Water Research X 24 (2024) 100244

Contents lists available at ScienceDirect

Water Research X

journal homepage: www.sciencedirect.com/journal/water-research-x

High resolution data visualization and machine learning prediction of free chlorine residual in a green building water system

S. Wei [a], R. Richard [b], D. Hogue [a], I. Mondal [a,c], T. Xu [a], T.H. Boyer [a], K.A. Hamilton [a,c,*]

[a] School of Sustainable Engineering and the Built Environment, Arizona State University, Tempe, AZ 85281, United States
[b] Wilson & Company Engineers, United States
[c] Biodesign Center for Environmental Health Engineering, Arizona State University, Tempe, AZ 85281, United States

ARTICLE INFO

Keywords:
Premise plumbing
Opportunistic pathogens
Machine learning (ML)
Artificial intelligence (AI)

ABSTRACT

People spend most of their time indoors and are exposed to numerous contaminants in the built environment. Water management plans implemented in buildings are designed to manage the risks of preventable diseases caused by drinking water contaminants such as opportunistic pathogens (e.g., *Legionella* spp.), metals, and disinfection by-products (DBPs). However, specialized training required to implement water management plans and heterogeneity in building characteristics limit their widespread adoption. Implementation of machine learning and artificial intelligence (ML/AI) models in building water settings presents an opportunity for faster, more widespread use of data-driven water quality management approaches. We demonstrate the utility of Random Forest and Long Short-Term Memory (LSTM) ML models for predicting a key public health parameter, free chlorine residual, as a function of data collected from building water quality sensors (ORP, pH, conductivity, and temperature) as well as WiFi signals as a proxy for building occupancy and water usage in a "green" Leadership in Energy and Environmental Design (LEED) commercial and institutional building. The models successfully predicted free chlorine residual declines below 0.2 ppm, a common minimum reference level for public health protection in drinking water distribution systems. The predictions were valid up to 5 min in advance, and in some cases reasonably accurate up to 24 h in advance, presenting opportunities for proactive water quality management as part of a sense-analyze-decide framework. An online data dashboard for visualizing water quality in the building is presented, with the potential to link these approaches for real-time water quality management.

1. Introduction

Most children and adults spend 8 h or more per day in commercial and institutional (C&I) buildings (e.g., offices, schools), and even longer in elder care facilities or large apartment complexes (Klepeis et al., 2001). Traditionally, a focus has been placed on managing energy use related to heating, ventilation, and air conditioning within buildings, especially within sustainability programs (Bravo et al., 2020; Park et al., 2020). However, over the past few decades there has been greater visibility of high-profile incidents wherein water quality within buildings puts people at risk from lead, copper, waterborne pathogens like *Legionella* spp., disinfection by-products (DBPs), or unaesthetic water (Abokifa et al., 2020; Allen et al., 2017; Baum et al., 2016). *Legionella* spp. is the also a leading cause of drinking water-associated outbreaks in the United States and is of epidemiologic importance globally (CDC, 2018; WHO, 2007).

"Green" buildings are beneficial for sustainability but can have water quality issues due to stagnant water and low-flow fixtures, although causal linkages are difficult to assess (Logan-Jackson et al., 2023; Rhoads et al., 2016; Rhoads and Hammes, 2021). Due in part to regulatory requirements (that vary by country and jurisdiction) and logistical inputs, facilities managers have typically relied on limited risk management tools to improve the quality of water in their buildings (e.g., water flushing or changing the water heater temperature set point). Building water management plans are recommended but there is limited science-based information on how to design and implement them in practice (NASEM, 2019). Many different water quality management documents are available (Julien et al., 2020; Singh et al., 2020a);

* Corresponding author at: School of Sustainable Engineering and the Built Environment, Arizona State University, 1001 S McAllister Ave, Tempe, AZ 85281, United States.
E-mail address: kerry.hamilton@asu.edu (K.A. Hamilton).

https://doi.org/10.1016/j.wroa.2024.100244
Received 9 March 2024; Received in revised form 23 July 2024; Accepted 25 July 2024
Available online 26 July 2024

however, new proactive, science-based approaches are needed to advance building water quality due to logistical considerations such as facilities management personnel time, monitoring costs, heterogeneity among buildings requiring tailored approaches, and other factors. Similar approaches in other water quality monitoring contexts have indicated that optimized water quality network monitoring can reduce the time and cost associated with detecting pollutants while meeting water quality goals (Zhu et al., 2019).

Real-time water quality monitoring in premise plumbing environments has been suggested as the first step toward an approach for proactively managing water quality in buildings (Aden and Boyer, 2022; Kropp et al., 2022; Richard et al., 2020; Saetta et al., 2021). However, real-time monitoring of health-relevant parameters such as pathogens, DBPs, and metals is often not feasible due to technological limitations or cost, and monitoring is often limited to disinfectant residual (e.g., free chlorine) or common water quality parameters such as temperature, pH, conductivity, etc.. Typically, water quality monitoring involves frequent hands-on calibration or troubleshooting in most cases, and sensors may not be within budget or feasibility considerations for many buildings as it is currently performed.

Nevertheless, a ready-to-use, data-to-analysis framework for online premise plumbing water quality sensors could help to manage water quality and protect public health. Automated data visualization and analysis could alleviate stresses on building management staff. This approach is currently underutilized in practice. Both mechanistic and data-driven approaches have been used to address water quality (Supplemental Table S1). For example, Saetta et al. (2021) used data mining methods to predict chlorine residuals in premise plumbing using low-cost sensors (Saetta et al., 2021). Machine learning and artificial intelligence (ML/AI) techniques have been used to predict various water-related events in buildings (Kropp et al., 2022) but have not yet been bridged with physics-based or water quality predictive models (Heida et al., 2022; Palmegiani et al., 2022). Models using EPANET or Simdeum are under development to examine the impacts of local-scale premise plumbing hydraulics on water quality phenomena but have not yet been applied in routine practice (Clements et al., 2023; Ghasemzadeh, 2023). Water quality applications of sensors and ML in environmental studies are complicated by factors such as data non-stationarity due to sensor drift and calibration, changing hydraulics and/or water use patterns, and complexities surrounding missing data and time series information (Zhu et al., 2023).

The objectives of this work are therefore to: (1) propose a data collection and visualization framework for premise plumbing water quality sensors; (2) predict chlorine residual in premise plumbing using sensor data and ML models; (3) identify important variables and quantify accuracy of predictions as a function of lead time, which tests the ML model's ability to predict future chlorine concentration based on current sensor data; and (4) discuss the implications of ML for water quality management in buildings and needs for sensor accuracy and performance. Practical management includes stakeholder-centric goals for evaluating a warning signal when free chlorine is below a threshold, and predicting chlorine residual decreases ahead of time in order to anticipate and prevent a lapse in water quality.

2. Results and discussion

2.1. Water quality data and dashboard

An online dashboard was created to collect, organize, and visualize the high resolution sensor data (Supplemental Fig. S1). The dashboard allowed for real-time monitoring of pH, conductivity, temperature, ORP, DO, and free chlorine values for the 2nd, 3rd, and 7th floors of the building. The data are aggregated using a Google sheets database and Python was used with the Dash framework to create dynamic interactive graphical visualizations of the data. The graphs are on separate tabs for each floor and there is an option for the user to download the displayed data as a CSV file. The user can select between 3 data range options: 1 day, 2 days, or all available days of data stored locally since a predetermined starting time and has the option to show the mean value of each parameter on their respective graph.

Stakeholders for a building water quality dashboard include building managers, technical building staff like plumbers, and building occupants. Other individuals and organizations who might be interested in a building water quality dashboard include architects, municipal water department, and plumbing manufacturers. The dashboard in Supplemental Fig. S1 was designed to display data trends and could be tailored to different stakeholders. A user could monitor the dashboard for specific values of chlorine residual and note when the values are below a reference point (e.g., 0.2 mg/L) indicating that there may be insufficient residual present for microbial control and decide to perform actions such as flushing the fixtures. Additional information from other sensors could also indicate if a water quality change has occurred, which could trigger other actions related to other water system maintenance actions (e.g., examining the operation of the water softener, water heater, or other components). For example, as the use of dashboards has proliferated for wastewater quality data visualization (Naughton et al., 2023), their use could also be expanded more broadly for use in drinking water quality notifications at the facility level to provide advanced notification to building managers when parameters are out of their desirable range, and ultimately be used to trigger water quality management activities (e.g., collecting a grab sample or actuating a solenoid valve to flush a fixture in response to a water quality trigger) in response to water quality predictions in advance of an adverse water quality event.

Sensor data for chlorine, conductivity, oxidation–reduction potential (ORP), pH, and temperate collected over the course of the study are shown in Fig. 1. Data were collected from June 6, 2020 to May 31, 2021 with additional details on sensor measurements in Section 3.1. The water quality trends shown in Fig. 1 are typical for a building where water quality parameters can vary due to a combination of building characteristics, operational practices, and seasonal changes in water quality (Richard et al., 2021). For example, on a daily time scale, the free chlorine residual increases as fresh water from the distribution system enters the building when occupants are in the building using water, and then decays when occupants leave. In contrast, conductivity varies to a lesser degree on the time scale of days and instead changes seasonally. Although there are general trends for building water quality that can be expected, ML techniques have the potential to provide new insights and improved management of water use and water quality in buildings. 73,504 datapoints were available after cleaning. Erroneous values (e.g., 'NA' values, or signals for water quality variables recorded as '99') were omitted from the analysis.

2.2. Chlorine predictability with ML

Using five input variables (conductivity, temperature, ORP, pH, and WiFi network activity as a proxy of building occupancy), a Random Forest (RF) regression model was able to predict free chlorine concentrations with a 5-min lead time with a R^2 value of 0.9 and RMSE of 0.033 (ppm) evaluated on a test dataset withheld from training. The RF regression model outperformed a baseline multilinear regression model, which yielded a R^2 value of 0.265 and RMSE of 0.09 (ppm). Fig. 2 (a, c) shows the time series plot of predicted and in situ test data. The RF regression predicted values aligns well with the pattern of in situ data. Notably, the RF regression model tends to underestimate high values and overestimate low values, a known issue for various ML techniques as they are essentially interpolators. This could be problematic for low values in the case of free chlorine residual due to a desirable residual of 0.02–4 mg/L in the drinking water distribution system, with a value of $>$0.2 ppm at the distribution system point of entry cited for public health protection (USEPA, 2004; WHO, 2011). There is not currently a regulatory value specific to chlorine residual detection at the point of entry to

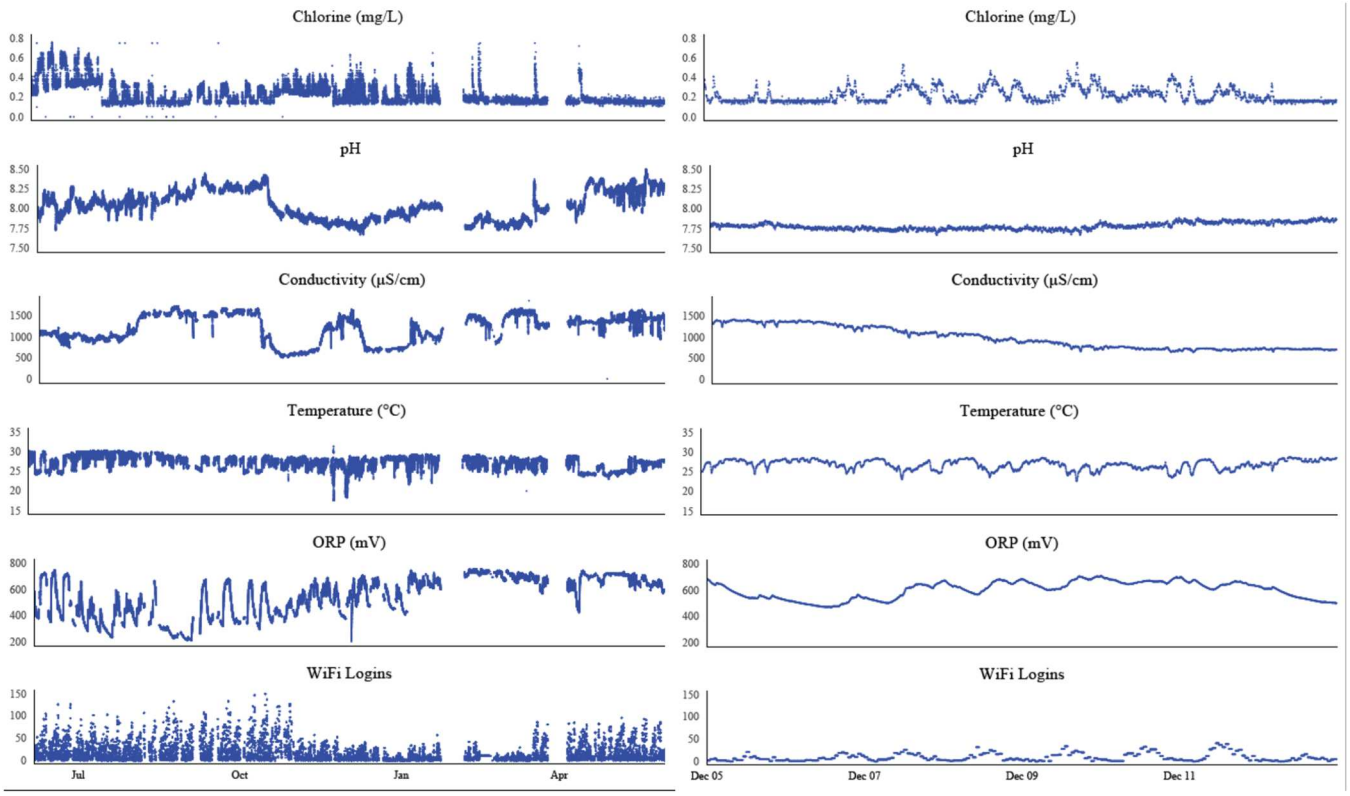

Fig. 1. Water quality data over course of the study (left) and representative of a short-term period during which additional analysis was performed (right). Note water quality changes occur on hly, daily, weekly, and seasonal time scales.

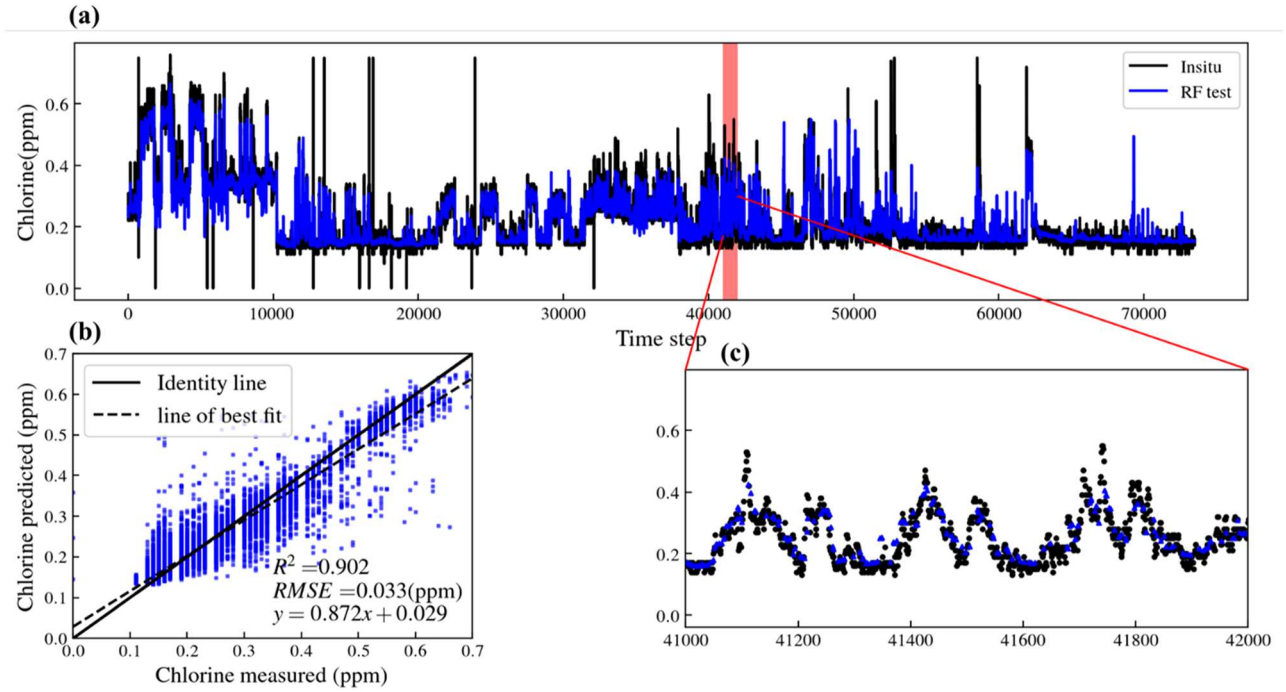

Fig. 2. Sensor-measured data and one time-step ahead ($(k = 1)$, 5 min lead-time) chlorine predicted by Random Forest regression (a) measured time series for entire 2020 and Random Forest regression predicted test data; (b) scatter plot of Random Forest regression predicted test data versus the measured test data, and (c) zoomed-in time series plot during a representative period.

buildings; however, 0.2 ppm is commonly used as a reference point. However, some outliers may be caused by sensor errors. Additionally, limitations in the sensitivity of sensors lead to artifacts which can be seen as discontinuous patterns in the data (see chlorine and pH data in Figs. 1 and 2).

In a separate test, we used a deep learning algorithm (Long Short-

Term Memory, LSTM) to predict free chlorine concentration using time series of the five input variables in a sequence-to-sequence (seq2seq) framework. Specifically, the LSTM model takes the sequential inputs, in this case, the time series of the five input variables, and predicts the chlorine concertation as a time series. The LSTM model yielded lower R^2 and higher RMSE than RF results (Supplemental Fig. S2), likely due to nonstationarity of input variables. Specifically, temporal variations were found in input variables but did not always result in changes in free chlorine concentration, while short spikes (including possible outliers) of free chlorine concentration may occur when input variables remain relatively stable. The mismatch of temporal variations between input and target is likely due to physicochemical and microbial processes that took place during the study duration, potentially related to stagnation, varying building operational conditions, seasonal temperature fluctuations, changes in source water quality, occupancy changes, and/or sensor data error (S Joshi et al., 2023). This suggests that sequential inputs, which incorporate historical information, may introduce noise into the model, making it challenging for the model to discern meaningful patterns between inputs and outputs when underlying processes are not fully captured by data.

Random Forest (RF) classification also shows promising accuracy in predicting free chlorine low values with an overall accuracy of 95.5 %. True Positive (TP) represents the correctly predicted true label "1", representing chlorine concentration lower than 0.2 ppm, while True Negative (TN) is vice versa. A False Positive (FP) occurs when the model predicts 0 but the label is 1, whereas False Negative (FN) is the misclassification of label 1 as 0. Our dataset had more labeled values of "1" than "0" (62 %:38 %). Given the minor imbalance and the focus on predicting "1" class, we chose not to process data to enforce class balance. For problems with more pronounced class imbalance and where the focus is to predict the class with fewer samples, e.g., an imbalance beyond 90:10 (He and Garcia, 2009) or more, it may be desirable to use downsampling and upweighting techniques (Drummond and Holte, 2003) to encourage the model to learn from the minor class.

2.3. Multiple lead time steps predictability

Fig. 3 illustrates the prediction accuracy of ML regression and classification models with different lead time-step (denotated as "*k* " in Section 3.3), i.e., how accurate the current inputs can predict the free chlorine in the future time steps. For both models, an increase in lead time-step results in slightly decreased accuracy despite fluctuations due to randomness inherent to the algorithm and induced in data splitting. RF regression shows robust performance, maintaining a low RMSE of 0.04 (ppm) even with a lead time of 288-time steps (24 h). Similarly, the accuracy of RF classification predicting labels 24 h later remains around 94.5 %. Because of autocorrelation in input variables and free chlorine concentration (Supplemental Fig. S3), the models were able to learn the relation between free chlorine concentration in the future and the five input variables at current time step and thus forecast free chlorine concentration with high accuracy. Fig. 4 shows the variable importance scores of RF regression and RF classification based on a permutation method described in Section 3.2. The importance scores were calculated for each lead time and then averaged. Conductivity resulted in the highest score (i.e., conductivity was most important for predicting free chlorine), while occupancy was the least important feature. This agrees with partial dependence plots (Supplemental Figs. S4 and S5), which show the marginal effect of each input variable on the model prediction. Specifically, the RF regression (Supplemental Fig. S4) tends to result in lower free chlorine values related with an increase in conductivity, while it is insensitive to changes in building occupancy. Similarly, the RF classification tends to have a high possibility of a free chlorine value <0.2 ppm (indicated as "label 1″) with an increase in conductivity (Supplemental Fig. S5). The variable importance score of temperature and ORP has a different ranking between RF regression and RF classification. The classifier model relies more on distinguishing the boundary (threshold-based label) compared to the regression model producing continuous output. Additionally, the partial dependence plots (PDP) and variable importance results of temperature and ORP result may be confounded by the correlation (Pearson correlation r is -0.57) between them (Fig. 5). This is expected given that temperature affects the rate of the chemical reactions that affect ORP, with warmer water tending to have a lower ORP than colder water (Nordstrom and Wilde, 1998). Correlations between pH and conductivity as well as occupancy (Pearson correlation r are 0.33 and 0.30) may reflect the effects of water usage on intermittent water chemistry. Free chlorine and conductivity demonstrated moderate correlation, which may be due to the presence of chloride ions formed from chlorine. In addition, to assess the necessity of using all sensors, we conducted two experiments using only the top 3 and top 2 most important variables from Fig. 4. This led to lower performance for both the RF regression and classification models (Supplemental Fig. S6). These comparisons demonstrate that information from a single sensor is less informative than using data from a combination of sensors to feed into a ML approach.

For comparison of the ML approaches with a baseline "simple" approach, multiple linear regression (MLR) and logistic regression were performed with five input variables and a lead time of one step. Their performance was compared to the RF regression and RF classification

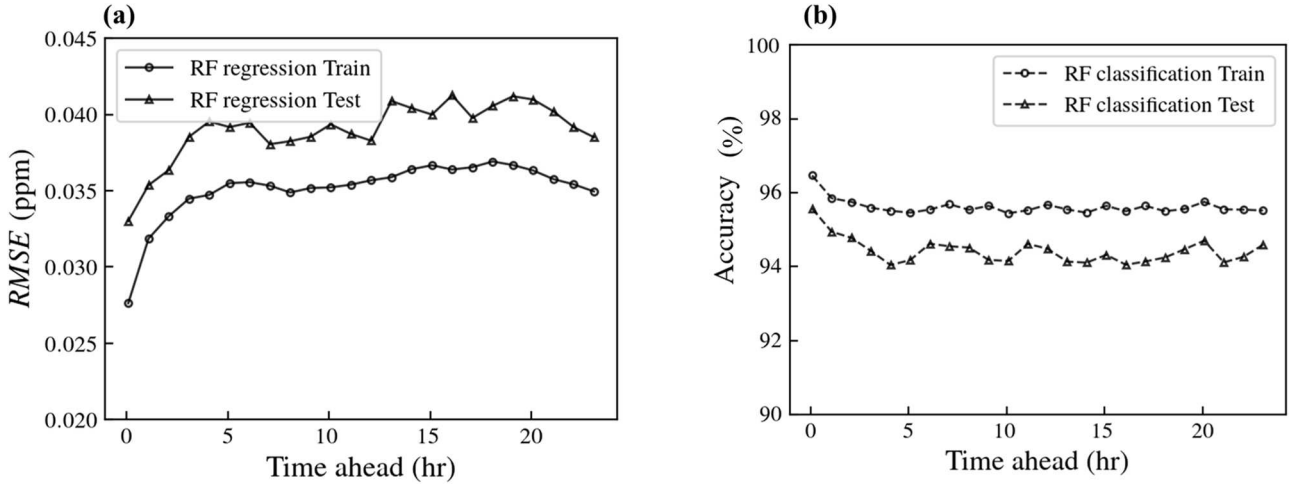

Fig. 3. Prediction performance of the Random Forest regression and classification model with lead-time ranging from 5 min to 24 h with a timestep of 1 h. (a) root-mean-square error (RMSE) for regression and (b) accuracy rate for classification.

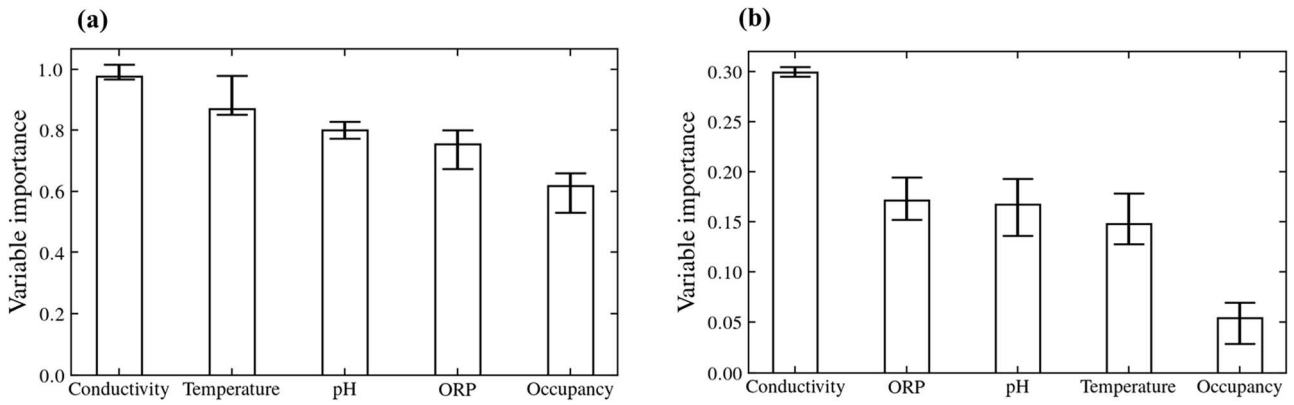

Fig. 4. Permutation based importance scores of Random Forest models for (a) score that based on decrease of R^2 for Random Forest regression; (b) score that based on decrease of accuracy for Random Forest classification. The important scores of features are median and error bar shows 95 % distribution of all experiments results that with lead - time ranging from 5 min to 24 h.

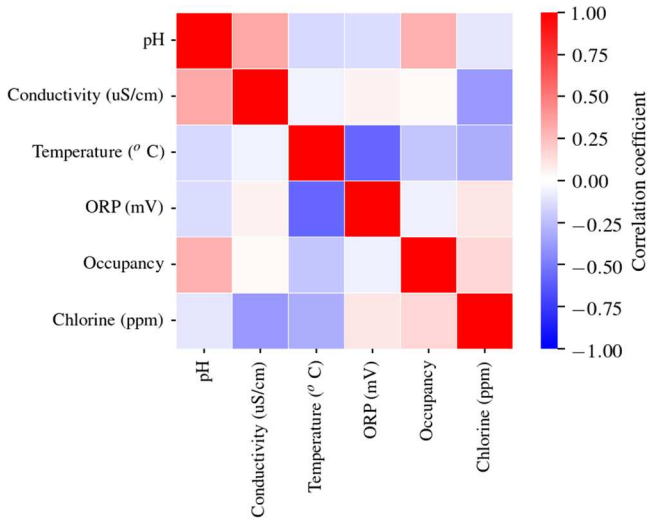

Fig. 5. Pearson correlation coefficient matrix among inputs (pH, conductivity, temperature, ORP, and occupancy) and target (free chlorine concentration).

models, respectively. The MLR model (Supplemental Fig. S7) model was evaluated on the same data as the RF and LSTM models and yielded an R^2 of 0.265 and RMSE of 0.09 ppm. The performance of the MLR model was substantially lower than the performance of the RF and LSTM models, suggesting the potential value of ML methods for predicting free chlorine concentration. Meanwhile, the logistic regression model (Supplemental Table S2), using the same data as RF classification, achieved an accuracy of 70.9 %, lower than RF classification. These comparisons demonstrate the need for using more sophisticated ML approaches for this problem.

2.4. Implications for building water quality management

The current work presents novel ML algorithms applied to predicting a key health-relevant variable for managing water quality in buildings, free chlorine residual. The ML models predicted free chlorine with reasonable accuracy (RMSE<0.042 ppm, accuracy>94 %) up to 24 h reliably in advance. A data dashboard for real-time water quality visualization was developed, showing 5-min data for water quality sensors on multiple building floors for pH, temperature, ORP, and free chlorine. The goal of the current analysis is to ultimately integrate the developed ML algorithms with the data dashboard for real-time water quality management, for example via notification of trained personnel, automated valve operation, or temperature setting control with a goal of some period of advance notice. As data dashboards have proliferated for use in wastewater testing and public health decision-making (Naughton et al., 2023), there is a need for this technology to be used for decision-making for drinking water quality and other aspects of the indoor built environment where the majority of human exposures will occur (Klepeis et al., 2001).

Current data visualization methods like those that involve data dashboards (Supplemental Information Fig. S1) often place the burden of decision-making on the stakeholder or user. While this type of approach provides current and historical data, it does not make use of data analysis in real time to form predictions, which is the goal of a "sense-analyze-decide" framework. A more advanced dashboard could show predictions and alerts, and then the stakeholder or user could perform, or be better to perform, a corrective action. The duration of corrective action would be dependent on the goals of the end user. Ideally, the visualization component of the sense-analyze-decide framework would be present at each stage of the process. A vision for this integration along with associated research needs is described in Table 2. While other work has sought to evaluate the needs of building facilities managers with respect to water quality (Rasheduzzaman et al., 2023, 2021; Singh et al., 2022, 2020b) there is no single data lead-time that is appropriate for all end users. Consultation with the facilities managers involved directly in this work indicated that at least a 1-h lead time, and up to 24 h for shift and activity planning would be beneficial for decision-making. User testing and consultation for the development of more applied user tools is needed to inform future work.

2.5. Limitations and uncertainty

While AI/ML frameworks present great promise for predicting water quality variables over time (Supplemental Table S1), there were several limitations of the current work, including sensor drift, stepwise short-term changes, shifting baseline values, the need for manual sensor calibration, and limitations related to free chlorine detection resolution, particularly at levels below 0.2 ppm. Additional validation of chlorine sensors would be beneficial to perform to quantify the uncertainty associated with each sensor and further validate the free chlorine measurements, e.g., using standard methods (Wendelken et al., 2009). Concordance between free chlorine sensor measurements and grab samples were demonstrated through periodic checks during the current study period against handheld Hach chlorine analyzers to within 10 %. The sensors in this study were used due to their ease of integration with microcomputers (e.g., Raspberry Pi) and sensor networks. Other examples of the sensors used in this study in other studies (Martinez Paz et al., 2024, 2022) have noted reliable results with infrequent calibration as specified by the manufacturer. Nevertheless, additional parallel comparisons between different quantification methods and sensor types

would be beneficial for calibrating and reality-checking sensor measurements over time as a function of different water qualities and user maintenance/calibration approaches.

The sensor flow cell maintained a continuous, but low, flow of water through the sensors. However, the character of the water flowing through the sensor was influenced by the flow or stagnation conditions happening in the area of the building where the sensor was located and could have influenced correlations among variables. As a result of uncertainty introduced by sensor limitations, relative comparisons of variables and trend detection may be better suited for analysis compared to absolute value thresholds as addressed in the current work.

The regression model predicts reference free chlorine concentration values, which helps stakeholders make informed adjustments to building water systems. On the other hand, the classification model offers early warnings by predicting whether free chlorine levels will meet critical thresholds, enabling immediate responses. Both models are necessary as they address different aspects of decision-making: regression offers a value that can be used as a reference for stakeholder long-term planning, and adjustments, while classification for immediate, threshold-based actions. By presenting both, we demonstrate how sensor data can support a wide range of decision-making needs, enhancing the overall utility of ML in monitoring chlorine concentrations.

The partial dependence plot (PDP) shows how changes in one feature affect the overall prediction and provide insights into the impact of uncertainty in the feature on model prediction uncertainty. When the PDP shows a steep slope in a specific range of a feature, uncertainties in that range have a higher impact on predictions, and vice versa for milder slopes. An example of this is shown in Fig. S4, where the PDP of conductivity has a steep and varying slope in the range of 500-1500 uS/cm. In particular, sensor error when conductivity is around 1300 uS/cm may cause predicted free chlorine concentration to fluctuate around the threshold (0.2 ppm).

As AI/ML methods proliferate for use in the water sector, there is a need to understand the implications of error propagation on model outputs. This could include a comparison of how the quality of the sensor, including sensor drift or error affects decision switchover points for water quality management (i.e., when, how often, and how much to flush the water system). Cost-benefit evaluations have indicated that combinations of lower-cost sensors and/or grab samples could achieve water quality benefits, indicating that there may be tradeoffs in the value of information between reducing error associated with a particular sensor vs. switching the combination of overall sensors used for prediction of a given water quality outcome like free chlorine residual (Saetta et al., 2021). Sensitivity analysis to examine the impact of sensor error on ML model fitting approaches, as well as the additional value of information approaches to weigh the cost of obtaining additional information or error reduction and dependency of the management approach on this cost would be beneficial (Luhede et al., 2024). This can help to inform considerations regarding the true applicability (robustness versus fragility) of the models. The need for understanding the specific drivers of uncertainty is increasingly addressed using ML models that combine physics-based approaches with data-driven models in environmental fields (Fung et al., 2021; Sayed et al., 2023; Zhao et al., 2023). Despite these limitations, the current work is a step forward for advancing a “sense-analyze-decide” framework for building water management that could operate within a multi-pronged digital building management system that advances sustainability and efficiency goals while protecting public health.

3. Methodology

3.1. Data collection and data visualization dashboard

Sensors were installed on the 2nd, 3rd, and 7th floors of the building and calibrated once per week (for pH, ORP, DO, and conductivity sensors). Data were recorded every 5 min for pH (AtlasScientific #ENV-40-pH), temperature (AtlasScientific #PT-1000), conductivity (AtlasScientific #ENV-40-EC-K1.0), oxidation–reduction potential (ORP) (AtlasScientific #ENV-40-ORP), dissolved oxygen (CO) (AtlasScientific #ENV-40-DOX), and free chlorine (Chemtrol PPMFC002). The sensors were installed and operated following manufacturer guidelines. The free chlorine sensor used in this work was a membrane-based, amperometric chlorine sensor that measured HOCl and OCl^-. The sensor had an operating pH range of 5.5 to 9.5 according to the manufacturer. Both HOCl and OCl^- diffused through the membrane where OCl^- was converted to HOCl in a low pH environment, and HOCl was detected. Therefore, the sensor signal was proportional to the sum of HOCl and OCl^- in the bulk water, and pH correction or buffer addition was not required. The sensors were installed in a flow cell that operated with continuous water flow. Calibrations for pH, ORP, DO, and conductivity sensors were performed weekly on Friday mornings during the study period. The manufacturer stated that calibration of the free chlorine sensors is not necessary, however, the gel in the sensors was checked every ~6 months and gel was replaced when necessary. In all cases, the sensor returned ± 10 % of the spiked value. The 5 to 20 min calibration periods are not expected to have introduced significant bias due to the small number of data points.

The sensor inputs were converted from analog to digital signals using manufacturer supplied circuits. Sensor data were wirelessly uploaded to a Google Sheets repository. Full description of the sensor systems are described previously (Saetta et al., 2021). The current work advances the analysis in Saetta et al. (2021) due to the inclusion of future time (forecasting) predictions and use of a more comprehensive dataset from June 2020- May 2021 (compared to September 2019- February 2020 in the original modeling effort). The current dataset is much larger and provided a more complete dataset with fewer missing values for the LSTM framework. Additional types of ML models were fit in the current analysis, including LSTM, a novel approach for predicting water quality data.

Occupancy data were collected as described previously (Aden and Boyer, 2022; Sayalee Joshi et al., 2023). Data from a 1-year period (June 2020 through May 2021) were used for model training, validation and test dataset. The validation data are used for tunning ML model hyperparameters and early stopping specifically for LSTM. The google sheets data were used to display the sensor data trends utilizing a dashboard style user interface hosted on a web server. The dashboard also allowed for user access and download of sensor data.

3.2. ML models

Experiments were conducted using three ML models: RF regression, RF classification, and LSTM. RF is a non-parametric ML algorithm that constructs an ensemble of decision trees. Each tree is constructed based on a bootstrap sample (i.e., sample with replacement) of the training dataset. Typically, about one third of training data is not selected in a bootstrap sample; these data points are referred to as out-of-bag (*oob*) data and can be used to assess the generalization performance of the tree. RF makes predictions by averaging outputs of all trees (Breiman, 2001). RF has achieved state-of-the-art performance in numerous applications (Díaz-Uriarte and Alvarez de Andrés, 2006; Wang et al., 2021; Wei et al., 2022) and is designed to alleviate overfitting issues attributed to an individual decision tree. RF also calculates variable importance scores, which characterize the importance of each input variable for predicting the target variable (free chlorine concentration in this study). For each input variable, the algorithm randomly permutes its value while keeping other variables unchanged, train the model on permutated data, and then evaluates model accuracy using the *oob* data. The variable importance score shows the decreased model performance from the baseline model trained with non-permuted data (Breiman, 2001; Molnar, 2020). In this study, the mean square error (MSE) is used to construct the trees, while R^2 is used to calculate variable importance

scores, following the default option in the scikit-learn package.

LSTM networks are a type of recurrent neural networks specifically designed for sequences such as sentences and time series (Graves et al., 2013). At a given time step, the network uses "gates" to process inputs, update the cell memory, which stores past information, and output prediction for the current time step. As such, LSTMs are capable of modeling temporal dynamics underlying sequential data in applications from various disciplines (Kratzert et al., 2018).

3.3. Data preparation and experiments

Data were cleaned by first removing any extraneous values caused by sensor errors. To differentiate between any hardware and software errors during the data collection periods, hardware errors (i.e., IOErrors) were assigned a value of 99. All other errors associated with storage or data transfer via network were reported as "NA". For instances where one or more sensors reported an error, all data points for that specific time point were excluded from analysis. WiFi login data and sensor data were combined based on time. WiFi data were reported hourly while sensor data were reported every 5 min. The WiFi data were assumed to be constant over the entire hour for which they were reported. The last reported WiFi data value was assigned to each time step (e.g., WiFi login count for 2:00PM was assigned to 2:05PM, 2:10PM, etc., until 3:00PM).

Each RF experiment utilized five inputs at a given time step (occupancy, ORP, pH, conductivity, and temperature, denoted as x_t) to predict the free chlorine concentration for future time steps y_{t+k}, where k is the lead time step. For both RF regression and classification, data pairs of $\{x_t, y_{t+k}\}$ were randomly split into training, validation, and test sets in a 6:2:2 ratio. Inputs are normalized to range [0,1] before being fed into models. To evaluate the performance of the models, we calculated root mean square error (RMSE) and the coefficient of determination R^2 for the regression models, and accuracy rate, calculated as the percentage of correctly predicted labels, and confusion matrix for the classification models (Table 1). In the RF classification experiment we preprocessed the target data into binary labels using a threshold of 0.2 (ppm) (WHO, 2011). A label of "1" represented any chlorine value lower than 0.2, while "0" represented free chlorine values larger than 0.2. Approximately 62 % of the targets were classified as label "1", while 38 % were labeled as "0". We used stratified sampling during the random split to ensure that the class proportions were the same for the train, validation, and test sets.

We used grid search to tune the hypermeters to minimize MSE (regression) or Accuracy (classification), indicating as the fraction of correct predictions, on validation data. For RF regression, the search space for hyperparameters are: maximum number features per tree: {2, 3, 4, 5}; minimum sample leaf size: {20, 40, 50, 60, 80}. For RF classification: maximum number features per tree: {2, 3, 4, 5}; minmumsample leaf size: {20, 40, 50, 60, 80}. The RF regression experiment had hyperparameters of number of estimators=800; maximum number features per tree=3; minimum sample leaf=20. The RF classification experiment had hyperparameters of number of estimators=800; maximum number features per tree=4; minimum sample leaf=20.

The LSTM regression model uses one-layer sequence-to-sequence LSTM with 512 hidden units. The model is trained using the Adam optimizer (Kingma and Ba, 2017) with a learning rate of 10^{-4} and MSE loss. A 30 % dropout is added at fully connected layers to prevent overfitting. The learning rate and dropout rate were tuned based on MSE loss on validation dataset. The grid search range for the learning rate is $\{5 \times 10^{-5}, 7 \times 10–5, 1 \times 10^{-4}, 5 \times 10^{-4}, 1 \times 10^{-3}\}$, while the range for dropout rates is {0.1, 0.2, 0.3, 0.4, 0.5}. Input and output data for this experiment are time series with a length of 60, i.e., we seek$[x_{t-59}, \ldots, x_t]$ to $[y_{t-59+k}, \ldots, y_{t+k}]$, k is the lead time step. In our implementation of the LSTM model, the output of each time step, y_{t+k}, is calculated using input variables up to t . The data were partitioned into chunks with no overlapping between chunks.

Table 1
Confusion matrix for one-time-step ahead prediction ($k = 1$) using Random Forest classification. Label 1 represents chlorine concentrations lower than 0.2 ppm, while label 0 represents the opposite. The model achieves an overall accuracy of 95.5 % (TP + TN) and a 4.5 % error rate (FP + FN).

		Predicted	
		Label 1	Label 0
Measured	Label 1	TP= 59.7 %	FN= 2.1 %
	Label 0	FP=2.4 %	TN=35.8 %

Table 2
Sense-analyze-decide framework vision and research needs for building water quality.

Framework step	Vision for future applicability in field settings	Research needs and challenges
Sense	-Data from sensors are displayed in real-time along with historical data in a data dashboard	-Additional calibration and validation of sensor measurements needed across various sensor types -Field testing of sensor performance under realistic conditions (e.g., various calibration and/or sensor maintenance regimes and water qualities), robust derivation of detection limits and collection of data to support uncertainty analysis -Performing user needs assessment to inform dashboard user interface
Analyze	-Predicted (i.e., future time points) data are displayed along with real-time and historical data. For example, a plot or time-series could show the chlorine concentration on the 3rd floor at real-time, 1 h prior to real-time, and 1 h into the future	-Evaluation of impact of sensor errors and uncertainty, and implications for error propagation, model fitting, and decision-making -Approaches for automating ML data storage and processing in real-time
Decide	-Forecasted predictions are linked to action (e.g., prompting a facilities manager to make a decision based on when chlorine is predicted to decrease below detection, or actuating a valve)	-Performing user decision-mapping to better understand user community needs and feedback loops with communication approaches -Linking decision trigger points to automated water quality management (e.g., performing automated flushing) -Evaluating the cost-benefit of different approaches, accounting for impacts on sustainability and health

Data availability

Data used for model training are provided in the supplementary materials. Source code are available for download from: https://github.com/GW-ASU/predict_chlorine_residual.

CRediT authorship contribution statement

S. Wei: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Data curation. **R. Richard:** Formal analysis, Data curation. **D. Hogue:** Formal analysis, Data curation. **I. Mondal:** Formal analysis, Data curation. **T. Xu:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Formal analysis, Conceptualization. **T.H. Boyer:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Methodology, Funding acquisition, Data curation. **K.A.**

Hamilton: Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Methodology, Funding acquisition, Formal analysis.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Kerry Hamilton reports financial support was provided by The Zimin Institute. If there are other authors, they 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

We acknowledge the Zimin Institute project "A data-driven approach for water safety plans in sustainable buildings to predict and prevent disease" for supporting this work. The authors are grateful to ASU building staff who provided logistical support for this work, as well as Carlos Levya who designed the data dashboard. The authors are also grateful to Mr. Aninda Ghosh for providing insights on ML model functionalities and Dr. Sayalee Joshi for providing information on sensor calibration and analysis.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.wroa.2024.100244.

References

Abokifa, A.A., Katz, L., Sela, L., 2020. Spatiotemporal trends of recovery from lead contamination in Flint, MI as revealed by crowdsourced water sampling. Water Res. 171, 115442 https://doi.org/10.1016/j.watres.2019.115442.

Aden, K., Boyer, T.H., 2022. Shift to remote learning degrades water quality in buildings. AWWa Water Sci. 4, e1316. https://doi.org/10.1002/aws2.1316.

Allen, J.M., Cuthbertson, A.A., Liberatore, H.K., Kimura, S.Y., Mantha, A., Edwards, M. A., Richardson, S.D., 2017. Showering in flint, MI: is there a DBP problem? J. Environ. Sci. 58, 271–284. https://doi.org/10.1016/j.jes.2017.06.009.

Baum, R., Bartram, J., Hrudey, S., 2016. The flint water crisis confirms that U.S. drinking water needs improved risk management. Environ. Sci. Technol. 50, 5436–5437. https://doi.org/10.1021/acs.est.6b02238.

Bravo, D., Bennia, A., Naji, H., Fellouah, H., Báez, A., 2020. General review of air-conditioning in green and smart buildings Revisión general sobre sistemas de acondicionamiento de aire en edificios ecológicos e inteligentes 35.

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

CDC, 2018. Legionella (Legionnaires' Disease and Pontiac Fever).

Clements, E., Irwin, C., Koestner, J., Taflanidis, A., Bibby, K., Nerenberg, R., 2023. Characterizing stochastic water age in premise plumbing systems using conventional and advanced statistical tools. Environ. Sci. Water Res. Technol. 9, 1182–1194. https://doi.org/10.1039/D2EW00872F.

Díaz-Uriarte, R., Alvarez de Andrés, S., 2006. Gene selection and classification of microarray data using random forest. BMC Bioinform. 7, 3. https://doi.org/10.1186/1471-2105-7-3.

Drummond, C., Holte, R.C., 2003. C4.5, Class Imbalance, and Cost Sensitivity: why Under-Sampling beats Over-Sampling 11, 1–8.

Fung, P.L., Zaidan, M.A., Timonen, H., Niemi, J.V., Kousa, A., Kuula, J., Luoma, K., Tarkoma, S., Petäjä, T., Kulmala, M., Hussein, T., 2021. Evaluation of white-box versus black-box machine learning models in estimating ambient black carbon concentration. J. Aerosol. Sci. 152, 105694 https://doi.org/10.1016/j.jaerosci.2020.105694.

Ghasemzadeh, K., 2023. Proactive Real-time Control of Multiple Interdependent Water Quality Variables in Buildings Water Networks. Arizona State University.

Graves, A., Jaitly, N., Mohamed, A., 2013. Hybrid speech recognition with deep bidirectional LSTM. In: Proceedings of the 2013 IEEE Workshop on Automatic Speech Recognition and Understanding. Presented at the 2013 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU). Olomouc, Czech Republic. IEEE, pp. 273–278. https://doi.org/10.1109/ASRU.2013.6707742.

He, H., Garcia, E.A., 2009. Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. 21, 1263–1284. https://doi.org/10.1109/TKDE.2008.239.

Heida, A., Mraz, A., Hamilton, M.T., Weir, M.H., Hamilton, K.A., 2022. Computational framework for evaluating risk trade-offs in costs associated with legionnaires' disease risk, energy, and scalding risk for hospital hot water systems. Environ. Sci. Water Res. Technol. 8, 76–97. https://doi.org/10.1039/D1EW00397F.

Joshi, S., Richard, R., Hogue, D., Brown, J., Cahill, M., Kotta, V., Call, K., Butzine, N., Marcos-Hernández, M., Alja'fari, J., Voth-Gaeddert, L., Boyer, T., Hamilton, K., 2023a. Water quality trade-offs for risk management interventions in a green building. Environ. Sci. Water Res. Technol. Rev.

Joshi, Sayalee, Richard, R., Levya, C., Harrison, J.C., Saetta, D., Sharma, N., Crane, L., Mushro, N., Dieter, L., Morgan, G.V., Heida, A., Welco, B., Boyer, T.H., Westerhoff, P., Hamilton, K.A., 2023b. Pinpointing drivers of widespread colonization of Legionella pneumophila in a green building: roles of water softener system, expansion tank, and reduced occupancy. Front. Water 4. https://doi.org/10.3389/frwa.2022.966223.

Julien, R., Dreelin, E., Whelton, A.J., Lee, J., Aw, T.G., Dean, K., Mitchell, J., 2020. Knowledge gaps and risks associated with premise plumbing drinking water quality. AWWa Water Sci. 2, e1177. https://doi.org/10.1002/aws2.1177.

Kingma, D.P., Ba, J., 2017. Adam: a Method for Stochastic Optimization. 10.48550/arXiv.1412.6980.

Klepeis, N.E., Nelson, W.C., Ott, W.R., Robinson, J.P., Tsang, A.M., Switzer, P., Behar, J. V., Hern, S.C., Engelmann, W.H., 2001. The national human activity pattern survey (NHAPS): a resource for assessing exposure to environmental pollutants. J. Expo. Sci. Environ. Epidemiol. 11, 231–252. https://doi.org/10.1038/sj.jea.7500165.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., Herrnegger, M., 2018. Rainfall–runoff modelling using long short-term memory (LSTM) networks. Hydrology and Earth System Sciences 22 (11), 6005–6022.

Kropp, I., Pouyan Nejadhashemi, A., Julien, R., Mitchell, J., Whelton, A.J., 2022. A machine learning framework for predicting downstream water end-use events with upstream sensors. Water Supply 22, 6427–6442. https://doi.org/10.2166/ws.2022.226.

Logan-Jackson, A.R., Batista, M.D., Healy, W., Ullah, T., Whelton, A.J., Bartrand, T.A., Proctor, C., 2023. A critical review on the factors that influence opportunistic premise plumbing pathogens: from building entry to fixtures in residences. Environ. Sci. Technol. 57, 6360–6372. https://doi.org/10.1021/acs.est.2c04277.

Luhede, A., Yaqine, H., Bahmanbijari, R., Römer, M., Upmann, T., 2024. The value of information in water quality monitoring and management. Ecol. Econ. 219, 108128 https://doi.org/10.1016/j.ecolecon.2024.108128.

Martinez Paz, E.F., Raskin, L., Wigginton, K.R., Kerkez, B., 2024. Toward the autonomous flushing of building plumbing: characterizing oxidation-reduction potential and temperature sensor dynamics. Water Res. 251, 121098 https://doi.org/10.1016/j.watres.2023.121098.

Martinez Paz, E.F., Tobias, M., Escobar, E., Raskin, L., Roberts, E.F.S., Wigginton, K.R., Kerkez, B., 2022. Wireless sensors for measuring drinking water quality in building plumbing: deployments and insights from continuous and intermittent water supply systems. ACS EST Eng. 2, 423–433. https://doi.org/10.1021/acsestengg.1c00259.

Molnar, C., 2020. Interpretable Machine Learning. Lulu.com.

NASEM, 2019. Management of Legionella in Water Systems. National Academies Press, Washington, D.C.. https://doi.org/10.17226/25474

Naughton, C.C., Holm, R.H., Lin, N.J., James, B.P., Smith, T., 2023. Online dashboards for SARS-CoV-2 wastewater data need standard best practices: an environmental health communication agenda. J. Water Health 21, 615–624. https://doi.org/10.2166/wh.2023.312.

Nordstrom, D., Wilde, F., 1998. Reduction-6.5 oxidation potential (electrode method). US Geol. Surv. 9. TWRI Book.

Palmegiani, M.A., Whelton, A.J., Mitchell, J., Nejadhashemi, P., Lee, J., 2022. New developments in premise plumbing: integrative hydraulic and water quality modeling. AWWa Water Sci. 4, e1280. https://doi.org/10.1002/aws2.1280.

Park, Sanguk, Park, Sangmin, Choi, M., Lee, S., Lee, T., Kim, S., Cho, K., Park, Sehyun, 2020. Reinforcement learning-based BEMS architecture for energy usage optimization. Sensors 20, 4918. https://doi.org/10.3390/s20174918.

Rasheduzzaman, M., Singh, R., Annapoorna Madireddy, L., Gurian, P.L., 2021. Conceptualization to development of a decision support tool to manage building water quality 925–936. 10.1061/9780784483466.084.

Rasheduzzaman, M., Singh, R., Haas, C.N., Olson, M.S., Gurian, P.L., 2023. A literature-engaged Delphi approach for water quality management in building water systems. AWWa Water Sci. 5, e1339. https://doi.org/10.1002/aws2.1339.

Rhoads, W.J., Hammes, F., 2021. Growth of Legionella during COVID-19 lockdown stagnation. Environ. Sci. Water Res. Technol. 7, 10–15. https://doi.org/10.1039/D0EW00819B.

Rhoads, W.J., Pruden, A., Edwards, M.A., 2016. Survey of green building water systems reveals elevated water age and water quality concerns. Environ. Sci. Water Res. Technol. 2, 164–173. https://doi.org/10.1039/C5EW00221D.

Richard, R., Hamilton, K.A., Westerhoff, P., Boyer, T.H., 2021. Physical, chemical, and microbiological water quality variation between city and building and within multistory building. ACS EST Water 1, 1369–1379. https://doi.org/10.1021/acsestwater.0c00240.

Richard, R., Hamilton, K.A., Westerhoff, P., Boyer, T.H., 2020. Tracking copper, chlorine, and occupancy in a new, multi-story, institutional green building. Environ. Sci. Water Res. Technol. 6, 1672–1680. https://doi.org/10.1039/D0EW00105H.

Saetta, D., Richard, R., Leyva, C., Westerhoff, P., Boyer, T.H., 2021. Data-mining methods predict chlorine residuals in premise plumbing using low-cost sensors. AWWa Water Sci. 3, e1214. https://doi.org/10.1002/aws2.1214.

Sayed, B.T., Al-Mohair, H.K., Alkhayyat, A., Ramírez-Coronel, A.A., Elsahabi, M., 2023. Comparing machine-learning-based black box techniques and white box models to predict rainfall-runoff in a northern area of Iraq, the Little Khabur River. Water Sci. Technol. 87, 812–822. https://doi.org/10.2166/wst.2023.014.

Singh, R., Chauhan, D., Fogarty, A., Rasheduzzaman, M., Gurian, P.L., 2022. Practitioners' perspective on the prevalent water quality management practices for legionella control in large buildings in the United States. Water 14, 663. https://doi.org/10.3390/w14040663.

Singh, R., Hamilton, K.A., Rasheduzzaman, M., Yang, Z., Kar, S., Fasnacht, A., Masters, S. V., Gurian, P.L., 2020a. Managing water quality in premise plumbing: subject matter

experts' perspectives and a systematic review of guidance documents. Water 12, 347. https://doi.org/10.3390/w12020347.
Singh, R., Hamilton, K.A., Rasheduzzaman, M., Yang, Z., Kar, S., Fasnacht, A., Masters, S. V., Gurian, P.L., 2020b. Managing water quality in premise plumbing: subject matter experts' perspectives and a systematic review of guidance documents. Water 12, 347. https://doi.org/10.3390/w12020347.
USEPA, 2004. National Primary Drinking Water Regulations: Surface Water Treatment Rule Subpart H-Filtration and Disinfection.
Wang, F., Wang, Y., Zhang, K., Hu, M., Weng, Q., Zhang, H., 2021. Spatial heterogeneity modeling of water quality based on random forest regression and model interpretation. Environ. Res. 202, 111660 https://doi.org/10.1016/j.envres.2021.111660.
Wei, S., Xu, T., Niu, G.Y., Zeng, R., 2022. Estimating irrigation water consumption using machine learning and remote sensing data in Kansas high plains. Remote Sens. 14, 3004. https://doi.org/10.3390/rs14133004.
Wendelken, S., Losh, D., Fair, P., 2009. Method 334.0: Determination of Residual Chlorine in Drinking Water Using An On-line Chlorine Analyzer. USEPA Office of Ground water and drinking water, Cincinnati, Ohio.
WHO, 2011. Guidelines For Drinking-Water quality, Fourth Edition [WWW Document]. WHO. URL. https://www.who.int/water_sanitation_health/publications/2011/dwq_guidelines/en/ (accessed 3.19.19).
WHO, 2007. Legionella and the Prevention of Legionellosis. World Health Organization, Geneva.
Zhao, L., Guo, Y., Mohammadian, E., Hadavimoghaddam, F., Jafari, M., Kheirollahi, M., Rozhenko, A., Liu, B., 2023. Modeling permeability using advanced white-box machine learning technique: application to a heterogeneous carbonate reservoir. ACS Omega 8, 22922–22933. https://doi.org/10.1021/acsomega.3c01927.
Zhu, J.J., Yang, M., Ren, Z.J., 2023. Machine learning in environmental research: common pitfalls and best practices. Environ. Sci. Technol. 57, 17671–17689. https://doi.org/10.1021/acs.est.3c00026.
Zhu, X., Yue, Y., Wong, P., Zhang, Y., Ding, H., 2019. Designing an optimized water quality monitoring network with reserved monitoring locations. Water 11, 713. https://doi.org/10.3390/w11040713.