Water Science & Technology

IWA PUBLISHING

Water Science & Technology Vol 90 No 3, 951 doi: 10.2166/wst.2024.233

Identification of illicit discharges in sewer networks by an SWMM-Bayesian coupled approach

Liyuan Yang, Biao Huang* and Jiachun Liu

School of Civil and Environmental Engineering, Ningbo University, Ningbo 315211, China
*Corresponding author. E-mail: huangbiao@nbu.edu.cn

BH, 0000-0001-9403-4508

ABSTRACT

Illicit discharges into sewer systems are a widespread concern within China's urban drainage management. They can result in unforeseen environmental contamination and deterioration in the performance of wastewater treatment plants. Consequently, pinpointing the origin of unauthorized discharges in the sewer network is crucial. This study aims to evaluate an integrative method that employs numerical modeling and statistical analysis to determine the locations and characteristics of illicit discharges. The Storm Water Management Model (SWMM) was employed to track water quality variations within the sewer network and examine the concentration profiles of exogenous pollutants under a range of scenarios. The identification technique employed Bayesian inference fused with the Markov chain Monte Carlo sampling method, enabling the estimation of probability distributions for the position of the suspected source, the discharge magnitude, and the commencement of the event. Specifically, the cases involving continuous release and multiple sources were examined. For single-point source identification, where all three parameters are unknown, concentration profiles from two monitoring sites in the path of pollutant transport and dispersion are necessary and sufficient to characterize the pollution source. For the identification of multiple sources, the proposed SWMM-Bayesian strategy with improved sampling is applied, which significantly improves the accuracy.

Key words: Bayesian-MCMC, illicit discharge, sewer network, source identification, SWMM-Bayesian

HIGHLIGHTS

- Identifying pollution sources can be applied for both instantaneous and continuous discharge scenarios.
- To characterize a single pollution source, data from two monitoring sites along the pollutant's path are necessary and sufficient.
- The strategic placement of monitoring sites and improved sampling enhance the effectiveness and precision of the Bayesian-SWMM approach for identifying multiple unauthorized discharge sources.

1. INTRODUCTION

In urban drainage systems, illegal connections and illicit discharges represent a common problem. Due to a variety of reasons, domestic and industrial sewage can be inadvertently discharged into storm sewers, leading to pollution in receiving waters and potentially to the dissemination of viruses, which poses a threat to public health and safety. Accidental spills or illicit discharges from manufacturing plants, which contain chemicals toxic to microorganisms, frequently occur in sanitary sewer systems (Li *et al.* 2017; Shao *et al.* 2021; Wu & Chen 2023). These illicit discharges reduce biomass activity and deteriorate the performance of wastewater treatment plants (WWTPs). Currently, in the service area of Zhenhai WWTP, Ningbo, China, water quality monitoring at a pumping station has revealed that persistent total phosphorus levels are particularly high; consequently, a pollution identification project is underway.

The EPA has introduced technical guidelines for illicit discharge detection and elimination in municipal storm drains (Brown *et al.* 2004), in which visual inspection and indicator sampling were considered effective in tracking and isolating the specific source (Irvine *et al.* 2011; Hachad *et al.* 2022). However, for unexpected pollution in sanitary and combined sewer systems, identifying the source proves to be significantly more challenging (Banik *et al.* 2017a, 2017b). Detecting abnormal discharge and/or water quality indicators is relatively straightforward; however, this information alone is insufficient to

pinpoint the source. To mitigate the adverse impacts effectively, it is imperative to identify the discharging source promptly. Owing to the complex structure of sewer networks, pinpointing the discharge location, let alone the discharging history, is exceedingly difficult (Shao *et al.* 2021). In such scenarios, methods capable of tracking down sources and reconstructing the discharge profile are in high demand. Currently, numerous Chinese cities, including Ningbo, are actively engaged in developing smart urban drainage systems, utilizing modern technologies such as the Internet of Things, big data, and artificial intelligence. Through the real-time monitoring of water quantity and quality, intelligent control of the urban drainage system can be achieved. Pollution source identification constitutes a critical component of such systems.

The identification of illicit connections and discharges is categorized as an inverse problem, marked by inherent complexity and potential for error (Laird *et al.* 2006; Levenspiel 2011; Williams *et al.* 2011; Moghaddam *et al.* 2021). Extensive research on source identification has been conducted in water distribution networks (Kessler *et al.* 1998; Ostfeld & Salomons 2005; Hart & Murray 2010; Liu *et al.* 2015; Berglund *et al.* 2020) and surface water bodies such as rivers, lakes (Cheng & Jia 2010; Wang *et al.* 2018), and groundwater (Skaggs & Kabala 1994; Alapati & Kabala 2000; Moghaddam *et al.* 2021). Only a few studies have focused on source identification in sewer systems, of which most rely on optimization processes (Jiang *et al.* 2013). Kim *et al.* (2013) introduced an optimization method based on artificial neural networks and developed an algorithm to search for pathogens in sewer systems. Banik et al. (2017a, 2017b) implemented a genetic algorithm integrated with the Storm Water Management Model (SWMM) to identify contaminant intrusions in sewer systems. Xu *et al.* (2021) combined a microbial genetic algorithm with the SWMM to trace illicit connections within a sewer system.

Optimization is a prevalent approach for addressing inverse problems, relying on minimizing the difference between observed and simulated data to find solutions. This approach, however, may not fully consider the uncertainties inherent in the data, which can affect the accuracy of the results (Shao *et al.* 2021). In sewer networks, uncertainties are amplified due to the variable degradation of chemical substances, adding complexity to source identification (Chen *et al.* 2012; Ramin *et al.* 2016). Consequently, identifying sources in sewer systems remains a formidable challenge. Recognizing the limitations of traditional optimization, the focus has shifted toward identifying the stochastic distribution of parameters, which offers a more comprehensive understanding of the problem than a single optimal solution (Yee & Flesch 2010; Wang & Harrison 2013). Recent advancements in Bayesian-based stochastic approaches (Wang & Harrison 2013; Wu *et al.* 2020) have reinvigorated efforts to solve the source identification problems in sewer systems. Bayesian methods are particularly advantageous because they can reconstruct source information using limited data, surveying the entire solution domain to provide statistically plausible solutions.

The Bayesian method, which can make full utilization of prior information and account for uncertainty, has been employed to solve the inverse problem and determine the probability of incidents, such as the approximate location of pollution sources and the time series of discharges. The results will be represented as a probability density function. The optimal estimation is derived from the known information about the pollution source, leading to the solution of the inverse problem (Neupauer & Wilson 1999). Bayesian inference methods have been applied to address source identification problems in various fields, notably within water distribution networks (Yang *et al.* 2009; Wang & Harrison 2013, 2014). Researchers have utilized the Bayesian-Markov chain Monte Carlo (Bayesian-MCMC) algorithm to trace the discharge process of multiple pollution sources in river channels (Yang *et al.* 2016; Wang *et al.* 2018). Notably, Shao *et al.* (2021) developed a stochastic source identification model by coupling Bayesian inference with SWMM to reconstruct the profile of an instantaneous pollution event in a sewer network. The Bayesian-SWMM model was demonstrated to be effective and accurate in identifying the unknown source parameters.

This study aims to apply the SWMM-Bayesian approach to investigate more practical scenarios, specifically continuous pollution discharge events (Grbčić *et al.* 2021) and the presence of multiple pollution sources (Yang *et al.* 2016; Wang *et al.* 2021). The SWMM was employed to model the hydraulic dynamics and water quality characteristics associated with identified sources of pollution within the sewer system. The Bayesian-MCMC technique was utilized for the analysis and optimization of model parameters, thereby facilitating the localization of pollution sources via sampling. The algorithm facilitates the simultaneous estimation of three key parameters of the illicit source, which are its location, mass, and discharge time, and devises a pragmatic approach for the identification. Following the validation of the methodology's efficacy, the study further examined situations with multiple sources and also proposed an improved SWMM-Bayesian approach for this purpose.

2. METHOD

The methodology employed in the current investigation comprises two core computational components, building upon the framework established by Shao *et al.* (2021). The first component employs a simulation based on the SWMM to elucidate the dynamics of flow and contaminant dispersion throughout the sewer system. Subsequently, the second element applies a Bayesian-MCMC inference technique to estimate the probabilities associated with the undetermined source parameters.

2.1. Pollutant transport modeling

The SWMM solver is commonly used to simulate the hydraulics and pollutant transport in urban drainage systems. Specifically, the model assumes that the pollutants are simulated as a continuous stirring tank reactor in the sewer network. Thus, it is a distributed discrete-time simulation model that computes all pollutant concentrations at each node and in each pipe after determining the hydraulic state at each computational time step size. SWMM solves for the transport of pollutants by solving the mass conservation equation for a complete discharge reactor, instead of the one-dimensional convective diffusion equation (Rossman & Huber 2016), written as

$$\frac{dVC}{dt} = (Q_{\text{in}} \cdot C_{\text{in}}) - (Q_{\text{out}} \cdot C) - KVC + X(J_x, M, T_d), \tag{1}$$

where V is the volume in the reactor, C is the concentration in the reactor, C_{in} is the influent concentration of the reactor, Q_{in} is the volumetric inflow rate, Q_{out} is the volumetric outflow rate, K is the first-order degradation coefficient, and $X(J_x, M, T_d)$ is the time-dependent source term that is a function of three parameters, including the illicit discharge node J_x, discharge mass m, and initial discharge time T_d.

2.2. Bayesian inference via MCMC

2.2.1. Bayesian statistics

The Bayesian inference approach is adopted in this study. The prior information delineates the probabilistic distribution of the source term parameter X prior to the acquisition of the concentration data Y. The prior distributions for the discharge node location, the discharge quantity, and the timing of discharge are presumed to follow a uniform distribution. For the junction node with illicit discharge, the probability function for the discrete uniform distribution is

$$pX_{J_x} = \frac{1}{K}, \tag{2}$$

where K is the total number of the junction nodes. For the discharge quantity and the timespan of discharge, the probability function can be expressed mathematically as

$$pX = \begin{cases} \dfrac{1}{X_{\max} - XE} & X_{\min} \le X \le X_{\max} \\ 0 & \text{others} \end{cases}. \tag{3}$$

The joint prior probability $pX(J_x, m, T_d)$ of the three illicit discharge source parameters can be calculated as follows:

$$pX(J_x,\ M,\ T_d) = p(X_{J_x})p(X_M)p(X_{T_d}). \tag{4}$$

The notation $P(Y|X)$ represents the conditional probability of observing the concentration monitoring data as Y, given that the parameter characterizing the discharge source is X. This is also referred to as the likelihood function, which quantifies the congruence between the model predictions and the actual observed values (Shao *et al.* 2021). Assuming that the observational noise can be characterized as white noise, it is further posited that this noise adheres to a normal distribution with a mean of zero and a standard deviation of σ. Consequently, the likelihood function can be represented as follows:

$$p(Y|X_i = (J_x,\ M, T_d)) = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^n} \exp\left\{ -\sum_{i=1}^{n} \frac{[Y_{\text{i}} - N_i(J_x, M, T_d|X)]^2}{2\sigma^2} \right\}, \tag{5}$$

where $Y = \{Y_1, Y_2, ..., Y_i, ..., Y_n\}$ denotes the set of actual observed concentration values, corresponding to the measurements obtained at the monitoring locations for the true source parameter, $N = \{N_1, N_2, ..., N_i, ..., N_n\}$ represents the set of numerically predicted concentration values produced at the monitoring sites as sampled by the Bayesian algorithm, and n is the total number of concentration data points.

The posterior probability denotes the probability of the distribution of the discharge source parameter X after obtaining the concentration value Y. According to Bayes' theorem, the following conclusions can be obtained:

$$p(X_i|Y) = \alpha \times \prod p(X_i) \times \frac{1}{\left(\sqrt{2\pi}\sigma\right)^n} \exp\left\{-\sum_{i=1}^{n} \frac{[Y_i - N_i(J_x, M, T_d|X)]^2}{2\sigma^2}\right\}, \tag{6}$$

where α is a proportional constant.

2.2.2. Metropolis–Hastings algorithm

Confronted with the challenge of resolving the posterior distribution in high-dimensional spaces, which is a formidable task for conventional parsing techniques, this study employs the MCMC method to facilitate the sampling process (Kass *et al.* 1998). The MCMC technique generates a sequence of sampling points whose distribution approximates the posterior probability. These points constitute a Markov chain, the equilibrium distribution of which, upon convergence, represents the sought-after posterior distribution.

The Metropolis–Hastings (M–H) algorithm is the most extensively utilized method within the suite of MCMC sampling techniques. It is the foundational algorithm from which other MCMC methods can be considered to be special cases or extensions (Hastings 1970). To construct a Markov chain that converges to the posterior probability distribution, one must select a proposal transition probability, denoted as $q(X_t, X_{t+1})$, along with a function $a(X_t, X_{t+1})$, where $0 < a(X_t, X_{t+1}) \leq 1$, for any pair (X_t, X_{t+1}) with $X_t \neq X_{t+1}$. Together, these define a transition kernel $p(X_t, X_{t+1})$ that governs the progression of the chain.

$$p(X_t, X_{t+1}) = q(X_t, X_{t+1})a(X_t, X_{t+1}). \tag{7}$$

The probability of acceptance at the sampling node is set as follows:

$$aX_t, X_{t+1} = \min\left\{\frac{p(X_{t+1})}{p(X_t)}, 1\right\}. \tag{8}$$

Within the M–H algorithm, the proposal distribution is often chosen based on the prior distribution of the parameters. A common choice for the transition probability function q is to select a uniform distribution $q(X_{t-\varepsilon}, X_{t+1-\varepsilon})$, or alternatively, a normal distribution $q(X_t, \varepsilon^2)$, in which ε denotes the sampling step size. This step size is indicative of the interval between successive values of the unknown parameter during sampling and is a critical parameter influencing the sampling efficiency.

The workflow of the M–H algorithm can be delineated as follows:

(1) *Initialization*: Generate an initial state or point X_0 based on prior information, such that $X_0 = X_t$. Using the current state X_t, draw a candidate sample X_{t+1} from the proposal distribution $q(X_t, X_{t+1})$.
(2) *Acceptance probability calculation*: Calculate the acceptance probability $a(X_t, X_{t+1})$ for the candidate state X_{t+1}.
(3) *Decision*: If the acceptance probability suggests that the candidate point is more probable than the current point, i.e., $p(X_{t+1}) > p(X_t)$, set a greater than 1; accept X_{t+1} as the new current state, assign $X_t = X_{t+1}$, and return to step (1) to draw a new sample. Conversely, if the candidate point is less probable, set a less than 1. The decision to accept the new point is then based on a random comparison with a. If rejected, maintain the current state X_t, and return to step (1) to draw a new sample.

Note that this description simplifies the decision-making process. In practice, a random number u from a uniform distribution $U(0,1)$ is generated, and if $u \leq a(X_t, X_{t+1})$, the new state X_{t+1} is accepted; otherwise, it is rejected, and the algorithm remains at the current state X_t. The algorithm then iterates from step (1) with the accepted or retained state to generate a sequence of samples. According to the conclusions obtained by Shao *et al.* (2021), the step sizes of the unknown parameters J_x, M, and T_d are 1, 10, and 1, respectively.

2.3. Coupling SWMM and Bayesian inference

The MatSWMM toolbox developed by Riaño-Briceño *et al.* (2016) was adopted in this study to couple the SWMM modeling and the Bayesian inference calculation. The Bayesian-MCMC algorithm described above was coupled with SWMM via the MATLAB platform and the flowchart of the coupling approach is given in Figure 1.

The main steps of the coupling are described as follows:

(1) *SWMM model construction*: Develop the SWMM model, specifying the 'true values' of various inputs such as $X(J_x, M, T_d)$. Create an initial input .inp file for the SWMM model; run the SWMM model to generate the .rpt file thereby obtaining the concentration value Y at the monitoring node.
(2) *Initial sampling*: Randomly generate an initial sample X_0 for the parameters of the discharge source based on the prior information.
(3) *SWMM simulation and likelihood calculation*: Rename the .inp file to a .txt file to edit the parameters.
(4) Locate the parameter values within the .txt file and replace them with the initial sample X_0. Rename the file back to .inp and run the SWMM simulation through MatSWMM. Extract the calculated concentration time series from the .rpt report file. Calculate the likelihood function $p(Y|X_0)$ by comparing the simulated concentration values with the 'true value' concentrations from step (1).

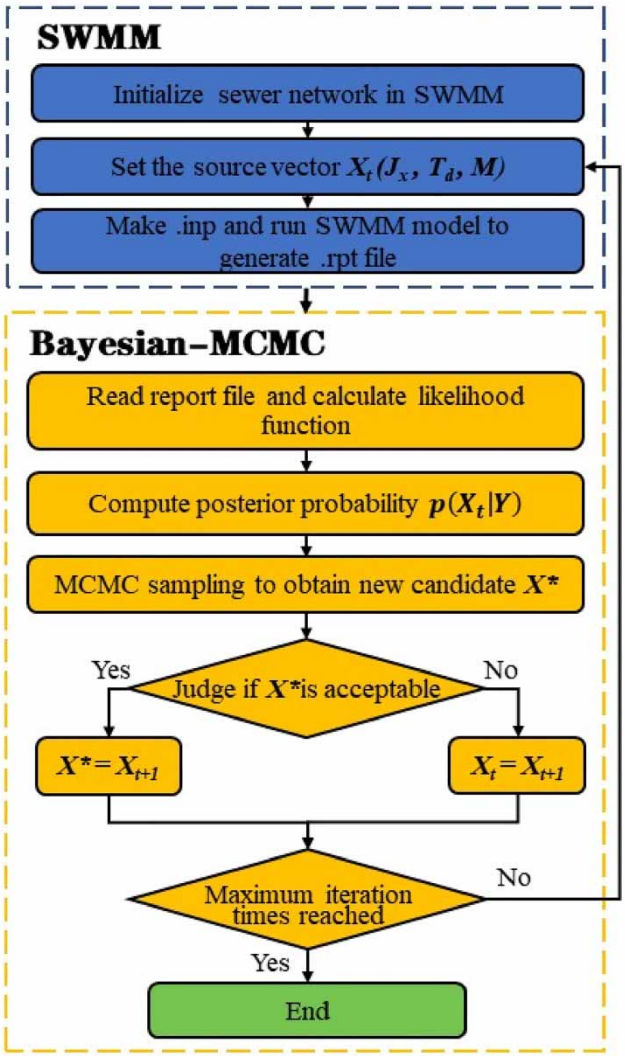

Figure 1 | Framework of the Bayesian-SWMM source identification method.

(5) *Parameter sampling*: Use the proposal distribution to sample a new test parameter X^* from the current parameter state $X_0 = X_t$. Repeat step (3) to calculate the likelihood function $p(Y|X^*)$ for the new sample X^*.
(6) *Acceptance check*: Draw a uniform random number u from the interval [0, 1]. If $u \leq a(X_t, X^*)$, accept the new sample and use it as the initial sample for the next iteration. Otherwise, reject X^* and retain X_t for the next iteration.
(7) *Iterate*: Repeat steps (3)–(5) until the desired number of iterations is reached, to obtain a collection of posterior samples of the discharge source parameters.
(8) *Posterior distribution analysis*: Analyze the posterior samples to infer the posterior distribution of the model parameters.

2.4. Performance evaluation

The effectiveness of the proposed method is evaluated using both the posterior distribution histogram and standard errors, similar to the approach utilized by Shao *et al.* (2021). The posterior distribution histogram acts as a summary statistic for the identified results, thereby illustrating the efficacy of the proposed inference method. Standard errors reveal the discrepancies between numerical predictions and analytical solutions. This study examines two types of standard errors: mean absolute error (MAE) and median absolute error (MedAE). The MAE and MedAE are defined as the mean and median relative differences between the numerical predictions and analytical values, respectively.

$$\text{MAE} = \frac{\text{Mean of MCMC sampled values} - \text{True Value}}{\text{True Value}} \times 100\%,$$

$$\text{MedAE} = \frac{\text{Medium of MCMC sampled values} - \text{True Value}}{\text{True Value}} \times 100\%.$$

For the identification process, errors ranging from 1 to 10% were introduced into the simulation data by the addition of white noise. This method aimed to replicate the variability and uncertainty commonly found in real-world data. Analysis revealed that these introduced errors had a minimal impact on the outcomes, indicating that the findings are robust against typical data perturbations encountered in practical scenarios.

2.5. Case study sewer network

The study site is located in Ningbo, China, encompassing a total area of approximately 76.5 hectares. The sewer network system comprises 20 pipe segments, 20 nodes, and 1 outlet. The diameters of the pipes range from 400 to 1,200 mm, and they feature a roughness coefficient of 0.01. The topology of the sewer network, including node numbering, is depicted in Figure 2. The sanitary sewer flow rate at each node was estimated according to the service area. The computed flow rate was input into the system at each node as a steady baseflow during dry weather. In all numerical simulations, the following hypotheses were posited: (1) flow attains a steady state within the sewer system prior to any illegal discharge; (2) the contaminant is assumed to follow a first-order decay reaction with a decay coefficient of 0.25 (Shao *et al.* 2021). Monitoring sensors were installed at downstream nodes to continuously record the contaminant concentration, yielding comprehensive time series data of the pollutant concentration.

In this study, a series of numerical experiments were designed to evaluate the performance of the pollution source identification algorithm under various conditions. Table 1 summarizes the key parameters and scenarios used in these experiments. Specifically, different types of pollution sources (continuous and instantaneous discharges) and multiple monitoring point configurations were considered.

Scenario A tests are conducted for single-point source pollution. For example, as the minimum time interval for inputting pollution time series in SWMM is 1 min, it is assumed that on a given day at 0:20 a.m., a pollution event occurs instantaneously at an upstream single node 13, where a total of 1,000 g of pollutants are discharged within 1 min for A1–A4. In tests A5–A8, it was assumed that on a certain day at 0:20 a.m., pollutants were continuously discharged at an upstream single node at a rate of 1,000 g/min for 30 min.

Scenario B tests involved water quality simulations for multiple-point source pollution. For example, in test B3, it was assumed that on a given day at 0:20 a.m., a pollution event occurred instantaneously at node 13, where 1,000 g of pollutants were discharged in 1 min, and simultaneously at node 19, another 1,000 g of pollutants were discharged in 1 min.

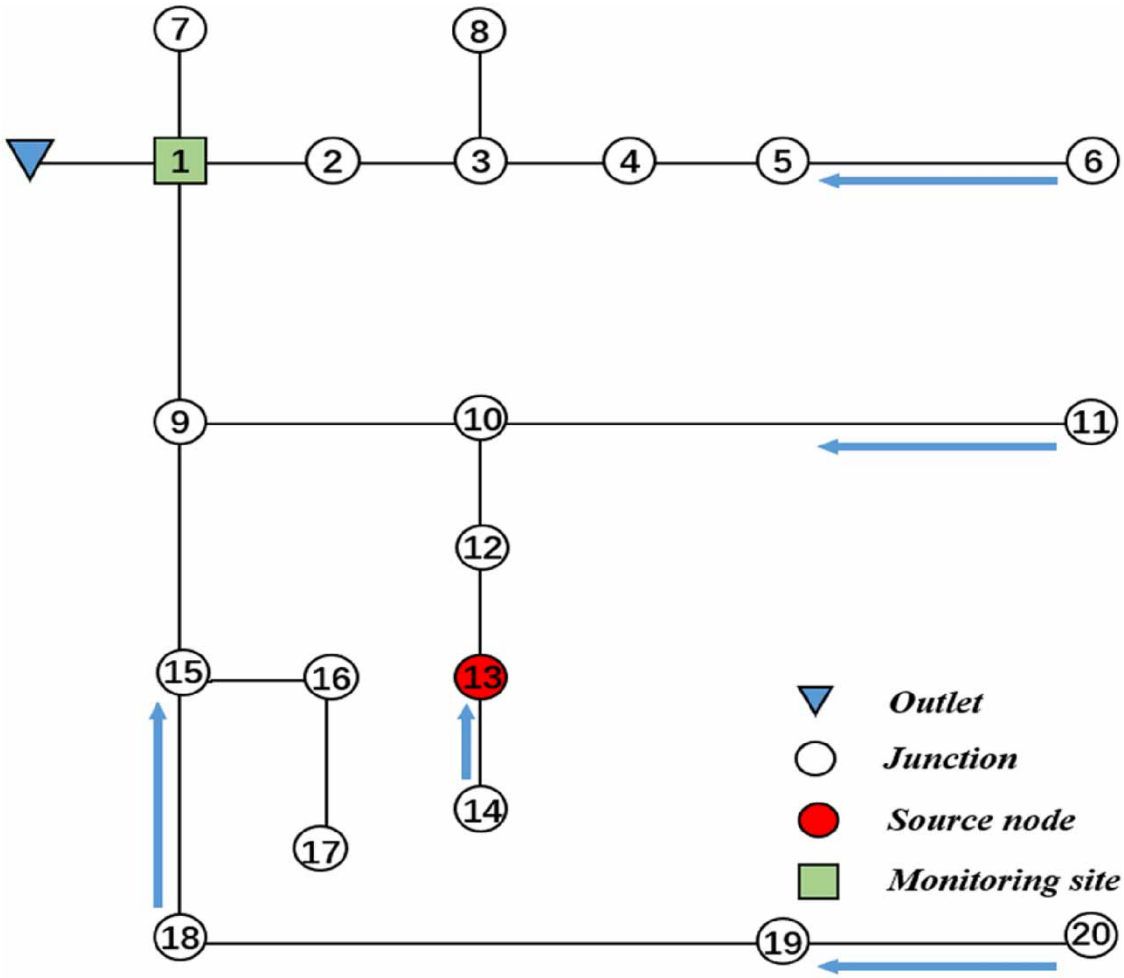

Figure 2 | Schematic of the sewer network layout under study.

3. RESULTS AND DISCUSSION

3.1. Inference of single source

3.1.1. Inference of single parameter

When considering a single unknown pollution source, it is assumed that among the three source parameters (J_x, M, T_d), only one parameter is unknown, while the other two parameters are known. For tests A1–A8, three cases were established to infer each of these three parameters and the monitored concentration profile was used to infer the unknown source parameters. The posterior probability histograms for identifying the discharge nodes of A1–A8 are illustrated in Figure 3. The agreement between the statistical extrapolation of results and the true values demonstrates the applicability and accuracy of the approach. The analysis results show that when there is only one unknown source parameter, J_x, the inference method can effectively explore the entire parameter domain and construct a reasonable posterior distribution for the unknown parameter. Specifically, the posterior distribution is highly concentrated in the region surrounding the true value. During the inference process, the sampling method focuses not only on values with high probabilities but also traverses values with low probabilities, indicating that the proposed sampling method has good global convergence capabilities.

The statistical median and mean of the identified discharge mass and time were calculated and subsequently compared with the actual values in Table 2. Errors associated with the discharge node are not presented, as the node represents a discrete variable and, consequently, the associated errors lack physical significance. Comparisons presented in Table 2 suggest that the Bayesian inference results generally align with the actual solution, indicating the effectiveness of the proposed

Table 1 | Simulated cases with various conditions

Scenario No.	Source type	Discharge mode	J_x		T_d (min)	M (g or g/s)
A1	Single	Instantaneous	5		10, 20, 30	800, 1,000, 1,200
A2			13		10, 20, 30	800, 1,000, 1,200
A3			16		10, 20, 30	800, 1,000, 1,200
A4			19		10, 20, 30	800, 1,000, 1,200
A5		Continuous	5		10, 20, 30	800, 1,000, 1,200
A6			13		10, 20, 30	800, 1,000, 1,200
A7			16		10, 20, 30	800, 1,000, 1,200
A8			19		10, 20, 30	800, 1,000, 1,200
B1	Multiple	Instantaneous	5	13	20	1,000
B2			5	16		
B3			13	19		
B4		Continuous	5	13	20	1,000
B5			5	16		
B6			13	19		

method. As reported by Shao *et al.* (2021), for the inverse problems with only one unknown parameter of a single source, inference utilizing the Bayes-MCMC method remains valid and accurate irrespective of whether the illicit source discharges instantaneously or continuously into the sewer system. The efficacy of the Bayes-MCMC method in thoroughly searching the parameter space is contingent upon the transfer probability density function and its sampling step size. The selection of the transfer probability function distribution is governed by the specific problem at hand, as there is no universal standard; thus, detailed problem-specific analysis is required.

3.1.2. Inference of multiple unknown parameters

When confronted with the absence of details for all three parameters related to the illicit discharge source, the scenario becomes complex owing to the myriad potential parameter combinations, varied hydraulic conditions, and other influential factors. Such complex interactions may yield similar concentration profiles at monitoring points, considerably diminishing the accuracy of concurrent inference for all three unknown parameters, in contrast to the identification in scenarios with only one parameter. The efficacy of Bayesian inference hinges significantly on the likelihood function, given its direct correlation with the number of unknown variables, as shown in Equation (6). An increase in unknowns is anticipated to adversely affect the accuracy of the determined results.

As illustrated in Figure 4, when identifying the three unknown parameters for scenarios A2 and A4, the identified range of results significantly increases, compared to the inference results in Figure 3. The posterior probability density distributions for the three simultaneously identified unknown parameters are dispersed, complicating the task of drawing definitive conclusions about the discharge source compared to the results depicted in Figure 3. Typically, the highest probability does not exceed 0.25, which is significantly lower than the probabilities calculated in single-parameter cases.

To quantitatively evaluate convergence, an analysis of the standard error of the identification results is conducted, as presented in Table 3. Compared to Table 2, both the mean error and median error are larger than in the case where only one unknown parameter is identified. According to the governing equation of the pollution transport process, the monitored concentration at any node is a function of the initial concentration, and travel time, which can be expressed in terms of travel distance and flow velocity. For a sewer network, there are numerous possible combinations of these three parameters that produce the same monitoring readings. Consequently, Bayesian inference fails to effectively discriminate among these sources.

3.1.3. Identification by adding monitoring sites

As outlined in the preceding section, pinpointing a source with three unknowns presents a significant challenge. The initial sampling generated by the Bayesian-MCMC algorithm is random. The initial point determines where the algorithm

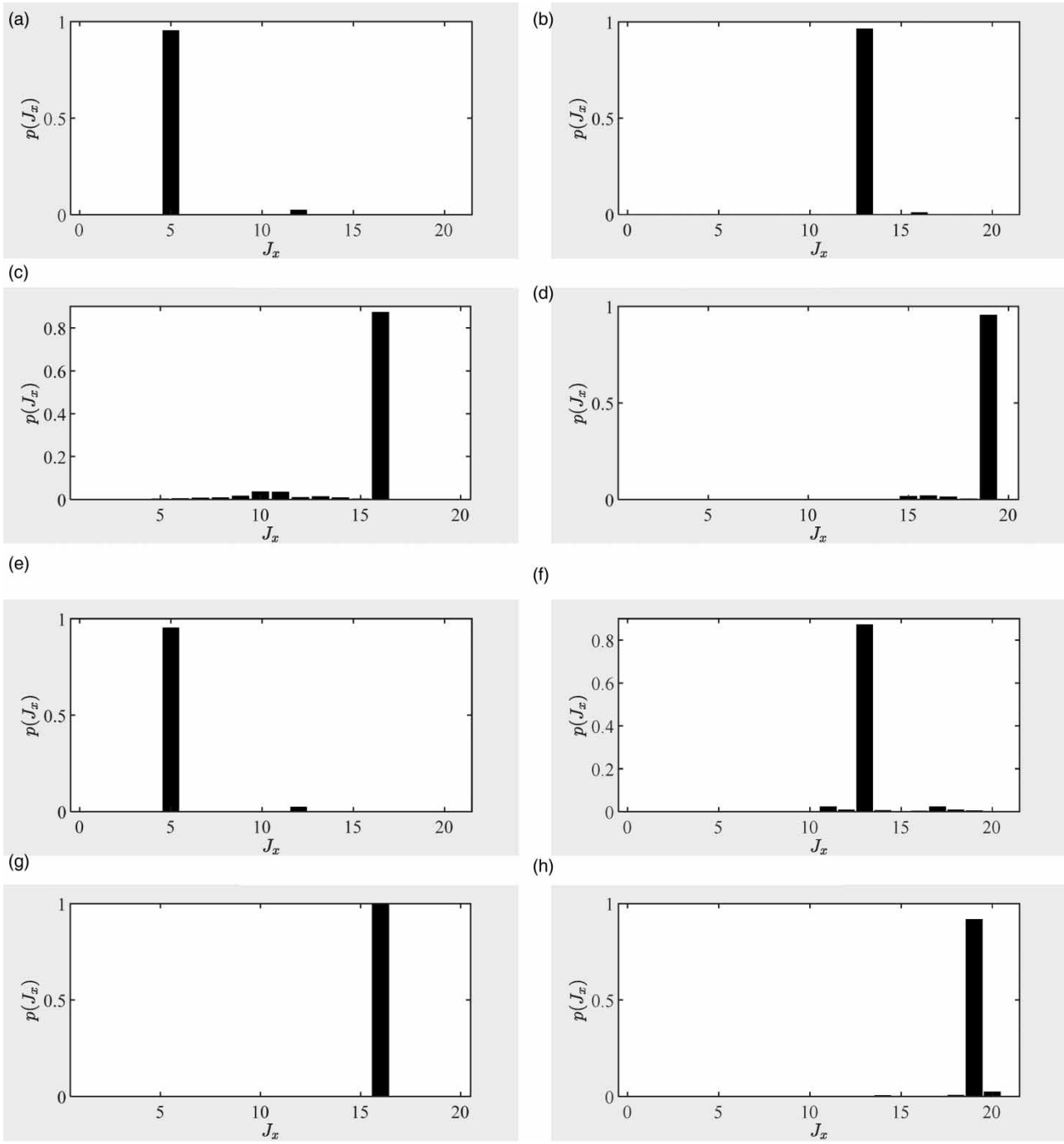

Figure 3 | Posterior probability histograms for the discharge node identification: (a) A1, (b) A2, (c) A3, (d) A4, (e) A5, (f) A6, (g) A7, and (h) A8.

commences its iterations, which significantly influences the iterative process of the MCMC algorithm and may lead to instability sometimes. The addition of a new monitoring site upstream of node 1 allows for more targeted data collection. The data help ascertain whether pollution plumes from non-hotspot areas are likely to pass through or bypass the new monitoring node. This targeted approach significantly reduces uncertainty in our model, thereby enhancing both the precision and reliability of the inferential algorithm. Utilizing the sewer network's topology in conjunction with the frequency statistics of potential nodes offers a straightforward and efficient strategy.

Table 2 | Comparison of the predictions and true values when inferring one unknown parameter

			The proposed SWMM-Bayesian predictions			
Scenario No.	**Parameter**	**True value**	**Mean value**	**MAE (%)**	**Median value**	**MedAE (%)**
A1–A4	T_d (min)	10	10.11	1.1	10.01	0.1
		20	20.12	0.6	20.006	0.03
		30	30.51	1.7	30.15	0.5
A5–A8	T_d (min)	10	10.181	1.81	10.002	0.02
		20	20.202	1.01	20	0
		30	30.48	1.6	30	0
A1–A4	M (g)	800	806.8	0.85	800	0
		1,000	1002	0.2	1,001.3	0.13
		1,200	1,219.2	1.6	1,204.8	0.4
A5–A8	M (g/min)	800	804.56	0.57	800.96	0.12
		1,000	1,000.1	0.01	100.1	0.01
		1,200	1,204.92	0.41	1,200.24	0.02

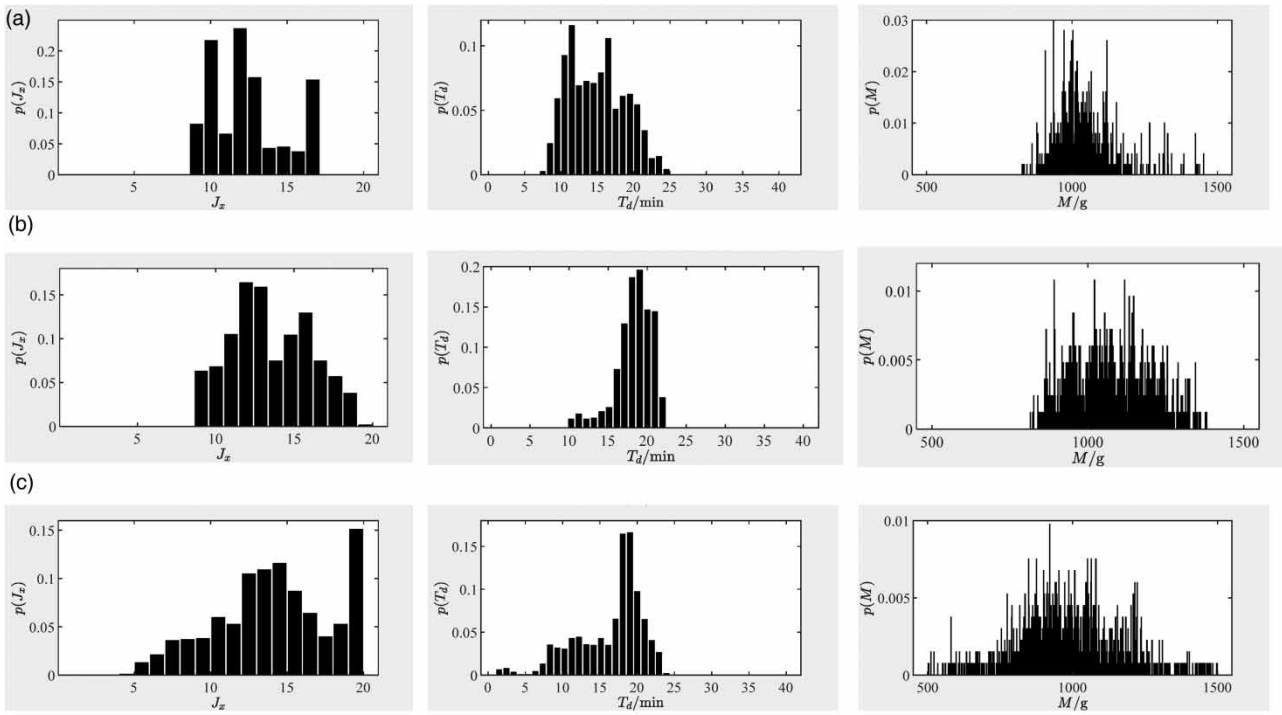

Figure 4 | Histograms of the posterior probability of simultaneously identifying the three parameters of one pollutant source: (a) A2, (b) A3, and (c) A4.

The frequency of the discharge node falling in each region of the prior information is counted. In B1, the probability that the true discharge node falls between nodes 9 and 16 is the largest; in B2, the probability that the true discharge node falls between nodes 9 and 18 is the largest. In B3, the probability that the true discharge node is between nodes 9 and 20 is the largest.

A new monitoring site was established at node 9 to identify the three source parameters for cases A2–A4. To refine the results, additional Bayesian-MCMC inference was conducted, focusing on the nodes that showed large probability distributions in the initial analysis, specifically targeting nodes 10–20. The results are shown in Figure 5. When compared with

Table 3 | Comparison of predicted and true values when inferring three unknown parameters for A2–A4

Scenario No.	Parameter	True value	The proposed SWMM-Bayesian predictions			
			Mean value	MAE (%)	Median value	MedAE (%)
A2	T_d (min)	20	14.866	25.67	15.396	23.02
	M (g)	1,000	1,081	8.1	1,140	11.4
A3	T_d (min)	20	17.27	13.65	18.5	7.5
	M (g)	1,000	1,118	11.8	1,071	7.1
A4	T_d (min)	20	16.792	16.04	18.18	9.1
	M (g)	1,000	858	14.2	899.1	10.09

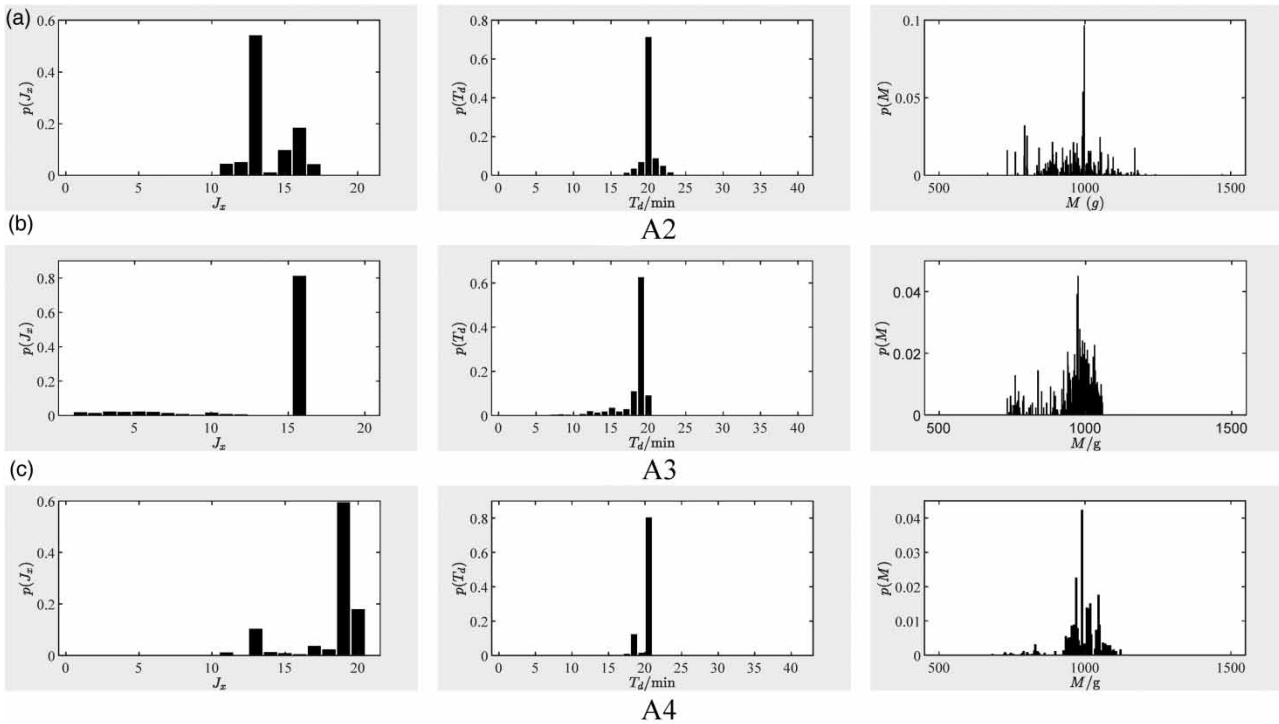

Figure 5 | Histograms of the posterior probability of the parameters for A2–A4: (a) A2, (b) A3, and (c) A4.

Table 4 | Comparison of predicted and actual values of pollution source parameters for A2–A4

Scenario No.	Parameter	True value	The proposed SWMM-Bayesian predictions			
			Mean value	MAE (%)	Median value	MedAE (%)
A2	T_d (min)	20	20.44	2.2	20.04	0.2
	M (g)	1,000	956.9	4.31	973	2.7
A3	T_d (min)	20	19.67	1.65	19.84	0.8
	M (g)	1,000	973.7	2.63	988.4	1.16
A4	T_d (min)	20	19.46	2.7	20	0
	M (g)	1,000	1,075	7.5	1009	0.9

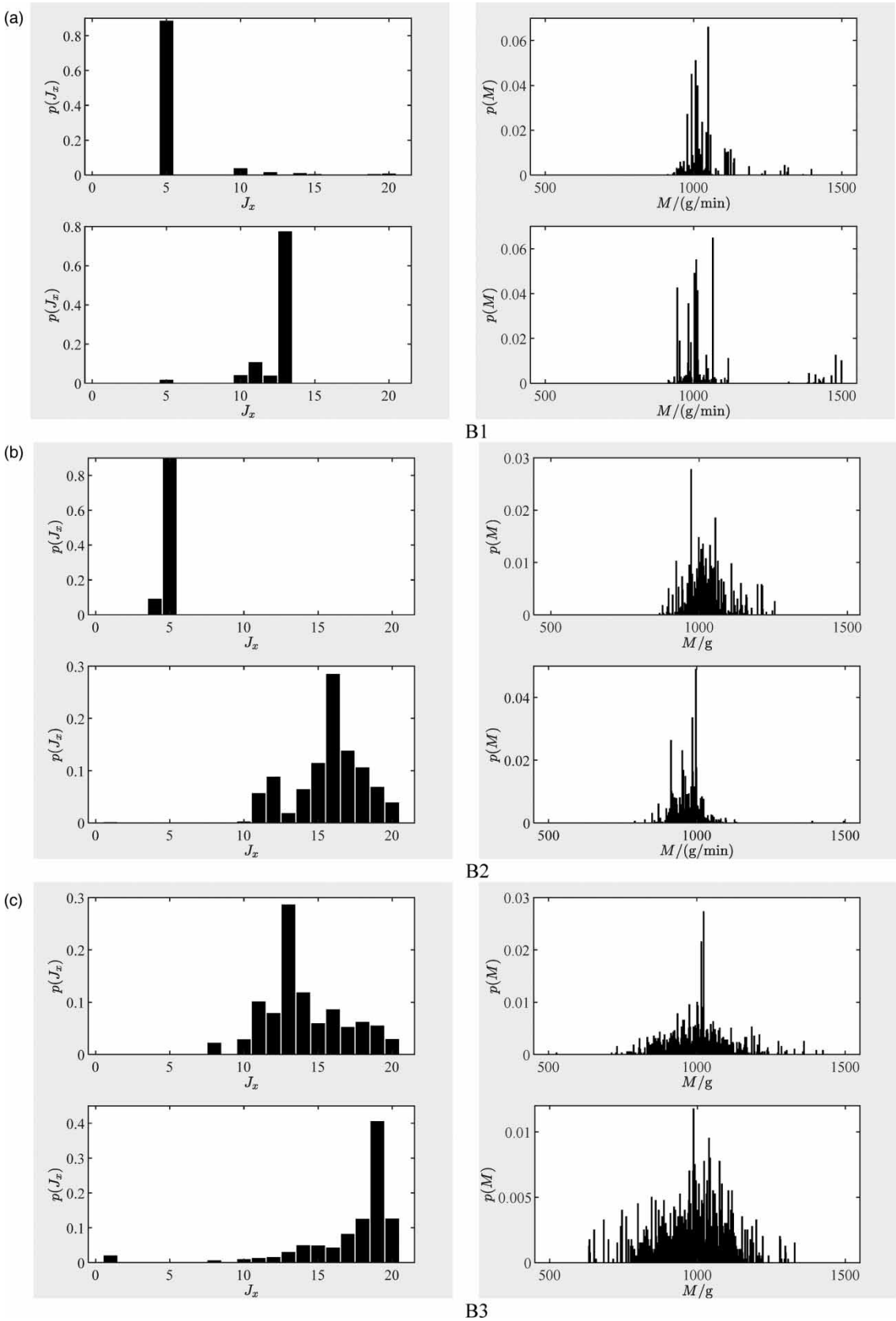

Figure 6 | Histograms of the posterior probability of the parameters: (a) B1, (b) B2, and (c) B3.

the results from Figure 4, a significant reduction in uncertainty for all parameters is clearly evident. This demonstrates that the additional monitoring site significantly enhances the Bayesian inference performance, particularly in pinpointing the source location. Table 4 shows the statistics of the identified results with the actual values. Both the median and mean errors are significantly reduced with the inclusion of one additional monitoring site. This validates the proposed source inference method's ability to identify the illicit discharge, given that adequate monitoring data are available.

3.2. Inference of multiple sources

Due to the possibility of multiple illicit sources in the sewer network, it is necessary to extend the approach to include multi-node source scenarios. Previous studies found that the accuracy of source identification decreases significantly as the number of pollution point sources increases (Yang *et al.* 2016; Wang *et al.* 2018; Wang *et al.* 2021). The following presents the case of identifying two pollution sources. The posterior probability histograms for identifying the discharge node and discharge mass of B1–B3 are illustrated in Figure 6.

Traditional M–H algorithms for random sampling of more complex parameter spaces are usually unable to search over all ranges. Therefore, based on inverse problems with potentially nonunique solutions, it is incomplete and perhaps even misleading to solve the inverse problem based on samples generated by random wandering that fall into local optimal solutions. To avoid the result of the local optimal solution, a scheme is designed to make it escape from the local optimum when it falls into the local optimum, and iterations are performed several times using different initial values to increase the chance of finding the global optimal solution (Zhu *et al.* 2009; Yang *et al.* 2018). Each iteration may converge to a different local optimal solution, but the probability of finding the global optimal solution can be improved by multiple attempts, and the specific improved MCMC method is as follows: record the likelihood of each sample's simulated concentration value and the observed value; in the iterative process, if the current candidate sampling value is not accepted, the sample with the largest likelihood will be used as the initial value for the next iteration; use different initial values for multiple iterations, so that the calculation can be directly used to find a more accurate initial value, thus increasing the likelihood of finding the global optimum and avoiding the large fluctuations in the Markov chain.

For the case of multiple-point sources, it is assumed that the number of sources is 2. Pollutants are discharged simultaneously at two upstream nodes, and the monitoring of the downstream concentration change process is carried out at node 1. The unknown parameter discharge nodes J_x and discharge mass M of the two sources are simultaneously reasoned using the improved SWMM-Bayesian algorithm. It can be seen in the *a posteriori* histogram in Figure 6 that both unknown parameters converge to the true value.

The SWMM-Bayesian improvement algorithm is applied to the scenario of two pollution sources occurring in the drainage network system, and the unknown parameter emission nodes and emission intensities are traced back to the working conditions B4–B6; the *a posteriori* probability histograms are obtained as shown in Figure 6. According to the error statistics in Table 5, the errors of the results of inverse pollutant discharge mass do not exceed 10%, which is enough to prove that the proposed Bayesian-MCMC improvement method has global convergence ability and effectiveness, and is able to converge well near the real value in both the instantaneous discharge mode and the continuous discharge mode. In practical applications, whether the discharge nodes can be traced back accurately or not is crucial, and the improvement method can provide the information of the discharge node efficiently (Figure 7 and Table 6).

Table 5 | Comparison of predicted and actual values of pollution source parameters for B1–B3

			The proposed SWMM-Bayesian predictions			
Scenario No.	**Parameter**	**True value**	**Mean value**	**MAE (%)**	**Median value**	**MedAE (%)**
B1	M (g)	1,000	1,000	0	1,006.5	0.65
			1,054	5.4	1,044	4.4
B2	M (g)	1,000	1,013	1.3	1,013	1.3
			917.2	8.28	968.7	3.13
B3	M (g)	1,000	984.7	1.53	992	0.9
			989	1.1	986	1.4

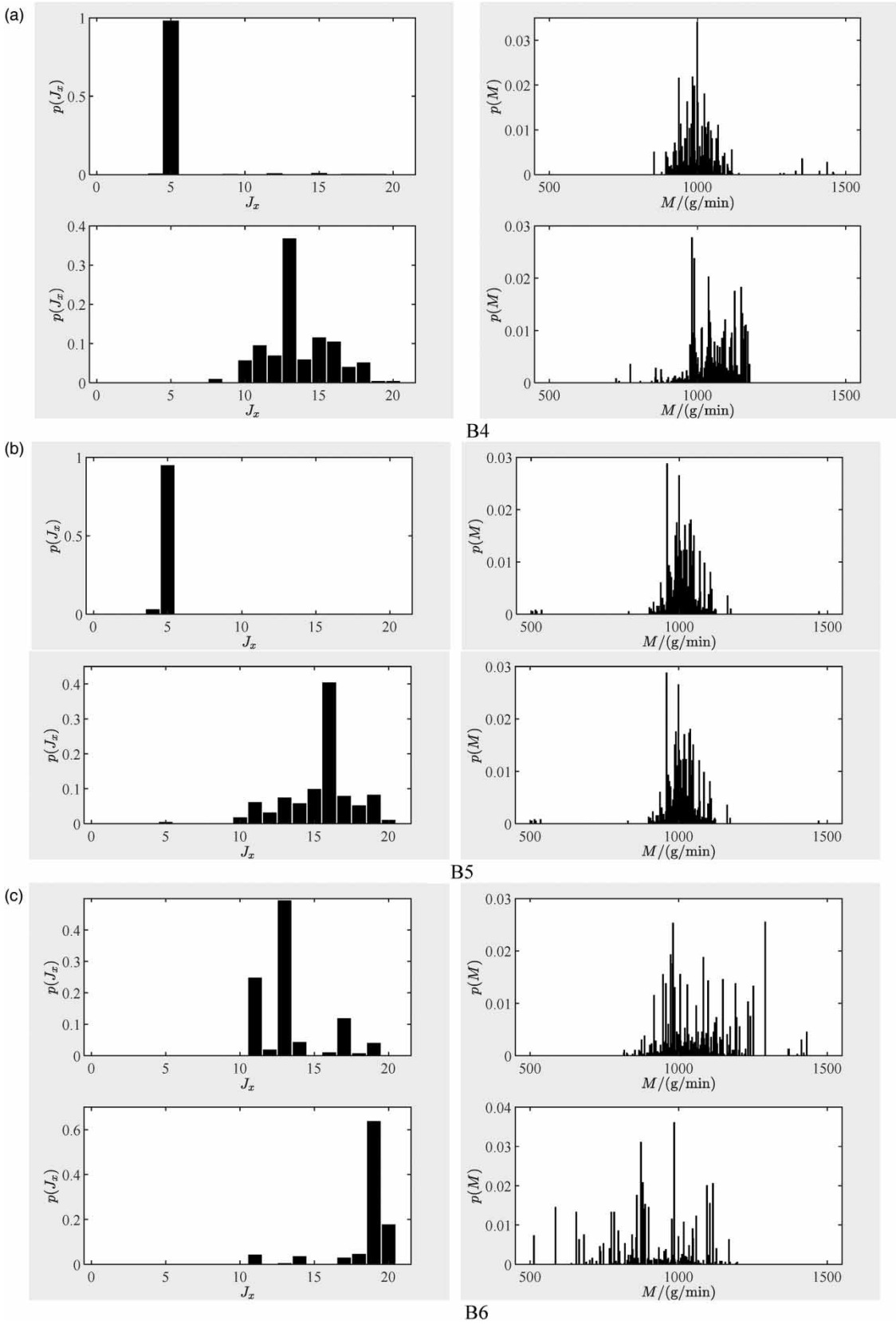

Figure 7 | Histograms of the posterior probability of the parameters: (a) B4, (b) B5, and (c) B6.

Table 6 | Comparison of predicted and actual values of pollution source parameters for B4–B6

			The proposed SWMM-Bayesian predictions			
Scenario No.	**Parameter**	**True value**	**Mean value**	**MAE (%)**	**Median value**	**MedAE (%)**
B4	M (g/min)	1,000	1,036	3.6	1,015	1.5
			1,047	4.7	1,010	1.0
B5	M (g/min)	1,000	1,026	2.6	1,013	1.3
			1,030.2	3.02	1,024.6	2.46
B6	M (g/min)	1,000	1,089.9	8.89	1,098	9.8
			931	6.9	958	4.2

3.3. Discussion

Through the investigation of tests A5–A8, in which the discharge was continuously released for a specific time period, it was found that the identification of the pollution source is as accurate as in instantaneous dumping cases A1–A4. Both instantaneous and continuous pollution releases are deterministic processes for pollution transport and dispersion. The parameters involved in solving the inverse problem do not change substantially, indicating that the accuracy and reliability of the statistical inference achievable in both cases are of the same order. However, it is important to note the assumption of known start times of emissions in continuous pollution release cases.

The effectiveness of introducing an additional monitoring site for single-point source identification problems has been confirmed by prior research (Wang *et al.* 2018; Shao *et al.* 2021). The main reason is that it facilitates the reconstruction of the one-dimensional dispersion process of pollutants within the sewer network. As elucidated by Fischer (1968) with regard to the 'routing procedure', the concentration of a dispersing cloud is channeled through the sewer network, analogous to the routing of a flood. Given that the entire process is deterministic, a comparison of the upstream and downstream profiles can subsequently be employed to resolve the inverse problem. Consequently, the presence of two monitoring sites enables the reconstruction of the physical process via the concentration profiles. Therefore, in cases involving a single pollution source, the identification of the illicit discharge can be ascertained given two strategically positioned monitoring sites. For instance, the potential hotspot area can be delineated using frequency statistics. This information, when integrated with an analysis of the sewer network topology, allows for the determination of the subsequent monitoring station, thereby enabling an update to the statistical inference results. This iterative process is continued until information pertaining to all pollution sources is ascertained with the requisite degree of confidence.

In the case of identification challenges with multiple-point sources, the effectiveness of the SWMM-Bayesian identification method is limited not only by the multiple combinations of parameters but also by the superposition of the process lines of discharge concentrations at multiple points, resulting in inferred results that are easily limited to local optimal solutions. Thus, the normal distribution is employed as the proposal distribution for MCMC sampling to enhance algorithmic efficiency. In addition, the likelihood of each sample relative to observed values is recorded. During the iterative process, if the current candidate sample is not accepted, the sample with the highest likelihood is selected as the initial value for the subsequent iteration. Repeating iterations using these more accurate initial values increases the probability of identifying the global optimum and reduces substantial fluctuations in the early stages of the Markov chain. Consequently, the SWMM-Bayesian method shows good agreement with the true values.

In the present study, an idealized steady-state flow was assumed to simplify the initial analysis and model development. However, the importance of considering unsteady flow conditions should be noted for future investigations to more accurately reflect the dynamic nature of sewer flows. In addition, the limitations of the proposed strategy for placing upstream monitoring points must be acknowledged, particularly in scenarios where sudden pollution events occur, and the pollution plume may have already traversed these points by the time they are identified.

4. CONCLUSIONS

A SWMM-Bayesian source identification method was developed and successfully applied in a sewer network. Given a set of monitoring concentration data, the proposed method combines the Bayesian inference theory with an MCMC sampling

method to produce probability distributions of the unknown source parameters. By coupling Bayesian-MCMC inference with SWMM simulations, the dispersion process of the unknown source can be accurately reconstructed.

In the context of identifying a single illicit discharge, this method accurately determines the true value of unknown parameters. It was found that the accuracy of the proposed method for both instantaneous and continuous discharge scenarios is satisfactory. In practical scenarios where pertinent information regarding the illicit discharge is often unknown, the precision of identification significantly decreases, resulting in merely an approximate determination of the source range. This limitation can be mitigated by strategically placing additional water quality monitoring stations in areas of high probability.

In scenarios involving multiple sources, the reduced precision of the method is attributable not only to the myriad combinations of undetermined parameters but also to the superimposition of concentration profiles. To address this challenge, an improved sampling method was employed. This method fully utilizes the likelihood function to optimize the initial value of the iterative process, avoids interference from local optimal solutions, and demonstrates that the proposed improvement enhances the global search capability.

FUNDING

The study was supported by the National Key R&D Program of China (2022YFC3203200) and the Key Research and Development Program of Ningbo (2023Z216).

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Alapati, S. & Kabala, Z. J. 2000 Recovering the release history of a groundwater contaminant using a non-linear least-squares method. *Hydrological Processes* **14** (6), 1003–1016.

Banik, B. K., Alfonso, L., Di Cristo, C., Leopardi, A. & Mynett, A. 2017a Evaluation of different formulations to optimally locate sensors in sewer systems. *Journal of Water Resources Planning and Management* **143** (7), 04017026.

Banik, B. K., Di Cristo, C., Leopardi, A. & de Marinis, G. 2017b Illicit intrusion characterization in sewer systems. *Urban Water Journal* **14** (4), 416–426.

Berglund, E. Z., Pesantez, J. E., Rasekh, A., Shafiee, M. E., Sela, L. & Haxton, T. 2020 Review of modeling methodologies for managing water distribution security. *Journal of Water Resources Planning and Management* **146** (8), 03120001.

Brown, E., Caraco, D. & Pitt, R. 2004 *Illicit Discharge Detection and Elimination: A Guidance Manual for Program Development and Technical Assessments*. Water Permits Division, Office of Water and Wastewater, US Environmental Protection Agency, Seattle, WA, USA.

Chen, H., Teng, Y., Wang, J. & Song, L. 2012 A framework for pollution characteristic assessment and source apportionment of heavy metal contaminants in riverbed sediments: A case study. *Fresenius Environmental Bulletin* **21** (5), 1110–1117.

Cheng, W. P. & Jia, Y. 2010 Identification of contaminant point source in surface waters based on backward location probability density function method. *Advances in Water Resources* **33** (4), 397–410.

Fischer, H. B. 1968 Dispersion predictions in natural streams. *Journal of the Sanitary Engineering Division* **94** (5), 927–943.

Grbčić, L., Kranjčević, L. & Družeta, S. 2021 Machine learning and simulation-optimization coupling for water distribution network contamination source detection. *Sensors* **21** (4), 1157.

Hachad, M., Lanoue, M., Duy, S. V., Villemur, R., Sauvé, S., Prévost, M. & Dorner, S. 2022 Locating illicit discharges in storm sewers in urban areas using multi-parameter source tracking: Field validation of a toolbox composite index to prioritize high risk areas. *Science of the Total Environment* **811**, 152060.

Hart, W. E. & Murray, R. 2010 Review of sensor placement strategies for contamination warning systems in drinking water distribution systems. *Journal of Water Resources Planning and Management* **136** (6), 611–619.

Hastings, W. K. 1970 Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57** (1), 97–109.

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, S., Zhang, Y., Wang, P. & Zheng, M. 2013 An almost-parameter-free harmony search algorithm for groundwater pollution source identification. *Water Science and Technology* **68** (11), 2359–2366.

Kass, R. E., Carlin, B. P. & Neal, G. R. M. 1998 Markov chain Monte Carlo in practice: A roundtable discussion. *American Statistician* **52** (2), 93–100.

Kessler, A., Ostfeld, A. & Sinai, G. 1998 Detecting accidental contaminations in municipal water networks. *Journal of Water Resources Planning and Management* **124** (4), 192–198.

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** (13), 4630–4638.

Laird, C. D., Biegler, L. T. & van Bloemen Waanders, B. G. 2006 Mixed-integer approach for obtaining unique solutions in source inversion of water networks. *Journal of Water Resources Planning and Management* **132** (4), 242–251.

Levenspiel, O. 2011 *Tracer Technology: Modeling the Flow of Fluids*, Vol. 96. Springer Science & Business Media, New York, NY, USA.

Li, T., Winnel, M., Lin, H., Panther, J., Liu, C., O'Halloran, R., Wang, K., An, T., Wong, P. K., Zhang, S. & Zhao, H. 2017 A reliable sewage quality abnormal event monitoring system. *Water Research* **121**, 248–257.

Liu, S., Che, H., Smith, K., Lei, M. & Li, R. 2015 Performance evaluation for three pollution detection methods using data from a real contamination accident. *Journal of Environmental Management* **161**, 385–391.

Moghaddam, M. B., Mazaheri, M. & Samani, J. M. V. 2021 Inverse modeling of contaminant transport for pollution source identification in surface and groundwaters: A review. *Groundwater for Sustainable Development* **15**, 100651.

Neupauer, R. M. & Wilson, J. L. 1999 Adjoint method for obtaining backward-in-time location and travel time probabilities of a conservative groundwater contaminant. *Water Resources Research* **35** (11), 3389–3398.

Ostfeld, A. & Salomons, E. 2005 Optimal early warning monitoring system layout for water networks security: Inclusion of sensors sensitivities and response delays. *Civil Engineering and Environmental Systems* **22** (3), 151–169.

Ramin, P., Libonati Brock, A., Polesel, F., Causanilles, A., Emke, E., de Voogt, P. & Plosz, B. G. 2016 Transformation and sorption of illicit drug biomarkers in sewer systems: Understanding the role of suspended solids in raw wastewater. *Environmental Science & Technology* **50** (24), 13397–13408.

Riaño-Briceño, G., Barreiro-Gomez, J., Ramirez-Jaime, A., Quijano, N. & Ocampo-Martínez, C. 2016 MatSWMM – An open-source toolbox for designing real-time control of urban drainage systems. *Environmental Modelling & Software* **83**, 143–154.

Rossman, L. A. & Huber, W. C. 2016 *Storm water management model reference manual volume III–water quality*. US Environmental Protection Agency, Cincinnati, OH, USA.

Shao, Z., Xu, L., Chai, H., Yost, S. A., Zheng, Z., Wu, Z. & He, Q. 2021 A Bayesian-SWMM coupled stochastic model developed to reconstruct the complete profile of an unknown discharging incidence in sewer networks. *Journal of Environmental Management* **297**, 113211.

Skaggs, T. H. & Kabala, Z. J. 1994 Recovering the release history of a groundwater contaminant. *Water Resources Research* **30** (1), 71–79.

Wang, H. & Harrison, K. W. 2013 Bayesian approach to contaminant source characterization in water distribution systems: Adaptive sampling framework. *Stochastic Environmental Research and Risk Assessment* **27**, 1921–1928.

Wang, H. & Harrison, K. W. 2014 Improving efficiency of the Bayesian approach to water distribution contaminant source characterization with support vector regression. *Journal of Water Resources Planning and Management* **140** (1), 3–11.

Wang, J., Zhao, J., Lei, X. & Wang, H. 2018 New approach for point pollution source identification in rivers based on the backward probability method. *Environmental Pollution* **241**, 759–774.

Wang, J., Liu, J., Wang, B., Cheng, W. & Zhang, J. 2021 A new method for multi-point pollution source identification. *Atmospheric and Oceanic Science Letters* **14** (6), 100098.

Williams, B., Christensen, W. F. & Reese, C. S. 2011 Pollution source direction identification: Embedding dispersion models to solve an inverse problem. *Environmetrics* **22** (8), 962–974.

Wu, H. & Chen, Q. 2023 An integrated approach using multi-source data for effective pollution risk monitoring of urban rivers: A case study of Hangzhou. *Water Science & Technology* **88** (2), 454–467.

Wu, W., Ren, J., Zhou, X., Wang, J. & Guo, M. 2020 Identification of source information for sudden water pollution incidents in rivers and lakes based on variable-fidelity surrogate-DREAM optimization. *Environmental Modelling & Software* **133**, 104811.

Xu, Z., Qu, Y., Wang, S. & Chu, W. 2021 Diagnosis of pipe illicit connections and damaged points in urban stormwater system using an inversed optimization model. *Journal of Cleaner Production* **292**, 126011.

Yang, Y. J., Haught, R. C. & Goodrich, J. A. 2009 Real-time contaminant detection and classification in a drinking water pipe using conventional water quality sensors: Techniques and experimental results. *Journal of Environmental Management* **90** (8), 2494–2506.

Yang, H., Shao, D., Liu, B., Huang, J. & Ye, X. 2016 Multi-point source identification of sudden water pollution accidents in surface waters based on differential evolution and Metropolis–Hastings–Markov Chain Monte Carlo. *Stochastic Environmental Research and Risk Assessment* **30**, 507–522.

Yang, H. D., Liu, B. Y. & Huang, J. H. 2018 Forecast model parameters calibration method for sudden water pollution accidents based on improved Bayesian-Markov chain Monte Carlo. *Control and Decision* **33** (4), 679–686.

Yee, E. & Flesch, T. K. 2010 Inference of discharging rates from multiple sources using Bayesian probability theory. *Journal of Environmental Monitoring.* **12** (3), 622–634.

Zhu, S., Mao, G., Liu, G. & Huang, Y. 2009 Improved MCMC method and its application. *Journal of Hydraulic Engineering* **39** (Z2), 1019–1023.

First received 19 March 2024; accepted in revised form 30 June 2024. Available online 11 July 2024