5-2022

Advancing Data Collection, Management, and Analysis for Quantifying Residential Water Use via Low Cost, Open Source, Smart Metering Infrastructure

Camilo J. Bastidas Pacheco
Utah State University

ADVANCING DATA COLLECTION, MANAGEMENT, AND ANALYSIS FOR QUANTIFYING RESIDENTIAL WATER USE VIA LOW COST, OPEN SOURCE, SMART METERING INFRASTRUCTURE

by

Camilo J. Bastidas Pacheco

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

DOCTOR OF PHILOSOPHY

in

Civil and Environmental Engineering

Approved:

_____	_____
Jeffery S. Horsburgh, Ph.D. Major Professor	David Rosenberg, Ph.D. Committee Member
_____	_____
David K. Stevens, Ph.D. Committee Member	Ruijie Zeng, Ph.D. Committee Member
_____	_____
Sarah Null, Ph.D. Committee Member	D. Richard Cutler, Ph.D. Interim Vice Provost of Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah

2022

ABSTRACT

Advancing Data Collection, Management, and Analysis for Quantifying Residential Water Use Via Low Cost, Open Source, Smart Metering Infrastructure

by

Camilo J. Bastidas Pacheco

Utah State University, 2022

Major Professor: Dr. Jeffery S. Horsburgh
Department: Civil and Environmental Engineering

Collecting and managing high temporal resolution (< 1 minute) residential water use data is challenging due to cost and technical requirements associated with the volume and velocity of data collected. This type of data has potential to expand our knowledge of residential water use and improve water management. Most studies collecting this type of data have been focused on applications of the data (e.g., developing and applying end use disaggregation algorithms) with much less focus on the cyberinfrastructure, or methods, used to collect and manage the data. The research in this dissertation is an investigation of open tools and systems to automate the process from high temporal resolution residential water use data collection to analysis, as well as residential water use practices and variability in Logan and Providence, Utah. Emphasis was placed on making the tools low cost, open source, and available to the public, so they can be reused, modified, improved, or used as a basis for future developments. Additionally, all data collected are publicly available. The principal outcomes of this work include new hardware and software for measuring and processing high temporal resolution water use data. New dataloggers were developed that collect data on top of, and without disrupting, existing

water meters. Software was developed for automating data transmission, management, archival, and analysis. Performance testing demonstrated scalability of the cyberinfrastructure to multiple hundreds of data collection devices. Using the hardware and software developed by this research, residential water use data was collected over a period of three years at 31 residential homes. In examining the data, we found significant temporal variability in indoor water use volume and timing and in the distribution of ends uses. Despite the fact that outdoor water use was the largest component of residential water use, we found that users were not significantly overwatering their landscapes. Opportunities for water conservation indoors and outdoors through adoption of more efficient fixtures and promoting conservation behaviors were identified.

(231 pages)

PUBLIC ABSTRACT

Advancing Data Collection, Management, and Analysis for Quantifying Residential Water Use Via Low Cost, Open Source, Smart Metering Infrastructure

Camilo J. Bastidas Pacheco

Urbanization, climate change, aging infrastructure, and the cost of delivering water to residential customers make it vital that we achieve a higher efficiency in the management of urban water resources. Understanding how water is used at the household level is vital for this objective. Water meters measure water use for billing purposes, commonly at a monthly, or coarser temporal resolutions. This is insufficient to understand where water is used (i.e., the distribution of water use across different fixtures like toilets, showers, outdoor irrigation), when water is used (i.e., identifying peaks of consumption, instantaneous or at hourly, daily, weekly intervals), the efficiency of water using fixtures, or water use behaviors across different households. Most smart meters available today are not capable of collecting data at the temporal resolutions needed to fully characterize residential water use, and managing this data represents a challenge given the rapidly increasing volume of data generated. The research in this dissertation presents low cost, open source cyberinfrastructure (datalogging and data management systems) to collect and manage high temporal resolution, residential water use data. Performance testing of the cyberinfrastructure demonstrated the scalability of the system to multiple hundreds of simultaneous data collection devices. Using this cyberinfrastructure, we conducted a case study application in the cities of Logan and Providence, Utah where we found significant variability in the temporal distribution,

timing, and volumes of indoor water use. This variability can impact the design of water conservation programs, estimations and forecast of water demand, and sizing of future water infrastructure. Outdoor water use was the largest component of residential water use, yet homeowners were not significantly overwatering their landscapes. Opportunities to improve the efficiency of water using fixtures and to conserve water by promoting behavior changes exist among participants.

ACKNOWLEDGMENTS

I am indebted to my advisor, Dr. Jeff Horsburgh, for his dedication and guidance throughout all these years. I will be forever grateful for his patience, the time he made available to me, for reviewing draft after draft, and for his high academic standards. I must also thank my committee members Dr. David Rosenberg, Dr. Sarah Null, Dr. David Stevens, and Dr. Ruijie Zeng for their guidance and suggestions. I would also like to thank those who contributed as coauthors on chapters of this dissertation.

I want to acknowledge the support from other colleagues, including: Nour Attallah, Josh Tracy, A.J., Juan Caraballo, and Amber Jones. I thank those who assisted during data collection, reproduced scripts, answered questions, helped me in any way, the residents who participated in our residential water use study, and Logan and Providence city. I would like to express my gratitude to the Utah Water Research Laboratory and the U.S. National Science Foundation (Grant CBET 1552444 - Cyberinfrastructure for Intelligent Water Supply (CIWS): Shrinking Big Data for Sustainable Urban Water) for the financial support provided during my studies. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Finally, I would like to thank my family members and friends. To my parents and brother, thanks for your support. To Angela, thanks for always being there. This work is dedicated to Gabriel, who is my greatest joy.

Camilo J. Bastidas Pacheco

CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

According to the United States Geological Survey, residential water use in Utah is the second largest in the United States, with an average of 169 gallons per capita per day (GPCD) (Dieter et al., 2018). The State of Utah Division of Water Resources (DWR) recently released a similar estimate of 168 GPCD for average residential water use in Utah during 2015 (Utah DWR 2020). Approximately 91% of Utah's population was living in urban areas in 2010 (The University of Utah, 2016), and since then urban populations have grown much faster than those in rural areas in the U.S. and globally (EPA, 2016; UN-Habitat, 2016). It has been estimated that Utah will need a $4.4 billion investment over a 20-year period to maintain the current level of service and meet the future water demands of its growing population (EPA, 2018). The increase in urban population density, the cost of delivering water to urban populations, and the variability in water resources availability related to climate change make it vital that we understand and efficiently manage urban water use.

Our ability to understand water use is limited by the temporal resolution of the data most commonly collected. In the U.S., metering of residential water use is common practice, and water meters are typically read monthly or quarterly for billing purposes. Monthly or coarser temporal resolution data are inadequate for understanding how water is used at the household level (Cole and Stewart, 2013; Gurung et al., 2015). Recently, "smart metering" devices have enabled data collection at higher temporal resolutions (Cominola et al., 2015). The term "smart meter" has been used to describe multiple different applications given the availability of different data collection technologies

(Boyle et al., 2013). In this dissertation, “smart meter” is used to denote devices capable of collecting high temporal resolution data (i.e., high sampling frequency) that can be integrated in efficient systems for data management (Cominola et al., 2015). Smart meters have potential to address existing gaps in residential water use knowledge (i.e., estimating peak timing, separating indoor versus outdoor water use, and identifying the distribution of water use across end uses) (Boyle et al., 2013; Cominola et al., 2018).

Water end use information, which can be derived from smart meting data, has potential to improve the accuracy of estimates of water demand price elasticity (Marzano et al., 2018); generate insights from observed trends in residential water use, which is not possible with the data available (Rockaway et al., 2011); assist in the design of water awareness campaigns (Abdallah and Rosenberg, 2014; Willis et al., 2010); improve existing campaigns to upgrade inefficient fixtures (Mayer et al., 2004; Suero et al., 2012); and is an essential input for water planning and management (Giurco et al., 2008). Most water use events at residential properties last on the order of seconds to minutes, and data at this, or finer, temporal resolutions is needed to quantify them (Nguyen et al., 2015). Key parameters for identifying events (e.g., duration, flow rate) are sensitive to the temporal resolution at which data is collected, which means that higher temporal resolution data can increase the accuracy of techniques used to identify and classify events (Cominola et al., 2018).

Obtaining data at these finer temporal resolutions presents several challenges in terms of data collection, storage, management and processing (Cominola et al., 2018). Most water meters operating today are not capable of collecting data at sub minute resolutions. In consequence, studies measuring water use at sub minute resolutions have

relied on different data collection devices and software (Cominola et al., 2015). These data collection devices are installed on top of existing meters, requiring additional sensors to record water use at a higher temporal resolution. The high temporal resolution data collected is later processed and analyzed to derive end use information. The associated cost of using such devices, which can cost as much as $2500 per device, can be prohibitive for many researchers and utilities. Recently, lower cost devices designed to collect data at higher temporal resolutions for the purpose of detecting leaks and providing information to water consumers have entered the market (e.g., Flume Inc., 2020; PHYN, 2020). However, these devices are proprietary and typically do not provide access to the raw data collected. Raw data is an essential input for researchers aiming at studding residential water use.

Without sufficient cyberinfrastructure for automating data management tasks, high temporal resolution data could be a barrier rather than an opportunity. The term "cyberinfrastructure" integrates hardware and software tools, as well as data networks that enable innovation (NSF, 2007). Available cyberinfrastructure for collecting, managing, and analyzing high temporal resolution remains scarce and proprietary given the cost and complexity of these applications. The closed source nature of these tools creates accessibility and interoperability issues that prevent advancement and reduce the adoption of open architectures (Hauser and Roedler, 2015; Robles et al., 2014). The tools and methods used for data management in past studies collecting this type of data are not fully described. At the utility level, dedicated information technology or data management staff would be needed to process and make use of the high volume of data generated and the new technologies needed (e.g., for databasing and data analytics).

Cyberinfrastructure for the urban water sector has been discussed in the past (Boyle et al., 2013; Hauser et al., 2016; Li et al., 2020; Liu and Nielsen, 2016; Makropoulos, 2017; Ye et al., 2016), yet actual implementations are scarce. Due to the lack of implementations, important information, such as performance metrics or guidance for implementation, are not commonly described (Li et al., 2020). The importance of cyberinfrastructure systems in shaping smart cities has long been identified (Hollands, 2008; Yan et al., 2013). The terms "smart city," "smart water systems," and other similar terms are commonly used to refer to cities that are implementing cyberinfrastructure to address urban challenges (Del-Real et al., 2021), but there is disagreement on the definition and the extent of such implementations (Albino et al., 2015; Esashika et al., 2021; Wissner, 2011). Cyberinfrastructure implementations are needed to test the resiliency, performance, network utilization, and computational requirements of smart water systems (Amaxilatis et al., 2020).

Open source cyberinfrastructure can help solve data management challenges and enable high temporal resolution data collection by researchers and utilities while laying the foundation for development of newer and better tools, as wells as standards for operation that increase interoperability. The overarching goal of the research presented in this dissertation was to advance the existing cyberinfrastructure for smart water metering applications and generate new information about water use in Logan and Providence, Utah. To guide the research, the following objectives were identified. Each of the objectives is addressed within one or more chapters of this dissertation.

Objective 1: Quantify residential water use at high frequency using a low cost, non-intrusive monitoring system.

While proprietary devices exist and have been used in past research to characterize residential water use at a high temporal resolution, cost and the proprietary nature of such systems remain significant barriers to the realization of such studies and to the advancement of the devices. The most widely used device in past residential studies (F.S. Brainard & Company, 2020) has an autonomy of less than eight days when collecting data at 5 second resolution, which significantly increases the cost of collecting data for periods longer than a few days. In other monitoring and sensing fields, the price of sensors and dataloggers is decreasing while their capabilities are increasing. Whereas existing commercially available devices in the field have not taken advantage of this trend, low-cost, open source dataloggers can exceed the current capabilities of the proprietary dataloggers and provide an open platform for constant improvement. Work under this objective was aimed at enhancing the availability of flexible hardware capable of collecting high resolution water metering data on top of magnetically driven water meters without upgrading existing metering infrastructure. While magnetically driven meters are the most common meters in the U.S., there is no publicly available information describing the types of meters that are currently installed across the country. We estimate that this number could be as high as 75 – 80% of all meters currently installed.

Objective 2. Develop open source cyberinfrastructure for high temporal resolution, residential smart metering data management.

Having the capability to manage and extract useful information from high temporal resolution water use data collected using the dataloggers generated from Objective 1 or similar devices is important as it reduces the burden and cost for

conducting research in this field. Additionally, cyberinfrastructure is the keystone for shaping smart grids in the water sector. Systems for water use analytics have usually been designed using multiple, connected software layers to achieve different objectives. Typically, software systems need to balance the benefits of multiple layers versus the complexity of the overall product; more layers give more flexibility but make the system more complex and potentially prone to errors. Complete implementations of cyberinfrastructure systems are rare, given the cost and complexity of these applications (Alvisi et al., 2019; Amaxilatis et al., 2020; Anda et al., 2013). The work under this objective focused on advancing the available software cyberinfrastructure for collecting, transmitting, storing, managing, and analyzing high resolution water metering data through investigation of inexpensive hardware and open source software solutions.

Objective 3. Investigate residential water use across groups of different consumption levels.

In the United States, residential end uses of water studies have been conducted sporadically, across a limited number of cities. There are important differences in how residents use water at the city level that highlight the importance of having local information when making water management decisions. Additionally, the temporal variability of indoor water use has not been fully evaluated in past studies. Furthermore, limited analyses of outdoor water use derived from end uses classification have been conducted. Characterizing outdoor water use from individually labeled events and assessing the variability of indoor water use require longer data collection periods than those that have been used in prior studies that examined end uses of water (the most common data collection period in prior studies is 2 weeks). Work under this objective

focused on conducting a detailed residential water use study, coupling state-of-the-art data collection, management, and analysis tools to evaluate how water use varies for users at different levels of consumption and observe the temporal variability of end uses of water.

The outline of the rest of this dissertation is as follows. In Chapter 2, a new datalogging device used to collect high temporal resolution residential water use data is presented. Chapter 2 covers the design, calibration, and field testing of the datalogger. Chapter 2 mainly addresses Objective 1 but also contributes towards Objectives 2 and 3 by enabling the measurement of residential water use at high temporal frequency without disrupting the operation of existing meters and contributing towards the generation of the data used in the case studies presented in the other chapters.

In Chapter 3, we present the development of a cyberinfrastructure system designed to manage residential water use data collected from two contexts, single family residential properties and multi-unit residences on a college campus. This cyberinfrastructure was built by combining multiple, existing open source technologies and software tools developed for this specific application, including a new method for identifying and classifying end uses of water developed as part of this project (Attallah et al., 2021a). Chapter 3 addresses Objective 2 by demonstrating how the process from data collection to visualization and analysis can be automated.

Chapter 4 addresses Objective 3 by presenting a case study in Logan and Providence, Utah, were residential water use was analyzed over a sample of 31 residential properties for periods of time ranging between four and 22 weeks. Chapter 4 builds on the developments reported in Chapters 2 and 3 to demonstrate one of the

possible applications of the research products developed aimed at assisting investigations of residential water demand.

REFERENCES

Abdallah, A.M., Rosenberg, D.E., 2014. Heterogeneous residential water and energy linkages and implications for conservation and management. Journal of Water Resources Planning and Management 140, 288–297. https://doi.org/https://doi.org/10.1061/(ASCE)WR.1943-5452.0000340

Albino, V., Berardi, U., Dangelico, R.M., 2015. Smart Cities: Definitions, Dimensions, Performance, and Initiatives. Journal of Urban Technologies 22, 3–21. https://doi.org/10.1080/10630732.2014.942092

Alvisi, S., Casellato, F., Franchini, M., Govoni, M., Luciani, C., Poltronieri, F., Riberto, G., Stefanelli, C., Tortonesi, M., 2019. Wireless Middleware Solutions for Smart Water Metering. Sensors 19, 1853. https://doi.org/10.3390/s19081853

Amaxilatis, D., Chatzigiannakis, I., Tselios, C., Tsironis, N., Niakas, N., Papadogeorgos, S., 2020. A Smart Water Metering Deployment Based on the Fog Computing Paradigm. Appl. Sci. 10, 1965. https://doi.org/10.3390/app10061965

Anda, M., Le Gay Brereton, F., Brennan, J., Paskett, E., 2013. Smart metering infrastructure for residential water efficiency: Results of a trial in a behavioural change program in Perth, Western Australia, in: Information and Communication Technologies for Sustainability, 14 - 16 February, Zurich, Switzerland. https://researchrepository.murdoch.edu.au/id/eprint/22422/1/smart_metering_infrastructure.pdf

Amaxilatis, D., Chatzigiannakis, I., Tselios, C., Tsironis, N., Niakas, N., Papadogeorgos, S., 2020. A Smart Water Metering Deployment Based on the Fog Computing Paradigm. Applied Sciences 10, 1965. https://doi.org/10.3390/app10061965

Attallah, N.A., Horsburgh, J.S., Bastidas Pacheco, C.J., 2021. Tools for Evaluating, Developing, and Testing Water End Use Disaggregation Algorithms. Submitted for Publication.

Boyle, T., Giurco, D., Mukheibir, P., Liu, A., Moy, C., White, S., Stewart, R., 2013. Intelligent metering for urban water: A review. Water 5, 1052. https://doi.org/https://doi.org/10.3390/w5031052

Cole, G., Stewart, R.A., 2013. Smart meter enabled disaggregation of urban peak water demand: precursor to effective urban water planning. Urban Water Journal 10, 174–194. https://doi.org/10.1080/1573062X.2012.716446

Cominola, A., Giuliani, M., Castelletti, A., Rosenberg, D.E., Abdallah, A.M., 2018. Implications of data sampling resolution on water use simulation, end-use disaggregation, and demand management. Environmental Modelling and Software 102, 199–212. https://doi.org/https://doi.org/10.1016/j.envsoft.2017.11.022

Cominola, A., Giuliani, M., Piga, D., Castelletti, A., Rizzoli, A.E., 2015. Benefits and challenges of using smart meters for advancing residential water demand modeling and management: A review. Environmental Modelling and Software 72, 198–214. https://doi.org/10.1016/j.envsoft.2015.07.012

Del-Real, C., Ward, C., Sartipi, M., 2021. What do people want in a smart city?

Exploring the stakeholders' opinions, priorities and perceived barriers in a medium-sized city in the United States. International Journal of Urban Science 1–25. https://doi.org/10.1080/12265934.2021.1968939

Dieter, C.A., Maupin, M.A., Caldwell, R.R., Harris, M.A., Ivahnenko, T.I., Lovelace, J.K., Barber, N.L., Linsey, K.S. 2018. Estimated use of water in the United States in 2015. U.S. Geological Survey Circular 1441 https://doi.org/10.3133/cir1441

EPA, 2016. Urbanization and Population Change. URL: https://cfpub.epa.gov/roe/indicator.cfm?i=52 (accessed 5.5.21)

EPA, 2018. Drinking Water Infrastructure Survey and Assessment. https://www.epa.gov/sites/default/files/2018-10/documents/corrected_sixth_drinking_water_infrastructure_needs_survey_and_assessment.pdf

Esashika, D., Masiero, G., Mauger, Y., 2021. An investigation into the elusive concept of smart cities: a systematic review and meta-synthesis. Technology Analysis and Strategic Management 33, 957–969. https://doi.org/10.1080/09537325.2020.1856804

F.S. Brainard & Company, 2020. Meter-Master. URL: https://meter-master.com/product/model-20/ (accessed 9.25.21)

Flume Inc., 2020. Protect Your Home. URL: https://www.flumetech.com/ (accessed 9.26.21)

Giurco, D., Carrard, N., McFallan, S., 2008. Residential End use Measurement Guidebook. A Guide to Study Design, Sampling and Technology. https://opus.lib.uts.edu.au/bitstream/10453/35089/1/giurcoetal2008resenduse.pdf

Gurung, T.R., Stewart, R.A., Beal, C.D., Sharma, A.K., 2015. Smart meter enabled water end-use demand data: platform for the enhanced infrastructure planning of contemporary urban water supply networks. Journal of Cleaner Production 87, 642–654. https://doi.org/10.1016/j.jclepro.2014.09.054.

Hauser, A., Roedler, F., 2015. Interoperability: The key for smart water management. Water Supply 15, 207–214. https://doi.org/10.2166/ws.2014.096

Hauser, A., Sud, T., Nicolas Foret, C., Electric Stuart Combellack, S., Jonathan Coome, T., Quintilia Lopez, S., Elkin Hernandez, I., Water Salil Kharkar, D.M., Water Amin Rasekh, D., Michal Koenig, S., Remy Marcotorchino, Q., Nicolas Damour, S., 2016. Communication in Smart Water Networks SWAN Forum Interoperability Workgroup. https://www.swan-forum.com/wp-content/uploads/sites/218/2020/12/SWAN-White-Paper_Communication-Protocols.pdf

Hollands, R.G., 2008. Will the real smart city please stand up? Intelligent, progressive or entrepreneurial? City 12, 303–320. https://doi.org/10.1080/13604810802479126

Li, J., Yang, X., Sitzenfrei, R., 2020. Rethinking the Framework of Smart Water System: A Review. Water 12, 412. https://doi.org/10.3390/w12020412

Liu, X., Nielsen, P.S., 2016. A hybrid ICT-solution for smart meter data analytics.

Energy 115, 1710–1722. https://doi.org/10.1016/j.energy.2016.05.068

Makropoulos, C., 2017. Thinking platforms for smarter urban water systems: Fusing technical and socio-economic models and tools. Geological Society, London, Special Publications 408, 201–219. https://doi.org/10.1144/SP408.4

Marzano, R., Rougé, C., Garrone, P., Grilli, L., Harou, J.J., Pulido-Velazquez, M., 2018. Determinants of the price response to residential water tariffs: Meta-analysis and beyond. Environmental Modelling and Software 101, 236–248. https://doi.org/https://doi.org/10.1016/j.envsoft.2017.12.017

Mayer, P.W., B. DeOreo, W., Towler, E., Martien, L., M. Lewis, D., 2004. Tampa Water Department Residential Water Conservation Study: The Impacts of High Efficiency Plumbing Fixture Retrofits in Single-Family Homes.

Nguyen, K.A., Stewart, R.A., Zhang, H., Jones, C., 2015. Intelligent autonomous system for residential water end use classification: Autoflow. Applied Soft Computing 31, 118–131. https://doi.org/10.1016/j.asoc.2015.03.007

NSF, 2007. Cyberinfrastructure Vision for 21st Century Discovery. https://www.nsf.gov/pubs/2007/nsf0728/index.jsp

PHYN, 2020. Your Water Like You've Never Seen It. URL: https://www.phyn.com/technology/ (accessed 9.20.21)

Robles, T., Alcarria, R., Martin, D., Morales, A., Navarro, M., Calero, R., Iglesias, S., Lopez, M., 2014. An internet of things-based model for smart water management, in: 2014 IEEE 28th International Conference on Advanced Information Networking and Applications Workshops, IEEE WAINA 2014. IEEE Computer Society, pp. 821–826. https://doi.org/10.1109/WAINA.2014.129

Rockaway, T.D., Coomes, P.A., Rivard, J. and Kornstein, B. (2011), Residential water use trends in North America. Journal - American Water Works Association, 103: 76-89. https://doi.org/10.1002/j.1551-8833.2011.tb11403.x

Suero, F.J., Mayer, P.W., Rosenberg, D.E., 2012. Estimating and Verifying United States Households' Potential to Conserve Water. Journal of Water Resources Planning and Management 138, 299–306. https://doi.org/doi:10.1061/(ASCE)WR.1943-5452.0000182

The University of Utah, 2016. Fact Sheet August 2016. Utah at a Glance. https://gardner.utah.edu/wp-content/uploads/2016/08/UtahAtAGlance-Final1.pdf

UN-Habitat, 2016. Urbanization and Development: Emerging Futures, World Cities Report 2016. https://www.unhabitat.org/wp-content/uploads/2014/03/WCR- Full-Report-2016.pdf

UDWR. (2020). 2015 Municipal and Industrial Water Use Data: 2020 Version 3. Utah Division of Water Resources. https://drive.google.com/file/d/1aD9SorKQauIfiDW0wdMXlafd0VKsdX-F/view.

Willis, R.M., Stewart, R.A., Panuwatwanich, K., Jones, S., Kyriakides, A., 2010. Alarming visual display monitors affecting shower end use water and energy conservation in Australian residential households. Resources, Conservation and

Recycling 54, 1117–1127. https://doi.org/10.1016/j.resconrec.2010.03.004.

Wissner, M., 2011. The Smart Grid - A saucerful of secrets? Applied Energy 88, 2509–2518. https://doi.org/10.1016/j.apenergy.2011.01.042

Yan, Y., Qian, Y., Sharif, H., Tipper, D., 2013. A survey on smart grid communication infrastructures: Motivations, requirements and challenges. IEEE IEEE Communications Surveys & Tutorials, vol. 15, no. 1, pp. 5-20, First Quarter 2013. https://doi.org/10.1109/SURV.2012.021312.00034

Ye, Y., Liang, L., Zhao, H., Jiang, Y., 2016. The System Architecture of Smart Water Grid for Water Security, in: Procedia Engineering. Elsevier Ltd, pp. 361–368. https://doi.org/10.1016/j.proeng.2016.07.492

CHAPTER 2

A LOW-COST, OPEN SOURCE MONITORING SYSTEM FOR COLLECTING HIGH TEMPORAL RESOLUTION WATER USE DATA ON MAGNETICALLY-DRIVEN RESIDENTIAL WATER METERS[1]

Abstract

We present a low-cost (≈ $150) monitoring system for collecting high temporal resolution residential water use data without disrupting the operation of commonly available water meters. This system was designed for installation on top of analog, magnetically-driven, positive displacement, residential water meters and can collect data at a variable time resolution interval. The system couples an Arduino Pro microcontroller board, a datalogging shield customized for this specific application, and a magnetometer sensor. The system was developed and calibrated at the Utah Water Research Laboratory and was deployed for testing on five single family residences in Logan and Providence, Utah for a period of over 1 month. Battery life for the device was estimated to be over 5 weeks with continuous data collection at a 4 second time interval. Data collected using this system, under ideal installation conditions, was within 2% of the volume recorded by the register of the meter on which they were installed. Results from field deployments are presented to demonstrate the accuracy, functionality, and applicability of the system. Results indicate the device is capable of collecting data at a temporal resolution sufficient for identifying individual water use events and analyzing water use at coarser temporal

---

[1] Bastidas Pacheco, C.J., Horsburgh, J.S., Tracy, R.J., 2020. A Low-Cost, Open Source Monitoring System for Collecting High Temporal Resolution Water Use Data on Magnetically Driven Residential Water Meters. Sensors 20, 3655.

resolutions. This system is of special interest for water end-use studies, future projections of residential water use, water infrastructure design, and for advancing our understanding of water use timing and behavior. The system's hardware design and software are open source, are available for potential reuse, and can be customized for specific research needs.

2.1 Introduction

The vast majority of water meters used by water supply utilities today for quantifying residential water consumption are analog, magnetically driven, positive displacement meters. These meters use a nutating disc or a similar mechanism and measure water flow using the positive displacement principle. Water flows into a chamber in the meter causing the disk to nutate, and each nutation represents a fixed volume of water. The count of nutations is registered by the meters using a magnetically-driven register. Measurements made by these meters are typically within 0.25-0.5% of the actual value [1]. While these meters are highly accurate and have been used effectively for decades to quantify residential water use for billing purposes, they were designed to be read only periodically, typically monthly or quarterly. Because monthly resolution data provide little information about the distribution of use across end uses (e.g., toilets, showers, faucets, etc.) and the timing of use both within and outside a home, data at a higher temporal resolution must be collected to effectively identify and understand water use behavior. Smart meters have potential to meet this need while supporting automated billing processes. The term "smart meter" can be ambiguous [1]. In this article, the term is related to devices capable of collecting high temporal resolution data (i.e., high sampling frequency) that can be integrated in efficient systems for data

management [2]. However, replacing traditional meters with smart meters can be expensive, labor intensive, and disruptive. In consequence, collecting high temporal resolution water use data can be cost prohibitive for many utilities and researchers. Yet, doing so enables new opportunities for quantifying water use behavior at high temporal resolution [3–5].

Given that most water meters installed and operating today are not capable of recording high temporal resolution data, many past research studies requiring this type of data have relied on proprietary data collection devices and software that require operation by and input from trained analysts [2]. These data logging devices are installed on top of existing meters using different types of sensors to collect high temporal resolution water use data and, thus, to add smart metering capabilities. Collected data are then downloaded and processed to identify and disaggregate end uses of water. However, the associated costs can be prohibitive for many researchers and water utilities. For example, DeOreo et al. [6,7] used Meter-Master flow recorders [8] to collect 10 second resolution data for hundreds of households. This proprietary device is installed on positive displacement, magnetically driven meters and can collect high temporal resolution data, but a single unit can cost over $2,000 USD. Other authors have developed and tested different sensors and data recording devices to identify when a fixture is used within the house [9–12]. With these technologies, disaggregation of end uses requires the existence of a smart meter capable of high temporal resolution data collection. These devices can identify when, and in some cases where, a fixture is being used, but rely on post processing of the smart meter data to estimate volumes, flow rates and other characteristics of the events.

Some devices used in other studies to collect high temporal resolution data [9] are not available today because they have been sold to private partners [13].

There are commercially available smart meters that can collect data at the minute resolution [14], but this may be too coarse for some applications (e.g., identifying individual end use events) because some events have durations that last only seconds. Other devices are entering the market that are designed to collect data at higher temporal resolutions for the purpose of detecting leaks and providing information to water consumers [13,15]. These devices are proprietary (i.e., they are produced by commercial companies for sale, cannot be modified, and source code is not open), they are not interoperable, and they generally do not provide access to the raw data they collect, opting instead to provide water consumers with summary information designed to inform them about their water consumption. While promising for consumer applications, these devices are not well suited for research data collection. Thus, an openly available and affordable data collection device could solve one of the existing limitations to conducting research using high temporal resolution water use data.

Over the past several years, there has been a general reduction in prices of sensors and dataloggers; however, cost continues to be an important limitation for scientific research [16,17]. More recently, open-source electronics hardware, specifically Arduino, has been identified as a viable alternative for expensive, commercial instrumentation in scientific research [18]. Arduino is an open-source electronics prototyping platform that consists of both microcontroller hardware and the Arduino software for programming them [19]. In the field of water resources, Arduinos have been used in multiple applications that range from monitoring water quality in streams [20], promoting water

conservation [21], operating irrigation systems [22,23], and many other applications [24,25]. One of the strengths of the Arduino is the Integrated Development Environment (IDE) that includes extensive code libraries for developing measurement and control systems [26]. Arduinos are highly configurable computing devices that have expanded the development of customized applications in multiple fields. They can be transformed into autonomous systems, installed in tiny spaces, used in remote field locations, and they can be deployed without peripheral devices like monitors and keyboards [27]. The availability of Arduino-compatible development boards has helped create a new variety of inexpensive, open-source hardware for data logging applications [26].

In this paper, we describe an open source datalogger that uses an Arduino microcontroller board in combination with other commonly available hardware components to measure and record high temporal resolution water use data on analog, magnetically driven, positive displacement meters. Developed as part of a larger effort aimed at developing Cyberinfrastructure for Intelligent Water Supply (CIWS), the CIWS datalogger can be used with existing meters without affecting their functioning or their normal data collection activities, either manual or wireless. Thus, adding a CIWS datalogger to an existing, analog meter effectively transforms it into a smart meter capable of recording data at any temporal resolution required for a particular study. The hardware and software of the CIWS datalogger are open source, and they can be modified to fit specific research needs. The CIWS datalogger software uses existing Arduino code libraries, and new libraries were also developed for specific functions. The system presented is a low-cost alternative for collecting high temporal resolution residential water usage data. The main characteristics we sought to meet in the design of

this system included: ease of assembly, autonomous operation for approximately 6 weeks while recording data at high temporal resolution (< 5 seconds), low purchase and assembly cost, flexibility for customization, accuracy of measurements, and building from an open hardware and software platform.

This article is organized as follows: Section 2 describes the CIWS datalogger, its functioning principle, hardware design, software, and user interface. Section 3 presents the procedures we used to test and calibrate the device in a laboratory setting using multiple meters from different manufacturers along with calibration results. Section 4 discusses the results of a field deployment campaign we used to test the data collection capabilities and functioning of the device under normal operating conditions. The results of analyses conducted on the data collected are included in this section. Section 5 presents final discussion points, areas for improvement, and future work. The Hardware, Firmware and Data Availability section at the end of this article provides links to directories where readers can find: a) hardware designs along with instructions for performing all of the hardware modifications described and a diagram of connections; b) PCB designs and all information required to manufacture them; c) firmware code along with more detailed documentation about the organization and functioning of firmware; and d) data and scripts to reproduce calculations presented here.

2.2 System Description

The CIWS datalogger was designed to operate on top of existing, magnetically driven residential meters of common sizes (e.g., 1 in, 3/4 in, and 5/8 in). In this paper, the meter sizes are described in inches to match manufacturer specifications for how these meters are sold in the United States. The meters used to calibrate and test the CIWS

datalogger were manufactured by Neptune and Master Meter and were designed to operate at different flow rates depending on their size. For 3/4 and 5/8 in meters, the manufacturers report accuracy information for flow rates between 0.1 and 20 gallons per minute (GPM). For 1 in meters, the accuracy is reported between 0.35 and 50 GPM. We designed the CIWS datalogger to meet the following specifications: (a) operation on top of existing meters without requiring replacement of the meter and without affecting the function of the existing meter; (b) autonomous operation for longer than 20 days; (c) versatility to work with different meter brands and sizes without requiring in situ calibration; (d) sufficient accuracy and temporal resolution to allow the identification and classification of end uses of water in a residential home; (e) simplicity of use with an easily operable user interface; and (f) output data in an accessible format that is platform and software independent.

2.2.1 Principle of Functioning

Many existing residential water meters use a nutating disc or other similar device to measure water flow using the positive displacement principle. Water flows into a measurement chamber in the meter that obstructs the flow, and a nutating or rotating mechanism allows the passage of a fixed volume of water. Actuation of the chamber's fixed volume, or displacement, as a nutation or revolution of the measurement element represents passage of a fixed volume of water. The rate of revolution or nutation is proportional to the flow rate. The count of revolutions or nutations is recorded using a magnetically driven register. A magnet inside the register is paired with a spinning magnet inside of the meter's sealed housing. As water flowing through the meter causes the magnet inside the meter housing to rotate, the paired magnet inside the register also

rotates. These rotations are counted by the meter's register to record the count of pulses, which determines the flow volume and rate. The registers used with most meters provide only volume information, while some registers may also provide a flow rate. However, registers do not provide access to the magnetic pulse information from which the volume and flow rate are derived, and meter manufacturers do not typically publish the pulse resolution (volume of water per pulse) of their meters.

The magnets in the meter and register create oscillations in the magnetic field surrounding the meter as they rotate, generating peaks that can be measured with each revolution. Just like the meter's register, the CIWS datalogger counts the number of times the magnet inside the meter rotates. The main difference is that it detects the rotations using a magnetometer sensor mounted on the outside of the register. The magnetometer measures changes in the magnetic field as the magnet inside the meter rotates, and the datalogger then counts the peaks that occur in the magnetic field without modifying or affecting the regular function of the meter. The datalogger operates via a firmware code that has two main functions. First, it detects and sums the number of magnetic pulses (peaks) that occur during a time step. Second, it logs this value along with the corresponding date and time. The recording interval is configurable to allow for adequate identification and separation of short-duration water use events. Detailed descriptions of the hardware and software are provided in the sections that follow.

The magnetometer raw output is a linearly scalable integer between −128 and 127. For counting peaks in the magnetic field, scaling the raw signal does not provide any additional information as the exact value of the magnetic field is not of interest, but only the number of peaks. The raw magnetic field was observed in multiple experiments to

characterize it and define an algorithm that could potentially count peaks from any magnetically driven meter, independent of the position of the magnetometer sensor relative to the meter register. In these experiments, the magnetic field was sampled at 570 Hz, and some characteristics of the signal were observed. First, the magnetic field signal is weak, being contained by a small percentage of the ± 4 gauss range, which is the smallest range possible with the sensor selected. Second, peaks are closer to each other, in time, when the flow rate is higher, meaning the frequency of the signal is variable and proportional to the flow rate, which depends on the meter size, the pressure in the pipe, and the fixture through which water is being used. The maximum frequencies observed in our laboratory setting were below 50 Hz, although higher values may be possible in other settings. Third, the range and the average amplitude of the signal are different for every brand and model of meter and are also dependent on the position of the magnetometer relative to the meter. We observed values of the linearly scalable integer between −25 and 3 during laboratory experiments, but different values are possible depending on how the sensor is installed on the meter. Fourth, the value output by the magnetometer sensor can remain constant at any value within the observed range when the magnet stops spinning. Fifth, the signal is noisy. In the upper panel of Figure 2.1, the red dashed line presents the raw output of the magnetometer during a 2 s data recording interval. When starting the data collection, the valve controlling water flow through the meter was closed. It was then opened, and the variation in the signal can be observed as water flow was increased.

The approach designed to count peaks uses two thresholds, T1 and T2 (Figure 2.1). A pulse is counted when the signal goes above T1 and then sequentially below T2.

A single threshold approach is not sufficient given that the magnetic field can remain constant at any value within its range and the noise in the signal that can cause oscillations above and below a single threshold that do not represent true pulses. Additionally, observing local maximums within a fixed number of values is not possible due to the changing period in the signal as flowrates change. Since the average amplitude of the signal is not constant, these two thresholds would need to be calibrated for every installation (i.e., every type of meter and sensor location), which would limit the generalization of the application. To address this, an infinite impulse response (IIR) filter was added to process the magnetometer output to produce a signal with constant mean amplitude independent of the meter type and installation location of the magnetometer sensor. Digital filters, including IIR filters and finite impulse response (FIR) filters [28], are fundamental in processing signals to remove their unwanted parts. The main difference between them is that IIR filters are recursive and use feedback from the output in the filter structure, whereas FIR does not [29]. IIR filters have a higher computational economy because they require less memory and fewer arithmetic operations than FIR [30]. This makes them better suited for this application, which requires running the algorithm on the microcontroller in real time. Recording the raw data and processing it later in a centralized facility would require larger computational power given the volumes of data that are generated since the magnetic field is sampled at a 560−570 Hz rate. However, IIR filters need to be designed with extra care because they can become unstable [30].

IIR filters are typically expressed as a difference equation, which calculates a sample output at a time n based on past outputs and present and past inputs. The order of

a difference equation is defined by the number of past samples it uses [31]. A basic, first order, difference equation form is presented in Equation (1):

$$y_n = a * y_{n-1} + b_1 * x_n - b_2 x_{n-1} \quad (1)$$

where, y_n is the output filtered signal for the current time step, n; a is typically known as the feedback coefficient; y_{n-1} is the output filtered signal for the previous time step, n-1; b1 and b2 are the feedforward coefficients; and x_n and x_{n-1} are the inputs (raw signal) for the current and previous time steps, respectively [31]. If a is not zero, Equation 1 defines an IIR filter. The feedback and feedforward coefficients are predefined for classical filters, such as Butterworth, Chebyshev, or other designs [30]. For our application, the purpose of the filter was simply to output a signal with a constant average amplitude, rather than to pass or reject specified frequencies, as these filters are typically designed [29], which makes the problem simpler. The feedback and feedforward coefficients are typically calibrated to obtain the response desired. For our application, b1 and b2 were set to 1, and a was set to 0.95, resulting in Equation 2:

$$y_n = 0.95 * y_{n-1} + x_n - x_{n-1} \quad (2)$$

An IIR filter is stable if its response to an impulse approaches zero as n goes to infinity. With the parameters selected, y_n will decay gradually if the input is an impulse (i.e., a one followed by zeros). Then, Equation 2 represents a filter that produces a signal with constant mean amplitude equal to zero, that is independent of the meter type and installation location of the magnetometer, and that reduces noise while maintaining the shape of the input signal. These parameters were selected and tested to be valid for this application, but there are infinite configurations of a, b_1, and b_2 that would result in a stable filter that would satisfy our requirements (i.e., the output signal must have a

constant average amplitude and maintain the shape of the input signal). For example, keeping the same values of b_1 and b_2, any value for a larger than 0.95 but less than 1 will meet all requirements while maintaining stability. If we gradually select values for a less than 0.95, we will reach a point where the output signal will have fewer peaks than the original signal, any value of a larger than this number and less than 1 will meet our requirements. Having an output signal with the same shape as the input assures that the pulses counted in the output signal also exist in the input signal measured by the magnetometer. The existence of noise does not interfere with the two threshold approach to count pulses because the magnitude of the noise is much smaller than the magnitude of the overall signal. Therefore, finding the values of a, b_1, and b_2 that remove the most noise while maintaining the shape is not of interest, but could be easily done, if needed, in laboratory testing. The parameters selected are a valid and simple solution for our application. Figure 2.2 shows the frequency response of the filter designed. Signals with a frequency near 0 Hz are attenuated, whereas signals with higher frequencies pass through without any attenuation. Because these very low frequencies, especially any 0 Hz component, are so heavily attenuated, the signal loses its constant offset and becomes centered around zero. Lyons [32], provides a more detailed discussion around this type of filter. This filter operates adequately for the frequency ranges described above. Because this is a discrete filter, any frequencies above half of the sampling frequency will alias to a lower-frequency signal, and the resulting data will be faulty.

Having a signal with a constant mean amplitude, zero in this case, allows us to define fixed values for the thresholds – in our case T1 = 1 and T2 = -1. These values were selected based on observations made of the raw signal from multiple water meters and

have been proved valid in the field. Similar to the parameters in the filter, there are multiple options for T1 and T2 that would provide a valid solution. The process of counting pulses using two thresholds can be referred to as a digital Schmitt Trigger. The main function of the Schmitt Trigger is to convert the filtered signal, (blue, solid line, upper panel, Figure 2.1) into a clean, square wave (black, solid line, lower panel, Figure 2.1) from which pulses can be easily counted. The CIWS datalogger keeps track of time using a real-time clock (RTC) incorporated in the datalogging shield, and logs time and the count of pulses in regular, configurable, time step intervals.

2.2.2 Sample Output

The CIWS datalogger outputs a comma separated values (CSV) file including a 3 lines header with information about: 1) Site #, a 3 digit numerical ID used to keep track of where the logger is installed; 2) Datalogger ID #, a 3 digit numerical ID used to identify a datalogger, and; 3) Meter Resolution, a numeric value with 3 decimal places indicating the pulse resolution of the meter (gallons per pulse to match the meter's register units) where the logger is installed. The meter resolution is used for displaying volumes in the user interface. The logger registers data by keeping track of 3 variables: 1) Time, a datetime value including the date and time in format "Year-Month-Day Hour:Minute:Second"; 2) Record, a numerical ID used to keep track of the number of values logged, and; 3) Pulses, an integer indicating the number of pulses registered in a time interval. The datetime string format was chosen to be consistent with the International Standards Organization (ISO) 8601 standard for the representation of dates and times to make it easier to work with across computer operating systems, database programs, and programming languages. Figure 2.3 shows an example of a CSV file

obtained from the CIWS datalogger. In this example, the site and datalogger ID are 001 and meter pulse resolution is set to 0.033. Only the first 10 records are presented.

2.2.3. Hardware

The CIWS datalogger main components are a LIS3MDL digital output magnetic sensor [33], an Arduino Pro microcontroller board [34], and a custom sensor interface board assembled on an Adafruit datalogging shield [35].

2.2.3.1 Magnetometer Sensor

The LIS3MDL is an ultra-low-power, three-axis magnetometer that can operate at different gauss scales (± 4, ± 8, and ± 16 gauss). In this system, we use only one (the x) axis available on the sensor as the y and z axes do not provide additional information for this application. The sensor is configured to operate in the ± 4 gauss scale as the magnetic signal from the water meter is weak enough to be fully captured within this range. The highest flow rate we have observed at residential homes is ~80 liters per minute (LPM), for meters of smaller sizes (3/4 in, 5/8 in). With water flowing at that flow rate, we will observe ~40 pulses per second. Using the definition of pulses explained in the previous section, this means the signal will have ~40 positive peaks and the same number of negative peaks when water is flowing at this rate. Higher flow rates are possible. The Nyquist-Shannon sampling theorem establishes that if we want to properly characterize a signal, we must sample it with at least twice the input signal frequency [36]. The frequency of the magnetic signal from the magnetometer changes with the flow rate. In consequence, the sensor must continuously sample at a high rate in order to capture these changes.

The sensor has four system modes: continuous-measurement mode, single-measurement mode, and two idle modes, along with four operating modes: low-power, medium performance, high-performance, and ultra-high-performance. Multiple output data rate options are available by choosing an appropriate operating mode, ranging from 0.65 to 1,000 Hz [33]. The magnetic data are sent to different registers. For the sensor's x axis, two registers are used, OUT_X_H and OUT_X_L. These contain the most significant and least significant part of the magnetic signal on the x axis, respectively [33]. A FAST_READ option is also available to accelerate the process of reading data from the sensor. By selecting this option, only the OUT_X_H register data is sent [33]. The sampling frequency of the magnetic signal using the LIS3MDL magnetometer is then a function of the combination of the options selected for each one of these parameters.

In the CIWS datalogger, the sensor system mode is set to the continuous-measurement mode, the operating mode is set to medium performance, and the FAST_READ option is active, which results is a sampling frequency of approximately 560-570 Hz. Other configuration options to sample at 165 Hz and 300 Hz are available. These configuration options were tested in the laboratory for the range of flow rates we observed at residential homes using different meters. Results showed that sampling at 165 HZ and 300 Hz can capture the signal as accurately as sampling at the faster rate. Power consumption differences between these configurations were not estimated, but the slower sampling rates may consume less power and result in longer potential deployment times. The 570 Hz configuration settings were selected for field deployment as a higher sampling frequency results in a better characterization of the signal. This frequency has

proved to be sufficient to accurately capture the pulses associated with water flowing at the maximum flow rates we have observed in common residential size meters. The clock speed for the Inter-Integrated Circuit (I2C) controls the data transfer between the sensor and the datalogger and supports standard and fast sampling modes at 100 KHz and 400 KHz, respectively [33]. Multiple data transfers from the sensor to the datalogger were measured. Each transfer takes less than 0.7 ms; therefore, conducting 570 transfers in a second would take less than 0.4 s. Based on this data, we adopted the standard configuration after observing that it is fast enough to handle all data transfers between the sensor and the datalogger.

While it is acknowledged that the relatively weak magnetic signal produced by the water meters we tested occupies a small portion of the sensor's potential output range (i.e., a linearly scalable integer between -128 and 127 at the ± 4 gauss range), we were unable to find an inexpensive sensor with a range more suitable than the LIS3MDL. Additionally, the LIS3MDL draws less current than many other types of sensors (e.g., Hall Effect sensors) that do not provide a better range. Using an analog sensor would have required additional signal processing components and the use of the Arduino's onboard Analog-to-Digital Converter (ADC), which would have increased power consumption and reduced the autonomy of the datalogger. At a cost of less than $5 USD, the LIS3MDL magnetometer was the best and most practical sensor we could find for this application that worked well when paired with the filtering procedure described above.

2.2.3.2. Microcontroller board

The Arduino Pro is a microcontroller board based on the ATmega328 processor [34]. We chose it for this application because it is inexpensive, the absence of connectors and additional hardware components make it more customizable, the pin layout is compatible with Arduino Shields, and it is openly available. The Arduino Pro used in this system is the 3.3 V / 8MHz version. We made several modifications to minimize power consumption of our datalogger. Although the Arduino Pro has an integrated power regulator, we removed it and replaced it with a more efficient 3.3 V regulator installed on the data logging shield (Figure 2.4a.1). The power LED on the Arduino Pro was also removed from the board. The ATmega328 has the following peripherals: I2C, Timer 0, Timer 1, Timer 2, Serial Peripheral Interface (SPI), Universal Synchronous-Asynchronous Receive-Transmit (USART), and an Analog-to-Digital Converter (ADC) which are, by default, clocked by the microcontroller's system clock, causing them to consume power while not in use. The Arduino manufacturer added an eight-bit memory-mapped Power Reduction Register (PRR). The bits written to this register either activate or shutdown the clock signal to a specific peripheral. In our device, all of the peripherals mentioned are turned off using this register to reduce energy consumption. The ADC and all of its timers remain off for the entire operation of the device, while the firmware developed activates the SPI, the USART, and the I2C modules when needed and turns them back off when they are no longer in use.

2.2.3.3. Data Logging Shield

Adafruit's data logging shield was originally designed to work with the Arduino Uno. We adapted it to work with the Arduino Pro, which has fewer connections and is

more compact than the Arduino Uno. As purchased, the shield integrates a RTC for precise timing and an SD card memory slot for storage of observed data. However, several hardware modifications were needed on the logging shield to make it compatible with the Arduino Pro. First, we shorted the Serial Clock (SCL) and the Serial Data (SDA) jumpers on the bottom of the shield with solder, connecting the I2C bus to the Arduino Pro's I2C pins. Second, we shorted the input/output resistors (IOr), 3V, and 5V busses together. This connects the I2C pull-up resistors to the 3.3V bus. Since the Arduino Pro selected for the system is the 3.3 V version, this means that the pads listed as 5V are converted to 3.3V connections. Third, we removed the power LED from the logging shield, which reduces energy consumption. The voltage regulators on the logging shield were also removed as they are unnecessary given that a more efficient regulator was installed. A wake button was installed in the logging shield to provide a way for the operator to access the user interface designed to interact with the device (Figure 2.4a.2).

Figure 2.4b shows the diagram of the connection between the main components of the CIWS datalogger. The real-time clock (RTC) on the data logging shield keeps track of the current time. Every time step it generates a pulse on the shield's SQ pin, which is wired to the D3 pin on the Arduino Pro (visible on Figure 2.4). The RTC shares the I2C bus on the Arduino Pro with the magnetometer. When the Arduino receives the pulse at the selected time step (which can be modified to meet different research needs) from the RTC, it gathers date and time information from the RTC via the I2C bus and the number of pulses detected for the current time step and logs both in a CSV file stored on the SD card. During laboratory tests and deployments, 8 and 16 GB SD cards have been used interchangeably.

2.2.3.4. Deployment Hardware

For deployment, the sensor is wired to the screw terminals on the data logging shield, which is plugged in on top of the Arduino Pro (Figure 2.5a). The main connections between all the components of the system are represented in Figure 2.5, which is included to illustrate all the hardware elements. The system can be powered by any battery with a voltage equal or larger than 4V, although power consumption will be most efficient with 12V batteries. During testing and field deployment, a 12V 10Ah lead acid battery was used. Once built, the datalogger is encased in a waterproof box (Figure 2.5a). The magnetometer is attached to the meter's register by using a strap. The magnetometer can be installed in any place on the outside of the register, after which logging is started and the datalogger remains collecting data, as observed in Figure 2.5c.

Table 2.1 lists all components, source and approximate cost per unit, at the time of this writing, to build a CIWS datalogger. According to Table 2.1, the approximate cost to build a CWIS datalogger is around $150, which will slightly vary depending on the number of loggers built. Some parts, including cables and connectors, are only available in quantities larger than what is needed for a single datalogger. The costs presented in Table 2.1 were estimated after purchasing the materials to build 20 CIWS dataloggers. Part numbers and a specific link to each vendor are available in the project's GitHub repository.

2.2.3.5. Printed Circuit Board Design

As a final step in realizing our hardware design, we translated our prototype datalogger into a printed circuit board (PCB) design that can be used to reduce the time and effort required to manufacture the CIWS datalogger. This PCB design includes all of

the Arduino and datalogging shield components and simply needs to be connected to the LIS3MDL magnetometer sensor and the power source. We ordered a small run of five of these devices using our design from the PCBWay PCB manufacturing company (http://pcbway.com) and successfully tested them in the laboratory using the same procedures we used to test our prototypes (described below) to verify the correct functioning of these devices. The total cost for manufacturing and assembling a device (Figure 2.6) was $90 USD, which included manufacture of the PCB and placing of all of the components to create a finished product. This cost can be reduced if a larger number of devices is ordered. All of the information needed to manufacture this PCB design, including schematics showing how all the parts are connected; Gerber files containing configuration parameters, aperture definitions, and coordinate information for the location of parts; and a list of the materials required is publicly available in the project's GitHub repository. To connect a computer with this version of the CIWS datalogger, a micro-USB cable is used (Figure 2.6.a highlights this connector).

2.2.4. Firmware

The firmware for the CIWS datalogger is organized using a traditional, C-like Arduino programming approach [37] and was developed within the Arduino IDE, which is open source and freely available for Windows, Mac, and Linux operating systems [19]. Traditionally, C/C++ code is separated into a declaration or header (.h) file and implementation or source (.cpp) file [37] that, when precompiled together, are known as a library [38]. For the CIWS datalogger, multiple libraries were developed. For each of these libraries, the header and implementation files are available in the project GitHub repository, along with documentation about the functions developed within each library,

including their output types, variables created, and data formats. Table 2.2 lists the library names and their main functions. In addition to the libraries listed on Table 2.2, other existing Arduino libraries were used in the firmware, including the serial peripheral interface (SPI) library [39], the SD library [40], the Wire library [41], and multiple "AVR Libc" libraries [42].

The main datalogger firmware file, "Firmware.ino," calls all of the libraries mentioned to operate and control the CIWS datalogger. It is the starting point of the firmware and contains six functions: 1) setup(), 2) loop(), 3) INT0_ISR(), 4) INT1_ISR(), 5) storeNewRecord(), and 6) bcdtobin(). The setup() function is called once when the device is powered, and the loop() function runs continuously as long as the microcontroller is powered. The functions INT0_ISR() and INT1_ISR() are both interrupt service routines. An interrupt service routine is executed when an event in hardware occurs. The main loop() function checks these flags, and if they are set, responds accordingly. This is good practice as interrupt service routines need to be kept as short as possible [38]. Table 2.3 lists the main objective of these 6 functions that comprise the firmware of the CIWS datalogger and are included in the Firmware.ino file.

2.2.5. User Interface

We developed an interactive user interface within the datalogger's firmware code that allows users to execute basic functions needed to configure and operate the datalogger along with managing and retrieving logged data files. Through this interface, the datalogger can be configured to work with different meter brands and sizes, and users can also create simple deployment information like a site identifier that makes it easier to identify and manage datasets after they have been collected. The principle of functioning,

threshold values used in the Schmitt Trigger function, and data transmission rates remain constant regardless of the brand and size of the meter selected. However, the data recording frequency and the meter's pulse resolution can be stored in the datalogger's memory to accurately specify the volume of water associated with every observed peak (or "pulse") in the meter's magnetic field. Configuring the pulse resolution for a specific meter allows the user to observe volumes of water registered without interrupting data collection, which is useful in verifying correct deployment of the sensor.

The user interface can be accessed through any serial console emulator or using the Arduino IDE. After connecting a computer to the datalogger using a USB Transistor-Transistor Logic (TTL) serial cable or a USB cable with a Future Technology Devices International (FTDI) breakout module (in the case of the prototype datalogger – a standard USB micro cable is used for the PCB version), clicking the Wake button on the datalogger shield will allow access to the interactive user interface in the serial console. The message "Logger: ready" will be displayed on the screen, and the list of commands in Table 2.4 will be accessible. For actions with multiple options, an interactive menu will be displayed allowing users to choose the desired action/configuration.

2.3. Calibration and Implementation

The volume of water that passes through the meter per pulse measured by the datalogger, referred to in this document as the pulse resolution of the meter, must be defined in order to obtain an accurate estimation of water usage. Meter manufacturers generally do not publish this information. We determined the pulse resolution for a number of different brands and sizes of commonly used residential meters using an experimental testing facility in a laboratory setting. We chose meters used extensively by

municipalities in our surrounding area (e.g., Logan City and Providence City, UT), although our testing and calibration methods could be applied to any magnetically driven water meter. Laboratory experiments were conducted with our datalogger to ensure that it can accurately measure water use at the different flow rates commonly experienced with residential water use. In these experiments, water was passed through the meters at multiple flow rates ranging from 4.43 LPM to 86.78 LPM. The register for each meter was manually read before and after each run, allowing the volume of water used in each run to be determined by the difference in manual meter readings. The volume registered by the meter was then divided by the total number of pulses observed by the CIWS datalogger during the experimental run to calculate the meter pulse resolution, R (Equation 3):

$$R = \frac{V_m}{P} \quad (3)$$

where Vm = the volume of water that passed through the meter (liters) and P = the number of pulses observed by the datalogger.

This process was repeated multiple times at increasing flow rates, each of which resulted in an estimate of the meter's pulse resolution. We also verified our results using more than one meter of the same size and brand plumbed in series with separate dataloggers on each one. Table 2.5 presents the results of one of the calibration experiments conducted, where six runs at different flowrates and durations were observed. In this experiment, two Neptune T-10 meters of 1 in size were installed in series on the same pipe; measuring the same flow. During each run, manual meter readings from both Neptune T-10s (named M1 and M2 in Table 2.5) were taken before and after running water and were used to calculate the volume of water that passed

through each meter. A CIWS datalogger was installed on each of these meters. DL1 was installed on M1, and DL2 was installed on M2. The pulses counted by each of these dataloggers were logged for each of the six runs conducted. The volumes read manually on the meters and the pulses observed by the dataloggers were then used to calculate the pulse resolution of each meter. Continued experiments demonstrated that the pulse resolution of each meter is consistent across meters of the same size and brand and across flow rates, which can be also observed in Table 2.5 by comparing the calculated pulse resolution of each meter.

Using this procedure, we determined that the pulse resolution for a 1 in Neptune T-10 meter is 0.1257 L/pulse, which we calculated as the average of the pulse resolution values for both meters across the six runs conducted. The standard deviation of this value was 8.757 x 10-5, and the coefficient of variation was 0.26%. These values demonstrate that while there is some variability in the pulse resolution values across the meters and runs, it is small enough that the calculated pulse resolution value can be used across meters of the same model/size and across flowrates.

Similar experiments were conducted using 5/8 in Neptune T-10 meters and 1 in and 5/8 in Bottom Load (BL) Master Meter meters. Table 2.6 lists the calibrated pulse resolution values for all of the meters we tested. These pulse resolution values were used in all field deployments of the datalogger. To calculate the volume observed by a CIWS datalogger installed on a meter, the number of pulses recorded by the datalogger is multiplied by its corresponding pulse resolution value. Introducing the meter pulse resolution in the datalogger's user interface allows the user to visualize the volume the CIWS datalogger has registered since logging started in units of gallons (to match the

register's units). This function is useful when deploying datalogger for the first time in a meter, to ensure the installation was successful. As mentioned in Section 2.2, the output file includes the number of pulses and not volume.

After the pulse resolution for each meter was calculated, the dataloggers were further tested for accuracy under different flow scenarios – e.g., for capturing rapidly changing flow rates and events of short duration, as these are common situations in residential settings. We conducted two additional laboratory experiments. In Experiment 1 (Figure 72.a), the flow rate through the meter was varied by opening a flow controlling valve. The flow rate through the meter was increased in steps without interrupting the flow between flow rate increases. We then conducted a separate experiment (Experiment 2, Figure 2.7b) where the flow rate through the meter was increased in steps, but the flow-controlling valve was quickly closed between each flow rate change. Manual readings of the meter's register were taken before, during, and after each experiment to compare the volume registered by the meter's register with the volume registered by the datalogger. The flow rate signature of Experiment 2 is similar to the signature of the experiments conducted to calibrate the device (Table 2.5).

In both experiments presented in Figure 2.7 a CIWS datalogger was installed on top of a 1 in meter and another on a 5/8 in meter. The registers for both meters were read at the beginning and end of each experiment. The volume registered by the meter and the CIWS datalogger were 777.86 L and 779.33 L, respectively, with a percent error of 0.19%. For the 1 in meter, the volumes were 782.48 L and 779.99 L for the meter and the CIWS datalogger, respectively, with a percent error of -0.57%. The difference in volume recorded by the 5/8 in (777.86 L) versus the 1 in (782.48) meters is not a subject of this

investigation as the percent errors for the CIWS datalogger were calculated relative to the meter on which they were installed. Our goal was to ensure that the datalogger accurately reflects the corresponding meter reading. If the accuracy of the meter itself is compromised, so will be the accuracy of the measurements made by our datalogger. In Experiment 2, meters were read after each incremental increase in flow. Table 2.7 shows the volumes read by each meter and its corresponding CIWS datalogger along with the percent error for each step. The maximum error observed in this experiment was -0.58% for the 1 in meter and 0.61% for the 5/8 in meter. Multiple experiments of the same kind were conducted, and the error was less than 1.5% in all cases, with values being similar to the ones presented in Table 2.7.

2.4. Field Deployments

In addition to our laboratory testing, the CIWS datalogger was installed on the water meter at 5 houses in the cities of Logan and Providence, Utah between May and September 2019 to evaluate its performance under field conditions, Table 2.8 shows the dates the datalogger was installed on each site. The evaluation of the CIWS datalogger performance presented in this section is based on battery life, accuracy of the measurements, errors, and limitations observed in the field. Analysis of the data collected and potential products that can be derived from it are included to illustrate potential applications of the CIWS datalogger. Data was collected using a temporal resolution of 4 seconds to allow the identification and posterior classification of events of short duration. The anonymized datasets and the code used for the computations presented in this section are publicly available [43].

2.4.1. Battery Life

In each deployment, the voltage was measured before and during data collection. A 12V, 10Ah battery was used in each deployment. Batteries were fully charged before deployment, up to 13 V and were replaced at approximately 20% of charge (~11.58 V). From fully charged to 20% of charge, at the 5 sites installed, the average discharge time was between 5 and 6 weeks indicating that the devices could reasonably be used to collect a month of 4 second temporal resolution data before the batteries have to be replaced.

2.4.2. Limitations and Errors

An obvious limitation for the installation and operation of the CIWS datalogger is related to the accessibility of the water meter. In areas around Logan, UT, meters are installed underground within a covered meter pit to ensure that they do not freeze during the winter. We encountered meter pits of depths ranging from 20 to 80 cm during field deployments, see Figure 2.5c for a reference. The depth of the meter in the pit affects its accessibility. In cases where the meter is within the reach of the person installing the magnetometer sensor, the process is straightforward and can be successfully completed by the installer in a few minutes. In cases where the meter is deep enough that it is not within easy reach of the installer, installation requires tools to extend the installer's reach. In our field experiments, we found that ensuring proper placement of the magnetometer sensor and the proper functioning of the CIWS datalogger required some trial and error for meters that could not be easily reached. In this scenario, multiple visits to a same location and constant supervision of the data collected were required to ensure the accuracy of collected data. Once the datalogger was installed and functioning properly on

top of a water meter, few data collection problems were observed. Early in our field trials, several of the magnetometer sensors failed, presumably because of the humidity in the meter pit. These failures caused the dataloggers to stop working and created errors in the data collected. We were able to fix this problem by covering the sensor and all the wiring connections to it with potting material. After this modification was done, we did not observe additional sensor failures. The Datalogging Shield and the Arduino Pro were protected from humidity inside the waterproof box, in which a desiccant pack was added for extra protection from humidity. Two meter pits were completely flooded during the data collection period. In one case, the magnetometer failed after the flood. In the other, the device continued to work after the pit was dry. In both cases, the datalogging components were kept relatively dry inside the box and continued to work after they were dried. No other environmental factor has been identified to affect the measuring process or damage the CIWS datalogger.

Another error observed during the field deployments was related to writing data to the CSV file. Some files became corrupt, and significant data loss occurred. We were unable to trace the origin of this error completely, although memory related errors on the Arduino or power failures were identified as possible causes. In an effort to diagnose this error, a test was conducted by logging data over an extended period of time on multiple devices in the laboratory. Memory and battery on the device were tracked and logged into a CSV file during these experiments using the SD Arduino library [40]. Memory errors were discarded as the cause as we observed that memory handling was effective. Power failures while writing data to the SD card using Arduino-based devices have been identified by other authors to cause data loss [26]. This cannot be discarded as the

potential cause of the errors, but we were unable to diagnose them because we did not observe any issues during our laboratory testing. Although we were unable to fully diagnose these errors, the data loss problem was corrected by introducing an update into the datalogger firmware that checks to see if the data saved to the CSV file has errors. If errors are found, the CSV file is ended and a new one is automatically initiated. This firmware modification was introduced close to the end of the field data collection period, and the error was not observed again after it was implemented. As an additional safety measure to limit potential data loss in case of reappearance of the error, the firmware was modified so that a new CSV file is started every day, whereas in the original firmware a single CSV file was used for data collection periods of any length.

The Datalogging Shield has a SD card memory slot, which can fit SD / MMC storage within a range of 32 MB to 32 GB [35], in this application only SD cards were used. A week of data collected with a four-second recording interval is approximately 5 MB in size. Thus, data storage does not constitute a significant limitation for the system's autonomy. In the field campaign conducted to test the device, old and new CSV files were kept on the 16 GB SD cards for redundancy purposes. A 4 GB SD card is sufficient to handle multiple years of data, even if a smaller time step for data collection is selected.

2.4.3. Accuracy

We performed our calibrations using newly purchased meters. While we acknowledge that the performance and accuracy of the meter itself may change over time [44,45], given that the meter's register and our datalogger use the same spinning magnet to quantify flow through the meter, volume observed by the datalogger will match the volume recorded by the meter's register regardless of the meter's age. The meters

observed during the field deployment were different brands, types and ages. Because the CIWS datalogger does not directly measure flow through the meter, it can only accurately count the magnetic pulses from the meter. Thus, the accuracy values reported in this section assume that water use calculated by subtracting manual readings of the meter's register reflect the true value. The meter's register at each site was read periodically to allow comparison between the volume registered by the meter and the totalized volume observed by the CIWS datalogger.

In the initial phase of the field deployment process, the accuracy observed was lower due to inexperience reading water meters and difficulties in the sensor installation process that were previously discussed. Figure 2.8 shows the percent difference between the volume registered by the meter, calculated as the difference between two consecutive readings of the meter's register, and the volume registered by the CIWS datalogger, calculated as the total number of recorded pulses multiplied by the pulse resolution of the meter. All points calculated are presented in Figure 2.8 using a violin plot to present the distribution of the error values we observed in the field with the CIWS datalogger.

During laboratory experiments, volume calculations using the CIWS datalogger were all within ± 1.5% of the meter volumes. In laboratory conditions, we had easy access to the meter and volumes were calculated over relatively short periods of time when compared to the field deployments. The values observed in Figure 2.8 for field deployments range between ± 5% although most are within the ± 1.5% range, similar to what we observed in laboratory experiments. Most values outside this range were caused by errors or sensor installation problems. Sites 2, 3, and 4 were the first three sites installed in the deployment process and served as experimental sites. At site 4, the meter

is beyond the reach of the installer, which represented a problem during the installation process. When there is water use occurring in the home at the same time logging is started or stopped, small differences between the manual meter readings and the datalogger totals can be introduced given that it is hard to read the meter's register when it is moving. As the installers became more experienced, most problems were addressed, evidenced by the significantly smaller errors for sites 1 and 5, which were installed on a later date than the other 3 (see Table 2.8 for specific dates). Deployment periods where the CSV file became corrupt on the SD card are not included in Figure 2.8 as the water usage data in these files was not reliable.

2.4.4. Water Use

A data recording interval of seconds, rather than minutes or longer, enables the use of end-use disaggregation algorithms [46], limits the volume of leaked water that can go undetected, and decreases the error in the estimation of peak demand [14]. Disaggregation of end uses is a complex process, particularly for overlapping events [47–49,10]. The purpose of the analysis presented in this section is not to produce a disaggregation/classification algorithm but rather to demonstrate the potential for using data collected using the CIWS datalogger as input to such algorithms for disaggregation and classification of water end uses. The first step in this type of analysis is to identify water use events, followed by disaggregation of simultaneous or overlapping events, and finally classification of individual events by type. For simplification, water use events in this analysis were identified as periods of non-zero flow – an event starts when the pulse count is larger than zero and ends when the pulse counts is zero again. This simplified approach for separating events may lead to uncertain results when there are continuous

leaks where the pulse count does not return to zero between events. It also does not consider overlapping events.

Using this simplified approach, we identified 5838 events at Site 1, 2133 at Site 2, 73975 at Site 3, 2647 at Site 4, and 3777 at Site 5. In order to identify and label some of these events, the homeowner at Site 1 was asked to log the start time and type of water use events in their home. Table 2.9 lists a sample of the events logged by the homeowner, and Figure 2.9 shows the data for the date and time of these events. Figure 2.9.a shows two subsequent faucet events. Flow rates in these events are similar, but duration is different. Figure 2.9b, c, and d represent a shower, clothes washer, and toilet flush event, respectively. The flow rate and duration of water use events depend on the characteristics and setting of the fixtures and on personal preferences of the user. The oscillations between flow rates within each of the events are related to the data recording interval and the pulse resolution of the meter. Because only discrete pulses can be counted, when flow rates are relatively constant (e.g., within an event) the pulse counts within adjacent recording intervals may vary by $\pm$ 1 pulse, leading to the flow rate behavior shown in Figure 2.9. The homeowner at Site 1 labeled multiple events; however, not all the events of the same kind exhibit the same pattern in terms of flow rate or duration. Duration, volume, and flow rate have been used to identify end uses of water by finding similarities among events using multiple methodologies, ranging from visual identification to machine learning algorithms [47,48,50].

Some events have characteristics that make their identification easier than others. Events with a duration of 4 seconds (the temporal resolution of the data collection) or less and only one pulse are likely to be leaks. Events with duration and/or flow rates much

larger than most events at a site are likely to be outdoor irrigation events. Figure 2.10a shows leaks occurring at Site 5 in a period of approximately 12 hours when no other water use occurs. If all of the events lasting 4 seconds (or less) are assumed to be leaks, we can calculate the leak rate, resulting in: 6.7 L/d at Site 1, 2.2 L/d at Site 2, 48.4 L/d at Site 3, 0.5 L/d at Site 4, and 7.2 L/d at Site 5. Other studies have found that leaks represent, on average, 13% of the indoor water use and average 64.3 L/d [6]. However, their definition of leaks includes more events than those described here, and indoor water use is not fully assessed in this analysis. Figure 2.10b presents an irrigation event (identified by its long duration and large volume) at Site 1. Irrigation events will exhibit a different pattern depending on whether a manual or automated system is used, the number of "zones" that are irrigated, and the number and type of sprinkler heads within each "zone." At site 1, for the event presented, an automated sprinkler irrigation system is used with five different irrigation zones. We also observed overlapping events occurring (between 12:15 and 12:30).

From the total number of events identified using our simplified procedure, 39% (Site 1), 48% (Site 2), 91% (Site 3), 18% (Site 4), and 89% (Site 5) were classified as leaks, resulting in 85% of the total combined events being leaks. Of the remaining 15%, approximately only 1.3% had a duration larger than 25 minutes, which are likely to be irrigation or overlapping events. None of the participants reported having a swimming pool. Most (96%) of the non-leak events had a duration less than 10 minutes and an average flow rate less than 15 LPM. Figure 2.11 presents the duration and volume for these events at each site. Volume and duration alone do not seem to discriminate different types of events. However, when adding the average flow rate (colors of the points), levels

between the events begin to appear. Events at these shorter durations should include toilets, which have similar volumes, durations, and flow rates along with faucets and showers, which will have different duration and volumes but occur at similar flow rates. Dishwasher and clothes washer events should also be similar, although varying designs, manufacturers, and available cycles would contribute to differences. The distribution of events also varied among the sites, which could indicate differences in personal preferences, water fixtures, or both. Although a rigorous clustering analysis is beyond the scope of this paper, Figure 2.11 shows that even 3 calculated event attributes begin to illustrate differences in event types. All of the event statistics, including those that not shown in the figure (e.g., mode flow rate, maximum flow rate), could be used as factors in a more sophisticated clustering approach to classify each event into end use categories.

The high temporal resolution data allows for calculation of other important characteristics of water use, such as instantaneous peak, hour peak, daily average, and daily maximum water use. Table 2.10 shows these statistics for each of the sites. Data collection periods are not concurrent for all the sites, which could explain some of the difference observed in the per capita daily average. Utah daily average water use is approximately 632 L per capita [51]. Water usage has a large seasonal component corresponding to landscape irrigation, and the data collection period from this experiment is not long enough to capture the annual variability. The majority of sites exhibited high variability among daily, hourly, and sub-hour resolution, adding supporting evidence to the claim that there is water use patterns masked in coarser temporal resolution data, such as hourly, daily, or monthly values. Values such as the peak hour maximum daily use are typically estimated from coarser temporal resolution data, or calculated based on typical

characteristics of a household, which adds significant uncertainty to the management and design of water networks [52].

2.5. Discussion and Conclusions

A low-cost, open source datalogger for collecting high temporal resolution water use data has been presented. The system can be installed on top of existing, analog, magnetically-driven water meters without affecting their functionality. The hardware components we used to prototype the datalogger are readily available, and assembly is straightforward using supporting materials provided. For potential users who do not want to assemble dataloggers from components, we have provided a PCB design that can be used for commercial manufacture and assembly. All of these materials, along with the code of the open-source firmware developed for operating the datalogger are open source and available, making it possible for any researcher to use our datalogger design as presented here or to modify our design to develop their own systems. The CIWS datalogger can potentially work with a wide range of magnetically-driven, positive displacement meters existing worldwide, although validation and calibration of the datalogger with each meter type and size is required before extending its application beyond the specific water meters we tested. The logger can be configured to collect data at any temporal resolution required, which represents an improvement over other existing commercial products. The cost of a CIWS datalogger is significantly lower than other existing technologies for collecting high temporal resolution water use data, does not disrupt the functioning of the meter on which it is installed, and does not require plumbing or disruptive installation. Although we performed our calibrations in a controlled laboratory setting, calibration for other meter types could be achieved by

following the methodology presented in this paper with the datalogger installed on meters deployed in the field.

Battery life constitutes the biggest limitation in terms of autonomy of the CIWS datalogger. Using a 12V, 10Ah battery, we were able to get between 5 to 6 weeks of autonomous operation with a data recording interval of four seconds. However, this battery life has been sufficient for our data collection needs and exceeds that of proprietary dataloggers used in past studies where data was collected at a coarser temporal resolution. Adding a solar panel or an additional power source can extend the autonomy of the logger. Our results indicate there is a learning curve for reading existing meters and for developing the skills needed to properly install the sensor. Accuracy increases once this period of learning has elapsed. The differences in the volumes observed by the CIWS datalogger and the meter's register indicate that the system presented is accurate within approximately 2% of the meter readings, when properly installed on the Neptune T-10 and Master Meter BL meters we tested under field conditions. For new installers, or when the meter pit is deep, this value can be as large as 5% of the meter reading. The CIWS datalogger should work with any magnetically driven meter, although further testing and calibration of the pulse resolution parameters in other meters is recommended before installing it on meters outside the ones presented here.

For simplicity, and given the small size of our field deployment, we used the original CSV files recorded by the dataloggers to obtain the results presented. The results of the field testing campaign we conducted over 5 months indicate that the datalogger is accurate, reliable, and that it can withstand the temperature and humidity conditions

existing in underground meter pits during different periods of the year. The low cost (≈ $150) and ease with which the datalogger can be deployed and used makes it ideal for residential water use studies that may have been cost-prohibitive in the past. Results from the field campaign demonstrate that the four-second data recording interval enables identification of daily patterns in water usage, peak timing and volumes, and accurate identification and characterization of individual end use events. Enabling disaggregation of end uses is key to fully understanding how water is used inside a monitored home and for identifying opportunities for conservation, forecasting demand, and determining how water use patterns may change over time in response to population growth, demographic shifts, and improvements in technology.

Collecting high temporal resolution data can be expensive, labor intensive, and disruptive. Newer smart meters can enable high temporal resolution data collection, but analog, positive displacement meters are still the most common meters in use within the U.S. The CIWS datalogger can enable high temporal resolution water use data collection on these existing meters. The CIWS datalogger can be used by utilities for educational interventions, for assessing the outcome of conservation campaigns, for generating more accurate water demand forecasting, and for data collection in any projects that require collection of high temporal resolution water use data. Given the volume of data produced, deploying the CIWS datalogger at a wider scale will require the development of a data management system consisting of cyberinfrastructure that can enable organized transformation of the data collected into useful information. Indeed, future work will include advancing this data management cyberinfrastructure along with implementation of WiFi and/or cellular communication capabilities for the CIWS datalogger, which may

enable automated transmission of data from the meter into a water user's home or to a water utility's office.

Hardware, Firmware, and Data Availability

All of the hardware modifications, parts, PCB design, firmware code, and supplemental materials are available in the GitHub repository for the project at https://github.com/UCHIC/CIWS-MWM-Logger. A snapshot of this repository at the time of this writing was created for archival purposes and published in Zenodo [53]. The repository contains separate folders for Hardware, Firmware, and Tools. All of the firmware libraries (.h and .cpp files) and supplemental firmware documentation are available in the Firmware folder. The Hardware folder contains additional images of the logger, the hardware design, layout, PCB design, and instructions to perform the hardware modifications described in this article. The anonymized data collected at each site from our field testing campaign and R scripts used to produce the results presented here are published in HydroShare [43]. The HydroShare resource also includes the sample of the raw data discussed in Section 3 with R scripts for its analysis along with the data from the laboratory experiments presented.

Author Contributions

All authors contributed to the conceptualization of the work presented, to selection of the methodology used, and in testing and evaluating prototypes of the sensor and datalogger presented. C.B. wrote the initial draft of the paper. J.S.H. and R.J.T contributed to review and editing. R.J.T led hardware prototyping and firmware development with contributions from C.B. and J.S.H. C.B. collected the field data and

curated it in HydroShare. J.S.H. provided project supervision and funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding

This research was funded by the United States National Science Foundation under grant number 1552444. Any opinions, findings, and conclusions or recommendations expressed are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Acknowledgments

The authors would like to acknowledge the support of the Utah Water Research Laboratory at Utah State University for facilitating the laboratory work and providing technical support. We would like to acknowledge Logan City and Providence City for their cooperation and support in the realization of the field campaigns. The authors would also like to acknowledge support from Nour Atallah in the field deployment campaign and Daniel Henshaw for his contribution in the hardware and firmware development. We also want to acknowledge and thank the owners of the residential homes that participated in the data collection campaign.

Conflicts of Interest

The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

REFERENCES

1. Boyle, T.; Giurco, D.; Mukheibir, P.; Liu, A.; Moy, C.; White, S.; Stewart, R. Intelligent metering for urban water: A review. Water 2013, 5, 1052, https://doi.org/10.3390/w5031052.

2. Cominola, A.; Giuliani, M.; Piga, D.; Castelletti, A.; Rizzoli, A.E. Benefits and challenges of using smart meters for advancing residential water demand modeling and management: A review. Environmental Modelling & Software 2015, 72, 198–214, https://doi.org/10.1016/j.envsoft.2015.07.012.

3. Cardell-Oliver, R.; Wang, J.; Gigney, H. Smart meter analytics to pinpoint opportunities for reducing household water use. Journal of Water Resources Planning and Management 2016, 142, 04016007, https://doi.org/10.1061/(ASCE)WR.1943-5452.0000634.

4. Horsburgh, J.S.; Leonardo, M.E.; Abdallah, A.M.; Rosenberg, D.E. Measuring water use, conservation, and differences by gender using an inexpensive, high frequency metering system. Environmental Modelling & Software 2017, 96, 83–94, https://doi.org/10.1016/j.envsoft.2017.06.035.

5. Sønderlund, A.L.; Smith, J.R.; Hutton, C.; Kapelan, Z. Using smart meters for household water consumption feedback: Knowns and unknowns. Procedia Engineering 2014, 89, 990–997, https://doi.org/–10.1016/j.proeng.2014.11.216.

6. DeOreo, W.B.; Mayer, P.W.; Dziegielewski, B.; Kiefer, J. Residential End Uses of Water, Version 2; Water Research Foundation, 2016;https://www.waterrf.org/research/projects/residential-end-uses-water-version-2.

7. Beal, C.; Stewart, R.A. South East Queensland Residential End Use Study: Final Report; Urban Water Security Research Alliance, 2011; http://www.urbanwateralliance.org.au/publications/UWSRA-tr47.pdf

8. F.S. Brainard & Company Model 100EL and 100AF Flow Recorders Available online: https://meter-master.com/product/model-100el-100af/ (accessed on Mar 20, 2020).

9. Froehlich, J.; C. Larson, E.; Campbell, T.; Haggerty, C.; Fogarty, J.; N. Patel, S. HydroSense: Infrastructure-mediated single-point sensing of whole-home water activity. In Proceedings of the 11th international conference on Ubiquitous computing; Orlando, Florida, USA, 2009, https://doi.org/10.1145/–1620545.1620581.

10. Srinivasan, V.; Stankovic, J.A.; Whitehouse, K. WaterSense: water flow disaggregation using motion sensors. In Proceedings of the Third ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings; Seattle, Washington, USA, 2011, https://doi.org/10.1145/2434020.2434026.

11. Chen, J.; Kam, A.H.; Zhang, J.; Liu, N.; Shue, L. Bathroom activity monitoring based on sound.; Springer Berlin Heidelberg: Munich, Germany, 2005; pp. 47–61, https://doi.org/10.1007/11428572_4.

12. Fogarty, J.; Au, C.; Hudson, S.E. Sensing from the basement: a feasibility study of unobtrusive and low-cost home activity recognition. In Proceedings of the 19th annual ACM symposium on User interface software and technology; ACM: Montreux, Switzerland, 2006, https://doi.org/10.1145/1166253.1166269.

13. PHYN Your Water Like You've Never Seen It Available online: https://www.phyn.com/technology/ (accessed on Apr 5, 2020).

14. Cominola, A.; Giuliani, M.; Castelletti, A.; Rosenberg, D.E.; Abdallah, A.M. Implications of data sampling resolution on water use simulation, end-use disaggregation, and demand management. Environmental Modelling & Software 2018, 102, 199–212, https://doi.org/10.1016/j.envsoft.2017.11.022.

15. Flume Inc. Protect Your Home Available online: https://www.flumetech.com/ (accessed on Apr 15, 2020).

16. Sadler, J.M.; Ames, D.P.; Khattar, R. A recipe for standards-based data sharing using open source software and low-cost electronics. Journal of Hydroinformatics 2016, 18, 185–197, https://doi.org/10.2166/–hydro.2015.092.

17. Horsburgh, J.S.; Caraballo, J.; Ramírez, M.; Aufdenkampe, A.K.; Arscott, D.B.; Damiano, S.G. Low-cost, open-source, and low-power: but what to do with the data? Frontiers in Earth Science 2019, 7, https://doi.org/10.3389/feart.2019.00067.

18. Fisher, D.K.; Gould, P.J. Open-source hardware is a low-cost alternative for scientific instrumentation and research. Modern Instrumentation 2012, 1, https://doi.org/10.4236/mi.2012.12002.

19. Arduino - Software Available online: https://www.arduino.cc/en/main/software (accessed on Feb 21, 2020).

20. Rao, A.S.; Marshall, S.; Gubbi, J.; Palaniswami, M.; Sinnott, R.; Pettigrovet, V. Design of low-cost autonomous water quality monitoring system. In Proceedings of the 2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI); Mysore, India, 2013; pp. 14–19, https://doi.org/10.1109/ICACCI.2013.6637139.

21. Kuznetsov, S.; Paulos, E. UpStream: motivating water conservation with low-cost water flow sensing and persuasive displays. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; ACM: Atlanta, Georgia, USA, 2010, https://doi.org/10.1145/1753326.1753604.

22. Agrawal, N.; Singhal, S. Smart drip irrigation system using raspberry pi and arduino. In Proceedings of the International Conference on Computing, Communication & Automation; Noida, India, 2015; pp. 928–932, https://doi.org/10.1109/CCAA.2015.7148526.

23. Vellidis, G.; Tucker, M.; Perry, C.; Kvien, C.; Bednarz, C. A real-time wireless smart sensor array for scheduling irrigation. Computers and electronics in agriculture 2008, v. 61, 44–50–2008 v.61 no.1, http://dx.doi.org/10.1016/j.compag.2007.05.009.

24. Dai, B.; Chen, R.; Yang, W. Using Arduino to develop a Bluetooth electronic scale for water intake. In Proceedings of the 2016 International Symposium on Computer, Consumer and Control (IS3C); Xi'an, China, 2016; pp. 751–754, https://doi.org/10.1109/IS3C.2016.192.

25. Justo, P.D.; Gertz, E. Environmental Monitoring with Arduino 2012, https://learning.oreilly.com/library/–view/environmental-monitoring-with/9781449328603/.

26. Beddows, P.A.; Mallon, E.K. Cave Pearl Data Logger: A flexible Arduino-based logging platform for long-term monitoring in harsh environments. Sensors (Basel, Switzerland) 2018, 18, 530, https://doi.org–/10.3390/s18020530.

27. Cressey, D. The DIY electronics transforming research. Nature 2017, 544, 125–126, https://doi.org–/10.1038/544125a.

28. Tan, L.; Jiang, J. Chapter 6 - Digital Signal Processing Systems, Basic Filtering Types, and Digital Filter Realizations. In Digital Signal Processing (Third Edition); Tan, L., Jiang, J., Eds.; Academic Press, 2019; pp. 173–228 ISBN 978-0-12-815071-9.

29. Grout, I. Chapter 7 - Introduction to Digital Signal Processing. In Digital Systems Design with FPGAs and CPLDs; Grout, I., Ed.; Newnes: Burlington, 2008; pp. 475–536 ISBN 978-0-7506-8397-5.

30. Tan, L.; Jiang, J. Chapter 8 - Infinite Impulse Response Filter Design. In Digital Signal Processing (Third Edition); Tan, L., Jiang, J., Eds.; Academic Press, 2019; pp. 315–419 ISBN 978-0-12-815071-9.

31. Smith III, J.O. Introduction to Digital Filters: with Audio Applications; Stanford University: Center for Computer Research in Music and Acoustics (CCRMA); ISBN 978-0-9745607-1-7.

32. Lyon, R.G. Chapter 13 - Digital Signal Processing Tricks. In Understanding Digital Signal Processing (Third Edition); Pearson Education, 2011; ISBN 13: 978-0-13-702741-5.

33. STMicroelectronics LIS3MDL. Digital output magnetic sensor: ultra-low-power, high-performance 3-axis magnetometer Available online: https://www.st.com/resource/en/datasheet/lis3mdl.pdf (accessed on Apr 15, 2020).

34. Arduino Arduino PRO Available online: https://store.arduino.cc/usa/arduino-pro (accessed on Apr 15, 2020).

35. Adafruit Industries Adafruit Assembled Data Logging shield for Arduino Available online: https://www.adafruit.com/product/1141 (accessed on Feb 15, 2020).

36. Smith, S.W. Chapter 3: ADC and DAC. In The Scientist and Engineer's Guide to Digital Signal Processing; 1998.

37. Mitchell, S.R.; Guntheroth, K.; Green, D. The C++ Workshop. A new, interactive approach to learning C++ 2020, https://learning.oreilly.com/library/view/the-c-workshop/9781839216626/C14195_01_Final_SP_ePub.xhtml.

38. Purdum, J. Beginning C for Arduino. Second Edition 2015, https://learning.oreilly.com/library/–view/beginning-c-for/9781484209400/A334771_2_En_1_Chapter.html.

39. Arduino - SPI Available online: https://www.arduino.cc/en/reference/SPI (accessed on Mar 30, 2020).

40. Arduino - SD Available online: https://www.arduino.cc/en/reference/SD (accessed on Mar 30, 2020).

41. Arduino - Wire Available online: https://www.arduino.cc/en/reference/wire (accessed on Mar 30, 2020).

42. AVR Libc Home Page Available online: https://www.nongnu.org/avr-libc/ (accessed on Mar 30, 2020).

43. Bastidas, C.; Horsburgh, J.S. Supporting data for "A low-cost, open source, monitoring system for collecting high-resolution water use data on magnetically-driven residential water meters." HydroShare 2020, http://www.hydroshare.org/resource/1e752471bcf24f6da0f2e9b4df9a3d2f.

44. Neilsen, M.A.; Barfuss, S.L.; Johnson, M.C. Off-the-shelf accuracies of residential water meters. AWWA 2011, 103, 48–55, https://doi.org/10.1002/j.1551-8833.2011.tb11531.x.

45. Barfuss, S.L.; Johnson, M.C.; Nielson, M.A. Accuracy of In-Service Water Meters at Low and High Flow Rates; Water Research Foundation and U.S. Environmental Protection Agency, Denver, CO., 2011, https://www.waterrf.org/resource/accuracy-service-water-meters-low-and-high-flow-rates.

46. Nguyen, K.A.; Zhang, H.; Stewart, R.A. Development of an intelligent model to categorise residential water end use events. Journal of Hydro-environment Research 2013, 7, 182–201, https://doi.org/10.1016/j.jh–er.2013.02.004.

47. Pastor-Jabaloyes, L.; Arregui, F.J.; Cobacho, R. Water end use disaggregation based on soft computing techniques. Water 2018, 10, 21, https://doi.org/10.3390/w10010046.

48. Nguyen, K.A.; Stewart, R.A.; Zhang, H. An autonomous and intelligent expert system for residential water end-use classification. Expert Systems with Applications 2014, 41, 342–356, https://doi.org/10.1016–/j.eswa.2013.07.049.

49. Nguyen, K.A.; Stewart, R.A.; Zhang, H. An intelligent pattern recognition model to automate the categorisation of residential water end-use events. Environmental Modelling & Software 2013, 47, 108–127, https://doi.org/10.1016/j.envsoft.2013.05.002.

50. Aquacraft Trace Wizard description Available online: http://www.aquacraft.com/downloads/trace-wizard-description/ (accessed on Jan 15, 2020).

51. Dieter, C.A.; Maupin, M.A.; Caldwell, R.R.; Harris, M.A.; Ivahnenko, T.I.; Lovelace, J.K.; Barber, N.L.; Linsey, K.S.; Survey, U.S.G. Estimated use of water in the United States in 2015; Reston, VA, 2018; p. 76, http://pubs.er.usgs.gov/publication/cir1441.

52. Cole, G.; Stewart, R.A. Smart meter enabled disaggregation of urban peak water demand: precursor to effective urban water planning. Urban Water Journal 2013, 10, 174–194, https://doi.org/10.1080/1573–062X.2012.716446.

53. Horsburgh, J.S.; Tracy, J.; Bastidas, C. UCHIC/CIWS-MWM-Logger: Version 1.1.0. 2020, https://doi.org–/10.5281/zenodo.3832260.

Tables

Table 2.1. Parts required and costs to build a CIWS datalogger.

Part	Cost/ logger	Vendor
3.3 V 8 MHz Arduino Pro (ATmega328p Board)	$15.95	Sparkfun
Datalogging Shield	$15.95	Adafruit
LIS3MDL Magnetometer + Breakout Board	$4.95	Pololu
Male ICSP Headers (2x15 block)	$0.65	Mouser
Adafruit SOT23 Breakout Pack	$0.99	Mouser
Stripboard	$1.43	Mouser
Anderson Powerpole Connectors	$1.30	Amazon
MCP1703 3.3 V Regulator	$0.55	Mouser
Fuse	$0.17	All-Electronics
In-Line Fuse Holder	$0.45	All Electronics
Serial Extender Housing Pack	$0.16	Pololu
Screws	$0.17	Mouser
Nuts	$0.12	Mouser
Box Kit	$2.10	Mouser
Cable Glands	$1.95	Mouser
Battery Connectors	$1.08	Mouser
10 kOhm Resistor	$0.50	Mouser
1 uF Ceramic Capacitors	$0.92	Mouser
0.01 uF Ceramic Capacitor	$0.25	Mouser
Button	$0.16	Mouser
Spacers	$0.20	Mouser
2-Position Terminal Block	$0.90	USU ECE Store
5-Position Terminal Block	$4.11	Mouser
Wire (Battery – Positive)	$0.15	Mouser
Wire (Battery – Negative)	$0.15	Mouser
Wire (prototype board soldering)	$0.80	Mouser
Coin Cell (614-CR1220.IB)	$1.14	Mouser
Micro SD Cards with adapters	$9.95	Mouser
Pelican 1150 Case (with foam)	$31.96	Amazon
12 V 10 Ah Duracell Battery	$39.99	Batteries + Bulbs
Clasp	$0.739	Amazon
5-Conductor Cable	$9.11	Mouser
Female Headers (36 pin)	$0.59	Mouser
Strap Set: Gear Strapz (+5 Clasps)	$1.52	Amazon
Total Cost	$151.19	

Table 2.2. Code libraries developed for the CIWS datalogger firmware.

Library	Main objective
state	Initializing and keeping track of: • Whether the values registered have gone above or below the threshold values to trigger a pulse. • The number of pulses in the current time step, time stamp, and record number. • Whether the device is logging or not. • Whether the serial interface is active or not. • Whether the SD card has been initialized or not. • Whether the magnetometer data is ready or not. • Whether the configuration data is valid or not, including Site #, Datalogger ID, meter pulse resolution, and recording interval. • The current filename of the CSV output file. • Variables used for processing the magnetometer signal and variables used in multiple functions in other libraries.
detectPeaks	Detecting peaks in the magnetic field. Defines functions that are responsible for filtering the raw data from the magnetometer and applying the Schmitt Trigger.
magnetometer	Managing the LIS3MDL magnetometer. Defines the functions responsible for initializing and reading data from the LIS3MDL magnetometer.
RTC_PCF8523	Managing the RTC. Defines functions that are responsible for transferring data to and from the RTC, including configuration data such as the interval of data collection. Reads the date and time from the RTC that is printed in the output file.
configuration	Managing configuration data in the Electrically Erasable Programmable Read-Only Memory (EEPROM). Defines functions that: • Check if the EEPROM has configuration data. • Verify the correct functioning of the EEPROM. • Configure the writing, reading and the loading of data into the EEPROM data register.
powerSleep	Optimization of components for power management. Defines functions that set the Arduino Pro into standby mode, a low power consumption mode, and disable all the peripherals as described in the hardware section.
handleSerial	Operating the user interface. Define all the functions that allow the functioning and interaction with the user interface, described in the next section.

Table 2.3. Functions executed by the CIWS datalogger Firmware.ino file and main objective.

Function	Main objective
setup()	Executes the following tasks: • Initializes the system state data structure. • Initializes General-Purpose Input Output (GPIO) pins. • Initializes the magnetometer. • Initializes the real-time clock. • Sets up the magnetometer and real-time clock interrupt handlers. • Checks that the datalogger has valid configuration data. • Disables the clock for all unused Arduino peripherals. • Opens the serial interface if the serial activation button is pressed.
loop()	The datalogger firmware's main loop that performs the following actions: • Check if the serial activation button is pressed. • Run the serial menu. • Check if a data recording interval has passed. • Check if magnetometer data is ready. • Process incoming data to count peaks. • Starts a new CSV file everyday while logging data.
INT0_ISR()	An interrupt service routine that executes when the voltage on the Arduino's digital pin 2 transitions from low to high. The voltage signal on digital pin 2 is controlled by the magnetometer. When the magnetometer has new data ready to report, it sets the voltage on pin 2 high, causing INT0_ISR() to execute, indicating the main program to read data from the magnetometer sensor.
INT1_ISR()	An interrupt service routine that executes when the voltage on the Arduino's digital pin 3 transitions from high to low. The voltage signal on digital pin 3 is controlled by the RTC. When the interval of time has elapsed, the RTC sets the voltage on pin 3 low, causing INT1_ISR() to execute indicating the main program to read a new datetime from the RTC and store data in the SD card.
storeNewRecord()	Primarily responsible for storing data records to the datalogger's SD Card by performing the following actions: • Gather date and time information. • Activate the Serial Peripheral Interface (SPI) module's clock. • Open a CSV file on the SD Card. • Print the following fields separated by commas: Timestamp, Record Number, Pulse Count. • Close the file.

Function	Main objective
	• Increment the record number. • Deactivate the SPI module's clock. • Compares the number of bytes written with the number of bytes attempted and starts a new file if there are differences between these two values.
bcdtobin()	Responsible for converting the Binary Coded Decimal (BCD) data from the RTC into standard binary data. Takes as input a BCD value and a bitmask corresponding to the RTC register from which the BCD value came from. This conversion is accomplished by multiplying the top 4 bits by ten and adding that number to the bottom four bits.

Table 2.4. List of commands, actions, and brief description of the main functions available in the CIWS datalogger.

Command	Action	Description
c	Clean the SD card	Access an interactive menu that allows the user to delete files from the SD card.
d	View date/time	Display current date and time on the device.
e	Exit the serial interface	Exits the serial interface and puts the device back to sleep.
E	Eject the SD card	Allows the user to safely extract the SD card from the device for data transferring.
g	Set device configuration	Enter the configuration mode – site, file number, and meter pulse resolution are entered within this command.
h	Display help	Displays all of the configuration options available.
i	Initialize the SD card	Initializes the SD card, which must happen prior to starting to log data.
l	List all the files on the SD card	Lists all of the files currently on the SD card.
p	Print configuration data	Print the site number, file number, and meter pulse resolution.
R	Diagnose the RTC	Check configuration data of the RTC.
t	Change the time interval for data collection.	Allows to set the time interval for data collection. Values can range from 1 second to more than 15 minutes.
s	Start datalogging	Starts logging data using the configuration data and date/time on the device.
S	Stop datalogging	Stops the data logging process.
u	Update date/time	Allows the user to change the date and time on the device.
w	Print water flow data	Displays the volume of water measured by the device since logging started.

Table 2.5. Calibration results for the CIWS datalogger using two 1 in Neptune T-10 meters.

RUN		RUN 1	RUN 2	RUN 3	RUN 4	RUN 5	RUN 6
Duration (min)		7	3	3	2	2	2
Volume (L)	M1	30.9	32.2	44.6	59.1	95.3	173.5
	M2	30.9	32.0	44.6	59.1	95.2	173.7
Pulses	DL1	246	256	355	470	759	1382
	DL2	247	255	355	470	758	1372
Average Flow Rate (LPM)*		4.4	10.7	14.9	29.5	47.6	86.8
Pulse-Resolution (L/pulse)	M1	0.125565	0.125835	0.125505	0.125643	0.125582	0.125532
	M2	0.125210	0.125587	0.125612	0.125643	0.125598	0.126613

* The average flow rate is calculated using the average volume between the two meters used.

Table 2.6. Pulse resolution values resulting from calibration of the CIWS datalogger in the most popular meter models in Logan and Providence Cities, Utah.

Meter Brand and Model	Size (in)	Pulse resolution (L/pulse)
Neptune T-10	1	0.1257
	5/8	0.0329
Master Meter BL	1	0.1575
	3/4	0.0957

Table 2.7. Results from Experiment 2 for the 1 in and 5/8 in Master Meter.

Time	1 in Meter Volumes (L)			5/8 in Meter Volumes (L)		
	Meter	Datalogger	Error	Meter	Datalogger	Error
10:15	6.78	6.79	0.14%	6.89	6.92	0.37%
10:18	20.37	20.23	-0.66%	20.21	20.28	0.35%
10:21	9.43	9.42	-0.01%	9.43	9.48	0.61%
10:24	23.09	23.12	0.13%	23.05	23.12	0.27%
10:27	35.05	35.06	0.02%	34.94	35.14	0.56%
10:30	40.58	40.71	0.33%	40.43	40.67	0.59%
10:33	50.38	50.39	0.01%	50.12	50.32	0.39%
10:36	66.43	66.47	0.06%	66.06	66.29	0.35%
10:39	80.02	79.92	-0.13%	79.49	79.49	0.00%
10:42	94.71	94.62	-0.09%	94.33	93.98	-0.37%
10:45	102.24	102.16	-0.08%	101.64	101.42	-0.21%
10:48	115.08	115.23	0.13%	114.55	114.23	-0.27%
10:51	121.97	121.26	-0.58%	120.68	120.36	-0.27%
10:54	125.11	125.79	0.54%	125.22	124.97	-0.20%
10:57	135.56	135.59	0.02%	134.95	134.71	-0.18%
11:00	149.49	149.54	0.03%	148.84	148.71	-0.09%
11:03	175.98	176.18	0.11%	175.49	175.28	-0.12%

Table 2.8. Sites where a CIWS datalogger was installed, meter characteristics, and data collection period.

Site	Start date	End date	Meter Brand	Size	City
1	9/20/19	10/15/19	Master Meter	1”	Providence
2	5/31/19	7/17/19	Neptune	1”	Logan
3	5/28/19	7/9/19	Neptune	5/8”	Logan
4	5/17/19	6/17/19	Neptune	5/8”	Logan
5	6/3/19	7/17/19	Neptune	1”	Logan

Table 2.9. Events logged by the homeowner at Site 1 and main characteristics calculated from the high temporal resolution data collected.

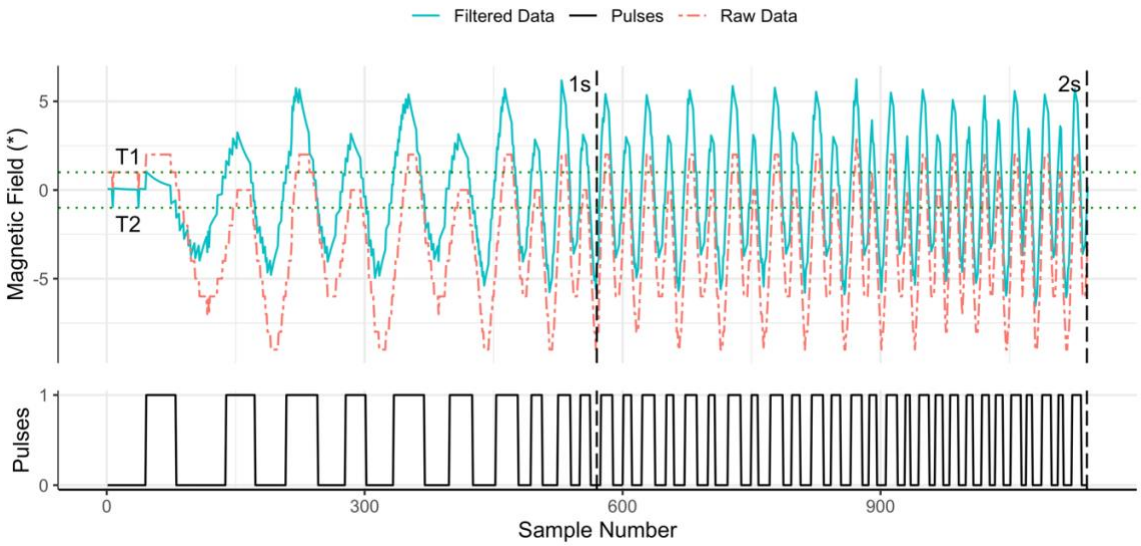
Date	Time	Type	Duration (min)	Volume (L)	Average flow rate (LPM)	Maximum flow rate (LPM)	Mode flow rate (LPM)
2019-10-09	17:05:18	Faucet	1.07	4.24	3.98	4.73	4.73
2019-10-09	17:07:40	Faucet	0.80	2.99	3.75	4.73	4.73
2019-10-14	10:58:19	Shower	10.60	67.76	6.40	11.81	7.08
2019-10-14	17:15:10	Clothes washer	7.60	109.02	14.35	21.27	18.89
2019-10-12	11:09:55	Toilet	1	10.41	10.41	16.54	14.20

Table 2.10. Water usage statistics calculated from the data collected.

Site	DAWU (L)	PCDU (L)	DSD (L)	DMWU (L)	MaxDU (L)	Peak—Hour (L)	Peak——Minute (L)	PeakInt (LPM)
1	1,630	326	2,188	1,283	12,979	5,351	131	158
2	3,308	1,654	4,048	2,331	13,106	3,872	134	83
3	1,145	573	673	1,102	2,458	667	30	34
4	897	224	1,316	389	5,311	1,718	56	64
5	14,512	7,256	10,667	12,084	44,225	10,506	190	128

DAU: Daily average water use.
PCDU: Per capita average daily use.
DSD: Standard deviation of daily use.
DMWU: Daily median water use.
MaxDU: Maximum daily water use.
PeakHour: Maximin hourly water use.
PeakMinute: Maximum minute water use.
PeakInt: Instantaneous peak, over every 4 seconds interval, in LPM.

Figures



(*) The magnetic field is expressed as an integer that varies from -128 to 127 for the range assigned (± 4 Gauss). This value was not scaled for this application.

Figure 2.1. Pulse detection process. The red dashed line represents the raw data collected by the magnetometer. The blue line represents the filtered signal (using Equation (2)), and the black line is the output of the digital Schmitt Trigger, the pulses that are counted and logged by the system. T1 and T2, the green dotted lines at 1 and −1, are the two thresholds used in the Schmitt Trigger function. At the sampling resolution selected, the CIWS datalogger collects 560−570 readings every second. The 1 and 2 s vertical dashed lines are superimposed on the plot at the location of the sample number that corresponds to those time steps.

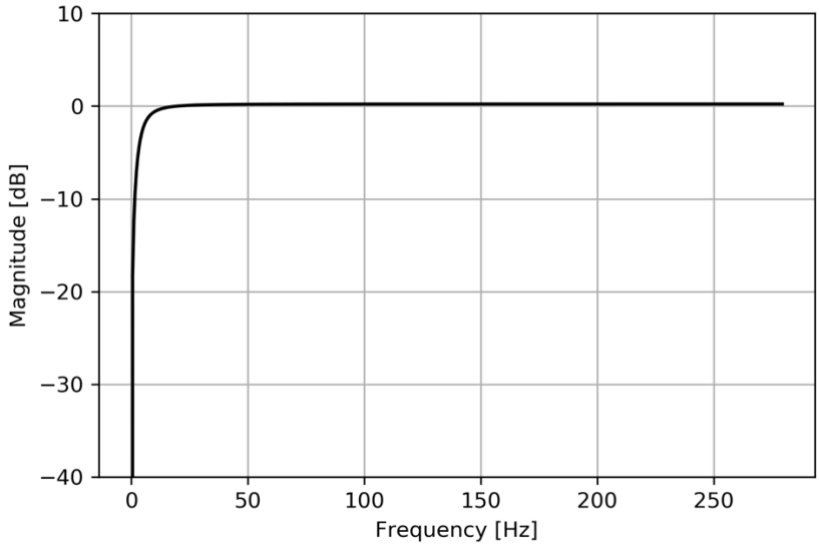


Figure 2.2. Frequency response of the IIR filter designed for this application.

```
Site #: 001
Datalogger ID #: 001
Meter Resolution: 0.033
Time,Record,Pulses
"20-3-16 17:53:34",1,1
"20-3-16 17:53:38",2,0
"20-3-16 17:53:42",3,0
"20-3-16 17:53:46",4,0
"20-3-16 17:53:50",5,38
"20-3-16 17:53:54",6,117
"20-3-16 17:53:58",7,119
"20-3-16 17:54:02",8,118
"20-3-16 17:54:06",9,119
"20-3-16 17:54:10",10,119
```

Figure 2.3. Sample output from the CIWS datalogger collecting data with a 4 second time interval. The data is stored in a CSV file, which is easily operable in multiple platforms and software.

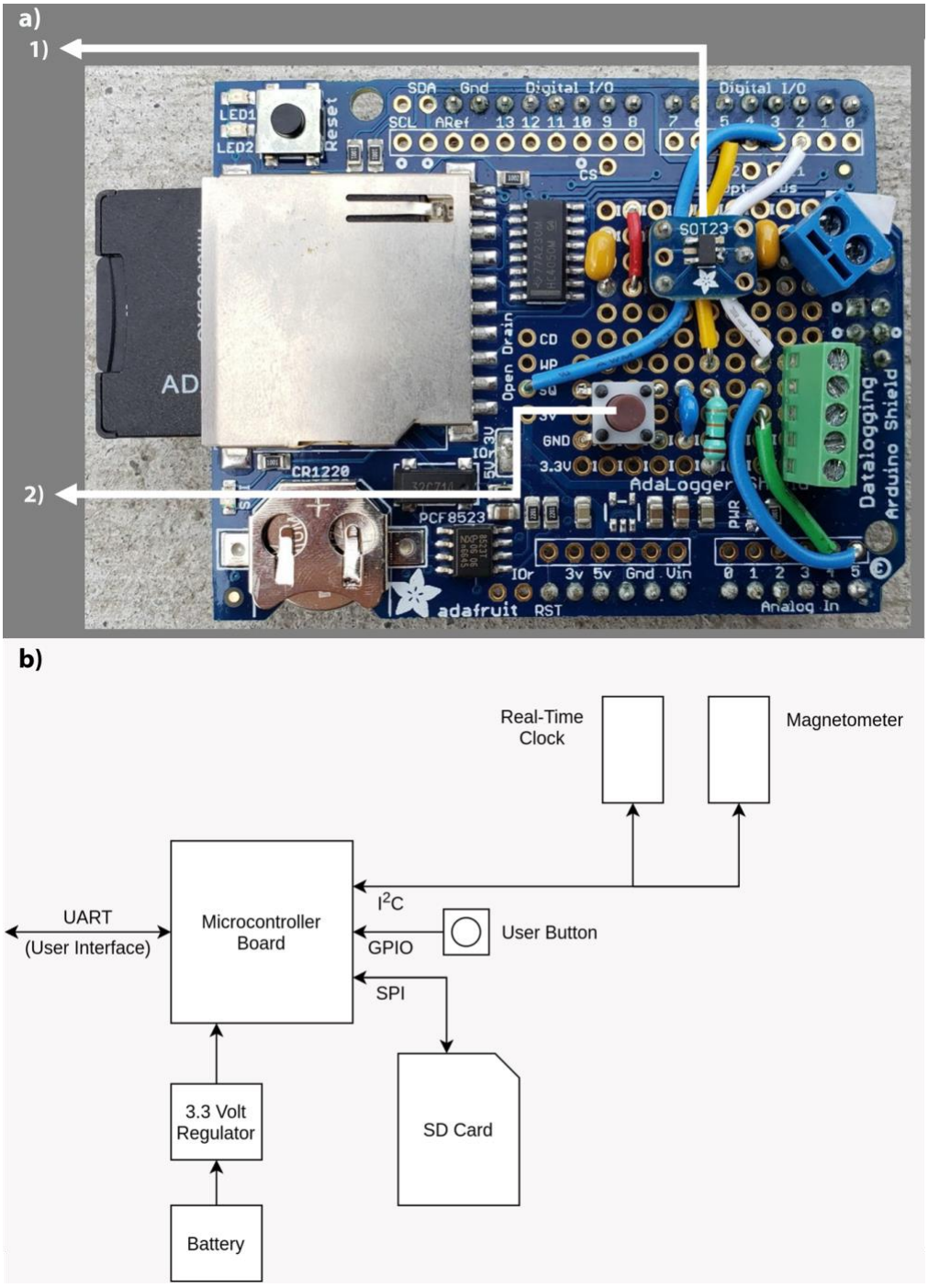


Figure 2. 4. a) Datalogging shield modified for the CIWS datalogger. Main external components added include: (1) more efficient power regulator installed on the Adafruit SOT23 Breakout Pack; and (2) wake power button installed to access the user interface when desired. Other modifications and components can be observed in this figure, including the SD card, pin connections described, terminal blocks, resistors, and capacitors added to the shield. b) Block diagram of the connections between main

components in the CIWS datalogger. A full diagram of connections is available on the project's GitHub.

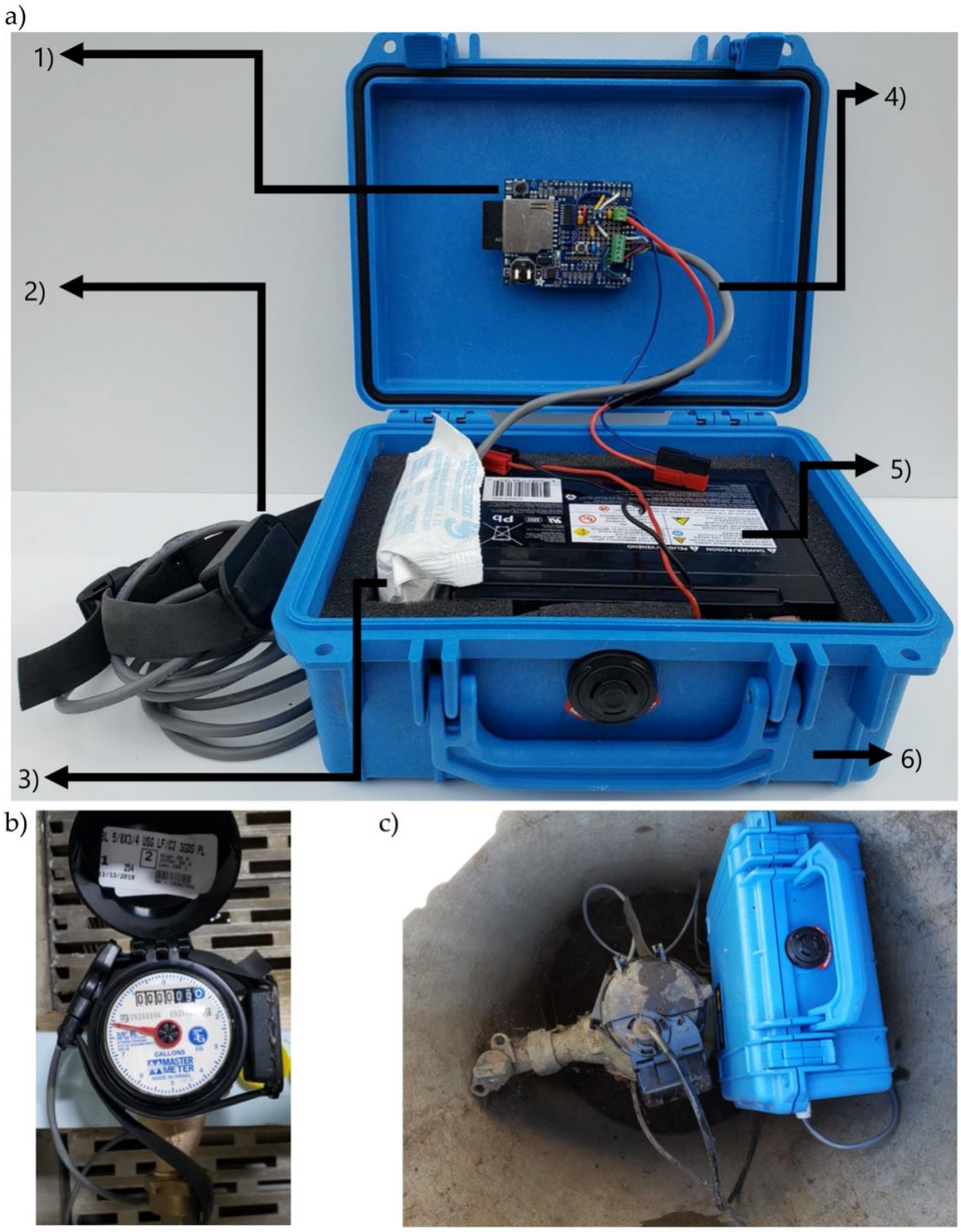


Figure 2.5. Assembled device ready for deployment. a) Main components: 1) Arduino Pro and Datalogging shield coupled together, 2) potted and encapsulated LIS3MDL magnetometer sensor, 3) desiccant pack, 4) cable connecting the magnetometer and the shield, 5) 12V 10 Ah battery, 6) waterproof box. b) Example of the sensor configuration when it is installed on a 5/8 in Master Meter meter (the orientation of the sensor does not affect the functioning of the device). c) Deployment of a CIWS datalogger on a meter pit (on a 1 in Neptune T-10 meter).

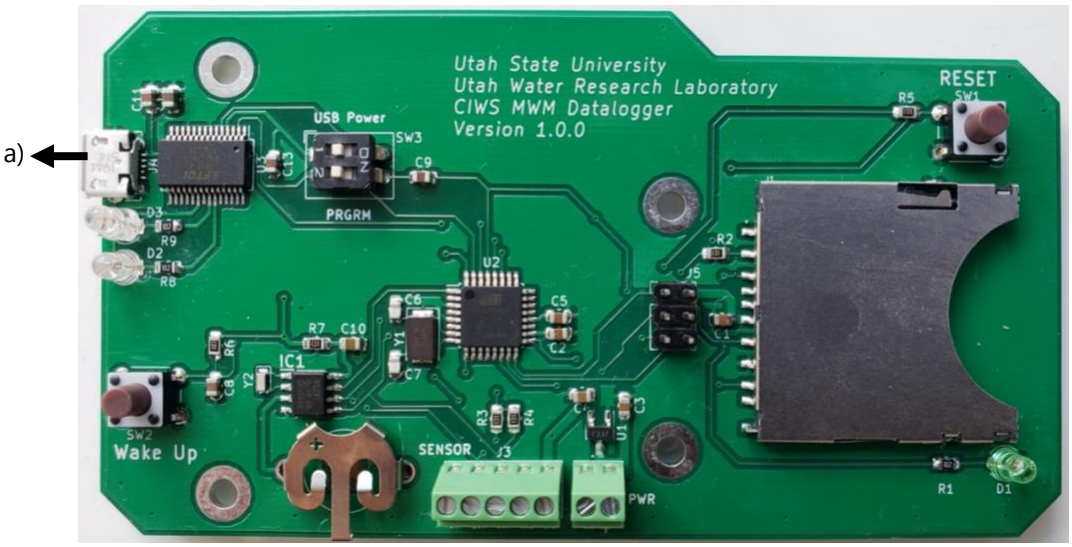


Figure 2.6. CIWS datalogger built using the PCB design developed. a) Micro-USB connector. Other important components of the datalogger are also observed: the wake up and reset buttons, the sensor and power connections are easily identifiable in the figure by reading the inscriptions included. The SD card adapter, the coin cell battery holder, and LED light are also visible.

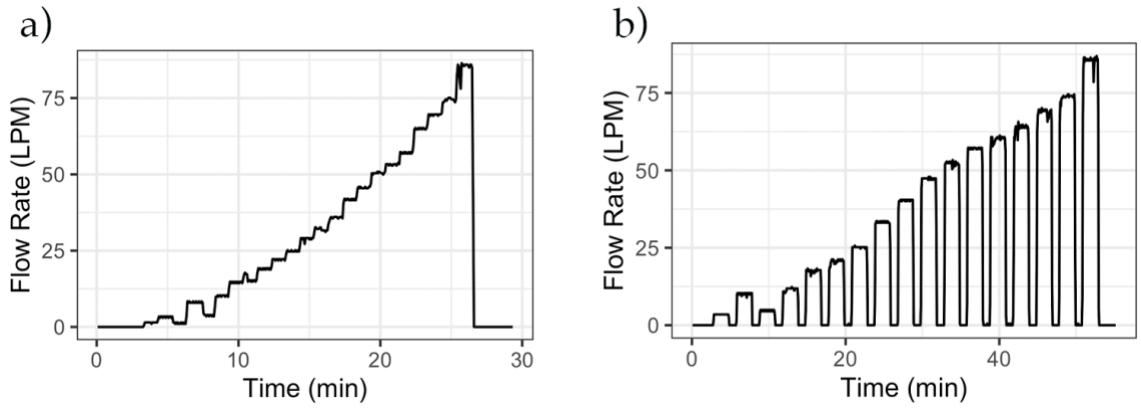


Figure 2.7. Flow signatures of the experiments used to verify the functioning of the CIWS datalogger. a) increasing the flow rate gradually; b) increasing the flow rate with intervals of no flow. Data shown are for a 5/8 in Neptune T-10 meter.

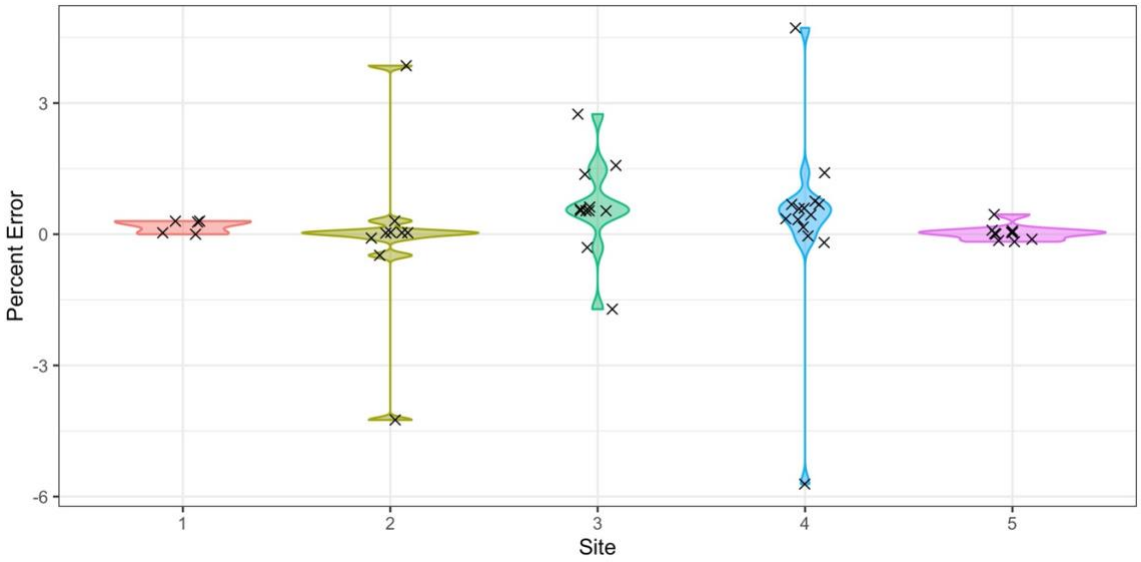


Figure 2.8. Percent difference between the volume registered by the meter (calculated as the difference between two consecutive, manual readings of the meter's register) and the volume registered by the CIWS datalogger (calculated as the number of observed pulses multiplied by the pulse resolution of the meter) for multiple deployment periods at each experimental site. The points in the figure represent individual percent errors computed for every field visit at each site. Values have been spread across the x axis for visualization purposes. The number of deployment periods was 5, 9, 11, 14, and 10 for each site, respectively.

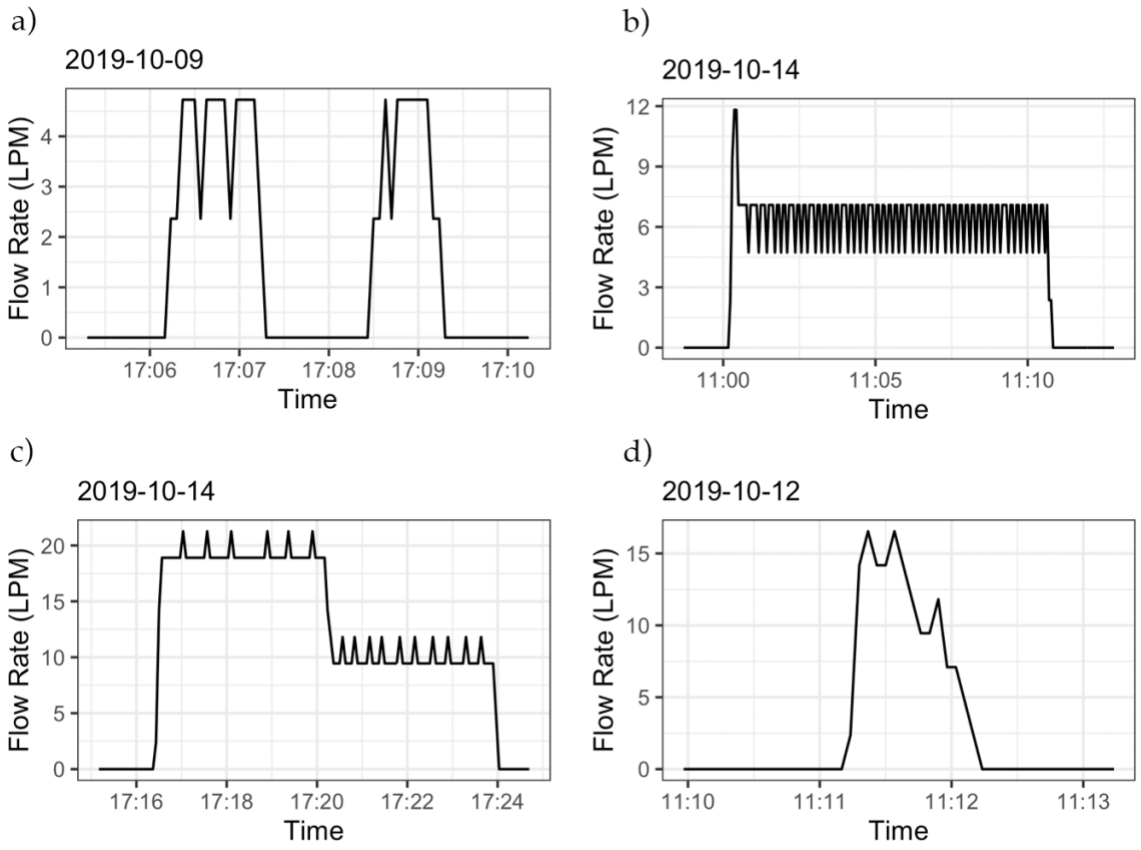


Figure 2.9. Flow rate signatures for events labeled by the homeowner at Site 1. Panel a) two subsequent faucet events, b) a shower event, c) a clothes washer event, and d) a toilet flush event.

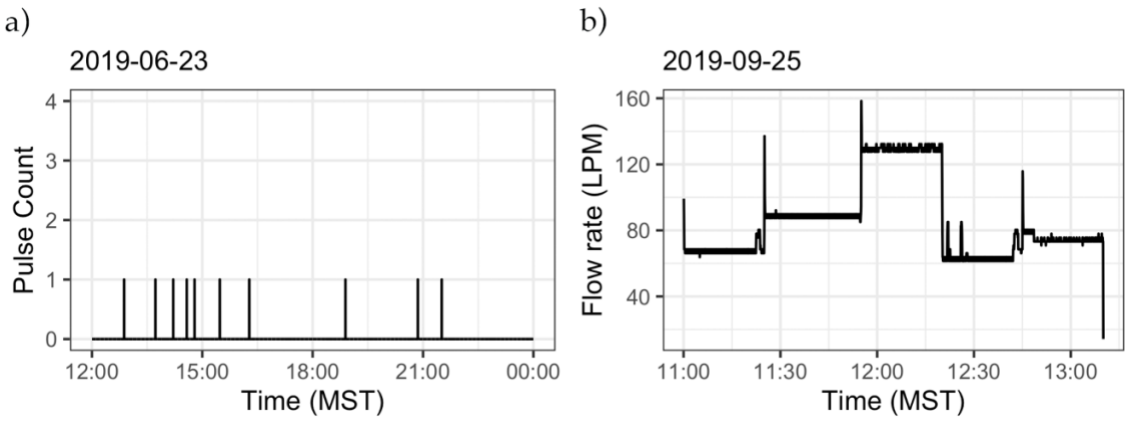


Figure 2.10. Sample of the events observed. a) Raw data collected at Site 5 on June 23, 2019 from noon to midnight. Multiple 4 second, single pulse events were observed at a time when no other water use occurs. b) An irrigation event at Site 1. At this temporal resolution, flow rates from different irrigation zones can be observed.

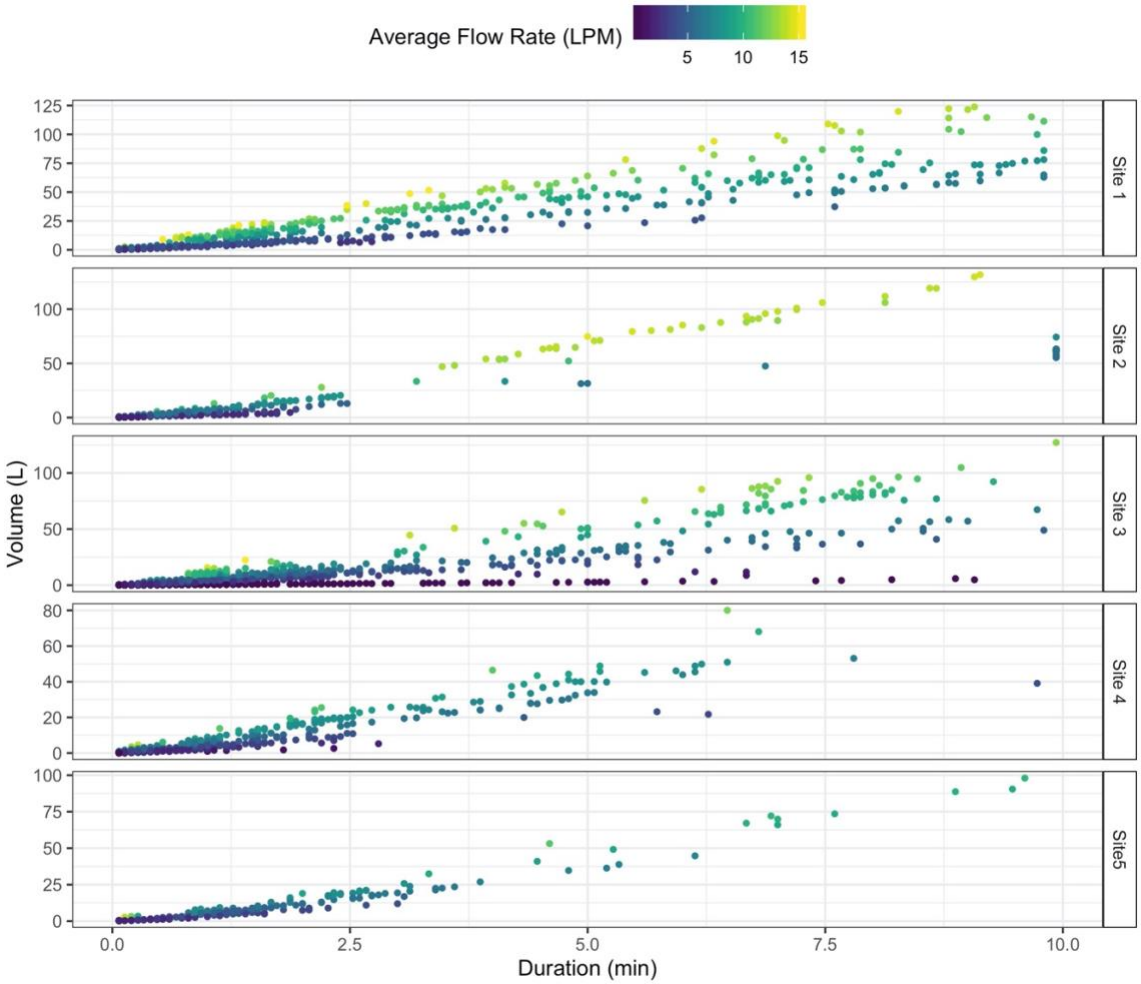


Figure 2.11. Duration (minutes) versus Volume (liters) of 23,478 events logged at the five sites. Events presented are those with a duration of less than 10 minutes and average flow rate less than 15 LPM (96% of all the events that are not leaks). Color is assigned to each event based on its average flow rate (LPM). These three values have been used by other authors to identify and classify end uses events.

CHAPTER 3

AN OPEN SOURCE CYBERINFRASTRUCTURE FOR COLLECTING, PROCESSING, STORING, AND ACCESSING HIGH TEMPORAL RESOLUTION RESIDENTIAL WATER USE DATA[1]

Abstract

Collecting and managing high temporal resolution residential water use data is challenging due to cost and technical requirements associated with the volume and velocity of data collected. We developed an open-source, modular, generalized architecture called Cyberinfrastructure for Intelligent Water Supply (CIWS) to automate the process from data collection to analysis and presentation of high temporal residential water use data. A prototype implementation was built using existing open-source technologies, including smart meters, databases, and services. Two case studies were selected to test functionalities of CIWS, including push and pull data models within single family and multi-unit residential contexts, respectively. CIWS was tested for scalability and performance within our design constraints and proved to be effective within both case studies. All CIWS elements and the case study data described are freely available for re-use.

3.1 Introduction

Achieving higher efficiency in urban water management and planning requires understanding of how water is used at the household level. Daily patterns in

---

[1] Camilo J. Bastidas Pacheco, Joseph C. Brewer, Jeffery S. Horsburgh, Juan Caraballo, An open source cyberinfrastructure for collecting, processing, storing and accessing high temporal resolution residential water use data, Environmental Modelling & Software, Volume 144, 2021.

consumption, potential for water savings and distribution of water use across end uses are essential inputs to water demand estimation, leak identification, design of programs to manage water demand, and water planning to ensure adequate supply (Giurco et al., 2008; Willis et al., 2011). Metering water use for billing purposes is a common practice in the United States, where meters are typically read monthly or quarterly. Our ability to characterize water demand is limited by the temporal resolution of the data collected. Higher resolution data can increase the accuracy of peak demand estimation and reduce leak volumes that can go undetected. Sub-minute resolution data is required to record and quantify end uses of water that have short duration (Cominola et al., 2018; Nguyen et al., 2015). However, obtaining this higher temporal resolution data at a scale larger than a few houses presents several challenges in terms of data collection, storage, management, and processing (Cominola et al., 2018), and doing it over an extended period of time can be unpractical (Cardell-Oliver, 2013).

Collecting a month of 10-s resolution data for a single meter, which is common in end uses of water studies (DeOreo et al, 2011, 2016; Mayer et al, 1999, 2004), produces more than 250,000 observations. Doing so at a water utility or municipality scale, which may have thousands of metered residential connections, presents obvious challenges associated with the volume of data that would be produced. Many utilities lack a dedicated information technology or data management staff, which means that new database management, software deployment, and data analysis tasks can be prohibitive. In these cases, and in the absence of sufficient cyberinfrastructure for automating data management tasks, high resolution data could be more of a roadblock for a water provider than a benefit. However, with adequate data collection and management tools,

utilities may be able to realize more of the potential benefits associated with high temporal resolution data. This includes quantifying water use behavior to better enable planning that ensures adequate supply, the promotion of water conservation behavior among users (Liu et al., 2015), improving customer service quality for utilities (Beal and Flynn, 2015), tipping the cost-benefit balance in the smart metering adoption case, which remains undefined (Cominola et al., 2018), and enabling the proliferation of scientific work in this field.

The term "cyberinfrastructure" integrates hardware and software tools, as well as data networks (NSF, 2007). Cyberinfrastructure can help solve data management challenges and enable more widespread collection of higher temporal resolution water use data for utilities and researchers. In a broader context, cyberinfrastructure is improving the communication of results from hydrological models (Souffront Alcantara et al., 2017), helping monitor watershed health parameters (Szwilski et al., 2018), assisting in the automation of comparing climate model results (Sun et al., 2020), and it is now ubiquitous in multiple scientific domains (Hachmann et al., 2018; Shams et al., 2018; Wegrzyn et al., 2020).

Smart meters have potential to solve one of the challenges in the pathway to an advanced water cyberinfrastructure, high resolution measurement of water use. The term "smart meter" can be ambiguous (Boyle et al., 2013). Within this article, it is used to denote devices capable of recording water use with high resolution (i.e., sub-minute frequency) that can be integrated in automated systems for data management. Nearly a decade ago, it was anticipated that use of smart meters would grow over time (Boyle et al., 2013), and they are, in fact, becoming more widely available and adopted. With this

emergence of smart meters, there has been an increase in the number of scientific publications using the high resolution data they produce for water demand analysis. Cominola et al. (2015) provide a comprehensive review. However, despite the increase in the number of publications using smart metering data to quantify end uses of water and water use behavior, the data management procedures, or tools, used in these studies are not well described, and most of the datasets used are not openly available (Di Mauro et al., 2020). In most of these studies, the focus has been on the tools and algorithms used for identifying water end uses and user behavior. Other components of the data management process are not described.

Available cyberinfrastructure for collecting, managing and analyzing this type of data remains scarce and of proprietary nature, with little available literature describing tools and procedures for data collection, management, and analysis. Meter manufacturers tend to have their own software systems designed for their metering technology, which complicates synthesis or integration of data from multiple systems and may help explain why research in this field has been conducted in a limited number of countries using a limited number of datasets. Many of these studies have used the same data logging device for data collection and the same software tool for end use analysis (Beal and Stewart, 2011; DeOreo et al, 2011, 2016; Mayer et al, 1999, 2004). Other studies have reused the same dataset to conduct different analyses. For example, Beal at al. (2013) present differences between perceived and actual water consumption, Willis et al. (2013) studied the impact of socio-demographic and efficient fixtures on water use, and Beal and Stewart (2011) presented end uses of water characteristics, all using the same dataset collected in Southeast Queensland, Australia.

The datalogging devices used in most high-resolution data collection studies lack communication capabilities, which limits the potential for automated integration with downstream cyberinfrastructure (e.g., telemetry, storage, management, and analysis applications). More recently, there has been increasing discussion around smart cities, smart grids, smart water networks and other related terms, despite there not being a wide agreement about their definition, what is meant by "smart," or the extension of their applications (Ardito et al., 2013; Hollands, 2008; Wissner, 2011). It is generally agreed that smart cities make use of information and communication technologies (ICT) in an attempt to assist cities in optimizing the use of their assets (Neirotti et al., 2014), water being one of the most important. Connectedness of data collection and its application is important in this context.

Advanced metering infrastructure (AMI) and ICT systems are vital for the successful deployment of a smart grid (Yan et al., 2013). In the energy sector, smart grids use smart technologies for metering, communication and automation and make use of digital information to improve reliability (U.S. Congress, 2007). The Internet of Things (IoT) has also been described as a potential enabler of smart grids in the water sector (Alghamdi and Shetty, 2016; Robles et al., 2014; Zanella et al., 2014), and, more recently, smart solutions that use IoT principles have been proposed (Amaxilatis et al., 2020; Stiri et al., 2019). Liu and Nielsen (2016) discussed existing technologies to develop an ICT system, or cyberinfrastructure, to enable smart meter analytics for the energy sector acknowledging the difficulties in processing and managing the large volumes of data generated. Similar systems have been proposed and discussed for water use analytics (Boyle et al., 2013; Li et al., 2020; Makropoulos, 2017; Moy De Vitry et al.,

2019), but few implementations have been published due to the cost and complexity of these applications (Alvisi et al., 2019; Amaxilatis et al., 2020; Anda et al., 2013). In one notable example, Chen et al. (2011) conducted analysis using data collected on a smart water service architecture deployed for billing purposes on the city of Dubuque, IA. This system collects data every 15 min providing more advanced analysis to water consumers and providers (Erickson et al., 2012).

While multiple high-level designs of a smart water network have been described (e.g., Hauser et al., 2016; Li et al., 2020; Ye et al., 2016), implementations are scarce. Most of the smart water systems designs we reviewed lacked a full demonstration or prototype implementation. In some cases, important elements, such as performance metrics and implementation guidance were not fully described (Li et al., 2020). When demonstrations were presented, the focus was primarily on the results of the specific case study (i.e., the lessons learned about water use and/or behavior) and not on the design and implementation of the tools used to complete the tasks. The limited availability of data and tools for the water sector constitutes a significant barrier for the development of research and prevents the advancement and implementation of smarter water grids at a large scale (Mutchek and Williams, 2014). The closed-source nature of existing data collection hardware and data management software creates accessibility and interoperability issues that prevent the progress of smart water grids while curtailing the adoption of open architectures (Hauser and Roedler, 2015; Robles et al., 2014). The development of open source cyberinfrastructure for managing high resolution data can lay the foundations for the development of newer and better tools for water utilities, as well as standards for operations that result in increased interoperability. All of these

actions could pave the road for more water demand research, and ultimately, advance technologies for the development of smart water grids.

Thus, in order to achieve the full potential of smart meters, cyberinfrastructure is needed to support utilization of the high resolution data they produce (Horsburgh et al., 2019; Mason et al., 2014). Developing effective cyberinfrastructure that can support both operational data collection and management (e.g., for billing, reporting and day-to-day management purposes) and exploration of data for research aimed at better understanding water use behavior is expensive and challenging (Stocks et al., 2019). Indeed, architectural designs and data structures for cyberinfrastructure supporting residential water use data must meet the needs of multiple users (i.e., water providers, water consumers, researchers) without disrupting a utility's necessary business functions. The research described here focused on the following research questions to advance the cyberinfrastructure and availability of software tools for collecting, managing and analyzing high resolution smart metering data: a) what is the general architecture for a cyberinfrastructure to support collection and management of high temporal resolution smart metering data, and b) how can that architecture be implemented to meet the needs of multiple potential users (e.g., water utilities, water consumers, researchers).

In this paper, we present a generalized architectural design for a Cyberinfrastructure for Intelligent Water Supply (CIWS) and a prototype implementation of each of the components within the architecture in support of multiple data collection, management and analysis case studies. The prototypes we developed demonstrate tools that are not currently available for researchers or utility managers and include: a) a data collection layer consisting of datalogging devices with data transmission capabilities,

which are modifications from our previous work (Horsburgh et al., 2017; Bastidas Pacheco et al., 2020); b) a data management and archival layer that receives, processes, and stores data; and c) a data analytics layer that enables calculation of common water use metrics (e.g., average hourly water use, instantaneous peak, and end uses of water disaggregation and classification). Components within these layers demonstrate the entire workflow consisting of data collection, communication, storage, management and archival, and visualization and analysis.

While CIWS was designed and implemented for research purposes, including appropriate mechanisms for protecting the identities of research participants where necessary, it facilitates implementation of high temporal residential water use analysis, which is of interest to not only researchers in the field, but also utility companies and water consumers and can provide information currently not available to them. The data collected and managed using CIWS is relevant for assessment and management of both water demand and for planning to ensure adequate water supply. We first describe the requirements for the system along with the overall architecture we designed to meet these requirements (Section 2). We then describe a set of case studies in which this overall architecture was prototyped and implemented using both existing and new open source hardware and software components (Section 3). Finally, we close with discussion and conclusions (Section 4).

3.2 Methods

3.2.1 CIWS Design and Overall Software Architecture

Our goal in developing CIWS was to create a generalized, modular architecture that can be used to automate the process from collection to analysis and visualization of

high temporal resolution water use data. In our case study applications of CIWS, we combined existing and developed new, open source hardware devices and software tools to demonstrate an integrated solution for high-resolution residential water use data collection, management, and analysis. The CIWS architecture and our prototype implementation were designed to address the following requirements. While we present our prototype implementations in this paper, there may be multiple implementations of the generalized architecture that meet these requirements.

a) An open architecture that could be implemented using a variety of technologies;
b) Open source software development to facilitate its deployment and use by other users, reduce costs, and provide a platform for future improvement by others while advancing financial feasibility of larger scale implementations;
c) A modular design, so each component of CIWS can be used, or advanced, independently;
d) Accept input data from different meters and measurement devices (sensors) to address heterogeneity in urban water meter technology;
e) Capacity to manage “push” and “pull” data retrieval from the metering devices depending on available communication technologies and storing of data in a centralized server;
f) Scalable to accommodate a large data volume while remaining responsive to queries for subsets of time series data of varying sizes;
g) Support production of analysis and insights that meet the needs of

different audiences.

In our review of the literature, we found that existing designs of smart components or cyberinfrastructure for managing water systems are not fully standardized. However, most systems described or implemented to date are composed of multiple layers working in connection to achieve the overall goal (Li et al., 2020). We found that the number, name and function of these layers was different in each design; however, we observed some similarities. In practice, the number of layers included in an architectural design comes down to tradeoffs between the benefits of modularity and separation of concerns that can be achieved versus the complexity and potential fragility introduced with a larger number of layers. Separate layers can be autonomous such that changes to one layer do not have to affect the other layers. However, a greater number of layers typically involves more components that can fail.

Our overall architectural design for CIWS adopts this multi-layer paradigm (Figure 3.1) and is composed of three main layers. The first layer is the Data Collection Layer and includes the physical instruments and sensors used to monitor water use. It has also been called the sensing layer (Ye et al., 2016), the physical layer (Hauser et al., 2016), or the instrument layer (Li et al., 2020). The second layer is the Data Management and Archival Layer, which handles data communication, parsing and archival. This layer has also been referred to as the network or function layer (Hauser et al., 2016; Li et al., 2020; Ye et al., 2016). The final layer is the Data Analytics Layer, which handles all the steps between queries to retrieve data from the archival component to final visualizations, analyses and presentations produced for utilities, water consumers, researchers, etc. (i.e., the consumers of the data). This layer has also been referred to as the application or the

data fusion and analysis layer (Hauser et al., 2016; Li et al., 2020; Ye et al., 2016). Some of the other systems reviewed include elements for real time monitoring and control of observed variables and processes within the system, resulting in architectural designs with a larger number of layers. Since these elements were not needed in our case study use cases, a three layer model met all of the requirements listed above. A system with more layers may become more fragile; therefore, our design includes the minimum needed to meet the design considerations.

The architecture for CIWS and our prototype implementations were developed with a research focus – e.g., collecting, storing and managing high resolution water use data to enable advanced study of residential water use behavior. This type of research may be carried out by utilities, universities, or other agencies involved in research related to or management of urban water supply and demand. The typical deployment size in this type of work has been around 50 houses per city; however, some studies have analyzed up to 762 sites (DeOreo et al., 2016). In the latter case, the data was not collected simultaneously at all sites. Our aim was to develop a system that can handle, at minimum, the number of simultaneous data collection sites within the range of deployments observed in the past (40-60 houses). In the following sections, we describe in more detail the high-level design for each of the architectural layers, their key components, and their basic functionality.

3.2.1.1. Data collection layer

Data collection refers to the actual measurement of the variable or variables of interest, in this case, high temporal resolution water use. Here, we define high temporal resolution data as data collected at a sub-minute resolution. Typical investigations of

water use behavior, such as separating and quantifying end uses of water within a home, require data to be recorded at 10-s or even finer resolution over data collection periods of weeks to months. With few exceptions, high temporal resolution data cannot be collected using existing, commercially available smart meters without adding additional hardware or software components (Cominola et al., 2018), which can be expensive (Horsburgh et al., 2017). Water metering technology typically consists of a physical meter that uses one of several measurement techniques paired with an analog or digital register on which a totalized volume of water use is recorded. Some registers, including those of commercially available smart meters, are capable of storing volume readings within internal memory; however, this is usually constrained to relatively short periods of time (e.g., weeks) at recording intervals longer than 1 min. Other registers report only the most recent volume reading and are designed for periodic (e.g., monthly or quarterly) readings either manually or automatically via radio. These practical limitations are driven by power, local data storage, and network bandwidth limitations of existing metering technology.

Some water use studies have added flow metering sensors directly on the water pipe leading to each appliance in a residential house (Kofinas et al., 2018; Di Mauro et al., 2019). Opting for this approach allows direct measurements of water use from each fixture, and by placing the measuring element inside the property, power and communications can be readily available. However, this approach is invasive and requires modifications to the plumbing in each home where data is collected, which can increase costs and limit the applicability of this methodology at a medium or large scale. Therefore, we opted to focus our efforts on datalogging devices that can be coupled with

the existing water meter available at the property. Datalogging devices designed to couple with existing meters are available (Bastidas Pacheco et al., 2020; F.S. Brainard & Company, 2020). These dataloggers essentially perform the same function as the meter's register, but have the capability of recording much more frequent observations over longer periods of time. To be fully integrated in a data management system like CIWS, the datalogging devices must also have communication capabilities. CIWS was designed to handle both push and pull data communication, making it adaptable for multiple scenarios. The term push is used to denote systems where the data is sent by each datalogger (client) to a centralized server, while pull refers to systems where a centralized server connects to each datalogger and requests data. Given the modular design of CIWS, it is possible to integrate dataloggers that lack communication capabilities, such as those used in most residential studies in the past. Under this scenario, a user can take advantage of the Data Management and Archival and Data Analytics Layers of CIWS, while using data files manually downloaded from the datalogging devices in the field.

3.2.1.2. Data management and archival layer

The Data Management and Archival Layer is responsible for the work required to process the data logged by the devices. The key component addressed in this layer relates to developing and using software elements to automate repetitive data management processes and enable an easier transition between large volumes of data collection and useful information generation. This layer is composed of multiple working elements (Figure 3.1). For push based data transmission, a listener service is required to receive the data sent by the dataloggers. In pull based data transmission, a request service is used to achieve the same task. Once the data is received, it must be verified, parsed and

transferred to a database component. The database component accepts and stores data for downstream analysis and decision making. Real-time monitoring of water use is typically not of interest in most research scenarios, where most data analysis happens after the data have been collected. Additionally, given the frequency with which observations are recorded (e.g., on the order of seconds), it is not practical to push or pull data every time a new observation becomes available. Based on this, CIWS was designed to collect and send files containing many observations rather than sending observations individually. This approach minimizes the communication load on the system because the data transfer process does not occur constantly, and it can be scheduled to meet specific needs.

The request service for pull based data transmission must execute the following tasks: a) connect to a datalogging device; b) check for new data files; c) request and transfer new files; d) read and parse the files, and e) upload the data into the database. Remotely accessing devices can be achieved using a variety of communication protocols like Secure Shell (SSH), which is a widely used method for similar tasks due to its simplicity, speed and security. In this model, the datalogging devices need to be powered on and connected to the network at the time the connection is established. Additionally, a key requirement is that each datalogging device must be located, addressed, and accessed directly, which also provides an opportunity for remote functionalities, such as software updates, troubleshooting, changing data collection settings, and others.

The listener service, which manages the data transferring process under the push model, must complete the following tasks: a) accept and validate the data sent from each datalogging device deployed, b) process incoming files, including parsing the information they contain, and c) saving the data received into the database. Under this approach, the

communication elements of the datalogger only need to be powered up and functioning for the time it takes to send the desired information to the listener service, which can contribute to lower power requirements. Additionally, there is no requirement for data logging devices to be uniquely addressed on a network as they can identify themselves within the content of the message they push to the listener service.

Multiple technologies that can potentially meet the data storage and accessibility design considerations (i.e., the database requirement) are available. The database must be able to manage large volumes of data and provide a platform for generating analytics of such data. The data managed by the system consist mainly of time series of flow observations, which are constantly being collected and written into the database. Thus, the databasing technology selected must provide: a) easy and fast querying between dates and times to enable manipulation of the data; b) high performance for read and write operations as the database is continuously being updated with new data and potentially accessed by multiple users; and c) scalability, as the volume of data to be stored in the database increases quickly as the monitoring network and time period over which data are collected grow. The database schema used to organize the data for CIWS was designed to maximize query efficiency while maintaining the ability to protect the privacy of water consumers by storing personally identifiable information outside of the database. Common queries to be conducted in projects where CIWS can be used include selecting all or part (time constrained) of the full resolution or time aggregated data for a single or multiple sites.

3.2.1.3. Data analytics layer

The Data Analytics Layer supports generalized interactions between data users

and the database for the purposes of visualization and analysis of the data. The necessary functions executed in this layer include: a) user authentication to access existing data, b) querying data from the data base, c) data manipulation and analysis, and d) generation of reports and visualizations of interest for different target audiences. For the purposes of this research, three main target audiences were identified as users of information produced by the Data Analytics Layer: water consumers, utility managers, and researchers. While these categories of users are not necessarily exhaustive or mutually exclusive, the information that would be useful to these different users and the methods used to interact with the data are not the same. For instance, an individual residential user would need to be able to access and interact with the data from their home in a practical and non-technical way that does not require specialized software. Past studies have evaluated residential users' preferences for water use feedback, finding that information about their prior water consumption, comparison of use with that of similar users, and details about their consumption can increase user understanding (Erickson et al., 2012; Liu et al., 2015).

Utility managers may want to access standardized plots or reports showing data from multiple users, and researchers may need much more freedom to formulate their own, custom queries to the database to subset, aggregate, or summarize data in useful ways. This implies that the Data Analytics Layer needs to support multiple mechanisms for accessing and interacting with the database. Authentication, authorization, and privacy for users with different privileges (read or write data in a database) to access online resources have been discussed for multiple applications (Christie et al., 2020; Heiland et al., 2015; Kim and Lee, 2017). High temporal resolution data products, such

as distribution and timing of end uses, can raise privacy concerns among water consumers that must be considered when designing data presentation tools (Froehlich et al., 2012). Aggregation and summarization techniques can be used to present information for multiple water consumers while protecting privacy, and authentication and authorization can be used to limit what data is available for different users. CIWS considers the use of anonymized datasets throughout the system by identifying water consumers with a unique identifier. Linkage with the personally identifiable information about each water consumer is stored separately and is only available to those who have appropriate privileges and are allowed match water consumers with their data.

3.2.2. Case study design and system testing

In order to evaluate the overall architecture design, we designed two case studies that demonstrate different aspects of the architecture presented in two distinct data collection environments. The first case study demonstrates data collection at individual single-family residential homes. It uses an autonomous datalogger with communication capabilities to collect high resolution water use data and demonstrates push-based transmission of the data to the Data Management and Archival Layer. The second case study demonstrates data collection within multi-unit residential structures on a University campus. It uses dataloggers with dedicated power supplies and network registrations to demonstrate pull-based transmission of the data to the Data Management and Archival Layer. In the second case study, we collected data for additional parameters needed to characterize the energy consumption related to hot water use. The collection of data for these parameters provides an example of CIWS flexibility. Both case studies share the same layers, but we describe the different elements used by each case study.

We created a full prototype implementation of the design layers presented in Figure 3.1 for each case study and deployed them in an operational environment. These prototypes and deployments were created to demonstrate proof-of-concept for data collection and management components, the shareability of components within the architecture regardless of the data transmission method, and generalizability for our architectural design. We tested the system developed for scalability by simulating an increased number of sites and larger volumes of data.

Python 3.7 was chosen to develop all of the code and software associated with our case studies given that it is freely available and open source, it is a high-level programming language with a vast number of libraries available to complete an important number of functions required in our application, and it could be used across all three layers of our architectural design. Using Python also helped us meet the first three requirements described above as the code can be easily shared, read and modified by other programmers and scientists, and can be deployed in different operating systems, which increases reuse possibilities.

3.2.2.1. Case study 1 description

Water use in single family residential homes is quantified, to a large extent, using analog, positive displacement water meters. The volume of water that has passed through the meter is usually the only variable recorded by this type of meter. In most cases, water meters are enclosed in underground pits of varying depth, limiting power supply availability. These meters are typically read monthly, quarterly or at coarser resolutions by the utility for billing purposes either manually or via a roving radio that receives the most recent volume observation from each meter when the roving radio passes within

range. Some more advanced networks include automated retrieval of the coarse resolution volume data, but very few have the capability to record and transmit high resolution data. Given that the vast majority of residential water meters in use today share these constraints, we chose this case study to demonstrate adding high resolution data collection and transmission capabilities to existing, analog water meters.

3.2.2.2. Case study 2 description

The Living Learning Community center (LLC) on Utah State University's (USU) campus was selected as a second case study for deploying CIWS within a set of multi-unit residential buildings. The LLC is one of USU's newer student housing options and houses approximately 500 students distributed among six dormitory buildings labeled building A – building F. The objective of this implementation was to characterize water and water-related energy use in five buildings (B–F). The importance of the water-energy nexus for optimizing conservation and sustainable management has been identified in the past (Hamiche et al., 2016; Kenway et al., 2016; Fang and Chen, 2017). However, collecting water and energy consumption data combined at a sufficient temporal resolution to analyze their relation is uncommon, and the methods for linking water and energy use are not well established. This case study demonstrates a methodology for collecting water and water-related energy data in a multi-unit residential setting. Buildings B–F host approximately 90 students each. Building A hosts administrative offices, has a much lower student occupancy, and was excluded from the study. We chose a pull based model for this case study given the availability of dedicated power at each data collection site and the availability of USU's campus Wi-Fi network to enable communications and data transmission.

Three water meters are present in the water supply system for each of these buildings - hot-water supply, cold-water supply, and hot-water return. To monitor water and water related energy use within each building, two characteristics of each meter were measured, flow and water temperature, resulting in a total of six variables collected per building (Table 3.1). The hot-water return is a feature of the LLC's innovative hot water recirculation system. Hot water is continually circulated from three boilers to the LLC buildings at a constant, base flowrate of approximately 3 gallons per minute (gpm) or 11.4 liters per minute (Lpm). Increases from this base flowrate constitute hot water use. Unused hot water returns to the one of the three boilers for reheating and eventual recirculation. Cold water is supplied in a typical on-demand basis.

3.3. Results and discussion

3.3.1. Case study 1: push based data collection for single family residential homes

We selected a single family residential property to test the CIWS functionality under a push based data retrieval model. We collected two weeks of data at this property, between January 15, 2021 and January 28, 2021, for the implementation described. All water use results presented are for this time period. This home had five occupants, three of ages between 10 and 25 and two between 40 and 60 during the data collection period. It was built in 2006, has three bathrooms and a total parcel area of approximately 12,000 ft2 (1114.8 m2). We chose push based data retrieval for this case study because it is enabled by heterogeneous networking – i.e., any datalogger device capable of high resolution data collection and sending data over an available data network could be used without the need for each device to be uniquely addressable on a network. Additionally, power requirements can be reduced given that data logging devices do not have to listen

for connections and requests from a centralized server but rather wake to transmit data on a user-configured schedule.

3.3.1.1. Data collection layer

At the property selected, a one inch (2.54 cm) Bottom Load (BL) Master Meter with an analog register was being used by the water utility to record monthly water use, transmit it to a roving receiver via a 3G radio and bill water usage. We added high temporal resolution data collection and transmission capabilities without affecting the normal operation of the utility's meter by installing a CIWS Water Meter Node (CIWS-WM-Node) datalogger to measure water use at a 4-s temporal resolution on top of the existing meter. The CIWS-WN-Node is an advanced modification of the CIWS datalogger (Bastidas Pacheco et al., 2020), which is an open source, Arduino-based datalogger that we designed to work with any magnetically-driven water meter. The CIWS datalogger uses a magnetometer sensor to measure the magnetic field around magnetically-driven residential water meters. It counts peaks in the magnetic field associated with movement of the magnetically-driven measurement element within the meter, and registers peaks as pulses that represent a fixed volume of water passing through the meter. These pulses are multiplied by a factor called the meter resolution (0.041619 gallons per pulse, or 0.1575 liters per pulse, for the case study meter), which is specific to each meter type, brand, and size, to obtain the volume of water that passed through the meter per unit of time. Meter pulse resolution values can be obtained from meter manufacturers or through a calibration procedure described by Bastidas Pacheco et al. (2020).

The CIWS-WM-Node we developed for this case study adds communication and

computational capabilities to the CIWS datalogger by coupling it with a Raspberry Pi Model B or Model B+ single-board Linux computer. The components of the CIWS datalogger control all of the datalogging functions, whereas the Raspberry Pi computer can be powered on a user defined schedule to process and transmit data. The Raspberry Pi runs a version of the Linux operating system called Raspberry Pi OS (previously called Raspbian). Although the Raspberry Pi is capable of interfacing with a number of different wireless communication options, including Wi-Fi, radio frequency, cellular 3G, LTE, Bluetooth, and satellite, we chose to use the Raspberry Pi's built in Wi-Fi capabilities for this case study because the homeowner's Wi-Fi network was easily accessible. In broader application, however, any Internet data connection compatible with a Raspberry Pi could be used.

The CIWS-WM-Node datalogger outputs a comma separated values (CSV) file including a three line header with a unique identifier for the site at which the datalogger is installed, a unique identifier for the datalogger, and the meter resolution for the meter on which it is installed. The datalogger records three variables during the logging process: Datetime, Record, and Pulses (Bastidas Pacheco et al., 2020). The CIWS-WM-Node datalogging device was configured to chunk the data files by day (i.e., a new CSV file is created for each day) and send data files once per day to the Data Management and Archival Layer via an HTTP POST request. This functionality was developed as a single Python script (data_transfer.py). When the Raspberry Pi is powered on, it can conduct any computation required, and the data_transfer.py script is executed to send data files to the Data Management and Archival Layer for further processing. After a file is successfully sent via HTTP, it is moved to a different folder in the datalogger's local

storage for backup.

3.3.1.2. Data management and archival layer

For our case studies, the Data Management and Archival Layer components were deployed within a VMWare ESXi server environment hosted at Utah State University on a single virtual machine (VM) running the Ubuntu Linux Server Version 18.04 (Bionic Beaver) operating system. Ubuntu is a free and open-source Linux distribution developed by Canonical Ltd. It is well supported, stable, and offers reliable file security. The VM was configured with a 64-bit architecture, four 2.3 GHz processor cores, eight GB of RAM, and 100 GB of hard disk space. We refer to this VM as the "Data Management and Archival server."

We developed three main components to complete the tasks described for this layer, the data posting service (DPS), the data loading service (DLS), and the operational database, each of which is described in the sections that follow. The DPS and the DLS were developed in a generalizable way to facilitate reuse and serve as the Network Listener shown in the center panel of Figure 3.1. However, some specific details were adapted to this implementation. For example, the data parsing works for the specific output format of the CIWS-Datalogger. The DPS and the DLS were deployed on the Data Management and Archival server and then configured via settings stored in a user-modifiable JavaScript Object Notation (JSON) file (named configuration.json) that details the information needed for their operation. For deployment, the configuration file must be placed in the same folder with the DPS and DLS.

3.3.1.2.1. Data posting service (DPS)

The DPS is a listener web service that receives and processes data files pushed to

the Data Management and Archival server from the CIWS-WM-Node dataloggers. The DPS works integrated with two common server technologies, the web server software that processes HTTP requests received by the server and a Web Server Gateway Interface (WSGI) that runs the DPS application in response to the requests. We chose NGINX (NGINX, 2021), which is a free, open source HTTP server, to serve as the web server software because of its high performance, stability, simple configuration, and low resource consumption. The WSGI was implemented using (Gunicorn, 2021), which is a Python WSGI HTTP server for Unix-like operating systems. Guidance for deploying the web server and WSGI software is available in the project's GitHub repository. The parameters included in the configuration files for the DPS and the DLS are described in Table 3.2.

The overall functioning of the DPS is as follows. Dataloggers send an HTTP POST request to the server that contains a data file (for our case study, one day of high resolution water use data for that home). These requests are received and handled by the NGINX web server, which passes them to the Gunicorn WSGI. Gunicorn then invokes and executes the DPS to authenticate the HTTP POST requests by using a token (client_token in Table 3.2), verifying the file type (CSV) and that the file does not already exist on the server, before moving it to a local folder on the server (source_directory in Table 3.2) for further processing by the DLS. The DPS is composed of three pieces of code: app.py which lists the functions needed to read the application configuration file, auth.py that lists all the functions for file authentication, and web_service.py which calls the previous two files and executes the tasks described. Figure 3.2 illustrates the processes described and lists the elements involved.

The DPS was implemented using Bottle (Hellkamp, 2021), which is a WSGI micro web-framework for Python. Bottle is simple, fast, lightweight, and works without additional dependencies, making it ideal for running small applications like the DPS. Bottle built-in functionalities, such as its simple URL routing capabilities and the convenient access to file uploads, were used to facilitate the development of the DPS and avoid dealing with low-level details of HTTP requests handling and routing. We implemented a very simple, token-based authentication for the HTTP POST requests in our prototype to avoid SPAM content being submitted to the DPS. More sophisticated and secure authentication and authorization processes could be integrated in the future, if needed to provide greater security. A log file keeps track of the requests received by the DPS and actions executed (the log file is located in a directory described in Table 3.2). The log file records successful and unsuccessful (e.g., a file that already exists is sent to the server multiple times, a request that is rejected by not having appropriate authentication credentials) posting attempts. All events are logged in a single file, named data_poster.log, which is limited to 5 MB in size. When a log file exceeds this size, it is saved adding a sequential number at the end (data_poster1.log initially) and the current logging continues in the original log file.

3.3.1.2.2. Data loading service (DLS)

We developed the DLS to read the files received from the dataloggers from the source directory on the server, parse the unique site identifier information from the header of the CSV file and insert the data into the database for archival and use by the Data Analytics Layer. The DLS also verifies that the data received does not already exist in the database by checking the unique site identifier and datetime values of the data to

avoid duplication of data in the database. The DLS uses the same configuration file as the DPS, described on Table 3.2. The DLS reads data files from a local/source directory and moves them to a local/target directory after successfully inserting the data into the operational database. If an error occurs, the files are moved to the quarantine directory. A log file records all the activity executed by the DLS, including any error observed in the process, such as invalid datetime stamps, invalid site identifiers, and attempts to load data that already exists in the database. This log file is named data_loader.log, and it is managed identically to the DPS log file. Both are located in the same folder (log_directory in Table 3.2).

We chose this implementation for several reasons. First, it enables preservation/archival of the original CSV data files recorded by the dataloggers. Second, the data are loaded into an operational database that is highly performant for querying and data retrieval in support of the Data Analytics Layer. Third, it enables all of the downstream components in the architecture to be used regardless of how the data files arrive on the server. For example, they can be automatically pushed to the server from the datalogger, pulled from the datalogger by the server (as in our second case study), or manually copied to the server in the case where data transmission is not automated. The DLS was implemented in a single Python script named loader.py.

3.3.1.2.3. Operational database

For the operational database component, we chose to use an existing technology given the availability of mature and robust database software. In our previous work related to investigating how to best manage large volumes of time series data, we tested the performance of four commonly used open source database technologies, including

MongoDB, MySQL, PostgreSQL, and InfluxDB (Brewer, 2020). Based on our tests, we chose to use InfluxDB (InfluxData, 2021) due to its time series oriented data structure, rapid query performance, and favorable disk space requirements when compared to the other software technologies. InfluxDB is a popular time series database designed specifically for time series data in applications that require handling high data write and query loads. It provides a powerful structured query language (SQL)-like query language and has both open source distributions that can be installed and used for free (e.g., as we did on our Linux VM) and cloud deployments that can be implemented with usage-based pricing. InfluxDB has been used in multiple IoT and other applications, where it has been tested for large datasets (Balis et al., 2017; Di Martino et al., 2019; Rinaldi et al., 2019). InfluxDB also offers extensive support for multiple programming languages, including Python and R, which are commonly used for data science. This made it straightforward for us to use Python to insert data and to execute queries from the Data Analytics Layer.

InfluxDB databases are organized around the concept of a measurement, which can be thought of as a "table" that contains an indexed column named time containing the timestamp of each data point, where each data point is a row in the table. Additional variables are stored in columns that can be tags or fields. The main difference is that tags are indexed and are not required in a data structure, whereas at least one field is required, fields are not indexed. The column names for tags and fields are defined as keys. Generally, it is recommended that data values are stored as fields, and metadata as tags to improve query performance. In our design for storing data in InfluxDB, the number of pulses recorded by the datalogger during each time interval is included as a field (key = pulses), and the site identifier (key = siteID) and the datalogger identifier (key =

dataloggerID) are included as tags (Table 3.3).

The data for all sites are stored in a single measurement within the Influx database. Raw data and quality controlled (QC) data are stored in separate measurements with the same structure. QC data is a copy of the raw data that is created after verifying that the volume registered by the datalogger is within ±5% of the volume registered by the meter (estimated using subsequent readings of the meter's register conducted during installation, during periodic site visits, and at removal of the datalogger). In some cases, known bad data were trimmed from the beginning and end of a valid deployment. Where the volume recorded by the datalogger did not match the volume recorded by the meter's register, the data were discarded and a new deployment was started. During our case study deployments, we did not observe any out of range, anomalous, or unreasonable pulse count values after this QC procedure. In consequence, additional QC modules were not implemented.

Calculating the volume registered by the meter requires manual meter readings. Because of this, the QC procedure we used required manual interaction by an analyst. Data was validated by an analyst and manually placed in the source_folder (Table 3.3) folder where the DPS, after reading the metadata included, parses the information and writes the data to the correct InfluxDB measurement. However, additional QC procedures could be implemented in the future. For example, it could be advantageous to automate QC procedures, which could be done where meter readings can be obtained automatically or where QC rules are defined that do not require meter readings. All queries and analysis were conducted using the QC data.

The database is the point of connection between the Data Analytics Layer and the

Data Management and Archival Layer, and its design must meet requirements from both layers to write and read data. Typically, database schemas are designed around the structure of the data to be stored and to facilitate the most common types of queries. This is usually a tradeoff between making it easy to insert data into the database while still providing highly performant queries. The simple database schema implemented in this case study (Table 3.3) mirrors the structure of the data files generated by the dataloggers, making it straightforward to insert data, but is also optimized to support the following queries: 1) selecting all of the data for a particular siteID; 2) selecting all of the data for a particular dataloggerID (e.g., to track the performance of a datalogger, which may be deployed at multiple sites at different times, and identify/correct any systematic errors); and 3) querying data for a specific time frame (e.g., between a beginning and ending date). Combining queries based on these three elements provides most of the functionality intended for CIWS and met all of the needs of our case study.

Additional queries intended to allow comparison of data across multiple sites may also be of interest. Our design separates the time series data, which are stored anonymously in the InfluxDB database, from household information, which is stored in a separate CSV file, named sites.csv. The data stored in InfluxDB do not contain any identifiable information, which removes privacy concerns from the time series data. The separate sites.csv file may include sensitive, personally identifiable information (e.g., names, addresses, etc.) along with any other descriptive characteristics (the version of the sites.csv file for this study published in HydroShare has been anonymized). Data managers may wish to maintain multiple versions of the sites.csv file (e.g., one with all personally identifiable information about data collection sites and one that has been

anonymized and could be released to a broader set of users). While this approach adds an additional step for certain types of queries (e.g., selecting data for all houses within a certain geographic area or of a certain built age) because the site information must be queried before the correct time series data can be retrieved, it provides a mechanism for protecting personally identifiable information and more flexibility for managing metadata about the sites. Removing or adding tags to existing measurements is significantly restricted in InfluxDB. In consequence, anonymizing the data stored in InfluxDB for publication is not needed, as the data stored is already anonymous. Queries against the time series data can always be executed using a siteID or set of siteIDs obtained via a prior query to the sites.csv file. It is also possible, but currently not implemented, to add all site metadata as tags in the InfluxDB measurement to eliminate this intermediate query step, if that is more convenient in a specific application.

Researchers and utility managers can access the data within the InfluxDB database with a non-administrator user account. InfluxDB allows for the creation of multiple non-administrator users and at least one administrator user. The administrator manages authorization for each non-administrator user. Non-administrator users can be restricted to write, read, or both. The free version of InfluxDB does not allow fine-grained authorization, which would be needed to restrict users to view only part of the data in a measurement. However, we did not see this as a significant drawback as high level users like researchers and/or utility managers would likely need to have unrestricted access to all of the data in an InfluxDB database. Furthermore, it is unlikely that the full resolution data would be provided to water consumers. Rather, a more likely scenario would be for a software application with a graphical user interface to be developed for

presenting water consumers with feedback about their consumption. Authentication and authorization of users could be handled separately by the software application in future deployments. Erickson et al. (2012) provide an example of an online water portal and discuss the privacy and user authorization concerns that impact the design of similar tools. Homeowners are typically presented with summary statistics and visualizations calculated for their property and may be provided with a summary-level comparison with other properties. However, they generally would not have access to view raw data for their own or other properties.

3.3.1.3. Data analytics layer

To illustrate the type of capabilities supported by the Data Analytics Layer, we developed Python tools that provide an example of the main aspects involved in this process: connection to the database, user authentication, and data retrieval via common queries. Once the data has been retrieved into a Python environment, it can be integrated with existing, and more advanced, data analysis and visualization tools. While it is beyond the scope of this paper to demonstrate all of the possible ways in which data can be retrieved from the database component and used within analytical applications, the tools we developed demonstrate the general patterns required for developing such tools and serve as a foundation on which others could be developed.

InfluxDB client programming libraries are available for several popular programming languages, including Python, Go, C#, Java, PHP, Ruby, Scala, JavaScript, and R, which simplifies software development using InfluxDB and facilitates desktop, mobile, and web application development. Using the Python client library for InfluxDB (InfluxDB, 2020), we first developed a set of functions for interacting with the InfluxDB

database. These functions were implemented within a single Python script called da_functions.py. This script connects to the database using a set of configuration parameters that are included in a JSON file named configuration.json, which is similar to the one used by the DPS and DLS applications. Parameters in the JSON file include: host, port, username, password, and database (as defined in Table 3.2). The functions we developed in da_functions.py (Table 3.4) use the existing capabilities of the InfluxDB Python client library along with specific parameters provided by the user (e.g., siteID, time, dataloggerID as defined in Table 3.3) to provide a simple application programming interface (API) for querying data from the database. We anticipate that these functions will meet many of the most common data requirements for most researchers and utilities. The functions generate a Pandas dataframe (McKinney, 2010) with the resulting data if a single siteID is provided, and a Python list of Pandas dataframes when multiple siteIDs are provided. If a start date or end date are not included, the function will download the entire record available. If only a start date is provided the function will return everything from that date to the end of the record, in the opposite case, it will retrieve data from the beginning of the record to the specified ending data. If measurement is not provided, the functions will query from the quality controlled data (QCData). Raw data can be downloaded by specifying measurement = ‘RawData.’ For time aggregated data, the function parameter can include any Influx supported aggregation function (e.g., mean, median max, min, sum). The time resolution of the aggregated data supports any InfluxDB duration type (e.g., ‘1m’ for 1 min data, ‘1h’ for hourly data, ‘1d’ for daily data, ‘1w’ for weekly data). All the arguments in both functions are Python keyword arguments. They must be preceded by their identifier (or name) when executing the

functions, i.e., get_data(site = 1) to return all the quality controlled data for siteID 1.

We then developed a Python Jupyter Notebook called data_analytics.ipynb that loads the functions listed and implements a basic workflow to produce metrics and analysis from the data collected. Jupyter Notebooks (Kluyver et al., 2016) allow creation and sharing of documents that contain live code, equations, visualizations and narrative text, which makes them ideal for prototyping visualizations and analyses for the Data Analytics Layer. The Notebook we developed imports data using the defined functions and then generates visualizations of common metrics of residential water use for presentation to water consumers. For example, Figure 3.3 shows the average hourly water use (blue solid line), and the boxplots show the distribution of hourly water use for the period of data collection at the residential home we monitored. We can notice two periods of higher water usage, one during the morning and the other early in the afternoon, corresponding with patterns typically observed in hourly residential water use data. During this period, no outdoor water use occurred; therefore, the figure represents indoor water use only. The Notebook then demonstrates calculation of summary water use information for the data collection period. For example, average daily water use was 170.2 gallons (644.3 L), leading to a per capita average daily water use of 34 gallons (128.7 L). The maximum daily water usage observed during the period was 292.7 gallons (1077.9 L), the instantaneous peak was 10 gpm, or 37.95 L per minute (Lpm), and the maximum hourly usage registered was 74.1 gallons (280.5 L).

Another analysis of special interest using high-temporal resolution data is the identification of end uses of water. We used an open source algorithm developed by (Attallah et al., 2021), available via the HydroShare repository (Attallah and Bastidas

Pacheco, 2021), within the Data Analytics layer to separate raw data into events and classify the resulting events into categories of end uses of water. The algorithm filters the data collected using a low-pass filter, making it easier identify single or concurrent events. Concurrent events are separated into single events, and the final table containing only single events is classified by using a combination of clustering to identify atypical or outlier events, and a fully-supervised machine learning methodology to assign labels to the remaining events. The machine learning model uses a Random Forest classifier (Liaw and Wiener, 2002) trained using a set of user-labeled and manually-labeled events to classify new events for individual residential homes (Attallah et al., 2021). We used the trained machine learning model to label the events generated during the data collection period at the residential home we monitored. While a potentially large number of analytics, visualization, and information can be generated from the labeled events, the Jupyter Notebook we developed presents a small subset of them (Figure 3.4) as an example of products that can be generated from the raw data.

At the observed home, toilet events account for 36.1% of the total indoor volume used, showers 26.3%, clothes washer 13%, faucets 12.4%, and bathtub events 11.1%. Unclassified events, defined as events lasting 4 s or less and consisting of a single "pulse" recorded by the meter (approximately 5 ounces, or 0.15 L of water), account for approximately 1% of total use. Unclassified events include very short water use events (e.g., ice making refrigerators, short faucet events) and leaks. Figure 3.4 shows the distribution of the volume a), flow rate b), and duration c) for each category of indoor water use. Unclassified events were excluded from Figure 3.4. Faucet events had a median flow rate of approximately 0.8 gpm (3 Lpm). Water-efficient bathroom faucets,

as defined by the United States (U.S.) Environmental and Protection Agency (EPA) in their Water Sense program (EPA, 2020), operate between 0.8 gpm at a pressure of 20 pounds per square inches (psi), or 137.9 Kilopascals (kpa), and 1.5 gpm (5.7 Lpm) at 60 psi (413.7 kpa). Compared to this EPA standard, the flowrates we observed from the faucets at the study property are efficient. A similar conclusion can be reached by comparing the median flow rate of shower heads at the study property (approximately 1.8 gpm, or 6.8 Lpm) with EPA Water Sense standards (limiting the maximum flow rate to 2.0 gpm, or 7.6 Lpm).

In previous studies from multiple U.S. cities, shower durations averaged 7.8 min (DeOreo et al., 2016). The average shower duration observed at the study property was approximately 8 min, with a median value of 6.3. Approximately 25% of the shower durations were longer than 9.5 min (Figure 3.4). The average gallons per flush (gpf) for toilets at the study property was 2.78 (10.5 L), significantly higher than the 1.28 (4.8 L) recommended by the EPA (EPA, 2020), indicating there is potential for reducing water usage by retrofitting the property with water-efficient toilets. There is relatively little variability in the durations of toilet and clothes washer events, as observed in Figure 3.4 c. For these events, the characteristics are dependent on the type, brand and setting used. Shower events reflect the largest variability, as expected, due to personal preferences of the different occupants of the property. Code to reproduce the results in this section and the raw data collected are publicly available in HydroShare (Bastidas Pacheco et al., 2021). The workflow that can be used to reproduce the results presented in this section consists of the following: a) InfluxDB is installed locally with instructions provided, b) the database described in Table 3.3 is created, c) the database is loaded with the raw data

provided using InfluxDB_Loading.ipynb, and then d) data_analytics.ipynb is executed on the database, producing all the results described.

3.3.2. Case study 2: pull based data collection within multi-unit residential buildings

For results of this case study, we present only the data collection and management infrastructure required. The specifics details about estimating and water-related energy use estimates using the data collected are reported elsewhere by Brewer (2020). The functionality of the Data Analytics Layer is independent of the selected data communication method (push or pull) because the Data Analytics Layer interacts only with the operational database. Given that the data collected by both case studies and the resulting database are similar, the considerations for implementing the Data Analytics Layer are equivalent to those of the first case study presented (e.g., ability to support queries, data privacy, etc.) and the technology of the implementation would follow the same process. To avoid duplication of results, we have chosen not to present an implementation of the Data Analytics Layer with this case study. However, similar functionalities related to this case study are discussed in our previous work (Brewer, 2020) and available in an online data resource (Brewer and Horsburgh, 2020).

3.3.2.1. Data collection layer

An enhanced version of the water meter datalogger presented by Horsburgh et al. (2017) was used to collect data for the variables listed in Table 3.1. This device was named the CIWS-EWM-Logger, where EWM denotes “electronic water meter” for the electronic output signal of the meter types it works with. The CIWS-EWM-Logger was designed to be installed on commercial water meters of the types typically used in multi-unit residential buildings and where a dedicated power source is readily available at the

meter's location. The CIWS-EWM-Logger also uses a Raspberry Pi 3 Model B or Model B + Linux computer running Raspberry Pi OS. The Raspberry Pi in this device controls the functioning of the datalogger and has integrated ethernet and Wi-Fi capabilities for connecting to a network while operating. Given the location of the water meters in utility closets with no wired ethernet ports, we chose to use Wi-Fi to enable communications with the dataloggers. Connecting a device to USU's Wi-Fi network requires registration of the device's hardware address, after which, each device is assigned a unique host name that is routable on USU's network. Thus, each datalogger could be located and connected to within the network, which allowed for remote work interactions with the datalogger. For example, the firmware of the loggers could be updated, their functioning could be evaluated in real time, and data could be pulled from them via SSH at any time. While this specific configuration relies on characteristics of USU's Wi-Fi network, we anticipate that Wi-Fi networks like USU's would be available in many application contexts. The functionality described here would function identically for wired ethernet connections.

The CIWS-EWM-Logger was specifically modified to read the output of each of the meters available on the LLC buildings along with water temperature values from three separate sensors. The CIWS-EWM-Loggers we deployed can be used with any water meter or sensor that has a 4–20 mA current loop output, analog voltage output, digital output readable by the Raspberry Pi via its General Purpose Input/Output (GPIO) ports, or pulsed output. The Master Meter Octave meters provide output through a 4–20 mA current loop module where the output current is directly proportional to the flow rate through the meter. The necessary transformations from current to voltage and then to flow rate were performed by the CIWS-EWM-Logger (Brewer, 2020), and a time series

of water flow in gallons per minute at a user-configurable temporal resolution was generated. The BLMJ meter outputs a pulsed signal (voltage) where every pulse represents a gallon of water that has passed through the meter. In this case, the count of pulses, which equals the number of gallons, was registered by the CIWS-EWM-Logger at the same user-configured temporal resolution. The DS18B20 digital thermometers provided digital 9-bit to 12-bit Celsius temperature measurements to an accuracy of ± 0.5 °C and were wired directly to the Raspberry Pi with a single wire for each sensor and do not require an external power supply.

The CIWS-EWM-Logger in each building logged data to a CSV file that was saved in a local directory within the Raspberry Pi's file system. For this deployment, data was collected at a 1-s time interval and includes the following columns: time (datetime of the measurement using the YYYY-MM-DD HH:MM:SS format), buildingID (B, C, D, E, or F), coldInFlowRate, coldInTemp, hotInFlowRate, hotInTemp, hotOutFlowRate and hotOutTemp with units indicated in Table 3.1. In the quality controlled data, the hot water return flow was transformed to gallons per minute for uniformity.

3.3.2.2. Data management and archival layer

To support pull based data retrieval, we developed an application called the Data Transfer Manager (DTM) to serve as the Request Service shown in Figure 3.1. It was implemented as a single Python script named transfer_manager.py and follows the same convention used by the DPS and the DLS, reading configuration data from a JSON file. As in the first Case Study, the DTM and the operational database were deployed on a VM with similar characteristics to the one described in Section 3.1.2. We used InfluxDB as the operational database for this case study as well given the similarity in the type of data

and requirements among both case studies and to show generalizability.

The DTM manages all data communications under the pull based model. Operation of the DTM was scheduled using Linux's native CRON functionality, which allows the user to specify how often the DTM program is executed. Upon being triggered by the scheduled CRON job, the DTM first reads the configuration file described in Table 3.5 and then proceeds through a list of defined tasks to manage transfer of data from each remote data collection site to the Data Management and Archival Layer:

1. Connect to each datalogger listed in the configuration file using Paramiko, a Python library that enables SSH connections for safely accessing network services over unsecured networks (Forcier, 2021).
2. Parse the datalogger's Linux file system for new datalog files and download them to the server with Secure File Transfer Protocol (SFTP), an extension of SSH that offers secure file transfer capabilities over any reliable data stream. Tasks 1 and 2 in this list are executed by a function named connect() in the transfer_manager.py Python script.
3. Upload new data into the InfluxDB database. This task is completed by the write_to_db() function in the transfer_manager.py Python script.

An additional function in the DTM, named send_error(), was developed to inform data managers about errors in the data transfer process. Errors are sent via Slack, a cloud-based instant messaging service (Slack Technologies, 2021). Messages are formulated as a JSON payload that is sent to a unique URL provided by Slack as a webhook. Information detailing which datalogger file caused the error is included in the message. Figure 3.5 describes the overall functionality of the DTM, indicating the key tasks

mentioned. For this case study, data transferring and parsing are executed by a single element (transfer_manager.py), which requires fewer moving parts and minimizes the amount of time between the data being retrieved from the remote dataloggers and having them show up in the operational InfluxDB database. This is a slightly different approach than the one presented for Case Study 1, which allows more flexibility in the system. The DTM can work concurrently on a user defined number of datalogger devices at the same time (connections in Table 3.5). The optimal number of threads is dependent on the number of CPU cores of the server. For our testing, we set the number of threads to 6, matching the number of dataloggers in the LLC buildings.

As in the first case study, the raw data and quality controlled data were stored in the same InfluxDB database in different measurements. Brewer (2020) describes the quality control procedures for the data collected in this case study. The database schema used for this case study is similar in structure to that of the first case study. The data included in the database copies all columns from the CSV files recorded by the dataloggers. BuildingID serves as the SiteID and is the only column stored as a tag. All additional variables (the recorded data values for each variable) are stored as fields.

3.3.3. Scalability and Performance Metrics

While we experienced no performance issues in the case study deployments, we performed scalability testing to investigate the performance of the system beyond the scale of our case studies. We conducted individual tests of the DPS, the DLS, and the DTM, simulating larger numbers of dataloggers and HTTP POST requests, in the case of the DPS and DLS, and a larger number of remote datalogger hosts, in the case of the DTM, to be processed by the system.

Scalability of the DPS is dependent upon its ability to handle many HTTP POST requests from many dataloggers posting data at the same time. The DPS was tested by sending multiple HTTP POST requests, each with a CSV file containing one day of randomly generated data with values recorded every 4 s (for consistency with the implementation of Case Study 1). The files were sent using a Python script implemented using the Asyncio library (Python Software Foundation, 2021a) from a MacBook Pro laptop computer with a 2.3 GHz 8-Core Intel Core i9 processor and 16 GB of memory. Asyncio is a library that can be used to write code that executes concurrently, allowing the code to send multiple simultaneous, or nearly simultaneous, requests to the DPS. There are limitations in the number of concurrent requests that can be sent from the same computer, as well as in the number of dataloggers that can send data at the exact same time in a filed deployment, considering computing power, speed of connection, and synchronization.

We simulated an increasing number of concurrent HTTP POST requests to the DPS (10, 50, 100, 200 and finally 500), and each operation was repeated ten times to characterize server/network variability. The total duration of each repetition, calculated as the end time of the last HTTP POST request minus the start time of the first request, on average, was 0.6 s, 2.05 s, 3.58, 6.91 s, and 16.7 s for 10, 50, 100, 200, and 500 requests, respectively. We observed no transmission errors or requests rejected by the server during our testing process. Figure 3.6 shows the durations of HTTP POST requests, separated by the batch size (10, 50, 100, 200, and 500) for each one of the 10 repetitions conducted. We observed that the median duration of POST requests was larger for the 10-request batches compared to all other batches, but longer durations were observed for

some requests in larger batches, which is expected as the DPS is busy with an increasing number of requests. Median times are consistent for batches with more than 50 POST requests. These times are affected by the processing power of the machine sending the request, the resources available on the remote server, and the speed and quality of the Internet connection but are provided here as an indicator of the performance of our prototype implementation. These tests indicate that the DPS can handle 500 nearly simultaneous POST requests in under 20 s with most individual requests being handled in under 0.2 s.

To test the DLS, we simulated different data loading scenarios ranging from loading one CSV file for a single site to loading one file for 500 sites. The testing procedure consisted of placing CSV files containing one day of data with values recorded every 4 s in the source directory and then executing the DLS. Each operation was repeated ten times. Table 3.6 presents the mean and standard deviation of each scenario along with the average time for loading a single file to facilitate comparisons. The DLS can load 1 day of data from 100 different sites in less than 50 s. There are differences between loading n files from the same site and loading 1 file from n sites, which can be explained by the way data are organized within the InfluxDB database. Although all of the data values are stored in the same InfluxDB measurement, InfluxDB logically groups data values by shared measurement, tag set, and field key. Writing data with multiple siteID tag values takes longer. Both scenarios are realistic applications. The first scenario (n files from 1 site) simulates loading data collected from dataloggers lacking communication technologies. The second scenario (1 file from n sites) represents a deployment like the one described in Case Study 1 with a larger number of sites.

We used the six dataloggers described in Case Study 2 to test the DTM. Each data logger sent 1 day of data during all tests. The functionality that allows the system to identify existing data or files was removed, allowing the system to upload existing CSV files and re-write existing data to the InfluxDB without restrictions. This configuration enabled us to simulate a larger number of connections by repeating dataloggers in the hosts list included in the DTM configuration file (described in Table 3.5). The number of dataloggers was gradually increased (6, 48, 96, and finally 480), and the DTM was executed ten times for each number of dataloggers, processing one CSV file containing one day of 1-s resolution data for each datalogger. The DTM was set to execute six threads at a time, meaning that it can be simultaneously connected to and downloading data from six dataloggers at a time, for consistency with the application of Case Study 2. During our testing, only 6 dataloggers were available, which meant that it was possible for the DTM to attempt connecting to and processing data from the same logger multiple times simultaneously. This can negatively affect the time reported if a host is not immediately available for processing when the system is trying to connect to it. Table 3.7 lists the duration and standard deviation after ten runs with an increasing number of datalogger hosts. Using our test configuration, it took less than 50 min for the DTM to process data from 480 hosts.

We tested the system up to, and with much larger numbers than the 40–60 sites in our design considerations and observed no real limitations for using CIWS in deployments roughly an order of magnitude larger, even with our relatively limited testing server. The DLS and the DTM include writing to the database as part of their tasks, and the times observed satisfy the stated requirements for our application. As a

final test, we tested the database by conducting standard queries from a Python environment, using the same laptop computer. We observed the amount of time required to downloaded one day, one week, and one month of data for 1, 5, and 10 sites along with the time required to load the data into a Pandas dataframe object (Table 3.8). All queries were conducted using the function get_data() described on Table 3.4. The timeit Python module (Python Software Foundation, 2021b) was used to repeat each query 10 times and measure execution times. Downloading one month of data (a common record length in studies collecting high resolution residential water use data) for ten sites into a Pandas dataframe takes less than 1 min. The log files and code to reproduce all the results of this section are publicly available in HydroShare (Bastidas Pacheco et al., 2021).

The cost of deploying CIWS to support data collection at residential houses using the equipment described for Case Study 1 can be broken down as follows: a) the cost of CIWS-NODE Datalogger devices, which is approximately $180 multiplied by the number of houses to be enrolled simultaneously, and; b) the cost of hosting a server with characteristics similar to our testing server (4 processor cores, 8 GB of memory, 100 GB of storage). At the time of this writing, hosting this machine using the Amazon Elastic Compute Cloud would cost approximately $57 per month (Amazon, 2021), although there are multiple hosting alternatives for the server that could be used and that would impact the cost estimate provided. The approximated cost of building the datalogger device used in Case Study 2 is $85.

3.4. Conclusions and future work

A complete cyberinfrastructure system that uses a layered approach to collect and manage high-temporal resolution water use data was developed and implemented. The

system was designed focusing on the scale of data collection that would be required for research projects conducted by utilities or other researchers. Having a standardized cyberinfrastructure like CIWS can increase the value of the data collected by allowing more straightforward data collection and management, as well as facilitating the analysis and understanding of data collected in different projects, cities and utilities. CIWS can be used to manage data collected or used for multiple purposes - e.g., collecting data to support estimates of design parameters for future home developments, guiding the planning of water conservation campaigns, assessing the effectiveness of rebate programs, assisting in the definition of utility rates, and defining future demand and infrastructure needs.

Our case studies showed that CIWS can work with any datalogging devices that generate CSV files containing time series of water use data, but it can also be used in the collection of other variables, as demonstrated in experimental Use Case 2. By integrating low cost data collection devices and open-source cyberinfrastructure we sought to increase the accessibility of tools for conducting high-temporal resolution data collection in support of residential water use studies. CIWS can reduce not only the cost of such studies, but also technical barriers by providing a framework to collect and manage the data.

CIWS can manage push and pull based data communication. Since each functionality is implemented separately, future users of CIWS can select push or pull, or a combination of both, depending on the needs and settings of their application. The work performed within the Data Management and Archival Layer depends on whether the push or pull model is used. In the pull case, the data is pulled from the device by a request

service, whereas in the push case the data is managed by a network listener web service that accepts incoming files and processes them. Both use the same database component, which means that the Data Analytics Layer can operate independent of how the data are transferred. The demonstrations we presented of the Data Analytics Layer serve as a proof of concept and show the foundation upon which more sophisticated tools could be built that can be used to communicate results with multiple interested parties.

We focused our design and implementation on a system that is capable of transferring high temporal resolution water use data from water meters to a centralized infrastructure for storage and subsequent analysis. In a research context, this is preferable, as researchers may not know at the outset of a study all of the specific analyses they may want to perform with the data and, thus, keeping all of the data is necessary. However, transferring large volumes of data to a centralized data management system poses challenges when scaling a system like this to larger deployments. While technically possible over Wi-Fi or cellular data networks, the availability of Wi-Fi is limited, and cost of data transfer over a cellular data network may be prohibitive. As an alternative, we are now investigating edge computing techniques using our CIWS-WM-Node datalogger to process the high resolution water use data on the logger to produce summary data products that are much smaller and can be transferred over a network with far less bandwidth and at lower cost. The tradeoff is that the full resolution data are never transferred or saved in the long term.

CIWS combines multiple open-source technologies. The modular design makes it easier to replace or update technology elements in the system, if needed. Similarly, additional tools can be added to system - e.g., more advanced analytics tools and

enhanced authentication protocols. The analytics presented show potential for conservation programs and can assist in the design of future urban water infrastructure. All of the components we developed are publicly available for reuse, and we envision future improvements to the system once the tools are used in other studies. The system testing, performance metrics, and deployment demonstrate that CIWS can meet and significantly exceed the design considerations in terms of scale and performance. We saw no impediment for using CIWS, or a similar system in larger deployments than the ones tested, by increasing the processing power of the virtual machine, or deploying multiple instances. The server we used for testing had only moderate system specifications and could either be run on private server hardware or could easily be hosted within a commercial cloud service provider at a reasonable monthly cost.

CIWS was designed for research purposes. In consequence, the primary users of the system are researchers interested in analyzing residential water use at high temporal resolution. However, CIWS facilitates the process of generating analyses that would be of interest to residents, utility managers, and city planners. Additionally, CIWS design and implementation provide a proof of concept for designing applications and interoperable solutions, assessing computational needs for similar systems, and for capitalizing on the benefits of such applications that would be of interest for utilities, smart water meter manufacturers, and policy makers.

Software and data availability

Name of Software: Cyberinfrastructure for Intelligent Water Supply (CIWS)

Developers: Camilo J. Bastidas Pacheco, Joseph C. Brewer, Jeffery S. Horsburgh, Juan Caraballo, Elijah West

Contact: jeff.horsburgh@usu.edu

Year First Available: 2021

Required hardware and software: We used open source dataloggers for the data collection efforts in this study. Datalogger hardware details are provided by Bastidas Pacheco et al. (2020) and Horsburgh et al. (2017). Data management and archival components of CIWS were designed to run on a Linux server and were tested using Ubuntu. The data analytics components we demonstrate require a computer running the Windows, Linux, or Macintosh operating system. Instructions for how to deploy the system are available in the project's GitHub repository.

Availability: Source code for the Data Management and Archival Layer software components described in this manuscript is freely available and can be downloaded from the CIWS Server GitHub repository (https://github.com/UCHIC/CIWS-Server). The src folder in that repository contains a folder named ciws_ci and a folder named data_transfer_manager where the elements related to Case Study 1 and Case Study 2 are located, respectively. The doc folder contains a deployment guide for CIWS. The data described in Case Study 1 and the source code of the Data Analytics Layer software are publicly available in HydroShare (Bastidas Pacheco et al., 2021) with instructions for reproducing the results presented in that section. The data described in Case Study 2 and tools used to analyze it are also publicly available in HydroShare (Brewer and Horsburgh, 2020). The log files from Section 3.3 (Scalability and Performance Metrics) and code used to generate the results presented are available in HydroShare (Bastidas Pacheco et al., 2021). Design files, instructions for assembly, and firmware for the open source dataloggers are available on the GitHub sites for the CIWS Water Meter Node datalogger

(https://github.com/UCHIC/CIWS-WM-Node) and the CIWS Electronic Output Water Meter datalogger (https://github.com/UCHIC/CIWS-EWM-Logger).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was funded by the United States National Science Foundation under grant number 1552444. Any opinions, findings, and conclusions or recommendations expressed are those of the authors and do not necessarily reflect the views of the National Science Foundation. We would like to acknowledge Providence City and Utah State University Housing and Facilities for their cooperation and support in the data collection efforts. The authors would also like to acknowledge support from Nour Atallah, Arle J. Beckwith, and Rob J. Tracy in the data collection efforts and Elijah West for his contribution in software development. We also acknowledge and thank the owner of the residential home and the students in the LLC buildings that participated in the data collection campaign.

REFERENCES

Alghamdi, A., Shetty, S., 2016. Survey toward a smart campus using the internet of things, in: Proceedings - 2016 IEEE 4th International Conference on Future Internet of Things and Cloud, FiCloud 2016. Institute of Electrical and Electronics Engineers Inc., pp. 235–239. https://doi.org/10.1109/FiCloud.2016.41

Alvisi, S., Casellato, F., Franchini, M., Govoni, M., Luciani, C., Poltronieri, F., Riberto, G., Stefanelli, C., Tortonesi, M., 2019. Wireless Middleware Solutions for Smart Water Metering. Sensors 19, 1853. https://doi.org/10.3390/s19081853

Amaxilatis, D., Chatzigiannakis, I., Tselios, C., Tsironis, N., Niakas, N., Papadogeorgos, S., 2020. A Smart Water Metering Deployment Based on the Fog Computing Paradigm. Appl. Sci. 10, 1965. https://doi.org/10.3390/app10061965

Amazon, 2021. Amazon EC2 Secure and resizable compute capacity to support virtually any workload. URL: https://aws.amazon.com/ec2/?nc2=h_ql_prod_fs_ec2&ec2-whats-new.sort-by=item.additionalFields.postDateTime&ec2-whats-new.sort-order=desc (accessed 5.5.21)

Anda, M., Le Gay Brereton, F., Brennan, J., Paskett, E., 2013. Smart metering infrastructure for residential water efficiency: Results of a trial in a behavioural change program in Perth, Western Australia, in: Information and Communication Technologies for Sustainability. ETH Zurich, Zurich. https://researchrepository.murdoch.edu.au/id/eprint/22422/

Ardito, L., Procaccianti, G., Menga, G., Morisio, M., 2013. Smart Grid Technologies in Europe: An Overview. Energies 6, 251–281. https://doi.org/10.3390/en6010251

Attallah, N.A., Horsburgh, J.S., Bastidas Pacheco, C.J., 2021a. Tools for Evaluating, Developing, and Testing Water End Use Disaggregation Algorithms. Manuscript submitted for publication.

Attallah, N., J. S. Horsburgh, C. J. Bastidas Pacheco 2021b. Tools for Evaluating, Developing, and Testing Water End Use Disaggregation Algorithms, HydroShare, http://www.hydroshare.org/resource/1521eba67f1d4571ac5fe2b8c5e01035

Balis, B., Bubak, M., Harezlak, D., Nowakowski, P., Pawlik, M., Wilk, B., 2017. Towards an operational database for real-time environmental monitoring and early warning systems, in: Procedia Computer Science. Elsevier B.V., pp. 2250–2259. https://doi.org/10.1016/j.procs.2017.05.193

Bastidas Pacheco, C. J., J. S. Horsburgh, J. Caraballo, N. Attallah, 2021. Supporting data and tools for "An open source cyberinfrastructure for collecting, processing, storing and accessing high temporal resolution residential water use data", HydroShare, http://www.hydroshare.org/resource/aaa7246437144f2390411ef9f2f4ebd0

Bastidas Pacheco, C.J., Horsburgh, J.S., Tracy, R.J., 2020. A Low-Cost, Open Source Monitoring System for Collecting High Temporal Resolution Water Use Data on Magnetically Driven Residential Water Meters. Sensors 20, 3655. https://doi.org/10.3390/s20133655

Beal, C., Stewart, R.A., 2011. South East Queensland Residential End Use Study: Final Report, Urban Water Security Research Alliance. https://research-repository-.griffith.edu.au/bitst-ream/handle/10072/46802/80687_2.pdf?sequence=1

Beal, C.D., Flynn, J., 2015. Toward the digital water age: Survey and case studies of Australian water utility smart-metering programs. Utilities Policy 32, 29–37. https://doi.org/10.1016/j.jup.2014.12.006

Beal, C.D., Stewart, R.A., Fielding, K., 2013. A novel mixed method smart metering approach to reconciling differences between perceived and actual residential end use water consumption. Journal of Cleaner Production. 60, 116–128. https://doi.org/10.1016/j.jclepro.2011.09.007

Boyle, T., Giurco, D., Mukheibir, P., Liu, A., Moy, C., White, S., Stewart, R., 2013. Intelligent Metering for Urban Water: A Review. Water 5, 1052–1081. https://doi.org/10.3390/w5031052

Brewer, J., J. S. Horsburgh (2020). Characterizing Water and Water-Related Energy in Multi-Unit Residential Structures with High Resolution Smart Meter Data, HydroShare, http://www.hydroshare.org/resource/b6bbdcd9b120430b9a54974a798961f1

Brewer, Joseph C., "Characterizing Water and Water-Related Energy Use in Multi-Unit Residential Structures with High Resolution Smart Metering Data" (2020). All Graduate Theses and Dissertations. 7976. https://doi.org/10.26076/669a-93b0

Cardell-Oliver, R., 2013. Water use signature patterns for analyzing household consumption using medium resolution meter data. Water Resour. Res. 49, 8589–8599. https://doi.org/10.1002/2013WR014458

Chen, F., Dai, J., Wang, B., Sahu, S., Naphade, M., Lu, C.T., 2011. Activity analysis based on low sample rate smart meters, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM Press, New York, New York, USA, pp. 240–248. https://doi.org/10.1145/2020408.2020450

Christie, M.A., Bhandar, A., Nakandala, S., Marru, S., Abeysinghe, E., Pamidighantam, S., Pierce, M.E., 2020. Managing authentication and authorization in distributed science gateway middleware. Futur. Gener. Comput. Syst. 111, 780–785. https://doi.org/10.1016/j.future.2019.07.018

Cominola, A., Giuliani, M., Castelletti, A., Rosenberg, D.E., Abdallah, A.M., 2018. Implications of data sampling resolution on water use simulation, end-use disaggregation, and demand management. Environmental Modelling & Software. 102, 199–212. https://doi.org/10.1016/j.envsoft.2017.11.022

Cominola, A., Giuliani, M., Piga, D., Castelletti, A., Rizzoli, A.E., 2015. Benefits and challenges of using smart meters for advancing residential water demand modeling and management: A review. Environmental Modelling & Software. 72, 198–214. https://doi.org/10.1016/j.envsoft.2015.07.012

DeOreo, W.B., Mayer, P.W., Dziegielewski, B., Kiefer, J., Foundation, W.R., 2016. Residential End Uses of Water, Version 2. Water Research Foundation.

https://www.waterrf.org/resource/residential-end-uses-water-version-2

DeOreo, W.B., Mayer, P.W., Martien, L., Hayden, M., Funk, A., Kramer-Duffield, M., Davis, R., Henderson, J., Raucher, B., Gleick, P., 2011. California single-family water use efficiency study, Report prepared for the California Dept. of Water Resources, Aquacraft Inc., Boulder, CO. https://cawaterlibrary.net/document/california-single-family-water-use-efficiency-study/

Di Martino, S., Fiadone, L., Peron, A., Vitale, V.N., Riccabone, A., 2019. Industrial Internet of Things: Persistence for Time Series with NoSQL Databases, in: Proceedings - 2019 IEEE 28th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises, WETICE 2019. Institute of Electrical and Electronics Engineers Inc., pp. 340–345. https://doi.org/10.1109/WETICE.2019.00076

Di Mauro, A., Di Nardo, A., Santonastaso G. F. ., and Venticinque, S., 2019. An IoT System for Monitoring and Data Collection of Residential Water End-Use Consumption, in 28th International Conference on Computer Communication and Networks (ICCCN), pp. 1-6, https://doi.org10.1109/ICCCN.2019.8847120

Di Mauro, A., Cominola, A., Castelletti, A., Di Nardo, A., 2020. Urban Water Consumption at Multiple Spatial and Temporal Scales. A Review of Existing Datasets. Water 13, 36. https://doi.org/10.3390/w13010036

EPA, 2020. The Water Sense Label. URL https://www.epa.gov/watersense, (accessed 01.05.2021).

Erickson, T., Podlaseck, M.E., Sahu, S., Dai, J.D., Chao, T., Naphade, M., 2012. The Dubuque Water Portal: Evaluation of the uptake, use and impact of residential water consumption feedback, in: Conference on Human Factors in Computing Systems - Proceedings. ACM Press, New York, New York, USA, pp. 675–684. https://doi.org/10.1145/2207676.2207772

Fang, D., Chen, B., 2017. Linkage analysis for the water–energy nexus of city. Appl. Energy 189, 770–779. https://doi.org/10.1016/j.apenergy.2016.04.020

F.S. Brainard & Company, 2020. Meter-Master. http://www.meter-master.com/products/pdf_products/MM100EL_cut.pdf

Forcier, Jeff, 2021. Welcome to Paramiko URL: http://www.paramiko.org/ (accessed 5.5.21)

Froehlich, J., Findlater, L., Ostergren, M., Ramanathan, S., Peterson, J., Wragg, I., Larson, E., Fu, F., Bai, M., Patel, S.N., Landay, J.A., 2012. The design and evaluation of prototype eco-feedback displays for fixture-level water usage data, in: Conference on Human Factors in Computing Systems - Proceedings. ACM Press, New York, New York, USA, pp. 2367–2376. https://doi.org/10.1145/2207676.2208397

Giurco, D., Carrard, N., McFallan, S., Nalbantoglu, M., Inman M., Thornton N., White, S. 2008. Residential End use Measurement Guidebook: A Guide to Study Design, Sampling and Technology. Prepared by the Institute for Sustainable Futures UTS

and CSIRO for the Smart Water Fund, Victoria. https://opus.lib.uts.edu.au/bitstream/10453/35089/1/giurcoetal2008resenduse.pdf

Gunicorn, 2021. Gunicorn. URL: https://gunicorn.org/ (accessed 5.5.21)

Hachmann, J., Afzal, M.A.F., Haghighatlari, M., Pal, Y., 2018. Building and deploying a cyberinfrastructure for the data-driven design of chemical systems and the exploration of chemical space. Molecular Simulation. 44, 921–929. https://doi.org/10.1080/08927022.2018.1471692

Hamiche, A. M., Stambouli, A. B., & Flazi, S. (2016), A review of the water-energy nexus, Renewable and Sustainable Energy Reviews, 65, 319–331, https://doi.org/10.1016/j.rser.2016.07.020.

Hauser, A., Roedler, F., 2015. Interoperability: The key for smart water management. Water Supply 15, 207–214. https://doi.org/10.2166/ws.2014.096

Hauser, A., Sud, T., Nicolas Foret, C., Electric Stuart Combellack, S., Jonathan Coome, T., Quintilia Lopez, S., Elkin Hernandez, I., Water Salil Kharkar, D.M., Water Amin Rasekh, D., Michal Koenig, S., Remy Marcotorchino, Q., Nicolas Damour, S., 2016. Communication in Smart Water Networks SWAN Forum Interoperability Workgroup. https://www.swan-forum.com/wp-content/uploads/sites/218/2020/12/SWAN-White-Paper_Communication-Protocols.pdf

Heiland, R., Koranda, S., Marru, S., Pierce, M., Welch, V., 2015. Authentication and authorization considerations for a multi-tenant service, in: SCREAM 2015 - Proceedings of the 2015 Workshop on the Science of Cyberinfrastructure: Research, Experience, Applications and Models, Part of HPDC 2015. Association for Computing Machinery, Inc, New York, New York, USA, pp. 29–35. https://doi.org/10.1145/2753524.2753534

Hellcamp, 2021. Bottle: Python Web Framework. URL https://bottlepy.org/ (accessed 5.5.21)

Hollands, R.G., 2008. Will the real smart city please stand up? Intelligent, progressive or entrepreneurial? City 12, 303–320. https://doi.org/10.1080/13604810802479126

Horsburgh, J.S., Caraballo, J., Ramírez, M., Aufdenkampe, A.K., Arscott, D.B., Damiano, S.G., 2019. Low-Cost, Open-Source, and Low-Power: But What to Do With the Data? Frontiers in Earth Science. 7. https://doi.org/10.3389/feart.2019.00067

Horsburgh, J.S., Leonardo, M.E., Abdallah, A.M., Rosenberg, D.E., 2017. Measuring water use, conservation, and differences by gender using an inexpensive, high frequency metering system. Environmental Modeling & Software. 96, 83–94. https://doi.org/10.1016/j.envsoft.2017.06.035

InfluxDB, 2020. API Documentation. URL https://influxdb-python.readthedocs.io/en/latest/index.html (accessed 2.1.21).

InfluxData, 2021, InfluxDB. URL https://www.influxdata.com/products/influxdb/ (accessed 5.5.21)

Kenway, S.J., Binks, A., Scheidegger, R., Bader, H.-P., Pamminger, F., Lant, P., Taimre, T., 2016. Household analysis identifies water-related energy efficiency opportunities. Energy Build. 131, 21–34. https://doi.org/10.1016/j.enbuild.2016.09.008

Kim, H., Lee, E.A., 2017. Authentication and Authorization for the Internet of Things. IT Prof. 19, 27–33. https://doi.org/10.1109/MITP.2017.3680960

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., Willing, C., 2016. Jupyter Notebooks—a publishing format for reproducible computational workflows, in: Loizides, F., Scmidt, B. (Eds.), Positioning and Power in Academic Publishing: Players, Agents and Agendas - Proceedings of the 20th International Conference on Electronic Publishing, ELPUB 2016. IOS Press BV, pp. 87–90. https://doi.org/10.3233/978-1-61499-649-1-87

Kofinas, D.T., Spyropoulou, A., Laspidou, C.S., 2018. A methodology for synthetic household water consumption data generation. Environ. Model. Softw. 100, 48–66, https://doi.org/10.1016/j.envsoft.2017.11.021

Li, J., Yang, X., Sitzenfrei, R., 2020. Rethinking the Framework of Smart Water System: A Review. Water 12, 412. https://doi.org/10.3390/w12020412

Liaw A, Wiener M: Classification and regression by randomForest. Rnews 2002, 2: 18–22.

Liu, A., Giurco, D., Mukheibir, P., 2015. Motivating metrics for household water-use feedback. Resources, Conservation and Recycling. 103, 29–46. https://doi.org/10.1016/j.resconrec.2015.05.008

Liu, X., Nielsen, P.S., 2016. A hybrid ICT-solution for smart meter data analytics. Energy 115, 1710–1722. https://doi.org/10.1016/j.energy.2016.05.068

Makropoulos, C., 2017. Thinking platforms for smarter urban water systems: Fusing technical and socio-economic models and tools. Geol. Soc. Spec. Publ. 408, 201–219. https://doi.org/10.1144/SP408.4

Mason, S.J.K., Cleveland, S.B., Llovet, P., Izurieta, C., Poole, G.C., 2014. A centralized tool for managing, archiving, and serving point-in-time data in ecological research laboratories. Environmental Modelling & Software. 51, 59–69. https://doi.org/10.1016/j.envsoft.2013.09.008

Mayer, P.W., B. DeOreo, W., Towler, E., Martien, L., M. Lewis, D., 2004. Tampa Water Department Residential Water Conservation Study: The Impacts of High Efficiency Plumbing Fixture Retrofits in Single-Family Homes. A Report Prepared for Tampa Water Department and the United States Environmental Protection Agency.

Mayer, P.W., DeOreo, W.B., Optiz, E.M., Kiefer, J.C., Davis, W.Y., Dziegielewski, B., Nelson, J.O., 1999. Residential End Uses of Water. American Water Works Association. https://www.sdu.dk/~/media/Files/Om_SDU/Institutter/ITI/Forskning/NATO%20ARW/Literature/Residential%20end%20uses_of%20water.pdf

McKinney, W., 2010. Data Structures for Statistical Computing in Python (SCIPY), in: van der Walt, S., Millman, J. (Eds.), Proceedings of the 9th Python in Science Conference. pp. 56–61. https://doi.org/10.25080/Majora-92bf1922-00a

Moy De Vitry, M., Schneider, M.Y., Wani, O., Manny, L., Leitao, J.P., Eggimann, S., 2019. Smart urban water systems: What could possibly go wrong? Environmental Research Letters. https://doi.org/10.1088/1748-9326/ab3761

Mutchek, M., Williams, E., 2014. Moving Towards Sustainable and Resilient Smart Water Grids. Challenges 5, 123–137. https://doi.org/10.3390/challe5010123

Neirotti, P., De Marco, A., Cagliano, A.C., Mangano, G., Scorrano, F., 2014. Current trends in smart city initiatives: Some stylised facts. Cities 38, 25–36. https://doi.org/10.1016/j.cities.2013.12.010

NGINX, 2021. About NGINX. URL: https://nginx.org/en/ (accessed 5.5.21)

Nguyen, K.A., Stewart, R.A., Zhang, H., Jones, C., 2015. Intelligent autonomous system for residential water end use classification: Autoflow. Appl. Soft Comput. 31, 118–131. https://doi.org/10.1016/j.asoc.2015.03.007

NSF, 2007. Cyberinfrastructure Vision for 21st Century Discovery. https://www.nsf.gov/pubs/2007/nsf0728/index.jsp

Python Software Foundation, 2021a. asyncio — Asynchronous I/O. URL: https://docs.python.org/3/library/asyncio.html (accessed 5.5.21)

Python Software Foundation, 2021b. timeit — Measure execution time of small code snippets. URL: https://docs.python.org/3/library/timeit.html (accessed 5.5.21)

Rinaldi, S., Bonafini, F., Ferrari, P., Flammini, A., Sisinni, E., Bianchini, D., 2019. Impact of data model on performance of time series database for internet of things applications, in: I2MTC 2019 - 2019 IEEE International Instrumentation and Measurement Technology Conference, Proceedings. https://doi.org/10.1109/I2MTC.2019.8827164

Robles, T., Alcarria, R., Martin, D., Morales, A., Navarro, M., Calero, R., Iglesias, S., Lopez, M., 2014. An internet of things-based model for smart water management, in: 2014 IEEE 28th International Conference on Advanced Information Networking and Applications Workshops, IEEE WAINA 2014. IEEE Computer Society, pp. 821–826. https://doi.org/10.1109/WAINA.2014.129

Shams, S., Goswami, S., Lee, K., Yang, S., Park, S.J., 2018. Towards distributed cyberinfrastructure for smart cities using big data and deep learning technologies, in: Proceedings - International Conference on Distributed Computing Systems. Institute of Electrical and Electronics Engineers Inc., pp. 1276–1283. https://doi.org/10.1109/ICDCS.2018.00127

Slack Technologies, 2021. Slack. URL: https://slack.com (accessed 5.5.21)

Souffront Alcantara, Michael A., Kesler, Christian, Stealey, Michael J., Nelson, E. James, Ames, Daniel P., and Jones, Norm L., 2017. Cyberinfrastructure and Web Apps for Managing and Disseminating the National Water Model. Journal of the American Water Resources Association (JAWRA) 54(4): 859- 871.

https://doi.org/10.1111/1752-1688.12608

Stiri, S., Chaoub, A., Bennani, R., Lakssir, B., Tamtaoui, A., 2019. Internet of things connectivity-based smart grids in Morocco: Proof of concept and guide to massive deployments, in: 5th IEEE International Smart Cities Conference, ISC2 2019. Institute of Electrical and Electronics Engineers Inc., pp. 129–135. https://doi.org/10.1109/ISC246665.2019.9071734

Stocks, K.I., Schramski, S., Virapongse, A., Kempler, L., 2019. Geoscientists' perspectives on cyberinfrastructure needs: A collection of user scenarios. Data Science Journal. 18, 1–15. https://doi.org/10.5334/dsj-2019-021

Sun, Z., Di, L., Cash, B., Gaigalas, J., 2020. Advanced cyberinfrastructure for intercomparison and validation of climate models. Environmental Modelling & Software. 123. https://doi.org/10.1016/j.envsoft.2019.104559

Szwilski, T.B., Smith, J., Chapman, J., Lewis, M., 2018. Cyberinfrastructure supporting watershed health monitoring and management. WIT Trans. Ecol. Environ. 228, 245–256. https://doi.org/10.2495/WP180231

U.S. Congress, 2007. Energy Independence and Security Act - SMART GRID. https://www.govinfo.gov/content/pkg/STATUTE-121/pdf/STATUTE-121-Pg1492.pdf, United States of America.

Wegrzyn, J.L., Falk, T., Grau, E., Buehler, S., Ramnath, R., Herndon, N., 2020. Cyberinfrastructure and resources to enable an integrative approach to studying forest trees. Evol. Appl. https://doi.org/10.1111/eva.12860

Willis, R.M., Stewart, R.A., Giurco, D.P., Talebpour, M.R., Mousavinejad, A., 2013. End use water consumption in households: Impact of socio-demographic factors and efficient devices. Journal of Cleaner Production 60, 107–115. https://doi.org/10.1016/j.jclepro.2011.08.006

Willis, R.M., Stewart, R.A., Williams, P.R., Hacker, C.H., Emmonds, S.C., Capati, G., 2011. Residential potable and recycled water end uses in a dual reticulated supply system. Desalination 272, 201–211. https://doi.org/10.1016/j.desal.2011.01.022

Wissner, M., 2011. The Smart Grid - A saucerful of secrets? Appl. Energy 88, 2509–2518. https://doi.org/10.1016/j.apenergy.2011.01.042

Yan, Y., Qian, Y., Sharif, H., Tipper, D., 2013. A survey on smart grid communication infrastructures: Motivations, requirements and challenges. IEEE Commun. Surv. Tutorials. https://doi.org/10.1109/SURV.2012.021312.00034

Ye, Y., Liang, L., Zhao, H., Jiang, Y., 2016. The System Architecture of Smart Water Grid for Water Security, in: Procedia Engineering. Elsevier Ltd, pp. 361–368. https://doi.org/10.1016/j.proeng.2016.07.492

Zanella, A., Bui, N., Castellani, A., Vangelista, L., Zorzi, M., 2014. Internet of things for smart cities. IEEE Internet Things J. 1, 22–32. https://doi.org/10.1109/JIOT.2014.2306328

Tables

Table 3.1. Variables measured, measuring device, and units of observation at each LLC building.

Measured Variable	Measuring device	Units
1) Hot-water supply flow 2) Cold-water supply flow	Master Meter Octave Ultrasonic water meter with 4-20 mA current loop outputs	gpm
3) Hot-water return flow	Master Meter Bottom Load Multi-Jet (BLMJ) water meter with pulsed output	pulses
4) Cold water supply temperature 5) Hot water supply temperature 6) Hot water return temperature	DS18B20 digital thermometer with digital output	^{o}C

Table 3.2. Parameters included in the configuration file for the data posting (DPS) and data loading (DLS) services. The configuration file follows the structure presented here.

Parameter		Description
log_directory		Directory where the log files are located.
source_directory		Directory where the files accepted by the DPS are placed. The DLS processes the files located in this directory.
target_directory		Directory where the CSV files will be moved to after the data is uploaded into the database for archival.
quarantine_directory		Directory where the CSV files will be moved to if an error occurs.
client_token		A public key used to generate upload tokens and authenticate upload requests.
secret_key		A private key used to generate the upload tokens.
database	name	Name of the InfluxDB database used.
	user	Username of the InfluxDB user used when connecting to the database.
	password	InfluxDB Password for the user selected.
	host	The host name of the server on which the InfluxDB database is installed.
	port	The Internet port number over which communications with the InfluxDB database server have been configured.

Table 3.3. InfluxDB database schema design in the push model implementation.

Influx Key	InfluxDB Type	Data Type	Example Value
time	Time Index	Timestamp	2020-01-01 00:00:01
siteID	Tag	String	“1”
pulses	Field	Integer	5
dataloggerID	Tag	String	“1”

Table 3.4. Functions implemented for querying data in the Data Analytics Layer.

Query	Python implementation
Get raw data for one or multiple sites, between specific dates, or the entire record.	get_data(site, startdate = None, enddate = None, measurement = ‘QCData’)
Get time aggregated data for one or multiple sites, between specific dates, or the entire record.	get_agg(site, function, t_res, startdate = None, enddate = None, Measurement = ‘QCData’)

Table 3.5. Parameters included in the configuration file for the DTM. The configuration file follows the structure presented here.

Parameter		Description
connections		The number of threads used for concurrent connection with hosts.
log_directory		Path where the log files are stored in the Data Management and Archival server (must have write permissions for that directory).
hosts		A list of datalogger host names or IP addresses to connect to.
database	name	Name of the InfluxDB database to connect to.
	user	Username for a user with permission to write data to the InfluxDB database.
	password	Password for for a user with permission to write data to the InfluxDB database.
	host	Database server hostname or IP address.
	port	The port number over which communications with the InfluxDB database server have been configured.
	measurement	Name of InfluxDB Measurement where the data will be saved.
sshinfo	username	Username used to connect to remote dataloggers via SSH.
	password	Password used to connect to remote dataloggers via SSH.
slack_webhook		Slack webhook to send error messages through the Slack messaging service.

Table 3.6. Results from the DLS testing. Every operation was repeated 10 times.

Load Operation	Average duration (seconds)	Standard deviation (seconds)	Average time for processing 1 file (seconds)
1 file from 1 site	0.37	0.06	0.37
10 files from 1 site	3.96	0.14	0.40
1 file from 10 sites	4.67	0.23	0.47
50 files from 1 site	19.92	0.67	0.40
1 file from 50 sites	23.87	0.33	0.48
100 files from 1 site	39.87	1.05	0.40
1 file from 100 sites	47.48	0.89	0.47
500 files from 1 site	195.19	2.98	0.39
1 file from 500 sites	240.70	3.00	0.48

Table 3.7. Results from the DTM testing.

Number of datalogger hosts	Average duration (seconds)	Standard deviation (seconds)
6	41.7	1.57
48	279.4	9.35
96	551.5	9.21
480	2,831	252

Table 3.8. InfluxDB downloading times for different queries. In all cases the data was downloaded and loaded into a Pandas dataframe.

Days of data	Number of sites	Average duration (seconds)	Standard deviation (seconds)
1	1	0.17	0.02
1	5	0.81	0.03
1	10	1.62	0.04
7	1	1.16	0.04
7	5	5.74	0.07
7	10	11.39	0.07
30	1	4.51	0.27
30	5	22.46	0.52
30	10	45.47	1.24

Figures

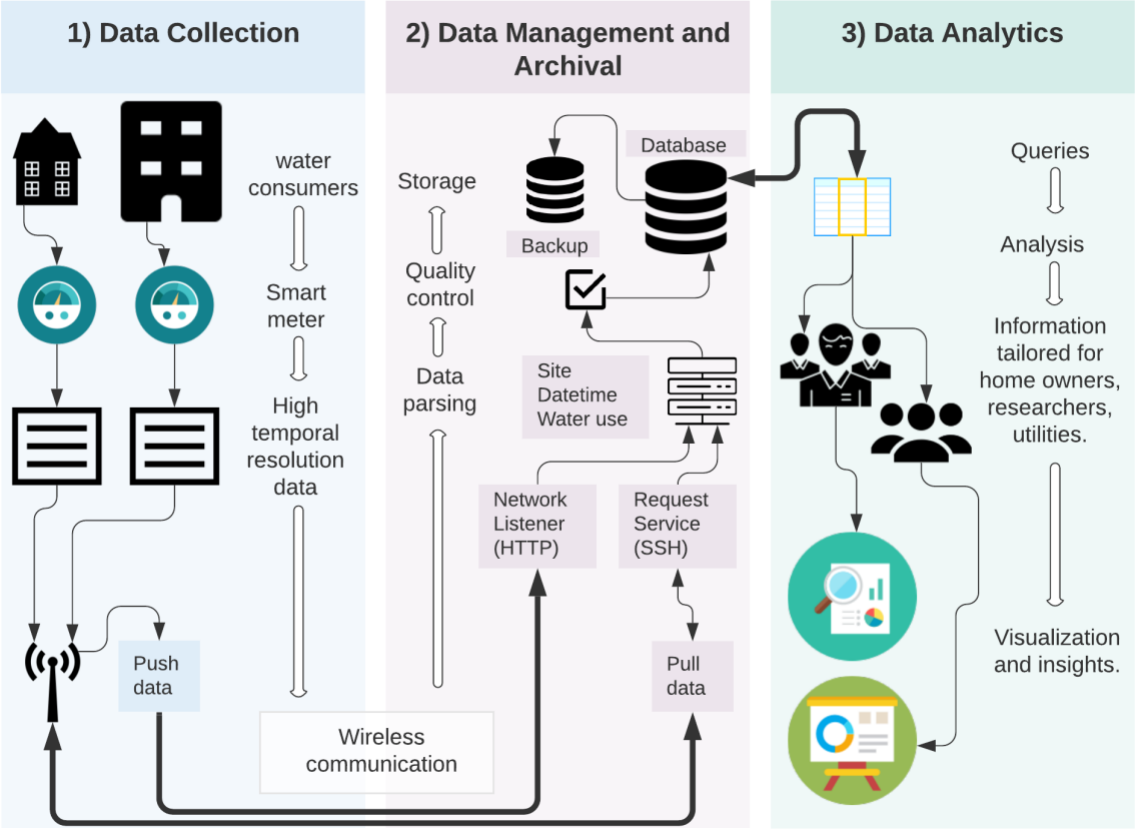


Figure 3.1. Overall architecture design of CIWS consisting of three main layers: 1) Data Collection, 2) Data Management and Archival, and 3) Data Analytics. Arrows are used to indicate data and workflow movement between components. White arrows indicate the flow of data and information and black arrows show the connection between elements and layers.

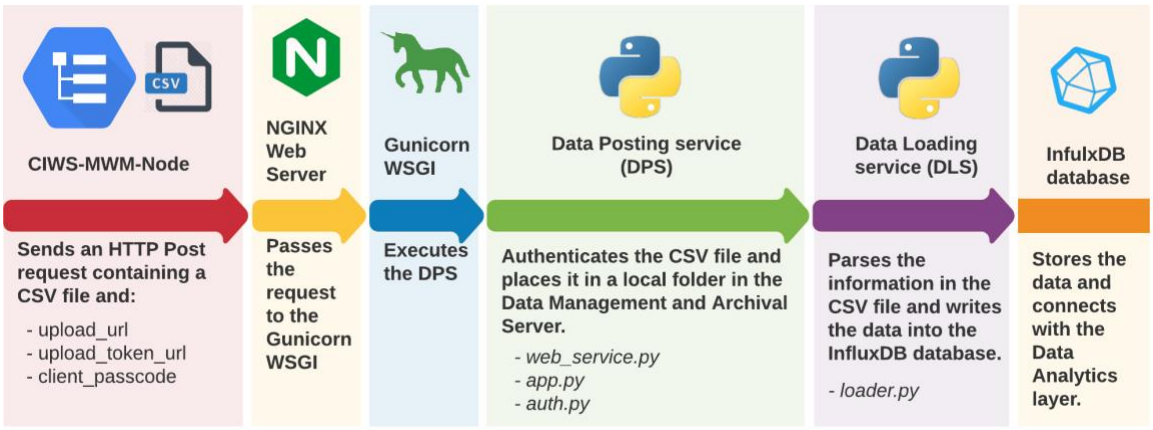


Figure 3.2. Workflow and elements of the data management process for the push based implementation of the CIWS.

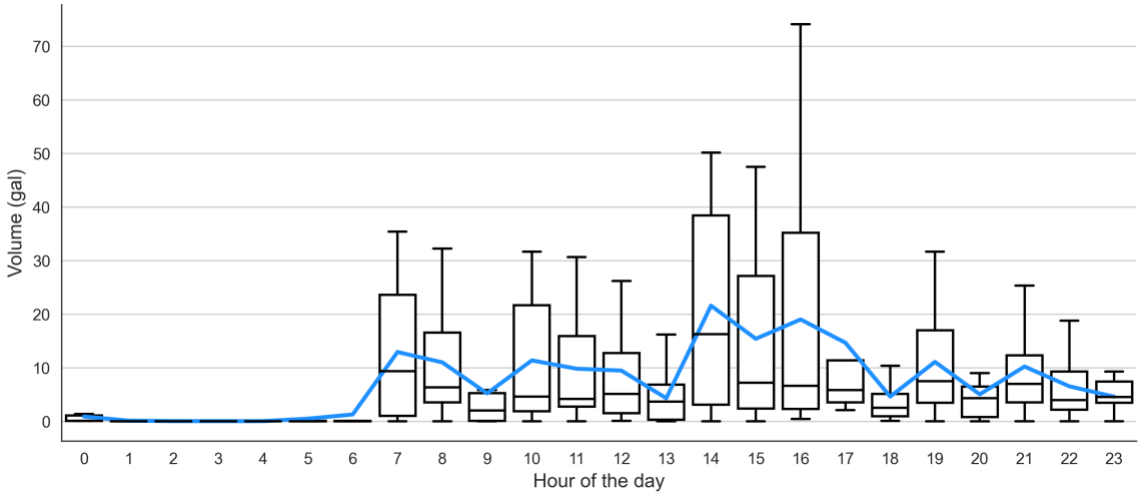


Figure 3.3. Hourly distribution of water use for the single family residential home between January 15, 2021 and January 28, 2021. The blue solid line shows the hourly average water use and the boxplot presents hourly water use variability.

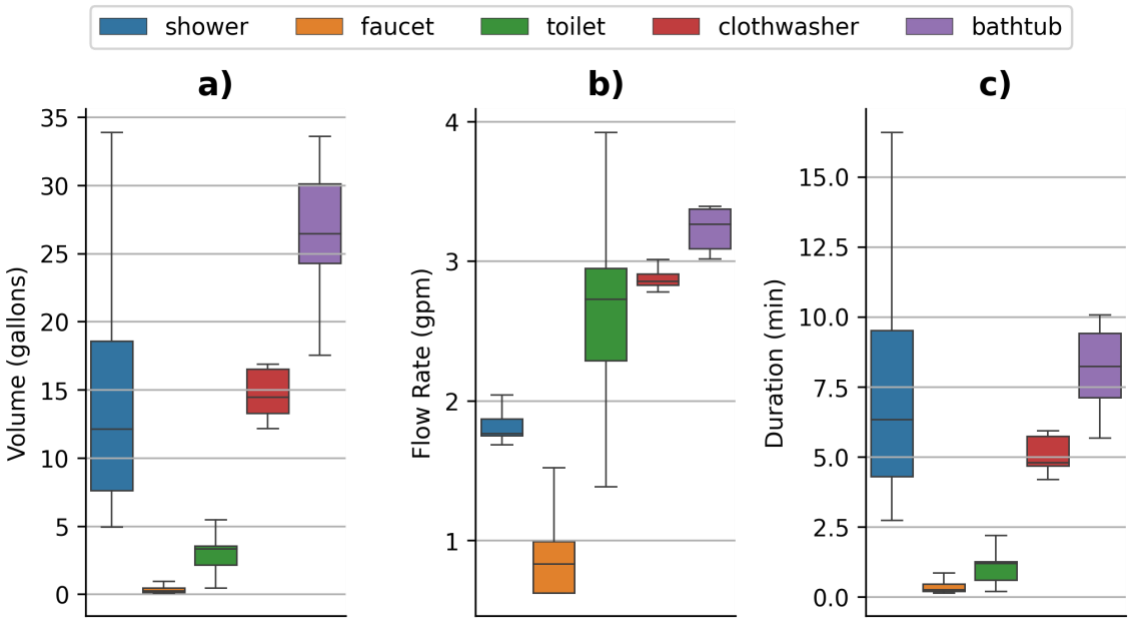


Figure 3.4. Illustrative examples of high-temporal residential water use data analytics for the case study home between January 15, 2021 and January 28, 2021. The figure presents boxplots of a) the volume of events, b) the flow rate of events, c) the duration of events. In all cases, the data is grouped by end use type. Outliers were removed to improve the quality of visualization for short duration and low volume events (faucet and toilet events).

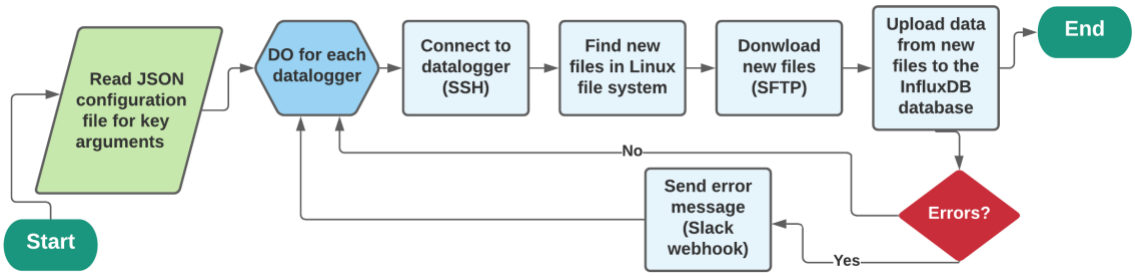


Figure 3.5. General functionality of the DTM.

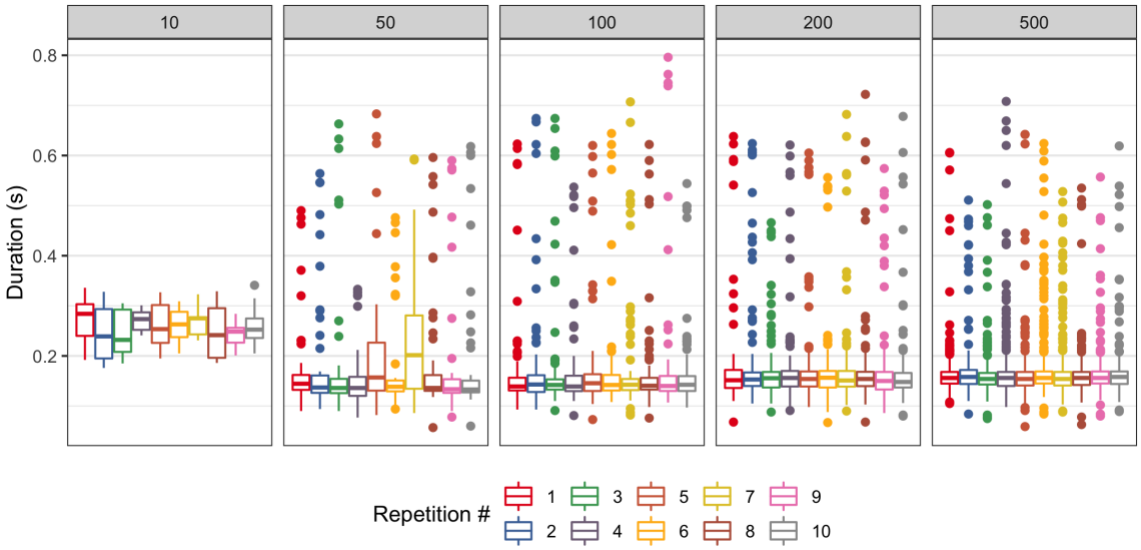


Figure 3.6. Boxplot of processing times, separated by the number of HTTP POST requests in the batch (10, 50, 100, 200, and 500) for each repetition, from 1 to 10. Duration is calculated as the final processing time minus the starting time of each individual POST request.

CHAPTER 4

VARIABILITY IN CONSUMPTION AND END USES OF WATER FOR RESIDENTIAL USERS IN LOGAN AND PROVIDENCE, UTAH, USA[1]

Abstract

Variation in water fixtures and appliances coupled with the different routines and preferences of users result in high levels of variability in residential water consumption. This study assessed differences in residential water use in terms of timing and distribution of end uses across residential properties. Past studies analyzing residential end uses of water have collected data for periods of time that may prevent observing temporal variations in indoor and outdoor water use practices. We examined indoor and outdoor residential water use at the household level by analyzing four to 23 weeks of 4-second resolution water use data at 31 single family residential properties in Logan and Providence, Utah, USA between 2019 and 2021. We identified and classified end uses of water for each property and analyzed monthly water use records to understand how water use varies for users at different levels of consumption. Our results indicate that indoor water use is influenced more by the frequency of use than by the characteristics of water using fixtures. At sites with longer data collection periods, indoor water use volume, timing, and distribution across end uses varied across homes and across weeks for which we collected data. We illustrate opportunities to conserve water indoors and outdoors by adopting more efficient fixtures (particularly toilets), promoting conservation behaviors related to shower durations, and reducing irrigation when rainfall occurs. All data and

---

[1] Co-authored by Camilo J. Bastidas Pacheco, Jeffery S. Horsburgh, and Nour A. Attallah

tools used in this study are freely available online for reuse.

4.1 Introduction

Residential water use in the state of Utah has been estimated at approximately 640 L per capita per day, which is the second largest in the United States (Dieter et al. 2018). Approximately 98% of the state's population is served by public water suppliers, one of the highest percentages in the country (Dieter et al. 2018). It is estimated that Utah will need a $4.4 billion investment, over a 20-year period, to maintain current service and meet future demands (EPA 2018). It was estimated that in 2010 91% of Utah's population was living in urban areas (The University of Utah 2016). This pattern is repeated in many areas within the United States, where the urban population grew much faster (500%) than the rural population (19%) between 1910 and 2010 (EPA 2016). The percent of the world's population living in urban areas increased from 43% to 54% between 1990 and 2015 (UN-Habitat 2016). With this increase in urban density and the costs associated with delivering water to urban populations, managing and reducing demand is vital for providing clean and safe water supply for the world's growing urban populations.

In order to accurately estimate and forecast urban water consumption, it is important to know the different daily patterns in consumption, the distribution of water use across end uses, how that distribution varies across time, and potential savings from different conservation programs or demand side measures (Willis et al. 2011). Conventional water use data (collected at monthly, bimonthly, or coarser resolutions) analyses leave knowledge gaps with regard to water use peak times and volumes along with detailed estimates of indoor versus outdoor water use. By one estimate, per capita

water use has decreased by 4.4% between 2010 and 2015 in the United States (Dieter et al. 2018). It is commonly assumed that decreases in residential water use are produced by the use of more efficient fixtures, yet few definitive statements can be made about this because little data exist to directly measure the performance and impact of retrofitted fixtures (Rockaway et al. 2010). Analyses derived from high temporal resolution data aimed at demand management, evaluation of fixture performance, or evaluation of conservation potential can address these gaps, yet this type of data has only been collected sporadically and over short periods of time, generating uncertainty about how generalizable and applicable the results obtained are.

Researchers are increasingly using smart meters and advanced analytics to monitor water use at finer temporal resolutions at the household level (Cominola et al. 2015). The potential for these technologies to address the existing gaps in residential water use knowledge is well recognized (Boyle et al. 2013). Smart meters sampling at high temporal frequencies can aid in the identification and quantification of individual water end uses, reveal water use behavior, and can also help detect and reduce the volume of leaks (Cominola et al. 2018). Furthermore, feedback to water users on their water use has the potential to motivate conservation behaviors (Cominola et al. 2021). For example, Fielding et al. (2013) noted significant differences in water usage for users receiving water use feedback derived from high temporal resolution (5 s) data.

"End uses" of water refers to the distribution of water usage across different uses (e.g., faucets, showers, toilets), and this information is needed to produce more accurate demand forecast modeling as well as identifying opportunities to improve water use efficiency (White et al. 2003). Water end use information can increase our understanding

of water use behavior, inform future water projections, and aid in the design and assessment of water conservation efforts. For example, incentives to upgrade inefficient fixtures/appliances (Mayer et al. 2004; Suero et al. 2012) or awareness campaigns targeting specific end uses would benefit from this information (Abdallah and Rosenberg 2014; Willis et al. 2010) by quantifying behavioral (frequency, duration) and technological (flow rate, volume) parameters for individual water use appliances. These parameters can be used to calculate potential or actual benefit of conservation measures and to identify the most effective strategies. Additionally, high resolution water use data can enable verification and calculation of accurate price elasticity estimations (Marzano et al. 2018).

High resolution (sub-minute) data is required to record and quantify end uses that have short duration (Nguyen et al. 2015). Typically, end use information is derived from high temporal resolution water use data by using algorithms that differentiate between the characteristics (duration, average flow rate, mode flow rate) of water use events. Several algorithms for disaggregating and classifying end uses of water have been developed by private companies (Aquacraft 1996) and by researchers (Attallah et al. 2021a; Froehlich et al. 2009; Nguyen et al. 2018; Pastor-Jabaloyes et al. 2018). Despite the number of algorithms described in the literature, opportunities to replicate or build from these tools remain limited due to the unavailability of code and/or data. Di Mauro et al. (2020) found that, from 41 datasets collected for assessing end uses of water at residential properties, only 4 (Beal and Stewart 2011; Makonin 2016; Vitter and Webber 2018; Kofinas et al. 2018) had an open access policy. In limited instances, flow trace data (i.e., the raw, high resolution data collected) and event files (i.e., end use events and their attributes extracted

from raw data) from past studies are available for purchase (Aquacraft 2016), including the events table resulting from the one of the largest end uses of water study conducted to date (DeOreo et al. 2016).

While the lack of available datasets is limiting, so is the duration of many of the collected datasets. Several past residential water use studies (Beal and Stewart 2011; DeOreo et al. 2016; Mayer et al. 1999) collected data for a period of two weeks. This relatively short data collection window may not allow observation of temporal variations in indoor water use volumes, timing, and distribution across end uses. It is likely also insufficient to assess outdoor water use practices. Furthermore, previous data were collected in small samples across a limited number of cities. Given that there are differences in how people use water at the neighborhood, city, and country levels (e.g., Inman and Jeffrey 2006), exploring temporal changes in water use, along with expanding available datasets for urban water planning and management motivates the importance of local case studies. High temporal resolution water use data is not available for Utah, which has to date limited the analyses that can be conducted to explain the large per capita water use observed at the state level.

To build on the results of prior studies, this paper focused on the following research questions: a) How do the distribution of indoor water use, frequency of use of indoor water using fixtures, indoor water use timing, and outdoor practices vary for users at different water consumption levels?, b) What is the efficiency of water using fixtures among the sample of residential homes analyzed and how do these values compare with previous studies?, and c) How do estimates of volume, the distribution across end uses, and timing of indoor water use change as the data collection period is increased beyond

the two weeks observed in the past? We analyzed water use at different temporal aggregations (monthly, daily, hourly, weekdays versus weekend), its subdivision between indoor and outdoor use, and the distribution of end uses to address these questions. We also show how the high temporal resolution data collected as part of this project can help researchers answer other questions. The data used are openly and freely available for reuse, providing an opportunity to expand the analyses and extend the research presented in this manuscript. The analyses we conducted convey new and key information that can assist water utilities and decision makers in Utah, and potentially other areas with similar characteristics (climate, landscape sizes, household occupancy, level of water use), in understanding how water is being used.

4.2 Methods

4.2.1 Study area and data used

This study combined data from multiple sources (Table 4.1). The area of study comprised the cities of Logan and Providence in northern Utah, USA. Monthly water use data was provided by the municipalities, and we collected high temporal resolution (4 second) data for 31 single family residential (SFR) properties, 19 in Logan and 12 in Providence. Logan and Providence have about 7,500 and 2,100 SFR connections, respectively. Logan City reads meters once per month, and Providence City reads meters once per month during the months of April through September. We calculated volumes for October through March for Providence by dividing the total winter volume measured (calculated using the September and May meter readings) by 6, resulting in the constant values shown in Figure 4.1 (included here as it provides context for the rest of the methods we selected) during those months.

Meter readings provided by Logan and Providence cities were collected on different days of the month, depending on the utility's working schedule. Thus, the volume of water used within a given month must be estimated from two meter readings. We calculated standardized monthly water use, from the first to the last day of each month, as follows:

$$V_n = \frac{V_{MR1}}{D_{MR1}} * D_{n-MR1} + \frac{V_{MR2}}{D_{MR2}} * D_{n-MR2} \quad (1)$$

where, V_n is the volume of water used for a month n. V_{MR1} is the water volume from the first meter reading (MR1) that contains water use for month n. D_{MR1} is the number of days covered by MR1 (i.e., the number of days since the previous meter reading), and D_{n-MR1} is the number of days within month n to which MR1 applies. V_{MR2}, D_{MR2}, and D_{n-MR2} have the same information for the second meter reading (MR2) that contains water use for month n. User ranking, monthly variation, and annual averages used when selecting participants to enroll in the study were derived from these standardized monthly values.

4.2.2 User enrollment

Participants were enrolled in this study using multiple methodologies. First, four households were enrolled by word of mouth to deploy and test data collection hardware and software (Bastidas Pacheco et al. 2020, 2021b), and these users have the longest data records. Second, we invited users based on their annual average water use (computed from monthly records) in an attempt to create a sample with participants from different water consumption levels so that we would have representatives ranking in the lower, mid, and higher end of water consumption. Prospective participants were sent a letter in

the mail inviting them to participate in this study. Of 200 letters sent, 11 participants responded positively and enrolled. Given the low response rate to mailed letters, an additional 16 participants were recruited and enrolled through word of mouth and targeted invitations. Originally, we intended to enroll 50 participants, but due to public health conditions associated with the COVID-19 pandemic, additional participation was limited.

Participation in this study was voluntary. Residents agreed to participate but did not necessarily know when data was being collected. In all cases, we conducted multiple data collection periods for each household (referred to as a site in this study). During enrollment, information was collected on water using fixtures in each home, the age of appliances, common time of the day of irrigation, and typical timing for the use of clothes washers. High temporal resolution data was collected while a small set of short duration events were registered (toilet flushes, opened and closed showerheads, faucets, bathtub faucets). Study sample household characteristics (n=31) are reported in Table 4.2, including length of the data record, number of occupants, irrigable area, building area, irrigation mode, volumetric pulse resolution of the meter, and annual average water use (for the same periods shown in Figure 4.1). The information for each site was obtained through different sources: 1) the survey conducted during enrollment, 2) publicly available data from the county, 3) analysis of the monthly water use records provided by each city, and 4) geographic information systems (GIS) analysis of high resolution imagery for Utah available from the Utah Geospatial Resource Center (UGRC 2021).

4.2.3 Data collection and management

High temporal resolution water use data for all sites was collected using the CIWS-Datalogger (Bastidas Pacheco et al. 2020) or the CIWS-Node Datalogger (Attallah et al. 2021b), which were attached to the existing meters at each site. These external dataloggers measure the magnetic field around magnetically-driven residential water meters and count peaks in the magnetic field associated with movement of the measurement element within the meter. They register peaks as pulses that represent a fixed volume of water passing through the meter. The volumetric pulse resolution (L/pulse) used in this study was determined in the laboratory (Bastidas Pacheco et al. 2020) and used for all meters of the same size and brand found in the field deployments. Other studies have self-calibrated this parameter on each meter when volumes (from the meter and datalogger) do not match (DeOreo et al. 2016), resulting in accuracies that are not directly comparable with ours. While we did not calibrate pulse resolution in this study, instead choosing to use only data that passed the quality control (QC) procedure described below, our field data logs provide the volumes recorded by the meters' registers and the raw pulse data we collected so that calibration methods could be applied to this data, if warranted for other studies.

Data was collected over a period of three years before and during the COVID-19 pandemic. We recorded the number of residents in each home during enrollment but did not collect any information related to the participants' schedules or employment status, nor did we assess any changes in these parameters due to the COVID-19 pandemic. We did not anticipate the COVID-19 pandemic; therefore, collecting the data that would allow us to fully assess the impact of COVID-19 on participants' schedules was not part

of the study design. We were limited in our ability to modify the Institutional Review Board (IRB) Protocol governing this study and were also constrained by a complete pause of all human subjects research implemented by our institution early in the pandemic. While we did not assess the impact of the COVID-19 pandemic on participants of this study, recent studies evaluating the impact of COVID-19 on water demand suggest that residential water use increased and some non-residential use (e.g., bars, restaurants, hotels, schools) decreased when stay-home orders were issued (Cooley et al. 2020; Meener et al. 2021; Cahill et al. 2021; Lüdtke et al. 2021). It is likely that these effects are evident in some of the data we collected but were not specifically analyzed.

We collected at least two weeks of data during months when irrigation was expected to occur (referred to as summer months, including May through October) (Figure 4.1), and two weeks during the rest of the months (referred to as winter months) at each site. In December, January, and February, access to meter pits was restricted by the municipalities due to cold temperatures, resulting is shorter records for those months. Log files included within the LogFiles folder in the HydroShare data repository (Bastidas Pacheco et al. 2021a) contain information about the data collection periods at each site, including the exact start and end time of each period, the volume registered by the meter's register and by our datalogger for each period, the percent error in volume for each period, the number of expected data values (computed using the start and end time of each data collection period, assuming one value was collected every 4 seconds), the number of recorded data values, the percent error in number of values logged, and an

indicator of whether outdoor water use was expected or not. The HydroShare resource also contains the high temporal resolution data for each site.

The data were managed using cyberinfrastructure described in Bastidas Pacheco et al. (2021b). The data management process involved collection and processing of raw 4 second resolution data, QC to ensure validity of the data, and storage in a centralized database for analysis. QC was initially conducted by comparing the volume recorded by the meter's register with the volume recorded by the installed datalogger for each data collection period. The volume recorded by the meter's register was calculated by subtracting manual meter readings made at the beginning and end of each data collection period. The volume of water registered by the dataloggers was calculated by multiplying the number of pulses recorded during each data collection period by the volumetric pulse resolution.

The initial condition of the QC process was based on the percent error of the volume read by the datalogger when compared to the volume calculated from manual readings of the meter's register. If the percent error was less than 5%, associated values were finalized without further review. In the opposite case, an additional review procedure was developed to determine whether portions of the data could be included in the analysis. This additional review was conducted on a daily basis. By visually examining the characteristics of the data (hourly and daily volumes and flow rates) for the period in question and comparing them with similar data (i.e., for the same site that were already validated), we were able to accept or reject portions of the data collected. The percent error of the number of data values collected when compared to the number of values expected was also considered in this procedure. In some cases, data were lost due

to an error associated with writing data to the datalogger's SD card (Bastidas Pacheco et al. 2020) and, consequently, the percent error in volume was not within the 5% threshold (e.g., only 5 days of data were recorded within a 10 day deployment resulting in a percent error of -50%, indicating that data was possibly good but incomplete). Analysts considered all the elements mentioned to accept or reject the raw data collected. Only data that passed these QC checks was used in this paper

4.2.4 End use classification

In many past studies that analyze the end uses of water, a single device measures total water use for a site, and the data are later disaggregated and classified (Al-Kofahi et al. 2012; Beal and Stewart 2011; DeOreo et al. 2011, 2016; Mayer et al. 1999; Meyer et al. 2020; Otaki et al. 2011; Roberts 2005). Less commonly, water use for each individual fixture is measured (Kofinas et al. 2018; Mauro et al. 2019). We adopted the first approach and used a single device (Attallah et al. 2021b; Bastidas Pacheco et al. 2020) to measure total water use, and data was disaggregated using an open source algorithm described in Attallah et al. (2021a). In summary, the disaggregation and classification process works in the following way: 1) an algorithm filters the raw data using a low-pass filter, facilitating the identification of single and overlapping events, 2) overlapping events are separated into single events using an iterative splitting process, 3) several features (e.g., average flow rate, mode flow rate, duration) are calculated for each event, 4) the events are classified using a combination of clustering to identify atypical or outlier events that are later labelled as "unknown" and a semi-supervised, machine learning methodology to assign labels to the remaining events (Attallah et al. 2021a). The machine learning model uses a Random Forest classifier (Liaw and Wiener 2002) trained using a

set of events manually labeled by a resident at one of our data collection sites to classify new events for individual residential homes.

Using the disaggregation and classification algorithm, water use was classified among the following end uses: irrigation, faucet, shower, toilet, clothes washer, bathtub, and unknown by using the most important features of each event (mode, average, root mean square and peak flow rate; duration; and volume) as identified by Attallah et al. (2021a). Dishwasher events were lumped with faucet events, as the features of these events were indistinguishable in our sample. A pool was only present in one of the participant sites, and pool-related events were likely labeled as irrigation by our algorithm. Additional uses, such as hose events, leaks, or those not described here were placed in the category that their features more closely resembled or were labeled unclassified. Additionally, we manually labeled all events with a duration of 4 seconds (the temporal resolution of the data) and volume equal to the meter pulse resolution (i.e., single pulse events) as unclassified (Attallah et al. 2021a). Indoor water use estimates were computed after filtering out irrigation events. Outdoor water use includes only those events labeled as irrigation. The accuracy of the method was characterized using data for a single site, and, under those conditions, the overall accuracy of the classification method was around 98% (Attallah et al. 2021a). This accuracy is expected to be maximum since the training and testing dataset for the machine learning algorithm contain events for the same site.

When using the algorithm to label events for new sites (those for which no user manually labeled events exist) it is expected that the accuracy will decrease given that the features of the unlabeled events may be different than the ones included in the training

dataset. We used a self-learning approach (Attallah et al. 2021a) to classify data from sites at which no manually labeled events were available (30 of our 31 participants). Using this approach, events were initially classified using the Random Forest algorithm trained using the manually labeled events. Events with a similarity score larger than 90% were then added to the training dataset (Attallah et al. 2021a). This process was repeated iteratively until there were no events with a similarity score larger than 90%. The revised Random Forest model for a site based on the enhanced training dataset was then used to classify all of the events for that site.

Without manually labeled events for each site, it is not possible to evaluate the accuracy of the classifications. In consequence, we applied a manual verification procedure consisting of examining the characteristics and raw data for a number of events of each end use type at each site. Events within each type were sorted according to their features in a step-by-step procedure (e.g., first sorting shower events by descending flow rate, then ascending flow rate) to observe differences between events at the endpoints and events in the middle of the distribution of each end use. This verification method assumes that events are generally labeled correctly, and is based on our observation that labeling errors are more likely to occur at the endpoints of the feature distributions of each end use. We are confident that the majority of events in each category are labeled correctly, yet at the endpoints of each distribution, where overlapping (similar features) exist (e.g., the lowest flow rate shower can overlap the highest flow rate faucets), there is uncertainty in the labelling process.

The number of events examined varied, depending on how similar events at the endpoints of the distribution for each end use were to the rest. Generally, we examined

between 10 to 30 events per end use and sort direction. The raw pulse data for a number of events (in the same 10 to 30 events range) was examined visually to verify similarities in events with the same label. This verification was performed for all event types at each site. The information and small set of labeled events for each site registered during enrollment allowed us to verify event labels (e.g., volume and flow rates observed in toilets and showers) and to find errors (e.g., the algorithm labeling bathtub events that were similar to clothes washer events in homes where bathtubs were not present or used). These events were not included as training data as they do not represent real events (with the exception of toilets), yet they provide an idea of the flow rate ranges for a specific site. Additionally, Attallah et al.'s (2021a) method seeks to classify end uses of water without the need for labelled events at each property. Thus, the manual evaluations we did were aimed at ensuring the quality of our analyses.

In some cases, events of different types can have similar features (e.g., short duration showers with low flow rates can appear similar to faucet events). When two events of different types have similar characteristics, it is not possible to differentiate them using existing methods, and they are assigned the same label. Metering of individual end uses can produce further data about the frequency at which this occurs and can provide further details about how this affects the accuracy of single point measure and disaggregation methods. Without meter data for individual end uses, it is not possible to assess how often or where these events occur. However, the tradeoff is that metering of individual end uses is expensive, invasive, and largely impractical at any scale.

As a last step in the verification procedure, we corrected the labels for some events using the following criteria: 1) misclassified events were re-labeled according to

where they were most likely to belong, based on the analyst's decision, considering all the elements described above; 2) when events were routinely misclassified by the algorithm, we filtered events of similar characteristics and applied the same corrected label to all. Without considering unclassified events, on average, changes were made to 6.3% of the labels assigned by the algorithm at each site. At sites 2, 4, 14, and 31, 15% to 18% of the algorithm assigned labels were reclassified. At these sites, the algorithm systematically made errors resulting from differences in the characteristics of the events occurring at those sites versus the manually labeled events used for training the algorithm.

4.2.5 Estimating outdoor irrigation efficiency

The time period for which data was collected at a site will influence the amount of outdoor water use captured (Figure 4.1). In order to obtain an estimate of outdoor irrigation efficiency comparable across different time periods of the year, the Landscape Irrigation Ratio (LIR) (Glenn et al. 2015) was calculated at weekly intervals for each site. The LIR is defined as the ratio between landscape water use and landscape water needs (Equation 2).

$$\text{LIR} = \frac{\text{Landscape Water Use}}{\text{Landscape Water Need}} \quad (2)$$

Landscape water needs were determined for each site based on a water budget (Equation 3) similar to Glenn et al. (2015):

$$\text{Landscape Water Need} = (K_c * ETo_i - P_i) \quad (3)$$

where K_c is the crop coefficient and ETo_i and P_i are the reference evapotranspiration and the precipitation for a given week (i), respectively in millimeters. Daily rainfall data and estimates of evapotranspiration from the Utah State University

(USU) Environmental Observatory weather station were used to estimate the landscape water need for all properties in Logan, and data from the Evans Farm weather station was used for properties in Providence (Bastidas Pacheco and Horsburgh 2021b). The crop coefficient represents the ratio between the reference evapotranspiration and the actual crop evapotranspiration (Doorenbos and Pruitt 1977). To determine this value, we assumed a uniform turfgrass surface for all sites and used K_c=0.8, similar to Endter-Wada et al. (2008). Typically, residential landscapes are composed of turfgrass and trees immersed in a turfgrass landscape (Kjelgren et al. 2000). Yet, there is limited information about crop coefficients for turfgrass (Romero and Dukes 2015), or landscapes with multiple plant species (White et al. 2004). The actual K_c, for each site is likely to be lower than the 0.8 used.

The landscape water use for a given week (i) in millimeters was computed for each site using Equation 4:

$$\text{Weekly Landscape Water Use}_i = 1{,}000 * \frac{\text{Weekly Outdoor Volume}_i}{\text{Landscape Area}} \quad (4)$$

where the $\text{Weekly Outdoor Volume}_i$ is the total volume of water used outdoors for week i (in cubic meters), and the Landscape Area is the area being irrigated (in square meters) at each site. Landscape areas were identified and manually digitized from high resolution aerial imagery for each site, and the areas were calculated using GIS. Using the LIR helps classify outdoor water use (Table 4.3).

Of the 31 sites enrolled in this study, participants at two sites (sites 21 and 22) moved into newly built homes between the time we collected winter and summer data. These sites had not yet developed their landscape when summer data was collected and

so outdoor water use was not assessed for those sites. All outdoor analyses presented in this paper are for the remaining 29 sites.

4.2.6 Indoor water use efficiency

To assess the efficiency of indoor water using fixtures, we compared the characteristics (showerhead and faucet flow rates, toilet volume used per flush) of existing fixtures at the 31 sites enrolled with the current federal standard, defined in the U.S. Energy Policy Act of 1992 (DOE 1992), and the U.S. Environmental Protection Agency's (EPA) WaterSense 'efficient' fixtures (EPA 2021). The 1992 Energy Policy Act (DOE 1992) set national water efficiency standards for toilets, faucets, and showerheads and has been in effect since 1994 in the U.S. The EPA WaterSense program labels water products using higher efficiency standards than the 1992 Energy Policy Act, achieving 20% more efficiency than average products in the same category (EPA 2021).

We divided faucet, toilet, and shower events into three categories: efficient (flowrate or volume per flush less than or equal to WaterSense specifications), compliant (flowrate or volume per flush larger than WaterSense specifications but less than or equal to the Federal Standard), and inefficient (flowrate or volume per flush larger than the Federal Standard). Events with small frequencies (< 5%) were not accounted for in the final assessment to reduce the impact of double toilet flushes, errors in the classification, or unintended use. We did not assess the efficiency of clothes washer events as this requires information about load sizes. We also did not assess bathtub events because there is no defined criteria about what an efficient bathtub event is.

4.2.7 Indoor water use observed for longer data collection periods

To explore how volumes, distribution across end uses, and timing of indoor water

use change as the length of the data collection period increases – our third research question – we examined high temporal resolution data for sites with records longer than 4 weeks (18 sites, including five sites with record lengths varying between 11 and 19 weeks). For this analysis, we used data collected during summer and winter months and removed irrigation events to focus on indoor water use. We quantified differences in the total volume used for indoor water use, the distribution across end uses, and hourly aggregated water use estimations at the weekly level for each site. Additionally, we quantified differences in the mean volumes (for faucet, shower, and bathtub events) and frequency of end use (for all end uses) events in winter versus summer months using Student's t-test (Student 1908) for sites that had at least 4 weeks of data during summer and winter months (10 sites).

4.3 Results and Discussion

SFR water use varies throughout the year in Logan and Providence, peaking in July with average monthly values, across all connections, per household, close to 125,000 L in Logan and 220,000 L in Providence (Figure 4.1). During winter months, SFR water use remained relatively constant in Logan, with per household monthly averages just below 20,000 L and below 30,000 L in Providence. Sociodemographic variables like differences in household and landscape sizes can likely account for the differences observed in water use between the two cities during winter months. However, we did not collect data at the city level, and the monthly data provided by the cities did not contain information that would allow us to further assess these differences. Outdoor water use drives the increase in residential water use observed during summer months, constituting the largest component of residential water use. Total annual water use did not vary

significantly from one year to the next during the period of data available for each city. Winter water use for Logan City (Figure 4.1) shows variations that are likely due to differences in indoor water use; however, when compared with the magnitude of the annual variation, changes during winter months appear minimal.

In order to place our sample of households in the context of single family residential water use in their city as well as other residential water use studies, we report brief general statistics about ranking and water use. Participant sites ranked between the 4^{th} and the 95^{th} percentile of annual SFR water use in each city (computed from monthly meter records). Appendix A provides additional information about participants' ranking and water use at each site. The average per capita daily water use among participants in this study computed from monthly water meter data was 695 Liters per capita per day (Lpcd), and the same figure computed from the high temporal resolution data we collected was 754 Lpcd. 73% of our high temporal resolution data was collected in summer months, which explains the difference between the estimations from monthly records and the high temporal resolution water use data collected. One recent estimate places per capita daily average residential water consumption in Utah at approximately 640 Lpcd (Dieter et al. 2018). However, this value was calculated by compiling data from different agencies, using coefficients in areas of the state where supply is not measured, and using population estimates that can impact the accuracy of this estimation (Milligan 2018) and thus hides hides a lot of variability within Utah. The State of Utah Division of Water Resources (DWR) estimated that in 2015, residential water use in Cache County, where Logan and Providence are located, was 784 Lpcd and that there is differences in water use across counties in the state (Utah DWR 2020).

It is well known that per capita averages, while useful for estimating total water demand at aggregated scales, provide little information about water use patterns or behavior within individual households – in particular because outdoor water use is not dependent on the number of occupants of a house. Households with a small number of occupants and a large landscape will have a larger per capita consumption. Given the differences in patterns of outdoor and indoor water use and conservation approaches, we analyzed indoor and outdoor water use separately. The following three subsections analyze indoor water use (frequency and volume of end uses, and timing) addressing the first research question.

4.3.1 Distribution of Indoor Water Use and Frequency of Use for Indoor Water Using Fixtures

The average daily per capita indoor water use among participants in this study was 174 Lpcd. Shower (31.2%), toilets (25.6%), and faucet (18.6%) events account for three quarters of the volume used indoors. Appendix A provides detailed information about indoor water use, the distribution of indoor water use across categories, specific features of each category, the frequency of use of fixtures along with comparison of these values with those obtained from past residential water use studies.

We did not observe a clear trend in the number of events (Figure 4.2b), or the average volume per occurrence (Figure 4.2a) corresponding to daily per capita average use (Figure 4.2c). However, in general, sites with higher indoor per capita consumption also had a larger number of toilet events per capita per day. Figure 4.2 indicates that indoor water use is a result of different combinations of behavioral (frequency, duration) and technological (flow rate, volume) parameters that impact the frequency and volume

used per event. A larger average per capita number of events per day could be the result of changes in occupancy for which we were unable to account (i.e., the number of residents changed during the course of our study), but more likely reflects differences in personal routines among the participants.

Site 9 (with the second largest per capita water use) had a larger indoor water consumption, in part due to a leak, indicating that not all indoor water use patterns observed were the result of intentional consumption. The leak was associated with more than 20,000 short duration and low flow rate events between July 26 and August 3, 2020. Because events associated with the leak were classified as faucet events, the average number of faucet events per capita at site 9 was 164.

We used the per capita daily average indoor water consumption to rank sites as low (< 33rd percentile), medium (33rd - 66th percentile), or high (>66th percentile) water users, depending on their percentile ranking of per capita daily average indoor water consumption. Figure 4.3 shows the same information presented in Figure 4.2 with values averaged for the three groups rather than separated for each individual site. There was less than 13% difference in the average volume used for all events across the three groups. A Kruskal-Wallis test (Kruskal and Wallis, 1952) with $p > 0.05$ showed that these differences were not statistically significant. Figure 4.3b shows that high consumption sites have a larger number of events per capita per day across all end uses. A Kruskal-Wallis test (Kruskal and Wallis, 1952) with $p < 0.05$ showed that the differences in frequency across all end uses were significant. These results indicate that the frequency of events, which is an indicator of behavior, has the largest influence on per capita daily water consumption at the group level.

The distribution among end uses remained relatively constant across the three consumption levels. As the largest indoor water use category, shower events accounted for 30.3%, 30.8%, or 31.6% of the total indoor per capita daily water use (for low, medium, and high consumption sites, respectively). Toilet events accounted for 21.8%, 20.8%, or 23.8% of the total indoor volume across levels (low, medium, and high, respectively). Faucets accounted for 17% of the total indoor water use across all categories.

4.3.2 Indoor water use timing

Hourly and daily aggregated values were considered to assess indoor water use timing and its variation across users of different consumption levels. Indoor water use timing is behavioral and is determined by personal preferences and schedules. We observed variation in the hourly distribution among participants' water use. Broadly, some sites (19.4%) had one period of higher consumption during the day, multiple periods of higher consumption (32.3% of the sites), or no obvious peaks with relatively similar water use throughout most of the day (48.4% of the sites). Figure 4.4 shows an example of these patterns with plotted values representing the percentage of total indoor water use that occurred during each hour of the day, e.g., the value plotted at 2:00 AM at any site was computed by totaling indoor water use between 2:00 AM and 3:00 AM, multiplying it by 100, and dividing this value by the total indoor water use for the same site. Site 14 had a single period of higher consumption occurring in the morning. During the peak hour (6:00 AM to 7:00 AM), residents of site 14 used more than 20% of their total daily volume. There are two periods of higher consumption at site 6, one between 4:00 AM to noon, and another one peaking between 7:00 PM and 8:00 PM; however, the

maximum value observed was slightly over 10%. Finally, Site 10 showed a relatively consistent pattern of consumption throughout the day with values for most hours varying between 4% and 6% between 5:00 AM and 9:00 PM. The timing pattern was independent of the consumption level. Sites within the low, medium, and high consumption categories followed all three patterns.

Participants consumed, on average, 21% more water during weekend days (Saturday and Sunday) than during weekdays (Monday to Friday). High consumption sites used 15.7% more water on weekend days compared with weekdays, while medium and low users used 19.6% and 28.9% more water, respectively. A smaller increase in the weekend versus weekday average per capita daily volume likely indicates longer presence at home during weekdays. This can partially explain the results observed in Figure 4.3. In general, the differences in the observed hourly and daily patterns are likely dictated to a large degree by the heterogeneity in the schedules of the occupants

4.3.3 Outdoor Water Use

We collected a combined 278 weeks of data between May and October at the 29 sites where outdoor water use was analyzed, recording 4,533,939 L of water use during these months. Approximately 83% of this volume was used for outdoor irrigation. Outdoor water use is largely driven by personal preferences, but in some instances can be required by homeowner associations. The volume used may also be impacted by the type of system used for irrigation (i.e., a hose, sprinkler system, automated timer, smart weather controller, soil sensors). While the level of technology used for irrigation is a personal preference, each type of system has a potential technological impact related to device performance. In our sample, eight sites irrigated using a hose (3, 6, 7, 8, 10, 16,

17, and 18 - some with an automated timer, others manually). All of these sites ranked below the 40th percentile for annual water use in Logan City (see Appendix A). The rest of the participants used a sprinkler system with automated controllers, and 88% of those sites ranked above the 40th percentile. These results are similart to those of past studies, which have found a strong correlation between the presence of automated sprinkler systems and higher water use (DeOreo et al. 2016; Mayer et al. 1999; Endter-Wada et al. 2008).

Figure 4.5b shows the average weekly outdoor volume for each site. Sites 27 and 5 had the largest outdoor water use, consuming, on average, more than 80,000 L per week, and had the largest and the third largest landscape areas (3,843 m^2 and 3,118 m^2, respectively). During six weeks (five in 2019 and one in 2020) the landscape irrigation needs were zero (rainfall supplied all the water landscapes needed), and any outdoor water use that occurred was unnecessary. Given that the LIR has an undefined value during these weeks (Equation 2), these six weeks were not included in Figure 4.5 but are addressed in the following paragraph. All hose irrigators used less than 8 m^3 of water per week, while 80% of sites with an automated sprinkler system used more than this value (Figure 4.5). The landscape areas for hose irrigators were not all smaller than sprinkler irrigated sites (ranking 1st, 2nd, 4th, 9th, 17th, 21st, 22nd, and 24th among the 28 sites presented in Figure 4.5). Using the LIR values determined for each week of irrigation (Figure 4.5c), there was one week where irrigation was excessive at sites 2, 9, 11 and 14 during week 36 of 2019, and at site 14 during week 38 of 2019. During week 36 of 2019 a rainfall event that supplied 97% of the landscape water needs was registered by the USU Environmental Observatory station, making outdoor water use inefficient during

that time period at those sites. In summary, outdoor water use was either efficient (62%) or acceptable (27%) during most the weeks collected, and excessive (5%) or inefficient (6%) during the rest.

During the six weeks where landscape water needs were zero (the LIR was undefined), we collected 33 full weeks of data across 20 sites. Figure 4.6 shows the number of full weeks of data collected at each of these 20 sites and the volume (average when more than one week was available) used during weeks where landscape irrigation needs were zero. Most sites (80%) reduced their outdoor water use between 11% to 90% in response to precipitation, when compared with the rest of the weeks. Nevertheless, precipitation can occur at the end of the week after all outdoor water has been applied, and the regular weekly intervals we used for our analysis did not attempt to account for this. Furthermore, homeowners would need additional information (landscape water needs, rainfall data, usage from their irrigation system) to accurately respond to precipitation events. Even with the reduction in outdoor water use observed, the total volume used for outdoor irrigation during weeks when the landscape water needs were met by precipitation (366 m^3) represents a large water conservation potential among participant sites.

Separating indoor from outdoor water use was more accurate when automated sprinkler systems were present, as the flow rates for automated irrigation events can be as much as twice the values observed for indoor events. Additionally, irrigation events produced by automated irrigation controllers have similar timing, flow rate, and duration. The flow rate of irrigation events was the highest among all end uses. At five sites (2, 5, 9, 19, 27), flow rates of irrigation events exceeded 70 Lpm. Irrigation events also had the

longest duration among all end uses, with an average of 42.1 minutes across all sites. Participants with automated sprinkler systems irrigated during early morning or late evening, which is within the recommended irrigation timing to reduce losses from evaporation, with the exception of sites 25 and 26 at which a few irrigation events were detected close to noon.

Individual sites were classified as low, medium, or high according to their monthly outdoor water use ranking, dividing at the 33rd and 66th percentile (computed for all SFR users by city, using the entire record of monthly data available, shown in Table 4.1). Monthly outdoor water use was computed as the difference between the average monthly water use during months when irrigation occurs (May through October, Figure 4.1) and the average monthly water use during months where irrigation is not expected (November through March). Using this procedure, 8 sites were ranked as low, 11 sites were ranked as medium, and the remaining 10 sites were ranked as high (Figure 4.7). Figure 4.7a shows the average monthly outdoor water use per landscape area, in millimeters. We collected two full weeks of data at site 13 during summer months, shown in Figure 4.5 and Figure 4.6, where outdoor water use was inefficient and unnecessary. Figure 4.7b shows the average outdoor monthly water use. Hose irrigators generally ranked lower than those who used a sprinkler system and applied less water per unit area, which is similar to past studies results (DeOreo et al. 2016; Mayer et al. 1999; Endter-Wada et al. 2008). Broadly, sites classified as medium and high applied water at similar rates, per unit of area. This indicates that the differences in outdoor water use observed were, in most cases, the result of the irrigation method used or the landscape area irrigated.

4.3.4 Efficiency of water using fixtures

To analyze the efficiency of water using fixtures among participant sites – our second research question – we examined the performance of showers, toilets, and faucets at each of the participant sites. The analyses presented in this paper focused primarily on the technological performance of fixtures (e.g., flow rates of showers and toilets, and the volume per flush used by toilets) rather than on behavioral aspects (e.g., frequency or duration of events). As an exception, we analyzed shower durations to highlight potential opportunities for conservation related to behavior. Most of our participant sites (28) had more than one bathroom. The efficiency analysis was conducted on events and not on average characteristics to observe differences in performance of different fixtures at the household level.

The federal standard for showerhead flow in the U.S. is 9.5 L/min (DOE 1992), while EPA WaterSense labeled showerheads use less than 7.6 L/min (EPA 2021). In terms of flow rate, we found inefficient shower events at 14 sites, compliant shower events at 29 sites, and efficient shower events at all 31 sites. At two sites, we observed only efficient shower events, indicating that all showerheads at those sites operate at or below the WaterSense standard. The remaining 29 sites had a mix of shower events across two or all three of the efficiency categories defined.

Shower durations are related to social norms, and, because of this, there is no consistent standard or guidance as to what shower duration is considered efficient. In these data, the distribution of shower durations was as follows: 25% of showers lasted less than 3.2 minutes, 25% lasted between 3.2 minutes and 5.87 minutes, 25% lasted between 5.87 minutes and 10 minutes, and the top 25% were longer than 10 minutes. The

average duration across all sites was 7.5 minutes, which is similar to the 7.8 minutes found in the 2016 Residential End Uses of Water Study (2016 REUWS) (DeOreo et al. 2016).

The federal standard for residential toilets in the U.S. is 6.1 L/flush (1.6 gallons per flush) (DOE 1992) while EPA WaterSense labeled toilets are designed to use 4.8 L/flush (1.28 gallons) or less (EPA 2021). The EPA allows some flexibility (±0.38 L = 0.1 gallon) in the values used for certifying toilets (EPA 2014), and we included this 0.38 L in the threshold definitions. We found inefficient toilet events at 30 of the participant sites, compliant toilet events at 21 sites, and efficient toilet events at only 7 sites. One site had only compliant toilet events, one site had compliant and efficient toilet events, and the remaining 29 sites had a mix of toilet events that included inefficient toilet events. The 2016 REUWS classified toilet events as efficient if they used less than 7.6 L. Using this higher threshold, they found that approximately 30% of homes had only efficient toilets, 28% had only inefficient toilets, and the rest had a combination of efficient and inefficient toilets.

The U.S. federal standard bathroom and kitchen faucet flow rate is 8.3 L/min (DOE 1992), and WaterSense labelled bathroom faucets use a maximum of 5.7 L/min (the EPA does not label kitchen faucets) (EPA 2020). The disaggregation and classification algorithm we used was unable to separate kitchen from bathroom faucets, and, while it is possible that higher flow rate faucet events are occurring in the kitchen, this cannot be guaranteed. 90% of the faucet events identified across all sites lasted less than 48 seconds and had a flow rate less than 5 L/min. Only 0.14% (1,136 occurrences) of faucet events had a flow rate larger than 8.3 L/min, indicating that faucet events

exceeding the federal standard for maximum flow rate were rare. Faucets were the most efficient category among those analyzed for efficiency. Faucets are likely replaced with a higher frequency than other water using appliances in a home, and the growing presence of water efficient faucets may explain why this category exhibits higher efficiency. The 2016 REUWS found similar results in terms of faucet event flow rates, with 99% of faucet events in that study having a flow rate less than 8.7 L/min.

4.3.5 Indoor water use observed for longer data collection periods

Indoor water use was relatively constant across weeks at some sites (e.g., sites 7, 12, 22), whereas differences in week-to-week volumes were observed at others (e.g., sites 2, 9, 18, 19) (Figure 4.8). In some cases, these variations occurred within subsequent weeks, as can be observed by analyzing the separation of points of similar color across sites in Figure 4.8. Some of these variations may be the result of changes in occupancy. For example, there are weeks with minimal water use at sites 5, 14, and 15, which indicates that occupants were likely not at home during these weeks. Additionally, changes in weekly schedules and personal preferences may have affected the amount of water used.

The daily timing of water use also varied between weeks, Figure 4.9 shows the total hourly water use week by week for site 19, i.e., the total water used within each hour summed across all the days of that week. The time of occurrence and the magnitude of peaks in water use varied from week to week, and this lack of pattern in hourly and weekly water use data was observed at most of the study sites. There are weeks at this site that follow each one of the broad patterns presented in Figure 4.4. Figure 4.10 shows weekly variations in indoor water use across end uses for the same site and the same

weeks. The percent of indoor volume used for shower events varied the most from week to week, moving from over 40% to 16% of total water use and ranking as the largest end use in some weeks while ranking 4th in others. The percent of indoor water use dedicated to bathtub events also exhibited changes, ranging from 6% to 20% and ranking from 2nd highest to 5th highest. The percentage of indoor water use dedicated to toilets varied between 25% and 36%. While we have generally reported use in terms of percentage of volume, the frequency with which end use events occurred also varied depending on the week.

We compared the mean frequency and mean volume of events between winter and summer months to determine whether there were seasonal differences in these values and whether differences were consistent. We compared 8 (or 6, as bathtub events were not present in four of these sites) parameters at each site. Of the 72 parameters analyzed across these 10 sites, there were significant differences (p-value < 0.05) in the mean of 19 (26% of the cases) of them, according to the t-test results. In 11 cases, we observed changes in the mean frequency, and in 8 cases we observed differences in the mean volume. The direction of these differences, by event type is reported in Table 4.4 (for frequencies) and Table 4.5 (for volumes). The inconsistency in the differences between frequency and volume of end uses events suggests that the observed differences are not generalizable. In some cases, the number of events or volume were larger in the summer, and the opposite occurs at other sites. Additionally, with the exception of faucets, no significant changes were observed at the majority of the sites analyzed.

As the number of weeks of data increased for a site, we observed only small variations in the average, mode, and peak flow rates for showerheads and faucets, and in

the volume per flush used in toilets. Thus, it appears that the technological performance of indoor fixtures can be accurately assessed with short data collection periods unless a fixture is replaced. However, capturing behavioral changes in indoor water use volumes and timing, along with developing a comprehensive representation of the distribution of indoor water use across end uses that were evident in our data requires longer data collection periods or a different study design. This lack of consistency in indoor water use patterns cannot be characterized using coarser resolution (e.g., monthly) data or when analyzing indoor and outdoor water use together. Other studies have pointed to similar results. For example, Rathnayaka et al. (2015) found differences in shower durations and frequency in summer versus winter months in 117 houses across two municipalities in Australia. Suero et al. (2012), who analyzed two weeks of data pre and four weeks of data post retrofitting with efficient appliances in 96 homes in the U.S., found differences in the frequency of use of toilets and clothes washers between their pre and post retrofit datasets. The seasonal differences we observed and those observed by prior studies indicate there are seasonal and shorter-term changes in the frequency, timing, and distribution of end uses. Longer, and continuous, periods of data collection are required to characterize these types of temporal variations (Figure 4.9), including changes in the distribution of water use across end uses (Figure 4.10) and the seasonal component of indoor water use (Rathnayaka et al. 2015). Additionally, we observed that indoor water use varies differently across sites (Figure 4.8), suggesting the record length needed to characterize indoor water use variability may be different across sites.

Collecting indoor water use data for short periods of time can generate parameters (volume, timing, and distribution across end uses) that may not be representative of water

consumption at a site given that water use depends on behavioral factors in addition to fixture performance. End use level data provide a basis for evaluating and designing water demand strategies (Beal and Stewart 2014), demand and infrastructure modeling (Blokker et al. 2010), and general planning (Willis et al. 2013). Using water use estimations resulting from data that do not capture a representative sample of water use may impact the accuracy of such applications and lead to the implementation of ineffective water management strategies, under or over dimensioning of infrastructure, and other issues. Yet, defining a fixed record length that secures a complete characterization of indoor water use across multiple residential properties is infeasible using currently available data, most of which are short duration. Further research is needed to define the effect of data record length on indoor water use estimations across different sites

4.4 Conclusions

The results presented here were derived from analysis of monthly water use data provided by two municipalities in northern Utah, USA and from 4-second temporal resolution data collected by the authors over a time span of three years at 31 homes in those two cities. Indoors, we found that total water use volume and the distribution across end uses varied across hours, days, and weeks. Our analysis of water usage across high, medium, and low water users revealed behavioral differences. While the distribution of indoor water use across end uses was similar for sites at all levels of consumption, sites with higher usage had a higher number of events per capita. Additional data, which could be collected via an additional survey, is needed to characterize the determinants of this behavior. The average daily per capita indoor water use varied considerably among

participating homes due to a combination of fixture characteristics, personal preferences, and differences in schedules. Showers and toilets were the largest indoor water using categories. All sites used more water during weekends compared to weekdays; however, sites at lower consumption levels had a higher percentage increase from weekday to weekend.

The data from this study demonstrate opportunity to improve toilet water use efficiency by either adjusting existing toilets or replacing them with more efficient toilets at 29 (93.5%) of the participant sites. This could be done through educational campaigns targeted at homeowners to explain how to adjust existing toilets or through rebate programs that encourage homeowners to replace existing toilets with efficient ones. Toilet age, installation characteristics, and valve status affect the volume used per flush. Even toilets manufactured under Federal standard specifications can perform outside their target range.

Approximately half of the participant sites had efficient showerheads when compared with high efficiency standards such as the EPA WaterSense (EPA 2020), and only one site had showerheads operating at flow rates above the federal standard (DOE 1992). Thus, the largest opportunity to reduce shower water use would be through promoting shorter duration showers given that 25% of all shower events lasted longer than 10 minutes. This may be difficult for a number of reasons, including identifying those with the highest opportunity to conserve and presenting them with effective information that may encourage conservation. There is also a shortage of longitudinal studies in the literature to assess the effectiveness and long-term effects of these types of campaigns. Bathtub events used significantly more water than showers, but were also less

frequent and not found or not used at 35% of the sites. Faucets were the most efficient indoor water use category.

In summer months, outdoor water use was the largest component of residential water use. The daily average per capita water use reported in this study (754 L) is affected by those sites using large volumes of water for landscape irrigation (6 sites used more than 50 m^3 in a week, on average, for landscape irrigation during our data collection campaign). Generally, outdoor water use volume per unit of irrigated area was similar across users at all consumption levels. Users that irrigated with an automated sprinkler system used larger volumes (in total and per unit of area) of water than those who irrigated with a hose. The total volume of outdoor water used at a site was mainly influenced by the irrigated area and the method used for irrigation.

Outdoor water use was "efficient" or "acceptable" according to the LIR categories during 89% of user-weeks, despite the large volumes used for outdoor irrigation. This indicates that most users are not significantly overwatering their landscapes according to the LIR. While we do not want to discount informational campaigns targeting at ensuring that people are not overwatering their landscapes, a more significant water savings may be achieved through campaigns aimed at reducing landscape water need by changing landscape size or composition. Furthermore, we found significant conservation potential (366 m^3 in a week across 20 sites) that would be realized if users did not irrigate when rainfall sufficient to meet landscape needs has occurred.

The total volume of water used, the distribution of use across end uses, and the timing of indoor water use varied from week to week such that data collection periods longer than those used in past studies and likely even those used in this study are needed

to fully characterize these changes. The temporal patterns of water use (peaks, timing of peaks) varied between weeks at all sites independently of their water consumption level. Daily indoor water use timing patterns can be difficult to determine, as they depend on personal and often variable schedules as was evident in our data. This type of variability is also not represented well in existing water demand modeling approaches, and doing so is an opportunity to improve these models.

Some of the general results of this study and the analyses included in Appendix A are similar to those of past studies, indicating that some aspects of residential water use are generalizable. However, our analysis of changes in the distribution of indoor water use across end uses, differences in weekly total use, differences in timing, and differences in outdoor water use across longer data collection periods convey new and key information that can assist water utilities and decision makers in Utah, and potentially other areas with similar characteristics (climate, landscape sizes, household occupancy, level of water use), in better understanding how water is being used. Participants in this study received detailed water use feedback comparing their annual usage with the rest of the SFR clients in their city; the performance of individual fixtures at their home; shower durations; outdoor water use; and opportunities for water conservation. Prior studies have shown that this type of specific information can motivate conservation behavior. Water managers in these cities can use the types of information generated by this study to assess demand, promote conservation, obtain insights about the real operational efficiency of fixtures within residential homes in Utah, design rebate programs, determine the effectiveness of such programs or other commonly applied strategies for managing demand, or to simply gain further insights into how and when are people using water.

Additionally, engineers and city planners can use the type of information we derived from the data we collected to increase the accuracy of water use estimations and assess infrastructure needs for future urban developments.

Data availability statement

The high resolution water use dataset containing the data for all 31 participant sites, the anonymized information collected for each site, the final end use events file, and log files indicating key information about each data collection period are publicly available in the HydroShare repository (Bastidas Pacheco et al. 2021a). The dataset with events manually labeled by the resident of site 19 from which our classification model was trained and tested is also available in HydroShare (Bastidas Pacheco and Horsburgh 2021b).

Reproducible results

The code used to generate all the results presented in this paper is available in HydroShare (Bastidas Pacheco and Horsburgh 2021b). Patricia Ayaa (Utah State University, Utah) downloaded and ran the code and reproduced the results presented.

Acknowledgments

This research was funded by the United States National Science Foundation under grant number 1552444. Any opinions, findings, and conclusions or recommendations expressed are those of the authors and do not necessarily reflect the views of the National Science Foundation. Additional financial support was provided by the Utah Water Research Laboratory at Utah State University. We acknowledge Logan and Providence Cities for their cooperation in the realization of the field data collection campaigns. The

authors also gratefully acknowledge the homeowners who participated in our data collection efforts.

REFERENCES

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

Al-Kofahi, S., Steele, C., VanLeeuwen, D., St. Hilaire, R. 2012. “Mapping land cover in urban residential landscapes using very high spatial resolution aerial photographs”. Urban For. Urban Green. 11, 291–301. https://doi.org/10.1016/j.ufug.2012.05.001

Aquacraft. (1996). “Trace Wizard description.” Accessed June 28, 2021. https://aquacraft.com/data-downloads/trace-wizard/

Aquacraft, 2016. “Project Downloads”. Accessed May 3, 2021. https://aquacraft.com/data-downloads/

Attallah, N.A., Horsburgh, J.S., Bastidas, C., Bastidas Pacheco, C.J. 2021a. “Tools for Evaluating, Developing, and Testing Water End Use Disaggregation Algorithms,” submitted, J. Water Resour. Plan. Manag.

Attallah, N.A., Horsburgh, J.S., Beckwith, A.S., Tracy, R.J. 2021b. “Residential Water Meters as Edge Computing Nodes: Disaggregating End Uses and Creating Actionable Information at the Edge”. Sensors 21. https://doi.org/10.3390/s21165310

Bastidas Pacheco, C.J., Attallah, N.A., Horsburgh, J.S., 2021a. “High Resolution Residential Water Use Data in Cache County, Utah, USA”. HydroShare. http://www.hydroshare.org/resource/0b72cddfc51c45b188e0e6cd8927227e

Bastidas Pacheco, C.J., Brewer, J.C., Horsburgh, J.S., Caraballo, J. 2021b. “An open source cyberinfrastructure for collecting, processing, storing and accessing high temporal resolution residential water use data”. Environ. Model. Softw. https://doi.org/10.1016/j.envsoft.2021.105137

Bastidas Pacheco, C.J., Horsburgh, J.S., 2021a. “Standardized Monthly Water Use Data for Logan and Providence Cities, Utah, USA.”’ HydroShare. http://www.hydroshare.org/resource/16c2d60eb6c34d6b95e5d4dbbb4653ef

Bastidas Pacheco, C.J., Horsburgh, J.S., 2021b. “Supporting data and tools for "Variability in Consumption and End Uses of Water for Residential Users in Logan and Providence, Utah, USA"”. HydroShare. http://www.hydroshare.org/resource/379d9e7037f04478a99d5aec22e841e6

Bastidas Pacheco, C.J., Horsburgh, J.S., Tracy, R.J. 2020. “A Low-Cost, Open Source Monitoring System for Collecting High Temporal Resolution Water Use Data on Magnetically Driven Residential Water Meters”. Sensors 20, 3655. https://doi.org/10.3390/s20133655

Beal, C., Stewart, R.A. 2011. “South East Queensland Residential End Use Study: Final Report”. Urban Water Security Research Alliance. Accessed October 15, 2021. http://www.urbanwateralliance.org.au/publications/UWSRA-tr47.pdf

Beal, C.D., Stewart, R.A. 2014. “Identifying Residential Water End Uses Underpinning Peak Day and Peak Hour Demand”. J. Water Resour. Plan. Manag. 140, 4014008. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000357

Blokker, E.J.M., G., V.J.H., van Dijk, J.C. 2010. “Simulating Residential Water Demand with a Stochastic End-Use Model”. J. Water Resour. Plan. Manag. 136, 19–26. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000002

Boyle, T., Giurco, D., Mukheibir, P., Liu, A., Moy, C., White, S., Stewart, R. 2013. “Intelligent metering for urban water: A review”. Water 5, 1052. https://doi.org/10.3390/w5031052

Cahill, J., Hoolohan, C., Lawson, R., & Browne, A. L. 2021. “COVID-19 and water demand: A review of literature and research evidence”. Wiley Interdisciplinary Reviews: Water e1570. https://doi.org/10.1002/wat2.1570

Cominola, A., Giuliani, M., Castelletti, A., Fraternali, P., Gonzalez, S.L.H., Herrero, J.C.G., Novak, J., Rizzoli, A.E. 2021. “Long-term water conservation is fostered by smart meter-based feedback and digital user engagement”. Clean Water 4, 29. https://doi.org/10.1038/s41545-021-00119-0

Cominola, A., Giuliani, M., Castelletti, A., Rosenberg, D.E., Abdallah, A.M. 2018. “Implications of data sampling resolution on water use simulation, end-use disaggregation and demand management”. Environ. Model. Softw. 102, 199–212. https://doi.org/10.1016/j.envsoft.2017.11.022

Cominola, A., Giuliani, M., Piga, D., Castelletti, A., Rizzoli, A.E. 2015. “Benefits and challenges of using smart meters for advancing residential water demand modeling and management: A review”. Environ. Model. Softw. 72, 198–214. https://doi.org/10.1016/j.envsoft.2015.07.012

Cooley, H., Gleick, P.H., Abraham, S., Cai, W. 2020. “Water and the COVID-19 Pandemic: Impacts on Municipal Water Demand”. Pacific Institute, Issue Brief. Accessed May 7, 2021. https://pacinst.org/wp-content/uploads/2020/07/Water-and-COVID-19_Impacts-on-Municipal-Water-Demand_Pacific-Institute.pdf

DeOreo, W.B., Mayer, P.W., Dziegielewski, B., Kiefer, J., Foundation, W.R. 2016. “Residential End Uses of Water, Version 2”. Water Research Foundation.

DeOreo, W.B., Mayer, P.W., Martien, L., Hayden, M., Funk, A., Kramer-Duffield, M., Davis, R., Henderson, J., Raucher, B., Gleick, P. 2011. “California single-family water use efficiency study, Report prepared for the California Dept. of Water Resources”. Aquacraft Inc.

Di Mauro, A., Cominola, A., Castelletti, A., Di Nardo, A. 2020. “Urban Water Consumption at Multiple Spatial and Temporal Scales. A Review of Existing Datasets”. Water 13, 36. https://doi.org/10.3390/w13010036

Dieter, C.A., Maupin, M.A., Caldwell, R.R., Harris, M.A., Ivahnenko, T.I., Lovelace, J.K., Barber, N.L., Linsey, K.S. 2018. “Estimated use of water in the United States in 2015”. U.S. Geological Survey Circular 1441 https://doi.org/10.3133/cir1441

DOE, 1992. Energy Policy Act.

Doorenbos, J., Pruitt, W.O., 1977. “Guidelines for predicting crop water requirements”. FAO Irrigation and Drainage Paper 24.

Endter-Wada, J., Kurtzman, J., Keenan, S.P., Kjelgren, R.K., Neale, C.M.U. 2008. “Situational Waste in Landscape Watering: Residential and Business Water Use in an Urban Utah Community”. J. Am. Water Resour. Assoc. 44, 902–920. https://doi.org/10.1111/j.1752-1688.2008.00190.x

EPA, 2021. “WaterSense Products”. Accessed May 7, 2021. https://www.epa.gov/watersense/watersense-products

EPA, 2020. “The WaterSense Label”. Accessed May 3, 2021. https://www.epa.gov/watersense

EPA, 2018. “Drinking Water Infrastructure Survey and Assessment”. Accessed May 3, 2021. https://www.epa.gov/dwsrf/epas-6th-drinking-water-infrastructure-needs-survey-and-assessment

EPA, 2016. “Urbanization and Population Change”. https://cfpub.epa.gov/roe/indicator.cfm?i=52

EPA, 2014. WaterSense ® Specification for Tank-Type Toilets Version 1.2. Accessed September 29. 2021. https://www.epa.gov/sites/default/files/2017-01/documents/ws-products-spec-toilets.pdf

Fielding, K.S., Spinks, A., Russell, S., McCrea, R., Stewart, R., Gardner, J. 2013. “An experimental test of voluntary strategies to promote urban water demand management”. J. Environ. Manage. 114, 343–351. http://dx.doi.org/10.1016/j.jenvman.2012.10.027

Froehlich, J., C. Larson, E., Campbell, T., Haggerty, C., Fogarty, J., N. Patel, S., 2009. “HydroSense: Infrastructure-mediated single-point sensing of whole-home water activity”. Proceedings of the 11th international conference on Ubiquitous computing. https://doi.org/10.1145/1620545.1620581

Glenn, D.T., Endter-Wada, J., Kjelgren, R., Neale, C.M.U., 2015. Tools for evaluating and monitoring effectiveness of urban landscape water conservation interventions and programs. Landsc. Urban Plan. 139, 82–93. https://doi.org/10.1016/j.landurbplan.2015.03.002

Inman, D., Jeffrey, P. 2006. “A review of residential water conservation tool performance and influences on implementation effectiveness”. Urban Water J. 3, 127–143. https://doi.org/10.1080/15730620600961288

Kjelgren, R., Kilgren, D., Rupp, L., Kilgren, D. 2000. “Water conservation in urban landscapes”. HortScience 35 https://journals.ashs.org/hortsci/view/journals/hortsci/35/6/article-p1037.xml

Kofinas, D.T., Spyropoulou, A., Laspidou, C.S. 2018. “A methodology for synthetic household water consumption data generation”. Environ. Model. Softw. 100, 48–66. https://doi.org/10.1016/j.envsoft.2017.11.021

Kruskal, William H., Wallis, W. Allen 1952. “Use of Ranks in One-Criterion Variance Analysis”. Journal of the American Statistical Association, 47:260, 583-621. https://doi.org/10.1080/01621459.1952.10483441

Liaw, A., Wiener, M. 2002. “Classification and Regression by randomForest”.

Lüdtke DU, Luetkemeier R, Schneemann M, Liehr S. 2021. ‘Increase in Daily Household Water Demand during the First Wave of the Covid-19 Pandemic in Germany”. Water 13(3):260. https://doi.org/10.3390/w13030260

Makonin, Stephen.”AMPds2: The Almanac of Minutely Power Dataset (Version 2)”. 2016. Accessed May 3, 2021. https://doi.org/10.7910/DVN/FIE0S4

Marzano, R., Rougé, C., Garrone, P., Grilli, L., Harou, J.J., Pulido-Velazquez, M. 2018. “Determinants of the price response to residential water tariffs: Meta-analysis and beyond”. Environ. Model. Softw. 101, 236–248. https://doi.org/10.1016/j.envsoft.2017.12.017

Mauro, A. Di, Nardo, A. Di, Santonastaso, G.F., Venticinque, S. 2019. “An IoT System for Monitoring and Data Collection of Residential Water End-Use Consumption”. 28th International Conference on Computer Communication and Networks (ICCCN). pp. 1–6. https://doi.org/10.1109/ICCCN.2019.8847120

Mayer, P.W., B. DeOreo, W., Towler, E., Martien, L., M. Lewis D., 2004. “Tampa Water Department Residential Water Conservation Study: The Impacts of High Efficiency Plumbing Fixture Retrofits in Single-Family Homes”.

Mayer, P.W., DeOreo, W.B., Optiz, E.M., Kiefer, J.C., Davis, W.Y., Dziegielewski, B., Nelson, J.O. 1999. “Residential End Uses of Water”. American Water Works Association.

Menneer, T.; Qi, Z.; Taylor, T.; Paterson, C.; Tu, G.; Elliott, L.R.; Morrissey, K.; Mueller, M. 2021. “Changes in Domestic Energy and Water Usage during the UK COVID-19 Lockdown Using High-Resolution Temporal Data”. Int. J. Environ. Res. Public Health 18, 6818. https://doi.org/10.3390/ijerph18136818

Meyer, B.E., Jacobs, H.E., Ilemobade, A. 2020. “Extracting household water use event characteristics from rudimentary data”. J. Water Supply Res. Technol. 69, 387–397. https://doi.org/10.2166/aqua.2020.153

Milligan, M. 2018. “Glad You Asked: Does Utah Really Use More Water Than Any Other State? – Utah Geological Survey”. Accessed October 19, 2021. https://geology.utah.gov/map-pub/survey-notes/glad-you-asked/does-utah-use-more-water/

Nguyen, K.A., Stewart, R.A., Zhang, H., Jones, C. 2015. “Intelligent autonomous system for residential water end use classification: Autoflow”. Appl. Soft Comput. 31, 118–131. https://doi.org/10.1016/j.asoc.2015.03.007

Nguyen, K.A., Stewart, R.A., Zhang, H., Sahin, O. 2018. “An adaptive model for the autonomous monitoring and management of water end use”. Smart Water 3, 5. https://doi.org/10.1186/s40713-018-0012-7

Otaki, Y., Otaki, M., Sugihara, H., Mathurasa, L., Pengchai, P., Aramaki, T. 2011.

"Comparison of residential indoor water consumption patterns in Chiang Mai and Khon Kaen, Thailand". J. AWWA 103, 104–110. https://doi.org/10.1002/j.1551-8833.2011.tb11457.x

Pastor-Jabaloyes, L., Arregui, F.J., Cobacho, R., 2018. "Water end use disaggregation based on soft computing techniques". Water 10, 21. https://doi.org/10.3390/w10010046

Rathnayaka, K.; Malano, H.; Maheepala, S.; George, B.; Nawarathna, B.; Arora, M.; Roberts, P. 2015. "Seasonal Demand Dynamics of Residential Water End-Uses". Water 7, 202-216. https://doi.org/10.3390/w7010202

Roberts, P., 2005. Yarra Valley Water 2004 residential End Use Measurement Study.

Rockaway, T.D., Coomes, P., Rivard, J., Kornstein, B., Foundation, W.R. 2010. "North America Residential Water Usage Trends Since 1992". The Water Research Foundation.

Romero, C.C., Dukes, M.D. 2015. "Review of Turfgrass Evapotranspiration and Crop Coefficients", in: 2015 ASABE / IA Irrigation Symposium. https://doi.org/10.13031/irrig.20152145395

Student, 1908. "The probable error of a mean". Biometrika, pp.1–25.

Suero, F.J., Mayer, P.W., Rosenberg, D.E. 2012. "Estimating and Verifying United States Households' Potential to Conserve Water". J. Water Resour. Plan. Manag. 138, 299–306. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000182

The University of Utah, 2016. Fact Sheet August 2016. Utah at a Glance. https://gardner.utah.edu/wp-content/uploads/2016/08/UtahAtAGlance-Final1.pdf

UGRC 2021. "Aerial Photography Products for Utah". Utah Geospatial Resource Center. Accessed May 3, 2021. https://gis.utah.gov/data/aerial-photography/

UN-Habitat 2016. "Urbanization and Development: Emerging Futures, World Cities Report 2016".

Utah DWR, 2020. "2015 Municipal and Industrial Water Use Data: 2020 Version 3." Utah Division of Water Resources. https://drive.google.com/file/d/1aD9SorKQauIfiDW0wdMXlafd0VKsdX-F/view.

Vitter, J. S., and Michael Webber. 2018. "Water Event Categorization Using Sub-Metered Water and Coincident Electricity Data" Water 10, no. 6: 714. https://doi.org/10.3390/w10060714

White, R., Havalak, R., Nations, J., Thomas, J., Chalmers, D., Dewey, D. 2004. "How Much Water is Enough? Using PET to Develop Water Budgets for Residential landscapes". Texas A&M University. https://hdl.handle.net/1969.1/6100

White, S., Robinson, J., Codell, D., Jha, M., Milne, G. 2003. "Urban water demand forecasting and demand management: research needs review and recommendations" Water Services Association of Australia.

Willis, R.M., Stewart, R.A., Giurco, D.P., Talebpour, M.R., Mousavinejad, A. 2013. "End use water consumption in households: Impact of socio-demographic factors

and efficient devices". J. Clean. Prod. 60, 107–115. https://doi.org/10.1016/j.jclepro.2011.08.006

Willis, R.M., Stewart, R.A., Panuwatwanich, K., Jones, S., Kyriakides, A. 2010. "Alarming visual display monitors affecting shower end use water and energy conservation in Australian residential households". Resour. Conserv. Recycl. 54, 1117–1127. https://doi.org/10.1016/j.resconrec.2010.03.004

Willis, R.M., Stewart, R.A., Williams, P.R., Hacker, C.H., Emmonds, S.C., Capati, G. 2011. "Residential potable and recycled water end uses in a dual reticulated supply system". Desalination 272, 201–211. https://doi.org/10.1016/j.desal.2011.01.022

Tables

Table 4.1. Datasets used in the present study, source, coverage, and availability.

Dataset	Source	Coverage	Availability
Monthly Water Use for Logan City	Logan City	Jan 2017 – Dec 2018	Anonymized and standardized monthly values are available in HydroShare (Bastidas Pacheco and Horsburgh, 2021a)
Monthly Water Use for Providence City	Providence City	Jan 2018 – Dec 2019	
Parcel and building area for properties in Logan and Providence	Cache County	Updated and maintained by Cache County	Available in HydroShare for participant sites (Bastidas Pacheco et al., 2021)
Aerial photography for the area. Hexagon (1 ft or 6 in) and Google (6 in) Licensed Imagery, and High Resolution Orthophotography (1 foot or better) (UGRC, 2021).	Utah Geospatial Resource Center	Collected between 2012 and 2021.	Available to use by Utah agencies and educational institutions in web and desktop mapping applications.
High temporal resolution (4 second) water use data	Collected by the authors	Collected between 2019 and 2021	Anonymized version available in HydroShare (Bastidas Pacheco et al., 2021)
Characteristics of each residence participating in the study	Surveyed by the authors, combined with county data	Surveys conducted during enrollment	
Daily rainfall and evapotranspiration data for the USU Environmental Observatory and Evans Farm weather stations	Utah Climate Center	Jan 2019 – Apr 2021	Publicly available in HydroShare (Bastidas Pacheco and Horsburgh, 2021b)

Table 4.2. Data collection period and characteristics of each site where data was collected.

Site-ID	Length of QC data record (Weeks)	Number of Occupants	Irrigable Area (m^2)	Building Area (m^2)	Irrigation mode	Volumetric pulse resolution (L/pulse)	Annual average water use (m^3)
2	21.6	2	643	140	Sprinkler System	0.1257	397.9
3	22.5	2	1,408	138	Hose	0.0329	234.9
4	9.3	4	1,015	136	Sprinkler System	0.0329	647.7
5	16.4	2	3,118	169	Sprinkler System	0.1257	1786.0
6	6.2	3	294	101	Hose	0.0329	96.2
7	16.8	3	241	104	Hose	0.1257	55.2
8	6.4	2	1,789	160	Hose	0.1257	720.6
9	9.3	2	509	173	Sprinkler System	0.1257	602.4
10	6.2	2	824	102	Hose	0.1257	149.0
11	12.4	4	827	136	Sprinkler System	0.1257	401.3
12	9.9	2	1,744	156	Sprinkler System	0.1257	181.2
13	7.8	2	742	239	Sprinkler System	0.1257	1507.5
14	10.4	2	2,005	315	Sprinkler System	0.1257	1099.8
15	9	6	405	171	Sprinkler System	0.1257	392.2
16	7.4	3	1,162	151	Hose	0.0329	247.4
17	8.5	3	1,451	92	Hose	0.0329	341.0
18	8.7	1	410	74	Hose	0.0329	233.5
19	23.1	5	982	128	Sprinkler System	0.1575	854.1
20	5.2	4	1,202	177	Sprinkler System	0.1575	942.2
21	6.8	6	NA	NA	Sprinkler System	0.1575	NA
22	8	7	NA	NA	Sprinkler System	0.1575	NA
23	5.7	6	1,108	144	Sprinkler System	0.1575	809.2
24	8.1	8	1,276	279	Sprinkler	0.1575	1308.0

Site-ID	Length of QC data record (Weeks)	Number of Occupants	Irrigable Area (m^2)	Building Area (m^2)	Irrigation mode	Volumetric pulse resolution (L/pulse)	Annual average water use (m^3)
					System		
25	7.4	6	914	282	Sprinkler System	0.1575	614.4
26	6.7	6	3,592	117	Sprinkler System	0.0962	644.0
27	6.7	7	3,842	299	Sprinkler System	0.0962	2248.6
28	4.8	6	1,846	337	Sprinkler System	0.1575	1747.6
29	4.8	3	700	133	Sprinkler System	0.1575	573.3
30	5	6	1,250	137	Sprinkler System	0.1575	716.2
31	4.6	3	827	154	Sprinkler System	0.1257	695.7
32	4.6	2	862	104	Sprinkler System	0.0329	730.1

Notes: The length of the record presented here is the sum of all individual data collection periods that passed quality control. Water use records for site 21 and 22 were not available (NA).

Table 4.3. Category benchmarks for the LIR (Glenn et al., 2015).

Benchmark category		LIR value
Justifiable water use	Efficient	$LIR \leq 1$
	Acceptable	$1 < LIR \leq 2$
Unjustifiable water use	Inefficient	$2 < LIR \leq 3$
	Excessive	$LIR > 3$

Table 4.4. Number of sites and changes observed in the mean frequency of events (summer versus winter).

Change observed	Faucet	Shower	Toilet	Bathtub	Clothes Washer
Larger frequency of events (summer versus winter)	4	1	2	0	2
Smaller frequency of events (summer versus winter)	1	1	0	0	0
No significant change	5	8	8	6	8

Table 4.5. Number of sites and changes observed in the mean volume of events (summer versus winter)

Change observed	Faucet	Shower	Bathtub
Larger mean volume (summer versus winter)	2	0	1
Smaller mean volume (summer versus winter)	4	1	0
No significant change	4	9	5

Figures

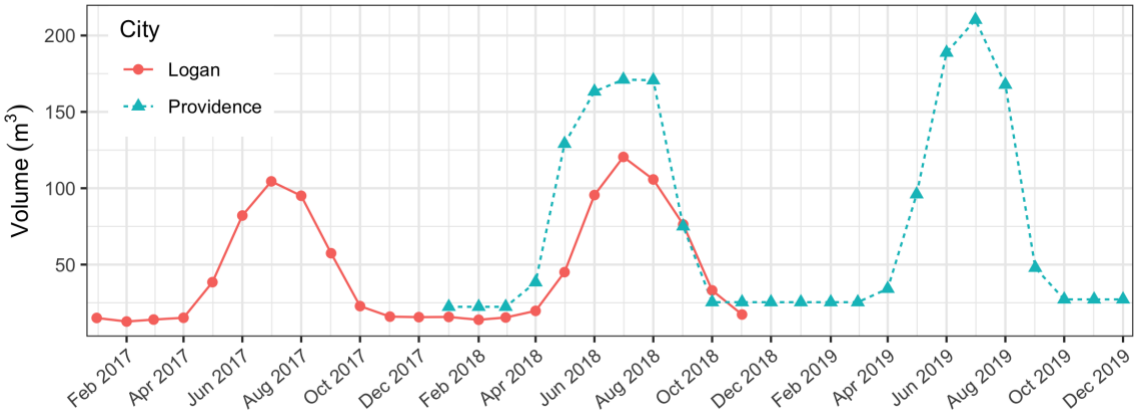


Figure 4.1. Average monthly water use per household across all residential customers in Logan and Providence, Utah between 2017 and 2019 calculated from billing data for 7,522 and 2,113 connections, respectively.

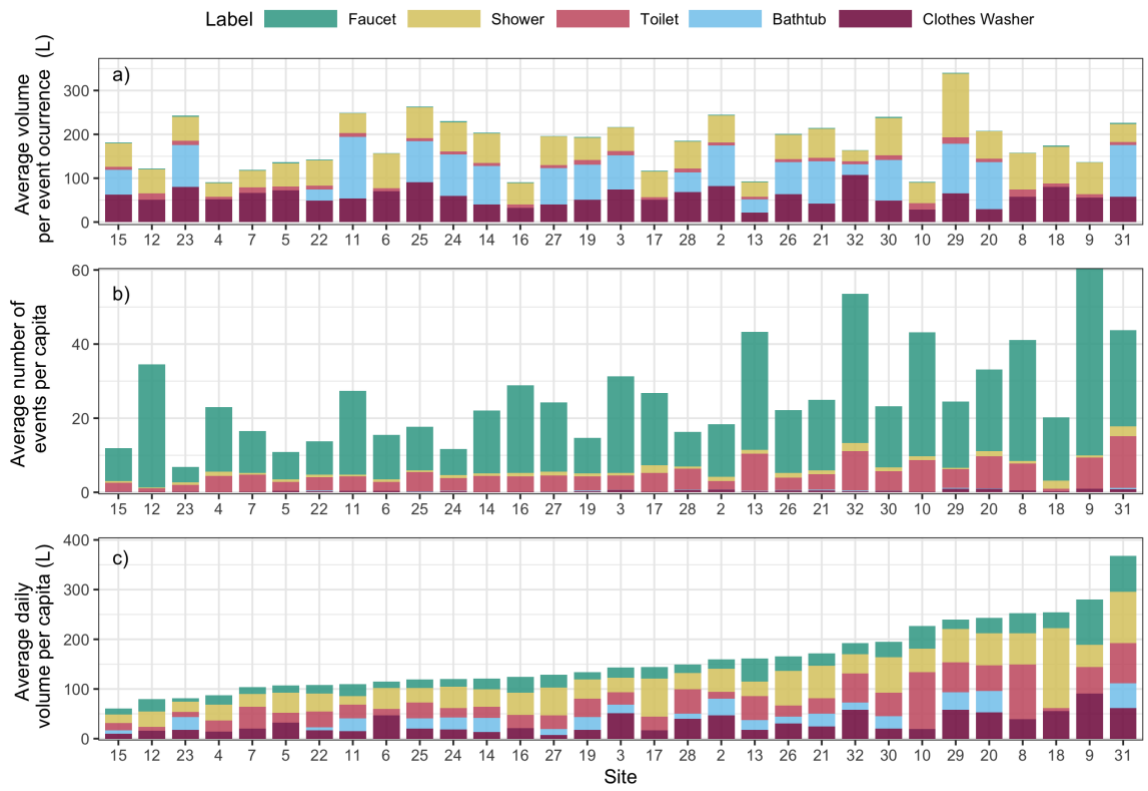


Figure 4.2. Indoor water use summary by site: a) average water use volume per event occurrence, b) average number of events per capita per day, and c) average daily indoor water use per capita and distribution among end uses. Note: The average number of faucet events per capita per day at site 9 is 164 (the y axis at panel b is limited at 60 for visualization purposes).

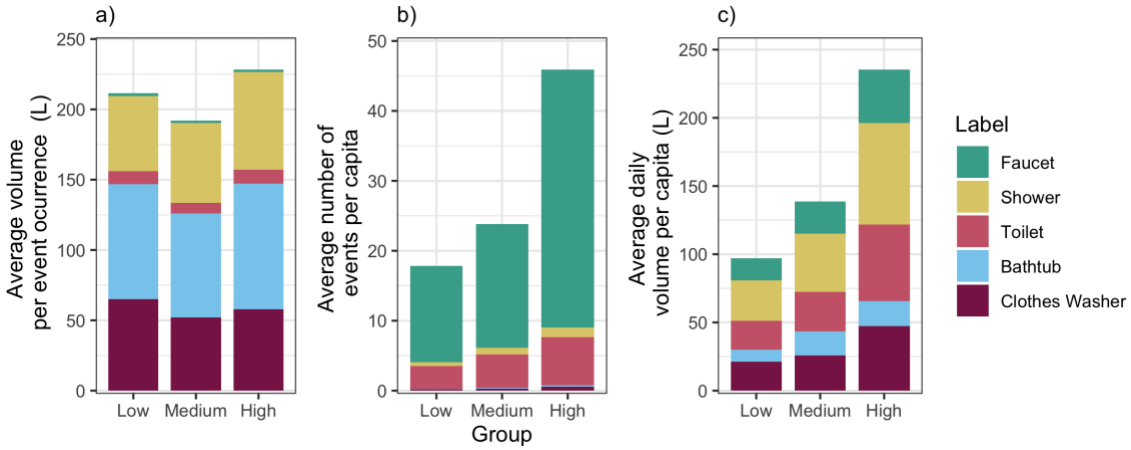


Figure 4.3. Indoor water use summary by group for low, medium, and high water users: a) average water use volume per event occurrence, b) average number of events per capita per day, and c) average daily indoor water use per capita and distribution among end uses.

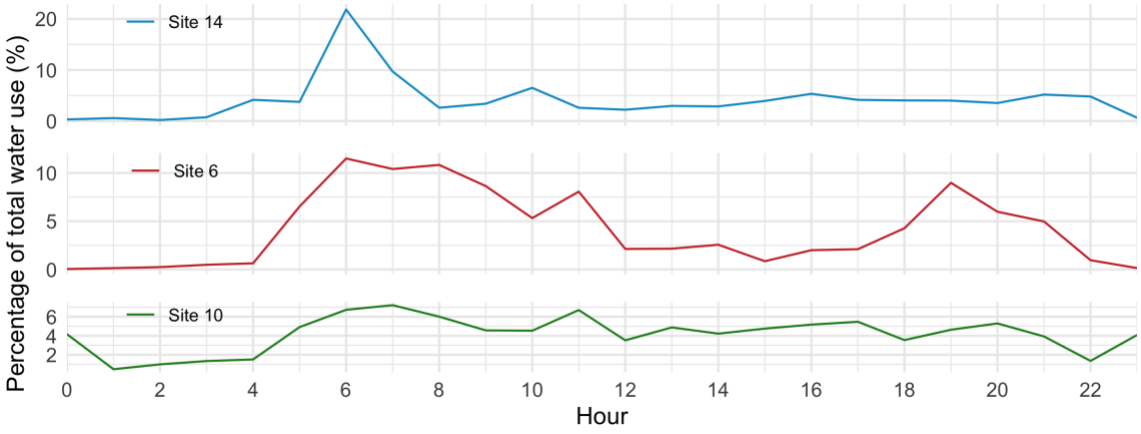


Figure 4.4. Examples of hourly distribution (in percentage) of total indoor water use: 1) a single period of higher consumption (Site 14), 2) multiple periods of higher consumption (Site 6), and relatively similar water use throughout the day (Site 10).

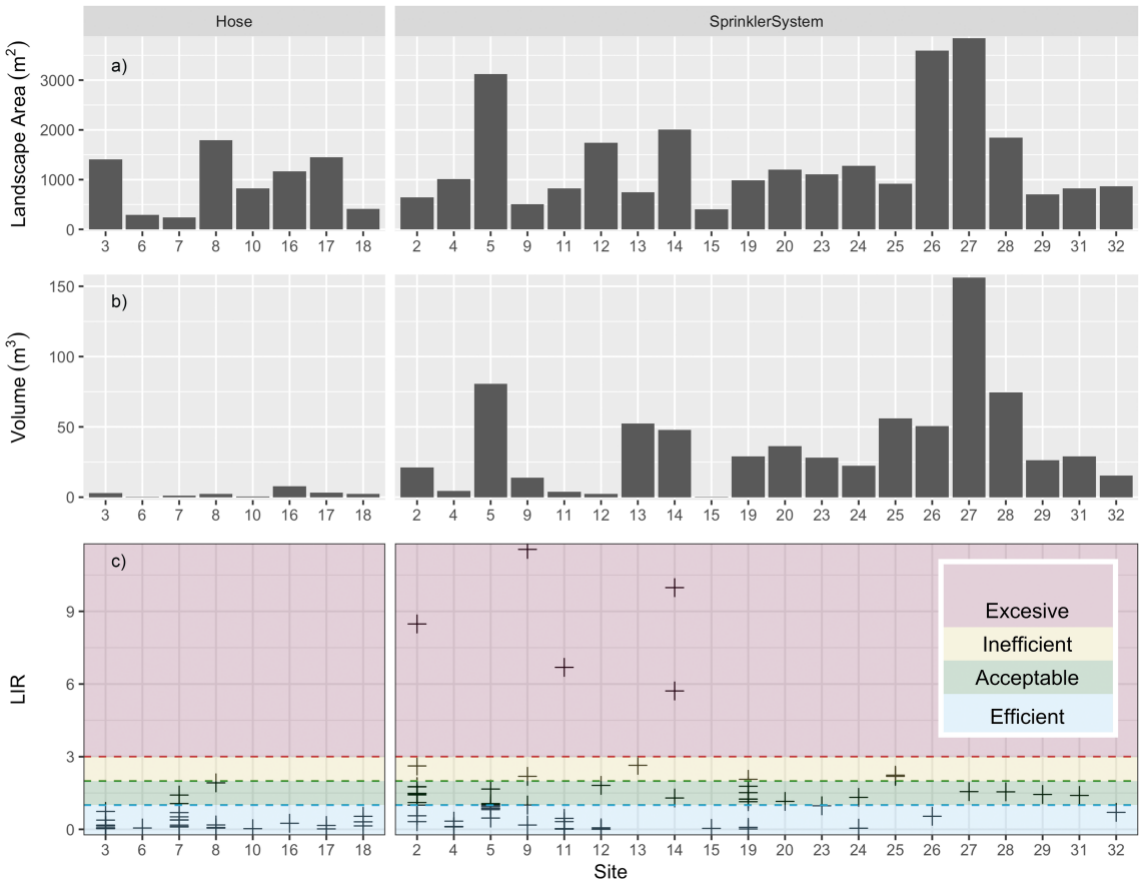


Figure 4.5. Weekly outdoor water use information (excluding weeks where the landscape water needs were zero) and landscape area: a) landscape area, b) average weekly outdoor water use volume, and c) weekly LIR values for each site.

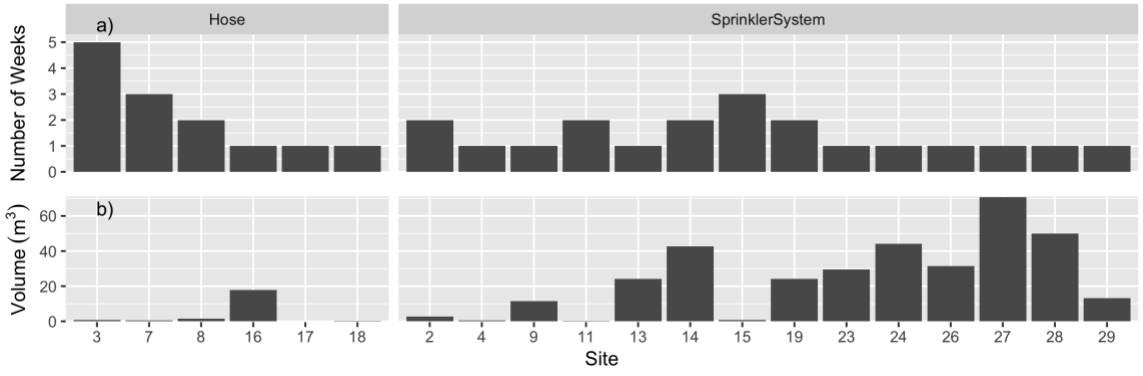


Figure 4.6. Outdoor water use measured during weeks when landscape irrigation need was zero: a) Number of weeks of data collected, and b) average volume used.

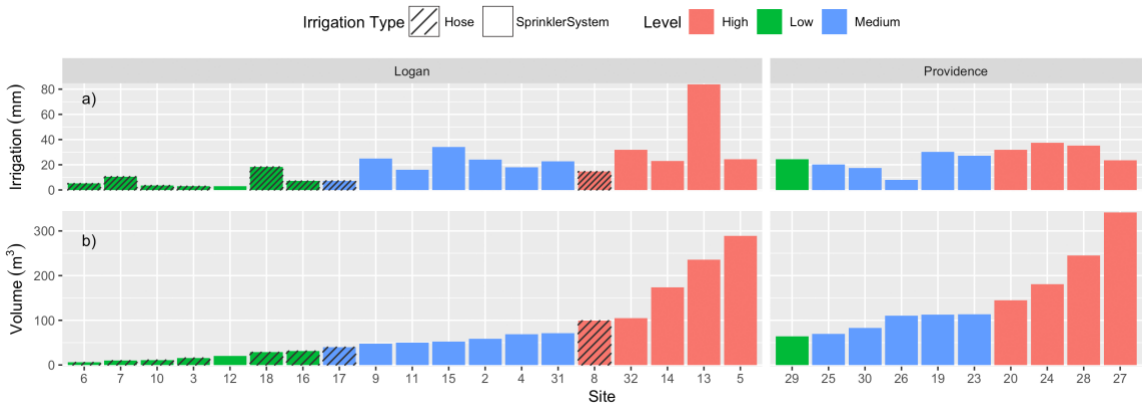


Figure 4.7. Outdoor water use analysis from monthly records: a) outdoor water use per unit area, and b) average monthly outdoor water use.

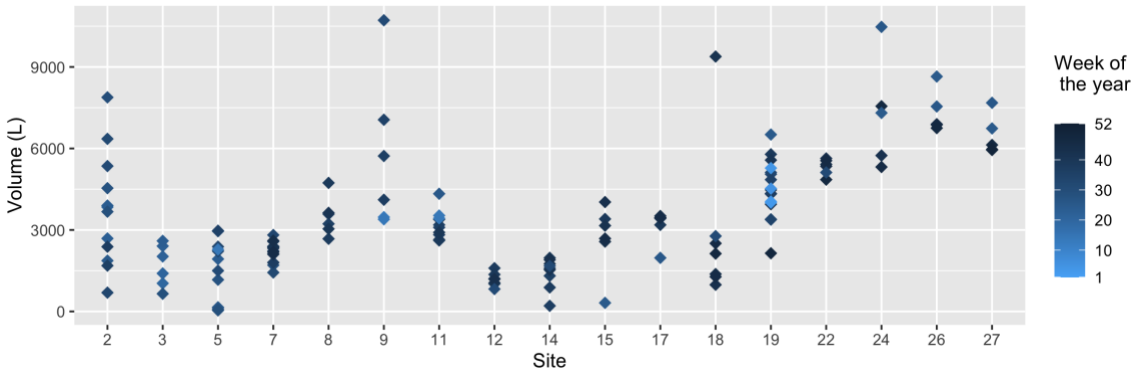


Figure 4.8. Indoor weekly water use volumes for sites with a data record longer than four weeks. The point color indicates the week of the year.

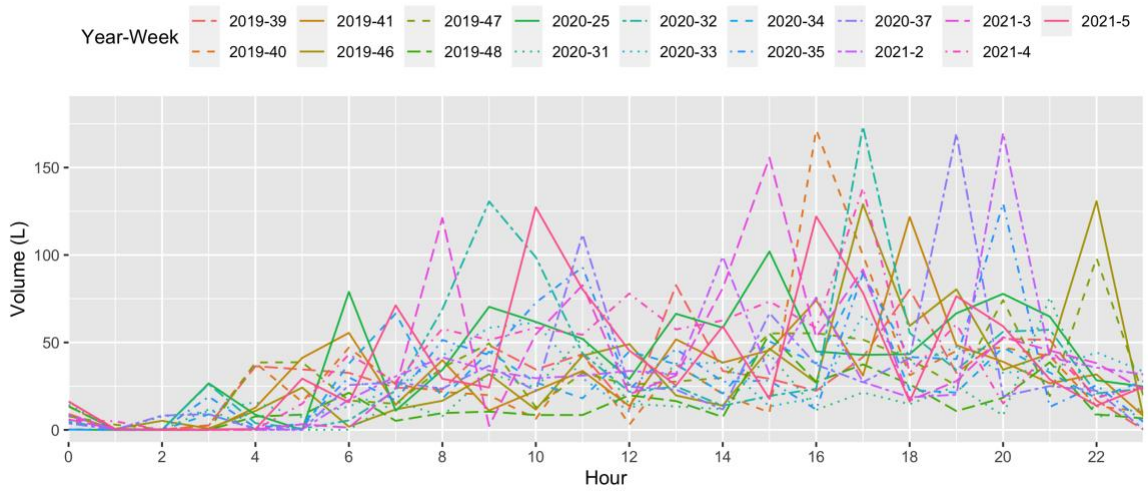


Figure 4.9. Total hourly indoor water use for the 17 full weeks of data at site 19. Values for each hour include all water used during that hour, e.g., the value plotted at 4:00 AM includes all water use between 4:00 AM and 5:00 AM. The week of the year is indicated in the labels (YYYY-WW).

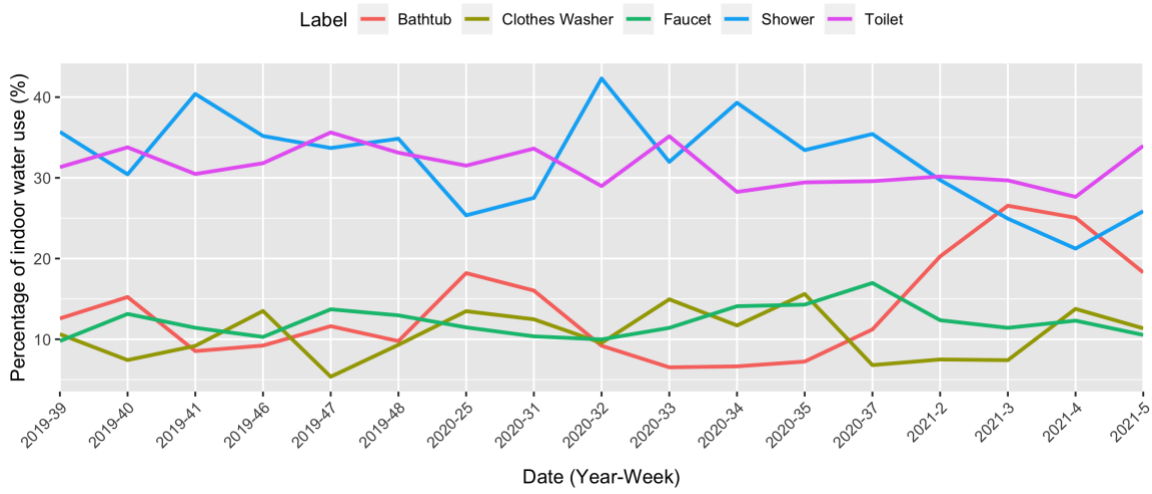


Figure 4.10. Weekly percentages of indoor water use by end use for the 17 full weeks of data at site 19.

CHAPTER 5

SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

The research presented in this dissertation sought to address the need to better understand how water is used at the household level as well as the growing need for open source tools that support collection and management of data that enables observing water use behavior and characteristics (i.e., high temporal resolution water use data). These needs are driven by rapidly growing urban populations, climate change and variability, uncertainty in urban water supplies, and aging infrastructure that will inevitably need to be replaced. Water management decisions require data, and the types of data that are most commonly recorded by water managers do not meet all of their needs. While tools to measure water use data at a high temporal resolution are not new, they have traditionally been proprietary and private, which has prevented the advancement of the tools and the widespread collection of data. There is consensus among researchers in this field that the benefits from open source tools, implementation, case studies, and data of this kind will contribute towards achieving the goals of more efficient residential water use and better informed management.

The collective hardware and software tools required to enable more informed management of urban water resources through high resolution data collection make up Cyberinfrastructure that make these goals more attainable. The significance of the work presented in this dissertation includes presentation of a design and implementation of hardware and software tools that enable recording and managing high temporal resolution data as well as case studies that demonstrate the suitability of these tools for addressing existing gaps in data collection, management, and analysis aimed at better quantifying

residential water use. The design and implementation of hardware and software tools was guided by the need to address interests of researchers and water managers while generating information that is also useful for water users and decision makers. All the work presented in this dissertation was part of the Cyberinfrastructure for Intelligent Water Supply (CIWS) project funded by the U.S. National Science Foundation.

Chapter 2 presented the design and implementation of a residential water use datalogging device, called the CIWS Datalogger, designed to work on top of existing, analog, magnetically driven, positive displacement, residential water meters. The CIWS Datalogger can collect data at a variable time resolution (selected by the user) and can be deployed to the field autonomously for approximately 5 weeks when collecting data at a 4 s time interval. This exceeds the capabilities and autonomy of devices used in past research projects analyzing residential end uses of water. Extending the autonomy of this type of device allows data collection campaigns designed to answer research questions that require more than a few days of data collection. Battery life remains the limiting factor in the autonomy of this type of device, which constitutes a barrier in the implementation of advanced metering infrastructure and the implementation of smart water networks.

The CIWS Datalogger is a low cost (~$150) device, which facilitates its use in cases where cost limits the collection of high temporal resolution data. The only other device currently on the commercial market with similar capabilities sells for more than $2500 per unit. Additionally, the CIWS Datalogger was built using open source electronic hardware and firmware allowing modification and advancement of the device itself. In fact, a device that advances the CIWS Datalogger functionality, adding wireless

communication and edge computational capabilities, already exists (Attallah et al., 2021b). Data collected using the CIWS Datalogger is, under ideal installation conditions, within 2% of the volume read by the register of the meter on which it is installed, making it as accurate as any other existing similar device.

Chapter 3 presented CIWS, an open source cyberinfrastructure that automates the process from data collection to analysis and presentation of high temporal residential water use data. The chapter includes the design and a prototype implementation that was tested in two case studies, one in a single family residential (SFR) context, and the other in residential buildings that host Utah State University (USU) students. CIWS has three main architectural components: first, the sensors and dataloggers for water use monitoring; second, the data communication, parsing and archival tools; and third, the analyses, visualization and presentations of data produced for different audiences. For the first component, the CIWS Computational Node (Attallah et al., 2021b) was used and integrated into CIWS operation. For the second component, we designed software tools, considering the individual characteristics in terms of power availability and type of communication network installed, of each case study. CIWS was adapted to manage pushing data (in the SFR case) and pulling data (in the USU case study), which demonstrate the flexibility in CIWS design. Future users of CIWS can select push or pull, or a combination of both since these functionalities were implemented separately. The USU case study also demonstrated the flexibility of CIWS to manage a variety of data by incorporating temperature and water use data from different meters. In both implementations, the data is stored in an open source database implementation.

For the third component, we developed an application programming interface

(API) that connects to the databases and generates a set of analyses that are of interest for researchers, water managers, and homeowners. This API works independent of how the data are transferred to the database and provides a proof of concept showing the foundation upon which more sophisticated tools could be built. Researchers interested in answering specific research questions can access the raw data collected through the API. Additionally, analytic tools that estimate multiple statistics of interest (e.g., peak time and volume, maximum hourly water use) were developed. An existing open source tool to calculate end uses of water (Attallah et al., 2021a) was integrated into CIWS modules for analyses and workflow. The system was tested for scalability and performance, and the results indicated that it could easily handle the scale of data collected for common research projects and could be adapted to meet the needs of much larger deployments. The current version of CIWS could be used to collect and manage the data required to assist in the design and implementation of water conservation programs, rebate programs, water demand estimation and forecasting, and design of future urban water infrastructure. We envision future improvements to the system once it is used in additional studies.

Chapter 4 analyzed the variability, in terms of timing and distribution of end uses, of residential water use in a sample of 31 SFR properties in the cities of Logan and Providence, in Utah. The data used in this study were collected and managed using the hardware and software described in Chapters 2 and 3. The analyses presented were derived from 4 to 23 weeks of high temporal resolution water use data for each participant site collected using the CIWS datalogger at a 4 s temporal resolution between 2019 and 2021. We found that outdoor water use was the largest component of residential water use among participants, accounting for approximately 84% of the volume we

measured between April and October. Despite its large contribution to overall water use, in most cases residents were not grossly overwatering their landscape. We found that most residents irrigated in early morning or late evening, which is recommended to reduce losses due to evaporation. The differences observed in outdoor water use among participant sites were produced by differences in the irrigation method used (i.e., hose versus sprinkler systems) and the irrigated landscape area, with automated sprinkler systems using more water than hose irrigators and increasing water use with larger irrigated landscapes.

Showers and toilets were the two largest indoor water use categories, among the five observed (showers, toilets, faucets, clothes washer, and bathtubs), accounting for 31.3% and 25.6% of the total indoor water used volume, on average. There is opportunity to conserve water by increasing the efficiency of water using fixtures and promoting conservation behavior. The variability of indoor water use volume and timing observed was the result of a combination of factors: 1) differences in schedule among occupants of a house, 2) characteristics of water using fixtures at home, and 3) personal preferences. We found significant temporal variability (day to day and week to week) in the distribution of end uses, volume, and timing of indoor water use for users with longer (> 4 weeks) data collection periods. Temporal indoor water use timing parameters can be difficult to determine, as they depend on personal schedules, yet they are needed for the accurate design of residential water use infrastructure. Further research aiming at characterizing this variability is needed to fully understand, and accurately predict residential water use.

This dissertation presented novel hardware and software that advance existing

tools for collecting, managing, and analyzing high temporal resolution residential water use data. While the tools presented either exceed the capabilities of existing tools or represent one of the very limited number of existing tools (or both), there are still several areas that can be advanced. Extending the autonomy of the datalogging device presented can further reduce the cost of collecting data for longer periods of time. Currently this type of data is collected at a scale of a few days per site. A device with an autonomy in the scale of a few months would enable longer data collection campaigns designed to address existing gaps in our understanding of residential water use (e.g., the day to day and week to week variability we observed). These devices are suitable for research projects but are, currently, not viable for longer term (i.e., years) deployment, or are unpractical at the utility system level because they are not fully integrated with the metering systems used by utilities.

The functionalities of smart water networks (e.g., the temporal resolution of the data collected, the data transmission frequency and schedule), for research projects and at the utility level, are constrained by the type of smart meter (or datalogger device) installed. For example, the dataloggers used in the SFR case study allow only one way communication (i.e., they can transmit the data collected but cannot receive data) and transmit data once a day (which increases the amount of time a leak can go undetected), limited by power constrains. Additionally, existing water meters are not capable of collecting sub-minute resolution data and commonly use low data rate transmission systems in an effort to conserve power. In order to expand the current functionalities of smart meters, and devices similar to the CIWS datalogger, this power limitation must be addressed.

Research investigating a different water metering paradigm may be beneficial in defining the shape of future smart water networks. For example, water metering devices could be moved inside residential properties, where power and Internet connectivity is readily available, enabling collection of higher temporal water resolution data, edge computing for data processing, real time data transmission, and more advanced functionalities via two way communication. Alternatively, water meters could harness energy from water flowing (Li and Chong, 2019) or could be reconfigured to enable use of solar panels to extend battery life, allowing the same functionalities described above. In either scenario, and even with current trends of water metering, we are generating larger volumes of data, making systems like CIWS vital in order to obtain the expected benefits from the data measured. Interoperable solutions are needed to enable the progress of smart water networks, especially across different metering systems and manufacturers, and open architectures and standards for data management can lead to advancement in this area (Hauser and Roedler, 2015). Edge computing (Paltoglou et al., 2008; Shi et al., 2016) can be used to calculate all relevant parameters at the meter location and reduce the amount of data transmitted along with associated costs. Further research and implementations are needed to define the computational capabilities required to operate systems like CIWS, the tradeoffs between raw data transmission versus edge computing, and the variables of interest that need to be generated by the system.

Our case study produced new insights into residential water use and generated data and information currently not available for Utah. It is known that higher temporal resolution data increases the accuracy of end use disaggregation techniques (Cominola et

al., 2018). Accurate estimation of end uses of water can help water mangers identify fixture/appliance characteristics and performance and water use behaviors, which could increase the efficiency of existing demand management programs (rebates, retrofit, technical assistance). The only available open source tool for end use estimation (Attallah et al., 2021a) was used in this study . The advancement of this tool, and the development of new methods, may further increase the accuracy of the estimations presented.

Additionally, further research is needed to define the tradeoffs between the accuracy of end use estimation and the temporal resolution at which data is collected. Most end use studies up to this point have been based on data that are regularly spaced in time and that aggregate “pulses” from a water meter (where each pulse represents a fixed volume of water) within each recorded time interval. Based on preliminary work with a modified version of the CIWS Datalogger programmed to record the timestamp of each individual pulse, there may be significant opportunity for identifying and classifying water use events by simply examining the pulse rate and/or spacing between pulses that make up an event. While this may simplify identification and classification of events, it could produce more data that would have to be managed. Additional case study applications are needed to demonstrate the benefits of, promote the development of, and encourage wider spread adoption of hardware and software cyberinfrastructure systems that permit collection, management, and analysis of high temporal resolution water use data, including estimation of end uses.

REFERENCES

Attallah, N.A., Horsburgh, J.S., Bastidas, C., Bastidas Pacheco, C.J., 2021a. Tools for Evaluating, Developing, and Testing Water End Use Disaggregation Algorithms, submitted for publication.

Attallah, N.A., Horsburgh, J.S., Beckwith, A.S., Tracy, R.J., 2021b. Residential Water Meters as Edge Computing Nodes: Disaggregating End Uses and Creating Actionable Information at the Edge. Sensors 21. https://doi.org/10.3390/s21165310

Cominola, A., Giuliani, M., Castelletti, A., Rosenberg, D.E., Abdallah, A.M., 2018. Implications of data sampling resolution on water use simulation, end-use disaggregation, and demand management. Environmental Modeling and Software. 102, 199–212. https://doi.org/https://doi.org/10.1016/j.envsoft.2017.11.022

Hauser, A., Roedler, F., 2015. Interoperability: The key for smart water management. Water Supply 1 February 2015; 15 (1): 207–214. https://doi.org/10.2166/ws.2014.096

Li, X.J., Chong, P.H., 2019. Design and Implementation of a Self-Powered Smart Water Meter. Sensors 19, no. 19: 4177. https://doi.org/10.3390/s19194177

Paltoglou, G., Salampasis, M., Satratzemi, M., 2008. A Comparison of Centralized and Distributed Information Retrieval Approaches, in: 2008 Panhellenic Conference on Informatics. pp. 21–25. https://doi.org/10.1109/PCI.2008.18

Shi, W., Cao, J., Zhang, Q., Li, Y., Xu, L., 2016. Edge Computing: Vision and Challenges. IEEE Internet Things J. 3, 637–646. https://doi.org/10.1109/JIOT.2016.2579198

APPENDICES

Appendix A. Water use rankings, Indoor water use statistics, and comparison with past studies

Figure. A1 (a) and (b) show the percentile ranking of annual water use for each participating site, computed for the last two years of data available in each city, years 2017 and 2018 for Logan, and 2018 and 2019 for Providence. Despite the combined user sampling approach used (targeted invitations, word of mouth) and the relatively small number of participating homes, the sample contains a broad range of percentile rankings and annual water use volumes. Percentile rankings were not consistent from year to year, indicating that there is significant interannual variability in water use that is not determined solely by climatic conditions driving outdoor use. Figure. A1 (c) presents the participating sites' per capita daily average water use for the same two years. Occupancy was registered during enrollment (2019-2021), and monthly water use data were recorded during previous years, therefore, changes in occupancy during this period were not accounted for. Figure. A1 (c) shows that our sample includes users that differ from per capita average values presented in the text of the article.

Table A1 shows the average per capita daily volume used for each indoor category and the percentage of indoor water use that each category represents. Short events lasting less than 4 seconds with a single recorded pulse (unclassified) are the most common indoor event (79.2% of all indoor events were in this category) but represent only 4.73% of the average indoor water consumption. This category includes leaks, very short duration events (e.g., faucets and refrigerators with ice makers), and other events that we were not able to separate or identify because they all had the same volume and

duration. “Unknown” events included outlier events identified during clustering that we were unable to classify and represented approximately 1.51% of the total indoor volume.

Showers were the largest indoor water use in our study, representing 31.2% of total indoor water use. Toilets were the second largest end use across all sites at 25.6% of total indoor use, although toilets were the largest water end use in 13 of the 31 homes. In contrast, the 2016 Residential End Uses of Water Study (REUWS) (DeOreo et al. 2016) found that toilets were the largest indoor end use, consuming 24% of indoor volume, followed by showers (19%). The South East Queensland Residential End-Use Study (SEQREUS) (Beal and Stewart 2011) conducted in Australia found that showers consumed 29.5% of the indoor volume and toilets 16.5%. These results indicate that the distribution of water use across end uses is different across individual residential homes as well as regionally.

The number of per capita faucet events, showers, and toilet flushes in our study was 23.2, 0.97, and 5, respectively. The 2016 REUWS found similar results in terms of per capita daily frequency of toilet flushes (5) and faucet events (20) but lower shower frequency (0.69) per capita per day. The 2016 REUWS used a much larger sample of homes (763 homes across nine cities in the U.S. versus 31 homes in two neighboring cities in this study), but the average household occupancy was 2.7, which is much lower than the 3.8 in our study. The number of bathtub events per capita per day in our study was 0.12, higher than the 0.05 encountered by the REUWS. The frequency of clothes washer events in our study was considerably less than that of REUWS. Assuming each load has 2 cycles (wash and rinse) we estimated 0.19 clothes washer events per capita per day versus 0.3 in the 2016 REUWS.

Showers

High use associated with showers can be the result of personal preferences (longer and/or more frequent showers) or the presence of less efficient fixtures (showerheads operating at higher flow rates). Figure. A2 shows the average flow rate and duration of shower events for each site. Site 31 had the largest average shower head flow rate and the largest per capita daily average shower use, despite having lower shower durations. Site 18 had the second largest daily per capita shower consumption. This site had a much lower shower head flow rate but higher shower durations. Site 17 had lower duration and showerhead flow rate but a higher number of showers per day, ranking in third place for daily shower volumes. At site 27, median shower duration was 15 minutes. The average shown in Figure. A2 was increased by three events that had a duration longer than 80 minutes at flow rates in the same range as all showers. Without additional information, it was not possible to identify if these were erroneous events (i.e., incorrectly labeled as showers), and they remained labelled as showers.

Toilets

Figure. A3 shows the volume distribution of toilet flushes for all sites. Toilets are a mechanical end use (i.e., the flow rate and duration do not depend on user preferences and are expected to be similar for each flush). We observed some variability in the volumes shown on Figure. A3. Some sites had a multimodal distribution (e.g., site 16, 19, and 30) that is the result of having multiple toilets with different characteristics. For example, site 19 had toilets that used approximately 8.3 and 13.2 L/flush. The average flush at this site used 10.8 L, but this value could range lower or higher depending on

which toilet is used more frequently. This is true for every site with a multimodal distribution.

The values at the extremes of each violin plot are typically events with flow rates similar to most toilet events but with different duration. These may be half flushes, double flushes, a toilet valve remaining open longer than normal due to flapper valve malfunction, or other uses being misclassified. At some sites with multiple toilets, the distribution shows a single mode but with higher variability. This is the case at sites 20 and 23, which have 3 and 4 toilets, respectively. In these cases, it is likely that toilets perform similarly enough that the volumes mix, giving the appearance of a single mode distribution.

Toilet and faucet events happening simultaneously (i.e., washing hands before the toilet tank is done refilling) is common and can be identified when examining the raw data. However, given the low flow rate of faucets, attempting to automatically separate them tends to make the algorithm too sensitive towards classifying single events as overlapping. For the purposes of this study, we decided to not separate toilet events from short duration faucet events happening simultaneously which means that these events are lumped as toilet events.

Faucets

Faucet events were the third largest category of indoor water use by volume. This category includes kitchen and bathroom faucets, hose bibs, and other short duration and low flow rate events that do not fit other categories. Faucet events were the second most common events, behind the unclassified category, which can also include very short duration faucet events. The characteristics of dishwasher cycles were indistinguishable

from faucet events. Thus, they were labeled as and lumped with faucet events. Future improvements of our classification method could include identification of cycles for dishwashers and clothes washer events, as has been described in other methods (Nguyen et al. 2018). Most faucet events were short (93% last less than 1 minute) and low volume (80% use less than 2 L). Figure. A4 shows flow rates (a) and duration (b) of faucet events across all sites. Sites 5, 18 and 23 have the largest faucet events duration. Site 9 shows the smallest flow rate variability for faucet events, 91% of the faucet events at this site have a flow rate less than 2.2 L/min. Sites 8 and 31 have the highest median faucet flow rates among all participants.

Clothes Washers

Despite clothes washers being a mechanical end use, identification and classification of clothes washer events are not straightforward. Clothes washers can have different configurations such that the volume of water used can vary depending on the load size and cycle selected. Additionally, the flow rate can vary depending on the temperature of water used – hot, cold, or both. According to DeOreo et al. (1996; 2019), the average per load volume used decreased from 155 L in 1996 to 117 L in 2016, and this change was attributed to the adoption of more efficient appliances. However, it is not clear how clothes washer cycles were grouped together in these studies. Other methods used to classify end uses of water use a time span of two hours to aggregate and identify clothes washing cycles (Nguyen et al. 2018), adding all events with clothes washer characteristics in this time span to a single load. How clothes washer events are aggregated to loads has a significant impact on the statistics reported.

For this study, we did not aggregate clothes washer cycles (i.e., we identified individual clothes washer cycles but did not aggregate them into multi-cycle loads). The average water consumption per clothes washer cycle was 60 L. If we assume a load consists of one wash and one rinse cycle, the average (120 L) is close to the 117 L reported by DeOreo et al. (2019). Figure. A5 shows the volume distribution of clothes washer events for all sites. We observed large variability in the volume used in clothes washer events at most sites, which we attribute to different load sizes and clothes washer settings. Site 24 had 2 clothes washers, whereas all other sites had a single clothes washer. Sites with a large number of events with similar volumes (e.g., sites 19, 20, 29) are likely doing laundries without constantly modifying the appliance settings. Site 32 had the highest volume per clothes washer event values among all sites (Figure. A5) and ranked fifth overall for average per capita daily clothes washer use.

Bathtubs

Bathtub events can use up to 265 L of water, and the time at which the drain is plugged (before or after the temperature is adjusted) can increase this volume (EPA 2021). Figure. A4 shows that bathtub events were not found at 11 sites (35% of participating homes). The average volume used in bathtub events among the remaining participants was 77 L, similar to the 76 L per bath found on the REUS study (DeOreo et al. 2016). The average flow rate at which bathtubs were filled was 14.5 L/min, and the average duration of these events was 5.6 minutes.

REFERENCES

Beal, C., Stewart, R.A. 2011. “South East Queensland Residential End Use Study: Final Report”. Urban Water Security Research Alliance. Accessed October 15, 2021. http://www.urbanwateralliance.org.au/publications/UWSRA-tr47.pdf

DeOreo, W.B., Mayer, P.W., Dziegielewski, B., Kiefer, J., Foundation, W.R. 2016. “Residential End Uses of Water, Version 2”. Water Research Foundation.

EPA. (2021). “The WaterSense Current: Summer 2017”. Accessed May 3, 2021. https://www.epa.gov/watersense/watersense-current-summer-2017

Nguyen, K.A., Stewart, R.A., Zhang, H., Sahin, O. 2018. “An adaptive model for the autonomous monitoring and management of water end use”. Smart Water 3, 5. https://doi.org/10.1186/s40713-018-0012-7

Tables

Table A.1. Indoor per capita end use expressed in liters per capita a day (LPCD) and percent of indoor use by end use.

End Use	Average per capita use (LPCD)	Percent of indoor water use
Shower	54.3	31.2%
Toilet	44.6	25.6%
Faucet	32.4	18.6%
Clothes Washer	24.1	13.62%
Bathtub	7.72	4.44%
Unclassified	8.23	4.73%
Unknown	2.62	1.51%

Figures

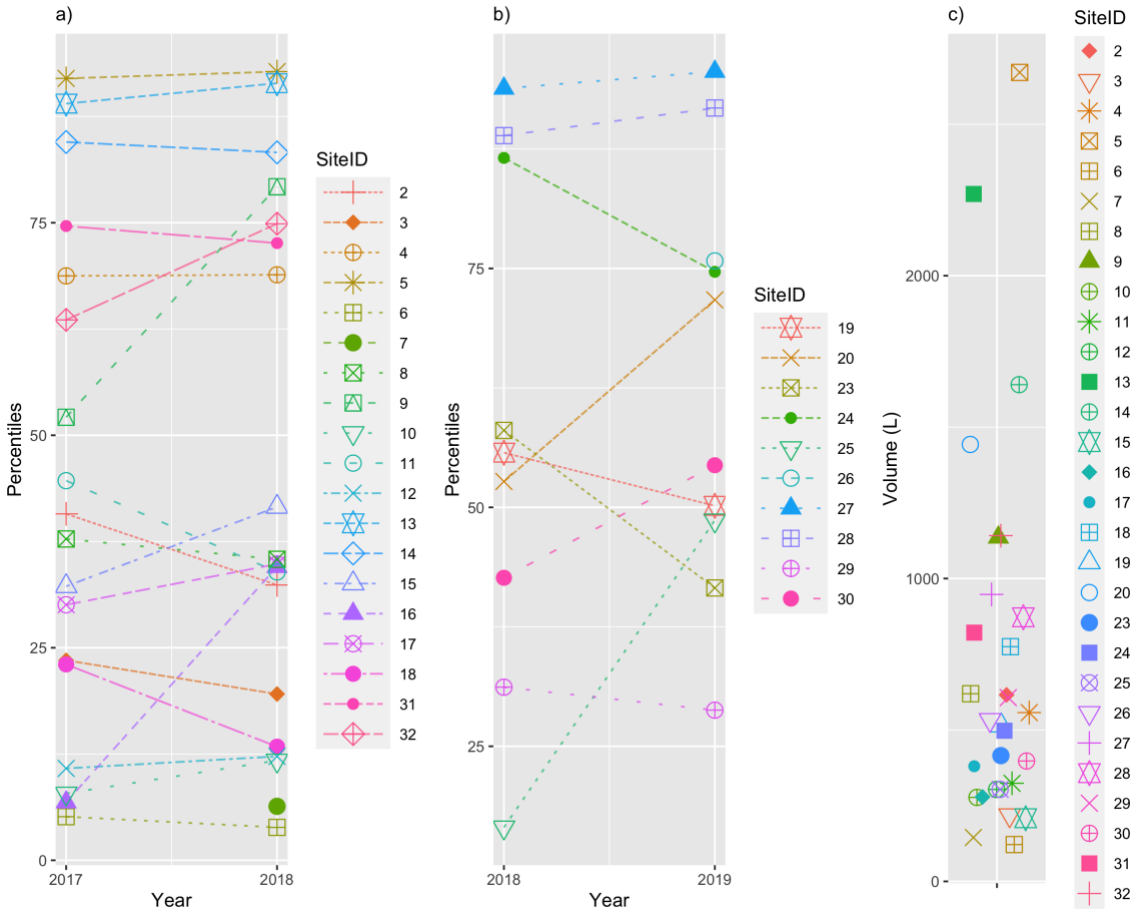


Figure A.1. Annual water use ranking of the participants in the high-temporal resolution study in a) Logan (2017-2018) and b) Providence (2018-2019). Panel c) shows average per capita daily water use volume, in L, for all participants computed from monthly records. Participants for which we had less than one year of monthly billing data (sites 21 and 22 who moved during the study) were removed from all plots.

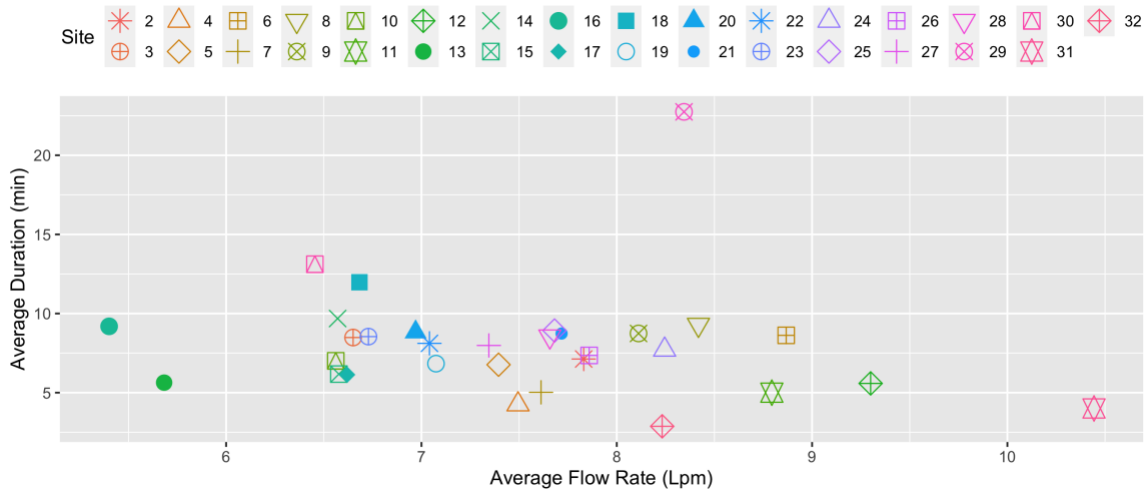


Figure A.2. Average flow rate and duration of shower events.

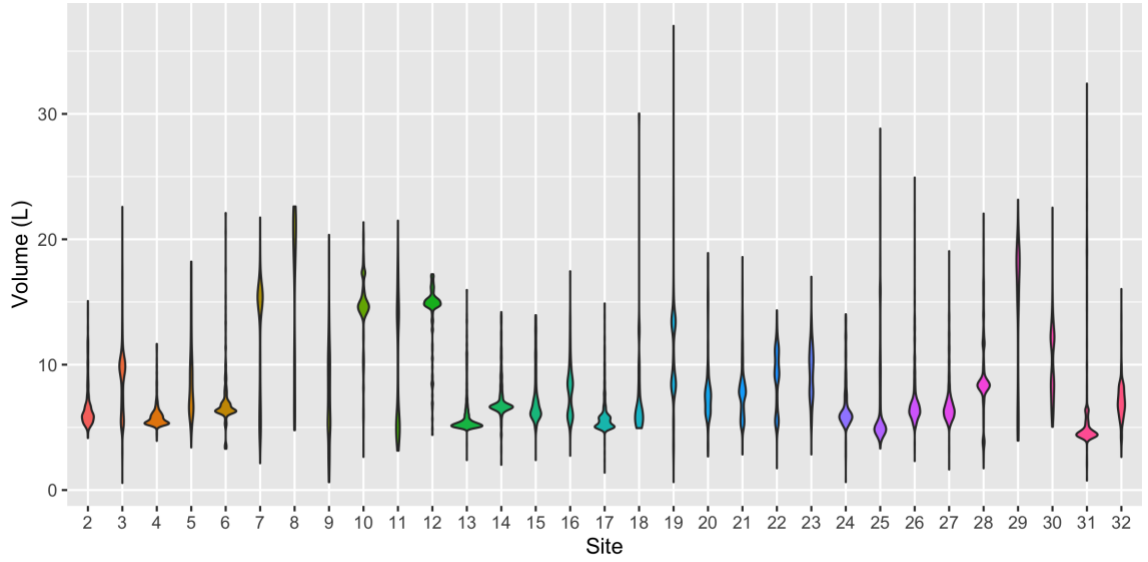


Figure A.3. Volume distribution of toilet flush events for all sites.

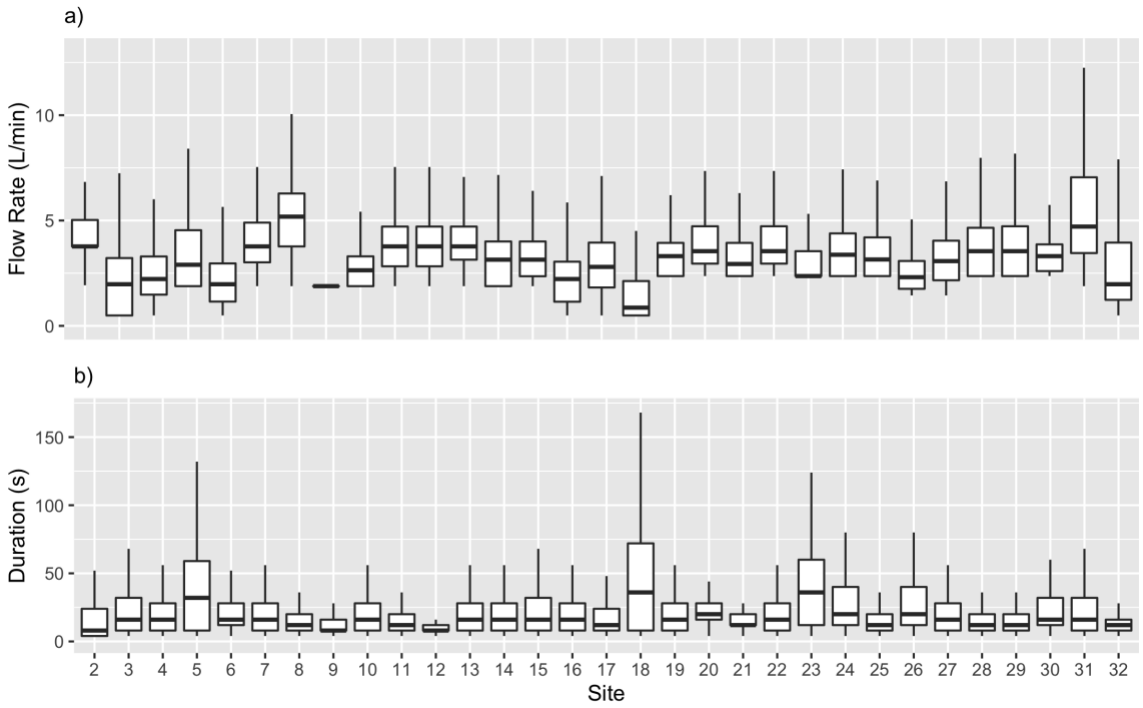


Figure A.4. Boxplots of flow rate (a) and duration (b) of faucet events across all participant sites. Outliers were removed for visualization purposes.

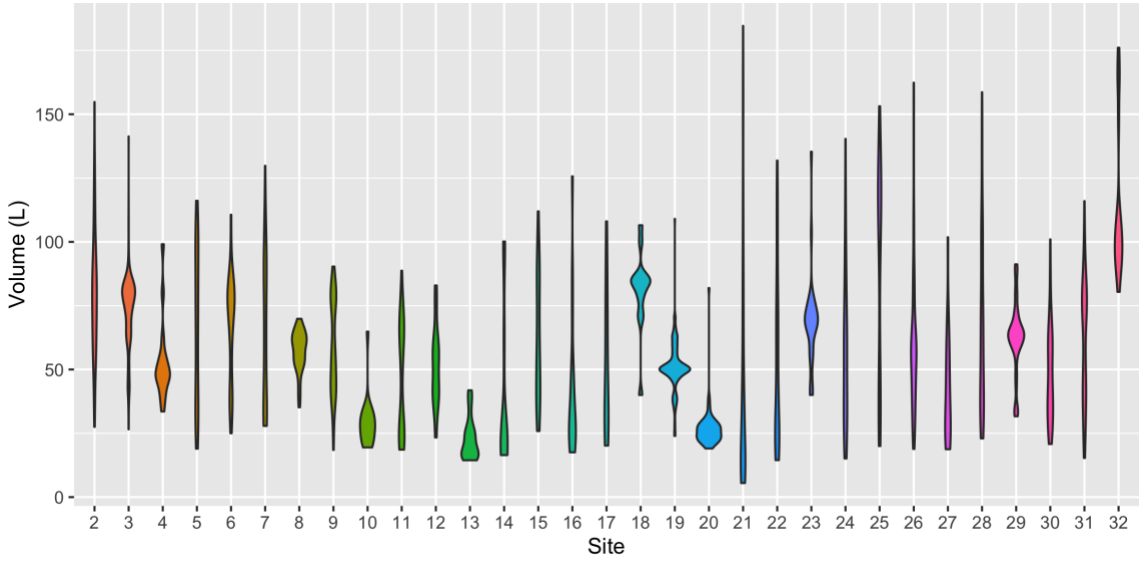


Figure A.5. Volume distribution of clothes washer events for all sites.

CURRICULUM VITAE

Camilo J. Bastidas Pacheco
Department of Civil and Environmental Engineering
Utah Water Research Laboratory
Utah State University, 8200 Old Main Hill, Logan, UT 84322-8200 USA
Phone: (435) 754-5722
Email: camilo.bastidas@usu.edu

Education

Ph.D. Civil and Environmental Engineering 2021
Utah State University, Logan, UT.
Dissertation: Advancing Understanding of Residential Water Use Via Low Cost, Open Source, Smart Metering Infrastructure.
Advisor: Jeffery S. Horsburgh.

M.S. Water Management and Hydrological Science 2017
Texas A&M University, College Station, TX.
Thesis: Characterizing Residential Water Use in College Station and Assessing the Effectiveness of Educational Programs to Reduce Outdoor Water Use.
Advisor: Ronald A. Kaiser.

M.S. Statistics (All but thesis) 2010
Simon Bolivar University, Caracas, Venezuela.

B.S. Hydrometeorological Engineering 2008
Central University of Venezuela, Caracas, Venezuela.

Research Interests

- Water resources management
- Water supply and demand
- Hydroclimate analysis
- Availability and access to water
- Hydroinformatics
- Hydrological processes
- Data science
- Risk assessment and management

Professional, Research, and Teaching Experience

Graduate Research Assistant 01/2018 - 12/2021
Utah Water Research Laboratory, Utah State University. Logan, Utah, USA.

- Designed, developed, calibrated, and implemented smart metering devices and cyberinfrastructure to collect, manage, and analyze high-temporal resolution residential water use data.
- Conducted a residential end uses of water study in two cities in Cache Valley, Utah.

HECRAS Modeling and Bathymetry Specialist 01/2020 - 02/2020
Food and Agriculture Organization of the United Nations. Kabul, Afghanistan.

- Provided training in multiple hydrological GIS applications, field data collection methods, analysis of hydrological extreme events and HECRAS dam-break modeling in a local reservoir.

HECRAS Modeling Specialist 07/2018 - 08/2018

Food and Agriculture Organization of the United Nations. Kabul, Afghanistan.

- Delineated flood hazard maps for dam-break scenarios in two important dams in the country and provided training to local experts.
- Analyzed a glacial lake outburst flood (GLOF) event and drafted recommendations for risk monitoring in a GLOF-prone region.

Hydrology Specialist 07/2016 - 06/2017

United Nations Office for Project Services. Guadalajara, Mexico.

- Collected, validated and analyzed hydrometeorological data from local monitoring networks and remote sensing sources.
- Supervised and executed fieldwork; inspection of automatic weather stations and stream gauges, streamflow measurement, water quality assessment, and bathymetric surveys of water reservoirs.
- Applied and interpreted the WEAP hydrological model and its results. Analyzed and defined different scenarios for the water balances that represent current and future scenarios in the area.

Graduate Research Assistant 11/2014 - 07/2016

Texas A&M AgriLife / Texas A&M University. College Station, Texas, USA.

- Estimated individual landscape water requirements based on near real-time meteorological data.
- Analyzed residential water use, designed and evaluated the results of educational interventions to reduce outdoor water use in College Station, Texas.

Senior Analyst of Effluents 06/2014 - 08/2014

Petroleum of Venezuela S.A. Caracas, Venezuela.

- Evaluated the state of the tools and methods user for effluent discharge measurements in the industry and defined technical recommendations to their advancement.

Adjunct Instructor 10/2010 - 04/2012

Maritime University of the Caribbean / Environmental Engineering. Catia la mar, Venezuela.

Courses taught:

- Physical Hydrology
- Atmospheric Science

Engineer 03/2009 - 06/2014

National Institute of Hydrology and Meteorology. Caracas, Venezuela.

- Conducted and supervised hydrometeorological field campaigns; bathymetry, surveying, streamflow measurements, water quality sampling, aquifer tests, and risk assessments.
- Designed and installed automated networks of hydro-climatological observations, early warning systems, and provided training for their operations and management.

- Developed hydrological and hydrodynamic models to assist in the planning and management of water resources.
- Coordinated hydro-electrical potential evaluation strategies, cloud seeding campaigns, and hydro-meteorological risk mapping efforts.
- Developed different levels of quality-controlled hydro-climatological databases for final users.

Analyst 09/2008 - 03/2009

National Laboratory of Hydraulics. Caracas, Venezuela.

- Calibrated a hydrological model on multiple watersheds to evaluate the availability of water resources in a data-scarce context.

Intern Summer 2007

Ministry of Environment. Caracas, Venezuela.

- Elaborated a water balance for the state of Amazonas, coupling it with the water balance for the nation, part of the UNESCO Hydrological International Program – Latin America and the Caribbean.

Publications

Journal Papers - Published

Bastidas Pacheco, C.J., Brewer, J., Horsburgh, J.S, Caraballo J., (2021). An open source cyberinfrastructure for collecting, processing, storing and accessing high temporal resolution residential water use data. https://doi.org/10.1016/j.envsoft.2021.105137.

Bastidas Pacheco, C.J., Horsburgh, J.S., Tracy, R.J., (2020). A low-cost, open source monitoring system for collecting high-resolution water use data on magnetically-driven residential water meters. Sensors, https://doi.org/10.3390/s20133655.

Journal Papers - in Review:

Bastidas Pacheco, C.J, Horsburgh, Jeffery S., Attallah, Nour A., (2022) Variability in Consumption and End Uses of Water for Residential Users in Logan and Providence, Utah, USA

Attallah N, Horsburgh J.S., Bastidas Pacheco, C.J., (2022). Tools for Evaluating, Developing, and Testing Water End Use Disaggregation Algorithms.

Theses

Bastidas Pacheco, C.J., (2018). Characterizing Residential Water Use in College Station and Assessing the Effectiveness of Educational Programs to Reduce Outdoor Water Use. Master's thesis, Texas A&M University. Available electronically from http://hdl.handle.net/1969.1/173299

Hardware, software and data

Bastidas Pacheco, C. J., Atallah, N., Horsburgh, J. S. (2021). High Resolution Residential Water Use Data in Cache County, Utah, USA, HydroShare, http://www.hydroshare.org/resource/0b72cddfc51c45b188e0e6cd8927227e

Bastidas Pacheco, C. J., Horsburgh, J. S (2021). Supporting data and tools for "Variability in Consumption and End Uses of Water for Residential Users in Logan and Providence, Utah, USA", HydroShare, http://www.hydroshare.org/resource/379d9e7037f04478a99d5aec22e841e6

Bastidas Pacheco, C. J., Horsburgh, J. S. (2021). Standardized Monthly Water Use Data for Logan and Providence Cities, Utah, USA., HydroShare, http://www.hydroshare.org/resource/16c2d60eb6c34d6b95e5d4dbbb4653ef

Bastidas Pacheco, C. J., J. S. Horsburgh, J. Caraballo, N. Attallah (2021). Supporting data and tools for "An open source cyberinfrastructure for collecting, processing, storing and accessing high temporal resolution residential water use data", HydroShare, https://doi.org/10.4211/hs.aaa7246437144f2390411ef9f2f4ebd0

Attallah, N., Bastidas Pacheco, C. J. (2021). Supporting data and tools for "Tools for Evaluating, Developing, and Testing Water End Use Disaggregation Algorithms"http://www.hydroshare.org/resource/3143b3b1bdff48e0aaebcb4aedf02feb

Bastidas Pacheco, C.J., Horsburgh, J. S. (2020). Supporting data for "A low-cost, open source, monitoring system for collecting high-resolution water use data on magnetically-driven residential water meters ", HydroShare, https://doi.org/10.4211/hs.4de42db6485f47b290bd9e17b017bb51

Horsburgh, J.S.; Tracy, J.; Bastidas Pacheco, C.J. (2020) UCHIC/CIWS-MWM-Logger: Version 1.1.0. https://doi.org–/10.5281/zenodo.3832260.

Skills

Languages: English (fluent), Spanish (native), French (B - intermediate).

Technology / Software: R, Python, ArcGIS, HECRAS, HECHMS, SWAT, WEAP, SQL, Linux, InfluxDB, databasing.

Professional Meetings Presentations

Oral Presentations

Bastidas Pacheco, C. J., Horsburgh, J. S., Brewer, J. C., Tracy, R. J., and Caraballo, J.: Advancing open source cyberinfrastructure for collecting, processing, storing and accessing high temporal resolution residential water use data, EGU General Assembly 2021, 19–30 Apr 2021, EGU21-6031, https://doi.org/10.5194/egusphere-egu21-6031, 2021.

Attallah, N., Horsburgh, J., and Bastidas Pacheco, C.J.: Advancing the cyberinfrastructure for smart water metering: A new open source water end use disaggregation algorithm, EGU General Assembly 2021, 19–30 Apr 2021, EGU21-3347, https://doi.org/10.5194/egusphere-egu21-3347, 2021.

Horsburgh, J. S., Bastidas Pacheco, C.J., Atallah, N., Brewer, J., Consalvo, P., Vause, N., Consalvo, P., Whitfield, T., Carmellini, A., Tracy, J. (2019). Cyberinfrastructure for Intelligent Water Supply: Measuring Water Use, Conservation, and Socio-Demographic Differences Using an Inexpensive, High

Frequency Metering System, 2019 CUAHSI Hydroinformatics Conference, Provo, UT, 29-31 July.

Bastidas Pacheco, C.J., Horsburgh, J. S. (2018). Advancing Understanding of Residential Water Use via Low Cost, Open Source Smart Metering infrastructure, 2018 Water Smart Innovations Conference, Las Vegas, NV, 1-5 October.

Posters

Bastidas Pacheco, C.J., Horsburgh, J. S., Tracy, J. (2019). A low-cost, open source, monitoring system for collecting high-resolution water use data on positive displacement residential water meters, 2019 CUAHSI Hydroinformatics Conference, Provo, UT, 29-31 July.

Bastidas Pacheco, C.J. (2015). Evaluating Educational Strategies to Reduce Outdoor Water Use, 2015, Water Smart Innovations Conference, Las Vegas, NV, 07-10 October.

Awards and Honors

Good Neighbor Scholarship, for the academic year 2015-2016. Texas A&M University.

Honorable Mention, to the undergraduate special degree work entitled “Flood Hazard Map for the Manzanares River Basin”.

Professional Affiliations

American Geophysical Union.

American Water Works Association.

American Society of Civil Engineers.

European Geophysical Union