EPA Public Access

Author manuscript

J Water Resour Plan Manag. Author manuscript; available in PMC 2020 January 01.

About author manuscripts

Submit a manuscript

Published in final edited form as:

J Water Resour Plan Manag. 2019 ; 145(1): . doi:10.1061/(ASCE)WR.1943-5452.0001005.

Quantifying hydraulic and water quality uncertainty to inform sampling of drinking water distribution systems

David Hart¹, J. Santiago Rodriguez², Jonathan Burkhardt³, Brian Borchers⁴, Carl Laird^{1,2}, Regan Murray³, Katherine Klise¹, Terranna Haxton³

¹Sandia National Laboratories, Albuquerque, NM,

²Purdue University, West Lafayette, IN

³U.S. Environmental Protection Agency, Cincinnati, OH

⁴New Mexico Institute of Mining and Technology, Socorro, NM

Abstract

Sampling of drinking water distribution systems is performed to ensure good water quality and protect public health. Sampling also satisfies regulatory requirements and is done to respond to customer complaints or emergency situations. Water distribution system modeling techniques can be used to plan and inform sampling strategies. However, a high degree of accuracy and confidence in the hydraulic and water quality models is required to support real-time response. One source of error in these models is related to uncertainty in model input parameters. Effective characterization of these uncertainties and their effect on contaminant transport during a contamination incident is critical for providing confidence estimates in model-based design and evaluation of different sampling strategies. In this paper, the effects of uncertainty in customer demand, isolation valve status, bulk reaction rate coefficient, contaminant injection location, start time, duration and rate on the size and location of the contaminant plume are quantified for two example water distribution systems. Results show that the most important parameter was the injection location. The size of the plume was also affected by the reaction rate coefficient, injection rate and the injection duration, while the exact location of the plume was additionally affected by the isolation valve status. Uncertainty quantification provides a more complete picture of how contaminants move within a water distribution system and provides more information when using modeling results to select sampling locations.

Publisher's Disclaimer: DISCLAIMER

Publisher's Disclaimer: The U.S. Environmental Protection Agency (EPA) through its Office of Research and Development funded and managed the research described herein under Interagency Agreement (IA # DW8992403601) with the Department of Energy's Sandia National Laboratories. It has been reviewed by the Agency but does not necessarily reflect the Agency's views. No official endorsement should be inferred. EPA does not endorse the purchase or sale of any commercial products or services.

Publisher's Disclaimer: Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

INTRODUCTION

Drinking water utilities are responsible for the delivery of safe, potable water to consumers. As the water is transported from the treatment plant through the distribution system to the customers, the quality of the water can deteriorate due to reactions within the bulk water and on pipe walls. Water utilities are required to take samples within the distribution system to ensure public health protection and compliance with safe drinking water regulations. Taking water quality samples is a common practice for a variety of reasons. Kirmeyer et al. (2002) highlighted five general sampling categories that are used for normal system operations. These sampling categories are related to regulations, operations, maintenance, improvements and customer complaints.

Regulatory sampling aims to gather a broad understanding (or representation) of water quality within a drinking water distribution system (WDS) (Kirmeyer et al. 2002). The specific objectives, number of samples and sampling methods of regulatory sampling programs vary depending on the rule and the regulatory body for which the samples are being collected. Operational sampling is used to aid in understanding, tracking or improving water quality within the system. This type of sampling is geared towards ensuring disinfection or other treatment technologies are working as expected. Maintenance sampling helps identify the areas in the system which need to be flushed or which pipes in the distribution system need to be replaced. Improvement sampling examines the effects on water quality from changes in source water or construction. Customer complaints can trigger sampling to determine the cause of the complaint (Kirmeyer et al. 2002).

In contrast to the sampling categories identified above, security or emergency response sampling focuses on determining if a contaminant is present in a distribution system, where it might have originated and how far it has spread. These samples seek to understand the effect of an incident and help guide emergency response actions. Within this category, three main types of sampling objectives exist: confirmatory, plume extent and clearance. Confirmatory sampling can be used to confirm that a contamination incident has occurred or is occurring in the WDS. The sample would most likely be taken after an alert from a water quality surveillance and response system (e.g., water quality sensor or customer complaint). In addition, the confirmatory sample should be used to identify the composition or concentration of a contaminant. Plume extent sampling can be used to identify the injection location, the extent of the contamination plume and the required decontamination area. Clearance sampling can be used to confirm that the decontamination procedure worked and that the system is no longer contaminated. This type of sampling can be used to help inform decisions about when the system can be restored to normal service. Sampling provides data that can support public health decisions as well.

WDS modeling methodologies and tools can be used to plan and inform sampling strategies for all of the previously mentioned sampling categories. To use the WDS infrastructure models to support real-time decision making and response to contamination incidents, users must have a high degree of confidence in the results. These infrastructure models are subject to various types of uncertainty. For example, infrastructure models are often calibrated to one customer demand scenario (e.g., average or maximum day) or to ensure availability of

fire flow. This type of model can be useful for addressing some operational objectives, but would not capture realistic minute to minute variability in demands or changes in demand that might occur under a public health order (e.g., do-not-use or do-not-drink). These variations from the average/maximum demand scenarios can create uncertainty in how the model results should be used and interpreted.

Hydraulic model parameters that cause uncertainty include (1) customer demands at each node and time, (2) operational controls (e.g., valve settings, pump curves), (3) representation of infrastructure (e.g., missing pipes or junctions, effective pipe diameters) and (4) initial conditions (e.g., tank levels, pump statuses). Water quality model parameters that might cause uncertainty include (1) the type of contaminant, (2) the reaction dynamics of the contaminant, (3) the amount of contaminant injected, (4) the contaminant injection location, (5) the time of the contaminant injection and (6) the duration of the contaminant injection.

Uncertainty quantification techniques have been applied to model calibration efforts by many researchers (Lansey et al. 2001; Lansey and Basnet 1991; Bargiela and Hainsworth 1989; Sanz and Pérez 2014). Generally, these efforts have focused on calibrating specific model parameters, such as roughness coefficients or disinfectant concentrations, using controlled field experiments or other known system states. Yang and Boccelli (2013) examined the effects of hydraulic uncertainty on contaminant propagation. In their paper, the effect of contaminant mass loading and arrival times was studied. In addition, important considerations for temporal demand aggregation were provided. Uncertainty in the position (i.e., which nodes are contaminated) and extent of a contaminant plume (i.e., how many nodes are contaminated) and a measure to quantify that uncertainty on a system-level scale, have not previously been defined. Quantifying this uncertainty through simulation studies will help identify some of the parameters that should be targeted when seeking to decrease uncertainty through sampling.

This paper investigates uncertainty related to the following model parameters: (1) customer demands variation, (2) isolation valve status, (3) bulk reaction rate coefficient, (4) injection start time, (5) injection duration, (6) injection location and (7) injection rate. The objective of this study is to identify which of these parameters has the greatest influence on the contamination plume and discuss how this information can be used to select sampling locations to identify the edge of the contamination plume. To determine how each source of uncertainty affects the extent of the contamination plume, this paper uses a full factorial design of experiments (Plackett and Burman 1946) through simulation. The uncertainty quantification techniques used in this study report the effect of different parameter values on the predicted extent of contamination.

This study did not modify operating controls, tank levels, pump curves or valve settings to create different operating regimes. In addition, the distribution system models were assumed to be calibrated for an average day. The parameters that were chosen considered in this work are consistent with those commonly used by water security research (Yang and Boccelli 2013; Klise et al. 2014), and were kept to a subset of all parameters to maintain computational tractability.

METHODS

Monte Carlo simulation was used to examine the uncertainty in contamination plumes based on full factorial design of experiments using several different hydraulic and water quality parameters. The following section describes the simulation study, which includes the two example WDS, the model parameters and the metrics used in the analysis.

Simulation Study

The smaller WDS model, KL, used in this study is composed of 935 junctions, 1,274 pipes and 1 reservoir (Figure 1). Figure 1 also contains results which are discussed in the Results section. The model file was obtained from the University of Kentucky's Water Distribution System Research Database (<http://www.uky.edu/WDST/database.html>). Because the model did not have a variable demand, it was modified to include the customer usage pattern from the larger WDS model described below. Versions of the system have been used in the research studies of Kang and Lansey (2012) and Bi et al. (2015). The system represents a small-sized drinking water utility with a daily production of 11,800 m³/day (3.12 MGD) and approximately 15,600 residents assuming an average usage rate of 200 gallons per day per person. The total length of pipe in the system is 253 km. The residence time in the system is 16 hours on average.

The larger WDS model, N6, used in this study is composed of 3,322 junctions, 3,828 pipes, 2 control valves, 61 pumps, 1 reservoir and 34 tanks (Watson et al. 2009) (Figure 2). Figure 2 also contains results which are discussed in the Results section. The system represents a medium-sized drinking water utility with a daily production of 115,000 m³/day (30.4 MGD) and approximately 150,000 residents assuming an average usage rate of 200 gallons per day per person. The total length of pipe in the system is 640 km. The residence time in the system is 62 hours on average. Customer demand varies according to a single pattern which is repeated over a 24 hour cycle. Davis and Janke (2011) and U.S. EPA (2010) provided additional characteristics of the system as their Network 4.

EPANET was used for all hydraulic and water quality simulations. A Python™ package was written utilizing the Water Network Tool for Resilience (WNTR) package to modify the WDS model and run a large number of EPANET simulations efficiently (Klise et al. 2016). WNTR provided access to the compiled EPANET libraries. Additionally, SciPy (Jones et al. 2016) was used for the stochastic simulations and the h5py (Collette 2016) package was used for managing HDF5 (The HDF Group 2016) formatted file input/output. The pattern time step was set to 1 hour, the hydraulic time step was set to 15 minutes, the water quality time step was 1 minute and the reporting time step was 15 minutes. These values were selected to provide an adequate range in model results without requiring a prohibitive amount of simulation time. EPANET's "statistic maximum" option was used to keep file sizes reasonable, as the analyses used only the maximum concentration value and did not need the time at which the maximum concentration occurred. The final set of simulations were executed in parallel on a computational cluster of Linux based servers and took approximately one month to complete (33.9 CPU-years compute time). The simulations on the smaller KL model produced 4.6 GiB of compressed HDF5 files, while the execution on

the N6 model produced 30 GiB of data. These results were post-processed using Python and MATLAB® (The MathWorks, Inc. 2015).

Table 1 summarizes the parameters and their values that were used in the study. The first two parameters are stochastic parameters which were used to create the realizations of different hydraulic conditions, while the remaining are direct model parameters. All are described in more detail in the following paragraphs.

To simulate uncertainty in demands, a Gaussian distribution was used to modify the demand multiplier for each pattern step at each non-zero demand node. Demands, $\hat{d}_{tj}$, were calculated using Equation 1:

$$\hat{d}_{tj} = \max(Y_{tj}, 0.0)m_{tj}b_j \quad (1)$$

$$\forall j \in \mathcal{J}, \quad t = 0, 1, \dots, T$$

where $\mathcal{J}$ is the set of all nodes (excluding tanks or reservoirs), T is the number of pattern time steps, m_{tj} is the demand multiplier for a specific time step t and node j , b_j is the base demand for node j and Y_{tj} is a random number sampled from a normal distribution with mean 1.0 and standard deviation σ for every node j and pattern time step t . The variable σ was set to values of 0, 0.2 and 0.4 for this study (Table 1). Any demand that became negative due to this modification was set to zero.

In this study, the total system demand was assumed to be accurately represented in the WDS model at each time. The temporary demands were rescaled to match the original total system demand, as shown in Equation 2. This allowed for spatial variation in demands while keeping total usage constant. An example of two demand patterns for a node in a WDS model is shown in Figure 3.

$$d_{tj} = \hat{d}_{tj} \frac{\sum_{i \in \mathcal{J}} m_{ti} b_i}{\sum_{i \in \mathcal{J}} \hat{d}_{ti}}, \quad t = 0, 1, \dots, T, \quad \forall j \in \mathcal{J} \quad (2)$$

The final demands, d_{tj} , were used in the EPANET simulations. This was accomplished by changing the input file's base demand to 1, and the d_{tj} values were inserted as new demand multipliers for the patterns of each of the non-zero demand nodes. Thus, each node was assigned a unique pattern for the full length of the simulation.

The uncertainty in isolation valve status was simulated by closing a random selection of pipes in the system. Most WDS models typically do not contain all of the isolation valves within the system. However, they are a source of uncertainty because these valves can break or can be closed during repair and not be reopened. EPANET allows individual pipes to be fixed open or closed, and this technique was used to model valve closures during stochastic simulation.

The stochastic isolation valve statuses were created as follows: for each pipe, a random number was sampled from a uniform (0,1) distribution. If the random number was below the threshold, κ in Table 1, then the pipe was closed. Only pipes with a diameter of 12 inches or less were considered for closure, since it was assumed that utility staff would be aware of closed isolation valves on larger pipes. The κ parameter was tested at values of 0, 0.0025 and 0.0050, representing 0%, 0.25% and 0.5% of total pipes being closed. Any simulation that caused isolation of part of the distribution system was discarded as infeasible, and a new set of closed pipes was selected; this was repeated until the specified number of successful simulations were completed.

A contamination plume within a WDS is also affected by the reaction dynamics within the system. Davis et al. (2014) studied 11 possible contaminants with bulk reaction rate coefficients ranging from 0.003 to 8,000/day. Three bulk reaction rate coefficients were selected for this study, with K_b assigned as either 0 (no decay), 0.1/day or 5/day (Table 1). These values of K_b were set as the global bulk decay coefficient in the EPANET input file, assuming a first order bulk reaction. Higher reaction rate coefficients were also investigated, but since the exponential decay was so rapid, the resulting contamination plumes were small and did not persist, making it unlikely that sampling would be required.

The plume is also influenced by when the contaminant was injected, how long the contaminant was injected, where the contaminant was injected and how much of the contaminant was injected. To determine the effect of time of day, the injection start time, t_{inj}^0 in Table 1, was set to 24:00, 32:00 or 40:00. The injection duration, t_{inj} , was simulated as 1 hr, 12 hr or 24 hr. The values of t_{inj}^0 and t_{inj} were used to define the contaminant injection pattern in EPANET. The injection location, a_{inj} , was simulated at all of the non-zero demand (NZD) nodes (nodes with demands greater than zero) plus all reservoirs and tanks. Both systems only have one reservoir, and only N6 contains tanks. EPANET simulates a source located at a tank or reservoir such that the injection takes place at a point just outside the tank or reservoir, not directly into the tank or reservoir.

For a constant mass injection rate, r_{inj} , the amount of the injected contaminant is linearly related to the observed concentration; a ten-fold increase in initial mass yields a ten-fold increase in concentration. This means that impacts due to changes in injected mass can be calculated after the fact by simply scaling concentrations. In the EPANET simulations, a constant mass injection rate of 1 kg/min was assumed for this study. Examination of uncertainty in injection rate was accomplished during post-processing by adjusting the concentrations to achieve values for r_{inj} of 0.1, 1, 10 and 100 kg/min. The threshold for contamination, τ , of a node or pipe was set to 0.01 mg/L.

This injection information, along with a_{inj} and contaminant pattern, was used to define the contamination incident under the “sources” section of the EPANET input file. For all simulations, the duration was set for a period of 10 days to ensure the contaminant reached its maximum extent, even when the the injection start time was 40:00 and the injection duration was 24 hours.

Fifty stochastic simulations were generated for each set of parameter values, in which the demand model and isolation valve closure parameter values were based on a random number selected from a distribution. Based on the parameters listed in Table 1 for the KL system, a full factorial study should include 12,150 hydraulic and 7,581,600 water quality simulations ($3(\text{demand}) * 3(\text{valve}) * 3(\text{reaction}) * 3(\text{start}) * 3(\text{duration}) * 624(\text{location}) * 50(\text{stochastic})$). However, the final simulation study included 10,827 hydraulic and 6,756,048 water quality simulations (132,300 fewer simulations) since the combination of $\sigma = 0$ and $\kappa = 0$ did not create different stochastic systems for the hydraulic simulation. Likewise, a full factorial simulation study for N6 would include 20,120,400 water quality simulations; however, the final simulation study included 17,929,512 water quality simulations.

Impact and Uncertainty Metrics

Four metrics were used to quantify how model parameters affected the predicted contamination plumes. The first two metrics defined the direct impact of a specific simulation; these were extent of contamination (EC) and population impacted (PI) and were calculated for every simulation. The EC metric was the length of pipe that was contaminated. The PI metric was defined as the number of customers in the contaminated zone. Since changing model parameters resulted in plumes that covered different physical areas, the standard deviations for the EC or PI metrics were not necessarily representative of the uncertainty in the plume's impact. For EC in particular, several plumes had nearly the same total EC but due to differences in the parameters or in the stochastic hydraulics, the simulations contaminated completely different parts of the system.

To quantify this uncertainty in plume position, two additional metrics, the unknown extent of contamination (uEC) and unknown population impacted (uPI), were developed. To describe these additional metrics conceptually, the following example considers a set of simulations with the same injection node. Nodes that are contaminated in every simulation in the set, within some tolerance, are placed in the "contaminated zone." Nodes that are never contaminated in any of the simulations in the set are placed in the "uncontaminated zone." The remaining set of nodes, nodes that are contaminated in at least one of the simulations in the set, are placed in the "unknown status zone." The uEC and uPI metrics are used to quantify the "size" of the unknown status zone. Figure 4 illustrates the differences between EC and uEC. In this example, two different simulations of a contaminant injection at the same node (red) are depicted in Figures 4a and b. Since the hydraulics were different for Simulation 1 and Simulation 2, the contaminated nodes (yellow and blue, respectively) were different as well. To identify the uncontaminated, contaminated and unknown status zones, the two simulation results were overlaid in Figure 4c. The white nodes were not contaminated in any of the simulations (uncontaminated zone), the green nodes were contaminated in both simulations (contaminated zone) and the yellow and blue nodes were contaminated in at least one of the simulations (unknown status zone). The contaminated nodes for each simulation were used to calculate the EC metric and the nodes in the unknown status zone were used to calculate the uEC metric as described later.

The mathematical formulation for the four metrics is described here. For a specific simulation x , the contamination plume, $\mathcal{P}(x, \tau)$, contains any node that ever experienced a contaminant concentration greater than a threshold, τ , as stated in Equation 3:

$$\mathcal{P}(x, \tau) = \left\{ j \in \mathcal{J} \left| \max_t c_{tj}(x) > \tau \right. \right\} \quad (3)$$

where $c_{tj}(x)$ is the concentration at time t and node j for simulation x as calculated by EPANET.

For a set of simulations $\mathcal{X}$, $\mathcal{S}(\mathcal{X}, \tau)$ is defined as the set of all nodes contaminated in at least one of the simulations and is given by Equation 4, which is the union of all of the nodes in the contaminated and unknown status zones.

$$\mathcal{S}(\mathcal{X}, \tau) = \bigcup_{x \in \mathcal{X}} \left\{ j \in \mathcal{P}(x, \tau) \right\} \quad (4)$$

Rather than requiring a node to be contaminated in every simulation in order to be part of the contaminated zone, if a node is contaminated in at least $(1 - \alpha)$ percent of the simulations, it is then considered part of the contaminated zone. The threshold of contamination, τ , provides a lower limit of a node being placed in the unknown status zone, while the α threshold provides an upper limit. (As mentioned previously, τ was set to 0.01 mg/L for the simulations). The unknown status zone, $\mathcal{U}(\mathcal{X}, \tau, \alpha)$, for a set of simulations $\mathcal{X}$ is calculated based upon Equation 5:

$$\mathcal{U}(\mathcal{X}, \tau, \alpha) = \{ j \in \mathcal{S}(\mathcal{X}, \tau) | \mathbb{P}(j \in \mathcal{P}(x, \tau) | x \in \mathcal{X}) < (1 - \alpha) \} \quad (5)$$

where $\mathbb{P}(j \in \mathcal{P}(x, \tau) | x \in \mathcal{X})$ is the probability that a node is contaminated during a randomly selected simulation x from $\mathcal{X}$. For this study, α was set to 0.05, meaning that any node with a 95% probability or greater of receiving contaminated water was considered “contaminated” and therefore is not part of $\mathcal{U}(\mathcal{X}, \tau, \alpha)$. Because of the probabilistic approach used here, it should be noted that it is possible for nodes in the uncontaminated zone to be contaminated and nodes in the contaminated zone not to be contaminated. An upper confidence limit on the probability of contamination can be calculated using a binomial distribution based on the number of stochastic scenarios.

The extent of contamination, $EC(x, \tau)$, for a specific simulation, x , is the sum of the lengths of all pipes within the contamination plume that have received water with a contaminant concentration greater than the threshold of contamination, τ . Since EPANET does not output the concentration within pipes, node concentrations are used here along with the length of adjacent pipes. A simple approach was taken by assigning half of each pipe length to its two

endpoint nodes. Thus, the length of contaminated pipe is half the sum of the lengths of all pipes attached to contaminated nodes. The EC metric can then be written as Equation 6:

$$EC(x, \tau) = \sum_{j \in \mathcal{P}(x, \tau)} L_j \quad (6)$$

where L_j is the length of pipes assigned to node j . For Figure 4a, the contaminated nodes in Simulation 1 were F, H, I, J, L, N and O and the number of pipes connected to each nodes was 2, 3, 4, 2, 2, 3 and 1, respectively. Assuming that each pipe length was 1 m, then the EC metric calculated by Equation 6 was 8.5 m. The contaminated nodes in Simulation 2 (Figure 4b) were A, E, F, H, I and J, which were connected to 1, 3, 2, 3, 4 and 2 pipes, respectively. The EC metric for Simulation 2 was calculated as 7.5 m assuming a pipe length of 1 m each.

The unknown extent of contamination, $uEC(\mathfrak{X}, \tau, \alpha)$, for a set of simulations, $\mathfrak{X}$, is the sum of the lengths of pipes assigned to nodes within the unknown status zone and is defined by Equation 7:

$$uEC(\mathfrak{X}, \tau, \alpha) = \sum_{j \in \mathcal{U}(\mathfrak{X}, \tau, \alpha)} L_j \quad (7)$$

In Figure 4c, the nodes in the contaminated zone were F, H, I and J, the nodes in the uncontaminated zone were B, C, D, G, K and M and the nodes in the unknown zone were A, E, L, N and O. Using the number of pipes connected to the nodes in the unknown zone (1, 3, 2, 3 and 1, respectively) and Equation 7, the uEC metric was 5 m assuming the pipe lengths were 1 m each.

The PI and uPI metrics are defined similarly to the EC and uEC metrics. The population impacted, $PI(x, \tau)$, is the sum of the population at each node, Π_j , within the contaminated zone and is defined by Equation 8. The unknown population impacted, $uPI(\mathfrak{X}, \tau, \alpha)$, is defined by Equation 9:

$$PI(x, \tau) = \sum_{j \in \mathcal{P}(x, \tau)} \Pi_j \quad (8)$$

$$uPI(\mathfrak{X}, \tau, \alpha) = \sum_{j \in \mathcal{U}(\mathfrak{X}, \tau, \alpha)} \Pi_j \quad (9)$$

where Π_j represents the population at node j . The population at node j , Π_j , was calculated to be the original daily demand at node j divided by the demand per user, in this simulation a value of 200 gallons per person per day.

RESULTS

The EC, PI, uEC and uPI metrics were analyzed using MATLAB® main-effects method, ‘maineffectsplot’ (The MathWorks, Inc. 2015), to determine the parameters that resulted in the most uncertainty. For each parameter listed in Table 1, the main effects method calculates the average for each metric for the different values of the parameter. The results of these averages are then plotted for each parameter, providing a visual estimation of the affect of a specific parameter on a metric. Lines with an increasing slope show an increase in uncertainty with an increase in the parameter value. Parameters with greater effect have a higher absolute slope.

KL Simulation Results

The results for all metrics and parameters for the KL simulations are shown in Figure 5. The scale for the vertical axis was kept consistent for all the parameters, allowing the relative importance of the parameters and values to be compared. The injection location results have been sorted by the EC metric and are displayed in that order for all of the plots. As can be seen in Figure 5, the injection location parameter, a_{inj} , dominated all of the other parameters in its effect on the four metrics. It can also be seen that the correlation between EC and PI was very high; the correlation coefficient for these two metrics was 0.9948, and between uEC and uPI it was 0.9953. Because these metrics were so highly correlated, only the EC and uEC metrics are used and discussed in the remaining text and graphics.

Three example injection locations were chosen to highlight the results. These locations are marked with red diamonds on the map in Figure 1, in which the nodes are sized by the average EC of an injection occurring at the node and colored by the uncertainty in the contamination extent (uEC). These example injection locations were chosen to include locations that had the maximum average EC (KL-1, the reservoir), a large uEC (KL-326) and EC and uEC values between the other two example injection locations (KL-1427). Summary statistics for the EC and uEC metrics for these locations are provided in Table 2. The statistics are presented for the average over all injections locations with specific results for the three example nodes. Figure 6 shows the main effects plots for the KL system for (a) the average of all injection locations, (b) an injection at KL-1 (the reservoir), (c) an injection at KL-326 and (d) an injection at KL-1427 for the EC (blue square) and uEC (red circle) metrics with the effects of the injection location parameter removed.

Figure 6a shows the average of the EC and uEC metrics across all injection locations. For the EC metric (Figure 6a), the injection rate, reaction coefficient and the injection duration parameters had, averaged across all locations, the greatest effect on the size of the contamination extent (i.e., number of nodes contaminated). This general trend seemed to be true for the example injection locations (Figure 6b and d); for the injection at KL-326 (Figure 6c), the isolation valve closure parameters also had a greater effect. The uncertainty in plume extent (uEC) was affected significantly by uncertainty in the isolation valve closure and injection rate parameters, κ and t_{inj} . The other general trend is that decreasing the injection rate, t_{inj} , or increasing the rate of decay, K_d , decreased EC and uEC as both

parameters change the total mass within the system. The injection start time, t_{inj}^0 , and the stochastic demand model parameter, σ , had very small effect on all metrics.

These individual injection location results also highlight why the injection location parameter was so dominant. While the isolation valve closure parameter had the greatest effect on uncertainty when averaged across all injection locations, the injection rate parameter was of equal importance for an injection at KL-1. The reaction rate parameter had the opposite effect on uncertainty when an injection occurred at KL-326 and KL-1427. For an injection at KL-1427, the injection rate had more of an effect on uncertainty in the middle of the range than at either extreme. These very different behaviors suggest caution in making overarching statements based upon the average effects.

N6 Simulation Results

The results for all metrics and parameters for the N6 simulations are shown in Figure 7 using the same figure characteristics as Figure 5. Figure 7 shows that the injection location parameter dominated all of the other parameters in its effect on each of the four metrics. In Figure 7, it can be seen that more than half of the injection locations affected a small number of people and contaminated only small areas of the system. In contrast, only a few locations affected a large number of people and contaminated a large area of the system. For example, injecting at a node near the source affected the majority of the WDS, while an injection at a node on the edge of the system affected only a few nodes. Figure 7 also shows similar trends in the EC metric and the PI metric for all six parameters. The uEC and the uPI metrics again follow similar trends to each other which are different than the EC and the PI metric trends. Therefore, only the EC and uEC pair of metrics are shown for the rest of the figures.

When the injection location was removed from consideration, the trends related to the other parameters became clearer. Three example injection locations were used in the remainder of this section to illustrate the results. The reservoir, N6-3323, was selected as it had the largest total impact EC and low uncertainty in the contamination extent uEC. The other injection locations selected were N6-188, which had a relatively large extent of contamination EC and high uncertainty in uEC, and N6-1718, which had a relatively small extent of contamination EC and low uncertainty in uEC. Figure 2 shows these three nodes outlined by red diamonds on the system schematic. Nodes are sized by the average EC of an injection occurring at the node and colored by the uncertainty in the contamination extent uEC. The summary metric statistics for N6 simulations are included in Table 2.

Figure 8 displays the main effects in EC (blue square) and uEC (red circle) for (a) the average of all injection locations, (b) an injection at the reservoir, N6-3323, (c) an injection at N6-188 and (d) an injection at N6-1718. The overall trends for EC (blue square) were similar to those from the KL system, since the injection rate, reaction coefficient and injection duration parameters were the most significant. All of these parameters influenced the total area of the system that was ever contaminated. Increasing the injection rate provided more contaminant that could spread further into the system before becoming too diluted. High reaction rate coefficients decreased the concentration below the contamination threshold more quickly. Increasing the injection duration ensured that the contaminant

remained in the system longer and therefore increased the area affected. The demand, isolation valve closure, and injection start parameters had a smaller effect on EC. Isolation valve closures, changes in flow pattern due to demand variation and injection start times changed the arrival time or concentration of the contaminant downstream of the injection location, but eventually the same number of nodes were contaminated in the WDS model.

All of the parameters, except the injection start time, appeared to affect the uncertainty in the contamination plume extent, uEC, metric (red circle) when averaged across all locations (Figure 8a). For the N6 model, the effect of the parameter on the uEC metric depended on the injection location. For example, the injection rate parameter had the greatest effect on uEC metric for an injection at the reservoir (Figure 8b), while the other parameters had little effect. However, if the injection occurred at N6–1718, then all of the parameters appeared to have a significant effect on the uEC metric (Figure 8d).

Parameter Interactions

The data was also analyzed using the N-way analysis of variation (ANOVA) approach up through second order interactions (Rice 1995) by applying the MATLAB ‘anovan’ commands to the data. Figure 9 shows bar charts of the contribution to variation for different parameters and combinations using the η^2 statistic, which is the sum of squares of the parameter divided by the total sum of squares for all variation. The η^2 statistic is a measure of effect size and can be used to compare the relative affect of parameters. The “error” term of the ANOVA contained all effects due to experimental errors as well contributions due to higher order parameter combinations and any “hidden parameters.” The demand model parameter, σ , and the isolation valve closure model parameter, κ , were not measures of the actual demands or valves closed, but their variation. The specific node and time demand values and actual valve statuses would have created millions of additional parameters. Therefore, these parameters were wrapped up in the error term and “hidden” from the ANOVA model.

Figure 9 shows the sum of the main effects and all interaction effects of the EC and uEC metrics for the different parameters (and 1st order combinations), in which the cross parameter contribution is included in both of the separate parameter columns. For example, the contribution for the cross parameter $\sigma \times \kappa$ was included in the columns for both σ and κ . The figure shows the results by WDS and metric, and includes the contribution due to the a_{inj} parameter. If the assumption that most of the “error” contribution came from the actual hydraulic conditions (i.e., demands and valve status) holds true, then it accounted for 41% and 31% of the variability in uEC for KL and N6, respectively. This is significant, as these are the two parameters in which a utility can determine ahead of time or during real-time monitoring; the other parameters all relate to the specific contamination scenario. Having a single, accurate hydraulic model also makes real-time calculations much more tractable.

From Figure 9, the injection location was the most significant parameter in the variation of both EC and uEC metrics. Variation in pipe closures was a significant parameter for uEC, but not for EC when evaluating across all injection locations. As discussed previously, the injection rate, reaction coefficient and injection duration parameters affected the EC and uEC metrics, since these parameters determined how far the contaminant traveled before it

became too dilute to detect (see Figures 6 and 8). The length of the simulation, 10 days, might have been the reason that injection start time was almost insignificant.

Figure 10 shows the 1st and 2nd order contributions to the variance in the (a) EC and (b) uEC metrics for each of the example injection locations for each WDS model. Only those parameters that provide greater than 6% of the variation for at least one location and metric are shown. Parameters are ordered from left to right by decreasing maximum average contribution across all points. As with Figure 9, the “Error” term is made up of the influence of actual demands and valves closed by the stochastic models.

The results presented in Figure 10 reinforce the importance the injection location has on both the total contamination and also on which parameters contribute most to the uncertainty. For example, the injections simulated at KL-325 and N6-188 showed a very small influence from the stochastic effects (i.e., the actual demands and valves closed), but uncertainty in the number of valves closed, κ , was the most significant parameter in the uncertainty. This is the opposite of the results for injections occurring at KL-1 and N6-3323.

While the most important parameter was the injection location and the second most important parameter varied by injection location, the results highlighted the parameters that did not have as much effect on the metrics. Uncertainty in the injection time had very little affect on the uncertainty in the total extent of contamination. However, the injection time, t_{inj}^0 , would be important during a real-time response. In addition, the uncertainty in the injection duration had a very small affect on the uEC.

DISCUSSION AND CONCLUSIONS

The results of this case study indicated that the injection location was the largest source of uncertainty in predicting the size of a contamination plume (EC) and the total area of the plume (uEC). If a drinking water utility wanted to take samples in its WDS to determine the extent of a contamination plume, and therefore the area in need of remediation, these results indicated that determining where the contaminant was injected would help reduce uncertainty about where the plume is now and therefore where to sample. If the utility were to use a WDS model to assist in finding sampling locations, narrowing down a set of injection locations to a smaller likely set would reduce the uncertainty in model output. Several approaches for identifying the source of a contamination incident have been presented in the literature (e.g., De Sanctis et al. 2010; Mann et al. 2012; Perelman and Ostfeld 2013; Yang and Boccelli 2014; Wagner et al. 2015; Seth et al. 2016). Previous source localization efforts have not included additional levels of uncertainty beyond unknown injection characteristics. Results reported here show that additional sources of uncertainty can impact the contaminant plume, and inclusion of those parameters will improve plume estimation and sample site selection. Source localization approaches do not assume that the true system state is in the set of precomputed solutions, which makes accounting for uncertainty in the scenario a part of the challenge. Future source localization research could benefit from reporting a measure that describes the range of uncertainty where an algorithm is effective.

The results also showed that the reaction coefficient, the injection rate and the injection duration parameters had a large effect on the uncertainty of EC at most locations. If the contaminant is known or can be identified quickly using field tests, the value of the bulk reaction parameter can be determined and this source of uncertainty can be decreased. In many cases, the injection duration will be unknown during an incident and so the uncertainty due to this parameter can probably not be reduced.

The uncertainties due to the injection location, the reaction rate coefficient and the injection duration parameters can only be addressed during a particular contamination incident because these parameters will vary with each incident and cannot be predicted ahead of time. However, uncertainties due to isolation valve status can be addressed to a large degree prior to an incident. An asset management program in addition to a valve testing and repair program can help to ensure that the status of isolation valves is known and correctly represented in a WDS model.

Sampling throughout the WDS during an incident can also help to determine, and therefore reduce uncertainty about, the size and location of a contamination plume. Figure 11 shows that as the number of samples in the WDS increases, the unknown status zone of the plume decreases. In other words, a larger number of samples provides more information about which areas of the WDS need to be treated or decontaminated to ensure public health protection. In this figure, all of the parameter combinations previously simulated from Table 1 were considered as possible contamination scenarios, and the unknown status zone was then determined. Two samples were taken at random locations at hour 40, and the yes/no results of the samples were used to update the nodes in the unknown status zone by removing all simulations that did not match the results at those two sample locations. This zone is quite large as can be seen in Figure 11a, encompassing 619 km of pipe, essentially the entire network. Figure 11b shows that if six random samples were taken instead, the uEC is only 202 km and the number of nodes that are almost certainly contaminated has significantly increased.

This shows the potential for sampling to be useful in determining the extent of the contamination plume and to assist in pipe decontamination efforts. If optimization techniques were used to improve the sampling locations, the unknown status zone would be expected to decrease even further providing more accurate information about the location and size of the plume. Additionally, this shows that even if a real-time hydraulic model is unattainable or still contains significant uncertainty at specific locations, the sampling methods used are still able to reduce the uncertainty in the plume extent quickly and efficiently.

Since the injection location was identified as the most significant parameter, simulations that incorporate uncertainty in the injection location are needed (e.g., simulating injections at every location in the system). For the N6 simulations in this study, nearly twenty million simulations were necessary in order to cover all the possible contamination scenario variability. This research generally showed that demand variability was not a significant source of uncertainty, but accounted for 50 simulations per other combination of parameters. If that number were only 25 simulations, then the total number of simulations needed would

be cut in half. If the number of simulations were kept consistent, then approximately 10 million simulations could be used to investigate different reaction rates, more valve closures, or some other parameter that might be more relevant to a given system. While a large set of simulations might be necessary to investigate uncertainty in the contamination plume and where to sample to identify the exact location, these results can help to inform which parameters variations are more important to explore thoroughly.

The results of this study are limited by a few factors. The number of isolation valve closures was limited to only 0.5% of pipes, the maximum that was feasible for the models tested here. Larger networks, or all-pipe models, may have more redundancies within the network and would only be impacted by larger values of this parameter. Similarly, the demand variation was limited by the assumption that total system demand was represented accurately. In reality, the total system demand could be expected to vary from that given in the model due to seasonal patterns or specific changes in demand due to customers responding to the contamination incident. A similar simulation study needs to be conducted for additional WDS models in order to verify that the same trends across parameters apply. Uncertainty in the magnitude of the demand uncertainty had a very small impact, although the actual demands did. It would be worth exploring whether this holds true for other demand models; if so, then having multiple average daily use patterns that are more accurate than a global pattern would decrease uncertainty in plume extent considerably without necessarily needing highly accurate, continuous demand information. Future research could also focus on limiting the computational requirements by finding ways to eliminate scenarios, particularly if much faster network analysis methods or node/pipe characteristics can be used.

Overall, this study showed that there are several sources of uncertainty that can affect the size and location of a contamination plume in a WDS. The size of the plume was affected by the injection location, reaction rate coefficient, injection rate and the injection duration, while the exact location of the plume was affected by the injection location, isolation valve status and the injection rate. These sources of uncertainty can be grouped into hydraulic uncertainties and water quality uncertainties. Hydraulic uncertainties can be reduced through rigorous asset management, geographic information systems (GIS) mapping, automated meter reading and maintenance and repair programs. Water quality uncertainties might be reduced by additional security and the installation of backflow preventers by reducing the number of possible injection locations. In addition, these water quality uncertainties can also be reduced during a contamination incident when information is available about the type of contaminant and its reaction rate, the location or region of the contaminant injection, and the timing of the incident. This case study highlights the value of having a topologically accurate WDS model to assist with responding to contamination incidents. Determining the injection location and estimating the hydraulics can assist a drinking water utility in identifying where the best location is for sampling.

REFERENCES

- Bargiela A and Hainsworth GD (1989). "Pressure and flow uncertainty in water systems." *Journal of Water Resources Planning and Management*, 115(2), 212–229.

- Bi W, D. GC and Maier HR (2015). "Improved genetic algorithm optimization of water distribution system design by incorporating domain knowledge." *Environmental Modelling & Software*, 69, 370–381.
- Collette A (2008–2016). "Hdf5 for python, <<http://www.h5py.org>>.
- Davis MJ and Janke R (2011). "Patterns in potential impacts associated with contamination events in water distribution systems." *Journal of Water Resources Planning and Management*, 137(1), 1–9.
- Davis MJ, Janke R, and Magnuson ML (2014). "A framework for estimating the adverse health effects of contamination events in water distribution systems and its application." *Risk Analysis*, 34(3), 498–513. [PubMed: 24102461]
- De Sanctis A, Shang F, and Uber J (2010). "Real-time identification of possible contamination sources using network backtracking methods." *Journal of Water Resources Planning and Management*, 136(4), 444–453.
- Jones E, Oliphant T, Peterson P, et al. (2001–2016). "SciPy: Open source scientific tools for Python, <<http://www.scipy.org/>>.
- Kang D and Lansey K (2012). "Revisiting optimal water-distribution system design: issues and a heuristic hierarchical approach." *Journal of Water Resources Planning and Management*, 138(3), 208–217.
- Kirmeyer GJ, Friedman M, Martel K, Thompson G, Sandvig A, Clement J., and Frey M (2002). *Guidance manual for monitoring distribution system water quality*. American Water Works Association.
- Klise K, Sirola J, Hart D, Hart W, Phillips C, Haxton T, Murray R, Janke R, Taxon T, and Laird C (2014). "Water security toolkit user manual version 1.2." *Report no.*, Sandia National Laboratories (SNL-NM), Albuquerque, NM (United States).
- Klise KA, Bynum M, Moriarty D, and Murray R (2016). "Water network tool for resilience (WNTR) user manual, version 0.1." Report No. EPA/600/R-17/264, U.S. Environmental Protection Agency.
- Lansey KE and Basnet C (1991). "Parameter estimation for water distribution networks." *Journal of Water Resources Planning and Management*, 117(1), 126–144.
- Lansey KE, El-Shorbagy W, Ahmed I, Araujo J, and Haan CT (2001). "Calibration assessment and data collection for water distribution networks." *Journal of Hydraulic Engineering*, 127(4), 270–279.
- Mann AV, McKenna SA, Hart WE, and Laird CD (2012). "Real-time inversion in large-scale water networks using discrete measurements." *Computers & Chemical Engineering*, 37, 143–151.
- Perelman L and Ostfeld A (2013). "Bayesian networks for source intrusion detection." *Journal of Water Resources Planning and Management*, 139(4), 426–432.
- Plackett RL and Burman JP (1946). "The design of optimum multifactorial experiments." *Biometrika*, 33(4), 305–325.
- Rice JA (1995). *Mathematical Statistics and Data Analysis*. Duxbury Press, Belmont, CA, 2nd edition.
- Sanz G and Pérez R (2014). "Demand Pattern Calibration in Water Distribution Networks." *Procedia Engineering*, 70, 1495–1504.
- Seth A, Klise KA, Sirola JD, Haxton T, and Laird CD (2016). "Testing contamination source identification methods for water distribution networks." *Journal of Water Resources Planning and Management*, 142(4), 04016001.
- The HDF Group (1997–2016). "Hierarchical Data Format, version 5. /HDF5/.
- The MathWorks, Inc. (2015). "MATLAB Release R2015b.
- U.S. EPA (2010). "Assessing potential impacts associated with contamination events in water distribution systems: A sensitivity analysis." Report No. EPA/600/R-10/061, U.S. Environmental Protection Agency, Washington, DC.
- Wagner DE, Neupauer RM, and Cichowitz C (2015). "Adjoint-based probabilistic source characterization in water-distribution systems with transient flows and imperfect sensors." *Journal of Water Resources Planning and Management*, 141(9), 04015003.
- Watson J-P, Murray R, and Hart WE (2009). "Formulation and optimization of robust sensor placement problems for drinking water contamination warning systems." *Journal of Infrastructure Systems*, 15(4), 330–339.

- Yang X and Boccelli DL (2013). "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*, 140(8), 04014017.
- Yang X and Boccelli DL (2014). "Bayesian approach for real-time probabilistic contamination source identification." *Journal of Water Resources Planning and Management*, 140(8), 04014019.

Fig. 1.

Map of the KL WDS. Nodes are sized by the average EC of an injection occurring at the node and colored by the uncertainty. Nodes outlined with red diamonds are used as example injection locations, from top: reservoir KL-1, junctions KL-326 and KL-1427.

Fig. 2.

Map of the N6 WDS. Nodes are sized by the average EC of an injection occurring at the node and colored by the uncertainty. Nodes outlined with red diamonds are used as example injection locations, from left: reservoir N6-3323, junctions N6-188 and N6-1718.

Fig. 3.

Example of demand variation model with two stochastic realizations. Each node has demand modified independently for unique demand patterns for each hydraulic simulation.

Fig. 4.

Depiction of the extent of contamination (EC) and unknown extent of contamination (uEC) metrics in which (a) Simulation 1 (yellow) and (b) Simulation 2 (blue) with similar EC metrics form (c) the simulation set that is used to calculate the uEC metric based upon the non-overlapping parts of each plume, where overlapping parts are green. Injection occurs at the red node.

Fig. 5.

Main effects plots for all of the KL simulations for the (a) EC (blue square) and uEC (red circle) metrics and (b) PI (blue square) and uPI (red circle) metrics, in which the metrics are on vertical axis and the parameters are on the horizontal axis.

Fig. 6.

Main effects plots for the KL system for (a) the average of all injection locations, (b) injection at KL-1 (reservoir), (c) injection at KL-326 and (d) injection at KL-1427 for the EC (blue square) and uEC (red circle) metrics with the effects of the injection location parameter removed.

Fig. 7.

Main effects plots for all of the N6 simulations for the (a) EC (blue square) and uEC (red circle) metrics and (b) PI (blue square) and uPI (red circle) metrics, in which the metrics are on vertical axis and the parameters are on the horizontal axis.

Fig. 8.

Main effects plots for the N6 system for (a) the average of all injection locations, (b) injection at N6-3323 (the reservoir), (c) injection at N6-188 and (d) injection at N6-1718 for the extent of contamination EC (blue square) and uncertainty extent of contamination uEC (red circle) metrics with the effects of the injection location parameter removed.

Fig. 9.

For all injection locations combined, the sum of the η^2 for each parameter and the η^2 for the interaction of that parameter with all other parameters, taken from the ANOVA analysis. The total across columns will be greater than 1.0, as, for example, the columns for a_{inj} and κ both include the contribution from $a_{inj} \times \kappa$. The “Error” category includes both higher-order interactions and possible hidden parameters.

Fig. 10.

For the six specific locations chosen from KL (shades of blue) and N6 (shades of red), the η^2 contributions from ANOVA analysis of the (a) extent of contamination (EC) and (b) unknown extent of contamination (uEC) metrics.

Fig. 11.

The effects of sampling in changing the size of the contaminated and unknown status zones for (a) two random sample locations and (b) six random sample locations.

TABLE 1.

Experimental Design Parameters

Parameter	# Values	Variable	Values
Stochastic demand model	3	a	0, 0.2, 0.4
Stochastic valve closure	3	κ	0, 0.0025, 0.0050
Reaction coefficient	3	K_b	0, -0.1, -5.0
Injection start [†]	3	t_{inj}^0	24:00, 32:00, 40:00
Injection duration	3	t_{inj}	1 hr, 12 hr, 24 hr
Injection rate [*]	4	r_{inj}	0.1, 1.0, 10.0, 100.0 kg/min
Injection location	N_{nzd}^*	a_{inj}	NZD nodes, tanks, reservoirs

[†]Start times indicate time from start of simulation (not clock time)

^{*}For the KL model, $N_{nzd}=624$; for the N6 model, $N_{nzd}=1656$

^{*}Injection rate was varied in the post-processing step

TABLE 2.

Summary Statistics for the EC and uEC Metrics in kilometers of Contaminated Pipe for the Two WDS Models, KL and N6

Model, injection node	$\overline{EC}$	$EC_{50\%}$	EC_{max}	$\overline{uEC}$	$uEC_{50\%}$	uEC_{max}
<i>KL (all)</i>	<i>41</i>	<i>26</i>	<i>250</i>	<i>28</i>	<i>9</i>	<i>228</i>
a_{inj} : KL-1 (Reservoir)	181	223	250	15	< 1	83
a_{inj} : KL-326	32	< 1	175	104	120	196
a_{inj} : KL-1427	38	40	52	9	6	39
<i>N6 (all)</i>	<i>32</i>	<i>4</i>	<i>634</i>	<i>13</i>	<i><1</i>	<i>610</i>
a_{inj} : N6-3323 (Reservoir)	248	109	634	24	4	138
a_{inj} : N6-188	130	121	383	91	36	396
a_{inj} : N6-1718	24	20	186	40	21	184