Improving the Interpretation of Data-Driven Water Consumption Models via the Use of Social Norms

Renee Obringer, Ph.D.[1]; Roshanak Nateghi, Ph.D.[2]; Zhao Ma, Ph.D.[3]; and Rohini Kumar, Ph.D.[4]

Abstract: Water is essential to improving social equity, promoting just economic development and protecting the function of the Earth system. It is therefore important to have access to credible models of water consumption, so as to ensure that water utilities can adequately supply water to meet the growing demand. Within the literature, there are a variety of models, but often these models evaluate the water consumption at aggregate scales (e.g., city or regional), thus overlooking intra-city differences. Conversely, the models that evaluate intra-city differences tend to rely heavily on one or two sources of quantitative data (e.g., climate variables or demographics), potentially missing key cultural aspects that may act as confounding factors in quantitative models. Here, we present a novel mixed-methods approach to predict intra-city residential water consumption patterns by integrating climate and demographic data, and by incorporating social norm data to aid the interpretation of model results. Using Indianapolis, Indiana as a test case, we show the value in adopting a more integrative approach to modeling residential water consumption. In particular, we leverage qualitative interview data to interpret the results from a predictive model based on a state-of-the-art machine learning algorithm. This integrative approach provides community-specific interpretations of model results that would otherwise not be observed by considering demographics alone. Ultimately, the results demonstrate the value and importance of such approaches when working on complex problems. DOI: 10.1061/(ASCE)WR.1943-5452.0001611.

Author keywords: Urban water consumption; Mixed-methods study design; Statistical learning theory; Socio-environmental systems; Socio-hydrological systems.

Introduction

Globally, cities are facing pressing challenges, including rapid urbanization and intensifying climate change. These challenges will likely undermine urban water management around the world. Traditionally, urban water management has been supply-focused, meaning that cities have focused on increasing their water supply to meet the demand (Gleick 2003). Recently, however, many cities have started to integrate demand management with their existing policies (Mitchell 2006; Luo et al. 2015; Wang et al. 2020). As droughts become more frequent and intense (Dai 2011; AghaKouchak et al. 2015; Bruss et al. 2019; Ault 2020), it is likely that these integrated management strategies will become increasingly important in many regions. To ensure that integrated strategies can be implemented successfully, there is a need to better understand the nuances of residential water consumption, particularly within cities.

Many water consumption models aimed at characterizing intra-city water consumption focus on demographics and housing characteristics, which are often correlated with water consumption. For example, in a case study conducted in Reno, Nevada, Viñoles et al. (2015) found that length of residency was associated with higher water consumption. The authors reasoned that this was due to the growth in physical and social capital that participants experienced the longer they lived in the area (Viñoles et al. 2015). A similar study, conducted in Phoenix, Arizona, found that longer residence times increased water consumption and that longer-term residents were more likely to believe in the idea that Phoenix is an oasis (Harlan et al. 2009). Interestingly, a recent study conducted in Southern California found that water consumption decreased as residents occupied a house for longer periods of time (Bolorinos et al. 2020), which suggests that there may be significant differences between areas, possibly based on localized norms. In addition to length of residency, a number of studies have demonstrated an increase in water consumption as household income increases (Harlan et al. 2009; Shandas and Parandvash 2010; Ghavidelfar et al. 2017). Many of these studies attributed this to the larger homes and lot sizes that are often associated with higher incomes. Recently, Cominola et al. (2018) adopted a segmentation approach to classify the water and electricity demand profiles in Los Angeles, California. The authors found that increased water consumption was driven by large house sizes, more occupants, and intensive outdoor uses (Cominola et al. 2018), echoing previous work that leveraged linear models. Going beyond income and housing characteristics, a recent study focused on the impact of the COVID-19 pandemic on water consumption (Li et al. 2021). The authors found that while California's urban water consumption decreased while residents

[1]Assistant Professor, Dept. of Energy and Mineral Engineering, Pennsylvania State Univ., Hosler Bldg., University Park, PA 16802; The National Socio-Environmental Synthesis Center, Univ. of Maryland, 1 Park Place, Annapolis, MD 21401; Environmental and Ecological Engineering, Purdue Univ., 500 Central Dr., West Lafayette, IN 47907 (corresponding author). ORCID: https://orcid.org/0000-0002-4471-4131. Email: obringer@psu.edu

[2]Associate Professor of Industrial Engineering and the Director of the Laboratory for Advancing Sustainable Critical Infrastructure, School of Industrial Engineering, Purdue Univ., 315 N. Grant St., West Lafayette, IN 47907. ORCID: https://orcid.org/0000-0003-4569-9233

[3]Professor, Dept. of Forestry and Natural Resources, Purdue Univ., 195 Marsteller St., West Lafayette, IN 47909. ORCID: https://orcid.org/0000-0002-9103-3996

[4]Senior Scientist, Dept. of Computational Hydrosystems, Helmholtz Centre for Environmental Research—UFZ, Permoserstr.15, Leipzig 04318, Germany. ORCID: https://orcid.org/0000-0002-4396-2037

Note. This manuscript was submitted on March 12, 2021; approved on June 28, 2022; published online on September 27, 2022. Discussion period open until February 27, 2023; 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.

were working from home, the residential sector experienced an increase in consumption (Li et al. 2021). This suggests that remote work (one of the demographics included in the present study) may impact water consumption, although it is not as common of a predictor as income or housing characteristics. These studies highlight the need to consider a wide array of demographic variables when modeling water consumption.

In addition to the work on demographics-based models, there is a significant body of literature dedicated to understanding the role that social norms play in determining water consumption. For example, several studies have demonstrated that environmental awareness often leads to increased water conservation, especially if there are social norms that encourage pro-environmental behaviors (Pinto et al. 2011; Willis et al. 2011; Beal et al. 2016; Ramsey et al. 2017). However, the norms that influence water conservation may vary within a given city or lead to different reactions from different segments of a population. For example, Bhanot (2017) found that competitive messaging about water consumption among neighbors led to lower consumption for those that were already among the more efficient users, while the households that consumed more water (and thus had a lower rank) were likely to increase their water consumption when presented with competitive messaging. The author attributed this to the "last place effect" and subsequent demotivation to work towards reducing water use (Bhanot 2017). In a similar study, Cominola et al. (2021) used smart meter data to evaluate the impact of getting feedback on water consumption via a mobile application, including comparisons with peer households. The authors found that households that had access to the application reduced their water consumption in both the short- and long-term (Cominola et al. 2021). Social influences can also arise through media consumption. Quesnel and Ajami (2017) found that increased news media coverage and subsequent internet searches led to a reduction in water consumption during an extreme drought. Overall, these studies demonstrate the importance of social norms in explaining trends in water consumption in the context of groups (i.e., segments of populations), as well as the benefit of integrating these norms within local water management policies.

Going beyond demographics and social norms, some studies consider the climate impacts on water consumption. For example, Ashoori et al. (2016) found that residential water consumption in Los Angeles was sensitive to the climate, namely, temperature and precipitation. Specifically, Ashoori et al. (2016) found that precipitation was especially significant in predicting water consumption single family homes, with more precipitation leading to lower water consumption. This was tied to the increased outdoor water use common to single family homes in the area. Using higher resolution data, Balling et al. (2008) further demonstrated that climate played a role in determining water consumption at the census tract level. The authors found that water consumption increased across the city during periods of higher-than-normal temperatures, as well as lower-than-normal precipitation (Balling et al. 2008). In a study conducted in San Francisco, California, the authors found that higher temperatures were correlated with higher water consumption across income levels, while precipitation was found to be insignificant (Quesnel and Ajami 2017). Finally, a recent study used future climate scenarios to predict water consumption (Rasifaghihi et al. 2020). Rasifaghihi et al. (2020) found that future increases in temperature and changes in precipitation patterns are likely to lead to increased seasonal water consumption, while base water consumption is likely to remain constant. Many of these studies considered only temperature and precipitation as climatic variables and often used linear regression as the main modeling technique (Balling et al. 2008; Ashoori et al. 2016; Rasifaghihi et al. 2020). In fact, in a seminal review by House-Peters and Chang (2011), the authors found that studies focused on the climate impact on water consumption primarily focused on precipitation and temperature, with only a few studies considering wind speed and evapotranspiration. Recent work, however, has provided evidence that additional climatic variables, such as relative humidity and dew point temperature, are needed to accurately model water consumption (Obringer et al. 2020a). Moreover, it has been shown that the relationship between climate and water consumption is nonlinear (Obringer et al. 2019, 2020a; Wongso et al. 2020), which calls for more complex modeling techniques than linear regression.

Much of the previously discussed literature has focused on evaluating the impact of a single data type (e.g., demographics or climate variables). That being said, there have been a number of models aimed at integrating multiple data types into a single analysis of urban water consumption. For example, Ashoori et al. (2016) included price and population in addition to the climate variables to model water consumption in Los Angeles. However, this study focused on sector-level water use (e.g., single-family residential or commercial), and did not account for the intra-city differences beyond housing type. In this sense, the study may have overlooked potential cultural or demographic indicators of water consumption. Several studies have integrated climate variables and demographics to characterize the climate sensitivity of intra-city water consumption. House-Peters et al. (2010), for example, found that the census blocks with newer homes and more educated residents tended to be more sensitive to climatic conditions (e.g., drought). Similar results were found in a study by Balling et al. (2008), which demonstrated that the water consumption in census tracts with large lots and higher income was more sensitive to the climate. These studies primarily relied on linear models and did not account for the cultural aspects that may also have been driving water consumption. Finally, Ramsey et al. (2017) evaluated the impact of demographics and social norms on water consumption in India, but did not account for the various climate influences that may shape water consumption. Given that water consumption is multi-faceted, with a number of different influences, including demographics, climate, and social norms, it is important to build predictive models that consider a variety of input variables, including the socio-demographic characteristics of the population and other social variables when available. Moreover, models that account for the intra-city water consumption patterns will enable practitioners to develop community-specific plans to curb water consumption.

Here, we present a data-driven model to predict intra-city residential water consumption, accounting for the variability in demographics and climate, while leveraging social norms to aid the interpretation of the model results. In particular, we combine demographic variables measured by the census, such as education level and household income, with high resolution climate data, including precipitation, temperature, and relative humidity. These quantitative variables are used to create a data-driven model of water consumption, the results of which are then interpreted through the novel incorporation of qualitative social norms data. This work advances the growing body of work surrounding the use of data-driven models within water resources research by (1) focusing on a non-linear modeling technique, which has been shown to be effective in other scenarios (Obringer and Nateghi 2018; Wongso et al. 2020), (2) expanding the included climate variables beyond precipitation and temperature, which better captures consumption trends (Obringer et al. 2019), and (3) emphasizing intra-city patterns, which are underexplored in comparison to larger, sector-level studies. Additionally, the use of social norms as a tool for improving interpretation of model results aids in bridging the gap between quantitative and qualitative work. The integration of quantitative and qualitative data is a key aspect of socio-environmental systems

research, particularly as qualitative data can be used to provide insight to quantitative measurements (Elsawah et al. 2020). The data-driven model is based on observational data, but can be used to predict water consumption at the census tract level, assuming the demographics and climate input data are available. We then expand this quantitative model by leveraging qualitative data on social norms to provide insights to the model results that cannot be inferred from demographics or climate data alone. This study aims to increase the scientific understanding of the driving factors behind urban water consumption, allowing water utilities to implement community-specific demand management techniques. In the following sections, we first discuss the data and methods used within this study. Then, we delve into the results and discussion. Finally, we conclude with a summary of the implications for practitioners.

Data and Methods

Site Description

This study considered the city of Indianapolis, Indiana as a test case to demonstrate the value of the integrative approach. Indianapolis is a Midwestern city, which generally experiences mild spring and autumn months, with more extreme summer and winter months. Most of the region is considered to have a temperate climate (the Köppen classification is humid continental). In terms of urban form, similar to many Midwestern cities, Indianapolis is more sprawling than in other areas of the country (Ewing and Hamidi 2014; Hamidi et al. 2015). Finally, the region is considered to be water-rich, based on the number of water resources, both on the surface and below ground. These conditions often lead to higher water consumption (Harlan et al. 2009; Shandas and Parandvash 2010; Ghavidelfar et al. 2017). That being said, there is still interest in enacting demand management strategies within the city, so as to reduce the load on existing infrastructure and ensure adequate water supply in the event of a drought. Little work within the urban water demand management literature has focused on the Midwest, likely owing to the large availability of water resources when compared to other locations, such as the Southwestern United States. Nonetheless, the region has experienced some significant droughts in the recent past (Basara et al. 2019), which have encouraged the water utility company to focus more intently on demand management and drought preparation. The lack of previous work in this region, paired with the past experiences with drought, make Indianapolis an ideal case study for testing the novel integrative approach to modeling urban water consumption.

Data Description

There were four main categories of data collected for this study: (1) water consumption data; (2) demographic data; (3) climate data; and (4) social norms data.

The residential water consumption data served as the response variable (i.e., dependent variable) in the analysis. The data were obtained from the Indianapolis water utility based on monthly metering. To protect consumer privacy, the consumption values were aggregated to the census tract level, such that each data point represented the total water consumption within each census tract for each month in 2018.

The demographic data were obtained from the 2018 American Community Survey conducted by the US Census Bureau (US Census Bureau 2018). These data contained 72 variables obtained directly from the Census Bureau, without pre-selection, which are outlined in Table 1. It should be noted that most variables listed in Table 1 had multiple levels, which were considered to be separate variables for the analysis. For example, the birth rate variable had three levels based on different age groups (i.e., the birth rate for the population below 18 years old, 19–34 years old, and above 35 years old), which were used as predictors (i.e., independent variables) in the study. To limit bias within the modeling framework, each of the 72 demographic variables (Table 1) were considered in the initial analysis. This ensured that we were not inserting any bias into the model by only selecting certain variables from the start. Later, we implemented a process to iteratively remove variables to reduce complexity (see "Methods" section). Each variable was obtained for each census tract (defined by the 2010 Census) in the city, each of which has an average population of 4,000.

The climate data were obtained from the PRISM Climate Group through the Northwest Alliance for Computational Science and Computing (Prism Climate Group 2018). The data are available at a 4-km spatial scale and include precipitation, dry bulb temperature (i.e., the ambient air temperature), dew point temperature (i.e., the temperature at which the air is fully saturated with water vapor), and vapor pressure deficit, which was used to calculate relative humidity. In particular, we considered the total precipitation (mm), the minimum, average, and maximum dry bulb temperature (°C),

Table 1. Demographic variables from the 2018 American Community Survey considered in this study

Variable category	Description
Birth rate	Birth rate separated by age group
Education level	Percent of the population that has achieved various levels of education
Income level	Percent of the population with various levels of household income
Household unit type	Percent of the population that belongs to various types of households (e.g., single person, married couple, family with kids, etc.)
House type	Percent of population that resides is various types of houses (e.g., detached, attached, mobile, etc.)
House value	Percent of population that resides in houses of various values
Language	Percent of population that speaks various languages at home
Marital status	Percent of population that identifies as various martial statuses (e.g., married, divorced, single, etc.)
Place of birth	Percent of population that was born outside of the US, separated by continent
Age	Percent of population in various age groups
Race	Percent of population with various racial identities
Poverty rate	Poverty rate
Work commute	Percent of population that uses various modes of transportation to get to work (e.g., car, bus, work from home, etc.)

Source: Data from US Census Bureau (2018).
Note: There were 72 demographic variables in the study, which are presented here as categories. Under birth rate, for example, the variables included the birth rate for the 15–19 age range, 20–34 age range, and so on, as defined by the US Census.

the average dew point temperature (°C), and the average relative humidity (%). These variables were selected based on previous work on water consumption (Ashoori et al. 2016; Obringer et al. 2019). The climate variables were subject to the same variable selection process as the demographic variables (see "Methods" section). Although the climate data are not directly linked to the census tracts, the 4-km resolution is high enough to conduct an intra-city analysis on the impact of climate variables on water consumption. To transform the climate data to the same scale as the census tracts, we used a nearest-neighbor approach to associate each 4-km grid point from the PRISM data with a census tract. Thus, each census tract had one value for each climatic variable, corresponding to the nearest grid point from the 4-km PRISM dataset.

Finally, the social norms data were collected via semi-structured interviews conducted in Indianapolis. These interviews focused on assessing resident awareness of water conservation programs, as well as the expectations they and others held regarding water conservation (see the Supplemental Methods for interview questions). The existence of a social norm (as opposed to a personal value) was determined by asking three styles of questions, defined by Bicchieri (2016): (1) expectations of oneself, (2) expectations of others, and (3) other's expectations of oneself. A respondent from a neighborhood without clear social norms on water conservation, for example, might have expectations of themselves (i.e., a personal value), but will not have expectations of others nor will they feel that others have expectations of them in terms of water conservation. On the other hand, the presence of a social norm will lead people to expect a certain behavior of other people, as well as feeling that other people expect that behavior of them. Interviewees were selected via a snowball sampling approach, in which we interviewed an initial group of people and asked them to nominate other people within their social groups to be interviewed (Neumann 2011), who were then interviewed and asked to provide additional interviewees, and so forth. Our initial group of interviewees consisted of people that held leadership roles within their neighborhood associations. The interviewees came from a variety of neighborhoods, ranging from quasi-suburban to central downtown. We followed standard qualitative sampling procedures that enabled us to reach data saturation, a point when additional interviews no longer reveal new insights relevant to the research questions (Fusch and Ness 2015; Saunders et al. 2018; Guest et al. 2020). For a generally non-controversial topic with a demographically homogeneous population, data saturation is expected around 12 interviews (Guest et al. 2006). In our case, we felt confident that data saturation was reached before the 15th interview. Therefore, after conducting a total of 15 interviews in different neighborhoods, we completed the interview process and no longer sought additional interviews (Guest et al. 2006). These interviews were associated with distinct neighborhoods throughout Indianapolis, which we were able to geographically connect to census tracts. Some of the larger neighborhoods covered multiple census tracts, while others were contained within a single tract. These interviews were used to help interpret the quantitative model results within those tracts.

Methods

The primary analysis was based on supervised learning (i.e., predictive modeling), a subset of statistical learning theory, followed by an interpretation phase using the qualitative data (see Fig. 1). For more information on supervised learning, see the Supplemental Methods. In particular, this study leverages the random forest algorithm (Breiman 2001), which has been successfully used in a number of sustainability and resilience-focused studies on the energy sector (Mukherjee and Nateghi 2019; Mukherjee et al. 2018; Lokhandwala and Nateghi 2018) and the water sector (Obringer and Nateghi 2018; Wongso et al. 2020). A more detailed account of the algorithm can be found in the Supplemental Methods.

There are four main steps in the modeling framework, as shown in Fig. 1. The first is data collection and pre-processing. The data were collected as described above and aggregated into seasons, as variability in residential water consumption tends to be seasonal (e.g., water consumption rises in the summer due to outdoor activities). Due to the heavy-tailed nature of the water consumption values, it was necessary to separate the data into moderate- and high-intensity datasets, which were analyzed separately. Often, the high-intensity users are of particular interest to water utilities since they tend to use a disproportionate amount of water and thus have the potential to see large savings from conservation (Rosenberg 2007; Suero et al. 2012; Abdallah and Rosenberg 2014). However, these high-intensity users are also likely to have different key predictors of their consumption than the majority of the population. In this sense, we have split the dataset based on these levels of intensity, so that we may better understand the impact of the predictor variables across the city. To perform this separation, we drew from previous work, namely Balling et al. (2008) and Mukherjee and Nateghi (2017), to classify the lower 75% of the data as 'moderate-intensity' and the remaining 25% as 'high-intensity'. It is important

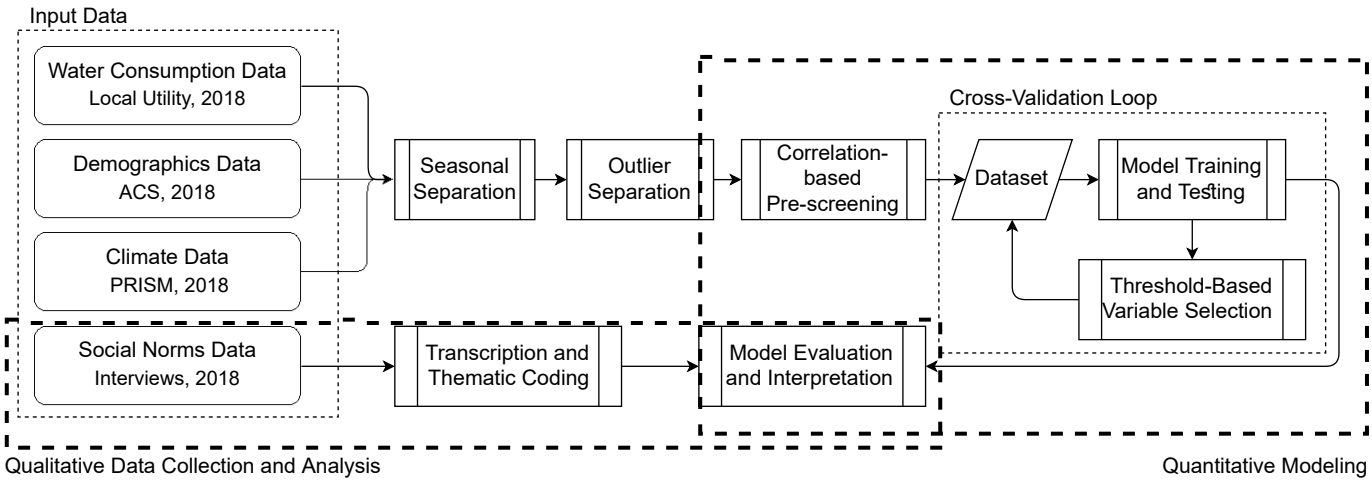

Fig. 1. The integrated modeling approach employed in this study.

J. Water Resour. Plann. Manage., 2022, 148(12): 04022065

to note that the moderate-intensity group also includes low-intensity users, but for brevity, we will refer to this group as moderate-intensity throughout this text. This allowed us to model the majority of census tracts with accuracy, as these are likely to be of interest to water utilities, as well as uncover the general trends present within the city. The high-intensity analysis, results of which are included in the Supplemental Materials, contains many rural census tracts, which may be better predicted by different variables than the urban and suburban areas. Additionally, the social norms data were collected in the central part of the city, which is included in the moderate-intensity analysis. In this sense, the novel integration of quantitative modeling with interpretation from qualitative data remains unchanged by the decision to separate the moderate- and high-intensity users.

Variable Selection

There were two phases of variable selection in this analysis: (1) a correlation-based pre-screening; and (2) a threshold-based variable removal. This process helps reduce the complexity of the model while maintaining a high predictive accuracy. The first phase of variable selection (correlation-based pre-screening) was performed prior to testing and training the model. Although multicollinearity does not pose a problem to the model's predictive accuracy, the presence of a high number of highly-correlated variables (see Fig. S1) can create 'masking effects' and confound model inferencing (Shmueli 2010; Mukherjee and Nateghi 2017; Obringer et al. 2020b). Masking effects often confound model inferencing by causing a number of highly correlated variables to be of high importance, even though the variables are likely all explaining the same thing (due to the correlation). This means that other important variables might be 'masked' by the highly correlated variables. Removing these correlated variables can allow for more insightful interpretation without impacting model performance. Moreover, large amounts of predictor variables increase computation time and unnecessarily increase the complexity of the model. In the variable selection analysis, we keep only the variables X that contribute the most towards predictive accuracy and remove the highly-correlated variables Y that have a correlation score to variables X greater than 0.6. For example, home ownership, a key predictor of water consumption, is highly correlated with detached houses and marital status. Since home ownership contributed more to the overall accuracy than the detached house or marital status variables, it was kept in the dataset, while the variables that were highly-correlated with home ownership (detached house and marital status) were removed. This method of variable selection is often referred to as a high correlation filter. Here, we implemented this filter automatically through computational code so that there is limited opportunity to introduce bias to the system.

Following the first phase of variable selection, the model was trained and tested using the updated variable list. This step involved developing a model for each season, considering the total residential water consumption in each census tract as the response variable. The model was developed using five-fold cross validation, in which 20% of the data were held out for testing purposes (Hastie et al. 2009). This process was done iteratively, such that a different 20% was held out of the model in each iteration. The result is a robust model that is not biased towards one particular area of the city. Following this initial training and testing, the second variable selection step was performed—threshold-based selection. Here, the final variables were selected based on a threshold analysis. In particular, any variable that was in the 90th percentile in terms of relative importance was kept for the final model. In this study, relative importance was determined by the increase in predictive error (mean squared error; MSE) that would occur, should the variable be removed from the model. In other words, the variables that contributed the most to predictive accuracy were included in the final model, while those that made very little contribution were removed. Removing the variables not only reduces the complexity of the model (and therefore reduces predictive error through the bias-variance trade-off) (Hastie et al. 2009), but it also ensures that only the most important variables are considered in the final interpretation phase, avoiding any issues of 'masking' (Shmueli 2010). Finally, threshold-based variable selection using relative importance as the metric has been widely used in previous predictive modeling studies that focused on model interpretability (Genuer et al. 2010; Obringer et al. 2020a).

The approach discussed above relies on selecting a correlation threshold, which can introduce bias into the model. To avoid this bias, it is possible to implement an automated process, such as variable selection using random forests (VSURF), which implements a threshold and conducts an importance analysis with minimal input from the user (Genuer et al. 2015). However, this method is very computationally expensive, particularly with large datasets, such as the one considered in this study. Additionally, it may result in a predictor set that has high levels of multicollinearity, which will not impact predictive accuracy but limit the interpretability and add unnecessary complexity to the model (Shmueli 2010; Mukherjee and Nateghi 2017; Obringer et al. 2020b). As such, it may be preferable to leverage the two-stage approach discussed here to improve computational efficiency and increase interpretability.

Model Training and Testing

After the variables were selected, the model was re-run through the training and testing process. The results were then subsequently analyzed by evaluating variable importance and partial dependence (Hastie et al. 2009), as well as comparing the modeled water consumption with the social norms data collected via the interviews. The variable importance analysis was performed using the variables selected via the process outlined above. This type of analysis demonstrates magnitude of importance, but not the direction of the relationship (i.e., positive or negative). To comprehensively assess the relationship between water consumption and the predictor variables, we performed a partial dependence analysis. A partial dependence analysis holds all predictor variables constant except the one of interest, which allows the user to determine the impact that the variable of interest has on the average value of the response variable (Hastie et al. 2009).

Qualitative Data Analysis

A key focus of this research was the integration of qualitative data with quantitative modeling, following a mixed-methods study design. One common use of qualitative data is to enhance or clarify quantitative results based on quantitative surveys or models. Specifically, when the quantitative component of a study reveals unexpected patterns that the research does not quite understand or need additional explanations, using qualitative interviews with targeted individuals could help illuminate the unexpected results (Creswell and Clark 2017; Schoonenboom and Johnson 2017). More information on this style of study design can be found in the Supplemental Methods. In this study, we conducted a total of 15 interviews with residents around the city to discuss their views on water conservation, both personally and within their neighborhood. The interviews were coded and analyzed thematically to discern perceptions of water conservation within various neighborhoods (Hsieh and Shannon 2005). We paid particular attention to the presence (or lack thereof) of social norms related to water conservation. Here, by matching the neighborhood of each interviewee with the associated census tract(s), we used the social norms data to further interpret the quantitative model results in a select number of census tracts. In particular, the qualitative data, though smaller in extent when compared to the quantitative model results, were used to interpret differences between the

modeled water consumption and actual water consumption. For example, in some areas, the model predicted higher consumption than reality. Here, the qualitative data was used to provide context as to why the area was using less water than expected. Ultimately, the use of qualitative data as an interpretation tool led to deeper insight into the model results than what would be possible with solely using demographic and climate data. Through this approach, we were able to use qualitative data collection and analysis (i.e., semi-structured interviews) to aid the interpretation of quantitative modeling (i.e., supervised learning) and ultimately gain deeper insight into intra-city water consumption patterns.

Results

In this section, we discuss the results from the moderate-intensity analysis (described in the section "Methods"). In particular, we first show the model performance, including measures of error and the difference between the actual and predicted water consumption values. Then, we discuss the variable importance in terms of predictive accuracy. Finally, we discuss the results of the qualitative interviews, which are then used to interpret the predicted residential water consumption. Results of the high-intensity analysis are presented in Figs. S5–S10.

Model Performance

The statistical performance of a model is often measured in terms of out-of-sample prediction error, as well as the ability of the model to explain the variance of the actual data. Table 2 outlines the model performance across each season. In particular, the table contains values for out-of-sample (i.e., the test sample) R^2, root mean squared error (RMSE), and the normalized root mean squared error (NRMSE). The R^2 values can be used to evaluate the ability of the model to fit the data (i.e., the amount of variance explained by the model). These performance measures indicate that the model can adequately capture the variance in the water consumption data across the seasons, with R^2 values from 0.77 to 0.83 (Table 2). The other two measures of model performance (RMSE and NRMSE) represent the prediction error. The NRMSE is the normalized form of the RMSE, providing a unitless measure of prediction error. Here, lower values indicate lower prediction error. The results therefore demonstrate that the model has high predictive accuracy (low error) across the seasons (Table 2). In particular, the summer season performs the best, which is a critical time for demand management, as it generally represents tpeak usage.

Additionally, the NRMSE can be used to gauge uncertainty. Our model, for example, results in an NRMSE ranging from 0.096 in the summer to 0.114 in the spring (see Table 2). This means that the average error in our model is 9.6% to 11.4%, depending on the season. In other words, the prediction results may be about 10% more or less than the actual values. That being said, there remains uncertainty in the model, which can present a challenge when applying the framework. For example, a 9.6% error in the model is relatively low, but translates to about 2.57 million liters of water (see Table 2)—this could cause issues for utility companies that plan to allocate a certain portion of water to the city but end up with an unexpected deficit or surplus. Nonetheless, the model performs well and improves upon previous work. In terms of the variance, our results indicate that the model explains 77% to 83% of the variance in the actual data, depending on the season (see Table 2). This is a significant improvement over previous work conducted in Phoenix, Arizona, in which the average R^2 value was 0.25 (Balling et al. 2008). The improvement is likely due to the use of a nonlinear model that does not require any strict parameterizations of the relationship between the dependent variable (water consumption) and the independent variables (demographics and climate).

Table 2. Measures of model performance, including R^2, root mean squared error (RMSE), and normalized root mean squared error (NRMSE), for each season

Season	R^2	RMSE (L)	NRMSE
Spring	0.77	2,454,930	0.114
Summer	0.83	2,571,736	0.096
Fall	0.80	2,440,708	0.109
Winter	0.79	2,506,419	0.112

Note: Each measure represents the out-of-sample (i.e., test sample) model performance averaged across the five cross validation iterations.

Fig. 2 shows the differences between predicted and actual water consumption in each census tract over the course of 2018, which can be used to visualize the spatial variations in predictive accuracy. One can, for example, evaluate where the model over-predicts residential water consumption (blue shades) and which areas the model under-predicts water consumption (red shades), as well as the magnitude of those over-/under-predictions. The figure shows relatively small differences across the study area with some seasonal differences. In particular, the summer and fall months include more extreme differences (> 7.5 million liters) than the winter and spring seasons. Around 4% of the census tracts in the summer model have extreme differences between the actual and predicted data, compared to 2% in the winter model. Likewise, the fall model shows 3% of tracts with extreme differences (> 7.5 million liters), compared to 2% in the spring model. These differences could be due to the increase in outdoor water consumption during the summer and fall seasons, compared to the winter and spring seasons. In most of these extreme cases, the model predicted less water consumption than the observations (red tracts in Fig. 2). This is likely due to housing characteristics (lot size, house age, etc.), which were not included in this study. It is possible that adding these variables would increase the predictive accuracy in certain tracts, but that is beyond the scope of this study.

It is notable that the model performs better in the central areas of the city, where water consumption tends to be less than the outer, more suburban areas (see Fig. S11). The suburban areas likely have similar demographics to some of the more central areas (house value, income, etc.), but different end-uses (e.g., outdoor landscaping). Considering the generalized data-driven model architecture for the entire study domain, if two different tracts have similar demographic data, the model will predict similar water consumption. Overall, the majority of the tracts are being predicted accurately, with minimal differences between the predicted and actual values—a benefit for water utilities and policymakers interested in better understanding the intra-city water consumption patterns.

Variable Importance and Partial Dependence

Following the variable selection process outlined in the section "Data and Methods," five to six variables per season were selected for the final model. The final variables are shown in Fig. 3. Among the important variables, home ownership was found to be the most influential across all seasons. This means that the percentage of houses that are owned within a census tract is crucial for predicting the water consumption within that tract. Furthermore, house value and household income are repeatedly among the most important variables, indicating a close relationship between socioeconomic status and water consumption. Additionally, the percentage of families with kids in a given census tract was found to be important for predicting water consumption. Education was also shown to

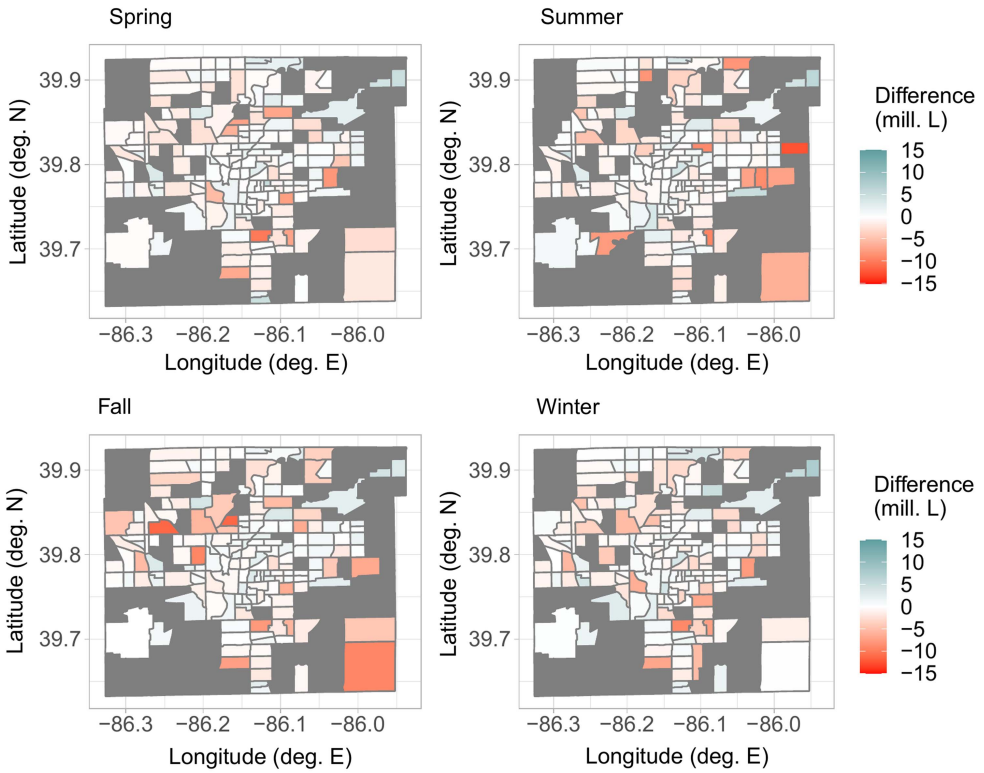

Fig. 2. Differences between the actual and predicted water consumption within each census tract for each season during the year 2018. Blue represents an over-prediction of the water consumption, while red represents an under-prediction. Note that the grayed-out areas represent tracts that were con-sidered part of the high intensity dataset (see Fig. S6).

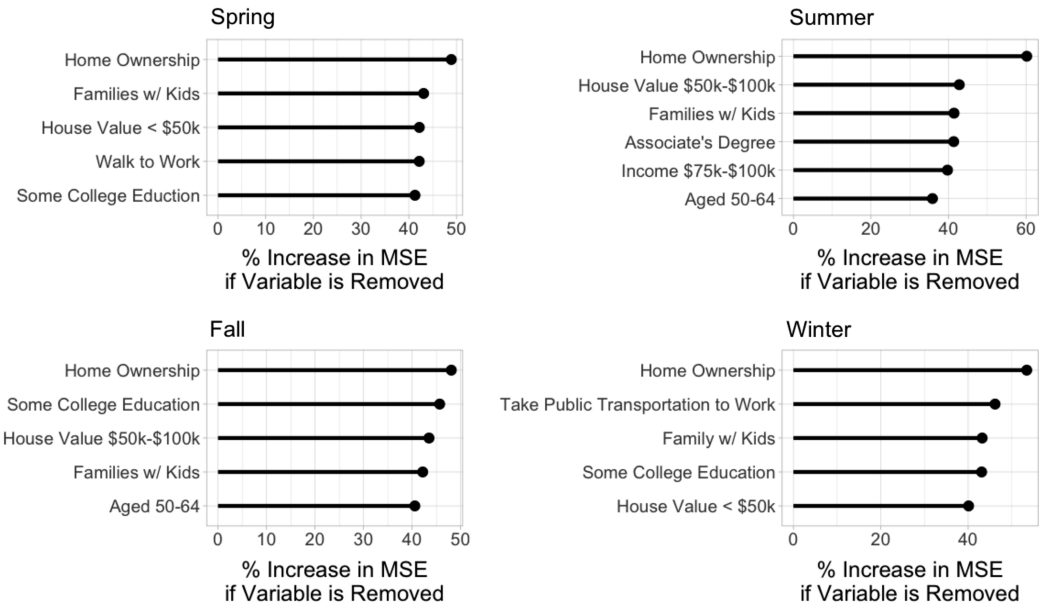

Fig. 3. Variable importance within each season.

play an important role in predicting total water consumption within a census tract. In particular, in the spring, fall, and winter months, the percent of residents with some college education was an important predictor, while in the summer months, the percent of residents with an associate's degree was found to be important. Finally, the percentage of people that walk to work was important in the spring months. Notably, none of the climate variables remained following the final threshold-based variable selection. This suggests that while changes in climate may be important at the larger inter-city scale (Obringer et al. 2019), they are less important for predicting intra-city water consumption within the city of Indianapolis. This is may be due to the variability in the data. In other words, the climate does not vary as much (within the city) as compared to the demographics. It is possible, then, that an analysis

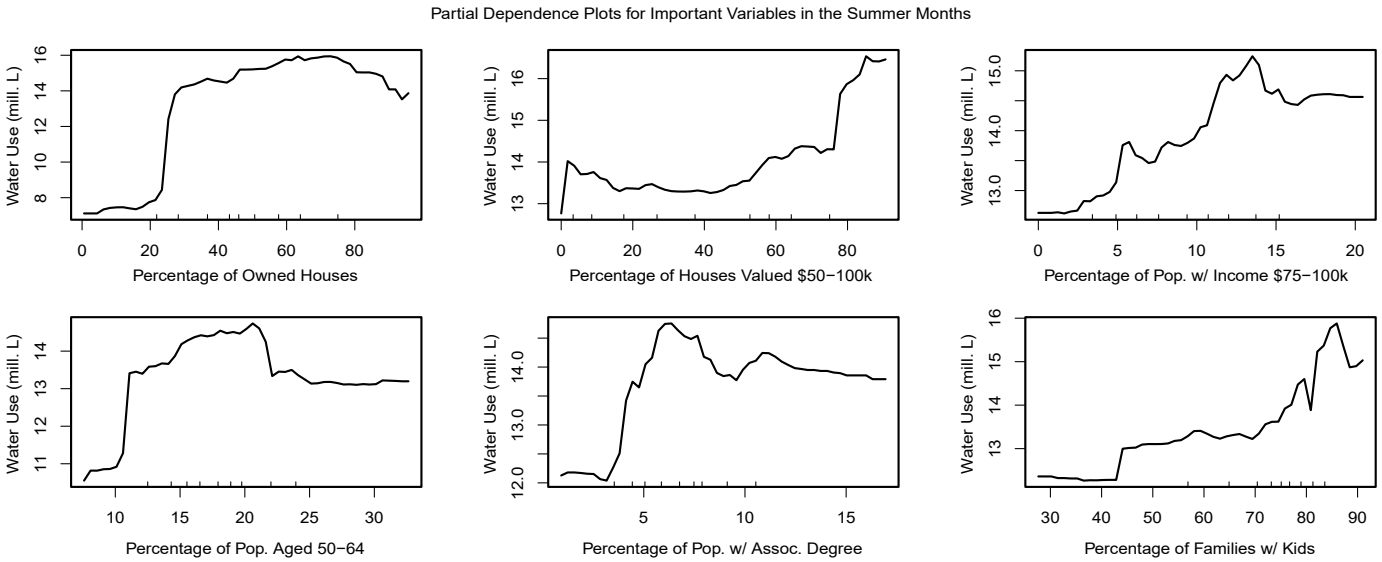

Fig. 4. Partial dependence plots for the most important variables during the summer months.

conducted in a more climatically-variable area may be more sensitive to the changes in intra-city climate, as previously shown in Phoenix (Balling et al. 2008).

The variable importance plots (Fig. 3) indicate the most crucial independent variables for predicting intra-city water consumption, but they do not indicate the direction of the relationship (i.e., positive or negative correlation). To understand how the important variables impact water consumption, we can use partial dependence plots. Fig. 4 shows the results of the partial dependence analysis for the summer months (see Figs. S2–S4 for the other seasons included in the analysis). In particular, the figure shows the partial dependence for the six variables shown in Fig. 3. Fig. 4 shows that as the percentage of home owners increase, the water consumption increases as well. Similarly, as the percentage of families increase, so does total water consumption within the census tract. In terms of socioeconomic indicators, as the percentage of houses valued between $50,000 and $100,000 increases, there is an initial reduction in water consumption. However, this drop is followed by a steady increase in water consumption as the percentage of lower-valued homes increases. Finally, the percentage of people with associate's degrees was an important variable in the summer months (Fig. 3). The partial dependence plot indicates that at first, water consumption increases with the percentage of associate's degrees, but then decreases around 5%. The partial dependence plots allow us to attach a direction to the sensitivity of important variables and begin to understand the relationship between the predictor and response variables.

Qualitative Interviews

The interviews were coded and analyzed thematically to discern perceptions of water conservation within various neighborhoods. We paid particular attention to the presence (or lack thereof) of social norms related to water conservation. In general, most interviewees indicated that there were no strong social norms regarding water conservation within their neighborhoods. Rather than expressing an expectation on their neighbors in terms of water conservation, the interviewees discussed having personal values that were not shared by the neighborhood as a whole, as evident in the following quote: "I think it's very much on a house to house basis, there are people doing individual things, but I don't know in our neighborhood that there is necessarily like a grassroots effort." That being said, there were a few respondents who indicated their neighborhoods did in fact have social norms surrounding water conservation. These interviewees indicated that environmentally-friendly activities, such as recycling, driving electric vehicles, and using rain barrels, were popular among the residents. Moreover, they stated that those behaviors were expected and that conservation would often be brought up at neighborhood gatherings. One particular interviewee discussed pro-environmental behavior and mindsets in their neighborhood by saying: "I think that from a community perspective, at a micro-level, [the neighborhood] is going to be much more socially conscious of environmental conservation efforts, if you will, than Indiana as a whole or even Indianapolis. I think we're somewhat unique in that regard and I certainly think that is the reality, especially from the interactions that I have had, not just in person, but on social media." By geographically connecting these neighborhoods with the census tract(s) within the same designated area we use the interview results to interpret the quantitative model findings discussed above.

Discussion

Within the body of literature on water demand modeling, there have been a number of studies aimed at determining the various factors that ultimately impact water consumption. For example, Sankarasubramanian et al. (2017) found that higher education and income often led to higher adoption of efficiency techniques. House-Peters et al. (2010) found that water consumption depended on several housing characteristics (e.g., outdoor space and house size), as well as education levels. Finally, Ashoori et al. (2016) found that temperature and precipitation were important factors for predicting water consumption, particularly in single family homes. The work presented here shares some similarities with previous findings, as well as some differences, which are discussed below.

For example, our study did not include housing characteristics, but did include home ownership, which was found to be important across all four seasons. Fig. 3 shows that the percentage of home ownership was the most important variable across the entire year—indicating a potential demographic variable for the water utility to use for targeted conservation initiatives. Fig. 4 demonstrates that as the number of owned houses increases, the total water consumption does as well, which is expected, as home ownership usually means larger lot sizes and houses, leading to higher water

consumption, especially in the summer when landscaping is popular (House-Peters et al. 2010; Sankarasubramanian et al. 2017). Likewise, as the number of families with kids increases, so does water consumption (Fig. 4). This is likely due to the increased number of people within a household, which leads to more water consumption. This finding is aligned with previous studies on the subject (Worland et al. 2018). The percentage of families with kids was also an important predictor in the high-intensity user group (see Fig. S5), indicating that this may be an optimal demographic predictor across the city. These findings suggest that efforts to limit water consumption ought to be targeted at areas in which home ownership is significant, rather than areas with a majority of renters, as well as areas that are primarily made up of families.

One of the particularly novel results of our analysis was the importance of walking to work in predicting the water consumption, which has not, to our knowledge, been reported prior to this work. The percentage of people that walk to work is not a variable that many would intuitively connect to water consumption, thus it is possible that previous analyses have simply not included variables related to commutes. This is an advantage of starting with a large dataset (e.g., 72 demographic variables) and doing an automated variable selection procedure—the algorithm was able to use a variable to make a prediction that otherwise might not have even been included. Given that the percentage of people who walk to work is positively correlated with the percentage of single people and people in their twenties while negatively correlated with the percentage of families and home ownership (see Fig. S1), it is likely that walking to work is an indirect proxy for location within the city. In particular, the population that walks to work is primarily located in the city center (see Fig. S11), which is also an area that contains a lot of apartments and smaller houses with little to no outdoor space. In this sense, it may not be walking to work which is the predictor of water consumption, rather walking to work is representative of one's location within the city and thus, the predominant style of living, which may be the true predictor of water consumption. This echoes previous work, which demonstrated that suburban households (which are unlikely to be associated with walking to work) tend to consume more water due to outdoor landscaping, as well as larger house sizes (Balling et al. 2008; House-Peters et al. 2010). Conversely, in the high-intensity user group, which contains a number of rural census tracts, the percentage of people that worked from home and the percentage of people that drive to work were found to be important predictors in the summer and winter season, respectively (see Fig. S5). The partial dependence analysis suggests that as the number of people that work from home increases in these high-intensity tracts, the water consumption increases (see Fig. S8). This could be due to increased use of indoor water while people are home most of the day, as shown by a recent study focused on the impact of remote work (Li et al. 2021). Another possibility is that the increased use of water is being used primarily outdoors for work-related purposes, such as farming, since the majority of the high-intensity census tracts are rural. Additional research is needed to better understand these impacts on the high-intensity user base and how they might be leveraged to encourage water conservation.

A common predictor of water consumption is household income. A number of studies have found positive relationships between these two variables, with higher income often leading to higher water consumption (Harlan et al. 2009; House-Peters et al. 2010; Worland et al. 2018). Similarly, in our model, during the summer months, the percentage of houses valued between $50,000 and $100,000 (the median house value for the city is $130,000) was found to be positively related to water consumption. In particular, there is a notable inflection point when 30% of the homes within the tract are valued between $50,000 and $100,000 in which the water consumption begins to steadily increase. In fact, census tracts with percentages of lower-valued houses above this threshold tend to also have higher percentages of low income households (less than $50k a year), as shown in Fig. S11. Previous work has shown that lower-income households are less likely to have efficient appliances (Sankarasubramanian et al. 2017), which may contribute to the increased water consumption shown in Fig. 4. In the high-intensity group, income was also shown to be important, with higher percentages of affluence leading to higher consumption of water (see Figs. S7–S10). This aligns with previous research into the connection between income and water consumption (Harlan et al. 2009; House-Peters et al. 2010; Worland et al. 2018). Finally, education levels were found to be important across the seasons. In particular, the percentage of people with associate's degrees was found to be important in the summer months. Previous work has suggested that education levels were a significant predictor of water use efficiency (Sankarasubramanian et al. 2017), which may explain the decrease once the census tract reaches 5% of the population having associate's degrees. Overall, the important variables presented here not only echo previous studies, but also provide valuable information on how to target conservation within cities.

The majority of previous literature has focused on quantitative modeling of water consumption, which has its own limitations in terms of interpretation. That is, one cannot interpret the results beyond the correlation or the predictive accuracy. One way to move past this limitation is to using qualitative data as a source of insight (Elsawah et al. 2020). Here, we used qualitative interviews to assess the presence of social norms in several census tracts, shown in Fig. 5. Visually, these census tracts cluster into two groups—the center-top and the central group, with the center-top group containing more census tracts where the water consumption was overpredicted than the central group. The central group represents the city center, while the center-top group is still relatively close to the city center, but larger properties (often with large yards) and detached houses are common. Normally, these neighborhoods would have higher than usual water consumption—given the percentage of home

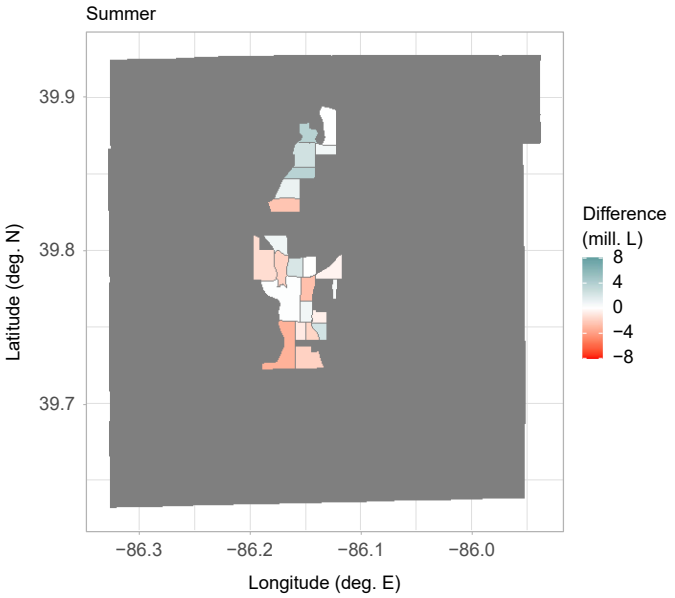

Fig. 5. Differences in the predicted and actual water consumption values during the summer months for the specific census tracts that geographically correspond with the neighborhoods in which we conducted interviews. Note that some neighborhoods cover multiple census tracts.

ownership and families. However, the model over-predicted the consumption, indicating that in reality these neighborhoods consume less water than their demographics alone would suggest, based on the rest of the study area. This finding was confirmed by the interview results, discussed above, in which interviewees from the center-top group indicated that there were social norms within their neighborhoods that encouraged water conservation practices. In particular, the interviewees discussed the prevalence of using rain barrels for outdoor landscaping, rather than relying on water from the tap. This information regarding the local water conservation practices cannot be obtained through the usual data sources (i.e., census data, housing records, etc.), but it is, nonetheless, critical to understanding intra-city water consumption patterns. Based on the demographics of the area (e.g., presence of families, detached homes, higher income, etc.), water utilities may assume that they need to run a conservation program in these neighborhoods, when in fact, conservation-based norms already exist and are leading to less consumption than neighborhoods with similar demographics. By integrating an understanding of local social norms into the interpretation of quantitative modeling results, utility companies may be able to focus their efforts on areas of the city in which demand management interventions will be more effective.

Limitations of the Study

There are a few limitations of the study. First, the present study only uses 15 interviews to conduct the interpretation analysis. While the sample size was guided by the data saturation point in our semi-structure interview process (Guest et al. 2006), we recognize that it represents a small fraction of the population in Indianapolis. A broader data coverage over the city, rather than focusing on a few central neighborhoods, could contribute to more inclusive and successful conservation interventions. It would be especially important to extend this analysis to the high-intensity users, as they tend to have a disproportionately high level of water consumption and also represent an opportunity to greatly reduce water consumption (Rosenberg 2007; Suero et al. 2012; Abdallah and Rosenberg 2014). However, conducting interviews on such large scales is labor intensive and may be infeasible for research teams, policymakers, and water and other resource management practitioners due to time and monetary constraints. In place of interviews, a survey method may be more beneficial in future studies for determining large-scale attitudes towards water conservation, as well as any social norms that are present within different areas of the city. Surveys have their own set of challenges, however, and must be carefully designed to ensure unbiased results. Although a survey was outside the scope of the current study, the results from the interviews can be used to develop future survey questionnaires and large-scale datasets. Similar work has been done in other areas of the country, although not specifically focused on social norms (White et al. 2019). In addition, in our current study, social norm data was only collected and used a posteriori to improve interpretation of model results, but not used to build the mathematical models. There is great potential for including social norm variables in future models and use relevant large-scale data as model inputs.

Another limitation of the study is the lack of landscaping variables in the study, such as lot size or irrigation requirements. These variables play a major in role in determine total water consumption, particularly in the summer months (House-Peters and Chang 2011). Moreover, landscaping variables, or more broadly outdoor water use, is likely to be more influential among the high-intensity group, which could improve the interpretation of the model results from these tracts. Often this data is collected from a variety of sources, such as real estate websites or remote sensing images. However, the lack of unified public database for the city of Indianapolis led us to not include the data in the quantitative analysis—although some interviewees discussed irrigation habits within their neighborhoods. This exclusion of landscaping variables likely impacted the predictive accuracy of the model, especially for suburban census tracts, which are likely to have larger yards that require irrigation. Future work should seek to include these variables, particularly if the end goal is to improve predictive accuracy.

Finally, this study leveraged a two-stage approach for variable selection that relied on a pre-detremined correlation threshold. While this process has been leveraged in several previous studies (Genuer et al. 2010; Mukherjee and Nateghi 2017; Obringer et al. 2020b), the correlation pre-screening may add bias, particularly if it involves expert opinions in the decision-making process. To minimize this opportunity for added bias, we maintained a strict computational criterion for the pre-screening filter, which did not rely on expert opinions. Additionally, the two-stage process implemented in this analysis was found to be more computationally efficient, as well as more interpretable, when compared to a heuristic-based algorithm. That being said, in the future, there is likely to be a shift towards automated variable selection, particularly as algorithms become more efficient. There are a growing number of algorithms, such as the variable selection using random forest (VSURF), which is similar to the threshold-based approach implemented here (Genuer et al. 2015). These algorithms can be used with a variety of algorithms, from simple linear regression (Li et al. 2013) to more complex tree-based models (Galelli and Castelletti 2013). In the future, researchers may opt to implement these procedures to further limit any potential for added bias.

Conclusions

This study sought to improve the interpretation of water consumption models through the integration of data-driven modeling with qualitative data collected via semi-structured interviews. This integrated modeling approach, which pulls from both data science and social science, represents a step towards deeper integration of these two fields, which is a challenge facing socio-environmental systems research (Elsawah et al. 2020). Using Indianapolis, Indiana as a case study, we showed that the developed model can reliably predict the actual consumption across the seasons, particularly during the summer months, which is when water consumption peaks in the study area. Additionally, we demonstrated that variables such as the percentage of home ownership and families with kids were key predictors of total water consumption in a given census tract. Looking at differences within the study area, the results indicated that the modeled water consumption was representative of the actual water consumption across the city. A few exceptions were found on the outer edge of the study area, where lots tend to be larger, possibly indicating that in these areas' water consumption is more dependent on housing characteristics (lot size, house age, etc.) than demographics. Additionally, there were some areas in the city in which the model predicted more water consumption than in reality. Using resident interviews to supplement these results, we showed that social norms regarding water conservation may play a role in limiting water consumption in certain areas around the city. This highlights the need for utilities and policymakers to consider the conservation-focused social norms of communities when trying to implement water conservation initiatives and plan for future water supply needs. That being said, this study mainly focused on presenting the model results, rather than delving deep into the implications of using this model in a practical sense. Therefore, future work should further investigate the impact of this model on reducing water

consumption. In this sense, researchers and practitioners could test scenarios in which this model could be deployed to determine optimal areas around the city for water saving measures, then tested to evaluate any changes to overall water consumption. This would further add to the practicality of the presented model. Future researchers could also look into segmentation analysis, which was briefly mentioned in the Introduction. This type of analysis creates groups of residents based on shared characteristics and can be used to further conservation policies. Although the model was developed and applied in Indianapolis, the methodology presented is general and can be applied to a number of other locations around the world, although different cities are likely to have different important variables. For example, Balling et al. (2008) found intra-city water consumption patterns in Phoenix to be sensitive to climate, while our study showed that the climatic changes within the city were not as important for explaining the differences between census tracts. To test the current model in different areas would require a more intensive analysis with similar data across a number of cities. Although this is outside of the scope of the present study, it is an area of interest for future research. Finally, the integrated, mixed-methods modeling approach presented here, which allowed us to draw a number of conclusions that went beyond single-method approaches, demonstrates not only the possibility of integrating data science and qualitative social science, but also the importance of such integrative frameworks when faced with complex challenges, such as water conservation.

Data Availability Statement

The data and code used in this study have been archived with Zenodo (https://doi.org/10.5281/zenodo.6452575) and are also available on GitHub (https://github.com/reneeobringer/WaterConsumptionAnalysis).

Reproducible Results

Masooma Batool (Helmholtz Centre for Environmental Research—UFZ) downloaded, installed, and ran the code using the input data set and reproduced results in Table 2 and Figs. 2–4.

Acknowledgments

The authors would like to acknowledge support from NSF Grants #1826161 and #1832688, as well as the National Socio-Environmental Synthesis Center (SESYNC) under funding received from NSF Grant #DBI-1639145. The authors would also like to acknowledge support from the Graduate School at Purdue University, the Bilsland Dissertation Fellowship, the Purdue University Center for the Environment, and the Purdue Climate Change Research Center.

Supplemental Materials

Supplemental Methods, Table S1, and Figs. S1–S11 are available online in the ASCE Library (www.ascelibrary.org).

References

Abdallah, A. M., and D. E. Rosenberg. 2014. "Heterogeneous residential water and energy linkages and implications for conservation and management." J. Water Resour. Plann. Manage. 140 (3): 288–297. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000340.

AghaKouchak, A., D. Feldman, M. Hoerling, T. Huxman, and J. Lund. 2015. "Water and climate: Recognize anthropogenic drought." Nature News 524 (7566): 409. https://doi.org/10.1038/524409a.

Ashoori, N., D. A. Dzombak, and M. J. Small. 2016. "Modeling the effects of conservation, demographics, price, and climate on urban water demand in Los Angeles, California." Water Resour. Manage. 30 (14): 5247–5262. https://doi.org/10.1007/s11269-016-1483-7.

Ault, T. R. 2020. "On the essentials of drought in a changing climate." Science 368 (6488): 256–260. https://doi.org/10.1126/science.aaz5492.

Balling, R. C., P. Gober, and N. Jones. 2008. "Sensitivity of residential water consumption to variations in climate: An intraurban analysis of Phoenix, Arizona." Water Resour. Res. 44 (10): 1–11. https://doi.org/10.1029/2007WR006722.

Basara, J. B., J. I. Christian, R. A. Wakefield, J. A. Otkin, E. H. Hunt, and D. P. Brown. 2019. "The evolution, propagation, and spread of flash drought in the Central United States during 2012." Environ. Res. Lett. 14 (8): 084025. https://doi.org/10.1088/1748-9326/ab2cc0.

Beal, C. D., T. R. Gurung, and R. A. Stewart. 2016. "Demand-side management for supply-side efficiency: Modeling tailored strategies for reducing peak residential water demand." Sustainable Prod. Consumption 6 (Apr): 1–11. https://doi.org/10.1016/j.spc.2015.11.005.

Bhanot, S. P. 2017. "Rank and response: A field experiment on peer information and water use behavior." J. Econ. Psychol. 62 (Oct): 155–172. https://doi.org/10.1016/j.joep.2017.06.011.

Bicchieri, C. 2016. Norms in the wild: How to diagnose, measure, and change social norms. Oxford, UK: Oxford University Press.

Bolorinos, J., R. Rajagopal, and N. K. Ajami. 2020. "Mining the gap in long-term residential water and electricity conservation." Environ. Res. Lett. 16 (2): 024007. https://doi.org/10.1088/1748-9326/abbfc2.

Breiman, L. 2001. "Random forests." Mach. Learn. 45 (1): 5–32. https://doi.org/10.1023/A:1010933404324.

Bruss, C. B., R. Nateghi, and B. F. Zaitchik. 2019. "Explaining national trends in terrestrial water storage." Front. Environ. Sci. 7 (Jun): 85. https://doi.org/10.3389/fenvs.2019.00085.

Cominola, A., M. Giuliani, A. Castelletti, P. Fraternali, S. L. H. Gonzalez, J. C. G. Herrero, J. Novak, and A. E. Rizzoli. 2021. "Long-term water conservation is fostered by smart meter-based feedback and digital user engagement." npj Clean Water 4 (1): 1–10. https://doi.org/10.1038/s41545-021-00119-0.

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.

Creswell, J. W., and V. L. P. Clark. 2017. Designing and conducting mixed methods research. 3rd ed. Thousand Oaks, CA: SAGE.

Dai, A. 2011. "Drought under global warming: A review." WIREs Clim. Change 2 (1): 45–65. https://doi.org/10.1002/wcc.81.

Elsawah, S., et al. 2020. "Eight grand challenges in socio-environmental systems modeling." Socio-Environ. Syst. Modell. 2 (Jan): 16226. https://doi.org/10.18174/sesmo.2020a16226.

Ewing, R., and S. Hamidi. 2014. Measuring sprawl 2014. Salt Lake City: Smart Growth America.

Fusch, P., and L. Ness. 2015. "Are we there yet? Data saturation in qualitative research." Qual. Rep. 20 (9): 1408–1416. https://doi.org/10.46743/2160-3715/2015.2281.

Galelli, S., and A. Castelletti. 2013. "Tree-based iterative input variable selection for hydrological modeling." Water Resour. Res. 49 (7): 4295–4310. https://doi.org/10.1002/wrcr.20339.

Genuer, R., J.-M. Poggi, and C. Tuleau-Malot. 2010. "Variable selection using random forests." Pattern Recognit. Lett. 31 (14): 2225–2236. https://doi.org/10.1016/j.patrec.2010.03.014.

Genuer, R., J.-M. Poggi, and C. Tuleau-Malot. 2015. "VSURF: An R package for variable selection using random forests." R J. 7 (2): 19–33. https://doi.org/10.32614/RJ-2015-018.

Ghavidelfar, S., A. Y. Shamseldin, and B. W. Melville. 2017. "A multi-scale analysis of single-unit housing water demand through integration of water consumption, land use and demographic data." Water Resour. Manage. 31 (7): 2173–2186. https://doi.org/10.1007/s11269-017-1635-4.

Gleick, P. H. 2003. "Global freshwater resources: Soft-path solutions for the 21st century." Science 302 (5650): 1524–1528. https://doi.org/10.1126/science.1089967.
Guest, G., A. Bunce, and L. Johnson. 2006. "How many interviews are enough?: An experiment with data saturation and variability." Field Methods 18 (1): 59–82. https://doi.org/10.1177/1525822X05279903.
Guest, G., E. Namey, and M. Chen. 2020. "A simple method to assess and report thematic saturation in qualitative research." PLoS One 15 (5): e0232076. https://doi.org/10.1371/journal.pone.0232076.
Hamidi, S., R. Ewing, I. Preuss, and A. Dodds. 2015. "Measuring sprawl and its impacts: An update." J. Plann. Educ. Res. 35 (1): 35–50. https://doi.org/10.1177/0739456X14565247.
Harlan, S. L., S. T. Yabiku, L. Larsen, and A. J. Brazel. 2009. "Household water consumption in an arid city: Affluence, affordance, and attitudes." Soc. Natl. Resour. 22 (8): 691–709. https://doi.org/10.1080/08941920802064679.
Hastie, T., R. Tibshirani, and J. Friedman. 2009. The elements of statistical learning: Data mining, inference, and prediction. 2nd ed. New York: Springer.
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.
House-Peters, L. A., and H. Chang. 2011. "Urban water demand modeling: Review of concepts, methods, and organizing principles." Water Resour. Res. 47 (5): 1–15: https://doi.org/10.1029/2010WR009624.
Hsieh, H.-F., and S. E. Shannon. 2005. "Three approaches to qualitative content analysis." Qual. Health Res. 15 (9): 1277–1288. https://doi.org/10.1177/1049732305276687.
Li, D., R. A. Engel, X. Ma, E. Porse, J. D. Kaplan, S. A. Margulis, and D. P. Lettenmaier. 2021. "Stay-at-home orders during the COVID-19 pandemic reduced urban water use." Environ. Sci. Technol. Lett. 8 (5): 431–436. https://doi.org/10.1021/acs.estlett.0c00979.
Li, G., H. Lian, S. Feng, and L. Zhu. 2013. "Automatic variable selection for longitudinal generalized linear models." Comput. Stat. Data Anal. 61 (May): 174–186. https://doi.org/10.1016/j.csda.2012.12.015.
Lokhandwala, M., and R. Nateghi. 2018. "Leveraging advanced predictive analytics to assess commercial cooling load in the U.S." Sustainable Prod. Consumption 14 (Apr): 66–81. https://doi.org/10.1016/j.spc.2018.01.001.
Luo, T., R. Young, and P. Reig. 2015. Aqueduct projected water stress county rankings. Washington, DC: World Resources Institute.
Mitchell, V. G. 2006. "Applying integrated urban water management concepts: A review of Australian experience." Environ. Manage. 37 (5): 589–605. https://doi.org/10.1007/s00267-004-0252-1.
Mukherjee, S., and R. Nateghi. 2017. "Climate sensitivity of end-use electricity consumption in the built environment: An application to the state of Florida, United States." Energy 128 (Jun): 688–700. https://doi.org/10.1016/j.energy.2017.04.034.
Mukherjee, S., and R. Nateghi. 2019. "A data-driven approach to assessing supply inadequacy risks due to climate-induced shifts in electricity demand." Risk Anal. 39 (3): 673–694. https://doi.org/10.1111/risa.13192.
Mukherjee, S., R. Nateghi, and M. Hastak. 2018. "A multi-hazard approach to assess severe weather-induced major power outage risks in the U.S." Reliab. Eng. Syst. Saf. 175 (Jul): 283–305. https://doi.org/10.1016/j.ress.2018.03.015.
Neumann, L. 2011. Social research methods: Qualitative and quantitative approaches. 7th ed. Boston: Pearson Education.
Obringer, R., R. Kumar, and R. Nateghi. 2019. "Analyzing the climate sensitivity of the coupled water-electricity demand nexus in the Midwestern United States." Appl. Energy 252 (Oct): 113466. https://doi.org/10.1016/j.apenergy.2019.113466.
Obringer, R., R. Kumar, and R. Nateghi. 2020a. "Managing the water–electricity demand nexus in a warming climate." Clim. Change 159 (2): 233–252. https://doi.org/10.1007/s10584-020-02669-7.
Obringer, R., S. Mukherjee, and R. Nateghi. 2020b. "Evaluating the climate sensitivity of coupled electricity-natural gas demand using a multivariate framework." Appl. Energy 262 (Mar): 114419. https://doi.org/10.1016/j.apenergy.2019.114419.
Obringer, R., and R. Nateghi. 2018. "Predicting urban reservoir levels using statistical learning techniques." Sci. Rep. 8 (1): 1–9. https://doi.org/10.1038/s41598-018-23509-w.
Pinto, D. C., W. M. Nique, E. da Silva Añaña, and M. M. Herter. 2011. "Green consumer values: How do personal values influence environmentally responsible water consumption?" Int. J. Consumer Stud. 35 (2): 122–131. https://doi.org/10.1111/j.1470-6431.2010.00962.x.
Prism Climate Group. 2018. "Climate datasets." Accessed August 1, 2022. https://prism.oregonstate.edu/.
Quesnel, K. J., and N. K. Ajami. 2017. "Changes in water consumption linked to heavy news media coverage of extreme climatic events." Sci. Adv. 3 (10): e1700784. https://doi.org/10.1126/sciadv.1700784.
Ramsey, E., E. Z. Berglund, and R. Goyal. 2017. "The impact of demographic factors, beliefs, and social influences on residential water consumption and implications for non-price policies in urban India." Water 9 (11): 844. https://doi.org/10.3390/w9110844.
Rasifaghihi, N., S. S. Li, and F. Haghighat. 2020. "Forecast of urban water consumption under the impact of climate change." Sustainable Cities Soc. 52 (Jan): 101848. https://doi.org/10.1016/j.scs.2019.101848.
Rosenberg, D. E. 2007. "Probabilistic estimation of water conservation effectiveness." J. Water Resour. Plann. Manage. 133 (1): 39–49. https://doi.org/10.1061/(ASCE)0733-9496(2007)133:1(39).
Sankarasubramanian, A., et al. 2017. "Synthesis of public water supply use in the United States: Spatio-temporal patterns and socio-economic controls." Earth's Future 5 (7): 771–788. https://doi.org/10.1002/2016EF000511.
Saunders, B., J. Sim, T. Kingstone, S. Baker, J. Waterfield, B. Bartlam, H. Burroughs, and C. Jinks. 2018. "Saturation in qualitative research: Exploring its conceptualization and operationalization." Qual. Quantity 52 (4): 1893–1907. https://doi.org/10.1007/s11135-017-0574-8.
Schoonenboom, J., and R. B. Johnson. 2017. "How to construct a mixed methods research design." Kölner Zeitschrift für Soziologie und Sozialpsychologie 69 (S2): 107–131. https://doi.org/10.1007/s11577-017-0454-1.
Shandas, V., and G. H. Parandvash. 2010. "Integrating urban form and demographics in water-demand management: An empirical case study of Portland, Oregon." Environ. Plann. B: Plann. Des. 37 (1): 112–128. https://doi.org/10.1068/b35036.
Shmueli, G. 2010. "To explain or to predict?" Stat. Sci. 25 (3): 289–310. https://doi.org/10.1214/10-STS330.
Suero, F. J., P. W. Mayer, and D. E. Rosenberg. 2012. "Estimating and verifying united states households' potential to conserve water." J. Water Resour. Plann. Manage. 138 (3): 299–306. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000182.
US Census Bureau. 2018. American community survey. Washington, DC: US Census Bureau.
Viñoles, M. V., K. Moeltner, and S. Stoddard. 2015. "Length of residency and water use in an arid urban environment." Water Resour. Econ. 12 (Oct): 52–66. https://doi.org/10.1016/j.wre.2015.09.004.
Wang, H., D. Bracciano, and T. Asefa. 2020. "Evaluation of water saving potential for short-term water demand management." Water Resour. Manage. 34 (10): 3317–3330. https://doi.org/10.1007/s11269-020-02615-3.
White, D. D., E. K. Rauh, A. Sullivan, K. L. Larson, A. Wutich, D. Linthicum, V. Horvath, and K. L. Lawless. 2019. "Public attitudes toward urban water sustainability transitions: A multi-city survey in the western United States." Sustainability Sci. 14 (6): 1469–1483. https://doi.org/10.1007/s11625-019-00658-z.
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.
Wongso, E., R. Nateghi, B. Zaitchik, S. Quiring, and R. Kumar. 2020. "A data-driven framework to characterize state-level water use in the U.S." Water Resour. Res. 56 (9): 1–17. https://doi.org/10.1029/2019WR024894.
Worland, S. C., S. Steinschneider, and G. M. Hornberger. 2018. "Drivers of variability in public-supply water use across the contiguous United States." Water Resour. Res. 54 (3): 1868–1889. https://doi.org/10.1002/2017WR021268.