Uncertainty Quantification of Contaminant Plume Spread in Water Distribution Systems

by

David Blaine Hart

Submitted in Partial Fulfillment
of the Requirements for the Degree of
Master of Science in Mathematics
with Specialization in Industrial Mathematics

New Mexico Institute of Mining and Technology
Socorro, New Mexico
December, 2016

I would like to dedicate this thesis to Sean McKenna, my mentor at Sandia National Laboratories since I was an undergraduate student. He championed me with my managers and coworkers, and I have always looked up to him both as a person and as what a scientist should be.

Most importantly, I dedicate this to my wife, Sara, who has supported me and cheered me on throughout this process. She has put up with several years of my working and going to school, and I could never have done it without her help.

David Blaine Hart
New Mexico Institute of Mining and Technology
December, 2016

ABSTRACT

Ensuring safe drinking water is an issue of concern across the world. Contamination events may be deliberate or accidental, of man-made or organic causes. Once an event is detected, effective response strategies are needed. Source localization and contamination impact have been previously studied under the assumption of known hydraulics. Hydraulics and other parameters are subject to uncertainty and the impact of this uncertainty on resulting contamination was examined. Metrics that can be used to quantify uncertainty in impact were developed. Simulation studies were used to attempt to quantify the uncertainty; Source code to perform the analyses on an arbitrary network was developed. The location of the contaminant source overwhelmed all other sources of uncertainty in both the extent of contamination and in the contaminant position within the network. After location, the main parameter impacting total extent of contamination is the reaction coefficient. The main parameter affecting the uncertainty in plume position is network topology. Methods to incorporate this information into existing source-inversion and manual sampling planning routines are proposed.

Keywords: UNCERTAINTY QUANTIFICATION; WATER DISTRIBUTION NETWORKS; CONTAMINATION RECOVERY; SOURCE LOCALIZATION

ACKNOWLEDGMENTS

I would like to thank my committee and professors at New Mexico Tech, and Sandia National Laboratories (SNL) University Programs Special Degree Program which made it possible for me to complete this research and coursework.

I would also like to thank the team members on the collaborative SNL and U.S. Environmental Protection Agency (EPA) water security project: Terra Haxton, Regan Murray, and Jonathan Burkhardt at the EPA National Homeland Security Research Center; Katherine Klise, Dylan Moriarty, and Carl Laird at SNL; and Michael Bynum and Santiago Rodriguez at Purdue University. The water security project is funded through an interagency agreement between Sandia National Laboratories and the U.S. Environmental Protection Agency.

Sandia National Laboratories is a multi-mission laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

This thesis was typeset with LaTeX[1] by the author.

[1]The LaTeX document preparation system was developed by Leslie Lamport as a special version of Donald Knuth's TeX program for computer typesetting. TeX is a trademark of the American Mathematical Society. The LaTeX macro package for the New Mexico Institute of Mining and Technology thesis format was written for the Tech Computer Center by John W. Shipman.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AMR	automatic meter reading
DOE	design of experiments
EC	extent of contamination
eCDF	empirical cumulative distribution function
EPA	U.S. Environmental Protection Agency
i.i.d.	independent identically distributed
MPI	message passing interface
NZD	non-zero demand
PI	population impacted
PRP	Poisson rectangular pulse
PSI	pounds per square inch
SCADA	supervisory control and data acquisition
SP	sensor placement
uEC	unknown extent of contamination
uPI	unknown population impacted
UQ	uncertainty quantification
WDS	water distribution system
WNTR	Water Network Tool for Resilience
WQ	water quality
WST	Water Security Toolkit

LIST OF SYMBOLS

Algorithm Notation

$b := a$	assign value in variable a to variable b
$\mathbf{A} \odot \mathbf{B}$	element-wise multiplication of matrices (or vectors) $\mathbf{A}$ and $\mathbf{B}$
$\mathrm{L}\mapsto\mathrm{F}(\dots)$	call function F from library L
$b.c$	member variable c of the object (structure) b
$\textsc{Func}(x, y)$	a function with parameters x and y
$b.c \leftarrow a$	assign value in a to member c of object b

Water Distribution System (WDS) Network Model

$\mathfrak{G}$	the WDS network model object; contains as members (as $\mathfrak{G}.-$) all the following variables and the variables in the *Matrices and Vectors* section, below
$\mathcal{E}$	graph edges; includes pipes, valves and pumps
$\mathcal{N}$	graph nodes; includes junctions, tanks, and reservoirs
$\mathcal{H}$	hydraulics (calculated demands, velocities, flows)
$\mathcal{J}$	junctions; nodes which are neither tanks nor reservoirs
$\mathcal{P}$	pipes; edges that are neither pumps nor valves
t_{total}	total simulation time
$\mathcal{T}_{hyd}$	hydraulic solution times
$\mathcal{T}_{pat}$	multiplier pattern start times
$\mathcal{T}_{wq}$	water quality solution times
$\mathcal{T}_{rpt}$	EPANET model output times
n_N, n_J, n_P	number of nodes, junctions, and pipes, respectively
n_{pat}, n_{rpt}	number of pattern and output steps, respectively
$\Delta t_{hyd}, \Delta t_{wq}$	maximum step size for hydraulic and water quality solvers
$\Delta t_{pat}, \Delta t_{rpt}$	step size for patterns and output reporting, respectively

Matrices and Vectors

$\mathbf{b}$	base demand; (size n_J)
$\mathbf{C}$	concentration in mg/L; (size $n_{rpt} \times n_N$)
$\mathbf{d}$	pipe diameter; (size n_P)
$\mathbf{e}_j$	index vector; a column vector where every value is 0, except in position j, where it is 1

$\boldsymbol{\lambda}$	node-based pipe length; (size n_N)
$\boldsymbol{\ell}$	pipe length; (size n_P)
$\mathbf{M}$	demand pattern multipliers; (size $n_{pat} \times n_J$)
$\boldsymbol{\pi}$	population; (size n_J)
$\mathbf{Q}$	final calculated demands; size $n_{pat} \times n_J$
$\mathbf{s}$	pipe status [open or closed]; (size n_P)

Parameters

K_b	bulk reaction coefficient (day^{-1})
β_f	topology model parameter; fraction of pipes to close
σ_d	demand model parameter; standard deviation as percentage of original demand
δ_d	maximum diameter pipe to include in pipe closure model
Δt_inj	injection duration in hours
x_inj	injection location as node index
$t_{0,\mathsf{inj}}$	injection start time in hours after $t_{hyd} = 0$
τ_c	concentration threshold; level required for element to be "contaminated"
α_c	determines if a contamination status is "unknown"

Results

$\mathfrak{P}$	contaminant plume (set of contaminated nodes)
ξ	scenario; the parameters and results for a specific simulation
Ξ	scenario set; a set of simulations, typically with one or more parameters in common
$\mathcal{Z}_\mathcal{U}$	uncontaminated zone (set of nodes)
$\mathcal{Z}_\mathcal{C}$	contaminated zone (set of nodes)
$\mathcal{Z}_\mathcal{S}$	unknown status zone (set of nodes)

This thesis is accepted on behalf of the faculty
of the Institute by the following committee:

Brian Borchers

_____

Academic Advisor

_____

Research Advisor

William Stone

_____

Committee Member

Oleg Makhnin

_____

Committee Member

Carl Laird

_____

Committee Member

I release this document to New Mexico Institute of Mining and Technology

David B. Hart Nov 16, 2016

_____

Student Signature Date

CHAPTER 1

INTRODUCTION

There are many reasons that water utilities may need to quantify the uncertainty of chemical concentration within a water system. Understanding the system's dynamics can help calibrate treatment processes, optimize operations and anticipate possible physical problems; these benefits can be implemented for the everyday, normal operation of the system. However, it is the worst-case scenarios where uncertainty analysis will provide the greatest help. Cases of treatment failures, such as in Flint, Michigan,[2] or source water contamination, like in West Virginia,[14] occur more often than we would hope. Intentional contamination is another possibility, one we hope never happens – but one we must consider. In all these cases, water demand, network flow, and water quality reactions can change drastically, and determining the uncertainty in the real water quality becomes very important.

The goal of the research presented here is to aid in response and recovery from one of these abnormal events – it is not intended for real-time, rapid source localization. However, because contamination events are rare, any tools must also give utilities benefits during normal operations as well. Being able to quantify uncertainty, not just identify it, will provide benefits in both realms.

Real-time modeling of network hydraulics is an emerging technology; research and tool development for calibrating models in real-time using hydraulic data is already being undertaken. However, most utilities do not yet have this technology installed. This research examines how an average demand, extended period model can be used to efficiently quantify uncertainty in water quality during normal operations using limited water-quality sampling information. Once uncertainty during normal operations has been quantified, the change in uncertainty during abnormal events can be examined.

1.1 Background

This section, on water distribution system (WDS) network models, is addressed to a reader without a background in drinking water systems. The basics of WDS networks and modeling will be given to help such a reader understand the simulation studies and unique challenges in the drinking water domain.

Drinking water utilities are unique among critical infrastructure systems. Their product, potable water, has a direct health impact on their customers. Providing clean water is therefore of equal importance to satisfying demand. Drinking water must be treated to kill or remove harmful organisms and impurities at the upstream plant and must also stay clean as it travels through pipes, tanks, and other machinery in the distribution system. This yields two separate, but related models – the hydraulic model and the water quality model.

Water utilities frequently manage both the "upstream" water treatment facilities and the actual distribution network: the pipes, tanks, and other associated physical mechanisms used to transport the clean water to the customer. Occasionally, the "upstream" water may be purchased from a wholesaler who moves pretreated water to the sources – tanks and reservoirs – used by the distribution network. However, these upstream systems can be modeled as simple inputs in the distribution network model. Water utilities may also handle "downstream" water, i.e., storm drains, gray water and sewers. Again, these models can be disconnected from the drinking water network, as they can be lumped in with all other the consumption by customers (they should be physically disconnected too, for safety's sake).

Part of the reason model isolation is possible, and a critical difference from other types of water systems, is the high pressure within distribution pipe networks. The American Water Works Association (AWWA) recommends a minimum pressure of 20 pounds per square inch (PSI) at all points of the network under any simulated conditions.[1] This pressure is much higher than the pressure in attached drainage systems, which should make cross-contamination unlikely (though it can occur). The pipe network of a drinking water system is also highly branched. This creates situations where flow directions change rapidly, and the actual upstream source of the water may be unknown. In-pipe hydraulic sensors are expensive and hard to maintain – most pipes are buried under streets, after all. Water quality sensors are possibly harder to maintain, as several types require reagents or can only work under limited pressures. When compared to highly-monitored process control systems such as chemical plants, the entire state of the distribution network is essentially unknown (see discussion of real-time models in section 1.2).

Another complicating factor is that the network model a utility possesses is likely out-of-date. Some utilities only map the pipes above a certain diameter – six-inch mains and larger, for example. Some utilities have been running for well over one-hundred years, and it is possible that the existence of certain pipes or connections may be unknown. Others may try to map all pipes, but with the continuous construction most cities see, renovation or repair projects, or even from simple human error, a network that is up-to-date one day will be at least slightly out-of-date the next day.

Finally, most network models are based on 'average demand.' This means that the water usage at each node within the network is based on the average usage by customers living or working at that node. Real-time usage for every service node is not generally available. Improvements in automatic meter read-

ing (AMR) technology are bringing the water world closer to this point, and software based on real-time hydraulic measurements is actively being developed, but most water utilities do not have this technology available to them at this time.[16] The primary source of uncertainty in water distribution models comes from the unknown – and currently unknowable – real time water use.

1.1.1 Hydraulic model

The most widely used hydraulic solver for drinking water systems is the EPANET[26] software. EPANET solves the hydraulics using the Todini method[27]. The topology of the network can be given as a directed graph, $(\mathcal{N}, \mathcal{E})$, where $\mathcal{N}$ are the vertices (nodes), and $\mathcal{E}$ are the edges (links). The solver is demand driven, which means that, if demand cannot be met due to insufficient pressure within the network, the solver will fail.

The following physical and non-physical parameters are necessary to model the water network hydraulics. Where appropriate, a comparison to electrical circuits is made to help describe relationships; hydraulic head can be loosely imagined as voltage, water flow as current, usage demand as energy use, and pipes as wires. However, these comparisons are *only* for illustrative purposes, as the actual equations do not coincide. The description of the hydraulic and water quality model, below, are paraphrased from the EPANET 2.0 User's Manual[26].

Nodes Considering the water distribution network as a graph, a node is a vertex on the graph. Junctions, tanks, and reservoirs are the physical elements of the WDS that are also nodes within the EPANET model.

Junctions Junctions are points where two or more pipes meet. They are also the points where water enters or leaves the system. Every junction has a base demand, $\mathbf{b}_{(n_J)}$, given in units of volume per time, describing a constant flux at that node. Total demand, $\mathbf{Q}_{(n_J \times n_{pat})}$ is calculated using the base demand and a time-dependent scaling factor (see "Muliplier patterns," below). In the electrical conceptual model, energy is being used (or provided) at these points in the circuit. Junction information is stored in the indexed set $\mathcal{J}$ such that $\mathcal{J} \subseteq \mathcal{N}$.

Population at a junction, $\boldsymbol{\pi}_{(n_J)}$, is not stored or tracked in the EPANET input files. Other researchers have used a linear relationship between the daily demand and the population at the junction.[4] This study assumes a usage of 200 gallons per person per day. Thus, the population can be calculated as

$$\boldsymbol{\pi} = \mathbf{Q}\vec{1}\,\frac{1}{n_{pat}\Delta t_{pat}}\,\frac{1}{200}. \tag{1.1}$$

Reservoirs "Reservoirs are nodes that represent an infinite external source or sink of water to the network. They are used to model such things as lakes, rivers, groundwater aquifers, and tie-ins to other systems. Reservoirs can also serve as water quality source points."[26] Upstream water treatment plants are modeled in these systems as reservoirs, and the hydraulic head of the aquifer can be changed on a pattern to model output from the plants. In this case, most disinfectants, such as chlorine, will enter the system at the reservoirs. These can be imagined as constant voltage sources with infinite current or as grounds.

Tanks Tanks are used within water distribution systems to store water for high-demand situations and to provide additional water pressure – like the battery in a hybrid car. Many utilities will fill tanks overnight – when electricity is cheaper and demand for water is lower – and then allow the tank to drain during the day to provide water to customers as demand goes up. Tanks are specified using a volume curve to determine storage and elevations to determine hydraulic heads. Tanks also are frequent sources for water quality changes.

Links Considering the water distribution network as a digraph, the edges of the graph are called links in the EPANET model. The physical elements of the WDS that can be links are the pipes, valves, and pumps.

Pipes Pipes transport water between junctions. They are links in the graph, and are stored in the indexed set $\mathcal{P}$, such that $\mathcal{P} \subseteq \mathcal{E}$. Pipes have roughness, diameters, $\mathbf{d}$, and lengths, $\boldsymbol{\ell}$, specified as part of the network topology. These are used to compute the headloss, flow rate, and velocity in a pipe. In the electrical system analogy, pipes are wires with a resistance determined by the roughness and flow rate, and the headloss is the voltage drop.

The roughness coefficient must be determined empirically, and can vary with age and pipe material; however, it is not a parameter of variation in this research. Each pipe can also be assigned a minor loss, which accounts for turbulence at bends, fittings, and junctions. These losses are not varied in this thesis.

Most pipes are bidirectional edges within the graph, but pipes can also have check valves that restrict flow to a single direction. Pipes can also set to "closed," stopping all flow. In the accompanying simulation studies, pipe status will be set at the beginning of the simulation and remain constant.

Pumps Pumps increase the hydraulic head across a link. Pumps are used to increase pressure in the system, fill tanks, or move water from one part of the system to another. EPANET automatically shuts off a pump if the hydraulic solver requires the pump to exceed its operational limits, or if the pump drains all the water out of the adjacent pipe, or if water tries to flow backwards through the pump. Unfortunately, the reality is that pumps will not automatically shut off

in many cases, and would instead burn out their motor. EPANET gives only a warning when the automatic shutoff occurs. The uncertainty quantification (UQ) process will instead discard the scenario as infeasible. Pump links are directed edges in the network graph.

Valves There are eight different kinds of valves, but most are not used in the case studies. In general, valves are used to reduce pressure or reduce flow. Isolation valves, used to isolate a section of a pipe or network during repairs or construction, are not represented as valves in the network graph. Isolation valves are modeled by setting a pipe status to open or closed.

Non-physical elements The non-physical elements in an EPANET model modify behavioral options of different nodes or links. Most of these characteristics will not be modified during this study; however, many water distribution systems try to optimize these parameters when designing expansions or trying to minimize energy usage.

Multiplier pattern Demand by customers varies in time. EPANET defines demand multipliers, **M**, which scale the base demand at a junction based on a repeating pattern. Most EPANET model inputs define patterns on a 24-hour cycle, but there is no requirement to do so, nor is there a requirement that any two patterns contain the same number of steps; all patterns will repeat if the simulation is longer than the patterns length. There **is** a requirement that the pattern step size, $\mathcal{T}_{pat}$, will be constant for all patterns.

Many EPANET network models use only a few patterns, applied to all junctions; the usage is scaled for a specific hour, based on some monthly average. In this thesis, the simulation studies will have a separate pattern for each junction. The patterns will be created with n_{pat} entries, one for each time within the simulation, eliminating the need to repeat a pattern. An example of a multiplier pattern, from the network used in the simulation study, is presented in Figure 1.1.

Volume curves Volume curves are used to model the storage capacity of a tank as a function of water level. While it is easy to idealize a tank as a cylinder with a radius that varies as a function of height, this is rarely the case in reality. An empirical function $V(h)$ is given as points on the volume curve instead, and linear interpolation is used for heights not directly defined. In this study, all tanks were defined as cylinders.

Pump curves Pump curves are used to describe the "relationship between the head and flow rate that a pump can deliver at its nominal speed setting."[26] In electrical terms, the curve describes a non-linear, concave relationship between

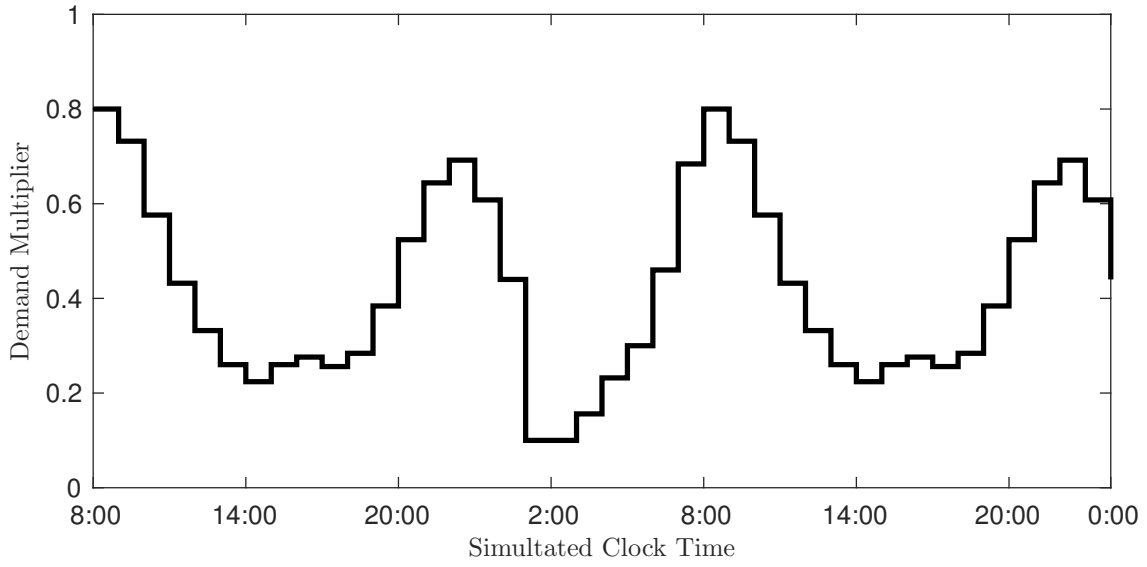

Figure 1.1: Example multiplier pattern. The pattern is constant for one-hour increments and repeats itself after 24-hours.

current and voltage, like a transformer. The end of the pump curve, where the head gain is zero, describes the maximum flow the pump is capable of handling. To try to push more flow will burn out the pump, causing EPANET warnings and errors described above. In this study, burning out a pump will be considered an infeasible scenario.

The network used for the simulation study uses only three-point pump curves. The EPANET model uses three points to fit a nonlinear model of the form $h_G = A - Bq^C$, where h_G is hydraulic head and q is the flow through the pump. An example curve from the network used in the simulation study is shown in Figure 1.2 to illustrate this type of curve.

Items not used The following items are important to WDS modeling, but are not present in the network used in the simulation study.

Headloss curves These are used to describe the loss in head as a function of flow rate. These act like a variable resistor that changes resistance based on the current applied. Headloss curves apply only to certain types of valves.

Emitters These elements are used to simulate pipe leaks, fire hydrants or sprinker systems. These are generally added to the model manually for a specific fire flow study.

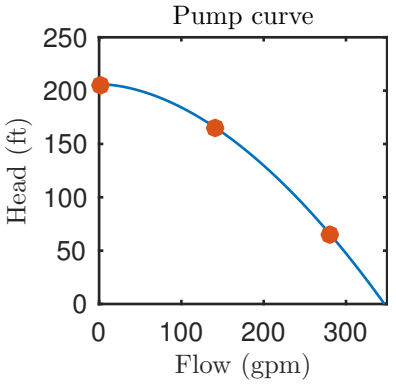

Figure 1.2: Example three-point pump curve.

Efficiency curves These curves relate pump efficiency (in terms of electrical energy) to flow rate, are used in energy consumption calculations. This feature is used to minimize cost for filling tanks and moving water and has no impact on the hydraulic or water quality models.

1.1.2 Water quality model

The water quality simulation models chemical species transport and reaction within the network. There are several reaction models that describe different chemicals. For example, chlorine has a well developed model that describes concentration as the water travels through the network; the model uses both time-based decay as well as loss due to reaction with pipe walls and other elements. This is in stark contrast to a biologic contaminant that will increase in concentration unless treated, or a completely non-reactive species such as fluoride.

Water quality is solved in EPANET using 'packets' of water. Each packet has a single concentration value and volume, and will be split when flow splits, and averaged when flow merges. The water quality step should be kept small enough for realistic transport. Because a single link will contain a varying number of packets at any given period, the concentration of a link is not well defined. It is frequently represented as the midpoint between the concentrations assigned at the two endpoint nodes, though this does not account for a high-concentration packet within the pipe. However, as concentration within pipes is not the concentration of interest for this study, the midpoint value is sufficient for graphically conveying information.

There is an alternative water quality model by Mann et al. [23], that is significantly faster than the EPANET model because it is a reduced order, linearized approximation. Despite being a reduced order model, it produces comparable results to the full model; however, for consistency with other studies, the EPANET model will be used here. There is another alternative model, called the EPANET

Multispecies Extension (EPANET-MSX). This extension allows multiple chemical species and their interactions to be modeled within the EPANET water quality model. This extension is significantly more computationally intensive, and will not be used.

1.2 State of technology: real-time models and data availability

There is ongoing effort to integrate EPANET with on-line monitoring to provide real-time updates to the hydraulic model. One such effort is the EPANET Real-Time Extension (EPANET-RTX). The tool optimizes the internal hydraulic model to best match the observed pressure conditions. *In situ* pressure sensors are becoming more common, but are still generally limited to a small fraction of nodes within a network. EPANET-RTX does not use water quality data in its optimization, nor does it provide quantified uncertainty on the input demands or the water quality results.

Real-time data acquisition is improving, but many utilities do not have the infrastructure in place to take advantage of those improvements yet. AMR is generally used only to gather monthly water usage for billing. AMR is capable of providing more frequent measurements, but as a low powered system, data communication requires the reader to be nearby. On a per-house basis, most demand is of short duration, such as flushing a toilet or using a sink, and so usage is still generally aggregated. Research by Yang and Boccelli [30], has shown that the aggregation length can play a significant role in determining contaminant flow direction and the time of arrival.

While real-time hydraulic measurements are becoming more common, real-time concentration measurements are still very limited. Water "quality" is a rather nebulous condition, and is typically described by surrogates such as chlorine concentration, electrical conductivity, pH, as well as specific regulatory measurements, such as total coliform counts or organic matter. The sensors to measure these values are sensitive, expensive, and generally don't work at high pressures, making them hard to install and maintain. Bypass lines can be used to solve the pressure issue, but continuous sensors are still very uncommon outside of the treatment plants or nearby tanks and reservoirs. Most utilities use manual sampling to satisfy regulatory requirements.[11, 12, 13] This involves sampling from specific points within the network over a month's period of time, and analyzing the samples in the lab. Rapid sensors and kits are available for field use for certain surrogate parameters.

Because of these challenges, there is continuing research in the development of tools to determine the source of a contaminant, plume size and location, and sampling direction during a contamination incident. These tools must be fast enough to run in real-time, or must have taken sufficient uncertainty into account to be useful under abnormal conditions even though solutions were precomputed. This thesis is aimed at providing methods to quantify the uncertainty

in plume position and identify the most significant contributing factors in reducing this uncertainty.

1.3 Literature review

Hatchett et al. [15] described the software tool EPANET-RTX [16] which uses known information from a WDS supervisory control and data acquisition (SCADA) to calibrate a hydraulic model based on real data. The software uses real-time information regarding network inflows, tank levels, pumps and valve settings to calculate the real-time system-wide demand. This is used as input to a demand estimation model to estimate instantaneous nodal demands. Forward predictions are compared to observations to calibrate the model. The software is being used at multiple locations and development is ongoing.

Yang and Boccelli [30] described the impact of temporal demand aggregation on water quality within a WDS. They examined three levels of demand aggregation (1 hour, 10 minutes, and 1 minute) and the effect this had on arrival time and time to maximum loading at all nodes. Demands were generated using the PRPsym[7] software, an implementation of the Poisson rectangular pulse (PRP) model proposed by Buchberger and Wu [8] to simulate stochastic residential demands. Yang and Boccelli focused specifically on short-duration, deliberate contamination incidents, which are more difficult to detect than a continuous injection. They found that there were significant differences, of up to 1000 minutes, in the arrival times of the one-hour and the one-minute demand simulations. They concluded that traditional deteriministic demand modeling approaches can be highly inaccurate in the predictions of contamination spread and arrivals.

Demand modeling has received significant attention in the water distribution system modeling literature, and is still an active research area. In addition to the PRP model described by Buchberger and Wu, numerous other stochastic and deterministic models for demand have been proposed. Davidson and Bouchart [9] and Kang and Lansey [18] both discussed the need for stochastic demands to accurately model water distribution system behavior.

Al-Jasser [3] examined the effects of service age on chlorine decay within the pipe network. They found that reactions within cast iron, and to a lesser extent steel pipes, were most significantly impacted by service age. The wall decay constants were found to vary from -92% to +431% compared to new pipes. This information helps provide bounds on parameter variation.

Blokker et al. [5] described a method of "bottom-up" demand estimation, creating stochastic demands at the household, or service node level rather than distributing the demands based on average system-level demands with a single, average demand pattern for each node.

Jonkergouw et al. [17] described calibration of the hydraulic model using chlorine measurements within the system. The calibration is performed on the system-wide demand multiplier pattern values, the wall correlation coefficient

parameters, and the source chlorine parameters. Calibration was done using a stochastic optimization algorithm followed by a derivative optimization algorithm to minimize the difference between the observed and modeled chlorine.

Xie et al. [29] examined the optimal placement of chlorine sensors for the calibration of chlorine decay model coefficients. A mixed integer programming method was used to determine the best placement. Calibration was done on two simulated case studies and one field data set from a small system. Calibration under the ideal simulated scenario resulted in recovery of the true parameters with error $\leq 2.3\%$. They concluded that using water quality data is an effective way to enhance hydraulic model calibration when combined with traditional hydraulic observed data.

1.4 Uncertainty quantification problem

This thesis will approach quantifying the uncertainty in the final contaminant plume from a deliberate or accidental contamination event. The uncertainty in the actual concentration values will not be quantified; rather, the binary contamination status – whether the concentration exceeded a certain threshold – will be used instead. Typical impact measures include the extent of contamination and population impacted.

While these metrics describe the total impact on a system, they do not aid in determining where the plume actually traveled. To this effect, two new impact measures will be proposed and evaluated. These impact measures will be used to quantify the uncertainty in the plume position by describing the size of the "edge" of the plume, the parts of the network where contamination is uncertain. These metrics will be computed from concentration results from a sample EPANET water distribution network model. Several different hydraulic and water quality parameters and injection scenarios will be considered.

CHAPTER 2

METHODS

In this chapter I will describe the methods used to perform the analyses on the WDS model. I will start with a discussion of the approach and the algorithms developed to provide uncertainty within the input parameters. I will then present the impact metrics that quantify the output of the EPANET model in a scalar or vector form; this includes the development of two new metrics for quantifying uncertainty in the position and system impact of a contamination event. Finally, I will describe the specifics for the simulation studies and the analysis techniques to be used.

2.1 Algorithm development

When demand and pipe status are per-node and per-pipe parameters, the number of parameters is far larger than the possible set of observations. A straightforward sensitivity analysis of all these individual parameters is computationally intractable, as there are simply too many parameters, all of which are continuously variable. Instead of per-node and per-pipe variation, the hydraulic parameters of demand and pipe closures will be varied as two sets of independent identically distributed (i.i.d.) random values. The bulk reaction coefficient has a logarithmic analytical solution which only needs three different values to compute at each node, and the injection scenario is a reasonable sized discrete parameter set. This reduces the parameter set to six factors, two hydraulic and four water quality related.

2.1.1 Hydraulic uncertainty models

Using a stochastic demand model is consistent with current and existing research [5, 18, 30]. However, there are many different views on how to generate such demands, and developing an exact model is likely system dependent. Real-time hydraulic measurements and frequent AMR will help reduce the uncertainty in actual demands in the system, but uncertainty will still remain. This uncertainty comes from the fact that, even with AMR reporting hourly from every house – a nearly unheard-of feat – it is still an hourly aggregation. The one

measurement that can be counted on is total system demand, based on outflow to the distribution system from reservoirs or pipelines.

Because developing an exact demand model is outside the scope of this thesis, a simple model will be used in the simulation study. It will be assumed that

- the system demand patterns were appropriately created based on average use for a specific demand class (residential, industrial, etc.),
- the correct demand pattern and base demand were assigned to each node to correctly represent the average demand at the node, and
- the total of junction demands accurately represents the total system demand on an hourly basis.

Given these assumptions, the demand modification model was developed as Algorithm 1. The demand at each node and time was adjusted by sampling from a normal distribution, with mean 1.0 and standard deviation σ_{d}. Because there is a chance that the random number will negative, especially for large σ_{d} values, a truncation is needed at zero; negative demands in the model would represent back-flow, which should be impossible. Once all the stochastic modifications are made, the demands are rescaled so that the total demand is equal to the original demand.

Algorithm 1 Demand modification model.

function MODIFYDEMANDS($\mathbf{b}, \mathbf{M}, \sigma_{\mathsf{d}}, \mathbf{Q}$)
  $\mathbf{\Psi} :=$ ZEROS(n_{pat}, n_{pat})
  $\mathbf{Y} :=$ NORMAL($0,\ {\sigma_{\mathsf{d}}}^2,\ [n_{pat},\ n_J]$) ▷ $n_{pat} \times n_J$ *i.i.d. samples from normal distribution*
  $\mathbf{B} :=$ diag($\mathbf{b}$)
  $\mathbf{Q} := \mathbf{M}\,\mathbf{B}$
  $\mathbf{Q} := \mathbf{Y} \odot \mathbf{Q}$
  $\mathbf{Q} :=$ MAX($\mathbf{Q},\ \mathbf{0}$)
  for $j = 1, 2, \ldots, n_{pat}$ **do**
    $\psi_{j,j} := \|\mathbf{B}^\top \mathbf{M}^\top \mathbf{e}_j\|_1 \ / \ \|\mathbf{Q}^\top \mathbf{e}_j\|_1$
  end for
  $\mathbf{Q} := \mathbf{\Psi}\,\mathbf{Q}$
  return Q ▷ *total demand per node per pattern step*
end function

The topology uncertainty model was limited to uncertainty in closures of isolation valves, and is given in Algorithm 2. I did not find any research available that provides an estimate of how many incorrectly marked valves there are in a typical network model. It is unlikely, but not impossible, that major lines are incorrectly noted, but the status for smaller isolation valves could easily be wrong,

which is the reason topology must be considered as a source of uncertainty. Anecdotal evidence gained through interviews with industry personnel indicate that in a large city there could be thousands of valves in the network that are shut, or partially shut, without anyone knowing. If any part of the network were to be fully isolated, it would become known very quickly, so any modification to the topology of the model must still satisfy demand to all nodes.

EPANET models typically only explicitly contain valves that are operated on a regular basis, and isolation valves do not fit that description. Because of this fact, isolation valve closures are modeled by closing an entire pipe in EPANET. The pipe closure model uses a random closure percentage from a uniform distribution. A threshold parameter is used to convert the random numbers into 0/1 status values. If a pipe is closed, it is closed for the entire simulation.

This brute-force method of closing pipes generates infeasible scenarios – network isolation and infeasible hydraulics occur with increasing frequency as the threshold is increased. While it may seem faster to do a graph theory analysis to find isolated parts of the network, the need to ensure that the hydraulics will solve, even without isolated areas, makes it more efficient to simply run the hydraulic model first and see if it solves. If the model does not solve, EPANET will return an error message and the hydraulic scenario is discarded and a replacement is generated.

Algorithm 2 Pipe closure model.

```
1:  function MODIFYTOPOLOGY(s, d, β_f, δ_d)
2:      for p = 1, 2, ..., n_P do
3:          if d_p ≤ d_max then
4:              y := UNIFORM(0, 1)                ▷ sample from uniform distribution
5:              if y ≤ β_f then
6:                  s_p := 0
7:              end if
8:          end if
9:      end for
10:     return s
11: end function
```

The hydraulic simulation combines the demand and topology uncertainty models with the EPANET hydraulic model, and can be implemented using Algorithm 3. The resulting system hydraulics are used as input to the water quality simulation. The input parameters and their theoretical ranges are listed in Table 2.1.

2.1.2 Water quality uncertainty models

Water quality uncertainty that is not due to hydraulics comes from the contamination profile. The contamination profile consists of the answers to the

Algorithm 3 Hydraulic simulation with uncertainty.
Note that the while loop only exits after a successful EPANET hydraulic solution is obtained; unsuccessful EPANET execution results in both the modified demands and topology being discarded.

function DoHydraulics($\mathfrak{G}, \sigma_d, \beta_f$)
  $\mathcal{H} := \emptyset$
  repeat ▷ *run until a hydraulic solution is found*
    $\mathbf{Q}$:= ModifyDemands($\mathfrak{G}.\mathbf{b}$, $\mathfrak{G}.\mathbf{M}$, σ_d)
    $\mathfrak{G}.\mathbf{M} \leftarrow \mathbf{Q}$ ▷ *assign new total demands as multipliers,*
    $\mathfrak{G}.\mathbf{b} \leftarrow \mathbf{1}$ ▷ *so need to set the new base demands to 1*
    $\mathbf{s}$:= ModifyTopology($\mathfrak{G}.\mathbf{s}$, $\mathfrak{G}.\mathbf{d}$, β_f, $d_{max} = 12''$)
    $\mathfrak{G}.\mathbf{s} \leftarrow \mathbf{s}$
    $\mathcal{H}$:=EPANET2$\mapsto$RUNHYDRAULICS($\mathfrak{G}$)
      ▷ *EPANET returns null set if hydraulic simulator fails*
  until $\mathcal{H} \neq \emptyset$
  return $\mathcal{H}, \mathfrak{G}$ ▷ *return hydraulics and modified network;*
    ▷ *these will be used by EPANET water quality model*
end function

questions "what?" "where?" "when?" "how?" The first question refers to the water chemistry and reactions, the remaining refer to the injection profile.

Water chemistry uncertainty is limited in this thesis to the bulk reaction coefficient, K_b. The bulk reaction coefficient varies according to the contaminant, with a value of 0 indicating a non-reactive tracer, negative values indicating decay, and positive values indicating organic growth (biologics). Biologic contaminants were not considered. Values for K_b are in units of inverse time for first-order decay.

Complex reactions (such as with chlorine) and wall reaction coefficients were not used. Chemical reaction models require a different version of EPANET, the multispecies EPANET-MSX extension; such reactions are complex and must be written for a specific contaminant – they are not conducive to a more general study. Wall reactions depend on both the contaminant and the state of the pipe surface. There are a number of studies in the literature which demonstrate methods to calibrate wall reaction coefficients for a specific system[22, 3]; in a simulation study, assigning wall coefficients would require arbitrary assignment of pipe qualities, limiting generality. Values for K_b are available for real contaminants, allowing the simulation to sample from a realistic range of values.

The injection profile is composed of the injection strength, source location, start time and duration. Because the reactions are first-order decay or no decay, and are independent of all other parameters, the concentrations at every node and time scale linearly with changes in the initial concentration. This simplifies the model, allowing injection strength to be held constant instead of being set as a model parameter; any evaluation of the effect of initial concentration can be performed through post-model execution scaling.

The first injection profile parameter that must be modified within the model is the injection location, x_{inj}. Injections can take place at any reservoir, tank, or non-zero demand (NZD) junction in the network. A junction with no demand consumes no water, and is considered to be buried and inaccessible as a point of (deliberate) contamination. When considering accidental contamination from pipe breaks or other disasters, this may not be a valid assumption; however, limiting injections to NZD nodes is a reasonable, repeatable and clearly defined way to limit the parameter space. The remaining two parameters are time-based. The first is the injection start time, $t_{0,\text{inj}}$, and the second is the injection duration, Δt_{inj}. The start time is relative to the start of the simulation, and the duration is relative to the start time.

The contamination uncertainty parameters are all variables with a definable range of values of uniform likelihood. The values were selected to provide reasonable coverage, rather than from random distributions, with a factorial design for these four independent parameters. The injection locations were selected randomly from the NZD nodes and tanks one time, and this resulting set of locations used for all simulations. The input parameters and their theoretical ranges are listed in Table 2.1. The algorithm for the water quality parameter modifications is presented in Algorithm 4.

Algorithm 4 Water quality simulation with uncertainty.

function DOQUALITY($\mathfrak{G}, \mathcal{H}, K_{\text{b}}, x_{\text{inj}}, t_{0,\text{inj}}, \Delta t_{\text{inj}}$)
  $\mathfrak{G}.K_{\text{b}} \leftarrow K_{\text{b}}$
  define an EPANET source, $\mathcal{I}$
  clear quality for all nodes
  set $\mathcal{I}$ source type to "MASS"
  set $\mathcal{I}$ source quantity to 1000.0 mg/min
  set $\mathcal{I}$ source node to x_{inj}
  set $\mathcal{I}$ start time to $t_{0,\text{inj}}$
  set $\mathcal{I}$ duration to Δt_{inj}
  EPANET2$\mapsto$RUNQUALITY($\mathfrak{G}, \mathcal{H}, \mathcal{I}$)
  $\mathbf{C}$:= READEPANETBINARY
  return C
end function

2.2 Impact measures

To perform UQ, there must be both input parameters and output measures that can be quantified. Looking first at the output measures, the concentration at each node (and in each pipe) is the output of the EPANET water quality simulation that is of import here. Any impact measure must combine concentration at all locations and times to allow uncertainty to be quantified.

Table 2.1: Input parameters, model method, and ranges of values.

Description	Parameter	Model	Theoretical Range
Demand modification	σ_{d}	Alg. 1	$[0\ ,\ \infty)$
Pipe closures	β_{f}	Alg. 2	$[0\ ,\ 1)$
Bulk reaction	K_{b}	choice	$(-\infty\ ,\ 0]$
Injection location	x_{inj}	choice	$x \in \mathcal{N}$
Injection start	$t_{0,\mathsf{inj}}$	choice	$[0\ ,\ \infty)$
Injection duration	Δt_{inj}	choice	$[0\ ,\ \infty)$

For a specific simulation run, ξ, there are two metrics that will be used to describe the results of a specific simulation: extent of contamination (EC) and population impacted (PI). These are typical impact measures previously used in the literature to evaluate sensor placement (SP) design [28] and disaster effects [19]. The extent of contamination is the total length of the pipes that have been contaminated by a specific event. The population impacted is the total number of people who may have been exposed to contaminated water or who will certainly be affected by remediation efforts, for example, due to pipe replacement. The EC and PI metrics are very useful in sensor placement, as the goal of SP is to minimize impact prior to detection.

The EC and PI metrics are less useful in plume localization; the goal is not to minimize the extent of the plume, it is to find the actual edges of the plume. As noted by Yang and Boccelli [30], changes in flow path can impact not only the arrival time and concentration levels, but can significantly change to location of the plume. To account for this positional uncertainty, two new, but related metrics are proposed here – unknown extent of contamination (uEC) and unknown population impacted (uPI). These metrics will encompass only those parts of the network where the plume extent is uncertain, in other words, the "unknown edge" of the contaminant plume. As this edge is reduced, the actual plume position becomes more clear.

2.2.1 Simulation impact metrics

To calculate the metrics, the concentration matrix, $\mathbf{C}$, must be computed using EPANET. The output file from EPANET does not give the concentration for the pipes, only nodes. Because extent of contamination is a measure of pipe lengths, a heuristic must be used to estimate the actual length contaminated. Other authors have defined EC using the average contamination across the pipe taking flow direction into account. Given the quantity of data already being recorded, flow direction was not saved, and a different method is required. A length, λ, is assigned to each node equal to one half the sum of the length of all adjacent pipes; this uses the conservative assumption that if a node has the contaminant present, at least part of all adjacent pipes will need decontamination.

The plume extent is not the same as the extent of contamination. Plume extent varies with time based on the current position of the contaminant; effectively, pipes are removed from the plume as it passes by. The EC metric, however, is based on the maximum concentration seen at a node across all times. To calculate the maximum concentration seen, use the row selection vector $\mathbf{e}_j$ to choose the appropriate junction and compare the ℓ_∞-norm to a threshold of contamination, τ_c. Any node that is contaminated gains membership in the plume, $\mathfrak{P}$, which is a set of nodes.

$$\mathfrak{P}(\mathbf{C}, \tau_c) = \left\{ j, \forall j \in \mathcal{N} \;\middle|\; \left\| \mathbf{C}\mathbf{e}_j \right\|_\infty \geq \tau_c \right\} \tag{2.1}$$

$$EC(\mathfrak{P}, \tau_c) = \sum_{j \in \mathfrak{P}} \boldsymbol{\lambda}_{\mathsf{idx}(\mathcal{N}, j)} \tag{2.2}$$

The specific results for a simulation are designated as follows. The superscript notation, $v^{(\xi)}$, is used to indicate that variable v belongs to the parameters and results of simulation ξ.

$$\mathring{\mathfrak{P}}_{\tau_c}^{(\xi)} = \mathfrak{P}(\mathbf{C}^{(\xi)}, \tau_c) \tag{2.3}$$

$$\mathring{EC}_{\tau_c}^{(\xi)} = EC(\mathring{\mathfrak{P}}_{\tau_c}^{(\xi)}, \tau_c) \tag{2.4}$$

Similarly, the population impacted, PI, can be calculated by adding all people who live at junctions that have seen contaminated water. Tanks and reservoirs have zero population by definition, so the iteration can be limited to junctions. This is different from the population exposed (PE) health metric. The PE metric takes into account the actual mass of contaminant consumed by people and the toxicity of the contaminant; the PI metric merely looks at the total population that will be subject to any disruption, whether death or merely inconvenience. The population at each junction is contained in the vector $\boldsymbol{\pi}$.

$$PI(\mathfrak{P}, \tau_c) = \sum_{j \in \mathfrak{P}} \boldsymbol{\pi}_{\mathsf{idx}(\mathcal{J}, j)} \tag{2.5}$$

2.2.2 Uncertainty metrics

There is no physical reason these two types of measures would be correlated – people live in a three-dimensional world, best represented as a distribution across a continuous surface. The WDS network has numerous restrictions on it, and is better represented as a graph.

The uEC and uPI metrics can only be calculated when results of multiple simulations are present; let Ξ be the set of simulations to be used. Conceptually, the contaminant plumes that result from each simulation can be laid over one

another to see the most likely areas of contamination. This map can be divided into three parts: the contaminated zone, $\mathcal{Z}_{\mathcal{C}}$, where every node has a very high probability of being contaminated, the uncontaminated zone, $\mathcal{Z}_{\mathcal{U}}$, where every node has a very low probability of being contaminated, and the unknown status zone, $\mathcal{Z}_{\mathcal{S}}$, where the probability of contamination is in the middle. The uEC and uPI metrics describe the size of the unknown status zone.

All of the random variables used for the hydraulics are independent; the simulation results are therefore also independent. The sample probability, $\mathbb{P}$, can be calculated as the number of simulations where a node is contaminated divided by the total number of simulations. This is acceptable for determining the size of the contaminated zone, and the threshold α_c is used to determine a cutoff probability for membership.

$$\mathcal{Z}_{\mathcal{C}}(\Xi, \tau_c, \alpha_c) \;=\; \left\{ j, \forall j \in \mathcal{N} \,\middle|\, \mathbb{P}\left(j \in \mathfrak{P}_{\tau_c}^{(\xi)} \middle| \xi \in \Xi \right) > (1 - \alpha_c) \right\} \tag{2.6}$$

The membership of the uncontaminated zone is comprised of all nodes that are contaminated in zero simulations.

$$\mathcal{Z}_{\mathcal{U}}(\Xi, \tau_c) \;=\; \left\{ j, \forall j \in \mathcal{N} \,\middle|\, \nexists\, \xi \in \Xi \text{ such that } j \in \mathfrak{P}_{\tau_c}^{(\xi)} \right\} \tag{2.7}$$

It is important to note that the probability of contamination for nodes in this zone is always non-zero. An upper bound on the probability of contamination for nodes in this zone can be calculated using a binomial fit with $p \leq 0.5$ and $n = |\Xi|$. Figure 2.1 shows the $\hat{p}$ value for probability of contamination at a node with one simulation showing contamination and the upper value of the 95% confidence interval for the $\hat{p}$ when no simulation shows contamination. The upper bound on the confidence interval for $\hat{p}$ when no simulations show contamination is higher than the estimated $\hat{p}$ with a single simulation showing contamination. By specifying a desired upper bound on the probability of contamination for nodes in the uncontaminated zone, the number of stochastic trials needed can be computed. In this case, a parameter like α_c should be added to $\mathcal{Z}_{\mathcal{U}}$ and the zone membership reformulated in a manner similar to $\mathcal{Z}_{\mathcal{C}}$.

The unknown status zone is comprised of all nodes that are members of neither the contaminated nor uncontaminated zone. While $\mathcal{Z}_{\mathcal{S}}$ is defined as set operation in Equation 2.8, the algorithm used to create the set improves efficiency by comparing the sample probability to zero rather than computing the otherwise unused set $\mathcal{Z}_{\mathcal{U}}$.

$$\mathcal{Z}_{\mathcal{S}}(\Xi, \tau_c, \alpha_c) \;=\; \mathcal{N} - \mathcal{Z}_{\mathcal{C}}(\Xi, \tau_c, \alpha_c) - \mathcal{Z}_{\mathcal{U}}(\Xi, \tau_c) \tag{2.8}$$

Where the EC and PI metrics operate on the set of nodes within a plume, the uEC and uPI metrics perform the same operations on the set of nodes within

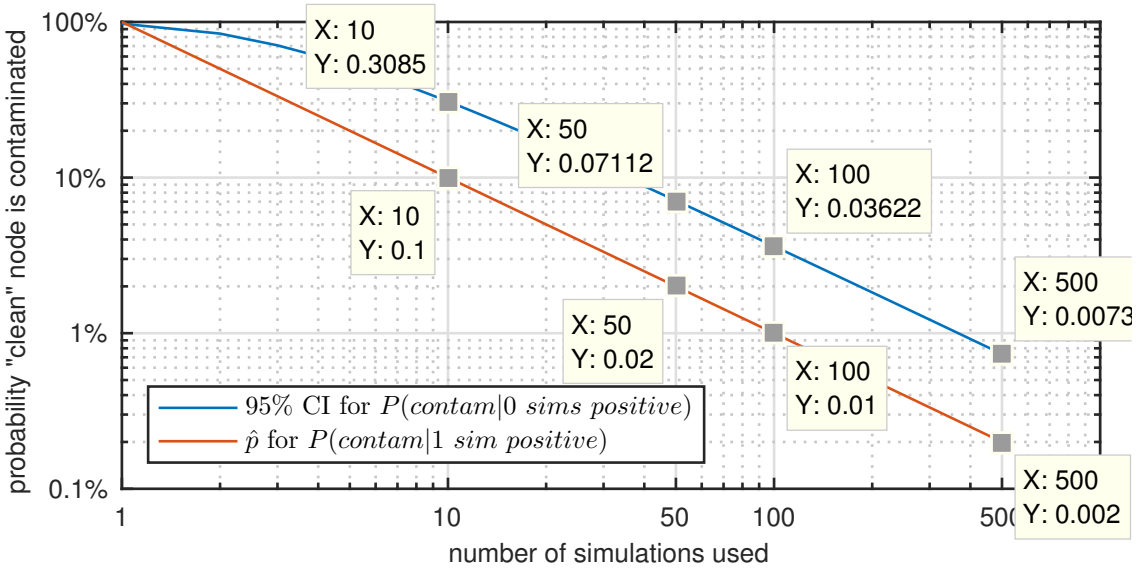

Figure 2.1: Probability of contamination within $\mathcal{Z}_{\mathcal{U}}$.

the unknown status zone defined by a set of simulations, Ξ; the parameters to $\mathcal{Z}_{\mathcal{S}}$ are omitted in the summation for readability.

$$uEC(\mathcal{Z}_{\mathcal{S}}, \tau_c, \alpha_c) = \sum_{j \in \mathcal{Z}_{\mathcal{S}}} \boldsymbol{\lambda}_{\mathsf{idx}(\mathcal{N},j)} \tag{2.9}$$

$$uPI(\mathcal{Z}_{\mathcal{S}}, \tau_c, \alpha_c) = \sum_{j \in \mathcal{Z}_{\mathcal{S}}} \boldsymbol{\pi}_{\mathsf{idx}(\mathcal{N},j)} \tag{2.10}$$

2.3 Simulation studies

Analysis will be completed using a partial-factorial experimental design with stochastic simulation providing the independent trials. This approach was taken after doing a full Monte Carlo analysis was determined to be too expensive (computationally) with the full parameter space. Identification and quantification of the most significant parameters to help limit the parameter space was needed first, as it was discovered that the methods used by others did not work as well applied to this problem.

For example, the UQ studies for water distribution networks, such as those by Yang and Boccelli [30], Blokker et al. [6], Jonkergouw et al. [17], and Pasha and Lansey [25], either are looking at calibration or a specific "injection" events – usually involving chlorine. In both cases, the source location, time, and mass

are known *a priori*, a very different scenario than with an unanticipated contamination incident. While future work can address fully varying input parameter space, it was too soon to do so here. Thus, a full Monte Carlo simulation or other techniques such as latin-hypercube sampling will have to wait. The experimental design method will analyze correlation between different metrics, look for parameters that can be eliminated or reduced in future work, and examine main effect and interaction effects for the different parameters.

Simulations were run on "Net6," a skeletonized network of a real city of around 300-thousand occupants; the network was taken from Watson et al. [28]. The network is shown in Figure 2.2 with the selected injection locations indicated. The network contains one reservoir, at a low elevation, and the majority of water is pumped to tanks from where it is distributed to customers. Water age, a measure of the turnover rate within the system, is generally less than 24 hours, except at the furthest extents of the network. Pipe junctions without an assigned demand are assumed to be buried pipe intersections, not accessible as points of contamination; this is a common assumption (see [20]).

Prior to running the Monte Carlo simulations, test simulations were executed to ensure that parameter values were reasonable. Several parameters that were not studied in the UQ study were explored to ensure the model was configured for sufficient accuracy or to ensure true independence of the parameter. A large number of testing and preparatory runs ($\sim$5 million forward model calls) were completed during this exploration. The parameters chosen for the study runs are presented in Table 2.2.

2.3.1 Physical parameters

There are five physical parameters that were discussed in Chapter 3 – node demand standard deviation, pipe closure fraction, pipe closure limiting diameter, injection strength and bulk reaction coefficient. As discussed previously, the down-stream concentration is directly proportional to the input concentration. Injecting at positive-demand nodes requires using the EPANET "MASS" injection model, which is a constant rate of injection defined in mg/min. A value of 1000.0 mg/min was used as the injection strength for all simulations to ensure concentrations were sufficiently large to avoid mass-balance and tolerance errors within EPANET.

After experimentation and discussion with other researchers, the limiting pipe diameter, δ_d, was fixed at twelve inches. Anecdotal evidence from other researchers is that several percent of the isolation valves in a system may be incorrectly marked in the network model. In the Net6 model, however, it quickly became apparent that randomly selecting pipes to close would isolate parts of the network at target fractions, β_f, of even a tenth of a percent; those that did not isolate pieces of the network would frequently fail due to hydraulic errors, and between these two issues less than one in 1000 produced a topology that would converge.

Limiting the pipes to those with a diameter of less than twelve inches allowed for the increase of the useful target fraction range while simultaneously decreasing the number of random iterates needed to find a valid topology. The parameter for the pipe closure model, β_f, was set to 0, 0.0025, and 0.005, representing 0, 0.25% and 0.5% of pipes less than 12" in diameter.

The demand modification model was artificially constructed. While there are some publications[21, 24] that have examined per-household hourly usage, the Net6 model is a skeletonized model – households are aggregated into a single node. The aggregation of commercial, residential, and other usage makes the likelihood of the aggregated demand being zero very small. The number of times the demand was truncated at zero was thus a useful measure of the realistic range for the σ_d parameter. Experimentation showed that for $\sigma_d = 0.5$ there were generally around 3 000 node-time steps, $\approx 5\%$ of the total node-time step pairs, where demand was truncated for each stochastic simulation, so this was considered the top of the realistic range for this network model. For non-skeletonized models this value could be very different. The values selected were 0, 0.2, and 0.4; this choice allowed for one case ($\sigma_d = 0.2$) where there were generally only one or two time steps that were truncate and one case ($\sigma_d = 0.4$) where the number of truncations occurred much less than 5% of the time.

The final physical parameter was the bulk reaction coefficient. A study by Davis et al. [10] identified bulk reaction coefficients for a dozen contaminants of interest. The decay coefficients ranged from 0/day to as high as 8000/day. Experimentation with different values found that values of 100/day or higher would decay too fast to be of use in simulation. Additionally, it was discovered that the concentration could be calculated analytically based on a node-specific relationship with concentrations from three K_b values. The coefficients for the equation are different for every node and time step, but can be generalized according to the following relationship,

$$c(j,t,K_b) \;=\; c(j,t,K_{b0}) + \frac{c(j,t,K_{b1}) - c(j,t,K_{b2})}{\log|K_{b1}| - \log|K_{b2}|} \log|K_b|, \qquad (2.11)$$

where $K_{b0} = 0.0$ and K_{b1}, K_{b2} are any other two negative bulk reaction parameters. This relationship is only true for chemical parameters and would not be true for biologic parameters (where growth would occur).

While a normal log-linear relationship would only need two values, the round-off errors make the equation more accurate if the intercept and slope are calculated using different data. Values for K_b of 0/day, -0.1/day, and -5.0/day were chosen. This brackets the values for the most common contaminants and provides the three points necessary to vary decay coefficients after the fact.

2.3.2 Temporal parameters

The two temporal parameters for the UQ are the injection start time and the injection duration. There are other temporal parameters in the simulation:

the pattern step size, Δt_{pat}, impacts the number of demand modifications needed; the report step, Δt_{rpt}, needs to be smaller than the patter step, but not so small that the results become too large to process; the simulator step sizes, Δt_{hyd} and Δt_{wq}, affect the accuracy of the concentrations. Experimenting with the Δt_{hyd} and Δt_{wq} parameters showed that while the accuracy of concentrations increased slightly with decreasing step sizes, there was a limit on the improvements due to EPANET's internal tolerances and memory constraints. Additionally, decreasing step sizes for the solvers did not impact the EC or uEC metrics unless the report size was significantly decreased; because the report size was set to fifteen minutes to keep storage reasonable, keeping the solver step size in the minute range instead of the seconds range was appropriate. The Δt_{hyd} was set to 5 minutes and the Δt_{wq} was set to 1 minute. The patterns defined by the Net6 network input file are one-hour patterns, so Δt_{pat} was kept at 1 hour.

The injection start time was limited to three values, in keeping with the number of values that were chosen for the physical parameters and because the factorial combination of values was increasing steadily. Start times of hour 0, 8 and 16 within the simulation were selected to provide good coverage of different demand types (representing 8am, 4pm and 10pm simulated clock times within the demand patterns). The injection durations were set to 1 hr, 12 hrs, and continuous through the end of the simulation (24 to 40 hours depending on start time). The continuous injection and one-hour injection are frequently used scenarios in water quality research. The 12-hour injection provided a midpoint.

2.3.3 Positional parameter

The only positional parameter is also a discrete parameter – the injection location. Because there are over 1 600 NZD nodes in the system, plus tanks and the reservoir, time and size constraints were not conducive to running all possible locations at this time. The factorial combination of the physical and temporal parameters already provided 243 combinations – not including the number of Monte Carlo runs. A subset of locations was chosen, instead.

A reservoir is always a likely target or source of contamination, so the reservoir was chosen to be part of the subset. Ninety nine additional nodes were chosen at random from the tanks and NZD nodes to be injection locations. The random selection was then examined on a map to ensure that coverage of the network looked appropriate. The chosen locations are presented in Figure 2.2.

One note should be made regarding the EPANET model and injections at tanks. The model does not contaminate the tank directly; instead, an imaginary point just outside the tank is used for injection, and contamination only flows away from the tank, not into it. This is not true for contaminated water flowing into the tank – such water does mix with water in the tank – but MASS injections at the tank while it is filling contaminate a single packet of water that stays outside the tank until it starts draining, when it then moves downstream. This

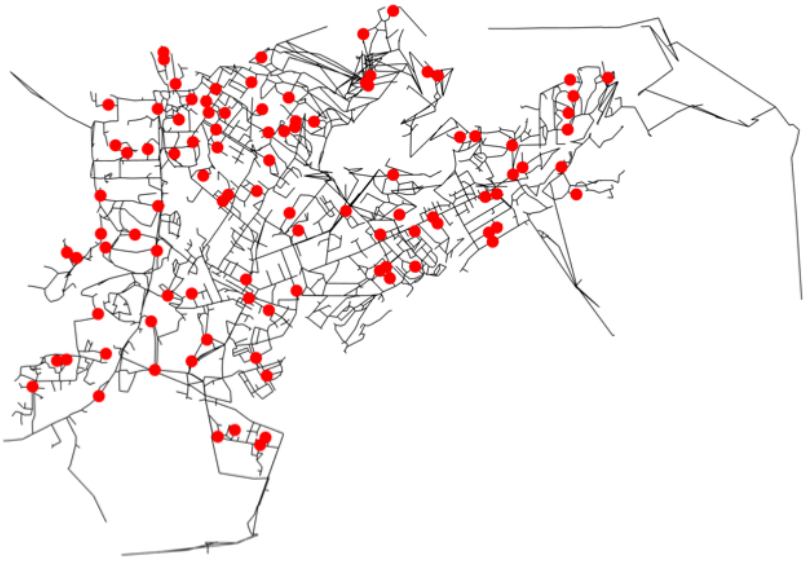

Figure 2.2: The network used for the simulation study. Injection locations are highlighted.

is an unfortunate feature of EPANET, but for repeatability by others it was not corrected in the software library that was used.

2.3.4 Threshold parameters

The output metrics require threshold parameters. The threshold for concentration, τ_c, was set to 10^{-6} mg/L. This was near the numerical limit for non-zero concentrations, due to EPANET's tolerances, and was three orders of magnitude smaller than the lowest maximum concentration for a single simulation; constant mass injections result in smaller concentrations based on the volume of water flowing past the node where the injection takes place. The probability threshold, α_c, was set to 0.05.

2.4 Software and implementation

The methods described were implemented in a Python library which makes use of the Water Network Tool for Resilience (WNTR)[19] Python library and the EPANET 2.0 toolkit. The WNTR library was used to parse the EPANET input file and create an object with the network and options. The WNTR EPANET simulator (which calls the EPANET library) was used as a starting point, but the simulator was rewritten completely by the end of the project. The EPANET toolkit

Table 2.2: Parameter values chosen for the simulation study. Values marked by $\star$ indicate the range is specifically for the Net6 model.

Description	Parameter	Realistic Physical Range	Chosen Values
Demand modification	σ_{d}	$0 \leq \sigma_{\mathsf{d}} < 0.5^{\star}$	0, 0.2, 0.4
Pipe closures	β_{f}	$0 \leq \beta_{\mathsf{f}} \lll 1$	0, 0.0025, 0.005
Pipe closure diameter	δ_{d}	$(0, \max(\mathbf{d})]$	12 in. (*fixed**)
Bulk reaction	K_{b}	$[-5/\Delta t_{rpt}\ ,\ 0]$	0, -0.1, -5.0 day^{-1}
Injection location	x_{inj}	x in NZD nodes	100 nodes
Injection start	$t_{0,\mathsf{inj}}$	$[0\ ,\ t_{\mathsf{total}})$	0, 8, 16 hrs
Injection duration	Δt_{inj}	$(0\ ,\ t_{\mathsf{total}} - t_{0,\mathsf{inj}})$	1 hr, 12 hrs, continuous
Concentration threshold	τ_{c}	$[0, \max(\mathbf{C})]$	1×10^{-6}
Probability threshold	α_{c}	$[0, 0.1]$	0.05

allows for programmatic access to the EPANET hydraulic and water quality simulator, ensuring consistency with other research efforts in the field.

The software was designed to use an input file that specifies the parameter values or ranges, the number of Monte Carlo iterates, and parallelization options. The main program parses these options and executes simulations in a subprocess or using a message passing interface (MPI) library. The outline of the program is presented in Algorithm 5.

Once all simulations have been completed, and the results saved, the uncertainty metrics can be calculated. The maximum concentration seen at each node was output in a matrix, one row per simulation. This data was processed in MATLAB to collect data into scenario sets, Ξ, and then calculate the uEC and uPI metrics. This process was coded into Python, but the visualization functions were written in MATLAB and required the same basic information, so the metrics were calculated in MATLAB directly.

The modified demands and pipe statuses are stored with the result concentration values in a compressed binary file. The file format is Hierarchical Data Format version 5 (HDF5). The HDF5 format was developed for storing extremely large datasets for astrophysics and climate modeling, but is convenient due to its portability (binary compatibility between platforms is not common), the abundance of libraries for different programming languages, and its efficiency. The configuration file for the UQ main program is written in JavaScript Object Notation (JSON), a hierarchical text-formatted notation that is widely used and which has libraries for most languages. It is also more human-readable than most other configuration file formats.

The Python library runs approximately 4 500 lines of original code; in addition, debugging and bug fixes were applied to the WNTR library.

Algorithm 5 Main program

procedure MAIN_MPI
  if primary processor **then** ▷ *MPI designates a master processor*
    read config file and load parameter settings
    calculate number of Monte Carlo iterates per processor, n_{pp}
    start MPI processes ▷ *pass information the master read to child processes*
  end if
  $\mathfrak{G} :=$ WNTR$\mapsto$WATERNETWORKMODEL($filename$) ▷ *parse input file*
  for $i = 0, 1, \ldots, n_{pp} - 1$ **do**
    ▷ *Next, perform a factorial combination of parameter values from config file*
    for all $\sigma_{\mathsf{d}}, \beta_{\mathsf{f}}, K_{\mathsf{b}}, x_{\mathsf{inj}}, t_{0,\mathsf{inj}}, \Delta t_{\mathsf{inj}}$ in UQ ranges **do**
      start subprocess ▷ *necessary to ensure Python/DLL memory clears*
      $\xi_{i,cpu} :=$ DOSIMULATION($\mathfrak{G}, \sigma_{\mathsf{d}}, \beta_{\mathsf{f}}, K_{\mathsf{b}}, x_{\mathsf{inj}}, t_{0,\mathsf{inj}}, \Delta t_{\mathsf{inj}}, \tau_{\mathsf{c}}$)
    end for
  end for
  save all $\xi_{\cdot,cpu}$ results
end procedure

function DOSIMULATION($\mathfrak{G}, \sigma_{\mathsf{d}}, \beta_{\mathsf{f}}, K_{\mathsf{b}}, x_{\mathsf{inj}}, t_{0,\mathsf{inj}}, \Delta t_{\mathsf{inj}}, \tau_{\mathsf{c}}$)
  $\mathcal{H}, \mathfrak{G}^{\star} :=$ DOHYDRAULICS($\mathfrak{G}, \sigma_{\mathsf{d}}, \beta_{\mathsf{f}}$)
  $\mathbf{C} :=$ DOQUALITY($\mathfrak{G}^{\star}, \mathcal{H}, K_{\mathsf{b}}, x_{\mathsf{inj}}, t_{0,\mathsf{inj}}, \Delta t_{\mathsf{inj}}$)
  $\mathring{\mathfrak{P}} := \mathfrak{P}(\mathbf{C}, \tau_{\mathsf{c}})$
  $\mathring{EC} := EC(\mathring{\mathfrak{P}}, \tau_{\mathsf{c}})$
  $\mathring{PI} := PI(\mathring{\mathfrak{P}}, \tau_{\mathsf{c}})$
  ξ.params $\leftarrow \langle \mathfrak{G}^{\star}, \sigma_{\mathsf{d}}, \beta_{\mathsf{f}}, K_{\mathsf{b}} x_{\mathsf{inj}}, t_{0,\mathsf{inj}}, \Delta t_{\mathsf{inj}} \rangle$ ▷ *save modified network with param's*
  ξ.results $\leftarrow \langle \tau_{\mathsf{c}}, \mathring{\mathfrak{P}}, \mathring{EC}, \mathring{PI}, \mathbf{C} \rangle$ ▷ *save concentrations and initial metric calculations*
  return ξ
end function

CHAPTER 3

RESULTS AND DISCUSSION

In this chapter I will present the results of the analyses. First, an analysis of the performance of the new impact measures, uEC and uPI, will be presented. Following this, the results from the large-scale simulation study described in Section 2.3 will be presented along with discussion.

3.1 Analysis of impact measures

As discussed in Chapter 2, there are two pairs of impact measures that were used, one based on the length of pipe contaminated and one based on the population impacted. Because of the number of simulations and the size of the results, it would be convenient if the majority of the analysis could be performed using only one of the pairs; this requires a very high degree of correlation between the two measures and is contingent on the network configuration. To evaluate if this simplification is possible on the network used, the correlation between all pairs of metrics was examined.

Figure 3.1 shows the empirical cumulative distribution function (eCDF) for each of the four metrics. The correlation coefficient, R, between the direct contamination metrics, EC and PI, and the uncertainty metrics, uEC and uPI, are both 0.998, as shown in Table 3.1. The correlation coefficient values of $R \approx 0.6$ show a not-insignificant correlation between the direct and uncertainty metrics, a relationship that also seems consistent with the eCDFs in Figure 3.1. The values EC and uEC were plotted against each other in log-log space, and the result is shown in Figure 3.2. The plot shows that there is some correlation between EC and uEC, but this correlation does not support a direct transformation between the direct and uncertainty metrics.

The network used, Net6, shows sufficient correlation between the pipe length and population metrics that simplification to a single metric type is reasonable. The remainder of the results will show only the EC and uEC results, but it should be remembered that on a different network the same comparison of metrics would be necessary and may indicate that the metrics will need to be analyzed separately.

Table 3.1: Correlation coefficient R for the four impact metrics; $\text{std}(R) < 10^{-4}$ for all cases where the stochastic simulations could be used to calculate R for each of 50 samples.

R	PI	uEC	uPI
EC	0.998	0.580	0.592
PI		0.572	0.586
uEC			0.998

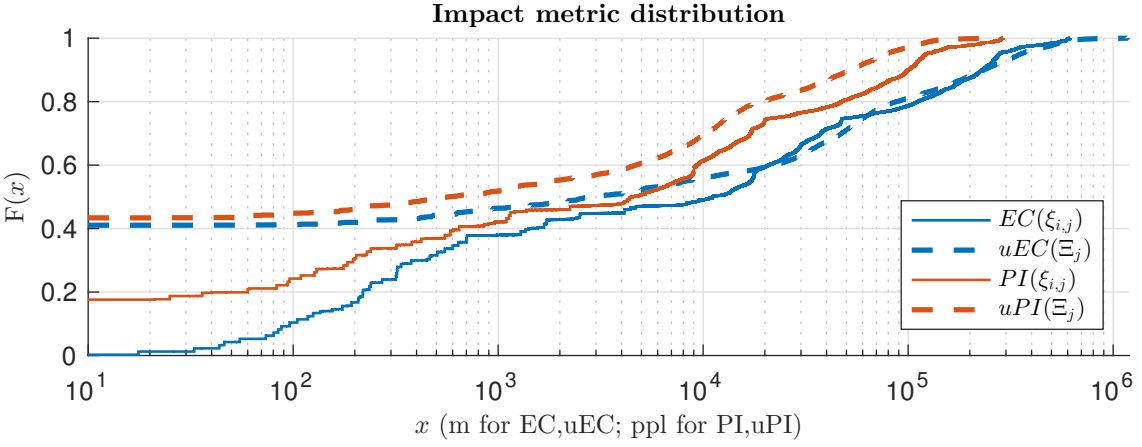

Figure 3.1: Empirical cumulative distribution function (eCDF) for the EC and PI of $\xi_{i,j}$; the standard deviation in EC and PI across the stochastic simulations in a set, Ξ_i; and the uEC and uPI for each Ξ_i. The impact measure axis is truncated for readability, but no information is lost – all metrics have a single, small x value where the eCDF jumps from 0 to the same F(x) value evaluated at $x = 10$

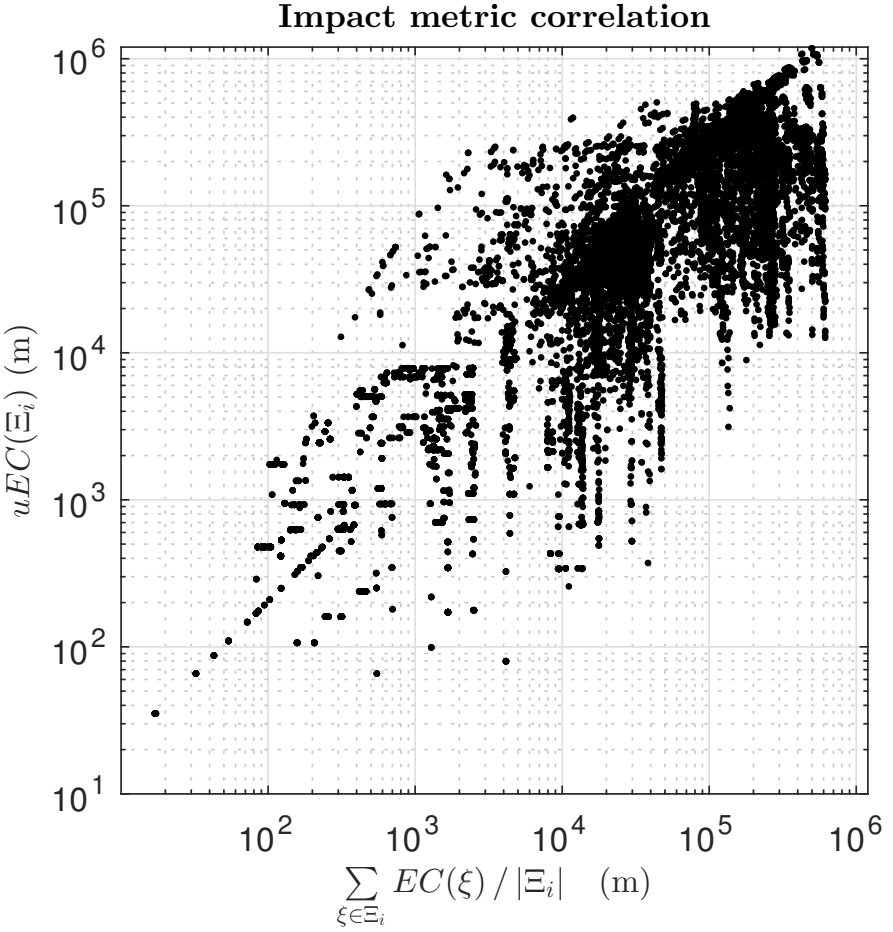

Figure 3.2: Correlation between the EC and uEC metrics as a function of average EC for all members of a simulation set, Ξ, and uEC for that set.

Table 3.2: Parameters used for uEC vs M calculations.

Symbol	σ_{d}	β_{f}	K_{b}	$t_{0,\mathsf{inj}}$	Δt_{inj}	x_{inj}
R	0.2	0.0025	0/day	0	1 h	Reservoir-3323
T	0.4	0.0025	0/day	16	end of sim.	Tank-3331
J	0.4	0.0	0/day	0	1 h	Junction-1451

3.2 Sensitivity to number of simulations

The uncertainty metrics are based on probabilities and take the threshold parameter, α_{c}. Because of this, the number of samples will play a role in the results. A small study was performed to examine this relationship in more depth. Five hundred stochastic simulations were run for specific injection nodes (Reservoir-3323, Tank-3331, and Junction-1451) with the parameter values shown in Table 3.2. The mean value for the uEC metric was calculated using a block-averaging method,

$$\overline{uEC}(M) \;=\; \sum_{n=0}^{M-1} \sum_{j=0}^{M-n-1} \frac{uEC(\Xi_{n:n+j}(\alpha_{\mathsf{c}}))}{M-n}, \tag{3.1}$$

where $M = 500$ is the number of simulations and $\Xi_{i:j}$ represents the integer numbered simulations i through j. The results were calculated using three different threshold values, $\alpha_{\mathsf{c}} = 0, 0.01, 0.05$, with the physical meaning that a node is assigned to $\mathcal{Z}_{\mathcal{C}}$ if it is contaminated in 100%, 99%, or 95% of simulations, respectively. Figure 3.3 shows results of $\overline{uEC}(M)$ for the three nodes as a function of M.

When $\alpha_{\mathsf{c}} = 0$, the value for uEC increases monotonically; changes occur when a new node is contaminated for the first time and when a node is uncontaminated for the first time. In both cases, the node moves from its original zone into $\mathcal{Z}_{\mathcal{S}}$, increasing the uEC. When $\alpha_{\mathsf{c}} > 0$, the number of permitted "non-contaminated" simulations a node can have while staying in $\mathcal{Z}_{\mathcal{C}}$, k, changes as the rounding function $k = \lfloor \alpha_{\mathsf{c}} n \rfloor$. The uEC as a function of n follows the unthresholded curve at first, but as k increases, the scenarios move back from $\mathcal{Z}_{\mathcal{S}}$ into $\mathcal{Z}_{\mathcal{C}}$, and a sawtooth pattern appears.

Though the uEC for sample tank and junction appear to approach some maximum when $\alpha_{\mathsf{c}} = 0$, the only bound on the limit for uEC is the total size of the pressure zone – an area where the boundary conditions prevent exit. Such a situation would represent the case when $\mathcal{Z}_{\mathcal{C}} = \mathcal{Z}_{\mathcal{U}} = \emptyset$. For values of $\alpha_{\mathsf{c}} > 0$, this is no longer true. By lowering k, there will always be some non-zero proportion of the nodes in the network that will meet the criteria for membership in $\mathcal{Z}_{\mathcal{C}}$. As the number of simulations increases, the value for uEC will approach this value. The largest $\overline{uEC}(M)$ occurs when $M \lessapprox {\alpha_{\mathsf{c}}}^{-1}$.

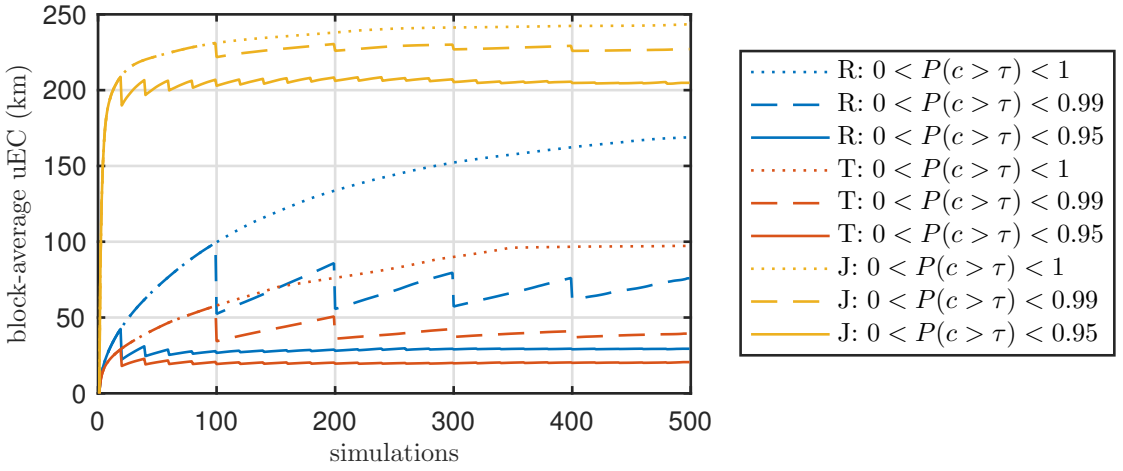

Figure 3.3: Change in uEC based on the threshold, α_c, and the number of stochastic simulations, M. The sawtooth lines are due to junctions with $M - k$ contaminated simulations moving from $\mathcal{Z}_S$ to $\mathcal{Z}_C$ as $(M - k)/M$ becomes larger than α_c. R is the reservoir, T is Tank-3331, and J is Junction-1451.

3.3 Simulation study results

The total number of parameter values for a factorial experiment design was 24 300, and each of the hydraulic parameter combinations was evaluated with 50 stochastic hydraulic simulations; the only exception was the combination of $(\sigma_d = 0, \beta_f = 0)$, where stochastic simulation would produce identical models every time. The result was 401 successful hydraulic simulations and 1 082 700 water quality simulations. For the large values of β_f, it took up to twenty hydraulic simulation failures (with the typical value between three and eight) to find a valid topology. Simulation took approximately fifty CPU-days (an average of 4 seconds per water quality (WQ) simulation).

For each of the factorial parameter combinations $\langle \sigma_d, \beta_f, K_b, t_{0,inj}, \Delta t_{inj}, x_{inj} \rangle_i$, $i = 1, 2, \ldots, 24\,300$, the simulation set $\Xi_i = \{\xi_{i,j} | j = 1, 2, \ldots, 50\}$. The EC and PI for each simulation were calculated and are noted as $EC_{i,j}$ and $PI_{i,j}$, respectively. The uEC and uPI metrics were calculated for each Ξ_i, and are referenced as uEC_i and uPI_i. Calculations are also performed on sets of scenarios that have a single parameter value in common; these sets are designated using the notation $\Xi\langle p = v, \ldots \rangle$, where p is a parameter and v is the associated value; any unspecified parameter varies across all values. For example, the set of simulations where demand and closure model parameters are zero is written $\Xi\langle \sigma_d = 0, \beta_f = 0 \rangle$. Likewise, examples of vectors of results are $\mathbf{PI}_i$, $\mathbf{EC}\langle p = v \rangle$, and $\mathbf{uEC}\langle p = v \rangle$.

The distribution of values for EC and uEC were presented in Figure 3.1. The distribution of the EC values for each of the parameter values taken individually is presented in Figure 3.4. For the hydraulic and WQ parameters – σ_d, β_f,

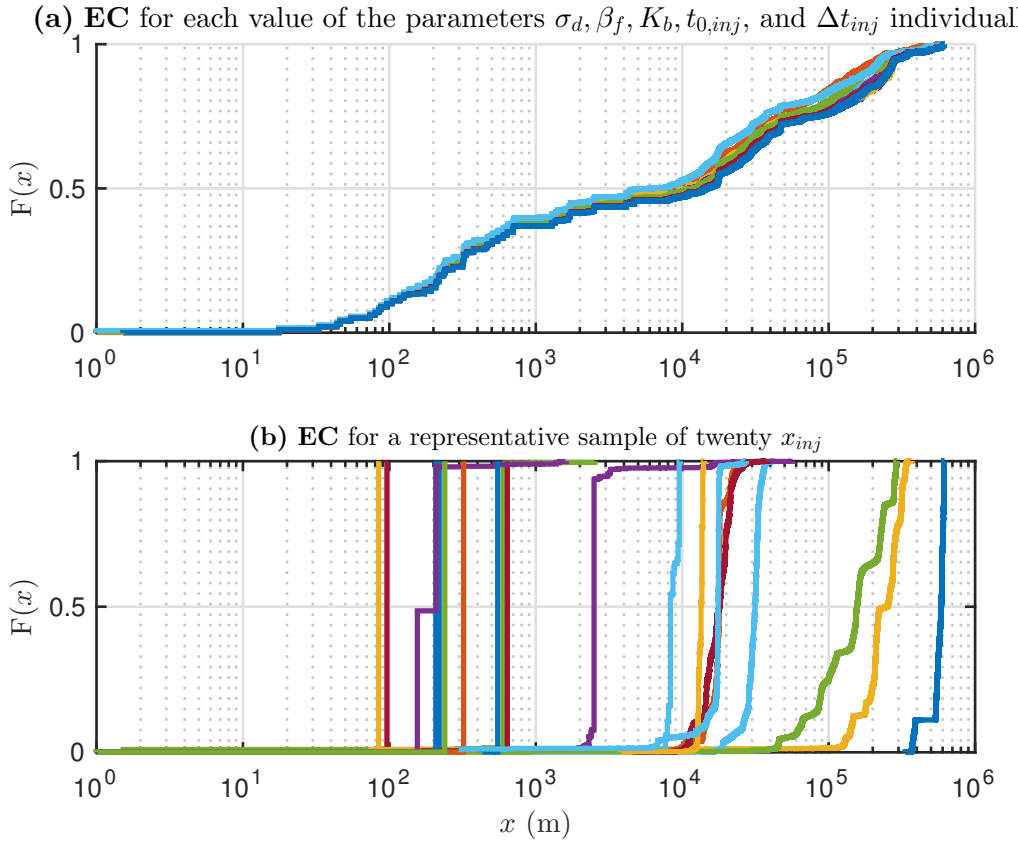

Figure 3.4: (a) eCDF of EC for each of the five hydraulic and WQ parameters set to a specific value while all other parameters are allowed to vary. (b) eCDF of EC for each injection location, where all other parameters fully vary.

K_b, $t_{0,inj}$, and Δt_{inj}– there is a change in the EC distribution at large EC values, but the shape of the distributions are very similar to the overall distribution. For the $t_{0,inj}$ parameter, the shape and magnitude of the distributions change significantly. The location has such an overwhelming effect that performing uncertainty quantification on the plume extent if source location is unknown becomes uninformative; source localization must be the first step in reducing uncertainty.

3.3.1 Main effects by location

The results of the simulations are presented in this section as plots structured in the following manner:

- each row of plots corresponds to an impact metric, EC on top, uEC on the bottom,
- each column of plots corresponds to an input parameter,
- for each value of the parameter, a box-and-whisker plot is drawn, with a target, $\odot$, representing the median, notches indicating a confidence interval around the median, and whiskers of length 1.05 times the full height of the lower and upper quartile boxes,
- a red line connecting the mean impact metric value for each of the parameter's values; this mimics a traditional main-effects plot.

The number of samples per location parameter value is equal to $\approx 4\,000$ for EC and ≈ 80 for uEC– the number for $\langle \sigma_d = 0 \rangle$ and $\langle \beta_f = 0 \rangle$ being slightly less than the rest – and 100 times larger in Figure 3.5, where all locations are combined.

In Figure 3.4 the distribution of EC was presented by parameter in the form of an eCDF. Figure 3.5 shows the distributions for both EC and uEC for the simulations with all locations combined. The maximum and minimum y-axis range is set to the full extent of the mean values if they were plotted by location. This shows once again that the injection location dominates all other parameters. The reaction coefficient, K_b, is the most significant non-location parameter influencing EC, and pipe closure, β_f, is the most significant influencing uEC. This provides a good sanity check on the results, as faster decay should decrease extent of contamination and topology changes should increase uncertainty. Further analysis requires looking at each injection location individually.

Analysis was conducted on all hundred locations; only the results from a representative sample of the locations will be presented. One of the critical points is the entry point of water into the system, in this case at the reservoir. The main effects plot is presented in Figure 3.6. Consistent with the preliminary network-wide results, the K_b parameter show the greatest impact on the EC metric, with a change of over 100 km of pipe in the average extent of contamination between $K_b = 0/\text{day}$ and $K_b = -5/\text{day}$. Likewise, the β_f parameter is clearly the most important parameter affecting uEC, and the distribution of uEC values is much tighter around the mean for β_f than it is for the other parameters. The K_b, $t_{0,inj}$, and Δt_{inj} parameters have the opposite influence on uEC that they have on EC.

The average uEC from the reservoir is much smaller than the average EC. This is not true everywhere. Figure 3.7 shows the results for Junction 13, located only a few pipe links away from the reservoir. It has a similar maximum EC of nearly the entire network. However, its position in the network gives it a maximum uECthat is far larger than for the reservoir, with the $\mathcal{Z}_S$ for some simulations encompassing nearly the entire network. Figure 3.8 shows the extent of the possible contamination scenarios for injections at these two nodes. For contrast, the zones for injections from Junction 1451 (used in Figure 3.3) and Junction 1718 (discussed following) are shown in Figure

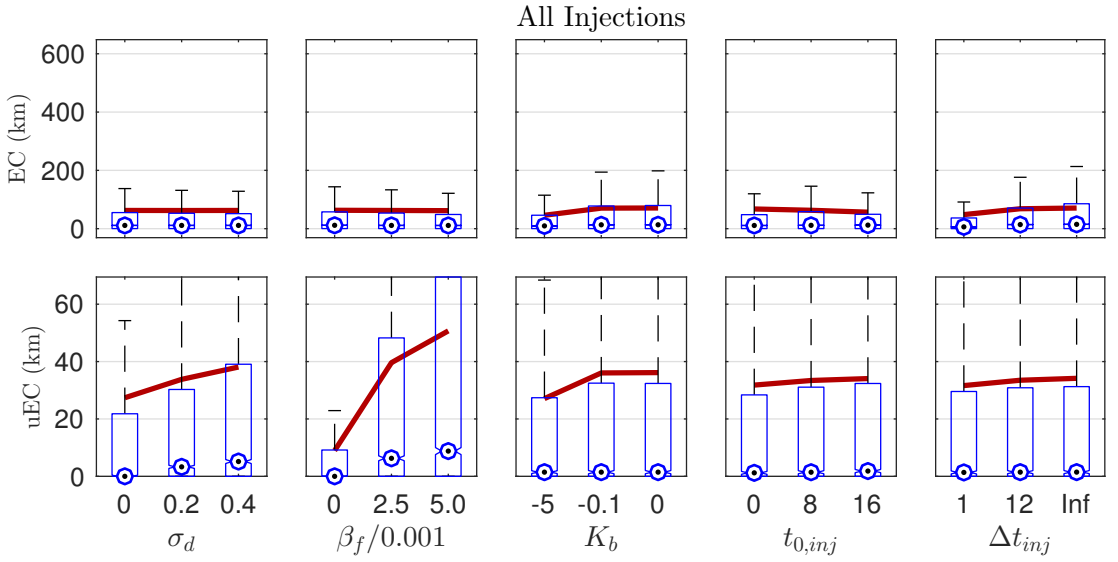

Figure 3.5: Combined box-and-whisker and main effects plot for all locations analyzed jointly.

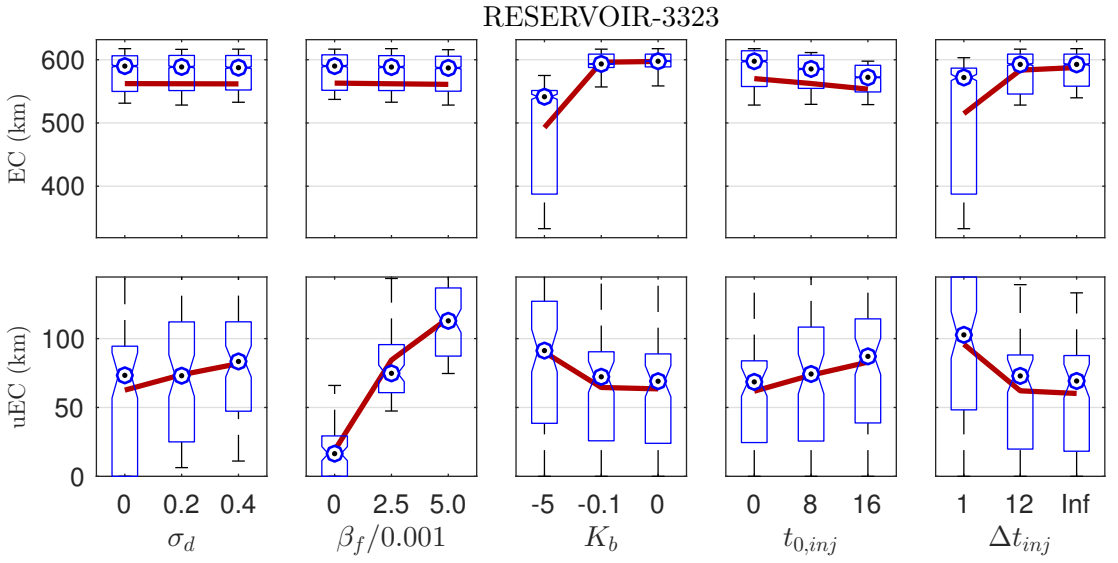

Figure 3.6: Combined box-and-whisker and main effects plot for injections at the reservoir.

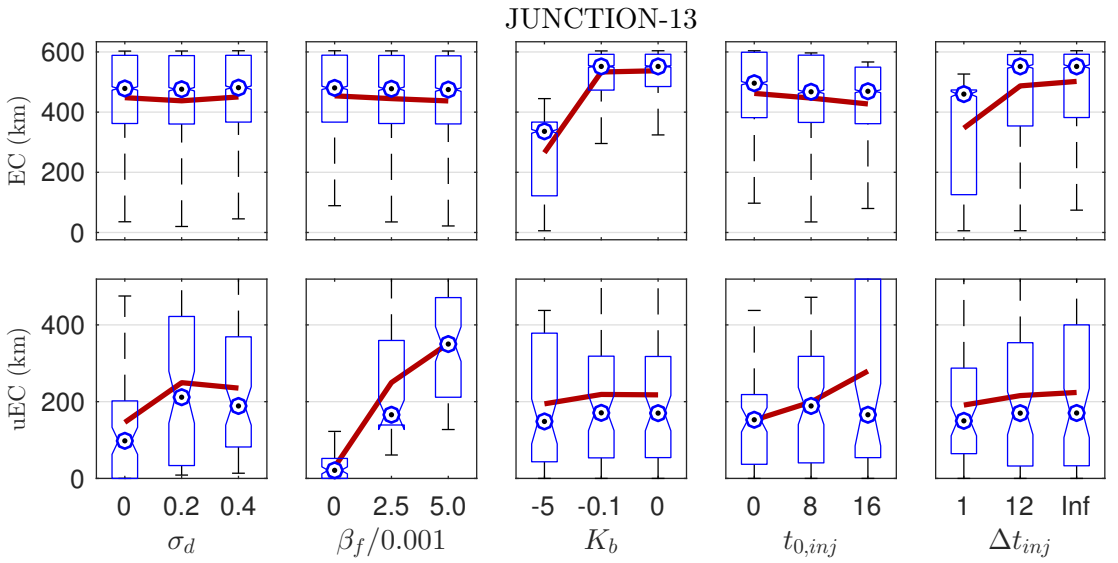

Figure 3.7: Combined box-and-whisker and main effects plot for injections at Junction 13.

3.3.2 Parameter interaction

The two-way interaction effects are, on average, small compared to the main effects. More than the main effects, the interaction effects are highly location dependent. The total EC for reservoir has an interaction effect for the $K_b \times \Delta t_{inj}$ that is as large as the Δt_{inj} effect alone; there are no significant interactions which affect uEC at the reservoir. By contrast, a few junctions, such as Junction 1718, have no significant two-way interactions for EC, but show strong interaction between $t_{0,inj} \times \Delta t_{inj}$ and $\sigma_d \times \beta_f$ for uEC.

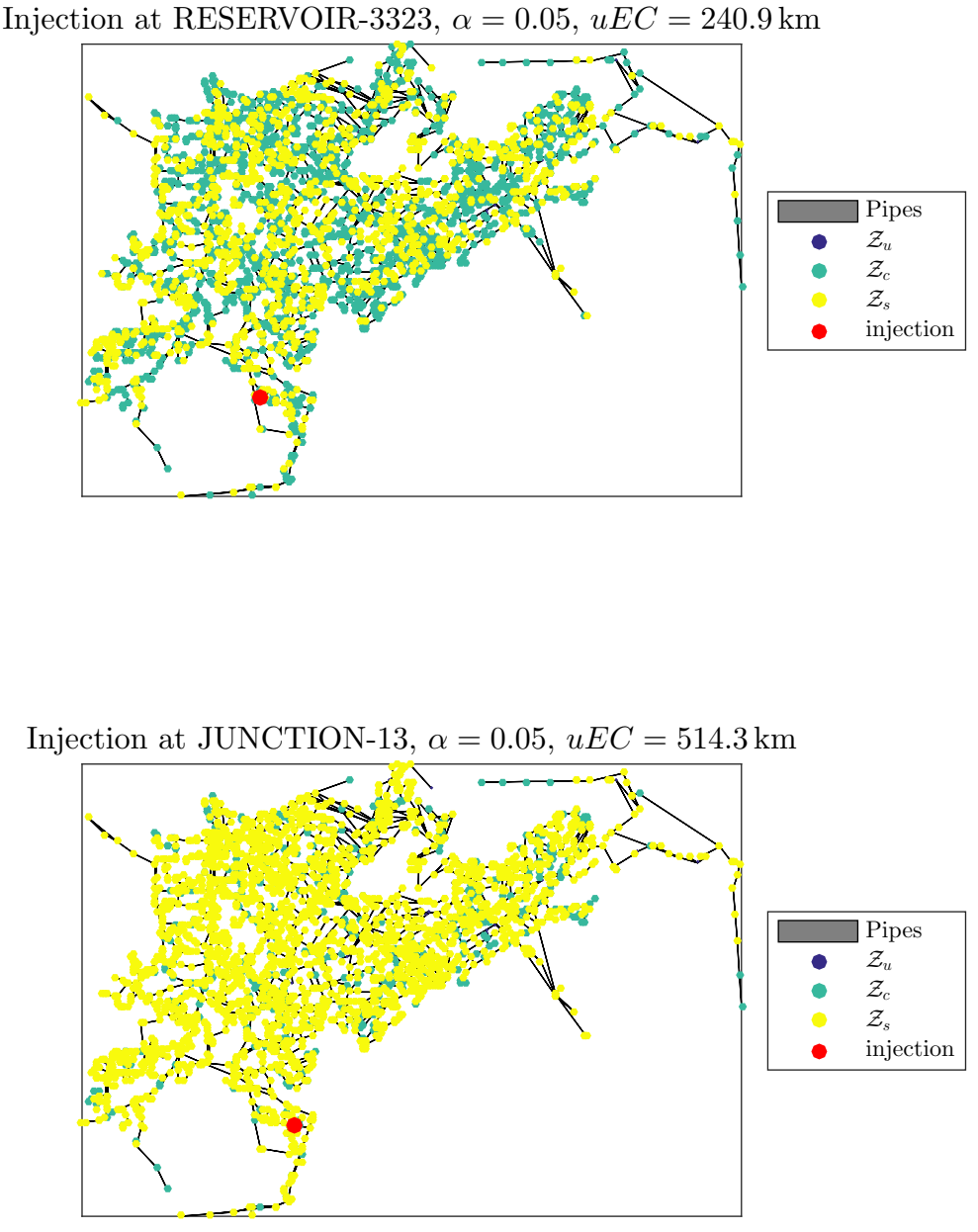

Figure 3.8: Uncontaminated, contaminated, and unknown status zones for injections from the reservoir and from junction 13.

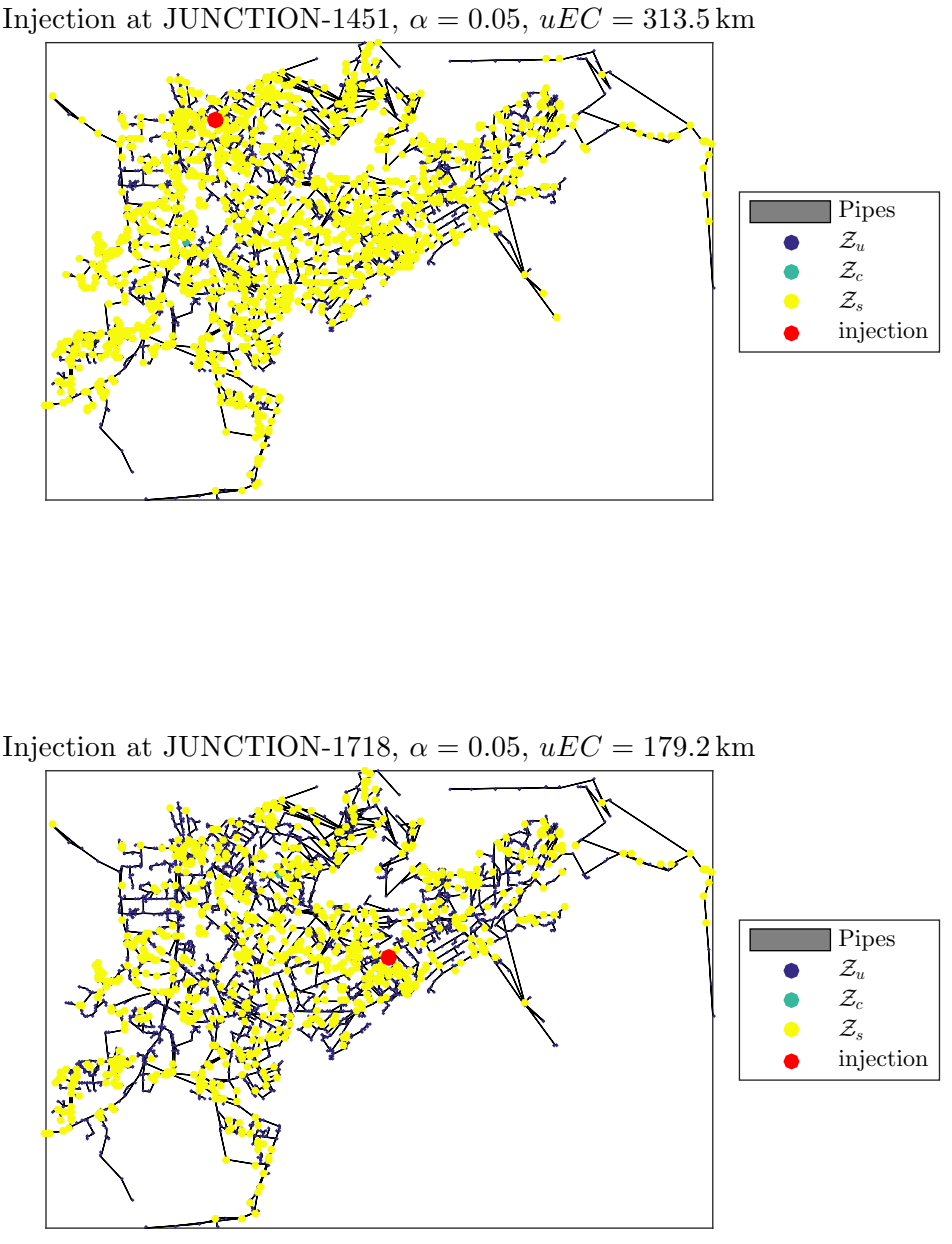

Figure 3.9: Uncontaminated, contaminated, and unknown status zones for injections from the junction 1451 and junction 1718.

CHAPTER 4

CONCLUSIONS

4.1 Conclusions

Two new contamination incident impact metrics were developed that were both shown to effectively describe the uncertainty in plume position. These metrics, uEC and uPI, can be used in minimization routines, as they measure the size of the unknown status zone, where other metrics have a bias towards either underestimating or overestimating contamination impacts, as they would minimize the size of the contaminated or uncontaminated zone. A software tool to perform an analysis of the uncertainty in contamination impact was developed. The tool is modular and has the ability to plug in different parameter models, such as an atypical demand model, to perform UQ.

The location of the source of contamination overwhelms all other parameters when trying to quantify uncertainty. The sensitivity of contaminant plume to other parameters is location dependent, making network-wide quantification of uncertainty misleading. Qualitative analysis of the most important remaining factors is possible, however, and the results provide insight into the effects of uncertainty on contamination incidents after location uncertainty is removed.

Demand variation did not play as large a role in the uncertainty of the plume extent as anticipated, but due to the assumptions and limitations of the demand model used in the simulations no generalizations can be made from this result. Demands were simulated as independent identically distributed Gaussian variation from the original demand; spatial correlation of demands is a very important aspect that was not explored in this work.

For total EC, regardless of injection location, the reaction coefficient dominates the extent of the network that will be effected. For long-duration injections where usage patterns begin to repeat, the injection start time loses its importance in evaluating total impact. When evaluating the uncertainty in plume position, uEC, network topology takes over as the dominating parameter. Start time and duration still play a role, but the reaction coefficient becomes less important.

4.2 Recommendations for future work

4.2.1 Atypical demands and demand correlation

The development of realistic and justifiable models for abnormal demand patterns is still needed. Once these are developed, UQ for plume extent and position need to be reexamined. Parameters that may need examination include the speed of change to abnormal demand patterns and the incorporation of real-time hydraulics. Real-world examples of changes in the system demand due to "do-not-use" and "boil-water" public safety alerts would be helpful in directing this work, all of which would involve water utility cooperation and involvement.

Atypical demands could easily break the EPANET hydraulic simulator. Analysis will be needed to determine if there is a level of skeletonization that is needed to ensure numerical stability, and what effect this would have on demand uncertainty models.

The impact from spatially correlated demands needs to be explored. In many ways, such work would directly lead into atypical demands, as do-not-use orders apply to spatially continuous regions. It is highly likely that correlated demands could be as important, or more important, than topology changes. Analysis from real-world demand records with their associated location is necessary to develop good correlation models for use with simulations.

4.2.2 Contamination response tools

There are several major takeaways from this UQ work that directly apply to contamination response. The first is that source localization is critical – all other factors are eclipsed by the need to find the right location. Secondly, performing sufficient Monte Carlo simulations of possible topology scenarios, and to a lesser extent demand changes, is important when trying to decrease uncertainty in the plume position within the network. Third, contamination start time and duration play a large role in both extent of contamination and plume path, but simulations can be split into sequential, non-overlapping short duration simulations. Finally, starting concentration and bulk reaction coefficient uncertainty can be added in post Monte Carlo simulation, which would then decrease computational costs.

All these elements influence how a sampling algorithm should be developed. Efficient sampling methods should decrease plume uncertainty while also locating the source and identifying the start time and duration. The bulk reaction coefficient should be assignable at the point the contaminant is identified to update the sampling process. Bayesian methods used to update the probability of a specific scenario – i.e., location, start time, duration, and plume location – should take into account the bounds developed on probability of contamination for simulated non-contaminated nodes when evaluating samples taken in the field.

Assuming normal demands for system state prior to public announcement or do-not-use orders, the Monte Carlo runs can be performed off-line prior to any incident, then kept in reserve. Once the scenario is sufficiently identified, using real-time simulations with real-time hydraulics should be investigated, but the uncertainties in topology would not be eliminated through use of real-time hydraulics, even if such measurements truly existed for every service node in the network; therefore, Monte Carlo topologies would still need be performed in real time to ensure the best plume localization.

REFERENCES

[1] *Distribution System Requirements for Fire Protection*. Number M-31 in AWWA Standards. American Water Works Association, Denver, CO, 2008. ISBN 978-1-58321-580-7.

[2] Events That Led to Flints Water Crisis. *The New York Times*, Jan. 2016. URL `http://www.nytimes.com/interactive/2016/01/21/us/flint-lead-water-timeline.html?_r=0`.

[3] A. O. Al-Jasser. Chlorine decay in drinking-water transmission and distribution systems: pipe service age effect. *Water Research*, 41:387–96, Jan. 2007. doi:10.1016/j.watres.2006.08.032.

[4] J. Berry, E. Boman, L. A. Riesen, W. E. Hart, C. A. Phillips, and J. P. Watson. User's Manual TEVA-SPOT Toolkit 2.5.2. Technical Report EPA 600/R-08/041B, U.S. Environmental Protection Agency, Cincinnati OH, Sept. 2012.

[5] E. J. M. Blokker, H. Beverloo, A. J. Vogelaar, J. H. G. Vreeburg, and J. C. van Dijk. A bottom-up approach of stochastic demand allocation in a hydraulic network model: a sensitivity study of model parameters. *Journal of Hydroinformatics*, 13:714–728, 2011. doi:10.2166/hydro.2011.067.

[6] M. Blokker, J. Vreeburg, and V. Speight. Residual chlorine in the extremities of the drinking water distribution system: the influence of stochastic water demands. *Procedia Engineering*, 70:172–180, 2014.

[7] S. Buchberger, J. Carter, Y. Lee, and T. Schade. Random demands, travel times, and water quality in deadends. Technical Report 90963F, American Water Works Association Research Foundation, Denver, CO, 2003.

[8] S. G. Buchberger and L. Wu. Model for Instantaneous Residential Water Demands. *Journal of Hydraulic Engineering*, 121(3):232–246, 1995. ISSN 0733-9429. doi:10.1061/(ASCE)0733-9429(1995)121:3(232).

[9] J. W. Davidson and F. J.-C. Bouchart. Adjusting Nodal Demands in SCADA Constrained Real-Time Water Distribution Network Models. *Journal of Hydraulic Engineering*, 132(1):102–110, 2006. ISSN 0733-9429. doi:10.1061/(ASCE)0733-9429(2006)132:1(102).

[10] M. J. Davis, R. Janke, and M. L. Magnuson. A Framework for Estimating the Adverse Health Effects of Contamination Events in Water Distribution Systems and its Application. *Risk Analysis*, 34(3):498–513, 2014. doi:10.1111/risa.12107.

[11] EPA. Sampling Guidance for Unknown Contaminants in Drinking Water. Technical Report EPA-817-R-08-003, U.S. Environmental Protection Agency, Washington, DC, 2008.

[12] EPA. Water Security Initiative: Evaluation of the Sampling and Analysis Component of the Cincinnati Contamination Warning System Pilot. Technical Report EPA-817-R-14-001G, U.S. Environmental Protection Agency, Washington, DC, 2014.

[13] EPA. Water Quality Surveillance and Response System Primer. Technical Report EPA 817-B-15-002, U.S. Environmental Protection Agency, Washington, DC, 2015.

[14] T. Gabriel. Thousands Without Water After Spill in West Virginia. *The New York Times*, Jan. 2014. URL `http://www.nytimes.com/2014/01/11/us/west-virginia-chemical-spill.html`.

[15] S. Hatchett, D. Boccelli, T. Haxton, R. Janke, A. Kramer, A. Matracia, S. Panguluri, and J. Uber. How accurate is a hydraulic model. In *Proceedings of the 2010 Water Distribution Systems Analysis Symposium, Tucson, Ariz*, 2010. doi:10.1061/41203(425)123.

[16] S. Hatchett, J. Uber, D. Boccelli, T. Haxton, R. Janke, A. Kramer, A. Matracia, and S. Panguluri. Real-time distribution system modeling: development, application, and insights. In *Proceedings of the eleventh international conference on computing and control for the water industry, Exeter, UK. Centre for Water Systems, University of Exeter, Exeter*, 2011.

[17] P. M. Jonkergouw, S.-T. Khu, Z. S. Kapelan, and D. A. Savi. Water quality model calibration under unknown demands. *Journal of Water Resources Planning and Management*, 2008. doi:10.1061/(asce)0733-9496(2008)134:4(326).

[18] D. Kang and K. Lansey. Real-Time Demand Estimation and Confidence Limit Analysis for Water Distribution Systems. *Journal of Hydraulic Engineering*, 135(10):825–837, Oct. 2009. doi:10.1061/(ASCE)HY.1943-7900.0000086.

[19] K. Klise, Bynum, D. Moriarty, and R. Murray. A software framework for assessing the resilience of drinking water systems to disasters with an example earthquake case study. *Environmental Modeling and Software*. Submitted, contact author for information.

[20] C. Laird, L. Biegler, B. van Bloemen Waanders, and R. Bartlett. Contamination Source Determination for Water Networks. *Journal of Water Resources Planning and Management*, 131(2):125–134, Mar. 2005. doi:10.1061/(ASCE)0733-9496(2005)131:2(125).

[21] S. P. Linger. *Spatial-temporal neighborhood-scale urban water demand estimation*. Master's Thesis, University of New Mexico, 2011. URL `http://hdl.handle.net/1928/12805`.

[22] S. H. Maier, R. S. Powell, and C. A. Woodward. Calibration and comparison of chlorine decay models for a test water distribution system. *Water Research*, 34:2301–2309, 2000.

[23] A. V. Mann, G. A. Hackebeil, and C. D. Laird. Explicit Water Quality Model Generation and Rapid Multiscenario Simulation. *Journal of Water Resources Planning and Management*, 140(5):666–677, May 2014. doi:10.1061/(ASCE)WR.1943-5452.0000278.

[24] S. A. McKenna, B. V. B. Waanders, C. D. Laird, S. G. Buchberger, Z. Li, and R. Janke. Source location inversion and the effect of stochastically varying demand. *Proc. ASCE World Water and Environmental Resources*, 10, 2005.

[25] M. Pasha and K. Lansey. Effect of parameter uncertainty on water quality predictions in distribution systems-case study. *Journal of Hydroinformatics*, 12:1–21, 2010. doi:10.2166/hydro.2010.053.

[26] L. A. Rossman. EPANET 2: Users Manual. Technical Report EPA/600/R-00/057, U.S. Environmental Protection Agency, Cincinnati, OH, Sept. 2000. URL `https://www.epa.gov/water-research/epanet`.

[27] E. Todini and S. Pilati. A gradient method for the analysis of pipe networks. Leicester Polytechnic, UK, Sept. 1987.

[28] J.-P. Watson, R. Murray, and W. E. Hart. Formulation and optimization of robust sensor placement problems for drinking water contamination warning systems. *Journal of Infrastructure Systems*, 15(4):330–339, 2009. doi:10.1061/(ASCE)1076-0342(2009)15:4(330).

[29] X. Xie, B. Zeng, and M. Nachabe. Sampling design for water distribution network chlorine decay calibration. *Urban Water Journal*, 12:190–199, 2013. doi:10.1080/1573062x.2013.831911.

[30] X. Yang and D. L. Boccelli. 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, Aug. 2014. doi:10.1061/(asce)wr.1943-5452.0000359.

Uncertainty Quantification of Contaminant Plume Spread in Water Distribution Systems

by

David Blaine Hart