ENVIRONMENTAL RESEARCH
INFRASTRUCTURE AND SUSTAINABILITY

ACCEPTED MANUSCRIPT • OPEN ACCESS

Is smart water meter temporal resolution a limiting factor to residential water end-use classification? A quantitative experimental analysis

To cite this article before publication: Zahra Heydari *et al* 2022 *Environ. Res.: Infrastruct. Sustain.* in press https://doi.org/10.1088/2634-4505/ac8a6b

Is smart water meter temporal resolution a limiting factor to residential water end-use classification? A quantitative experimental analysis

Zahra Heydari[1], Andrea Cominola[2,3], and Ashlynn S. Stillwell[1]

[1]Civil and Environmental Engineering, University of Illinois Urbana-Champaign; [2]Chair of Smart Water Networks, Technische Universität Berlin, Berlin, Germany; [3]Einstein Center Digital Future, Berlin, Germany.

Abstract

Water monitoring in households provides occupants and utilities with key information to support water conservation and efficiency in the residential sector. High costs, intrusiveness, and practical complexity limit appliance-level monitoring via sub-meters on every water-consuming end use in households. Non-intrusive machine learning methods have emerged as promising techniques to analyze observed data collected by a single meter at the inlet of the house and estimated the disaggregated contribution of each water end use. While fine temporal resolution data allow for more accurate end-use disaggregation, there is an inevitable increase in the amount of data that needs to be stored and analyzed. To explore this tradeoff and advance previous studies based on synthetic data, we first collected 1-second resolution indoor water use data from a residential single-point smart water metering system installed at a 4-person household, as well as ground-truth end-use labels based on a water diary recorded over a 4-week study period. Second, we trained a supervised machine learning model (random forest classifier) to classify six water end use categories across different temporal resolutions and two different model calibration scenarios. Finally, we evaluated the results based on three different performance metrics (micro, weighted, and macro F1 scores). Our findings show that data collected at 1- to 5-second intervals allow for better end-use classification (weighted F-score higher than 0.85), particularly for toilet events; however, certain water end uses (e.g., shower and washing machine events) can still be predicted with acceptable accuracy even at coarser resolutions, up to 1 minute, provided that these end use categories are well represented in the training dataset. Overall, our study provides insights for further water sustainability research and widespread deployment of smart water meters.

Keywords: smart water meter, temporal resolution, residential water use, water sustainability, supervised machine learning

1. Introduction

Strong emphasis on sustainability in water use has been increasingly brought to light by growing population and urbanization (Cosgrove and Loucks 2015), coupled with climate change impacts on water resources (Jabaloyes et al. 2018; Karamouz and Heydari 2020). With existing limitations on water resource availability, new developments to increase water storage and supply are often physically or economically constrained. Therefore, better management of existing water resources has become an issue of paramount importance (Mazzoni et al. 2021). Public utilities are now investing significant resources and efforts in the development and implementation of water management strategies, both on the supply and the demand side, to ensure future water security (Jain and Ormsbee 2002; Herrera et al. 2010). On the demand side, these strategies include water saving technologies, new water policy regulations, rebate programs for water-efficient devices, leakage management, and source substitution (e.g., replacing non-potable end-uses with grey, recycled, or harvested rainwater (Dixon et al. 1999)) (Gleick et al. 2003; Inman and Jeffrey, 2006; Stewart et al. 2013; Cominola et al. 2015; Ntuli and Abu-Mahfouz 2016).

Beside their direct effect on water resources, residential water conservation and efficiency strategies can help save water-related energy required for water treatment, distribution, and heating (Srinivasan et al. 2011). Residential end uses are responsible for more than 70% of all water-related energy use (Escriva-Bou et al. 2018). However, the effectiveness of these measures hinges on an accurate estimate of water demand from detailed understanding of how and when water is used in the residential sector. Access to high resolution water consumption data can help improve our knowledge of water demand, identify specific fixture/appliance end uses (e.g., toilet, shower, washing machine, outdoor irrigation), or detect anomalies, such as leaks (Luciani et al. 2019). Smart water meters, which can provide the fine resolution data necessary to discern end uses, have been proven essential in supporting water conservation and efficiency measures in practice (Britton et al., 2008).

Conventional residential water meters typically collect coarse resolution data and require manual readings, limiting the understanding of household-scale water use characteristics and its patterns in time. Conversely, smart (or digital) water meters enable the collection and automated reporting of fine resolution water use data, thereby allowing planners and utilities to better understand demand patterns and enact management strategies. Smart metering can help the development of accurate demand characterization and forecasts and, hence, improve the operation and long-term planning of water supply and distribution systems (Stewart et al. 2018), or promote durable conservation behaviors (Cominola et al., 2021). In addition, detailed knowledge about water consumption at the household level can also translate into financial savings for home occupants, especially when complemented with information about individual end uses (e.g., Blokker et al. 2010).

Obtaining information on residential end uses is not a trivial problem. Information about residential water demand at the end-use level could, in principle, be obtained through direct measurements via intrusive monitoring, i.e., by installing sub-meters at all household end uses. However, this approach is often practically or economically infeasible from a utility perspective and would likely be rejected by home occupants due to its intrusive nature. Instead, water utilities are increasingly installing residential smart water meters that collect fine resolution water consumption data at the service line or entrance into the home, providing aggregate water data, which are so far primarily used for billing purposes (Fogarty and Hudson 2006; Froehlich et al. 2009). Similarly to previous experiences in the electricity sector, limits to directly collecting water-use data at the residential end uses has motivated the development of several non-intrusive disaggregation approaches, which have the advantage of allowing the decomposition of a signal measured at the household level (i.e., aggregate water use) into the individual contribution of each end use (Cominola et al. 2017; Di Mauro et al. 2020; Bethke et al. 2021).

Several state-of-the-art disaggregation techniques require additional sensing on the premise plumbing infrastructure and/or a manual characterization of each end use (Fogarty and Hudson 2006; Kim et al. 2008). These techniques can be intrusive, expensive, and time consuming, thus they are not easy to develop or replicate at large scales (Froehlich et al. 2009, 2011; Srinivasan et al. 2011; Ellert et al. 2015; Ntuli and Abu-Mahfouz). Other disaggregation techniques use only flow (or volume) data collected at the household water inlet point. They can classify end uses in a non-intrusive way, with the accuracy of results varying across different data sampling temporal resolutions (e.g., 1-10 seconds vs. minutes; Clifford et al. 2018; Vitter and Webber 2018). Understanding the tradeoff between the value of the information provided by fine-resolution data and the economic and operational costs of the metering system is crucial to inform the design of future metering networks and associated analytics to facilitate customer data interpretation.

The availability of fine-resolution smart meter data generates several opportunities for advancing water demand management. Sub-minute sampling resolution is essential for most water end-use disaggregation algorithms to provide a reliable categorization of household level water use into different fixtures/appliances (e.g., shower, toilet, dishwasher, etc.) (Willis et al. 2010; Nguyen et al. 2013; Abdallah and Rosenberg, 2014; Horsburgh et al. 2017; Cominola et al. 2018). However, high resolution metering inevitably increases the amount of data the water utility must collect, process, and manage. Sampling at 1-second resolution, for instance, implies replacing the typical 12 monthly readings per user with over 31.5 million data readings. Large amounts of data can compromise hardware and software performance due to issues with meter power sources, battery life, telemetry network capacity, data gaps, and billing software, besides requiring utilities to acquire new skill sets for their employees (Stewart et al., 2010; Suero et al. 2012).

Among the existing literature that has already explored the implication of data sampling resolution on water end use disaggregation (e.g., Wonders et al. 2016), Cominola et al. (2018) developed an analysis based on synthetic time series of water end use generated with STREaM, the STochastic Residential water End-use Model. Their model relied on statistical distributions of end-use characteristics derived from a large dataset of disaggregated water end-uses from over 300 single-family households in nine U.S. cities (DeOreo, 2011). STREaM generated synthetic time series of water end uses with diverse sampling resolutions, which were analyzed with a multi-resolution assessment framework to identify potentially critical thresholds in data resolution for different aspects of information retrieval and demand management. While such studies tend to make up for the shortness of (or limited access to) data through stochastic modeling to generate synthetic disaggregated water use data, a data gap remains with limited availability of ground-truth water end-use observations from real-world data (Di Mauro et al. 2020; Di Mauro et al. 2021).

Here, we address the challenge of testing if and how the theoretical results obtained in the literature on synthetic data change when similar analysis is replicated directly on real-world data. Compared to synthetic data, real-world data might be characterized by higher signal noise, data gaps, and limited dataset size for model calibration. We build on the above modeling efforts through collection and analysis of observed data from a monitored study home in the Midwest United States, exploring the tradeoffs between data sampling resolution and performance in water end-use classification. We examine different data sampling resolutions and explore water end use disaggregation results by aggregating 1-second water consumption data from a 4-person study household to coarser resolutions. We evaluate a set of performance metrics regarding water end-use classification using supervised machine learning informed by ground-truth end-use labels obtained from a water diary recorded over a 4-week study period. Findings from our multi-resolution assessment can support further research and assist utilities in quantifying the benefits associated with second-to-minute data sampling resolutions and the costs of adopting and maintaining fine-resolution metering infrastructures.

The major contributions of this work include:

- Training and testing a water end-use classification model on real-world observation data obtained with a single-point smart meter for a 4-person household coupled with labels from a water diary.
- Quantifying the effects of temporal data sampling resolution on the performance of water end-use classification.
- Analyzing the tradeoff between end-use classification performance and data sampling resolution under two scenarios characterized by different model calibration strategies.

2. Material and Methods

2.1. Metering setup, data collection, and temporal aggregation

In this study, we used data from a single-point smart water metering system installed at a 4-person, single-family, fully-detached residence in the Midwest United States, collecting 1-second resolution flow rate data over a 4-week study period from September 3 to October 1, 2021. Aggregate indoor household water use data were collected from a custom ally® electromagnetic flow meter provided by Sensus, installed on the main water supply pipe into the residents' home. In addition to measuring flow rate (gal/min), the meter also sensed temperature (K) and pressure (psi) data at a 1-second resolution. Although these pressure and temperature data are useful to water system operations, they are not as valuable to demand disaggregation due the large impact the distribution system has on these variables. We validated this assumption through feature analysis based on correlation and data visualization (see Figures S15-S18 in the Supporting Information). Consequently, we focused our analysis on flow rate data. The water meter writes data to a computer running a script that parses the raw data into a suitable format for further analysis. A data acquisition system connected to the water meter parsed the raw data into a timestamped comma separated value (csv) format for further analysis.

To examine the effects of data sampling temporal resolution on water end-use classification, we aggregated the 1-second resolution time series to resolutions of 5 seconds, 10 seconds, 30 seconds, and 1 minute. The 1-minute resolution has been recognized as a critical threshold for certain end-use data analytics in the electricity sector (Armel et al. 2013), Similarly, a previous study based on analysis of synthetic data identified the same threshold as critical for end-use disaggregation in the water sector (Cominola et al. 2018). Here, we test these findings with an experimental study based on real-world data and aim to identify a similar critical data sampling resolution threshold for water end-use classification in the residential sector. Meanwhile, since the study is only based on a 4-person household, we preliminarily compare water consumption patterns with a larger study to ensure the study home is representative of larger scale behavioral patterns.

During the study period, the home occupants manually recorded a water diary of labeled end uses. In this study, six types of indoor water end uses contributed to the total household water demand: faucets, toilets, showers, refrigerator, dishwasher, and washing machine. We used a written water diary over the 4-week study period to collect ground truth end use data for model training and validation. The 4-week period was selected based on previous studies and practicality (Beal et al. 2011; DeOreo et al. 2016; Horsburgh et al. 2017). The water diary included end use labels, start time, and date that were completed by the household occupants. More details about the diary are reported in the Supporting Information, including the water diary template (Figure S19) and an example of completed recordings (Figure S20). This data collection included only factual data such that this work was determined not to meet the definition of human subjects research and, therefore, did not require Institutional Review Board (IRB) approval. Documentation of this IRB decision is available upon request. Limitations that naturally arose during the water diary process were as follows:

- Events that occupants would forget to fill in the diary could not be labeled after the disaggregation of the data.
- Start times listed in the diary would sometimes correspond to events that occurred 1-2 minutes before the reported time, implying that occupants would sometimes fill in the diary after the event.
- Specifically for faucet events, occupants mentioned occasionally leaving the faucet on to avoid reporting multiple events, resulting in long faucet durations that can represent atypical behavior in the model training process.
- The water diary was completed manually and was unreadable for some events.
- Some reported events did not match the information received from the meter.

In addition to these limitations, a power outage created a 2-day data gap in the smart water meter dataset, where the water diary was completed but measured water flow was missing.

2.2. End-use disaggregation

The end-use disaggregation step separates concurrent water use events along with single events, that, aggregated on the axis of time, would give the original time series collected at the single-point residential water meter. While end-use disaggregation and end-use classification sometimes coalesce into one concept in literature, in this study we consider disaggregation as the first step of the end-use classification process (Nguyen et al. 2018). Single events are defined as those that occur in isolation (e.g., dishwasher only), while combined or concurrent events have simultaneous occurrences of water usage (e.g., a toilet flush during a shower). A single *i*-th water use event E_i can be quantitatively characterized by a vector of features F_i, which include values of, e.g., start time, end time, average flow, and volume of that event. Separating and identifying overlapping, or concurrent, water use events is a significant challenge in residential water studies, and the accuracy of existing smart meter disaggregation models decreases significantly when these types of events are encountered (Cominola et al. 2015). Concurrent events occur often, especially during longer duration events such as showers or outdoor irrigation. Thus, disaggregating concurrent events from one another by leveraging information on the characteristics of individual fixtures or by learning the patterns of individual end uses is essential for the purpose of creating a comprehensive water profile for the household.

In this analysis, we used the disaggregation model from Bethke et al. (2021), developed based on Nguyen et al.'s (2013) method of separating concurrent events by calculating the vector gradients of the flow rate data to identify start and end times of overlapping events. Once we separated events with the above disaggregation approach, we manually labeled each appliance/fixture water event based on the water diary and examined the events further with the classification model described below. We repeated this process for every resampled resolution as well as the original 1-second data. At coarser resolutions, the performance of the disaggregation model deteriorated when detecting multiple short duration events happening simultaneously (e.g., hand washing), or short duration events happening on top of a long duration event. Therefore, in addition to naturally having fewer observations at coarser resolutions, the number of events that we were able to match with the diary also decreased (Figure S21).

2.3. End-use classification

After disaggregating the original water use time series, we labeled each event by matching with the water diary. We then trained a random forest (RF) classifier to perform appliance/fixture end-uses classification, using the disaggregated water events resulting from the previous step of end-use disaggregation. The classification algorithm allocated each data point (i.e., a *i*-th water use event E_i) in the dataset to one of the labeled classes, after training on tuples of events and associated features (E_i, F_i).

RF models have been presented by Breiman (2001) as classical ensemble learning algorithms and have shown to be outstanding predictive models in classification tasks (Herrera et al. 2010; James et al.2013). Random forests are built using the same fundamental principles as decision trees and bagging (Bootstrap Aggregation). Bagging introduces randomness into the tree building process by building many trees on random subsets of the training data with replacement; this process is also known as bootstrapping. Bagging then aggregates the predictions across all the trees, which reduces the variance of the overall procedure and improves predictive performance (Géron 2019). However, bagging trees could result in tree correlation that limits the effect of variance reduction. Random forests help reduce variance by injecting more randomness into the training process (Hastie et al. 2009). The random forest algorithm is a bagging algorithm that draws random bootstrap samples from the training set. However, while bagging provides each tree with the full set of features, random forests have a random feature selection that makes trees more independent of each other, which often results in better variance-bias tradeoffs (Table S1) (Friedman et al. 2001; Probst et al. 2019). In this study, the two features of average flow and duration were eventually selected to build the final models, based on the results of our feature importance analysis (Figure S22). Therefore, the search for the split variable was limited to a random subset of the two chosen features. Feature importance was performed based on permutation-based feature importance (Breiman 2001) by evaluating which features contributed the most to the generalization power of the model.

To understand the mechanism used by RF models, it is necessary to understand the construction of classification decision trees. The goal of such a tree is to partition data into small and homogeneous groups. When travelling down the tree, data are split into possible responses called nodes that symbolize the branches of a tree. To perform each partitioning operation, a decision is based on an index (e.g., the Gini index), which allows RF models to partition the nodes of each tree into more homogenous groups that contain a larger proportion of one class in each subsequent node (Kuhn & Johnson, 2013). The Gini index is calculated as in Eq. 1, where C is the total number of classes in the model and p_{nk} is the probability of the occurrence of class k at node n. In this study, six different classes were evaluated based on typical household end uses: faucets (f), toilets (t), showers (s), refrigerator (r), dishwasher (d), and washing machine (w). The sum of all probabilities at a certain node is equal to one (see Eq. 2):

$$\mathrm{G} = \sum_{k \epsilon C} p_{nk}(1 - p_{nk}) \qquad (Eq.1)$$

$$p_t + p_s + p_f + p_r + p_w + p_d = 1 \qquad (Eq.2)$$

Other metrics similar to the Gini index can be used to build decision trees, including cross entropy and misclassification error. However, the Gini coefficient is the most commonly used metric in the literature (James et al. 2013). Moreover, according to Raileanu and Stoffel (2004), the frequency of disagreement of the Gini index and entropy is only 2% of all cases, yet entropy is generally slower to compute because it requires a logarithmic function. For the above reasons, we used the Gini coefficient in this study.

Besides Gini, the RF algorithm involves several other hyperparameters that can be tuned to optimize model performance. While studies have shown that RF models are less sensitive towards tuning than other algorithms such as support vector machines (Probst et al. 2019), modest performance gains can still be valuable considering the limitations that naturally come with a small dataset. Using grid search, we gave ranges to RF hyperparameters to exhaustively try all possible combinations and select the best hyperparameter combination. Minimum sample at each leaf (2- 5), minimum sample split (2, 5, 8, 12) number of sub-features (1, 2), maximum depth (3-10), and the number of trees (10, 20, 50, 100, 200) were initially given to the grid for hyperparameter tuning.

2.4. Model calibration and data sampling resolution scenarios

We considered two scenarios for calibration analysis of the classification model: the "1-second only calibration" (Scenario 1), and "multi-resolution calibration" (Scenario 2).

The "1-second only calibration" (Scenario 1): In this scenario, the RF model was trained only on the measured data at the 1-second resolution. Extended time series of 1-second resolution water use data are not usually available from utility records, but they can be collected in small-scale customized and experimental smart meter installations. With this scenario, we test whether investing efforts and resources in gathering a small model calibration dataset at sub-minute resolution is worth the potential gain of model disaggregation accuracy at coarser resolutions. Our assumption behind this scenario is that the features of water use events can be more accurately learned from data collected at higher resolutions. In the 1-second trained RF model scenario, we split the labeled data into train (70% of the data) and validation (30% of the data) datasets. The validation set was used to assess the model performance on the 1-second trained data. Then, the entire resampled dataset from all other resolutions were separately used as test sets to compare the performance of the model on coarser resolutions.

The "multi-resolution calibration" (Scenario 2): In this scenario, we trained different RF models for each resolution (5 seconds, 10 seconds, 30 seconds, and 1 minute) on their own dataset and compare their performances both with one another and with Scenario 1. In this scenario, we examine the value of retraining the RF model specifically for different temporal resolutions to quantify differences in performance between sampling resolution and, comparatively with Scenario 1, across different model training strategies. To retain the value of limited data and improve generalizability of the models, we implemented a k-fold cross-validation strategy (Hawkins et al. 2003). We thus split the training set into k subsets, called folds, and then iteratively fit the model k times, each time training the data on k-1 folds and evaluating on the remaining single fold (representing the validation data). In this study, we fit the model with $k = 10$. At the end of training, we averaged the performance across all validation folds as the final performance metric for the model.

2.5. Performance metrics

RF is a noise robust technique. However, when considering imbalanced problems, canonical machine learning algorithms generally tend to be biased towards the majority group.

This behavior happens because such algorithms consider the number of objects in each group to be roughly similar (Krawczyk, 2016, Ribeiro and Reynoso-Meza 2020). However, the minority class is often the most important when dealing with skewed distributions, and a performance metric should be chosen in a way to overcome such bias. While we do not directly balance the dataset used in this study because of its limited size, in this analysis we evaluate and compare the model performance using different formulations of the F1-score (FS). Specifically, we compare (i) micro-FS, which is a global metric attributing equal importance to each sample, thus giving emphasis on common labels, (ii) macro-FS, which attributes equal importance to each class, and (iii) weighted FS, which computes the weighted average of the FS values obtained for individual classes. While using these metrics does not solve class imbalance, we examine different F-score formulations to see whether our classifier gets biased towards well represented classes or not.

Micro-FS (usually referred to as simply FS) is a global performance metric that puts more emphasis on the most represented labels in the data set since it gives each sample the same importance. Labels that are underrepresented in the dataset may not be intended to influence the overall micro-FS heavily if the model is performing well on the other more common classes. Micro-FS (Eq. 3) is defined as the harmonic mean of the precision (Eq. 4) and recall (Eq. 5):

$$Micro\text{ - }FS = 2 \times \frac{(precision \times recall)}{(precision + recall)} \qquad (Eq.3)$$

$$Precision = \frac{TP}{TP + FP} \qquad (Eq.4)$$

$$Recall = \frac{TP}{TP + FN} \qquad (Eq.5)$$

where true positives (TP) are the number of correctly classified positive instances, false positives (FP) are the number of negative instances incorrectly classified as positive, and false negatives (FN) are the number of positive instances incorrectly classified as negative.

Macro-FS (short for macro-averaged F1 score) is used to assess the quality of classification in problems with multiple classes. The macro-FS gives the same importance to each class, with low values for models that only perform well on the common classes while performing poorly on the classes with less data. The macro-FS is defined as the mean of class-wise FS in Eq. 6:

$$Macro\text{ - }FS = \frac{1}{N}\sum_{i=1}^{N} FS_i \qquad (Eq.6)$$

where i is the class index and N is the number of classes/labels.

The weighted-average FS (Eq. 7) is calculated by taking the mean of all per-class F1 scores while considering the number of actual occurrences of each class in the dataset.

$$Weighted\text{ - }FS = \frac{1}{H_i}\sum_{i=1}^{N}|i| \times FS_i \; (Eq.7)$$

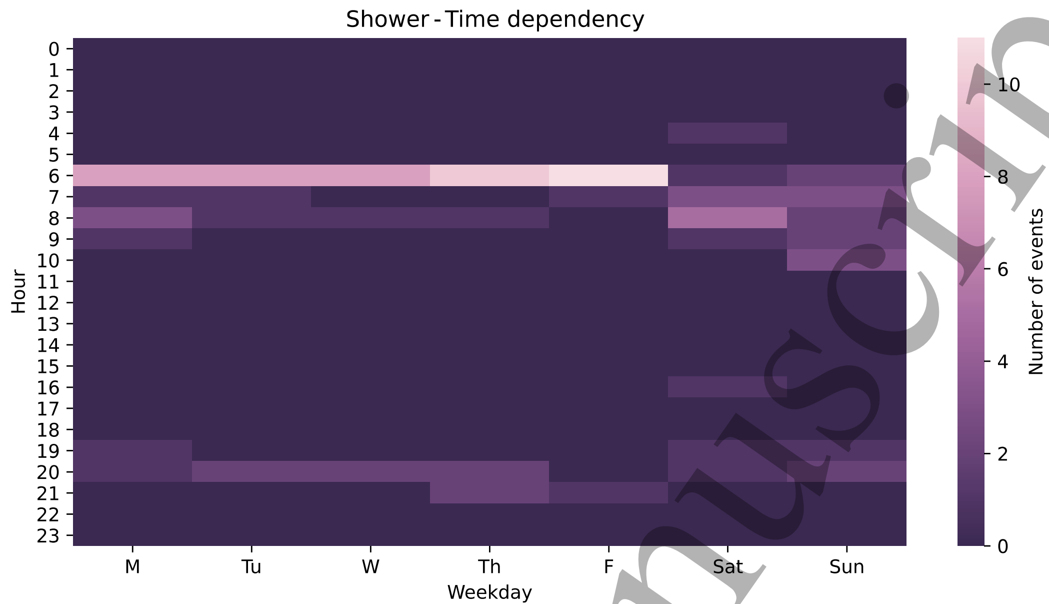
where i and N are as above, and H is the total number of aggregated elements across all classes (Cominola et al. 2021).

The weighted-FS formulation modifies the macro-FS to account for class imbalance, while imbalance is not considered in micro-FS and macro-FS.

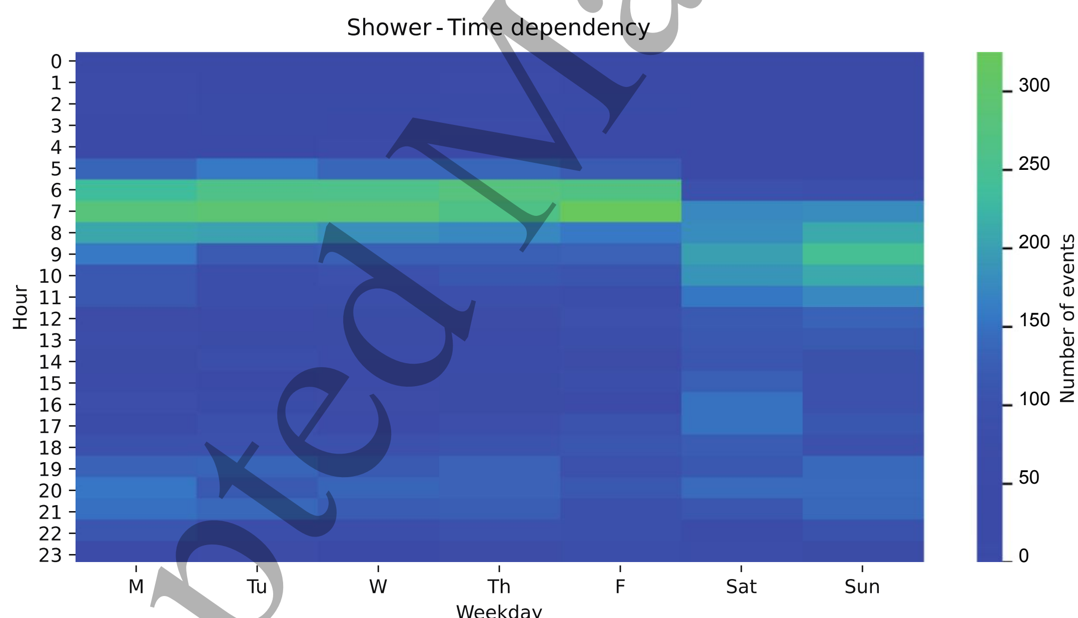
3. Results and discussion

3.1 Data characterization - time of the day visualization

To make sure our study home could be a proper representative of a larger study scale, we initially visualized the time-of-day and day-of-week distribution of three major classes of events (shower, washing machine, and dishwasher) to find regular patterns of consumption similar to those displayed in larger datasets. Much of the occupants' water consumption occurs during typical weekday mornings and evenings. Figure 1(a) depicts shower end use distribution throughout the week and time of the day in our study home. The results show that showers have a more sporadic pattern of use on weekends while during weekdays most of them occur during regular morning and evening peak hours. These behavioral patterns align with the time-of-day and day-of-week distribution of showers reported in an analysis of water end use data gathered for 762 U.S. households (Cominola et al. 2020), shown in Figure 1(b). The time-of-day and day-of-week distribution figures for the washing machine (Figure S1) and dishwasher (Figure S2) are also shown in the Supporting Information, with similar results. Washing machine events are observed mostly during weekends with a wide distribution throughout time of the day, while dishwashers are typically used throughout the week, either mornings or evenings. Comparison of the results show similar patterns between our study home and the larger study of U.S. households used in Cominola et al. (2020), demonstrating the potentially transferrable nature of our study home results. Similar widespread end-use data would help water planners and managers understand water consumption patterns, consumer behavior, and temporal variability. Decreasing consumption during peak time on a widespread scale could contribute to lowering overall peak demand for the local utility and reduce pressure on existing water infrastructure.



(a)



(b)

Figure 1. Time-of-day and day-of-week analysis: (a) Results from shower end use in this study home, 4-week study with 1-second resolution data; (b) Results adapted (with permission) from Cominola et al. (2020) from shower end-uses in 762 U.S. homes, 2-week study period with 10-second resolution data.

3.2. Comparative multi-resolution scenario analysis

The overall RF model performance across different resolutions in both calibration scenarios is presented in Figure 2. Grey lines represent Scenario 1 (1-second only calibration) and blue lines represent Scenario 2 (multi-resolution calibration). The micro-FS, weighted-FS, and macro-FS are represented with dashed, solid, and dotted lines, respectively. We observe that Scenario 2 gives higher performance across different temporal resolutions regardless of the performance metric. For both 1-second and 5-second resolutions, the micro-FS and weighted-FS values are similar: 0.91 and 0.89 for the micro- and weighted-FSs, respectively, at the 1-second resolution, and 0.87 and 0.85 for the micro- and weighted-FSs, respectively, at the 5-second resolution. The macro-FS generally shows the lowest values for all resolutions for both scenarios. We observe a mild decrease in performance metrics with coarser temporal resolutions in Scenario 2, while performance metrics decrease notably for resolutions coarser than 5 seconds in Scenario 1, dropping as low as 0.2 for the 1-minute resolution.

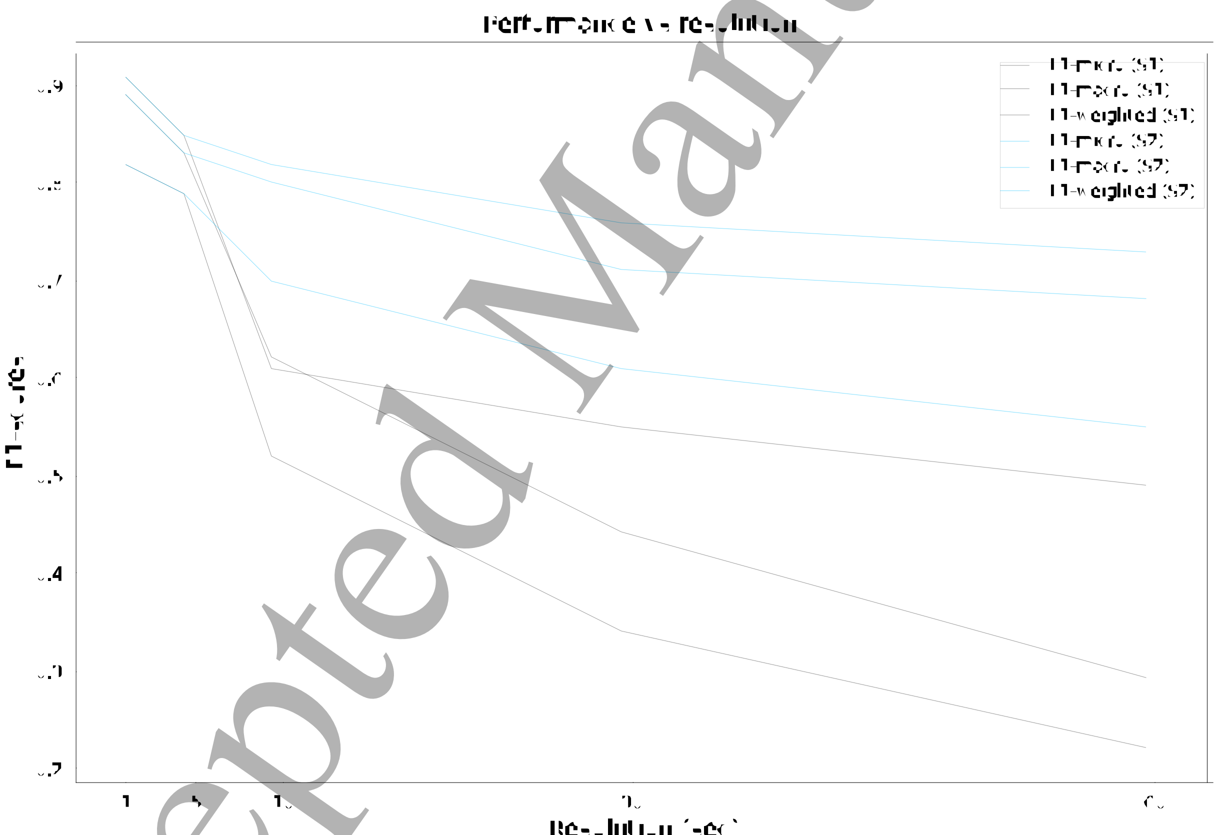
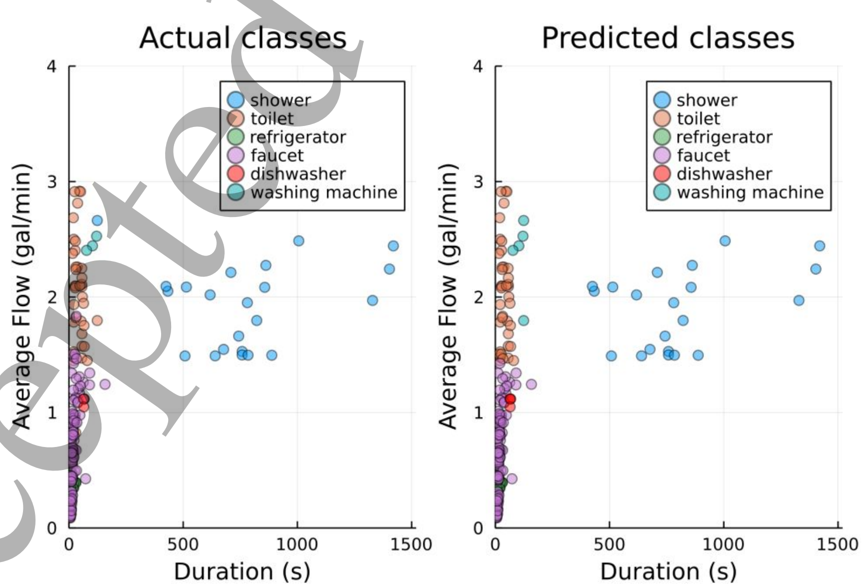


Figure 2. F1-score vs resolution curves for different F1-score formulations for Scenario 1 (grey lines) and Scenario 2 (blue lines). The micro-FS (dashed lines), weighted-FS (solid lines), and macro-FS (dotted lines) are represented.

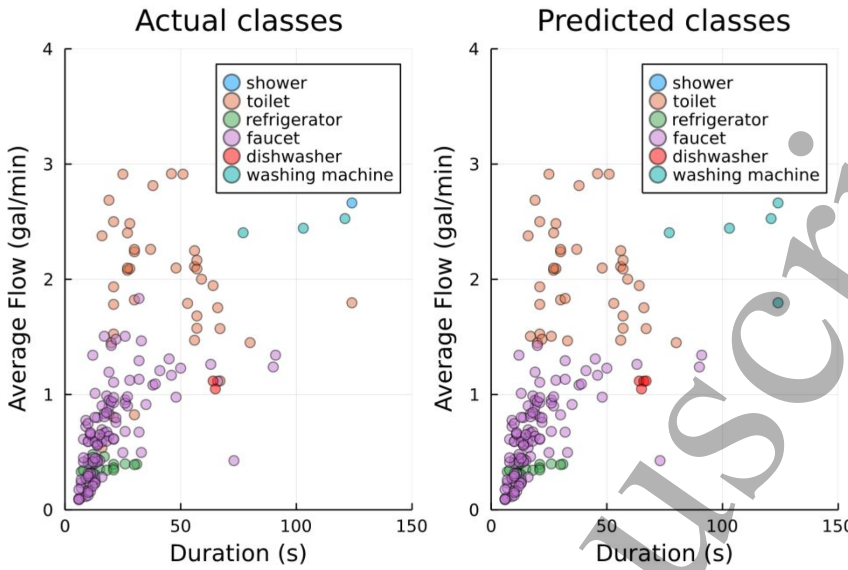
Overall, our results indicate that the RF models learned end use event features better when trained at the same data sampling resolution that they are tasked to use to classify unseen events, provided that a training dataset with labelled events at that resolution is available. If classification models are trained for application on data measured at the same resolution (Scenario 2), those models can perform at an acceptable level of performance even at coarser resolutions, depending on the relative importance of different end use classes. This observation has important implications related to the tradeoffs between fine-resolution data collection and increased data analytics needs. For instance, if a utility wants an estimate of water consumption by the main indoor water uses in households (e.g., toilets and showers), the 1-minute resolution model still provides an acceptable performance (weighted-FS equal to 0.73). This performance is lower than the FS of 0.89 obtained for the 1-second resolution model, but this loss in model accuracy is balanced by the benefit of gathering, storing, and analyzing fewer data observations at the coarser temporal resolution. Conversely, if detailed information on all end uses is required, only the 1-second and 5-second resolutions provide high performance predictions on all end use classes; for less represented end uses, performance is compromised at coarser resolutions.

3.3. Detailed end-use classification results

Our detailed RF model validation results are presented in Figure 3, where the predicted classes (right) are compared to the actual classes (left). Figure 3(a) represents the entire 1-second resolution set of events, while Figure 3(b) zooms in on shorter duration events for clarity. The average flow rate (gal/min) and duration (s) were used as identifying features for our model. Of the total 654 events labeled, we used 196 events as a validation set. The model predicts the test set with an accuracy of 92% and a weighted-FS of 0.89, which is noteworthy given the fact that the training dataset had limited observations in some classes such as dishwasher and washing machine. The model correctly predicts 179 events out of 196 total events of the test set.



(a)



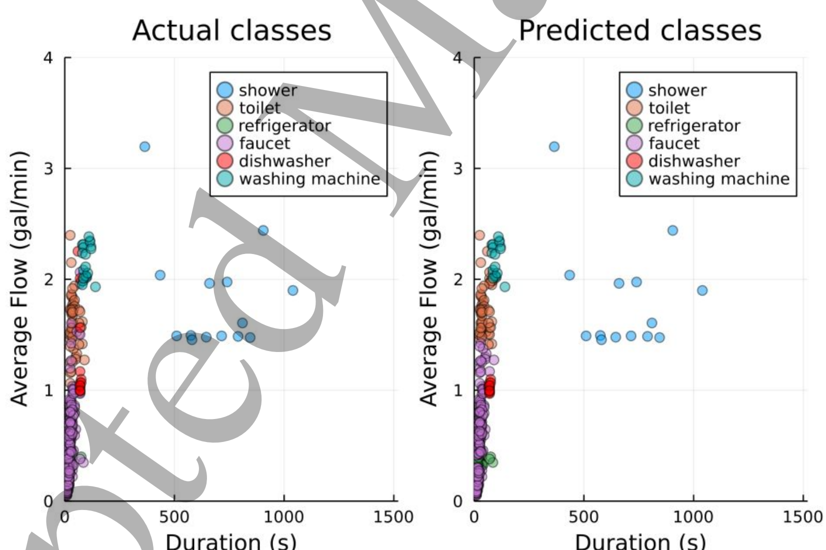
(b)

Figure 3. Actual and predicted water end-use classes. Predicted classes are obtained as results of the RF classifier on the 1-second resolution test set: (a) shows the entire dataset with durations ranging from 1-1500 seconds, and (b) shows the same results zoomed in on events within a duration range of 150 seconds (excluding 23 shower events) for clarity.

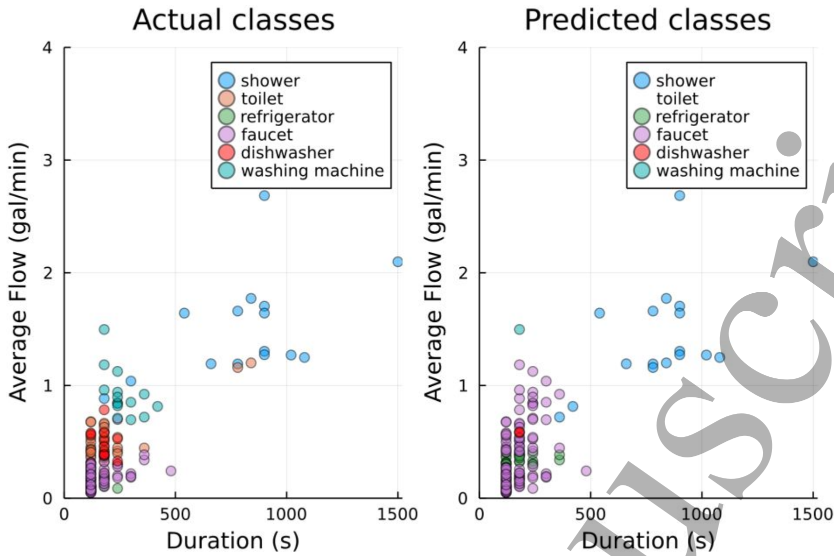
Yet, the high model performance in all classes might overrepresent the overall ability of our RF models to classify unseen end use events. Our results might imply that, due to the fine temporal resolution of the data, the model discerns the constant range of duration and average flow of those end uses with automatic water consumption cycles (e.g., washing machine, dishwasher) and detects them correctly. However, since our study represents a single household only, the model might be overfitting on data from automatic appliances due to the invariance of duration and flow in these specific automatic appliances, thus results on these specific end uses may not be generalizable.

It is important to note that, while individual toilet uses are typically homogeneous in terms of water consumption volume and duration, even considering dual-flush systems, the combination of toilet and bathroom faucet uses are difficult to detect and disaggregate because such uses are often almost simultaneous (e.g., use of toilet and consequent handwashing in a same minute). Although temporal resolutions finer than 1 minute reduce disaggregation errors (Mazzoni et al. 2021), we were not able to disaggregate all toilet events followed by faucets. Rather, we labeled the mentioned events as toilets since we attributed the subsequent faucet use due to the toilet use. As a result, toilets have a wider range of flow and duration, as shown in Figure 3.

Figure 4 shows the classification results for Scenario 1 (1-second only calibration) applied to the resampled 5-second (Figure 4(a)) and 1-minute resolutions (Figure 4(b)), respectively, selected as examples at the two extremes of the considered spectrum of data resolutions. We report our analysis results in both U.S. customary units (gal/min) and SI units (L/min). In comparing different temporal resolutions, coarser resolutions tend to compress data points on the vertical axis (i.e., decrease average event flow) and extend their range on the horizontal axis (i.e., increase event duration) due to temporal averaging. For example, toilet events that originally ranged from 1.7-3 gal/min (6.4-11.4 L/min) average flow in the measured 1-second resolution tend to shift to 1-2.5 gal/min (3.8-9.5 L/min) in the 5-second resolution and decrease further to 0.4-0.8 gal/min (1.5-3 L/min) in the 1-minute resolution. The duration of events increases with coarser temporal resolution to an extent that the total volume of events is the same as to the volume in the original 1-second resolution measurements. The mentioned shifts in values of end-use features leads to decreased model performance with coarser temporal resolutions, up to a point where, as shown in Figure 4(b), the model can no longer detect any toilet events. The model still correctly predicts showers and a few washing machine events at the 1-minute resolution; however, the model application to the 1-minute data predicts most other end uses as a faucet under Scenario 1. Similar Scenario 1 classification results for the 10- and 30-second resolutions are presented in the Supporting Information (Figures S3-S4) along with the zoomed in figures of the 5-second and 1-minute resolutions for a detailed view (Figures S5-S8).



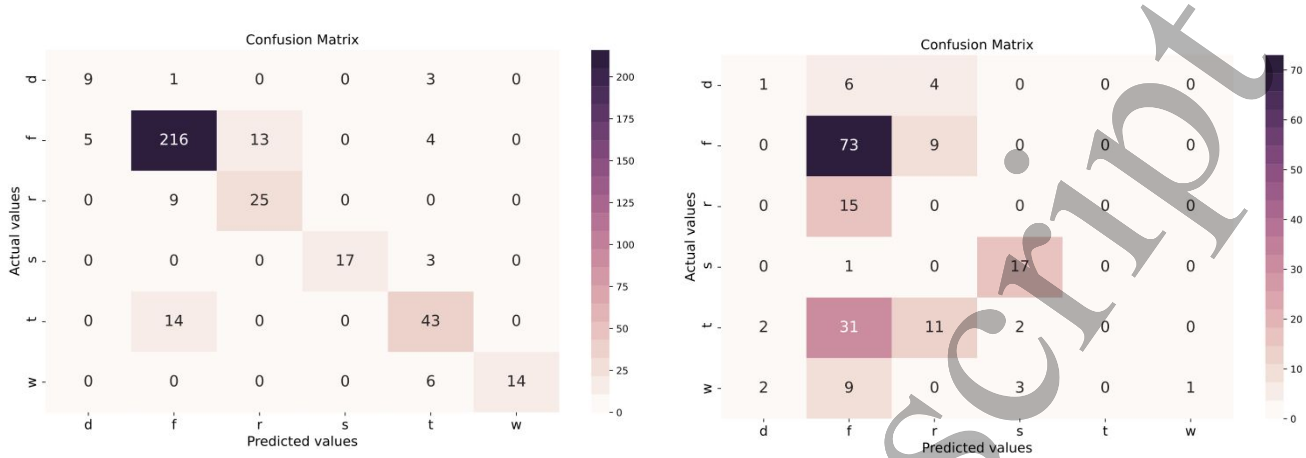
(a)



(b)

Figure 4. Actual and predicted water end-use classes. Predicted classes are obtained as results of the RF classifier trained on the 1-second data (Scenario 1) and applied to the (a) 5-second resolution test set, and (b) 1-minute resolution test set.

Figure 5 shows the confusion matrices of water end use classification across the events of our 4-person study household for Scenario 1. Faucets (f) account for the most frequent end uses, followed by toilets (t). The matrices show the total number of events labeled for each resolution, the actual classes, and the predicted classes by the model. The results for the 5-second resolution show that of 382 total events that we were able to match with the water diary, 324 events were classified correctly (Figure 5(a)). The main misclassifications were in predicting 14 actual toilet end uses as faucets and 4 actual faucet end uses as toilets. This misdetection mostly occurs for data that fall in the area with average flows of 1-1.5 gal/min (3.8-5.7 L/min) and durations of 25-50 seconds (see Figure S5 in the Supporting Information). For the 1-minute resolution (Figure 5(b)), only 187 events had corresponding end uses in the water diary due to disaggregation errors where the model was not able to separate concurrent events because of loss of information that naturally accompanies coarser resolutions. Out of these 187 events, 92 were classified correctly. The classification model predicts 135 events as faucets. While only 73 of these events are actually faucets, they still account for 40% of the prediction accuracy, motivating consideration of F1-score metrics due to the imbalanced dataset.

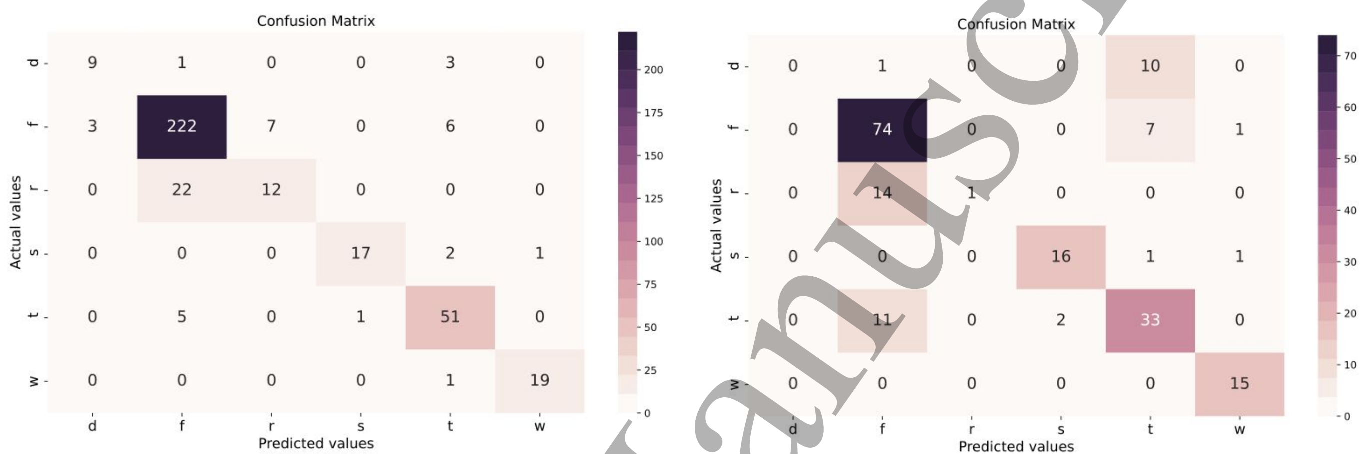


(a) (b)

Figure 5. Confusion matrices for Scenario 1 (1-second trained random forest model): (a) 5-second resolution with 382 total events; (b) 1-minute resolution with 187 total events. Matrix rows show the actual classes and columns show the predicted classes for the following end uses: w (washing machine), s (shower), f (faucet), t (toilet), r (refrigerator), and d (dishwasher). Cell color is proportional to the number of events in that cell.

Figure 6 shows the confusion matrices of end use water consumption for Scenario 2, the multi-resolution calibration, for the 5-second (Figure 6(a)) and 1-minute (Figure 6(b)) resolutions. These results illustrate how the percentage of correct predictions changes from the 5-second resolution to the 1-minute resolution. Compared to Scenario 1 (Figure 5(a)), the 5-second resolution performance in Scenario 2 has either slightly improved or stayed the same, with the exception of refrigerator events (r). The 1-second resolution trained model in Scenario 1 had a better performance in predicting 5-second resolution refrigerator events. The prediction of toilets improved notably from 43 out of 57 to 51 out of 57 events. The main misclassifications were in predicting 5 actual toilet end uses as faucets and 6 actual faucet end uses as toilets. This misdetection mostly occurs for data that fall in the area with average flows of 1-1.5 gal/min (3.8-5.7 L/min) and durations of 25-50 seconds (see Figure S5 in the Supporting Information). Overall, the 5-second resolution has a high performance under both scenarios, with performance metrics slightly less than those of the 1-second resolution (as shown in Figure 2). In the 1-minute resolution, our model correctly predicts 139 of 187 labeled events, having the highest prediction accuracy in washing machine (100%), faucet (90%), and shower (88%) events. These results imply that if any of the aforementioned end uses are of importance, the 1-minute resolution can still be informative.

With further investigation of the diagonals of the confusion matrices, we see how Figure 6(b) has ameliorated in comparison to Figure 5(b), increasing correct predictions from 93 to 139. The 1-minute resolution model is still not able to discern refrigerator faucet events (r) from tap faucet events (f); however, this misclassification is not a critical issue since the refrigerator faucet is a faucet in nature. A noteworthy observation is that although the 1-minute resolution model under Scenario 2 incorrectly classifies one shower and one actual faucet event as a washing machine (i.e., false positive, FP in Eq. 4), it does not label any other actual washing machine event as other events (i.e., false negative, FN in Eq. 5), which leads to a higher Recall in this specific class (100%) than in the 1-second resolution model (86%) (refer to Figure S12) and the 5-second resolution model (95%), with the tradeoff of lower Precision (88% versus 95% in the 1-minute and 5-second resolutions, respectively). Additional confusion matrices at other temporal resolutions are available in Figures S9-S14 of the Supporting Information.



(a) (b)

Figure 6. Confusion matrices for Scenario 2 (random forest model trained at the resampled temporal resolutions): (a) 5-second resolution with 382 total events, and (b) 1-minute resolution with 187 total events. Matrix rows show the actual classes and columns show the predicted classes for the following end uses: w (washing machine), s (shower), f (faucet), t (toilet), r (refrigerator), and d (dishwasher). Cell color is proportional to the number of events in that cell.

In general, misclassifications do not cause significant degradation in predicting total water consumption if they are infrequent and roughly symmetric across the diagonal (Srinivasan et al. 2011). For example, if toilet events are misclassified as faucet events while the same (or nearly the same) number of faucet events are misclassified as toilet events, these misclassifications can cancel out in terms of the accurate total number of events for those classes.

4. Broader implications

Overall, our study contributes to the literature showing that smart water meters provide water utilities with more accurate and less labor-intensive information, enabling better knowledge on changing water demands (Gurung et al. 2015; Stewart et al., 2018). High resolution temporal and spatial water consumption data have undeniable social and technical benefits. Smart metering contributes to more accurate water demand forecasting, demand management strategies, and better informed utility operations and planning strategies (McDaniel and McLaughlin 2009; Cominola et al. 2015; Salomons et al. 2020). Detailed water consumption patterns, which enable researchers to investigate the relationships between human behaviors and the water cycle as part of a broader socio-environmental scale, can be now obtained with advanced analytics, enabled by fast paced computing power improvement and metering technology allowing data collection with unprecedented temporal and spatial granularity (Flint et al. 2017; Zipper et al. 2019). While these advances support greater understanding of water consumption patterns and water-related human behaviors, we also acknowledge that there are potential privacy concerns regarding individuals and communities that need to be addressed and appreciated. Water consumption information transformed from the meter acts as an information side channel (McDaniel and McLaughlin 2009), exposing household habits and behaviors. End-uses like showers and toilets have detectable water consumption signatures, making end use classification information prone to potential privacy abuse. Consequently, well established privacy policies would benefit utilities in appropriate water demand management. Additionally, researchers have an ethical responsibility to protect participant confidentiality.

Recent studies have addressed privacy issues in both the water and energy sectors and presented solutions to overcome privacy related constraints to maximize the potential of granular data (Khurana et al. 2010; Molina-Markham et al. 2010; Gurstein 2011; Amin 2012; Cole and Stewart, 2013; Harter et al. 2013; Sankar et al. 2013; Helveston 2015; Park and Cominola 2020; Salomons et al. 2020). For instance, smart meter data can be used without invading individual privacy by aggregating data to coarser spatial or temporal scales as presented in our study. Nevertheless, as shown in this study, aggregation limits the ability of end-use classification, or any water consumption related research, to explore fine-scale behavioral dynamics for better demand modeling. Therefore, any research intersecting with human behavior should prioritize confidentiality (e.g., via anonymized data collected over a large sample of households) while providing sufficient information to enable future improvements in that field. While the formulation of privacy and security protection strategies is not within the scope of this study, we acknowledge that privacy and security considerations must be addressed and proactively planned for prior to collecting data throughout the research process so that modern metering technologies could be leveraged to their full extent while securing customer privacy (Meyer 2018).

From the findings of this study, we can identify the following limitations and opportunities for future research. First, future studies could focus on assessing how our results generalize when data from a larger household sample or homes from different socio-demographic, geographical, and climate contexts are available. Second, in this study we only considered six classes of indoor water uses from a 4-person household. Further research could include outdoor water use and test end-use disaggregation capabilities on houses with different sizes. Third, as highlighted in the methods, end use datasets are often imbalanced, i.e., the number of events in each end use class might vary substantially. While here we considered class imbalance *a posteriori*, by assessing the disaggregation results with different formulations of the F-score, an alternative approach to be tested when larger datasets are available is to balance the classes *a priori* (i.e., before performing the classification), e.g., by oversampling/undersampling, which would solve the problem of class imbalance. Finally, while here we only considered RF classifiers and a specific approach for disaggregation, future studies could comparatively assess the performance of different models, possibly accounting for multi-class events.

5. Conclusion

In this analysis, we presented a supervised approach to classify residential water consumption end use events and tested it on data collected in a 4-person household through consideration of multiple temporal resolutions by measuring water use data with a 1-second resolution smart water metering system and labeling events based on a water diary for a 4-week study period. We investigated two different scenarios of model calibration in evaluating the effect of temporal resolution on end use classification performance. The first scenario consisted of training a random forest classifier on the original 1-second resolution data only and testing it on other labeled temporal resolution datasets (i.e., 5 seconds, 10 seconds, 30 seconds, 1 minute). In this scenario, our model exhibited high overall performance on the 1-second and 5-second resolution water use events and classified certain classes of end uses with fairly good accuracy for the 10-second resolution. The performance decreased notably for the 30-second and 1-minute resolutions.

The second scenario consisted of training separate models for each temporal resolution using k-fold cross-validation. We saw that coarser temporal resolutions ameliorated in this second scenario, with F1-score performance metrics as high as 0.89 for certain end use classes at the finer resolutions. A weighted F1-score above 0.85 was obtained in this scenario for disaggregation tasks performed at 1- and 5-second resolutions.

Our results reveal detailed information that can help utilities and residents make informed water conservation and efficiency decisions based on detailed knowledge on water demands. The analysis of classification model performance versus temporal resolution considering different F1-scoore formulations provides insight for future water management regarding the selection of an efficient monitoring resolution based on priorities and data management capabilities.

In addition, our approach performed end use disaggregation of data aggregated at different temporal resolutions that are closer to the resolutions of commercial smart water meters (i.e., 1 minute). Thus, while making use of data collected at a finer resolution (e.g., 1 second) might not be available to water utilities due to data management and analysis tradeoffs, we demonstrate possible model extensions to broader and further contexts in the field of residential water demand monitoring.

Ultimately, disaggregating and classifying water events obtained from residential smart water meter data reveals detailed information about how water is consumed within households. Understanding the overall water consumption profile and performance of different resolutions presents opportunities for improved residential water conservation and efficiency and long-term water resource sustainability (Attari 2014; Inskeep and Attari 2014; Horsburgh et al. 2017; Goulas et al. 2022). Our study presents an experimental example of how using smart water meter data can provide end use information to pinpoint opportunities for improved efficiency within residential buildings.

599 **Acknowledgments**

600 Thanks to the study home occupants for their participation in recording the water diary. The custom ally® water meter used in this analysis was provided by Sensus. Thanks to Marie-Philine Becker (Technische Universität Berlin) and Riccardo Taormina (TU Delft) for sharing the results reproduced in Figures 1(b), S1(b), and S2(b). This work was supported by the National Science Foundation, grant CBET-1847404; the opinions, findings, and conclusions or recommendations expressed here are those of the authors and do not necessarily reflect the views of the National Science Foundation.

607 Author contributions: Z.H. analyzed data, created and implemented the classification models and quantified performance metrics, and analyzed and summarized results; A.C. assisted with creating the classification models and performance metrics and contributed to scoping the analysis; A.S.S. formulated the analysis scope, supervised the research, and acquired funding to support the analysis; all authors contributed to writing and reviewed the manuscript.

612 Data availability: The labeled end-use data used in the study are available at https://stillwell.cee.illinois.edu/data/.

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*(3), 288-297.

Amin, S. M. (2012, July). Smart grid security, privacy, and resilient architectures: Opportunities and challenges. In *2012 IEEE Power and Energy Society General Meeting* (pp. 1-2). IEEE.

Armel, K. C., Gupta, A., Shrimali, G., & Albert, A. (2013). Is disaggregation the holy grail of energy efficiency? The case of electricity. *Energy policy*, *52*, 213-234.

Attari, S. Z. (2014). Perceptions of water use. *Proceedings of the National Academy of Sciences*, *111*(14), 5129-5134.

Beal, C. D., Gurung, T. R., & Stewart, R. A. (2016). Demand-side management for supply-side efficiency: Modeling tailored strategies for reducing peak residential water demand. *Sustainable Production and Consumption*, *6*, 1-11.

Beal, C., Stewart, R., Huang, T., & Rey, E. (2011). *South East Queensland residential end use study*. Brisbane, Australia: Urban Water Security Research Alliance.

Bethke, G. M., Cohen, A. R., & Stillwell, A. S. (2021). Emerging investigator series: disaggregating residential sector high-resolution smart water meter data into appliance end-uses with unsupervised machine learning. *Environmental Science: Water Research & Technology*, *7*(3), 487-503.

Blokker, E. J. M., Vreeburg, J. H. G., & Van Dijk, J. C. (2010). Simulating residential water demand with a stochastic end-use model. *Journal of Water Resources Planning and Management*, *136*(1), 19-26.

Boyle, T., Giurco, D., Mukheibir, P., Liu, A., Moy, C., White, S., & Stewart, R. (2013). Intelligent metering for urban water: A review. *Water*, *5*(3), 1052-1081.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Clifford, E., Mulligan, S., Comer, J., & Hannon, L. (2018). Flow-signature analysis of water consumption in nonresidential building water networks using high-resolution and medium-resolution smart meter data: Two case studies. *Water Resources Research*, *54*(1), 88-106.

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*(3), 174-194.

Cominola, A., Becker, M.-P., and Taormina, R.: Benchmarking machine learning algorithms for Non-Intrusive Water Monitoring, EGU General Assembly 2020, Online, 4–8 May 2020, EGU2020-16119, https://doi.org/10.5194/egusphere-egu2020-16119, 2020

Cominola, A., Giuliani, M., Castelletti, A., Fraternali, P., Gonzalez, S. L. H., Herrero, J. C. G., ... & Rizzoli, A. E. (2021). Long-term water conservation is fostered by smart meter-based feedback and digital user engagement. *npj Clean Water*, *4*(1), 1-10.

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.

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.

Cominola, A., Giuliani, M., Piga, D., Castelletti, A., & Rizzoli, A. E. (2017). A hybrid signature-based iterative disaggregation algorithm for non-intrusive load monitoring. *Applied energy*, *185*, 331-344.

DeOreo, W. B. (2011). Analysis of water use in new single family homes. *By Aquacraft. For Salt Lake City Corporation and US EPA*.

DeOreo, W. B., Mayer, P. W., Dziegielewski, B., & Kiefer, J. (2016). *Residential end uses of water, version 2*. Water Research Foundation.

Di Mauro, A., Cominola, A., Castelletti, A., & Di Nardo, A. (2021). Urban Water Consumption at Multiple Spatial and Temporal Scales. A Review of Existing Datasets. *Water*, *13*(1), 36.

Di Mauro, A., Di Nardo, A., Francesco Santonastaso, G., & Venticinque, S. (2020). Development of an IoT System for the Generation of a Database of Residential Water End-Use Consumption Time Series. *Environmental Science Proceedings*, *2*(1), 20.

Escriva-Bou, A., Lund, J. R., & Pulido-Velazquez, M. (2018). Saving energy from urban water demand management. *Water Resources Research*, *54*(7), 4265-4276.

Figueiredo, M., Ribeiro, B., & de Almeida, A. (2013). Electrical signal source separation via nonnegative tensor factorization using on site measurements in a smart home. *IEEE Transactions on Instrumentation and Measurement*, *63*(2), 364-373.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). 4.3: Linear Discriminant Analysis. *The elements of statistical learning*, *1*, 106-119.

Froehlich, J., Larson, E., Saba, E., Campbell, T., Atlas, L., Fogarty, J., & Patel, S. (2011, June). A longitudinal study of pressure sensing to infer real-world water usage events in the home. In *International conference on pervasive computing* (pp. 50-69). Springer, Berlin, Heidelberg.

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. " O'Reilly Media, Inc.".

Goulas, A., Goodwin, D., Shannon, C., Jeffrey, P., & Smith, H. M. Public perceptions of household IoT smart water ‘event’meters in the UK–implications for urban water governance.

Gurstein, M. B. (2011). Open data: Empowering the empowered or effective data use for everyone?. *First Monday*.

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.

Hartter, J., Ryan, S. J., MacKenzie, C. A., Parker, J. N., & Strasser, C. A. (2013). Spatially explicit data: stewardship and ethical challenges in science. *PLoS Biology*, *11*(9), e1001634.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Random forests. In *The elements of statistical learning* (pp. 587-604). Springer, New York, NY.

Hastie, T., Tibshirani, R., Botstein, D., & Brown, P. (2001). Supervised harvesting of expression trees. *Genome Biology*, *2*(1), 1-12.

Hawkins, D. M., Basak, S. C., & Mills, D. (2003). Assessing model fit by cross-validation. *Journal of chemical information and computer sciences*, *43*(2), 579-586.

Helveston, M. N. (2015). Consumer protection in the age of big data. *Wash. UL Rev.*, *93*, 859.

Herrera, M., Torgo, L., Izquierdo, J., & Pérez-García, R. (2010). Predictive models for forecasting hourly urban water demand. *Journal of hydrology*, *387*(1-2), 141-150.

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 modelling & software*, *96*, 83-94.

Inskeep, B. D., & Attari, S. Z. (2014). The water short list: The most effective actions US households can take to curb water use. *Environment: Science and policy for sustainable development*, *56*(4), 4-15.

J. Fogarty, C. Au, and S. Hudson. Sensing from the basement: a feasibility study of unobtrusive and low-cost home activity recognition. In Proceedings of UIST, 2006.

J. Froehlich, E. Larson, T. Campbell, C. Haggerty, J. Fogarty, and S. Patel. Hydrosense: infrastructure-mediated singlepoint sensing of whole-home water activity. In In UbiComp, 2009.

Jain, A., & Ormsbee, L. E. (2002). Short-term water demand forecast modeling techniques—CONVENTIONAL METHODS VERSUS AI. *Journal-American Water Works Association*, *94*(7), 64-72.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer.

Karamouz, M., & Heydari, Z. (2020). Conceptual design framework for coastal flood best management practices. *Journal of Water Resources Planning and Management*, *146*(6), 04020041.

Khurana, H., Hadley, M., Lu, N., & Frincke, D. A. (2010). Smart-grid security issues. *IEEE Security & Privacy*, *8*(1), 81-85.

Kowalski, M., & Marshallsay, D. (2003, April). A system for improved assessment of domestic water use components. In *II International Conference Efficient Use and Management of Urban Water Supply*.

Makonin, S., Popowich, F., Bajić, I. V., Gill, B., & Bartram, L. (2015). Exploiting HMM sparsity to perform online real-time nonintrusive load monitoring. *IEEE Transactions on smart grid*, *7*(6), 2575-2585.

Mantovani, R. G., Horváth, T., Cerri, R., Junior, S. B., Vanschoren, J., & de Carvalho, A. C. P. D. L. F. (2018). An empirical study on hyperparameter tuning of decision trees. *arXiv preprint arXiv:1812.02207*.

Mayer, P. W., DeOreo, W. B., Opitz, E. M., Kiefer, J. C., Davis, W. Y., Dziegielewski, B., & Nelson, J. O. (1999). Residential end uses of water.

Mazzoni, F., Alvisi, S., Franchini, M., Ferraris, M., & Kapelan, Z. (2021). Automated Household Water End-Use Disaggregation through Rule-Based Methodology. *Journal of Water Resources Planning and Management*, *147*(6), 04021024.

McDaniel, P., & McLaughlin, S. (2009). Security and privacy challenges in the smart grid. *IEEE security & privacy*, *7*(3), 75-77.

Meyer, M. N. (2018). Practical tips for ethical data sharing. *Advances in methods and practices in psychological science*, *1*(1), 131-144.

Molina-Markham, A., Shenoy, P., Fu, K., Cecchet, E., & Irwin, D. (2010, November). Private memoirs of a smart meter. In *Proceedings of the 2nd ACM workshop on embedded sensing systems for energy-efficiency in building* (pp. 61-66).

Nguyen, K. A., Zhang, H., & Stewart, R. A. (2013). Development of an intelligent model to categorise residential water end use events. *Journal of hydro-environment research*, *7*(3), 182-201.

Ntuli, N., & Abu-Mahfouz, A. (2016). A simple security architecture for smart water management system. *Procedia Computer Science*, *83*, 1164-1169.

Park, S., & Cominola, A. When Privacy Protection Meets Non-Intrusive Load Monitoring: Trade-off Analysis and Privacy Schemes via Residential Energy Storage.

Pastor-Jabaloyes, L., Arregui, F. J., & Cobacho, R. (2018). Water end use disaggregation based on soft computing techniques. *Water*, *10*(1), 46.

Piga, D., Cominola, A., Giuliani, M., Castelletti, A., & Rizzoli, A. E. (2015). Sparse optimization for automated energy end use disaggregation. *IEEE Transactions on Control Systems Technology*, *24*(3), 1044-1051.

Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: data mining and knowledge discovery*, *9*(3), e1301.

Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: data mining and knowledge discovery*, *9*(3), e1301.

Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: data mining and knowledge discovery*, *9*(3), e1301.

Rahimpour, A., Qi, H., Fugate, D., & Kuruganti, T. (2017). Non-intrusive energy disaggregation using non-negative matrix factorization with sum-to-k constraint. *IEEE Transactions on Power Systems*, *32*(6), 4430-4441.

Raileanu, L. E., & Stoffel, K. (2004). Theoretical comparison between the gini index and information gain criteria. *Annals of Mathematics and Artificial Intelligence*, *41*(1), 77-93.

Ribeiro, V. H. A., & Reynoso-Meza, G. (2020). Ensemble learning by means of a multi-objective optimization design approach for dealing with imbalanced data sets. *Expert Systems with Applications*, *147*, 113232.

Salomons, E., Sela, L., & Housh, M. (2020). Hedging for privacy in smart water meters. *Water Resources Research*, *56*(9), e2020WR027917.

Sankar, L., Rajagopalan, S. R., Mohajer, S., & Poor, H. V. (2012). Smart meter privacy: A theoretical framework. *IEEE Transactions on Smart Grid*, *4*(2), 837-846.

Sønderlund, A. L., Smith, J. R., Hutton, C. J., Kapelan, Z., & Savic, D. (2016). Effectiveness of smart meter-based consumption feedback in curbing household water use: Knowns and unknowns. *Journal of Water Resources Planning and Management*, *142*(12), 04016060.

Srinivasan, V., Stankovic, J., & Whitehouse, K. (2011, November). Watersense: Water flow disaggregation using motion sensors. In *Proceedings of the Third ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings* (pp. 19-24).

Stewart, R. A., Nguyen, K., Beal, C., Zhang, H., Sahin, O., Bertone, E., ... & Kossieris, P. (2018). Integrated intelligent water-energy metering systems and informatics: Visioning a digital multi-utility service provider. *Environmental Modelling & Software*, *105*, 94-117.

Stewart, R. A., Willis, R., Giurco, D., Panuwatwanich, K., & Capati, G. (2010). Web-based knowledge management system: linking smart metering to the future of urban water planning. *Australian Planner*, *47*(2), 66-74.

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*(3), 299-306.

Vitter, J. S., & Webber, M. E. (2018). A non-intrusive approach for classifying residential water events using coincident electricity data. *Environmental modelling & software*, *100*, 302-313.

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*(12), 1117-1127.

Wonders, M., Ghassemlooy, Z., & Hossain, M. A. (2016). Training with synthesised data for disaggregated event classification at the water meter. *Expert Systems with Applications*, *43*, 15-22.

Y. Kim, T. Schmid, Z. Charbiwala, J. Friedman, and M. Srivastava. Nawms: nonintrusive autonomous water monitoring system. In Sensys, 2008.

Zipper, S. C., Stack Whitney, K., Deines, J. M., Befus, K. M., Bhatia, U., Albers, S. J., ... & Schlager, E. (2019). Balancing open science and data privacy in the water sciences. *Water Resources Research*, *55*(7), 5202-5211.