Learning Water Quality Functions from Data for Monitoring and Controlling Water Quality in Networks

Kia Zadeh
Industrial Engineering Ph.D. Student
School of Computing and Augmented Intelligence (SCAI)
Arizona State University
Tempe, AZ 85281, USA
kghasemz@asu.edu

Treavor Boyer
Professor
School of Sustainable Engineering and the Built Environment
Arizona State University
Tempe, AZ 85281, USA
thboyer@asu.edu

Pitu Mirchandani
Professor
School of Computing and Augmented Intelligence (SCAI)
Arizona State University
Tempe, AZ 85281, USA
pitu@asu.edu

Abstract

It is envisioned that future Commercial and Institutional (CI) buildings will be installed with a multitude of sensors that will provide water quality measures in real-time. Since water quality does not change instantly, one needs to predict how water quality will change over time and space, and proactively control water quality using chemical additives, or removing contaminants from water, to keep its quality within the regulated safe range for drinking. Therefore, the control systems must first develop water quality functions that can be used in predictive models and that can subsequently be used in a feedback control system to make optimal control decisions on additives, filtering, flushing, pressure and temperature adjustment, and other available control actions. This paper focuses on the development of these predictive models. Water flow including its physical properties (temperature, pressure, etc.), its chemistry (chlorine, pH levels, etc.), and its contaminants are simulated using the easily accessible water network modeling software EPANET. Resultant spatial-temporal water quality data from the simulation model was used to develop macro water quality functions using nonlinear regression and machine learning methods. The computational evaluation using supervised learning shows that these models can predict the water quality well.

Keywords
Predictive Models, Nonlinear Regression, Machine Learning, Water Quality Management, Water Network Modeling.

1. Introduction

Commercial and institutional (CI) buildings account for nearly 50% of municipal water use, whereas current water systems have done an inadequate job in balancing the efficient use of water with the health risks of people. Most people spend 8 hours or more per day in CI buildings (e.g., schools, office buildings, hospitals). Moreover, while we have made significant progress optimizing energy use, comfort, and security within CI buildings, we have done an inadequate job in balancing efficient use of water within CI buildings with the risks of people getting sick from drinking or contacting poor quality water (Elfland. et al, 2010; Edwards et.al 2011; Rhoads et al., 2015; Rhoads et al., 2016; Buse et al. 2017). For example, in 2013–2014, over 1000 Americans became ill and 13 died from drinking/contacting poor quality water in CI buildings (Benedict, et al. 2017). Additionally, "In the U.S., reported cases of Legionnaires' disease, which is a serious type of pneumonia caused by bacteria in the water, have increased by nearly four and a half times since 2000" (CDC 2017)]. The source of Legionnaires' disease is the growth of Legionella bacteria in the water system of CI buildings. Moreover, the Center for Disease Control and Prevention (CDC), the U.S. Department of Health & Human Services, and the U.S. Environmental Protection Agency (EPA) require healthcare and assisted-living facilities to implement water management programs to deliver safe water (USEPA 2012); however, the most effective way to do this is not known now. Routine flushing of building water pipes decreases water stagnation, prevents bacteria growth, metal leaching, and elevated copper and lead levels in building water systems – but it wastes water. The above situation motivates the central hypothesis in this paper: a well-designed, quality-monitored, water delivery network within a CI building, appropriately enhanced with data-gathering technologies and connected proactive water quality controls, can sustainably deliver safe water.

1.1 Objectives

Municipal water coming into CI buildings is supposed to be of acceptable quality but the operators of the water delivery in CI buildings have little or no control over its exact quality. For example, the "age "of water depends on how often the building is occupied. Furthermore, the dissolved chemicals (e.g., chlorine and bromides), pH levels, contaminants (e.g., metals such as lead and copper and bacteria such as Legionella) in the incoming water are supposed to be within the regulated safe range for drinking, but the exact quantities depend on the city's water system sources and locations. Therefore, the water quality delivered at the taps and water fountains in the building depends on the quality of water supply coming into the building network, the demand and water usage by water users, and the quality management controls such as flushing, filtration, and chemical injections. The objectives for designing a water quality management system include (1) monitoring and identifying the quality of the water coming into the systems and being delivered to users, and (2) proactively applying available control actions to provide water to users with sufficiently guaranteed water quality. This paper focuses on the first objective, setting aside the feedback control issues for future research. To address this objective, the research team identified the following three sub-objectives or tasks:

A. Based on reported literature develop a microsimulations model (the research team used a commonly accepted simulation software, EPANET (USEPA 2000), for a five-story building being planned) which simulates water flows including its physical properties (temperature, pressure, etc.), its chemistry (chlorine, pH levels, etc.) and contaminants.
B. Use the developed model to synthesize water quality data through the pipes and at the taps, using a few different sources characterized by different water quality functions (the research team used three of them)
C. Conduct data analytics on the collected spatial-temporal water quality data to (i) identify the underlying water quality function (WQF) using machine learning and computational statistics, (ii) approximate the WQF to be used in a look-ahead feedback control policy, and (iii) conduct computational experiments and evaluation of the models developed.

2. Literature Review

Legionella contamination of drinking water in Flint, Michigan is a prime example of the need for better water quality management (Byrne et al. 2019; Garner t al. 2019). Also, it has been shown that Legionella is more resistant to chlorine than most bacteria in water (Kuchta et al. 1983). Even though some might suggest using high amounts of chlorine, Bond et al. (2014) have shown that high levels of chlorine, as well as high levels of bromide, can increase the disinfection byproducts (DBPs) formation, especially the formation of Trihalomethanes (especially THM4). Chowdhury et al. (2011) have linked the high level of THMs in drinking water to human cancer such as bladder and colorectal cancer. There are many DBPs formation models, but very few use large enough datasets that their results are reproducible (Ged et al. 2015). These authors examine different THM4 formation models whether they include bromide, or the ones that do not include bromide on a large dataset to examine the strength of their prediction.

Fisher et al. (2017) discuss the complexity of chlorine decay and associated byproduct formation in drinking water and the need for detailed mathematical or simulation modeling throughout the pipe walls and bulk water for their analyses. In our current model, based on the paper from Georgescu et al (2012), we have assumed the chlorine level of the input water (from the city pipeline to CI) to be equal to the chlorine level of the upstream water treatment facility. In the CI water pipes, on the other hand, the wall-reaction rate is not consistent and has an inverse relationship with the concentration of chlorine (Fisher et al. 2011).

The use of EPANET is common practice for water quality researchers. Kohpaei and Sathasivan (2011) have shown that chlorine decay has an initial fast reaction followed by a slower one. Monteiro et al. (2014) show that modeling the first and nth-order decay kinetics for chlorine decay in EPANET has a sufficient level of accuracy. Monteiro et al. (2014) discuss that these results have been achieved after calibration of the wall decay coefficient which is in line with an aim of this paper which is to use Machine Learning and/or Statistical Methods (MLSM) to develop a macro WQF model that can show the state of the system *x(z,t)*, which represents the state of the water quality (chlorine level, bacteria level, metal contaminants, etc.) at any point z in the water network, at any time *t* given the "control inputs to the system" *u(z,t)* (such as flushing, adjusting water temperature and pressure, chlorine injection, etc.).

3. Methods

To motivate our methods and data collection, it will be useful to preview the architecture of the predictive feedback control system for managing CI building water quality. Figure 1 shows a schematic diagram of the system where *y(z,t)* are the measurements from sensors (this could include periodic manual sampling). This paper focuses on the prediction module. (We remark that the authors are concurrently exploring approaches for designing the control module which will be reported in a later paper.) Observe that some prior information is needed to obtain a good model fit.

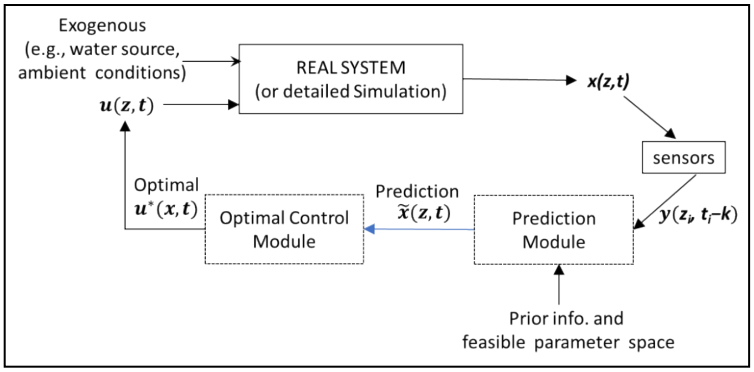


Figure 1: Schematic for the Predictive Control of Water Quality

With the exploding growth of machine learning prediction models, an increase in the use of machine learning models to predict the water quality variables is natural (Muhammad et al. 2015; Haghiabi et al. 2018; Ahmed et al. 2019). For example, Nejjari et al. (2014) use a Genetic Algorithm (GA) to minimize the difference between the model-predicted values and field-measured data, and GA has been used to estimate the unknown parameters by comparing the measured and simulated chlorine concentration at the monitored nodes within the distribution system. Another case of using an advanced machine-learning technique is in Saetta et al. (2021) where gradient boosting machines have been used to predict the chlorine residuals throughout the building plumbing network. In this paper, we study Least Absolute Shrinkage and Selection Operator (***LASSO***) regression, as well as ***non-linear regression***, (**NLR**) to estimate and predict the complex water quality variables' reactions and interactions for the prediction module within the system for a real-time control a building's water quality.

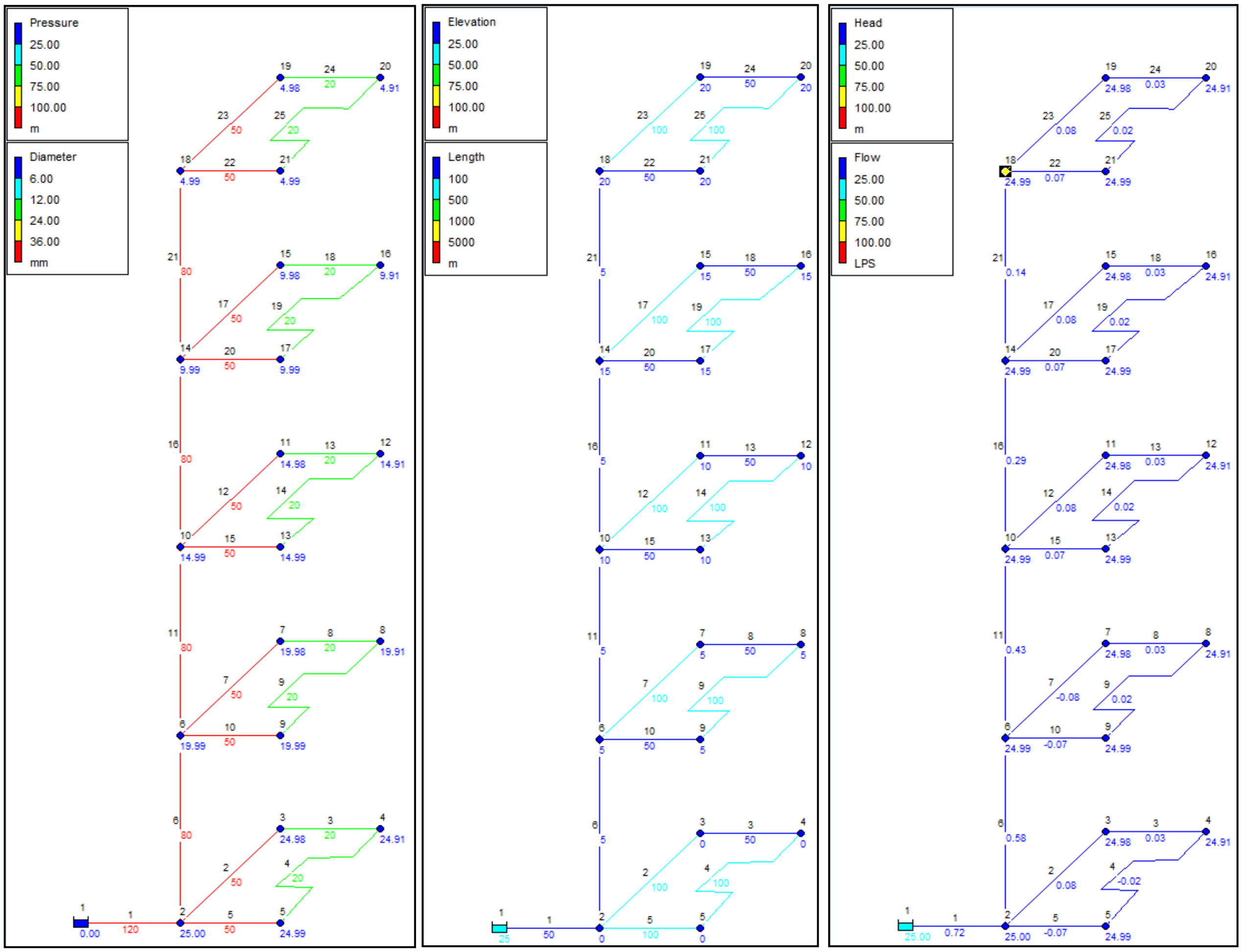
3.1 EPANET Input Model

As mentioned earlier, a base EPANET input model was developed to synthesize water quality data that can be used for ground truth for designing prediction and control modules. This EPANET model is for a five-story building with the most general pipeline properties used in CI buildings. Figure 2 below, shows some of the features of the input model. It has been assumed that each floor has 3 demand points (bathrooms, water fountains, kitchen, etc.). Furthermore, to observe the water quality entering the floor, a junction has been added to the pipeline where it enters the floor. Water demand was assumed cyclical, where it peaks during the day and reaches its minimum after midnight.

3.2 EPANET-MSX Model

EPANET was designed mostly with the hydraulic variables of water flow in mind. To model complex reactions between multiple chemical and biological species in both the bulk flow and at the pipe wall, EPANET Multi-Species eXtension (MSX) was developed as an extension to EPANET, as both a stand-alone executable program as well as a toolkit library of functions that programmers can use to build customized applications. The research team imported the library of functions from the EPANET-MSX toolkit into *Python* software, which enabled it to run multiple simulations and thus provides control over randomly changing the input parameters without changing each file manually. Hence, a large number of scenarios were simulated to collect synthesized data for the development of models that estimate, predict, and control water quality.

The Python code loads an EPANET input file as well as an MSX file and then changes the desired random input parameters. During each EPANET simulation scenario, the input parameters get randomly selected from given distribution based on data and literature sources to create an unbiased water quality data set; *pH, Bromide, DOC, TOC*, *water temperature,* and *water age* coming to the CI building, are some of the input parameters that are changed randomly at each iteration. A Python interface was developed that transfers the EPANET-MSX output data to a Comma-Separated Values (CSV) file, which is standard for many MLSM algorithms. Transferring the data points to a CSV file enables the researchers to remove early data points during the *warm-up period* to use only equilibrated conditions for the model developments to follow.



(a) (b) (c)

Figure 2: (a) Water pressure at the junctions and the pipelines' diameter at the links
(b) Junctions' elevation and pipelines' length (c) Water head at the junctions and water flow at the links

The first water quality function (WQF) that we assumed to characterize a building's water quality, which we will refer to as model WQF-A, is a model offered by Amy et al. (1998) in an EPA published book on DBPs, where THM4 is estimated based on the following chemical and quality variables: chlorine (*Cl_2*), bromide (*Br^-*), *pH*, dissolved organic carbon (*DOC*), temperature (*T*), age or time (*t*). Amy et al.'s (1998) empirical model were developed based on water treatment plant data and even though it captures logical aspects of water chemistry (e.g., increasing the chlorine level, leads to an increase in the THM4 level, etc.), it needs good approximate prior information to produce common THM4 levels observed in CI buildings.

$$THM4 = 0.0412(DOC)^{1.10}(Cl_2)^{0.152}(Br-)^{0.068}(T)^{0.61}(pH)^{1.60}(t)^{0.260} \qquad \text{(WQF-A)}$$

The second WQF we experimented with, which we will refer to as model WQF-B, is a model offered by Harrington et al. (1992), which is based on a very robust database from several water sources in Amy et al. (1987) with the many of the same water chemical and quality variables in WQF-A model plus a variable for Ultraviolet absorption at 254 nm wavelength (*UV254)* and uses total organic carbon (*TOC*) which, as mentioned before, includes 90% *DOC*.

$$THM4 = (0.00309 * 252)\,(TOC\,.\,UV254)^{0.44} * (Cl_2)^{0.409} * \,(t)^{0.265} * \,(T)^{1.06}\,(pH - 2.6)^{0.715}\,(Br - \,+\, 1)^{0.03} \quad \text{(WQF-B)}$$

3.3 Supervised Learning Models Selection

The goal of supervised learning is to train a model so that it can predict the output when it is given new data. In this research, LASSO regression and NLR were studied to estimate and predict the complex water quality variables. Even though other supervised learning models such as decision tree regression, and random forest regression have also been used to compare the quality of estimations, the later models are not considered here since for feedback control approach (that is currently being designed) requires algebraic forms. To test the effectiveness of any model fit, the synthesized data was first partitioned into two sets in an 8:2 ratio, where the larger set referred to as "*training data*" was used to train the model and smaller set to test the model which we refer to as "*test data*" below. Next, the R^2 level of fit was computed to evaluate the goodness of fit of the predicted data in comparison with input data (a good fit mean R^2 is very close to 1).

LASSO regression is a modification of linear regression that includes a penalty term (α) (which is decided by the modeler) for the sum of absolute values of the coefficients. Since the underlying functions and equations in water quality systems are very complex, LASSO regression offers a simple and fast computable prediction model, making it very attractive for real-time control applications (Kley-Holsteg and Ziel 2020). The LASSO regression model used on WQF-A synthesized data for WQF-A first suggested the removal of chlorine (Cl_2) as a regressor due to its high correlation with the age (t) variable. The LASSO regression used on WQF-B shows similar results, and it also suggested removal of Bromide (Br^-) when penalty α was made higher.

NLR can create more complex models which can be used to predict some unobservable water quality variables in real-world data, such as the age of the water. In addition, NLR can be used to calibrate the parameters of its "base model" that is required as prior information to narrow the search for the underlying coefficients (we discuss this further in section 5 on results). Also, due to the complexity of NLR, the search for best-fit parameters often requires a prior range for valid estimated parameters as well as an initial guess (seed); NLR then computes the near-best or optimal parameters that fit the data assuming the given base form. Since the water chemical and quality are correlated (e.g. the chlorine decay has time as its independent variable), a hierarchical three-phase approach has been developed. In *Phase 1*, we estimate the first multiplicative coefficient in the WQF because, during the simulation phase, we observed that this coefficient, which normally is not well estimated in the WQF models, had to be adjusted to produce reasonable results. Then keeping the coefficient fixed, in *Phase 2*, we estimate the initial water age (t) as well as its exponent since water age is affected by both the initial age of input water and stagnation of water in the CI building. Finally, in *Phase 3*, we estimate the test of values of the other parameters in the WQF.

4. Data Collection

Currently, there are no available datasets that provide information on all the end-user water variables in CI buildings. Therefore, the **input** values used in the EPANET-MSX file were randomly selected from distributions based on some real data as well as given Box and Whisker (BW) plots from literature (see, e.g., Figure 3). In the following subsections, we will discuss each distribution for input water variables.

4.1 pH Distribution

To find distribution for pH in CI buildings, we have tried fitting a normal distribution N(8, 0.197) to our measurement data (376 data points) from 11 different buildings at Arizona State University using the ARENA input analyzer. Based on the Kolmogorov–Smirnov test we accepted the hypothesis that the collected pH data approximately follows a normal distribution N(8, 0.197).

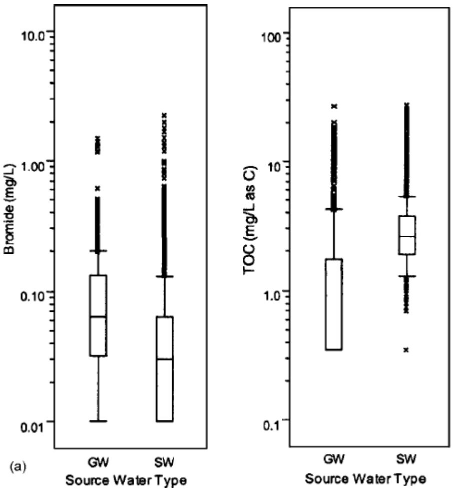


Figure 3: Bromide and TOC Levels in Ground Water (GW) and Surface Water (SW) BW plot (Obolensky et al. 2007)

4.2 Bromide (mg/L) Distribution

According to Obolensky et al.'s (2007) dataset on water treatment plans, the mean and median of the Surface Water (SW) bromide levels is 0.06 and 0.03, respectively (note that we assumed that our CI buildings have SW sources). Since there is no actual distribution given in the paper for the SW bromide levels, but only the BW plots (Figure 3), a random distribution with the same statistical attributes as the BW plot was developed and sampled for the simulation runs.

4.3 DOC (mg/L) Distribution

In the literature, there are no data set for Dissolved Organic Carbon (DOC) in CI buildings, but experts believe that Total Organic Carbon (TOC) contains about 90% DOC. Therefore, the TOC distribution based on the BW plot (Figure 4) from Obolensky et al (2007), multiplied by a factor of 0.9, was used for the distribution of the random inputs in the simulations.

5. Results and Discussion

5.1 EPANET-MSX Simulation Results

After running the EPANET-MSX simulation for 20 random scenarios with both models (WQF-A and WQF-B) we synthesized 722840 data points for each WQF model. The synthesized data appeared well behaved based on observed trends such as (a) the prevailing chlorine levels increase and the THM4 levels decrease during the high demand hours when stagnation is low, and conversely, and (b) the chlorine levels decrease and the THM4 levels increase (THM4 formation) during the least demand hours, i.e., after midnight.

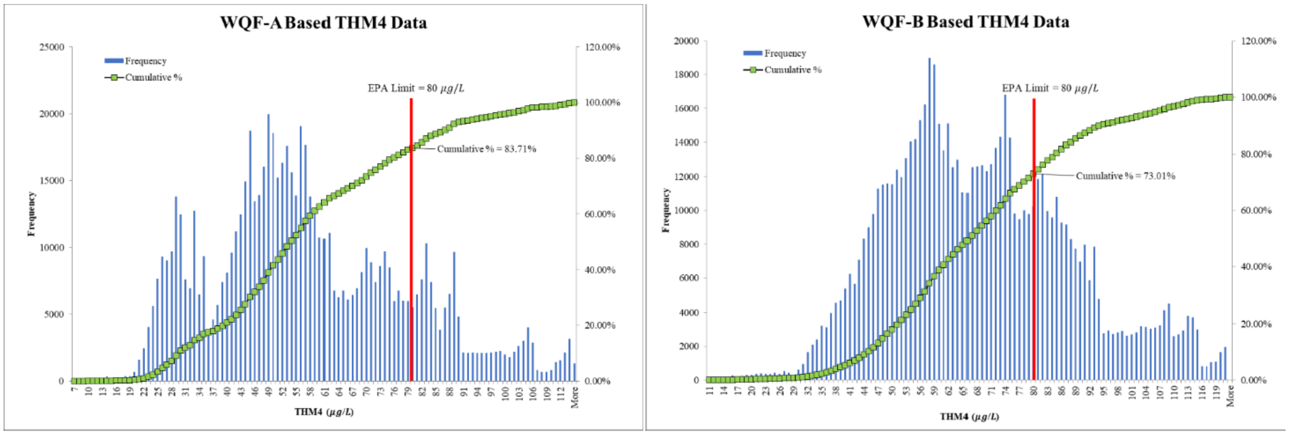


Figure 4: The THM4 Histogram for synthesized data based on WQF-A and WQF-B models

Given that all the input parameters were based on distributions from real-world data, the synthesized data is a good representation of the water quality in CI buildings. From Figure 4, we can see that 16.29% and 26.69% of the water delivered at the demand points is higher than the acceptable limits imposed by EPA. This further **motivates** the need for developing a predictive control system that assures THM4 to always be within safe THM4 limits for drinking.

5.2 LASSO Regression Results

As mentioned before, the LASSO regression model is based on linear regression that includes a penalty (α) for the sum of absolute values of the regression coefficients. The following results (see Figure 5) are based on the test data (144568 data points which are 20% of the total synthesized data points for WQF-A).

The LASSO regression models for WQF-A, for various α penalties, plotted as dots in Figure 5, are as follow:

(a) $THM4 = 130 \ + 42.3 * Cl_2 + 2.27 * T - 2.76 * \mathrm{t} + 8.08 * pH + 0.111 * Br^{-} + 28.7 * DOC$
(with α $=$ 0.0001) (R^2 = 0.97057)
(b) $\mathrm{THM4} = \ 794 \ + \ \ 0 \ \ * Cl_2 + \ 2.2 \ * \mathrm{T} - 9.56 * \mathrm{t} + 8.05 * pH \ + 0.111 * Br^{-} + 28.7 * DOC$
(with α $=$ 0.0001) (R^2 = 0.9705)
(c) $\mathrm{THM4} = \ 749 \ + \ \ 0 \ \ * Cl_2 + 2.17 * T - 9.05 * t + 7.68 * pH \ + \ 0.11 \ * Br^{-} + 28.7 * DOC$
(with α $=$ 0.0001) (R^2 = 0.97047)
(d) $\mathrm{THM4} = \ 298 \ + \ \ 0 \ \ * Cl_2 + 1.89 * T - 3.95 * t + 4.01 * pH \ + 0.103 * Br^{-} + 28.5 * DOC$
(with α $=$ 0.0001)(R^2 = 0.96809)

The R^2 values indicating the measure of fit for our selected α levels are quite high. For α level 0.001 and above the LASSO regression detects the high correlation between chlorine level and water age (t) and hence we may select model (b) which gives R^2 = 0.9705 while estimating the coefficient of chlorine as 0.

We repeated this experiment WQF-B and applied LASSO regression on WQF-B test data (144568 data points which are 20% of the total synthesized data points for WQF-B). The LASSO fits results were similar with high R^2 values (plotted points were similar to Figure 5) are given below:
(a) $\mathrm{THM4} = 1590 + 33.6 * \mathrm{Cl_2} + 4.3 * \mathrm{T} - 17.8 * \mathrm{t} + 12.6 * \mathrm{pH} + 8.52 * \mathrm{Br^-} + 12.4 * \mathrm{TOC} + 297 * \mathrm{UV254}$
(with $\alpha = 0.0001$) ($R^2 = 0.97405$)
(b) $\mathrm{THM4} = 2140 + 0 * \mathrm{Cl_2} + 4.04 * \mathrm{T} - 23.1 * \mathrm{t} + 12.5 * \mathrm{pH} + 7.03 * \mathrm{Br^-} + 12.4 * \mathrm{TOC} + 296 * \mathrm{UV254}$
(with $\alpha = 0.001$) ($R^2 = 0.97400$)
(c) $\mathrm{THM4} = 2100 + 0 * \mathrm{Cl_2} + 4.24 * \mathrm{T} - 22.6 * \mathrm{t} + 11.2 * \mathrm{pH} + 8.52 * \mathrm{Br^-} + 12.4 * \mathrm{TOC} + 281 * \mathrm{UV254}$
(with $\alpha = 0.01$) ($R^2 = 0.97335$)
(d) $\mathrm{THM4} = 1700 + 0 * \mathrm{Cl_2} + 3.93 * \mathrm{T} - 17.6 * \mathrm{t} + 0.131 * \mathrm{pH} + 0 * \mathrm{Br^-} + 14.5 * \mathrm{TOC} + 115 * \mathrm{UV254}$
(with $\alpha = 0.1$) ($R^2 = 0.97057$)

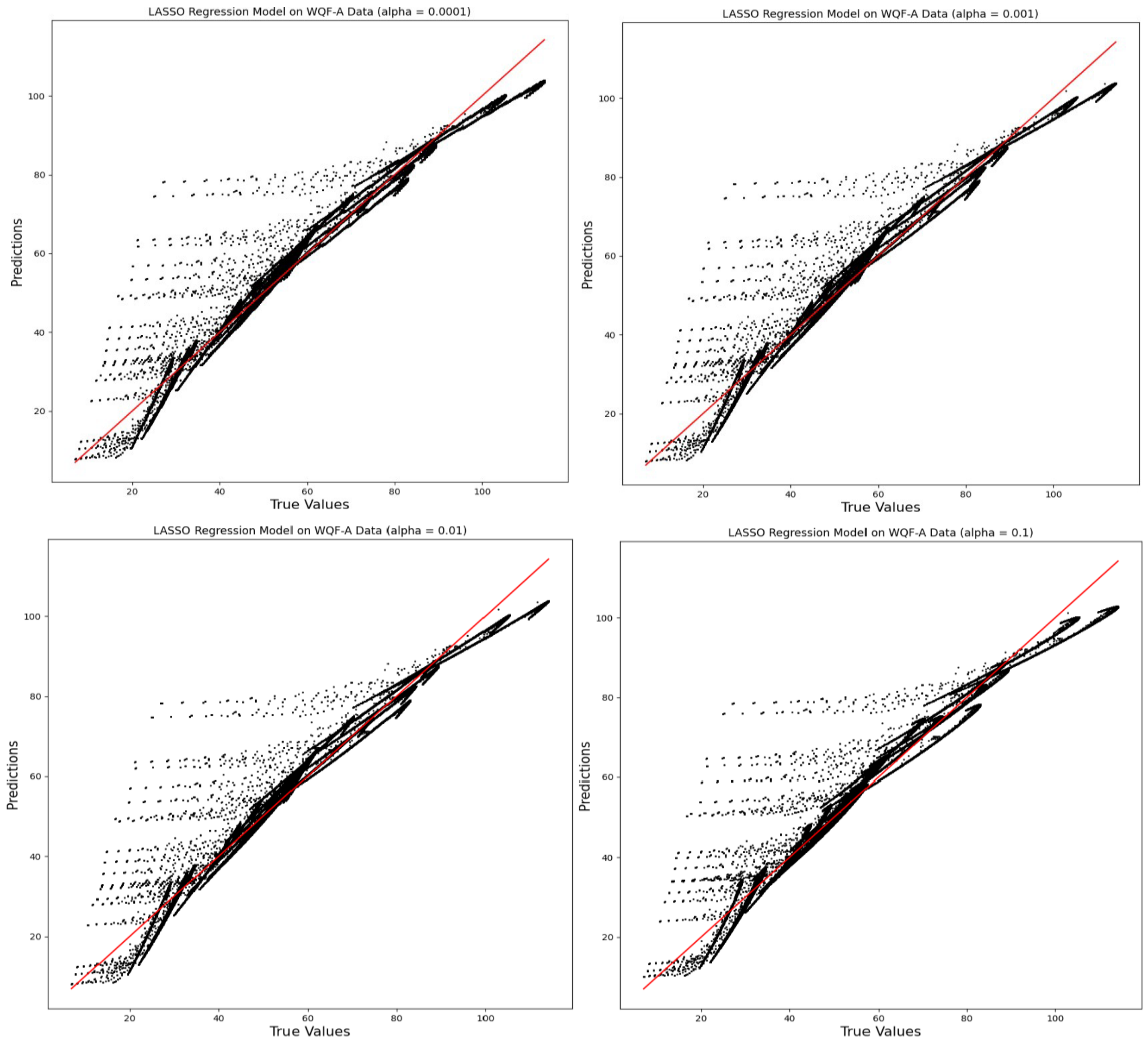


Figure 5: Comparing Lasso regression prediction for WQF-A with perfect prediction line at 45° (see red line)

It is interesting to point out that LASSO for both WQF-A and WQF-B removes the highly correlated regressor, chlorine when $\alpha > 0.0001$. Moreover, in WQF-B with $\alpha = 0.1$, the coefficient of bromide is estimated as zero implying that bromide (Br^-) does not impact THM4 in this case.

5.3 Non-Linear Regression (NLR) Results

As discussed earlier, we can use NLR to estimate and calibrate the parameters of the WQFs. The age (*t*) of the water flowing into the building is a water quality variable that cannot be obtained from a sensor and it is very complex to estimate or measure. In this section, we first will try to obtain a prior estimate of the initial water age which, for the EPANET simulations was drawn from a normal distribution with a mean of 100 hours. The goal is to converge to the value of the initial water age that is known to us (since we have used it to create the synthesized data), but it is unknown to the NLR model. Due to the complexity of the feasible region for finding the optimal *K* values, NLR requires a base equation form, seed (an initial guess), and lower and upper bounds on the parameters that need to be estimated. The seed (except for the seed of initial water age which can be arbitrary), was chosen as the original value of the parameters proposed in WQFs. The upper and lower limits of the parameters were set within 50% of the seed, except for Br^{-} which is modeled quite differently in WQF-A and WQF-B. Finally, since some water chemical and quality variables are highly correlated, we developed and applied the three-phase approach as described in subsection 3.3.

5.3.1 WQF-A Based Non-Linear Regression Model

In this subsection, the non-linear regression has the prior base form model for WQF-A was

$$THM4 = K_1 * (DOC)^{K_2} * (Cl_2)^{K_3} * (Br-)^{K_4} * (T)^{K_5} * (pH)^{K_6} * (t + K_7)^{K_8}$$

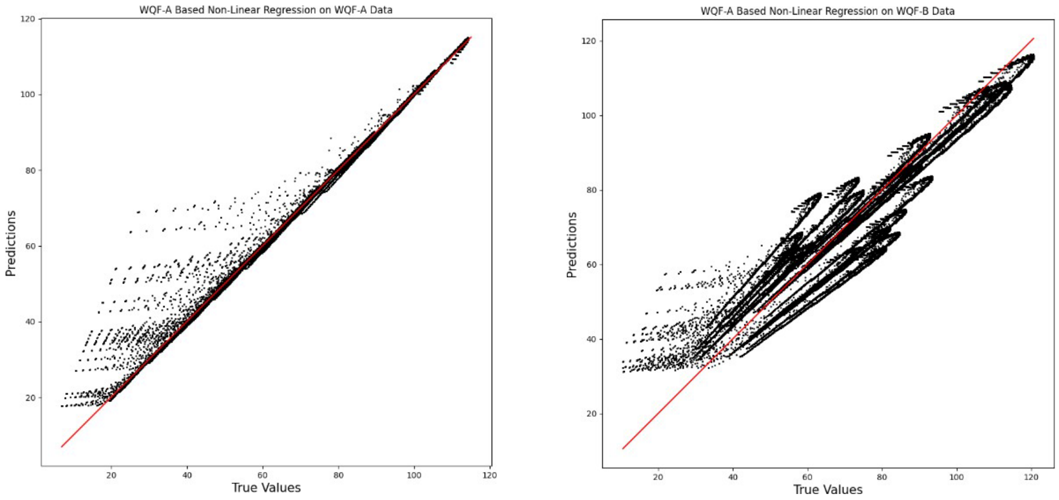
where the K_i parameters need to be estimated from data. To check the model adequacy of WQF-A based non-linear regression, we have trained the model on WQF-A synthesized data, and tested the model on test data as shown in Figure 6(a). In *Phase 1*, we estimated the first coefficient K_1 regression as 0.0406 which is very close to the input coefficient of 0.0412 for WQF-A. Fixing the estimated K_1, in *Phase 2*, we estimated the age $t + K_7$ of the input water to the building and its exponent K_8; estimates of K_7 and K_8 were 84.3 and 0.27, respectively. The random input used in the simulation was $t = 94.9$ hours, and exponent K_8 was 0.27, and this we note these estimation errors related to age are negligible. Finally, in *Phase 3*, we estimated the other parameters of the WQF-A giving us the estimate:

$$THM4 = 0.0406 * (DOC)^{1.1} * (Cl_2)^{0.2} * (Br-)^{0.0674} * (T)^{0.675} * (pH)^{1.53} * (Age + 84.3)^{0.27} \ (R^2 = 0.99291)$$

Next, to check the validity of the above WQF-A base, we trained it on WQF-B synthesized data using the base of WQF-A model (recall there is no $UV254$ in WQF-A base) and obtained the following fitted model (Figure 6).

$$THM4 = 0.042 * (TOC)^{0.664} * (Cl_2)^{0.228} * (Br-)^{0.00229} * (T)^{0.915} * (pH)^{1.51} * (Age + 84.2)^{0.27} \ (R^2 = 0.84453)$$

Thus, the WQF-A based NLR performs quite well as its R^2 is quite high. The estimation for initial water age is 84.2 is close to the value of 99.8 used in the WQF-B simulation. Moreover, NLR correctly detects the insignificance of bromide (Br^-) and estimates the exponent of bromide as 0.00229, which is close to the exponent 0.03 used for input WQF-B simulations. Note that there is a linear relationship between DOC and TOC; it has been observed that TOC contains about 90% DOC, meaning that the two values can be used interchangeably as long as the ratio is preserved in the calculations.



(a)

(b)

Figure 6: (a) Prediction results of WQF-A based NLR on WQF-A synthesized data ($R^2 = 0.99291$)
(b) Prediction results of WQF-A based NLR on WQF-B synthesized data ($R^2 = 0.84453$)

5.3.2 WQF-B Based Non-Linear Regression Model

In this subsection, the NLR has the base form equation fromWQF-B:

$$THM4 = K_1 * (TOC)^{K2} * (UV254)^{K3} * (Cl_2)^{K4} * (Age + K_5)^{K6} * (T)^{K7} * (pH - 2.6)^{K8} * (Br^- + 1)^{K9}$$

As before, we trained WQF-B based NLR on WQF-B synthesized data and then evaluated the prediction on the test data (see Figure 7a). In Phase 1, we estimated coefficient K_1 as 0.772 which is very close to the input coefficient value of 0.77868. Fixing K_1 the estimated coefficients from Phase 2, were K_5 (for age) and its exponent K_6, as 91.8 and 0.27, respectively. These are close to the random inputs for the synthesized data where K_5 was 99.8 hours and K_6, was 0.27. Finally, in Phase 3, we estimated the other parameters of the WQF-B as shown below.

$$THM4 = 0.772 * (TOC)^{0.438} * (UV254)^{0.441} * (Cl_2)^{0.434} * (Age + 91.8)^{0.27} * (T)^{1.1} * (pH - 2.6)^{0.662} * (Br^- + 1)^{1.2e-18} \quad (R^2 = 0.99307)$$

Next, to check the validity of the base form of WQF-B, we trained it on synthesized WQF-A data where we assumed the mean value of *UV254* from Obolensky et al. (2007) (recall there is no *UV254* in the synthesized WQF-A data) and tested it on WQF-A test data (see as shown in Figure 7(b). Using the same three-phase approach we have:

$$THM4 = 0.652 * (DOC)^{1.13} * (UV254)^{0.319} * (Cl_2)^{0.205} * (Age + 96)^{0.267} * (T)^{0.682} * (pH - 2.6)^{0.735} * (Br^- + 1)^{0.045} \quad (R^2 = 0.98855)$$

The WQF-B based NLR performs extremely well on WQF-A data, with R^2 almost as high as the prediction for WQF-B synthesized data. The estimation for initial water age is 96 which is very close to the input value 94.9 used for the WQF-A simulation. Moreover, the NLR correctly detects the insignificance of bromide (Br^-) in WQF-B and estimates the exponent of bromide as 0.045 which is close to the exponent of bromide in WQF-A, 0.068.

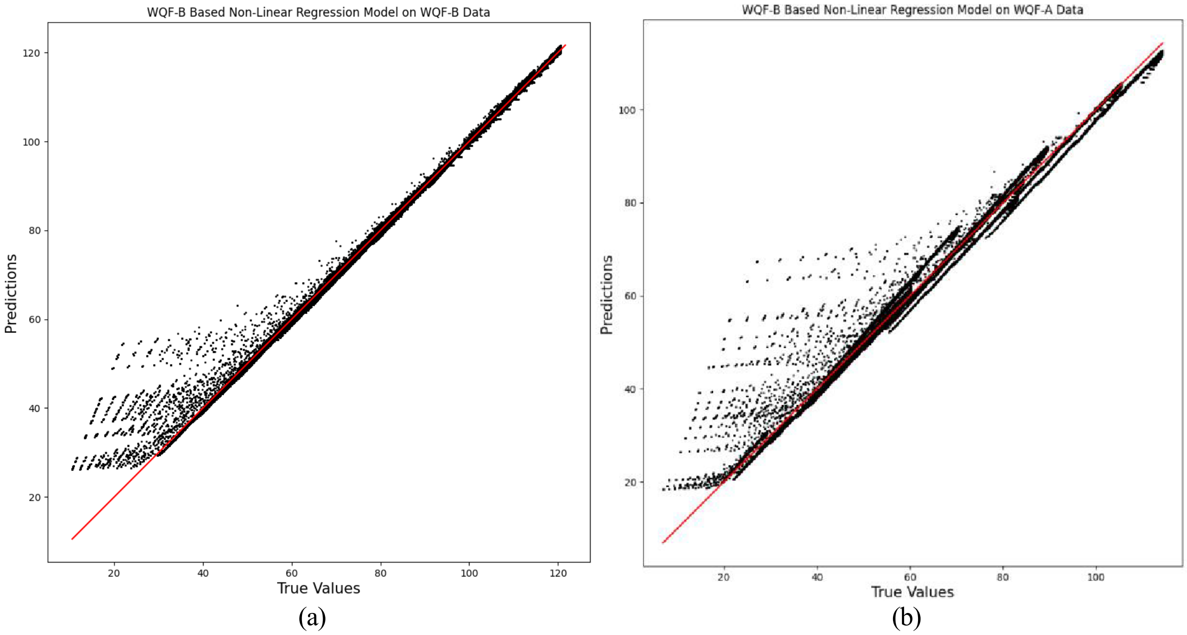


Figure 7: (a) Prediction results of WQF-B based NLR on WQF-B synthesized data ($R^2 = 0.99307$)
(b) Prediction results of WQF-B based NLR on WQF-A synthesized data ($R^2 = 0.98855$)

6. Conclusions

In this paper, we have developed a method to use sensor data to predict the water quality at the outputs (taps and water fountains) of CI water networks. The primary goal of our approach is to first identify the quality of the water coming into the CI building, and then estimate and predict the expected quality of water at the pipes' outputs. Our eventual goal is to use this method to develop a predictive feedback control scheme to determine controls such as flushing, adjusting water temperature and pressure, filtration, chlorine injection, etc., to assure that water is of acceptable quality at the pipe network outputs.

From extensive literature on the normal contents and chemistry of drinkable water in building networks, we hypothesized a water quality function (WQF) which we referred to as WQF-A. The literature develops many such models based on data and curve-fitting of the data each with an associated R^2 measure of fit. Developing expressions for water contents and chemistry in both surface water and ground water tanks, city pipes, and buildings' pipes are currently active research among environmental engineering research. We simply assumed one of them which included chemical and physical properties of interest in measuring the quality of drinking water. To test the methodology developed, we assumed another WQF, which we referred to as WQF-B, again based on our review of literature in this field.

A microsimulation model for a five-story water flow, using the well-accepted modeling software EPANET, was developed that simulated the water flow in the building with associated physical properties (temperature, flow rates, pressure, etc.), its chemistry (chlorine, pH levels, THM, TOC, etc.) and metal compounds contaminants. In our first set of experiments, we characterized water quality using WQF-A. It was clear that due to the high correlation of some variables in WQF-A the fitting of data using single-pass Machine Learning and/or Statistical Methods (MLSM) approach was unstable and highly dependent on the "seed" for starting the optimization process needed for the curve fitting. Hence we developed a *three-phase approach* where at Phase 1 we estimated the first multiplicative coefficient K_1 of the WQF. Keeping this estimated coefficient fixed, we then estimated in Phase 2 the age of in-flow water characterized by $(t + K_7)^{K_8}$. Finally, in Phase 3 we estimated the values of the other parameters in the WQF. The computational experiments indicated that our estimates were close to what the actual inputs were in our simulations, with an R^2 of approximately 0.97 to 0.99.

We repeated these computational experiments using input characterization WQF-B. This allowed the research team to conclude that the NLR method indeed is successful in estimating the WQF. Since it is envisioned that in our predictive feedback control system, currently being developed and tested, we need to look ahead only for short times (in 1-2 hours) we compared these estimation methods with a commonly used approach in machine learning, *LASSO Regression*. Computational experiments with LASSO regression are also evaluated in this paper. LASSO also performed quite well, fitting with R^2 close to 0.97.

Previewing our framework for the eventual development of a real-time predictive feedback control system for water in CI Buildings, our future research directions include (1) Development of acceptable water quality characterization which can be envisioned as a *target subspace* in the water quality state, defined by ranges of variables of the contents, physical attributes and chemistry of the drinking water, (2) Development of a predictive feedback control schemes to guarantee that water at the outputs in the target subspace, and (3) Evaluation of the developed predictive feedback control system, in simulations and through pilot testing.

Acknowledgments

We thank Derek Hogue (Ph.D. student of the Environmental Engineering Program at Arizona State University, ASU) for assisting with the literature review and sharing some chemical aspects of water quality. We also thank Katelin Aden (undergraduate Environmental Engineering at ASU) for sharing data from her data collection efforts. We acknowledge the research support of the NSF grant with Award Number: 2027444.

References

Ahmed, A. N., Othman, F. B., Afan, H. A., Ibrahim, R. K., Fai, C. M., Hossain, M. S., & Elshafie, A.. Machine learning methods for better water quality prediction. *Journal of Hydrology*, 578, 124084. (2019)

Amy, G. L., Chadik, P. A., & Chowdhury, Z. K. Developing models for predicting trihalomethane formation potential and kinetics. *Journal-American Water Works Association*, *79*(7), 89-97. (1987)

Amy, G.L., Siddiqui, M., Ozekin, K., Zhu, H.W., Wang, C., *Empirical based models for predicting chlorination and ozonation byproducts: Haloacetic acids, chloral hydrate, and bromate*. Report cx 819579. U.S. Environmental Protection Agency, Washington, D.C. 1998

Benedict, K. M.; Reses, H.; Vigar, M.; Roth, D. M.; Roberts, V. A.; Mattioli, M.; Cooley, L. A.; Hilborn, E. D.; Wade,
T. J.; Fullerton, K. E., Surveillance for waterborne disease outbreaks associated with drinking water—United States, 2013–2014. *MMWR. Morbidity and mortality weekly report* 2017, *66*, (44), 1216.

Bond, T., Huang, J., Graham, N. J., & Templeton, M. R. Examining the interrelationship between DOC, bromide and chlorine dose on DBP formation in drinking water—a case study. *Science of the Total Environment*, 470, 469- 479 (2014).

Buse, H. Y.; Ji, P.; Gomez-Alvarez, V.; Pruden, A.; Edwards, M. A.; Ashbolt, N. J., Effect of temperature and colonization of Legionella pneumophila and Vermamoeba vermiformis on bacterial community composition of copper drinking water biofilms. *Microbial Biotechnology* 2017, *10*, (4), 773-788.

Byrne, B. G., McColm, S., McElmurry, S. P., Kilgore, P. E., Sobeck, J., Sadler, R., and Swanson, M. S. Prevalence of infection-competent serogroup 6 Legionella pneumophila within premise plumbing in Southeast Michigan. *MBio*, *9*(1), e00016-18. (2018).

CDC, *Developing a Water Management Program to Reduce Legionella Growth & Spread in Buildings: A Practical Guide to Implementing Industry Standards*. 2017.

Chowdhury, S., Champagne, P., & McLellan, P. J. Models for predicting disinfection byproduct (DBP) formation in drinking waters: a chronological review. Science of the Total Environment, 407(14), 4189-4206. (2009).

Chowdhury, S., Rodriguez, M. J., & Sadiq, R. Disinfection byproducts in Canadian provinces: associated cancer risks and medical expenses. *Journal of hazardous materials*, 187(1-3), 574-584. (2011).

Edwards, M.; Parks, J.; Griffin, A.; Raetz, M.; Martin, A.; Scardina, P.; Elfland, C., Lead and Copper Corrosion Control in New Construction. *Water Research Foundation* 2011.

Elfland, C.; Scardina, P.; Edwards, M., Lead-contaminated water from brass plumbing devices in new buildings.
American Water Works Association. Journal 2010, *102*, (11), 66.

Fisher, I., Kastl, G., Sathasivan, A., & Jegatheesan, V. Suitability of chlorine bulk decay models for planning and management of water distribution systems. *Critical Reviews in Environmental Science and Technology*, *41*(20), 1843-1882. (2011).

Fisher, I., Kastl, G., & Sathasivan, A. New model of chlorine-wall reaction for simulating chlorine concentration in drinking water distribution systems. *Water Research*, *125*, 427-437. (2017).

Garner, E., Brown, C. L., Schwake, D. O., Rhoads, W. J., Arango-Argoty, G., Zhang, L., ... & Pruden, A. Comparison of Whole-Genome Sequences of Legionella pneumophila in Tap Water and in Clinical Strains, Flint, Michigan, USA, 2016. *Emerging infectious diseases*, *25*(11), 2013. (2019).

Ged, E. C., Chadik, P. A., & Boyer, T. H. Predictive capability of chlorination disinfection byproducts models. *Journal of Environmental Management*, *149*, 253-262. (2015).

Georgescu, A. M., & Georgescu, S. C. Chlorine concentration decay in the water distribution system of a town with 50000 inhabitants. University" Politehnica" Of Bucharest Scientific Bulletin, Series D: Mechanical Engineering, 74(1), 103-114. (2012).

Harrington, G. W., Chowdhury, Z. K., & Owen, D. M. Developing a computer model to simulate DBP formation during water treatment. *Journal-American Water Works Association*, *84*(11), 78-87. (1992).

Haghiabi, A. H., Nasrolahi, A. H., & Parsaie, A. Water quality prediction using machine learning methods. *Water Quality Research Journal*, 53(1), 3-13. (2018).

HHS, Requirement to reduce Legionella risk in healthcare facility water systems to prevent cases and outbreaks of Legionnaires' Disease (LD). *S&C 17-30-Hospitals/CAHs/NHs* 2017, *revised 06.09.2019*,

Kley-Holsteg, J., & Ziel, F. Probabilistic multi-step-ahead short-term water demand forecasting with Lasso. *Journal of Water Resources Planning and Management*, *146*(10), 04020077. (2020).

Kohpaei, A. J., & Sathasivan, A. (2011). Chlorine decay prediction in bulk water using the parallel second order model: An analytical solution development. *Chemical Engineering Journal*, *171*(1), 232-241.

Kuchta, J. M., States, S. J., McNamara, A. M., Wadowsky, R. M., & Yee, R. B. Susceptibility of Legionella pneumophila to chlorine in tap water. *Applied and Environmental Microbiology*, *46*(5), 1134-1139. (1983).

Monteiro, L., Figueiredo, D., Dias, S., Freitas, R., Covas, D., Menaia, J., & Coelho, S. T. Modeling of chlorine decay in drinking water supply systems using EPANET MSX. *Procedia Engineering*, *70*, 1192-1200. (2014).

Muhammad, S. Y., Makhtar, M., Rozaimee, A., Aziz, A. A., & Jamal, A. A. Classification model for water quality using machine learning techniques. *International Journal of software engineering and its applications*, 9(6), 45-52. (2015).

Nejjari, F., Puig, V., Pérez, R., Quevedo, J., Cugueró, M. A., Sanz, G., & Mirats, J. M. Chlorine decay model calibration and comparison: application to a real water network. *Procedia Engineering*, *70*, 1221-1230. (2014).

Obolensky, A., Singer, P. C., & Shukairy, H. M. Information collection rule data evaluation and analysis to support impacts on disinfection by-product formation. *Journal of Environmental Engineering*, *133*(1), 53-63. (2007).

Rhoads, W. J.; Ji, P.; Pruden, A.; Edwards, M. A., Water heater temperature set point and water use patterns influence Legionella pneumophila and associated microorganisms at the tap. *Microbiome* 2015, *3*.

Rhoads, W. J.; Pruden, A.; Edwards, M. A., Survey of green building water systems reveals elevated water age and water quality concerns. *Environmental Science-Water Research & Technology* 2016, *2*, (1), 164-173.

Saetta, D., Richard, R., Leyva, C., Westerhoff, P., & Boyer, T. H. Data-mining methods predict chlorine residuals in premise plumbing using low-cost sensors. *AWWA Water Science*, *3*(1), e1214. (2021).

USEPA, WaterSense at Work: Best Management Practices for Commercial and Institutional Facilities. *Office of Water*
2012, *EPA 832-F-12-034*, 308 pp.

USEPA, *EPANET [website]*. https://www.epa.gov/WATER-RESEARCH/EPANET: accessed 03/05/2017, 2000.

Biographies

Kia Zadeh is a Ph.D. student in the Industrial Engineering Program in the School of Computing and Augmented Intelligence at Arizona State University. He earned his B.S. in Industrial Engineering from the University of Tehran. Kia has researched on scheduling and routing of Unmanned Autonomous Vehicles (UAVs) using a network modeling approach. His research interests include simulation, optimization, machine learning, and applied statistics.

Treavor Boyer [Ph.D. and MS, Environmental Engineering, University of North Carolina, Chapel Hill] is a Professor, in the School of Sustainable Engineering and the Built Environment, Arizona State University. He is the Program Chair of the Environmental Engineering in the Ira A. Fulton Schools of Engineering. His research is broadly focused on water sustainability and spans drinking water and wastewater treatment, and natural aquatic systems. His engineering passion is to develop robust approaches to the treatment of water at various stages in its lifecycle to maximize water conservation, recover valuable materials, sequester harmful contaminants, minimize the production of waste byproducts, and advance the water–energy–food nexus. He strongly supports taking a systems-thinking approach to water quality and treatment that considers global drivers such as urbanization, climate change, biogeochemical cycles, sustainable engineering, and disruptive innovation. In 2012, he earned a National Science Foundation CAREER award for his efforts toward using urine source separation and treatment as catalysts for new research directions in wastewater treatment and resource recovery. Other research interests include innovative ion exchange treatment and regeneration for small drinking water systems, and impacts of sea-level rise and seawater intrusion on drinking water treatment and disinfection.

Pitu Mirchandani [SM and ScD in Operation Research, MIT] is a Professor of Computing and Augmented Intelligence, and AVNET Chair of Supply Chain Networks, in the Ira A. Fulton Schools of Engineering at the Arizona State University. He is the Director of the Advanced Traffic and Logistics Algorithms and Systems Laboratory (ATLAS), the Chief Scientist of the DHS Center for Accelerating Operations Efficiency, and a Senior Global Futures Scientist, ASU's Julie Ann Wrigley Global Futures Laboratory. Mirchandani has an extensive background in Optimization (including network optimization, AI, and complexity guaranteed heuristics), Predictive Analytics (including applied statistics, estimation methods, and machine learning), and Stochastic Control. He has used his expertise to develop concepts of Dynamic Stochastic Networks and their management and control, and is interested in developing models and systems for making strategic/tactical/operational decisions in stochastic networked environments, with applications typically related to transportation and logistics. He is a Fellow of INFORMS and a Fellow of IEEE.