Contaminant source characterization in water distribution systems using binary signals

Jitendra Kumar, E. Downey Brill, G. Mahinthakumar and S. Ranji Ranjithan

ABSTRACT

This paper presents a simulation–optimization-based method for identification of contamination source characteristics in a water distribution system using filtered data from threshold-based binary water quality signals. The effects of quality and quantity of the data on the accuracy of the source identification methodology are investigated. This study also addresses the issue of non-uniqueness in contaminant source identification under various data availability conditions. To establish the robustness and applicability of the methodology, numerous scenarios are investigated for a wide range of contamination incidents associated with two different networks. Results indicate that, even though use of lower resolution sensors lead to more non-unique solutions, the true source location is always included among these solutions.

Key words | contaminant warning system, evolutionary algorithms, non-uniqueness, water distribution systems, water quality sensors

Jitendra Kumar (corresponding author)
Computer Science and Mathematics Division,
Oak Ridge National Laboratory,
One Bethel Valley Road,
PO Box 2008, MS 6016,
Oak Ridge, TN 37831,
USA
E-mail: *kumarj@ornl.gov*

E. Downey Brill
G. Mahinthakumar
S. Ranji Ranjithan
Department of Civil, Construction and Environmental Engineering,
North Carolina State University,
Campus Box 7908,
Raleigh, NC 27695,
USA

INTRODUCTION

Water distribution systems (WDSs) are vital for supplying safe drinking water to the public, and they are vulnerable to contamination that can be introduced into the system either deliberately or accidentally. The fate and transport of a contaminant and the extent of the contaminant spread through the network depend on the characteristics of the network topology and the resulting hydraulic conditions of the network.

Networks of sensors can be used as an early warning system (EWS) to detect sudden deterioration in water quality. They are meant to supplement conventional water quality monitoring by providing timely information on unusual threats to a water supply system. The goal of an EWS is to reliably identify a contamination event in the distribution systems in time to allow an effective targeted response that reduces or avoids entirely the adverse effects of contamination on the system. The contaminant source characterization (i.e. identifying the contaminant source location and injection pattern) problem has been approached as an inverse problem by many researchers (e.g. Waanders *et al*. 2003; Laird *et al*. 2005, 2006). These approaches attempt to reconstruct a contamination injection event by matching the estimated contaminant concentration patterns with the observations at contaminant sensor locations. Most of the investigations reported in the literature on contaminant source characterization assume the availability of continuous concentration measurements from ideal contaminant-specific sensors.

Waanders *et al*. (2003) developed a non-linear programming-based approach for source inversion and applied it to a network in an industrial area. This optimization approach was used to find contaminant releases at nodes that resulted in contaminant concentrations matching as much as possible the sensor observations, which were simulated using EPANET. Laird *et al*. (2005) developed an origin-tracking algorithm to solve a non-linear programming-based optimization problem. They (Laird *et al*. 2006) extended it to consider the issue of non-uniqueness in the solution space by formulating the optimization model as a mixed-integer quadratic program. Guan *et al*. (2006) applied a reduced gradient method to solve an optimization model of the inverse contaminant characterization problem.

doi: 10.2166/hydro.2012.073

Preis & Ostfeld (2008) coupled a model tree with linear programming to solve the contaminant characterization problem. Model trees were constructed using EPANET-based simulated values of maximum concentration and time of its occurrence in the network for a set of random contaminant injection events. These model trees are then used as a surrogate for EPANET simulations and coupled with a linear programming-based optimization model to search for the best fitting contaminant characteristics.

While these studies offer new algorithms and methodologies for contaminant source detection, their applicability is contingent on the availability of precise and continuous sensor measurements. According to the state-of-the-art review of technologies by the United States Environmental Protection Agency (USEPA) for EWS, it is impractical and probably unproductive to focus on specific drinking water contaminants when designing an EWS (USEPA 2005, Section 3.1.2). The quality, precision and reporting frequency of the sensor observations may vary among utilities but an EWS is required to be able to detect the occurrence of a contamination event. An EWS methodology (READiw – Real time Event-Adaptive Detection, identification and warning) developed by Yang *et al.* (2009) focused on adaptively filtering out background data to detect anomalies in water quality measurements. This filtering can reduce the high rate of false-negative and false-positive detection in conventional sensors. If such algorithms are programmed into emerging water quality sensors, the resulting signals may have binary or V- or U-shaped patterns. Thus there is an increasing need for interpreting these types of filtered signals, including using them in the characterization of contaminant sources.

A few studies have focused on contamination characterization methods that use discrete measurements of contaminants in WDS. For example, Preis & Ostfeld (2008) extended the contaminant characterization investigations to consider different types of sensor signals, including a Boolean signal, indicating the presence of contamination at a sensor. They solved the inverse problem using a genetic algorithm that attempts to match the Boolean signal of the contaminant observations.

Because the amount of information is reduced in a Boolean signal versus a continuous signal, the solutions to the contaminant characterization problem will have a high degree of non-uniqueness and be sensitive to many related sensor parameters and uncertainties. In comparison to earlier work, this paper presents a new methodology that identifies a set of non-unique solutions in a single execution of the optimization algorithm to reconstruct the contamination characteristics based on binary signals. This study also conducts an extensive sensitivity investigation to evaluate the robustness of the approach when considering the sensitivity of the binary signal to the detection threshold level that acutely affects the quality of the information. Sensitivity of a sensor may impact its utility as well as cost (USEPA 2005, Section 3.1.7). Thus, in addition to lowering of information quantity due to inherent data filtering in a binary signal, the sensitivity of the sensor affects the quality of the information used to estimate the source characteristics. Accuracy and precision of an inverse solution are also affected by the amount of observations available in the network, which depends on the number and the placement of the sensors in the network. The applicability of the proposed new method considering these factors and sensitivities is investigated for an array of contamination events in two different illustrative WDS networks.

DATA FILTERING EFFECTS IN BINARY SIGNALS

This section illustrates the effects of filtering when obtaining a binary signal from continuous concentration signals that would have been observed at an ideal sensor. Figure 1(a) shows the continuous concentration patterns at a sensor. The filtered signals (shown by the vertical bars) at different levels of sensor sensitivity are shown in Figures 1(b–d) that correspond to threshold values of 0.01, 0.1 and 0.2 mg/l, respectively. As evident in the graphs, the duration over which the contaminant is detected decreases as the sensor sensitivity decreases. Another effect of data filtering is the potential lack of discrimination of the magnitude of the contaminant concentration at the sensor. For example, the sensors with threshold values of 0.01 and 0.1 mg/l cannot differentiate the larger and smaller peaks (Figures 1(b, c)), and the sensor with a threshold value of 0.2 mg/l completely misses the smaller peak. Such effects of data filtering in binary signals can potentially contribute to non-uniqueness and lack of precision in the source characteristics being identified.

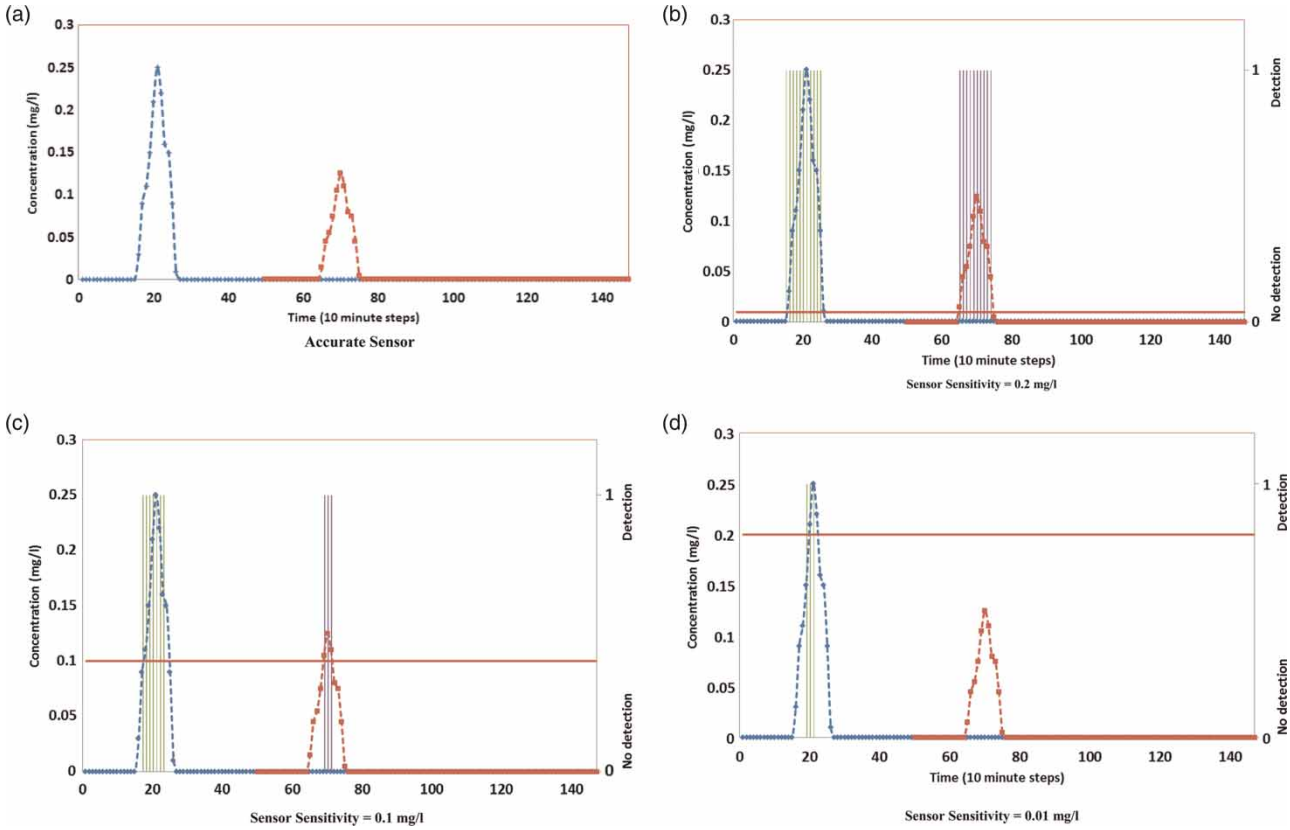

Figure 1 | Effect of sensor threshold on the filtering of data to produce a binary signal. The true signal is shown in all figures for comparison.

CHARACTERIZATION OF THE CONTAMINANT SOURCE USING BINARY SIGNALS

This paper focuses on developing and testing an approach that uses the filtered binary signals to identify the location of a contaminant source and the injection pattern. The contaminant source characterization problem is posed as an inverse problem where the source characteristics are treated as unknowns. For each potential set of source characteristics (i.e. the location and injection pattern), the binary signals are simulated and compared to the observations. The prediction error is then used as a measure of goodness of fit to direct the search for a set of source characteristics that best fits the observations. The resulting search problem is solved using a coupled simulation–optimization approach.

The following characteristics of the contaminant source are decision variables for the optimization problem:

1. Location (L) of the contaminant source in the network. Location of the source is assumed to be at one or more of the nodes in the WDS network.
2. Start time of the contaminant injection into the system (T_0). Start time is measured as the time elapsed from the start time of the simulation.
3. The pattern of contaminant injection ($C_0 = \{C_{0,1}, C_{0,2}, \ldots, C_{0,D}\}$) into the system measured in terms of milligrams of contaminant injected per minute ($C_{0,j}$) into the system (mg/min). Injection rate (mg/min) is assumed to be constant during each water quality timestep of the simulation. No assumptions are made regarding the magnitude, pattern or duration measured in number of time increments (D) of the contaminant injection.

The optimization model for this inverse problem is defined below (Equations (1)–(7)), assuming that the contamination scenario involves only a single source;

however, it could be extended to scenarios involving simultaneous contamination at multiple sources:

$$\text{Minimize } E = \sum_{i=1}^{N_s} \sum_{t=1}^{T_s} \left| BS^s_{i,t} - BS^a_{i,t} \right| \quad (1)$$

$$\text{subject to } BS^s_{i,t} = \begin{cases} 0, & \text{if } C^s_{i,t}(L,\ T_0,\ C_0) < C_{th} \\ 1, & \text{if } C^s_{i,t}(L,\ T_0,\ C_0) \geq C_{th} \end{cases} \forall i = 1,\ \cdots N_s;\ \forall t = 1,\ \cdots T_s \quad (2)$$

$$BS^a_{i,t} = \begin{cases} 0, & \text{if } C^a_{i,t} < C_{th} \\ 1, & \text{if } C^a_{i,t} \geq C_{th} \end{cases} \forall\, i = 1,\ \cdots N_s;\ \forall t = 1,\ \cdots T_s \quad (3)$$

$$1 \leq L \leq N \quad (4)$$

$$1 \leq T_0 \leq T_{detect} \quad (5)$$

$$0 \leq C_{0,j} \leq C_{max} \forall\, j = 1,\ \ldots,\ D \quad (6)$$

$$1 \leq D \leq (T_s - T_0) \quad (7)$$

where $BS^s_{i,t}$ and $BS^a_{i,t}$ are the simulated binary signal and the actual binary signal, respectively, for given source characteristics at sensor i at any time t, N is the total number of nodes in the network, N_s is the number of sensors in the network, T_s is the total time for which contaminant mixing and transport in the network is being simulated and D is the duration (number of time increments) for which the source is active. The simulated concentration of the contaminant at sensor i at time t is given by $C^s_{i,t}(L,\ T_0,\ C_0)$ for a given solution as a function of location (L), time, start time of the contamination injection (T_0) and the contaminant injection pattern (C_0). C_{th} is a constant defining the threshold concentration level for determining detection and generating a binary signal (equals '1' if concentration is above C_{th} and equals '0' otherwise). $C^a_{i,t}$ is the actual measured concentration of the contaminant at sensor i at time t.

The objective function (Equation (1)) is a measurement of error represented as the total number of 'mis-hits' at the sensors. A simulated expected signal ($BS^s_{i,t}$) for the given values of L, T_0 and C_0 will be considered a mis-hit if it does not match the filtered binary value of the actual signal ($BS^a_{i,t}$). Specifically, there is a mis-hit when there is a simulated detection signal ($BS^s_{i,t} = 1$) when the actual signal was of no-detection ($BS^a_{i,t} = 0$) or vice versa. The optimization problem is to obtain the values of L, T_0 and C_0 that minimize the total number of mis-hits between the simulated and actual sensor signals. Equation (2) filters the simulated concentration levels for a given set of source characteristics at the sensors into detection/no-detection binary signals, depending on the sensitivity of the sensor for the particular contaminant.

Equations (4)–(7) describe the constraints for the optimization problem. The contaminant source location (L) can be any node in the network (Equation (4)). The injection start time (T_0) is greater than or equal to the start time of the solution (timestep '1') and less than or equal to the time of the first detection (T_{detect}) at any of the sensors (Equation (5)). The contaminant injection rate ($C_{0,t}$) can assume any value between zero and any large specified maximum concentration (C_{max}, 15,000 mg/min in this study) (Equation (6)). The source injection pattern is allowed to be of any length between the injection start time (T_0) and the end of the simulation (T_s) (Equation (7)).

METHODOLOGY

The solution to the optimization model represented by Equations (1)–(7) would ideally identify the true contaminant source location and the injection pattern. Because of data limitations, however, the solution to Equations (1)–(7) is likely to yield non-unique source characteristics to fit the observations equally well. Specifically, there may be data quality limitations resulting from, for example, measurement errors and filtering effects, as well as quantity limitations resulting from, for example, discrete measurements and limited number of observations. Therefore, the solution approach should be able to conduct a global search and be able to identify simultaneously the set of potential source characteristics that provide an equally good fit to the given observations. The niched co-evolution strategies (NCES) search procedure (Zechman & Ranjithan 2004, 2007) was used to solve the above model. Also, given the possibility of error in modeling of the physical distribution network, it is also important to identify near-optimal solutions that might fit the real observations. The

NCES procedure also provides this capability. In this work, the original NCES (implemented in Java) was re-implemented (in C) to provide improved computational performance as well as enhanced capability to use parallel computing resources for the WDS simulations (Kumar 2010).

NCES conducts a global search using evolution strategies (ES), a generalized heuristic optimization approach. An ES works in a continuous space and has the capability to self-adapt its major algorithmic parameters such as selection and mutation. ES starts with a randomly initialized population of μ individuals. A probabilistic mutation operator is applied to produce ν new solutions each generation. The next set of individuals is selected from either a combined set of parent and offspring solutions (denoted as $(\mu + \nu)$ selection) or from the set of offspring alone (denoted as (μ,ν) selection).

The standard ES is extended in NCES to generate a set of alternative solutions that performs equally well in terms of fitness, i.e. the objective function being optimized. The search for a set of solutions in NCES is based on the concept of cooperative co-evolution to evolve simultaneously a set of subpopulations for exploring the decision space. Each subpopulation is driven to a region in the solution space that has good fitness and is far away from other solutions with good fitness. This helps to identify different contamination source locations and mass injection patterns, if any, that can produce the same or similar observations at the sensors (i.e. within comparable prediction error values), leading to insights about potential non-uniqueness or imprecision about the estimate of the true source characteristics. The Euclidean distance between the location of an individual and the location of the centroid of every subpopulation is calculated, and the distance from the closest centroid is used as the distance metric. For the contaminant source identification problem, identifying sources at different possible locations was recognized to play an important role. Thus for all reported scenarios the distance measures in terms of location are used during the selection step. Distance between two nodes in the network can be calculated in terms of Euclidean length, or as the connection length based on the shortest distance along the pipes connecting the two nodes, or in terms of average hydraulic lengths of flow (or contaminant travel times) between the two nodes. In this work, the distance was defined in terms of Euclidean distance in two dimensions (i.e. horizontal planar distance) since it worked equally well for the problems considered, based on preliminary tests, and is the least computationally demanding approach.

While one subpopulation focuses its search on finding the optimal solution based on the prediction error (E in Equation (1)), the other subpopulations search for alternatives based on the current best prediction error value as well as the degree of difference among the alternative solutions. In the results presented in this paper, every solution that is within one mis-hit from the best prediction error value is considered as a viable alternative solution.

NCES was applied and tested on two water distribution networks to study: (1) the effect of the sensitivity of binary signals (quality of data) on the accuracy and non-uniqueness of the solutions to such source identification problems; (2) the effect of the quantity of available data (i.e. number of installed sensors) on the solutions to source identification problems and (3) the robustness of the method when applied to a set of different contamination scenarios and networks.

Representation of variables

A contamination event was characterized by three types of decision variables: location of the contaminant source, start time of the contaminant injection, and the rate and pattern of contaminant injection. The following representation schemes were used to encode them:

1. Location of contaminant source: Every node in a pipe network is assumed to be a potential location for contaminant injection. The nodes are sequentially numbered, and an integer variable was used to encode the node number, bounded by 1 and the total number of nodes in the network (see Equation (4)).
2. Start time: An integer variable was used to represent the time when the contaminant injection began. This number refers to the discrete timestep in the EPANET simulation of the contaminant fate and transport. The value of the integer variable is bounded by 1 (i.e. start time of the simulation) and the timestep when the first sensor was triggered (see Equation (5)).
3. Contaminant injection pattern: The contaminant injection pattern is represented as an array of real values (Equation (6)), each denoting the mass loading at a timestep after the start of injection. As the duration of injection is not predefined, the length of this array varies, enabling each solution

to represent contaminant injection patterns of different durations (Equation (7)). The injection pattern is encoded as a linked-list, along with a special feasibility-preserving mutation operator as described below.

Genetic operators

Except for the integer variable representing the start time, the variables representing the node index of contamination location and contaminant injection pattern contain problem-specific characteristics, which require the use of special adaptive mutation operators as described below.

1. Location of contaminant source. When mutating an integer value that represents a node index in the network, a new node index value from among the feasible index set is chosen using a Gaussian-like adaptive mutation (Beyer & Schwefel 2002). The new node index is selected considering topology adjacency and node connectivity. A list of one-node-away connected neighbors, two-node-away connected neighbors and so on was first generated to capture the connectivity and adjacency of each node in the WDS network. Then a mutation parameter specifies a random node-connectivity distance between the current node and the child node, and the index for a new node within that distance from the current node is selected randomly from the connectivity/adjacency-preserving list.
2. Start time. A simple adaptive Gaussian mutation operator (Beyer & Schwefel 2002) for a bounded integer variable is used.
3. Contamination injection pattern. A special mutation operator was designed to operate on this linked-list of real values representing the injection pattern. To allow the length of this linked-list to be dynamically variable, three different mutation operations were used. One is for the addition of link-items and the other is for the deletion of link-items, both of which change the duration of the pattern. The third mutation operation is a standard Gaussian mutation that changes a real value in the linked-list to modify the contaminant injection rate at the corresponding timestep. While no restrictions are applied to the length of injection pattern, the operator implicitly bounds the injection loading to not extend beyond the end of simulation time. The value of the injection rate during any timestep has a lower bound of zero and a specified upper bound. The mutation operator allows the injection rate to take the value of zero, thus enabling the representation of continuous as well as intermittent contaminant injection patterns.

ILLUSTRATIVE WATER DISTRIBUTION NETWORK EXAMPLES

Networks analyzed

Two networks were modeled and investigated to demonstrate the proposed simulation–optimization method for contamination events representing conservative contaminants. Data for both problems are available in the literature. Using such networks is necessary to compare new methods directly to earlier ones (a practice that has been of great benefit in WDS research through the years). The first case (Network 1) is a test network included with the standard EPANET (Rossman 2000) distribution. Network 1 (Figure 2(a)) consists of 97 nodes,

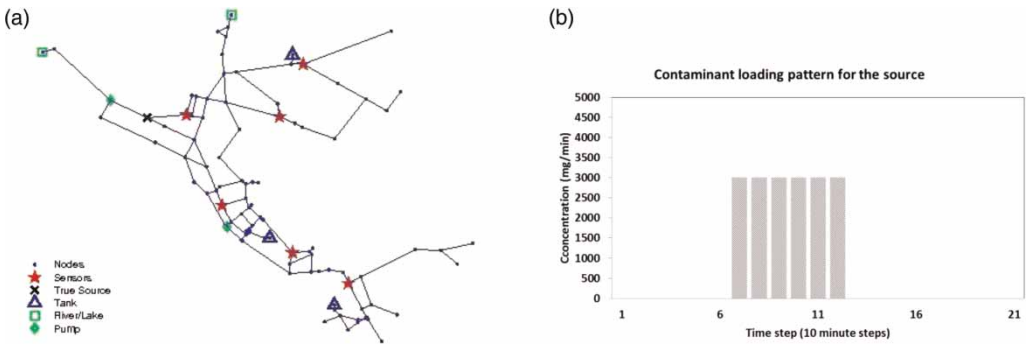

Figure 2 | Layout of the Network 1 water distribution network. (a) Network, (b) injection pattern.

including two water supply sources, three tanks and 117 pipes. The network is assumed to have a maximum of six water quality sensors. EPANET was used for simulating the hydraulic and contaminant transport in the network for a period of 24 h starting at 12:00 am. Hydraulic simulations were carried out after every hour and were considered to be at steady state during each hour. Water quality transport simulations were carried out using 10 min timesteps and the measurements were recorded every 10 min.

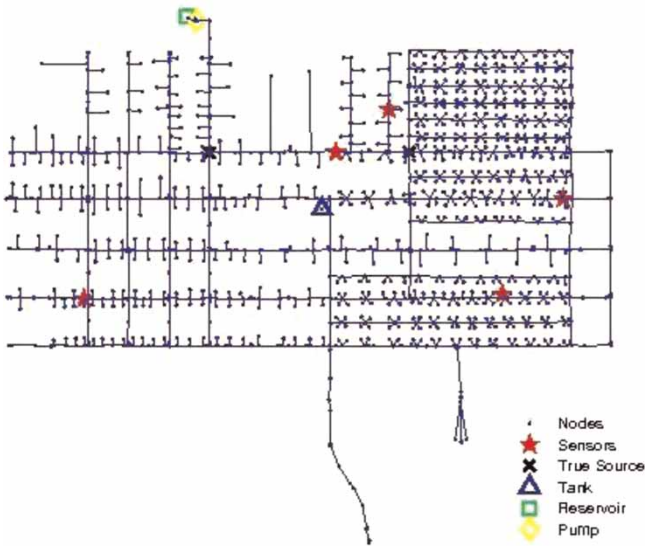

Figure 3 | Layout of the Micropolis water distribution network.

The second test network considered is for the virtual city of Micropolis (Brumbelow *et al.* 2007), which represents a realistic municipality capturing the evolution of the infrastructure over many decades. This is a publicly available WDS dataset intended and used for research studies without compromising public security (Figure 3). Micropolis mimics a small city of approximately 5,000 people in a typically rural region that is served by a single elevated storage tank with a capacity of 440,000 gallons (1,670,000 l). The WDS consists of 1,236 nodes, 575 mains (for a total length of 10.1 miles), 486 service and hydrant connections (total length of 7.1 miles), 197 valves, 28 hydrants, eight pumps, two reservoirs and one tank. The 458 demand nodes are composed of 434 residential, 15 industrial and nine commercial/institutional users with diurnal demand patterns that are defined at hourly time intervals.

Scenarios analyzed

Network 1 was used to construct several hypothetical contamination scenarios (Scenarios 1–11, Table 1) designed to investigate various aspects of the source characterization problem that is to be solved based on binary signals from a set of sensors.

Scenario 1 was designed to serve as a benchmark for an ideal case in which continuous and accurate concentration measurements of the contaminants are assumed to be

Table 1 | Details of the scenarios analyzed

Scenario	True source location: Node No. (Network)	Sensor type	Sensor sensitivity (mg/l)	No. of sensors	Contaminant injection pattern
1	12 (Network 1)	Ideal	–	6	Continuous (1 h)
2	12 (Network 1)	Binary	0.2	6	Continuous (1 h)
3	12 (Network 1)	Binary	0.1	6	Continuous (1 h)
4	12 (Network 1)	Binary	0.01	6	Continuous (1 h)
5	12 (Network 1)	Binary	0.1	6	Continuous (1 h)
6	12 (Network 1)	Binary	0.1	3	Continuous (1 h)
7	12 (Network 1)	Binary	0.1	1	Continuous (1 h)
8	12 (Network 1)	Binary	0.1	6	Intermittent
9	12 (Network 1)	Binary	0.1	6	Intermittent
10	20 (Network 1)	Binary	0.1	6	Continuous (1 h)
11	54 (Network 1)	Binary	0.1	6	Continuous (1 h)
12	787 (Micropolis)	Binary	0.1	5	Continuous (1 h)
13	653 (Micropolis)	Binary	0.1	5	Continuous (1 h)

available at the water quality sensors. The sum of the squared error between the actual and simulated concentrations at the sensors was used as the metric to be minimized in the optimization problem. Equation (8) represents the mathematical formulation of the objective function for this scenario:

$$\text{Minimize } E = \sum_{i=1}^{N_S} \sum_{t=1}^{T_S} (C_{i,t}^{S}(L,\ T_0,\ C_0) - C_{i,t}^{a})^2 \text{mg}^2/\text{l}^2 \tag{8}$$

In Scenarios 2, 3 and 4, the sensitivity of the sensors was set at 0.2, 0.1 and 0.01 mg/l, respectively, to examine the impact of increasing sensor sensitivity on the accuracy and degree of non-uniqueness of the solutions to the source characterization problem. The effect of the quantity of available data was studied in Scenarios 5, 6 and 7 using observation data from six sensors, three sensors and one sensor, respectively. Scenarios 8–11 were designed to test the methodology for a range of different contaminant injection instances and to assess the robustness and general applicability of the method. While a continuous and uniform contaminant injection (i.e. a constant mass injection pattern active for 1 h) scenario was used for Scenarios 2–7, the ability to solve the problem for non-uniform and discontinuous contaminant injection scenarios was analyzed in Scenarios 8 and 9, respectively. To study the performance of the procedure when the contamination source location varies, Scenarios 10 and 11 were analyzed with sources located in different regions of Network 1. For Scenarios 1–11, the contamination injection was assumed to start at 1:00 am (i.e. at timestep 7) (Figure 2(b)) and was assumed to be active for a 1 h period. To evaluate the applicability of the proposed procedure for other networks, Scenarios 12 and 13 (described later) were analyzed using a contamination event in the Micropolis network.

For each potential solution (i.e. the source characteristics) in the NCES search, the EPANET model was run to predict the concentration values at the sensors. For all scenarios except Scenario 1, these concentration values were then filtered and converted to a simulated binary signal indicating the presence or absence of contamination at each sensor. The actual binary signal from a sensor was generated for the true contamination event again using EPANET simulations; however, in real applications, these observations would be the sensor readings obtained from measurements. When the simulated signal and the actual binary signal at each sensor and each observation time point do not match, the mis-hit count was incremented by one, as expressed in Equation (1).

In most scenarios involving Network 1 (Scenarios 1–5 and 8–11), the search was conducted to identify up to five alternative potential solutions by including five subpopulations in NCES. For the Network 1 scenarios with fewer sensors (Scenarios 6 and 7) and the larger Micropolis network (Scenarios 12 and 13), 50 subpopulations were used. For all scenarios, the number of subpopulations was chosen to be sufficiently large to accommodate the maximum possible number of non-unique solutions that differ in source locations and have no more than one mis-hit higher than that of the best solution. Subpopulation 1 was designed to search for the solution with the minimum number of mis-hits, while the other subpopulations were set to search for alternative solutions that are maximally different from each other and have no more than one mis-hit over the lowest number of mis-hits (i.e. the minimum objective function value for Subpopulation 1). The NCES algorithmic parameters used in this study are shown in Table 2.

Table 2 | Algorithmic parameter settings for the NCES algorithm

Parameter	Setting
Population size (μ)	200 (Network 1), 400 (Micropolis)
Individuals created by mutation (ν)	200 (Network 1), 400 (Micropolis)
Number of subpopulations	5 (Network 1, Scenarios 1–5 and 8–11), 50 (Scenarios 6–7 and Micropolis)
Mutation strength for self adaptive mutation	0.5
Number of generations	100 (Network 1), 50 (Micropolis)
Feasibility criteria	Objective function value of best individual of first subpopulation +1
Number of random trials	50 (Network 1), 10 (Micropolis)

RESULTS

For each scenario reported in this section (i.e. Scenarios 1–11), 50 random trials were carried out to assess the robustness of the solution approach. For each trial, prediction error was calculated using Equation (1) for all scenarios except for Scenario 1 where Equation (8) was used. The summary statistics as well as a detailed description of a representative solution are presented, and the following statistics are also reported: (1) prediction error in Subpopulation 1 (which is tasked to search for the minimum error solution); (2) number of solutions with non-unique source locations and (3) locations of all the feasible non-unique solutions identified in the 50 random trials. In the ensuing results and discussion, every feasible alternative with a different source location is counted as a non-unique solution, but the corresponding start time is reported as part of the results.

Ideal water quality sensor measurements (based on Scenario 1)

The results for Scenario 1 described below are based on the sets of five alternative solutions obtained for each of the 50 trials. Solutions with a prediction error that is no more than 110% of the best prediction error were accepted as feasible alternative solutions. In every trial, the first subpopulation, designated to search for the solution with the least prediction error, always identified the true source location (i.e. Node 12). The median prediction error of the best solution was 0.217 mg^2/l^2 with a range of 0.088–0.476 mg^2/l^2. Figure 6 shows the statistical distribution of the prediction error of this solution among the trials. The small standard deviation in prediction error among the 50 trials shows that the NCES algorithm gives consistent performance from trial to trial (i.e. the algorithm is reasonably robust). The source location and the contaminant injection start time for the set of feasible alternative solutions obtained for a typical trial are presented in Table 3 and the injection patterns are shown in Figure 4. No more than three feasible alternative solutions (with source locations at Node 10, Node 12 or Node 86) were identified in the 50 trials. The median number of feasible alternative solutions was one with a range of 1–3 (the corresponding mean was 1.10 with a standard deviation of 0.36), illustrating that no other alternative

Table 3 | The source location and injection start time of the non-unique solutions obtained for a typical trial for Scenarios 1–4

Scenario	Solutions (location, start time)	No. of mis-hits
1	(12, 7)	0.195[a]
2	(12, 7)	0
	(12, 7)	1
	(86, 10)	0
3	(12, 8)	0
	(86, 9)	1
	(12, 8)	1
4	(12, 7)	0
	(12, 6)	1

[a]Sum of squared error mg^2/l^2 since binary signals were not used in this scenario.

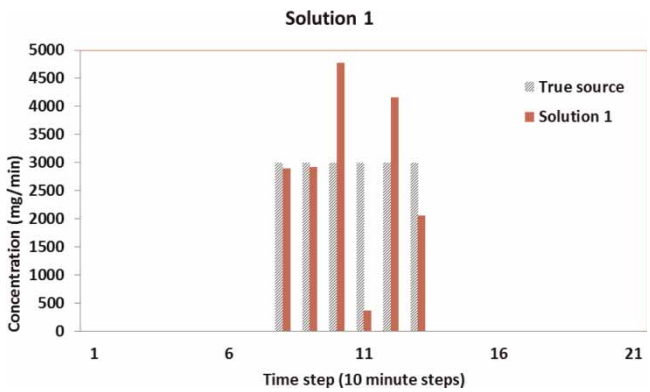

Figure 4 | Contamination injection patterns for a typical solution for Scenario 1 and the true source.

besides the best solution was found for most trials (Figure 7). In the 50 trials, feasible solutions were identified at locations 10, 12 and 86 with an average error of 0.476, 0.228 and 0.309 mg^2/l^2, respectively. Even in the case of ideal sensors with error-free continuous measurements, the number of sensors (six in this scenario) and their placements in the network are insufficient to precisely identify the source characteristics, resulting in non-unique solutions at three locations in the network. The two alternative locations (Nodes 10 and 86) (Figure 5(a)) identified in the 50 trials are adjacent to the true source location (Node 12) in the network.

Effects of varying data quality (based on Scenarios 2, 3 and 4)

Scenarios 2–4 were run with different values of sensor sensitivity to study the effect of data quality (due to different

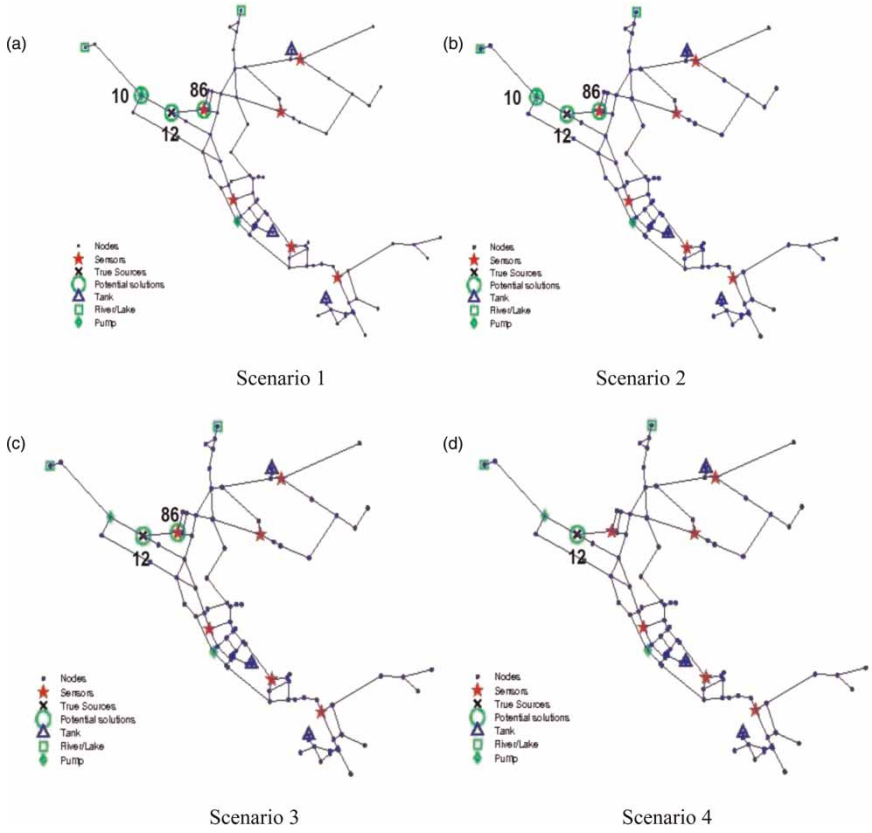

Figure 5 | Potential contamination source locations identified by the set of non-unique solutions for Scenarios 1–4.

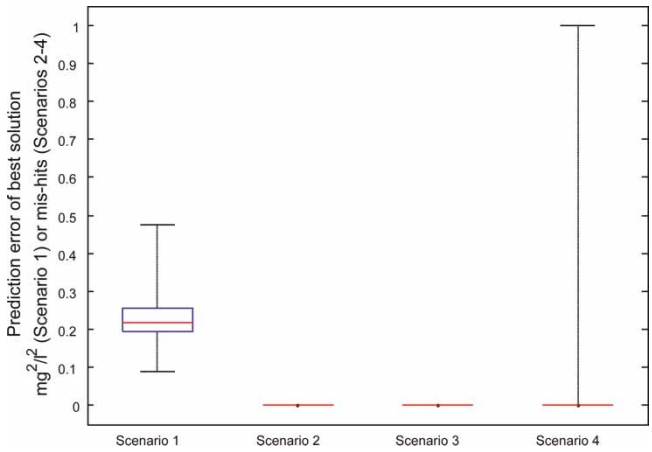

Figure 6 | Prediction error distribution of the best solutions obtained for the 50 trials. Horizontal bar (in red) indicates the median, the box indicates the mid 50% and the whiskers indicate the range. The absence of box and whisker for Scenarios 2 and 3 indicates that for nearly all trials the prediction error was zero mis-hits.

degrees of data filtering in the sensor) on the level of non-uniqueness. In each case, 50 random trials were run to assess the number of alternative solutions and the relative locations of the source nodes associated with the non-unique solutions.

Scenario 2: sensor sensitivity = 0.2 mg/l

In each of the 50 trials, at least one solution among the set of alternative solutions matched the observation data perfectly (i.e. no mis-hit) as shown in Figure 6. Table 3 shows a representative result for this scenario. In this case, the location and start time of the best solution were identical to those of the true source but the injection pattern was different from that of the true source. The source location (Node 86) of a non-unique solution is adjacent to the true source location, but with a perfect prediction of the observation data. Another solution with the source location and start time the same as those of the true source was identified, but it had a different contaminant injection pattern, resulting in one mis-hit in the observation data. Based on 50 trials, the median number of feasible alternative locations was two with a

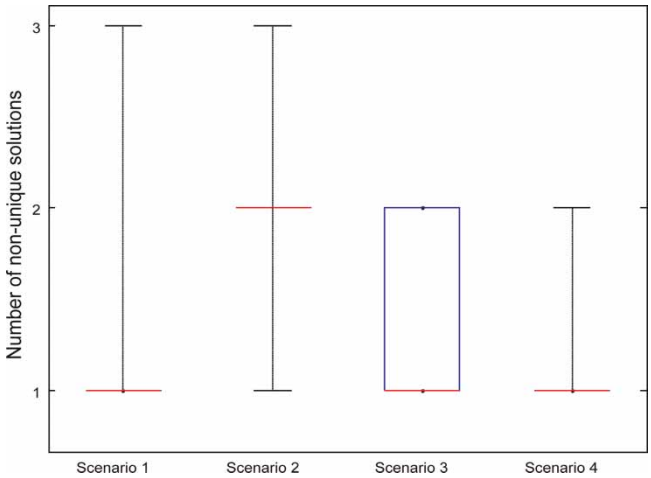

Figure 7 | Distribution of the number of non-unique locations obtained for the 50 trials. Horizontal bar indicates the median, the box indicates the mid 50% and the whiskers indicate the range. The absence of box for Scenarios 1, 2 and 4 indicate that most solutions corresponded to the median.

range of 1–3 (Figure 7); the corresponding mean was 2.12 with a standard deviation of 0.48. The three feasible locations over all trials are shown in Figure 5(b). This demonstrates that, for a low resolution sensor, one could expect a higher number of non-unique locations.

Scenario 3: sensor sensitivity = 0.1 mg/l

The key aspects of the solutions for a representative trial for this scenario with improved sensor sensitivity are shown in Table 3. A solution identified at the true source location (Node 12) had a start time of 8 (instead of the true start time of 7), but perfectly matched the observation data. The second solution was identified at an adjacent location (Node 86) and matched the observation data with only one mis-hit. The third solution that was identified at the true source location matched the observation data with a single mis-hit. A solution perfectly matching the observed data (i.e. no mis-hit) was always identified among the set of feasible solutions obtained for each of the 50 trials (Figure 6). The two locations (i.e. Nodes 12 and 86) of alternative solutions are shown in Figure 5(c). Based on the 50 trials, the median number of alternative locations was just 1 with a range of 1–2 (Figure 7); the corresponding mean was 1.46 with a standard deviation of 0.50. When compared to Scenario 2, this represents a reduction in non-uniqueness due to the increased resolution of the sensor.

Scenario 4: sensor sensitivity = 0.01 mg/l

Table 3 shows a summary of the key characteristics of a representative solution obtained for this scenario with the most sensor sensitivity. The best solution in the first subpopulation perfectly matched the source location and start time of the true source. One other solution was identified, which had the same source location as the true source but with a start time of 6 (instead of 7). In the 50 trials, the best solution was found with a median prediction error of zero mis-hits with a range of 0–1, as shown in Figure 6. All trials identified the true source location as the only feasible location (Figures 5(d) and 7). This represented a further reduction in non-uniqueness due to the increased sensitivity of the sensor.

Comparison of Scenarios 2, 3 and 4

The Scenario 2, 3 and 4 results, as summarized in Figures 6 and 7, show the trends in the accuracy of solutions and degree of non-uniqueness as the level of sensitivity of the sensors increases. The first subpopulation, which was dedicated to search for the best solution, was able to obtain a perfectly matching solution (zero mis-hits) in all trials of Scenarios 2 and 3 and in most trials of Scenario 4 (Figure 6). This is because it is easier to match the observation data perfectly in low-resolution sensors (Scenarios 2 and 3) due to the filtering effect. Conversely, one would expect the solutions to be more error-prone (thus show more non-uniqueness) with decreasing resolution in filtering of the sensor signals. As the sensor sensitivity increases, the number of non-unique solutions decreases (Figure 7). Correspondingly, the prediction error at the true source decreases while at the other potential sources the prediction error increases (Table 4), helping to rule out the non-unique solutions that are not at the true source location.

Effect of varying data quantity (based on Scenarios 5, 6 and 7)

Scenarios 5, 6 and 7

The amount of data available for characterizing the source has a significant impact on the source identification

Table 4 | The average prediction error (number of mis-hits) for non-unique sources located at Nodes 10, 12, and 86

Location	Scenario 2	Scenario 3	Scenario 4
Node 10	1.0	N/A[a]	N/A[a]
Node 12	0.671	0.523	0.347
Node 86	0.550	1.0	N/A[a]

[a]No feasible solution with no more than one mis-hit was found.

problem, which was studied using several scenarios with different numbers of sensors. Scenarios 5, 6 and 7 include six sensors, three sensors and one sensor, respectively, and the sensitivity of all the sensors was assumed to be 0.1 mg/l. Scenario 5 is the same as Scenario 3 and, for Scenarios 6 and 7, 50 subpopulations were used. All other parameters were the same as in previous scenarios. It was observed that, as the amount of data decreases with the decreasing number of sensors, the number of alternative solutions that fit the observations (thus the degree of non-uniqueness) increases. The median number of alternative locations was 1, 5 and 5 for the six-sensor, three-sensor and one-sensor cases, respectively. Figures 8–10 show the alternative source locations in the network for Scenarios 5–7, respectively, and the overall number of locations where alternative sources were found was 2, 13 and 13 for Scenarios 5–7, respectively.

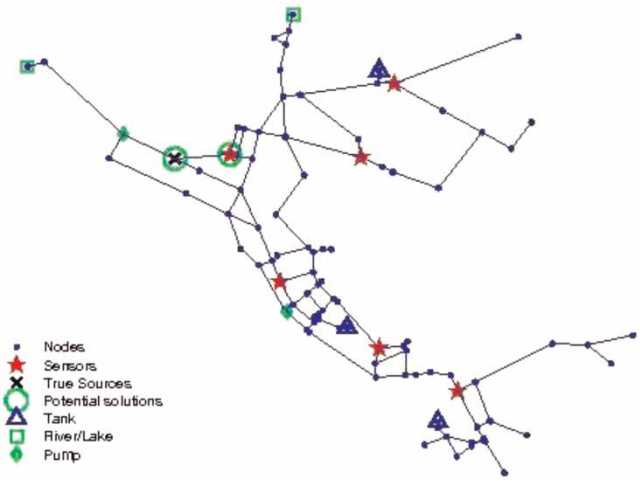

Figure 8 | Potential contamination source locations identified for the six-sensor case (Scenario 5).

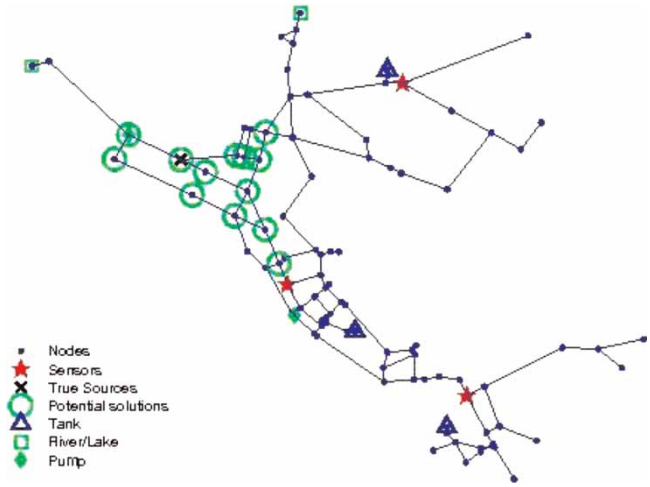

Figure 9 | Potential contamination source locations identified for the three-sensor case (Scenario 6).

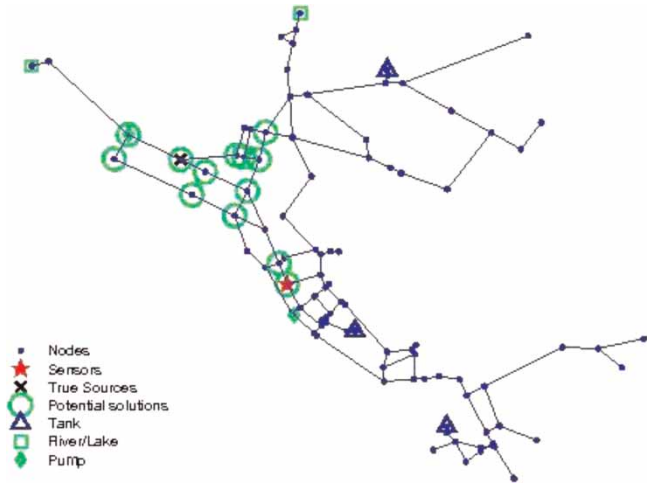

Figure 10 | Potential contamination source locations identified for the one-sensor case (Scenario 7).

In the three-sensor scenario (Scenario 6), contamination was detected at only one sensor, which was the same location as where the sensor was placed in the one-sensor scenario (Scenario 7). Thus, active detection observations were the same in both cases. The no-detection signal at the other two sensors in Scenario 6 is additional information that is available and used when reconstructing the source characteristics. This leads to some differences in the overall results of the two scenarios, resulting in slightly different sets of nodes for the 13 locations.

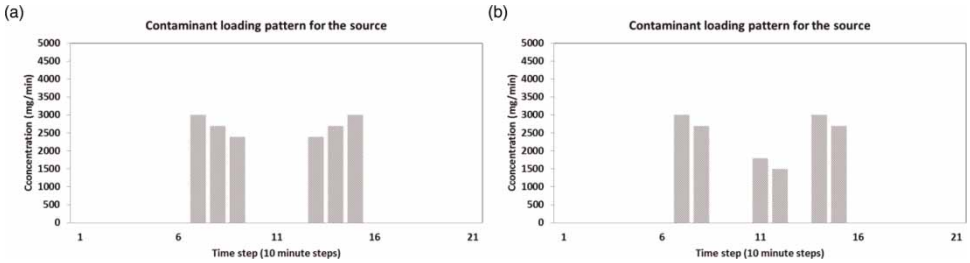

Figure 11 | Intermittent contaminant injection pattern used in Scenario 8(a) and Scenario 9(b).

Effect of different contamination scenarios (based on Scenarios 8–11)

Scenarios 8 and 9: intermittent contaminant injection patterns

To test the applicability of the method to different source characteristics, different mass injection patterns were considered in Scenarios 8 and 9, where two different intermittent contaminant injection scenarios were used (Figure 11). For both scenarios, Node 12 was chosen as the true source location and six sensors with a sensitivity of 0.1 mg/l were assumed to be present in the network (Table 1).

For both the scenarios, feasible solutions contained the true source location (Node 12) or an adjacent node (Node 86). For Scenario 8, the best solution in the first subpopulation had an average prediction error of 1.0 mis-hits over the 50 trials and a standard deviation of 1.069 mis-hits. The median number of non-unique solutions was 1 with a range of 1–2. Similarly, for Scenario 9, the average prediction error of the best solution was 0.64 mis-hits (with a standard deviation of 0.942 mis-hits), and the median number of non-unique solutions was 1, with a range of 1–2. Comparing the two scenarios, one can surmise that the two injection patterns are not sufficiently different to cause a change in source location because the same two locations (Nodes 12 and 86) were identified as possible solutions in both scenarios.

Scenarios 10 and 11: different source locations

To test the ability to identify a contaminant source at different locations in the network, in Scenarios 10 and 11 the source was placed at Node 20 and Node 54, respectively (Figures 12 and 13). The contaminant injection pattern was the same as in Scenarios 1–7. Six sensors with a sensitivity of 0.1 mg/l were placed at the same locations as before in Scenarios 1–4.

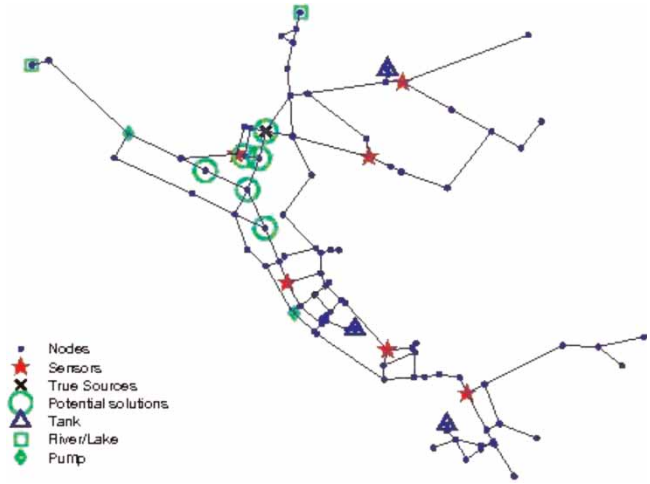

Figure 12 | Potential contamination source locations identified for Scenario 10.

For Scenario 10, where the source was located at Node 20, a solution with the true source location and a set of non-unique solutions with five other locations (Nodes 13, 16, 17, 18 and 85), all of which are close to the true source location (Table 5, Figure 12), were found. A perfect-fit (i.e. no mis-hit) solution was identified by the first subpopulation during each of the 50 trials. As shown by Solution 1 in Table 5, a perfect fit in the objective space does not necessarily imply that the solution obtained in the decision space is the same as the true case. In this case, both the location (Node 18) and start time (11) were different from the true source characteristics (located at Node 20, with a start time of 7). This is expected since the loss of information in

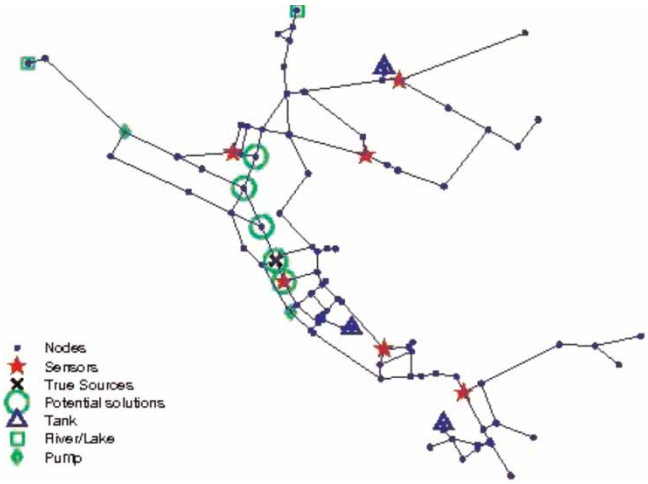

Figure 13 | Potential contamination source locations identified for Scenario 11.

Table 5 | The set of solutions obtained for a typical trial for Scenario 10

Solution	Location	Start time	No. of mis-hits
1	18	11	0
2	17	12	1
3	13	11	1
4	85	11	1
5	20	8	1

Table 6 | The set of solutions obtained for a typical trial for Scenario 11

Solution	Location	Start time	No. of mis-hits
1	54	7	0
2	18	2	1
3	17	2	1
4	17	5	1
5	17	1	1

a binary signal contributes to non-unique solutions that perfectly match the observation data. The median number of alternatives was 3, with a range of 2–5, which signifies a high degree of non-uniqueness.

In Scenario 11 where the true source was located at Node 54, non-unique solutions were identified at the true source location along with four other locations (Nodes 16, 17, 18 and 88) (Table 6, Figure 13). A perfect-fit solution was identified by the first subpopulation in all 50 random trials. The median prediction error of solution was 1 mis-hit with a range of 0–1. The median number of alternative locations was 3, with a range of 1–5, again signifying a high degree of non-uniqueness. Although the same set of sensors are used as in the other scenarios, the high level of non-uniqueness exhibited by Scenarios 10 and 11 underscore the importance of sensor placement for contamination detection and source characterization associated with a wide range of contaminant source locations.

Testing for a different water distribution network (based on Scenarios 12 and 13)

For simulating the contamination scenario in the Micropolis network, water quality simulations were carried out using 5 min timesteps with EPANET. Five sensors with a sensitivity of 0.1 mg/l were included in the network (Figure 14). Presented here are two scenarios, Scenario 12 and Scenario 13, with the contamination source located at Node 787 and Node 653, respectively. In both scenarios the contamination source was assumed to be active for 1 h starting at 12:00 am (i.e. at timestep 1) with an injection pattern as shown in Figure 15. For each scenario, the source characterization method was applied with 50 subpopulations to search for a maximum of 50 alternative (non-unique) solutions. Solutions within a single mis-hit from the best solution were

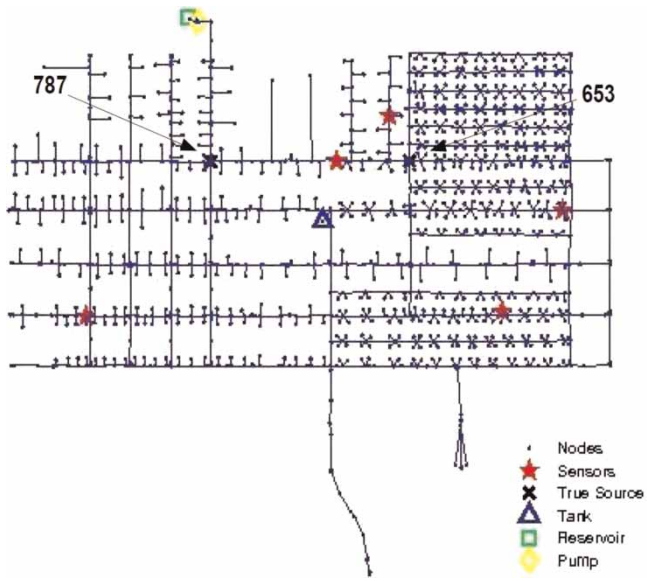

Figure 14 | Layout of the Micropolis water distribution network.

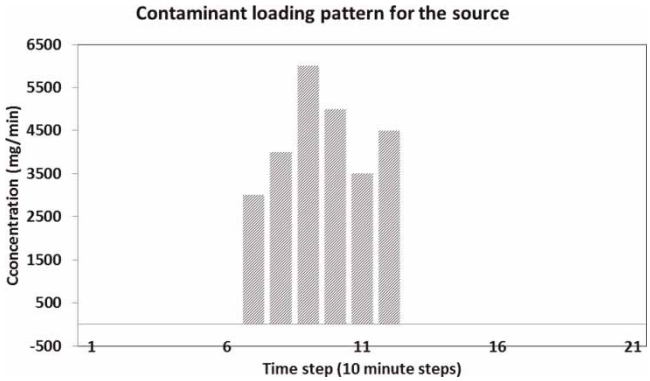

Figure 15 | Contamination injection pattern for the true source in Scenarios 12 and 13.

accepted as a feasible solution to the problem. The statistics of the solutions obtained for 10 random trials are reported for each scenario.

Scenario 12

In this scenario, the contaminant source was located at Node 787 with a contaminant injection pattern as shown in Figure 15. Because the contaminant source is close to the source of water in the network, the contaminant spreads out to a large section of the system and is detected at all five water quality sensors. In each trial, the best solution had a prediction error of 0 mis-hits (perfect match). The median number of alternatives was 14, with a range of 13–14. In each trial, the true source location was always identified and among the 10 trials, 13 other non-unique source locations were identified (Figure 16).

Scenario 13

In this scenario, the contaminant source was located at Node 653 with a contamination injection pattern as shown in Figure 15. Again, 10 random trials were carried out. The true source location was identified by the solutions for all 10 trials. As in Scenario 12, the best solution in each trial had a prediction error of 0 mis-hits. In all trials, the median number of alternatives was 35, with a range of 33–40. Among the 10 trials, in addition to the true source location, 7 alternative locations were identified (Figure 17).

Results for Scenarios 12 and 13 demonstrate that the methodology described in this paper is able to identify

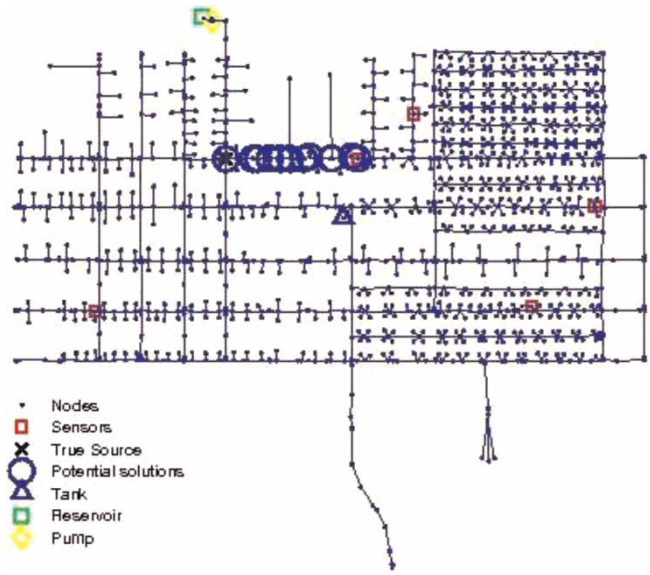

Figure 16 | Potential contamination source locations (the true source location and 13 others) identified in the 10 random trials for Scenario 12.

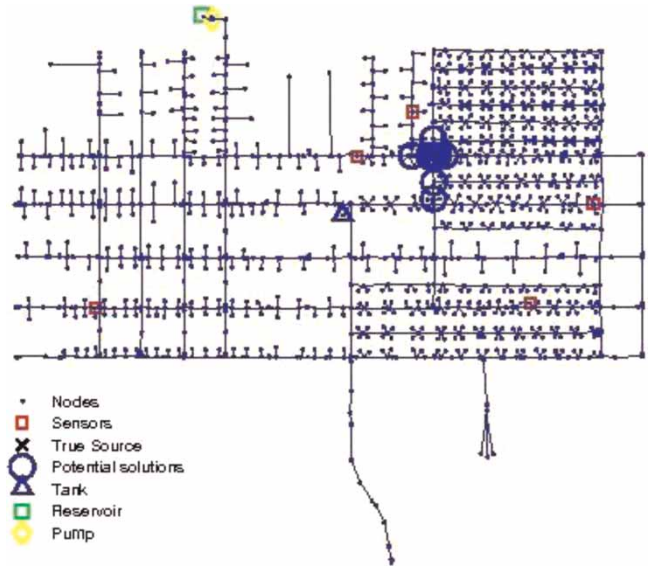

Figure 17 | Potential contamination source locations (the true source location and seven others) identified in the 10 random trials for Scenario 13.

potential solutions to the contaminant source characterization problem for the Micropolis example, which represents a medium-sized water distribution network. Besides consistently identifying the true source locations, it was also able to identify a set of non-unique solutions that are expected to fit the limited observations. It is interesting to note that the set of non-unique source

locations is concentrated around the true source location, potentially helping to pinpoint an area of the network where the true source is located.

COMPUTATIONAL IMPLEMENTATION

The methodology described above was implemented as part of an integrated parallel simulation–optimization framework designed to use high performance computing resources. A customized parallel EPANET 'wrapper' that allows for multiple but independent evaluations of EPANET on parallel computing resources was developed using the Message Passing Interface (MPI) (Gropp *et al.* 1999). The ES-based optimization algorithm was also designed and implemented for execution in a parallel computing environment. The parallel ES-based optimization algorithm and the MPI-enabled parallel EPANET simulator were coupled within the parallel simulation–optimization framework. All results presented in the paper were carried out using this parallel framework implemented on a Linux-based cluster computing environment. For the simulations performed in this study, typical computation time for the smaller network (97 nodes, 100,000 EPANET simulations per trial) was 166.00 s and the time for the larger network (1,257 nodes, 1,000,000 EPANET simulations per trial) was 8,820.00 s using 12 processors of an Opteron cluster. Details of the parallel framework and computational results can be found in Chapters 5, 6 and 7 of the thesis by Kumar (2010).

CONCLUSIONS

This paper presents an evolutionary algorithm-based methodology for characterizing a pollutant source during a contamination event in a WDS that is being monitored using a set of sensors that report binary signals indicating the presence/absence of contamination. The source characteristics were reconstructed using these binary signals.

The accuracy of contaminant source characterization depends primarily on the amount and quality of the data available. The effects of sensor sensitivity on the accuracy and degree of non-uniqueness were studied. It was observed that the decrease in sensitivity of the sensors leads to a reduction in accuracy and an increase in the level of non-uniqueness in source prediction. Further, a decrease in the quantity of data resulted in an increase in the level of non-uniqueness. Extremely low quantity and quality of the available data may make it impossible to reasonably resolve the contamination source characteristics (i.e. precisely predict the source characteristics), but still help to pinpoint a small region of the WDS network where the source could potentially be found. The methodology was tested for a range of different contamination scenarios to establish its robustness. The methodology was able to resolve the contaminant source to a small subset of possible locations while identifying the true source in all cases.

The methods and results of these investigations that were based on binary signals from water quality sensors can be applied to process filtered information that is utilized and analyzed in other instances such as event detection methods (Guralnik & Srivastava 1999; Hart *et al.* 2007; McKenna *et al.* 2007; Uber *et al.* 2007). The solution of the source identification problem is highly computing-intensive, especially for larger networks. Methods such as network backtracking (Shang *et al.* 2002; De Sanctis *et al.* 2010) and logistics regression modeling (Liu *et al.* 2011b) can be used as prescreening tools to reduce the search space and reduce the computational efforts required by the optimization algorithms. Even when the number of sensors available in a network is small, these prescreening tools can effectively eliminate many nodes as potential source locations, efficiently reducing the complexity of the contamination characterization problem. While the focus of the work presented in this paper was to address the contamination source identification problem, the ES-based optimization methodology for identifying a set of alternative solutions can applied for the other WDS problems where a similar non-uniqueness issue is encountered; e.g. water quality sensor placement problems (Ostfeld *et al.* 2008; Tryby *et al.* 2010).

New sensor technologies provide the capability for the real-time collection of sensor network data through SCADA systems. An extended adaptive dynamic version (ADOPT) of the ES-based optimization methodology (Liu *et al.* 2011a) can effectively use the real-time data to provide the best estimate for the contaminant source characteristics

using the data available at any given point of time; the estimate of the contaminant characteristics can be updated with streaming sensor data. The ADOPT method assumes, however, the availability of continuous sensor data. The methods and results from the binary signal-based source characterization can be integrated with the ADOPT approach to investigate dynamic source characterization using binary signals. While almost all past studies and the work presented in this paper treat the contaminant as conservative, more work is needed to understand the effects of contaminant reactivity on the ability to solve the contaminant characterization problem. Preliminary results (Kumar 2010) suggest that the approach is computationally tractable.

ACKNOWLEDGEMENTS

This work was supported by the National Science Foundation (NSF) under grant no. CMS-0540316 from the Dynamic Data Driven Analysis System (DDDAS) program. This document describes activities performed under contract no, De-AC0500OR22750 between the US Department of Energy and Oak Ridge Associated Universities. All opinions expressed in this report are the authors' and do not necessarily reflect policies and views of the US Department of Energy or the Oak Ridge Institute for Science and Education.

REFERENCES

Beyer, H. G. & Schwefel, H. P. 2002 Evolutionary strategies: a comprehensive introduction. *Natural Comput.* **1** (1), 3–52.

Brumbelow, K., Torres, J., Guikema, S., Bristow, E. & Kanta, L. 2007 Virtual cities for water distribution and infrastructure system research. In: *World Environmental and Water Resources Congress* (ed. K. C. Kabbes), Tampa, FL. ASCE, Reston, VA.

De Sanctis, A. E., Shang, F. & Uber, J. G. 2010 Real-time identification of possible contamination sources using network back tracking methods. *J. Wat. Res. Plann. Mngmnt.* **136** (4), 444–453.

Gropp, W., Lusk, W. & Skjellum, A. 1999 *Using MPI: Portable Parallel Programming with the Message Passing Interface*, 2nd edition. MIT Press, Cambridge, MA.

Guan, J., Aral, A. M., Masila, A. L. & Grayman, W. M. 2006 Identification of contaminant sources in water distribution systems using simulation–optimization method: Case study. *J. Wat. Res. Plann. Mngmnt.* **132** (4), 252.

Guralnik, V. & Srivastava, J. 1999 Event detection from time series data. In *Proc. International Conference on Knowledge Discovery and Data Mining (KDD-99)*, San Diego, CA, pp. 33–42.

Hart, D., McKenna, S. A., Klise, K., Cruz, V. & Wilson, M. 2007 CANARY: a water quality event detection algorithm development tool. In: *World Environmental and Water Resources Congress* (ed. K. C. Kabbes), Tampa, FL. ASCE, Reston, VA.

Kumar, J. 2010 *Asynchronous Hierarchical Parallel Evolutionary Algorithm-Based Framework for Water Distribution Systems Analysis*. PhD thesis, North Carolina State University, Raleigh, NC. Online. Available from http://repository.lib.ncsu.edu/ir/bitstream/1840.16/6388/1/etd.pdf (accessed 23 April 2012).

Laird, C. D., Biegler, L. T. & van Boleman Waanders, B. G. 2006 A mixed-integer approach for obtaining unique solutions in source inversion of drinking water networks. *J. Wat. Res. Plann. Mngmnt.* **132** (4), 242–251.

Laird, C. D., Biegler, L. T., van Boleman Waanders, B. G. & Bartlett, R. A. 2005 Contamination source determination for water networks. *J. Wat. Res. Plann. Mngmnt.* **131** (2), 125–134.

Liu, L., Ranjithan, S. & Mahinthakumar, G. 2011a Contamination source identification in water distribution systems using an adaptive dynamic optimization procedure. *J. Wat. Res. Plann. Mngmnt.* **137** (2), 183–192.

Liu, L., Sankarasubramanian, A. & Ranjithan, S. 2011b Logistic regression analysis to estimate contaminant sources in water distribution systems. *J. Hydroinf.* **13** (3), 545–557.

McKenna, S. A., Hart, D., Klise, K., Cruz, V. & Wilson, M. 2007 Event detection from water quality time series. In: *World Environmental and Water Resources Congress* (ed. K. C. Kabbes), Tampa, FL. ASCE, Reston, VA.

Ostfeld, A., Uber, J. G., Salomons, E., Berry, J. W., Hart, W. E., Phillips, C. A., Watson, J.-P., Dorini, G., Jonkergouw, P., Kapelan, Z., di Pierro, F., Khu, S.-T., Savic, D., Eliades, D., Polycarpou, M., Ghimire, S. R., Barkdoll, B. D., Gueli, R., Huang, J. J., McBean, E. A., James, W., Krause, A., Leskovec, J., Isovitsch, S., Xu, J., Guestrin, C., VanBriesen, J., Small, M., Fischbeck, P., Preis, A., Propato, M., Piller, O., Trachtman, G. B., Wu, Z. Y. & Walski, T. 2008 The battle of the water sensor networks (BWSN): a design challenge for engineers and algorithms. *J. Wat. Res. Plann. Mngmnt.* **134** (6), 556–568.

Preis, A. & Ostfeld, A. 2008 Genetic algorithm for contaminant source characterization using imperfect sensors. *Civil Eng. Environ. Syst.* **25** (1), 29–39.

Rossman, L. A. 2000 *EPANET 2 Users Manual*. National Risk Management Research Laboratory, Environmental Protection Agency, Cincinnati.

Shang, F., Uber, J. G. & Polycarpou, M. M. 2002 Particle backtracking algorithm for water distribution system analysis. *J. Environ. Eng.* **128** (5), 441–450.

Tryby, M. E., Propato, M. & Ranjithan, S. R. 2010 Monitoring design for source identification in water distribution systems. *J. Wat. Res. Plann. Mngmnt.* **136** (6), 637.

Uber, J. G., Murray, R., Magnuson, M. & Umberg, K. 2007 Evaluating real-time event detection algorithms using synthetic data. In: *World Environmental and Water Resources Congress* (ed. K. C. Kabbes), Tampa, FL. ASCE, Reston, VA.

United States Environmental Protection Agency (USEPA) 2005 *Technologies and Techniques for Early Warning Systems to Monitor and Evaluate Drinking Water Quality: A State-of-the-art Review*. USEPA, Washington, DC.

Waanders, B. G., Bartlett, R. A., Bigler, L. T. & Laird, C. D. 2003 Nonlinear programming strategies for source detection of municipal water networks. In: *World Environmental and Water Resources Congress* (eds P. Bizier & P. DeBarry), Philadelphia, PA. ASCE, Reston, VA.

Yang, Y. F., Haught, R. C. & Goodrich, J. A. 2009 Real-time contaminant detection and classification in a drinking water pipe using conventional water quality sensors: techniques and experimental results. *J. Environ. Mngmnt.* **90**, 2494–2506.

Zechman, E. M. & Ranjithan, S. 2004 An evolutionary algorithm to generate alternatives (EAGA) for engineering optimization problems. *Eng. Optim.* **36** (5), 539–553.

Zechman, E. M. & Ranjithan, S. 2007 Generating alternatives using evolutionary algorithms for water resources and environmental management problems. *J. Wat. Res. Plann. Mngmnt.* **133** (2), 156–165.

First received 8 June 2011; accepted in revised form 27 September 2011. Available online 16 February 2012