frontiers | Frontiers in Environmental Science

TYPE Original Research
PUBLISHED 15 May 2024
DOI 10.3389/fenvs.2024.1401942

Intelligent optimal layout of drainage pipe network monitoring points based on information entropy theory

Min He, Yibo Zhang, Zhaoxi Ma* and Qin Zhao

Department of Civil Engineering and Architecture, Xi'an University of Technology, Xi'an, China

OPEN ACCESS

EDITED BY
Ming Zhong,
Sun Yat-sen University, China

REVIEWED BY
Ying Zhu,
Xi'an University of Architecture and Technology, China
Xiang Zhang,
Wuhan University, China

*CORRESPONDENCE
Zhaoxi Ma,
1200710004@stu.xaut.edu.cn

RECEIVED 18 March 2024
ACCEPTED 26 April 2024
PUBLISHED 15 May 2024

CITATION
He M, Zhang Y, Ma Z and Zhao Q (2024), Intelligent optimal layout of drainage pipe network monitoring points based on information entropy theory.
Front. Environ. Sci. 12:1401942.
doi: 10.3389/fenvs.2024.1401942

The rapid expansion of urban drainage pipe networks, driven by economic development, poses significant challenges for efficient monitoring and management. The complexity and scale of these networks make it difficult to effectively monitor and manage the discharge of urban domestic sewage, rainwater, and industrial effluents, leading to illegal discharges, leakage, environmental pollution, and economic losses. Efficient management relies on a rational layout of drainage pipe network monitoring points. However, existing research on optimal monitoring point layout is limited, primarily relying on manual analysis and fuzzy clustering methods, which are prone to human bias and ineffective monitoring data. To address these limitations, this study proposes a coupled model approach for the automatic optimization of monitoring point placement in drainage pipe networks. The proposed model integrates the information entropy index, Bayesian reasoning, the Monte Carlo method, and the stormwater management model (SWMM) to optimize monitoring point placement objectively and measurably. The information entropy algorithm is utilized to quantify the uncertainty and complexity of the drainage pipe network, facilitating the identification of optimal monitoring point locations. Bayesian reasoning is employed to update probabilities based on observed data, while the Monte Carlo method generates probabilistic distributions for uncertain parameters. The SWMM is utilized to simulate stormwater runoff and pollutant transport within the drainage pipe network. Results indicate that (1) the relative mean error of the parameter inversion simulation results of the pollution source tracking model is linearly fitted with the information entropy. The calculation shows that there is a good positive linear correlation between them, which verifies the feasibility of the information entropy algorithm in the field of monitoring node optimization; (2) the information entropy algorithm can be well applied to the optimal layout of a single monitoring node and multiple monitoring nodes, and it can correspond well to the inversion results of the tracking model parameters; (3) the constructed monitoring point optimization model can well realize the optimal layout of monitoring points of a drainage pipe network. Finally, the pollution source tracking model is used to verify the effectiveness of the optimal layout of monitoring points, and the whole process has less human participation and a high degree of automation. The automated monitoring point optimization layout model proposed in this study has been successfully applied in practical cases, significantly improving the efficiency of urban drainage network monitoring and reducing the degree of manual participation, which has important practical significance for improving the level of urban water environment management.

KEYWORDS

pipe network, optimization of monitoring points, information entropy, Bayesian reasoning, pollution source tracking

1 Introduction

With the continuous expansion of the urban scale, the complexity of urban drainage pipe networks is also increasing. Enhancing the monitoring of these networks is crucial for protecting drainage facilities and enhancing urban information management (Shi et al., 2021). The drainage pipe network plays a vital role in urban stormwater management by ensuring water safety, reducing flood control pressure, alleviating water scarcity, and enhancing water quality (Mirauda and Ostoich, 2020; Groothuis et al., 2021). However, with rapid industrialization, environmental pollution issues have become more pronounced, posing new challenges to urban drainage systems. Among them, due to the high cost of wastewater treatment, the phenomenon of illegal discharge or leakage of sewage from the pipeline network occurs from time to time, resulting in a serious threat to the active sludge of the wastewater treatment plant and resulting in a variety of problems, such as the treatment of water quality not being up to the standard and the deterioration of the quality of the discharged water body. Strengthening drainage network monitoring is one of the most effective measures to solve the urban underground drainage problem (Guo et al., 2018; Meyer et al., 2018).

At present, there are many research studies on drainage network monitoring, which mainly focus on drainage network information management and system design (Tang et al., 2018). Wei et al. (2020) studied the establishment of an urban waterlogging pre-warning system; using the terminal monitor placed in the inspection well to obtain the actual water depth of each monitoring point when the rainstorm occurs, the inundation situation of each point can be dynamically displayed on the computer. Omamageswari et al. (2021) developed a real-time monitoring system of drainage pipe networks based on the Internet of Things technology.

In addition to the establishment of online monitoring systems to solve the problem of realizing drainage network information technology, another key point lies in the optimization and layout of the monitoring points. Due to limitations in human, material, and financial resources, the layout goal of monitoring points is to use as few monitoring points as possible to obtain comprehensive, efficient, and accurate information related to the monitoring targets (Liu et al., 2018). At present, there are relatively limited studies on the optimal layout of drainage pipe network monitoring points. Most of the existing optimization methods use the empirical formula approach, which is solved by statistical cluster-based analysis or by constructing a multi-objective planning model. Guo et al. (2022) proposed an automatic machine learning (AutoML) technique based on the genetic algorithm to construct a rapid early warning and comprehensive analysis model for urban flooding. For that, the local topography and waterlogging risk are considered, and the AutoML models can be used in the area without monitoring the water level, quickly predict waterlogging depths, and provide spatial grid results for rapid early warning. Banik et al. (2015) studied the optimization of water quality monitoring points in the drainage pipe network. Under the multi-objective programming model, the genetic algorithm was used, and the optimal water quality monitoring points were determined. Based on entropy theory, combined with multi-objective optimization tools and the numerical model, Yazdi (2017) proposed a location method for new sewage stations of urban drainage networks. The modeling framework provides the best compromise between the maximum possible information and the minimum common information between stations, which is applied to the main surface-water collection system in Tehran to determine the new best monitoring point considering the cost. It should be noted that although machine learning algorithms such as genetic algorithms can quickly complete the processing of pipe network data, the objectives in multi-objective optimization are constrained by each other, and it is easy to produce a locally optimal arrangement scheme.

In summary, the current research on the optimization of monitoring points usually considers the water quality monitoring of the water supply network as the optimization goal. The only research on the optimization arrangement of drainage network monitoring points is mostly concentrated on the development of urban waterlogging monitoring systems, and the research on the optimization of urban drainage pipe network monitoring points aiming at the identification of pollution sources is rarely carried out. In addition, the previous optimization method of drainage pipe network monitoring is mainly for the fuzzy clustering method and the dynamic progress method; although set up on the principle of clustering, but still largely dependent on human judgment, correlation recognition is only an auxiliary function, and more dependent on artificial analysis and understanding of the network topology, the final layout results are heavily influenced by artificial factors. A large number of monitoring data may be invalid (Zhao et al., 2018). In addition, most of the current studies optimize the layout of the monitoring points by calculating only a set of monitoring schemes. Few studies have been carried out on how to quantitatively evaluate the obtained monitoring programs, so it is impossible to quantitatively assess the merits and disadvantages of the obtained monitoring programs (Retzer et al., 2009; Keum et al., 2019). Therefore, making full use of the model to strengthen the simulation of drainage network operation under multiple scenarios can reduce the influence of subjective location uncertainty and realize automatic identification of the monitoring points and quantitative and objective evaluation of the monitoring scheme (Blasone et al., 2007; Wu et al., 2021).

In this study, aiming at the optimization of monitoring points with pollution source identification as the goal, the SWMM pipe network model was established. Under the condition of determining the initial monitoring time, monitoring interval time, and monitoring times, the SWMM model was used for simulation, and the posterior probability density under different monitoring conditions was calculated, and the results were put into the information entropy calculation. Information entropy is used to quantify the information of each node in the drainage network, and the location of the monitoring points is gradually selected. Finally, the monitoring cost and monitoring accuracy are comprehensively considered to complete the selection of the number of monitoring points, and the optimization design process of the monitoring scheme is finally completed. Combining the information entropy theory and the SWMM model, the uncertainty of pollution source identification is quantified, and the problem of excessive reliance on human judgment inherent in the existing methodologies is overcome to realize the optimization of drainage pipe network monitoring points. The results are expected to provide accurate information and data for the related departments to manage drainage pipe networks, improve our understanding of complex pipe network systems, and provide theoretical support for scientific and effective prevention and control of water pollution events.

2 Methodology

2.1 Bayesian reasoning

In the case of Bayesian statistics, all unknown parameters are regarded as random variables, and their distributions are obtained from the known information (Vrugt, 2016). Therefore, Bayesian statistics provides a rigorous method for uncertainty analysis and can provide key information for management decision-making. Bayesian reasoning is based on the following formula (Martino and Elvira, 2018):

$$p(X|Y) = \frac{p(X)p(Y|X)}{p(Y)} = \frac{p(X)p(Y|X)}{\int_X p(X)p(Y|X)dX} \propto p(X)p(Y|X), \tag{1}$$

where $p(X|Y)$ is the *a posteriori* probability distribution function of X, indicating the obtained observation value, and Y is the distribution law of parameter X; $p(X)$ refers to the *a priori* probability distribution function, which represents the distribution law of parameter x obtained from data and experience; $p(Y|X)$ is the likelihood function, which indicates the fitting degree between the model parameters and the observed data (Zhang et al., 2016).

The optimal design of the drainage network monitoring scheme mainly includes the optimization of the number and location of the monitoring well points (Jiang et al., 2021; Sambito and Freni, 2021). Assuming that the initial monitoring time is t and the monitoring value obtained from the monitoring scheme s is still recorded as y, then the Bayesian formula can be rewritten as the following:

$$p(x|y,S) = \frac{p(x|S)p(y|x,S)}{\int p(x|S)p(y|x,S)dx}, \tag{2}$$

Since the prior distribution $p(x|S)$ of the parameters represents a preliminary understanding of the unknown parameters and is not affected by the monitoring design scheme s, that is, $p(x|S) = p(x)$, Equation 2 becomes:

$$p(x|y,S) = \frac{p(x)p(y|x,S)}{\int p(x)p(y|x,S)dx}, \tag{3}$$

The normalization constant $\int p(x)p(y|x,S)dx$ represents the probability of the occurrence of monitoring data y obtained from the monitoring design scheme s, which can be abbreviated as $p(y|S)$, that is:

$$p(y|S) = \int p(x)p(y|x,S)dx, \tag{4}$$

2.2 Optimal design of the monitoring points based on information entropy

In the information theory, information entropy is defined as the measure of uncertainty, which has been widely used to estimate the uncertainty of variables (Li et al., 2012). It can be regarded as the negative expected value of the logarithm of the probability density function of a variable (Parvan, 2010). Among them, the uncertainty definition of a random variable is used to describe the amount of information of the variable. The more information required to describe the variable, the greater the uncertainty of the variable and the greater the information entropy (Monache et al., 2008; Retzer et al., 2009; Mooselu et al., 2020). Based on the above research progress, this paper establishes the SWMM pipe network model; uses the SWMM model to simulate the result under the condition of determining the initial monitoring time, monitoring interval, and monitoring times; takes the information entropy as the evaluation index to calculate the optimal monitoring schemes under different monitoring types; screens and selects the monitoring scheme that takes into account the monitoring cost and monitoring accuracy; and finally, operates with the selected monitoring scheme To summarize, the optimal design of a drainage pipe network monitoring network that can provide accurate information is very important to improve our understanding of complex pipe network systems and reduce heavy and unnecessary capital expenditure (Zhang et al., 2015; Yang et al., 2020). The mathematical expression of information entropy is as follows:

$$H(x) = -\int_a^b f(x)\ln f(x)dx, \tag{5}$$

where x is the random variable; $f(x)$ is the probability density function of a random variable with *a priori* range of [a, b].

The specific coupling calculation process of Python programming of a monitoring point optimization model based on information entropy includes the following five steps: generation of unknown parameters,* inp file modification, extraction and storage of the time series, calculation of the conditional probability function, and calculation of information entropy. The automatic modification of the SWMM model and the extraction of the time series are directly realized in Python program without opening SWMM software. The optimization model is also the coupling calculation of the information entropy algorithm and the SWMM model. The time series extracted by the SWMM model is directly used for the calculation diagram of the likelihood functions $p(y^i|x^i)$ and $p(y|x^i)$. The specific calculation principle of information entropy of the monitoring nodes of a pipe network is as follows:

Based on the above coupling principle, Python language programming is used to directly calculate the SWMM model to extract the pollutant concentration monitoring value y, and the inversion model is used to invert the unknown parameter x, where the posterior probability density function of x is $p(x|y,S)$, s is the monitoring scheme, and the information entropy of the posterior distribution of parameter X is similarly defined as follows:

$$H(S,y) = -\int p(x|y,S)\ln p(x|y,S)dx, \tag{6}$$

The left end of Equation 6 contains the monitoring value y, but due to the influence of the actual conditions, the actual monitoring value y cannot be obtained in the optimization node stage of the monitoring points. Therefore, it is assumed that the monitoring value y is a random variable, and the corresponding probability density function is $p(y|S)$ (Kaynar and Ridder, 2010). Therefore, in order to obtain the calculated information entropy only affected by the monitoring scheme, both sides of Equation 5 are multiplied by $p(y|S)$, and then y is integrated to obtain the expectation of information entropy $H(S,y)$:

$$\begin{aligned} E(H(S,y)) &= -\int\left[\int p(x|y,S)\ln p(x|y,S)dx\right]p(y|S)dy \\ &= -\int\int p(x|y,S)p(y|S)\ln p(x|y,S)dxdy \end{aligned}, \tag{7}$$

According to Equation 7, $E(H(S,y))$ is only affected by the monitoring scheme S, which is recorded as $E(S)$. By understanding the relevant principles of information entropy, we can solve $E(S)$. According to the concept of information entropy, the unknown parameter x is inversed when using the pollutant concentration monitoring value y* obtained from s*. At this time, the information entropy of x posterior distribution is the smallest, indicating that the uncertainty of X is also the smallest, and the inversion effect is the best (Liu et al., 2013; Wang et al., 2016). In Equation 7, the calculation of the solution algorithm is complex, and it is difficult to calculate the explicit result of this expression. Therefore, this paper uses the Monte Carlo method to approximate the solution. First, Equation 7 is rewritten as follows using Equation 2:

$$\begin{aligned} E(H(S,y)) &= -\int\left[\int p(x|y,S)\ln p(x|y,S)dx\right]p(y|S)dy \\ &= -\int\int p(x|y,S)p(y|S)\ln p(x|y,S)dxdy, \end{aligned} \tag{8}$$

The prediction uncertainty of the expected value of information entropy can be divided into two parts as follows:

$$\begin{aligned} E(S) &= U(C) + U(S) \\ &= \left(-\iint p(y|x,S)p(x)\ln p(x)dxdy\right) \\ &+\left(-\iint p(y|x,S)p(x)\left[\ln p(y|x,S) - \ln p(y|S)\right]dxdy\right), \end{aligned} \tag{9}$$

where $E(S)$ represents the total uncertainty of model prediction, $U(C)$ is the parameter uncertainty derived from the inside of the model, and $U(S)$ is the conceptual model uncertainty derived from among the models.

In Equation 9, the prior distribution information entropy of the inversion parameter x is as follows:

$$\begin{aligned} U(C) &= -\iint p(y|x,S)p(x)\ln p(x)dxdy \\ &= -\ln p(x), \end{aligned} \tag{10}$$

According to Equation 10, the larger the prior range of inversion parameter x, the larger the calculated value of information entropy of parameter x, and the greater the uncertainty of the inversion result of parameter X. When the prior probability distribution $p(x)$ remains unchanged, the information entropy $-\ln p(x)$ remains unchanged. Therefore, in order to obtain the minimum value of $E(S)$, the second half is extracted. As shown in Equation 1, only the minimum value of $U(S)$ needs to be calculated to obtain the minimum value of information entropy of the corresponding monitoring scheme (Ainslie et al., 2009; Yan and Zhou, 2019).

$$U(S) = -\iint p(y|x,S)p(x)\left[\ln p(y|x,S) - \ln p(y|S)\right]dxdy, \tag{11}$$

According to Keum et al. (2018), the value of U(S) is less than 0 in equation 11. Therefore, according to the concept of information entropy and formula 8, the use of monitoring points in the Bayesian formula reduces the uncertainty of the inversion parameter X.

Equation 12 is solved by the Monte Carlo method as follows:

$$U(S) \approx -\frac{1}{N}\sum_{i=1}^{N}\left[\ln p(y^i|x^i,S) - \ln p(y^i|S)\right], \tag{12}$$

First, n samples are randomly selected from the unknown parameter x *a priori* distribution $P(X)$, which are recorded as x^i (i = 1, 2, . . ., N). For each i∊N, using Python program to modify the unknown parameters of the SWMM model with samples, the simulation is used to extract the conditional probability density function under the condition of the monitoring scheme s, and then one sample y^i is randomly selected from the function $p(y|x^i,S)$, with a total of n samples. Each group of x^i and y^i is replaced into Equation 1, and $p(y^i|x^i,S)$ is calculated in the conditional probability density function of Equation 12. It can be seen from Equation 5 that in Equation 12, where $p(y^i|S) = \int p(x)p(y^i|x,S)dx$, the Monte Carlo method is still used to calculate this integral.

$$p(y^i|S) \approx -\frac{1}{N}\sum_{j=1}^{N}p(y^i|x^i,S), \tag{13}$$

Therefore, as long as the monitoring design scheme s is fixed, the approximate information entropy of this monitoring scheme can be obtained through equations 8–11.

2.3 MCMC method

The Metropolis–Hastings (MH) algorithm is a typical Markov chain Monte Carlo (MCMC) sampling method. In this study, the Metropolis–Hastings algorithm will be used. Its essence is to construct an appropriate Markov chain and generate the next state under a given state (Abellán et al., 2011; Wang and Chen, 2013). Assuming that the target distribution is $f(x)$, a sample with sample size T will be generated from this distribution. The iterative steps of the Metropolis–Hastings algorithm are shown in Figure 1. $x^{(t)}$ is a vector, which represents the value of the *t*th iteration in the algorithm (Wang and Jin, 2013; Zeng et al., 2016).

It is assumed that the unknown parameters of the model are days, where $\bar{C}_i$, $C_i(\theta)$, and $p(\bar{C}_i|\theta)$ are the measured value, predicted value, and likelihood function of the *i*th measuring point; $\varepsilon = C_i(\theta) - \bar{C}_i$ is the measurement error; and i = 1–6. It is assumed that ε obeys the mean value 0, and the standard deviation is σ. The likelihood function of pollutant traceability, i.e., the conditional probability density function, can be expressed as follows:

$$p(\bar{C}_i|\theta) = \frac{1}{(2\pi\sigma^2)^{6/2}}exp\left(\sum_{i=1}^{6}\left[-\frac{(C_i(\theta) - \bar{C}_i)^2}{2\sigma^2}\right]\right), \tag{14}$$

According to the Bayesian theorem, the posterior probability density function of the model parameters is as follows:

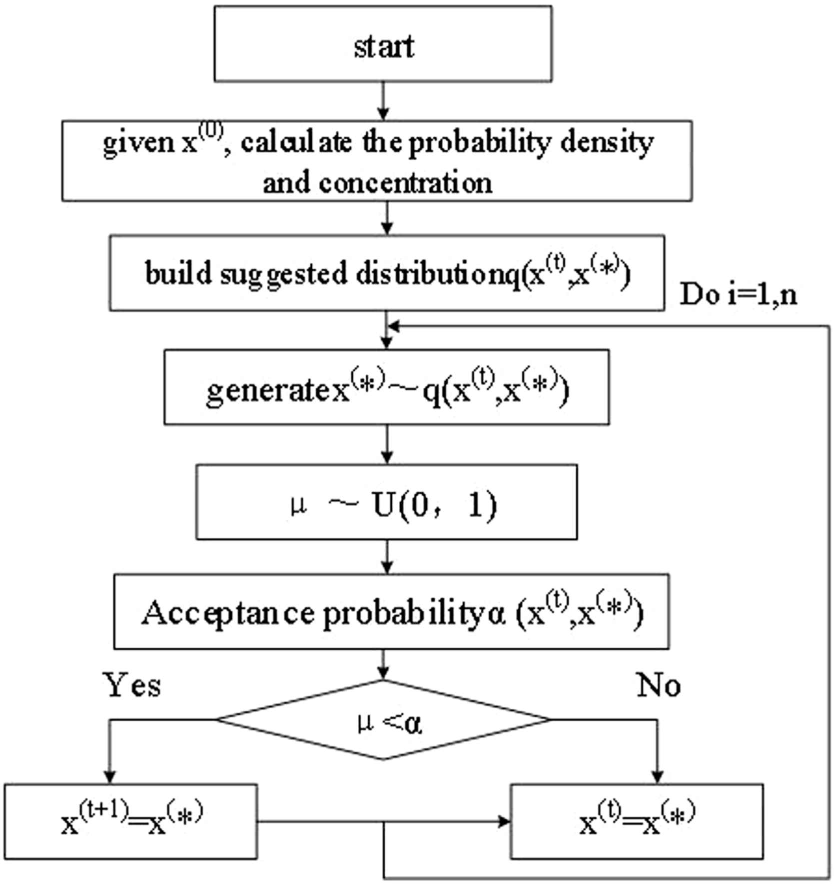

FIGURE 1
Flowchart of the Metropolis-Hastings sampling algorithm.

$$p(\theta|\bar{C}_i) = \frac{p(\theta)}{(2\pi\sigma^2)^{6/2}} exp\left(\sum_{i=1}^{6}\left[-\frac{(C_i(\theta)-\bar{C}_i)^2}{2\sigma^2}\right]\right), \tag{15}$$

3 Results and discussion

3.1 Optimal layout of the monitoring points of a drainage network

3.1.1 Verification of the monitoring optimization effect

There are 32 nodes in the pipe network in the study area. If the optimal combination of different numbers of monitoring points is carried out, there will be $\sum_{i=1}^{32} C_{32}^{i} = (2^{32}-1)$ monitoring schemes. Solving the information entropy of all the alternative monitoring schemes will produce a large workload, the number of pipe network monitoring points will increase, and the monitoring cost will increase exponentially.

In order to reduce the workload and consider the monitoring cost and coverage, on the basis of limiting the five monitoring nodes, this paper optimizes the number and layout of the monitoring nodes in this area and formulates the monitoring scheme by means of progressive addition of nodes based on the minimum information entropy. First, the single node monitoring scheme is optimized, the corresponding monitoring scheme is specified, the single monitoring node with the lowest information entropy from the 32 alternative monitoring nodes is selected, and then the combined monitoring scheme of the two monitoring nodes of the pipe network with the lowest information entropy based on the single monitoring node with the lowest information entropy is screened, and so on. Finally, combined with the monitoring cost, the optimal combined monitoring scheme of multiple monitoring points is determined to ensure that the monitoring nodes accurately cover the whole research area. The information entropy calculation is generalized as follows:

$$\min E(S) = -\ln p(\alpha) + \min U(S), \tag{16}$$

where s represents the monitoring node scheme.

The information entropy $E(S)$ of different monitoring schemes is obtained according to Equations 7, 11, and 12. In order to verify the optimal design effect of the monitoring wells based on the Bayesian formula and information entropy, the monitoring scheme is evaluated by taking the information entropy $E(S)$ and the mean value of relative error $MRE(S)$ of the inversion results as the index. As shown in Table 1, from parameters α, in the prior distribution, 20 groups of parameters are randomly and evenly obtained by the MH sampling method, and these real parameters are simulated by the SWMM model to produce 20 ×

TABLE 1 Simulation parameter values of prior distribution.

Serial number	Time (min)	Node	Pollutant loads (g)
1	38	5	954,764
2	145	32	910,473
3	42	2	979,596
4	19	8	923,168
5	103	5	952,507
6	86	2	916,292
7	78	25	919,011
8	171	20	967,767
9	1	25	919,284
10	97	8	975,506
11	140	6	912,276
12	130	23	978,355
13	96	2	913,304
14	160	12	965,248
15	148	6	976,633
16	61	12	945,836
17	63	22	948,136
18	108	20	965,090
19	83	20	943,419
20	95	10	975,916

32 groups of concentration monitoring values. Combined with the monitoring values, the MCMC algorithm is used to invert the parameters α, and the length of each Markov chain is 10,000. A posteriori statistic is carried out on these 10,000 groups of samples to obtain the parameter posteriori mean estimation MS. The real parameter a in Table 1 is substituted into the $MRE(S)$ expression as follows:

$$MRE(S) = \left(\sum_{j=1}^{m}\sum_{k=1}^{n}\frac{|M_S(j,k) - \alpha(j,k)|}{\alpha(j,k)}\right/ (m \times n), \tag{17}$$

where MRE is the posterior mean error between the posterior mean estimation M and the real parameters α, m is the number of assumed pollution parameter groups, n is the number of parameters, j is the real pollution parameter of group J α, and K represents the kth type of parameter α.

First, the first monitoring point is calculated. According to formula 13, the information entropy and *a posteriori* mean error of 32 monitoring schemes under S1 are obtained. The calculation results are shown in Table 2. The linear fitting between $MRE(S1)$ and $E(S1)$ is shown in Figure 2. There is a good positive linear correlation between $MRE(S1)$ and $E(S1)$. The linear fitting equation is PRE(S1) = 0.02703(S1)-0.47148, where the determination coefficient R2 = 0.8569. It shows that the information entropy $E(S1)$ is an effective measure of the accuracy of the parameter inversion results. The smaller the value of $E(S1)$, the higher the accuracy of the parameter inversion results.

TABLE 2 Information entropy and posterior mean of the monitoring points.

Type	Project number	S1	E (S1)	MRE (S1)
Monitoring point	1	1	19.994	0.076
	2	2	19.994	0.076
	3	3	19.807	0.068
	4	4	20.183	0.083
	5	5	19.808	0.068
	6	6	20.183	0.083
	7	7	20.183	0.081
	8	8	19.808	0.073
	9	9	20.001	0.078
	10	10	20.032	0.08
	11	11	20.003	0.079
	12	12	19.991	0.074
	13	13	19.990	0.073
	14	14	19.650	0.061
	15	15	20.041	0.079
	16	16	19.998	0.075
	17	17	20.183	0.083
	18	18	19.977	0.071
	19	19	19.659	0.068
	20	20	19.637	0.064
	21	21	20.183	0.078
	22	22	19.865	0.073
	23	23	19.402	0.06
	24	24	19.952	0.075
	25	25	19.988	0.081
	26	26	19.993	0.075
	27	27	19.917	0.074
	28	28	19.262	0.057
	29	29	19.951	0.069
	30	30	19.235	0.057
	31	31	19.229	0.056
	32	32	19.276	0.058

3.1.2 Multi-monitoring point layout of a pipe network

It can be seen from Table 2 that the J31 node has the smallest information entropy, so it is regarded as the first monitoring node selected in the multi-monitoring scheme. Subsequently, the second monitoring node in the multi-monitoring scheme is selected, the selected monitoring node J32 and the other 31 alternative monitoring nodes are combined, respectively, to obtain 31 combination forms, and the information entropy of these

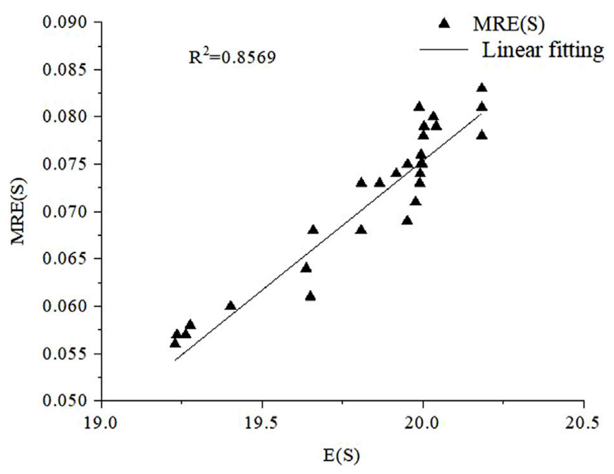

FIGURE 2
Linear fitting diagram of E (S) and MRE (S) (monitoring point = 1).

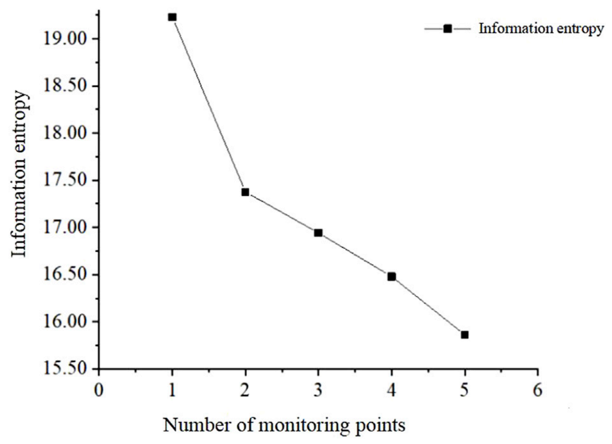

FIGURE 3
Minimum information entropy changing curve of different monitoring numbers.

TABLE 3 Optimal design scheme for multiple monitoring points.

Number	Number of monitoring	Monitoring scheme	Minimum information entropy
1	2	(31, 14)	17.376
2	3	(31, 14, and 27)	16.941
3	4	(31, 14, 27, and 30)	16.577
4	5	(31, 14, 27, 30, and 23)	15.461

31 combined monitoring schemes is calculated. The monitoring scheme of the multi-monitoring combination is shown in Table 3.

It can be seen from Table 3 that there are monitoring schemes with the minimum information entropy under different numbers of monitoring points. Through the monitoring optimization verification of the information entropy algorithm in Section 2.1, it can be seen that information entropy has a positive linear correlation with the posterior mean estimation. Therefore, only the monitoring scheme with the minimum information entropy needs to be selected as the final optimal monitoring layout scheme of the drainage pipe network. Combined with Table 3, the variation curve of information entropy $E(MP)$ of the optimal scheme in different monitoring types with the number of monitoring logs is drawn below, as shown in Figure 3. Under the optimal scheme conditions of each monitoring type, the information entropy decreases with the increase in the number of monitoring nodes. The information entropy of two monitoring points is significantly less than that of one monitoring point, and the information entropy greater than that of two monitoring points decreases to a certain extent, but the difference is small. In practice, on the premise of accurate identification of pollution sources, other factors are usually considered, including monitoring coverage and monitoring cost. Therefore, this example selects two monitoring schemes for pollution source identification, including the following two schemes:

1. When the selection of monitoring nodes is greater than 1, the monitoring point scheme will take into account the inversion accuracy, monitoring point coverage, and monitoring cost. With the increase in the number of monitoring wells, the coverage of the monitoring points is wider and the monitoring is more accurate, but the monitoring cost increases significantly. It can be seen from the figure that the information entropy of the combination of two monitoring points is not different from that of 3–5 groups of monitoring points. Therefore, considering the detection cost, the J31 and J14 combination monitoring methods are selected as the monitoring scheme.
2. Based on the monitoring well scheme with high inversion accuracy, since the information entropy of 3–5 groups of monitoring wells is not different, it is considered that the inversion accuracy is similar, so the J31, J14, and j27 monitoring combination is selected for this scheme.

3.2 Inversion verification of the pollution source tracking model

It is assumed that the value of model parameters α=(Jx, M, T) =(15, 950000, 60) is true, where Jx is the node number, M denotes pollutant loads, and T denotes time. We use the three monitoring schemes selected in the previous section and the MH algorithm to realize the inversion of pollution source parameters, in which the length of each Markov chain is 10,000. According to the calculation, the posterior distribution range of each monitoring

10.3389/fenvs.2024.1401942

TABLE 4 Posterior statistical results of inversion.

Number	Programs	Information entropy	Parameter	Parameter truth	Posterior range
1	31	19.229	Jx	15	[1, 2]/[15, 16]
			M	950000	[920000, 970000]
			T	60	[56, 60]
2	31, 14	18.915	Jx	15	8/10/[14, 15]
			M	950000	[920000, 970000]
			T	60	[60, 61]
3	31, 14, 27	18.821	Jx	15	[15, 16]
			M	950000	[930000, 970000]
			T	60	[59, 60]

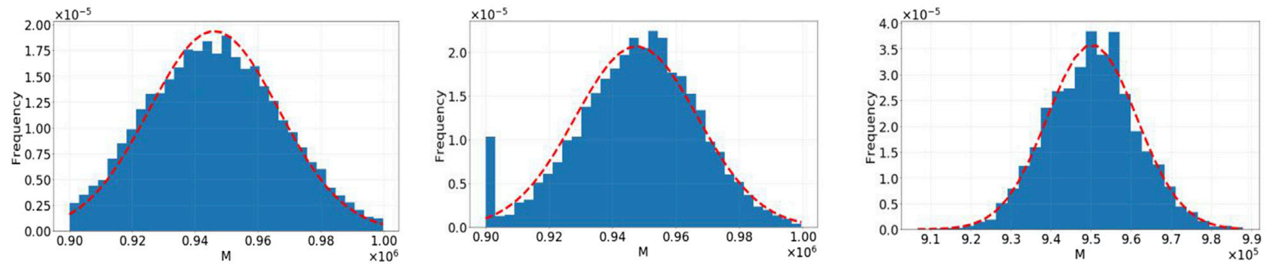

FIGURE 4
Histogram of parameter M in scheme 1.

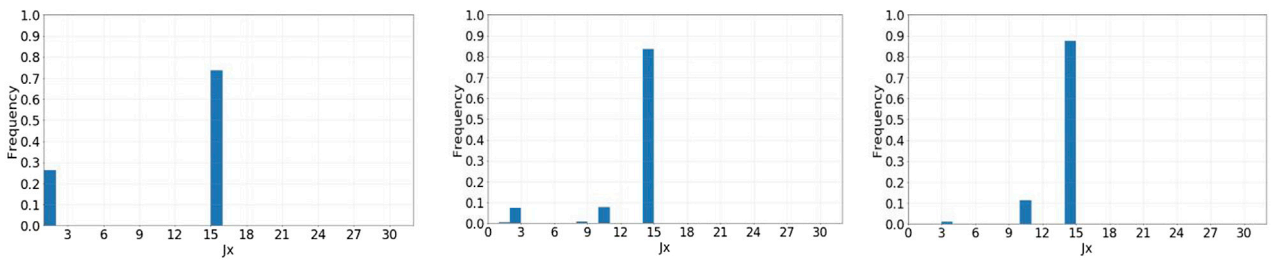

FIGURE 5
Histogram of parameter Jx in scheme 1.

scheme is shown in Table 4. The distribution histogram is shown in Figures 4–6.

It can be seen from Tables 4 and 5 and Figures 3 and 5 that the relative error decreases with the increase in the number of monitoring points, and the mean error of the parameters Jx, M, and T at the two monitoring points decreases by 6.774%, 0.177%, and 0.441%, respectively, compared with that of one monitoring point. The mean value errors of parameters JX, m, and t at the three monitoring points are reduced by 0.993%, 0.012%, and 0.224%, respectively, compared with that of the two monitoring points. The median error of the parameter m at two monitoring points is 0.265% lower than that at one monitoring point. The median error of the parameter m at the three monitoring points is 0.056% lower than that at the two monitoring points. It can be seen that the error of the two monitoring points decreases more than that of one monitoring point, and the error of three monitoring points decreases more slowly than that of two monitoring points, which is consistent with the trend of information entropy, indicating that information entropy can accurately describe the uncertainty of the model.

Among them, the prior range of inversion nodes and inversion time from scheme 1 to scheme 3 is also gradually reduced. Due to the high pollution value, the posterior distribution range is not significantly reduced. However, the mean value error is reduced to a certain extent, which is consistent with the change trend of monitoring scheme information. It can be seen from Table 5 that the real value

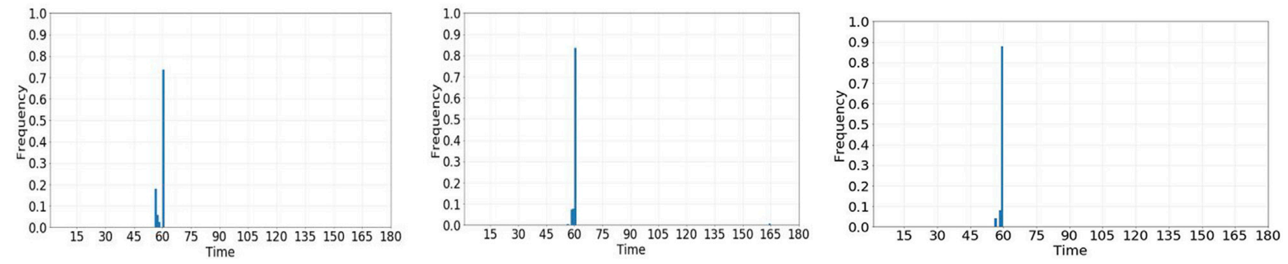

FIGURE 6
Histogram of parameter T in scheme 1.

TABLE 5 Error statistics of the monitoring schemes.

Programs	Parameter	Parameter truth	True value probability (%)	Mean value	Mean error (%)	Median	Median error (%)
31	Jx	15	75.84	13.136	12.427	15	0
	M	950000	—	946280.722	0.392	945969.923	0.424
	T	60	73.51	59.182	1.363	60	0
31, 14	Jx	15	83.62	14.152	5.653	15	0
	M	950000	—	947953.931	0.215	948493.259	0.159
	T	60	83.54	60.553	0.922	60	0
31, 14, 27	Jx	15	87.41	14.301	4.660	15	0
	M	950000	—	948069.155	0.203	949023.129	0.103
	T	60	87.78	59.581	0.698	60	0

probabilities of the parameters Jx and T in scheme 2 are increased by 7.78% and 10.03%, respectively, compared with the two parameter probabilities in scheme 1, and the real value probabilities of the parameters Jx and T in scheme 3 are increased by 3.79% and 4.24%, respectively, compared with the two parameter probabilities in scheme 2. It is further determined that the smaller the information entropy of the monitoring scheme, the higher the inversion accuracy and the smaller the uncertainty of the posterior distribution. However, the more the number of monitoring nodes, the more the cost and the less the cost of single-well monitoring. The information entropy difference between scheme 2 and scheme 3 is very small, and the average value of the relative error of the posterior sample mean of the three parameters is also approximately equal. Adding a monitoring point for monitoring will significantly increase the cost. For this case, if it is necessary to comprehensively consider the monitoring cost and the posterior distribution range of the parameters, scheme 2 is considered to be the best monitoring scheme.

4 Conclusion

Aiming at a series of problems, such as the chaos of the urban drainage supervision system, the frequent occurrence of illegal discharge and leakage of industrial enterprises, the serious pollution impact on the sewage treatment plant, and the destruction of the surrounding environment in the discharge area, this paper puts forward the monitoring point optimization model based on information entropy. This model combines Bayesian reasoning, the Monte Carlo method, the SWMM model, and information entropy to realize the automatic layout of drainage network nodes. Finally, the pollution source tracking model is used to verify the effectiveness of the optimal layout of the monitoring points.

In this paper, information entropy is used as the optimization index for the optimal design of the monitoring well scheme. First, the Bayesian formula is combined with information entropy as the optimization index of the monitoring scheme, and the average value of relative error between information entropy and inversion results is used as the index to evaluate the monitoring scheme; that is, the relative mean error of the parameter inversion simulation results of the pollution source tracking model is linearly fitted with information entropy. Through calculation, it is known that there is a good positive linear correlation between them, which verifies the feasibility of the information entropy algorithm in the field of monitoring node optimization. The optimization results based on the information entropy algorithm under different monitoring schemes are further studied. The smaller the uncertainty of pollution source identification, the smaller the information entropy calculation result. By calculating the information entropy of multiple monitoring points, the optimization of monitoring points of drainage pipe networks is finally realized. The results show that the information entropy algorithm can be well-applied to the optimal layout of a single monitoring node and multiple monitoring nodes and can correspond well to the inversion results of the tracking model parameters. According to the above research, the constructed monitoring point optimization model can well realize the optimal layout of the monitoring points of drainage pipe networks.

Therefore, this paper proposes the optimal design of monitoring points based on the Bayesian information entropy theory, which is helpful in realizing the optimization of monitoring points aiming at pollution source identification and increasing the accuracy of pollution source identification. The research results are helpful for the efficient layout of the monitoring points of urban drainage networks. As an important part of urban ecological environment, urban water environment has an important impact on the governance and improvement of the ecological environment. This study specifically addresses the phenomenon of pollutant smuggling and leakage; however, various water pollution events in real life are intricate and complex. In both domestic and foreign research, quickly tracking water intrusion outside the drainage network poses a significant challenge. Many issues in this area stem from manual search methods and the lack of a well-structured monitoring point arrangement. For such events, further in-depth studies and research are needed in the future.

Data availability statement

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

Author contributions

MH: conceptualization and writing–original draft. YZ: data curation and writing–review and editing. ZM: formal analysis and writing–original draft. QZ: funding acquisition and writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This research was funded by the National Natural Science Foundation of China (grant numbers 51878556, 52009102). The Natural Science Basic Research Program of Shaanxi (Program No. 2023-JC-YB-219).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abellán, J., Baker, R. M., and Coolen, P. A. F. (2011). Maximising entropy on the nonparametric predictive inference model for multinomial data. *Eur. J. Operational Res.* 212, 112–122. doi:10.1016/j.ejor.2011.01.020

Ainslie, B., Reuten, C., Steyn, D. G., and Zidek, J. V. (2009). Application of an entropy-based Bayesian optimization technique to the redesign of an existing monitoring network for single air pollutants. *J. Environ. Manag.* 90, 2715–2729. doi:10.1016/j.jenvman.2009.02.016

Banik, B. K., Alfonso, L., Torres, A. S., Mynett, A., Di Cristo, C., and Leopardi, A. (2015). Optimal placement of water quality monitoring stations in sewer systems: an information theory approach. *Procedia Eng.* 119, 1308–1317. doi:10.1016/j.proeng.2015.08.956

Blasone, R. S., Vrugt, J. A., Madsen, H., Rosbjerg, D., Robinson, B. A., and Zyvoloski, G. A. (2007). Generalized likelihood uncertainty estimation (GLUE) using adaptive Markov chain Monte Carlo sampling. *Adv. Water Resour.* 31, 630–648. doi:10.1016/j.advwatres.2007.12.003

Groothuis, P. A., Mohr, T. M., Whitehead, J. C., Cockerill, K., and Gu, C. (2021). Measuring the direct and indirect effect of scientific information on valuing storm water management programs with a hybrid choice model. *Water Resour. Res.* 57 (5). doi:10.1029/2020WR027552

Guo, X. C., Li, M., Zhao, D. Q., et al. (2018). Research and progress on optimal layout of monitoring points in urban drainage network. *Water supply drainage China* 34, 26–31. doi:10.19853/j.zgjsps.1000-4602.2018.04.006

Guo, Y., Quan, L., Song, L., and Liang, H. (2022). Construction of rapid early warning and comprehensive analysis models for urban waterlogging based on AutoML and comparison of the other three machine learning algorithms. *J. Hydrology* 605, 127367. doi:10.1016/j.jhydrol.2021.127367

Jiang, X., Ma, R., Wang, Y. X., Gu, W., Lu, W., and Na, J. (2021). Two-stage surrogate model-assisted Bayesian framework for groundwater contaminant source identification. *J. Hydrology* 594, 125955. doi:10.1016/j.jhydrol.2021.125955

Kaynar, B., and Ridder, A. (2010). The cross-entropy method with patching for rare-event simulation of large Markov chains. *Eur. J. Operational Res.* 207, 1380–1397. doi:10.1016/j.ejor.2010.07.002

Keum, J., Awol, F. S., Ursulak, J., and Coulibaly, P. (2019). Introducing the ensemble-based dual entropy and multiobjective optimization for hydrometric network design problems: EnDEMO. *Entropy* 21, 947. doi:10.3390/e21100947

Keum, J., Coulibaly, P., Razavi, T., Tapsoba, D., Gobena, A., Weber, F., et al. (2018). Application of SNODAS and hydrologic models to enhance entropy-based snow monitoring network design. *J. Hydrology* 561, 688–701. doi:10.1016/j.jhydrol.2018.04.037

Li, C., Singh, V. P., and Mishra, A. K. (2012). Entropy theory-based criterion for hydrometric network evaluation and design: maximum information minimum redundancy. *Water Resour. Res.* 48, W05521. doi:10.1029/2011WR011251

Liu, B. J., Chen, X. H., Lian, Y. Q., and Wu, L. (2013). Entropy-based assessment and zoning of rainfall distribution. *J. Hydrology* 490, 32–40. doi:10.1016/j.jhydrol.2013.03.020

Liu, X. C., Zhang, H. F., Song, X. M., Gao, Y., Yang, Y. S., Wang, C. Y., et al. (2018). Monitoring and analysis of the current situation of the drainage network in a city of northern China. *IOP Conf. Ser. Earth Environ. Sci.* 186, 012033. doi:10.1088/1755-1315/186/6/012033

Martino, L., and Elvira, V. (2018). Group importance sampling for particle filtering and MCMC. *Digit. Signal Process.* 82, 133–151. doi:10.1016/j.dsp.2018.07.007

Meyer, A. M., Klein, C., Fünfrocken, E., Kautenburger, R., and Beck, H. P. (2018). Real-time monitoring of water quality to identify pollution pathways in small and middle scale rivers. *Sci. Total Environ.* 651, 2323–2333. doi:10.1016/j.scitotenv.2018.10.069

Mirauda, D., and Ostoich, M. (2020). MIMR criterion application: entropy approach to select the optimal quality parameter set responsible for river pollution. *Sustainability* 12 (5), 2078. doi:10.3390/su12052078

Monache, L. D., Lundquist, J. K., Kosovic, B., Johannesson, G., Dyer, K. M., Aines, R. D., et al. (2008). Bayesian inference and Markov chain Monte Carlo sampling to reconstruct a contaminant source on a continental scale. *J. Appl. Meteorology Climatol.* 47, 2600–2613. doi:10.1175/2008JAMC1766.1

Mooselu, M. G., Liltved, H., Nikoo, M. R., Hindar, A., and Meland, S. (2020). Assessing optimal water quality monitoring network in road construction using integrated information-theoretic techniques. *J. Hydrology* 589, 125366. doi:10.1016/j.jhydrol.2020.125366

Omamageswari, M., Mohanraj, A., Jeeva, S. C., Reddy, A. K., and Penchalaiah, U. (2021). Iot based smart drainage monitoring and cleaning system for solid waste materials. *Mater. Today Proc.* 3. doi:10.1016/j.matpr.2020.11.940

Parvan, A. S. (2010). Critique of multinomial coefficients method for evaluating Tsallis and Rényi entropies. *Phys. A Stat. Mech. its Appl.* 389, 5645–5649. doi:10.1016/j.physa.2010.08.040

Retzer, J. J., Soofi, E. S., and Soyer, R. (2009). Information importance of predictors: concept, measures, Bayesian inference, and applications. *Comput. Statistics Data Analysis* 53, 2363–2377. doi:10.1016/j.csda.2008.03.010

Sambito, M., and Freni, G. (2021). Strategies for improving optimal positioning of quality sensors in urban drainage systems for non-conservative contaminants. *WATER* 13, 934. doi:10.3390/w13070934

Shi, Z. B., Shen, Q. R., Tan, Q., and Li, T. (2021). Development of integrated flooding early warning and rainfall runoff management platform for downtown area of shanghai. *Sustainability* 13 (20), 11250. doi:10.3390/su132011250

Tang, X. Z., Shu, Y. Q., lian, Y. Q., Zhao, Y., and Fu, Y. (2018). A spatial assessment of urban waterlogging risk based on a Weighted Naive Bayes classifier. *Sci. Total Environ.* 630, 264–274. doi:10.1016/j.scitotenv.2018.02.172

Vrugt, A. J. (2016). Markov chain Monte Carlo simulation using the DREAM software package: theory, concepts, and MATLAB implementation. *Environ. Model. Softw.* 75, 273–316. doi:10.1016/j.envsoft.2015.08.013

Wang, G. S., and Chen, S. L. (2013). Evaluation of a soil greenhouse gas emission model based on Bayesian inference and MCMC: model uncertainty. *Ecol. Model.* 253, 97–106. doi:10.1016/j.ecolmodel.2012.09.010

Wang, H., and Jin, X. (2013). Characterization of groundwater contaminant source using Bayesian method. *Stoch. Environ. Res. Risk Assess.* 27, 867–876. doi:10.1007/s00477-012-0622-9

Wang, H. Q., Lin, G., and li, G. L. (2016). Gaussian process surrogates for failure detection: a Bayesian experimental design approach. *J. Comput. Phys.* 313, 247–259. doi:10.1016/j.jcp.2016.02.053

Wei, M., She, L., and You, X. Y. (2020). Establishment of urban waterlogging pre-warning system based on coupling RBF-NARX neural networks. *Water Sci. Technol.* 82 (9), 1921–1931. doi:10.2166/wst.2020.477

Wu, H. J., Chen, B., Ye, X. D., Guo, H., Meng, X., and Zhang, B. (2021). An improved calibration and uncertainty analysis approach using a multicriteria sequential algorithm for hydrological modeling. *Sci. Rep.* 11, 16954. doi:10.1038/s41598-021-96250-6

Yan, L., and Zhou, T. (2019). Adaptive multi-fidelity polynomial chaos approach to Bayesian inference in inverse problems. *J. Comput. Phys.* 381, 110–128. doi:10.1016/j.jcp.2018.12.025

Yang, Y. H., Sun, L. F., Li, R. N., Yin, J., and Yu, D. (2020). Linking a storm water management model to a novel two dimensional model for urban pluvial flood modeling. *Int. J. Disaster Risk Sci.* 11, 508–518. doi:10.1007/s13753-020-00278-7

Yazdi, J. (2017). Optimization of hydrometric monitoring network in urban drainage systems using information theory. *Water Sci. Technol.* 76, 1603–1613. doi:10.2166/wst.2017.226

Zeng, X. K., Wu, J. C., Wang, D., Zhu, X., and Long, Y. (2016). Assessing Bayesian model averaging uncertainty of groundwater modeling based on information entropy method. *J. Hydrology* 538, 689–704. doi:10.1016/j.jhydrol.2016.04.038

Zhang, J. J., Li, W. X., Zeng, L. Z., and Wu, L. (2016). An adaptive Gaussian process-based method for efficient Bayesian experimental design in groundwater contaminant source identification problems. *Water Resour. Res.* 52, 5971–5984. doi:10.1002/2016WR018598

Zhang, J. J., Zeng, L. Z., Chen, C., Chen, D., and Wu, L. (2015). Efficient Bayesian experimental design for contaminant source identification. *Water Resour. Res.* 51, 576–598. doi:10.1002/2014WR015740

Zhao, C. S., Shao, N. F., Yang, S. T., Xiang, H., Lou, H., Sun, Y., et al. (2018). Identifying the principal driving factors of water ecosystem dependence and the corresponding indicator species in a pilot City, China. *J. Hydrology* 556, 488–499. doi:10.1016/j.jhydrol.2017.11.031