Optimal sensor placement with mitigation strategy for water network systems under uncertainty

Rajib Mukherjee\textsuperscript{a}, Urmila M. Diwekar\textsuperscript{a,\textsuperscript{*}}, Ashok Vaseashta\textsuperscript{b}

\textsuperscript{a} Vishwamitra Research Institute, 2714 Crystal Way, Crystal Lake, IL, 60012, United States
\textsuperscript{b} International Clean Water Institute, Manassas, VA 20112, United States

A R T I C L E   I N F O

Article history:
Received 20 November 2015
Received in revised form 22 November 2016
Accepted 17 March 2017
Available online 21 March 2017

Keywords:
Sensor placement
Water network
Stochastic optimization
Impact assessment
Uncertainty Principle

A B S T R A C T

Contamination of water in a distributed water network can cause serious public health disaster. The contamination can occur at one part of the network and may spread to different regions depending on the flow pattern of water. It is essential to monitor contamination using distributed water networks and a practical way to do so, is through the application of sensor networks. Sensors are used for timely detection of the presence of various contaminants in water. To accurately and optimally map the contamination distribution, it is quintessential to place sensors in optimal positions in the network whereby the goal is to minimize the net effect from contamination – inadvertent or intentional. The demand of water at different locations in the network can be uncertain leading to various flow patterns. The location of contamination within the network can also be uncertain. With these uncertainties, the propagation of the contaminant in the network varies accordingly, thus affecting the impact from contamination. An optimal placement of the sensors should account for these uncertainties. A mitigation response of contamination, as detected by a sensor network is manifested in a decreased demand due to public information as a function of distance from the sensor. The overall impact from contamination is estimated by including impacts, as found in the downstream of the sensor in the network. In this work, the most efficient network of sensors is designed by including demand and location uncertainty as well as response from downstream from the sensor. The problem is solved by formulating a nonlinear, stochastic mixed integer program. Generally stochastic formulation requires repeated function calculations. In our research, a reweighting approach is used instead of repeated function simulation to estimate an expected value of the impact. The reweighting scheme along with a novel sampling technique has been implemented through a Better Optimization algorithm for Nonlinear Uncertain Systems (BONUS). The results are compared with that from TEVA SPOT, which neither takes into account demand uncertainty nor downstream response from the source of contamination (attack). Inclusion of downstream response and demand uncertainty on optimal sensor placement in water network systems show better sensor layouts with minimum impacts from contamination.

1. Introduction

Distributed water network consists of nodes and extended pipelines connecting such nodes and covering a large area with varied population. Due to open infrastructure, water networks have risk of contamination and hence security breach that may lead to catastrophic health impacts. In order to minimize these risks, it is essential for timely detection of such contamination incidents and impose a control action that can minimize risks asso-

\textsuperscript{*} Corresponding author.

E-mail addresses: rajib@vri-custom.org (R. Mukherjee), urmila@vri-custom.org (U.M. Diwekar), prof.vaseashta@ieee.org (A. Vaseashta).

http://dx.doi.org/10.1016/j.compchemeng.2017.03.014
0098-1354/© 2017 Elsevier Ltd. All rights reserved.
Most of these articles published so far discuss about contamination propagation and subsequent risk minimization in a steady state environment with no uncertainty associated with population distribution leading to demand variation at the nodes. As a result the analysis mentioned above can result in suboptimal solutions.

Optimal sensor placement literature also gave rise to information related to identification of source of contamination (Laird et al., 2005, 2006) and leak (Casillas et al., 2013; Sarate et al., 2014). In Laird et al. (2005, 2006) the goal was to identify the exact source of contamination with the help of spatially distributed sensors that measures water quality. We address an inverse problem in which the unknowns are pollutant injections in the network. However the problem is ill posed and have multiple solutions, and can only be used for early warning systems.

An optimal sensor network for impact minimization will have sufficient number of sensors at appropriate locations so that contamination can be detected before a substantial amount of hazard can propagate. To make the solution robust it is essential to consider uncertainties in the problem. Uncertainty can manifest itself through (a) changing population density/water demands at various junctions (nodes) or (b) through varying probability of contamination at different nodes. Berry et al. (2005) considered typical water demand throughout the day with a fixed set of patterns with no uncertainty. The problem is solved using a deterministic optimization method. Shastri and Diwekar (2006) have used demand uncertainty and used stochastic optimization. Both Berry et al. (2005) and Shastri and Diwekar (2006) have the objective of minimizing the fraction of population exposed to contamination. Berry et al. (2006) had used a temporal integral programing method where the objective was to minimize the impact of contamination. Here again they have assumed that the demand follows a set of patterns and each pattern is constant for one or more hours. They have also assumed that an alarm is raised when contamination is detected, all consumption is prevented, and the network is completely shut down as a response to the alarm.

The present work proposes a new approach to solve the sensor placement problem using impact minimization. It uses a demand uncertainty and contamination (attack) location uncertainty, characterized in terms of probability distributions. Thus, we have an optimization under uncertainty problem where there are uncertainties involved in the data and/or model. This problem can be categorized as stochastic programming problem (Diwekar, 2003).

In this investigation, the demand uncertainty is sampled using our efficient Hammersley Sequence Sampling (HSS) (Kalagnanam & Diwekar, 1997; Diwekar & Kalagnanam, 1997). We have used the model network from EPANET-MSX (Rossman, 1993; Rossman, 2000; Shang, Uber, & Rossman, 2008) to simulate the system’s steady state behavior for different base cases (demand). In stochastic programming, the system needs to be evaluated for different cases and at different iterations. The solution to this problem is achieved through a Better Optimization for Nonlinear Uncertain System (BONUS) algorithm (Sahin & Diwekar, 2004) where a reweighting scheme is used to avoid multiple simulation of process behavior for different sample cases and at different steps.

We looked into the problem of sensor placement to detect contamination incident with a different approach. When contamination is detected in a water supply network, two main actions can be taken: 1) a mitigation strategy where the population is alerted not to use water and 2) close the water supply in order to minimize health impact (Di Nardo et al., 2013). The effectiveness of closing a water network depends on the possibility of closing pipes to disconnect network sectors. In an ideal situation, it is assumed that there will be zero demand as an immediate response to a contamination incident. In reality, there is a high level of demand uncertainty associated with the response from a contamination incident. Detection of contamination with a sensor will initiate a warning signal in the network resulting in a reduction of the downstream demand. In the present investigation, this variability is analyzed using techniques from statistical methods to determine which nodes should be measured using a network of sensors so that the net impact from a contamination incident can be minimized. The proposed problem formulation and solution method gives results that are optimal and can be translated to an actual water distribution network where there are uncertain demands. The net impact of contamination is found to be minimum when sensors are located considering response from the downstream nodes and demand uncertainty as compared to the conventional methods including TEVA-SPOT, a sensor placement optimization tool for water security that uses location uncertainty alone.

The next section provides the problem formulation. This is followed by a section on solution procedure. The sections explaining the solution procedure for the proposed problem has discussion on the BONUS algorithm and method of impact estimation. The last section presents the results from the proposed formulation, and its comparison with other formulations using a case study network and then draw conclusions.

### 2. Problem formulation

A sample water network system is shown in Fig. 1. This network is based on the “Example Network 1” of EPANET. There are 12 nodes in the network, comprising two pumping stations, one storage tank, and nine consumption points (numbered as 10, 11, . . . , 32). Each consumption point can have uncertain demands. Sensors can be located at the nodes with demands or in between two nodes.

The sensor placement problem can be defined in multiple ways. A fixed number of sensors can be optimally located so that the net impact is minimized. In this case the overall sensor cost will limit the maximum number of sensor. For a fixed number of sensors, the sensor network can consider the cost of the sensors versus benefits from the sensor in contamination detection and human health protection (Bagajewicz et al., 2004; Bhushan et al., 2008; Shastri & Diwekar, 2006; Li & Upadhyaya, 2011). A common practice is to compare the health cost with sensor cost. An example of formula based on cost derived from Shastri & Diwekar, 2006 is given in Eqs. (1) to (6).

\[
\text{Minimize}(\text{cost}_t + \sum_{i} l \sum_{j} \text{jsij} + (\text{cost}_{op} + \alpha + \sum_{i} l \sum_{j} p \sum_{j} \text{jsij}))
\]  
(1)
Subject to:
\[ C_{pi} = 1 = 1, \ldots, n; \ p = 1, \ldots, P \]  
(2)
\[ s_{ij} = s_{ij} = 1, \ldots, n - 1; i < j \]  
(3)
\[ C_{ij} = C_{pk} - S_{ij}, \quad (i, k, j) \in E; \text{s.t.} f_{ijp} = 1 \]  
(4)
\[ \sum_{(i,j) \in E: j < i} s_{ij} \leq S_{\text{max}}, \ s_{ij} \in \{0, 1\} \]  
(5)

EPANETModelEquations(BlackBox)

where

- \( C_{ij} \): contamination indicator, 1 if j is contaminated by an attack at i or 0 at a flow pattern p
- \( \delta_{ij} \): demand at location j for a flow pattern p
- \( s_{ij} \): Sensor indicator, 1 if sensor present between i and j, otherwise 0
- \( E \): Set of nodes.
- \( \alpha \): attack probability (assumed to be same at all nodes for different flow patterns)
- \( \cos \ t, \cos \ s \): cost of sensor
- \( \cos \ t, pop \): human health cost

In the above formulation, the probability of an attack (\( \alpha \)) at different nodes during flow pattern p is considered to be equally likely. The demand \( \delta_{ij} \) at node j while flow pattern p is active is considered to be uncertain. The contamination indicator \( C_{ij} \) indicates if node j is contaminated by an attack at node i during flow pattern p by taking value of 1 and 0 otherwise. \( s_{ij} \) is a decision variable indicating presence of sensor. It takes value of 1 if a sensor is placed between node i and j and 0 otherwise. A flow pattern p between \( f_{ijp} \) node i and j is represented by a binary parameter \( f_{ijp} \), takes value of 1 if there is positive flow from i to j during flow pattern p and 0 otherwise.

The problem considers uncertainty through variable population density at different nodes that leads to demand uncertainty \( \delta_{ij} \) with a known normal probability distribution. Shifts in demand at various nodes with time also contributed in different flow patterns in the problem. The contamination indicator \( C_{ij} \) and flow pattern \( f_{ijp} \) values are obtained from the EPANET-MSX simulator for various demand distributions and under different contamination conditions. In solving the optimization problem, different constraints are used which ensures that when a node is directly attacked, it is contaminated and contamination only propagates from node i to node j if node i is contaminated, there is positive flow along a directed edge from i to j, and there is no sensor on that edge. The \( S_{\text{max}} \) is the other constraint that enforces a limit on the total number of sensors.

The population at any node is considered at risk if it consumes contaminated water. Thus demand uncertainty rather than population uncertainty is used for analysis. It is also assumed that the demand uncertainty is proportionate to the population uncertainty. In the above formulation, costs of sensors, \( \cos \ t, \cos \ s \) and human health cost associated with each person affected by the contamination \( \cos \ t, pop \) are also considered along with uncertainties \( \delta_{ip} \) in demand density and location of attack. The goal of the above optimization problem is to find the optimal sensor configuration so as to minimize the human health cost in treatment of expected fraction of population at risk from an attack.

The second formulation used for our analysis is that from Berry et al. (2006). They have used the impact on population for sensor placement. It is given in Eq. (7) to (12). A fixed set of flow pattern is used to estimate the impact \( d_{ai} \). During estimation of impact, Berry et al. (2006) also assumes that all consumptions can be prevented as a response to the alarm from an attack incident. The witness indicator \( x_{ai} \) used in their formulation ensures that the sensor at node i observe contamination from an attack incident a. A witness indicator makes sure that sensors are only located on the nodes where there is a presence of contaminants. The resulting objective function needs to minimize the total impact from different attack incidents.

Minimize \[ \sum_{a} \alpha_{a} \sum_{i} d_{ai}x_{ai} \]  
(7)

Subject to:
\[ \sum_{i \in L_{a}} x_{ai} = 1 \forall a \in A \]  
(8)
\[ x_{ai} \leq s_{i} \forall a \in A; i \in L_{a} \]  
(9)
\[ 0 \leq x_{ai} \leq 1, \ \forall a \in A; i \in L_{a} \]  
(10)
\[ \sum_{s_{i} \in (0, 1)} S_{\text{max}, i} \in L_{a} \]  
(11)

EPANETModelEquations(BlackBox)

where

- \( \alpha_{a} \): weight of an attack incident
- \( d_{ai} \): total impact from an attack incident a when contamination is detected at i
- \( L_{a} \): Location contaminated by incident a
- \( s_{i} \): Sensor indicator, 1 if sensor present at location i, otherwise 0
- \( A \): Set of contamination incident
- \( x_{ai} \): Witness indicator, 0 if contamination is not witnessed at i from an attack incident a

To make a robust analysis for sensor placement, we not only need to minimize the net impact from sensor placement but also consider demand uncertainties and variable flow patterns. The present formulation includes demand uncertainty and variable flow pattern in net impact assessment. Furthermore, the present formulation considers a mitigation strategy by informing public about contamination, thereby, reducing the downstream demands from the node that has been contaminated.

2.1. Demand model using mitigation strategy

Sensor placement theory hitherto was based on a complete control strategy where it was assumed that as soon as a sensor detects a contamination incident, there is complete shutdown in the water network and any further consumption of contaminated water can be prevented. In a mitigation strategy, it is assumed that there will be some impact even with the placement of sensor and a complete shutdown may not be possible. As soon as contamination is detected at some node, the public warning is issued at the nodes downstream to the contaminated node as mentioned in Di Nardo et al. (2013). We also assumed that as the flow continued downstream, it would take some time for the contamination to reach the next node and so on. The information will also take some time to propagate. Thus, the downstream demand will decrease. The actual amount of decreased demand depends on the real world information network. In the present case study, we assumed that the demand to be half of previous node. The demand distribution in the downstream nodes may vary with different assumption but the approach we are presenting here is applicable to all those cases.

In our case study, we assumed demand distribution to be normal distribution. We assumed that initial demand with \( \mu \pm 3 \sigma \) spread where \( \sigma = 1/3 \mu \) for this case study. If the node has a sensor and it detects contaminants, then we assume that the demand
Table 1
The mean and standard deviation of demand distribution as a function of distance.

<table>
<thead>
<tr>
<th>Consecutive Nodes in a Pattern</th>
<th>Demand from the sensor, (ii)</th>
<th>Mean, (\mu) demand</th>
<th>Demand standard deviation, (\sigma)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>(\mu_0/\mu(0+2) = \mu_0/2.0)</td>
<td>(\sigma_0/\sigma(0+2) = \sigma_0/2.0)</td>
<td>(\sigma_0/\sigma(0+2) = \sigma_0/3.0)</td>
</tr>
<tr>
<td>1</td>
<td>(\mu_1/\mu(1+2) = \mu_1/4.0)</td>
<td>(\sigma_1/\sigma(1+2) = \sigma_1/4.0)</td>
<td>(\sigma_1/\sigma(1+2) = \sigma_1/5.0)</td>
</tr>
<tr>
<td>2</td>
<td>(\mu_2/\mu(2+2) = \mu_2/5.0)</td>
<td>(\sigma_2/\sigma(2+2) = \sigma_2/5.0)</td>
<td>(\sigma_2/\sigma(2+2) = \sigma_2/6.0)</td>
</tr>
<tr>
<td>3</td>
<td>(\mu_3/\mu(3+2) = \mu_3/6.0)</td>
<td>(\sigma_3/\sigma(3+2) = \sigma_3/6.0)</td>
<td>(\sigma_3/\sigma(3+2) = \sigma_3/7.0)</td>
</tr>
</tbody>
</table>

is reduced by half shifting the mean and changing the standard deviation of demand distribution by half for the next node. For consecutive downstream nodes, the mean and standard deviation of demand distribution are further reduced as a function of distance. The example mean and standard deviation of the demand variation as a function of sensor node and successive downstream nodes is given in Table 1.

### 2.2. Optimization with mitigation strategy

To make a robust analysis for sensor placement, we need to minimize the net impact from the sensor node as well as downstream nodes. Thus, the net impact estimating function should be used as an objective of minimization to determine optimal sensor location. This is represented by Eq. (13) below. In the present formulation we considered the scenario where an impact from a contamination can be mitigated as compared to shutting down the complete network as soon as it is detected contamination. In this approach, the impact is reduced in the sensor node and further reduced in subsequent downstream nodes as the information reaches the population at the nodes.

\[
\text{Minimize } \sum_k \alpha_k \left( \sum_i \sum_p \sum_j l_{ipj} R_{ipj} \delta_s (1 - s_i) \right) + \sum_j \sum_i \sum_p \sum_j l_{sipj} R_{sipj} \delta_{ip} (1 - s_i) \tag{13}
\]

Subject to:

\[
\sum_{i \in E} s_i \leq S_{max}, s_i \in (0, 1); (i) \in E \tag{14}
\]

EPANETModelEquations(BlackBox) \tag{15}

where

- \(\alpha_a\): weight of an attack incident \(a\)
- \(\delta_p\): frequency for a flow pattern \(p\)
- \(R_{ipj}\): downstream indicator, 1 if \(j\) is downstream relative to \(i\) or 0 from a flow pattern \(p\)
- \(l_{ipj}\): Impact without sensor at \(j\) when sensor detects contaminant at \(i\) from an attack incident \(a\)
- \(l_{sipj}\): Impact with sensor at \(j\) when sensor detects contaminant at \(i\) from an attack incident \(a\)
- \(s_i\): Sensor indicator, 1 if sensor presents at location \(i\), otherwise 0
- \(E\): Set of nodes for sensor location

In a water network, direction of flow of water depends on the demand at various nodes. The flow pattern (direction of flow to and from different nodes) changes with the change in demand at the nodes. The dispersion of the contamination in the network depends on the complex relation between flow patterns and point of contamination insertion inside the network. Contamination at a point in the network can only affect downstream locations in the network. The actual impact depends on the contaminated node or nodes and downstream nodes. Since the downstream nodes relative to different nodes depends on demand pattern at the nodes, our present objective function has a downstream indicator \(R_{ipj}\) for each node. The downstream indicator is represented by a set of matrices, each of which represents different flow patterns. For the range of uncertain demand at various nodes used in our simulation, each of those downstream indicator matrices appears with different frequency \(\delta_p\) due to uncertainties. To determine the downstream indicator \(R_{ipj}\) at each flow pattern and the frequency of each flow pattern \(\delta_p\), uniform uncertain demand is used where the distribution range is wide enough so that all scenarios are considered. The objective function as given in Eq. (13) includes all aspects of water network as described above including uncertain demand at the nodes and frequency for different flow patterns along with impact and downstream indicator.

Contamination in the network is assumed to be inserted through a nodal point. The location of the node where contamination may occur is also uncertain. The uncertainty in location of contamination is incorporated through the weight of an attack incident \(a\) (\(\alpha_a\)). Since contamination may occur from any nodes, the weight of an attack incident at different nodes is considered to be equally likely. But the impact from contamination at different nodes (with and without sensor) varies according to the location of contamination.

Other than the use of uncertain demand at the nodes, the present formulation also uses a novel approach for impact estimation both with \((l_{sipj})\) and without sensor \((l_{ipj})\). For estimating impact, it does not assume that a contamination incident can be stopped immediately once a sensor detects it (Berry et al., 2005; Berry et al., 2006; Shastri & Diwekar, 2006). The total impact is measured over a certain period of time before it is assumed that it can be mitigated. For all practical purpose, due to sensor accuracy and/or delay in response, a contamination incident cannot be stopped immediately as soon as a sensor identifies it. The demand at the downstream nodes is expected to decrease with narrower distribution as we move away from the sensor node.

Unlike contamination indicator \(C_{ipj}\), as given in Eq. (1), the downstream indicator \(R_{ipj}\) in the present formulation is not affected by the presence (absence) of a sensor and only changes for different flow patterns. We have also estimated the effect of uncertainties, where the same formulation as in Eq. (13) is considered with no uncertainty in demand and contamination location. In that case a single demand is used for optimal sensor location. There will be no flow pattern variation. Thus there is only one downstream indicator \(R_{ipj}\) and the frequency of the flow pattern is one. The downstream nodes are still considered for impact estimation. The above formulation leads to a stochastic MINLP problem where the nonlinear part comes from the calculation of downstream indicator \(R_{ipj}\) as well as impact estimation with and without sensor \((l_{ipj} \text{ and } l_{sipj})\) respectively as obtained from EPANET-MSX.

### 3. Solution procedure

The problem of sensor network design under demand uncertainty forms a stochastic mixed integer nonlinear programing where the binary variables are constrained such that the total number of sensor should be less than or equals to the maximum number of specified sensor. The uncertain demand at the nodes is simulated by the generation of four hundred normally distributed sample points. At each node the demand uncertainty is assumed to be nor-
Fig. 2. Pictorial representation of stochastic optimization problem.

Optimal distribution with mean \( \mu \) and standard deviation \( \sigma \) (spread \( \mu \pm 3\sigma \), where \( \sigma = 1/3 \mu \) for the current formulation).

The problem is solved using the following five steps:

1. Specifying uncertainties in model parameters (demand at various nodes as well as attack probabilities) in terms of probability distributions.
2. Sampling the population distribution using the Hammersley sequence sampling (HSS) using initial uniform distribution.
3. Propagating the effects of uncertainties through the model (EPANET-MSX) to find the effect on output response (concentration of the contaminant at different nodes).
4. The impact is estimated from the concentration of the contaminant and population at each node.
5. A nonlinear, stochastic mixed integer problem is formulated to find the optimal position of the sensors that can minimize the impact.

The aim of an optimization problem is to calculate the value of the decision variable that optimizes the objective function within the given constraints. Stochastic optimization is a type of an optimization which deals with uncertainties. The objective function in a stochastic optimization problem is expressed in terms of some probabilistic representation (e.g., expected value, variance, fractiles, most likely values). Along with the decision variables, it also has uncertain variables or parameters. A generalized stochastic optimization problem (Diwekar, 2008) where the decision variables and uncertain parameters are separated, can then be viewed as:

\[
\text{Optimize } P_1 (x, u) \quad (16)
\]

subject to \( P_2 (h (x, u)) = 0 \) \quad (17)

\( P_3 (g (x, u) \geq 0) \geq \infty \) \quad (18)

where \( u \) is the vector of uncertain parameters and \( P \) represents the cumulative distribution functional such as the expected value, mode, variance or fractiles. In our sensor placement problems, we use the expected value of the objective function for stochastic optimization. Uncertainty in the attack probability is present in the objective function while demand probability function is embedded in the EPANET model. The probability of satisfying demand is assumed to be 1.

A generalized way of solving stochastic nonlinear programming (SNLP) or SMINLP problems is to use sampling based methods as shown in Fig. 2. A sampling loop can be embedded within the optimization model to capture the uncertainty for the decision variables as shown in Fig. 2. This can be computationally expensive as the model will have to re-run for each sampling point. General techniques for these types of optimization problems (Fig. 2) determine a statistical representation of the objective, such as maximum expected value or minimum variance. Once embedded in an optimization framework, the iterative loop structure emerges where decision variables are determined, a sample set based on these decision variables is generated, the model is evaluated for each of these sample points, and the value of the probabilistic objective and constraints are evaluated. The sheaf number of model evaluations rises significantly causing this method ineffective for even moderately complex models. We found that for our case study one model run for EPANET on an average requires 12 s. If we use it for stochastic optimization using sampling, it requires 640 min for the inner sampling loop in Fig. 2 (for each optimization iteration). Further, the black box nature of the EPANET is a big problem to use mathematical programming methods for optimization. We circumvent the problem by using the BONUS algorithm.

Better Optimization of Nonlinear Uncertain System (BONUS) algorithm was developed by Sahin and Diwekar in 2004. BONUS works in probabilistic space. In the BONUS approach, initial uniform distributions (between upper and lower bounds) are assumed for decision variables. These uniform distributions together with specified probability distributions of uncertain variables form the base distributions for analysis. BONUS samples the solution space of the objective function only at the beginning of the analysis by using the base distributions. As decision variables change, the underlying distributions for the objective function and constraints change, and the proposed algorithm estimates the objective function and constraints values based on the ratios of the probabilities for the current and the base distributions (a reweighting scheme), which are approximated using kernel density estimation techniques. For more details, refer to Diwekar and David (2015). Thus, BONUS avoids sample model runs in subsequent iterations. The model involved in BONUS algorithm can be a black box. In BONUS the derivative information required for stochastic nonlinear programming problems is also evaluated using reweighting schemes thus avoiding sample model runs for derivative calculations. In order to improve computational efficiency further, we use the efficient Hammersley sequence sampling (HSS) (Kalagnanam & Diwekar, 1997; Diwekar & Kalagnanam, 1997) for the initial base distributions.

3.1. Uncertainty propagation

For BONUS base distributions, the concentration profiles are generated for four hundred sample points with varying demands using EPANET-MSX. The contamination is used as a tracer to find the concentration at different nodes. The concentration is then used for impact assessment using the formulation of containment consumption by the population. The expected value of the impact under uncertainty is calculated using the reweighting scheme by assuming a normal distribution of demand around the mean \( \mu \) with standard deviation \( \sigma \).

When a node with a sensor detects a contaminant, it is assumed that the response from the attack incident will generate a decreased demand with less variation. It is also assumed that response from the attack will also create decreased demand at the downstream nodes from the sensor node with successively reduced mean and variance as we move further away from the sensor node. For the upstream nodes, the impact from contamination is always zero. A detail description is given in impact estimation section.

Different attack incidents are generated by changing the node of injection of the contaminant. The stochastic nature of the problem leads to different patterns. The upstream and downstream nodes are different for different patterns. For each possible combination, the impact at each node for different demand distribution is estimated from the value of model output response (contaminant) at
each node (mg/L) multiplied by the population at that node (person) and average volume of water consumed per person (L/person). This will give the impact in terms of mass of contaminant consumed (mg). A stochastic mixed integer nonlinear programming problem is formulated to find the optimal position of the sensors that can minimize the impact resulting from different attack incidents in the entire network. At each iteration, a reweighting scheme is used according to the BONUS algorithm to generate the expected value of impact at the nodes. The final outcome is the position of the sensors for minimizing the net impact on the network. The sensor placement problem is discrete but sensor placement is manifested into different demand distributions at the downstream nodes. The problem remains SMINLP but mathematical programming methods with derivative calculations can be applied because of the BONUS reweighting scheme.

3.2. Reweighting scheme for BONUS

The BONUS algorithm (Sahin & Diwekar, 2004; Diwekar & David, 2015) uses the reweighting approach, based on one of the various reweighting schemes proposed by Hesterberg (1995). In BONUS algorithm, samples taken from a uniformly distributed sample space is compared to a new reference distribution to create a set of distribution weights. These weights are then used to reweight the output distribution function. This reweighting approach is beneficial for eliminating the need to simulate the model at the new reference distribution as shown in Fig. 3. Using the reweighting scheme, the EPANET-MSX simulation model only needs to compute the water network process behavior for the set of initial uniform distribution only. For different combination of potential sensor networks at different optimization iterations, the reweighting approach will generate the output distribution function thereby significantly reducing the overall computational time. The reweighting scheme is used to generate the output distribution function of the impact at various nodes. This approach has been extensively used for various stochastic optimization problems in literature (Sahin & Diwekar 2004; Salazar et al., 2011; Lee & Diwekar 2012) and details of the approach are provided elsewhere (Diwekar & David, 2015).

Fig. 3 illustrates the reweighting scheme. On the first iteration, a set of N sample points uniformly distributed across a d dimensional sample space are used to perform N simulation replications of the water network (i.e., at various demands at the nodes). Let \( f_0(x_i) \) be the probability density function associated with the base input distribution for the input variable (in the present problem demand variable at the nodes) \( x_i = 1, 2, \ldots, \text{SN} \). Following the simulation of the water network at iteration \( t = 0 \), let \( f_0(y_i) \) be the base probability distribution function associated with the output variable (which in the present case is the impact at the nodes obtained from contamination concentration) \( y_i = 1, 2, \ldots, \text{SN} \), respectively, where \( y_i = h(x_1; x_2; \ldots; x_{\text{SN}}) \) is the nonlinear transformation from each demand variable, \( x_i \), to the concentration profile \( y_i \).

BONUS reweighting scheme is used to find effect of different demand distributions. With a new input distribution such as the response from an attack incident at a sensor node or a node located at downstream from the sensor node. The redefined distribution, \( f_t(x_i) \), at iteration \( t \) is used to create a set of weights \( W_t(x_i) = f_t(x_i)/f_0(x_i), i = 1, 2, \ldots, \text{SN} \) (19) which gives the likelihood ratio between the redefined and base distributions. These weights along with the known output distribution for the base sample set are used to approximate the probabilistic behavior of the new output distribution as illustrated below.

Given the input variables act independently, these weights are used to construct the resulting distribution for the downstream intermediate or output variables at iteration \( t \) by multiplying the associated weights \( W_t(x_i) \) with the base distribution \( f_0(y_i) \).

\[
 f_t(y_j) = f_0(y_j) \prod_{i=1}^{\text{SN}} (1 + \gamma(y_j) W_t(x_i) + 1), j = 1, 2, \ldots, \text{SN} \) (20)

where \( \gamma(y_j) = 1 \) if variable \( y_j \) is (not) downstream of \( x_i \).

The reweighting scheme has helped to eliminate the need to regenerate output distribution at a new set of \( N \) sample input points through simulation of the EPANET at each iteration \( t \). Various underlying distributions corresponding to downstream positions relative to the sensor node are evaluated using the reweighting scheme. Thus the BONUS reweighting algorithm provides an efficient method for calculating the impact resulting from several reconfigurations of a network of sensors.

3.3. Impact estimation

In order to account for the uncertain demand, samples are generated for each node with a nonzero demand. Initially a node is selected and methyl chloride (\( \text{CH}_3\text{Cl} \)) is inserted. Methyl chloride will flow downstream from the node where it has been injected and will appear at the downstream nodes. The concentration profile of methyl chloride as a function of time at all the nodes is simulated using EPANET-MSX. This process is repeated by changing the node for contamination injection until all the nodes with nonzero demands are used as source of contamination.

Due to the mitigation strategy, the location of a sensor will generate a demand distribution function different from that without a sensor. We use the mitigation strategy for all downstream nodes from the sensor node where there will be a separate distribution depending on the relative distance from the sensor node. Initially, at the four hundred uniformly distributed sample points, the impacts at the nodes are calculated as a base case. The base case is then used to estimate the impact as sensor nodes, downstream nodes and nodes without sensor by the reweighting scheme.

The normal demand distribution at the sensor node and downstream nodes are generated with \( \mu \pm 3\sigma \) spread. When the mitigation strategy is considered but demand uncertainty is not considered, a constant demand is taken in the sensor node and downstream nodes with no variance. We assumed demand to be proportional to population, the proportionality constant is arbitrarily chosen.

The impact from contamination at a node is calculated from the product of the concentration, population at that node and average volume consumed per person. The reweighting approach is used to find the expected value of impact at all of the nodes for different demand distributions. The expected impact at each node: one without sensor, one with sensor and several downstream nodes with respect to a reference sensor node is used for optimal sensor placement using a stochastic mixed integer nonlinear programing that minimizes the net impact from contamination. The upstream and downstream of the nodes are determined by the flow patterns. For each flow pattern, there is a binary matrix representing the upstream and downstream of each node. This matrix is called the downstream indicator matrix. Similarly, a distance indicator matrix represents the position of each node relative to the sensor node. In the present case, the distance indicator matrix only identifies the sequence of the downstream nodes. The mixed integer programing uses the downstream indicator matrix and distance matrix to find appropriate impact from the downstream for optimal sensor location. Example of downstream indicator matrix and distance indicator matrix is shown in results and discussions section.
Since from a stochastic algorithm any node can be a sensor node, the relative position of all other nodes with respect to the sensor node is identified and the impacts on the nodes are calculated accordingly. The outcome of the mixed integer program selects those nodes for sensor location where the net impact on the population is minimum. By eliminating the need to regenerate a new set of N sample points through simulation of the water network, the BONUS reweighting algorithm provides an efficient method for calculating the impact resulting from several reconstructions of a network of sensors. A schematic diagram representing the steps involved in the process of finding optimal sensor location is shown in Fig. 4.

The present approach of decreased demand with time and distance is based on the assumption that the impact of contamination cannot be stopped immediately but it will occur over a period of time. Davis and Janke (2008) have shown that if the contaminant concentration varies over a period of time, the actual time of exposure is important for proper impact assessment. Thus, the probabilistic model for the timing of consumption of tap water developed by Davis and Janke (2009) is useful for health impact assessment with the present assumption of impact mitigation. The present model assumes a fixed set of probabilistic distributions for the downstream nodes. We will use an approach based on the par-
ticular timing of exposure as suggested by Davis and Janke (2009) for impact assessment from the downstream nodes in future.

4. Case study

The sample water network system used in the present problem is shown in Fig. 1. In the present water network system Node 10 is considered to have zero demand and eight other nodes have nonzero uncertain demands. Uncertain demand samples are generated for all the eight nodes with nonzero demands. The attack probability is considered to be fixed and equal at all the nodes with nonzero demand. The contaminant is inserted from one of the nodes and the concentration profile of the contaminant at various nodes is observed over a period of 48 h. The injection location is varied to include all eight nodes that have nonzero demand and subsequent impact is estimated. In order to account for the uncertain demand, four hundred samples are generated for each node with a nonzero demand. The population at the nodes are allowed to vary with a uniform distribution from zero to the maximum population possible.

Initially a node is chosen and methyl chloride (CH3Cl) with a concentration of 20000 mg/minute is inserted for four hours. The concentration profile with time at all the nodes is simulated for all four hundred sample cases. For all the nodes, the concentration reaches a steady state after few hours. The constant concentration at different nodes is used for impact estimation at the sample points. This process is repeated by changing the node for contamination injection until all the eight nodes with nonzero demands are used. Thus we have (400 × 8) sample points where each sample point gives concentration at different nodes for our analysis.

For the given network (Fig. 1), seven different demand distributions are expected at each node: one for without sensor, one with sensor and five for acting as subsequent downstream nodes of the sensor node depending on the relative position of the nodes with respect to the sensor node. Population distribution at various conditions at node 11 with uncertainty is shown in Fig. 5 as an example. We assumed demand to be proportional to population. Thus Fig. 5 represents the demand distribution at the node. Seven values of expected impact at each node is estimated for seven different demand distributions and used in optimal sensor placement.

We need to generate the base case scenario, which is the uniform demand distribution. For all other cases BONUS can be used. 400 sample points are taken from the base case uniform demand distributions. The 400 samples from uniform distribution is used for each attack incident and every sample point gives the impact at the nodes. As the injection is repeated at 8 nodes, 400 × 8 samples are used for impact estimation generated using EPANET-MSX.

To verify the validity of the reweighting approach, the EPANET-MSX model was first simulated using the uniform distribution across each input variables. The initial uniform distribution has to encompass all demands. Then, the model was simulated for concentration using the normal distributions at each of the input variables with standard deviation according to Table 1. Simultaneously, BONUS reweighting scheme is used for estimating concentration using a normal distribution. No significant difference in the calculation of the concentration at the nodes resulted between the two simulation approaches. This occurred primarily because (a) there were sufficient number of sample points taken (N = 400), (b) the Hammersley sampling technique covers the sample space uniformly, and (c) the reweighting approach only undergoes one iteration when computing the concentration for a given set of input variable distributions.

5. Results and discussions

The simulation is performed with four different methods;

- Cost optimization using objective as shown in Eq. (1). Performed Stochastic optimization with uncertain demand. (A)
- Using TEVA SPOT that performs impact minimization. It uses objective as shown in Eq. (2) and does not use demand uncertainty or consider downstream nodes for analysis. It takes into account contamination location uncertainty with different attack incident. (B)
- Deterministic optimization with impact minimization using objective as shown in Eq. (3) implementing impact on downstream nodes but without uncertain contamination location (single attack incident a with α = 1) and demand (single Rdp with frequency one) (C)
- Stochastic optimization with impact minimization using objective as shown in Eq. (3) implementing impact of downstream nodes and demand uncertainty (multiple Rdp) and contamination location uncertainty with multiple attack incident. (D)

The analysis is performed on the water network shown in Fig. 1. It is a modified version of the first example network from EPANET. It is assumed that the location of the sensor will redistribute the demand pattern at the node and subsequent downstream nodes due to the response from a contamination incident. The demand uncertainty is applied to all eight nodes having a nonzero demand. This is because the sensor can be located in any one of the nodes where the demand is nonzero and location of sensor will change the distribution at that node and subsequent downstream nodes. Demand uncertainty at the nodes can generate 9 possible flow patterns (scenario) in a modified EPA network 1 as found from using a normal demand distribution at different nodes with population mean μ and standard deviation σ. This distribution property is similar to that found without a sensor at the node (see Table 1). Each scenario will determine the upstream and downstream for each node. Through sampling (400 samples), the frequency of each scenario is generated using EPANET.

The matrix representing the position of each and every node relative to the sensor node is shown in Table 2. Here each row represents the sensor node and each column the relative position of other nodes relative to the sensor node. The relative position of the nodes relative to the sensor node as given in Table 2 defines the properties of demand distribution at that node as given in Table 1. For example node 3 is at a relative position of 5 from node 1. Node 3 will have a pattern defined by n in Table 1.

Contaminants are inserted from different nodes to create multiple attack scenarios. This will result in various concentration level at the nodes from different attack incidents. With the help of relative position obtained from Table 2 and the corresponding distribution from Table 1, the expected values of contaminants at the sensor and subsequent downstream nodes are obtained using the reweighting approach. This process is repeated for different attack incidents. For a given flow pattern, a downstream indicator (Rdp) will determine if an attack at a particular node can contaminate another node in the network. The downstream indicator depends on the flow pattern, which again depends on the demand scenario at different nodes. Downstream indicator is represented by the binary matrix where the upstream of a node is represented by 0 and downstream of a node by 1. Nine different matrices are used to represent downstream indicators as obtained from nine possible flow patterns of the example network in the range of uncertain demand distribution. An example matrix representing the downstream indicator is shown in Table 3. Impact from the sensor node and subsequent downstream nodes is used for net impact estimation until a downstream node is not a sensor node by itself. The
optimal sensor position is obtained by minimizing the net impact from different attack incidents each of which is estimated from an ensemble average over different flow patterns possible in the given water network.

The stochastic programing is performed with varying the maximum number of sensors in the network. The optimal position of four sensors as obtained from different methods (A, B, C and D) is shown in Fig. 6. Method A is used by Shastri and Diwekar (2006) and method B is used in TEVA SPOT (Berry et al., 2008). Method A uses cost optimization under demand uncertainty. For cost optimization, the cost per affected person is assumed to be $1,80,000 and the cost of sensor is assumed to be $15,00,000. The sensor costs and cost per affected person are decided so as to show a tradeoff between the two parts of the objective function in Eq. (1). This is

### Table 2
Position of a node relative to the sensor node.

<table>
<thead>
<tr>
<th>Sensor node</th>
<th>Relative position</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>12</td>
<td>3</td>
</tr>
<tr>
<td>13</td>
<td>4</td>
</tr>
<tr>
<td>21</td>
<td>3</td>
</tr>
<tr>
<td>22</td>
<td>4</td>
</tr>
<tr>
<td>23</td>
<td>5</td>
</tr>
<tr>
<td>31</td>
<td>3</td>
</tr>
<tr>
<td>32</td>
<td>5</td>
</tr>
</tbody>
</table>

### Table 3
Matrix representing upstream (by 0) and downstream (by 1) nodes for a given flow pattern.

<table>
<thead>
<tr>
<th>Sensor node</th>
<th>Upstream/ Downstream node</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
</tr>
<tr>
<td>21</td>
<td>1</td>
</tr>
<tr>
<td>22</td>
<td>0</td>
</tr>
<tr>
<td>23</td>
<td>0</td>
</tr>
<tr>
<td>31</td>
<td>1</td>
</tr>
<tr>
<td>32</td>
<td>1</td>
</tr>
</tbody>
</table>
following the method shown in Shastri and Diwekar (2006). The cost of health is a function of number of people affected by contamination. The result shows that method A tries to minimize the cost from health hazard by locating the sensors around the nodes that covers maximum fraction of the population.

Optimal sensor position using TEVA SPOT (Method B) tries to locate the sensor at the nodes where the impact on population is maximal. Impacts on the downstream nodes are not considered for assessment. Demand uncertainty is also not considered in TEVA SPOT though uncertainty in the form of contaminant location is incorporated through different attack incidents. Method C uses impact assessment from sensor nodes and downstream nodes for analysis without considering any uncertainty. Method D is the focus of this paper and uses impact assessment from sensor nodes and downstream nodes as well as demand and contaminant location uncertainty for analysis. Thus we have method A with demand uncertainty, method B with contaminant location uncertainty, method C without uncertainty and method D with both contaminant location as well as demand uncertainty. For method D with population uncertainty, the population mean and standard deviation at node 11 used for analysis is shown in Fig. 5. Similar distributions from other nodes are used for our analysis.

Method A uses the link between two nodes for sensor location. For all other methods, the sensors are located at the nodes. In order to compare the result from method A with other methods, we have chosen one of the nodes connected by the link for sensor location that has either a nonzero demand or is an optimal node as chosen by other methods. From our analysis we have found that in some of the cases, two different methods have chosen the same nodes. But the sequences in which the nodes are chosen are different in different methods. This is evident as we increase the number of nodes. The choice of nodes for optimal sensor placement as obtained from different methods is shown on Table 4. Here the maximum number of allowable sensor is changed from one to four.

Results from the present formulation are significantly different from that of either method A, B (TEVA SPOT) or C. Comparing method D with either A, B or C, we can say that consideration of impacts from downstream nodes as well as uncertainty affect the choice of sensor nodes. Comparing method C and D as seen in Table 4 we can say that uncertainty has affected choice of sensor location. Out of eight nodes to choose from, there is no node that has been chosen by all the four methods. When we have to choose optimal position for one sensor, method B has chosen 31 while method D has chosen 32. With two sensors nodes, method B has chosen 31 and 32 while method D has chosen 13 and 32. With three sensors onwards, their choice is entirely similar. This could be due to the fact that method B has included location uncertainty in their impact assessment. Thus if the number of sensor increases, then we can witness similar selection but when few sensors has to be optimally placed, a difference can be found from using demand uncertainty at the nodes.

Impacts from sensor location as obtained from different methods are calculated with the present formulation as in Eq. (3). Considering the fact that the present formulation has used four important factors, namely, downstream impact, no complete mitigation immediately after contamination detection, and demand uncertainty at the nodes as well as contamination location uncertainty, results from method D are considered to be most optimum. Choice of any other nodes will give a suboptimal result. This is shown in Fig. 7. Our goal is to minimize the net impact after sensor placement. It is seen that the net impact after sensor location is minimum from present formulation using method D for first few sensors. The nodes chosen from method A (cost minimization) shows maximum impact for any number of sensors when compared with those from TEVA SPOT (B) or present formulation (C or D). Nodes chosen by method C shows higher impact with any number of sensors compared to that from TEVA SPOT and present formulation (D). This shows that consideration of uncertainty is very important in optimal sensor location. In the present network we have found that the impacts from B and D are similar for three sensor nodes onwards. For two or less sensor nodes, choice of nodes with the present formulation is the best as it reduces maximum impacts from contamination.

Just like the present formulation, optimal sensor position using TEVA SPOT also tries to minimize the impact on population. But their analysis does not take into account the uncertain demand distribution or downstream consideration. Results from TEVA SPOT are based on the assumption that an alarm of contamination will immediately stop any further consumption. The present problem is trying to minimize the impact of contamination by positioning the sensors at those nodes whereby maximum mitigation possible in the network. Thus the present simulation holds a pragmatic approach where the demand is expected to decrease with a narrower distribution from an alarm of contamination.

In order to check the effect of uncertainties, method C is used where there is no uncertainty in demand or location. It is evident from the result that the incorporation of downstream effect may not make a significant change since method C shows more impact than B or D. On the other hand incorporation of uncertainty seems to have most significant effect. Method D that incorporates demand
uncertainty, location uncertainty as well as downstream effect is better than all other methods. Thus, inclusion of uncertainty, especially contamination location uncertainty will show performance better than A. C. For best performance at least with few sensors, inclusion of demand uncertainty is also essential. Thus we can conclude that inclusion of downstream effect as well as different form of uncertainty is necessary for optimal sensor placement as all other methods will always be suboptimal. Table 5 presents the efficiency of BONUS to solve this problem for 1–4 sensors.

<table>
<thead>
<tr>
<th>Number of sensors</th>
<th>(A)</th>
<th>(B)</th>
<th>(C)</th>
<th>(D)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11</td>
<td>31</td>
<td>22</td>
<td>32</td>
</tr>
<tr>
<td>2</td>
<td>11, 22</td>
<td>31, 32</td>
<td>22, 32</td>
<td>13, 32</td>
</tr>
<tr>
<td>3</td>
<td>11, 22, 31</td>
<td>13, 31, 32</td>
<td>22, 23, 32</td>
<td>13, 31, 32</td>
</tr>
<tr>
<td>4</td>
<td>11, 12, 22, 31</td>
<td>13, 23, 31, 32</td>
<td>12, 22, 23, 32</td>
<td>13, 23, 31, 32</td>
</tr>
</tbody>
</table>

Table 5

Efficiency of BONUS algorithm.

<table>
<thead>
<tr>
<th>No. of Sensors</th>
<th>Optimization Iterations</th>
<th>EPANET runs without BONUS</th>
<th>EPANET runs with BONUS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15</td>
<td>48000</td>
<td>3200</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>16000</td>
<td>3200</td>
</tr>
<tr>
<td>3</td>
<td>15</td>
<td>48000</td>
<td>67200</td>
</tr>
<tr>
<td>4</td>
<td>21</td>
<td>Total</td>
<td>179200</td>
</tr>
<tr>
<td>% Savings</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

6. Conclusions

This paper presents a method for impact assessment from contamination incidents about different demand scenario using a reweighting scheme for BONUS. The optimization problem is formulated as a stochastic mixed integer nonlinear programming problem, where the objective is to minimize the overall impact from different attack scenario using sensors, subject to a given budget constraint. The placement of each sensor is determined by solving the optimization problem. The advantage of our approach based on impact assessment in solving the sensor placement problem is the use of nonlinear demand distribution in impact estimation. Further, we used information based mitigation strategy if and when the sensor detects the contamination.

The present formulation accounted for uncertain location of contamination injection in the network and uncertain demand distribution in case of an attack. There are other uncertainties associated within the network. These include type of contamination, propagation rates of contamination, pipe conditions, and so on. However, the optimization is based on the present formulation. The demand uncertainty will lead to the variation in flow pattern and the uncertainty in location and concentration of contaminant will lead to various concentrations of contaminants at different nodes in the network. In our research, we have incorporated these uncertainties to make the optimal sensor placement more robust. Generally a stochastic non-linear programming method calculates the probabilistic objective function by repeated evaluation for each sample at every iteration. We have avoided this computational burden by using the reweighting scheme based BONUS algorithm. The hypothetical condition of stopping the network completely from any further consumption once an attack has been detected is avoided. Future research can be extended to include various other objective functions, larger and more realistic networks, and modifying the mitigation strategy further to include information flow to people.

Acknowledgements

One of the authors, Prof. Ashok Vaseashta, would like to acknowledge support from NATO- Science for Peace and Security program 984403 “Technical Advances to Detect and Remove Contaminants in Water for Safety and Security”.

References


