CLIMATE CHANGE EFFECTS ON URBAN WATER RESOURCES: AN INTERDISCIPLINARY APPROACH TO MODELING URBAN WATER SUPPLY AND DEMAND

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Renee Obringer

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

May 2020

Purdue University

West Lafayette, Indiana

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF DISSERTATION APPROVAL

Dr. Roshanak Nateghi, Chair

School of Industrial Engineering and Environmental and Ecological Engineering

Dr. P. Suresh Rao

Lyles School of Civil Engineering

Dr. Zhao Maa

Department of Forestry and Natural Resources

Dr. Rohini Kumar

Department Computational Hydrosystems, Helmholtz Centre for the Environment - UFZ

Approved by:

Dr. Linda Lee

Head, Ecological Science and Engineering Interdisciplinary Graduate Program

To Mom and Dad, your unending support is the reason why this dissertation exists

ACKNOWLEDGMENTS

First and foremost, I would like to express my sincere gratitude and upmost appreciation for my advisor, Dr. Roshanak Nateghi. When, as a third year PhD student, I found myself in need of a new advisor, Dr. Nateghi took me on as a student. Under her mentorship, I was able to flourish as a researcher and make great strides towards my future career. I would like to thank her for believing in me throughout these past few years and for all the guidance that she has provided me.

I would also like to thank my committee members, Dr. Suresh Rao, Dr. Zhao Ma, and Dr. Rohini Kumar for their valuable insights into my research. I would like to thank Suresh in particular for his support from the beginning of my time here at Purdue. Suresh is someone who regularly pushes his students to think outside the box, which has been invaluable for the evolution of my own research. I want to thank Zhao for her patience and mentorship, especially while teaching an engineer how to do qualitative social science work. Finally, I would like to thank Rohini for his recommendations on the climate science aspects of my research.

I am incredibly grateful to Dr. Linda Lee, the head of the Ecological Science and Engineering Interdisciplinary Graduate Program. Her support throughout my time at Purdue has meant more to me than I can put into words. I would also like to thank Christal Musser, Deirdre Carmicheal, and the rest of the staff in the Office of Interdisciplinary Graduate Programs for their support. If all graduate programs offered even a fraction of OIGP's support to students, academia would be a very different place.

I also want to acknowledge Dr. Nina Robinson, Dr. John Sutherland, Jill Wable, Cresta Cates, and the rest of the administration and staff within Environmental and Ecological Engineering.

Any acknowledgement section would be incomplete without mentioning my family and friends. My parents, Doug and Cindy Obringer, have been supportive throughout my entire life, but especially so during my PhD. I am very fortunate to have parents that have given me their unconditional love and support, for which I am very grateful. I would also like to thank my partner, Thomas, who has patiently listened to all my complaints and celebrated my successes.

Additionally, I would like to thank all of my friends that have kept me sane throughout this process. Special thanks to Camila, Xing, Apu, Praneet, Oscar, and Taisha, who are currently spread out around the world, but have always been there to offer some words of encouragement.

Finally, I want to acknowledge my funding throughout my PhD. In particular, the Andrews Fellowship (2015-2017), NSF grant #1728209 (2017-2018), the Purdue University Center for the Environment seed grant (2018-2019), and the Bilsland Dissertation Fellowship (2019-2020). These grants have made my dissertation research possible.

TABLE OF CONTENTS

Page

Page

Page

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

ABSTRACT

Obringer, R. Ph.D., Purdue University, May 2020. Climate change effects on urban water resources: An interdisciplinary approach to modeling urban water supply and demand. Major Professor: Roshanak Nateghi.

Urban populations are growing at unprecedented rates around the world, while simultaneously facing increasingly intense impacts of climate change, from sea level rise to extreme weather events. In the face of this concurrent urbanization and climate change, it is imperative that cities improve their resilience to a multitude of stressors. A key aspect of urban resilience to climate change is ensuring that there is enough drinking water available to service the city, especially given the projections of more frequent and intense droughts in some areas. However, the study of climate impacts on urban water resources is fairly nascent and many gaps remain. In this dissertation, I aim to begin to close some of those gaps by adopting an interdisciplinary approach to studying water availability. First, I focus on urban water supply, and in particular, reservoir operations. I employ a variety of methods, ranging from data science techniques to traditional hydrological models, to predict the reservoir levels under a variety of climate conditions. Following the analysis of water supply, I shift focus to urban water demand. Here, I include interconnected systems, such as electricity, to evaluate and characterize the impact of climate on water demand and the benefit of considering system interconnectivities. Additionally, I present an analysis on the projection of water and electricity demand into the future, based on representative concentration pathways of CO_2. Finally, I focus on the human dimension to the demand studies. By studying the social norms surrounding water conservation in urban areas, as well as the demographics, I built a predictive model to estimate monthly water consumption at the census tract-level. Through these interdisciplinary studies,

I have made progress in filling knowledge gaps related to the impact of climate change on urban water resources, as well as the impact of people on these water resources.

1. INTRODUCTION

There are many phenomena by which future scientists will characterize the next few decades, among them will be rapid urbanization across the world [1], as well as the unprecedented climatic changes that are unfolding. As humans, we will have to come to terms with these challenges and learn to adapt to, and hopefully mitigate, some of the more deleterious impacts associated with simultaneous urbanization and climate change. One of the areas that will require an immense amount of scientific innovation is access to water resources. With a limited amount of freshwater available and an ever-growing population, providing adequate water supply will be a serious challenge. The pressure brought on by a growing urban population will be exacerbated by climate change, which is expected to lead to intense droughts in some regions and intense precipitation in others. In this sense, part of the world will struggle with having too little water, while the other will have too much—making the task of urban water management increasingly difficult. Given these challenges that society will have to face, there is a pressing need to further the scientific understanding of the impacts of climate change on urban water resources, as well as the development of models that can be used to evaluate those impacts.

In this dissertation, I present a data-driven, interdisciplinary approach to studying both water supply and demand in the context of urban areas. In particular, I will focus on three main gaps that exist within the literature on urban water availability: (i) the lack of model comparison in water supply studies, which leads to methodologies and analyses that are siloed; (ii) the lack of consideration of interconnected systems in water demand analyses, which can lead to suboptimal management decisions by water utilities and policymakers; and (iii) the exclusion of behavioral data in many engineering studies on urban water demand, which can create issues for demand projections, as well as successful intervention implementation.

1.1 Research Goal

The goal of this research is to evaluate urban water resources, with a particular focus on the impact of climate change and human behavior. I aimed to achieve this goal through the integration of data science, hydroclimatology, and normative behavior science, such that the final result is an interdisciplinary analysis of the relationship between people, water, and climate.

1.1.1 Objectives

In order to address the aforementioned gaps in the literature and to achieve the goal described above, I have determined three main objectives, which are outlined below.

I. Create and compare several models to predict reservoir levels based on the relevant hydroclimatic conditions.

II. Build a multi-outcome model to characterize the climate sensitivity of the coupled water and electricity demand and predict future demand under climate change.

III. Integrate behavioral data surrounding water conservation into a data-driven model to predict water consumption.

1.1.2 Scope

The scope of this dissertation was to focus on the climate impacts, with human behavioral data only being introduced later (see Chapter 4). In this sense, there was little consideration of the non-climatic factors that also play a role in urban water resources, such as socioeconomic status, housing characteristics, urban and regional culture, etc. Additionally, there are various research domains used throughout the analyses included in this dissertation. For example, some work was done solely in

the Midwestern United States, while other analyses considered other regions around United States. All of the analyses, however, use data collected within urban areas around the United States.

1.2 Background

As discussed earlier, the focus of this dissertation is urban water resources. Water is fundamental to human life—it is essential to improving social equity, promoting just economic development, and protecting the function of the earth system. In fact, global freshwater use has been identified as one of nine planetary boundaries regulating the safe operating space of Earth to support humanity [2]. Thus, freshwater management is one of the most pressing global challenges for sustainable development in the Anthropocene [3]. This is especially true in a world that is becoming increasingly urban. Traditionally, cities have followed a 'hard path' towards water management. That is, there was a focus on finding more sources of water, instead of working towards reducing the water consumption within the city (which is known as a 'soft path') [4]. In recent years, however, cities have begun to integrate both supply and demand management policies (i.e., hard and soft paths toward urban water management) [5]. This practice has been successful in many parts of the world, especially those that are drought-prone. In fact, integrated water management will likely become increasingly important as droughts, especially those related to human activities (i.e., anthropogenic droughts), become more frequent and intense in some regions [6]. However, in order to ensure integrated water management is successful, a deeper understanding of the impacts of climate change and human behavior on urban water resources is needed.

1.2.1 On the Relationship between Water Supply & Climate

One of the major challenges to urban water management is understanding how climate change will impact water supply. In fact, depending on the sources of water,

the climate impacts may be more pronounced in certain cities. For example, surface water sources (e.g., reservoirs) are highly reliant on precipitation and streamflow, which make them more susceptible to droughts [7]. Moreover, most water supply reservoirs have been created by dams. Recently, it was shown that dams, though they help stabilize the long-term fluctuations in streamflow, ultimately fail to mitigate the impacts of climate change on water supply [8]. Therefore it is necessary not only to build an understanding of the impacts of climate change on these surface water resources, but also develop mitigation and adaptation policies that go beyond the physical infrastructure. In this dissertation, I focus on the former challenge, and in particular, the problems associated with predicting water levels in urban reservoirs under climate change.

The research tools for analyzing and predicting reservoir levels range from complex land surface models to simpler water balance models. For example, McDonald et al. used a gridded simulation model that included hydrological processes, climate change data from Global Climate Models (GCMs), and demographic data to predict future water availability in cities located in the developing world [9]. This study found that by 2050, 250 million urban dwellers will likely experience water shortages. In a similar study, the authors analyzed the vulnerability of water resources to climate change on a global scale [10].

Another approach to modeling reservoir volume is to use a water balance model. These types of models tend to be more simple than the land surface models, but are built on the same concepts. There are various types of water balance models, such as the model employed by Tarroja et al. In this study the authors assumed that the change in reservoir volume could be explained by streamflow in (and therefore basin precipitation), water withdrawals, streamflow out, and evaporation. By using downscaled climate data from GCMs, the authors were able to project the water availability in the California reservoir network under climate change [11]. Another common approach is the 'abcd model'. The abcd model is a series of equations representing the various inputs and outputs to a reservoir. In a study by O'Hara

and Georgakakos, they included reservoir release, basin precipitation, water imports, and reservoir surface evaporation as their inputs and outputs. Using San Diego as a case study, the authors were able to predict the water availability under future climate change scenarios [12]. These studies, however, require a lot of data to model the change in reservoir volume. Recent developments, therefore, have focused on modeling wetland volume using a water balance model that is not data intensive. These studies take advantage of the stochastic nature of precipitation to create an analytical probability density function that is not dependent on large amounts of data [13,14]. Building a model that follows a similar methodology is critical for cities that do not have the infrastructure in place to collect the all of the data needed to run a traditional water balance model.

A final and less conventional approach to water supply modeling is statistical learning. The algorithms that fall under statistical learning theory often lead to accurate predictions and have been used in several studies to predict reservoir volume. For example, Ficchi et al. utilized these algorithms to predict the volume of reservoirs created by dams on the Seine River. The focus of this study was predicting flood conditions, so as to prepare downstream municipalities for potential dangers [15]. Likewise, a similar study in California focused on predicting reservoir levels throughout the state [16]. Each of these studies demonstrated the power of statistical learning to predict reservoir levels. Given the success and availability of these different types of modeling techniques, it is important to consider multiple methodologies when modeling urban water supply, as different models may be complementary and provide important information on the reservoir in question.

1.2.2 On the Relationship between Water Demand & Climate

The other major challenge in urban water management involves the demand practices. The demand arguably more difficult to manage, since it deals with people and their behaviors. However, there is an added challenge, which is related to the climatic

conditions that are, in part, responsible for how people decide to use water. Finally, demand management is further complicated by the interconnectivity between urban systems, such as water and electricity. Often, this interconnectivity is thought of from a supply-side point of view—water is needed to generate electricity and electricity is needed to treat and distribute water [17]. However, there is also evidence to suggest that water and electricity demand are also interconnected [18, 19]. This interdependence between water and electricity, referred to as the water-electricity nexus, has gained much attention, especially in regions susceptible to droughts and heatwaves. In these situations (i.e., a concurrent drought and heatwave), the water supply is often limited due to the drought, while electricity needs increase due to the heatwave (and the subsequent increase in air conditioning use). Moreover, in the US, the majority of electricity is generated via thermoelectric power, which requires cooling water [20]. In this sense, if there is already a reduction in available water supply due to the drought, the electric generators may not be able to provide adequate electricity, resulting in rolling blackouts or planned outages [21]. Similar issues can arise in the food and tourism sectors, especially if the region is highly dependent on income from agriculture or water-based recreation.

There are a number of studies that focus on the supply-side perspective (i.e., the water needed for electricity generation and vice versa). One study, for example, projected electricity supply and demand across the US, and found that approximately 20 metropolitan regions will likely see severe water shortages due to the need for increased electricity generation [22]. Specifically, the authors used population growth, utility-estimated capacity increases, and anticipated summer water deficits as metrics to predict the future state of the system [22]. There are many other studies that focus on the impact of water stress and scarcity on electricity generation [20, 23, 24], but there are fewer studies that focus on the electricity used for water (and wastewater) treatment, and even fewer that focus on the residential end-use water-electricity nexus. However, there are many household activities, such as heating water or washing clothes, that require both water and electricity, making it a crucial metric

for understanding urban water availability. There have been a few studies that focus on residential demand nexus and its importance when trying to increase the prevalence of water conservation within a community. For example, Ruddell and Dixon found that converting residential landscaping from mesic (grass) to xeric (drought-tolerant) led to a significant change in the microclimate. Although, the landscaping change reduced the amount of water used, the change in microclimate caused an increase in air conditioning use (the xeric landscaping reduced the amount of moisture in the air, causing a spike in temperature), which ultimately led to more water being used for electricity generation [25]. Similarly, a study in Brazil found that when households implemented rain barrels to reduce their dependence on the centralized water system, the overall electricity consumption increased by 4% [26]. The authors explained that this increase in electricity consumption was due to the diseconomies of scale, and for rain barrels to be effective, they ought to be paired with other conservation measures to reduce the need for centralized sewage (e.g., graywater reclamation). These studies, among others, demonstrate the need to assess both water and electricity demand when trying to reduce residential water use. However, the inclusion of climate in these demand nexus studies is rare, especially those that go beyond the impact of precipitation and temperature.

The limited number of studies that do consider the impact of climate on the water-electricity nexus, generally employ simplistic measures, such as the change in precipitation or temperature, to determine the impact [27]. In one study, for example, the authors considered the impact of precipitation and temperature on the nexus, but failed to consider other critical variables such as evaporation, which can greatly impact water resources [28]. A similar study, which focused on modeling residential water and electricity consumption, included the effect of temperature, but not humidity, which plays a major in role in the *experienced* temperature [29]. Due to the difference between actual and experienced temperature, it is likely that the water-electricity demand nexus is dependent on more variables than just precipitation and temperature. Understanding which variables are important and how they are

related to the coupled water-electricity demand profile is critical for future demand management.

1.2.3 On the Relationship between Water Demand & People

As alluded to previously, human decisions present a challenge in urban water management. Often, utilities try to reduce water demand by 'nudging' their constituents to more sustainable behaviors. However, these nudges may backfire, potentially leading to increased water use throughout the city. In other words, demand management often requires knowledge of the attitudes, beliefs, and values present within a community that lead to various behaviors surrounding water conservation. Many of these attitudes and values are embedded in social norms that we, as people, follow every day, often unknowingly. There are many studies that demonstrate the importance of accounting for social norms in interventions that seek to increase sustainable practices or conservation among a certain population. For example, one study found that by bringing consumers' attention to the amount of electricity they used in comparison to their neighbors, there was a decrease in electricity use equivalent to that which would result from an 11-20% increase in price [30]. This intervention is fairly common and is based in social norm research—the comparison between peers activated a social norm, which allowed the electric utility to reduce demand without raising prices. In another study, a water utility employed a comparison system similar to the one described above—consumers got to see how much water they were using compared to their neighbors, which, again, activated a social norm that led to reduced water use among the residents [31]. These studies, and others like them, demonstrate the importance of social norms in both electricity and water conservation programs. However, there are few studies that integrate the data collected from social norm studies with computational models to predict water demand. The inclusion of this data will likely increase the accuracy of the models, especially at the intra-city scale, as collective behavior plays a major role in the resource consumption of a city.

As described above, there are several gaps in the research on urban water availability. The proposed work will focus on filling a few of those gaps by: (i) comparing different models for predicting urban reservoir levels; (ii) including both interconnected systems and a wider array of climate variables when evaluating urban water demand; and (iii) integrating human behavior data into a data-driven model predicting urban water demand.

1.3 Organization

There are five chapters in this dissertation. The second chapter will discuss work done on the supply side of urban water management. In particular, I will present results from multiple studies using different methodologies to predict the water volume in reservoirs used for urban water supply. The third chapter will then focus on the demand side of the equation. In this chapter, I will present work done on the water-electricity demand nexus, including the development of a model to simultaneously predict water and electricity use, as well as the results of that model being used to make projections of future water and electricity use. The next chapter will continue to focus on water demand, but will look more deeply into the human dimension. I will present results of qualitative interviews and an integration of this data into a computation model for predicting water consumption. Finally, I will conclude this dissertation in chapter five with a summary of the work presented herein, as well as the future work to be done in order to fully understand the impact of climate change on urban water resources.

2. PREDICTING URBAN WATER SUPPLY

A version of Section 2.2 has been previously published in *Scientific Reports*: https://doi.org/10.1038/s41598-018-23509-w.

2.1 Introduction

One of the major tasks for urban water managers is maintaining the reservoirs that provide the city's drinking water as well as reacting to changes brought on by various hydroclimatic phenomena. For example, after a large rain event, water may need to be released downstream to avoid flooding, and during a long-term drought, water use restrictions may need to be implemented to reduce the impact of the drought and conserve water. Preparation for these events is key if one wants to reduce the impacts of flooding or water stress, both of which can cause major ecological, economic, and societal problems. Understanding and predicting urban floods and droughts, often referred to as hydrological droughts [7], is a major focus of the urban resilience community. An important step to improve urban resilience is to understand and predict urban reservoir responses under the various hydroclimatic conditions that lead to flooding and droughts, so that water managers can implement the necessary mitigation policies (e.g., controlled releases or water use restrictions). Moreover, urban water supplies are especially at-risk to future hydrological extremes because of the unprecedented urban growth that is happening around the world. Currently, about 50% of the world population lives in cities, and the World Bank has projected that by 2050, this number will grow to 65% [1]. When paired with a changing hydrological

environment, including an increased likelihood of droughts [32], rapid urban growth puts cities and their watersheds in a vulnerable position.

To minimize these vulnerabilities, water managers must be aware of the likelihood of any major changes in reservoir level that may affect water availability, so that they can begin to prepare and, hopefully, minimize any negative effects. The typical approach to predicting hydrological extremes in reservoirs is through probabilistic analyses. Most notably, de Araújo and Bronstert used a simple volume equation to assess changes in a Brazilian reservoir [33]. They included inputs such as precipitation and streamflow and outputs such as withdrawals and infiltration. This analysis demonstrated the correlation between reservoir level and drought severity. That is, as the drought increases in severity, the reservoir levels decrease. De Araújo and Bronstert also found that small, isolated systems cannot cope with long-term droughts, making it important for cities with these systems to be proactive in their drought planning [33]. Finally, the authors found that hydrological droughts are often out of phase with meteorological droughts, which provides evidence towards the need to evaluate droughts in reservoirs separately than the typical meteorological (i.e., precipitation-based) droughts if we are to improve urban water system resilience.

There are a few studies that have gone beyond the basic volume equation and done predictive studies on reservoirs, including that of Ficchi et al. This study focused on predicting reservoir levels for flood applications on the Seine River [15], where there are several small reservoirs that are designed to control the streamflow and prevent flooding downstream. Prior to this study, the reservoirs were managed based on historical averages, however, when the streamflow was significantly different than the average, this method failed to prevent flooding downstream. The authors leveraged a tree-based model that used weather data from the European Centre for Medium-Range Weather Forecasts (ECMWF) as the input to predict water levels. They found that the model could adequately predict high-flow conditions within the next nine days, which would allow water managers to implement the regulating features and therefore reduce the risk of flooding downstream. However, when they

repeated their analysis for low-flow scenarios, the authors found that they could not accurately predict droughts, likely because droughts require longer forecasts, which cannot always be made with the meteorological data used in this study. Similar work was done by Yang et al. on reservoir discharges in California [16]. In this study, the authors used two different types of tree-based algorithms: classification and regression trees (CART) and random forest, to predict the outflow of the reservoirs. The outflow in this study was the controlled release of water back into the river if the reservoir levels got too high. The results showed that random forest was able to successfully predict when controlled releases should occur, based on the reservoir storage, precipitation, reservoir inflows, runoff, snowpack, and downstream river conditions. Additionally, the authors leveraged cross-validation to avoid overfitting of the model. The predictions of the cross-validated model outperformed the basic run, demonstrating the importance of performing cross-validation during the model selection process. Finally, the authors showed that random forest was also able to predict the storage trajectory of the reservoirs.

Although these studies demonstrate the benefits of various types of hydrological models, there has been little comparison work done. Often, the choice of model is dictated by one's discipline, leading to a silo effect with regard to urban water supply management. In this chapter, I present the results from a few studies aimed at comparing models used to predict urban reservoir levels. First, I present the results from a study comparing different statistical learning models. Next, I show results comparing the most accurate (in terms of prediction) statistical learning model with a stochastic water balance model. Finally, I wrap up with a discussion on the comparison between models, including the pros and cons of all the methods considered in this chapter.

2.2 Statistical Learning Model Analysis

The objectives of this section were to test the performance of different statistical learning techniques in predicting the water levels in Lake Lanier (Atlanta, Georgia, USA) based on the current hydroclimatic conditions and city characteristics, and to determine the best model for the task. We hypothesized that the random forest model would perform the best, as it has been previously used in hydrological studies [15,16].

In the following sections, I discuss the data and statistical learning techniques used to predict water levels in urban reservoirs. Then, I show the results of the various models in terms of predictive accuracy. Finally, focusing in on the most accurate model, I discuss some interpretations of the results and the implications on urban water management.

2.2.1 Data and Methods

To determine the optimal statistical learning model for predicting urban reservoir levels, the city of Atlanta was selected as an initial case study. Later, the cities of Indianapolis and Austin were considered to test the generalizability of the selected model. Finally, the section wraps up with a discussion of the methodology.

2.2.1.1 Site Description

The main focus for this study was Atlanta, Georgia, although Indianapolis, Indiana and Austin, Texas were also included in the analysis. The city of Atlanta obtains nearly 90% of its water from Lake Sidney Lanier [34], a reservoir located northeast of the city on the Chattahoochee River. Atlanta itself is located in the northern part of the state of Georgia, which is in the southeastern United States. For a visual depiction of the location of the city and reservoir, see Appendix A. Atlanta is in a semi-humid climate zone, yet the region regularly experiences severe droughts that are accompanied by drops in reservoir levels [35]. Atlanta is a major metropolitan

area in the United States, currently home to over 470 thousand people within the city limits and 5.7 million people in the metropolitan area [36]. The Atlanta population is heavily dependent on Lake Lanier for its drinking water, making it imperative that it is secured for the future, a task which is complicated by the water laws in the Chattahoochee River basin [35]. If the water managers in charge of Atlantas water supply have knowledge on the hydroclimatic conditions that may lead to reduced water supply, they can better prepare while maintaining adequate supply downstream. The relative dependence on a single source as well as a climate that is prone to occasional severe droughts makes Atlanta an ideal location to study the viability of machine learning techniques to predict urban reservoir water levels. Additionally, the Eagle Creek reservoir, which serves the city of Indianapolis, and Lake Travis, which serves the city of Austin, were included in the analysis in order to test the generalizability of the results obtained from Atlanta. Both reservoirs, like Lake Lanier, are major sources for the cities they serve. The main difference between the cities are the climates and water usage patterns, making them ideal for studying the transferability of the results.

2.2.1.2 Data Description

Data for this study was obtained from several government agencies, including the US Army Corps of Engineers (USACE), the US Geological Survey (USGS), and the National Centers for Environmental Information (NCEI). Specifically, we obtained the reservoir level data from USACE [37], streamflow data from USGS [38], and precipitation, humidity, and temperature data from NCEI [39]. Additionally, we obtained population data from the US Census Bureau [36], water use data from the North Georgia Water Planning district [34], soil moisture from the NOAA Climate Prediction Center [40], and ENSO data from NOAA [41]. The streamflow data was collected from two locations: one 20 miles upstream of the reservoir (USGS site 02331600) and one 30 miles downstream of the city (USGS site 02338000). The

meteorological data was collected from the Atlanta Hartsfield International Airport, which maintains a long-running and accurate weather station southwest of the city (about 45 miles southwest of the reservoir). This station was selected due to the longevity of the data record and the relative quality of the data. Although it is not exactly positioned next to the reservoir, it is close enough that most of the meteorology will not change much between the two locations. The population and water use data are both limited to the city itself, not the metropolitan area. This was done because the North Georgia Planning District specifically separated the city of Atlanta from the remainder of the district. Finally, the soil moisture was collected from the CPC, which is a gridded product. We selected the grids surrounding the reservoir and averaged them to obtain a soil moisture in the area.

The data included daily values from 1965-2016 (those that were not initially daily measurements, were scaled to that resolution). During this period, the reservoir level ranged from 1050.8 to 1076.2 feet with a mean of 1067.1 feet. Likewise, the streamflow (into the reservoir) ranged from 66 to 15800 ft^3/s with a mean of 766 ft^3/s, while the discharge (downstream of the city) ranged from 852 to 58600 ft^3/s with a mean of 3900 ft^3/s. The dew point temperature ranged from -13.6 to 75.7°F with a mean of 49.7°F, the relative humidity ranged from 23.3 to 100.0% with a mean of 68.0%, and the precipitation ranged from 0 to 7.0 inches with a mean of 0.13 inches. Finally, the soil moisture ranged from 271.3 to 673.2 mm/m with a mean of 470.2 mm/m. The distribution of these variables and others can be seen in Figure 2.1.

In this study, the response variable was the reservoir level and the predictors were: streamflow (into the reservoir), precipitation, population, water usage, discharge (downstream of the city), ENSO index, soil moisture, dew point temperature, and relative humidity. These predictors were selected based on a thorough review of the literature on the subject. Specifically, we chose to include streamflow into the reservoir, as it is the most likely determinant of reservoir level. That is, there is unlikely to be a higher reservoir level if the streamflow is lower than usual. For similar reasons, we chose to include population and water usage as predictors. Given

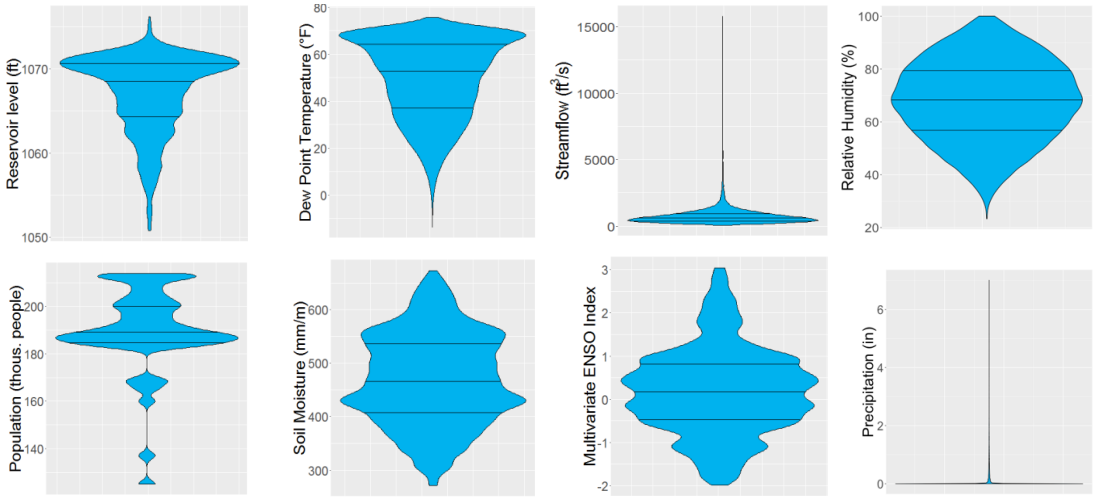

Fig. 2.1.: Violin plot showing the density of six variables used in the statistical learning water supply model: reservoir level, dew point, streamflow, humidity, population, soil moisture, ENSO index, and precipitation. Discharge and water use plots are not shown because they have similar patterns to the streamflow and population plots, respectively.

that one of the main uses of Lake Lanier is providing drinking water for the city of Atlanta [35], it is logical that the population using the water as well as the amount of water consumed will be important variables in predicting reservoir level. Additionally, we included the discharge downstream of the reservoir, which includes both wastewater and overflow discharges. These are related to reservoir level and it was thought that they may provide additional information about reservoir level. Moreover, the streamflow, withdrawals, and discharge were all used in previous studies [16, 33]. We also included a few meteorological variables, including precipitation, dew point temperature, and humidity. These variables were selected as atmospheric measures of the hydrological cycle. In other words, we wanted to include several atmospheric variables that either inherently a part of the water cycle (i.e., precipitation) or closely related (i.e., dew point temperature and humidity, which both have influence over evaporation). Atmospheric variables are easily measured and made available to the public, and, as demonstrated by Ficchi et al. [15], are necessary for predicting reser-

voir levels. Therefore, we decided to include them in our analysis. The decision to include soil moisture as a predictor followed similar reasoning, though it is not an atmospheric variable, it does have a role in the hydrological cycle. Specifically, soil moisture is used a proxy for storage, similar to reservoir level. Therefore, it is reasonable to say that water stored in the soil is not water stored in the reservoir, which affects the water level. Finally, we included the El Niño/Southern Oscillation index because of its known effects in the southeast region of the United States [42]. The ENSO has major climate impacts across the United States, so it is likely that there will be some changes in reservoir level depending on the strength of the El Niño (or La Niña). Overall, the predictors were selected using knowledge gained from the literature review as well as that previously known to the authors. Further details can be found in Appendix A, along with a correlation matrix of the variables.

2.2.1.3 Statistical Models and Analysis

Supervised learning is branch of statistical learning theory in which the response variable guides the learning process. It has been extensively applied to areas ranging from risk and resilience analysis to hydrological modeling [43–46]. Mathematically, supervised learning technique can be described as: $y = f(X) + \epsilon$, where y represents the process of interest (the reservoir level in this study), X represents the series of input variables used to estimate the response (see Appendix A for the variable list), and the noise $\epsilon \sim N(0, \sigma^2)$ represents the irreducible error [47]. The goal of supervised learning is to leverage data and estimate a statistical response surface $\hat{f}(X))$ such that the loss function $L = \int \Delta[\hat{f}(X), f(X)]dX$ is minimized over the entire domain of the independent variable X. Here, Δ represents a measure of distance (e.g., Euclidean distance) between the estimated and actual response functions [47].

Parametric vs. Non-Parametric Models

Supervised learning models vary widely in their degree of complexity, stability, flexibility and interpretability, and can be categorized as parametric, semi-parametric or non-parametric methods. The most popular approach is parametric modeling (e.g., generalized linear regression models) where a parametric function is fitted to the training data (e.g., via mechanisms such as least-squares), such that: $\hat{f}(X) = g(X|(\hat{\beta}_j)_1^p)$. The advantage of parametric modeling is that by assuming a functional form, estimating the complex shape of the response function can be simplified as estimating a set of β parameters, which renders the method simple to compute and interpret. However, such an approach is 'inflexible' and often fails to approximate the true function accurately (since the dependencies in real data are rarely linear). *Non-parametric* models, on the other hand, do not make assumptions about the shape of the function f. Instead, they harness the power of the input data to approximate the function. While they have the advantage of not assuming unrealistic functional form and thereby better approximating the true function, they can be very data-intensive [47].

In this study, we employed several statistical models to predict the reservoir level based on the predictors, these models ranged from parametric to non-parametric. Specifically, we used the: (1) generalized linear model (GLM) [48], (2) generalized additive model (GAM) [49], (3) multivariate adaptive regression splines (MARS) [50], (4) classification and regression trees (CART) [51], (5) bagged classification and regression trees [52], (6) random forest [53], (7) support vector machine (SVM) [54], and (8) Bayesian additive regression trees (BART) [55] methods. These methods were chosen to ensure a variety of algorithms were tested. Descriptions and mathematical representations of these algorithms can be found in Appendix A.

We included linear models such as GLM and more complex additive models such as GAM and MARS, tree-based models, such as CART, random forest and BART, and more complex data-miners, such as SVM. In this way, we can ensure that we have tested the performance of a wide range of statistical learning algorithms, and

not limited to the scope of tree-based models alone. The rational for including a linear parametric model such as generalized linear models is that, as described above, they are highly interpretable and lend themselves easily to statistical inferencing. Generalized additive models and multivariate adaptive regression splines were included because these models relax some of the rigid assumptions associated with generalized linear models, which allows them to achieve higher predictive accuracy compared to the GLM.

A number of tree-based models were included, namely because previous studies have leveraged these algorithms for hydrological applications. Beyond hydrological modeling, tree-based models are widely popular in many different areas because they generally capture the structure of the data well, have an intuitive structure, and lend themselves to interpretations. Regression trees are generally thought of as 'low-bias, high-variance' techniques, meaning while they capture the structure of the data (i.e., they have a low bias), they are not stable and minor perturbations of input data can lead to significantly different tree structures (i.e., they have a high variance). To reduce the variance of tree-based models and improve their stability, meta-algorithms such as boosting and bagging (i.e., bootstrap aggregation) can be leveraged to improve the predictive performance. Bagging trees, as done in the bagged CART model, consists of taking bootstrap samples of the input data and developing a tree model for each sample and then aggregating all of the trees. However, while model averaging is an effective variance reduction technique, its effectiveness is limited if the aggregated trees are correlated to one another. The random forest algorithm addresses this limitation by adding another layer of randomness to the model through randomly sampling a subset of variables for each tree, which reduces the correlation among the trees. Random forest is therefore a low-bias, low-variance technique that yields robust estimates, even in the presence of outliers and noise. The Bayesian additive regression tree method is another robust ensemble-of-trees approach, where the meta-algorithm boosting is applied to the trees. Boosting differs from bagging in that each tree is

used to fit the unexplained variability of the previous tree, ultimately improving the final models variance.

Finally, the support vector machine is a theoretically grounded and powerful machine learning algorithm that leverages hyperplanes to classify the feature space by maximizing the distance between the nearest training data points of any class to the hyper-plane (boundary). To account for non-linearity, the algorithm uses kernel functions to project the non-linear feature space to higher dimensions; using kernel functions, however, significantly reduces the interpretability of the model (particularly in a regression setting). Detailed theoretical foundations and mathematical formulations of the above-mentioned methods are included in Appendix A. It should be noted that there is a host of other flexible, non-parametric machine learning algorithms such as artificial neural networks and generic programming that can account for non-linearities in the data. However, since the goal in this paper is not only prediction, but also making statistical inferencing, such models fall outside the scope of the present analysis. More specifically, while methods such as artificial neural networks can provide robust predictions, due to the transformations of the input space in the inner layers, statistical inferencing cannot be easily implemented [47].

Model performance was assessed based on randomized 5-fold cross validation, such that each fifth of the data was used as a test set for the remaining data. The final error was calculated by averaging the root-mean-squared error (RMSE) of each of the folds. RMSE was chosen as the main measure of error because it penalizes larger deviations more heavily, making it a suitable choice for applications in which large prediction errors are highly undesirable. RMSE represents the out-of-sample (test data) error of the model and is calculated using equation 2.1.

$$RMSE = \sqrt{\frac{\sum (x_P - x)^2}{n}} \tag{2.1}$$

where x_P represents the predicted values, x represents the actual values, and n is the number of observations.

The final model was selected based on the best (lowest) RMSE, and then confirmed through a series of pairwise t-tests. In other words, we compared the results from the model with the lowest RMSE to the two models with the next lowest RMSE values. The t-tests were performed after the Shapiro-Wilk test failed to reject the null hypothesis that the data were normally distributed. The purpose of the pairwise t-tests was to determine if there was a statistically significant difference between the results of each model.

2.2.2 Results & Discussion

Following the methodology described above, the optimal statistical learning model was selected based on predictive performance. In this section, I first show the results from all the models, as well as the model selection process. Then, I discuss the inferencing and analysis performed with the previously selected model.

2.2.2.1 Predictive Performance

The performance of the model was assessed based on the out-of-sample RMSE (see Equation 2.1). This means that the measure of error was calculated on the test set, or the data that was originally held out during the training phase. Generally, the predictive accuracy is lower in this set than the training set (i.e., in-sample) since it contains data that the model has not seen before. In this sense, the out-of-sample RMSE is a way to determine the *predictive* accuracy of a given model. The results from this initial analysis can be found in Table 2.1.

2.2.2.2 Model Selection

The random forest model had the lowest out-of-sample error (1.45) but was closely followed by the support vector machine (RMSE of 1.91) and Bayesian additive regression tree (RMSE of 2.04) models. Therefore, before selecting the random forest

Table 2.1.: Results from the initial performance analysis of the statistical learning models used to predict urban reservoir levels.

Model	In-Sample RMSE	Out-of-Sample RMSE
GLM	3.83	3.83
GAM	3.83	3.16
MARS	3.38	3.41
CART	3.29	3.32
Bagged CART	3.18	3.21
Random Forest	0.66	1.45
SVM	0.64	1.91
BART	1.97	2.04
Null (Mean-Only)	4.59	4.59

model as the final model, we ran two pairwise t-tests to determine if the differences between the models (i.e., random forest vs. SVM and random forest vs. BART) were statistically significant. A Shapiro-Wilk test was performed to confirm the normality of the data prior to performing the t-tests. The results of the t-tests confirmed that the random forest model outperformed the other models in a statistically significant way. That is, both tests demonstrated that the differences between the random forest RMSE and the other RMSE values were statistically significant. Specifically, the t-test between random forest and SVM had a p-value of 1.364×10^{-5} and the t-test between random forest and BART has a p-value of 2.396×10^{-4}.

It is interesting to note that random forest outperformed the more theoretically grounded and complex models (i.e., SVM and BART). This is can be explained by the bias-variance tradeoffas bias decreases, variance increases. It is the goal in statistical learning is to simultaneously minimize both bias and variance. Complex methods tend to do well at minimizing bias, but not variance. Therefore, the best model may be a less complex model that is better able to minimize variance without losing too much accuracy to the increase in bias. Random forest, as described in the previous

section, was designed to minimize the bias and variance, making it a powerful predictive model, even when compared to SVM and BART. The random forest model also showed an improvement of 68% over the null model (i.e., the mean-only model), demonstrating the ability of the model to predict reservoir level beyond the historical averages. This supports our initial hypothesis that the random forest model would perform the best. In addition to having a small error, which demonstrated high predictive accuracy, the random forest model also had a high goodness-of-fit, as demonstrated in Figure 2.2. This figure shows the actual reservoir levels plotted against the fit of the training data (a) and the predicted values of the test data (b), with a 45° line for reference.

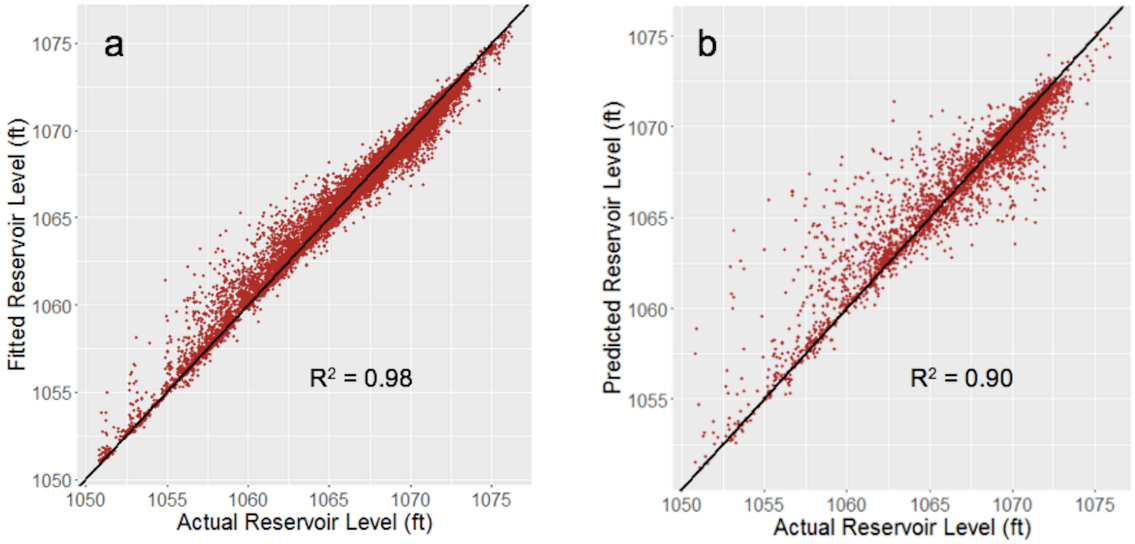

Fig. 2.2.: Actual reservoir levels compared to (a) the fitted values and (b) the predicted values using the random forest model. A 45° line has been plotted for reference.

2.2.2.3 Variable Importance

In addition to knowing which model performs the best, it is important to understand which predictors are contributing the most to the predictive accuracy. That is, which predictors most greatly affect the reservoir level. Variable importance is

measured by ranking the predictors based on their contribution to the out-of-sample accuracy. That is to say, the larger the decrease in accuracy after the removal of a predictor, the more important that predictor is to the final model. As shown in Figure 2.3, the most important variables were the streamflow (into the reservoir), dew point temperature, and population, followed by soil moisture and the El Niño/Southern Oscillation (ENSO) index. Conversely, precipitation was the least important variable when trying to predict reservoir level. Therefore, one could remove precipitation from the model and not lose significant predictive accuracy. In fact, the predictive performance of the model may increase, since the removal of an unrelated variable will reduce the complexity of the model and improve the bias-variance trade-off.

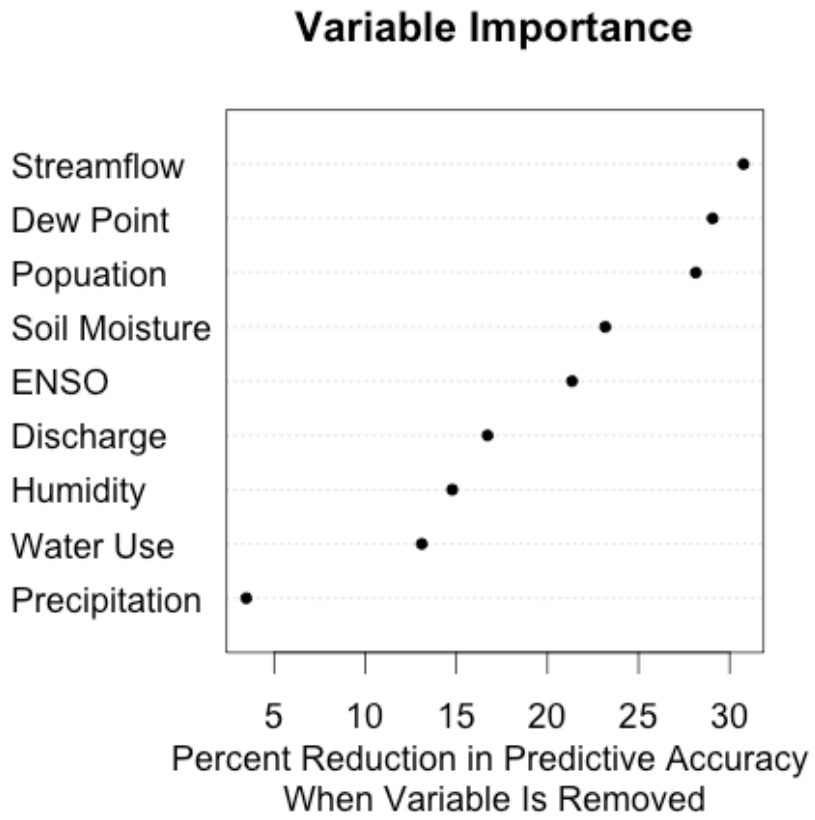

Fig. 2.3.: Predictors ranked by importance in the water supply model. The higher values represent higher contribution to predictive accuracy.

2.2.2.4 Partial Dependence

Partial dependencies are a useful measure for assessing the relationship between the individual predictors and the response variable in nonparametric models [56]. In this project, the partial dependencies were calculated using equation 2.2, as described by Friedman et al. [56].

$$\bar{f}(x) = \frac{1}{n}\sum_{i=1}^{n} f(x, x_{iC}) \tag{2.2}$$

where x is the variable of interest and x_{iC} represents the other variables.

Results from the partial dependence analysis can be used to determine the effects of individual variables on the response, without the influence of the variables. In the case of the streamflow into Lake Lanier, which was the most important variable in predicting the reservoir level, the partial dependence plot for the streamflow in Atlanta is as expected (see Figure 2.4a). Low streamflow means low reservoir levels, but there is a point in which additional streamflow does not influence the water level. This threshold is near the capacity of the reservoir (around 1070 feet), so it is indicative that the managers are releasing water to keep the level at a manageable level. Another important variable was the dew point temperature. The dew point temperature is the temperature at which the air is fully saturated with water vapor. In this study, the mean daily dew point temperature was used as a predictor. As shown in Figure 2.4b, as the dew point increased the reservoir level also increased. A higher dew point is indicative of more moisture in the air, leading to less evaporation and more water staying in the reservoir. In this sense, water managers can assess the state of their water resources by evaluating the streamflow and the mean dew point temperature—a low streamflow with high dew point might not be too damaging, but a low streamflow and low dew point could be cause for concern. Finally, the ENSO intensity was also a relatively important variable in the random forest model. The partial dependence plot can be seen in Figure 2.4c, where there is strong trend towards the presence of an El Niño leading towards higher reservoir levels. The effects

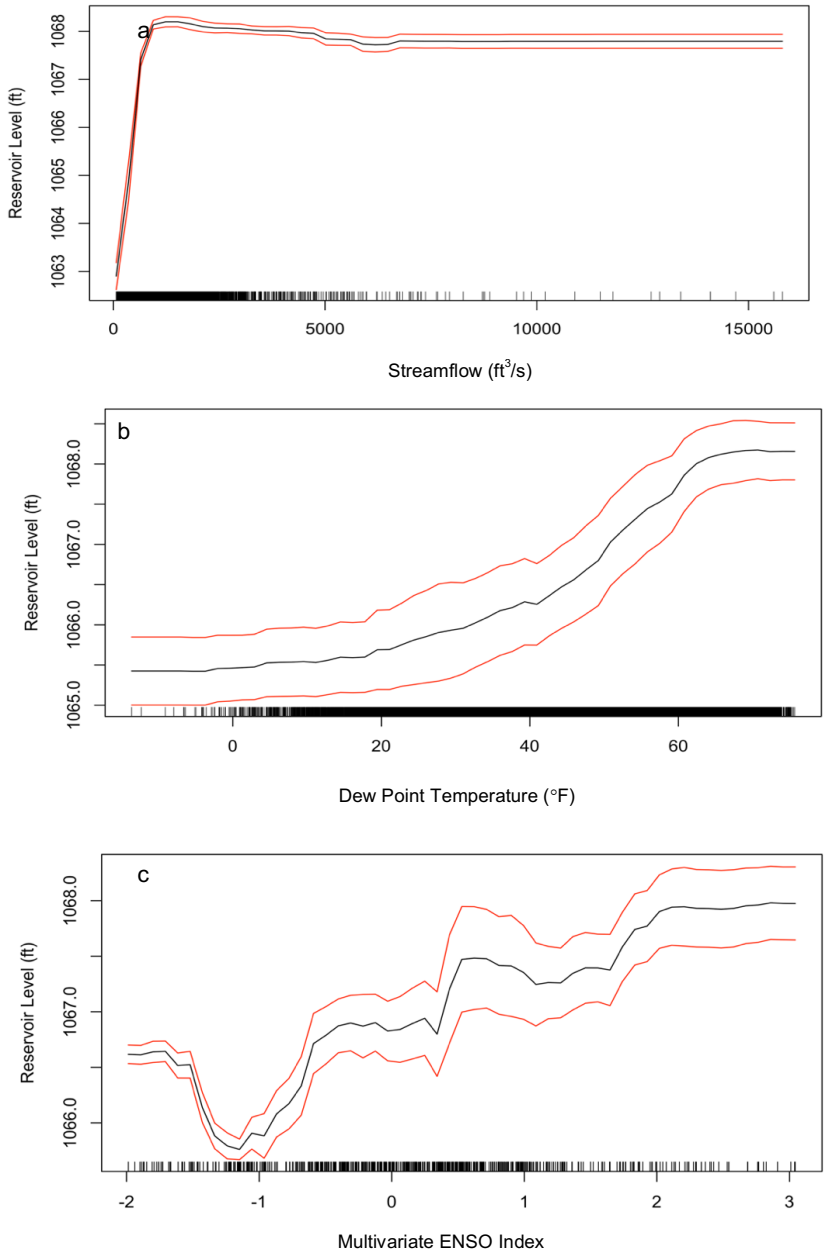

Fig. 2.4.: A selection of the partial dependence plots for the water supply model. The variables shown are: (a) streamflow into Lake Lanier, (b) dew point temperature, and (c) multivariate ENSO index, each with a 95% confidence band and data distribution notches along the x-axis.

of an El Niño in the southeastern United States are increased precipitation and cooler

temperatures [42], which would ultimately lead to more water entering and staying in the reservoir.

2.2.2.5 Comparison of Results to Other Cities

As demonstrated above, the random forest model was the best model for predicting the water level in Lake Lanier. However, this result may be specific to Lake Lanier. To test this site specificity, we ran the same random forest model in two other reservoirs: Eagle Creek (Indianapolis, IN) and Lake Travis (Austin, TX). Similar to Lake Lanier in Atlanta, both reservoirs serve as the main source of drinking water for their respective cities. Likewise, both regions have experienced drought years and wet years within the study period. We found that the predictions from the random forest model greatly outperformed the prediction made by the null (mean-only) model in both cases. Specifically, using the random forest model led to a 55% improvement for Eagle Creek and 92% for Lake Travis. This indicates that the use of the random

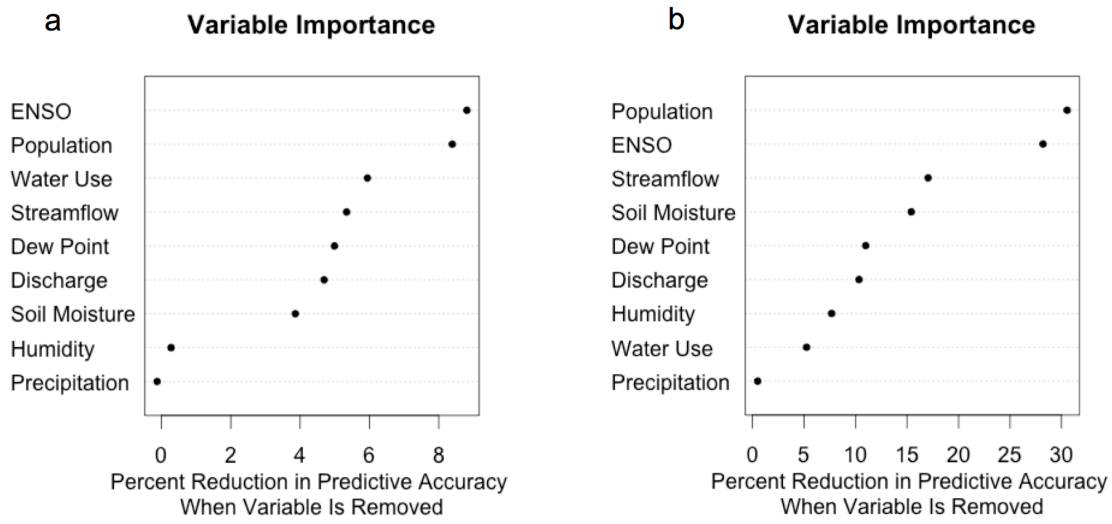

Fig. 2.5.: Variable importance plots for the water supply model for (a) Eagle Creek (Indianapolis, IN) and (b) Lake Travis (Austin, TX)

forest model for predicting reservoir levels can be transferred to other cities. The main difference between the three cities was the important variables in the random forest model. As discussed earlier, the most important variables in the Lake Lanier analysis were streamflow, dew point temperature, and population. However, in the Eagle Creek reservoir, the ENSO index, population, and water use were the most important variables (see Figure 2.5a). Finally, in the Lake Travis analysis, population, ENSO index, and streamflow were the most important variables (see Figure 2.5b). This shows that although the random forest model is transferrable between different cities, there is still a need for site-specific studies to determine the important predictors.

2.2.3 Summary

This study focused on determining the most accurate statistical learning technique for predicting reservoir levels based on the current hydroclimatic conditions. We hypothesized that random forest would perform well, as tree-based methods have been successful in predicting reservoir conditions [15, 16]. The results support this hypothesis and extend the practice of applying supervised learning techniques to long-term analyses, especially those focusing on drought. Though initially focused on Lake Lanier in Atlanta, the results were generalizable in two other cities with different climates and water use patterns. This indicates that implementing a random forest model in urban water management scenarios can be useful, especially if one wants to understand the important variables effecting reservoir levels. That being said, our analysis showed that, although the model was the same, the important variables differed between reservoirs. Each reservoir is going to be different, therefore it is important that future studies include a site-specific analysis to determine the important variables. In Atlanta, the most important variables are streamflow and dew point temperature. This means that a deviation from normal, specifically a decrease, in either one or both variables could be indicative of a decrease in reservoir

level that has the potential to develop into a hydrological drought. This is crucial information for water managers who must make decisions about drought declaration and water use restrictions. In Indianapolis and Austin, the climates are different than Atlanta and therefore, the important variables are different. Specifically, both reservoir levels in Indianapolis and Austin have a high dependency on population and ENSO. The relative importance of population in all three cities is likely because as the cities grow, more water is consumed, even in the presence of conservation measures.

The ENSO index is another variable that affects all three cities, although it is more important in Indianapolis and Austin. The ENSO is a large-scale climate process that is formed in the Pacific Ocean when there are major shifts in sea surface temperature. Although the ENSO occurs in the Pacific Ocean, it has major effects on the climate around the world, including changes in precipitation over the US that may lead to drought. We used the NOAA Multivariate ENSO Index (MEI) to describe ENSO intensity. This index runs from -2 to 3, with more negative values indicating a strong La Niña and more positive values indicating a strong El Niño [41]. The effect of ENSO on reservoir levels is an important one, as it is a predictable climatic phenomenon, and therefore, knowing the effects on water availability could greatly impact a city's ability to prepare for a potential drought. Interestingly, in all three cities, precipitation was the least important variable. This is likely because we used daily precipitation data to predict the reservoir level of that same day. It is more likely that a weekly accumulation of precipitation will have a greater impact on the reservoir level than the precipitation that day. Overall, this study demonstrated the ability of the random forest model to accurately predict reservoir levels, given the current hydroclimatic conditions and city characteristics, for three different cities.

2.3 Stochastic Water Balance Model Analysis

In contrast to the random forest algorithm, the water balance model is based in traditional hydrological modeling. At its simplest, this model is an input-output

model aimed at calculating the change in storage of a body of water. There are inputs, such as streamflow (and precipitation, and outputs, such as discharge and evaporation. These inputs and outputs can be combined to estimate the change in storage over a given time step. The objective of this study was to develop a model to evaluate changes in urban reservoir storage that has a physical basis beyond that of a computational algorithm. In the following sections, I discuss the data and methods used to develop this water balance model, as well as the results from applying the model in a variety of water supply reservoirs.

2.3.1 Data and Methods

To build and evaluate the water balance model, nine reservoirs were selected as case studies. These reservoirs are fairly spread out across the country, and are considered to be managed for a variety of purposes (e.g., drinking water supply, flood control, recreation, etc.). In this section, I first discuss these nine reservoirs, followed by the data collection and methodology.

2.3.1.1 Site Description

There were nine main reservoirs selected for this part of the study, as shown in Table 2.2. These reservoirs were chosen, in part, due to the availability of data. Additionally, they are all managed reservoirs used as water supply for their respective cities. Finally, the reservoirs are in different regions across the United States, allowing for the transferability of the water balance model to be assessed.

2.3.1.2 Data Description

There were six main variables collected for this analysis. The dependent variable was water level (which was ultimately transformed into reservoir volume) in the reservoirs listed in Table 2.2. This data was collected from a variety of sources, ranging

Table 2.2.: Reservoirs considered in the development of the water balance model for urban reservoirs.

Reservoir	Location	Purpose[1]
Chester Morse Lake	Seattle, WA	Water Supply
South Fork Tolt Reservoir	Seattle, WA	Hydroelectric, Water Supply
Falls Lake	Raleigh, NC	Flood Control, Water Supply, Recreation
Lake Mead	Las Vegas, NV	Irrigation, Hydroelectric, Water Supply
Lake Travis	Austin, TX	Irrigation, Hydroelectric, Water Supply
O'Shaughnessy Reservoir	Columbus, OH	Hydroelectric, Water Supply, Recreation
Hoover Reservoir	Columbus, OH	Water Supply
Lake Hefner	Oklahoma City, OK	Water Supply
Eagle Creek Reservoir	Indianapolis, IN	Flood Control, Water Supply

[1]Based on the US Army Corps of Engineers Inventory of Dams [57].

from the US Geological Survey to the US Army Corps of Engineers. The independent variables were split into inputs and outputs to the reservoir. The inputs were precipitation and streamflow into the reservoir, collected from the NCEP Reanaysis dataset [58] and US Geological Survey [59], respectively. The outputs were evaporation, water withdrawals, and streamflow out of the reservoir, collected from the NCEP Reanalysis dataset [58], local utilities, and US Geological Survey [59], respectively. Additionally, the infiltration into groundwater was considered as an output, however, there are no data sources detailing this loss. Therefore, the infiltration was estimated based on a ratio calculated between evaporation and infiltration, based on previous work on balancing the water within reservoirs [33]. Although data was collected for the entire year, this analysis only considers the summer months, as that is a critical time for predicting reservoir levels, and there are less confounding factors,

such as snow or ice melt, which may not be included in the input/output variables. Additionally, the seasonal variability plays a major role in many of reservoirs selected in this study, so separating the data into seasons was an important step.

2.3.1.3 Methodology

In this study, a water balance equation was developed to evaluate the change in storage. As shown in Equation 2.3, the inputs were precipitation (P) and streamflow into the reservoir (Q_{in}). The outputs were evaporation (E), water withdrawals (L), streamflow out of the reservoir (Q_{out}), and infiltration into groundwater (I).

$$\Delta V = (P + Q_{in}) - (E + L + Q_{out} + I) \tag{2.3}$$

It is important to note that the infiltration term was calculated as a ratio that related infiltration to evaporation on a per-reservoir basis. This was one by back-calculating the infiltration term using the known volume data for every step in the time series. Then, a ratio of infiltration to evaporation was calculated for every point and the average was considered in the final equation. In other words, I in Equation 2.3 can be written as $x \times E$, where x is the average ratio between the back-calculated infiltration and known evaporation. Here it is necessary to mention that this back-calculation did not exclude any error that may be caused by the data or variability in the system. In this sense, the infiltration terms might be higher than expected. However, this does not necessarily have an impact on the equation, as the word *infiltration* could be replaced with *unaccounted for losses*, which would include infiltration, errors, variability in the system, etc. The ratios calculated in this study are listed in Table 2.3. In particular, Eagle Creek and South Fork Tolt have large ratios. Both reservoirs are in areas where infiltration is expected to be large, however, the large ratio most likely indicates a significant amount of unaccounted for losses beyond the infiltration. This could indicate issues in the data or in the assumptions. For example, in this study the streamflow was originally obtained as instantaneous

readings in ft^3/s and later aggregated to daily values. In order to aggregate, it was assumed that the average instantaneous reading could be applied to the entire day. That is, if the average instantaneous discharge was 50 ft^3/s, it was assumed that in every second of day, 50 ft^3/s flowed out of the reservoir. It is possible that due to the management of Eagle Creek and South Fork Tolt, that the discharge fluctuates over the course of the day, potentially being much higher than the assumed daily value. This would, in turn, lead to losses that were unaccounted for, and hence, the high ratio.

Table 2.3.: Empirical ratios of infiltration to evaporation.

Reservoir	Ratio
Chester Morse Lake	1.8
South Fork Tolt	38
Falls Lake	3.6
Lake Mead	4.6
Lake Travis	1.9
O'Shaughnessy	5.2
Hoover Reservoir	5.5
Lake Hefner	0.25
Eagle Creek	35

Using Equation 2.3 to calculate the change in storage, a time series was built containing the *modeled* reservoir volume. The modeled reservoir data was compared to the actual reservoir data using statistical moments. In particular, the mean and coefficient of variation (CV) were considered. Additionally, both a t-test and a Kolmogorov-Smirnov test were performed to test for any statistically significant differences between the means and distributions of the modelled and actual data, respectively. Ideally, there would be no statistically significant differences between the modelled and actual data. This would indicate that the water balance model is adequately representing the real system. Should the means match but the CVs differ, it is indicative that there is inherent variability in the system that is not being

captured in the data, whether that be through infiltration, evaporation, or any of the other variables.

An important note is that Equation 2.3 calculates the change in storage (i.e., volume), while the data collected was for the stage (i.e., water level). In order to convert the stage data to volume, the bathymetry data was obtained from the US-ACE National Dam Inventory [57]. Using equations outlined in the literature [13,14], I was able to estimate the volume based on the stage. This method, however, could result in an erroneous estimation of volume due to the propagation or error from the bathymetry data. That being said, since the data is collected and maintained by the US Army Corps of Engineers, it is likely that the data is accurate. Additional errors could be introduced at the measurement level, though, especially since stage measurements could be different at different points in the day and at different locations within the reservoir. In order to manage the temporal fluctuations, the average value was calculated, based on sub-daily readings. Unfortunately, there is little to be done about the spatial fluctuations in stage. This is certainly a limitation in the data collection and something to be aware of when analyzing the results of the study.

In addition to evaluating the efficacy of the water balance model, I analyzed the differences between upstream and downstream flows within these nine reservoirs. Based on a study by Ferrazzi and Botter [8], I estimated the mean and CV of the streamflow in and out the reservoir and compared the differences. It has been shown, for example, that reservoirs used primarily for water supply cause the mean to decrease downstream (compared to upstream), while the CV increases [8, 60]. This is indicative of overall reductions in downstream flows, but also more erratic flows, which ultimately have impacts on the ecology and hydrology further downstream [8]. In this chapter, I follow this same method to evaluate the differences in these reservoirs and explore the implications of these findings alongside the modeling results.

2.3.2 Results & Discussion

In this section I will discuss the results of the modeling work, as well as the test on streamflow consistencies. Starting with the streamflow analysis, I will present the results from the moment analysis and discuss the implications in terms of the type of reservoir and the impacts on the local hydrology. Then, using this initial discussion on reservoir types as a backdrop, I will present the results from the water balance model and discuss any differences between the reservoirs, in terms of practical use and the local climate.

2.3.2.1 Streamflow Analysis

Understanding the impact of urban reservoirs on the local hydrology is critical to informing policy on the creation of new reservoirs or the maintenance of current ones. In this section, the downstream impacts of nine reservoirs were analyzed using statistical moments, which are shown in Table 2.4. Previous work has shown that dams tend to reduce the mean streamflow, but increase the coefficient of variation [60].

Table 2.4.: Moment analysis on the streamflow in and out (i.e., streamflow and discharge, respectively) of the reservoirs. Note that there is no outflow from Lake Hefner because it is a terminal reservoir for drinking water supply.

Reservoir	Streamflow		Discharge	
	Mean (ft^3/day)	CV	Mean (ft^3/day)	CV
Chester Morse Lake	9568206	1.27	1679596	0.607
South Fork Tolt	2144326	1.36	5826577	0.082
Falls Lake	9154625	2.43	24761513	1.71
Lake Mead	1276361096	0.22	1218669324	0.13
Lake Travis	812224933	2.4	64943497	1.12
O'Shaughnessy	5388000	0.83	8297280	0.45
Hoover Reservoir	481528	2.4	12685553	0.075
Lake Hefner	5025325	1.74	—	—
Eagle Creek	14480526	1.38	7716017	1.85

In other words, downstream of reservoirs, one can expect to see reduced means but increased CVs in flow. In the reservoirs considered in this study, this pattern only holds true for Eagle Creek. In fact, although the change in mean varies across the reservoirs, most lead to a reduction in CV. This indicates that the dams are actually reducing the variance in the streamflow, rather than increasing it. This reduction in CV might be due to the nature of the reservoir. Recent work, for example, found that in the Eastern United States, reservoirs used for water supply result in less streamflow variability [8]. This reduction was attributed to the need for storage, as well as the constant withdrawal of water. In other words, water supply reservoirs are maintained such that they remain close to the same level in the long-term (under normal conditions). This means that any water being withdrawn needs to be replenished by the streamflow and only the 'left over' water will be released downstream. Since water supply operations run year-round, this creates a situation where the downstream releases are somewhat constant, thus the reduction in streamflow variability.

In the reservoirs studied here, there is a similar pattern. For example, Chester Morse Lake, Lake Mead, and Lake Travis are the main sources of water supply for Seattle (WA), Las Vegas (NV), and Austin (TX), respectively. These reservoirs all create reductions in streamflow mean and CV, as expected of water supply reservoirs. South Fork Tolt Reservoir and O'Shaughnessy Reservoir are also primary sources of water (Seattle, WA and Columbus, OH, respectively), but also provide additional services, namely hydroelectric power generation. In terms of hydroelectric power generation, it is important to maintain a steady flow to ensure consistent generation. These reservoirs therefore have to balance the need to store enough water for the supply but also discharge enough to generate electricity. In this sense, one can expect and increase in mean (for the power generation) and a decrease in CV (to keep a steady flow). Interestingly, the two reservoirs that are used for flood control, Falls Lake (Raleigh, NC) and Eagle Creek (Indianapolis, IN), have opposite behaviors—Falls Lake increases the mean and reduces the CV, while Eagle Creek reduces the mean and increases the CV. This is likely due to the differences in the *primary* function

of the reservoirs. Falls Lake, for example, is primarily used as water supply for the city of Raleigh, NC, but will occasionally be used as flood control. In this sense, the managers need to maintain storage to ensure the water supply, but are more likely to discharge more water than the inputs to limit the chances for overflow. As such, although they may be discharging more to remain at a certain water level, it is likely to be a certain amount every day due to the management for water supply. Eagle Creek, on the other hand, is primarily used for flood control, with only minor withdrawals for water supply. Therefore, the reservoir is kept at roughly the same level until a major event happens, after which a significant amount of water is released downstream. This creates overall reductions in the mean, but increases in the variance, since these releases are only occurring after major events and not on a regular basis. Overall, this analysis demonstrated the importance of considering the reservoir purpose when evaluating the impact that reservoirs will have on the surrounding environment. It is also important, however, to go beyond the impact that reservoirs have and also study the reservoirs themselves.

2.3.2.2 Water Balance Modeling Results

The main goal of this chapter is to compare various ways to assess urban reservoir storage, one of those ways to to build a water balance model that takes into account the various inputs and outputs to the system. Here, nine reservoirs were modeled using the water balance method. The results from the moment analysis can be found in Table 2.5. In addition the the moment analysis, several statistical tests were performed on the data. The results from these tests, namely Welch's t-test and the Kolmogorov-Smirnov test, are shown in Table 2.6.

The water balance model (see Equation 2.3) was applied with varying degrees of success across the nine reservoirs, based on the moment analysis. In most reservoirs, with the exception of South Fork Tolt, the means were fairly close. In fact, in over half of the reservoirs studied, there was less than a 5% difference between the mean of

the actual volume and the modeled volume. Of these reservoirs, Falls Lake (Raleigh, NC) had the closest match, with only 0.16% difference between the means. Other well-modeled reservoirs were Hoover Reservoir (0.29%), Lake Mead (0.6% difference), Lake Travis (1.02% difference), Lake Hefner (3.3% difference), and O'Shaughnessy Reservoir (3.9% difference), as shown in Table 2.5. There model also modeled the mean fairly well in Eagle Creek (7.1% difference) and Chester Morse Lake (13% difference), but did poorly in South Fork Tolt (41.9% difference). This poor result may be related to the large ratio (see Table 2.3) used in the analysis. South Fork Tolt was one of two reservoirs, the other being Eagle Creek, that had exceptionally large infiltration to evaporation ratios. Earlier I hypothesized that this was probably due to unaccounted for losses, potentially in the discharge from the reservoir. It is interesting to note that Chester Morse Lake and South Fork Tolt are incredibly close, geographically, and both serve as the water supply for the city of Seattle, WA. Logically, one would expect the model to perform well in both cases, since they are very similar. In fact, one of the few differences between these two reservoirs are they purposes. As shown in Table 2.2, Chester Morse Lake is only used for water supply, while South Fork Tolt is also used as a sources of hydropower. In fact, Chester Morse Lake provides nearly 80% of Seattle's drinking water, while South Fork Tolt is used

Table 2.5.: Results from the moment analysis on both the actual and modeled reservoir volume.

Reservoir	Actual Data		Modeled Data		Difference (%)	
	Mean (ft^3)	CV	Mean (ft^3)	CV	Mean	CV
Chester Morse Lake	2.78×10^9	0.11	3.17×10^9	0.26	13	81.5
South Fork Tolt	1.99×10^9	0.186	1.31×10^9	0.295	41.9	45.7
Falls Lake	4.32×10^{11}	0.0015	4.31×10^{11}	0.0017	0.16	7.1
Lake Mead	1.06×10^{12}	0.085	1.07×10^{12}	0.056	0.6	40.9
Lake Travis	7.43×10^{10}	0.06	7.51×10^{10}	0.054	1.02	10.6
O'Shaughnessy	6.22×10^8	0.021	5.98×10^8	0.027	3.9	23.2
Hoover Reservoir	2.92×10^9	0.046	2.91×10^9	0.046	0.29	0.12
Lake Hefner	2.79×10^9	0.129	2.70×10^9	0.22	3.3	52.6
Eagle Creek	9.55×10^8	0.073	1.02×10^9	0.93	7.1	171

to supplement the remaining 20%. This difference may explain the discrepancies between the two reservoirs. That being said, the Chester Morse Lake and South Fork Tolt models are the two that are most different from the actual data, in terms of mean, which may be indicative of a geographic issue. These two reservoirs also had two of the largest differences in CV, though not the largest. In fact, Lake Hefner and Eagle Creek, which were fairly well represented in terms of the mean, did not represent the CV well, with 52.6% and 171% differences, respectively. Interestingly, Hoover Reservoir (Columbus, OH) was best represented in terms of the modeled CV. However, in this reservoir, as well as O'Shaughnessy Reservoir, the dataset was limited to only three months of data (July - September 2016). This small dataset meant that there was very little variability to model in the first place. The reservoirs with longer recording periods, especially those in areas that have experienced drought, would have been more difficult to model due to the added complexity. In fact, looking at Lake Mead, a 0.6% difference in mean and a 40.9% difference in CV is actually quite good and indicates a model that is able to adequately represent the system. Moreover, since the CV values are so small (due to the managed nature of the reservoirs), even small changes constitute a large difference. For example, the difference between the actual and modeled CV for Lake Mead is less than 0.03, but that is a 40% difference. With this in mind, it is important to not just look at the percent difference, but also take into account the scale of the actual CV values. Additionally, it is important to evaluate the statistical significance of the differences between the actual and modeled data, for which we can use a variety of different analyses.

For this analysis, I selected the Kolmorogov-Smirnov test to evaluate the differences in the distribution of the data [61, 62] and Welch's t-test to evaluate the differences in the mean [63]. In the Kolmorogov-Smirnov test, the D statistic is computed, which represents the maximum difference between the empirical cumulative distribution functions of the two datasets. The larger the D statistic, the larger the difference between the two datasets. Additionally, the null hypothesis for the two-sample test is that the two samples come from the same distribution. Thus a p-value

Table 2.6.: Results from the statistical tests between the actual and modeled reservoir volume. The Kolmogorov-Smirnov test evaluates the difference between the distributions of the data and the t-test evaluates the difference in means. In both tests, a p-value less than 0.01 indicates there is a statistically significant difference in the distribution or mean, depending on the test.

Reservoir	Kolmogorov-Smirnov test		Welch's t-test	
	D statistic	p-value	t statistic	p-value
Chester Morse Lake	0.465	$< 2.2\times10^{-16}$	14.18	$< 2.2\times10^{-16}$
South Fork Tolt	0.61	$< 2.2\times10^{-16}$	-42.8	$< 2.2\text{e}\times10^{-16}$
Falls Lake	0.497	$< 2.2\times10^{-16}$	-24.34	$< 2.2\times10^{-16}$
Lake Mead	0.268	$< 2.2\times10^{-16}$	2.13	0.034
Lake Travis	0.36	6.04×10^{-13}	1.9	0.057
O'Shaughnessy	0.69	5.77×10^{-8}	-7	1.49×10^{-9}
Hoover Reservoir	0.059	1	-0.026	0.7959
Lake Hefner	0.33	$< 2.2\times10^{-16}$	-4.67	3.20×10^{-6}
Eagle Creek	0.496	$< 2.2\times10^{-16}$	2.31	0.021

below the significance level rejects the null hypothesis and indicates a statistically significant difference between the two distributions. Welch's t-test is a variation of the student's t-test [64] that relaxes the assumption of equal variance between the samples. Given the differences between the CVs of the actual and modeled data (see Table 2.5), Welch's test was selected over the student's t-test. The null hypothesis in this test is that the two means are equal, therefore, a p-value below the significance level leads to a determination that there is a statistically significant difference in the means. Ultimately, these tests can indicate similar information as the moment analysis, but have the advantage of representing the statistical significance of the results.

The results of this analysis, which are shown in Table 2.6, indicate that there were statistically significant differences in the distributions and means of the actual and modeled data in the majority of the reservoirs. In fact, using a significance value of 0.01, only Hoover Reservoir was found to have no difference in the distribution. On the other hand, Lake Mead, Lake Travis, Hoover Reservoir, and Eagle Creek were shown to have no difference in the mean. However, given the small sample size

of the Hoover Reservoir sample, the results of the statistical test are not entirely trustworthy. In this sense, it can be said that of the reservoirs with a large enough sample size, none of the distributions were equal. In terms of the mean, there were more similarities, however, it is interesting to note some major discrepancies between the percent difference calculations and the t-test results. For example, Falls Lake had a 0.16% difference in means between the actual and modeled data (see Table 2.5), but the t-test called for a rejection of the null hypothesis, indicating a significant difference between the means. This could be a consequence of the reservoir size. The average volume of Falls Lake is 432 billion ft^3, so even a small percentage like 0.16% represents a difference in 691 million ft^3. In this sense, though there is little difference between the means, the sheer size of the reservoirs means that a small percentage is a huge number. That being said, overall, the water balance model does perform well. Moreover, the reservoirs that were modeled well are in a variety of different climate zones and geographic areas, which demonstrates the generalizability of the model.

2.3.3 Summary

In this section I discussed the development of a water balance model, which I applied to nine difference reservoirs across the United States. The model considered a variety of inputs and outputs that impact the reservoir storage. Ultimately, the results showed that the model performed fairly well in most reservoirs, although it failed to represent the reservoirs located in Pacific Northwest. In particular, Lake Mead (Las Vegas, NV) and Falls Lake (Raleigh, NC) were best represented by the water balance model. Evaluating the moments of the actual volume data and the modeled data demonstrated that the mean is much better represented by the water balance model than the variance. This is indicative of some natural variance that is not included in the model. This may be caused by the aggregation to daily data, which assumed that an instantaneous value (such as streamflow) could be applied for the entire day. It is probable that there is some intra-daily variability that is missing

from this analysis. Additionally, the results from the Kolmorogov-Smirnov test and Welch's t-test were presented. The Kolmogorov-Smirnov test indicated that there were statistically significant differences in the actual and modeled distributions for all the reservoirs, except Hoover Reservoir (Columbus, OH), which had too few data points to be considered accurate. In terms of the Welch's t-test, there were mixed results, with about half of the reservoirs showing no difference between the actual and the modeled mean. Ultimately, however, the results show an acceptable amount of similarity between the actual and modeled data, indicating the ability of the water balance model to represent urban reservoir volumes.

2.4 Model Comparison

The goal of this chapter was to compare different models used to predict urban reservoir storage and demonstrate the complementary nature of this models. In this section, I compare the results from the random forest algorithm (see Section 2.2) and the water balance model (see Section 2.3). Following a similar methodology as above, I will present the moment analysis and the results of the t-test and Kolmogorov-Smirnov test. However, in addition to looking at the ability of the model to evaluate the level within the observational space, I will also look at projections. These projections were based on an assumption that the input data for a particular day would be equal to the climatological mean for that same day. The climatological mean, though unlikely to be exact, would be one of the tools available for reservoir managers to assess future reservoir conditions. Additionally, I will discuss the pros and cons of the various methods and the ways in which they might be complementary.

2.4.1 Comparison in the Observational Space

In order to compare the results from the water balance model with those from the random forest model, I first calculated the mean and CV for each of the nine reservoirs considered in Section 2.3. As shown in Table 2.7, the random forest model represents

Table 2.7.: Results from the moment analysis on both the actual and modeled reservoir volume (random forest method).

Reservoir	Actual Data		Modeled Data		Difference (%)	
	Mean (ft^3)	CV	Mean (ft^3)	CV	Mean	CV
Chester Morse Lake	2.77×10^9	0.04	2.77×10^9	0.039	0.04	3.5
South Fork Tolt	2.00×10^9	0.047	1.98×10^9	0.039	1.12	19.3
Falls Lake	4.07×10^{11}	0.0069	4.07×10^{11}	0.0034	0.057	67.8
Lake Mead	1.07×10^{12}	0.014	1.06×10^{12}	0.019	1.23	26.9
Lake Travis	7.43×10^{10}	0.014	7.43×10^{10}	0.008	0.082	60.3
O'Shaughnessy	6.22×10^8	0.021	6.22×10^8	0.014	0.045	41.6
Hoover Reservoir	2.92×10^9	0.046	2.92×10^9	0.021	0.016	76.4
Lake Hefner	2.79×10^9	0.05	2.79×10^9	0.016	0.01	103
Eagle Creek	9.55×10^8	0.023	9.51×10^8	0.017	0.43	29.7

the mean quite well, with less than a 2% difference in each reservoir. The CV, on the other hand, is more variable. In some cases, such as Chester Morse Lake and South Fork Tolt, the difference is less than 20%. In other cases, including Lake Hefner and Falls Lake, the difference is well over 50%. Interestingly, Chester Morse Lake and South Fork Tolt were among the most poorly modeled reservoirs in the water balance analysis, while Lake Hefner and Falls Lake were modeled fairly well. This indicates that the water balance model and random forest model might be better suited for different reservoirs or regions. Chester Morse Lake and South Fork Tolt, are located in the Pacific Northwest. The fact that these two geographically similar reservoirs are poorly predicted in the water balance model, but accurately predicted in the random forest model indicates that there is likely some nonlinear relationship between the reservoir volume and the inputs/outputs. One of the benefits of predictive modeling algorithms is the ability to leverage of nonlinear relationships, which can later inform additional models, such as the water balance model.

Looking at the statistical tests used to evaluate the data (Table 2.8), there appears to be less statistically significant differences between actual data and the random forest model results. Considering a significance level of 0.01, the results of the Kolmogorov-Smirnov test indicate that there is no statistically significant differences

Table 2.8.: Results from the statistical tests between the actual and modeled reservoir volume (random forest method). The Kolmogorov-Smirnov test evaluates the difference between the distributions of the data and the t-test evaluates the difference in means. In both tests,a p-value less than 0.01 indicates there is a statistically significant difference in the distribution or mean, depending on the test

Reservoir	Kolmogorov-Smirnov test		Welch's t-test	
	D statistic	p-value	t statistic	p-value
Chester Morse Lake	0.073	0.64	0.103	0.92
South Fork Tolt	0.178	0.0019	-2.72	0.0066
Falls Lake	0.155	0.0034	-1.21	0.229
Lake Mead	0.271	9.70×10^{-10}	-8.83	$< 2.2\times10^{-16}$
Lake Travis	0.422	0.00056	-0.328	0.744
O'Shaughnessy	0.222	0.336	-0.107	0.915
Hoover Reservoir	0.265	0.186	0.018	0.985
Lake Hefner	0.25	1.75×10^{-7}	-0.022	0.982
Eagle Creek	0.179	0.0032	-2.12	0.035

in the distributions of Chester Morse Lake, O'Shaughnessy Reservoir, and Hoover Reservoir. This is considerably better than the water balance model, in which all the reservoirs were found to have statistically significant distributions. However, the small sample size of O'Shaughnessy and Hoover Reservoirs is not ideal for these statistical tests, so the results may not be the same as they would be given a larger sample size. Similar to the results for the percent difference between the actual and modeled data, the Welch's t-test results indicate little difference between the means of the datasets. In fact, in all the reservoirs other than South Fork Tolt and Lake Mead, the test indicates that there is no statistically significant differences between the means. Lake Mead was accurately predicted by the water balance model, with a t-test result indicating a slight significant difference, which would be considered not significant under a different significance level. This result suggests that the level of management in the Colorado River Basin, and subsequently in Lake Mead, might allow for the better representation using a water balance model, while some of the less managed reservoirs are better modeled by the random forest model.

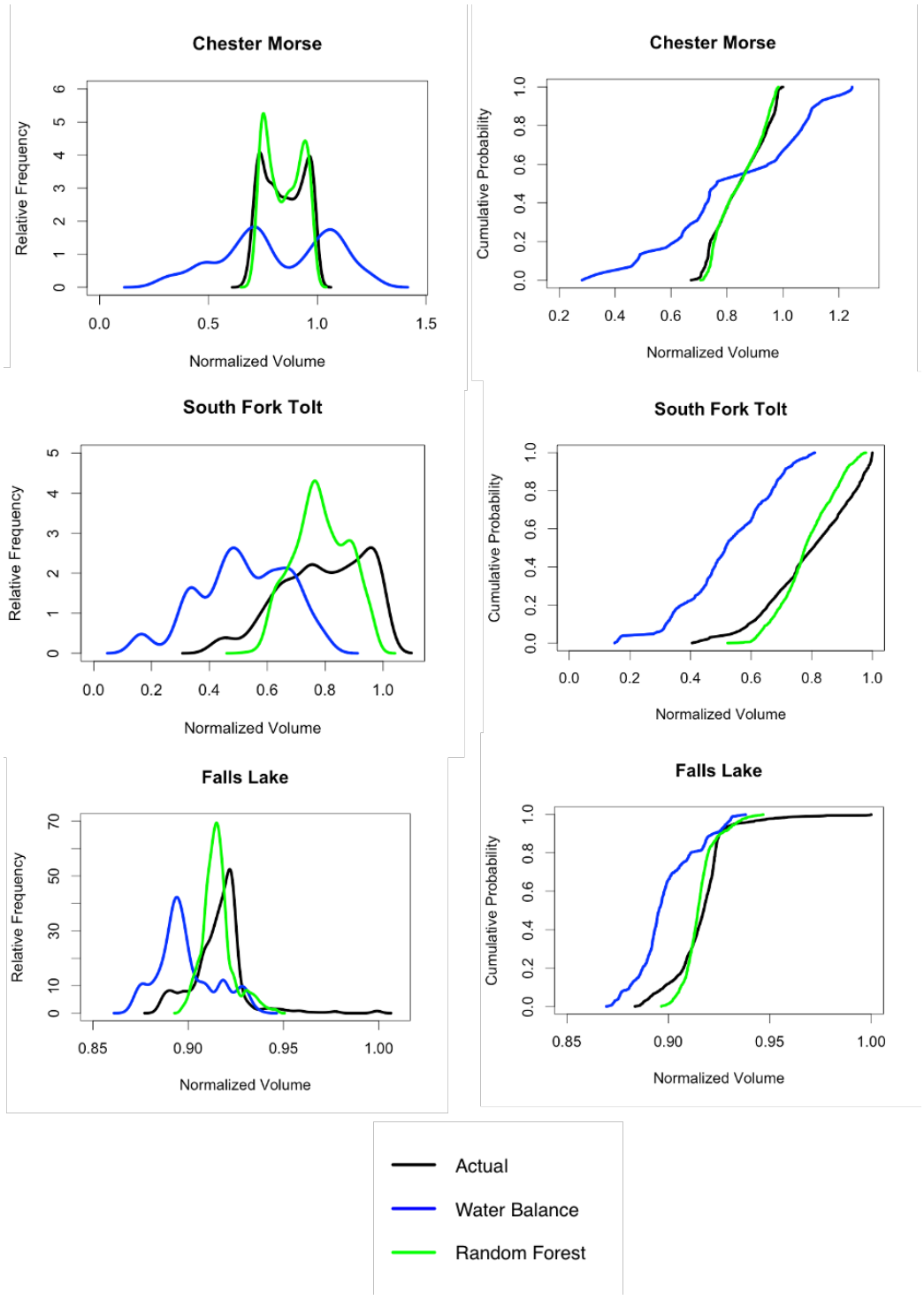

Fig. 2.6.: Empirical pdfs and cdfs for Chester Morse Lake, South Fork Tolt, and Falls Lake. The black lines represent the actual data, while the blue and green lines represent the results from the water balance and random forest models, respectively.

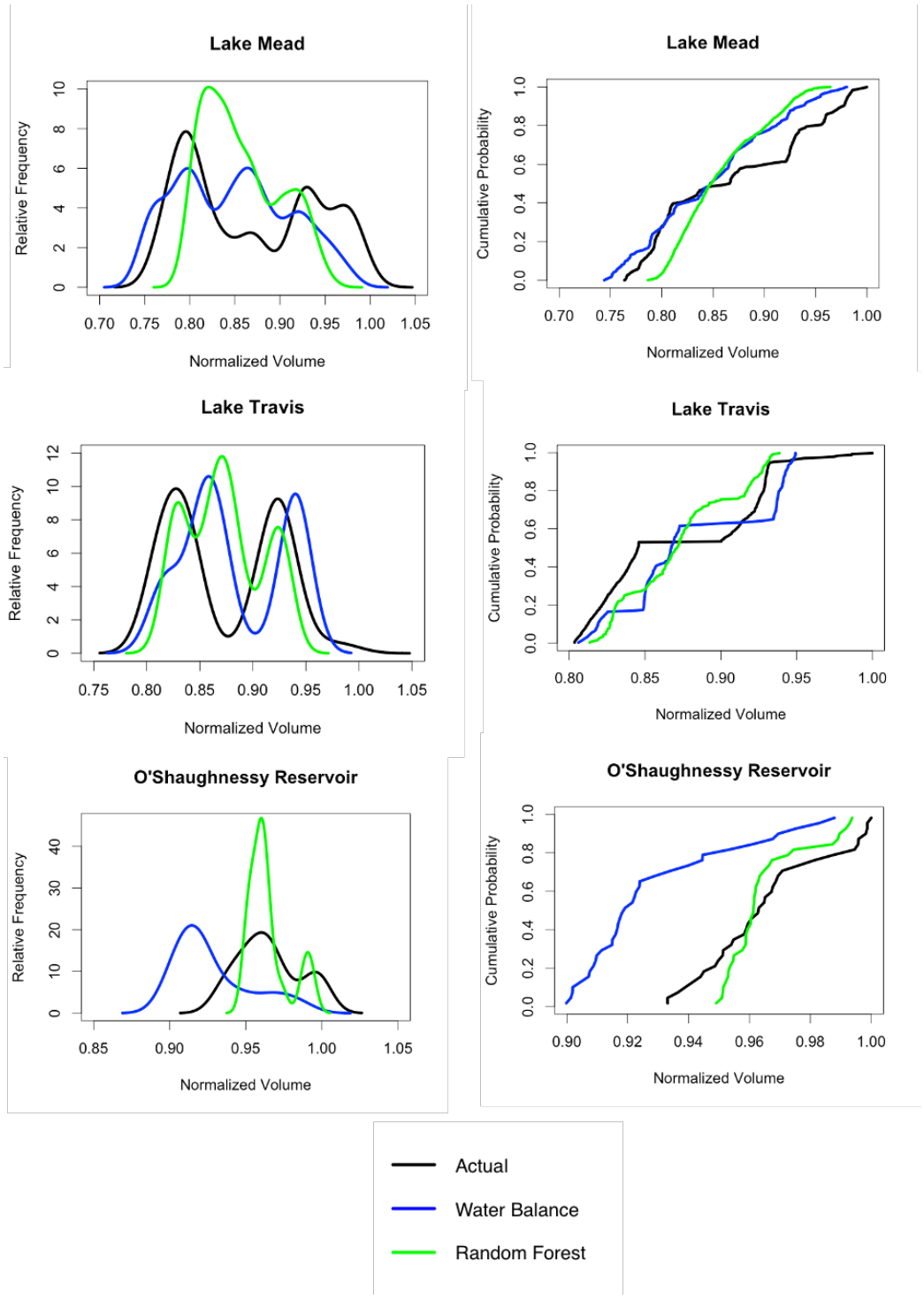

Fig. 2.7.: Empirical pdfs and cdfs for Lake Mead, Lake Travis, and O'Shaughnessy Reservoir. The black lines represent the actual data, while the blue and green lines represent the results from the water balance and random forest models, respectively.

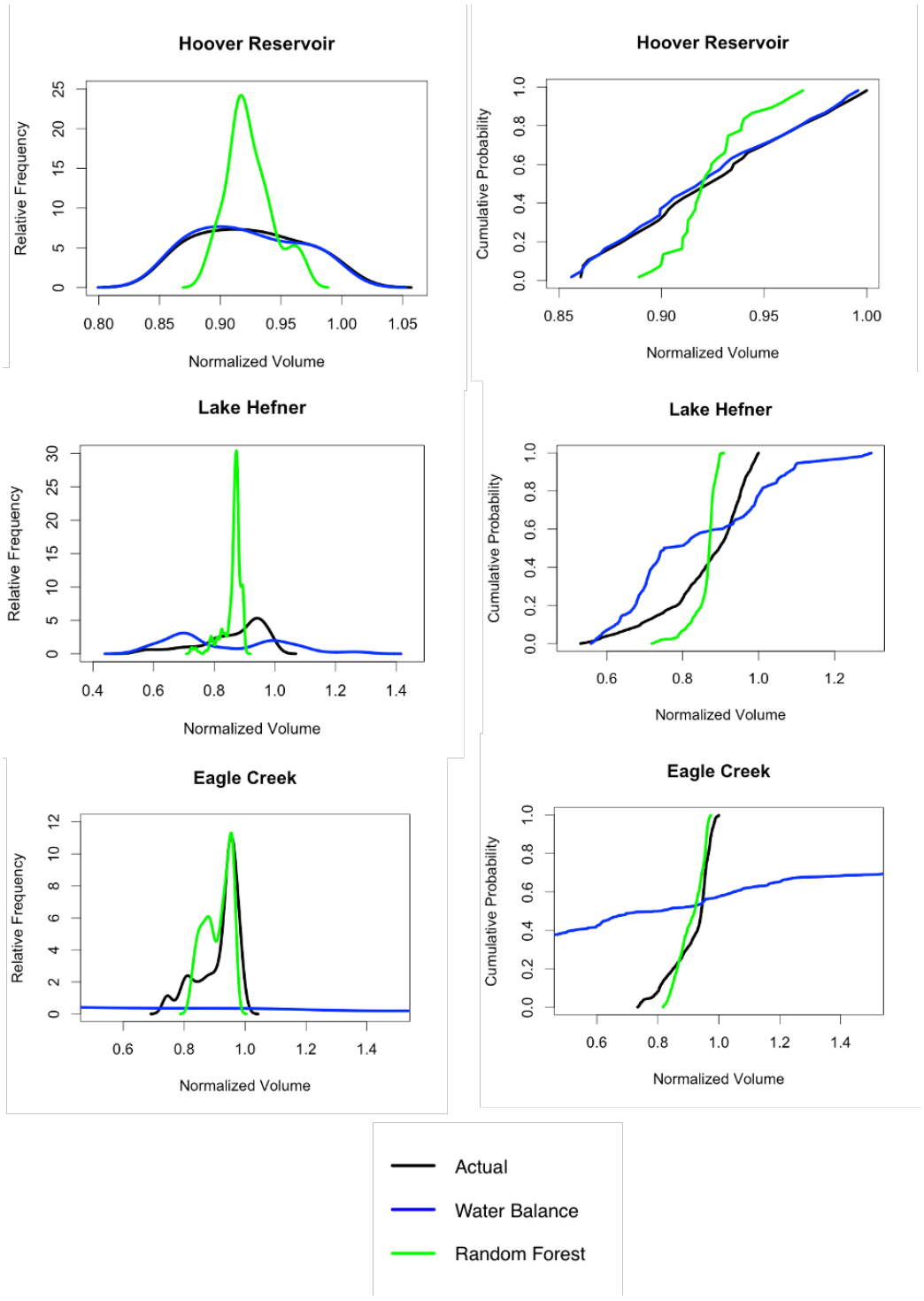

Fig. 2.8.: Empirical pdfs and cdfs for Hoover Reservoir, Lake Hefner, and Eagle Creek. The black lines represent the actual data, while the blue and green lines represent the results from the water balance and random forest models, respectively.

These results are further demonstrated in Figures 2.6 - 2.8, which show the empirical probability distribution functions (epdfs) and cumulative distribution functions (ecdfs). In Figure 2.6, for example, we can see that the random forest method tends to match the actual data better than the water balance model for Chester Morse Lake and, to a lesser extent, South Fork Tolt and Falls Lake. It is interesting to note that while the water balance model is overpredicting the variance in Chester Morse Lake, it does accurately represent the bimodal nature of the pdf. The variances of South Fork Tolt and Falls Lake, on the other hand, appear to be modeled well by the water balance model. In these reservoirs, the means have modeled as less than reality. In this sense, model selection would be based on the research goal—explanatory studies might only focus on matching the mean, while predictive studies would aim to match both the mean and variance.

Similar results are shown in Figure 2.7, which shows results from Lake Mead, Lake Travis, and O'shaughnessy Reservoir. Here, the benefits of using the water balance method to model Lake Mead is further demonstrated, especially in terms of the variance. O'Shaughnessy Reservoir is not well-modeled by either technique, which might be a response of having limited data points. I would hypothesize that, given a large sample size, the water balance model, which seems to have better matched the magnitude and shape of the actual data, would perform better. Additionally, complex predictive modeling techniques, such as random forest, tend to also perform better with increased sample sizes. Interestingly, both models do a good job of representing the actual data in Lake Travis, especially given the bimodal nature of the data. The water balance model appears to do a better job of matching the magnitude and location of the peaks, although shifted slightly towards larger volumes than the actual data.

Finally, in Figure 2.8, the results from Hoover Reservoir, Lake Hefner, and Eagle Creek are shown. The water balance model outperforms the random forest model in both Hoover Reservoir and Lake Hefner, although the results from Lake Hefner indicate a much larger variance than reality. In fact, the random forest results are not

even close in terms of the variance of the actually data. In Eagle Creek, the random forest model performs well compared to the actual data, with the water balance model performing the worst out of all the reservoirs. In this plot, the epdf and ecdf are nearly horizontal. In fact, if one were to zoom out to the full extent of the plots, the epdf would have an x-axis range of -1 to 3, indicating that the reservoir level drops below zero at some point, but is also holding 3 times the maximum volume at another. This may be due to the large infiltration ration considered in Section 2.3. However, a larger issue may be the data in this case. It is hard to say for certain, however, since the primary purpose of this reservoir is flood control, it is possible that the intra-daily variation is critical to accurate modeling. It is also likely, that there is some nonlinear aspect to the reservoir, which is being harnessed by the random forest model to better predict the torage. Overall, both models do well at representing the reservoir volume within the observational space, but going forward, it is critical to be able to use these models to make projections about the future of water supply reservoirs.

2.4.2 Comparison in the Projection Space

Often models are developed in the observational space, but then used to make future projections. This can introduce error to the results, especially if the original model wasn't developed with prediction in mind [65]. In order to test the *predictive* capabilities of the water balance and random forest models, I held out the last four months of data. I used the climatological mean (i.e., the average historical conditions for a given day) to predict the reservoir volume over these four months. Unfortunately, due to the lack of data points, Lake Travis, O'Shaughnessy Reservoir, and Hoover Reservoir were not included in this analysis. The reason for leaving out O'Shaughnessy and Hoover Reservoirs was discussed earlier, as part of the analysis in the observational space. Lake Travis, however, had enough data points (> 200) to accurately model the observations. That being said, holding out four months of data

effectively halved the dataset, leading to more points being held out than included for training. Due to this disparity, Lake Travis was not considered within the projection analysis.

Considering the reservoirs that were included in the projection analysis, Table 2.9 shows the moment analysis between the actual volume and the projected volume using the water balance method. The results indicate that the water balance model does a fairly good job of predicting the mean reservoir volume. In fact, the largest differences between the actual and projected mean are in South Fork Tolt and Eagle Creek, which were among the poorly represented reservoirs in the observational space. Similarly, in Falls Lake and Lake Mead, which were accurately represented by the water balance model in the observational space, there is little difference between the actual and projected means. The CV, however, was not well modeled by the water balance method in most reservoirs. Lake Mead is one exception, with only a 12% difference between the actual and projected CV. It is possible that due to the highly-managed nature of Lake Mead, the water balance method is sufficient for both modeling observations and making projections. However, the remainder of the

Table 2.9.: Results from the moment analysis on both the actual and projected reservoir volume (water balance method). Note that no projections were made for Lake Travis, O'Shaughnessy Reservoir, or Hoover Reservoir, since there were insufficient data points.

Reservoir	Actual Data		Modeled Data		Difference (%)	
	Mean (ft^3)	CV	Mean (ft^3)	CV	Mean	CV
Chester Morse Lake	2.64×10^9	0.112	2.50×10^9	0.087	5.7	25.1
South Fork Tolt	1.89×10^9	0.186	1.49×10^9	0.089	24.0	70.9
Falls Lake	4.10×10^{10}	0.02	4.01×10^{10}	0.0025	2.4	156
Lake Mead	9.77×10^{11}	0.0055	9.84×10^{11}	0.0062	0.84	12.3
Lake Travis	—	—	—	—	—	—
O'Shaughnessy	—	—	—	—	—	—
Hoover Reservoir	—	—	—	—	—	—
Lake Hefner	2.85×10^9	0.054	3.07×10^9	0.029	7.3	59.0
Eagle Creek	1.01×10^9	0.017	7.74×10^8	0.183	26.2	166

Table 2.10.: Results from the moment analysis on both the actual and projected reservoir volume (random forest method). Note that no projections were made for Lake Travis, O'Shaughnessy Reservoir, or Hoover Reservoir, since there were insufficient data points.

Reservoir	Actual Data		Modeled Data		Difference (%)	
	Mean (ft^3)	CV	Mean (ft^3)	CV	Mean	CV
Chester Morse Lake	2.64×10^9	0.112	2.87×10^9	0.095	8.43	16.8
South Fork Tolt	1.89×10^9	0.186	2.11×10^9	0.123	11.0	40.4
Falls Lake	4.10×10^{10}	0.02	4.10×10^{10}	0.0056	0.19	114
Lake Mead	9.77×10^{11}	0.0055	1.04×10^{12}	0.029	6.43	136
Lake Travis	—	—	—	—	—	—
O'Shaughnessy	—	—	—	—	—	—
Hoover Reservoir	—	—	—	—	—	—
Lake Hefner	2.85×10^9	0.054	2.84×10^9	0.0065	0.53	157
Eagle Creek	1.01×10^9	0.017	9.57×10^8	0.042	5.19	84.8

reservoirs do not do too well at representing the variance of the system, especially in Falls Lake and Eagle Creek. It is likely that the use of climatological means led to better representation of the means, while failing to represent the variance of the system. That being said, most reservoirs were well-represented by the water balance model, indicating its use as both an *explanatory* and *predictive* model. The results from applying the Kolmogorov-Smirnov test and Welch's t-test to the water balance model projections can be found in Appendix A.

In addition to using the water balance model to make projections, I also considered the random forest model. Random Forest is an algorithm that was developed for predictive purposes, so one would expect that the results would be an improvement to the water balance model projections. As indicated by Table 2.10, however, this is not always the case. For example, the average volume of Lake Mead was better represented by the water balance model than the random forest model. This is to be expected, since Lake Mead appears to be more accurately modeled using linear inputs and outputs than the complex random forest model. The random forest model also fails to match the CV of the actual data, while the water balance model was able to

represent the variance. This is further indicative of the linear nature of Lake Mead, as well as the regulated nature of the reservoir—since the inputs and outputs have been dictated by laws, the climatological mean is a good assumption for summer inputs and outputs. It is interesting that the random forest model performed so poorly in the projection space for Lake Mead, as the model performed well in the observational space. This may be due to the lack of memory in the random forest process. In other words, random forest predicts each value separately, such that there is no reliance on the previous values. This introduces a lot of variance to the model, with the possibility for rapid changes in predicted volume within a few days of each other. This is one of the reasons for doing cross validation—to reduce the variance [47]. Depending on the period used to make the projections, the random forest output could be more or less variable. It is likely that Lake Mead falls into the former case, given the training data was dominated by a long-term drought that had ended previous to the projection period. A solution to this issue would be to use a cross-validation scheme, however, in an effort to replicate a process that might be available to a practitioner, it was decided to use the previous data to predict the future data in one iteration. The statistical analysis results (i.e., the Kolmogorov-Smirnov test and Welch's t-test) from using the random forest model to make projections can be found in Appendix A.

Finally, going beyond the numbers, it is possible to visually inspect the differences in the empirical probability distribution functions and the empirical cumulative distribution functions. Figures 2.9 and 2.10 show the results from the six reservoirs included in the projection analysis. In Figure 2.9, for example, it is shown that Chester Morse Lake is better represented by the water balance model in the projection space, which is different than the observational space. In Figure 2.6, Chester Morse Lake was better modeled by the random forest method. That being said, the random forest does not perform terribly, when compared to the water balance model, just not as well. South Fork Tolt and Falls Lake, on the other hand, were better predicted by the random forest model in both the observational and projected space. This suggests a

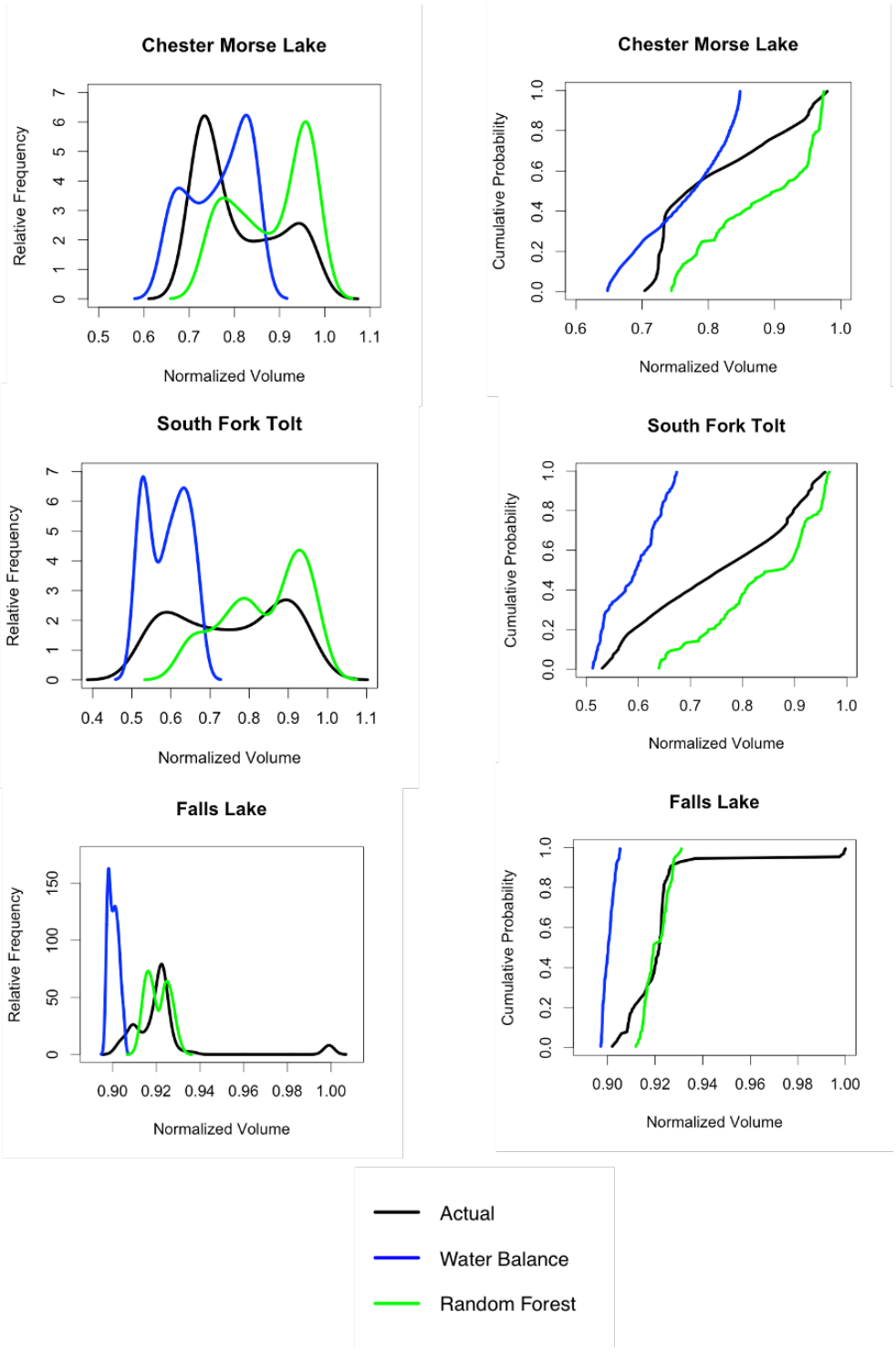

Fig. 2.9.: Empirical pdfs and cdfs for the projections of Chester Morse Lake, South Fork Tolt, and Falls Lake. The black lines represent the actual data, while the blue and green lines represent the results from the water balance and random forest models, respectively.

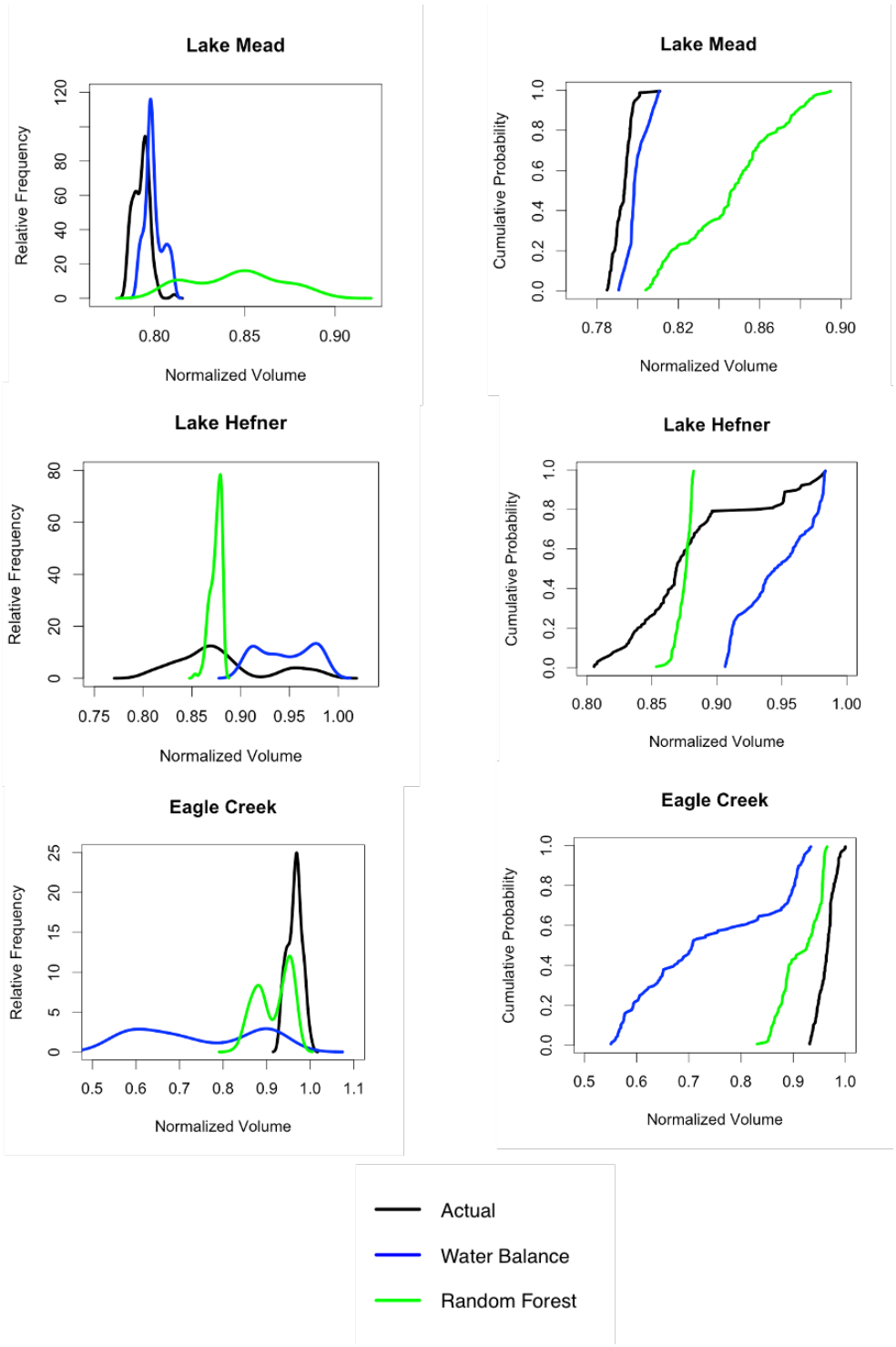

Fig. 2.10.: Empirical pdfs and cdfs for the projections of Lake Mead, Lake Hefner, and Eagle Creek. The black lines represent the actual data, while the blue and green lines represent the results from the water balance and random forest models, respectively.

significant nonlinear relationship between the predictor variables and the reservoir volumes that is present both in the observational and projections spaces.

In Figure 2.10, we can see that the water balance model does a better job a projecting the reservoir volumes in both Lake Mead and Lake Hefner, although the patterns appear to be the opposite of one another. For example, the random forest overpredicts the variance in Lake Mead, while it underpredicts the variance in Lake Hefner. The Lake Mead results were expected, based on the moment analysis, and are likely a function of the regulated nature of the reservoir. The Lake Hefner projections are also no surprising, given the similarity to the pattern observed in the observational space (see Figure 2.8), in which the random forest significantly underpredicted the variance in volume. Finally, we see that the random forest model better represents the projected volume, when compared the water balance model, although it is not perfect. In Eagle Creek, both models overestimated the variance in the reservoir volume, possibly due to the use of climatological means, which do not provide a sense of the variety in input conditions that could happen during the course of the summer months. In fact, the use of mean (as opposed to median) could have skewed the results towards more extreme conditions than common, which may have impacted all of the projections analyses. That being said, however, some reservoirs were better represented by different models, suggesting that that neither method should be used indiscriminately.

2.4.3 Discussion

In this chapter, I presented the results from two different models: a traditional water balance model and a random forest model. More broadly, these models can be described as explanatory and predictive, respectively. The main difference between explanatory and predictive modeling is the focus on minimizing bias and variance, as opposed to just bias [65]. In other words, an explanatory model which seeks to *explain* some phenomenon will focus on minimizing the bias, thus resulting in the

most accurate representation. Predictive modeling, in which the aim is to *predict* some phenomenon, attempts to minimize bias *and* variance, potentially sacrificing some accuracy to gain more precision in the results [65]. When applying these two types of models to urban reservoir systems, there were significant differences between the results. For example, within the observational space, the water balance model performed better than the random forest model in Lake Mead, Lake Travis, Hoover Reservoir, and, to a lesser extent, Falls Lake. These reservoirs are all primary water sources for the cities of Las Vegas (NV), Austin (TX), Columbus (OH), and Raleigh (NC), respectively, so it is possible that leads to characteristics that more easily modeled by the water balance equations. However, Chester Morse Lake is also a primary drinking water source, and the volume was better modeled by the random forest model. More likely, there are different factors that play a role. For example, Chester Morse Lake and South Fork Tolt are both located in the Pacific Northwest and are both better modeled by the random forest model. One of the benefits of using the random forest model is the ability to consider nonlinear relationships between the predictors and the reservoir volume. The performance of the random forest model in this region suggests that there is a nonlinear aspect to the reservoir inputs and outputs, perhaps due to the climate of the region. Interestingly, O'Shaughnessy Reservoir and Hoover Reservoir are located in central Ohio (both are within 20 miles of each other), but have different responses to the models. O'Shaughnessy Reservoir is not predicted well by either model, although the random forest better matches the mean, while the water balance better matches to variance. In Hoover reservoir, the water balance model outperforms the random forest model in terms of both mean and variance. These discrepancies may be due to differences in purpose—O'Shaughnessy Reservoir is used for both hydroelectric and water supply, while Hoover Reservoir is only used for water supply. It is also possible, however, that the sample size was too small to adequately assess modeling capabilities. With just a few months of data, the sample size is smaller than preferred for use in complex models, such as random forest. Additionally, it is hard to judge the validity of the water balance model in

these two reservoirs, when only a few data points have been tested. It is possible that if more data had been available, that the water balance model would have performed differently, especially in Hoover Reservoir.

Lake Mead, on the other hand, had a large sample size and is well-represented by the water balance model, both in the observational and projection spaces. This is likely a result of the management practices in the Colorado River Basin. Essentially, there is a specific amount that must enter and leave Lake Mead on a daily basis, per regulations. The linear nature of this directed process allows for the water balance model to accurately represent the system. Interestingly, the water balance model, though considered an explanatory model, also does well at projecting the future reservoir volume. In fact, Lake Mead was the reservoir that was best projected by the water balance model, and one of the few in which the water balance model outperformed the random forest model. This performance can be linked to the similarity between the climatological mean of the input variables and the actual values of the variables on any given day. This is logical, given the regulation of streamflow in and out of the reservoir, as well as a fairly constant climate in terms of precipitation (or rather lack thereof) and evaporation. The poor performance of the random forest model might be due to the lack of memory within the model. In other words, the random forest model does not build on the previous values, which can result in a lot of variance in the output.

Looking at the other reservoirs used in the projection analysis, however, the random forest method tends to perform better, especially in South Fork Tolt, Falls Lake, and Eagle Creek. This is to be expected however, since random forest is considered a predictive model. This means that the random forest algorithm works to minimize the variance in addition to the bias, effectively improving the predictive accuracy. However, given the data-driven nature of the random forest model, it relies on the availability of large datasets to make predictions. This could create issues for reservoirs in data sparse areas. In this sense, although the random forest model performs

well in terms of predictive accuracy in most cases, it might not be ideal for all reservoirs. In some areas, the water balance model might be the preferred choice.

2.4.3.1 Pros and Cons of the Different Models

Both the water balance and random forest model have positives and negatives. For example, the random forest model requires large sample sizes—including both response and predictor data, while the water balance model can be performed with just a few data points. This could be critical for areas in the developing world or more rural areas that don't have the sensing capabilities or infrastructure to collect the necessary predictor data. Additionally, the random forest model, like all predictive modeling algorithms, requires response data to train the model. This could create issues in reservoirs where the reservoir level data is not collected regularly. In fact, one of the key drivers behind the selection of reservoirs for this study was the availability of volume data. It was significantly easier to find input data, but high resolution water level data was scarce, even in urban reservoirs within the US. An attempt to move the random forest modeling beyond the study reservoirs would be difficult given the data sparseness. In this sense, the use of the random forest model is limited by the data availability, while the water balance model can be used in all situations.

That being said, it is hard to determine which model is better, especially since different metrics of 'better' would lead to different conclusions. For example, if the models were to be compared on a quantitative basis, one could use some measure of error, such as the root mean squared error (RMSE), or a comparison of statistical tests, such as the Kolmorgorov-Smirnov test. In this chapter, I compared the statistical test results, as well as the differences in the moments between the actual and modeled data. This comparison is purely quantitative and can be used to determine efficacy on the basis of model accuracy. Basing the determination on accuracy alone however, would lead to the conclusion that the random forest algorithm was the best model. However, as a machine learning algorithm, random forest is mathematically-based,

rather than physically-based. These mathematical algorithms have the potential to base the prediction on relationships that are not physically possible [66]. Therefore, it is critical to also adopt other metrics to ensure the best model is used in a given analysis.

The model determination can be done on the basis of qualitative assessments, such as the computational efficiency or the number or the assumptions or parameterizations within a model. In this chapter, the random forest model is considered to be slightly less computationally efficient in that it takes longer to run and also requires more computational resources. However, the water balance model makes more assumptions about the data. The random forest model, being a non-parametric algorithm, makes no assumptions about the data [47, 53]. In this sense, depending on the preferred metric, one model could be selected over the other.

Finally, there is the possibility to select the model on a more philosophical level. For example, the model that is more user friendly may be preferred. This gets into the desired user of the model. In practice, it is unlikely that reservoir managers will have the knowledge of machine learning algorithms. Moreover, they are less likely to have the computational programs or knowledge to run these algorithms. Therefore, the water balance method may be preferred, even if the accuracy is slightly less than the random forest model.

Given the number of ways to determine the optimal model, it is important to be transparent about the reason behind using the model and metrics used to assess the models for that particular purpose. In order to do this, however, one needs to be aware of the pros and cons of the different techniques and be able to apply them in different situations. In this sense, the water balance model and random forest algorithm are complementary.

2.5 Conclusion

The purpose of this chapter was to evaluate the capabilities of multiple models to predict urban reservoir volume. First, I used statistical learning theory to test multiple algorithmic models. Through this work, I compared a variety of parametric, semi-parametric, and non-parametric models. Ultimately, the results showed that the random forest algorithm provided the most accurate representation of urban water supply.

Then, moving into a more traditional hydrological space, I used a water balance model to assess the changes to urban water supply over time. This model was shown to be accurate in terms of the mean and coefficient of variance. However, the results of the statistical tests, including the Kolmorogov-Smirnov test and Welch's t-test, indicated that there were statistically significant differences in both the distributions and means of the modeled and actual data. This could indicate problems with the parameter estimations (namely, infiltration to groundwater) or the bathymetry assumptions. Additional work is needed to test these remaining hypotheses.

Finally, the random forest and water balance models were compared, in both an observational and projection space. The results indicated that the water balance model performed well in the observational space, but not in the projection space. This is to be expected, though, since the water balance model is primarily an *explanatory* model, meaning it was not designed to accurately predict future states. The random forest model, however, is a *predictive* model, and as expected, performed much better in the projection space than the water balance model. This indicates the need to assess different modeling needs and apply different techniques. Depending on the study goals and available data, it could be beneficial to use the water balance model, the random forest model, or both. Ultimately, this means that researchers need to be aware of different modeling techniques and test various methods before moving forward with the best model for the task at hand.

3. ANALYZING THE WATER-ELECTRICITY DEMAND NEXUS

A version of Section 3.2 has been previously published in *Applied Energy*: https://doi.org/10.1016/j.apenergy.2019.113466. A version of Section 3.3 has been published in *Climatic Change*: https://doi.org/10.1007/s10584-020-02669-7. The introduction is a combination of both manuscripts.

3.1 Introduction

The water-electricity nexus is a concept dating back to the late 1980's, however applying the concept to urban areas began around 2010's [67]. Since the release of these studies and reports, there have been many initiatives surrounding the water-electricity nexus calling for researchers to evaluate the nexus and its impacts at various spatiotemporal scales and for numerous applications. The idea behind studying the nexus, as opposed to studying water and/or electricity in isolation, is that the two systems are interrelated and studying them separately will likely lead to (i) attenuated effects in efficiency and conservation programs to reduce residential energy and water consumption, (ii) overestimating price elasticity of demand, and (iii) designing ineffective demand response programs. On the other hand, considering their co-benefits in conservation measures has demonstrated potential to achieve savings at no net cost in some regions [68]. Moreover, simulation tools that have been built in isolation (i.e., tools that simulate only water or electricity) have been shown to result in significantly different consumption patterns than their integrated counterparts [69].

There are a variety of ways to study the water-electricity nexus, including water for electricity analyses and electricity for water analyses. To understand water for electricity, researchers frequently evaluate the water that is used during electricity generation [17]. An estimated 90% of the electricity in the US comes from thermoelectric power plants, which require water for cooling [20]. The amount of water withdrawn by these plants accounted for 40% of the water withdrawals in the US during 2005 [70], making these plants a crucial aspect to studying water availability in the US, especially during heatwaves and droughts. Higher temperatures and drought conditions have been shown to increase electricity demand, which ultimately leads to increased water withdrawals by thermoelectric generators [20], especially if the generators are coal-fired or cooled using open-loop technologies [71]. The remainder of the electricity in the US comes from other sources, including hydropower, which also requires a significant amount of water resources. Although hydropower is often used for grid stabilization, it can be significantly effected by increased rates of evaporation that accompany droughts [72]. Given that droughts are expected to increase [32], it is crucial that models represent the interdependencies between water and electricity, even in non-thermoelectric power plants. Electricity for water analyses, on the other hand, focus on quantifying the electricity it takes to treat and distribute water [17]. It was estimated that in 2012, water utilities in the United States consumed 38,100 GWh of electricity [73], which will likely increase as utilities continue to expand to keep up with urban growth. Given that water-related electricity use is expected to increase in states that are already water stressed, such as Florida, Texas, and Arizona [74], analyses that focus on the water-electricity nexus are becoming increasingly important.

In addition to the supply-based interdependencies discussed above, there are many aspects of water and electricity use that are interconnected. For example, watering landscapes, washing clothes, taking hot showers, and using a dishwater all require both water and electricity. These dependencies are critical for both electric and water

utilities trying to reduce peak load to lower the likelihood of supply inadequacies and service disruption risks, and reduce operations and maintenance cost [75].

In comparison to the studies of water-electricity *supply* nexus, research on the water-electricity *demand* nexus is more nascent [19]. The majority of the work on the demand-side has primarily focused on human behavior and specific tasks (e.g., heating water or using a dishwasher [76], as well as outdoor activities such as landscaping [25]). These studies provide a wealth of information on people's behaviors and the coupling between the urban water and electricity systems, but there is very little work on the subject that takes climate variability and change into account. The handful of studies that do consider climate, employ only simple and limited measures (e.g., change in precipitation or temperature) to determine the impact [27]. For example, one study performed by Venkatesh et al. (2014) demonstrated the value of precipitation and temperature on raw water sources [28], but did not include other key factors, such as evaporation. Similarly, a study by Mostafavi et al. (2018) considered temperature when modeling resdiential water and energy consumption, but did not include potentially important variables, such as relative humidity [29]. In fact, the climate measures impacting the water-electricity nexus likely go beyond simple measures such as precipitation and temperature that have yet to be explored. In particular, the El Niño/Southern Oscillation cycle, which has been shown to impact the water-energy-food nexus [77], has not been included in urban water-electricity demand nexus studies. This combination of limited research on system interdependencies on the demand side, as well as the simplistic view of climate impacts could lead to misinformed managment decisions.

These suboptimal management decisions are undesirable under the current conditions, but as urban areas continue to grow in population and the climate continues to change, they could be disastrous. In fact, with estimates that 70% of the world population will live in urban areas by 2050 [1], utility companies could be experiencing a significant increase in demand, without taking climate into account. By taking climate variability into account, any stress caused by the increase in demand

will be exacerbated [78–80]. For example, electric grids have been designed to handle specific peak loads, but under climate change, peak loads will likely exceed the capacity margins more frequently [81–83]. Given that these peaks in usage tend to be more sensitive to variations in climate than average usage [84], it is likely that electric utilities will experience a dangerous level of stress, that could result in blackouts and shutdowns, if they do not prepare adequately [80, 85, 86]. This will be especially true for the residential sector, which is more sensitive to climate variability than the commercial and industrial sectors [87]. Therefore, it is crucial that electric utilities have access to accurate and credible models that adequately characterize the climate sensitivity of residential electricity use, as it represents the sector that is most likely to be affected by climate change. Moreover, electricity use is affected by water use, especially in the residential sector [19], making it imperative that these models also account for the impact of climate change on water use.

Water utilities, unlike electric utilities, have the ability to store resources for later use. However, as climate change progresses, droughts are likely to become more intense, potentially reducing storage capabilities of reservoirs that are mainly used for public drinking water supply [32, 88]. Moreover, increased temperatures usually lead to increased water use within the residential sector [89, 90], which will put additional stress on the water supply reservoirs. These impacts of climate change are not experienced in isolation, rather, they are interconnected, such that the impacts on the water sector will affect the electricity sector, and vice versa. For example, in the event of a drought, the water supply reservoirs may experience a significant drop in storage, which will put pressure on the water utility to maintain a certain service level under limited supply. This water supply will be put under additional pressure by the increase in demand that follows higher temperatures and drought conditions. Furthermore, there will be even more pressure brought on by the electric utility which will require an increasing amount of water for cooling generators in regions where the electricity is generated by thermoelectric technology, as is the case in 90% of the United States [20, 86]. In this sense, the nexus leads to increased stress on both water

and electricity utilities, especially under climate change [91, 92]. This creates a need for water and electric utilities to work together to prepare for climate change and make decisions that are the best for both sectors.

In order to ensure infrastructure managers and urban planners can make the best decisions now and in the future, there needs to be increased development of accurate, credible and accessible models that take system interdependencies into account [93,94]. However, there are only a few models that project the water-electricity nexus into the future and they often only use a small subset of climate variables, generally precipitation and temperature [28, 29, 95]. In this chapter, I present the results from two main projects focused on filling this gap. First, I present the initial model developed to predict the climate-sensitive portion of the water-electricity demand nexus at the city-scale. Then, I apply the model to a regional analysis focused on projecting the water and electricity use into the future using climate change scenarios.

3.2 Multivariate Model Development

The central goal of this section is to comprehensively assess the climate sensitivity of the urban water-electricity demand nexus, which has largely been overlooked in previous studies. The proposed framework is designed to handle multiple interdependent response variables. Since the coupled water-electricity nexus model takes the correlation between the response variables into account, it was hypothesized that this multivariate modeling framework would predict the water and electricity use better than similar univariate models. To test this hypothesis, the framework was applied to six large-range cities in the Midwestern United States and evaluated the impacts of climate variability on the demand nexus. It was also hypothesized that both local climatic variables, such as precipitation and temperature, and large climatic drivers, such as the El Niño/Southern Oscillation index, would be important predictors of end-use demand for water and electricity.

In the following sections, I first discuss the data and methodology used to build and test the model. Then, I show the model performance of the multivariate model, followed by a comparison with a similar univariate model. Finally, I wrap up with a discussion on important variables and differences between the cities.

3.2.1 Data and Methods

To demonstrate the applicability of the proposed approach, the Midwest region in the United States was selected as a case study. In this section, we will first describe the study sites and the input data used for the analyses presented in this paper, and will then delve into the proposed methodology for assessing the coupled water-electricity nexus in the case study areas.

3.2.1.1 Site Description

In this study, the focus was on the northern and eastern parts of the Midwest, including Ohio, Indiana, Illinois, Wisconsin, and Minnesota. Within this study area, depicted in Figure 3.1, six cities of varying population sizes were selected: Chicago (IL), Columbus (OH), Indianapolis (IN), Minneapolis (MN), Cleveland (OH), and Madison (WI). These cities were selected in order to capture a variety of different sizes, while still focusing on some of the most populous cities in the region. In fact, the population ranges from 255,000 people in Madison to 2,716,000 people in Chicago. Moreover, each city, though they have different demand patterns, will likely experience similar impacts of climate change due to their geographical proximity. In particular, it is likely that the Midwest region as a whole will have higher temperatures and more precipitation as CO_2 levels continue to rise [96], which will in turn affect the urban water-electricity demand nexus.

3.2.1.2 Data Description

The data for this study was obtained from four main sourcesthe US Energy Information Administration (EIA), National Centers for Environmental Information (NCEI), National Oceanic and Atmospheric Administration (NOAA), and local water utilities. Specifically, monthly residential electricity use was obtained from the EIA [97], meteorological and climate data from the NCEI [39] and NOAA [41], and residential water use was obtained through records requests to local water utilities. The meteorological data was collected from several meteorological towers stationed around each city and aggregated to get an average monthly value for each city between 2007 and 2016. Specifically, there were four active towers in Chicago, Columbus, and

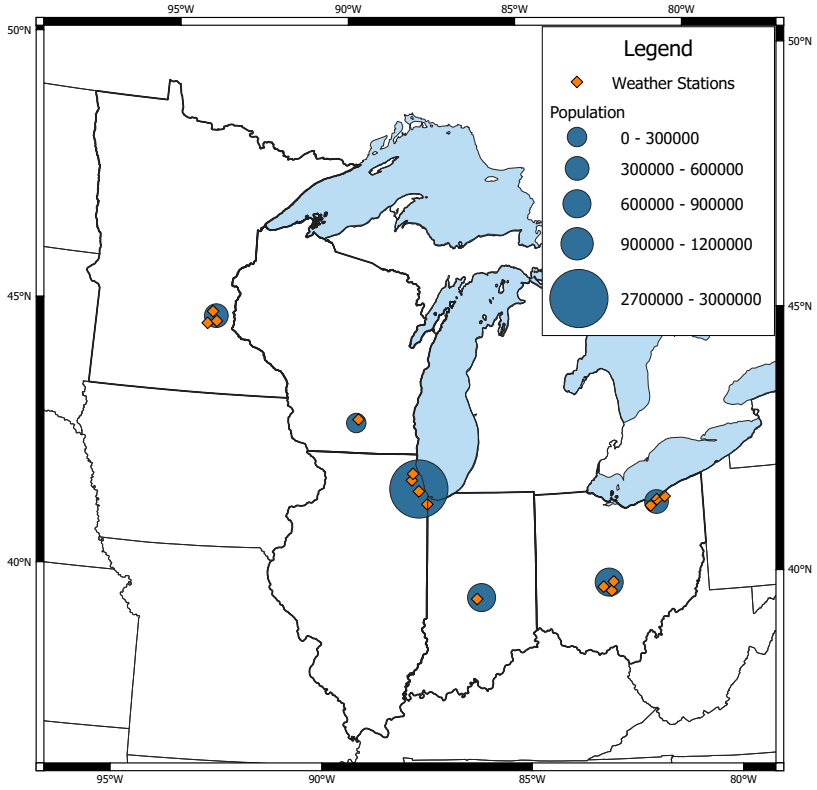

Fig. 3.1.: Study area: Midwestern region of the United States. The blue circles represent the cities included in the regional analysis, sized relative to population, and the orange diamonds represent the weather stations that were used as sources for the observational data.

Minneapolis, three in Cleveland, and one in Indianapolis and Madison (see Figure 3.1). Meteorological variables used in the analysis included temperature (dry bulb and dew point), relative humidity, wind speed, and precipitation. The El Niño/Southern Oscillation strength index was also included in the analysis, as a large-scale climatic driver that has been shown to impact the climate of the Midwest [98].

Table 3.1.: The input variables used for developing the coupled water-electricity demand nexus model. Each variable was collected from January 2007 through December 2016 and aggregated to the monthly time scale.

Variable Type	Variable Name	Units	Source
Response	Monthly Water Use	L/cap	Local Utilities
	Monthly Electricity Use	MWh/cap	EIA-861M [97]
Predictor	Average Maximum Dry Bulb Temperature	°C	NCEI [39]
	Average Dew Point Temperature	°C	NCEI [39]
	Average Relative Humidity	%	NCEI [39]
	Average Maximum Relative Humidity	%	NCEI [39]
	Average Wind Speed	m/s	NCEI [39]
	Average Maximum Wind Speed	m/s	NCEI [39]
	Accumulated Precipitation	cm	NCEI [39]
	El Niño/Southern Oscillation index	–	NOAA [41]

In this study, there were two response variables: residential electricity use and residential water use, both normalized by the number of customers reported by the utility. Often water and electricity are provided by separate utilities, with potentially different service areas, this normalization allowed us to compare these two variables regardless of the differences in service area. Additionally, the response data was adjusted for seasonality to ensure that the results were demonstrating the effect of

climate on the water-electricity demand nexus, independent of the natural seasonality present in the usage patterns. In the seasonality adjustment, the time series were decomposed and the seasonality components were subtracted from the original time series [99] (see Appendix B for more information). There were also eight meteorological and climatic predictors (see Table 3.1), that were included in the initial model run. There was a focus on variables that are easily measured by meteorological stations due to the availability of such data, as well as the results of previous studies, which showed the importance of meteorological variables on water and electricity demand. For example, Balling et al. (2008) showed the impact of precipitation and temperature on water consumption [89]. Similarly, Mukherjee and Nateghi demonstrated the impact of temperature and wind speed on electricity consumption [80]. Both average and maximum values of meteorological variables were included to establish which statistic (i.e., maximum or mean) would better capture the intensity of the signals in the water and electricity demand data. Similarly, it has been shown that the El Niño/Southern Oscillation plays an important role in affecting hydroclimatic processes across the US, and in particular, the Midwestern region [98], making it an important variable to include in the analysis of the climate impact on residential water and electricity use.

3.2.1.3 Methodology

The interconnectivity between water and electricity use has been well documented throughout the literature [67], with a few studies focusing on the impacts of climate [27]. However, this is the first time, to our knowledge, that the impact of climate on the water-electricity nexus has been evaluated through a *multivariate* framework based on statistical learning theory. The advantages of this framework include (i) assessing the role of a wider range of climatic variables on the water-electricity demand nexus than previous studies, and (ii) leveraging a robust, non-parametric technique to assess the climate-sensitivity of both water and electricity use simultaneously, while

taking their complex and non-linear interactions into account. Moreover, the required inputs to the modeling framework are readily available, such that utility managers, researchers, or other interested parties can easily apply the model to their city or cities of interest.

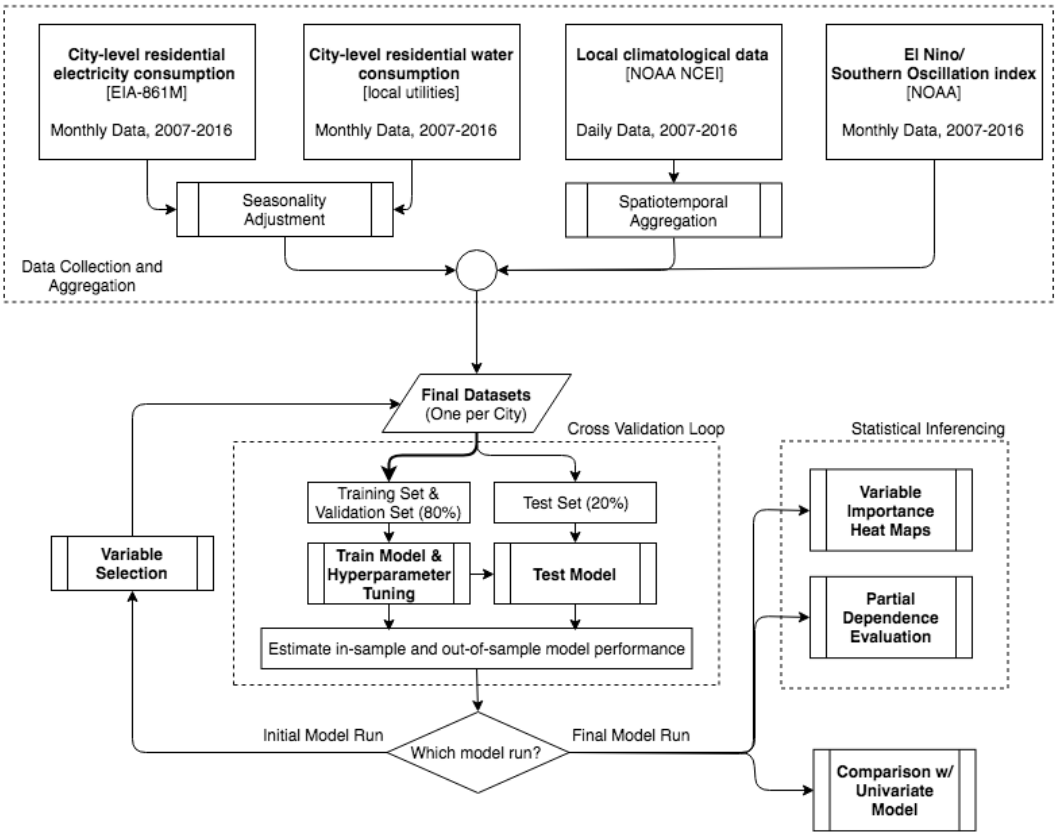

Fig. 3.2.: Schematic of the modeling process used in the initial model development phase of the demand analysis.

There are four main steps in the modeling process: (1) data collection, preprocessing and aggregation, (2) model training and testing, (3) statistical inferencing, and (4) comparative analysis with a univariate model. A schematic of this process can be seen in Figure 3.2. The first step was to collect the data, normalize the response variables and implement seasonality adjustments (as described in Section 3.2.1.2), and to aggregate the meteorological data spatially across weather stations and temporally

from daily to monthly values. The initial model training and testing was performed—within a 5-fold cross validation loop—with all the predictor variables (see Table 3.1). Cross validation, which is a standard process for ensuring the model is robust and validating the predictions, was used for both model hyperparameter tuning as well as model performance assessment. The initial model runs were then followed by a variable selection step to establish the key predictors. Finally, the statistical inferencing was performed using the results from the final best model that included the reduced input variable set, based on the variable selection step. Each of these steps will be described in further detail in the following sections.

Supervised Learning Theory

The algorithm used throughout this study fall into a larger category of statistical learning theory known as 'supervised learning'. Supervised learning algorithms are built to predict target variable(s) of interest (i.e., the response variable(s)), given a number of predictor variables. Supervised learning can be mathematically described as:

$$Y = f(X) + \epsilon \tag{3.1}$$

where Y is the response variable(s) of interest, X is the series of predictor variables used to predict the response, and ϵ is the irreducible error ($\epsilon \sim N(0, \sigma^2)$) [47]. In supervised learning, the aim is to predict the response variable(s) such that the the expected error is minimized as shown below [47].

$$\min \frac{1}{N} \sum_{i}^{N} \Delta[\hat{f}(X_i), f(X_i)] \tag{3.2}$$

Here $\hat{f}(X_i)$ and $f(X_i)$ represent the estimated and true functions, respectively, and Δ represents some measure of distance (e.g. the Euclidean or Manhattan distance).

Among the wide library of supervised learning algorithms, tree-based methods are one of the most popular non-parametric learning techniques [47]. Tree-based models offer competitive predictive accuracy compared to most of the state-of-the art

statistical machine learning algorithms [100], and lend themselves more easily to interpretation and inferencing compared to other "black box" algorithms, such as deep learning and support vector machines [47]. In this paper, a multivariate extension of an ensemble-of-trees approach was implemented, as described below.

Algorithm Description

The proposed framework is based on an advanced supervised learning technique—based on an ensemble-of-trees approach—that leverages the covariance structure of multiple response variables to better estimate the complex interactions between the target variables. Specifically, the predictive model of the coupled residential water and electricity demand was developed based on a multivariate extension of the gradient boosted regression trees algorithm [**?**].

Gradient boosted regression trees is an ensemble-of trees method that takes advantage of the boosting meta-algorithm to increase the predictive accuracy [**?**]. The boosting meta-algorithm works by sequentially fitting models (in this case decision trees), where in each iteration more weight is given to the better classifiers and the misclassified points in order to reduce the overall loss function and enhance the predictive accuracy. Boosting is represented mathematically in the equation below.

$$G(x) = \sum_{m}^{M} \alpha_m C_m(x) \tag{3.3}$$

Here $G(x)$ is the final ensemble model, M is the total number of iterations to be completed, α_m is the weight of each prediction, and C_m is the tree models fitted to the input variable x at iteration m.

In this paper, multivariate tree boosting, which extends gradient boosted regression trees to a multivariate (i.e., multi-response) case, is leveraged. Thus, the multivariate extension of the algorithm enables the simultaneous prediction of multiple response variables [101]. Specifically, this algorithm iteratively builds trees by minimizing the squared error loss for each response variable and maximizing the covariance

discrepancy in the multivariate response. In other words, at each iteration, a prediction is made for each response variable, such that the loss function is minimized and the covariance discrepancy between the current and previous predictions is maximized. This allows each subsequent prediction to be incrementally more accurate than the previous, while ensuring the predictors that account for the most covariance in the nexus of the response variables are selected. The steps of the algorithm are summarized below:

Algorithm 1 Multivariate Ensemble Tree Boosting Algorithm [101]

1: **for** m *in* $1, \ldots, M$ steps (regression trees) **do**
2: **for** r in $1, \ldots, R$ quantitative response variables (e.g., water and electricity demand) **do**
3: train tree $m^{(r)}$ to residuals, and estimate the covariance discrepancy $D_{m,r}$
4: **end for**
5: Select the response $y^{(r)}$ corresponding to the regression tree that yielded the maximum $D_{m,r}$
6: Update residuals by subtracting the predictions of the tree fitted to $y^{(r)}$, multiplied by step-size.
7: **end for**

This algorithm has been tested in a few multivariate predictive applications, ranging from psychological well-being [101] to multi-dimensional infrastructure resilience assessment [85], and it was hypothesized would be a good candidate for electricity-water nexus modeling.

Variable Selection

Per Occam's razor, it is desirable to establish the simplest model (containing a subset of input variables) that best captures the data dependencies and covariance. In other words, variable selection was conducted to reduce model complexity via retaining only the most important or influential predictors in the final model. In this

framework, variable selection was based on establishing the relative influence of each variable, via measuring the sum of squared errors obtained on any split of a given predictor, summed over all trees in the the prediction model [47]. The calculated sums of squared errors provide a basis for ranking the predictor variables. Thus, the relative influence is related to the amount of reduction in total error that can be attributed to a given predictor—the higher the reduction in error, the more influential (and important) the variable is in the model. For multi-dimensional response variables, the univariate relative influence is first measured for each independent variable and for each response. Summing the importance over all response variables renders a 'global' measure of influence for the independent variables across all target variables.

In this study, the variables were selected for the final model if they had a relative influence greater than 5% in at least 4 of the 6 cities. Using this threshold, the following five predictors were retained in the final model: average maximum dry bulb temperature, average dew point temperature, average relative humidity, average wind speed, and the El Niño/Southern Oscillation index. These variables were used in the final model run and subsequent inferencing and analysis.

Statistical Inferencing and Analyses

The statistical inferencing for the multi-dimensional water-electricity nexus model—developed using the multivariate tree boosting algorithm described above—was conducted using the following methods: (1) evaluating the model performance (i.e., model goodness-of-fit and predictive accuracy), (2) assessing the covariance explained by each predictor on individual response variables and identifying the clusters of input variables that jointly influence one or both response variables, (3) visualizing the partial dependence between the important predictors and the response variables, and (4) comparing the multivariate model performance to a similar univariate model.

- *Model Performance*

To evaluate model fit and predictive accuracy, the algorithm was run—within the 5-fold cross validation loop—for each city simultaneously (Figure 3.2), resulting in

one prediction per city per response variable. The performance of the model was assessed using two statistical measures: the out-of-sample root-mean-squared error (RMSE) and the out-of-sample coefficient of determination (R^2). RMSE provides an absolute measure of error that heavily penalizes large deviations, making it ideal for prediction applications. The out-of-sample R^2 value demonstrates the fit of the model predictions made by the test dataset, which can be interpreted as the amount of variance explained by the predictor variables.

- *Heat Maps of the Covariance Structure*

The leveraged algorithm can help identify the pairs of the predictor variables that explain the variance in individual response variables and/or the covariance between multiple response variables. The hierarchical clustering technique can then be used to group the predictors that explain covariance in similar pairs of response variables, and the pairs of responses that are dependent on similar subsets of predictors; the results can then be illustrated as a heat map [101].

- *Partial Dependence*

A crucial aspect of statistical inferencing is determining the nature of the statistical relationship between the most important predictors and the response variables. For non-parametric models, partial dependency analyses are conducted to characterize the association between the inputs and the response variable(s). The partial dependence can be calculated using the following equation [47]:

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(x, x_C^{(i)}) \tag{3.4}$$

Where x is the predictor of interest and $x_C^{(i)}$ represents the other predictor variables that are not of interest. The estimated partial dependence, $\hat{f}(x)$, is the average value of the response variable, when only the predictor variable of interest is considered.

- *Model Comparison*

Finally, the results from the multivariate model were compared to results from a similar univariate model. Specifically, gradient tree boosting [**?**] was used to predict the water and electricity use as isolated variables. Gradient tree boosting is the basis for multivariate tree boosting [101], thus the main difference between the multivariate and univariate algorithms is the consideration of response variable dependencies. The purpose of this final analysis was to demonstrate the value of the multivariate framework, as this is the first time this coupled methodology has been applied to predicting the climate-sensitive portion of the water-electricity nexus.

3.2.2 Results

Following the modeling process outlined above (see Figure 3.2), the climate-sensitive portion of the interdependent water and electricity demand was estimated for each city in the study area. In this section, I will first describe the model performance, then discuss the results from the various statistical inferencing techniques, including the covariance explained evaluations and the partial dependence visualizations, before describing the comparison between the multivariate and univariate model performance.

3.2.2.1 Model Performance

To develop a predictive model of interdependent urban water and electricity demand, the multivariate tree boosting algorithm described in Algorithm 1 was leveraged. In the initial training of model, several independent variables that could potentially affect water and/or electricity demand were included (see Table 3.1). The final model included a reduced variable set based on the relative influence each predictor had over the predictive accuracy. The variables in the final model included maximum dry bulb temperature, average dew point temperature, average relative humidity, average wind speed, and the El Niño/Southern Oscillation index. The selected variables were similar to previous studies on the sensitivity of water demand [89] and electricity

demand [80].

Treatment of Seasonality

As part of the data preprocessing, the response variables were adjusted for seasonality. It has been shown that seasonality aids in the predictive accuracy, but in such a way that is misrepresentative of the actual system [99]. In other words, seasonality may mask the signals of long-term trends, such as those related to climate change. Here we present the results from the model performance using both the original dataset and the seasonally adjusted dataset to demonstrate the difference between them. Without the seasonality adjustment (i.e., the original dataset), the model performance was better (see Table 3.2 and Figure 3.3a), which aligns with previous work on the effect of seasonality on models. However, since the interest of this paper is the impact climate, an inherently long-term concept, the seasonality may be masking the true signal, thus including the seasonally adjusted dataset become important as well (see Table 3.3 and Figure 3.3b).

Measures of Model Performance

The performance of the final model was assessed based on the out-of-sample estimates of the coefficient of determination (R^2) and the root-mean-squared error (RMSE). These measures of error were calculated using the test set. Based on the R^2 values shown in Tables 3.2 and 3.3, demonstrate that climate variables alone can account for a significant fraction of the variability in the electricity and water demand—ranging from 43%-73% (i.e., R^2 values of 0.43-0.73) in the in-sample performance and 30%-71% (i.e., R^2 values of 0.30-0.71) in the out-of-sample performance, after seasonality was removed from the dataset.

Table 3.2.: The model performance for each city during the initial demand nexus model development phase using the original dataset (i.e., the dataset with seasonality intact). The in-sample measures were calculated using the same data used to train the model, while the out-of-sample measures were calculated using the test dataset, which was not included in the model training (see Figure 3.2).

City	Water Use				Electricity Use			
	in-sample R^2	in-sample RMSE	out-of-sample R^2	out-of-sample RMSE	in-sample R^2	in-sample RMSE	out-of-sample R^2	out-of-sample RMSE
Chicago	0.71	0.333	0.47	0.731	0.85	0.235	0.76	0.499
Columbus	0.82	0.285	0.78	0.619	0.89	0.218	0.84	0.496
Indianapolis	0.89	0.222	0.83	0.491	0.94	0.160	0.87	0.385
Minneapolis	0.88	0.221	0.81	0.468	0.91	0.197	0.83	0.431
Cleveland	0.51	0.452	0.31	0.876	0.81	0.306	0.77	0.566
Madison	0.79	0.290	0.71	0.623	0.85	0.226	0.77	0.450

Thus, while the previous literature primarily focused on explaining the variance in the demand as a function of socioeconomic and technological factors as well as cultural norms, in this study, there was a focus on isolating the effects of climate variability and demonstrated the significant role of climate in explaining the covariance of the water-electricity demand nexus.

The results summarized in Tables 3.2 and 3.3 indicate that a significant fraction of variability (i.e., relatively large R^2 values) in the water-electricity demand nexus can be explained by the input climate variables. This is futher demonstrated in Figure 3.3, which shows the predicted values plotted against the actual values for both the original dataset (Figure 3.3a) and the seasonally adjusted demand data (Figure 3.3b). The results are illustrative of the fact that climate variability is an important driver of water and electricity use in Midwestern cities.

Table 3.3.: The model performance for each city during the initial demand nexus model phase using the seasonally adjusted dataset. The in-sample measures were calculated using the same data used to train the model, while the out-of-sample measures were calculated using the test dataset, which was not included in the model training (see Figure 3.2).

City	Water Use				Electricity Use			
	in-sample R^2	in-sample RMSE	out-of-sample R^2	out-of-sample RMSE	in-sample R^2	in-sample RMSE	out-of-sample R^2	out-of-sample RMSE
Chicago	0.69	0.344	0.51	0.720	0.53	0.457	0.39	0.932
Columbus	0.63	0.416	0.62	0.894	0.49	0.500	0.31	0.975
Indianapolis	0.73	0.327	0.71	0.739	0.53	0.455	0.41	0.934
Minneapolis	0.69	0.333	0.55	0.761	0.50	0.467	0.42	1.113
Cleveland	0.44	0.490	0.23	0.910	0.46	0.509	0.34	0.943
Madison	0.54	0.444	0.34	0.925	0.43	0.512	0.30	1.003

3.2.2.2 Statistical Inferences from the Multivariate Model

One of the advantages of the proposed multivariate approach is the ability to determine the covariance explained by the predictors for each individual response variable and the nexus between response variables. This feature allows us to see what variables have the most impact on the water-electricity nexus and if those variables differ from those most greatly impacting water or electricity use alone.

Figure 3.4 shows the clustered heat maps of the covariance explained for each city. These heat maps are clustered via hierarchical clustering, which indicates which predictors are affecting the response variables in similar ways, as well as which response variables pairs are being influenced by similar subsets of predictors. Overall, assessing the covariance explained allows us to investigate the similarities and differences between the cities, as well as any differences between the isolated water use, isolated electricity use, and the water-electricity use nexus. The results from the heat maps demonstrate that although the model itself is generalizable across the different cities, as indicated by the model performance (see Table 3.3), the covariance explained by the variables will differ from city to city. For example, in the land-locked cities

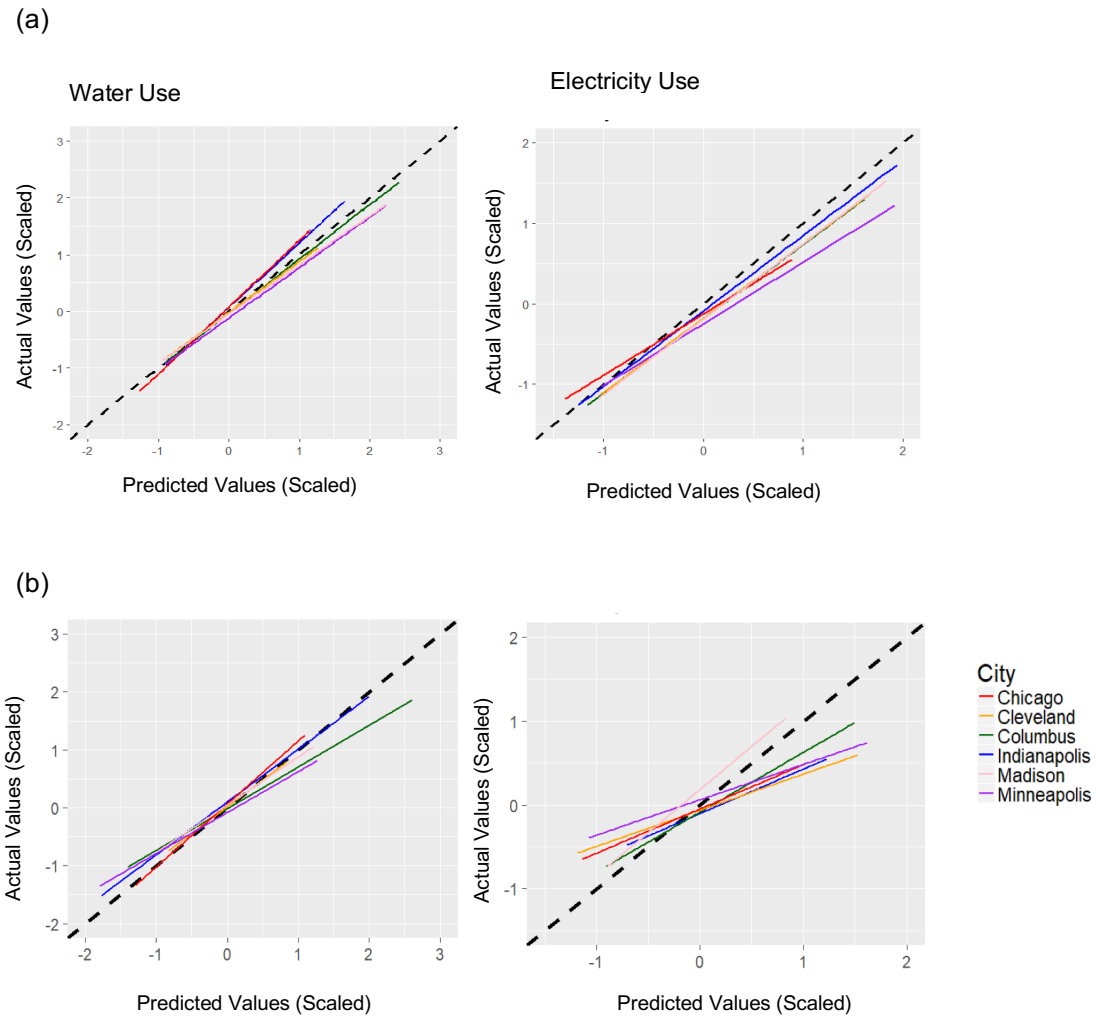

Fig. 3.3.: Out-of-sample model performance during the initial demand nexus model development phase for (a) the original dataset (i.e., the dataset with seasonality) and (b) the seasonally adjusted dataset with the multivariate model. The response variables, water and electricity use, have been scaled to account for different units of measurement. The lines are best fit lines plotted through the predicted versus actual points, with a 45° dashed line for reference.

of Columbus, Indianapolis, and Minneapolis, average relative humidity explains the most covariance in water use. This is different than the coastal cities of Chicago and Cleveland, where the ENSO index explains much of the water use and relative humidity has less of an impact.

Fig. 3.4.: Clustered heat maps showing the covariance explained by each predictor variable in each city in the initial model run, after the seasonality was removed from the dataset. The darker blues represent higher values of covariance explained, while the lighter blues represent less. The variables have been grouped using hierarchical clustering, a method used to group similar objects together. In this figure, predictors clustered together explain the covariance in similar outcome pairs, therefore, the position of the variables on the axes is different for each city due to each city has a different clustering outcome.

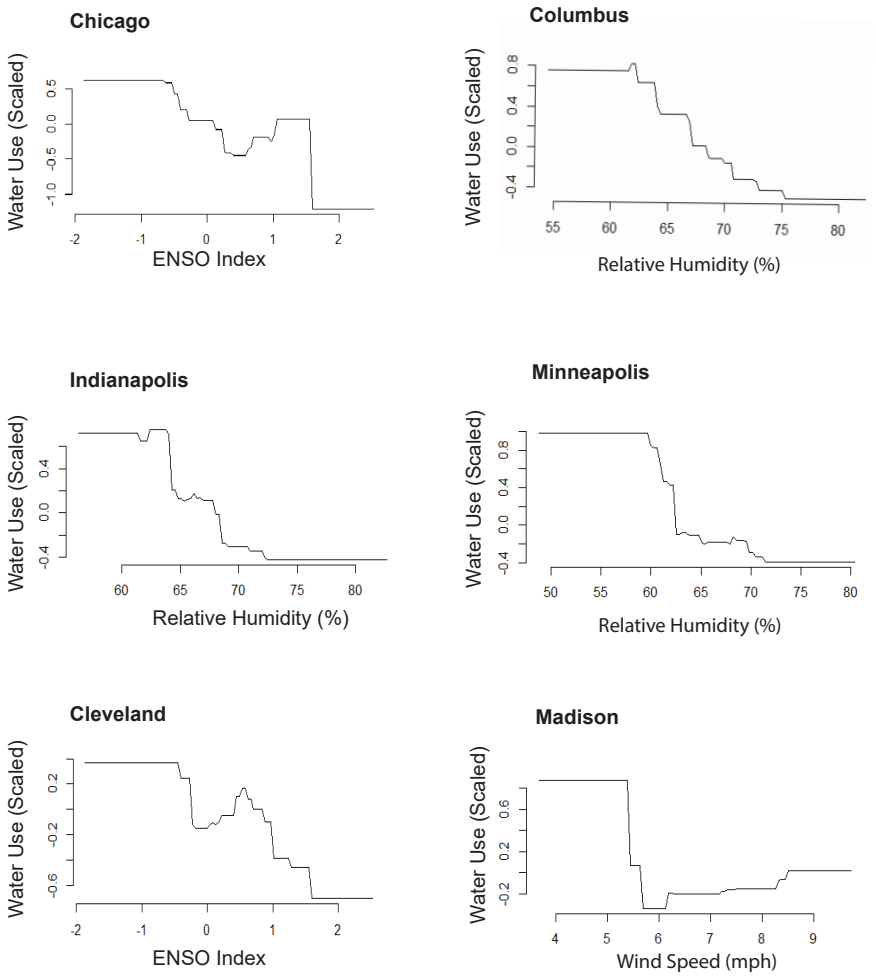

Fig. 3.5.: Partial dependence plots between the most important predictor variable and water use in each city included in the initial model development run. Note that the water use has been scaled, so there are no units.

The covariance explained, however, does not give any indication to the direction of the relationship between the predictors and the response variables—just the magnitude of it. Thus, it is necessary to perform other analyses to determine if higher relative humidity will lead to higher or lower water use in Indianapolis, for example. To answer this question, the partial dependence of the predictors on the individual response variables was evaluated. A selection of these partial dependence plots are shown in Figure 3.5 (additional partial dependence plots can be seen in Appendix B).

These plots show the relationship between the most important variables and water use in each city. In particular one can see that in the cities of Columbus, Indianapolis, and Minneapolis, as relative humidity increases, the water use decreases. A similar pattern appears in Chicago and Cleveland—as the El Niño gets stronger, the water use decreases. This suggests that utility managers trying to reduce water use in Columbus or Indianapolis should focus on the days with intermediate relative humidity, as that is when people are using the most water. Likewise, a manager in Chicago or Cleveland should focus their demand reduction efforts during the cold phase of the El Niño cycle (i.e., La Niña).

3.2.2.3 Univariate Model Comparison

One of the goals of this work was to demonstrate the power of including both water and electricity use in the model as interdependent response variables. This was done through a model performance comparison of the multivariate tree boosting model and a univariate version: gradient tree boosting. The results from the univariate model run are shown in Table 3.4.

Both approaches revealed that a significant fraction of the variability in the water and electricity use could be accounted for by climate variables alone. Additionally, the relative performance of the various cities matched between the univariate and multivariate models. For example, in both approaches, Indianapolis's water use was found to be most climate-sensitive, while Cleveland's revealed the least amount of

climate sensitivity (based on their estimated coefficients of determination). Overall, however, the multivariate model was better at capturing the climate sensitivity of two demands than the univariate model, with the exception of Cleveland's and Madison's water use.

The main difference between the univariate and multivariate models was the inclusion of response variable interdependencies within the multivariate model. This is indicative that, in most cases, the consideration of the interconnectivity between water and electricity use improves the final prediction of both water and electricity use. Of the cities tested as a part of this analysis, the climate sensitivity of water use in Cleveland and Madison—smallest cities included in this study—were better accounted for by the univariate model, which suggests a loose coupling between the climate-sensitive portion of the water and electricity use in those cities than the other cities studied. Additional research is necessary to determine the reason behind this reduced coupling between the climate-sensitive portion of the water and electricity demand.

Table 3.4.: The in-sample and out-of-sample model performance (R^2 and RMSE) of the univariate model, gradient tree boosting, for each city after the seasonality was removed from the data.

City	Water Use				Electricity Use			
	in-sample R^2	in-sample RMSE	out-of-sample R^2	out-of-sample RMSE	in-sample R^2	in-sample RMSE	out-of-sample R^2	out-of-sample RMSE
Chicago	0.60	0.437	0.50	0.747	0.36	0.600	0.32	0.981
Columbus	0.55	0.500	0.53	0.860	0.36	0.601	0.26	0.987
Indianapolis	0.62	0.429	0.64	0.732	0.39	0.577	0.29	0.938
Minneapolis	0.56	0.451	0.55	0.756	0.32	0.599	0.32	1.003
Cleveland	0.34	0.614	0.36	0.860	0.30	0.630	0.28	0.991
Madison	0.41	0.548	0.37	0.883	0.28	0.663	0.28	1.036

3.2.3 Discussion

This study focused on analyzing the water-electricity demand nexus based solely on climate variables. This allowed us to isolate the effect of climate on residential water and electricity use—a factor that is often not included in demand analyses. The results show that water use is more climate-sensitive in most of the cities included. This suggests that water use is more dependent on the climate than electricity use, which is an interesting finding, given the documented increase in electricity with increasing temperatures in the Midwest [85].

Given that the model performance for the electricity sector was more impacted by the seasonality adjustment than the water sector, the results suggest that in the Midwest, the long-term climatic conditions are more likely to drive changes in water use, while the short-term weather patterns are more likely to act as a driver for electricity use. That is not to say that climate is the only driver of changing water use, but rather it is a potentially important driver that has often been left out of many demand analyses. In this sense, water demand studies, which often focus on population, socioeconomic, and/or cultural factors, ought to also include climatic factors in their analyses. This will become especially important as researchers and practitioners try to predict water demand under climate change.

One of the main findings of this study was the importance of the El Niño cycle on the residential water and electricity demand in the region of interest. The ENSO index was consistently among the predictors that explained the most covariance in the response variables. Given that the El Niño cycle is a well-documented climate phenomenon that can be predicted relatively easily, it is an ideal variable for making more general or broad predictions. For example, a common ENSO-based prediction is the type of winter that a given region will have (e.g., a strong El Niño usually leads to warmer, drier winters in the Midwest [98]). This modeling framework allows us to make a simple, first order forecast for the demand nexus based on large scale climate predictor. In other words, the results suggest that a strong El Niño is more

likely to lead to lower water and electricity use . This knowledge would allow utility managers to prepare for the upcoming season based on the predicted El Niño strength that is determined on a monthly basis. The importance of the ENSO index also has implications for climate change. It is likely that El Niños will become stronger as sea surface temperature continues to increase [102], and the results suggest that if this holds true, water and electricity use in the Midwestern cities studied, will decrease as a result of the change in climate, should everything else in the cities remain constant. This assumption—that the population, socioeconomic breakdown, culture, etc. of a city will remain constant—is, of course, highly unlikely; however, the results demonstrate the importance of including climate variables in the overall analysis of water and electricity demand.

Finally, one of the goals of this study was to compare the results from the multivariate model, which considers the coupling between water and electricity demand, and a univariate model that is based on the same algorithm. The results demonstrate that the multivariate framework is able to better capture the climate-sensitivity of water and electricity use in most cases. Since both models were based on the same algorithm, the only difference between them being the inclusion of multiple interconnected response variables, the results suggest that system coupling are an important consideration for the prediction of water and electricity demand. Ultimately, the results indicate that there needs to be an increased effort to (i) consider the increasing role of climate drivers on demand and (ii) harness a multivariate framework to better account for the interdependent response variables in demand analyses.

3.2.4 Summary

The purpose of this study was to build a multi-response predictive model of the portion of the urban residential water-electricity demand nexus that was sensitive to climate, using the multivariate tree boosting algorithm. In this study, there were two response variables: water use and electricity use, and five main predictors. The model

was tested on six Midwestern cities of variable size, demonstrating the generalizability of the model to the region of interest. The results of the study indicated that a significant fraction of the water-electricity demand nexus can be explained by climate variability alone. Urban water and electricity demand are impacted by a number of factors, including population density, socioeconomic status, and cultural values, in addition to the climate. However, the role of climate has been understudied in comparison to other important drivers of urban water and electricity demand. For this reason, the goal in this study was to isolate the effects of climate and demonstrate the value of their inclusion in future analyses. The results indicated that water and electricity use are sensitive to climate variables, and will likely be affected by future climate change. The impact of the El Niño cycle was especially important in each city, as the variable consistently explained much of the covariance in the water-electricity nexus and in the individual response variables.

3.3 Regional Demand Forecasting

The previous section demonstrated the value of the developed multivariate model for predicting the climate-senstive portion of the water-electricity demand nexus. However, given that there is a significant seasonal aspect to water and electricity use, especially as it relates to the climatic conditions, I decided to update the initial model to account for the separate seasons. Additionally, given the large-scale nature of climate change, it is potentially more reasonable to build a regional model, rather than city-specific models. Therefore, the initial model was also updated to be regional, in addition to the separation of seasons.

With this in mind, the purpose of this section is two-fold: (1) to present an updated *regional model* for predicting the interconnected water and electricity use in different periods throughout the year, and (2) use that model to project the water and electricity use into the future under various climate change scenarios. The focus of this study is to isolate the impact of climate change on the water-electricity demand nexus,

therefore, only climate variables were considered as predictors within the modeling framework. Additionally, this study considers a wider array of climate variables than previously considered in other future projection studies. The Midwest region of the United States, which has several established cities of varying populations, was selected as the test region, however, the proposed modeling framework presented here could be applied to different regions.

3.3.1 Data and Methods

There are a growing number of frameworks being developed to model the coupled water-electricity nexus, however, there are few that take a variety of climate variables into account when making future projections. The proposed framework is novel in that it accounts for a larger array of climate variables to assess their impacts on the coupled water-electricity demand nexus at a regional scale. In this section, we will first describe the study area and the data used in the model before discussing the modeling process and analysis.

3.3.1.1 Site Description

In this study, the Midwestern region of the United States was selected as the study area (see Figure 3.1 in Section 3.2.1.1). Specifically, six established cities were chosen to be included in our regional model: Chicago (IL), Cleveland (OH), Columbus (OH), Indianapolis (IN), Madison (WI), and Minneapolis (MN). These cities, and the region as a whole, can expect to see higher temperatures and increased precipitation due to climate change [96], which will increase the vulnerability of the utility companies. Moreover, these cities have different water and electricity utilities, that do not always work together, which puts them at risk for disadvantageous management decisions in the face of climate change.

3.3.1.2 Data Description

There were two stages of data collection in this study: observational data (for model training, testing and validation) and climate model outputs (for conducting future projections). The first stage included response data (i.e., water and electricity use) from the US Energy Information Administration (EIA) and local utilities, as well as predictor data (i.e., climate variables) from the National Centers for Environmental Information (NCEI) and the National Oceanic and Atmospheric Administration (NOAA). Specifically, the response variables, residential electricity use and residential water use, were obtained through the EIA [97] and local utilities, respectively. The predictor variables were obtained from the local climatological dataset maintained by NCEI [39] in addition to the El Niño database maintained by NOAA [41]. This observational data, which is listed in Table 3.1 (in Section 3.2.1.2), was collected from January 2007 through December 2016 on a monthly time scale. The response variables (water and electricity use) were normalized by the service population, so as to make each city comparable in our regional model. Additionally, the response data was de-trended following a procedure that is well-established within the literature [80, 87, 103] to remove the trends associated with technological advancements as well as socioeconomic and demographic changes over time. This process, which is further described in Appendix B, is especially important for this study, since isolating the climate impact was one the main goals.

The second stage of the modeling process focused on making the future projections using the developed model. For this, climate data was taken from five CMIP5 global circulation models (GCMs), namely: the Geophysical Fluid Dynamics Laboratory - Earth Systems Model (GFDL-ESM2M), the Hadley Centre Global Environment Model (HadGEM2-ES), the Institut Pierre Simon Laplace Model (IPSL-CM5A-LR), the Model for Interdisciplinary Research on Climate - Earth Systems Model (MIROC-ESM-CHEM), and the Norwegian Earth System Model (NorESM1-M). These datasets included both the historical (1971-2005) and the projection time

frames (2006-2099). The projection data were considered for two extreme future emission scenarios that have end-of-century radiative forcings equal to 2.6 Wm^{-2} and 8.5 Wm^{-2}, denoted hereafter as RCP2.6 and RCP8.5 respectively. The GCM data was made available from the Inter-Sectoral Impact Model Intercomparison Project (ISI-MIP) [104] after screening through multiple GCMs from the CMIP5 archive (see the protocol-report available on www.isimip.org for more information). The climate data was downscaled and bias-corrected at a 0.5° global resolution using a trend-preserving approach based on the WATCH observation data [105]. Notably this projection data has been used in several impact assessment studies including the recent AR5 and SR1.5 reports of the Intergovernmental Panel on Climate Change (IPCC) [96, 106]. The data was extracted for the respective cities for each predictor variable included in the final model at a monthly time scale to be used in making future projections of the interconnected water and electricity use.

3.3.1.3 Modeling Framework

The modeling framework used in this study is similar to that presented in Section 3.2.1.3. In fact, the algorithm remained the same as that shown in Algorithm 1, but the data was aggregated as a region, then separated into seasons. Additionally, the analyses performed in this study focused more on the future projections of water and electricity demand.

There are three main steps to the modeling process, as shown in Figure 3.6: (1) data collection, aggregation, and preprocessing; (2) model training and testing with observational data; and (3) future projections using climate model output. In this first step, the data was collected as described above. The observational data was aggregated across the cities and grouped into three time periods according to a well-documented energy economy model known as the MARKet ALlocation (MARKAL) model [107]: Summer months (June-September), Winter months (December-March), and Intermediate months (April, May, October, November) , to account for the sea-

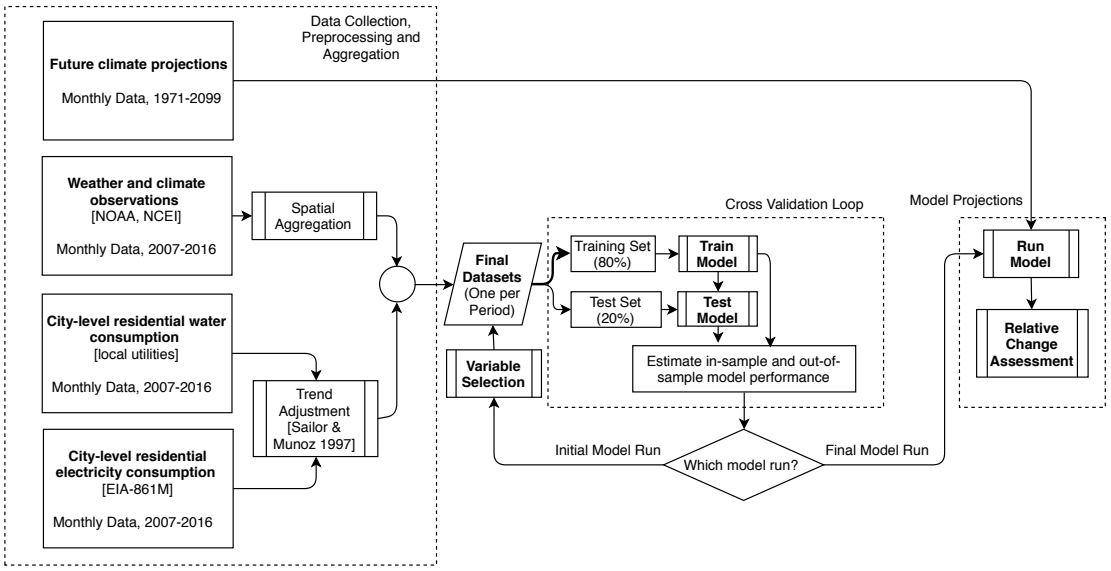

Fig. 3.6.: Schematic for the second iteration of the modeling framework used in the demand nexus analysis.

sonal fluctuations in the water-electricity demand nexus. These new datasets were the initial inputs to three separate models—one for each period.

The model training and testing was the next step. This step used the observational data to first determine the important predictor variables and then validate the predictive accuracy of the model. The important variables were selected based on a threshold criterion: any variable with a relative influence greater than 5% on either response variable in a given period was kept for that period's final model. Ultimately, five variables were kept: average maximum dry bulb temperature, average dew point temperature, average relative humidity, average wind speed, and accumulated precipitation.

The final step was to project the water and electricity use into the future using the climate data obtained from the global circulation models. The data was collected and separated into seasonal periods following the same process described for the observational data. Then, the model was run using the previously selected important variables.

3.3.1.4 Future Projection Analysis

The future projection analysis was performed in accordance with the recent IPCC SR1.5 report [106] to analyze the respective changes in water and electricity use for different global warming levels. Using the historical period (1971-2000) as the reference values, the percent change between the 30-year historical period and the 30-year future periods corresponding to three global warming levels (1.5, 2.0, and 3.0°C above pre-industrial levels) was calculated. A time-sampling approach [108], which has been recently adopted in several impact assessment studies [109–112], was used to identify the corresponding 30-year future periods. In this approach, the warming during the reference period, which was approximately 0.46°C warmer than the pre-industrial global mean temperature (1881-1910), was established based on several observational datasets [109, 113]. Using this offset value (i.e., 0.46°C), the 30-year periods were identified for each of the 10 GCM-RCP combinations (i.e., 5 GCMs $\times$ 2 RCPs) in which the global mean temperature increased by 1.04, 1.54, and 2.54°C respective to the reference period. These periods correspond to the 1.5, 2.0 and 3.0°C temperature thresholds used in the analysis. The future projections were obtained for each climate model simulation for two warming scenarios: low-warming (RCP2.6) and high-warming (RCP8.5). It should be noted that under the low-warming scenario, the 3.0 degree temperature threshold is not reached, and therefore was not included in the analysis.

3.3.2 Results

Following the modeling framework outlined in Figure 3.6, the interconnected water and electricity use was projected into the future under various climate change scenarios. In this section, we first discuss the model performance with the observational data and compare it to a conventional precipitation-temperature model before delving into the future projections of water and electricity use.

3.3.2.1 Model Performance

As described above, the first part of the analysis in this study was building the regional model for three different periods—summer, winter, and intermediate months. The main goal of this first task was to demonstrate the effectiveness of the proposed modeling framework that makes use of a larger array of climate variables than the baseline model that considers only precipitation and temperature. As shown in Figure 3.7, the *Selected Feature* model (i.e., the proposed model) tends to predict the water and electricity use more accurately than the *Baseline* model (i.e., the model that only considers precipitation and temperature, denoted 'precip-temp' in the figure). This is especially true in the extreme ends of the consumption patterns, where predictive accuracy is crucial. Moreover, the difference between the Selected Feature and Baseline models is more pronounced in the water use, for both summer and winter periods. This indicates that the additional variables considered in the Selected Feature model—dew point temperature, relative humidity, and wind speed—are more influential when predicting water use compared to electricity use. Figure 3.7 shows the results from the summer and winter periods; the results from the intermediate period can be found in Appendix B.

The improved performance of the Selected Feature model is further demonstrated in Figure 3.8, which compares the model performance measures for both models during the summer and winter periods (see Appendix B for the model performance during the intermediate period). Both the out-of-sample RMSE and out-of-sample R^2 were used to assess the model performance. RMSE is a measure of error, in which lower values are representative of a better prediction (i.e., less error). Often, RMSE is used to evaluate the predictive performance of the model. On the other hand, R^2 can be thought of as a measure that accounts for the percent of variance within the data that is explained by the model. In this sense, a value closer to 1 indicates that the model is explaining more variance in the data. That being said, R^2 is rarely used to assess *predictive* performance, as it is not a measure of error. Together,

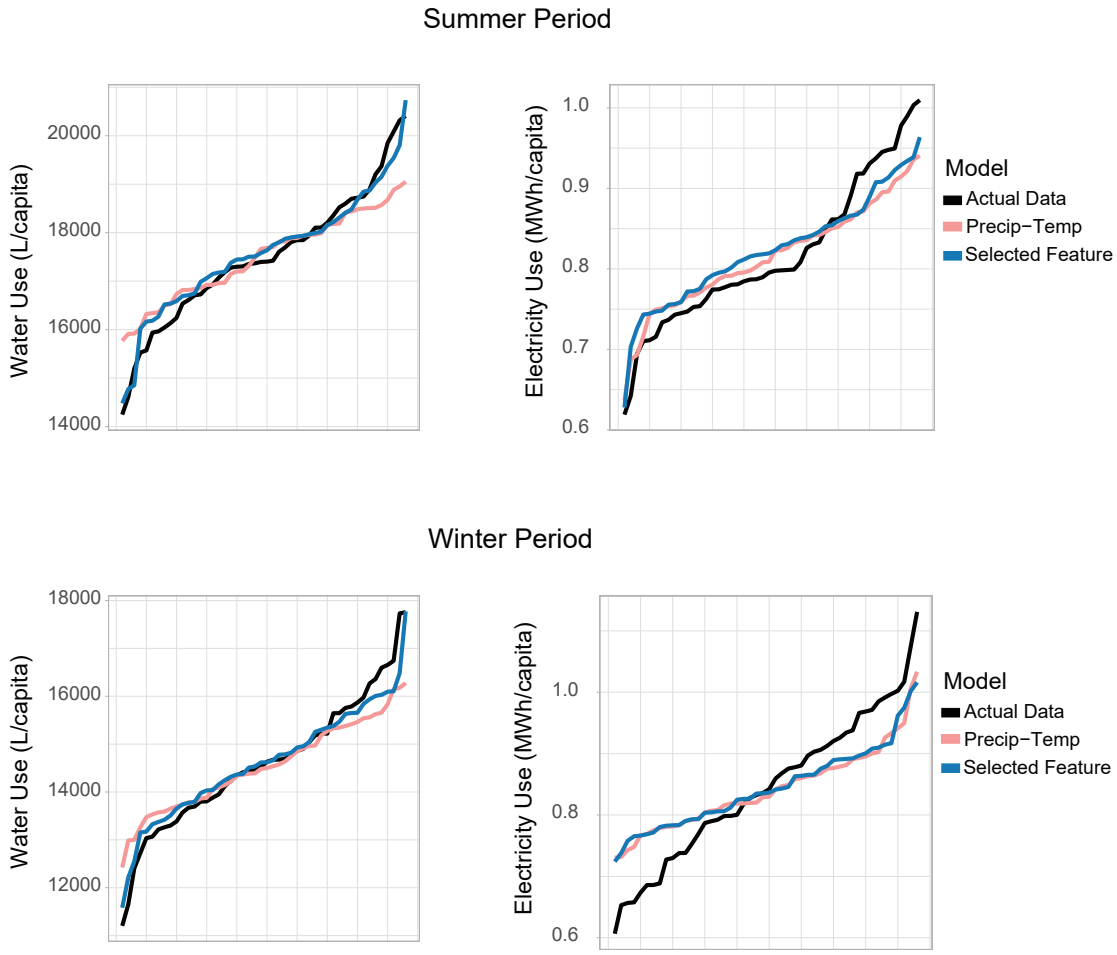

Fig. 3.7.: Observational data compared with the two types of demand nexus model runs: (1) a Baseline model that only considered precipitation and temperature (denoted 'Precip-Temp) and (2) the proposed Selected Feature model that considers a larger array of climate variables (denoted 'Selected Feature'). The results are presented for the summer and winter periods.

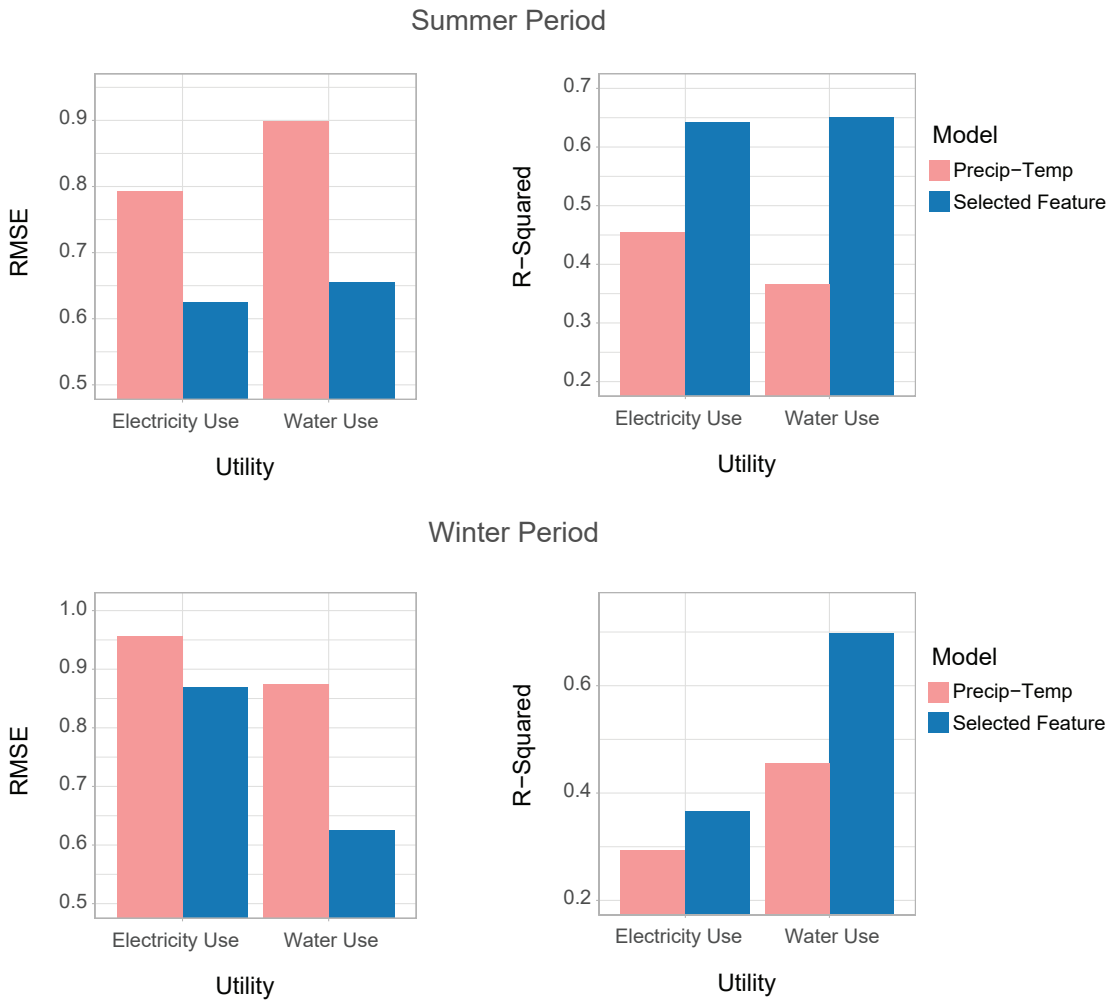

Fig. 3.8.: Out-of-sample model performance results (RMSE and R^2) from the two styles of demand nexus model runs: (1) a Baseline model that only considered precipitation and temperature (denoted 'Precip-Temp') and (2) the Selected Feature model that considers a larger array of climate variables (denoted 'Selected Feature'). The results are presented for the summer and winter periods.

however, RMSE and R^2 can be used to assess overall model performance—both from a predictive standpoint and the amount variance the model is able to capture.

3.3.2.2 Future Water and Electricity Use Projections

Following the analysis with the observational data, the selected feature model was used to make future projections of the climate sensitive portion of the water and electricity use in the region. The predictor variables were obtained from the five CMIP5 global circulation models discussed earlier. The purpose of this analysis was to show the potential change in water and electricity use due to climate change alone. In this sense, there was no consideration of technological changes or cultural shifts that would also have an impact on the water and electricity use. To evaluate the potential shifts in future water and electricity use, the percent change was calculated between the 'historical' period (1971-2000) and the 30-year period in which key temperature thresholds were reached within the model. The historical baseline data from 1971-2000 can be found in Appendix B. These thresholds—1.5, 2.0, and 3.0 °C—were selected based on several recent climate change assessment studies [106, 109–112]. Initially, the percent change was calculated based on all the model output, regardless of the future pathway scenario, followed by a scenario-specific (i.e., RCP2.6 and RCP8.5) calculation. Figures 3.9 and 3.10 show the results of this analysis for both the summer and winter periods (see Supplementary Figure S4 for the intermediate period projections).

In general, the water use is projected to increase after all three temperature threshold scenarios and in both periods (see Figure 3.9), but the electricity use is only projected to increase in the summer period (see Figure 3.10). For water use in particular, as the temperature continues to increase (i.e., higher thresholds are reached), the percent change in median water use also increases. In fact, in the summer period, the results indicate a relative increase in water use regardless of the temperature threshold or warming scenario. Given that the 1.5 degree threshold is approaching, these results demonstrate the necessity for Midwestern water utilities to prepare for increased summer demand in the near future. Similar results were shown for the summer electricity use in Figure 3.10. Interestingly, the model shows a median de-

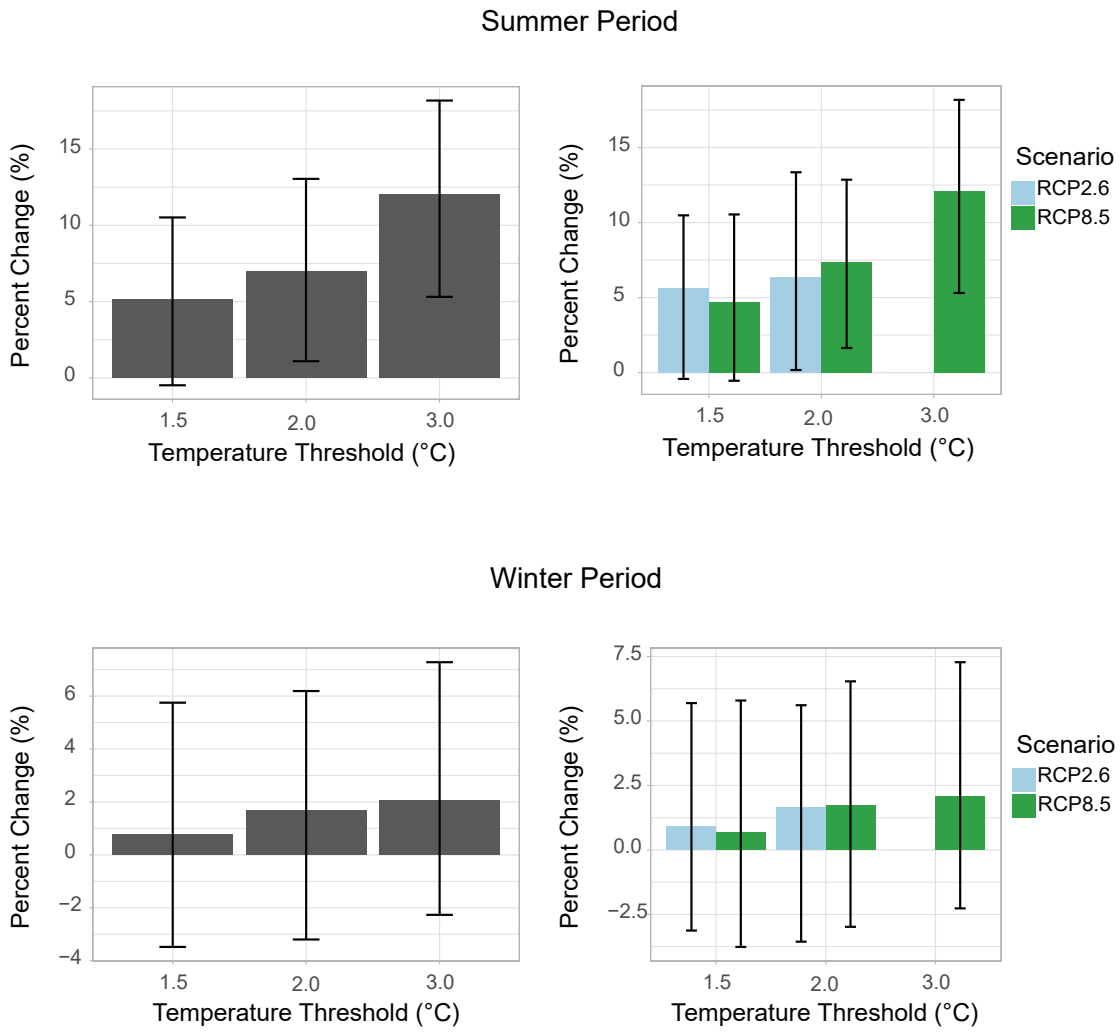

Fig. 3.9.: The median relative change in water use after three different temperature thresholds have been reached in the summer and winter periods. The error bars represent the interquartile range. The left plots show the aggregate of both warming scenarios and the right plots show the same change, separated by scenario. Note that under the low-warming scenario (RCP2.6), the 3.0°threshold is not reached before 2100, and thus is not shown in the figure.

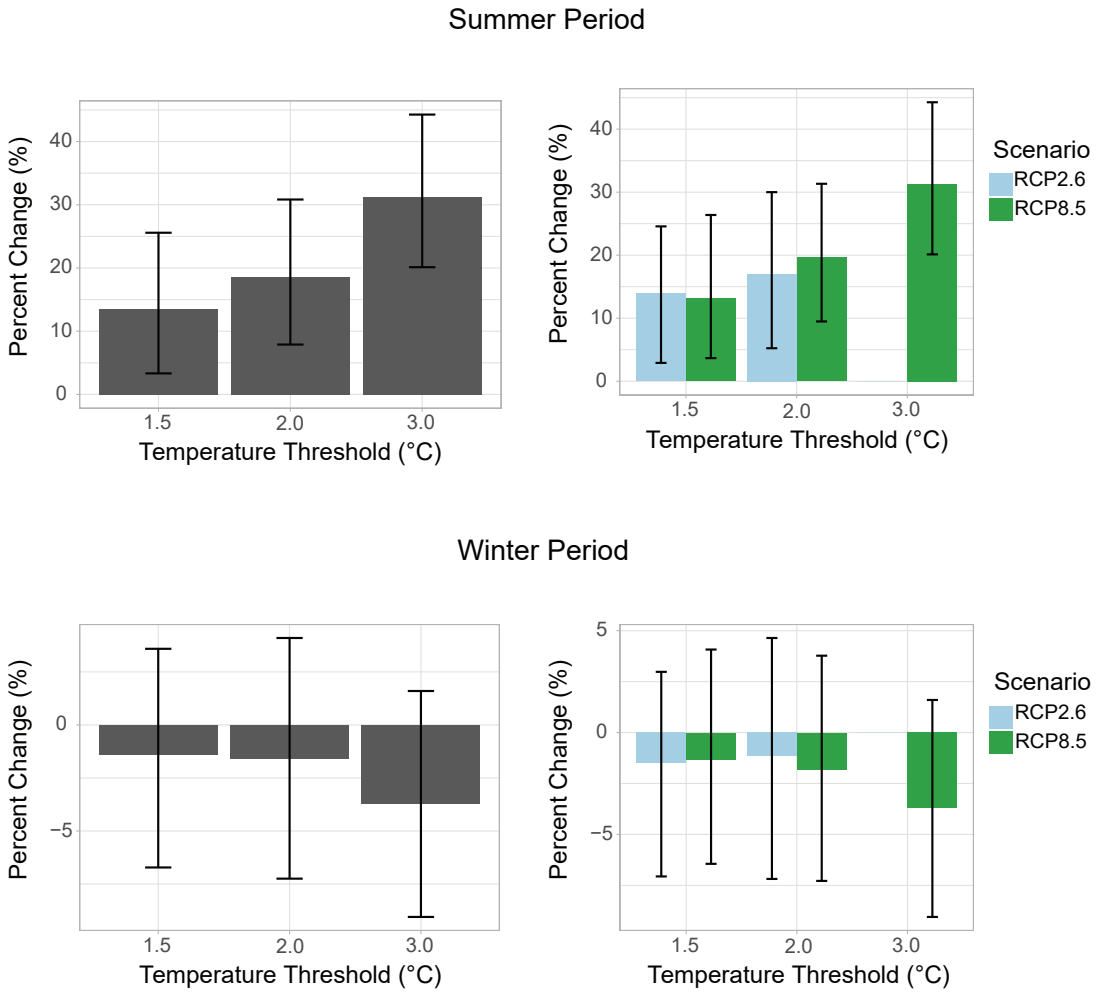

Fig. 3.10.: The median relative change in electricity use after three different temperature thresholds have been reached in the summer and winter periods. The error bars represent the interquartile range. The left plots show the aggregate of both warming scenarios and the right plots show the same change, separated by scenario. Note that under the low-warming scenario (RCP2.6), the 3.0°threshold is not reached before 2100, and thus is not shown in the figure.

crease in winter electricity use across all the thresholds and scenarios, but especially after the 3.0 degree threshold is reached. This is likely due to warming temperatures and a reduced need for space heating in the winter, which is a major contributor to electricity consumption during the winter months. However, this potential reduction in winter electricity use was not offset by the potential increase in summer electricity use, making it imperative that utilities begin to find ways to increase their supply capabilities. In addition to the results presented here, the model-specific projections can be found in Appendix B.

3.3.3 Discussion

This study focused on building a regional model to simultaneously project the climate sensitive portion of interconnected water and electricity use into the future under various climate change scenarios. There were two main parts of the analysis, the first of which was to build a rigorously tested predictive model (i.e., the Selected Feature model), using a variety of climate variables, and compare the predictive accuracy to the Baseline model that only considered precipitation and temperature. The results from this comparative analysis showed a significant improvement over the Baseline model when dew point temperature, relative humidity, and wind speed were included in addition to the standard dry bulb temperature and precipitation. In fact, initial results indicated that by including the average daily maximum values for relative humidity and wind speed, rather than the daily averages included here, there were additional improvements over the Baseline model—especially in the winter period. However, the GCM projections of these daily maximum variables are not readily available for downloading, nor are they easily extractable from the model output directly. Since the aim of this modeling framework is to provide practitioners with a tool to make projections for their own systems, it was decided to include the daily averages instead of the maximums, as the climate projections are easier to obtain. In future iterations, including these maximum values in the model may

lead to more accurate projections. Nevertheless, the selected feature model developed here did show significant improvement, especially on the extreme ends of the demand profile (see Figures 3.7 and 3.8).

The Selected Feature model was used in the second part of the study, which was to make future projections of water and electricity use based on future climate change scenarios. These results indicated a likely increase in both water and electricity use during the summer periods (see Figures 3.9 and 3.10), with minimal uncertainty. During the winter period, however,there was more uncertainty in the projections, although the model still showed a median increase in water use and a median decrease in electricity use. Previous work indicated that warmer temperatures led to increased water use [114], likely due to increased consumption for landscaping purposes. Landscaping, however, is generally only a summer demand pattern. It is possible that the increased temperatures allow for some winter landscaping in the more southern cities, which could explain the slight median increase in water demand. However, the large uncertainty bands make this determination difficult without further investigation beyond the scope of this study. In fact, given the range of possible winter temperature projections, as well as the variance introduced by seasonal shifts in the general climatic conditions, it is possible that winter water use will decrease along with electricity use. This decrease in winter electricity use may be due to the warming temperatures, which would lead to a decreased need for space heating in the winter months (but increased space cooling in the summer,hence the median increase in summer electricity use). Ultimately, this winter decrease paired with the summer increase could put additional pressure on the electricity utilities to cope with the seasonal fluctuations. Moreover, both cases represent a potential economic loss to the utility—in the summer, there is a higher chance for shortages, while in the winter, there is a higher chance for surplus, both of which are undesirable for electricity utilities.

In addition to the results indicated by the regional model projections, it is possible to use the regional model to predict the water and electricity use for specific cities. The results from these city-specific projections can be found in Figure 3.11, which

shows the results from the summer period projections for the 1.5 and 2.0 degree temperature thresholds. These two thresholds are the ones that are most likely to be passed in the near future—1.5 °C is projected to be reached around 2030 and 2.0 °C is projected to be reached around 2055—as well as being politically relevant at the international scale. The recent IPCC report, for example, recommended that warming levels be kept below 1.5 °C if the world is to avoid the most detrimental consequences of climate change [106]. But the 2015 Paris Agreement, which has been signed by the majority of countries around the world, argues for a 2.0 °C limit [115]. Either way, these are the main thresholds being discussed at the international level, and are therefore important for utility companies that will need to provide adequate services regardless of the temperature thresholds that are ultimately reached.

In each of these six cities, the patterns of future consumption are similar. Each city, for example, is projected to have increases in both summer water and electricity demand, although the summer electricity is projected to have a larger relative increase. Additionally, there are relative larger changes after the 2°C threshold than the 1.5°C threshold, which is to be expected. There are also some differences between the cities. For example, Chicago is more urban (as opposed to suburban) with less residential green space (i.e., yards) than the other cities on the list. This likely leads to lower summer water consumption for outdoor landscaping, and thus a somewhat lower increase in median water demand when compared to the other cities. Minneapolis is the northern-most city in the analysis, and likely to see less of a severe summer temperature increase then the other cities. This could explain the relatively lower increase in summer electricity than Indianapolis and Cleveland, for example. Interestingly, Columbus and Indianapolis, which are close in population and are geographically similar, are projected to experience different magnitudes of changes to the water and electricity demand profile, with Columbus projected to see less intense changes. This may be due to the sprawling nature of Indianapolis (Indianapolis is approximately 160 mi^2 larger than Columbus), which generally means more single family, detached homes. This would likely lead to increased water use (for

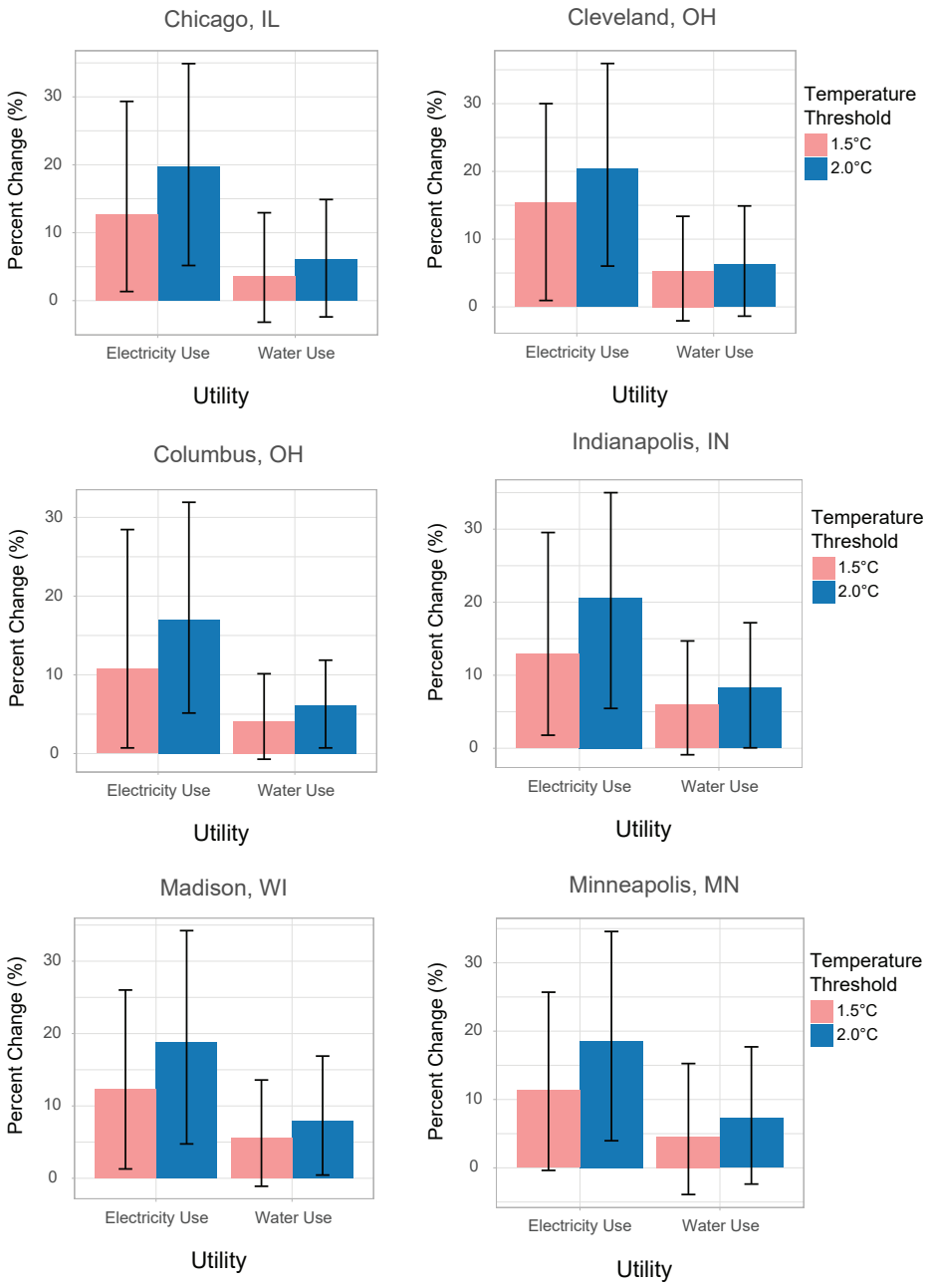

Fig. 3.11.: The median relative change in water and electricity use for the individual cities in the study region for the summer period following two of the three key temperature thresholds. The error bars represent the interquartile range.

landscaping) and electricity use (for space-cooling). In general, the cities all follow the same pattern, although there are some differences. That being said, even the small differences between the cities, as well as the differences between the thresholds can lead to large changes in the total demand.

If one focuses on Chicago, which is the largest city in the region, one can see that summer electricity is projected to increase significantly during the summer months with minimal uncertainty. In fact, after passing the 1.5 degree threshold, Chicago's electric utility could expect to see a 12% increase in per capita demand. Should the economy grow in accordance to the Shared Socioeconomic Pathway (SSP) scenario 1 (i.e., 'sustainable growth'), the population in Cook County (covering the city the Chicago), would be 5.39 million by 2030 [116] which roughly corresponds to the crossing of 1.5 degree threshold. Given a per capita demand estimate of 0.97 MWh/capita in the projected period, this would lead to an additional 745,000 MWh in monthly electricity demand during the summer months, only attributable to climate change. Without technological advances or cultural shifts towards conservation, this increase in demand will become more dramatic which will severely strain existing infrastructure. For example, if the SSP5 scenario is followed (i.e., 'fossil-fueled development'), the population is projected to be 5.71 million by 2030 [116], which corresponds to approximately 1.06 million additional MWh.

This need for technological advances or cultural changes only increases after the 2.0 degree threshold, which is expected to be crossed in the 2050's, should the current trend hold and no significant climate action plans be implemented. In fact, after this threshold the electricity demand in Chicago could increase by 1.6 million MWh (compared to the reference period) should warming not be capped at the recommended 1.5 degree threshold, assuming a population of 6.09 million in Cook County [116]. This intense increase demonstrates the benefit of following the IPCC recommendation and working to cap global emissions from the local utility perspective. Overall, these results signal the importance of making water and electricity use projections

and building models that can be adopted by utility managers that need to prepare for future demand shifts.

That being said, the above future changes in water and electricity demand in the Midwestern region of the United States include a fair amount of uncertainty. The uncertainty presented here as the interquartile range (i.e., the error bars in Figures 3.9 and 3.10) demonstrate relatively larger uncertainty during the winter season in both water and electricity use. In fact, the percent change spans over both positive and negative values—leading to highly uncertain projections. This makes preparing for the future more difficult, since it cannot be said, for certain, what will happen. Although the signal is stronger in the summer months (i.e., there is a demonstrable increase in usage across all scenarios), there is still some uncertainty. Part of this uncertainty comes from the climate models themselves (see Appendix B for the model-specific projections). As discussed earlier, only five climate models were selected, as they are most often used within the literature [96, 106], which introduces bias into the study. However, these are pitfalls that occur with any future projection study. Moreover, this modeling framework has been developed such that it can be applied at a broader scale with a larger number of climate models included. In this sense, although the uncertainty is present, the results can still be interpreted as potential pathways forward, should the outcome of the climate models come to pass.

3.3.4 Summary

The goal of this study was to build a data-driven, regional model to evaluate the impact of future climate change on the coupled water and electricity demand nexus. The modeling framework leverages the multivariate tree boosting algorithm to simultaneously predict the interconnected water and electricity demand in the residential sector. There were two response variables: monthly water and electricity use, and five final predictor variables: maximum dry bulb temperature, average dew point temperature, average relative humidity, average wind speed, and accumulated

precipitation. The proposed Selected Feature model proved to be more accurate than the Baseline model, which only included maximum dry bulb temperature and accumulated precipitation. Many demand projection studies in the past have used only this standard Baseline model, which tended to underpredict the higher demand levels. Accurately predicting these higher demand levels, which represent the peak load, is crucial for utility managers. The results presented here indicate that including additional variables, such as relative humidity and wind speed, could greatly improve the predictive accuracy of peak load forecasting models, which will be beneficial for practitioners.

Additionally, the modeling framework was used to make future projections of the water and electricity demand, given the output from several global circulation models. The results from the projection analysis showed that the summer water and electricity demands can be expected to increase due to climate change. This means that, ultimately, utilities will either need to rely on technological advances or cultural shifts to limit these increases in demand or spend a significant amount of money to expand their supply capacities. On the other hand, the winter demands were slightly more uncertain, but there is a potential that winter electricity use will decrease due to climate change. This will introduce the additional challenge of managing fluctuations, especially for electric utilities, which lack the storage capabilities of most water utilities.

3.4 Conclusions

The purpose of this chapter was to evaluate the impact of climate change on the water-electricity demand nexus. Previous work has focused on the supply-side of the water-electricity nexus [17, 67, 73, 74]. The focus on the demand nexus, on the other hand, has only recently become a topic of interest [19, 25, 76]. Therefore, one of the objectives of this chapter was to develop a data-driven framework that can be used to model the water-electricity demand nexus. When compared to a univariate model

(i.e., one response variable), the developed multivariate framework was shown to be more accurate, as it considered the interdependencies between water and electricity consumption. The results indicate that modeling water and electricity use as a joint variable, rather than in isolation, provides measurable benefits and may be able to aid in the decision processes of water and electric utilities.

Additionally, the climate impact was of special interest, since much of the literature considers only basic variables, such as temperature and precipitation, when assessing the effect of climate on the demand profile [27–29]. It was hypothesized in this work that other variables, such as relative humidity and wind speed, which contribute to the *experienced* temperature would also be important contributors to changes in the demand nexus. This hypothesis proved to be true, especially during the peak demand periods, when accurate predictions are critical for ensuring adequate supply.

Finally, this chapter sought to use the developed multivariate framework to project the water and electricity use into the future. First, this was done using direct outputs from the CMIP5 climate models as input variables. When applied to the Midwest region of the United States, the model projected the median summer electricity and water demand to increase by 19% and 7%, respectively—solely attributable to climate change. In this sense, in order to maintain a similar demand structure, there will need to be significant efforts towards technological advancements in efficiency or cultural shifts toward conservation.

4. EVALUATING THE HUMAN DIMENSION OF WATER DEMAND

4.1 Introduction

It is well known within the social science community that water use is affected by social norms in such a way that water conservation can be encouraged through strategic activation of certain social norms present in a community [31, 75, 117, 118]. However, social norms are rarely integrated or even considered in engineering studies that are interested in water demand. This lack of integration could be the reason behind the failure of some intervention programs to establish long-term behavioral changes in communities that are facing or likely to face intense droughts. The conceptualization of social norms has taken many different forms within the literature, depending on the focus of the study. For this project, a social norm can be thought of as a socially-enforceable behavior or rule that the majority of people within a given group follow, potentially unconsciously [119]. In other words, a norm is a collective behavior or rule that people will not only follow, but also implement social sanctions or punishment for those that deviate from the behavior or rule. This sanctioning or punishment is a powerful deterrent for those who might choose to deviate and often the main difficulty encountered when trying to change social norms. It is also this social enforcement that differentiates a social norm from a personal values or norms [120].

Within social norms there are two categories: descriptive and injunctive norms. Descriptive norms are norms about which people have only empirical expectations, while injunctive norms require both empirical and normative expectations [119]. In this case, empirical expectations refer to what we expect others to do, and normative expectations refer to what we expect others to think we ought to do. Therefore, for

a descriptive norm to exist, there only needs to be an expectation that others will behave a certain way. A common example of this type of a norm is a fashion or fad. We expect others to follow the trend, so we do as well, but there is not an expectation of others thinking we ought to follow the trend. However, for an injunctive norm to exist, there needs to be both an empirical expectation that people will behave a certain way and a normative expectation that others believe we ought to behave a certain way. In a water use study, for example, a descriptive norm might be the expectation to reduce water use when you see your neighbors are conserving (this is a common intervention method used by utility companies). An injunctive norm, however, would include this empirical expectation alongside an emoticon that is specific to how well the user is doing (e.g., a smiling face for below-average users), which indicates that people think you ought to be reducing water [31]. It is likely that both types of norms are important for water conservation and the success of intervention methods aimed at increasing water conservation across a given neighborhood or city.

Some of the main challenges when researching social norms is identifying and measuring them. Often people are not conscious of the norms they are following, making it difficult to identify which norms are being activated in a given situation [121]. Moreover, once the norm is identified, the questions that aim to measure the impact of the norm may lead to inaccurate answers due to respondent's desire to please the experimenter or preserve their self-image [119]. To ensure accuracy it is important to maintain anonymity (and to make sure the respondents are confident in the anonymity) and to provide open-ended questions that do not guide respondents to certain answers. When asking questions aimed at identifying social norms, it is critical to focus on both empirical and normative expectations, as both are important factors in determining the presence and impact of norms [119]. Additionally, it is important to determine the personal normative beliefs of the respondents, as these may or may not align with the social norm. In the extreme case of differing personal and social norms, people will often continue to follow and even participate in social sanctioning of a social norm that they think is wrong. This is often caused by pluralistic ignorance,

or when people assume that their personal attitudes are different than their peers' even though their public actions are similar [121]. This is more often a challenge for studies interested in changing social norms than identifying their presence and impact, but it is still important to be aware that personal norms are not always in line with social norms when interviewing respondents.

Understanding the human dimension of water demand, and in particular, the impact that social norms have at the neighborhood-level, is critical for ensuring adequate water supply in the years to come. Moreover, integrating this data with engineering models will be important, given the relevance of both disciplines in the study of water resources. In this chapter, I present the results from a study aimed at integrating this data. First, I discuss the results from semi-structured interviews focused on water conservation. Then, I show results from applying a common statistical learning technique to predict water consumption at the census tract-level. Finally, I wrap up with a discussion on using the interview results to improve the modeling capabilities and explain anomalies in the predictive modeling results.

4.2 Semi-Structured Interviews

One of the common ways to identify the presence of social norms is through semi-structured interviews. Interviews are a type of qualitative research method that is often used as the first step in a research approach [122]. Interviews can be used to narrow the focus and determine the types of questions or analysis needs to be done in the next step. For example, one may conduct interviews and use those interviews to inform a survey or other research method. Interviewees should be sampled from the reference population, however, when used as the first stage in a multi-step research approach, it is not necessary to make the selection random [123]. A common approach to sampling for interviews is through snowball sampling. With this technique, the researcher selects a few key interviewees to begin with and asks them to recommend one or two additional interviewees and so on until the desired number of interviews

is reached [123]. Often, the desired number of interviews is based on the saturation point, or the point when additional interviews are not providing any new information about the topic [124].

4.2.1 Methods

In this section, I discuss the methods used during the semi-structured interview phase. First, I will discuss the research domain, then I will delve deeper into the interview methodology.

4.2.1.1 Site Description

The research domain for this study was the city of Indianapolis. Indianapolis was selected for a number of reasons. First, it is fairly close in proximity to Purdue, making it easier to conduct the semi-structured interviews. Second, Citizen's Energy, the water utility for the city, has worked closely with Purdue in the past, providing data and other resources for research. This relationship allowed me to obtain high resolution water consumption data, that would otherwise be difficult to find. This data, which is discussed later, is the basis for the second part of this study. Finally, Indianapolis is in the process of revitalizing several neighborhoods and there has been a general movement towards framing neighborhoods as communities, rather than geographic areas. Ultimately, I hypothesized that the social norms at the neighborhood level would be important due to this revitalization.

4.2.1.2 Methodology

The first part of this study was to conduct semi-structured interviews on water use with select neighborhood leaders around Indianapolis. The interviewees were sampled via snowball sampling [123], with the initial interviewees being selected from neighborhood associations. The interviews were recorded and anonymity was ensured, so

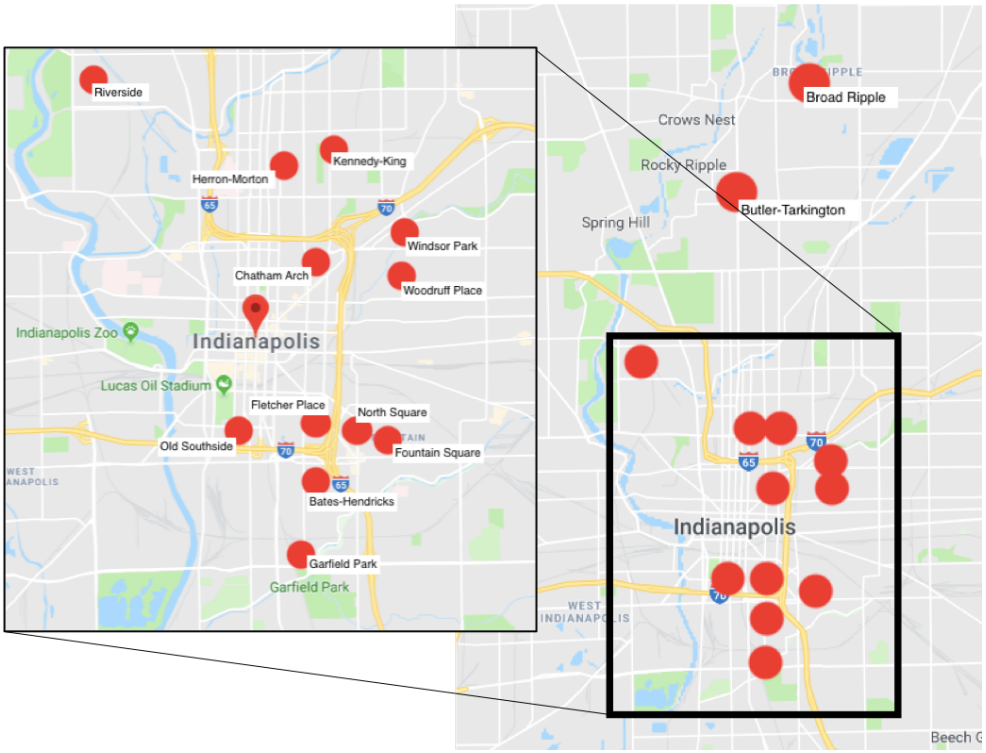

Fig. 4.1.: Location of neighborhoods in which semi-structured interviews were conducted.

as to encourage participants to provide their honest opinions on water conservation within their home and neighborhood. Ultimately, 15 interviews were conducted in 14 neighborhoods around Indianapolis (see Figure 4.1). The neighborhoods were geographically spread out and of a variety of socioeconomic and demographic backgrounds. The participants were asked a series of open-ended questions surrounding four main topics: (i) general thoughts about water conservation; (ii) awareness of local conservation initiatives; (iii) personal beliefs about conservation; and (iv) thoughts on others' beliefs about and conservation. The interview questions, which can be found in Appendix C, focused on determining the prevalent empirical and normative expectations within the community, and therefore the descriptive and injunctive norms that are important. The interviews were recorded and transcribed. Then, the transcripts were coded based on keywords that were then synthesized into common themes surrounding water conservation. These themes ranged from awareness (or

lack thereof) of water conservation issues and programs to expectations of others to conserve water.

4.2.1.3 Results

Here I present the themes determined from interview transcripts. The themes were separated into general themes that could be applied to very generally to water conservation in Indianapolis, and question-specific themes that delve into the specific answers provided throughout the interviews.

General Themes

For the most part, people seemed to be more likely to conserve electricity than water. For example, the majority of people interviewed had heard of and participated in programs sponsored by Indianapolis Power and Light (IPL), which included energy assessments, receiving LED lightbulbs, discounts on smart thermostats, and air conditioning management. People were mainly concerned with stormwater management and the tunnel Citizen's Energy is putting in as opposed to water conservation. That being said, there was a sizable amount of people with rain barrels, which they use to supplement outdoor water use. The reasoning behind getting a rain barrel varied, including runoff management, water conservation.

In general, people leaned toward environmental reasons over economic for conservation. That being said, people with older homes tended to focus on the economic impact since their homes did not have the insulation of modern houses. A few people mentioned solar panels or other environmentally-mindful options they are considering or doing in their home, but weren't necessarily related to conservation. There were also several people from 'working class' neighborhoods or neighborhoods being gentrified, in which the interviewees expressed that their neighbors are not likely to be thinking about conservation beyond limiting usage to reduce their bill.

Finally, the interviewees did not express many expectations of others to conserve. Most people said that they hoped it was something people were doing, but not something they expected. However, a few people mentioned that if there were government mandates to reduce water use, they would expect people to follow those. Moreover, some people expressed various exceptions to their original statement of not having expectations. For example, a few interviewees indicated that they expect people to be economical or to not waste things, in a more general sense. Another said that they expect people with children to adopt environmentally-mindful behaviors. Similarly, one person stated that they only expected conservation practices to be adopted by their close family members, and in particular, their children. Finally, one person indicated that their expectations were limited to people with the means to implement some of the more expensive technologies. Going beyond personal expectations, most people did not feel that others held any expectations that they (the interviewees) should conserve. Those who did believe others had expectations indicated that it was because they are thought of as environmentalists or because they are seen as a leader of a group.

Question-Specific Themes

In addition to the general themes discussed above, I also looked into the themes on a question-by-question basis. These themes, presented in Table 4.1, include both primary and secondary themes that occurred throughout the neighborhoods considered in this study. Overall, there seemed to be a lot of agreement across the study area for the different questions.

Table 4.1.: Themes for each of the 11 questions in the interview protocol (see Appendix C for the questions).

Question Number	Norm Type	Major Themes	Secondary Themes
1	—	IPL programs; nothing from Citizens	—
2	—	Rain barrels	IPL programs
3	—	Rain barrels, LEDs, and programmable or smart thermostats, efficient appliances	Christmas lights lead to higher bills
4	—	Online billing, little attention except when higher than usual	self-auditing process
5	—	Electricity conservation more popular than water conservation. Environmental context is popular	Economic context popular among those with older homes
6	—	Prices would likely not have an impact unless the increase was drastic. Incentives are good. Mandates to conserve water during a drought were acceptable.	Mandates should be somewhat selective, as water can be the only way some lower income people can stay cool in the summer

7	Descriptive	In general, others are not thinking about conservation	'Socially conscious' neighborhoods or neighborhoods with younger people are more likely to be thinking about conservation
8	Descriptive	Rain barrels, LEDs, Nest thermostats, efficient appliances	Reconnecting our waterways (ROW), Keep Indianapolis Beautiful
9	Injunctive	No expectations, but many *hope* people are doing what they can	Expectations to follow mandates/not be wasteful.
10	Injunctive	No expectations from others	Some expectations based on actions/words or leadership positions
11	Injunctive	Everyone has a positive reaction	Potential to learn from others

4.2.2 Discussion

Overall, the people I interviewed were more likely to be aware of and participate in electricity conservation than water conservation inside the home. Outdoor water use was not very common, especially since there is a popular rain barrel program throughout the city, which is put on by the public library system. Moreover, the interviewees seemed be more environmentally focused with regard to conservation than economically focused, with the exception of the people in older homes that are less efficient. However, few interviewees thought their neighbors felt or thought the same

way they did. A number of people spoke about the average socioeconomic status of their neighborhoods, identifying them as 'working-class' or 'being gentrified', as the reason they didn't think others thought the same way. They indicated that people are more concerned with being able to pay their bills than to upgrade appliances so that they are more efficient or buying a rain barrel. There were a few interviewees that spoke about their neighborhood being a community rather than a typical neighborhood. These community-focused neighborhoods may have stronger social norms than other neighborhoods, where the norms are more likely to emerge from some other identity or group. These community-focused neighborhoods were also the neighborhoods in which the interviewee expressed that they felt an expectation from others to conserve. That being said, the interviewees were all involved in the neighborhood association, so it is likely that they are all fairly affluent. In particular, they all have the time and resources to take on an unpaid leadership role. In this sense, there could be some bias introduced to the study. However, it is likely that residents in the lower socioeconomic brackets within these neighborhoods are still conserving, if only to reduce their bills.

4.3 Modeling Water Consumption

Based on the interview results, people in different areas across the city tend to think about water conservation differently. People from the older neighborhoods tended to think about conservation from an economic standpoint, since their houses were not as efficient as the newer homes. Additionally, respondents from the more suburban neighborhoods, where yards are commonplace, discussed the need to maintain their landscaping, but also expressed interest in rain barrels to supplement the treated water that they would use outside. This brings up an interesting question—does the water consumption actually vary across the city? If so, do social norms play a role, or is it just differences in demographics and housing characteristics? In this section, I will present results from a study that focuses on census tract-level water

consumption. Using monthly data collected during 2018, I will predict the intra-city water consumption based on a series of demographic variables. Ultimately, I will compare the results of the interviews to the predictive accuracy of the computational model.

4.3.1 Data & Methods

In this section, I will first discuss the data collected for the census tract-level study. Then, I will present the methodology used to make predictions of intra-city water consumption.

4.3.1.1 Data Description

For this part of the study, the water consumption data for the city of Indianapolis was collected at the census tract-level for each month in 2018. The data was obtained from Citizen's Energy, the water utility for the city of Indianapolis. Additionally, predictor data was collected from the US Census Bureau. Since the water consumption data was aggregated by census tract, the demographic data collected from the Census Bureau was directly aligned with the response data (water consumption). Table 4.2 shows the predictor variables considered throughout the course of this study. Initially, climate data was also included in the analysis, however, due to the lack of granularity, it was not suitable to evaluating intra-city differences. Therefore, after a brief analysis, which proved that the city-level climate data had no impact on the accuracy of the census tract-level predictions, the data was not considered in the remainder of the study.

4.3.1.2 Methodology

In this study, predictive modeling was used to evaluate the intra-city water consumption. In particular, I leveraged the random forest algorithm [53]. This model

Table 4.2.: Predictor variables considered in this study, separated into demographics and climate categories.

Variable Name	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 Type	Percent of population that is part of various types of households (e.g., families, married couple, single parent, 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.)

is both flexible and interpretable, making it ideal for the analysis of large datasets. The model was developed in 3 steps. First, I trained and tested the model considering a variety of demographic data. Then, I selected the important variables and reran the model. This step focused on improving predictive accuracy and limiting the complexity of the model. Finally, following the analysis on demographics, I assessed the impact of the social norms. This was done through a deeper analysis in the census tracts that correspond to the various neighborhoods from which I interviewed residents. By looking at any anomalies in the predicted data compared to the actual data, I was able to determine if there was a behavioral aspect (i.e., a social norm) to water consumption. In other words, if considering demographics leads to under

or overprediction in a given area, then it is likely that there is a social norm that is leading to such behavior.

4.3.2 Results & Discussion

In this section, I will discuss the results of the study. First, I will focus on the demographics-only model. I will evaluate the ability of the model to predict water consumption based solely on demographics and housing characteristics. Additionally, I will determine the important variables and the relationship of those variables with water consumption. Following a discussion on the final model development and final variable selection process, I wrap up with a discussion on any anomalies found in the data and the impact that social norms may be having on water consumption.

4.3.2.1 Impacts of Demographics on Intra-City Water Consumption

When considering the intra-city water consumption, it is likely that demographics and housing characteristics play a significant role. Here I present a model that evaluates the ability of these variables to predicting water consumption at the census tract-level.

In total, 72 demographic variables were considered in the initial stages of this study. Of those variables, the initial analysis showed that several were important across all twelve months. The top three variables, in terms of importance, are shown in Table 4.3. The most important variable for predicting water consumption across all the months, was home ownership. This indicates that if the percentage of home owners in a tract was removed from the analysis, there would be a significant reduction in predictive accuracy. This is intuitive, since home owners tend to have larger residences with yards, as opposed to renters, which are more likely to live in apartments or multi-family homes. This is especially true in Indianapolis, which is more sprawling than similar sized cities (e.g., Columbus, OH has a similar population but is 150 mi^2 smaller in size). This suggests that Indianapolis residents are more likely

Table 4.3.: Top three most important variables in each month of the initial demographics model. Importance was determined based on the percentage of increase in predictive error caused by removing the particular variable [53].

Month	Variable 1	Variable 2	Variable 3
January	owned house	detached house	family household
February	owned house	detached house	family household
March	owned house	family household	detached house
April	owned house	family household	detached house
May	owned house	family household	detached house
June	owned house	family household	couple household
July	owned house	couple household	married status
August	owned house	family household	detached home
September	owned house	family household	detached home
October	owned house	detached home	family household
November	owned house	family household	detached house
December	owned house	family household	detached house

to live in more sprawling neighborhoods that rely heavily on water consumption for outdoor landscaping and recreation. It is interesting that home ownership is the most important in the winter months as well, as the main usage during winter is standard indoor activities. However, it is possible that home owners have larger houses than renters, which would ultimately lead to increased water use throughout the home.

The second and third most important variables, in most months, are the type of house and the type of household. In particular, having a detached house and a family are important predictors of water consumption. This would be opposed, for example, to living in an attached house or apartment complex and living as a couple without kids or a single person. Again, these results are somewhat intuitive, since detached houses tend to have larger yards that require upkeep. Additionally, more people living in a house would lead to more water consumed. Interestingly, in June and July, a household made up of a couple without kids is an important predictor of water consumption. Overall, the variable analysis suggests that housing characteristics, as well as the household type, are important predictors of water consumption.

In addition to evaluating the important variables, it is essential to assess the model performance. In Table 4.4, the results from the model performance are shown for each month. The R^2 represents the goodness-of-fit, or the variance explained by the model. The normalized root mean squared error (NRMSE) is a measure of prediction error. Overall, the results indicate that using demographic variables alone accounted for only 30% of the variance in the data (see R^2 values in Table 4.4). This is not indicative of a poor model, however, since in Section 3.2, it was shown that climate variables account for over 60% of the variance in water consumption. Therefore, it is to be expected that the demographics-only model would only account for the remaining 40%. Moreover, it is likely that social norms, as well as unseen factors, also play a role in explaining some of variance in the data, however, those factors were not included in this study. Unfortunately, due to the lack of high resolution climate data, it is hard to capture the intra-city effects of climate on water consumption. That being said, the NRMSE is close to zero, indicating that the predictive error is low. As such, it may be possible to use the demographic-only model to make predictions about future water consumption. Moreover, it is likely that the current model is

Table 4.4.: Model performance for the initial demographics model. Measures of model performance include R^2 (goodness-of-fit) and normalized RMSE (measure of error).

Month	R^2	Normalized RMSE
January	0.29	0.114
February	0.29	0.114
March	0.28	0.112
April	0.28	0.115
May	0.30	0.102
June	0.33	0.110
July	0.33	0.104
August	0.32	0.115
September	0.32	0.105
October	0.32	0.117
November	0.29	0.114
December	0.29	0.109

too complex, which could be increasing the variance in the modeled data, effectively increasing the predictive error. It is important, therefore, to reduce the number of predictor variables considered in the final model.

4.3.2.2 Final Model Development & Results

In order to develop a final model, I ran through a variable selection process, described below. I also evaluated the updated model performance and considered the differences between the predicted and actual data. The results from this final model are discussed below.

Variable Selection

An advantage of using predictive modeling is that, for the most part, correlation between prerictos variables does not affect the predictive accuracy of a model [65]. However, large datasets of predictor variables can increase the complexity and make interpretation difficult. Moreover, some predictor variables that are highly correlated could mask the effect of other non-correlated variables. For example, if household income and house value are highly correlated, both could end up as top predictors, effectively overshadowing a less correlated variable that might also be important. Correlation plots of the predictor variables before and after variable selection can be found in Appendix C.

Given the large dimensionality of the predictor dataset, it is beneficial to try to reduce the number of predictors, either through variable selection, which will be discussed later, or through algorithms such as principal component analysis (PCA). Using PCA to evaluate the predictor dataset, it was shown that 17 components (out of 72) were needed to explain 95% of the variance in the data. If only 90% of the variance needs to be explained, just 10 components are required. In this sense, there is an opportunity to reduce the number of predictor variables considered in the study without sacrificing any accuracy. A biplot showing the relationship between the

various predictors and the top 2 principal components can be found in Appendix C. Going forward, however, I opted to use a variable selection process, since PCA involves data transformations that are not as interpretable. Understanding the relationships between real variables and water consumption is critical for practitioners that are interested in managing the residential demand.

In this study, the variable selection process was based on a threshold analysis. Using the 90^{th} quantile in percent of increased mean squared error as the threshold, variables were kept if their score was above the threshold and removed if it wasn't. This reduced the variable count from 72 to less than 10, which allows for a more manageable analysis, as well as significantly reducing the complexity. These variables are shown in Figures 4.2 and 4.3.

For the most part, the final variables resemble the previously important variables, although there are some differences. For example, in some months, the percentage of family households became more important than the percentage of home owners. In most of these instances, however, the percentage of home owners remained in the top 2 most important variables. Similarly, the percentage of detached homes was still considered to be an important variable, although it dropped rank in some months.

There are some interesting changes, however, that represent variables that were previously lower on the list, but may have gained importance after other correlated variables were removed through the variable selection process. For example, in February, the percentage of people that take public transit to work became the second most important variable to predicting water consumption. This could be indicative of two different phenomena: an association between lower income and reduced water use, or an association of between environmentally-mindful ideals and reduced water use. In Indianapolis, the main form of public transit is the bus system, which is primarily used by lower income residents that may not have access to a car. In fact, during the semi-structured interviews, many of the most environmentally-mindful interviewees mentioned that they would like to ride the bus more frequently, but that it was inconvenient. With this in mind, it is likely that the association between public transit

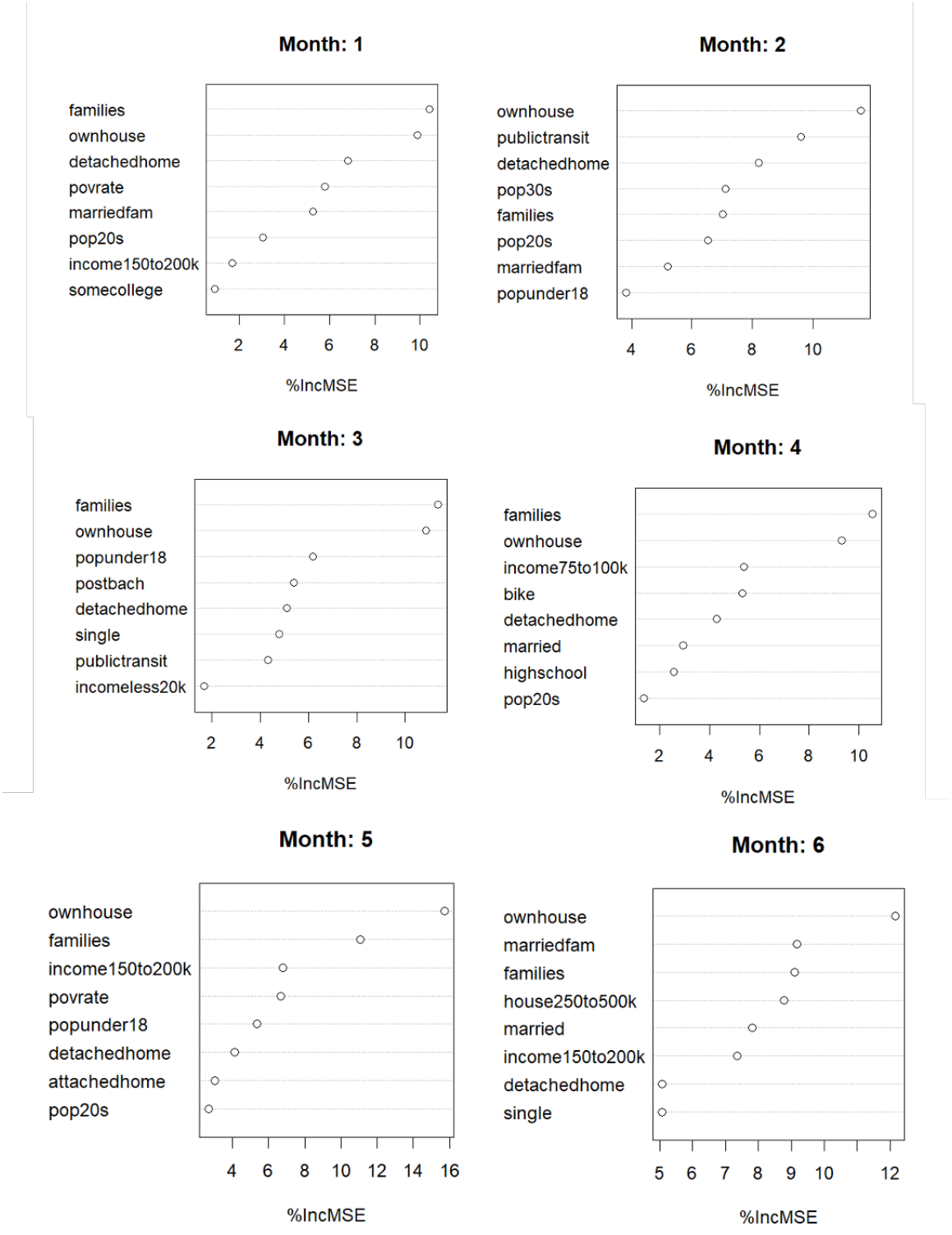

Fig. 4.2.: Important variables in the final model for January through June.

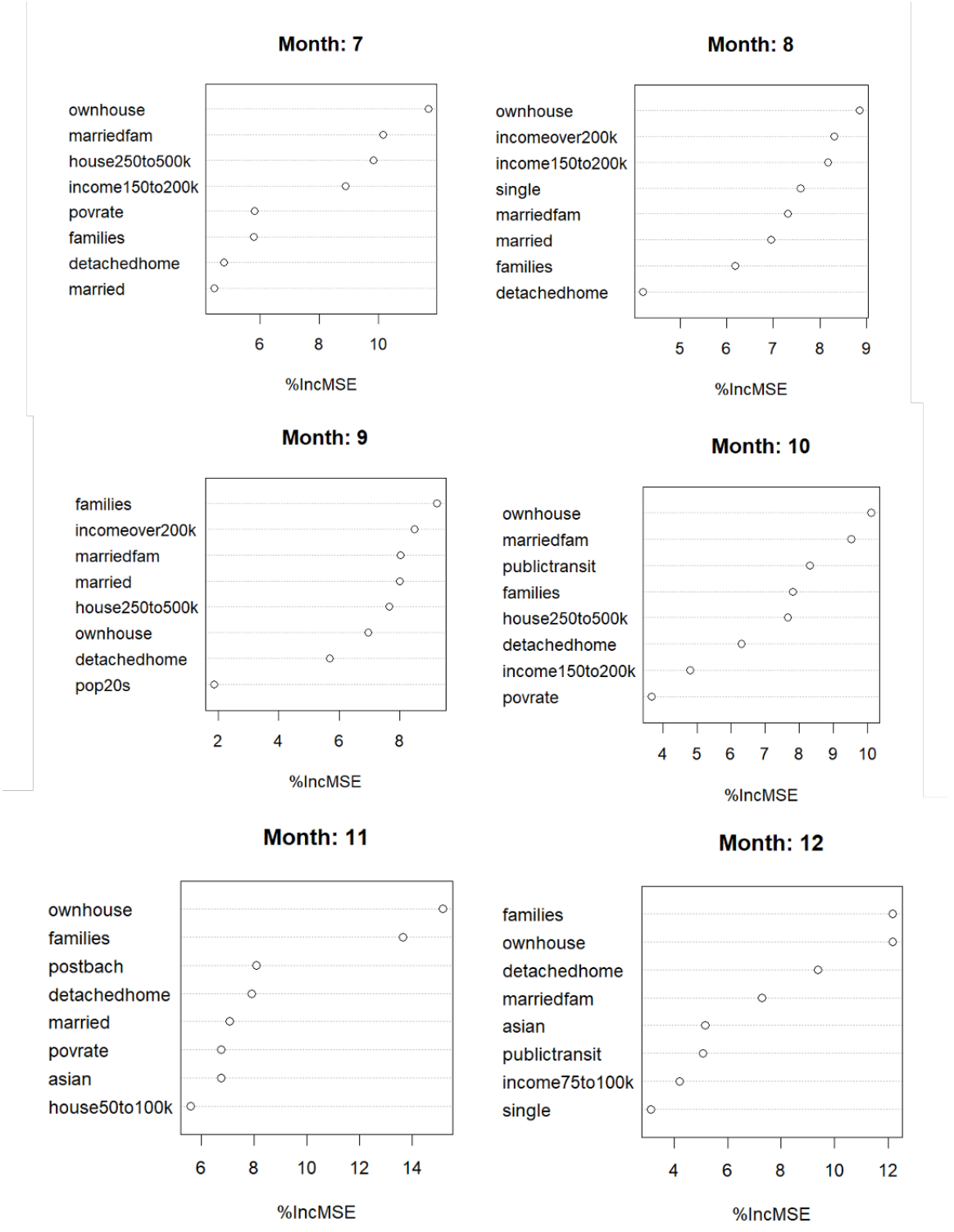

Fig. 4.3.: Important variables in the final model for July through December.

use and water consumption is linked to household income and potentially size, both of which play a role in water demand. Similar results can be found in several other months, such as April, May, July, August, and September, where household income and the poverty rate have become increasingly important predictors. Overall, the variables that are related to housing characteristics, income, or household size tend to play important roles in predicting water consumption within Indianapolis.

Using partial dependence plots, it is possible to assess the nature of these relationships and determine further evidence to some of the hypotheses discussed above. In partial dependence plots, each of the predictor variables, except for the one of interest, are held constant to assess the impact of a single variable. For more information on partial dependence, see Section 2.2.2.4. In Figures 4.4 and 4.5, a selection of partial dependence plots are shown. The variables were chosen based on their prevalence and relative importance compared to other variables. In particular, Figure 4.4 shows the variables that were important in the majority of months, while Figure 4.5 shows the results considering a subset of variables that were important in a few months, but not the majority.

In Figure 4.4, it is shown that as the percentage of home owners, detached houses, and families increases, the water consumption also increases for each month. Since home ownership and detached houses tend to come with larger yards and increased landscaping needs, it is logical that water consumption would increase in the summer months. As for the winter months when water consumption is limited to indoor uses, it is likely that that owned houses and detached houses are larger than apartments, which would ultimately lead to higher indoor water use. It is also possible that home owners are more likely to leave faucets dripping during the colder winter periods, since burst pipes would be their responsibility, rather than a rental agency. The trend towards increased water consumption with increased percentage of families is also to be expected, since increasing the number of people in a household or area will inevitably lead to higher water use. Interestingly, household income also tends to lead to higher water consumption, but more so in summer months. In particular,

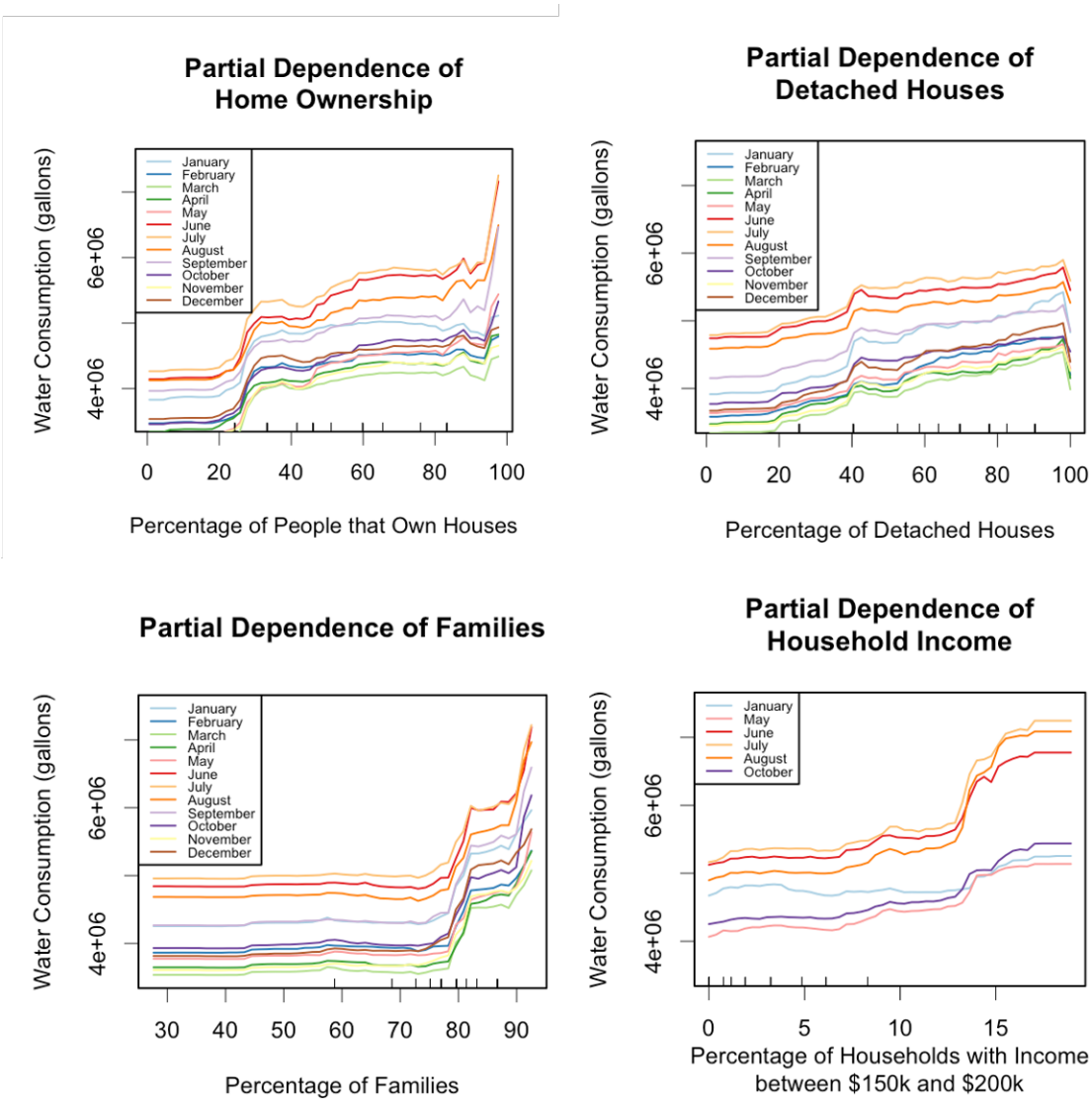

Fig. 4.4.: Partial dependence plots for home ownership, detached houses, families, and household income.

Figure 4.4 shows that there is a threshold between 10 and 15% of households with an income between $150,000 and $200,000, after which there is a sharp increase in summer water consumption, but no similar jump in winter water consumption. This could be tied to higher income residents having larger houses and larger lot sizes, which would require increased water for landscaping. But it could also be indicative of the status quo surrounding the desire to have manicured lawns, particularly in

more affluent neighborhoods. In other words, there may be pressure to maintain a certain style of lawn based on one's socioeconomic class. If this is true, members of that group may not give into the pressure if most of their neighbors are not within their perceived social group. However, after a critical number of neighbors are part of that group, a person might give into that pressure to maintain their status. In Figure 4.4, this critical number appears to be around 13% of households in a given census tract.

Looking beyond the variables that were important in the majority of months, Figure 4.5 shows the partial dependence on public transit and house value, which were only important in a few months. In particular, the plot of public transit use indicates that to a point, increased reliance on public transportation to get to work leads to less water consumption. As discussed above, this could be related to lower household income or environmentally-mindful practices. However, based on the interviewees' feelings on the Indianapolis bus system, as well as the partial dependence plot of household income, the most likely reason is the association between using public transit and lower household income. In other words, if use of public transportation

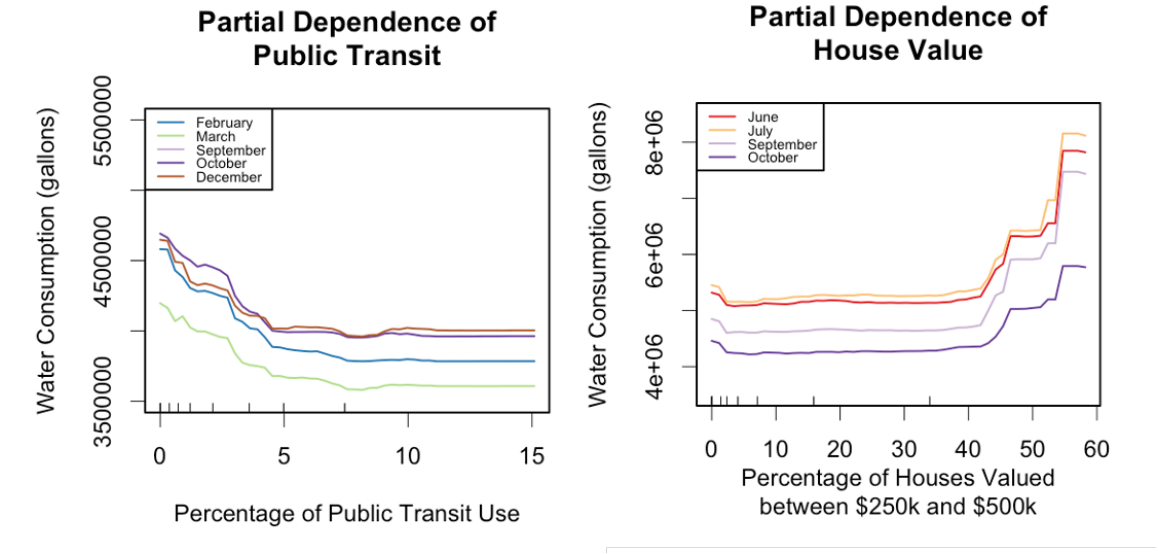

Fig. 4.5.: Partial dependence plots for use of public transportation and house value.

is linked to household income, then as household income drops, so does water consumption. This could be due to smaller houses and lot sizes or due to conservation for economic reasons. Either way, it is apparent that household income plays a major role in water consumption. Finally, looking at house value, the partial dependence plots demonstrate that there is little change until more than 40% of the people living in a census tract have houses valued between $250,000 and $500,000. Similar to the household income results discussed above, the higher valued houses are likely to be larger and have more green space that requires additional landscaping. Ultimately, this would lead to higher water consumption, especially in the warmer months, which is when house value becomes important.

Model Performance

In addition to looking at the important variables and partial dependence plots, I assessed the model performance of the final model. The results of this assessment can be found in Table 4.5. Following the variable selection process, the model accuracy improved. In fact, based on the R^2 values, the final model was able to capture 30-

Table 4.5.: Model performance for the final model. Measures of model performance include R^2 (goodness-of-fit) and normalized RMSE (measure of error).

Month	R^2	Normalized RMSE
January	0.35	0.111
February	0.35	0.106
March	0.31	0.110
April	0.33	0.110
May	0.32	0.096
June	0.38	0.104
July	0.39	0.098
August	0.40	0.108
September	0.36	0.102
October	0.40	0.112
November	0.33	0.113
December	0.32	0.107

40% of the variance in the data, which is an improvement over the initial analysis. Moreover, this suggests that a model with demogrphics and climate variables would explain upwards of 90% of the variance, effectively capturing any trends in the data and allowing for a more accurate representation of the system. In terms of NRMSE, the values were further reduced in the final model, indicating the predictive power of using the reduced variable subset. Using this final model, the summer months (May-September) were predicted most accurately, which is the critical time for the water utility. In this sense, improving the predictive accuracy during these months is crucial to the planning processes and ensuring the utility can provide enough supply to match the demand.

One of the key aspects of this model is the focus on intra-city differences in water consumption. In this sense, it is possible to assess the ability of the model to accurately predict the water consumption in various census tracts, potentially signaling problematic areas within the model. Additionally, evaluating intra-city differences can help the utility determine areas in which they ought to focus on in terms of demand management.

Model Results

Considering the differences between the predicted and actual values is important for evaluating the intra-city accuracy of the model. For example, it is likely that some census tracts have more accurate predictions than others. In Figures 4.6 and 4.7, the anomalies are plotted over the entire study area for each month (see Appendix C for maps depicting the actual and predicted water consumption). In these figures, census tracts are filled with red to represent underpredictions and blue to represent overpredictions, with varying shades indicating the severity of the inaccurate prediction. Overall, there seems to be many similarities between the months. For example, the model is regularly underpredicting the water consumption of the census tracts in the bottom left corner of the county. In the upper right corner, on the other hand, there are a few census tracts for which the model overpredicts

the water consumption. Additionally, the upper left and bottom right corners stand out has having more extreme over or underprediction issues than most of the central tracts. These areas of larger prediction errors are all located in some of the more rural areas of the county. It is likely that there are larger lot sizes and potentially agricultural activity that would require increased water consumption than the central census tracts with the similar demographics. Additionally, since these census tracts tend to be larger, indicating a less dense population, there might be a wider variety of housing values and incomes than the central tracts, which are more likely to have homogeneous populations. This homogeneity could be leading to higher predictive accuracy in some of the more central tracts. This is further confirmed by the fact that many of these large census tracts fluctuate between over and underpredicting, while the central tracts remain fairly consistent throughout the year. This is indicative of both varying demographics that are having more or less influence throughout the year, as well as potential shifts in consumption patterns due to agriculture. These agricultural consumption patterns would be different than standard suburban patterns, since agriculture would require significantly larger amounts of water and also serve a different purpose. In other words, a homeowner is likely watering their lawn in an effort to keep it green, but a farmer is going to be watering crops. Moreover, a farmer wouldn't necessary irrigate their crops everyday, but rather on certain days based on the recent precipitation patterns.

It is interesting to note that the city center remains consistent throughout the year, albeit with a slight overprediction. This is likely due to the built-up nature of the urban environment. Since there are less lawns and landscaping needs in the center of the city than in the suburbs, indoor uses are the primary driver of consumption. In these cases, demographics, particularly household income and household size, might be more influential than climate variables or lot size, for example. It is notable that the model is overpredicting the water consumption for the majority of the city, indicating that there are additional factors that limit the water consumption beyond what the demographics would suggest. These factors could be related to the climate,

but they could also be related to the personal norms of the residents. These norms could cause people to limit water consumption, especially within neighborhoods.

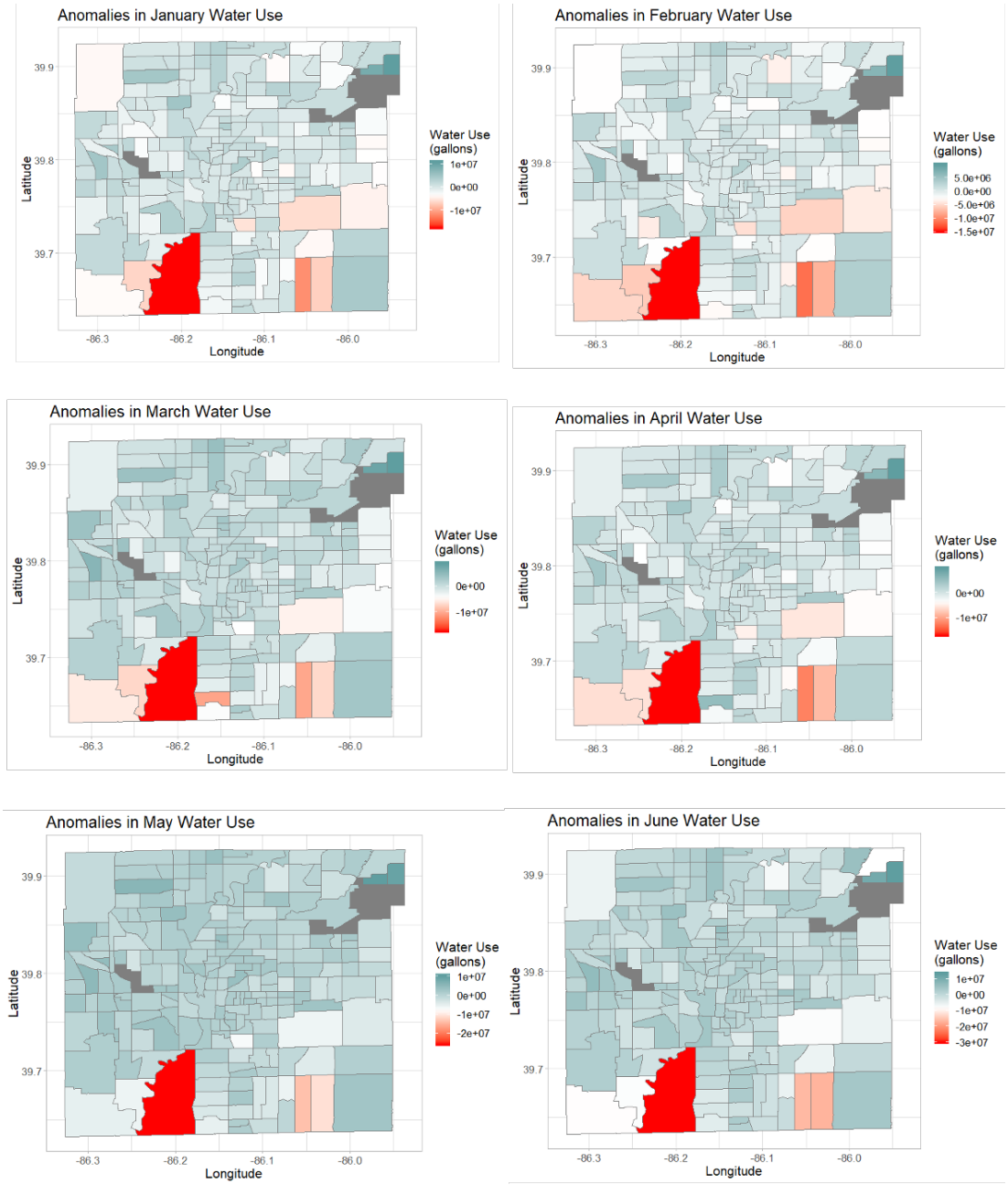

Fig. 4.6.: Anomalies in the predicted water consumption for January through June. Shades of red represent underpredictions, while blue shades represent overpredictions. The grey areas represent tracts without any water consumption data.

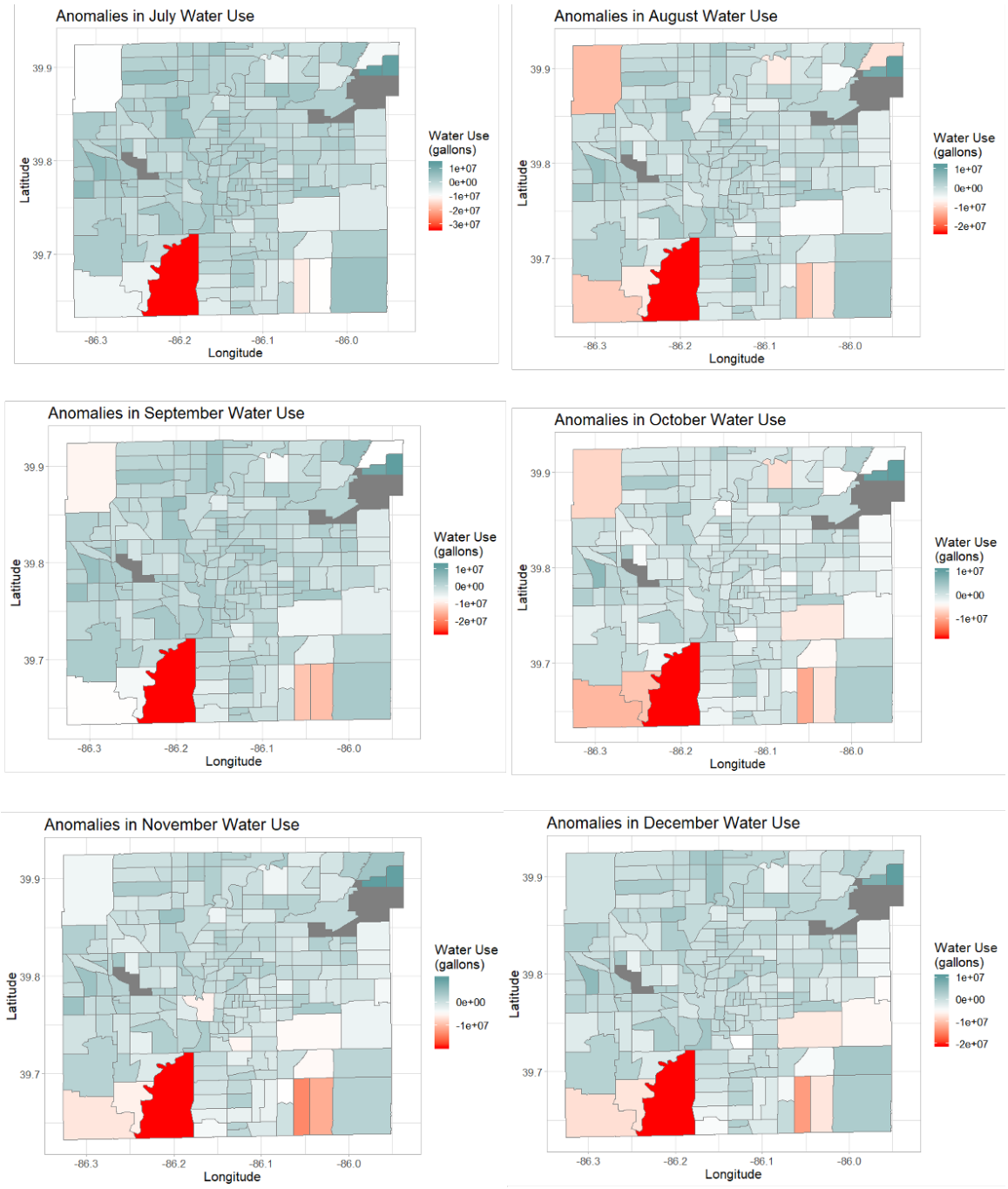

Fig. 4.7.: Anomalies in the predicted water consumption for July through December. Shades of red represent underpredictions, while blue shades represent overpredictions. The grey areas represent tracts without any water consumption data.

Given that most of the census tracts are slightly overpredicted, with a few extreme outliers, the analysis might be further improved by removing these census tracts from the dataset. Removing these poorly predicted tracts would likely reduce the variance in the data and lead to improvements in the predictive accuracy across the city. Reducing the predictive error in some of the more populous census tracts would lead to significant benefits from the utility perspective, which would likely outweigh the costs of not having predictions in some of the less populated tracts. This analysis, however, was beyond the scope of this chapter. The extreme variation in the water consumption values, however, should be kept in mind during future analyses.

4.3.2.3 Comparison with Interview Results

As cities continue to focus on neighborhood revitalization, it is likely that the social norms of one's neighborhood will become increasingly important. It is possible that this growth of social norms will impact water conservation measures within certain neighborhoods. Moreover, these norms likely account for any variance not explained by demographics or climate variables within the water consumption data. Using the results from the interviews, it is possible to make some inferences about the nature of the norms, as well as their impact on the modeling result.

For example, the neighborhoods of Butler-Tarkington and Broad Ripple are located north of the city center. This area is more urban than the suburbs, while having more green space and larger lot sizes than the city center. Notably, interviewees in these neighborhoods expressed more environmentally-mindful views, including a focus on conservation, than most of the more centrally located interviewees. Furthermore, these interviewees felt that most of their neighbors felt the same, considering the prevalence of rain barrels and participation in utility-run conservation programs. This may explain the overprediction shown in Figure 4.8—if the residents in this area regularly use rain barrels for landscaping and actively try to reduce their consumption, their demographics model will predict that they are using more water

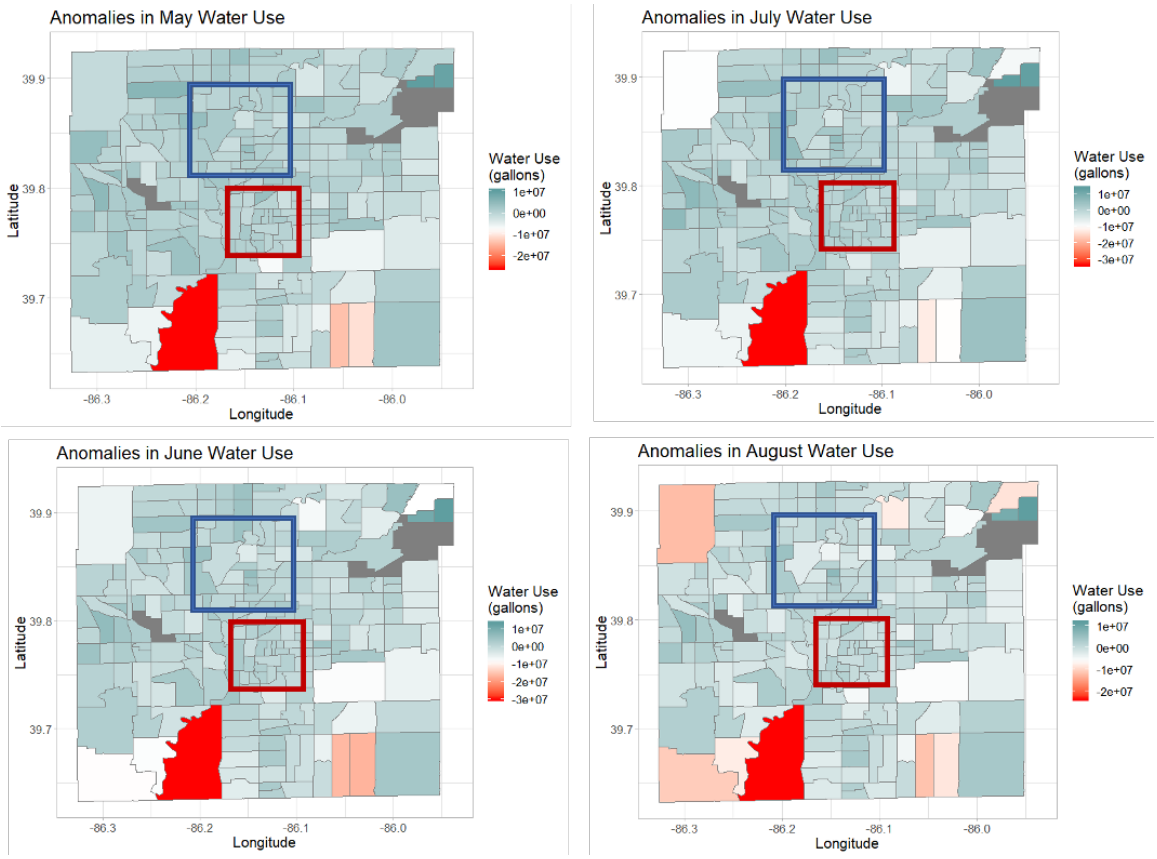

Fig. 4.8.: Summer anomalies in water consumption, with the census tracts associated with Butler-Tarkington and Broad Ripple highlighted in blue and the remaining central neighborhoods in red.

than in reality. It is interesting to note that Butler-Tarkington and Broad Ripple are among the more affluent neighborhoods in the area, which would suggest higher water consumption, based on Figures 4.4 and 4.5. Looking at the census tracts that fall within these neighborhoods, an average of 9% of households have an income between $150,000 and $200,000, which is shown on Figure 4.9. There is, however, a wide range of the household income levels between the different census tracts. For example, there are four census tracts within Broad Ripple, which range from 1-20% of households with income between $150,000 and $200,000. Likewise, Butler-Tarkington includes six census tracts with 3-20% of households with income between $150,000 and $200,000. The high end of this range is the minority, though, with the majority

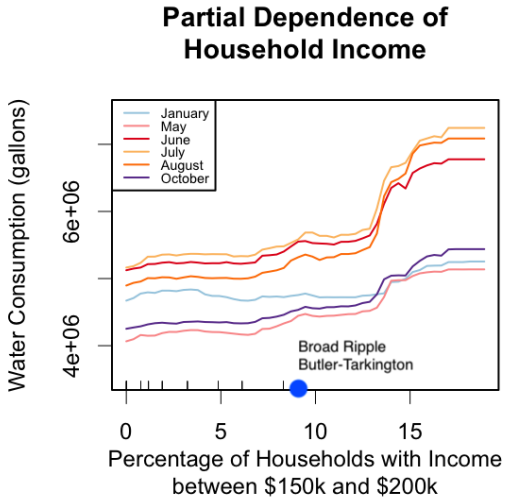

Fig. 4.9.: Partial dependence on household income, with the average percentage of households with income between $150,000 and $200,000 in Broad Ripple and Butler-Tarkington neighborhoods marked with a blue dot.

of both neighborhoods having less than 12% of the households with income between $150,000 and $200,000. These values are below the critical number discussed above, which may also explain the reduced water consumption in the summer time within these neighborhoods.

There is a similar pattern in the central area of the city, although the magnitude of overprediction is slightly less than that in the neighborhoods further out. The interviewees in these central neighborhoods still indicated an awareness of conservation efforts, and many spoke about the efforts that they do themselves, including the use of rain barrels where applicable. Different from the northern neighborhoods though, interviewees in the central neighborhoods did not think that other people had similar thoughts, suggesting that conservation norms are not as strong in the central locations. This could perhaps be due to nature of downtown neighborhoods and the lack of the community focus that was mentioned by the interviewees from the northern neighborhoods. That being said, the model still overpredicts the water consumption,

which might indicate a prevalent personal norm towards conservation, but perhaps not one that originated at the neighborhood level.

Overall, it is likely that social norms are playing a role in people's decisions on how and when they consume water. Recognizing these social norms and their impact on water consumption could help utilities improve their demand predictions, as well as tailor interventions to meet the needs of various neighborhoods. For example, neighborhoods with larger lots and more landscaping needs might benefit more from a rain barrel initiative than an efficient appliance initiative. Furthermore, in a neighborhood that is more close-knit, which is likely to have more influential social norms, an initiative that provides external proof of participation would likely have better success than internal measures. In other words, if people can see that their neighbors are participating in a given program, such as using rain barrels, then it might encourage them to participate as well. Understanding the social norms at play and how they impact water consumption is therefore important for utilities and policymakers interested in demand management.

4.4 Conclusion

The goal of this chapter was to evaluate the impact of social norms and demographics on water consumption. Previous work has focused on city-level water consumption, however, there is a need to assess intra-city changes. In this study, I conducted semi-structured interviews with residents of Indianapolis and used the results to determine the influence of social norms on water conservation. The results indicated that most interviewees were cognizant of their water consumption and actively tried to reduce it, mainly through limiting outdoor use and supplementing with water collected via rain barrels. However, most people, with the exception of those from a few neighborhoods, felt that other people in their neighborhood did not feel or act the same way with regard to water conservation. This indicated that perhaps there is a widespread personal norm towards environmentally-mindful practices, but

within the neighborhoods, there is not a shared norm around conservation. This was especially true in the central neighborhoods, where there is less green space and therefore less opportunities for external water conservation measures, such as rain barrels.

Using the results from the interviews as guidance, I compared the results from a computational, demographics-based model of water consumption with the prevalence of social norms. Overall, the computational model overpredicted the water use in most of the census tracts. However, that overprediction was more extreme in the neighborhoods of Butler-Tarkington and Broad Ripple, which are among the more suburban neighborhoods, as well as represent neighborhoods from which interviewees indicated a presence of social norms, especially with regard to outdoor water use. The fact that the model overpredicted the water consumption suggests that there is an outside factor that is limiting people's water use compared to other neighborhoods of similar demographics. It is likely that these outside factors are related to the prevalence of rain barrels that are used in lieu of outdoor water consumption. The upsurge of rain barrels may be linked to the norms within the neighborhood, which would encourage people to fall in line with their neighbors and start using rain barrels more frequently. Ultimately, using these results, utilities could work to tailor intervention methods to specific areas of the city, where norms may be more or less important. For example, in the neighborhoods mentioned above and neighborhoods like them, outdoor interventions, such as the rain barrels, would allow people to see who is participating in the program and the opt in to the program themselves. In other areas closer to the city center, or neighborhoods with less prevalent social norms, it might be more effective to focus on indoor interventions, such as efficient appliances. In general, however, utilities and policymakers interested in demand management should take both demographics and social norms into account to not only make predictions, but also plan interventions.

5. CONCLUSIONS AND RECOMMENDATIONS

People, water, and climate are highly interconnected. As urban areas continue to grow in population, they will require more water to supply the people of the city. This need will be further increased by intensifying climate change, which will not only change the demand profile for water, but also change the supply. Given that water is necessary for human life, it is imperative that we understand this nexus of people, water, and climate and work to improve the resilience of water resources. In this dissertation, I sought to further this understanding of the climate impacts on water resources, as well as develop practical tools that can be used to evaluate the state of water resources, now and in the future.

Throughout my dissertation, I had several hypotheses within the various research projects. For example, in Chapter 2, I initially hypothesized that there would be uniform behavior across the reservoirs. In other words, I expected that the water balance model would either over or underpredict the actual volume in all the reservoirs. Instead, I found that it varied from reservoir to reservoir, possibly due to the geographical location or the purpose of the reservoir. In Chapter 3, I hypothesized that including the interdependence between water and electricity demand in the model would improve the accuracy. This was proven to be true, at least in the study region considered in the analysis. Additionally, in Chapter 3, I hypothesized that including a wider array of climate variables would lead to further improvements to the model when compared to a common baseline model. This hypothesis was also shown to be true in the study region. Finally, in Chapter 4, I expected the demographic variables to play a significant role in predicting water consumption. However, the results indicated that only 30-40% of the variance in the data could be explained by the demographics. It is likely that the remainder can be explained by the climate vari-

ables, as well as the social and personal norms present in the city. These hypotheses, successful and rejected, were critical to the research process.

Within the rest this chapter, I will first discuss the conclusions of this dissertation, as well as the implications to society. Then, I delve into my recommendations, including future work and study limitations. Finally, I wrap up with a discussion on the applicability of the work presented in this dissertation beyond urban water systems.

5.1 Conclusions

In this dissertation, the impact of climate change on urban water resources was explored. I first started with a focus on water supply, using both statistical learning theory and a more traditional input/output model to evaluate the change in volume within urban reservoirs. This results indicated that different types of reservoirs, as well as reservoirs in different climate zones, called for different techniques to be used to evaluate volume. In fact, the highly managed Lake Mead was best represented by the water balance model, while the random forest model was best for the reservoirs in the Pacific Northwest. Using these models to make projections into future conditions, the random forest method performed better, with the exception of Lake Mead. This was expected though, since the random forest model is a *predictive* model, rather than *explanatory*. That being said, there are pros and cons to using both the water balance and random forest model. For example, although the random forest model is best for making projections, the amount of data required could make it infeasible for cities to run such a model. On the other hand, the water balance model can be used with minimal data, but using it to make projections could be misrepresentative of the actual system. Overall, it is important to test both kinds of models, when possible, and use the best one for a given situation. In fact, in some cases, using a combination of both models could be beneficial for understanding the changes to urban reservoirs.

Following the study on water supply, I presented results from a series of studies on water demand. Considering the water-electricity nexus, I explored the impact of climate change on this nexus, as well as the benefits of considering system interdependencies. Through these studies, I first demonstrated that using a multi-outcome model to predict the climate-sensitive portion of the water-electricity demand nexus provided significant improvements to the predictive accuracy. Often, utilities operate in isolation (i.e., the water utility does not consult with the electric utility to make operational decisions). These results suggest that this practice could be leading to decisions based on less accurate predictions, potentially creating issues of supply inadequacy down the line. Moving forward with the multi-outcome model, I assessed the important climate variables. In particular, I compared a baseline model that only considered precipitation and temperature, a common occurrence in both practice and research, with a model that included a wider array of climate variables. This latter model considered relative humidity and wind speed, which are important for understanding *experienced* temperature and its impact on the demand structure. The results indicated that the model was improved by the inclusion of relative humidity and wind speed, especially when trying to predict the peak load. The peak load is of special importance because it represents the maximum amount of water or electricity that a utility will have to supply for any given point in time. Utilities that are relying on precipitation and temperature to make such predictions are likely to significantly underestimate this peak load, potentially creating situations where the supply won't be able to meet the demand. Finally, using the multi-outcome model with a wider array of climate variables, I projected the coupled water and electricity demand in to the future. The results showed that the Midwest region could expect to see significant increases in summer water and electricity use under different climate change scenarios. This is likely due to the increased temperatures and more variable precipitation, which will lead to increased electricity use (via the increased need for air conditioning) and water use (via increased landscaping and recreational needs). The winter demand shifts were more uncertain, but the model projected a median

decrease in electricity use and a median increase in water use over the winter season. These shifts may be caused by the more moderate winters, which would reduce the need for electric space heating, as well as provide some potential for landscaping on the tail ends of the winter season. Ultimately, these changes will require utilities, city planners, and policymakers to rethink their plans for ensuring adequate supply under climate change, as most policies do not consider the climate impacts to the coupled water-electricity nexus.

Finally, I delved deeper into the intra-city differences in water consumption and evaluated the potential impacts of social norms on water conservation. The results indicated that while most individuals were aware of conservation programs, as well as the steps they could take to reduce their water consumption, most thought that their neighbors did not think like they did. This suggests a larger cultural norm towards conservation, but less impact from neighborhood-specific norms. The exception were the neighborhoods of Butler-Tarkington and Broad Ripple, in which interviewees expressed a social aspect to conservation practices, and in particular the growth in use of rain barrels to supplement landscaping needs. This suggests that these neighborhoods are more closely linked, thus have more influential social norms. Additionally, the use of outdoor intervention measures seemed to improve the salience of the norm, since people were confronted with what their neighbors were doing to conserve water on a regular basis. Comparing the results from the social norm analysis with a computational prediction of water consumption based on demographics revealed that in the areas where social norms were prevalent, the model greatly overpredicted the water consumption. In other words, based on the demographics, the neighborhoods should have been consuming a lot more water, but something prevented them from doing so in reality. This something was likely a social norm that encouraged conservation, especially in the summer months. Using this knowledge, utilities and policymakers will be able to tailor innterventions to specific neighborhoods, as well as take into account a larger number of variables in their demand projections. Ultimately, this

could lead to better demand management practices that work to create long-lasting ideals on water conservation.

In conclusion, this dissertation sought to explore the impacts of climate change on urban water supply and demand through an interdisciplinary lens. Starting with the water supply, I predicted reservoir volume using two commonly used methods. This work demonstrated that different reservoir purposes and locations lead to different methods performing well. This indicates that by using one method without testing others, researchers and practitioners may be making decisions based on poor predictions. Then, focusing on the water demand, I first demonstrated the impact of climate change, before shifting focus to the human dimension. Ultimately, this work showed that there is a need to consider both climatic and non-climatic forcings on water consumption, as both play a role. Moreover, delving into the intra-city differences, social norms were shown to influence water consumption, causing the demographics-only model to overpredict the water use in certain parts of the city. Overall, this work aimed to combine data science, climatology, and social science to better understand the impact that climate change will have on urban water systems. These systems are critical to the future of cities and this work will certainly aid in the improvement of these systems, and ultimately, the building of resilience to climate change and related disasters.

5.2 Recommendations

Going forward, there remain a few gaps that still need to be filled, especially with regard to the limitations of this study. I recommend that future work seek to rectify these limitations, which are discussed below.

5.2.1 Study Limitations

One limitation of this study was the relatively small study regions considered throughout. For example, in Chapter 2, I discuss the results from nine reservoirs

around the United States. However, in order to test some of the hypotheses generated within the study, such as the role of climate zone or reservoir purpose in determining the optimal model, more reservoirs will need to be tested. Considering additional reservoirs will be challenged by the availability of data. Since both methods rely on the input and output data, there needs to be gauges or other modes of data collection available in the area of interest and that data needs to be publicly available. This includes reservoir level data, water withdrawal data, streamflow data, and weather data, all of which may come from different sources. Moreover, many reservoirs that I considered initially had some data, but not all of it. For example, several large reservoirs run by the Army Corps of Engineers were considered for inclusion in this study. These reservoirs had all of the data except the water withdrawals. Looking deeper into these sites, it was difficult to determine who had rights to withdraw water and how much they would be withdrawing. The lack of data led to the exclusion of these large reservoirs. Collecting this data is possible, but would require a deeper dive into specific reservoirs that might be tied to many different cities and towns.

Additionally, in Chapter 2, the projections were made based on climatological mean. This is a good estimation when the projection is being made for the next season, but not for the next decade. This issue of lead times is something that should be explored, especially with regard to different stakeholders. For example, a reservoir manager would be most interested in the next season's storage, but an investor would likely want to know the next decade's outlook in order to plan future supply. In order to project the reservoir storage a decade into the future, one would need projections of the inputs and outputs to the reservoir. This would ultimately require the integration of a number of different models. For example, the precipitation and evaporation data could be obtained from the global climate models (GCMs), while the water consumption data can be estimated based on population growth models. The tricky variables would be the streamflow in and out of the reservoir. This could potentially be obtained from physics-based hydrological simulations, or estimated based on the

precipitation and statistical models of streamflow. Either way, making projections at different lead times will require the integration of more data from a variety of sources.

Another limitation was the lack of climate data included in the analysis for Chapter 4. As discussed within the chapter, the available climate data was a single value for the entire city, so when included in an intra-city analysis, the variables were determined to have no impact. Technically, the climate plays a major role in water consumption, as shown in Chapter 3, but without higher resolution data to include in the analysis, adding climate data to an intra-city model had no effect on the final water consumption. This is a significant limitation, as the climate data would likely improve the predictive accuracy and allow for a more holistic analysis of water consumption.

Finally, in Chapter 4, the semi-structured interviews were not performed over the entire county area, making it difficult to interpret the results beyond the few neighborhoods considered. This was done primarily due to lack of participants from those outside neighborhoods. Initially, I sent interview requests to a number of different people from around the city, but only those presented here accepted. Ideally, I would have gotten more responses and been able to have a more spatially diverse group of interviewees.

5.2.2 Future Work

That being said, the limitations of this study present an opportunity to continue to work in this area and potentially improve upon the results presented here. For example, there are a number of novel data sources (e.g., remote sensing) that can be utilized for evaluating urban reservoir levels. By tapping into these sources, one could expand the study presented in Chapter 2 beyond the nine reservoirs, potentially covering the entire US. Moreover, using high resolution climate data, such as the PRISM project from the University Center for Atmospheric Research, could significantly improve the model developed in Chapter 4. Additional research should be

done on the impact of the extreme variation among the different census tracts. It is likely that by removing the outlier tracts, the predictive accuracy could be improved in the rest of the city. Finally, developing a simulation tool using agent-based modeling would be a way to model consumer behavior in a more dynamic setting, rather than a static analysis presented here. This would require additional data, which could be collected via surveys. Surveys would potentially lead to a more spatially heterogeneous dataset, as they could easily be sent to people living in a number of different areas around the city. Overall, this study is not without limitations, however, I would recommend future researchers to take these limitations as opportunities to expand upon and improve the results presented here.

5.3 Applicability Beyond Urban Water Systems

The work presented in this dissertation was primarily focused on urban water systems. However, the methodologies used and the tools developed are applicable to many areas. For example, the multivariate framework developed in Chapter 3 can be used to evaluate different interconnected systems. For example, on the energy demand side, electricity and natural gas demand are interconnected. Using the framework outlined in Chapter 3, it was possible to evaluate this nexus and improve predictive accuracy beyond the univariate model [125]. Beyond urban systems, the multivariate framework can be applied to measures of resilience, which are often interconnected. For example, when predicting death and damage rates of tsunamis, the results for the multivariate model were different than that of the univariate. In fact, although both models were fairly accurate, choosing one over the other would lead practitioners to arrive at different conclusions [126]. This work on multi-outcome modeling can be expanded to encompass the food-water-energy nexus or even compounding disasters. In general, multi-outcome modeling allows us to better represent complex systems in a data-driven framework.

On a more philosophical level, the results in this dissertation demonstrated the complementary nature of different methodologies that are often seen as opposites. For instance, engineering studies and social science studies are often performed separately, even if they are focused on the same application. In fact, it is likely that researchers from these two disciplines might not even be aware of the work being done within the other discipline. This siloed practice may be leading to misinformed decisions, if those decisions are based on a study that excluded part of the equation (e.g., an infrastructure study that did not take into account how people use the infrastructure). The work presented in Chapter 4 focused on integrating social science research methods, namely qualitative interviews, with a statistical learning model that can be used to predict water consumption. This is a critical first step towards creating a boundary object that can overlap both engineering and the social sciences. Ultimately, creating a boundary object will be critical for improving resilience and solving the grand challenges that are facing society today.

Similarly, in Chapter 2, I presented multiple models that can be used to understand urban reservoir levels. Beyond studying water resources, there are lessons to be learned and applied in other fields. For example, it is critical to understand multiple methodologies, since one might be better than the others in certain situations. Additionally, being aware of the pros and cons of various methodologies used within different fields is important. If one goes about research and only looks into the methodologies that they are comfortable with or that are prevalent in their field, it is likely that they will miss out on a technique that may be more applicable in a given situation. At worse, using one technique over others could result in errors being propagated throughout the process without the knowledge of the researcher. The ability to keep an open mind is critical for research, especially in interdisciplinary work.

Overall, this dissertation sought to use interdisciplinary thinking to better understand the impact of climate change on urban water resources. However, the models used and conclusions drawn can be applied across many applications and can guide

future endeavors into interdisciplinary research. Interdisciplinary research is incredibly important and I truly believe that we will not be able to solve many of the worlds problems without it.

REFERENCES

REFERENCES

[1] The World Bank, "Cities and Climate Change: An Urgent Agenda," Tech. Rep., 2010.

[2] J. Rockström, W. Steffen, K. Noone, Å. Persson, F. S. I. Chapin, E. Lambin, T. Lenton, M. Scheffer, C. Folke, H. J. Schellnhuber, B. Nykvist, C. de Wit, T. Hughes, S. van der Leeuw, H. Rodhe, S. Sörlin, P. Snyder, R. Costanza, U. Svedin, M. Falkenmark, L. Karlberg, R. Corell, V. Fabry, J. Hansen, B. Walker, D. Liverman, K. Richardson, P. Crutzen, and J. Foley, "Planetary Boundaries: Exploring the Safe Operating Space for Humanity," *Ecology and Society*, vol. 14, no. 2, Nov. 2009.

[3] D. Butler, S. Ward, C. Sweetapple, M. Astaraie-Imani, K. Diao, R. Farmani, and G. Fu, "Reliable, resilient and sustainable water management: The Safe & SuRe approach," *Global Challenges*, vol. 1, no. 1, pp. 63–77, 2017.

[4] P. H. Gleick, "Global Freshwater Resources: Soft-Path Solutions for the 21st Century," *Science*, vol. 302, no. 5650, pp. 1524–1528, Nov. 2003.

[5] V. G. Mitchell, "Applying Integrated Urban Water Management Concepts: A Review of Australian Experience," *Environmental Management*, vol. 37, no. 5, pp. 589–605, May 2006.

[6] A. AghaKouchak, D. Feldman, M. Hoerling, T. Huxman, and J. Lund, "Water and climate: Recognize anthropogenic drought," *Nature News*, vol. 524, no. 7566, p. 409, Aug. 2015.

[7] A. K. Mishra and V. P. Singh, "A review of drought concepts," *Journal of Hydrology*, vol. 391, no. 1, pp. 202–216, Sep. 2010.

[8] M. Ferrazzi, R. Vivian, and G. Botter, "Sensitivity of Regulated Streamflow Regimes to Interannual Climate Variability," *Earth's Future*, vol. 7, no. 11, pp. 1206–1219, 2019.

[9] R. I. McDonald, P. Green, D. Balk, B. M. Fekete, C. Revenga, M. Todd, and M. Montgomery, "Urban growth, climate change, and freshwater availability," *Proceedings of the National Academy of Sciences*, vol. 108, no. 15, pp. 6312–6317, Apr. 2011.

[10] C. J. Vörösmarty, P. Green, J. Salisbury, and R. B. Lammers, "Global Water Resources: Vulnerability from Climate Change and Population Growth," *Science*, vol. 289, no. 5477, pp. 284–288, Jul. 2000.

[11] B. Tarroja, A. AghaKouchak, R. Sobhani, D. Feldman, S. Jiang, and S. Samuelsen, "Evaluating options for Balancing the Water-Electricity Nexus in California: Part 1 – Securing Water Availability," *Science of The Total Environment*, vol. 497-498, pp. 697–710, Nov. 2014.

[12] J. K. O'Hara and K. P. Georgakakos, "Quantifying the Urban Water Supply Impacts of Climate Change," *Water Resources Management*, vol. 22, no. 10, pp. 1477–1497, Oct. 2008.

[13] J. Park, G. Botter, J. W. Jawitz, and P. S. C. Rao, "Stochastic modeling of hydrologic variability of geographically isolated wetlands: Effects of hydro-climatic forcing and wetland bathymetry," *Advances in Water Resources*, vol. 69, pp. 38–48, Jul. 2014.

[14] L. E. Bertassello, P. S. C. Rao, J. Park, J. W. Jawitz, and G. Botter, "Stochastic modeling of wetland-groundwater systems," *Advances in Water Resources*, vol. 112, pp. 214–223, Feb. 2018.

[15] A. Ficchi, L. Raso, P.-O. Malaterre, D. Dorchies, M. Jay-Allemand, F. Pianosi, P.-J. van Overloop, and G. Thirel, "Short Term Reservoirs Operation On The Seine River: Performance Analysis Of Tree-Based Model Predictive Control," *International Conference on Hydroinformatics*, Aug. 2014.

[16] T. Yang, X. Gao, S. Sorooshian, and X. Li, "Simulating California reservoir operation using the classification and regression-tree algorithm combined with a shuffled cross-validation scheme," *Water Resources Research*, vol. 52, no. 3, pp. 1626–1651, 2016.

[17] S. Derrible, "Urban infrastructure is not a tree: Integrating and decentralizing urban infrastructure systems," *Environment and Planning B: Urban Analytics and City Science*, vol. 44, no. 3, pp. 553–569, May 2017.

[18] A. Maas, C. Goemans, D. T. Manning, J. Burkhardt, and M. Arabi, "Complements of the house: Estimating demand-side linkages between residential water and electricity," *Water Resources and Economics*, p. 100140, Mar. 2019.

[19] A. Escriva-Bou, J. R. Lund, and M. Pulido-Velazquez, "Saving Energy From Urban Water Demand Management," *Water Resources Research*, vol. 54, no. 7, pp. 4265–4276, 2018.

[20] B. R. Scanlon, I. Duncan, and R. C. Reedy, "Drought and the water–energy nexus in Texas," *Environmental Research Letters*, vol. 8, no. 4, p. 045033, Dec. 2013.

[21] M. T. H. van Vliet, S. Vögele, and D. Rübbelke, "Water constraints on European power supply under climate change: Impacts on electricity prices," *Environmental Research Letters*, vol. 8, no. 3, p. 035010, Jul. 2013.

[22] B. K. Sovacool and K. E. Sovacool, "Identifying future electricity–water tradeoffs in the United States," *Energy Policy*, vol. 37, no. 7, pp. 2763–2773, Jul. 2009.

[23] J. Macknick, R. Newmark, G. Heath, and K. C. Hallett, "Operational water consumption and withdrawal factors for electricity generating technologies: A review of existing literature," *Environmental Research Letters*, vol. 7, no. 4, p. 045802, Dec. 2012.

[24] F. Ackerman and J. Fisher, "Is there a water–energy nexus in electricity generation? Long-term scenarios for the western United States," *Energy Policy*, vol. 59, pp. 235–241, Aug. 2013.

[25] D. M. Ruddell and P. G. Dixon, "The energy–water nexus: Are there tradeoffs between residential energy and water consumption in arid cities?" *International Journal of Biometeorology*, vol. 58, no. 7, pp. 1421–1431, Sep. 2014.

[26] A. S. Vieira and E. Ghisi, "Water-energy nexus in low-income houses in Brazil: The influence of integrated on-site water and sewage management strategies on the energy consumption of water and sewerage services," *Journal of Cleaner Production*, vol. 133, pp. 145–162, Oct. 2016.

[27] K. L. Lam, S. J. Kenway, and P. A. Lant, "Energy use for water provision in cities," *Journal of Cleaner Production*, vol. 143, pp. 699–709, Feb. 2017.

[28] G. Venkatesh, A. Chan, and H. Brattebø, "Understanding the water-energy-carbon nexus in urban water utilities: Comparison of four city case studies and the relevant influencing factors," *Energy*, vol. 75, pp. 153–166, Oct. 2014.

[29] N. Mostafavi, F. Gándara, and S. Hoque, "Predicting water consumption from energy data: Modeling the residential energy and water nexus in the integrated urban metabolism analysis tool (IUMAT)," *Energy and Buildings*, vol. 158, pp. 1683–1693, Jan. 2018.

[30] H. Allcott, "Social norms and energy conservation," *Journal of Public Economics*, vol. 95, no. 9, pp. 1082–1095, Oct. 2011.

[31] S. P. Bhanot, "Rank and response: A field experiment on peer information and water use behavior," *Journal of Economic Psychology*, vol. 62, pp. 155–172, Oct. 2017.

[32] A. Dai, "Drought under global warming: A review," *WIREs Climate Change*, vol. 2, no. 1, pp. 45–65, 2011.

[33] J. C. de Araújo and A. Bronstert, "A method to assess hydrological drought in semi-arid environments and its application to the Jaguaribe River basin, Brazil," *Water International*, vol. 41, no. 2, pp. 213–230, Feb. 2016.

[34] Atlanta Regional Commission, "Water Metrics Report," Metropolitan North Georgia Water Planning Dsitrict, Tech. Rep., 2011.

[35] T. M. Missimer, P. A. Danser, G. Amy, and T. Pankratz, "Water crisis: The metropolitan Atlanta, Georgia, regional water supply conflict," *Water Policy*, vol. 16, no. 4, pp. 669–689, Aug. 2014.

[36] US Census Bureau, "Quick Facts: Atlanta, Georgia," Tech. Rep., 2016.

[37] US Army Corps of Engineers, "Lake Lanier Level Data," Tech. Rep.

[38] US Geological Survey, "USGS 02338000 Chattahoochee River Near Whitesburg, GA," Tech. Rep., 2017.

[39] NOAA National Centers for Environmental Information, "Local Climatological Data (LCD)," 2019.

[40] H. van den Dool, J. Huang, and Y. Fan, "Performance and analysis of the constructed analogue method applied to U.S. soil moisture over 1981–2001," *Journal of Geophysical Research: Atmospheres*, vol. 108, no. D16, 2003.

[41] K. Wolter and M. S. Timlin, "Measuring the strength of ENSO events: How does 1997/98 rank?" *Weather*, vol. 53, no. 9, pp. 315–324, 1998.

[42] J. McLeod and C. Konrad, "El Niño Impacts and Outlook: Southeast Region," Southeast Regional Climate Center, Tech. Rep., 2015.

[43] R. Nateghi, S. D. Guikema, and S. M. Quiring, "Comparison and Validation of Statistical Methods for Predicting Power Outage Durations in the Event of Hurricanes," *Risk Analysis*, vol. 31, no. 12, pp. 1897–1906, 2011.

[44] R. Nateghi, J. D. Bricker, S. D. Guikema, and A. Bessho, "Statistical Analysis of the Effectiveness of Seawalls and Coastal Forests in Mitigating Tsunami Impacts in Iwate and Miyagi Prefectures," *PLoS ONE*, vol. 11, no. 8, Aug. 2016.

[45] B. Grizzetti, A. Pistocchi, C. Liquete, A. Udias, F. Bouraoui, and W. van de Bund, "Human pressures and ecological status of European rivers," *Scientific Reports*, vol. 7, no. 1, pp. 1–11, Mar. 2017.

[46] K. Nishina, M. Watanabe, M. K. Koshikawa, T. Takamatsu, Y. Morino, T. Nagashima, K. Soma, and S. Hayashi, "Varying sensitivity of mountainous streamwater base-flow NO3- concentrations to N deposition in the northern suburbs of Tokyo," *Scientific Reports*, vol. 7, no. 1, pp. 1–9, Aug. 2017.

[47] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Science & Business Media, Aug. 2009.

[48] J. A. Nelder and R. W. M. Wedderburn, "Generalized Linear Models," *Journal of the Royal Statistical Society. Series A (General)*, vol. 135, no. 3, pp. 370–384, 1972.

[49] T. Hastie and R. Tibshirani, "Generalized Additive Models: Some Applications," *Journal of the American Statistical Association*, vol. 82, no. 398, pp. 371–386, Jun. 1987.

[50] J. H. Friedman, "Multivariate Adaptive Regression Splines," *The Annals of Statistics*, vol. 19, no. 1, pp. 1–67, 1991.

[51] L. Breiman, *Classification And Regression Trees*. Boca Raton, FL: Taylor & Francis Group, 1984.

[52] ——, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, Aug. 1996.

[53] ——, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[54] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sep. 1995.

[55] H. A. Chipman, E. I. George, and R. E. McCulloch, "BART: Bayesian additive regression trees," *The Annals of Applied Statistics*, vol. 4, no. 1, pp. 266–298, Mar. 2010.

[56] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[57] US Army Corps of Engineers, "National Inventory of Dams," US Army Corps of Engineers, Tech. Rep., 2020.

[58] E. Kalnay, M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, M. Chelliah, W. Ebusuzaki, W. Higgins, J. Janowiak, K. C. Mo, C. Ropelewski, J. Wang, A. Leetmaa, R. Reynolds, R. Jenne, and D. Joseph, "The NCEP/NCAR 40-Year Reanalysis Project," *Bulletin of the American Meteorological Society*, 1996.

[59] US Geological Survey, "USGS Current Water Data for the Nation," US Geological Survey, Tech. Rep., 2020.

[60] G. Botter, S. Basso, A. Porporato, I. Rodriguez-Iturbe, and A. Rinaldo, "Natural streamflow regime alterations: Damming of the Piave river basin (Italy)," *Water Resources Research*, vol. 46, no. 6, 2010.

[61] A. N. Kolmogorov, "Sulla determinazione empirica di una legge di distribuzione," *Giornale dell'Instituto Italiano degli Attuari*, vol. 4, pp. 83–91, 1933.

[62] N. Smirnov, "Estimate of deviation between empirical distribution functions in two independent samples," *Bulletin Moscow University*, vol. 2, no. 2, pp. 3–16, 1939.

[63] B. L. Welch, "The Generalization of 'Student's' Problem When Several Different Population Variances are Involved," *Biometrika*, vol. 34, no. 1-2, pp. 28–35, Jan. 1947.

[64] Student, "The Probable Error of a Mean," *Biometrika*, vol. 6, no. 1, pp. 1–25, 1908.

[65] G. Shmueli, "To Explain or to Predict?" *Statistical Science*, vol. 25, no. 3, pp. 289–310, Aug. 2010.

[66] A. Ganguly, E. Kodra, A. Agrawal, A. Banerjee, S. Boriah, S. Chatterjee, S. Chatterjee, A. Choudhary, D. Das, J. Faghmous *et al.*, "Toward enhanced understanding and projections of climate extremes using physics-guided data mining techniques," 2014.

[67] J. P. Newell, B. Goldstein, and A. Foster, "A 40-year review of food–energy–water nexus literature and its application to the urban scale," *Environmental Research Letters*, vol. 14, no. 7, p. 073003, Jul. 2019.

[68] M. D. Bartos and M. V. Chester, "The Conservation Nexus: Valuing Interdependent Water and Energy Savings in Arizona," *Environmental Science & Technology*, vol. 48, no. 4, pp. 2139–2149, Feb. 2014.

[69] Z. Khan, P. Linares, M. Rutten, S. Parkinson, N. Johnson, and J. García-González, "Spatial and temporal synchronization of water and energy systems: Towards a single integrated optimization model for long-term resource planning," *Applied Energy*, vol. 210, pp. 499–517, Jan. 2018.

[70] J. F. Kenny, N. L. Barber, S. S. Hutson, K. S. Linsey, J. K. Lovelace, and M. A. Maupin, "Estimated Use of Water in the United States in 2005," U.S. Geological Survey Circular 1344, 2009.

[71] T. A. DeNooyer, J. M. Peschel, Z. Zhang, and A. S. Stillwell, "Integrating water resources and power generation: The energy–water nexus in Illinois," *Applied Energy*, vol. 162, pp. 363–371, Jan. 2016.

[72] U. Lee, J. Han, A. Elgowainy, and M. Wang, "Regional water consumption for hydro and thermal electricity generation in the United States," *Applied Energy*, vol. 210, pp. 661–672, Jan. 2018.

[73] C. M. Chini and A. S. Stillwell, "The State of U.S. Urban Water: Data and the Energy-Water Nexus," *Water Resources Research*, vol. 54, no. 3, pp. 1796–1811, 2018.

[74] K. T. Sanders and M. E. Webber, "Evaluating the energy consumed for water use in the United States," *Environmental Research Letters*, vol. 7, no. 3, p. 034034, Sep. 2012.

[75] S. F. Hoque, *Water Conservation in Urban Households: THe ANswer to Water Shortage in the 21st Century*. IWA Publishing, 2014.

[76] A. K. Plappally and J. H. Lienhard V, "Energy requirements for water production, treatment, end use, reclamation, and disposal," *Renewable and Sustainable Energy Reviews*, vol. 16, no. 7, pp. 4818–4848, Sep. 2012.

[77] D. Conway, E. A. van Garderen, D. Deryng, S. Dorling, T. Krueger, W. Landman, B. Lankford, K. Lebek, T. Osborn, C. Ringler, J. Thurlow, T. Zhu, and C. Dalin, "Climate and southern Africa's water–energy–food nexus," *Nature Climate Change*, vol. 5, no. 9, pp. 837–846, Sep. 2015.

[78] L. Raymond, D. Gotham, W. McClain, S. Mukherjee, R. Nateghi, P. V. Preckel, P. Schubert, S. Singh, and E. Wachs, "Projected climate change impacts on Indiana's Energy demand and supply," *Climatic Change*, Jan. 2019.

[79] S. Mukhopadhyay and R. Nateghi, "Estimating climate — Demand Nexus to support longterm adequacy planning in the energy sector," in *2017 IEEE Power Energy Society General Meeting*, Jul. 2017, pp. 1–5.

[80] S. Mukherjee and R. Nateghi, "Climate sensitivity of end-use electricity consumption in the built environment: An application to the state of Florida, United States," *Energy*, vol. 128, pp. 688–700, Jun. 2017.

[81] M. Auffhammer, P. Baylis, and C. H. Hausman, "Climate change is projected to have severe impacts on the frequency and intensity of peak electricity demand across the United States," *Proceedings of the National Academy of Sciences*, vol. 114, no. 8, pp. 1886–1891, Feb. 2017.

[82] M. Lokhandwala and R. Nateghi, "Leveraging advanced predictive analytics to assess commercial cooling load in the U.S." *Sustainable Production and Consumption*, vol. 14, pp. 66–81, Apr. 2018.

[83] R. Nateghi, S. D. Guikema, Y. G. Wu, and C. B. Bruss, "Critical Assessment of the Foundations of Power Transmission and Distribution Reliability Metrics and Standards," *Risk Analysis*, vol. 36, no. 1, pp. 4–15, 2016.

[84] S. Mukherjee, C. R. Vineeth, and R. Nateghi, "Evaluating regional climate-electricity demand nexus: A composite Bayesian predictive framework," *Applied Energy*, vol. 235, pp. 1561–1582, Feb. 2019.

[85] R. Nateghi and S. Mukherjee, "A multi-paradigm framework to assess the impacts of climate change on end-use energy demand," *PLOS ONE*, vol. 12, no. 11, p. e0188033, Nov. 2017.

[86] J. Cronin, G. Anandarajah, and O. Dessens, "Climate change impacts on the energy system: A review of trends and gaps," *Climatic Change*, vol. 151, no. 2, pp. 79–93, Nov. 2018.

[87] S. Mukherjee and R. Nateghi, "A Data-Driven Approach to Assessing Supply Inadequacy Risks Due to Climate-Induced Shifts in Electricity Demand," *Risk Analysis*, vol. 39, no. 3, pp. 673–694, 2019.

[88] C. B. Bruss, R. Nateghi, and B. F. Zaitchik, "Explaining National Trends in Terrestrial Water Storage," *Frontiers in Environmental Science*, vol. 7, 2019.

[89] R. C. Balling, P. Gober, and N. Jones, "Sensitivity of residential water consumption to variations in climate: An intraurban analysis of Phoenix, Arizona," *Water Resources Research*, vol. 44, no. 10, 2008.

[90] N. Ashoori, D. A. Dzombak, and M. J. Small, "Modeling the Effects of Conservation, Demographics, Price, and Climate on Urban Water Demand in Los Angeles, California," *Water Resources Management*, vol. 30, no. 14, pp. 5247–5262, Nov. 2016.

[91] S. J. Pereira-Cardenal, H. Madsen, K. Arnbjerg-Nielsen, N. Riegels, R. Jensen, B. Mo, I. Wangensteen, and P. Bauer-Gottwein, "Assessing climate change impacts on the Iberian power system using a coupled water-power model," *Climatic Change*, vol. 126, no. 3, pp. 351–364, Oct. 2014.

[92] B. Gjorgiev and G. Sansavini, "Electrical power generation under policy constrained water-energy nexus," *Applied Energy*, vol. 210, pp. 568–579, Jan. 2018.

[93] B. Rachunok and R. Nateghi, "Interdependent Infrastructure System Risk and Resilience to Natural Hazards," *Proceedingsof the 2019 IISE Annual Conference*, Apr. 2019.

[94] B. A. Rachunok, J. B. Bennett, and R. Nateghi, "Twitter and Disasters: A Social Resilience Fingerprint," *IEEE Access*, vol. 7, pp. 58 495–58 506, 2019.

[95] L. L. Dale, N. Karali, D. Millstein, M. Carnall, S. Vicuña, N. Borchers, E. Bustos, J. O'Hagan, D. Purkey, C. Heaps, J. Sieber, W. D. Collins, and M. D. Sohn, "An integrated assessment of water-energy and climate change in sacramento, california: How strong is the nexus?" *Climatic Change*, vol. 132, no. 2, pp. 223–235, Sep. 2015.

[96] Intergovernemental Panel on Climate Change, "Climate Change 2013: The Physical Science Basis," Tech. Rep., 2013.

[97] US Energy Information Administration, "Form EIA-861M Sales and Revenue Data," 2019.

[98] D. Kluck and M. Woloszyn, "El Nino Impacts and Outlook: Midwest Region," National Oceanic and Atmospheric Administration, Tech. Rep., 2016.

[99] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*. OTexts, May 2018.

[100] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," *Proceedings of the 23rd international conference on Machine Learning*.

[101] P. J. Miller, G. H. Lubke, D. B. McArtor, and C. S. Bergeman, "Finding structure in data using multivariate tree boosting," *Psychological Methods*, vol. 21, no. 4, pp. 583–602, 2016.

[102] S.-W. Yeh, J.-S. Kug, B. Dewitte, M.-H. Kwon, B. P. Kirtman, and F.-F. Jin, "El Niño in a changing climate," *Nature*, vol. 461, no. 7263, pp. 511–514, Sep. 2009.

[103] D. J. Sailor and J. R. Muñoz, "Sensitivity of electricity and natural gas consumption to climate in the U.S.A.—Methodology and results for eight states," *Energy*, vol. 22, no. 10, pp. 987–998, Oct. 1997.

[104] L. Warszawski, K. Frieler, V. Huber, F. Piontek, O. Serdeczny, and J. Schewe, "The Inter-Sectoral Impact Model Intercomparison Project (ISI–MIP): Project framework," *Proceedings of the National Academy of Sciences*, vol. 111, no. 9, pp. 3228–3232, Mar. 2014.

[105] S. Hempel, K. Frieler, L. Warszawski, J. Schewe, and F. Piontek, "A trend-preserving bias correction - the ISI-MIP approach," *Earth System Dynamics*, vol. 4, no. 2, pp. 219–236, Jul. 2013.

[106] Intergovernemental Panel on Climate Change, "Global Warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emissionpathways, in the context of strengthening the global response to the threat of climate change," 2018, Tech. Rep.

[107] R. Loulou, G. Goldstein, and K. Noble, "Documentation for the MARKAL Family of Models," Energy Technology Systems Analysis Programme, Tech. Rep., 2004.

[108] R. James, R. Washington, C.-F. Schleussner, J. Rogelj, and D. Conway, "Characterizing half-a-degree difference: A review of methods for identifying regional climate responses to global warming targets," *WIREs Climate Change*, vol. 8, no. 2, p. e457, 2017.

[109] D. Jacob, L. Kotova, C. Teichmann, S. P. Sobolowski, R. Vautard, C. Donnelly, A. G. Koutroulis, M. G. Grillakis, I. K. Tsanis, A. Damm, A. Sakalli, and M. T. H. van Vliet, "Climate Impacts in Europe Under +1.5°C Global Warming," *Earth's Future*, vol. 6, no. 2, pp. 264–285, 2018.

[110] L. Samaniego, S. Thober, R. Kumar, N. Wanders, O. Rakovec, M. Pan, M. Zink, J. Sheffield, E. F. Wood, and A. Marx, "Anthropogenic warming exacerbates European soil moisture droughts," *Nature Climate Change*, vol. 8, no. 5, pp. 421–426, May 2018.

[111] A. Marx, R. Kumar, S. Thober, O. Rakovec, N. Wanders, M. Zink, E. F. Wood, M. Pan, J. Sheffield, and L. Samaniego, "Climate change alters low flows in Europe under global warming of 1.5, 2, and 3 °C," *Hydrology and Earth System Sciences*, vol. 22, no. 2, pp. 1017–1032, Feb. 2018.

[112] R. Singh and R. Kumar, "Climate versus demographic controls on water availability across India at 1.5 °C, 2.0 °C and 3.0 °C global warming levels," *Global and Planetary Change*, vol. 177, pp. 1–9, Jun. 2019.

[113] R. Vautard, A. Gobiet, S. Sobolowski, E. Kjellström, A. Stegehuis, P. Watkiss, T. Mendlik, O. Landgren, G. Nikulin, C. Teichmann, and D. Jacob, "The European climate under a 2\hspace0.167em°C global warming," *Environmental Research Letters*, vol. 9, no. 3, p. 034006, Mar. 2014.

[114] R. Obringer, R. Kumar, and R. Nateghi, "Analyzing the climate sensitivity of the coupled water-electricity demand nexus in the Midwestern United States," *Applied Energy*, vol. 252, p. 113466, Oct. 2019.

[115] UNFCC, "Adoption of the Paris Agreement, Proposal by the President," United Nations, Geneva, Switzerland, Tech. Rep., 2015.

[116] M. E. Hauer, "Population projections for U.S. counties by age, sex, and race controlled to shared socioeconomic pathway," *Scientific Data*, vol. 6, no. 1, pp. 1–15, Feb. 2019.

[117] M. V. Viñoles, K. Moeltner, and S. Stoddard, "Length of residency and water use in an arid urban environment," *Water Resources and Economics*, vol. 12, pp. 52–66, Oct. 2015.

[118] C. M. Jaeger and P. W. Schultz, "Coupling social norms and commitments: Testing the underdetected nature of social influence," *Journal of Environmental Psychology*, vol. 51, pp. 199–208, Aug. 2017.

[119] C. Bicchieri, *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms*. Oxford University Press, Dec. 2016.

[120] C. Horne, "The Relational Foundation of Norm Enforcement," in *The Complexity of Social Norms*, ser. Computational Social Sciences, M. Xenitidou and B. Edmonds, Eds. Cham: Springer International Publishing, 2014, pp. 105–120.

[121] C. Bicchieri, *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge University Press, Dec. 2005.

[122] B. K. Nastasi, J. H. Hitchcock, and L. M. Brown, "An Inclusive Framework for Conceptualizing Mixed Methods Design Typologies: Moving Toward Fully Integrated Synergistic Research Models," in *SAGE Handbook of Mixed Methods in Social & Behavioral Research*. 2455 Teller Road, Thousand Oaks California 91320 United States: SAGE Publications, Inc., 2010, pp. 305–338.

[123] H. R. Bernard and H. R. Bernard, *Social Research Methods: Qualitative and Quantitative Approaches*. SAGE, 2013.

[124] G. Guest, A. Bunce, and L. Johnson, "How Many Interviews Are Enough?: An Experiment with Data Saturation and Variability," *Field Methods*, vol. 18, no. 1, pp. 59–82, Feb. 2006.

[125] R. Obringer, S. Mukherjee, and R. Nateghi, "Evaluating the climate sensitivity of coupled electricity-natural gas demand using a multivariate framework," *Applied Energy*, vol. 262, p. 114419, Mar. 2020.

[126] R. Obringer and R. Nateghi, "Multivariate Modeling for Sustainable and Resilient Infrastructure Systems and Communities," *Proceedingsof the 2019 IISE Annual Conference*, May 2019.

APPENDICES

A. SUPPLEMENTARY INFORMATION FOR CHAPTER 2

A portion of this appendix has been previously published as supplementary material in *Scientific Reports*: https://doi.org/10.1038/s41598-018-23509-w.

Contents of this appendix include:

A.1 Methods

A.1.1 Generalized Linear Model (GLM

The generalized linear model (GLM) is an extension of linear regression that relaxes the normality assumption. In this model, the response is generated from an exponential distribution and then related to the predictors through a link function [48]. The GLM is defined by:

I A dependent variable Y that has a known distribution (i.e., normal, binomial, Poisson, or gamma), as shown below:

$$Y_i \sim f_{Y_i}(y_i)$$

$$f_{Y_i}(y_i) = exp[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)]$$

where θ and ϕ are the location and scale parameters, respectively.

II A set of independent variables x_i.

III A linking function $g(.)$ that relates the response variable to the predictors.

A.1.2 Generalized Additive Model (GAM)

The generalized additive model (GAM) is a further extension of linear regression, which in addition to relaxing the normality assumption as in the GLM also relaxes the linearity assumption, meaning that there could be local nonlinearities [49]. In the GAM, the response variable y has a distribution with mean $\mu = E[Y|x_1, X-2, ..., x_n]$ that is linked to the predictors through the link function:

$$g(\mu_i) = \alpha + \sum_{j=1}^{n} f_j(x_j)$$

where f_j is a smoothing function (i.e., a regression spline).

A.1.3 Multivariate Adaptive Regression Splines (MARS)

The multivariate adaptive regression splines (MARS) method is a semi-parametric procedure that combines recursive partitioning regression and spline fitting [50]. The model takes on the following mathematical form:

$$f(X) = \beta_0 + \sum_{j=1}^{n} \beta_j h_j(X)$$

where $h_j(X)$ is the linear spline, β_0 is the intercept, and β_j is the vector of coefficients, which are estimated by minimizing the sum of squares error. The MARS method also uses generalized cross validation (GCV) to avoid overfitting the model. This method penalizes complexity, which makes MARS especially applicable for high-dimensional datasets.

$$GCV = \frac{RSS}{N \times (1 - \frac{C}{N})^2}$$

where RSS is the residual sum of squares, N is the number of observations, and C is the effective number of parameters.

A.1.4 Classification and Regression Trees (CART)

The classification and regression tree (CART) method operates by iteratively partitioning the dataset into boxes in such a way that the residual sum of squares is minimized [51]. The partitioning is performed using the recursive binary splitting technique, an example of which is shown below:

$$R_1(j, s) = [X | X_j < s]$$

$$R_2(j, s) = [X | X_j \geq s]$$

where R_1 and R_2 are the partitioned boxes, X is the dataset, and s is the partitioning threshold.

A.1.5 Bagged Classification and Regression Trees

Bagging is a meta-algorithm that uses bootstrap aggregation to reduce the variance of the prediction5. The bagged CART method uses bootstrapping to iteratively run the CART method over a subset of the data, the final tree being an aggregation of all the iterations. The mathematical representation of bagging is:

$$\phi_B(x) = a\nu_B\phi(x, \mathcal{L}^{(B)})$$

where $\mathcal{L}^{(B)}$ is the subset of the data used in the bootstrapping procedure and $\phi(x, \mathcal{L}^{(B)})$ is the predictor formed from the bootstrapped sample. B represents the number of bootstrapped iterations.

A.1.6 Random Forest

Random forest is a tree-based ensemble method that builds B bootstrapped, decorrelated regression trees and then aggregates those trees to a single model [53]. The additional layers of randomness introduced in the random forest algorithm that leads to reduced correlation among the trees leads to further variance reduction and as a result improved performance over bagged-trees. The final model can be represented by the average of all the trees:

$$\hat{f}^B(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$$

where T_b is the regression tree and B is the number of bootstrapping iterations.

A.1.7 Support Vector Machines

Support vector machine (SVM) is an optimization technique, which allows for finding the global solution and therefore often leads to more accurate predictions [54]. The goal of the support vector machine algorithm is to find a hyperplane that

maximizes the margin between the two classes of data. The hyperplane can be found by:

$$\hat{f}_x = \sum_{i=1}^{n} \hat{a}_i K(x, x_i)$$

where $\hat{a}_i = (HH^T + \lambda I)^{-1} y$ (a transformation of the basis matrix H) and $K(x, x_i)$ is the kernel function (i.e., linear, radial, or polynomial).

A.1.8 Bayesian Additive Regression Trees

The Bayesian additive regression tree (BART) technique is an ensemble-based method that uses boosting to improve predictive accuracy [55]. Boosting, as opposed to bagging, fits a series of trees in which each tree is used to fit the variability not explained by the previous trees [56]. The BART method works by constraining each individual tree by implementing a regularization prior, creating a series of weak learners. The result is a sum of trees where each tree explains a different part of the whole:

$$Y = \sum_{j=1}^{n} g(x; T_j, M_j) + \epsilon$$

where T_j is a single regression tree, M_j is a set of parameter values, and ϵ is the error with distribution $N(0, \sigma^2)$.

A.2 Tables

Table A.1.: Data collected for the statistical learning study.

Variable Type	Variable Name	Minimum	Maximum	Mean	Units	Source
Response	Reservoir Level	1050.8	1076.2	1067.1	ft	[37]
Predictor	Precipitation	0.0	7.0	0.13	in	[39]
	Streamflow	66	15800	766.3	ft^3/s	[38]
	Discharge	852	58600	3900	ft^3/s	[**?**]
	Water Use	393760	497337	432022	gpcd	[34]
	Population	125000	214000	190000	people	[36]
	ENSO	-2	3	0.20	—	[41]
	Soil Moisture	271.3	673.2	470.2	mm/m	[40]
	Dew Point	-13.6	75.7	49.7	°F	[39]
	Rel. Humidity	23.3	100.0	68.0	%	[39]

Table A.2.: Tuning parameters used in the statistical learning models.

Model	Tuning Parameter*	Parameter Value	Parameter Description
GLM	family	Gaussian	link function
GAM	stepwise update	—	runs model in stepwise fashion
MARS	nk	8	max number of model terms
	nprune	7	max number of terms in pruned model
	degree	1	degree of interaction
	penalty	1	GCV penalty per knot
CART	—	—	—
Bagged CART	nbagg	25	number of bootstrap replications
Random Forest	ntree	30	number of trees to grow
	mtry	4	number of variables to sample at each split
SVM	kernel	radial	kernel
	cost	10	cost of constraint violation
	gamma	1	required parameter for radial kernels
BART	num_trees	10	number of trees to grow**
	num_burn_in	20	number of samples to be discarded as 'burn-in'
	q	0.99	quantile of the prior
	k	1	determines the prior probability
Null	—	—	—

*The names of the tuning parameters are specific to the packages used in R and may be different in other programming languages or libraries.
**BART is computationally expensive, so this value had to be constrained due to memory limitations.

Table A.3.: Results from the statistical tests between the actual and projected reservoir volume (water balance method). The Kolmogorov-Smirnov test evaluates the difference between the distributions of the data and the t-test evaluates the difference in means. In both tests,a p-value less than 0.01 indicates there is a statistically significant difference in the distribution or mean, depending on the test. Note that there is no data for Lake Travis, O'Shaughnessy Reservoir, or Hoover Reservoir, due to lack of data.

Reservoir	Kolmogorov-Smirnov test		Welch's t-test	
	D statistic	p-value	t statistic	p-value
Chester Morse Lake	0.341	1.18×10^{-6}	4.432	1.46×10^{-5}
South Fork Tolt	0.642	$< 2.2\times10^{-16}$	-11.97	$< 2.2\times10^{-16}$
Falls Lake	0.943	$< 2.2\times10^{-16}$	-12.96	$< 2.2\times10^{-16}$
Lake Mead	0.634	$< 2.2\times10^{-16}$	11.21	$< 2.2\times10^{-16}$
Lake Travis	—	—	—	—
O'Shaughnessy	—	—	—	—
Hoover Reservoir	—	—	—	—
Lake Hefner	0.797	$< 2.2\times10^{-16}$	13.45	$< 2.2\times10^{-16}$
Eagle Creek	0.992	$< 2.2\times10^{-16}$	-18.15	$< 2.2\times10^{-16}$

Table A.4.: Results from the statistical tests between the actual and projected reservoir volume (random forest method). The Kolmogorov-Smirnov test evaluates the difference between the distributions of the data and the t-test evaluates the difference in means. In both tests,a p-value less than 0.01 indicates there is a statistically significant difference in the distribution or mean, depending on the test. Note that there is no data for Lake Travis, O'Shaughnessy Reservoir, or Hoover Reservoir, due to lack of data.

Reservoir	Kolmogorov-Smirnov test		Welch's t-test	
	D statistic	p-value	t statistic	p-value
Chester Morse Lake	0.423	5.67×10^{-10}	6.42	7.02×10^{-10}
South Fork Tolt	0.301	2.93×10^{5}	5.6	6.38×10^{-8}
Falls Lake	0.22	0.00533	-0.997	0.3206
Lake Mead	0.992	$< 2.2\times10^{-16}$	23.67	$< 2.2\times10^{-16}$
Lake Travis	—	—	—	—
O'Shaughnessy	—	—	—	—
Hoover Reservoir	—	—	—	—
Lake Hefner	0.366	1.42×10^{-7}	-1.08	0.2816
Eagle Creek	0.602	$< 2.2\times10^{-16}$	-12.84	$< 2.2\times10^{-16}$

A.3 Figures

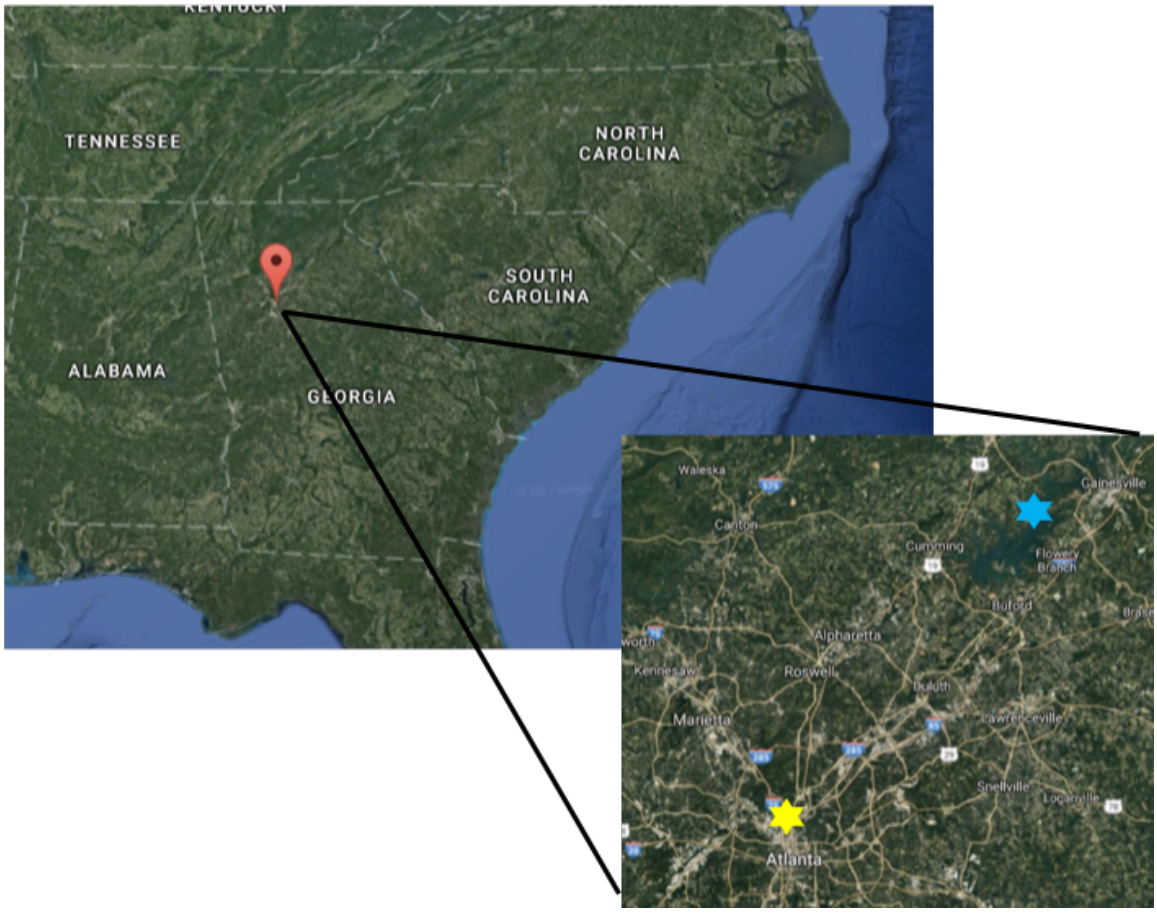

Fig. A.1.: Map showing the location of the city of Atlanta and Lake Sidney Lanier. The yellow star indicates the city and the blue star indicates the reservoir. Imagery: Landsat/Copernicus. Map data: Google.

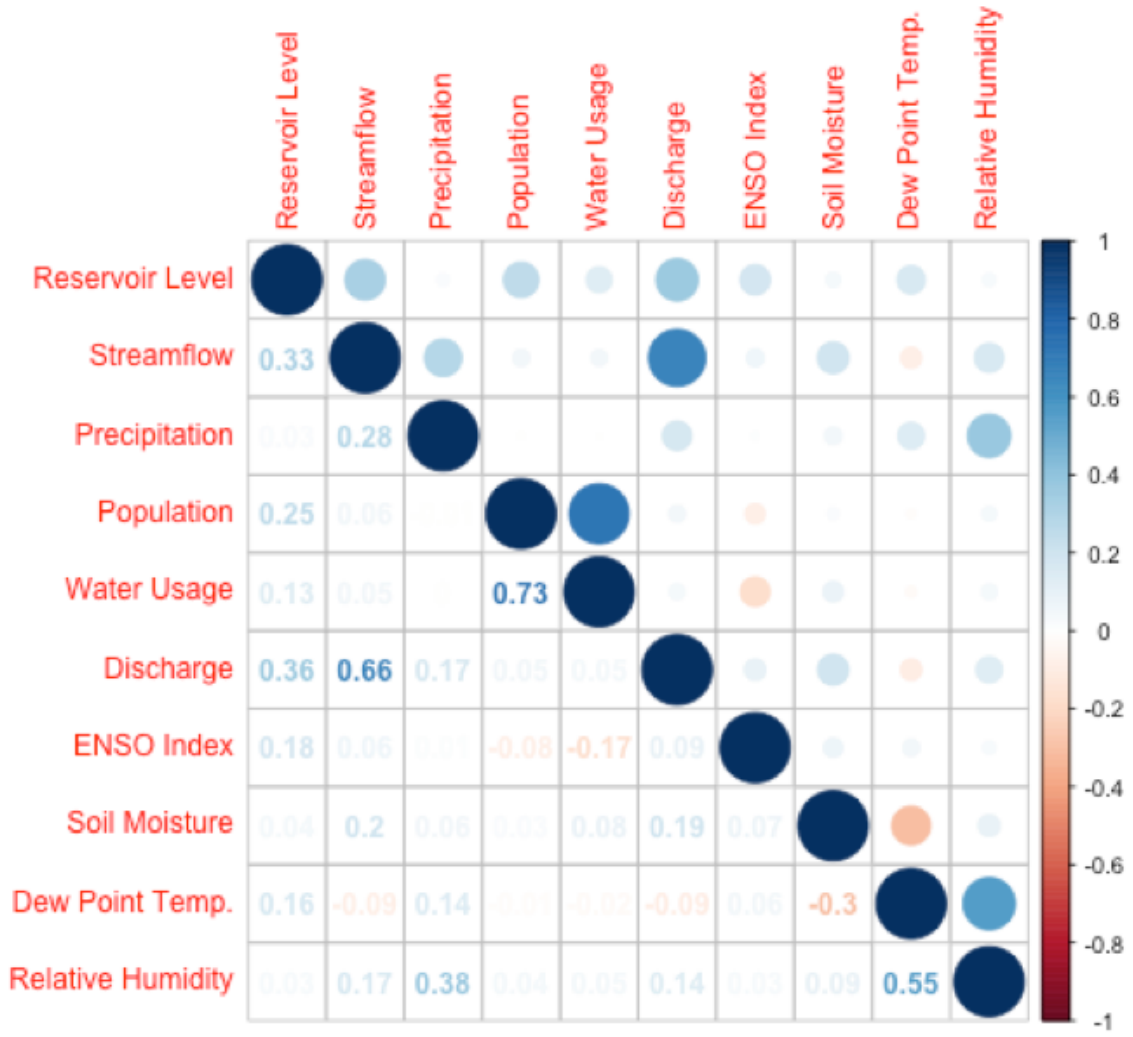

Fig. A.2.: Correlation matrix of variables used in study.

B. SUPPLEMENTARY INFORMATION FROM CHAPTER 3

Parts of this appendix have been previously published as supplementary material in *Applied Energy*: https://doi.org/10.1016/j.apenergy.2019.113466 and *Climatic Change*: https://doi.org/10.1007/s10584-020-02669-7.

Contents of this appendix include:

B.1 Methods

B.1.1 Removing the Seasonality

In order to remove the seasonality from the response variable dataset, we followed a common time series decomposition method. Specifically, we chose to decompose the time series based on rates of change. This method breaks a given time series into four main components: the trend component, the cyclical component, the seasonal component, and the irregular component [**?**]. For the purposes of this study, we were interested in removing the seasonal component, as it has been shown that seasonality improves the apparent predictive accuracy of models, which may be a misrepresentation of the true predictive accuracy.

To adjust for the seasonality in the dataset, we first determined that there was a significant seasonal component in the response data. This was done primarily through a spectral density analysis and visualized through periodograms (see Figure B.1). In this analysis, one finds the spike in spectral density and determines the corresponding frequency. The period of seasonality is then calculated as $T = 1/\omega$, where T is the period and ω is the frequency.

Once we determined that there was seasonality, we used the time series decomposition to isolate the seasonal component of the dataset. The isolated seasonal component was then subtracted from the fully composed dataset to create a new, seasonally adjusted dataset. This method follows a typical seasonality adjustment for additive time series [**?**].

B.1.2 Trend Adjustment

The response data (i.e., the electricity and water demand) was de-trended following the methodology described by Sailor and Muñoz (1997). Below is the process followed to de-trend the data.

1. Calculate the yearly average of the monthly data for the entire period of study

$$\bar{E} = \sum_{y=2001}^{2017} \sum_{m=1}^{12} E(m, y) \tag{B.1}$$

2. Calculate the adjustment factor

$$F_{adj}(y) = \frac{1}{\bar{E}} (\sum_{m=1}^{12} E(m, y)) \tag{B.2}$$

3. Calculate the trend adjusted data

$$E_{adj}(m, y) = \frac{E(m, y)}{F_{adj}(y)} \tag{B.3}$$

B.2 Figures

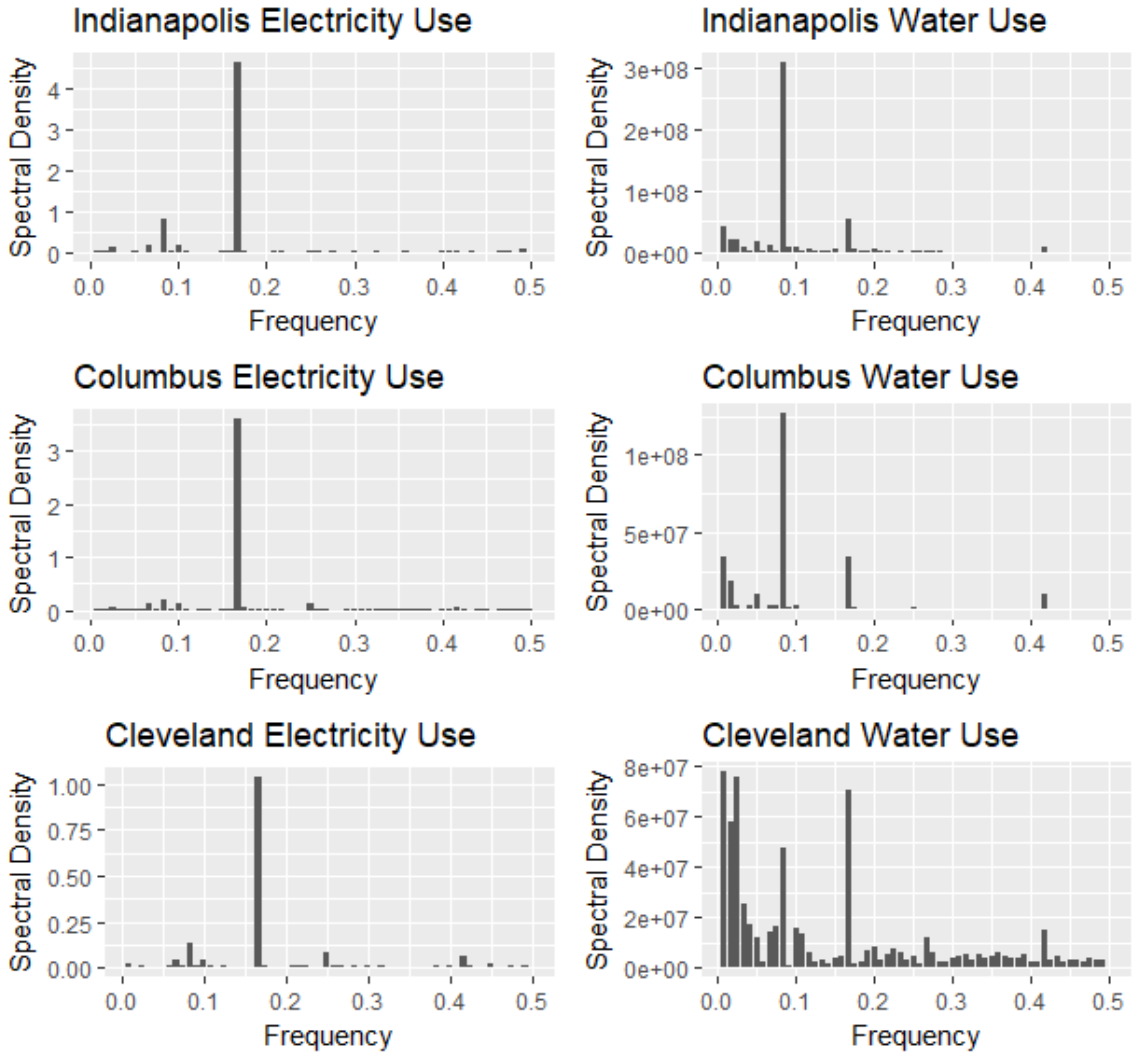

Fig. B.1.: Periodograms of a selection of the cities analyzed in this study. In these periodograms, the lone peaks demonstrate that seasonality is present at that frequency. Note that Cleveland has no apparent seasonality for water use.

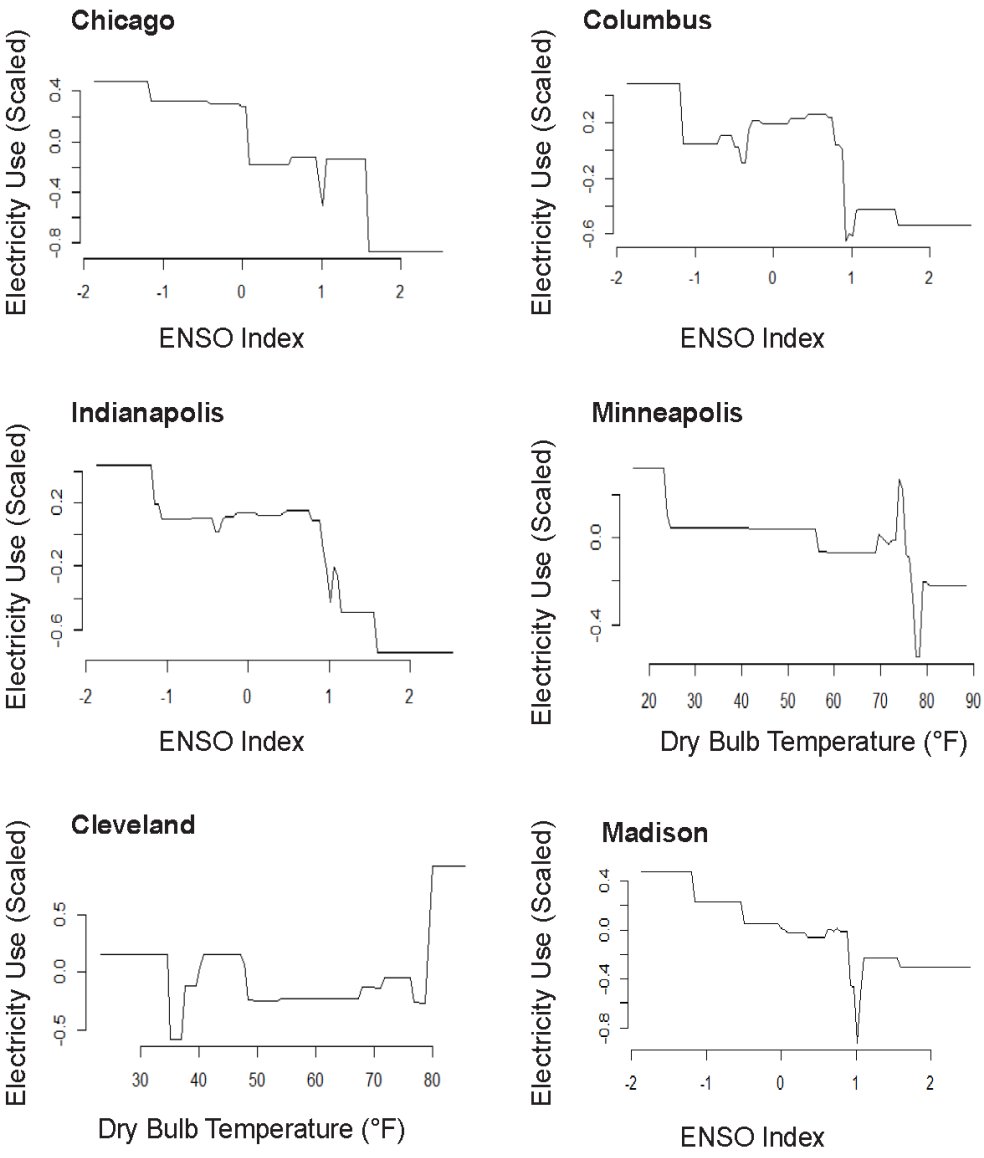

Fig. B.2.: Partial dependence between the electricity use in each city and the most important predictor variable.

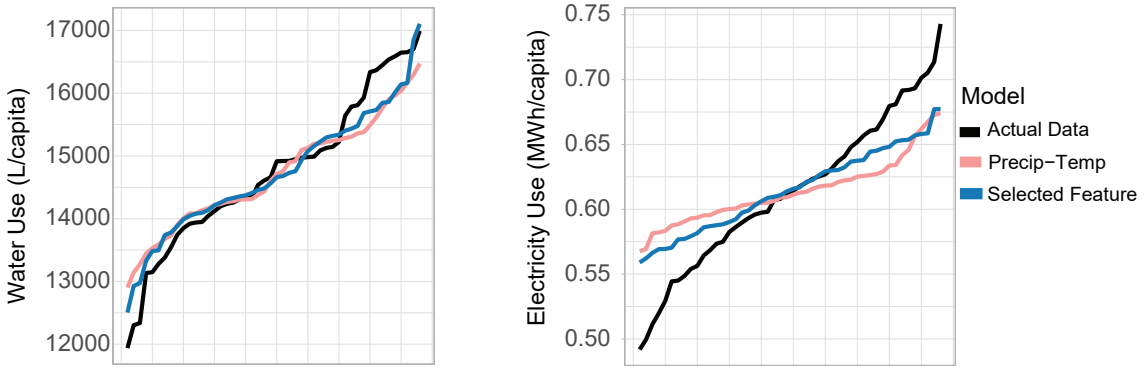

Fig. B.3.: Observational data compared to the model results for the intermediate time period (April, May, October, November). 'Precip-Temp' represents the baseline model that only considered precipitation and dry bulb temperature, while the 'Selected Feature' model was the model built for this study with a wider array of climate variables.

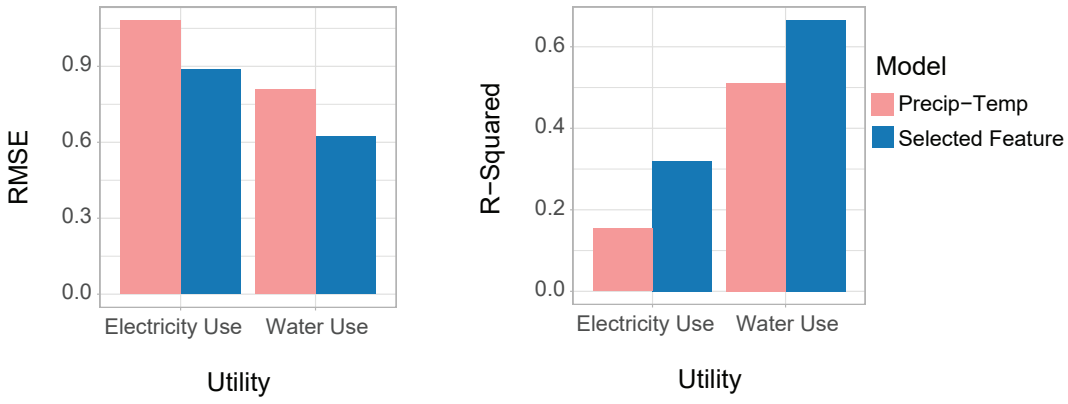

Fig. B.4.: Model performance results for the intermediate months.

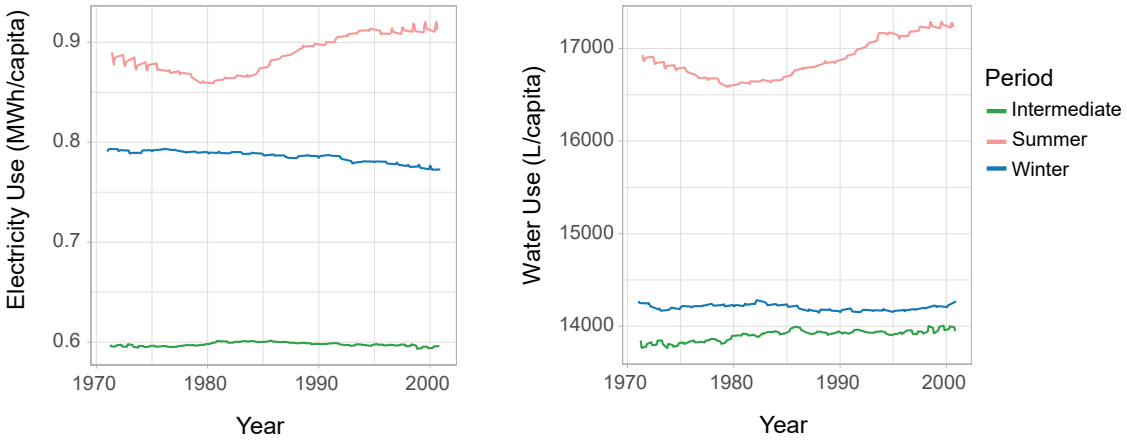

Fig. B.5.: Historical data (1971-2000) used as the baseline for the future projection analysis. The summer period included June-September, the winter period included December-March, and the intermediate period included the remaining months.

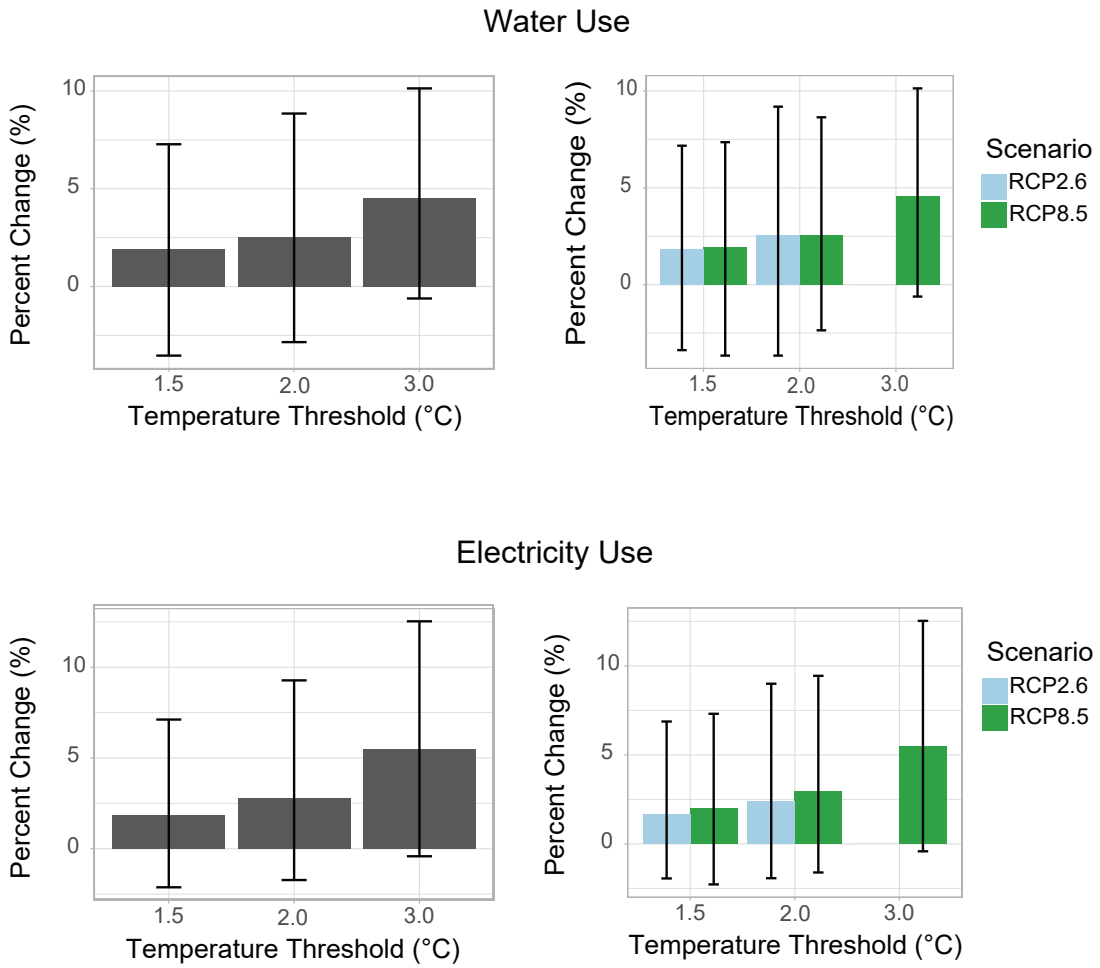

Fig. B.6.: Median relative change in water and electricity use for the intermediate period after three key temperature thresholds. The left panels represent the projections of all 10 GCM-RCP combinations (5 GCMS × 2 RCPs), while the right panels separated the the two RCP scenarios.

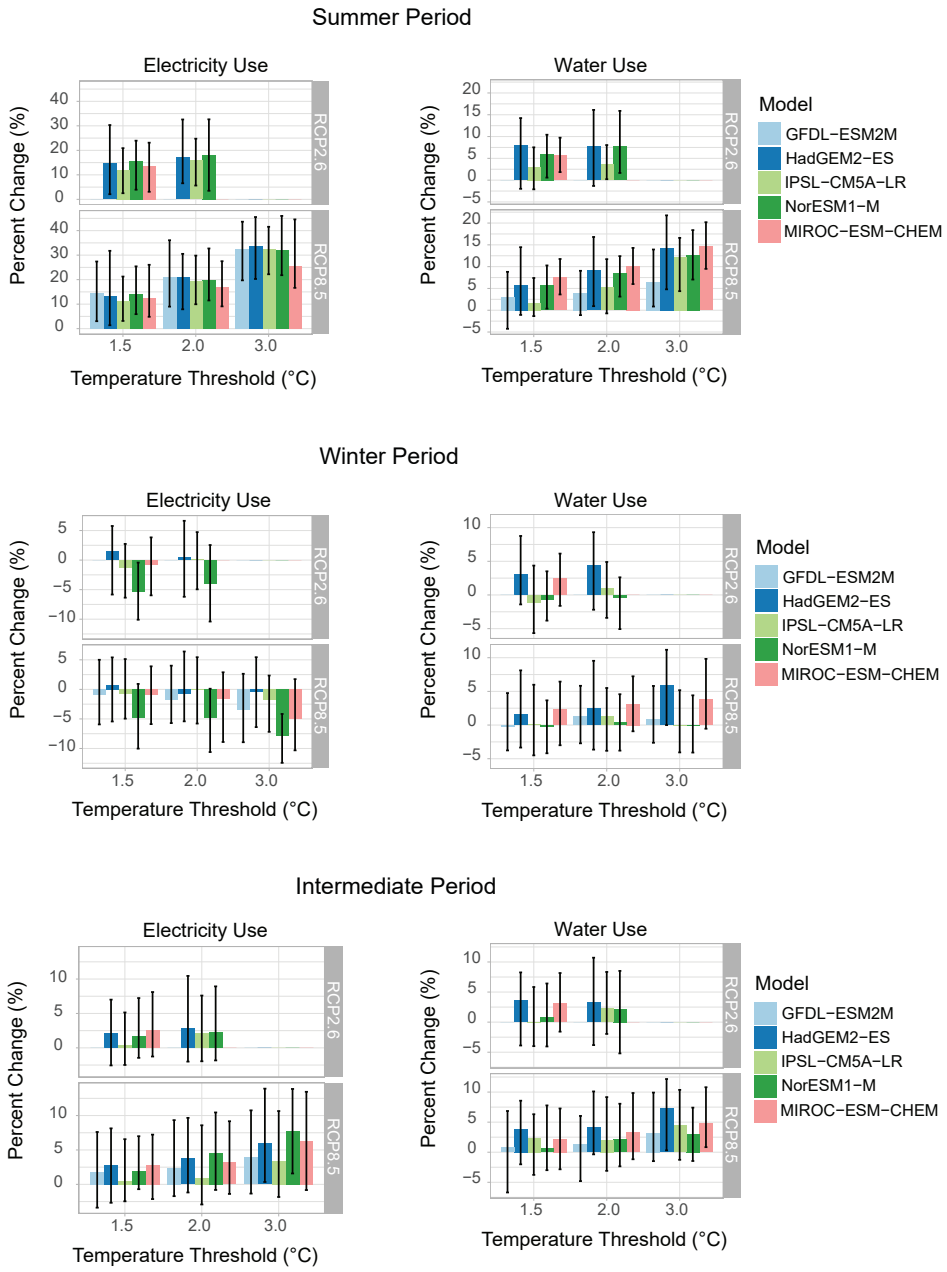

Fig. B.7.: Median relative change for water and electricity use in all three periods after three key temperature thresholds. Each GCM and RCP scenario has been separated, so that in each panel, RCP2.6 is on top and RCP8.5 is on the bottom.

C. SUPPLEMENTARY INFORMATION FOR CHAPTER 4

Contents of this appendix include:

C.1 Methods

C.1.1 Interview Protocol

Interview Protocol for the Project Titled The Influence of Norms on the Water-Electricity Demand Nexus: A Case Study in Indianapolis, Indiana

Interviewer:

Interviewee (first name only):

Interview Date/Time:

Interview Location:

Introduction

I am very grateful that you are taking time out of your day to do an interview with me. Thank you in advance. My name is Renee Obringer and I am a PhD student at Purdue University in the Division of Environmental and Ecological Engineering. I am working with Dr. Zhao Ma, a professor in the Department of Forestry and Natural Resources, and Dr. Roshanak Nateghi, a professor in the School of Industrial Engineering and the Division of Environmental and Ecological Engineering. Our research goal is to evaluate the effect of social norms on urban water and electricity use at the neighborhood level. During this interview, I would like to ask you a series of questions covering three main topics: your awareness of local water or electricity conservation programs, your personal beliefs about water or electricity conservation, and your perceptions of others beliefs about water and electricity conservation. This interview is entirely voluntary and should take about 60 minutes. Everything you tell me will be kept confidential and your name will not be revealed to anyone beyond the research team, that is myself, Dr. Ma, and Dr. Nateghi. For the purpose of data analysis, it will be helpful for me to record this conversation. Do you feel comfortable with this?

Again, thank you for participating in this interview. Unless you have any questions, we can go ahead and begin.

Section 1: Awareness of water and/or electricity conservation programs

To begin, I would like to ask you a few questions about your awareness of water and/or electricity conservation.

1. Have you heard of any programs offered by the utility company, city, state or other entity that encourage people to reduce their water and/or electricity use? Could you please describe these programs for me?
    - Prompt: If your water is provided by Citizens Energy, are you aware of the Be Water Wise campaign? If yes, have you participated in any of the water saving measures?
    - Prompt: If your electricity is provided by Indianapolis Power and Light, are you aware of the Ways to Save campaign? If yes, have you participated in any of the electricity savings measures, including the eScore home assessment, the PowerView, and the Heating and Cooling Rebates?
2. Have you heard of any initiatives specific to your neighborhood that involve water or electricity conservation?
    - Prompt: Has anyone brought up initiatives or conservation measures at your regular meetings?
    - Prompt: Has there been any interest in neighborhood water and electricity conservation in the past?

Section 2: Personal beliefs regarding water and/or electricity conservation

The next set of questions will focus on your personal habits and beliefs about water and electricity conservation.

1. Could you tell me about how you use water and electricity in and around your home, that is inside your home as well as any landscaping or outdoor activities that require water or electricity?
    - Prompt: Do you have an air conditioning system? Do you run it often? Is it controlled by a programmable, or smart, thermostat?
    - Prompt: Do you have a yard or garden that you maintain? How intensive would you say your outdoor maintenance routine is, with regard to water use?
    - Prompt: Do you have any efficient appliances? These may include energy star appliances, LED lightbulbs, or low-flow faucets.
    - Prompt: Can you think of any recurring instances that use a large amount of water electricity in your home? For example, do you regularly wash your car or take long showers? Do you keep your air conditioner running all day?
2. Could you describe the general bill-paying process in your place of residence?
    - Prompt: Do you pay for water and/or electricity directly?
    - Prompt: Do you participate in paper or paperless billing?
    - Prompt: Do you pay attention to the price and/or usage information on your bill?
    - Prompt: Are there other aspects of the bill that you pay attention to? Why?
3. Could you tell me about how you think about water and electricity conservation?
    - Prompt: Is it something that you think of regularly? Why or why not?
    - Prompt: When you think of it, is it in an economic context or an environmental context? Some other context? Could you please describe?
4. Can you think of a situation that would lead you to reduce your water and/or electricity use?
    - Prompt: What if water and/or electricity prices increased?
    - Prompt: What if there is an incentive program that offers rebates or reduced process for high efficiency or low flow appliances.

- Prompt: What if the city mandates limited water use for landscaping or car-washing?
- Prompt: What if the city or state was experiencing a major drought?

Section 3: Perceptions of others beliefs regarding water and/or electricity conservation

This final section will focus on your perceptions of the beliefs of your fellow residents of X neighborhood.

1. Do you think your friends and neighbors think about water and electricity conservation in a similar way that you do?
   - Prompt: If not, why not? What do you think your friends and neighbors think of water and electricity conservation? Do they think about it at all?
   - Prompt: Have you ever had a discussion about the amount of water or electricity you or others use, or the price of water or electricity, with your friends or neighbors? Can you describe those conversations? If you have not had them, why not?
2. Do you think your friends and neighbors in your area are doing anything related to water or electricity conservation?
   - Prompt: If not, why?
   - Prompt: If yes, what do you think they do in and around their home to conserve water or electricity?
   - Prompt: If unsure, why do you say that? (it is private matter, they dont pay attention, nothing visible to outsiders, etc.)
3. Do you expect others, that is your friends, neighbors, or people in your neighborhood, to conserve water or electricity?
   - Prompt: Who and why?
   - Prompt: Do you tend to associate with those people? Why or why not?

4. Do you feel others, that is your friends, neighbors, or people in your neighborhood, expect you to conserve water or electricity?
    - Prompt: Who and why?
    - Prompt: Do you often give into that expectation?
5. How would you react or feel if you found out that others, that is your friends, neighbors, or people in your neighborhood, were actively conserving water?
    - Prompt: Why and because of whom?
    - Prompt: Is it important for you to fit in? Is it important that you live in a neighborhood that shares your personal values?

That is all of the questions I have for you, but before we end, is there anything else you would like to share about water and/or electricity use in your home, among your friends and neighbors, and/or in your neighborhood.

Thank you very much for your time.

C.2 Figures

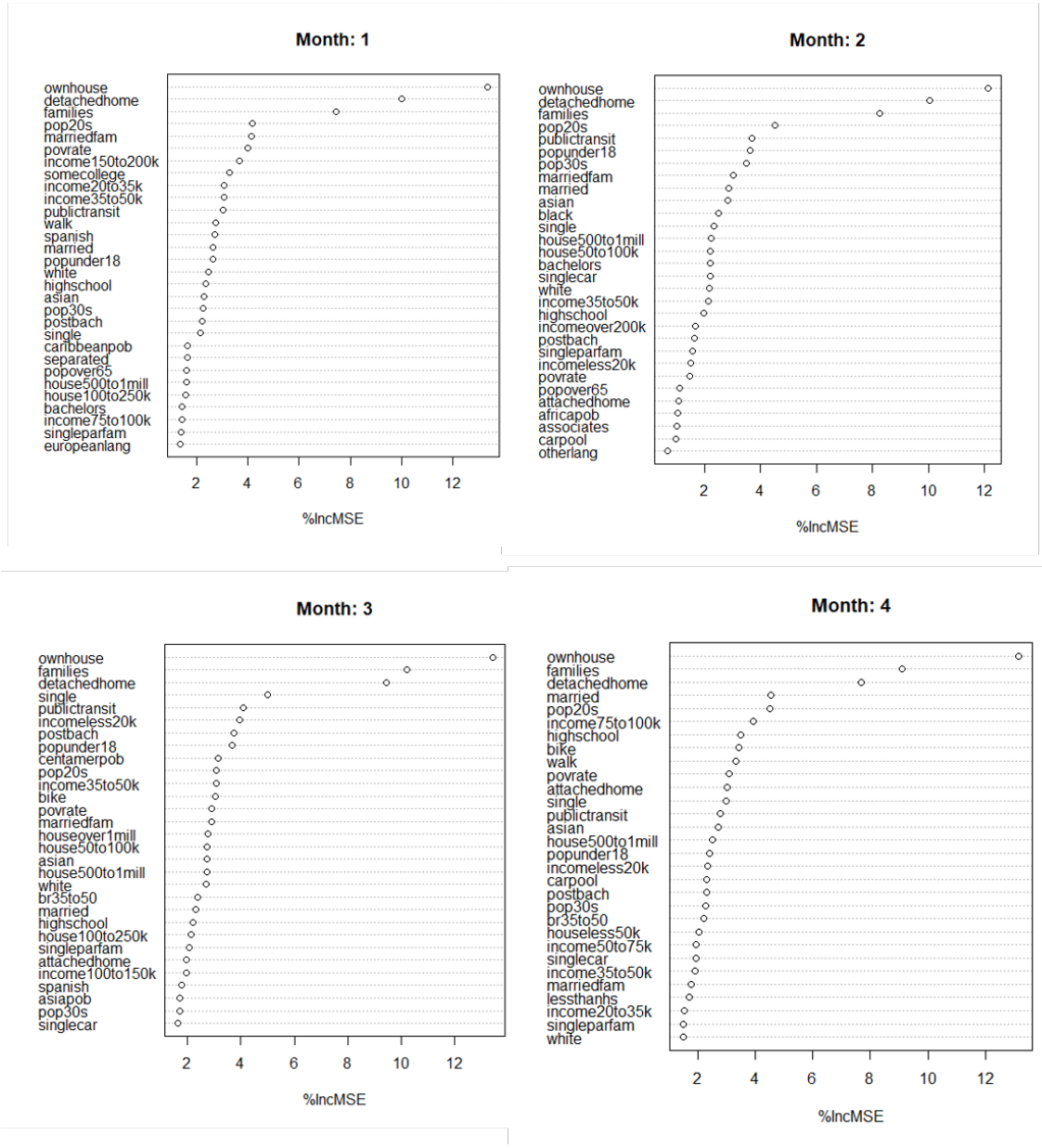

Fig. C.1.: Important variables in the demographics-only analysis of water consumption in January, February, March, and April.

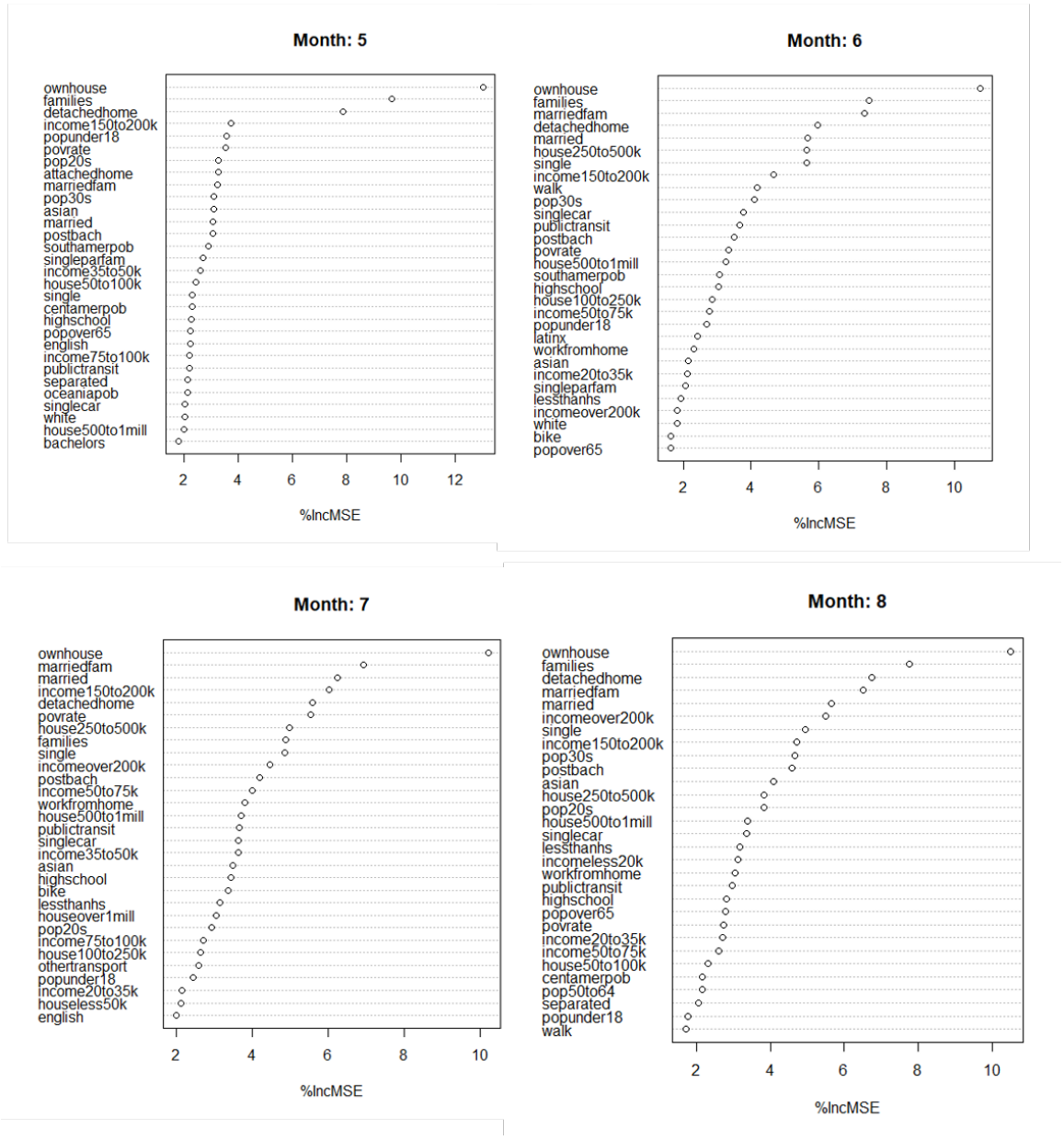

Fig. C.2.: Important variables in the demographics-only analysis of water consumption in May, June, July, and August.

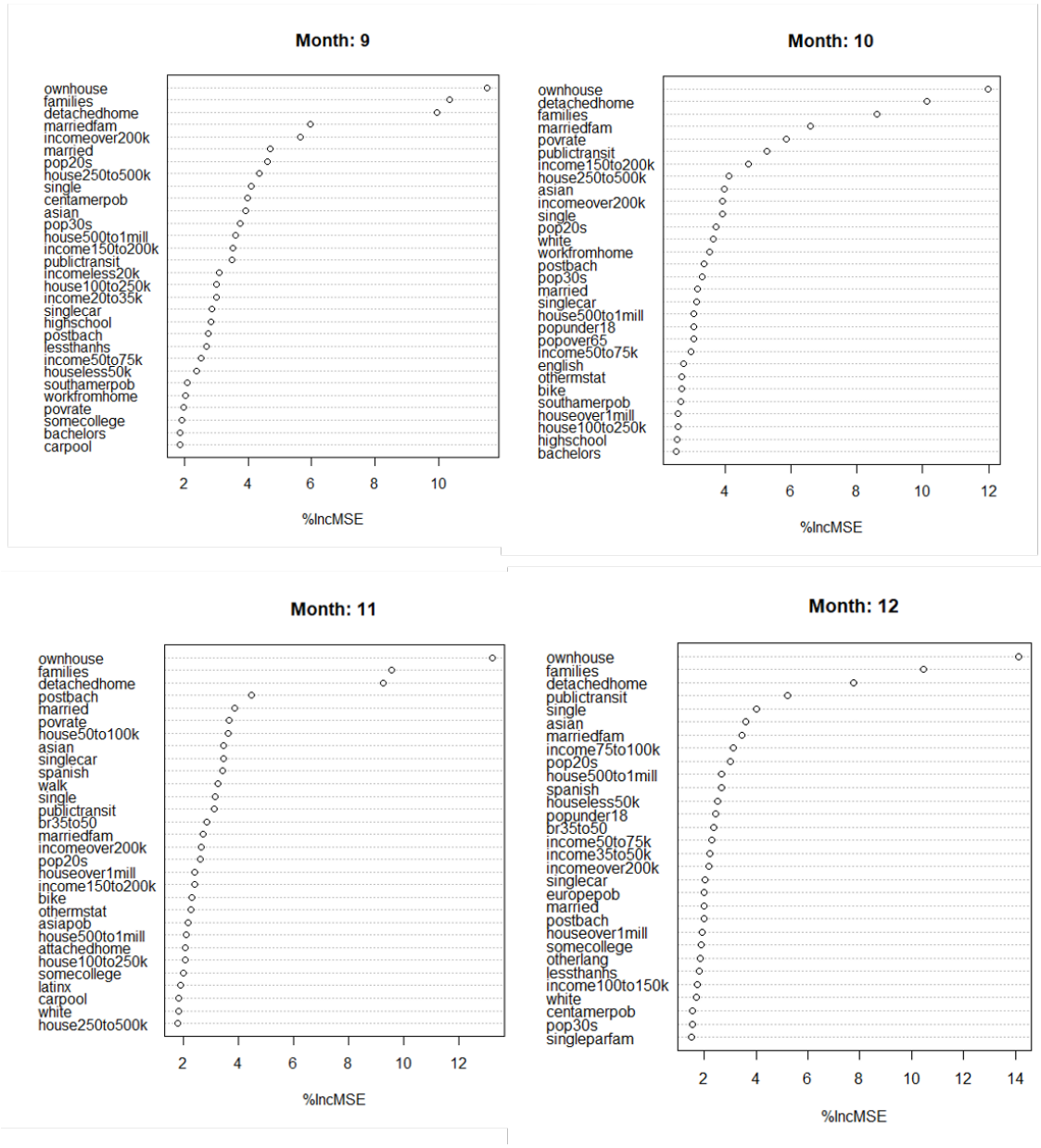

Fig. C.3.: Important variables in the demographics-only analysis of water consumption in September, October, November, and December.

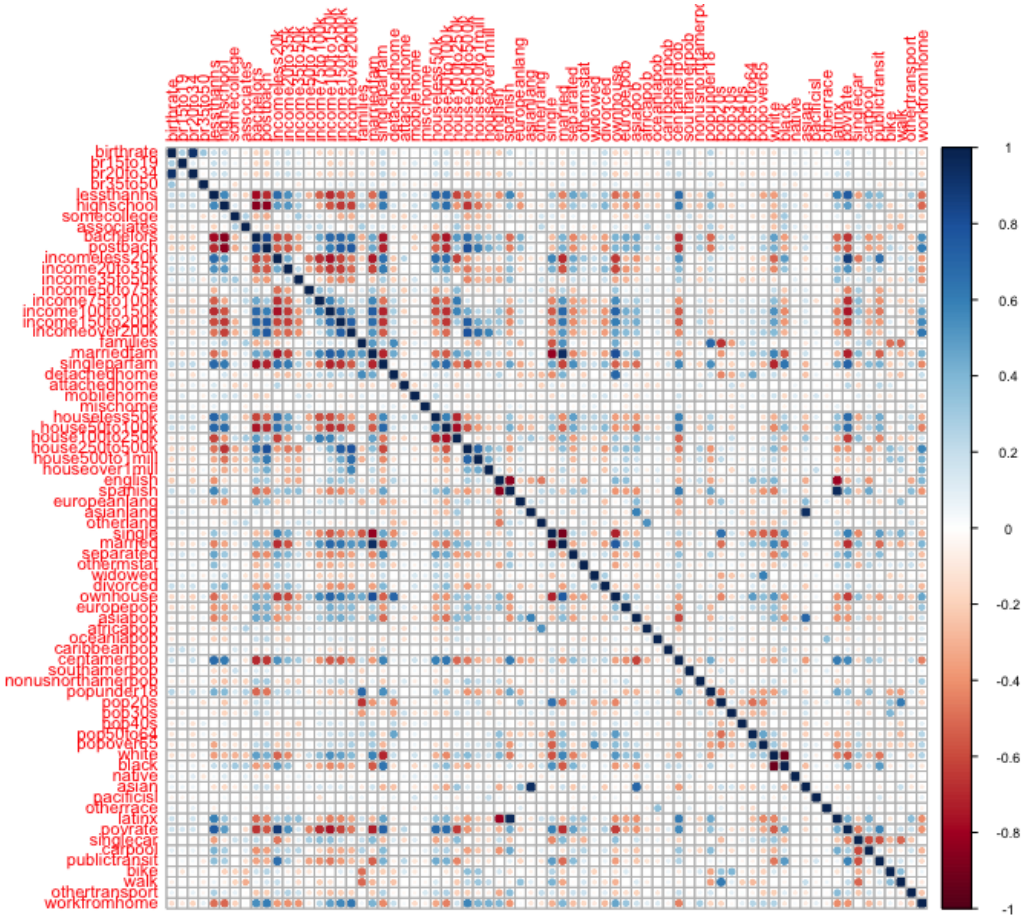

Fig. C.4.: Correlation plot of the predictor variables prior to variable selection.

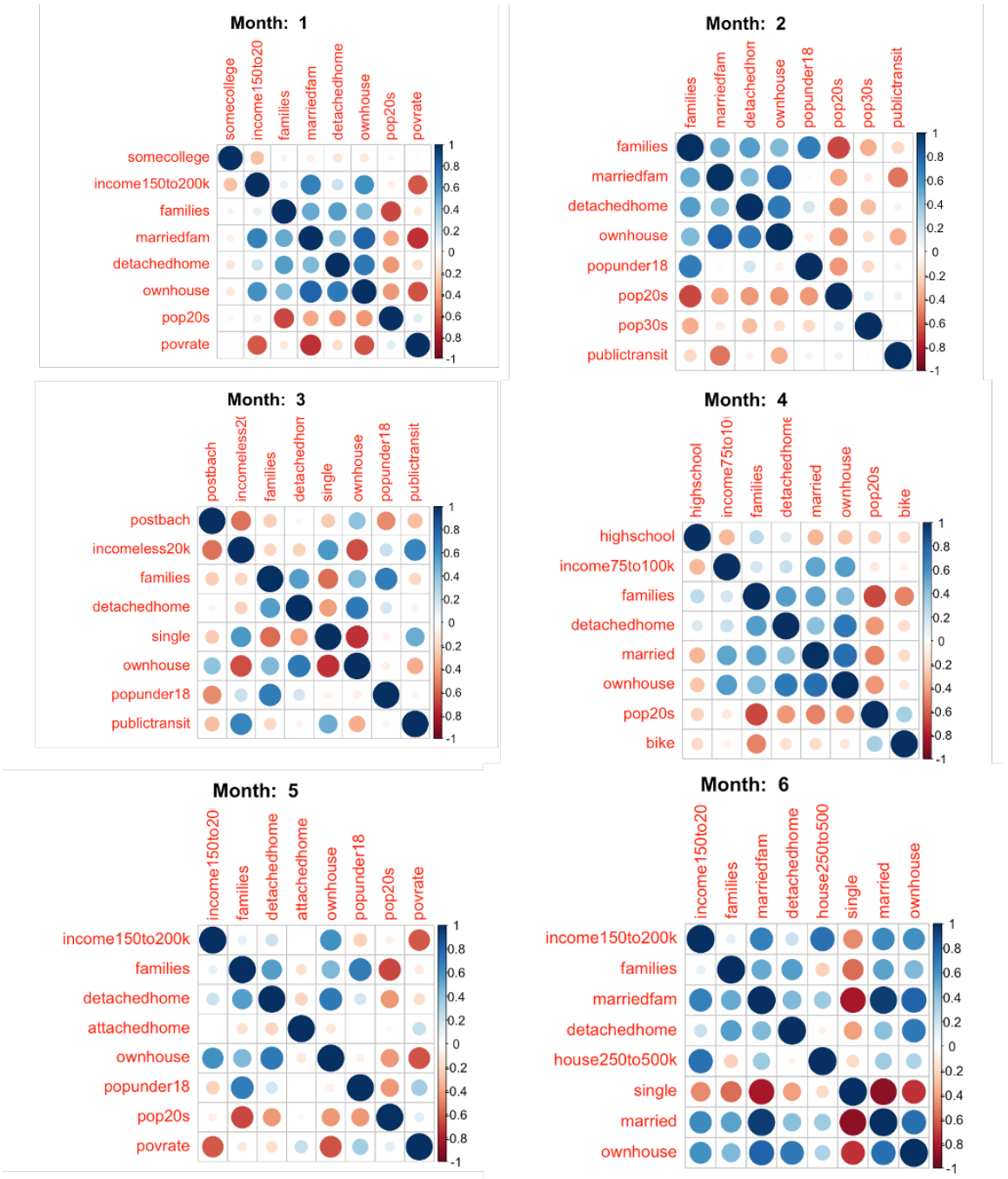

Fig. C.5.: Correlation plots of the predictor variables selected for January-June.

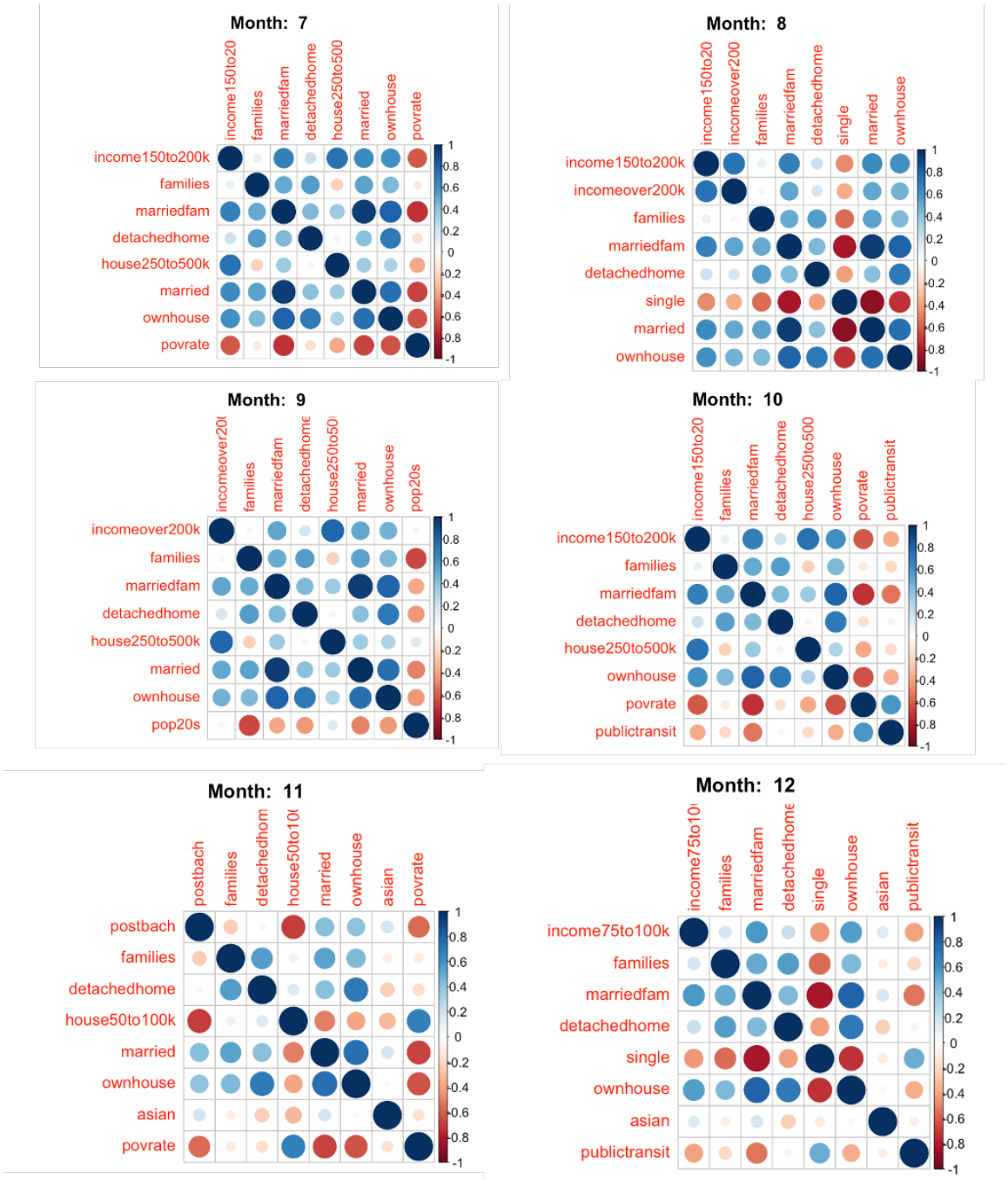

Fig. C.6.: Correlation plots of the predictor variables selected for July-December.

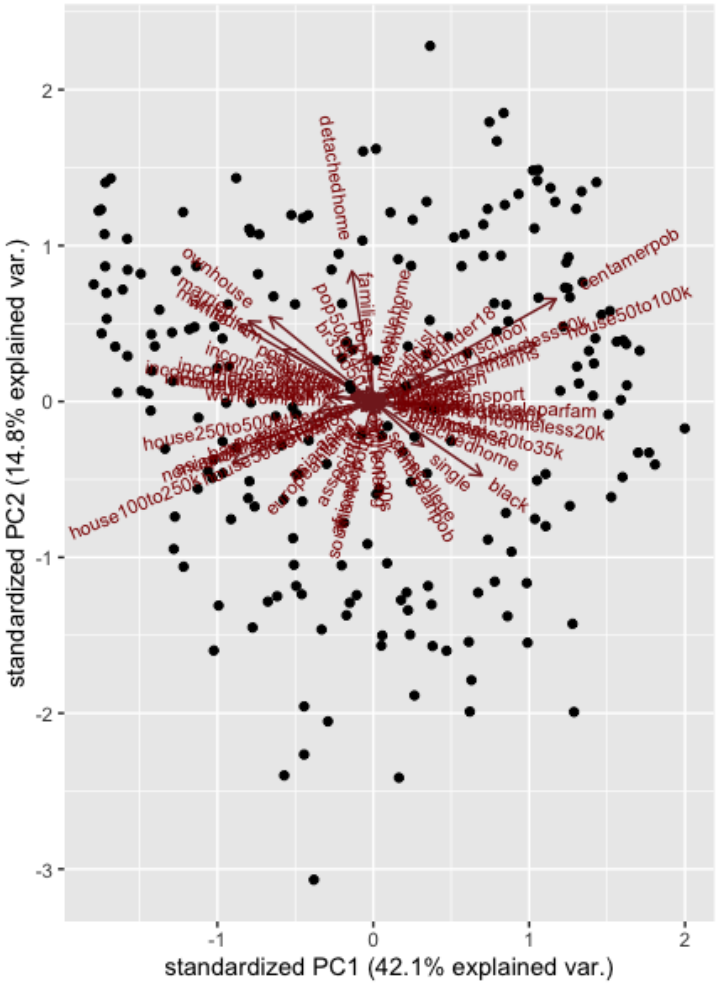

Fig. C.7.: Biplot showing the results of the principal component analysis.

Fig. C.8.: Actual and predicted water consumption in January, February, and March.

Fig. C.9.: Actual and predicted water consumption in April, May, and June.

Fig. C.10.: Actual and predicted water consumption in July, August, and September.

Fig. C.11.: Actual and predicted water consumption in October, November, and December.

VITA

VITA

Renee Obringer obtained her PhD in Environmental and Ecological Engineering from Purdue University. She was also a member of the Ecological Science and Engineering Interdisciplinary Graduate Program at Purdue, as well as a student affiliate of the Purdue Climate Change Research Center. Prior to attending Purdue, Renee obtained her B.S. in Environmental Engineering from Ohio State University (2015). Her research interests focus on understanding and evaluating the impact of climate change on urban systems, with an emphasis on water and electricity. More broadly, Renee harnesses methods from data science, climatology, and social science to study the nexus between climate change, people, and urban systems. Throughout her PhD, Renee published seven peer-reviewed articles in top journals, including Applied Energy, Climatic Change, and Scientific Reports. In addition to her publications, Renee has presented at a variety of conferences, including the Society for Risk Analysis annual meeting, INFORMS, the Institute for Industrial and Systems Engineers annual meeting, and the Behavior, Energy, and Climate Change conference. She has received a number of awards, including the Society for Risk Analysis Student Merit Award (2018), the College of Engineering Outstanding Graduate Student Award (2019), and the College of Engineering Outstanding Service Award (2020). Throughout her time at Purdue, she has been funded by the Andrews Fellowship and Bilsland Dissertation Fellowship as well as grants from NSF and the Purdue Center for the Environment.

An updated publication list can be found on her Google Scholar page: `https://scholar.google.com/citations?user=_iJ9gwwAAAAJ&hl=en`