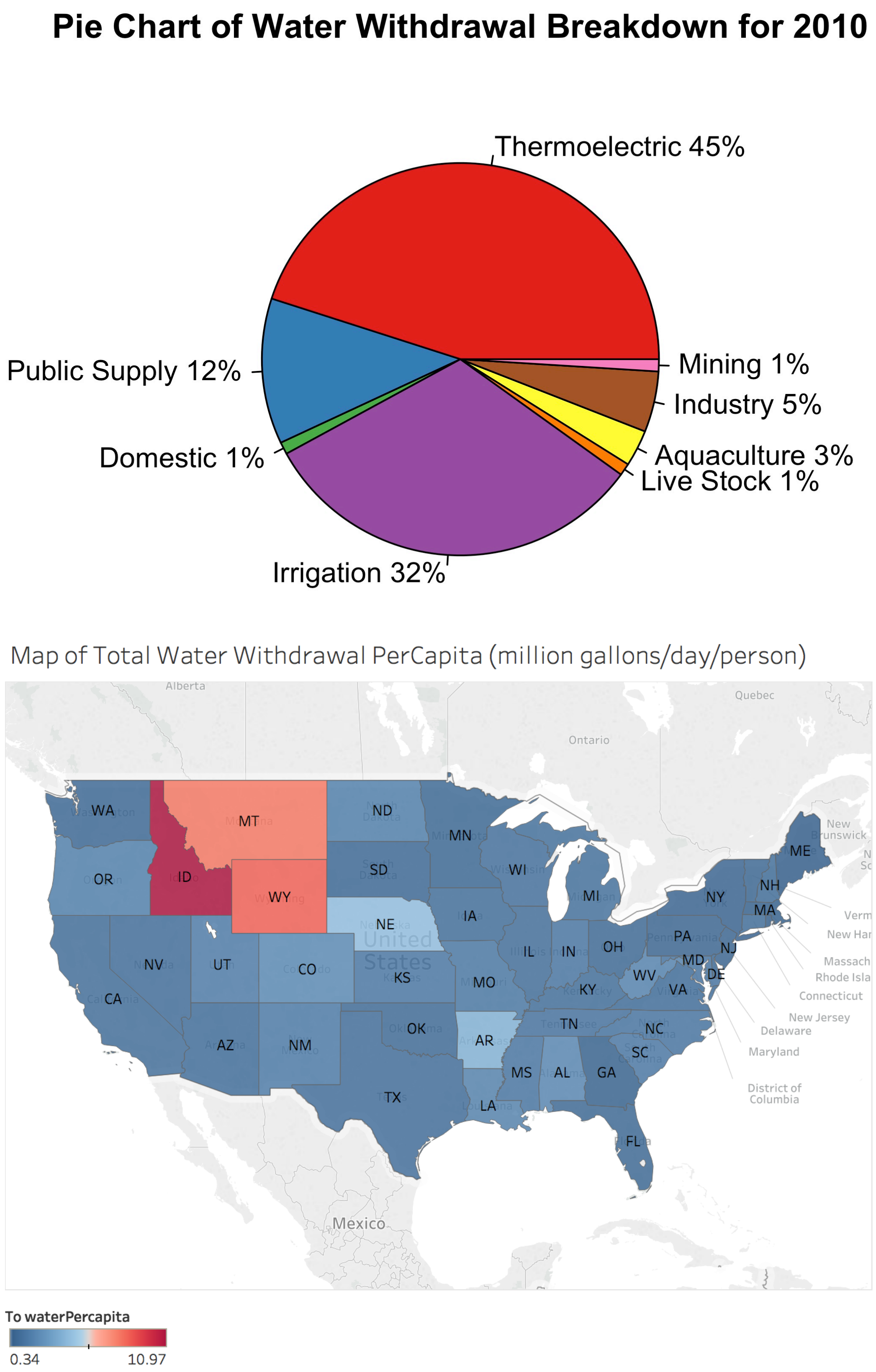


Figure 1. Top: The breakdown of US-wide water withdrawals across the eight major sectors during the period 2006-2010. Bottom: Spatial distribution of the U.S. wide per-capita water withdrawal (in million gallons per-day).

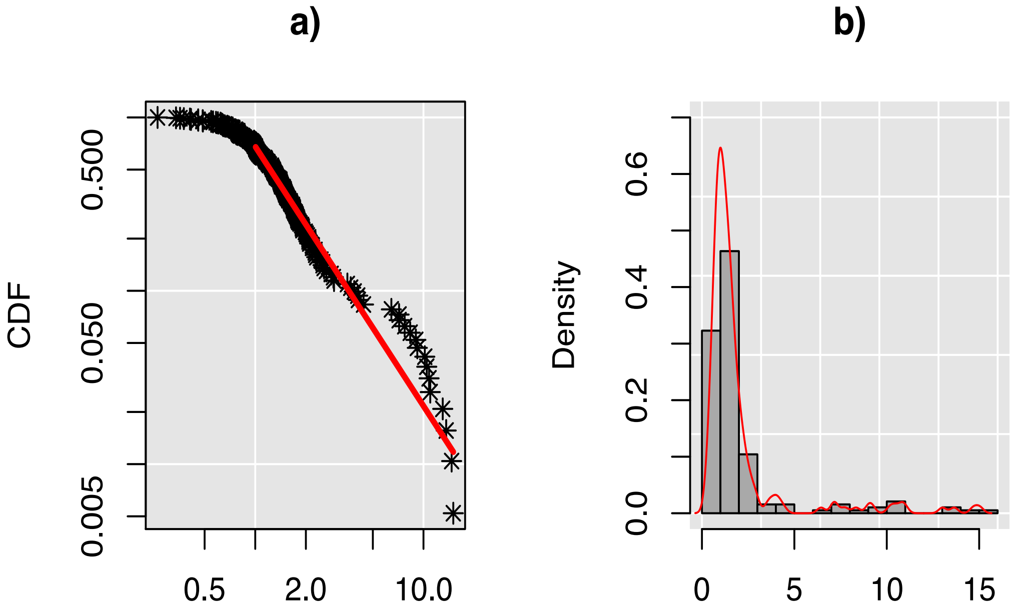


Figure 2. The empirical distribution of per-capita water withdrawals (in million gallons per day) for the period 1991-2010; (a) the red line shows that power-law fits the tail of the empirical cumulative distribution reasonably well (b) the histogram of per-capita water demand with overlain kernel density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

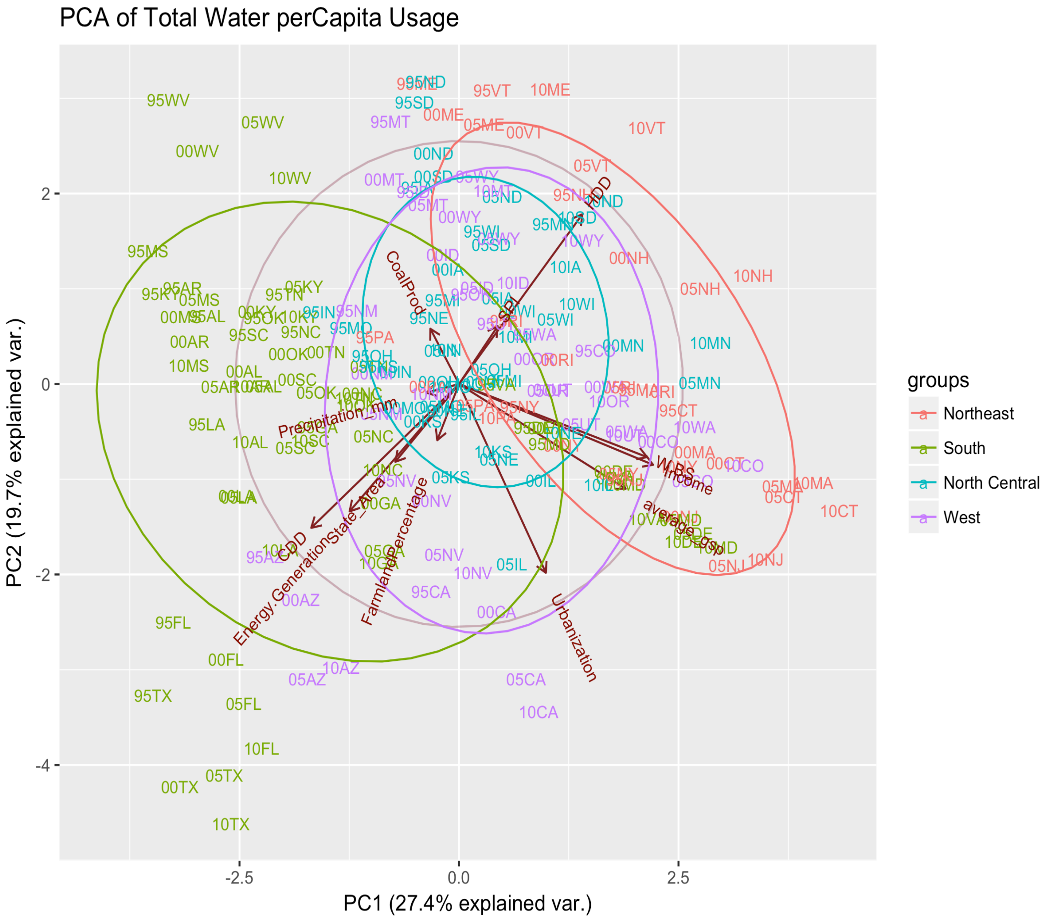
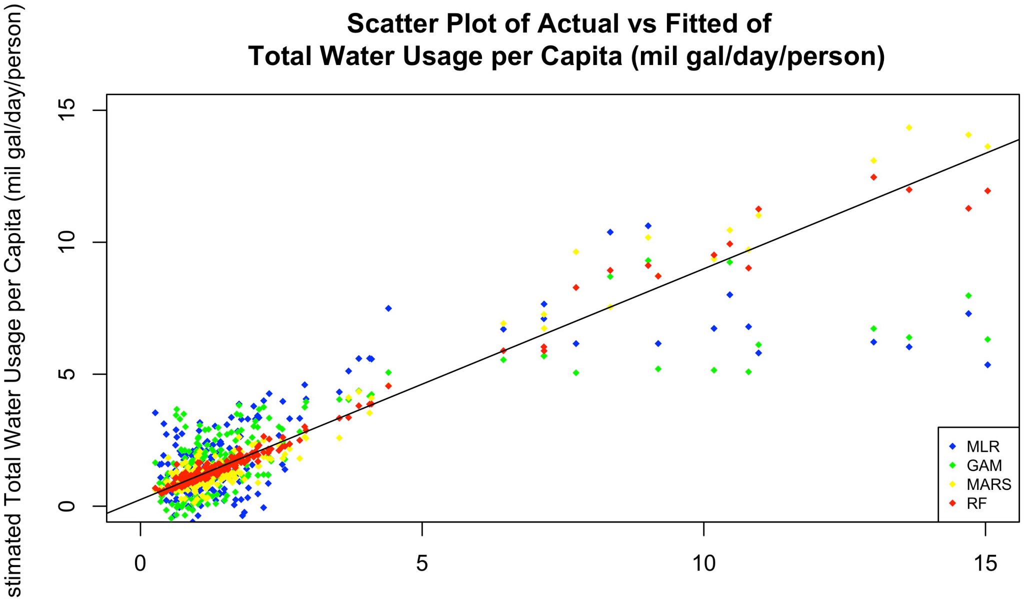


Figure 3. Principal Component Analysis (PCA) biplot of the per-capita water usage (in million gallons per-day) for the period 19952010. The states are color-coded based on their proximity to water bodies and the two digits next to the state codes indicate the year associated with the water use data for the state.



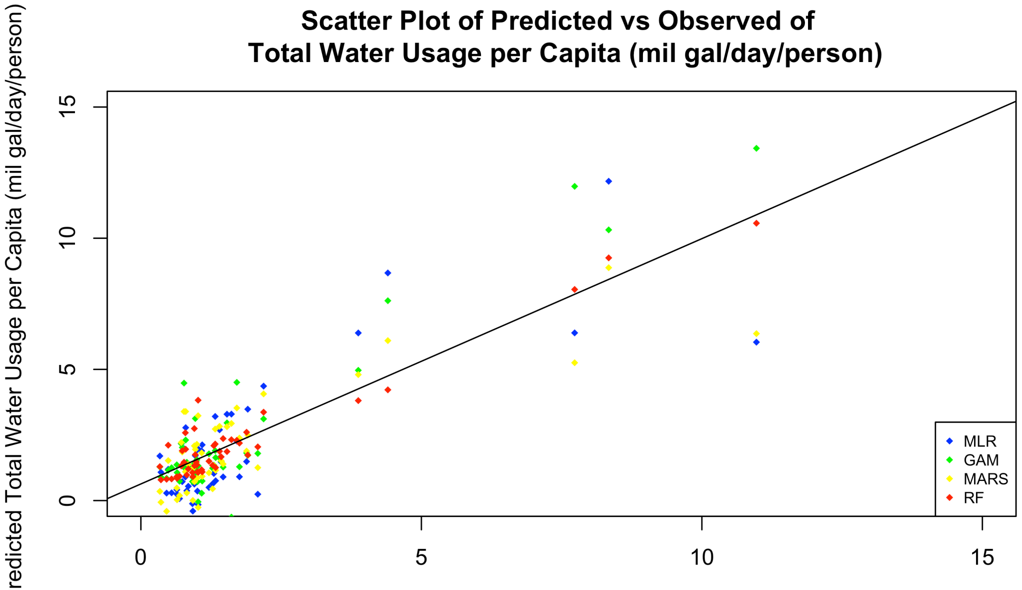


Figure 4. Top: Scatter plot of observed versus estimated values of per-capita water withdrawal (in million gallons per-day) using data of 1995-2010. Bottom: Scatter plot of observed versus predicted values of per-capita water usage (in million gallons per-day) using data of 2006-2010. In the latter case, the models were trained using data of 1995-2005, and the testing was conducted in an independent period of 2006-2010.

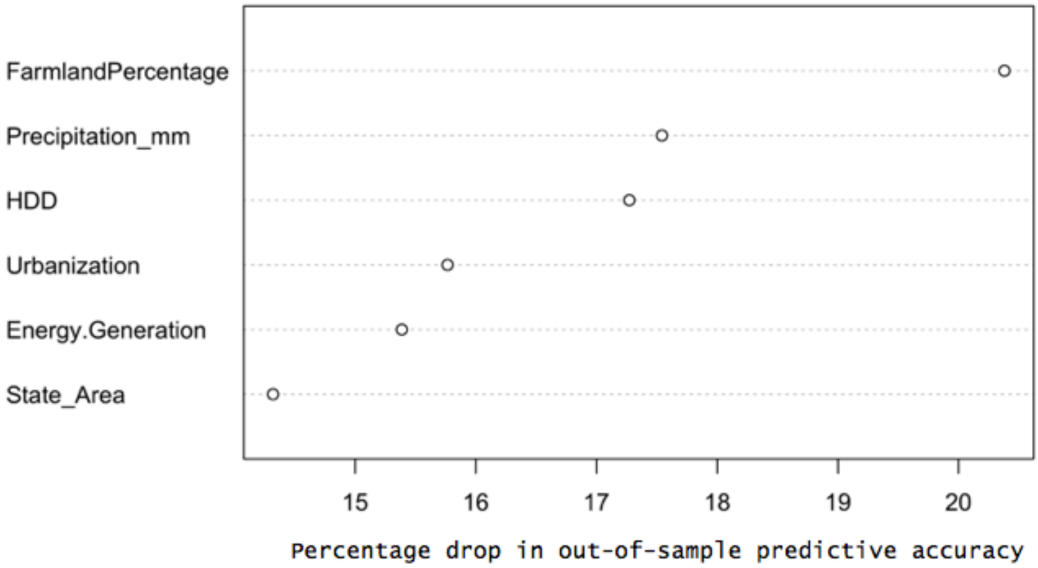


Figure 5. List of the most important predictors identified for the per-capita water withdrawal predictions, presented here as the percentage drop in the predictive accuracy for the out-of-sample datasets. The selected predictors are ranked from the most to least influential ones (top to bottom).

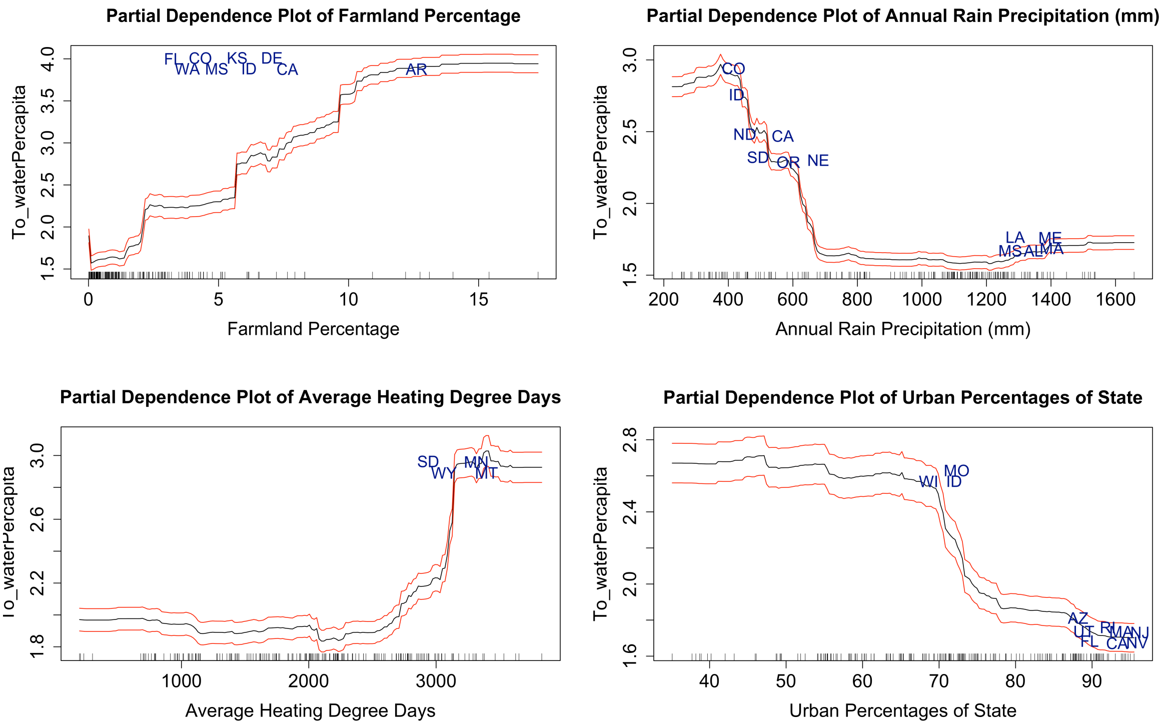


Figure 6. Partial dependency plot (PDP) for the fraction of irrigated farmland, annual precipitation, average heating degree days, and percentage of urban areas; depicting their sensitivity on the per-capita water withdrawal (in million gallons per-day). The two letters on the plot corresponds to the states, the black line the mean values, and the red lines the 95% confidence intervals.

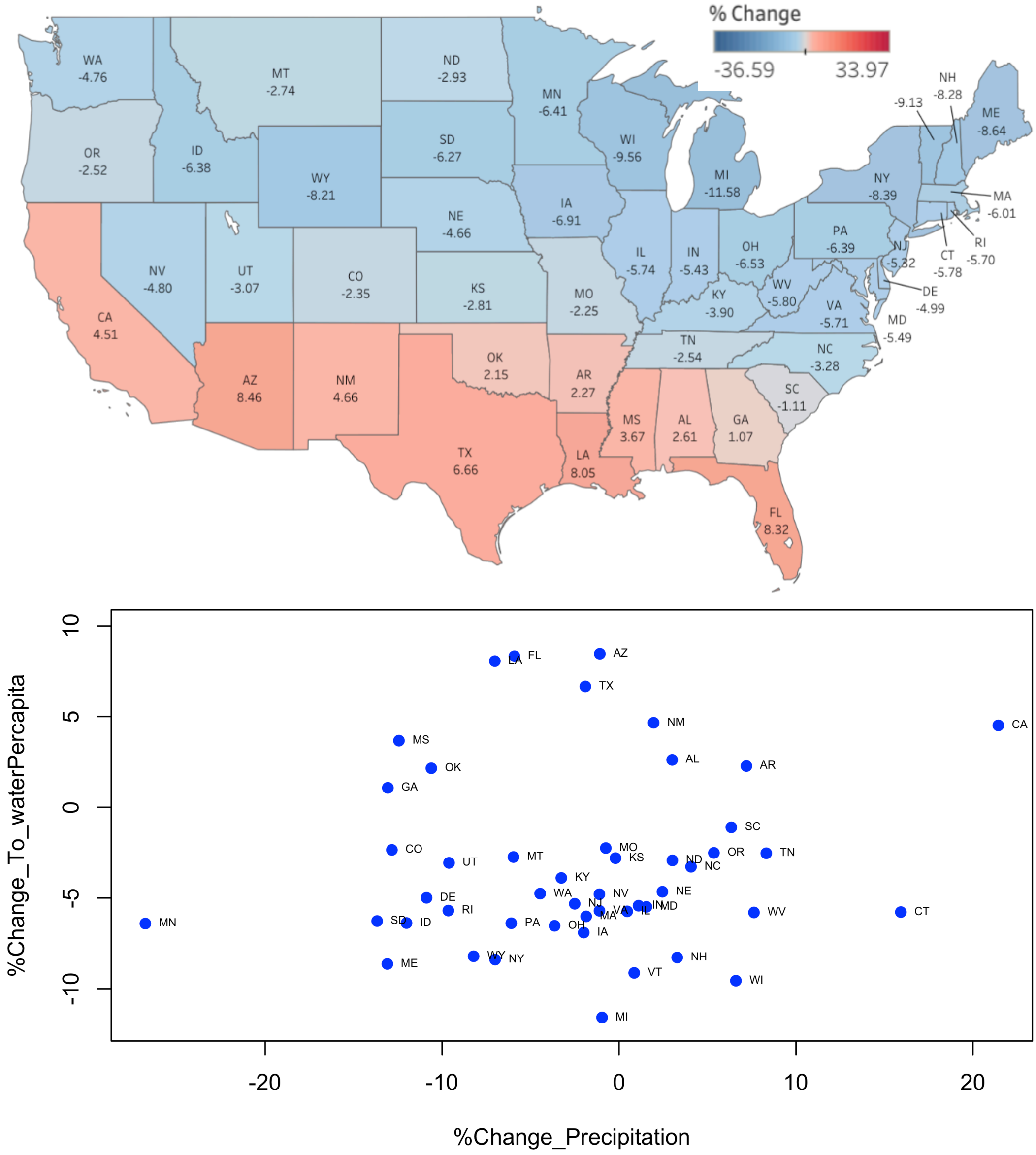


Figure 7. Spatial distribution of ensemble mean changes in the per-capita water withdrawal patterns across the U.S. in response to the future precipitation changes. Ensemble means were estimated based on the modeled WW values using the mean annual precipitation estimates from the five CMIP5 GCMs, while other predictors were kept constant at nominal values (see the corresponding texts for more details). Changes in the WW estimates corresponds to the future period (2070-2085) under the RCP8.5 scenario, relative to the reference estimates of the contemporary conditions (1995-2010). Bottom panel shows the scatter plot of percentage changes between precipitation and total per-capita water withdrawal.

A Generalized Linear Model (GLM)

GLM is the extension of Ordinary Linear Regression (OLR). GLM still retains all the assumption of OLR, it allows predictors to be categorical and allows interactions between predictors. The simplest form of GLM is define as one of the most widely used methods for function approximation. GLM can be mathematically summarized as below:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_n x_{i,n} + \epsilon_i \tag{A.1}$$

where the stochastic error is assumed to be normally distributed as: for all $\epsilon_i \sim N(0, \sigma^2)$. Each β_j describes the slope of predictor $x_{(i,j)}$. Sometimes transformation of original variables (such as polynomials) are used to improve the performance of the models. Multiple linear Regression (MLR) is popular because they can be easily fitted (even with limited data) and they are easily interpretable. However, their 'rigid' structure often fail to approximate the true function, especially when response is a complex (nonlinear) function of input variables. Their predictive accuracy is therefore often inferior to more flexible models (James, Witten, Hastie, & Tibshirani, 2013).

B Generalized Additive Models (GAM)

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 is a non-parametric (local parametric) fitting procedure where the conditional expectation of y is related to the input variables space as shown below:

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

where $f_j(x_{i,j})$ is a smoothing splines over the p-dimensional input space, with the number of observations running from $i = 1, \ldots, n$. GAM relaxes the linearity assumption of multiple linear regression with smoothing functions $f_j(x_{i,j})$. This allows for capturing the non-linear relationship between the predictors and the response variable. The flexibility of generalized additive model often result in better approximating the true function and therefore often outperform GAM in predictive accuracy.

C Multivariate Adaptive Regression Splines (MARS)

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

$$(x-t)_+ = \begin{cases} x - t & x > t \\ 0 & otherwise \end{cases} \tag{C.1}$$

$$(x-t)_- = \begin{cases} t - x & x < t \\ 0 & otherwise \end{cases} \tag{C.2}$$

And MARS has the function form of

$$f(X) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(X) \tag{C.3}$$

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).

D 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 over-fitting – and therefore have high variance. Bias-variance trade-off lies at the heart of developing models with high generalization power add references. Cross-validation is one of the most widely used methods in balancing bias and variance. We use the leave-one-out cross validation (LOOCV) to estimate predictive accuracy. The LOOCV procedure is defined as holding out one data as a test data and use the rest of the training data. Model generated from the training data is the used to predict the test data and we will calculate the MSE of that point. LOOCV MSE is defined by

$$LOOCVMSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2 \tag{D.1}$$

where i represents the iteration of one data left out, y_i represents the true value of the i^{th} iteration, $\hat{y_i}$ represents the predicted value and n the length of data.

Acronyms

BAU Business as Usual
CDD Cooling Degree Days (°F)
CPC Climate Prediction Center
CMIP5 Coupled Model Intercomparison Project (Phase 5)
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
GFDL-ESM2 Geophysical Fluid Dynamics Laboratory-Earth System Models
GLM Generalized Linear Model
GSP Gross State Product (millions of USD measured in 2009 real dollars)
HadGEM2-ES Met Office Hadley Centre Model-Earth System
HDD Heating Degree Days (°F)
IPSL-CM5A-LR Institut Pierre Simon Laplace Model-5 Component models
ISI-MIP Inter-Sectoral Impact Model Intercomparison Project
NEMS National Energy Modeling Systems
NOAA National Oceanic and Atmospheric Administration
MARS Multivariate Adaptive Regression Splines
MIROC-ESM-CHEM Model for Interdisciplinary Research on Climate-Earth System Models
NorESM1-M Norwegian Earth System Model 1 - Medium resolution
PDP Partial Dependence Plot
RCP Representative Concentration Pathway

RF Random Forest
SPI Standardized Prediction Index
U.S. United States
USD United States Dollar ($)
USGS United States Geological Survey

Acknowledgments

Funding for this project was provided by the NSF grant SEES#1555582 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 are also grateful to the Inter-Sectoral Impact Model Inter-comparison (ISI-MIP) Project for providing the climate model outputs used in this study. We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modeling groups for producing and making available their model output. All data-sets used in this study are collected from publicly available sources as detailed in Section 3. The source codes of the different statistical algorithms used in this study can be acquired from the corresponding authors, and they will be hosted on the on-line repository platform (Github).

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/;Lastaccessedon04/04/2017 .`*

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

Breiman, L. (2001).
Machine Learning, *45*(1), 5–32.

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.

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

Davies, E. G., 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.

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 resolution for 1948 to present. *Journal of Geophysical Research: Atmospheres*, *109*(D10).

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., … Tanaka, K. (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., … Tanaka, K. (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.* Springer New York.

Hayes, M., Svoboda, M., Wall, N., & Widhalm, M. (2010). The Lincoln Declaration on Drought Indices: Universal Meteorological Drought Index Recommended. *Bull. Amer. Meteor. Soc.*, *92*(4), 485-488.

IOWA. (2017). *Historical Urban Percentage of the Population for States. Source: Decennial Census, U.S. Census Bureau. Retrieved from Iowa State University, Iowa Community Indicators Program: Available at:* `http://www.icip.iastate.edu/tables/population/urban-pct-states`*. Last accessed on 04/04/2017.*

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

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.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomforest. *R news*, *2*(3), 18–22.

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.* Office of Scientific and Technical Information (OSTI). Lawrence Berkeley National Laboratory, Berkeley, CA.

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.* US Geological Survey.

McKee, T. B., Doesken, N. J., & Kleist, J. (1993). The relationship of drought frequency and duration to time scales. *Eighth Conference on Applied Climatology, Anaheim, California, 17-22 January 1993.*

NOAA. (2017). *National Oceanic and Atmospheric Administration (NOAA), Degree Days Statistics National Weather Service; Center for Weather and Climate Prediction. Retrieved fromm:* `http://www.cpc.ncep.noaa.gov/products/analysis_monitoring/cdus/degree_days/`*; Last accessed on 04/04/2017.*

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.

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. v., Wanders, N., Wada, Y., Bosmans, J. H., Drost, N., . . . others (2018). PCR-GLOBWB 2: a 5 arcmin global hydrological and water resources model. *Geoscientific Model Development*, *11*(6), 2429–2453.

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., Roo, A. d., Dirmeyer, P. A., Famiglietti, J. S., Hanasaki, N., . . . others (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. (2014). Global modeling of withdrawal, allocation and consumptive use of surface water and groundwater resources. *Earth System Dynamics*, *5*(1), 15.

Warszawski, L., Frieler, K., Huber, V., Piontek, F., Serdeczny, O., & Schewe, J. (2014). The Inter-Sectoral Impact Model Intercomparison Project (ISI-MIP): Project framework. *Proceedings of the National Academy of Sciences*, *111*(9), 3228–3232.