Spatial Demand Estimation: Moving Towards Real-Time Distribution System Network Modeling

1 Problem and Research Objectives

Water utilities must ensure that our water infrastructure is sustainable, robust and resilient to both long- and short-term forcing factors. While long-term factors (e.g., climate change, population shifts) are likely to be addressed through changes in infrastructure design, short-term factors (e.g., intrusion events, main breaks) can be addressed through real-time monitoring and decision support. Unfortunately, existing "real-time" decision support tools that, for example, are intended to assist with pump scheduling, are limited in practice as they require real-time estimates of the current and/or future states of the system (e.g., flow rates, water quality concentrations, etc), which are only observed at limited locations throughout the network. To compensate for the lack of observed data, estimates of the user demands – the driving factors for the underlying hydraulic and water quality dynamics – are required to simulate the system-wide states through a network model. Recent software developments have begun to integrate observed data with network models (e.g., IWLive [Innovyze]; Polaris [CitiLogics]), but only include simplistic demand estimation approaches and limited (if any) forecasting capabilities. Thus, there remains a critical need for real-time demand estimation and forecasting to fill the gap between data-model integration and the development of real-time decision support tools. The objective of this project, which is the next step in progressing towards our long-term goals, is to develop a composite demand-hydraulic model – one that couples a demand model with a network hydraulic solver – capable of being updated in real-time using observed hydraulic information.

2 Methodology

Our central hypothesis is that the observed hydraulic data commonly collected via utility SCADA systems can be used to estimate the expected values and uncertainty of a structured demand model that characterizes the temporal and spatial patterns of consumptive demands. Our rationale for developing a composite demand-hydraulic model that can be updated in real-time is to provide the framework for forecasting temporally and spatially correlated demands and the associated network hydraulics. In turn, these capabilities will provide for the development of real-time decision making tools associated with, for example, optimal pump scheduling to minimize energy costs. We will test our central hypothesis and associated objectives by pursuing the following: 1) the development of a composite demand-hydraulic model that will integrate a vectorized times series model with a network hydraulic solver; 2) the implementation of an expectation-maximization algorithm to estimate the demands and model parameters using limited observed hydraulic information; and 3) develop a clustering approach, based on water quality information, to reduce the parameterization of the demand estimation problem.

2.1 *Composite Demand-Hydraulic Model.* The proposed composite demand-hydraulic model will be formulated as a Dynamic Bayesian Network (DBN) – a generic framework

for representing complex conditional probabilistic models (Ghahramani, 1998; Koller and Friedman, 2009) – by integrating a time series demand model with a distribution system network hydraulic solver for estimating the hydraulic states (e.g., flow rate, pressure, tank levels, etc) of a distribution system. Figure 1 illustrates the linkage of the variables (boxed) and sub-models, and the conditional relationships of the DBN as the hydraulic states of the network are conditioned upon the distribution of demands, which are themselves conditioned on the demand model parameter estimates. The following further describes the demand and hydraulic sub-models with the estimation algorithm presented in the next section.

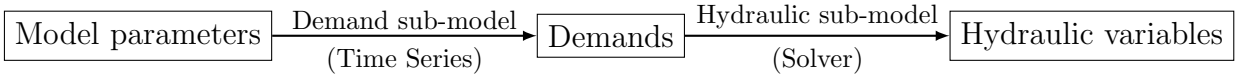


Figure 1: Variables (boxed) and sub-models of the proposed demand-hydraulic model.

Demand sub-model. The demand sub-model is proposed as a vectorized Seasonal Auto-Regressive Integrated Moving-Average (ARIMA) time series model (Box and Jenkins, 1976; Wei, 2006), which is capable of quantifying individual, or aggregated, user demands that are both temporally and spatially correlated. Ideally, the number of AR and MA parameters would be determined via standard model identification procedures (Box and Jenkins, 1976). However, in the case of the proposed model, the demand sub-model cannot be identified with regular methods because direct observations of individual, or aggregated, user demands are generally not available. Therefore, we will initially assume that the ARIMA model structure of the demand sub-model will have the same form as the total system demand.

As an example, Chen and Boccelli (2013) showed that a double seasonal autoregressive (AR(2)) model was sufficient to forecast the total system demand from a partner utility. Thus, assuming there are N individual, or aggregated, water users, the demands at time t can be denoted as a N-dimensional vector $\boldsymbol{q}_t = \left[q_{t(1)}, q_{t(2)}, ..., q_{t(N)}\right]^T$, and the resulting double seasonal AR(2) demand sub-model formally expressed as:

$$\nabla_{s_1}\nabla_{s_2}\left(\boldsymbol{q}_t - \boldsymbol{\phi}_1 \cdot \boldsymbol{q}_{t-1} - \boldsymbol{\phi}_2 \cdot \boldsymbol{q}_{t-2}\right) = \boldsymbol{a}_t \tag{1}$$

where ∇_s is the differencing operator defined as $\nabla_s \boldsymbol{q}_t = \boldsymbol{q}_t - \boldsymbol{q}_{t-s}$; s_1 and s_2 represent the lengths of weekly and daily demand periods ($s_1 = 168$ and $s_2 = 24$ for an hourly demand model); $\boldsymbol{\phi_1}$ and $\boldsymbol{\phi_2}$ are the vectors of the autoregressive parameters (e.g., $\boldsymbol{\phi}_1 = \left[\phi_{1(1)}, \phi_{1(2)}, ..., \phi_{1(N)}\right]$); and $\boldsymbol{a}_t$ is a multi-dimensional white noise process with mean 0 and covariance matrix $\boldsymbol{\Sigma}$ that accommodates the spatial correlation. While the time series model may vary for different systems, the proposed approach is generalizable for different vectorized seasonal ARIMA models.

Hydraulic sub-model. The hydraulic sub-model will be based on EPANET (Rossman, 2000) – a common distribution system network hydraulic and water quality solver. Assuming there are K online monitors in the distribution system (e.g., flow rate, pressure, etc), the observed SCADA data can be denoted by $\boldsymbol{Y}_t = \left[Y_{t(1)}, Y_{t(2)}, ..., Y_{t(K)}\right]^T$, and described as

$$\boldsymbol{Y}_t = \boldsymbol{h}(\boldsymbol{q}_t) + \boldsymbol{e}_t \tag{2}$$

where $\boldsymbol{h}(\cdot)$ represents the hydraulic solver that translates the demands into estimates of the observable hydraulic variables, and $\boldsymbol{e}_t$ is a vector of random measurement errors assuming independent Gaussian distributions. The inclusion of the sensor measurement error facilitates the calculation of conditional likelihoods of having observed the hydraulic measurements, which facilitate the estimation of the demands and demand model parameters.

2.2 *Demand and Parameter Estimation.* In the composite demand-hydraulic model, the intermediate variables of water demands are latent (i.e., no direct observations are available). Therefore, the parameters of the demand sub-model can not be estimated via regular maximum likelihood methods for time series analysis. However, the Expectation-Maximization (EM) algorithm (Dempster et al., 1977) was designed to compute maximum likelihood estimates (MLE) for models with incomplete observations. Specifically, an EM-Markov chain Monte Carlo (EM-MCMC) algorithm (Pasula et al., 1999) is proposed for the parameter estimation. Figure 2 illustrates the EM algorithm, which includes two steps – an E-step and an M-step – executed iteratively to determine the MLE of the demand model parameters.

The Expectation (E-) step generates the distribution of possible demands conditioned upon: 1) the observed hydraulic data, and 2) the current parameter estimates associated with the vectorized seasonal ARIMA model. Unfortunately, the resulting conditional demand distribution is not analytically tractable. Thus, the population of demand vectors will generated by an MCMC algorithm (Brooks et al., 2011; Gilks et al., 1995; Metropolis et al., 1953) – a sampling method for calculating statistics of complex, multi-dimensional probability distributions suitable for generating the conditional probability distributions associated with BNs (Koller and Friedman, 2009; Robert, 2007). We will implement the MCMC algorithm using the Differential Evolution Markov Chain (DE-MC) (Laloy and Vrugt, 2012) approach, which has been developed to improve the performance and convergence speed of the MCMC algorithm. The Maximization (M-) step then utilizes the distribution of demand vectors generated in the E-step to update the parameters of the demand sub-model. To update the parameters of the demand sub-model, a standard minimum sum of square (MSE) algorithm for ARIMA models (Box and Jenkins, 1976) will be used to generate the MLEs. The new MLEs replace the current demand model parameters to start a new EM cycle until the difference between MLEs generated by two consecutive cycles is less than some threshold. The resulting estimates of the EM algorithm have been shown to converge to the parameter estimates of a MLE algorithm (Dempster et al., 1977; Koller and Friedman, 2009).

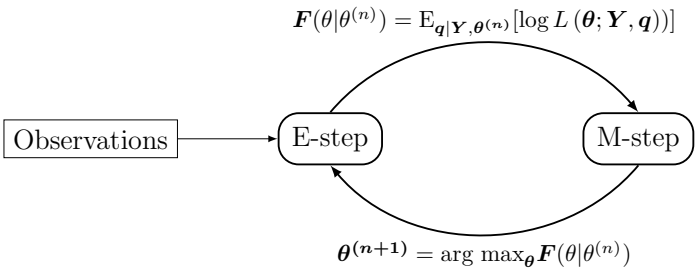


Figure 2: The EM algorithm. $\boldsymbol{\theta}$, $\boldsymbol{q}$, and $\boldsymbol{Y}$ denote the time series parameters, latent variables, and observable variables, respectively. $\boldsymbol{F}(\theta|\theta^{(n)})$ denotes the expectation of the log-likelihood function conditioned on best estimates in the nth iteration.

2.3 *Spatial Aggregation.* While the proposed composite demand-hydraulic model is capable of generating temporally and spatially correlated demands, estimating the model parameters for a realistic network model, in which the number of consumer nodes (N) can range

from 10^4 to 10^5, is not realistic. First, the computational burden would be too significant to perform the estimation in real-time. Second, the typical amount of observational data, K, is likely insufficient to accurately estimate the parameters at such fine spatial scales. Thus, an approach to group, or cluster, nodes that are assumed to behave similarly is required. To develop the clusters, we will utilize the approach of (Qin and Boccelli, 2015) that utilizes a backtracking algorithm (Shang et al., 2002) to determine the average hydraulic travel paths of every node within the network from the last 24-hours of a sufficiently long simulation. A correlation matrix generated based on the path information from each node, and a k-nearest-neighbor (knn) clustering algorithm (Larose, 2005) is used to identify nodes with similar path histories. The advantages of this algorithm are two-fold as the clustering approach will identify locations that are: 1) common to the larger, upstream flows, and 2) have similar residence time and hydraulic paths that should have longer-term benefits associated with water quality modeling. Figure 3 shows a portion of ten clusters from a test network using the hydraulic path correlation matrix and knn algorithm, as well as simulated chlorine signals from three randomly selected locations within each cluster. These results demonstrate the similarity in water quality signals resulting from similar hydraulic paths. While this approach must be performed on the network model prior to demand estimation, recent studies (van Thienen and Vries, 2013; Yang and Boccelli, 2013) have shown that the random nature of demands can impact the underlying residence times, but do no not significantly alter the hydraulic paths. Thus, determining the grouping of nodes using hydraulic path information should provide an adequate trade-off between spatial aggregation and model performance.

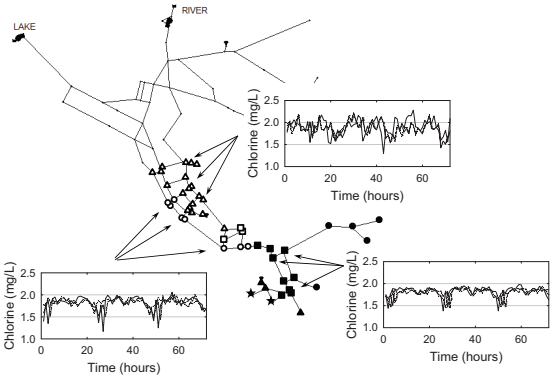


Figure 3: Water quality signals from locations within clusters identified using hydraulic path similarity [symbols] that demonstrate similar dynamics due to similar hydraulic paths.

3 Principal Findings and Significance

The following two sub-sections present the findings associated with the development and implementation of the composite demand-hydraulic model, and the analysis of the clustering algorithm for identifying nodes with similar water quality information.

3.1 Composite Demand-Hydraulic Model

3.1.1 *Case Study.* The "Net1" network included in the EPANET software package, shown in Figure 4(a), was used to demonstrate the capabilities of the composite demand-hydraulic parameter estimation algorithm. The network includes nine junctions (of which eight represent consumer nodes), twelve pipes, one reservoir, one pump, and one storage tank. In the original model, the eight customers belong to a single demand group with temporal changes in demands represented by a demand multiplier pattern that repeats every 24 hours. The

Table 1: Variables recorded within the virtual SCADA database

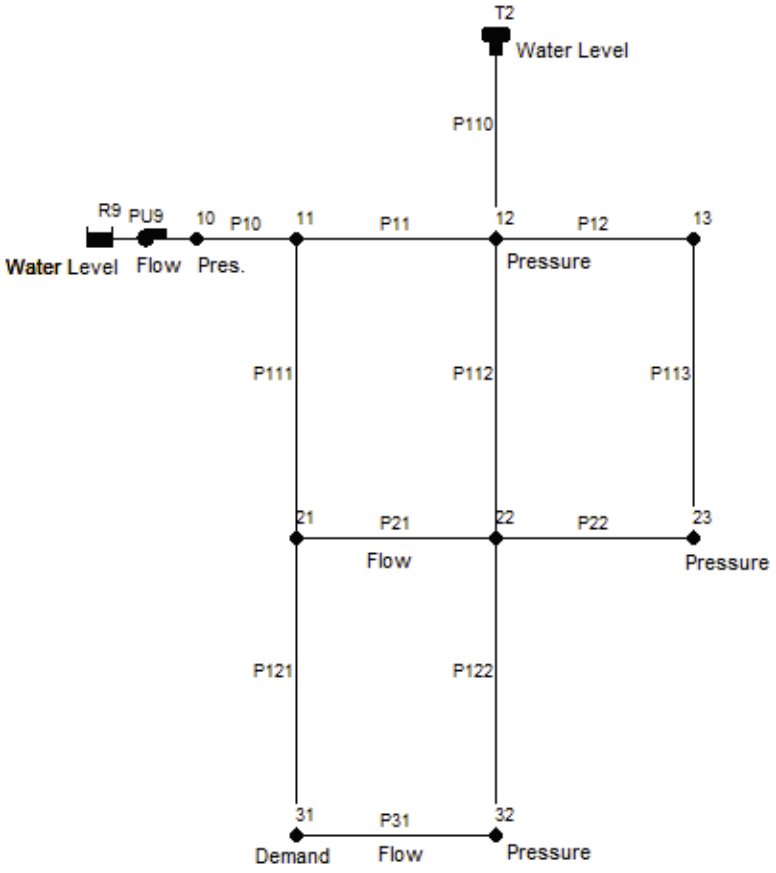
ID	Location	Variable	Unit
1	PU9	On/Off	
2	R9	Water level	Feet
3	T2	Water level	Feet
4	PU9	Flow rate	GPM*
5	P21	Flow rate	GPM
6	P31	Flow rate	GPM
7	10	Pressure	PSI†
8	12	Pressure	PSI
9	23	Pressure	PSI
10	32	Pressure	PSI
11	31	Demand	GPM

*GPM = Gallons Per Minute; †PSI = Pounds Per squared Inch

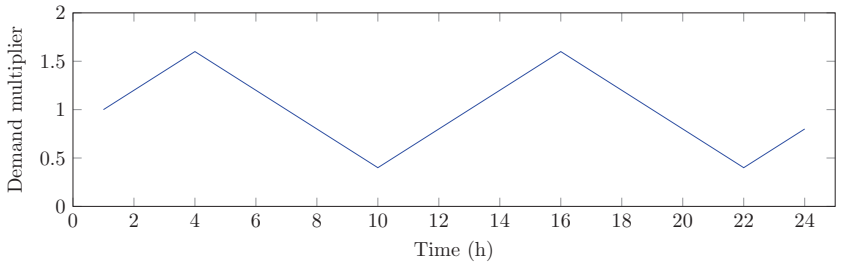
average total demand for the network is 1100 GPM and the max total demand is 1760 GPM. The selection of a small network was to facilitate a detailed analysis of the performance of the proposed estimation algorithm, which includes a correlation analysis of the water demands that would prove more difficult with a larger, realistic sized network model. For large-scale networks the methodology is still applicable, but the preliminary clustering and/or grouping of the nodes may be required to improve the computational performance.

To create a virtual SCADA database for the network, the observations of hydraulic states for a total duration of 504 hours (three weeks) were simulated in this study. During the generation process, random variations were added to the water demands that drive an extended period simulation (EPS) routine. Once the hydraulic states were computed, further random errors were added to the results to generate the actual measurements. To mimic the limited online data coverage of the real systems, only an incomplete set of hydraulic variables were written to the SCADA database. Table 1 lists the hydraulic variables monitored by the virtual SCADA system. Similar to a real SCADA system, the water levels of Reservoir R9 and Tank T2 were monitored in real-time. Three variables related to Pump station PU9 were also recorded: the on/off status, the flow rate, and the pressure on the discharge side. In addition, there were assumed to be three pressure transducers and three flow meters in the network producing hourly readings. Among the online flow meters, two were positioned along the Pipes 12 and 23, one was positioned to monitor the real-time water demand at Junction 31. The real-time water demands for the other seven consumers are unknown and will have to be estimated during the parameter estimation process. The 3-week simulated SCADA data were pre-computed and stored in a PostgreSQL 9.4 database.

A parameter estimation program was developed to implement the EM-MCMC algorithm described previously. The program was written in C/C++ and includes an EM core function along with miscellaneous routines for input/output processing. In the E-step, the demands are estimated based on the log-likelihoods of observing the measured flow rates, pressure and tank levels, and observing the estimated demands given the current parameter estimates for



(a)



(b)

Figure 4: (a) Map of the network "Net1"; the names of the nodes and links are shown close to their respective points and lines, and the text of "water level", "flow", "pressure", and "demand" represent the hydraulic variables recorded by the virtual SCADA system; and (b) the original demand pattern of Net1 with 24 hourly demand multipliers shown in the chart.

Table 2: Algorithmic parameters for the EM-MCMC algorithm

Name	Notation	Value
Start time of the estimation window	t_0	337
Estimation window size	W	168
MCMC chain size	M	10,000
Burn-in size	M_0	2,000
Standard deviation of the proposal density	σ_1	1
EM convergence threshold	ϵ_0	10

the time series demand model. In the M-step, the parameters of the time series demand model are estimated using the estimated demand information from the E-step. The algorithmic parameters used in this study are listed in Table 2. The estimation window was selected as the third week from the three-week simulated SCADA data. The "burn-in" size and total chain size of the MCMC were empirically selected to ensure the removal of initial boundary effects and proper chain mixing. The standard deviation of the proposal density is chosen to maintain an acceptance rate of 20% to 50% in the MCMC chains. The EM convergence threshold is set as 10, or an average shift of 0.05 for every parameter.

The parameter estimation program ran on a PC with a 2.5GHz Intel Sandy-Bridge CPU and 6 GB of memory. The program retrieved and processed the hydraulic measurements for the last week of the virtual SCADA database. In the test, 15 E-M cycles were required for the parameters to converge under the preset threshold, resulting in a total running time of around 80 minutes to estimate the complete set of demands during the third week.

3.1.2 *Results and Discussion*. The effectiveness of the proposed EM-MCMC algorithm relies on two factors. First, during the E-step, the MCMC sampler must produce a well-mixed chain of samples that represent the overall distribution of the water demands. Second, the EM iterations must converge to a point estimate of the parameters. If both criteria are met, the algorithm will produce the final time series model parameter estimates and demand estimates. Using the parameter and demand estimates, spatial and temporal correlations of the multivariate water demands can be analyzed for the network being studied.

3.1.2.1 MCMC Sampling. To illustrate the performance of the MCMC sampler, Figure 5 shows the Box-Whisker plots of individual water demands versus the size of the MCMC chain (including the burn-in samples) for the first time step in the first E-M iteration. The data are shown in seven subplots, each of which represents one dimension, or a single consumer node. For a subplot, the X-axis represents the number of sample points generated, and the Y-axis represents the value of water demands. For example, in the first subplot the fifth box from the left represents the empirical distribution of the first 5,000 samples comprising the MCMC chain. The upper edge, middle line, and lower edge in a box represent 25%, 50% (i.e., median), and 75% percentiles. The upper and lower whiskers represent the approximated extents for the population. Values are drawn as outliers if they are larger than $p_1 + 1.5 \times (p_3 - p_1)$ or smaller than $p_1 - 1.5 \times (p_3 - p_1)$, in which p_1 and p_3 are the 25th and 75th percentiles, respectively. The outliers are marked individually in the plot as "jittered"

dots.

All of the subplots in Figure 5 demonstrate similarities in fluctuations of the demand percentiles for the first two to three thousand samples. The visuals do not change as significantly after these initial periods, indicating that the statistics of the sampled water demands are stabilized. The analysis on the other time steps and in subsequent E-M iterations results in similar conclusions. Based upon the convergence information from these plots, the length of the MCMC chain used in this study was set at 10,000 with the first 2,000 samples discarded as the "burn-in" period and the last 8,000 samples utilized for computing the demand estimates.

3.1.2.2 EM Convergence. To evaluate the performance of the E-M algorithm, the parameters produced by successive E-M iterations are investigated. The changes to parameters are quantified using $\epsilon^{(r)} = \left|\Theta^{(r+1)} - \Theta^{(r)}\right|$, or the Euclidean distance between the parameters. In each E-M iteration, the hydraulic and demand likelihoods are calculated for evaluating the performance of the parameters. Figure 6 shows the changes in parameters and likelihoods during the test run of the parameter estimation algorithm. The X-axis is the number of E-M iterations, the left Y-axis is $\epsilon^{(r)}$ in base-10 logarithmic scale, and the right Y-axis is log-likelihood for the parameter estimate. The crisscross symbol denotes the parameter changes $\epsilon^{(r)}$. The circles, diamonds, and squares denote hydraulic likelihoods, demand likelihoods, and total likelihoods, respectively. Based on the information in Figure 6, the parameter estimates changed significantly during the first four iterations. Thereafter, the iterations yielded gradually smaller changes to the parameters. ϵ dropped below the pre-set threshold of 10 after the 15th iteration when the algorithm stopped. The refinement of the parameters is accompanied by the increasing likelihoods. The trend lines of the likelihoods show that major improvements happen in the first three iterations, which is mainly driven by the improvements associated with the hydraulic likelihoods. The demand likelihood showed a temporary decrease in the second iteration before increasing again. All three measures of likelihood continued to increase after the fourth iteration, but the gains in likelihoods became less significant as the parameter estimates converged to the final values. The overall analysis showed that the EM algorithm was effective in converging to a set of the parameters estimates for the composite model.

3.1.2.3 Demand Estimates. The EM iterations not only produced the parameter estimates but also estimates of the consumer demand and the hydraulic states of the system. In this study, the estimated demands and the "real" demands can be compared, because the simulated "real" water demands, though not exposed to the EM algorithm, have been recorded separately during SCADA data generation. Table 3 shows the estimation errors for the seven customers using measures of R^2 and AARE. Overall the estimated demand show reasonable match to the "real" demands, with AARE values ranging from 7.1% to 10.4%. Figure 7 shows the scatter plot of the estimated versus the "real" demands for best and worst performing consumer nodes for the one-week time span. The demands for Junction 11 and 32 are marked as plus signs and crosses, respectively. From the figure, the estimated demands generally matched the real demands, but the estimates for the high-demand hours are not as accurate as those for low-demand hours. Noticeably, the high demands at Junction 11 (greater than 200 GPM) are mostly underestimated. A possible explanation is that Junction

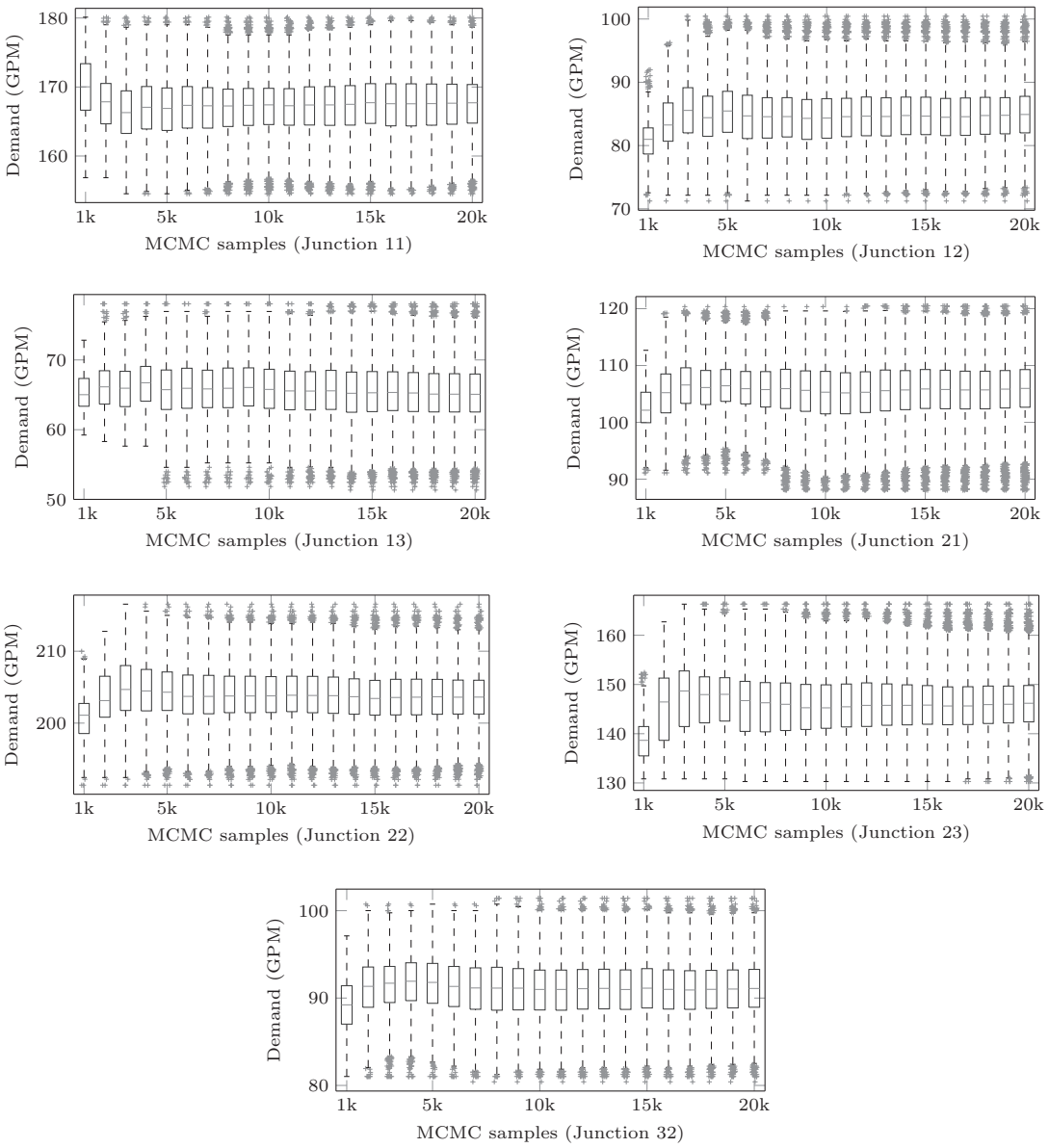


Figure 5: Distribution of the samples generated by the MCMC algorithm versus the chain size for all seven of the consumer nodes; the MCMC samples include those in the burn-in period.

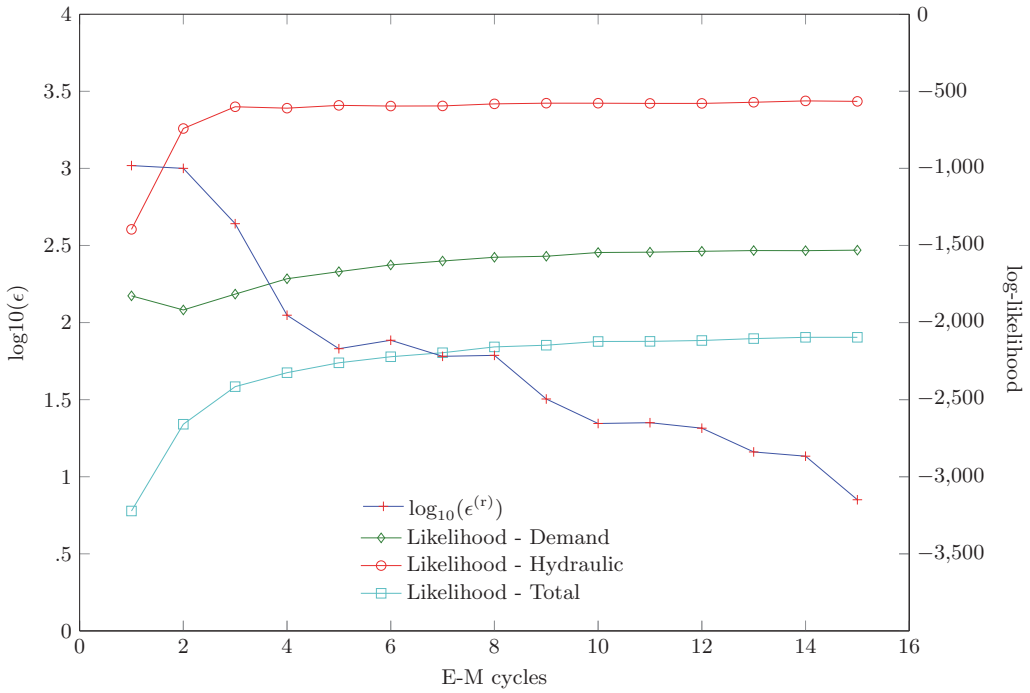


Figure 6: Convergence of the parameters during EM cycles

11 does not have pressure nor flow rate sensors, and the changes in high demands at Junction 11 will only yields limited impacts to the sensor readings located elsewhere. Therefore, using measurements from these sensors, the E-M algorithm could not provide accurate demand estimates in some hours. Installing more sensors around the area is expected to increase the accuracy of the demand estimates.

3.1.2.4 Temporal Correlations of Demand Estimates. The EM algorithm produced a time series of demands at each consumer node that can be analyzed with respect to the temporal correlation expected to be included in the data. Considering a single customer, the temporal correlations of the time series are studied by plotting the autocorrelation (AC) function. A

Table 3: Errors associated with the estimated water demands.

Customer	R^2	AARE*
Junc. 11	0.88	10.4%
Junc. 12	0.92	8.4%
Junc. 13	0.93	7.1%
Junc. 21	0.91	7.3%
Junc. 22	0.91	8.0%
Junc. 23	0.93	8.1%
Junc. 32	0.94	7.1%

*AARE = Average Absolute Relative Error

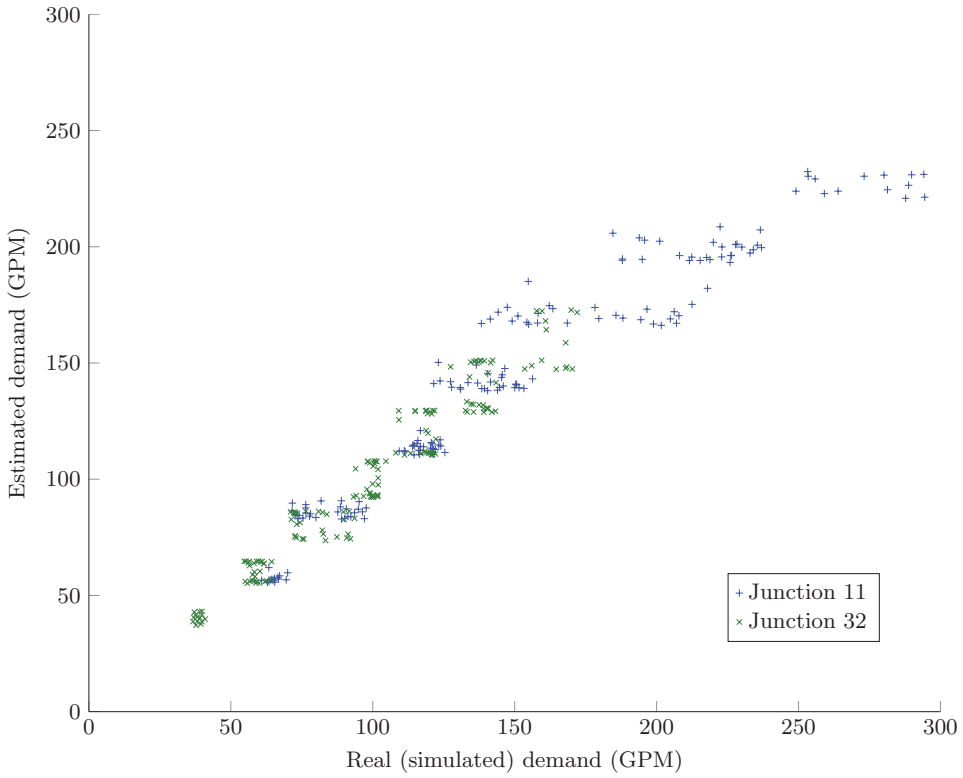


Figure 7: Comparison of estimated demands and "real" demands

lag-L autocorrelation coefficient is defined as

$$AC_L = \frac{\sum_j \left(x_j^{(i)} - \bar{x}^{(i)}\right)\left(x_{j+L}^{(i)} - \bar{x}^{(i)}\right)}{\sum_j \left(x_j^{(i)} - \bar{x}^{(i)}\right)^2} \tag{3}$$

in which $x_j^{(i)}$ is the demand for the i-th customers in time step j, $\bar{x}^{(i)}$ is the mean demand for the i-th customer. Figure 8 are the time series and autocorrelation plots of the demand estimates at Junction 11. The time series plot on the top shows that the estimates follow a general diurnal pattern with random deviations. The autocorrelation (AC) plot at the bottom shows that the series has both strong short-term (1- to 4-hour) correlations and strong periodic (24-hour, 48-hour, etc.) correlations. The auto-correlation gradually decays with longer lag-times. The analysis on the other customers show similar results. The results are also consistent with our previous study on uni-variate aggregated water demands (Chen and Boccelli, 2016) in which the same two types of correlations are identified. For the multi-variate time series introduced in this study, the temporal correlations presents important information that can be exploited for predicting water demands in future studies.

3.1.2.5 Spatial Correlations of Demand Estimates. The demand estimates generated by the EM algorithm were also employed to study the spatial correlations. The z-transformation is performed on the vectors of demand estimates to normalize the data, and the results are illustrated in Figure 9 using 1-d and 2-d histograms.

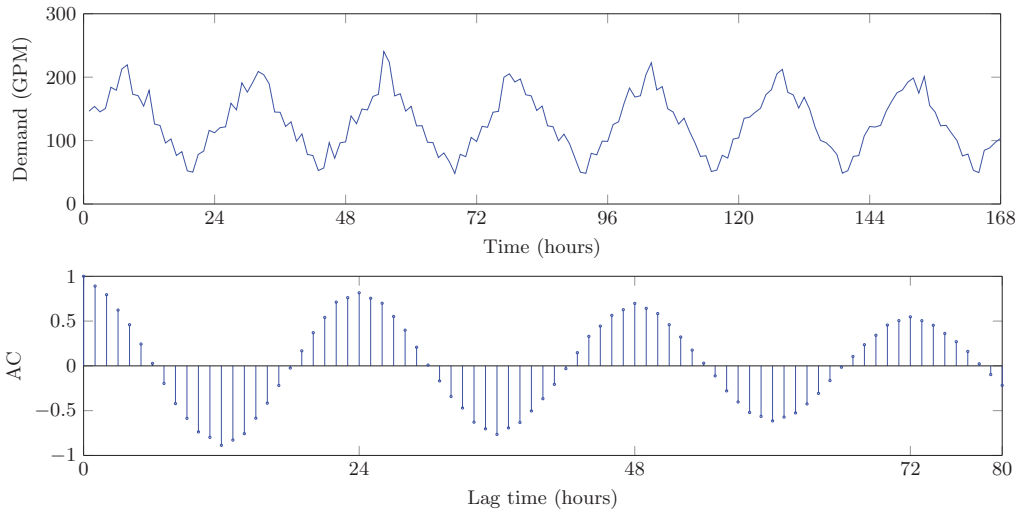


Figure 8: Time series plot and autocorrelation plot of demand estimates at Junction 32

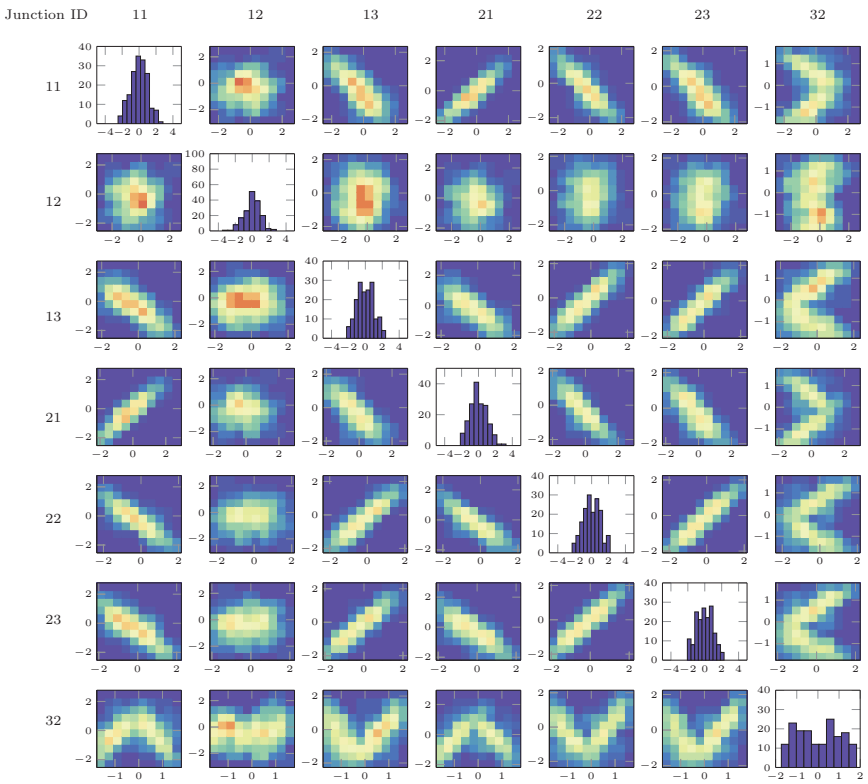


Figure 9: Spatial correlations of demand estimates

In Figure 9, the seven diagonal plots are (1-D) histograms of water demands for the seven customers. The X-axes are hour-in-a-day mean-adjusted and normalized water demands and the Y-axes are frequencies. From the plots, most demand estimates show centered distribution except for Junction 32. The off-diagonal subplots in Figure 9 present the correlations between different customers. Each plot is a 2-D histogram in which X- and Y-axes are demands for the pair of customers, and the colors represent frequencies of data points falling in the 2-D bins. The plots show that the 1st and 4th customers are positively correlated; the 3rd, 5th, and 6th customers are also mutually positively correlated. However, the customers across the two groups are negatively correlated with each other. From the network map, the 1-4/3-5-6 grouping is consistent with the network adjacency, which seems to suggest that customers can be roughly grouped in terms of their topological proximity. Moreover, the 2nd customer is relatively less correlated with any other customers. For the 7th customer, the estimates around the two peaks in the distribution show opposite types of correlations. The estimates in the lower part are positively correlated with the 1-4 group while the estimates in the higher part positively correlated to the 3-5-6 group. This result suggests that the characteristics of spatial correlations for Junction 32 vary with different hours in a week. The structure shown in Figure 9 reveals how the estimates of water demands are correlated under the given layout of customers and SCADA sensors.

3.2 Spatial Aggregation Based on Water Quality Characteristics

3.2.1 *Case Studies.* The proposed clustering algorithm was applied to two network examples, one small example network and one large real-world network.

Figure 10 shows the small network for Case Study 1, which is EPANET Example Network 3, that includes two sources (Lake and River). This network consists of 92 nodes, 3 tanks, 2 reservoirs, and 2 pumps that operate periodically. The three tanks float on the system, meaning that the flows into and out of that tanks are dependent on the source flow rates (inflows) and total system demand (outflows). For Case Study 1, a simulation lasting for 72 hours was applied to the system, and the last 24 hours concentration output collected to calculate the impact coefficients. The source species selected for each node is a conservative input, of which the concentration was set as 100 mg/L.

Figure 11 shows the real-world network model to be used as Case Study 2 that includes 12000 individual nodes (each represents, on average, eight service connections), three sources and two tanks. Two of the sources are the main feeds to the northern and southern portions of the system; the third source feeds a small portion of the network located in the east-northeast section of the network. For Case Study 2, a simulation lasting for 360 hours was applied to the system with the last 24 hours of output concentrations collected to calculate the impact coefficients. The source species selected for each node was assumed to be a conservative input with a concentration of 100 mg/L.

3.2.2 *Results and Discussion.* In this section, the results associated with the proposed clustering algorithm will be presented with an emphasis on the ability of the algorithm to cluster similar nodes and generate distinctly different clusters.

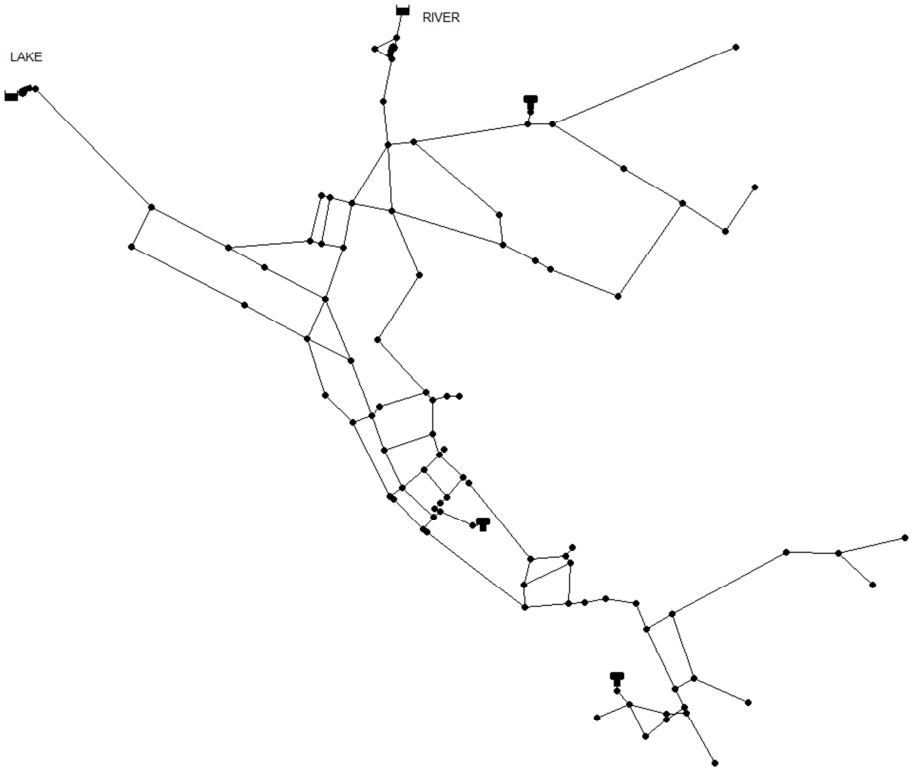


Figure 10: Small test network for the evaluation of the proposed clustering algorithm

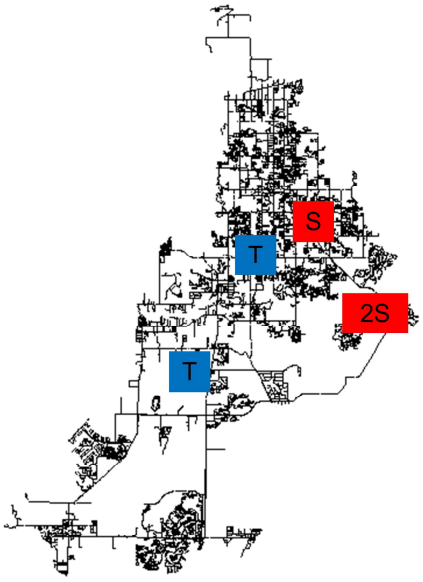


Figure 11: Real-world network model include the sources (S) and tanks (T)

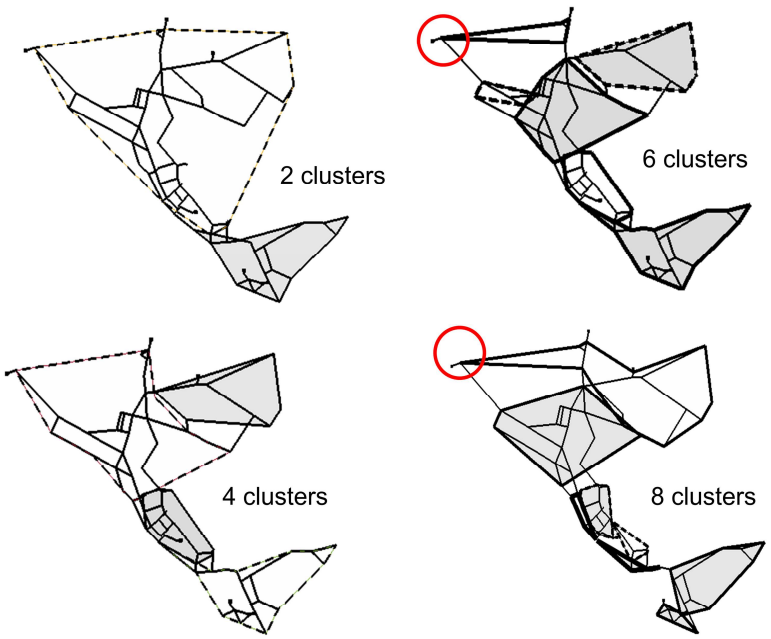


Figure 12: Results associated with clustering the small network; the node in the circle represents a zero-demand node that resulted in an outlier from the other clusters.

3.2.2.1 Case Study 1. Figure 12 illustrates the results when the network was divided into 2, 4, 6 and 8 clusters. By simple observation, the resulting clusters were spatially grouped into non-overlapping clusters by the similarity in the hydraulic flow paths. One outlier occurred (within the clusters of 6 and 8) in the northwest portion of the network (Node 10), which was a zero demand node connected to a pump that only operated intermittently resulting in periods of no flow passing that location. As a result, the resulting flow path for this node was significantly different than the nodes immediately downstream of that location and could be omitted as not meaningful.

To assess the similarity among nodes within a given cluster, water quality simulations were performed to allow the signals at the individual nodes to be evaluated. The hydraulic and water quality simulations were performed with a 504-hour duration and a water quality species entering the system at the two sources modeled as a first-order decay process (decay rate of -0.2s^{-1}). Figure 13 shows a plot of the source concentrations, which were applied to the sources. The reason for injecting the source water quality species was to assess the clustering algorithm as nodes within clusters should have similar concentration patterns. To evaluate the clustering algorithm, the water quality concentrations from each node in the network were collected at hourly intervals from hours 408 to 504. If the hydraulic paths among nodes within each cluster were truly similar, the expectation was that the concentration time series among the nodes within in each cluster would also be similar.

Using the results associated with separating the network into four clusters as an example, Figure 14 shows the concentration time series for each node within each cluster, the number of nodes within each cluster, and the variance of the means and standard deviations from each node in the cluster. By visual inspection, the nodes within each cluster are generally

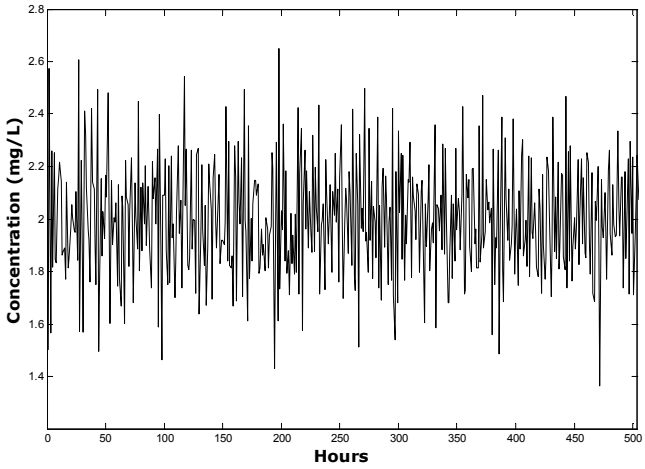


Figure 13: Influent concentrations applied to the two sources of the small test network.

observed to be similar to each other with common patterns in the time series dynamics. To better assess the similarities in water quality signals within each cluster, statistical measures, such as the mean and standard deviation, of the concentration signals were calculated for each node and used to compare similarities within and across clusters.

For assessing the similarities within a cluster, the expectation was that as the number of clusters increased the similarity of the water quality signals within each cluster would become more similar. Thus, increasing the number of clusters should also reduce the variability in the intra-cluster means and standard deviations. To test this hypothesis, the mean and standard deviation of the water quality signals from each node with every cluster was calculated; these data were used to estimate the variability of the means and standard deviations from each node within every cluster. Figure 15 summarizes these results illustrating the variance in the intra-cluster means and standard deviations, respectively, as box-and-whisker plots for clusters ranging from 2 to 15. The number of samples associated with each box-and-whisker plot is simply the number of clusters. The results from the intra-cluster analysis show that both intra-cluster variability of the means and standard deviations (Figure 15) of the water quality signals generally decreased as the number of clusters increased. However, an outlier appeared in both the intra-cluster variability of the means and standard deviations as the number of clusters were greater than 10. This outlier was related to one cluster, and the nodes in this cluster are located in the southern corner of network closest to the southern tank (when the number of clusters between 11 to 13 were analyzed, there were 6 nodes in this outlier cluster; when the number of clusters was greater than 13, there were 7 nodes in this outlier cluster; these nodes are circled in Figure 16). Figure 16 shows the concentration curves related to this outlier cluster when the total number of clusters to be analyzed was 12. The bold curve represents node 243, which is the source of the "outlier" in the clustering results. The reason why this cluster behaves as an outlier is that concentration of the node 243 (around 1.3 mg/L; Figure 16) is obviously lower than the concentrations of the other nodes within the cluster (around 1.8 mg/L). As this node is located at the end of the network and the demand is relatively low (around 6 GPM compared with around 20 to 50 GPM of the other dead-end nodes), the hydraulic residence time was considerably longer relative to

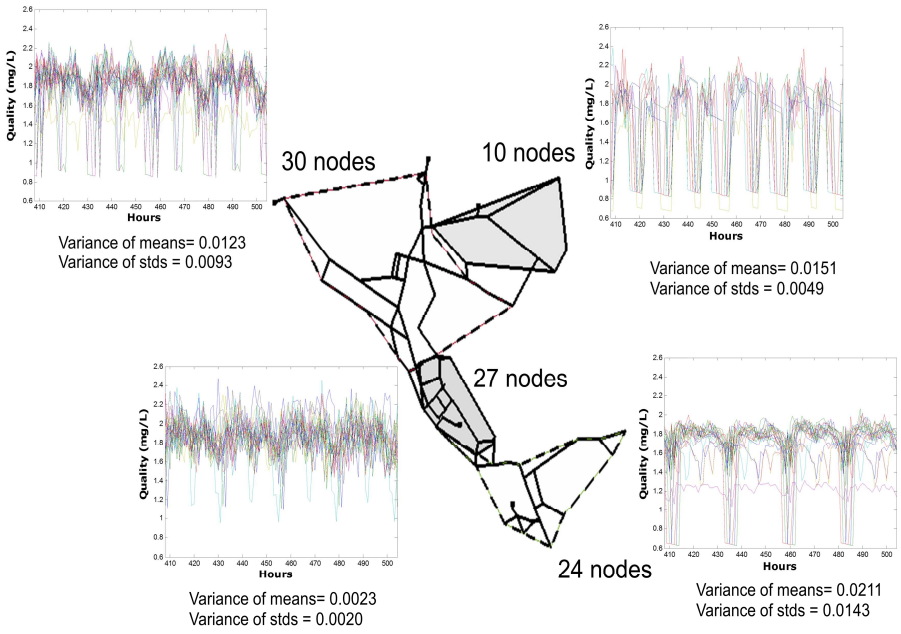


Figure 14: Plots of concentration versus time for the all nodes with the four clusters developed for the small test network.

the other nodes within the cluster. Thus, even though the path to get to all of these nodes was essentially the same, the longer residence time associated with Node 243 resulted in a significantly lower concentration leading to the larger variances in the means and standard deviations of the water quality signals within this cluster that caused the apparent outlier in Figure 15. Overall, the general trends in the intra-cluster analysis support the notion that as the number of clusters increased, the similarity between the nodes within each cluster became increasingly similar.

In addition to assessing the statistics within each cluster, the performance of the clustering algorithm can also be assessed by evaluating the statistics across the clusters. Thus, another approach for demonstrating the performance of the clustering algorithm was to evaluate the inter-cluster variability of the means of the concentration for nodes within the clusters. For the inter-cluster variability of the means, if the clustering algorithm truly separated the nodes, then the variability of the cluster means of concentrations should become greater as the number of clusters increased. Thus, for each cluster the overall mean of the concentrations from all of the nodes was calculated with the variability of those means representative of the differences between the clusters. Figure 17 presents the inter-cluster variability of the means and illustrates that as the number of clusters increased the variability in the cluster means became larger. These results suggest that the clusters were becoming more distinct with large variations in the inter-cluster variance as the number of clusters becomes very large. For the results with the large number of clusters, there are more clusters with only one node, which affects the inter-cluster variance.

3.2.2.2 Case Study 2. For the large network, the clustering results for two clusters are shown in Figure 18. While this network was successfully clustered into two groups, there appear to be overlapping area in the image. These regions are actually distinct and are an artifact of

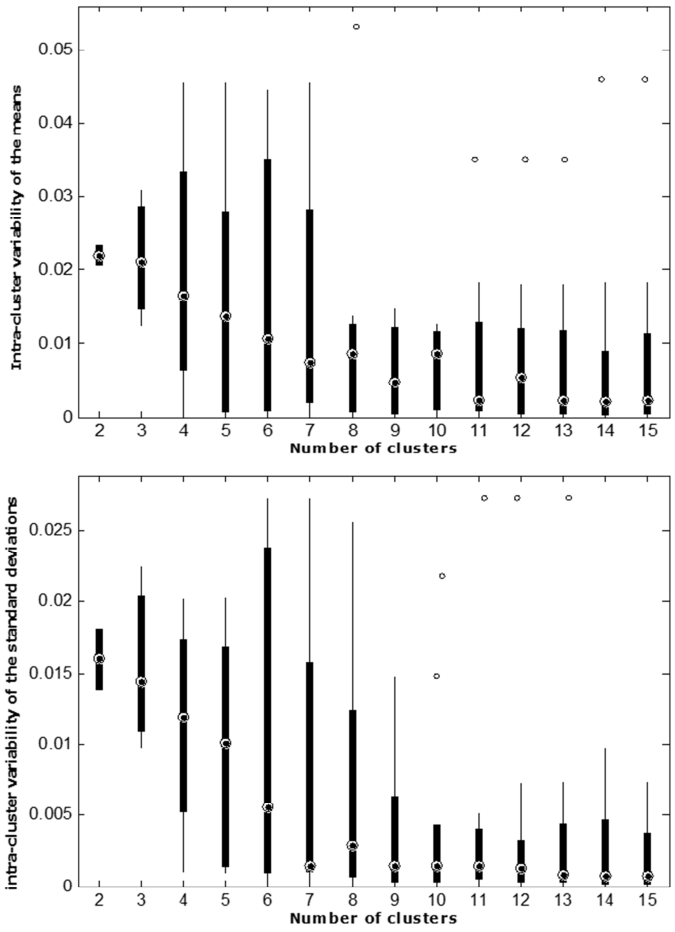


Figure 15: Box-and-whisker plots of the intra-cluster variability of the means and standard deviations for the small test network.

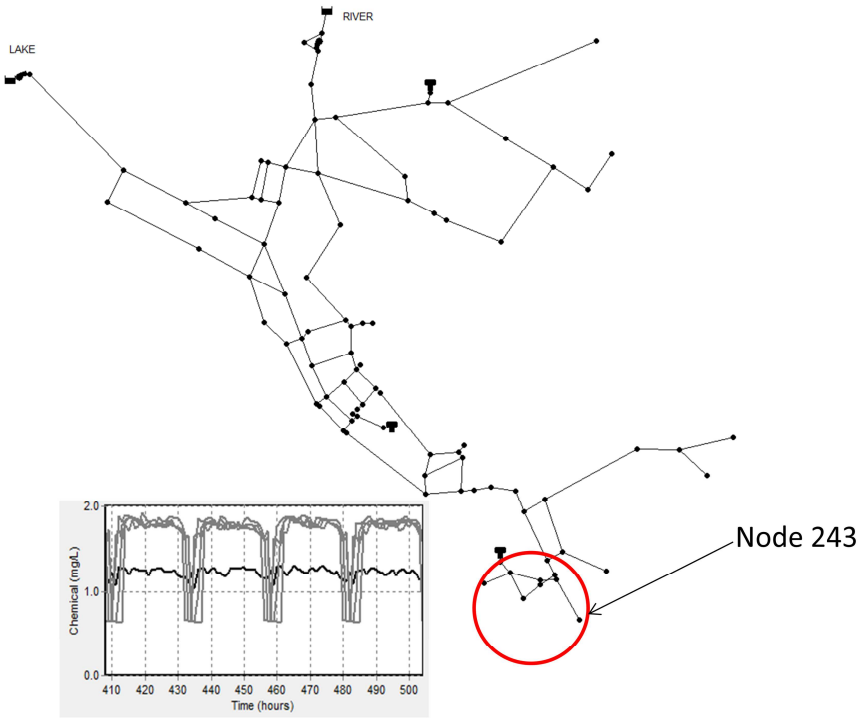


Figure 16: Location and concentration plots of the outlier cluster when the small test network is separated into 12 clusters.

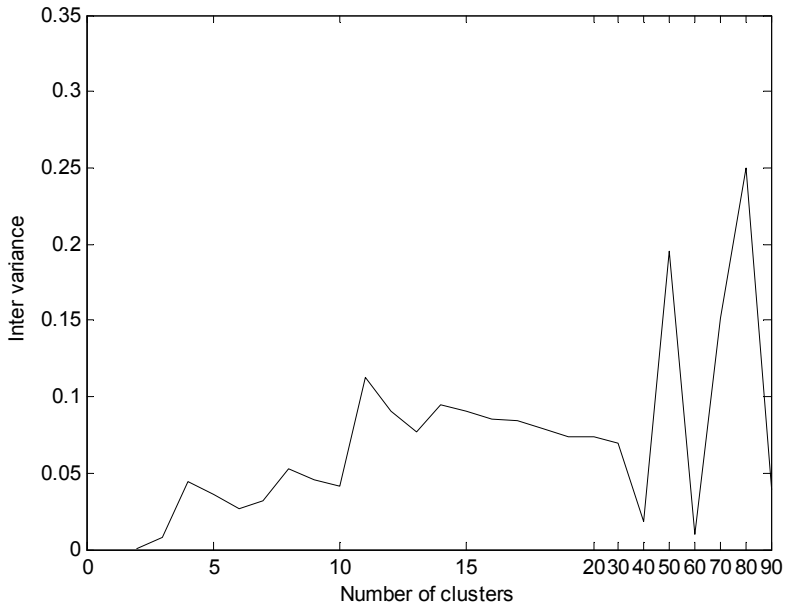


Figure 17: Inter-cluster variance of the mean concentration for an increasing number of clusters.

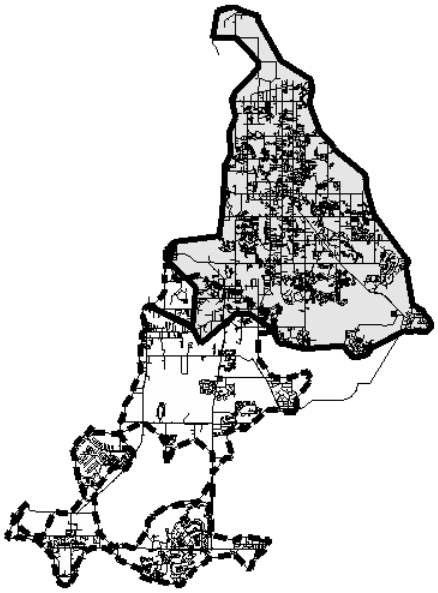


Figure 18: Separation of the realistic network into two clusters.

the current approach used to visualize the clusters.

To assess the similarity of hydraulic paths among nodes from the clustering algorithm, a water quality simulation was performed with a source concentration of 400 mg/L. The three water treatment plants were selected as sources for chemical injections. The hourly water quality concentrations were collected from hours 264 to 360. The variances in the intra-cluster means and standard deviations for clusters of 2, 4, 6, 8, 10, 20, 30, 40, and 50 are shown in Figures 19 and 20, respectively. From the results, the clusters of the large network lead to more similar clusters based on the similarity of hydraulic paths when clusters move towards finer ones, especially when number of clusters step into range exceeding 10 clusters.

Similar to inter-cluster variability analysis in Case Study 1, the same method was applied to the clustering results from this case study. Figure 21 presents the inter-cluster variability of the means, and the variability of the mean concentrations that demonstrate that both values increased as the number of clusters increased. These results suggest that the clusters continue to become more distinct as the number of clusters increased.

3.3 Significance

The results of this study are significant for two primary reasons. First, the development of the composite demand-hydraulic model was shown capable of estimating the demands and parameters of a time series model using limited hydraulic information. This result is the first approach to link an actual demand model to a network hydraulic model that will allow not only for demand estimation but demand forecasting to be performed. The latter of which will allow real-time decision making possible. Second, the proposed clustering algorithm was shown capable of grouping nodes based on similarities in water quality. This ability to group nodes will provide opportunities to reduce the scale of network demand estimation

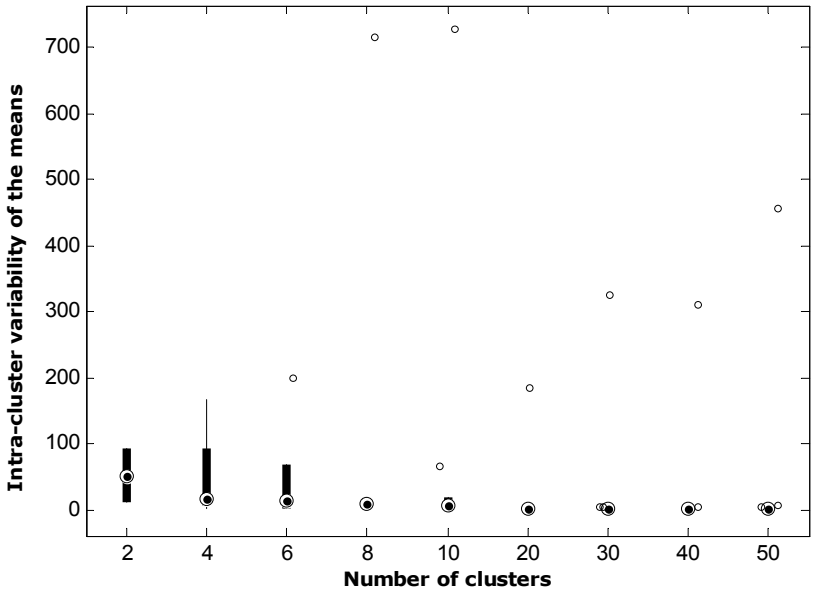


Figure 19: Box-and-whisker plots of the intra-cluster variability of the means.

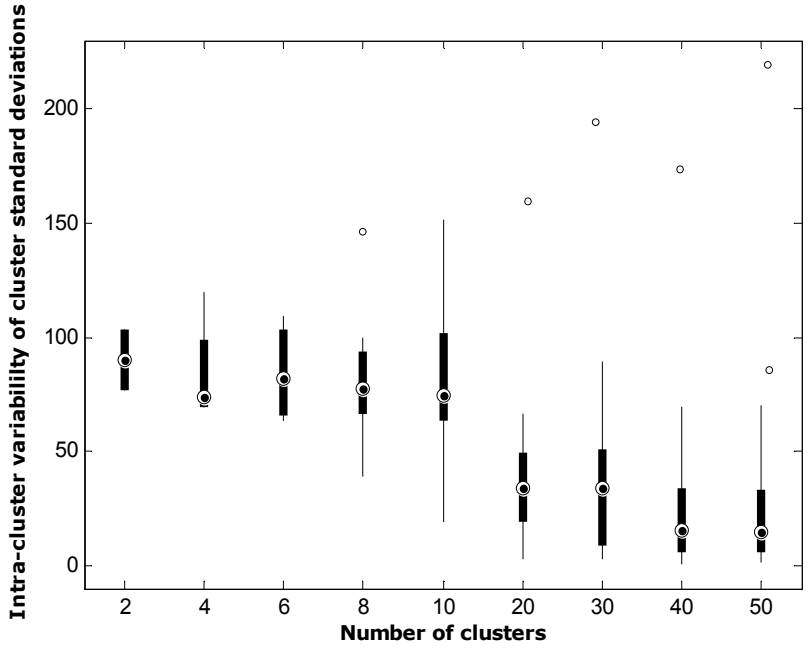


Figure 20: Box-and-whisker plots of the intra-cluster variability of the standard deviations.

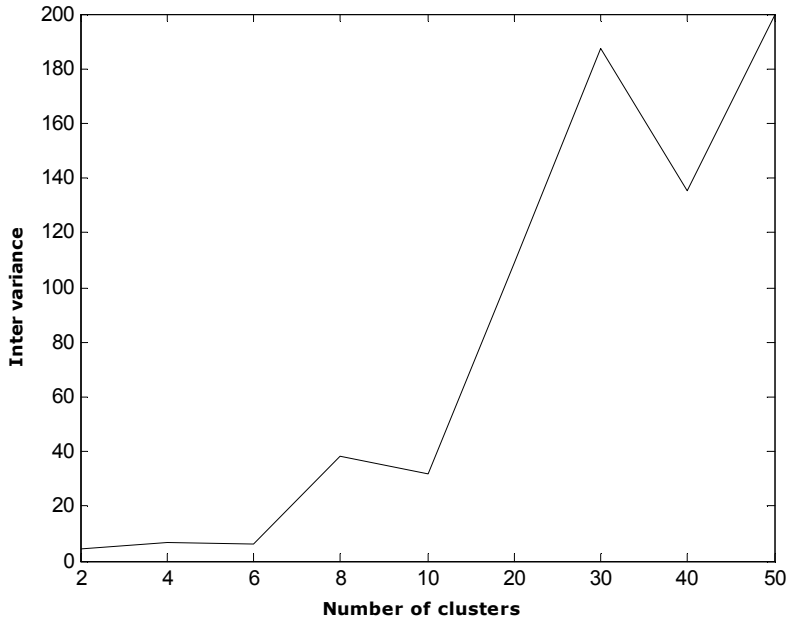


Figure 21: Inter-cluster variance of the mean concentration for an increasing numbers of clusters.

problems. That is, the clustering algorithm provides the capability to effectively reduce the scale of the demand estimation problem for realistic networks (e.g., Figure 18) to the scale of smaller network (e.g., Figure 4). Additionally, the clustering approach presented allows the grouping of nodes with similar water quality characteristics that can also help to reduce the problem scale of other applications such as locating sensors for contaminant warning systems or identifying regulatory sampling locations.

4 Publication Citations

Based upon the results presented in this report, we are expecting to publish four peer-reviewed manuscripts. The list of these potential manuscripts are included below. Publication 1 is close to being submitted; Publications 2 and 3 have been prepared as preliminary drafts; and Publication 4 requires finalization of the research results.

1. Qin, T. and Boccelli, D. L. (2016). "Network Clustering Using Hydraulic Path Similarity" *Urban Water Journal*, in preparation.

2. Chen, J. and Boccelli, D. L. (2016). "Stochastic Demand-Hydraulic Model: Methodology." *Water Resources Research*, in preparation.

3. Chen, J. and Boccelli, D. L. (2016). "Stochastic Demand-Hydraulic Model: Software Development and Application." *Environmental Modeling and Software*, in preparation.

4. Qin, T. and Boccelli, D. L. (2016). "Estimation of Water Demands Using an MCMC-MRF Algorithm." *Journal of Water Resources Planning and Management*, in preparation.

In addition to the peer-reviewer manuscripts, conference presentations that are directly based upon this research or a derivative of the research include:

1. Chen, J. and Boccelli, D. L. (2016). "A Framework for Real-Time Spatially Distributed Demand Estimation and Forecasting." *Smart Systems for Water Management: Modelling, Simulation, Analytics and ICT for Behavioural Change*, Monte Verità, Switzerland. (submitted)
2. Oliveira, P. J., Rana, S. M. M., Qin, T., Woo, H., Chen, J. and Boccelli, D. L. (2016). "Case Study: Evaluation of a Composite Demand-Hydraulic Modeling Framework." *2016 Water Distribution System Analysis Symposium*, Cartagena, Colombia. (accepted)
3. Chen, J. and Boccelli, D. L. (2015). "A Real-Time Demand-Hydraulic Model of Water Distribution Systems." *Proceedings of the World Water and Environmental Resources Congress*, ASCE, Austin, TX.
4. Qin, T. and Boccelli, D. L. (2015). "Grouping Water Demand Nodes by Similarity Among Flow Paths in Water Distribution Systems." *Proceedings of the World Water and Environmental Resources Congress*, ASCE, Austin, TX.

5 Number of Students Supported by the Project

The research has partially funded one Ph.D. student (Tian Qin) and one M.S. student (Lihe Wang) in Environmental Engineering.

6 Profession Placement of Graduates and Teaching Assistantship

Mr. Qin and Mr. Wang are both completing their graduate degrees at the University of Cincinnati. Dr. Chen graduated from the University of Cincinnati in 2015 and is currently employed by IDModeling (Dr. Chen was being funded from other sources, but was part of the study).

7 Awards or Achievements

Dr. Chen recently received an Honorable Mention (2nd place) in the 2016 University Council on Water Resources (UCOWR) Ph.D. Dissertation Award within the Natural Science and Engineering category.

8 Additional Funding for this Project

In addition to the funding provided by the Ohio Water Resources Center, additional funds from "A Comprehensive Field-Scale Distribution System Network Model Assessment and

Analysis: Hydraulics and Water Quality" (WaterRF, Sep 2010 – Aug 2012; T. Qin) and "Real-Time Distribution System Network Modeling and Fault Diagnosis" (NSF, Jul 2009 – Aug 2012; J. Chen) were utilized to fund portions of the research.

The current Ohio Water Resources Center funding also led to additional funding from the National Science Foundation for the project entitled "Data Assimilation and Forecasting for Real-Time Drinking Water Distribution System Modeling" (started Aug 2015).

9 References

Box, G. E. P. and G. M. Jenkins (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day series in time series analysis and digital processing. Oakland, Calif.: Holden-Day, Inc.

Brooks, S., A. Gelman, G. L. Jones, and X.-L. Meng (Eds.) (2011). *Handbook of Markov chain Monte Carlo*. Chapman & Hall/CRC.

Chen, J. and D. L. Boccelli (2013). Demand forecasting for water distribution systems. In *Proceedings, Computing and Control in the Water Industry*, Perugia, Italy.

Chen, J. and D. L. Boccelli (2016). Forecasting hourly water demands with seasonal autoregressive models for real-time application. *Water Resources Research*. in revision.

Dempster, A., N. Laird, and D. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1–38.

Ghahramani, Z. (1998). Learning dynamic Bayesian networks. *Adaptive Processing of Sequences and Data Structures*, 168–197.

Gilks, W., S. Richardson, and D. Spiegelhalter (1995). *Markov Chain Monte Carlo in Practice: Interdisciplinary Statistics*, Volume 2. Chapman & Hall/CRC.

Koller, D. and N. Friedman (2009). *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press.

Laloy, E. and J. Vrugt (2012). High-dimensional posterior exploration of hydrologic models using multiple-try dream (zs) and high-performance computing. *Water Resources Research 48*(1), W01526.

Larose, D. T. (2005). *Discovering Knowledge in Data: An Introduction to Data Mining*. Hoboken, NJ: Wiley-Interscience.

Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics 21*, 1087.

Pasula, H., S. Russell, M. Ostland, and Y. Ritov (1999). Tracking many objects with many sensors. In *International Joint Conference on Artificial Intelligence*, Volume 16, pp. 1160–1171.

Qin, T. and D. L. Boccelli (2015). Grouping water demand nodes by similarity among flow paths in water distribution systems. In *Proceedings of the World Water and Environmental Resources Congress*, Austin, TX. ASCE (accepted).

Robert, C. (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer Science+Business Media, LLC.

Rossman, L. A. (2000). *EPANET2 User's Manual*. Cincinnati, OH: Risk Reduction Engineering Laboratory, U. S. Environmental Protection Agency.

Shang, F., J. G. Uber, and M. M. Polycarpou (2002). Particle backtracking algorithm for water distribution system analysis. *Journal of Environmental Engineering 128*(5), 441–450.

van Thienen, P. and D. Vries (2013). Probabilistic backtracing of drinking water contamination events in a stochastic world. In *Proceedings, Computing and Control in the Water Industry*, Perugia, Italy.

Wei, W. W. S. (2006). *Time Series Analysis: Univariate and Multivariate Methods* (2^{nd} ed.). Pearson Education, Inc.

Yang, X. and D. L. Boccelli (2013). A simulation study to evaluate temporal aggregation and variability of stochastic water demands on distribution system hydraulics and transport. *Journal of Water Resources Planning and Management*. doi:10.1061/(ASCE)WR.1943-5452.0000359.