Predicting Residential Water Demand with Machine-Based Statistical Learning

Dongwoo Lee, S.M.ASCE[1]; and Sybil Derrible, A.M.ASCE[2]

Abstract: Predicting residential water demand is challenging because of two technical questions: (1) which data and variables should be used and (2) which modeling technique is most appropriate for high prediction accuracy. To address these issues, this article investigates 12 statistical techniques, including parametric models and machine learning (ML) models, to predict daily household water use. In addition, two data scenarios are adopted, one with only 6 variables, generally available to cities and water utilities (general scenario), and one with all 19 variables available from the Residential End-Use 2016 database (REU 2016 scenario). The results for the REU 2016 scenario indicate that ML models outperform linear models. In particular, gradient boosting regression (GBR) performs best with an R^2_{adj} of 0.69 compared to 0.54 for linear regression. The performance gap between ML and linear models becomes even wider for the general scenario with an R^2_{adj} of 0.60 for GBR compared to 0.33 for linear regression. The finding in this article can be useful to researchers, municipalities, and utilities seeking novel modeling techniques that can provide consistent modeling performance—i.e., high prediction accuracy—depending on data availability. Future work could include the development of new measures to increase the interpretability of ML models to better understand causal relationships between independent variables and daily household water use. **DOI: 10.1061/(ASCE)WR.1943-5452.0001119.**

Introduction

Being able to adequately model water demand is essential for municipalities and utilities to effectively meet consumer demand while managing the available supply of water (House-Peters et al. 2010). In fact, modeling water demand has been integral not only to water resource planning but also to urban infrastructure planning and policy decision-making. This is especially the case now as cities are expanding while simultaneously trying to consume less energy and fewer resources (Derrible 2016, 2017, 2018). Specifically in the water realm, efforts should be put into effectively modeling water demand of single-family households since they are the primary consumers of public-supply water use in North America (DeOreo et al. 2016).

Determining the right modeling approach (i.e., modeling algorithm) to predict household water demand is a challenging task because water demand can be affected by numerous factors, including technological, demographic, social, economic, and climate characteristics, and public policies (Donkor et al. 2012; Fricke 2014; House-Peters and Chang 2011). A wide variety of statistical techniques exists that can be used to model household end-use water demand. In general, parametric statistical models (also known as parametric models, such as linear regression) are the most commonly applied models to predict household water use. Recently, with the advent of data science and big data, the capabilities of available data mining techniques (also known as machine-based statistical learning, such as neural networks) seem virtually limitless (Ahmad et al. 2016, 2017; Ahmad and Derrible 2015; Derrible and Ahmad 2015; Golshani et al. 2018; Lee et al. 2018), and they offer new opportunities to model household water use, especially because they can capture unobserved patterns and nonlinear relationships (Friedman et al. 2001; Lee et al. 2018; Bishop 2006).

Both families of algorithms [parametric and machine learning (ML) models] have their own technical characteristics and methodological advantages (e.g., interpretability versus ability to capture nonlinearities) that need to be leveraged based on the context in which they are applied. For instance, although more intuitive and interpretable than other models, parametric models (e.g., linear regression) tend to generate more prediction errors than ML models. Furthermore, parametric models require a high degree of domain knowledge to construct and adjust a model's configuration while taking into account the underlying relationships between factors (e.g., to avoid collinearity issues). In contrast, ML models show a high degree of predictive accuracy with most data sets thanks to their ability to capture nonlinear and complex characteristics in the data, but many ML models are less interpretable than parametric models owing to their reliance on machine-based computation processes. Thus, selecting the right modeling algorithms is not trivial. In this article, 12 statistical learning algorithms are tested, including 4 parametric statistical models and 8 ML models (Table 1). To validate the models, a fivefold cross-validation (5-fold CV) process is applied (detailed later) (Friedman et al. 2001; Kohavi 1995; Zhou 2012), which is generally used to check the performance of ML models.

Beyond modeling, data availability can also be an issue. For example, in the United States, most municipalities and utilities do not have access to detailed water-use data sets such as longitudinal (i.e., multilevel) or time-series (e.g., smart metered data) data sets, although they have access to general individual- or household-level information from publicly accessible microdata. In this context, this article is purposely designed to investigate the role of data availability in modeling performance. For this, two data scenarios are elaborated, one that intentionally only includes commonly available variables (i.e., general scenario) and one that contains

[1]Postdoctoral Researcher, Complex and Sustainable Urban Networks Laboratory and Civil and Materials Engineering, Univ. of Illinois at Chicago, Chicago, IL 60607 (corresponding author). Email: dlee226@uic.edu

[2]Professor and Director, Complex and Sustainable Urban Networks Laboratory, Civil and Materials Engineering, and Institute for Environmental Science and Policy, Univ. of Illinois at Chicago, Chicago, IL 60607.

Note. This manuscript was submitted on April 4, 2018; approved on March 21, 2019; published online on October 30, 2019. Discussion period open until March 30, 2020; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Water Resources Planning and Management***, © ASCE, ISSN 0733-9496.

Table 1. Statistical methodologies used for two scenarios

Category	Regression methodologies
Parametric statistical learning algorithm (parametric models)	Ordinary least-squares regression (linear regression) Penalized: ridge regression (ridge) Penalized: lasso regression (lasso) Bayesian ridge regression (BRR)
Nonparametric statistical learning algorithm (ML models)	SVM with radial basis function (RBF) kernel (RBF-SVM) SVM with linear kernel (linear-SVM) Kernel ridge regression (KRR) Gradient boosting regression (GBR) Random forest (RF) regression K-nearest neighbor (KNN) regression Multilayer perceptron (MLP) regression GRNN

many variables from the Residential End Uses of Water survey (REU 2016 scenario) carried out by the Water Research Foundation (WRF). Specifically, the general scenario only includes six variables related to the demographic and economic characteristics of households and climate. These six variables were selected because they are mostly accessible to all US municipalities and utilities from micro-level public data sources or other city-level microdata samples. In contrast, the REU 2016 scenario includes 19 cross-sectional variables (i.e., for a single "typical" day), including variables related to water-saving behavior and detailed water-use patterns; technical details on how the REUW 2016 data set is used in this study is provided in the data preprocessing section.

Overall, this study will be useful to researchers, municipalities, and utilities seeking to model and forecast residential water use, especially when only limited data are available. The lessons from this study can be used for short- and long-term planning, especially in areas of rapid growth, and for routine operations by utilities. Moreover—although only partly done in this work—the models developed can also be used to infer the impact of individual variables on residential water use (e.g., determine which variables impact water use the most).

This study contains six sections. After this section, the section "Literature Review" briefly reviews the literature on the two technical aspects central to this work. The section "Research Design and Data Preprocessing" details the data and the analysis process. The section "Methodology" defines the 12 statistical learning algorithms and performance indicators selected for this article. In the section "Model Result and Discussion," the overall findings of the study are presented and discussed, and several future tasks are suggested. Finally, the final section concludes the article.

Literature Review

As mentioned, statistical algorithms can be broadly categorized into two families: parametric statistical models and ML models. Numerous studies focus on modeling household end-use water demand. In particular, parametric models have most commonly applied to predict household water use since they are easy to interpret and are based on strong predetermined assumptions (Arbues et al. 2010; Arbues and Villanua 2006; Brentan et al. 2017; Donkor et al. 2012; Goodchild 2003; Guhathakurta and Gober 2007; House-Peters et al. 2010; House-Peters and Chang 2011; Kenney et al. 2008; Kontokosta and Jain 2015). While parametric models are theoretically intuitive and easy to interpret (i.e., since they yield parameters), they also pose serious statistical issues.

First, parametric models have predetermined structures (e.g., residuals are assumed to follow a normal distribution), and a hypothetical test is performed to statistically validate the relationship (Hastie et al. 2009). Furthermore, a single parametric equation is globally employed and is supposed to hold over the entire data set (i.e., the same relationships are assumed to apply to everyone), while it is notoriously difficult for a linear parametric model to find a best-fitting mathematical function (Friedman et al. 2001; Hensher et al. 2005; Kuhn and Johnson 2013). To partially alleviate this issue, modeling algorithms incorporating clusters (i.e., generalized mixed-effect model) have been used by controlling detrimental effects (e.g., random and fixed) (House-Peters and Chang 2011; Wooldridge 2010). Nonetheless, these modeling algorithms are preferentially applied to multilevel data (e.g., longitudinal) that may not be accessible to many cities. Furthermore, finding the best-fitting configuration of a model while taking into account underlying interactions and relationships between variables (e.g., nonlinearity) is not trivial (Breiman et al. 1984; De'ath 2002; Elith et al. 2008).

As an alternative to parametric models, ML models have also been widely used in the urban infrastructure literature in general (Akbarzadeh et al. 2019; Derrible and Ahmad 2015; Golshani et al. 2018; Lee et al. 2018; Wisetjindawat et al. 2018) and specifically for household water use (Adamowski et al. 2012; Altunkaynak and Nigussie 2017; Al-Zahrani and Abo-Monasar 2015; Bai et al. 2014; Donkor et al. 2012; Firat et al. 2009, 2010; House-Peters and Chang 2011; Vitter and Webber 2018; Yurdusev et al. 2010). In general, ML models have been shown to have high predictive performance in a wide range of modeling applications thanks to significant advances in computational ability. Specifically, ML models can recognize nontrivial patterns from a data set that often result in high prediction accuracies.

In particular, artificial neural network (ANN) models have been widely applied to predict or forecast water consumption (Altunkaynak and Nigussie 2017; Firat et al. 2009, 2010). For instance, Firat et al. (2009, 2010) used six different ANN models to forecast monthly water consumption using time-series data and found that generalized regression neural network (GRNN) models performed best. Apart from ANNs, Bai et al. (2014) used a stepwise support vector machine (SVM) regression to forecast daily water consumption using time-series data, which is called a variable-structure SVM. Instead of using ML to directly model water demand, ML models are also applied to facilitate water demand analysis. For example, Vitter and Webber (2018) used a SVM classifier to classify specific water-use events (e.g., shower, clothes washing) in households by incorporating electricity consumption information that correlates to water consumption. Numerous other ML models have been applied to predict other urban resources (e.g., electricity and energy), including kernel-based methods, boosting methods, and bagging methods (Bansal et al. 2015; Kusiak et al. 2010; Lozano and Gutiérrez 2008; Robinson et al. 2017; Tso and Yau 2007).

In general, ML models include nonparametric (e.g., kernel) and complex structure (e.g., network) models that can capture nonlinear or complex relationships between various factors and target values (e.g., household water use). Furthermore, they generally provide higher predictive performances than parametric models in resource demand modeling (Al-Zahrani and Abo-Monasar 2015; Firat et al. 2009, 2010; Robinson et al. 2017) since nonparametric features in the model are trained by machine-based repetitive computation. Owing to their machine-based computation, however, ML models more commonly face overfitting issues, and they are also generally

less interpretable than parametric models (e.g., neural networks are often described as black-box models). To address the interpretability issues in ML models, several useful statistical measures exist. For instance, rule-based ensemble methods, such as gradient boosting regression (GBR), are able to examine the marginal effect of a feature on the predicted values of a learned model (a.k.a., partial dependence plot) (Doshi-Velez and Kim 2017; Friedman et al. 2001; Natekin and Knoll 2013).

In addition to modeling methodologies, the performance of water demand models largely depends on the quality of the data available that properly capture the relationship between water demand and the factors affecting the demand. Previous studies on household water demand modeling found that the most significant factors affecting water use include household demographic factors (e.g., size, income, and type) (Arbues et al. 2010; Arbues and Villanua 2006; DeOreo et al. 2016; Domene and Saurí 2006; Grafton et al. 2011; House-Peters et al. 2010; Mayer et al. 1999; Mazzanti and Montini 2006; Schleich and Hillenbrand 2009), climate factors (e.g., precipitation and temperature) (Donkor et al. 2012; Froukh 2001; Goodchild 2003; Guhathakurta and Gober 2007; House-Peters et al. 2010; House-Peters and Chang 2011; Jentgen et al. 2007; Kontokosta and Jain 2015; Lee et al. 2010, 2015; Schleich and Hillenbrand 2009), price, and detailed water-use and associated attitudinal information related to households (Arbues et al. 2010; Arbues and Villanua 2006; Cominola et al. 2018; DeOreo et al. 2016; Fricke 2014; Ghimire et al. 2015; Grafton et al. 2011; House-Peters et al. 2010; Kontokosta and Jain 2015; Vitter and Webber 2018; Willis et al. 2011). The first two factors (i.e., household demographics and climate factors) are generally available to cities and water utilities in the United States, which is not the case for detailed information on water use and its behavioral characteristics, which is rarely available.

Research Design and Data Preprocessing

Research Design

This article is designed to examine two common technical issues and investigate the modeling performances of 12 techniques under two data scenarios (see research framework in Fig. 1). Before the main analysis, this study conducts a thorough descriptive analysis to detect the presence of statistical issues in the data set, which is often the case for data that relate to resource consumption (e.g., water and electricity). Then the main analysis focuses on training 12 statistical learning algorithms (Table 1) on 70% of the data (i.e., train set) under 2 data scenarios (i.e., general and REU scenario). A 5-fold CV process is also applied to the training set. The learned models are then validated on the remaining 30% of the data (i.e., test set). The next section offers details on the data and the two modeling scenarios.

Data Preprocessing

This article uses the REU 2016 database (DeOreo et al. 2016) released by the Water Research Foundation (WRF). The REU 2016 study contains extensive household water-use information from 24 water utility companies across the United States and Canada. The REU database consists of four main data sets that come from two main sources: (1) household water use (e.g., 12 days of metered consumption) and billing information (e.g., annual water consumption) and (2) household survey responses (DeOreo et al. 2016). In particular, metered consumption was originally measured every 10 s for 2 weeks, but it was subsequently aggregated by day for 12 days (DeOreo et al. 2016).

This study mainly uses two data sets from the REU 2016 database: daily household water use ("REU2016_Daily_Use_Main")

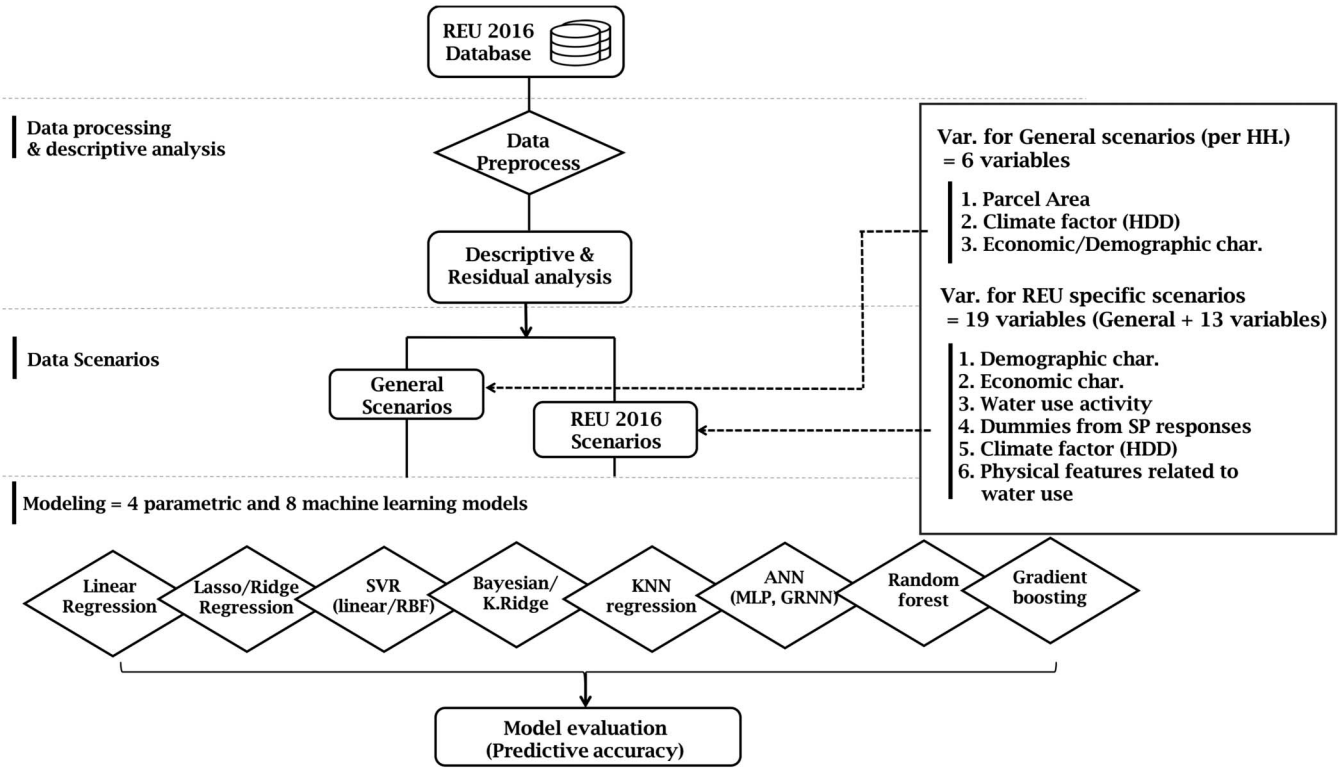

Fig. 1. Research Design.

J. Water Resour. Plann. Manage., 2020, 146(1): 04019067

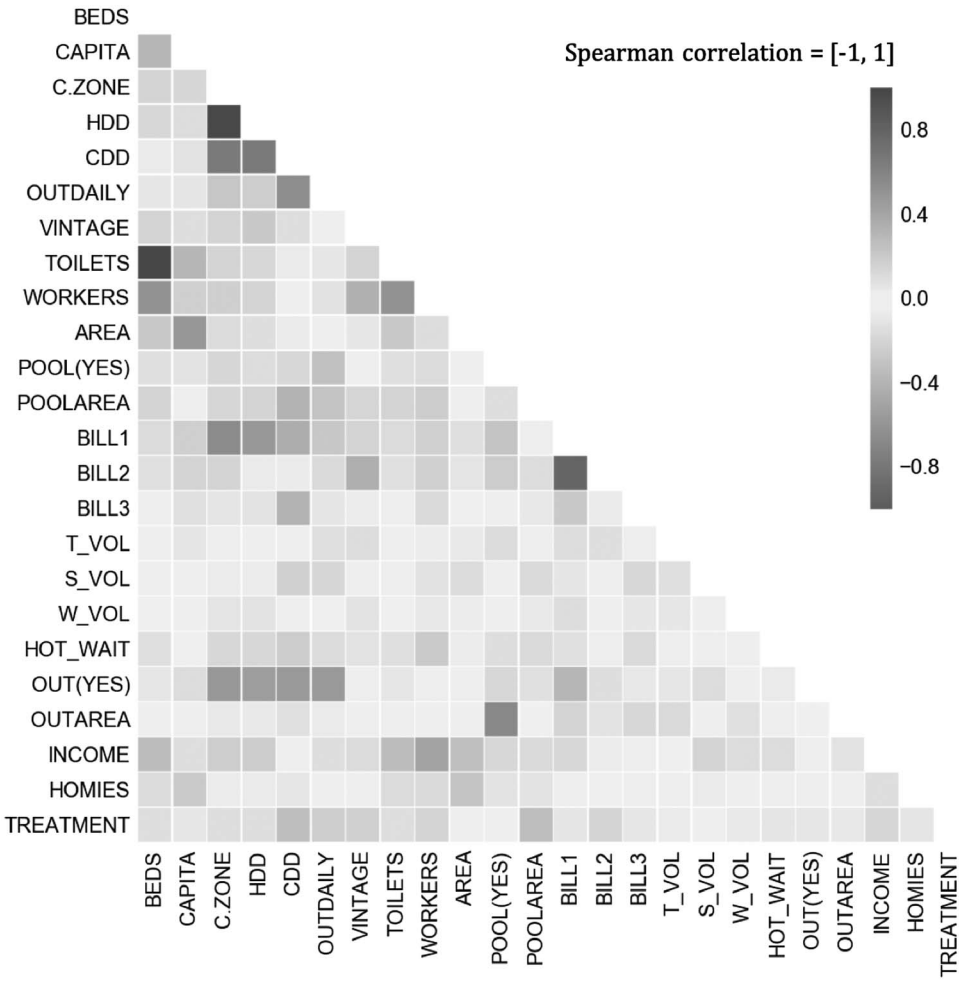

Fig. 2. Correlation of independent variables (predictors).

and mailed household survey information about demographics and water consumption behaviors ("REU2016_End_Use_Sample"). The daily household water-use data set includes nine utilities for a total of 771 households over 12 days. In total, the number of observations available is around 9,300 (some households include more than 12 days). Furthermore, the mailed survey data set contains detailed information on household demographic and economic characteristics as well as water consumption behaviors in the form of revealed preference (RP) and stated preference (SP).

From this database, this study purposely creates cross-sectional data by combining average daily household water use with household survey information, based on the given identification codes (KEYCODES). To calculate the average daily water use, the 12 recorded days of household daily water use are averaged. In essence, this data set is transformed into cross-sectional information to model one single "typical" day for the 771 households.

Subsequently, the combined data contained numerous missing values, and some variables contained redundant or interrelated information that could bias the results (i.e., collinearity issues). Therefore, multiple cleaning and variable selection processes were conducted initially. In particular, variables with a very low response rate (<10%–20%) were eliminated, and some variables having redundant or interrelated information were merged. In addition, household water demand also depends on climate conditions, which must be taken into account since not all households are located in the same geographic area. For this study, the number of heating degree days (HDDs), the number of cooling degree days (CDDs), and the climate zone (CZ) of each household were added to the data set from www.degreeday.net and from maps provided by the American Society of Heating, Refrigeration, and Air-Conditioning Engineers (ASHRAE).

To further study the relationship between the variables, Fig. 2 shows the correlation matrix between all independent variables. Specifically, no significant collinearity issues were detected that might lead to biased estimation; i.e., Spearman coefficient = 1.0. Nonetheless, as expected, HDD, CDD, and climate zone are strongly correlated, and the pairs between outdoor properties (e.g., pool) and climate conditions also show some correlation. Moreover, variables related to household size also show some correlation, such as number of toilets and bedrooms. In each case, only one of the variables that showed some correlations was selected; for instance, only HDD was selected. In the end, the data set contained 24 variables and 531 observations (i.e., single-family households). The full list of variables used is shown in Table 2.

Overall, the general scenario includes six variables: number of workers, household size, income, type, areas, and HDD. In particular, in the United States, these variables are available from publicly accessible micro data sets, such as the American Community Survey (ACS) and the Public Use Microdata Sample (PUMS), and from community-level household surveys. Although public micro data sets are mostly anonymous and only contain a limited number of samples, utilities and municipalities can use this information

Table 2. Variables used in two scenarios: general and REU 2016 specific

Variable type	Variable	Description	N	Mean	Standard
		General scenario: six variables			
Independent variable (x)	Capita	Number of people in household	531	2.73	1.44
	HDD	Heating degree days	531	4,098.92	2,432.28
	Employed adults	Number of workers in household	531	1.32	0.9
	Income	Household income ($10,000)	531	8.17	5.26
	Parcel area	Size of parcel area (m^2)	531	809.08	496.0
	Dummy outdoor	Existence of outdoor properties is 1, otherwise 0 (e.g., garden, tree, lawn, and pool)	531	0.6	0.49
		REU 2016 scenario: general scenario (6 variables) + 13 variables			
Independent variable (x)	Bedrooms	Number of bedrooms in household	531	3.38	0.87
	Outdoor area	Size of outdoor area (m^2)	531	320.98	330.14
	Pool area	Size of pool area (m^2)	531	15.72	17
	Homies	Person usually stay at home	531	1.01	0.85
	Vintage	Vintage of home	531	34.59	19.44
	Fixed charges	Fixed rates for water	531	17.6	9.57
	Marginal rate	Marginal rates for water	531	4.98	2.24
	Dummy treatment	Treatment system in household is 1, otherwise 0 (e.g., water softener or reverse osmosis system)	531	0.12	0.33
	Dummy pool	Household with pool (indoor or outdoor) is 1, otherwise is 0	531	0.11	0.32
	Dummy toilet flush	Average toilet flush is less than 7.58 l per flush is 1, otherwise is 0	531	0.45	0.5
	Dummy shower flow	Average shower flow is less than 7.58 L/min is 1, otherwise 0	531	0.51	0.5
	Dummy clothes load	Average washer load is less than 11 L per load is 1, otherwise 0	531	0.51	0.5
	Dummy hot water	Hot water wait in master bathroom is 1, otherwise 0	531	0.45	0.5
Dependent variable (y)	Trace daily[a]	Daily water consumption (L/day, lpd)	531	714.98	428.72

[a]Daily water consumption is transformed into $\log_{10}y$. See details in section "Descriptive Analysis" and Fig. 2.

based on existing statistical approaches that are widely used in a resource planning process—see details in Farooq et al. (2013), Guo and Bhat (2007), and Rosca et al. (2018). In contrast, the REU 2016 scenario includes all information in the general scenario and more detailed household level water-use information that describe household water consumption from RP and SP responses. For both scenarios, daily total household water consumption in liters per day is predicted, expressed as *TraceDaily*, that includes both indoor and outdoor water consumption [although only a limited number of households report outdoor properties (e.g., garden, lawn, pool)].

Descriptive Analysis

An early investigation of the distribution of household water use (i.e., *TraceDaily*) reveals that the variable is not normally distributed. Using ordinary least squares (OLS) regression, Fig. 3(a) shows that the distribution of household water use (y) is skewed to the right,

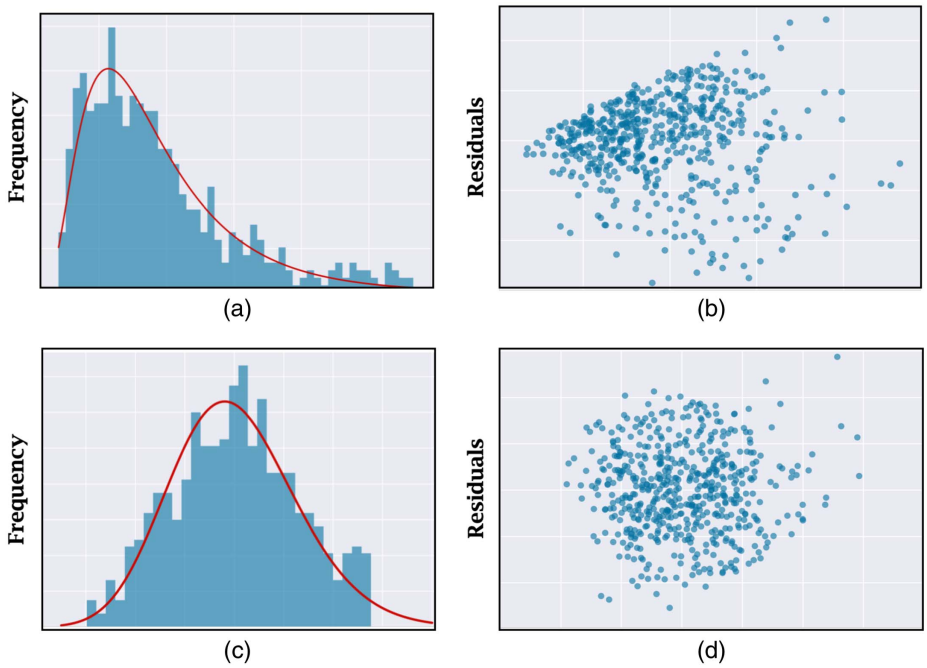

Fig. 3. (a) Distribution of dependent variables (y); (b) residuals of y; (c) distribution of dependent variables ($\log_{10}y$); and (d) residuals of $\log_{10}y$ (residuals are estimated by OLS).

which is common in lognormal distributions. Furthermore, the residual plots, Fig. 3(b), show the presence of a funnel-shaped pattern, suggesting a nonconstant variance in the error terms—i.e., heteroscedasticity—which violates the predetermined assumption in linear models (e.g., linear regression). To solve this heteroscedasticity issue, the actual y can be transformed using a concave function such as the logarithm function ($\log_{10} y$) or the square root of the actual $y(\sqrt{y})$ (James et al. 2013; Kuhn and Johnson 2013). This transformation shrinks the responses, which can alleviate the heteroscedasticity issue. In the literature, the log-transformed is most commonly used (Keene 1995; Kuhn and Johnson 2013; Robinson et al. 2017). As shown in Fig. 3(c), taking the log-transformed *TraceDaily* results in a normal distribution and randomly scattered residuals shown in Fig. 3(d). As a result, $\log_{10} y$ is used as the dependent variable instead of y.

Methodology

Parametric linear models (i.e., linear and penalized regression) assume either that the regression function $E(y|X)$ is linear in the inputs (X) to predict the output (Y) or that the linear model fits reasonably along with a flat hyperplane (Hastie et al. 2009; James et al. 2013; Kuhn and Johnson 2013). Thus, parametric linear models are simple and can sometimes outperform nonlinear models, especially for limited and sparse data (Hastie et al. 2009). In addition to the conventional linear regression technique (i.e., OLS), several parametric linear models introduce additional information or statistical assumptions, such as partial least squares (PLS), and penalized models, such as lasso and ridge regression to decrease the level of biases while preserving the predetermined assumptions (i.e., linearity). In contrast, numerous nonparametric learning models or ML models exist that can be adapted to the data without assuming that a linear regression function in $E(y|X)$ is linear. Due to differences between the two modeling categories, it is difficult to simply conclude which modeling technique is superior to the others because this largely depends on the purpose of the research and the intrinsic characteristics of the data used in the model.

Regardless of the algorithm, all have several common features. In particular, most statistical models estimate the relationship between a set of independent variables $\boldsymbol{x}$ with a dependent variable y while minimizing a loss function. For example, many models minimize the sum of squared errors (SSE), and are then evaluated by measuring how much they managed to minimize SSE, e.g., using the mean of squared errors (MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{N}(y_i - f(\boldsymbol{x}_i))^2 \tag{1}$$

where N is the total number of observations, $\boldsymbol{x}_i$ is a set of independent variable vectors for ith observation, and y_i a dependent variable for ith observation. In fact, the MSE can be decomposed into three parts:

$$E[MSE] = \varepsilon^2 + (\text{Bias})^2 + \text{Variance} \tag{2}$$

The first part (ε^2) consists of the unobserved errors that are impossible to eliminate in modeling. In the second term, "Bias" illustrates how well the estimated model can explain the relationship between x and y. The last term is the variance. Generally, the aim is to control the level of bias and variance when estimating a model. Specifically, more complex models (e.g., artificial neural networks) can have higher variances than models based on a linear assumption (e.g., linear regression), which can lead to overfitting. In contrast, simpler models can have lower variances, but they may not be able to fully infer the relationship between x and y, thereby resulting in underfitting. This trade-off between the two families of techniques is often referred to as the variance–bias trade-off (James et al. 2013; Kuhn and Johnson 2013). The following sections detail the 12 statistical learning techniques selected in this study.

Parametric Statistical Learning Algorithms

Linear Regression Model

Linear regression aims to explain the relationship between a set of independent variable vectors ($\boldsymbol{x}$) and a dependent variable (y) based on the linear function

$$y = f(X) = \beta_0 + \sum_{j=1}^{p}\beta_j X_j \tag{3}$$

where x_j is a vector for the jth independent variable, and β_j and β_0 are unknown parameters (coefficients and an intercept, respectively). This linear combination is estimated by minimizing the SSE between x and y Eq. (1), and it is also known as the standard OLS regression.

Penalized Models: Ridge and Lasso Regression

Penalized models aim to mitigate problems related to model variance when the number of independent variables increases in the standard OLS regression. Specifically, it is possible that highly correlated variables (i.e., collinearity) can greatly increase the variance, and such variance issues can increase the overall MSE. Thus, the family of penalized models, including ridge and lasso regressions, regulate the estimation process by adding a penalty to the SSE. Ridge regression adds an L_2 penalty in the SSE, which controls the trade-off between the variance and the bias. Specifically, this penalty sacrifices some bias, and it can reduce the variance that provides a lower MSE:

$$SSE_{L_2} = \sum_{i=1}^{N}(y_i - f(\boldsymbol{x}_i))^2 + \lambda\sum_{j=1}^{P}\beta_j^2 \tag{4}$$

where λ regulates the inflation of coefficient, and it must be calibrated through validation process.

In addition to the lasso regression, ridge regression (a.k.a., the least absolute shrinkage and selection operator modeling) has an L_1 penalty that substitutes the L_2 penalty in the ridge regression:

$$SSE_{L_1} = \sum_{i=1}^{N}(y_i - f(\boldsymbol{x}_i))^2 + \lambda\sum_{j=1}^{P}|\beta_j| \tag{5}$$

Modified Ridge Regression: Kernel and Bayesian Ridge Regression

Ridge regression is the simplest algorithm that can be kernelized or combined with probabilistic features (e.g., Bayesian). Specifically, $\boldsymbol{x}$ in Eq. (5) is substituted by the kernel function, $\emptyset$:

$$SSE_{L_2} = \sum_{i=1}^{N}(y_i - f(\emptyset_i))^2 + \lambda\sum_{j=1}^{P}\beta_j^2 \tag{6}$$

It is termed kernel ridge regression (KRR) since it uses the same loss function that is used in ridge regression. Alternatively, the context of Bayesian statistics can also be applied to ridge regression. Specifically, the prior information and the posterior mean of a model for the parameter (β_j) follow:

Prior: $\beta_j, \sim N(0, 1/\lambda)$; posterior: $\beta_j, \sim N(0, \sigma^2/\lambda)$ for all j (7)

All the modeling parameters are jointly estimated by maximizing the marginal log-likelihood function.

Nonparametric Machine-Based Statistical Learning Algorithms

Rule-Based Ensemble Models: Random Forest and Gradient Boosting Regression

Tree-based models estimate the relationship between X and y by partitioning the input based on specific rules (i.e., rule-based model). In particular, they provide a set of conditions and results that are highly interpretable and also easily include different types of variables without any assumptions and preprocessing. However, simple trees can have a highly unstable performance and tend to have higher variances than linear models (e.g., linear regression). Therefore, ensemble methods are generally preferred because they can reduce variance (James et al. 2013; Kuhn and Johnson 2013). This article adopts two popular ensemble methods: random forest (RF) regression and GBR.

Bagging algorithms, also called bootstrap aggregation techniques, build a large number of decorrelated trees using bootstrapping and then average them. Specifically, the bagging process in RF is as follows (Friedman et al. 2001):

1. Draw bootstrapped samples (size N) from the original data set.
2. Grow a regression tree for the bootstrapped samples and a subset of independent variables, and then recursively repeat the tree-growing process until the stopping criterion is reached (i.e., minimum node size).
3. Average all regression trees (size N) while reducing the overall model variance, which is also called bagging.
4. The RF model predicts y given $\boldsymbol{x}_i$:

$$y(\boldsymbol{x}_i) = \hat{f}_{RF}^{N}(x_i) = \frac{1}{N}\sum_{b=1}^{N} T_b(x_i)$$

where x_i = vector of independent variable; $T_b(x_i)$ = single regression tree grown by bootstrapped samples and a subset of variables; and N = total number of regression trees.

Gradient boosting regression uses another tree ensemble technique, known as a boosting algorithm. Although bagging algorithms (i.e., RF) also use multiple trees, boosting algorithms sequentially grow the trees. Specifically, each tree is grown by using information (i.e., poorly fitted observations) from previously grown trees, and different weights are assigned at each step (James et al. 2013). The general boosting process for GBR is as follows:

1. Initially set the number of trees (estimators), N, and number of splits (tree depth), D (stopping criteria).
2. A target (dependent) variable, $\hat{f}(x) = 0$, is initially set to zero, and residual (r_i) and target (dependent) variables (y_i) are assumed to be identical for all observations (i).
3. During the boosting process for each tree estimator (N number of trees), the following steps are repeatedly and sequentially conducted:
   a. Estimate a tree and compute the residual for each observation (computing negative gradient, r);
   b. Fit a regression tree $\hat{f}^b$ to the data $(\boldsymbol{x}, r)$); b denotes a single regression tree;
   c. Compute a new target value, $\hat{f}$, by adding in a regularized new tree:

$$\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x)$$

   d. Update the residuals:

$$r_i \leftarrow r_i - \lambda \hat{f}^b(x_i)$$

4. Sum the sequential trees that predict y given $\boldsymbol{x}$:

$$f_B(\boldsymbol{x}) = \sum_{b=1}^{N} \lambda \hat{f}^b(x)$$

Support Vector Machine

A SVM is a kernel-based method to find the optimal generalization boundaries for fitting y based on X. In fact, when used for regression, SVM inherits some properties from the SVM algorithm used for classification. Specifically, SVM adopts different kernel functions ($\emptyset$) to capture the relationship between X and y:

$$y(\boldsymbol{x}) = \sum_{m=1}^{M} \beta_m \emptyset_m(x) + \beta_0 \quad (8)$$

where $\emptyset$ = kernel function (also called a basis function) with M numbers. To estimate the parameters (β and β_0), the following kernel function is minimized:

$$\begin{aligned} \min \emptyset(\beta, \beta_0) &= \sum_{i=1}^{N} V_{\varepsilon}^{r} + \frac{1}{2}\|\beta\|^2 \\ &= \sum_{i=1}^{N} V(y_i - f(x_i)) + \frac{\lambda}{2}\sum \beta_m^2 \\ &\quad (\text{where } V_{\epsilon}^{r} = 0(\text{if } |r| \le \varepsilon), |r| - \varepsilon, \text{otherwise}) \end{aligned} \quad (9)$$

where V_{ϵ}^{r} measures the general errors from the support vectors selected by the model. The element ε is the threshold to manage the number of support vectors used for finding optimal bound, and λ is called the penalty parameter determining the flexibility of the model. Both ε and λ must be tuned to balance the variance–bias trade-off.

ANN Models: Multilayer Perceptron and Generalized Regression Neural Network

Artificial neural networks algorithmically construct a simplified model of the human brain to explain or infer the relationship between X and y. The multilayer perceptron (MLP) neural network has been widely adopted; it consists of three layers: input, hidden, and output. It is also called a backpropagation neural network (BPNN). In MLP, the hidden layer is able to capture nonlinear relationships between X (input layer) and y (output layer).

For regression problems, the MLP neural network takes input data and computes an output result based on the value of inputs ($\boldsymbol{x}$) and the corresponding weights ($\boldsymbol{w}$) using an internal activation function (f). The activation function is used to transfer inputs to outputs according to the functional form of the activation function. The weights are scaled values associated with the connections between neurons. The notion that MLP predicts y given $\boldsymbol{x}$ can be expressed as

$$y(\boldsymbol{x}, \boldsymbol{w}) = w_0 + f\left(\sum_{i}^{n} w_i \emptyset_i(\boldsymbol{x})\right) = w_0 + f\left(\sum_{i}^{n} w_i x_i\right) \quad (10)$$

where $\boldsymbol{x}$ = input vector; $\boldsymbol{w}$ = vector of associated weights; and $\emptyset_x$ = basis function. Here, the function f takes $\boldsymbol{x}$ as the basis function in the form of a linear combination. To estimate the weights, a backpropagation process is applied to minimize the

loss function (SSE) for the MLP by generally using the gradient descent method (GDM).

A GRNN is a feedforward network that is physically identical to the architecture of a BPNN (i.e., MLP)—three layers consisting of an input layer, a hidden [radial basis function (RBF)] layer, and an output layer. In contrast to the BPNN model, GRNN is formulated by a linear combination of input ($\boldsymbol{x}$) and associated weights ($\boldsymbol{w}$) through a RBF such as a Gaussian density function, g(x). To predict y given $\boldsymbol{x}$, the GRNN can be expressed as

$$y(\boldsymbol{x}, \boldsymbol{w}) = w_0 + f\left(\sum_i^n w_i \mathrm{g}_i(\boldsymbol{x})\right) \tag{11}$$

where $\boldsymbol{w}$ = vector of associated weights between output and hidden layers. The main difference between Eqs. (10) and (11) is the basis function that is changed from a linear to a Gaussian basis function.

Specifically, the basis function g(x) is conceptually obtained by calculating the distance between two vectors based on the Gaussian function whose outputs are inversely proportional to the distance from the mean:

$$\mathrm{g(x)} = \frac{1}{\sigma(2\pi)^{\frac{1}{2}}} \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{c_i}\|^2}{2\sigma^2}\right) \sim \exp(-\beta\|\boldsymbol{x} - \boldsymbol{c_i}\|^2) \tag{12}$$

where $\boldsymbol{x}$ = new input data samples to be classified with n variables; and $\boldsymbol{c}$ = mean of Gaussian distribution, also known as a prototype vector. To be specific, this equation computes the geometric distance between the new input vector and the prototype vector (i.e., mean of Gaussian distribution); thus, the similarity of the input vector and prototype vector is measured.

Model Specification and Evaluation

Variable Scaling

A set of variables in a data set is recorded based on varying scales and ranges. These numerical differences among variables may result in biased estimation, especially for some nonparametric models that are sensitive to scales (e.g., ANN and SVR). Therefore, scaled values (z_i) are preferred for both the independent and dependent variables in all models. Although variables do not need to be scaled for linear models, the same values are used in all models for consistency. In this study, the conventional min-max scaling technique is adopted; it is defined as

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \tag{13}$$

where z_i = scaled value of ith sample; x_i = original value of ith sample; and $\max(x)$ and $\min(x)$ = minimum and maximum value of x.

As mentioned earlier and as is common in data mining, all models are trained on 70% of the data and tested on the remaining 30% of the data. Furthermore, to detect any overfitting issues, a 5-fold CV analysis is conducted. That is, the training set (i.e., 70% of the data) is divided into five partitions, four of the five partitions are used for model training, and the remaining partition is used to evaluate the model. This process gives us five trained models, and the average and standard deviation of the performance of each model are calculated. Each model is then trained again on the full training set and tested against the test set. Although this two-step process adds some redundancy, it offers a statistically robust method to validate the results.

Model Evaluation Metrics

To evaluate the models, three metrics are used: mean absolute error (MAE), mean squared error (MSE), and adjusted r-squared (R^2_{adj}). MAE and MSE are primarily used to measure the deviation between the actual and predicted water consumption values:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{14}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{15}$$

where $\hat{y}_i$ = predicted value for ith household. In addition, R^2_{adj} is calculated to see how close the predicted values are to a fitted line or curve to overcome the limitations of the traditional R^2 indicator. In particular, R^2 increases whenever more independent variables are added; thus, more variables may appear to better fit the data, while this is not necessarily the case. R^2 can also be affected by the "noise" in the data. The traditional R^2 is defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2} \tag{16}$$

where n = number of observations. In contrast, the R^2_{adj} is adjusted by the number of variables in the model, and it can control increases of R^2 (Hastie et al. 2009). R^2_{adj} is therefore lower or equal to R^2, and it is defined as

$$R^2_{adj} = 1 - \frac{(1 - R^2)(n - 1)}{n - k - 1} \tag{17}$$

where k = number of independent variables; and $\bar{y}_i$ = mean of actual values.

Model Results and Discussion

This section contains the model validation results for the REU 2016 scenario first and then for the general scenario.

REU 2016 Scenario: Including 19 Variables

The performance of the 12 statistical models using the 5-fold CV analysis for the REU 2016 scenario is shown in Table 3. The table includes the MAE, MSE, and R^2_{adj} values. The first value in the table is the average performance, and the uncertainty value after the "±" is the standard deviation. All models use normalized data and predict the log-transformed of household water use. All are implemented with the Scikit-learn library built in Python (Pedregosa et al. 2011).

The results show that GBR outperforms the other models with a MAE of 0.098 and R^2_{adj} of 0.69. Furthermore, models containing probabilistic or nonparametric features such as KRR, RBF-SVR, RF, and GRNN perform better than other models. Parametric (linear) models such as ridge, linear-SVM, and linear regression also perform relatively well with a MAE of approximately 0.113 and R^2_{adj} of 0.55. On average, parametric models performed worse than ML models with a drop in R^2_{adj} of about 0.14. In addition to Table 3, Fig. 4 shows error plots comparing the predicted and actual water consumption values. In particular, the predicted values for GBR are more closely scattered to the standard regression line than the other models. Moreover, the scattered values of the other models (especially parametric models) have a slightly lower tangent to the fitted line than GBR—i.e., most models overestimate low consumption values and underestimate high consumption values. For instance, the linear regression model tends to overestimate lower values

Table 3. Cross-validation results of all models using all variables available in REU 2016

Statistical regression technique	MAE^N	MSE^N	R^2_{adj}
GBR	0.098 ± 0.01	0.017 ± 0.01	0.69 ± 0.09
RF regression	0.099 ± 0.02	0.017 ± 0.01	0.64 ± 0.10
RBF-SVM	0.110 ± 0.02	0.018 ± 0.00	0.62 ± 0.08
Bayesian ridge regression (BRR)	0.111 ± 0.02	0.018 ± 0.02	0.61 ± 0.10
GRNN	0.111 ± 0.01	0.019 ± 0.01	0.60 ± 0.11
Kernel ridge regression (KRR)	0.112 ± 0.01	0.021 ± 0.00	0.58 ± 0.07
Ridge regression (ridge)	0.113 ± 0.01	0.021 ± 0.01	0.55 ± 0.07
SVM with linear kernel (linear-SVM)	0.113 ± 0.02	0.021 ± 0.00	0.55 ± 0.07
Linear regression	0.113 ± 0.02	0.021 ± 0.01	0.54 ± 0.07
MLP regression (MLP)	0.115 ± 0.03	0.023 ± 0.02	0.52 ± 0.18
Lasso regression (Lasso)	0.116 ± 0.02	0.025 ± 0.01	0.49 ± 0.05
KNN regression (KNN)	0.119 ± 0.02	0.029 ± 0.01	0.41 ± 0.06

Note: MAE^N, MSE^N, and R^2 are calculated from normalized data.

(approximately bottom 30% in actual water consumption) and underestimate higher values (approximately ranging from the median to below the top 10%). This is partly because linear parametric models have strict assumptions of the error terms and the relations between independent variables. In general, they assume that covariance between independent variables is zero and that the error terms follow normal distributions with a mean of zero. These assumptions are likely to partly account for the poor prediction, especially when the dimensionality of variables is high (i.e., large number of independent variables).

In addition, MLP shows comparatively poor predictive performance because many values are underestimated or overestimated across the entire range of consumption values. Although similar patterns are seen in other models, MLP has proven to be more sensitive. this may be due to the fact that MLP with a backpropagation process generally requires large data sets, preferably at least 10 times larger than the number of weights in the network structure (Anthony and Bartlett 2009; Iyer and Rhinehart 1999). In this study, the MLP model only has 531 observations but 19 independent variables. Furthermore, the MLP model shows the largest variation during cross validation (standard deviation with ± 0.18 in R^2_{adj}), and this implies that it may not be optimized to the global minimum because of the algorithmic characteristics of GDM.

General Scenario: Including Six Variables

Similar to Table 3 for the REU 2016 scenario, the performance of the 12 statistical models for the general scenario are shown in Table 4. The model performances are systematically lower for

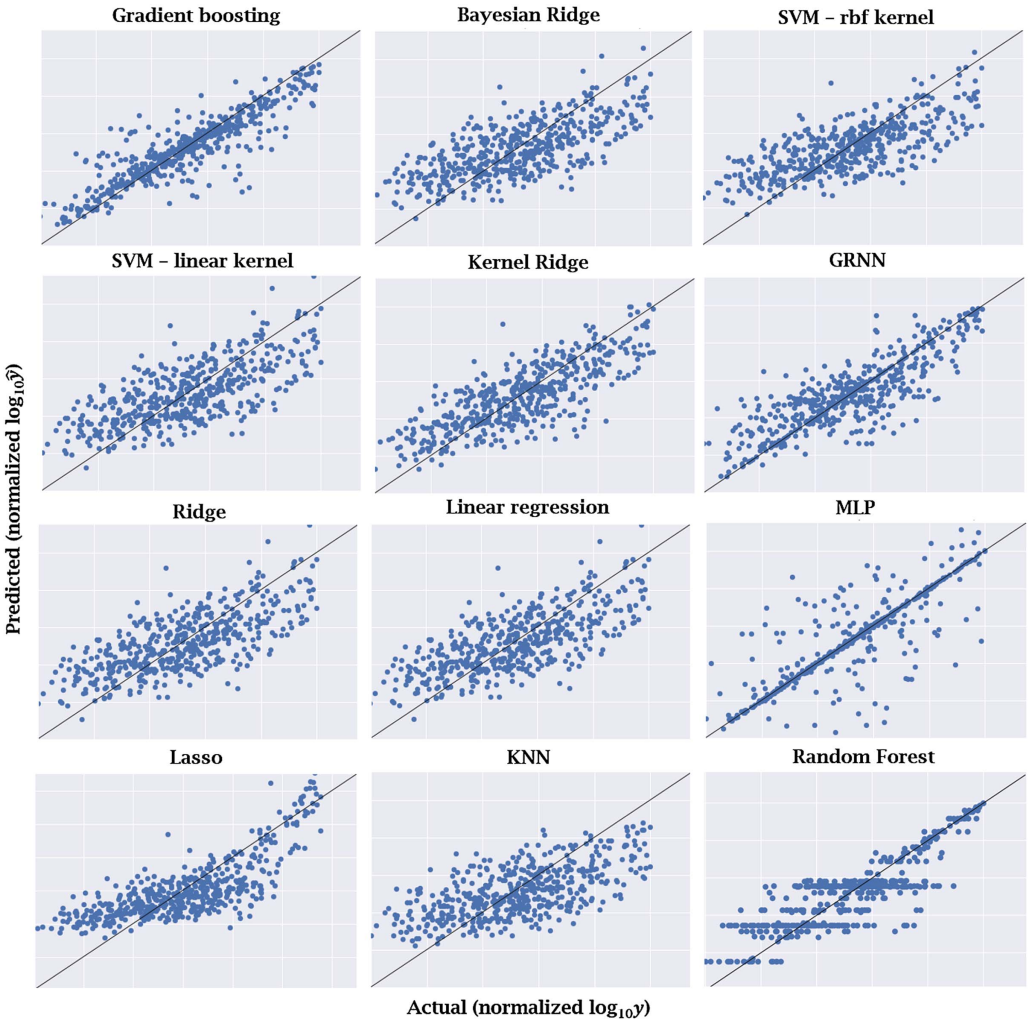

Fig. 4. Comparison of error plots for REU 2016–specific scenario [the horizontal is the logarithm of normalized actual water consumption ($\log_{10} y$) and the vertical axis is that of predicted consumption values].

Table 4. Cross-validation result for general scenario that only includes publicly available variables (six variables)

Statistical regression techniques	MAE^N	MSE^N	R^2_{adj}
GBR	0.128 ± 0.01	0.026 ± 0.01	0.60 ± 0.13
RF regression	0.130 ± 0.01	0.029 ± 0.01	0.51 ± 0.18
GRNN	0.131 ± 0.02	0.030 ± 0.01	0.50 ± 0.11
KRR	0.131 ± 0.02	0.030 ± 0.01	0.49 ± 0.12
MLP regression	0.132 ± 0.02	0.031 ± 0.01	0.48 ± 0.14
SVM (linear-SVM)	0.134 ± 0.02	0.033 ± 0.01	0.44 ± 0.10
KNN	0.135 ± 0.02	0.034 ± 0.01	0.43 ± 0.10
Ridge	0.137 ± 0.02	0.036 ± 0.00	0.39 ± 0.11
Bayesian ridge regression	0.137 ± 0.02	0.036 ± 0.00	0.37 ± 0.11
SVM regression (RBF-SVM)	0.139 ± 0.02	0.037 ± 0.00	0.34 ± 0.11
Linear	0.140 ± 0.02	0.037 ± 0.01	0.33 ± 0.11
Lasso	0.140 ± 0.02	0.037 ± 0.01	0.33 ± 0.08

the general scenario compared to the REU 2016 scenario, which is expected since fewer variables are used. The results also show that the performance gaps between the ML and parametric models increased significantly. For instance, the R^2_{adj} values for linear regression decreased from 0.54 to 0.33 (a 41% decrease) in contrast to the R^2_{adj} values for GBR, which only decreased from 0.69 to 0.60 (a 13% decrease). In addition, the error plots in Fig. 5 demonstrate that GBR shows similar patterns with Fig. 4, and the number of under- and overestimated samples increased only marginally. Consequently, these results suggest that the independent variables can still explain residential end-use water consumption behaviors relatively well. In particular, household type and size and climate factors appear to significantly affect water consumption. The results also show that MLP now performs relatively well in the general scenario, with only a 0.04 decrease in R^2_{adj} and a 0.016 increase in MAE. This is partially because the general scenario includes fewer variables than the REU 2016 scenario, thus requiring the training of fewer weights (which is important, as mentioned earlier).

Finally, in predictive modeling, it is generally desirable to gain an appreciation for the contribution of each variable, i.e., which independent variable contributes the most to explaining the dependent variables? For instance, the top-ranked model, GBR, has the potential to measure the relative importance of each variable, generally referred to as the variable importance (VI). The results of VI can be interpreted as the predictive power of independent variables. Based on the VI in GBR, income, household size, parcel area, the existence of outdoor properties, and climate make a greater contribution to the prediction of household water consumption. This also implies that the variables used in the general scenario are sufficient to provide an acceptable prediction performance.

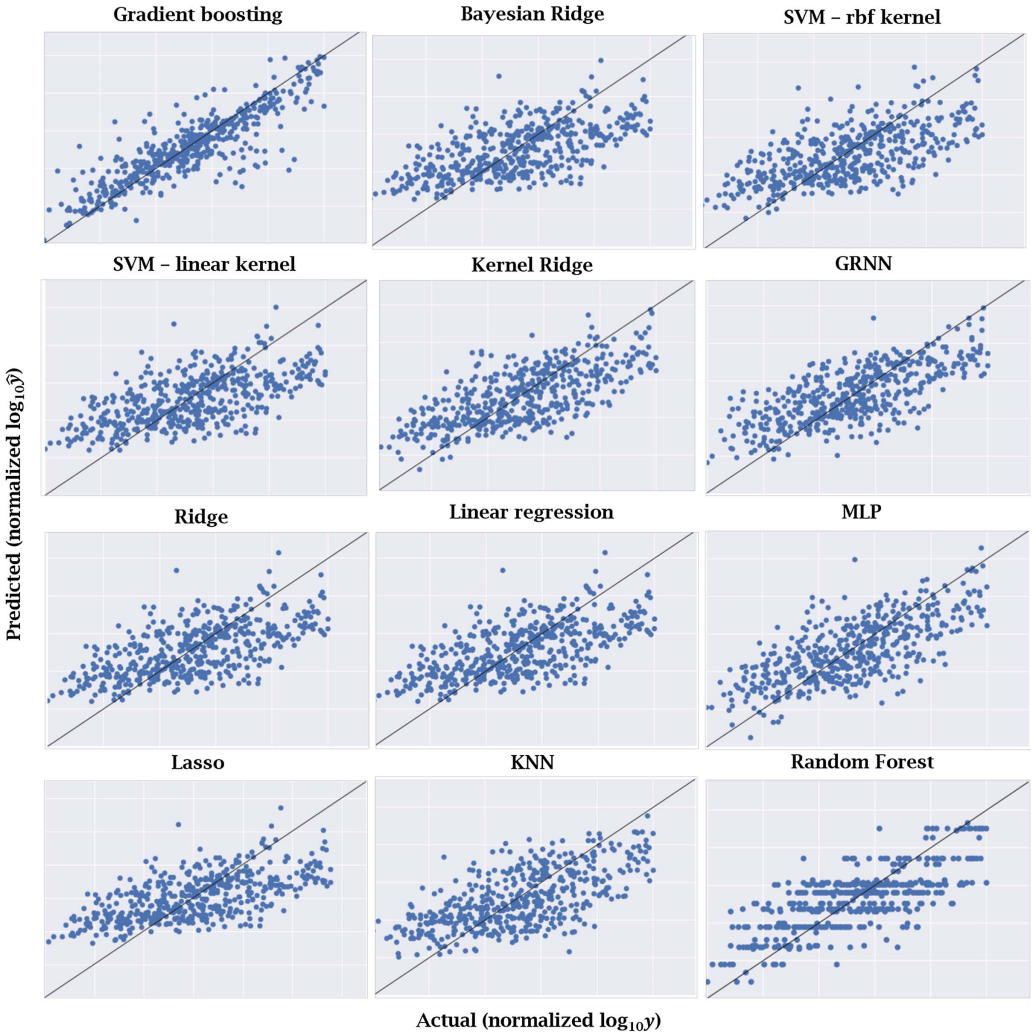

Fig. 5. Comparison of error plots for general scenario.

Technical Discussion

Generally, the results demonstrate that ML models outperform parametric linear models, such as linear regression, which is largely due to the algorithmic differences between the two families of techniques. For instance, GBR, SVM, KRR, RF, and ANN models are primarily designed to capture nonlinear and complex relationships between variables in regression problems. For that, they perform stochastic local optimization (e.g., kernel, boosting, and bagging) rather than single global optimization (Friedman et al. 2001). Thus, they are likely to decrease biases during the estimation process, and they can provide more accurate predictions than linear models.

Nonetheless, this local nonparametric learning process may generate overfitted models that can have high variances. For instance, ANN models using a high-dimensional input data set tend to be overfitted since they have too many weights that need to be optimized. These overfitting issues can be seen in any ML models. The simplest way to mitigate them is to set an early stopping rule that is widely used to interrupt the repetitive learning of a machine. When it comes to algorithmic features, regularization and shrinkage methods are added to the ML models to avoid overfitting (Friedman et al. 2001). Specifically, a regularization term within some models (e.g., GBR, lasso, ridge, and KRR) alleviates problems related to outliers (e.g., high biases) and high-dimensional inputs (e.g., high correlation) by introducing penalties while balancing the trade-off between the variance and bias. In addition to the regularization, GBR and ANN models contain a shrinkage parameter that can also control variances by sacrificing some biases. For example, this is applied to hyperparameters (i.e., weights) that can control skewed variables and outliers. These algorithmic features partly substantiate the performance gaps between the ML and parametric models used in this study, and GBR, in particular, possesses all the features mentioned earlier. In other words, ML models can be more appropriate for predicting residential water use, especially for data such as REU 2016 that inherently exhibit high variance (i.e., detrimental outliers).

In predictive modeling, the applicability of a model is also an important performance criterion, i.e., whether a model can easily be used for other municipalities or utilities. The availability of data on daily water use can vary dramatically across geographical locations, however. For instance, some cities may have access to longer time-series daily water use than the REU 2016 data sets that are only available for 12 days. This longer time-series data may include unobserved heterogeneity across households and other unobserved effects. To enhance future applicability, the modeling approaches discussed here are designed to alleviate these effects by controlling variables (a.k.a., covariates) that can incorporate seasonal and climate-related effects that affect water use. In addition, the top-ranked model, gradient boosting machine (GBM), possesses an algorithmic ability to handle heterogeneity issues, which is initially built in in rule-based models (see details in the "Methodology" section). Due to its nonparametric and rule-based properties, it is algorithmically well suited to handle mixed-type data that often show mixed effects (e.g., heterogeneity across observations) (Friedman 2001; Friedman et al. 2001). In addition to this algorithmic characteristic, GBM includes boosting machine processes that sequentially "forgive" poorly learned samples in a single tree structure by using multiple trees (i.e., estimators). This is one of the main technical reasons why GBM performs best among all modeling methods tested in this study. Therefore, the modeling performance will be consistent when the data get even longer than what is used in this article.

Despite a high degree of predictive power, the general criticism of many machine-based algorithms is their lack of interpretability, unlike parametric models, in which a domain expert can validate the parameters estimated. Several statistical measures exist to address this interpretability issue in ML models (Doshi-Velez and Kim 2017; Samek et al. 2017). Among numerous measures, a model-agnostic approach (i.e., not specific to a particular algorithm) is generally preferred since it can be applied to any predictive modeling process using ML models (Friedman et al. 2001; Lundberg and Lee 2017; Molnar 2018; Ribeiro et al. 2016). For example, and as shown in this article, GBR is able to identify the magnitude of the contribution of each variable by measuring the reduction in the overall error (i.e., bias and variance), called variable importance. Nonetheless, VI cannot represent the sensitivity of variables on dependent variables. In contrast, the marginal effect of an independent variable on the predicted values of a learned model can also be examined with ML models and is generally referred to as partial dependence (Friedman et al. 2001; Natekin and Knoll 2013; Semanjski and Gautama 2015; Lee et al. 2019). Although they are beyond the scope of this article, these interpretable features are straightforward. Not only do they provide valuable insights into the performance of a model, they can also help determine effective policies by assessing the contribution of individual variables and therefore help municipalities and utilities make better long-term and short-term decisions.

Conclusion

This article aimed to test two technical challenges in water demand modeling: determining which modeling technique is most appropriate and determining how much data and what variables are required for learning an acceptable model. Specifically, the performance of 12 statistical learning algorithms, including parametric and nonparametric models, were investigated to model household end-use water demand (i.e., the cross section of average daily water use), while taking into account two data scenarios. For the general scenario, only six variables were intentionally kept because they are commonly available in public micro databases and, thus, are accessible to all cities and water utilities.

The results for the REU 2016 scenario indicate that nonparametric ML models perform better than parametric linear models; specifically, GBR performed best. Furthermore, MLP showed relatively poor accuracy and the largest variation during cross validation, although it reportedly performed well in previous studies. This is likely due to the fact that GDM may have issues with finding optimal solutions while minimizing loss functions when the dimensionality of the input data (i.e., the number of variables) is small. In the general scenario, ML models perform adequately as well despite the data constraints (i.e., six variables), and here again GBR performed best. In contrast to the REU 2016 scenario, the performance gaps between the ML models and the linear models were even wider. In addition, linear models less accurately predicted under- and overestimated samples compared to the REU 2016 scenario.

The findings in this study can fill important technical knowledge gaps in predicting household water demand. Moreover, the study can be useful to municipalities and utilities that can adopt the same techniques (e.g., gradient boosting) on their own data set to predict water demand and to infer the importance of individual variables in their area. To further improve water demand prediction accuracy, future work can focus on simulating data sets that can provide more information with utilities to better capture household and individual water consumption behavior. In addition, as mentioned in the technical discussion, a single metric, such as predictive accuracy, is often not enough to be able to develop effective policies. Instead,

learned models should have both predictive power and be interpretable to take full advantage of the usability and adaptability of models in the future.

Data Availability Statement

The 2016 Residential End Use of Water survey (REU 2016) used in this study can be acquired from the Water Research Foundation (www.waterrf.org), and it is accessible to everyone upon request. The database is provided as a Microsoft Access file. Within the database, the authors used primarily "REU 2016_Daily_Use_Main" and "REU2016_End_Use_Sample," and these are combined with KEYCODES.

The Python codes used in this article were created mainly by using the Scikit-learn library (Pedregosa et al. 2011). Anyone with minimal experience in statistical modeling should be able to use the ML models used in this study easily by using this Python package or most other libraries available in Python and in other computer languages (e.g., R). The Python codes developed for this study are available from the corresponding author upon request.

Acknowledgments

The authors would like to acknowledge David Klawitter and Peter Cairo from the University of Illinois at Chicago for sharing information about REU 2016 data. This research is partly supported by the National Science Foundation (NSF) CAREER Award 155173 and by the NSF Cyber-Physical Systems (CPS) Award 1646395.

References

Adamowski, J., H. Fung Chan, S. O. Prasher, B. Ozga-Zielinski, and A. Sliusarieva. 2012. "Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in Montreal, Canada." *Water Resour. Res.* 48 (1): 1528. https://doi.org/10.1029/2010WR009945.

Ahmad, N., and S. Derrible. 2015. "Evolution of public supply water withdrawal in the USA: A network approach: Water withdrawal in the USA: A network approach." *J. Ind. Ecol.* 19 (2): 321–330. https://doi.org/10.1111/jiec.12266.

Ahmad, N., S. Derrible, and H. Cabezas. 2017. "Using fisher information to assess stability in the performance of public transportation systems." *R. Soc. Open Sci.* 4 (4): 160920. https://doi.org/10.1098/rsos.160920.

Ahmad, N., S. Derrible, T. Eason, and H. Cabezas. 2016. "Using fisher information to track stability in multivariate systems." *R. Soc. Open Sci.* 3 (11): 160582. https://doi.org/10.1098/rsos.160582.

Akbarzadeh, M., S. Memarmontazerin, S. Derrible, and S. F. Salehi Reihani. 2019. "The role of travel demand and network centrality on the connectivity and resilience of an urban street system." *Transportation* 46 (4): 1127–1141. https://doi.org/10.1007/s11116-017-9814-y.

Altunkaynak, A., and T. A. Nigussie. 2017. "Monthly water consumption prediction using season algorithm and wavelet transform-based models." *J. Water Resour. Plann. Manage.* 143 (6): 04017011. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000761.

Al-Zahrani, M. A., and A. Abo-Monasar. 2015. "Urban residential water demand prediction based on artificial neural networks and time series models." *Water Resour. Manage.* 29 (10): 3651–3662. https://doi.org/10.1007/s11269-015-1021-z.

Anthony, M., and P. L. Bartlett. 2009. *Neural network learning: Theoretical foundations*. Cambridge: Cambridge University Press.

Arbues, F., and I. Villanua. 2006. "Potential for pricing policies in water resource management: Estimation of urban residential water demand in Zaragoza, Spain." *Urban Stud.* 43 (13): 2421–2442. https://doi.org/10.1080/00420980601038255.

Arbues, F., I. Villanúa, and R. Barberán. 2010. "Household size and residential water demand: An empirical approach." *Aust. J. Agric. Resour. Econ.* 54 (1): 61–80. https://doi.org/10.1111/j.1467-8489.2009.00479.x.

Bai, Y., P. Wang, C. Li, J. Xie, and Y. Wang. 2014. "Dynamic forecast of daily urban water consumption using a variable-structure support vector regression model." *J. Water Resour. Plann. Manage.* 141 (3): 04014058. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000457.

Bansal, A., S. K. Rompikuntla, J. Gopinadhan, A. Kaur, and Z. A. Kazi. 2015. "Energy consumption forecasting for smart meters." Preprint submitted December 18, 2015. http://arxiv.org/abs/1512.05979.

Bishop, C. M. 2006. *Pattern recognition and machine learning*. New York: Springer.

Breiman, L., J. Friedman, C. J. Stone, and R. A. Olshen. 1984. *Classification and regression trees*. Boca Raton, FL: CRC Press.

Brentan, B. M., E. Luvizotto Jr., M. Herrera, J. Izquierdo, and R. Pérez-García. 2017. "Hybrid regression model for near real-time urban water demand forecasting." *J. Comput. Appl. Math.* 309 (Jan): 532–541. https://doi.org/10.1016/j.cam.2016.02.009.

Cominola, A., E. S. Spang, M. Giuliani, A. Castelletti, J. R. Lund, and F. J. Loge. 2018. "Segmentation analysis of residential water-electricity demand for customized demand-side management programs." *J. Cleaner Prod.* 172 (Jan): 1607–1619. https://doi.org/10.1016/j.jclepro.2017.10.203.

De'ath, G. 2002. "Multivariate regression trees: A new technique for modeling species-environment relationships." *Ecology* 83 (4): 1105–1117. https://doi.org/10.1890/0012-9658(2002)083[1105:MRTANT]2.0.CO;2.

DeOreo, W. B., P. W. Mayer, B. Dziegielewski, and J. Kiefer. 2016. *Residential end uses of water, version 2*. Denver: Water Research Foundation.

Derrible, S. 2016. "Urban infrastructure is not a tree: Integrating and decentralizing urban infrastructure systems." *Environ. Plann. B: Plann. Des.* 44 (3): 553–569. https://doi.org/10.1177/0265813516647063.

Derrible, S. 2017. "Complexity in future cities: The rise of networked infrastructure." Supplement, *Int. J. Urban Sci.* 21 (S1): 68–86.

Derrible, S. 2018. "An approach to designing sustainable urban infrastructure." *MRS Energy Sustainability* 5: E15. https://doi.org/10.1557/mre.2018.14.

Derrible, S., and N. Ahmad. 2015. "Network-based and binless frequency analyses." *PLoS One* 10 (11): e0142108. https://doi.org/10.1371/journal.pone.0142108.

Domene, E., and D. Saurí. 2006. "Urbanisation and water consumption: Influencing factors in the metropolitan region of Barcelona." *Urban Stud.* 43 (9): 1605–1623. https://doi.org/10.1080/00420980600749969.

Donkor, E. A., T. A. Mazzuchi, R. Soyer, and J. Alan Roberson. 2012. "Urban water demand forecasting: Review of methods and models." *J. Water Resour. Plann. Manage.* 140 (2): 146–159. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000314.

Doshi-Velez, F., and B. Kim. 2017. "Towards a rigorous science of interpretable machine learning." Preprint, submitted April 28, 2017. http://arxiv.org/abs/1702.08608.

Elith, J., J. R. Leathwick, and T. Hastie. 2008. "A working guide to boosted regression trees." *J. Anim Ecol.* 77 (4): 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x.

Farooq, B., M. Bierlaire, R. Hurtubia, and G. Flötteröd. 2013. "Simulation based population synthesis." *Transp. Res. Part B: Methodol.* 58 (Dec): 243–263. https://doi.org/10.1016/j.trb.2013.09.012.

Firat, M., M. E. Turan, and M. A. Yurdusev. 2010. "Comparative analysis of neural network techniques for predicting water consumption time series." *J. Hydrol.* 384 (1–2): 46–51. https://doi.org/10.1016/j.jhydrol.2010.01.005.

Firat, M., M. A. Yurdusev, and M. E. Turan. 2009. "Evaluation of artificial neural network techniques for municipal water consumption modeling." *Water Resour. Manage.* 23 (4): 617–632. https://doi.org/10.1007/s11269-008-9291-3.

Fricke, K. 2014. *Analysis and modelling of water supply and demand under climate change, land use transformation and socio-economic development*. Cham, Switzerland: Springer Theses, Springer International Publishing.

Friedman, J. 2001. "Greedy function approximation: A gradient boosting machine." *Ann. Stat.* 29 (5): 1189–1232. https://doi.org/10.1214/aos/1013203451.

Friedman, J., T. Hastie, and R. Tibshirani. 2001. *The elements of statistical learning*. New York: Springer.

Froukh, M. L. 2001. "Decision-support system for domestic water demand forecasting and management." *Water Resour. Manage.* 15 (6): 363–382. https://doi.org/10.1023/A:1015527117823.

Ghimire, M., T. A. Boyer, C. Chung, and J. Q. Moss. 2015. "Estimation of residential water demand under uniform volumetric water pricing." *J. Water Resour. Plann. Manage.* 142 (2): 04015054. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000580.

Golshani, N., R. Shabanpour, S. M. Mahmoudifard, S. Derrible, and A. Mohammadian. 2018. "Modeling travel mode and timing decisions: Comparison of artificial neural networks and copula-based joint model." *Travel Behav. Soc.* 10 (Jan): 21–32. https://doi.org/10.1016/j.tbs.2017.09.003.

Goodchild, C. 2003. "Modelling the impact of climate change on domestic water demand." *Water Environ. J.* 17 (1): 8–12. https://doi.org/10.1111/j.1747-6593.2003.tb00423.x.

Grafton, R. Q., M. B. Ward, H. To, and T. Kompas. 2011. "Determinants of residential water consumption: Evidence and analysis from a 10-country household survey: Determinants of residential water consumption." *Water Resour. Res.* 47 (8): W08537. https://doi.org/10.1029/2010WR009685.

Guhathakurta, S., and P. Gober. 2007. "The impact of the phoenix urban heat island on residential water use." *J. Am. Plann. Assoc.* 73 (3): 317–329. https://doi.org/10.1080/01944360708977980.

Guo, J. Y., and C. R. Bhat. 2007. "Population synthesis for microsimulating travel behavior." *Transp. Res. Rec.* 2014 (1): 92–101. https://doi.org/10.3141/2014-12.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. "Linear methods for regression." In *The elements of statistical learning*, 1–57. New York: Springer.

Hensher, D. A., J. M. Rose, and W. H. Greene. 2005. *Applied choice analysis: A primer*. Cambridge: Cambridge University Press.

House-Peters, L., and H. Chang. 2011. "Urban water demand modeling: Review of concepts, methods, and organizing principles." *Water Resour. Res.* 47 (5): W05401. https://doi.org/10.1029/2010WR009624.

House-Peters, L., B. Pratt, and H. Chang. 2010. "Effects of urban spatial structure, sociodemographics, and climate on residential water consumption in Hillsboro, Oregon." *J. Am. Water Resour. Assoc.* 46 (3): 461–472. https://doi.org/10.1111/j.1752-1688.2009.00415.x.

Iyer, M. S., and R. R. Rhinehart. 1999. "A method to determine the required number of neural-network training repetitions." *IEEE Trans. Neural Networks* 10 (2): 427–432. https://doi.org/10.1109/72.750573.

James, G., D. Witten, T. Hastie, and R. Tibshirani. 2013. *An introduction to statistical learning*. New York: Springer.

Jentgen, L., H. Kidder, R. Hill, and S. Conrad. 2007. "Energy management strategies use short-term water consumption forecasting to minimize cost of pumping operations." *Am. Water Works Assoc. J.* 99 (6): 86–94. https://doi.org/10.1002/j.1551-8833.2007.tb07957.x.

Keene, O. N. 1995. "The log transformation is special." *Stat. Med.* 14 (8): 811–819. https://doi.org/10.1002/sim.4780140810.

Kenney, D. S., C. Goemans, R. Klein, J. Lowrey, and K. Reidy. 2008. "Residential water demand management: Lessons from Aurora, Colorado." *J. Am. Water Res. Assoc.* 44 (1): 192–207. https://doi.org/10.1111/j.1752-1688.2007.00147.x.

Kohavi, R. 1995. "A study of cross-validation and bootstrap for accuracy estimation and model selection." In *Proc., 14th Int. Joint Conf. on Artificial Intelligence, IJCAI'95*, 1137–1143. San Francisco: Morgan Kaufmann Publishers.

Kontokosta, C. E., and R. K. Jain. 2015. "Modeling the determinants of large-scale building water use: Implications for data-driven urban sustainability policy." *Sustainable Cities Soc.* 18 (Nov): 44–55. https://doi.org/10.1016/j.scs.2015.05.007.

Kuhn, M., and K. Johnson. 2013. *Applied predictive modeling*. New York: Springer.

Kusiak, A., M. Li, and Z. Zhang. 2010. "A data-driven approach for steam load prediction in buildings." *Appl. Energy* 87 (3): 925–933. https://doi.org/10.1016/j.apenergy.2009.09.004.

Lee, D., S. Derrible, and F. C. Pereira. 2018. "Comparison of four types of artificial neural networks and a multinomial logit model for travel mode choice modeling." *Transp. Res. Rec.* 2672 (49): 101–112. https://doi.org/10.1177/0361198118796971.

Lee, D., J. Mulrow, C. J. Haboucha, S. Derrible, and Y. Shiftan. 2019. *Attitudes on autonomous vehicle adoption using interpretable gradient boosting machine*. Washington, DC: Transportation Research Record.

Lee, S.-J., H. Chang, and P. Gober. 2015. "Space and time dynamics of urban water demand in Portland, Oregon and Phoenix, Arizona." *Stochastic Environ. Res. Risk Assess.* 29 (4): 1135–1147. https://doi.org/10.1007/s00477-014-1015-z.

Lee, S.-J., E. A. Wentz, and P. Gober. 2010. "Space-time forecasting using soft geostatistics: A case study in forecasting municipal water demand for Phoenix, Arizona." *Stochastic Environ. Res. Risk Assess.* 24 (2): 283–295. https://doi.org/10.1007/s00477-009-0317-z.

Lozano, S., and E. Gutiérrez. 2008. "Non-parametric frontier approach to modelling the relationships among population, GDP, energy consumption and CO_2 emissions." *Ecol. Econ.* 66 (4): 687–699. https://doi.org/10.1016/j.ecolecon.2007.11.003.

Lundberg, S., and S.-I. Lee. 2017. "A unified approach to interpreting model predictions." In *Advances in neural information processing systems*, 4765–4774. Red Hook, NY: Curran Associates, Inc.

Mayer, P. W., W. B. DeOreo, E. M. Opitz, J. C. Kiefer, W. Y. Davis, B. Dziegielewski, and J. O. Nelson. 1999. *Residential end uses of water*. Denver: Water Research Foundation.

Mazzanti, M., and A. Montini. 2006. "The determinants of residential water demand: Empirical evidence for a panel of Italian municipalities." *Appl. Econ. Lett.* 13 (2): 107–111. https://doi.org/10.1080/13504850500390788.

Molnar, C. 2018 "Interpretable machine learning." In *A guide for making black box models explainable*. San Francisco: GitHub.

Natekin, A., and A. Knoll. 2013. "Gradient boosting machines, a tutorial." *Front. Neurorob.* 7: 21 https://doi.org/10.3389/fnbot.2013.00021.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, and V. Dubourg. 2011. "Scikit-learn: Machine learning in Python." *J. Mach. Learn. Res.* 12 (Oct): 2825–2830.

Ribeiro, M. T., S. Singh, and C. Guestrin. 2016. "Model-agnostic interpretability of machine learning." Preprint, submitted June 16, 2016. http://arxiv.org/abs/1606.05386.

Robinson, C., B. Dilkina, J. Hubbs, W. Zhang, S. Guhathakurta, M. A. Brown, and R. M. Pendyala. 2017. "Machine learning approaches for estimating commercial building energy consumption." Supplement *Appl. Energy* 208 (SC): 889–904. https://doi.org/10.1016/j.apenergy.2017.09.060.

Rosca, M., B. Lakshminarayanan, and S. Mohamed. 2018. "Distribution matching in variational inference." Preprint submitted February 19, 2018. http://arxiv.org/abs/1802.06847[cs,stat].

Samek, W., T. Wiegand, and K.-R. Müller. 2017. "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models." Preprint submitted August 28, 2017. http://arxiv.org/abs/1708.08296.

Schleich, J., and T. Hillenbrand. 2009. "Determinants of residential water demand in Germany." *Ecol. Econ.* 68 (6): 1756–1769. https://doi.org/10.1016/j.ecolecon.2008.11.012.

Semanjski, I., and S. Gautama. 2015. "Smart city mobility application: Gradient boosting trees for mobility prediction and analysis based on crowdsourced data." *Sensors* 15 (7): 15974–15987. https://doi.org/10.3390/s150715974.

Tso, G. K. F., and K. K. W. Yau. 2007. "Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks." *Energy* 32 (9): 1761–1768. https://doi.org/10.1016/j.energy.2006.11.010.

Vitter, J. S., and M. E. Webber. 2018. "A non-intrusive approach for classifying residential water events using coincident electricity data." *Environ. Modell. Software* 100 (Feb): 302–313. https://doi.org/10.1016/j.envsoft.2017.11.029.

Willis, R. M., R. A. Stewart, K. Panuwatwanich, P. R. Williams, and A. L. Hollingsworth. 2011. "Quantifying the influence of environmental and water conservation attitudes on household end use water consumption." *J. Environ. Manage.* 92 (8): 1996–2009. https://doi.org/10.1016/j.jenvman.2011.03.023.

Wisetjindawat, W., S. Derrible, and A. Kermanshah. 2018. "Modeling the effectiveness of infrastructure and travel demand management measures to improve traffic congestion during typhoons." *Transp. Res. Rec.* 2672 (1): 43–53. https://doi.org/10.1177/0361198118791909.

Wooldridge, J. M. 2010. *Econometric analysis of cross section and panel data*. Cambridge, MA: MIT Press.

Yurdusev, M. A., M. Firat, and M. E. Turan. 2010. "Generalized regression neural networks for municipal water consumption prediction." *J. Stat. Comput. Simul.* 80 (4): 477–478. https://doi.org/10.1080/00949650903520118.

Zhou, Z.-H. 2012. *Ensemble methods: Foundations and algorithms*. London: CRC Press.