


# Optimal water quality sensor positioning in urban drainage systems for illicit intrusion identification

Mariacrocetta Sambito, Cristiana Di Cristo, Gabriele Freni  and Angelo Leopardi

## ABSTRACT

In the last decade, the growth of the micro-industry in urban areas has produced an increase in the frequency of xenobiotic polluting discharges in drainage systems. Wastewater treatment plants are usually characterized by low removal efficiencies in respect of such pollutants, which may have an acute or cumulative impact on environmental and public health. To facilitate the early isolation of illicit intrusions, this study aims to develop an approach for positioning water quality sensors based on the Bayesian decision network (BDN). The analysis is focused on soluble conservative pollutants, such as metals. The proposed methodology incorporates several sources of information, including network topology, flows and non-formal 'grey' information about the possible locations of contamination sources. The methodology is tested using two sewer systems with increasing complexity: a literature scheme from the Storm Water Management Model (SWMM) manual and a real combined sewer in Italy. In both cases, the approach identifies the optimal sensor location gaining advantage from additional information, which reduces the computational effort needed to obtain the solution. In the real case, the application of the method yielded a better solution with regards to the real position of the implemented sensor network.

**Key words** | optimization methods, sensor location, uncertainty analysis, urban drainage, wastewater quality

**Mariacrocetta Sambito**  
**Gabriele Freni**  (corresponding author)  
School of Engineering and Architecture,  
University of Enna 'Kore', Cittadella Universitaria,  
94100 Enna,  
Italy  
E-mail: [gabriele.freni@unikore.it](mailto:gabriele.freni@unikore.it)

**Cristiana Di Cristo**  
Department of Civil, Architectural and  
Environmental Engineering,  
University of Naples Federico II,  
80125, Napoli,  
Italy

**Angelo Leopardi**  
Department of Civil and Mechanical Engineering,  
University of Cassino and Southern Lazio,  
Cassino 03043,  
Italy

## INTRODUCTION

### Wastewater quality and monitoring needs

In both water distribution systems and sewer systems, the monitoring of water quality is very important for preserving resources and public health. Monitoring physical, chemical and biological parameters increases the possibility of early detection of water quality deterioration and individuation of pollution sources. The quality of wastewater impacts the proper functioning of a sewer system and a wastewater treatment plant (WWTP) and the receiving water body in the case of combined sewer overflow (CSO) activation (Even *et al.* 2004). CSOs, which contain untreated domestic and industrial waste, toxic materials and debris, impact the physicochemical, biological, hydraulic and aesthetic status of

receiving water bodies. For example, overflows can cause oxygen depletion, increased turbidity and higher concentrations of micropollutants, heavy metals and pathogenic and faecal organisms in surface waters (Passerat *et al.* 2011). Xenobiotic substances, unlike organic substances, are only slightly affected by biological degradation processes. Metals often show a remarkable tendency for bio-accumulation and are unaffected by wastewater treatment and/or dangerous for common plant technologies.

Since the adoption of the Water Framework Directive 2000/60/EC, Member States in EU countries must apply local measures to address the pollution that affects their surface waters; thus, decreasing the occurrence of overflows and improving discharged water quality are important

parts of pollution-reducing strategies. Models for the characterization of the wastewater quality have been extensively investigated to assess the pollution load that is overflowed and/or transferred to WWTP. Boenne *et al.* (2014) showed that the contribution of WWTP overflows due to just two events of a few hours can produce an increase up to 22% of the measured nutrient load in a river. Jiang *et al.* (2018) asserted that the identification of polluting sources after river spill is critical to improving decision-making about the emergency response.

For these reasons, the implementation of a monitoring network is crucial for an efficient contamination prevention strategy in urban drainage systems, which involves the identification and elimination of illicit polluting discharges.

### Monitoring systems and polluting sources identification

The problem of polluting source identification was primarily investigated for pressurized distribution networks as a drinking water contamination event determines an immediate alarm for public health (Di Cristo & Leopardi 2008; Lifshitz & Ostfeld 2019). An abnormal polluting discharge into a sewer system usually has a lower impact on the general public even if it has a relevant impact in the environment.

The collection and the analysis of real data are indispensable to control wastewater quality and to identify the origin of the pollution. Montserrat *et al.* (2015) developed a methodology to evaluate the performance of combined sewer systems (CSSs) using low-cost monitoring to reduce the number and the impact of overflows. Using the measurements of various quality parameters, Srinivas *et al.* (2018) developed an approach for identifying the major sources of pollution in rivers using advanced hierarchical clusters and multivariate statistical analysis.

The development of specific sensors (Qin *et al.* 2012) facilitated online and real-time measurement of wastewater quality. Boenne *et al.* (2014) employed online high-frequency continuous measurements to assess the impact of wastewater quality downstream from a treatment plant on the receiving water body. They demonstrated that continuous *in situ* monitoring can furnish important information about pollution sources, even if the cost of the monitoring setup is higher than that of traditional sampling. Troutman *et al.* (2017) presented a data-driven identification/learning

toolchain to manage a large number of measurements for dynamic modelling and prediction of a combined sewer functioning. They identify a near-optimal time record for which measurements must be available to ensure an acceptable forecasting performance.

Regarding sewers, illicit discharges can easily enter systems via intentional or accidental dumping or spills as the networks are geographically dispersed and have multiple access points. For this reason, many countries have implemented regulations and projects to support actions for illicit discharge individuation in sewer systems (Irvine *et al.* 2011). Recently, Banik *et al.* (2017a) proposed a methodology for identifying an illicit intrusion in a sanitary or combined sewer system using online pollutant concentration measurements.

### Optimization of sensor location

The approaches proposed for illicit discharge individuation require the deployment of sensors. However, installing and operating measurement devices in sewers is expensive and limited by many constraints. The installation and maintenance costs can be reduced by optimizing the position of the sensors while simultaneously obtaining a reliable and inexpensive monitoring infrastructure. The sensor placement problem has been extensively investigated to design contamination warning systems in drinking water distribution networks (Rathi & Gupta 2016) and to monitor rivers (Lee *et al.* 2014).

Few studies address the sampling design in sewer systems. Kim *et al.* (2013) aimed their study at developing a decision-support model for identifying the location of the pathogenic intrusion in a real gravity sewer system as a means of facilitating rapid isolation and efficient containment using artificial neural networks (ANNs). The results showed that ANNs identified the location of the injection sites with 57% accuracy. Increasing the number of available sensors within the basin significantly improved the accuracy of the simulation results (from 57% to 100%). Recently, Banik *et al.* (2017b) developed and compared different multi- and single-objective optimization procedures to optimally locate sensors to detect illicit intrusion in sewer systems with objective functions expressed by the parameters entropy, detection time and reliability. The results show that the obtained sensor displacement in all cases is

more efficient than the existing displacement established without any optimization. In [Banik \*et al.\* \(2017c\)](#), the optimal placement of wastewater monitoring sensors that was formulated as a single objective optimization problem is solved using greedy algorithms. The results indicate the robustness of the methodology with respect to the detection of contaminants, the excellent performance of the single fitness function and the better efficiency of the greedy algorithm with respect to the genetic algorithm.

Many of the literature sensor location procedures assume that any node of a network can be a source with equal probability. [Weickgenannt \*et al.\* \(2010\)](#) presented an importance-based sampling method for selecting dangerous scenarios for a sensor location problem in a water distribution system. In this study, contaminant inputs in nodes located in highly populated areas are considered to be 'more important scenarios'. [Tinelli \*et al.\* \(2017\)](#) proposed a procedure based on practical considerations of network topology and operations for sampling the most representative contamination events in the sensor location problem of water distribution systems. The results indicate that the optimal sensor placement does not vary when only the selected sampled events are considered.

[Yazdi \(2018\)](#) proposed a methodology based on entropy theory and employed the differential evolution algorithm for identifying the best monitoring locations for detecting wastewater quality changes in sewers. The results indicate that the method improves the level of information with a limited number of sensors.

[Vonach \*et al.\* \(2018\)](#) presented a heuristic method for measurement site selection in a sewer system to obtain an efficient calibration of the hydrodynamic model.

### Bayesian decision networks

When optimal solutions are needed in an uncertain system, Bayesian approaches can be useful to make decisions (in this case for identifying the best sensor network) and assimilate information from the system in an upgradable and updatable way (in this case, gaining information from numerical simulations and data from the real system).

Bayesianism is the philosophy that asserts that to understand human problems while constrained by ignorance and uncertainty, the probability calculus is the single most

important tool for representing appropriate strengths of belief. The probability calculus enables us to represent the interdependencies that other systems require and enables the representation of any dependencies.

A Bayesian decision network (BDN), or Bayesian network, is an acyclic graphical structure that enables us to represent an uncertain domain and the conditional dependencies between independent variables and dependent variables in a probabilistic way. Bayesian networks are ideal for considering an event that occurred and predicting the likelihood that any one of several possible known causes was the contributing factor. The nodes represent a set of random variables from the domain ( $X = X_1, \dots, X_n$ ), while a set of directed arcs connects the pairs of nodes ( $X_i \rightarrow X_j$ ) for representing the direct dependencies among the variables. At least three distinct forms of uncertainty, with which an intelligent system that operates in the real world shall need to cope, exist: ignorance the limits of our knowledge, which cause us to be uncertain about many things; physical randomness or indeterminism and vagueness.

The BDN is a very robust and particularly useful method for assessing risk and uncertainty that provides a complete framework for analysing all cause and effect relationships ([Korb & Nicholson 2010](#)).

Few applications of BDN, which are related to urban water systems, are included in recent technical literature. These applications are aimed at guiding technical choices in uncertain domains to incorporate different sources of information into the decision. [Phan \*et al.\* \(2016\)](#) presented some applications of BDN to water resources with respect to spatial factors, water domains and the consideration of climate change impacts to guide management decisions.

[Kabir \*et al.\* \(2015\)](#) applied BDN to identify water system main failures considering the vulnerability and sensitivity of the system to failure and the global risk. In this case, collected data about failure were progressively incorporated in the Bayesian method to improve the selection.

[Freni & Sambito \(2017\)](#) proposed a probabilistic approach to the positioning of water quality sensors in urban drainage networks for identifying an illicit intrusion, which shows the progressive increase in the identification probability obtained by the Bayesian approach. In this work, the implementation of the pre-conditioning approach proposed by [Banik \*et al.\* \(2015\)](#), essentially depending on the

network topology, produced an improvement in term of computational efforts.

### Aim of the research

This study presents a methodology for solving a sensor location problem and individuating the source of an illicit intrusion in a sewer system. It is aimed at solving a sensor location problem, in which the positioning of fixed-type sensors is assumed. The analysis is performed with the hypothesis that each node of the network has the same probability of being the polluting source. Successively, the hypothesis that some nodes can be more frequently polluted than others is introduced. The different probabilities of a node of the contamination source are derived by the knowledge of the system topology, flows and possible polluting activities based on the grey information about the served area in terms of commerce and industry data. In this application, the capacity for incorporating all available information, to individuate the more risky scenarios, represents the main original aspect of the proposed methodology. Contrary to previous studies, the proposed approach considers the inclusion in the network of several sensors, taking into account the interaction and correlation among their responses. The preliminary analysis presented in Sambito *et al.* (2018) is completed.

The methodology is applied to two different networks with increasing size and complexity: the literature network Example 8 of the Storm Water Management Model (SWMM) application manual (Gironás *et al.* 2009), in which the contamination was analysed in wet weather conditions, and the real test case represented by the sewer system of Massa Lubrense (Italy), which was analysed in dry weather conditions. The paper is organized as follows. First, the sensor location problem formulation is presented, and the Bayesian approach that is employed to solve the problem is described. Second, the test cases are discussed, and the results are presented. Last, some conclusions are formed.

## MATERIALS AND METHODS

In the proposed methodology, the sensor location problem is solved using a Bayesian approach. The new information

from the analysis enables the operator to gain insight into the system once new contamination events are detected and identified. In this way, the approach is suitable for solving problems, in which data are initially collected and the operator plans to improve the monitoring strategy.

To solve the sensor location problem, two main components are required: a calibrated model for hydraulic and water quality simulations in sewer systems and a Bayesian solver for likelihood estimation and probability updating.

### Numerical simulation model

The EPA SWMM (5.022 version) was employed to simulate the urban drainage network and the propagation of contaminants in a sewer system. This model enables the user to select different mathematical models to describe the runoff formation and propagation in sewer systems (Gironás *et al.* 2009). The complete 1D Saint-Venant equations were applied to simulate the flow propagation into a sewer system by adopting an iterative explicit mathematical solver.

Water quality routing within conduit links assumes that the conduit behaves as a continuously stirred tank reactor (CSTR). Although a plug flow reactor assumption may be more realistic, the differences will be small if the travel time through the conduit is on the same order as the routing time step. The concentration of a constituent that exits the conduit at the end of a time step is obtained by integrating the conservation of mass equation and using average values for quantities that may change over the time step, such as flow rate and conduit volume. The quality of the water that exits the node is the mixture concentration of all water that enters the node. Water quality modelling within storage unit nodes and manholes follows the approach used for conduits.

The considered contaminant is assumed to be a soluble conservative xenobiotic, such as some heavy metals or soluble ionic compounds. This hypothesis was introduced as the intrusion of a conservative pollutant represents a more dangerous scenario.

The proposed sensor location approach has to be tested as polluting events occur. Considering that the position, magnitude and duration of contamination can be uncertain, each model application is given by a random simulation in which the contamination parameters are randomly set up

in terms of the contaminant mass, contamination duration and contamination node. The contaminant mass is randomly set between 0.01 kg and 0.5 kg; the contamination duration is randomly set between 0.25 hour and 3 hours. As mentioned in the Introduction, in Freni & Sambito (2017), all network nodes were considered to have the same probability to host the contamination event, while different probabilities were established in this analysis based on the information about the system and the served area.

Each sensor network configuration was investigated by 1,000 random contamination events, and its efficiency was evaluated by the uncertainty and isolation likelihood  $D$ , i.e., the probability of the sensor network to detect the presence and the origin of the contamination. The isolation likelihood is evaluated as the ratio between the number of events in which the network sensor was able to locate the contamination node and the total number of tested contamination events. The uncertainty is summarized by the probability that the sensor network is able to detect the contamination but unable to locate the source node. In the analysis, the likelihood and the reliability/uncertainty functions were slightly adapted from those presented in Preis & Ostfeld (2008) to comply with sewer networks instead of water distribution networks. According to Preis & Ostfeld (2008), the isolation likelihood  $F_1$  and detection reliability or redundancy  $F_2$  are expressed by the following equations:

$$F_1 = \frac{1}{S} \sum_{i=1}^S d_r \quad (1)$$

$$F_2 = \frac{1}{\sum_{i=1}^S d_r} \sum_{i=1}^S R_r \quad (2)$$

where  $S$  is the total number of analysed contamination events;  $d_r$  is 1 if the contamination was identified by the sensor network and is 0 otherwise; and  $R_r$  is 1 if the contamination was detected by at least two sensors and is equal to 0 otherwise. The indicator  $F_1$  (Equation (1)) provides information on the ability of the sensors' network to locate the contamination source, while  $F_2$  (Equation (2)) indicates the reliability of the sensor network (more than one sensor) in detecting an event. If the contamination is not confirmed by more than one sensor in the system, false positives may be present.

### Bayesian network approach for sensor location

As discussed in the Introduction, the BDN is used to guide decisions in an uncertain domain progressively incorporating information in the process. In this framework, in the present study:

- the upstream (independent) nodes of the BDN are related to contamination factors (position, magnitude, duration and starting time) and external factors, such as network characteristics, dry weather flows and wet weather flows;
- the intermediate nodes are related to the distribution of the contaminant concentrations and are connected to the upstream nodes by probabilistic arches depending on the model results, which are subjected to uncertainty;
- the downstream nodes are related to the likelihood of sewer manholes to be a suitable location for sensors alone and in combination with others and are connected to the intermediate nodes by probabilistic arches that express the detectability of the contamination and the reliability of the sensors.

Bayesian approaches start with the formulation of prior knowledge in terms of the probability of events to be representative of the truth (in this case, the probability that a sensor or group of sensors to be correctly located to identify the source of contamination). The system is solicited and investigated to obtain a series of events that can confirm or deny prior assumptions (in this case, a set of simulated events of contamination in which sensors were in place; each event was based on the probability distribution of the contamination in nodes). The number of events considered to be sufficient to verify and update prior knowledge (population of the update) is a parameter in the Bayesian approach. After this number is attained, the posterior probability is calculated by incorporating new information from the series of events in prior knowledge by the application of Bayes' theorem. The number of updates for which additional information does not significantly implement previous knowledge represents the efficiency of the approach in terms of rapid convergence to a stable sensor configuration. Once this asymptotical condition is satisfied, the sensor network can be evaluated in terms of the isolation likelihood and uncertainty. The maximum number of updates, robustness and uncertainty of the

approach depend on the complexity of the analysed problem.

This study investigated the effect of applying the pre-screening procedure by Banik *et al.* (2015) in reducing the number of nodes that should be considered as a possible source. In Freni & Sambito (2017), all nodes had an equal chance to be the origin of the contamination. In this study, the probability that the nodes are the source is assumed to be different using information about the system. This information is used to implement a pre-conditioning approach to assigning a prior sensor probability distribution to the BDN approach. Two different pre-conditioning strategies are implemented. The prior probability in the BDN analysis is assumed to be proportional to the wastewater volumes, or alternatively, proportional to the contaminant mass passed through each node.

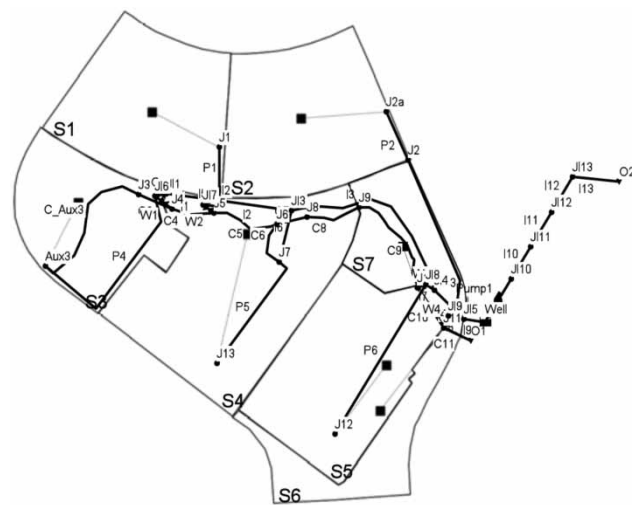
If the pre-conditioning approach is efficient, the BDN approach convergence to a stable sensor configuration may be faster (and reduce computational efforts), and the isolation likelihood may increase. After the prior sensor probability distribution is defined, the application of the BDN approach is affected by the population of events (in this case, the number of simulated contamination events) that is employed for each Bayesian update. A large number of events requires greater computational effort, but the information is used to update the probability distributions only if it is verified several times. A small number is required to achieve faster updates and possibly faster convergence to an asymptotic solution but introduces the risk that unreliable information may be used to update probability distributions. In this study, the tests are performed considering the following approaches (Prior A, B, C, D) to compare the results:

- Prior A: no pre-screening procedure and no prior knowledge (each node has an equal initial probability to be the location of a sensor).
- Prior B: no prior knowledge and pre-screening procedure based on network topology.
- Prior C: pre-screening procedure and prior knowledge based on water fluxes.
- Prior D: pre-screening procedure and prior knowledge based on the mass of contaminant that potentially passed through each node.

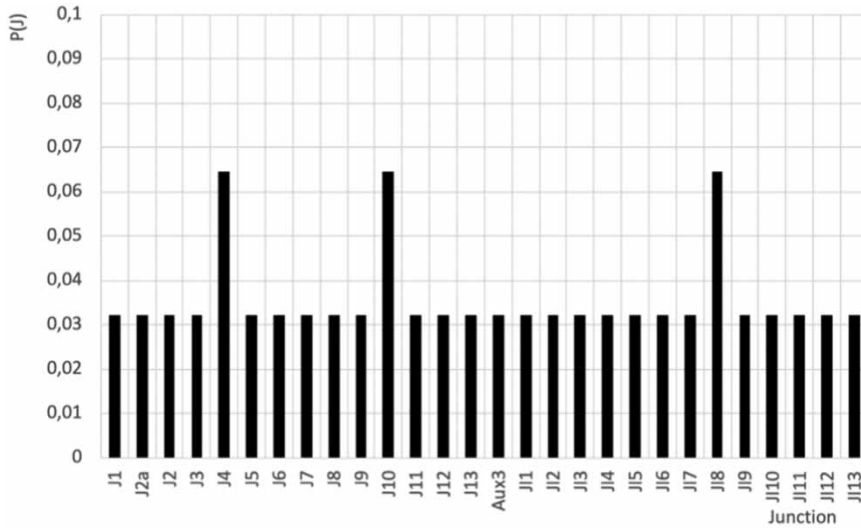
The results of the Bayesian approach are also compared with the results obtained by Banik *et al.* (2017b, 2017c) using the NSGA-II and the greedy algorithms. In the real case study, the efficiency of sensor disposition obtained through the Bayesian approach has been compared with the one of the real monitoring network.

## CASE STUDIES

The literature example, i.e., network Example 8 presented in the EPA SWMM reference manual, involves a combined sewer network that serves an area of 0.12 km<sup>2</sup> and consists of 31 nodes, 29 pipes and a pump (Figure 1). The network is characterized by two outfalls: the WWTP and the overflow. The nodes downstream of commercial/industrial activities are assumed to have a greater probability of being subject to an illegal spill of contaminants into the sewers. The probability distribution function of contamination, as reported in Figure 2, is not considered to be uniform. Three nodes (J4, J10 and J118) are hypothetically considered to host industrial activities, and a double probability of illicit contamination is estimated with respect to the other nodes. This hypothesis does not modify the general applicability of the method in the cases in which industrial nodes may



**Figure 1** | Example 8 network scheme: Jx and J1x (where x represents a sequential number) represent junctions (only Aux3 has a different name as dry weather and wet weather flows are split); Cx and 1x represent conduits of the combined system and dry weather main to the WWTP; and Ox are outflows to the river and the WWTP.



**Figure 2** | Probability density function of contamination in the nodes: three nodes have a double probability of contamination.

change with regards to number, contamination probability or location.

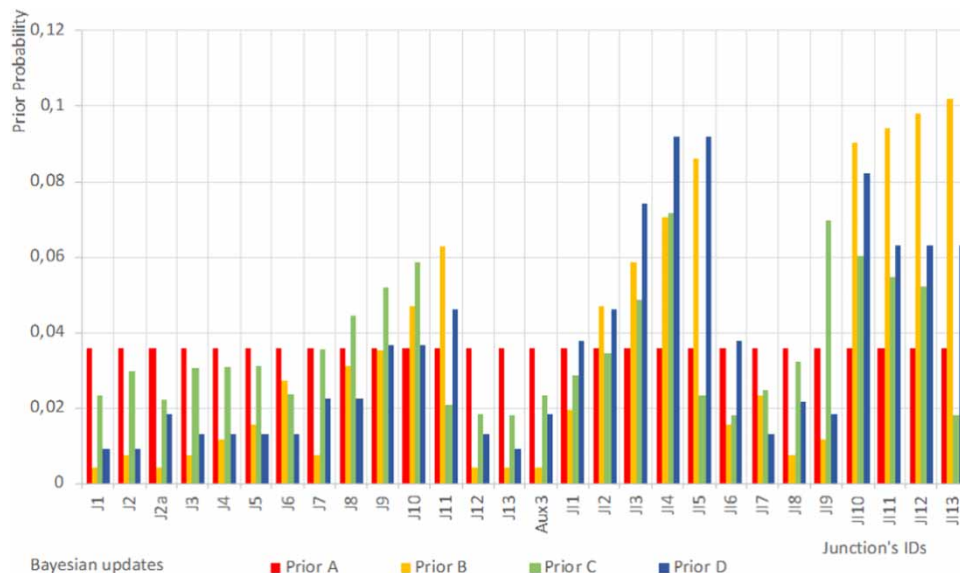
The real case study involves the sewer of Massa Lubrense (Figure 3), which is a town located near Naples, Italy. The sewer is a combined sewer system that covers a predominantly hilly area of 19.71 km<sup>2</sup> with a mean altitude of 121 m with respect to sea level. The system is divided into

12 subcatchments that serve 7,452 users who correspond to a population of 14,087 (2011).

The length of the network is 72 km, and the network consists of 1,909 circular conduits that connect 1,902 junctions, 14 pumps, 14 storage units and 1 treatment plant. The pipes have different cross sections and materials but approximately 80% of them are circular with diameters



**Figure 3** | Massa Lubrense network; the area with a double probability that it is the source of contamination is indicated by a circle. The dots represent the current location of 12 existing sensors that are used to calibrate the model.



**Figure 4** | Prior sensor location distribution in Example 8 network for the different approaches and one sensor.

that vary between 80 mm and 1,000 mm. Due to the variable altimetry of the area, 15 pumping stations exist, 15 outflows are located along the pipes and 10 outflows convey wastewater to the sea, while 5 outflows direct the flow into another pipe. The wastewater is carried to the treatment plant, which represents the final output. All geometric data, which are available on the website <http://www.progettosimona.it>, have been included in the SWMM input file. The daily average values of the dry weather flows in the 1,866 input nodes are estimated considering the population connected to each flow. The input file for the SWMM model has been calibrated using discharge measurements. The system has already installed 12 monitoring stations; their displacements were determined on the basis of practical considerations without any sensor location analysis. Similar to the literature example, the analysis was performed with the hypothesis that the circled area in Figure 3 has a double probability that it will be subjected to more contamination than any other node in the network: 45 nodes have a contamination probability of approximately 0.1%, and the other 1,857 nodes have a contamination probability of approximately 0.05%.

## ANALYSIS OF RESULTS

The Bayesian analysis is initially performed by assuming that any node in the network has the same probability that

it will be the location of a sensor (Prior A). Prior B was performed by applying the methodology proposed in Banik *et al.* (2015). In a second step, other prior knowledge scenarios are applied, in which a different contamination probability is assigned to the nodes. In Prior C and Prior D, a preliminary analysis is performed by Monte Carlo simulations, in which a single source of contamination is located in one of the nodes of the network according to the contamination probability that was previously assigned. In 50 random simulations, the average water and contaminant fluxes through nodes are calculated to assign the prior probability of sensor location to the BDN analysis.

### Example 8 network

The Bayesian analysis is performed for the Example 8 network considering the possible implementation of one, two or three sensors. As denoted in Banik *et al.* (2017b), with more than three sensors, the correlation among the measurements increases with a small increase in the information content.

Figure 4 shows the prior sensor location distribution for the BDN analysis in the case of a single sensor configuration for the Example 8 network for all performed tests. Relevant differences between non-informative distribution (Prior A) and the other distributions are observed, which highlights



the importance of a prior discrimination of the possible solutions in the application of the Bayesian approach. Sensors placed in the upper nodes of the network have a smaller probability of detecting contamination as the majority of contamination episodes occur in nodes located downstream of the sensors. Applying the pre-screening procedure based on the topological approach (Prior B), the prior sensor distribution probability is significantly different with respect to that of Prior A. The procedure is limited when several possible paths to the system outflow are present. In these cases, the flow dividers and their efficiencies in the separation of wet weather flows and polluting loads affect the ability of the sensors to detect the presence of contamination. Nodes JI10, JI11, JI12 and JI13 are located on the connection between the network and the WWTP and they are topologically located in the most downstream part of the network. The presence of dividers conveys only a small part of water volumes to these nodes, which renders them less relevant according to the Prior C approach. Comparing Prior C and Prior D, Prior C tends to overestimate the importance of nodes that receive combined sewer overflows (such as J12 and J13) and are characterized by larger volumes but usually small contaminant concentrations.

As indicated in Table 1, in other tests, the magnitude of the event population of each Bayesian update is modified to understand its impact on the analysis, especially with

respect to the minimum number of simulations, and highlight the most efficient configuration of the sensors. The modification also explains how the value of this BDN parameter can affect the selection of the most relevant nodes for sensor placement and the number of model simulations that are needed to obtain a stable configuration.

As an example, Figure 5 shows a comparison of the use of Prior A and Prior D for the placement of one sensor by adopting an event magnitude of ten for each update. Figure 6 shows the same comparison of adopting 25 events for each update. To make the graphs comparable, the analysis was stopped after 100 simulations, which corresponds to ten updates and four updates in the first case and second case, respectively.

Figure 5 shows some interesting results:

- The analyses do not differ in terms of selection of the best sensor location, which demonstrates that prior knowledge does not affect the final decision but affects the computational resources and data that are needed to attain the final stable distribution of likelihood.
- Using Prior D, after only two updates (20 simulations), the nodes with the highest likelihood of being selected (JI10) as a sensor location, are identified, and the selection does not change until the end of the analysis.
- Using Prior A, six updates (60 simulations) are necessary to achieve the same results obtained for Prior D after 20 simulations. Prior B and Prior C (not reported) performed equally well and obtained the same selection after four updates (40 simulations).

The comparison between Figures 5 and 6 shows that, even if the number of simulations is identical, the aggregation in the Bayesian update process produces some differences:

- The final selection is not affected by either of the adopted prior distributions. In both cases, three updates (75 simulations) are necessary to identify a stable candidate node for the best sensor location.
- Posterior distributions are affected by a large variability after four updates, which requires a larger number of simulations to obtain a robust solution.

Table 1 reports the results of other tests that were performed considering a maximum of three sensors and varying the event magnitude for each update from 10 to

**Table 1** | Results in terms of Bayesian probability after 1,000 simulations

**Example 8 network**

No. of sensors	No. of updates	Event pop. of updates	Highest sensor posterior probability	Prior probability of each sensor combination
1	10	100	0.1315	0.034
2	10	100	0.0076	0.0025
3	10	100	0.00068	0.00027
1	40	25	0.16	0.034
2	40	25	0.0085	0.0025
3	40	25	0.00075	0.00027
1	100	10	0.178	0.034
2	100	10	0.0085	0.0025
3	100	10	0.00077	0.00027

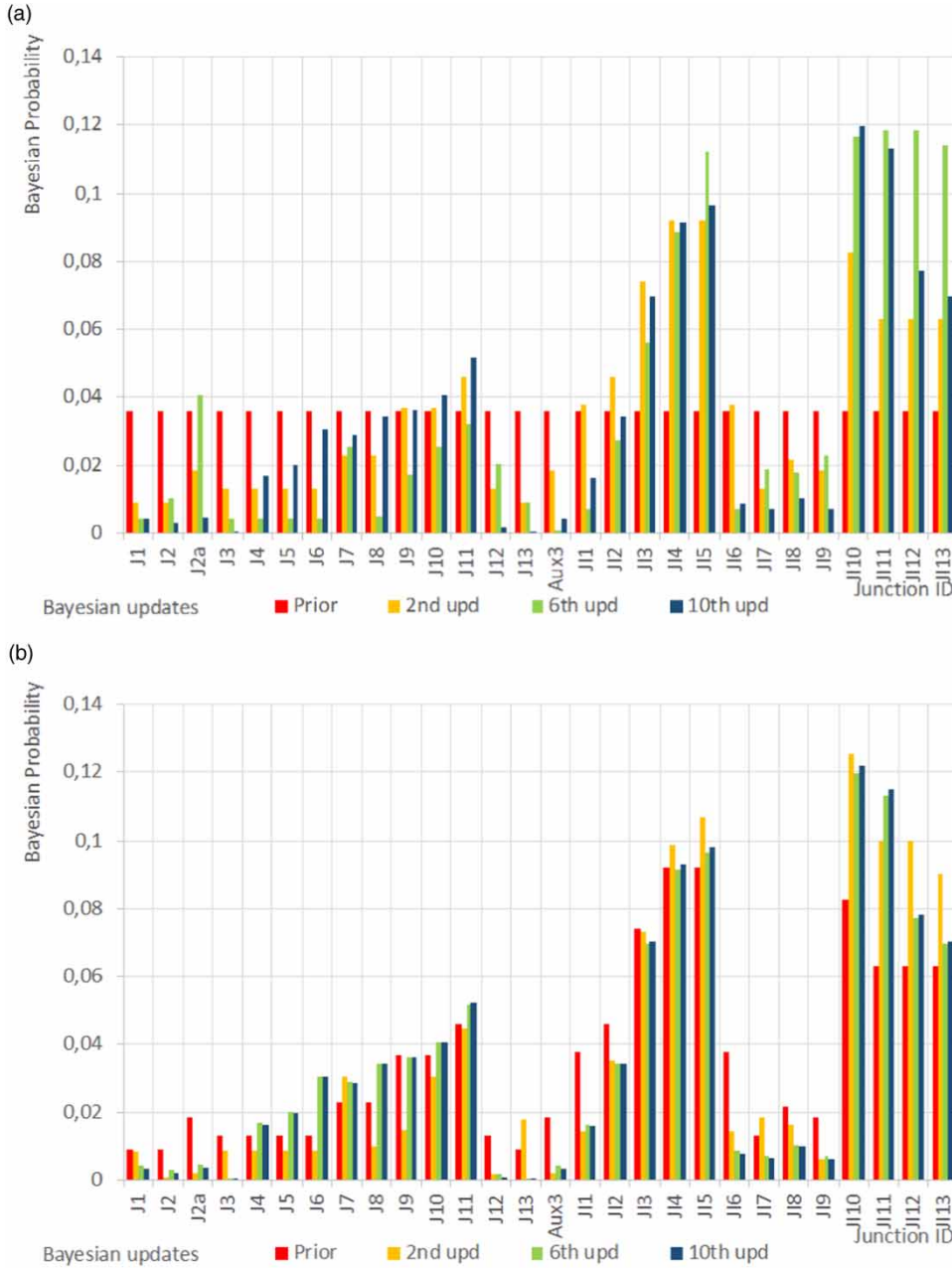
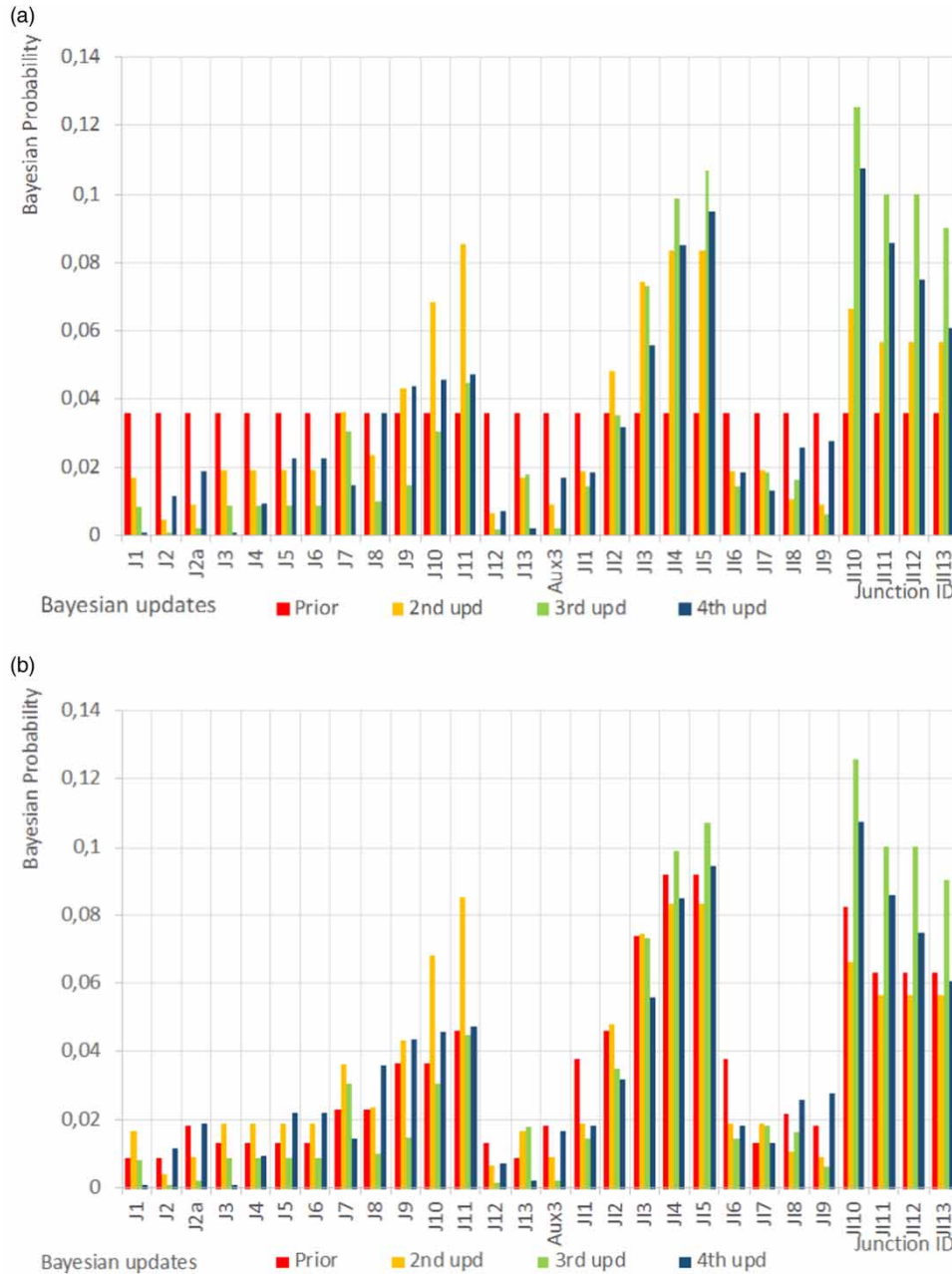


Figure 5 | Prior and posterior distribution after two, six and ten updates. Event population for each update = 10: (a) Prior A and (b) Prior D.

100. The results indicate that, for the Example 8 network, the BDN configuration that enables a more efficient selection of (in terms of Bayesian probability) sensors' location refers to the use of Bayesian updates with small populations (ten events). With ten events after ten updates (100 simulations) to a maximum of 100 updates (1,000 simulations),

the improvement is observed (not reported) to be limited, and the posterior distribution attains a stable configuration with very small variations independently from prior knowledge. This behaviour confirms that the Bayesian methods do not retain the memory of the initial assumptions if the number of performed updates is sufficiently large.



**Figure 6** | Prior and posterior distribution after two, three and four updates. Event population for each update = 25: (a) Prior A and (b) Prior D.

All applied approaches, independently from selected prior knowledge and the population of updates, attained the same final selection for the analysed network:

- 1 sensor: node J110,
- 2 sensors: nodes J11 and J111,
- 3 sensors: nodes J11, J15 and J110.

Starting from this result, the probability that the sensors positioned in these junctions will be able to identify the contamination source (isolation likelihood  $F_1$ ) is 47% with one sensor, 78% with two sensors and 84% with three sensors. With one sensor, 24% of the contamination events are undetected and 29% are detected; however, the origin of the contaminant is not discovered. With two sensors, only

11% of the events are not detected and an additional 11% of the events are detected without identifying the source. With three sensors, only 6% of the events are undetected even if 10% of the events are detected without identifying the source. Sensor network reliability  $F_2$  is relatively low with two and three sensors reaching 17% and 33%, respectively, which shows that the most relevant part of the contamination events was detected by a single sensor.

The use of a non-uniform informative a priori sensor location distribution reduces the simulation time. The use of distributions based on contaminant mass perform better than other methods, but the simple topological approach (obtained without the use of any additional simulation) can reduce the computational effort by one-third.

Considering the positioning of three sensors, two of the three selected nodes coincide with the ones individuated by two different procedures in Banik et al. (2017b). This comparison confirms the validity and robustness of the presented methodology.

### Massa Lubrense network

The BDN approach is successively applied to the real network of Massa Lubrense (Italy) considering three sensor configurations that involve 6, 12 (actual number of implemented monitoring stations for model calibration) and 18 sensors.

The analysis is performed to analyse the impact of the BDN update population and the ability of the best possible sensor locations. To reduce computational time, the analysis was performed starting from the informative distribution Prior D based on 100 random contamination simulations.

Table 2 reports the results in terms of the Bayesian probability in the various considered BDN configurations. In this case, the best sensor configuration and the efficiency of the sensor network in identifying the polluting source depend on the number of events and the number of procedure updates. The complexity of the analysed network considers that the best strategy is to increase the number of Bayesian updates, which reduces the population of each update: the use of 40 updates with 25 simulations each provides better results than the use of 10 updates with 100 simulations each. This finding can be explained by the complexity of

**Table 2** | Results in terms of the Bayesian probability in various BDN configurations (Massa Lubrense)

#### Massa Lubrense network

No. sensors	No. of Bayes updates	Event pop. of updates	Highest sensor posterior probability	Prior probability of each sensor combination
6	10	100	0.003	$5.26 \times 10^{-4}$
12	10	100	$2.45 \times 10^{-8}$	$1.53 \times 10^{-17}$
18	10	100	$6.44 \times 10^{-16}$	$2.21 \times 10^{-31}$
6	40	25	0.046	$5.26 \times 10^{-4}$
12	40	25	$3.23 \times 10^{-4}$	$1.53 \times 10^{-17}$
18	40	25	$4.32 \times 10^{-10}$	$2.21 \times 10^{-31}$
6	100	10	0.026	$5.26 \times 10^{-4}$
12	100	10	$6.65 \times 10^{-6}$	$1.53 \times 10^{-17}$
18	100	10	$2.15 \times 10^{-13}$	$2.21 \times 10^{-31}$

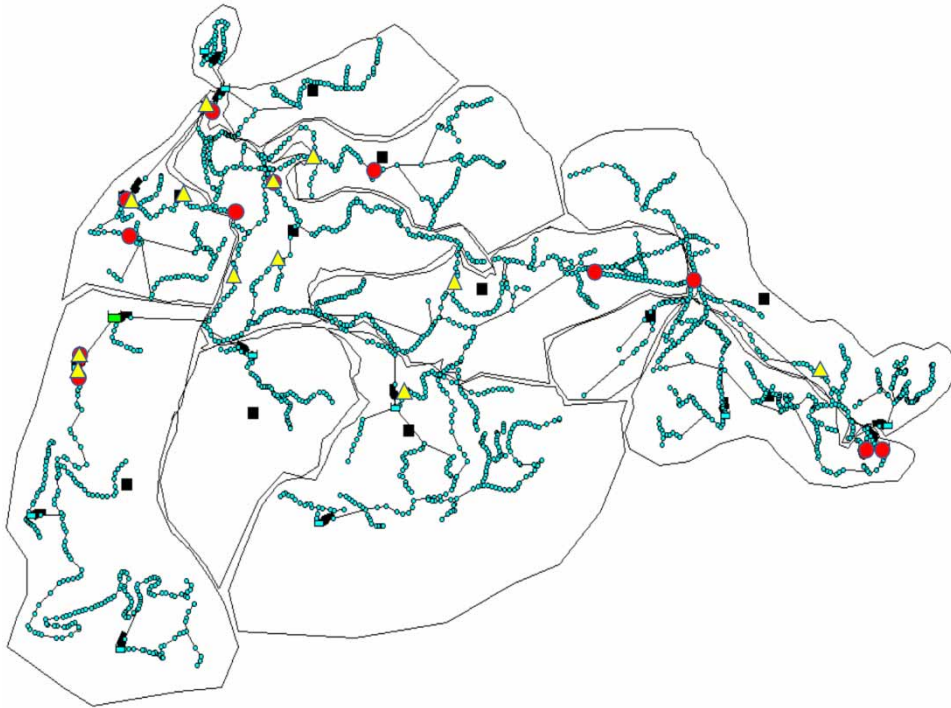
the system and the number of possible sensor combinations that should be considered.

Positively, the different analyses converge on the same set of sensor locations, which confirms the robustness of the proposed approach. Analysing the three configurations (6, 12 and 18 sensors), the probability that the sensors intercept the contamination (isolation likelihood  $F_1$ ) is 58% with 6 sensors, 81% with 12 sensors and 92% with 18 sensors.

Detection reliability  $F_2$  is substantially higher in the real case due to the higher number of sensors: 68% of the detected events are reported by two or more sensors in the first configuration (6 sensors) and the probability increases to 77% and 86% with 12 sensors and 18 sensors, respectively. A significant number of events remained undetected: 27% with 6 sensors, 6% with 12 sensors and 4% with 18 sensors.

Figure 7 shows the location of 12 sensors according to the proposed methodology (triangles). Comparing the obtained locations with the actual configuration of the monitoring network (large circles), some considerations can be obtained:

- The optimal configuration of sensors in the downstream part of the network (upper part of the figure) is consistent with the existing configuration with three sensors that are actually located in their optimal positions and two sensors that are only a few nodes from their optimal position.
- In the western part of the network, the methodology provided the same sensor locations of the real network,



**Figure 7** | Massa Lubrense network with the indication of the best location for 12 sensors (triangles) according to the methodology and compared with the actual position of the monitoring stations (large circles).

while a difference was observed in the central and eastern parts of the system; these differences can be primarily connected to the fact that the real sensor location was determined by simple geometric considerations for the network without integrating an analysis by numerical models and optimization methods.

## CONCLUSIONS

The contamination of surface waters represents one of the most important aspects of urban management for environmental and sanitary implications and social-economic issues that may arise. The goal of this study is to develop a decision-support approach for identifying the location of the water quality sensors and illicit intrusions in sewers. The study focuses on soluble conservative pollutants, such as heavy metals.

The analysis was based on a Bayesian approach to introduce data assimilation and identification probability

in the procedure. The system was simulated by SWMM for different random contamination scenarios. Progressive updates were performed considering contamination events and the need to evaluate the probability that the contamination source will be identified with the sensors located in a specific node. The proposed methodology showed a progressive increase in the identification probability obtained by the Bayesian update. Appropriate pre-screening and/or pre-conditioning approaches are relevant in terms of early identification of the most suitable sensor locations.

The following conclusions were obtained from the analyses:

- The value of the Bayesian probability of identifying the contamination source tends to increase as the number of Bayesian updates increase. However, the number of simulations in each update should be based on the complexity of the problem to be solved to ensure that more complex problems need a larger population for each update.
- As the number of updates increases, with the same number of sensors, the maximum value of the Bayesian

probability increases. However, the marginal increments are progressively lower.

- The use of a pre-screening procedure and/or the inclusion of knowledge about the system characteristics enables convergence towards an optimal solution in less time and with smaller computational resources. Even if this fact may be negligible for a small network, the impact on larger networks is relevant, which enables the same optimal configuration and half of the required number of simulations.
- The sensor locations obtained with the proposed methodology are able to identify a large number of contamination events: maximum of 84% events in Example 8 case with three sensors and maximum of 92% events in Massa Lubrense with 18 sensors.
- Even if a high number of sensors is required to obtain a high probability of contamination source identification, the BDN approach enables a progressive implementation of the sensor network depending on the water manager budget limitations. The comparison with the real configuration of the monitoring network in Massa Lubrense demonstrates that the methodology is also informative, which helps the water manager to understand the dynamics of the network and highlights the areas that require better monitoring.

The analysis shows the potential impact of the proposed methodology. However, further developments are needed to take into account non-conservative pollutants, such as biological contaminants that can have a large impact on the environment and public health. The methodology should be upgraded and tested to take into account the presence of multiple contamination sources and the possibility of deploying Lagrangian sensors carried by the flow. Additionally, the comparison with fault isolation approaches (Blanke *et al.* 2015) may be interesting from the perspective of transferring fault mode and effect analysis (FMEA)-type approaches to the water industry.

## REFERENCES

- Banik, B. K., Di Cristo, C. & Leopardi, A. 2015 [A pre-screening procedure for pollution source identification in sewer systems](#). *Procedia Engineering* **119** (1), 360–369.
- Banik, B. K., Di Cristo, C., Leopardi, A. & de Marinis, G. 2017a [Illicit intrusion characterization in sewer systems](#). *Urban Water Journal* **14** (4), 416–426.
- Banik, B. K., Alfonso, L., Di Cristo, C., Leopardi, A. & Mynett, A. 2017b [Evaluation of different formulations to optimally locate sensors in sewer systems](#). *Journal of Water Resources Planning and Management* **143**, 7.
- Banik, B. K., Alfonso, L., Di Cristo, C. & Leopardi, A. 2017c [Greedy algorithms for sensor location in sewer systems](#). *Water* **9**, 856. doi: 10.3390/w9110856.
- Blanke, M., Kinnaert, M., Lunze, J. & Staroswiecki, M. 2015 *Diagnosis and Fault-Tolerant Control*, 3rd edn. Springer-Verlag, Berlin, Heidelberg. New York. <https://doi.org/10.1007/978-3-662-47943-8>
- Boenne, W., Desmet, N., Van Looy, S. & Seuntjens, P. 2014 Use of online water quality monitoring for assessing the effects of WWTP overflows in rivers. *Environmental Science: Processes Impacts* **16**, 1510–1518.
- Di Cristo, C. & Leopardi, A. 2008 [Pollution source identification of accidental contamination in water distribution networks](#). *Journal of Water Resources Planning and Management* **134** (2), 197–202.
- Even, S., Poulin, M., Mouchel, J.-M., Seidl, M. & Servais, P. 2004 Modelling oxygen deficits in the Seine River downstream of combined sewer overflows. *Ecological Modelling* **173** (2–3), 177–196.
- Freni, G. & Sambito, M. 2017 Probabilistic approach to the positioning of water quality sensors in urban drainage networks. In: *International Conference of Urban Drainage*, Prague.
- Gironás, J., Roesner, L. A., Davis, J., Rossman, L. A. & Supply, W. 2009 *Storm Water Management Model Applications Manual*. National Risk Management Research Laboratory, Office of Research and Development, US Environmental Protection Agency, Cincinnati, OH, USA.
- Irvine, K., Rossi, M. C., Vermette, S., Bakert, J. & Kleinfelder, K. 2011 [Illicit discharge detection and elimination: low cost options for source identification and trackdown in stormwater systems](#). *Urban Water Journal* **8** (6), 379–395.
- Jiang, J., Shi, B., Huang, A., Wang, N. & Yuan, Y. 2018 [Inverse uncertainty characteristics of pollution source identification for river chemical spill incidents by stochastic analysis](#). *Frontiers of Environmental Science and Engineering* **12** (5), 6.
- Kabir, G., Tesfamariam, S., Francisque, A. & Sadiq, R. 2015 [Evaluating risk of water mains failure using a Bayesian belief network model](#). *European Journal of Operational Research* **240**, 220–234.
- Kim, M., Choi, C. Y. & Gerba, C. P. 2013 [Development and evaluation of a decision supporting model for identifying the source location of microbial intrusions in real gravity sewer systems](#). *Water Research* **47**, 4630–4638.
- Korb, K. B. & Nicholson, A. E. 2010 *Bayesian Artificial Intelligence*, 2nd edn. CRC Press, Boca Raton, FL, USA.
- Lee, C., Paik, K., Yoo, D. G. & Kim, J. H. 2014 [Efficient method for optimal placing of water quality monitoring stations for an](#)

- ungauged basin. *Journal of Environmental Management* **132**, 24–31.
- Lifshitz, R. & Ostfeld, A. 2019 Clustering for real time response to water distribution system contamination event intrusion. *Journal of Water Resource Planning and Management* **145**, 2.
- Montserrat, A., Bosch, L., Kiser, M. A., Poch, M. & Corominas, L. 2015 Using data from monitoring combined sewer overflows to assess, improve, and maintain combined sewer systems. *Science of the Total Environment* **505**, 1053–1061.
- Passerat, J., Ouattara, N. K., Mouchel, J.-M. & Rocher, V. 2011 Impact of an intense combined sewer overflow event on the microbiological water quality of the Seine River. *Water Research* **45**, 893–903.
- Phan, T. D., Smart, J. C. R., Capon, S. J., Hadwen, W. L. & Sahin, O. 2016 Applications of Bayesian belief networks in water resource management: a systematic review. *Environmental Modelling & Software* **85**, 98–111.
- Preis, A. & Ostfeld, A. 2008 Multiobjective contaminant sensor network design for water distribution systems. *Journal of Water Resources Planning and Management* **134**, 366–377.
- Qin, X., Gao, F. & Chen, G. 2012 Wastewater quality monitoring system using sensor fusion and machine learning techniques. *Water Research* **46** (4), 1133–1144.
- Rathi, S. & Gupta, R. 2016 A simple sensor placement approach for regular monitoring and contamination detection in water distribution networks. *KSCE Journal of Civil Engineering* **20** (2), 597–608.
- Sambito, M., Di Cristo, C., Freni, G., Leopardi, A. & Quintiliani, C. 2018 Pre-conditioning approach to Bayesian Decision Networks for water quality sensors positioning in urban drainage systems. In *13th Hydroinformatics International Conference*, 1–6 July 2018, Palermo, Italy.
- Srinivas, R., Singh, A. P., Gupta, A. A. & Kumar, P. 2018 Holistic approach for quantification and identification of pollutant sources of a river basin by analyzing the open drains using an advanced multivariate clustering. *Environmental Monitoring and Assessment* **190**, 720.
- Tinelli, S., Creaco, E. & Ciaponi, C. 2017 Sampling significant contamination events for optimal sensor placement in water distribution systems. *Journal of Water Resource Planning and Management* **143** (9), 1–10.
- Troutman, S. C., Schambach, N., Love, N. G. & Kerkez, B. 2017 An automated toolchain for the data-driven and dynamical modeling of combined sewer systems. *Water Research* **126**, 88–100.
- Vonach, T., Tscheikner-Gratl, F., Rauch, W. & Kleidorfer, M. 2018 A heuristic method for measurement site selection in sewer systems. *Water* **10** (2), 1–16.
- Weickgenannt, M., Kapelan, Z., Blokker, M. & Savic, D. A. 2010 Risk-based sensor placement for contaminant detection in water distribution systems. *Journal of Water Resources Planning Management* **136** (6), 629–636.
- Yazdi, J. 2018 Water quality monitoring network design for urban drainage systems, an entropy method. *Urban Water Journal* **15** (3), 227–233.

First received 9 February 2019; accepted in revised form 14 July 2019. Available online 13 September 2019