sensors

MDPI

Article

A Data-Based Framework for Identifying a Source Location of a Contaminant Spill in a River System with Random Measurement Errors

Jun Hyeong Kim [1], Mi Lim Lee [2] and Chuljin Park [1,*]

[1] Department of Industrial Engineering, Hanyang University, 222 Wangsimni-Ro, Seongdong gu, Seoul 04763, Korea

[2] College of Business Administration, Hongik University, 94, Wausan-ro, Mapo-gu, Seoul 04066, Korea

* Correspondence: parkcj@hanyang.ac.kr; Tel.: +82-2-2220-0476

Received: 28 June 2019; Accepted: 29 July 2019; Published: 1 August 2019

Abstract: This study addresses the problem of identifying the source location of a contaminant spill in a river system when a sensor network returns observations containing random measurement errors. To solve this problem, we suggest a new framework comprising three main steps: (i) spill detection, (ii) data preprocessing, and (iii) source identification. Specifically, we applied a statistical process control chart to detect a contaminant spill with measurement errors while keeping the false alarm rate at less than or equal to a user-specified value. After detecting a spill, we generated a nonlinear regression model to estimate a breakthrough curve of the observations and derive a characteristic vector of the estimated curve. Using the characteristic vector as an input, a random forest model was constructed with the sensor raising the first alarm. The model provides output values between 0 and 1 to represent the possibility of each candidate location being the true spill source. These possibility values allow users to identify strong candidate locations for the spill. The accuracy of our framework was tested on part of the Altamaha River system in Georgia, USA.

Keywords: source identification; sensor network; water quality monitoring; river system; statistical process control; random forest

1. Introduction

Water is a crucial resource for both public health and ecological life. Since the amount of fresh water is decreasing at the same time that population, industrialization, and environmental pollution are increasing, the importance of water quality monitoring is attracting more attention. Based on improvements to real-time sensor and data analysis technologies that enable people to monitor water quality more effectively, the problem of identifying the source location of a contaminant spill has also been extensively studied by researchers. Most previous studies related to this problem have addressed two types of water systems: groundwater and rivers. Optimization algorithms, such as linear/nonlinear programming and meta-heuristics, have been commonly used to identify contaminant spill locations in groundwater systems, as shown by Aral and Guan [1], Aral et al. [2], Gorelick et al. [3], Singh and Datta [4], and Sun et al. [5]. Additionally, statistical methods (such as a backward probability model approach [6,7] and a geostatistical approach [8]) and machine learning techniques including artificial neural network models have been used by Singh and Datta [9], Singh et al. [10], and Srivastava and Singh [11] to address similar problems.

Due to the size and complexity of the problem, fewer studies have been conducted for rivers than for groundwater systems. Boano et al. [12] employed a geostatistical approach that generates previous information about a pollutant. Chen et al. [13] applied multivariate statistical methods to

determine the spatial and temporal variations of water quality, then identified the contaminant source location. The backward probability method was also used by Ghane et al. [14] and Telci and Aral [15], while Lee et al. [16] recently provided a method based on random forest models to identify the source location of a contaminant spill in a river system. The random forest model is a famous classification model that consists of a set of tree-structured classifiers. It is known to have several advantages, such as computational efficiency, accuracy, and robustness, compared with other models. One may see Breiman [17] for an overview of the random forest model. For source identification in a river system, the random forest models provide values between 0 and 1, indicating the possibility that a candidate location is the true spill location, while the backward probability method provides only the rankings of candidate locations.

Many studies rely on hydrodynamic simulation models that provide contaminant concentration levels in a water system to develop and test their methods. It is difficult to apply these methods, however, because, in practice, the observed concentration levels often contain random measurement errors. As mentioned by Kim et al. [18], the results from a case considering observations with random measurement errors may be totally different compared with those from a case considering ideal observations generated by the simulation models without measurement errors (e.g., the concentration levels are exactly 0 when no spill event occurs). To obtain reliable results despite random measurement errors, one may need to control false alarm rates, which can be done by statistical process control (SPC) charts. A good review of SPC charts is provided by Montgomery [19]. Among the various SPC charts, we focus on the CUSUM chart developed by Kim et al. [20] due to the following advantages: (i) it enables users to derive a threshold value with a simple analytical method and (ii) it can deal with autocorrelated observations that follow a non-normal distribution.

This study, therefore, considers the problem of identifying a spill source location in a river system in the presence of random measurement errors. To the best of our knowledge, this is the first work considering random measurement errors in the source identification problem. To solve this problem, we suggest a new framework comprised of three main steps: (i) detecting a contaminant spill, (ii) preprocessing obtained contaminant spill data, and (iii) identifying the spill source location via random forest models. Specifically, we first use the CUSUM chart developed by Kim et al. [20] to detect a contaminant spill, where the false alarm rate is guaranteed to be less than or equal to a user-specified level. Then, we obtain a set of observation data including all information about the contaminant spill and use a robust locally-weighted regression model [3] to estimate the profile of the observations, called the breakthrough curve. Finally, using profile characteristics as inputs, we develop a classification method using random forest models to measure the possibility that each candidate spill location is truly the correct location of the contaminant source. Based on the possibilities, we can consistently identify strong candidate locations for the spill.

This paper is organized as follows. Section 2 describes the problem with notations and assumptions and provides a method for obtaining the observation data. In Section 3, we suggest a new framework to identify the source location of a contaminant spill in the presence of random measurement errors. Section 4 presents experimental results of a case study applied to a part of the Altamaha River system located in Georgia, USA. Concluding remarks follow in Section 5.

2. Background

2.1. Problem Description

We consider a river system including N possible candidate locations for a contaminant spill. We index the locations by integers starting from 1 and call these integers location indices. In order to monitor the water quality and detect the spill event, K number of sensors are installed at a subset of the possible spill locations to report the concentration levels of the contaminant at each discretized time t.

Let D denote the index set of all possible locations, i.e., $D = \{1, 2, \ldots, N\}$. For $2 \leq K \leq N$ given, the location of the K sensors are represented by a vector $\mathbf{y} = (y_1, \ldots, y_K)$, such that $y_j \in D$ for all

$j = 1, \ldots, K$. (We assume $y_1 < y_2 < \ldots < y_K$ to avoid the repetition of representations). At each time t, an observation $X_t(y_j)$ is returned by the sensor installed at y_j location, and the values from all K sensors at time t are represented by the observation vector:

$$\boldsymbol{X}_t(\mathbf{y}) = \begin{bmatrix} X_t(y_1) \\ \vdots \\ X_t(y_K) \end{bmatrix}. \tag{1}$$

Since the true spill location is unknown and a contaminant spill can occur at any of the candidate locations, the possibility that location $d \in D$ is the true source of the spill, $P(d)$, can be evaluated. Note that $0 \leq P(d) \leq 1$ for all $d \in D$. The closer $P(d)$ is to 1, the more likely that d is the source location of the spill. Let $\boldsymbol{P}$ denote a vector of $P(d)$ as follows:

$$\boldsymbol{P} = \begin{bmatrix} P(1) \\ \vdots \\ P(N) \end{bmatrix}. \tag{2}$$

The problem considered in this paper is to construct a data-driven framework for the purpose of identifying the source location of a contaminant spill. We develop a model that calculates $\boldsymbol{P}$ based on the recent observation vectors, $\boldsymbol{X}_{t-\omega+1}(\mathbf{y}), \ldots \ldots, \boldsymbol{X}_t(\mathbf{y})$ with a pre-specified window length ω, when K and $\mathbf{y}$ are given.

To construct the data-driven framework and test its performance, preparation of a large dataset is required. In the next subsection, we briefly explain the data acquisition method by incorporating random measurement errors with values obtained from a hydrodynamics simulation.

2.2. Data Acquisition with Simulation

This paper supposes that an observation $X_t(y_j)$ may contain some measurement error for all t and $j = 1, \ldots, K$. In order to obtain realistic $X_t(y_j)$ values reported by the sensors with random measurement errors, we adopt the model from Kim et al. [18] as follows:

$$X_t(y_j) = \xi_t(y_j) + \varepsilon_\xi, \tag{3}$$

where $\xi_t(y_j)$ is the true concentration level of the contaminant and ε_ξ is a random measurement error that depends on the value of $\xi_t(y_j)$.

For obtaining the realizations of $\xi_t(y_j)$ values, a popular simulation software package called the Storm Water Management Model (SWMM) provided by the United States Environmental Protection Agency, Durham, NC, USA, is used. The SWMM is developed by the United States Environmental Protection Agency to simulate hydrodynamics and contaminant transport in river systems under dynamic flow, including the various rainfall and watershed conditions described by Rossman [21]. The SWMM requires inputs of variable information related to the properties of random contaminant spills and rainfall events derived from historical data in addition to fixed information related to geologic/geometric properties and fundamental river system hydrodynamics. After executing the SWMM software using the inputs, we can obtain concentration levels from each candidate location at every inter-reporting time of the simulation clock.

As shown in Kim et al. [18], the measurement error ε_ξ is quantified based on two perspectives: accuracy (i.e., the bias of the measurement error) and precision (i.e., the variability of the measurement errors). Let μ_ξ denote the level of accuracy and σ_ξ denote the level of precision of the sensor at location y_j at time t. For given values of μ_ξ and σ_ξ, ε_ξ is assumed to be independent and identically distributed

along with the Laplace distribution. Therefore, the probability density function of ε_ξ is specified as follows:

$$f(\varepsilon_\xi \mid \mu_\xi, \sigma_\xi) = \frac{1}{\sigma_\xi \sqrt{2}} \exp\left(-\frac{\sqrt{2}\left|\varepsilon_\xi - \mu_\xi\right|}{\sigma_\xi}\right). \quad (4)$$

3. Methods

3.1. Overall Description of the Proposed Framework

In this subsection, we provide notations and an overall description of our framework to identify the source location of a contaminant spill in the presence of random measurement errors. A list of notations needed to describe our framework is as follows:

t	the discretized time index at which each sensor returns an observation;
y^*	the location index of the sensor that raises the first alarm;
τ_a^*	the point of time when the first alarm is on by the sensor at y^*;
τ_b^*	the point of time when the first alarm is off by the sensor at y^*;
x_{th}	the concentration level reported by the sensor at y^* at time τ_a^* (i.e., $X_{\tau_a^*}(y^*)$);
ω	a pre-specified window length for spill detection;
ℓ	a lag parameter pre-designated by users to determine τ_b^*;
$\psi(t)$	a nonlinear regression model for $X_t(y^*)$ in the time period $[\tau_a^*, \tau_b^*]$;
$\boldsymbol{B}(y^*)$	a vector representing the curvature characteristics of $\psi(t)$;
$D(y_j)$	the set of candidate spill locations that can be identified only by the sensor at y_j; and
$\phi(y_j)$	the random forest model corresponding to the sensor at y_j.

Figure 1 shows the overall description of our framework. Note that the framework consists of three main steps for (i) detecting a spill, (ii) preprocessing the set of obtained data, and (iii) identifying the source location with a classification model. Sections 3.2–3.4 provide details about each of the three steps, respectively.

Framework: Source Identification under Random Measurement Errors

(Step 0) **Initialization**
Set $\tau_a^* = \infty$, $\tau_b^* = \infty$, $t = 0$ and $\nu = 0$.
Set ℓ and ω as positive integers.
(Step 1) **Spill Detection (in Section 3.2)**
While no alarm is raised **do**
Obtain $\boldsymbol{X}_t(\mathbf{y})$.
Construct monitoring statistics based on $\{\boldsymbol{X}_{t-\omega+1}(\mathbf{y}), \cdots, \boldsymbol{X}_t(\mathbf{y})\}$.
Set $t \leftarrow t + 1$.
end while
Set y^* as the sensor location at which the first alarm is raised.
Update τ_a^* as the current time index t and set x_{th} as $X_{\tau_a^*}(y^*)$.
(Step 2) **Data Preprocessing (in Section 3.3)**
While $\nu < \ell$ **do**
Obtain $X_t(y^*)$.
If $X_t(y^*) < x_{th}$, $\nu \leftarrow \nu + 1$. Otherwise, $\nu = 0$.
Set $t \leftarrow t + 1$.
end while
Update τ_b^* as the current time index t.
Generate a nonlinear regression model $\psi(t)$ based on $\{X_t(y^*);\ t = \tau_a^*, \cdots, \tau_b^*\}$.
Derive $\boldsymbol{B}(y^*)$ from $\psi(t)$ for $= \tau_a^*, \cdots, \tau_b^*$.
(Step 3) **Source Identification (in Section 3.4)**
Select a random forest model $\phi(y^*)$ for source identification regarding $D(y^*)$.
Return $P(d)$ for all $d \in D(y^*)$ by using $\boldsymbol{B}(y^*)$ as the input of $\phi(y^*)$.
Return $P(d) = 0$ for all $d \notin D(y^*)$.

Figure 1. Overall description of the proposed source identification framework.

3.2. Spill Detection

Detecting a contaminant source location under measurement errors is more complicated than detecting one without measurement errors. If there is no measurement error, the observed value is the same as the true contaminant concentration level and a sensor can simply raise an alarm when the observed value exceeds 0. In this case, an alarm is always a true alarm providing notification that a spill event has happened somewhere upstream. With measurement errors, however, nonzero values can be reported by sensors even when no spill event has occurred. In this case, users should be aware of the risk of false alarms and must carefully determine the threshold value for triggering an alarm. (A high threshold value increases the chance of missing a spill event, while a low threshold value subjects users to frequent false alarms.)

In order to detect a contaminant spill under random measurement errors, a statistical process control (SPC) chart with a carefully designed threshold can be used as a monitoring tool. The SPC chart was originally developed to detect a change in the mean of monitoring statistics while maintaining a targeted rate of false alarms. Among the various SPC charts, this paper introduces the CUSUM chart developed by Kim et al. [20] because it provides a simple analytical method for identifying a threshold value even for autocorrelated observations that follow a non-normal distribution. The target false alarm rate is proportional to a target in-control average run length (ARL), denoted by ρ, which represents the average number of observations until a false alarm is raised when the true mean of monitoring statistics is the same. Based on the target ρ given (i.e., the target false alarm rate given), Kim et al. [20] provide the threshold value $\mathcal{H}$ by solving the following equation:

$$K\rho = \frac{1}{2\varsigma^2}\left\{\exp\left(\frac{2\varsigma(\mathcal{H}+1.166\sigma_0)}{\sigma_0}\right) - 1 - \frac{2\varsigma(\mathcal{H}+1.166\sigma_0)}{\sigma_0}\right\}, \tag{5}$$

where ς represents a reference parameter and σ_0 is the known standard deviation of the random measurement errors. We use $\varsigma = 0.1$, which is recommended by Kim et al. [20]. We then construct our CUSUM statistics $S_t^+\left(y_j\right)$ for the sensor at location y_j at time t with the monitoring window length ω as follows:

Constructing CUSUM monitoring statistics

Set $\tau = t - \omega$ and $S_{\tau-1}^+\left(y_j\right) = 0$.
While $\tau \leq t$ **do**
  $S_\tau^+\left(y_j\right) \leftarrow \max\left(0,\ S_{\tau-1}^+\left(y_j\right) + X_\tau\left(y_j\right) - k\sigma_0\right)$.
  $\tau \leftarrow \tau + 1$.
end while

Once the threshold value $\mathcal{H}$ from Equation (5) and the monitoring statistic $S_t^+\left(y_j\right)$ are ready, an alarm is raised by a sensor at location y_j at time t, when

$$S_t^+\left(y_j\right) \geq \mathcal{H}. \tag{6}$$

If the first alarm is raised by a sensor, we record the location of the sensor as y^*, the current time as τ_a^*, and the concentration level $X_{\tau_a^*}(y^*)$ as x_{th}, before proceeding to the data preprocessing step.

3.3. Data Preprocessing

The main purpose of the data preprocessing step is to prepare refined input data for the random forest models to identify the source of the contaminant spill in the next step. This preparation is done by modeling the breakthrough curve of the observed concentration levels over time after the alarm [16] and deriving the characteristics of the curve.

To model the breakthrough curve that contains the most important information about the spill event, we must first define the start and end times of the curve. Because of random measurement errors, positive values of observations can be returned by a sensor even without any spill event. Thus, the time period considered for the curve should not be chosen simply as a period having positive values for returned observations. Instead, our framework suggests a heuristic approach to defining the start time of the curve as τ_a^* and the end time as

$$\tau_b^* \equiv \min \{\ t \mid X_p(y^*) < x_{th} \text{ for } p = t-\ell, \ \ldots, \ t \text{ and } t\ > \tau_a^* \ \}, \tag{7}$$

where ℓ is a lag parameter pre-designated by users. See Figure 2 for examples of $X_t(y^*)$ over time, τ_a^* and τ_b^*.

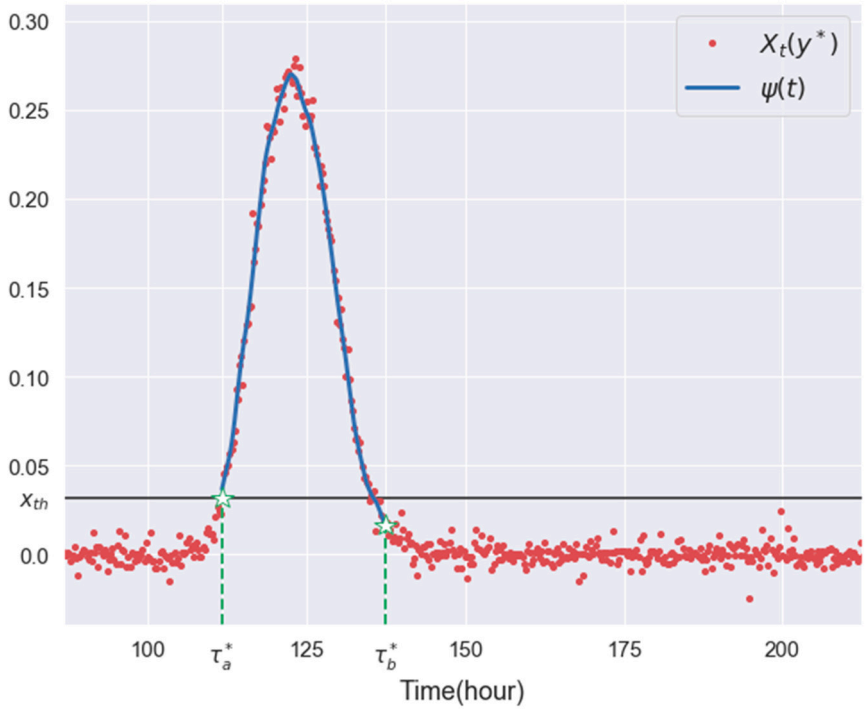

Figure 2. An example plot for $X_t(y^*), \tau_a^*, \tau_b^*$, and $\psi(t)$.

Since random measurement error is similar to a noise factor, instead of using raw data $\{X_t(y^*); t = \tau_a(y^*), \cdots, \tau_b(y^*)\}$ to represent the breakthrough curve, we used a nonlinear regression model to refine the curve. By reducing the impact of the random measurement errors with the regression model, a smoother and clearer breakthrough curve can be obtained. Among the various nonlinear regression models, we employ the robust locally-weighted regression model developed by Cleveland [22]. In this model, the independent variable is t and the dependent variable is $X_t(y^*)$ for $t = \tau_a^*, \cdots, \tau_b^*$. Let $\psi(t)$ denote the concentration level estimated by the regression model at t. In Figure 2, the solid curve provides an example of the curve $\psi(t)$ over time $t = \tau_a(y^*), \cdots, \tau_b(y^*)$.

After obtaining the estimated breakthrough curve $\psi(t)$, we need to derive its characteristics. We basically introduce two quantitative characteristics of $\psi(t)$, denoted by $B_1(y^*)$ and $B_2(y^*)$, which represent the total area and the time-averaged area between the horizontal axis and the estimated breakthrough curve $\psi(t)$, respectively [10]. Specifically, they are calculated as follows:

$$B_1(y^*) = \int_{\tau_a^*}^{\tau_b^*} \psi(t)dt, \tag{8}$$

$$B_2(y^*) = \frac{\int_{\tau_a^*}^{\tau_b^*} \psi(t)dt}{\left(\tau_b^* - \tau_a^*\right)}. \tag{9}$$

In addition to $B_1(y^*)$ and $B_2(y^*)$, we also consider the central statistical moment, standard deviation, skewness, and kurtosis of $\psi(t)$ as the main characteristics of the curve as described by Telci and Aral [15]. The first moment of the estimated breakthrough curve is denoted by $\mu(y^*)$ and calculated by

$$\mu(y^*) = \frac{\int_{\tau_a^*}^{\tau_b^*} t\ \psi(t)dt}{B_1(y^*)}. \tag{10}$$

The estimated kth central moment of $\psi(t)$, for $k = 2, 3, \ldots$, is denoted by $m_\psi^k(y^*)$ and calculated by

$$m_\psi^k(y^*) = \frac{\int_{\tau_a^*}^{\tau_b^*} (t - \mu(y^*))^k\ \psi(t)dt}{B_1(y^*)}. \tag{11}$$

The standard deviation, skewness, and kurtosis of the estimated curve, denoted by $B_3(y^*)$, $B_4(y^*)$, and $B_5(y^*)$ respectively, are calculated as follows:

$$B_3(y^*) = \sqrt{m_\psi^2(y^*)}, \tag{12}$$

$$B_4(y^*) = \frac{m_\psi^3(y^*)}{(B_3(y^*))^3}, \tag{13}$$

$$B_5(y^*) = \frac{m_\psi^4(y^*)}{(B_3(y^*))^4} - 3. \tag{14}$$

Based on Equations (8)–(14), the curvature characteristics of $\psi(t)$ are summarized into a five-dimensional vector as follows:

$$\boldsymbol{B}(y^*) = (B_1(y^*),\ B_2(y^*),\ B_3(y^*),\ B_4(y^*),\ B_5(y^*)).$$

3.4. Source Identification

To identify the location of the contaminant source, we employed the random forest model shown in Lee et al. [16]. As inputs of the random forest model, Lee et al. [16] used not only characteristics of the breakthrough curves, but also time differences in alarms between any pair of sensors. Obtaining such relative time information, however, is often time-consuming and not helpful for improving source identification with random measurement errors. Therefore, we used only the characteristics of the estimated breakthrough curve $\boldsymbol{B}(y^*)$ as inputs to our random forest models.

In this study, we partitioned the whole river region into K sub-regions, each of which was monitored by a single sensor. Recall that $D\left(y_j\right)$ is defined as a set of candidate spill locations that can be identified only by the sensor located at y_j. For example, when $N = 19$, $K = 2$ and $\mathbf{y} = (9, 19)$, $D(9) = \{1, 2, \cdots 14\}$ and $D(19) = \{15, 16, 17, 18, 19\}$ as in Figure 3. The sensor located at 9 raises the first alarm if a spill event occurs at any locations in $D(9)$, while the sensor located at 19 raises the first alarm if a spill occurs at any locations in $D(19)$. In our framework, a random forest model was constructed corresponding to each $D\left(y_j\right)$ by using simulated training data that includes random measurement errors. Let $\phi\left(y_j\right)$ denote the random forest model regarding $D\left(y_j\right)$.

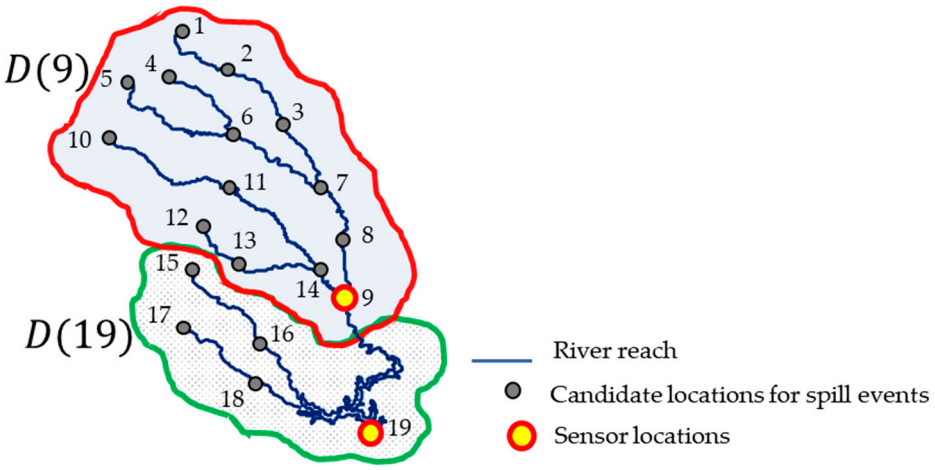

Figure 3. Examples of $D(y_j)$ when sensors are installed at locations 9 and 19.

The random forest model $\phi(y_j)$ consists of several tree classifiers. Suppose that we have Γ number of datasets sampled from the training data using a bootstrapping technique. Since a tree classifier is generated from each sample dataset, there are Γ number of tree classifiers in the random forest model. A tree classifier has internal nodes and terminal nodes. At each internal node, Λ number of input variables are randomly selected and linearly combined with coefficients. The linear combination of the selected input variables is checked if its value is above a certain threshold constant, and we move to the next node according to the result. To set the threshold constant and the coefficients of the linear combination at each internal node, a randomized node optimization algorithm is used [23]. Each terminal node represents one of the locations in $D(y_j)$ as the identified spill source location, and no additional decision or action is taken. Figure 4 shows a part of the tree classifier constructed for $D(19)$ from Figure 3 as an example. Notably, different tree classifiers have different structures (e.g., the numbers of nodes and arcs) and logic, with different threshold constants and linear combinations for each internal node of the classifiers. If an input vector is given in the model, each tree classifier individually nominates one of the candidate locations as the identified contaminant source location. The random forest model $\phi(y_j)$ collects the votes from all tree classifiers and calculates $P(d)$ for $d \in D(y_j)$ proportional to the number of votes from Γ. The model returns $P(d) = 0$ for $d \in D^c(y_j)$. Figure 5 depicts the overall structure of a random forest model.

As remarked on by Lee et al. [16], Γ (i.e., the number of tree classifiers) and Λ (i.e., the number of input variables randomly selected at each internal node) affect the performance of the random forest model. As Γ increases, the error of the model gradually decreases and converges to a certain value; therefore, we use a sufficiently large value for Γ. Like Lee et al. [16], we selected Λ value as that with minimum error among the possible integers in $[0.5\sqrt{M},\ 2\sqrt{M}]$ as well as $[0.5(log_2 M + 1),\ 2(log_2 M + 1)]$, where M is the number of variables in the input vector.

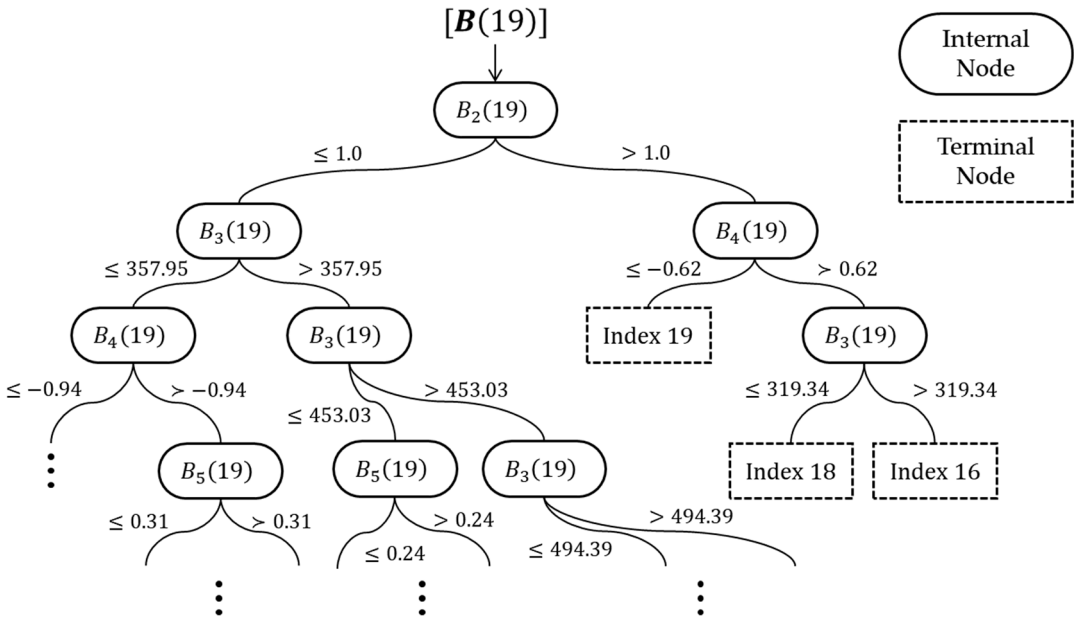

Figure 4. An example of part of a tree classifier.

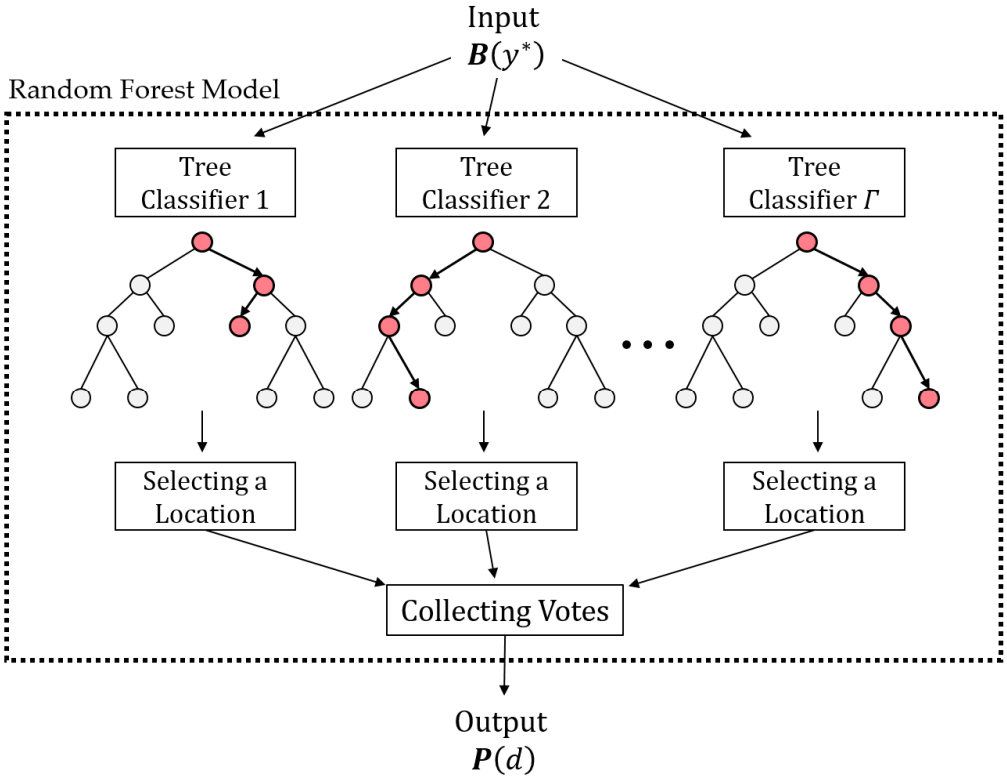

Figure 5. The overall structure of a random forest model for source identification.

4. Case Study

4.1. Simulation Setup for the Targeted River System

In our case study, we applied our framework to monitor a part of the Altamaha River system in the state of Georgia, USA. The whole river system is approximately 760 km long and has over 36,260 km^2 area with 60 reaches and 62 junctions. The fixed information for the SWMM, such as geometric, geologic, and hydrodynamic data, was obtained from the United States Geological Survey in the

National Elevation Dataset, and spill and rainfall events are considered variable information for the SWMM. Spill starting time and intensity followed uniform distributions in [0, 10] days and [10, 1000] grams per liter, respectively. The rainfall pattern was randomly selected among five rain patterns for each of 10 predesignated sub-catchments for the river. See Park et al. [24] and Telci et al. [25] for more details about the river system and the corresponding SWMM model.

For each run, the SWMM monitors hydrodynamics and contaminant levels of spills and rainfalls for 40 days. The related quantitative values (e.g., concentration levels, flow rates, and the amounts of overflows) are reported every 15 min in the simulation clock at each candidate location. Each training and test dataset is obtained by adding random errors whose density function is provided in Equation (4) to the concentration levels returned by the SWMM.

4.2. Experimental Setup

We consider a part of the Altamaha River as a targeted study area, as seen in Figure 6. The study area includes 53 candidate spill locations (i.e., $D = \{1, 2, \ldots, 53\}$), which are marked as gray circles in Figure 6. Based on research from Lee et al. [16], we set the number of sensors $K = 6$ and their locations, $\mathbf{y} = (9,\ 19,\ 26,\ 33,\ 46,\ 53)$. As mentioned in Section 3.4, the whole study area was divided into six sub-regions that were independently monitored by each sensor. The corresponding $D(y_i)$ values are represented in Figure 6.

For the random measurement errors, we consider two configurations to evaluate the impact of the errors: (i) low bias and low variability, denoted by (L, L), and (ii) high bias and high variability, denoted by (H, H). Specifically, we use $\mu_\xi = 0.002\xi_t(y_j)$ and $\sigma_\xi = 0.005 + 0.02\xi_t(y_j)$ for the (L, L) configuration and use $\mu_\xi = 0.05\xi_t(y_j)$ and $\sigma_\xi = 0.015 + 0.06\xi_t(y_j)$ for the (H, H) configuration as in Kim et al. [18]. For each of the random measurement error configurations, we searched the threshold value $\mathcal{H}$ of the CUSUM chart by using Equation (5) with $\rho = 1000$ (days) for each sensor (i.e., equivalent to the type I error, approximately 1.04×10^{-5}). Under each configuration of measurement errors, we identified the control limit, $\mathcal{H}$, for the CUSUM chart using Equation (5) when $\rho = 1000$. Thus, we set $\mathcal{H} = 0.228$ for the (L, L) configuration and $\mathcal{H} = 0.684$ for the (H, H) configuration. The monitoring window length for the CUSUM chart, ω, is set to 10 days and ℓ is set to 7.

Table 1 summarizes information about random forest models constructed by evaluating $D(y_j)$s. For each sensor location y_j, the random forest model $\phi(y_i)$ is constructed (and trained) by using a $\left|D(y_i)\right| \times 500$ number of simulation datasets where $|\cdot|$ represents the cardinality of a set. As mentioned in Section 3.4, we checked various values of Γ (i.e., the number of tree classifiers in a random forest model) and Λ (i.e., the number of input variables randomly selected at each internal node), and finally selected $\Gamma = 500$ and $\Lambda = 4$ for all models. We used the "StatsModels" package to generate each robust, locally-weighted regression and the "scikit-learn" package to train each random forest model in Python version 0.21.2 (provided by the Python Software Foundation, Beaverton, OR, USA) with a personal computer (Intel Xeon E5-1650 CPU; RAM 64GB). The average time required to generate each random forest model was approximately 10.7, 3.1, 3.9, 3.8, 10.5, 5.2 s. Table 1 also provides the out-of-bag (OOB) errors for each random forest model, which was calculated by the ratio misclassified datasets to the total number of training datasets for $\phi(y_i)$.

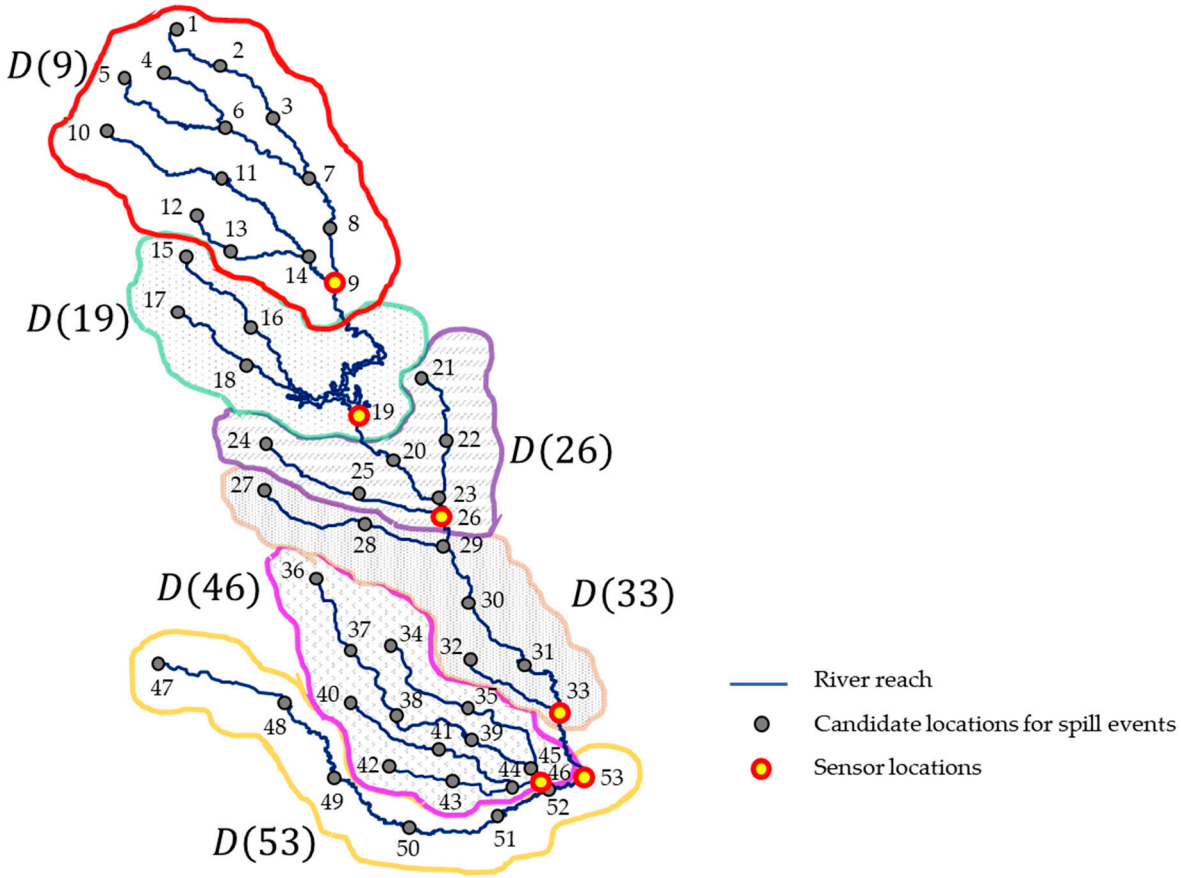

Figure 6. Targeted study area with six sensors.

Table 1. Random forest models with their OOB errors.

Sensor Location	Random Forest Model	Set of Candidate Spill Locations	% of OOB Error (L, L)	% of OOB Error (H, H)
9	$\phi(9)$	$D(9) = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14\}$	29.34	33.84
19	$\phi(19)$	$D(19) = \{15, 16, 17, 18, 19\}$	21.08	26.68
26	$\phi(26)$	$D(26) = \{20, 21, 22, 23, 24, 25, 26\}$	4.49	7.43
33	$\phi(33)$	$D(33) = \{27, 28, 29, 30, 31, 32, 33\}$	4.46	9.83
46	$\phi(46)$	$D(46) = \{34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46\}$	30.82	35.17
53	$\phi(53)$	$D(53) = \{47, 48, 49, 50, 51, 52, 53\}$	4.4	10.11

4.3. Results

We ran our framework on 53×100 test datasets (i.e., 100 spills occurred at each of 53 candidate locations) to evaluate the performance of source identification. Figures 7 and 8 show the identification accuracy rate (in y-axis) for the true spill location (in x-axis) under the (L, L) and (H, H) configurations, respectively. As the result of each test dataset, we can obtain a list of locations from the most promising to the least promising regarding the highest to lowest values of $P(d)$ returned by our framework. Figure 7a presents the percentage of times (i.e., the rate) that the correct source location is the first place on the list. Figure 7b,c presents the percentage of times (i.e., the rate) that the correct source location is within the top two or three places on the list, respectively. Figure 7a shows that spills occurred at about one-fifth of the locations are identified with a relatively low accuracy rate of less than 60% while the spills that occurred at about half of the locations are identified easily with an accuracy rate of more than 90%. However, regarding Figure 7b,c, high accuracy rates of more than 90% are achieved for all 53 locations. In sum, no matter where the spill occurs, users can find the true spill location with more

than 90% probability, if they visit the top three locations on the list. Figure 8 shows a similar pattern to Figure 7, but the identification accuracy rate in Figure 8 is slightly worse than that in Figure 7 because of higher bias and the variability of random measurement errors.

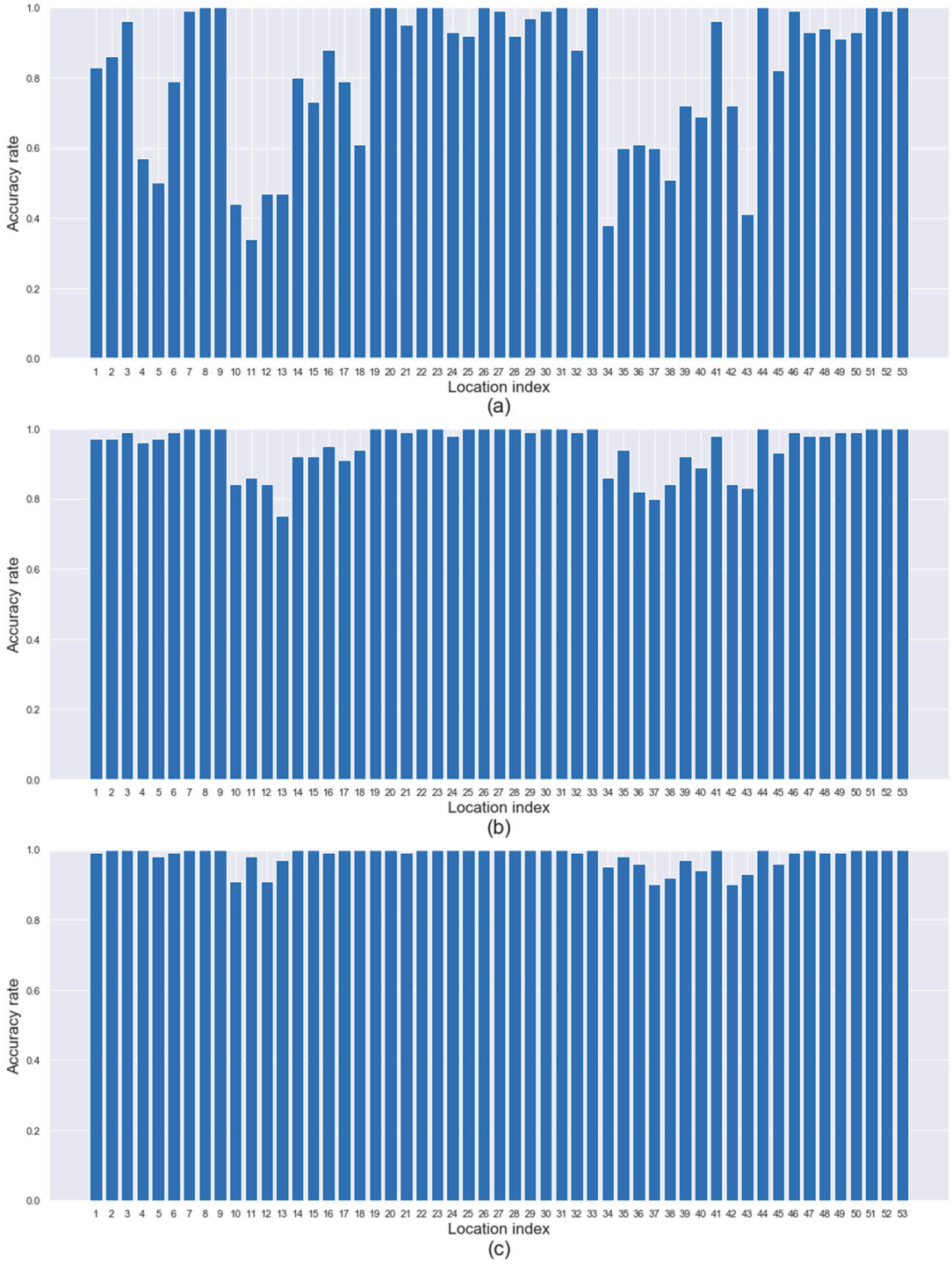

Figure 7. Identification accuracy regarding the top three locations under (L, L) configuration.

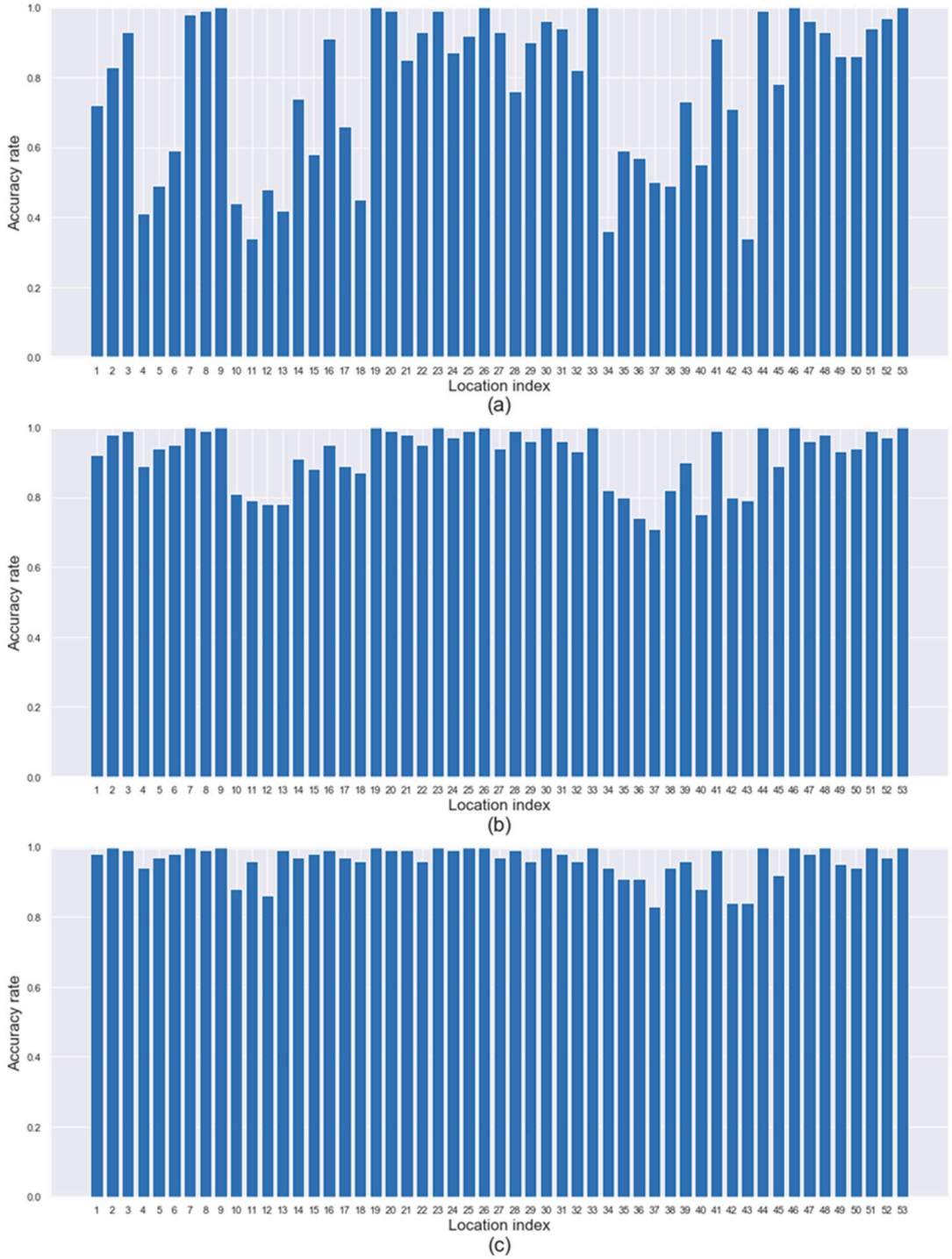

Figure 8. Identification accuracy regarding the top three locations under (H, H) configuration.

The random forest model is advantageous because it enables users to analyze the detailed possibility of the correct selection (e.g., $P(d)$). By comparing the obtained $P(d)$ values, users can pick stronger candidates for the spill quickly and efficiently. For example, see Figure 9 representing the averaged $P(d)$ values (denoted by $\hat{P}(d)$) at each location under the (L, L) configuration for some related pairs of locations. Let us consider the first pair (i.e., locations 4 and 5) in Figure 9, as an example. When a spill occurs at location 4 or 5, both locations 4 and 5 achieve significantly higher $\hat{P}(d)$ values than other locations. This implies that the spill at location 4 can often be confused with the spill at location 5. However, it implies that locations 4 and 5 can be recognized as promising

spill locations when compared with other locations, regardless of where the true spill location is. (Note that similar analyses can be done for other pairs in Figure 9.) Thus, users can distinguish a group of stronger candidates based on gaps among $P(d)$ values even when the true spill location is unknown, which is especially helpful in practical situations for finding unknown source locations under measurement errors.

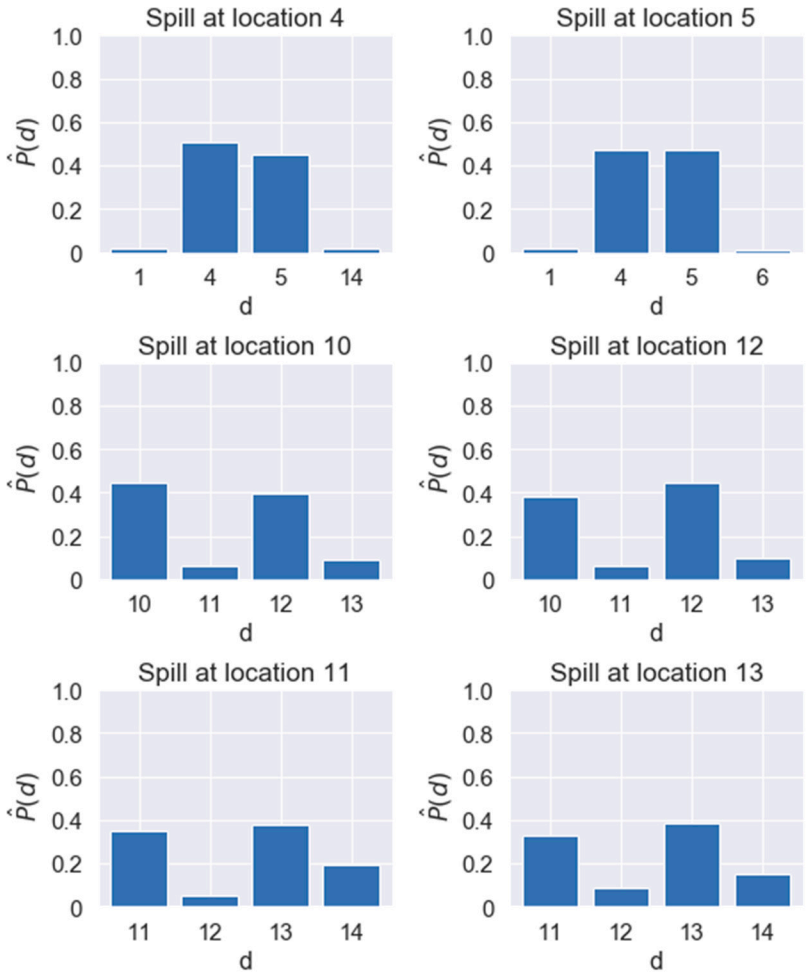

Figure 9. Examples of $\hat{P}(d)$ values for some related pairs of spill locations.

5. Conclusions

This study proposed a new framework to identify the source location of a contaminant spill in a river system with random sensor measurement errors. In the framework, one may first detect a contaminant spill using the CUSUM chart while the type I error of detection is controlled. The framework generates a nonlinear regression model for observation data with random errors to estimate a breakthrough curve, and derives characteristics of the estimated breakthrough curve for use as inputs to random forest models. After selecting one random forest model corresponding to the sensor that detected the spill, values between 0 and 1 are evaluated for the possibility of each candidate location being the true contaminant source location. Based on the detailed values, users can judge how likely each location is to be the true spill location, and analyze the gaps among the values to distinguish the most promising candidate locations instead of just picking locations regardless of the possibility. The test results of applying our framework to the Altamaha River system in the USA show that our framework performs well on the identification of a spill source location, even with measurement errors. In the test results, the true spill location is listed within the top three promising locations with more than 90% probability in most of the cases considered.

Our framework was originally proposed to identify a contaminant spill source location in a river system. However, beyond water monitoring systems, the framework can be applied to other areas including traffic and transportation, manufacturing process monitoring, and telecommunications.

While conducting this study, we determined that the identification accuracy of the framework can be improved if the exact starting and ending times of the breakthrough curve caused by the spill event are known. Further research on improving the accuracy of these time indices is ongoing.

Author Contributions: C.P., M.L.L., and J.H.K. designed the study performed the experiments, and analyzed the results; C.P. and M.L.L. wrote and revised the paper.

Acknowledgments: This research was supported by National Research Foundation of Korea (NRF) grants funded by the Korean government (MSIP) (No. 2016R1C1B2011462 and No. 2019R1F1A1061256) and 2019 Hongik University Research Fund.

Conflicts of Interest: The authors have no conflicts of interest to declare.

References

1. Aral, M.M.; Guan, J. Genetic algorithms in search of groundwater pollution sources. In *Advances in Groundwater Pollution Control and Remediation*, 2nd ed.; Aral, M.M., Ed.; Springer: Dordrecht, The Netherlands, 1996; Volume 9, pp. 347–369, ISBN 978-94-009-0205-3.
2. Aral, M.M.; Guan, J.; Maslia, M.L. Identification of contaminant source location and release history in aquifers. *J. Hydrol. Eng.* **2001**, *6*, 225–234. [CrossRef]
3. Gorelick, S.M.; Evans, B.; Remson, I. Identifying sources of groundwater pollution: An optimization approach. *Water Resour. Res.* **1983**, *19*, 779–790. [CrossRef]
4. Singh, R.M.; Datta, B. Identification of groundwater pollution sources using GA-based linked simulation optimization model. *J. Hydrol. Eng.* **2006**, *11*, 101–109. [CrossRef]
5. Sun, A.Y.; Painter, S.L.; Wittmeyer, G.W. A robust approach for iterative contaminant source location and release history recovery. *J. Contam. Hydrol.* **2006**, *88*, 181–196. [CrossRef] [PubMed]
6. Neupauer, R.M.; Lin, R. Identifying sources of a conservative groundwater contaminant using backward probabilities conditioned on measured concentrations. *Water Resour. Res.* **2006**, *42*. [CrossRef]
7. Neupauer, R.M.; Wilson, J.L. Numerical implementation of a backward probabilistic model of ground water contamination. *Groundwater* **2004**, *42*, 175–189. [CrossRef]
8. Sun, A.Y. A robust geostatistical approach to contaminant source identification. *Water Resour. Res.* **2007**, *43*. [CrossRef]
9. Singh, R.M.; Datta, B. Artificial neural network modeling for identification of unknown pollution sources in groundwater with partially missing concentration observation data. *Water Resour. Manag.* **2007**, *21*, 557–572. [CrossRef]
10. Singh, R.M.; Datta, B.; Jain, A. Identification of unknown groundwater pollution sources using artificial neural networks. *J. Water Resour. Plan. Manag.* **2004**, *130*, 506–514. [CrossRef]
11. Srivastava, D.; Singh, R.M. Breakthrough curves characterization and identification of an unknown pollution source in groundwater system using an artificial neural network (ANN). *Environ. Forensics* **2014**, *15*, 175–189. [CrossRef]
12. Boano, F.; Revelli, R.; Ridolfi, L. Source identification in river pollution problems: A geostatistical approach. *Water Resour. Res.* **2005**, *41*. [CrossRef]
13. Chen, Y.; Zhao, K.; Wu, Y.; Gao, S.; Cao, W.; Bo, Y.; Shang, Z.; Wu, J.; Zhou, F. Spatio-temporal patterns and source identification of water pollution in lake taihu (China). *Water* **2016**, *8*, 86. [CrossRef]
14. Ghane, A.; Mazaheri, M.; Samani, J.M.V. Location and release time identification of pollution point source in river networks based on the backward probability method. *J. Environ. Manag.* **2016**, *180*, 164–171. [CrossRef] [PubMed]
15. Telci, I.T.; Aral, M.M. Contaminant source location identification in river networks using water quality monitoring systems for exposure analysis. *Water Qual. Expo. Health* **2011**, *2*, 205–218. [CrossRef]
16. Lee, Y.J.; Park, C.; Lee, M.L. Identification of a contaminant source location in a river system using random forest model. *Water* **2018**, *10*, 391. [CrossRef]
17. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]

18. Kim, S.-H.; Aral, M.M.; Eun, Y.; Park, J.J.; Park, C. Impact of sensor measurement error on sensor positioning in water quality monitoring networks. *Stoch. Environ. Res. Risk Assess.* **2017**, *31*, 743–756. [CrossRef]
19. Montgomery, D.C. *Introduction to Statistical Quality Control*; Wiley: Hoboken, NJ, USA, 2009; ISBN 047151988X.
20. Kim, S.-H.; Alexopoulos, C.; Tsui, K.L.; Wilson, J.R. A distribution-free tabular CUSUM chart for autocorrelated data. *IIE Trans.* **2007**, *39*, 317–330. [CrossRef]
21. Rossman, L.A. *Storm Water Management Model User's Manual, Version 5.0*; U.S. Environmental Protection Agency: Cincinnati, OH, USA, 2004.
22. Cleveland, W.S. Robust locally weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc.* **1979**, *74*, 829–836. [CrossRef]
23. Ho, T.K. The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mech. Intell.* **1998**, *20*, 832–844. [CrossRef]
24. Park, C.; Telci, I.T.; Kim, S.-H.; Aral, M.M. Designing an optimal water quality monitoring network for river systems using constrained discrete optimization via simulation. *Eng. Optim.* **2014**, *46*, 107–129. [CrossRef]
25. Telci, I.T.; Nam, K.; Guan, J.; Aral, M.M. Optimal water quality monitoring network design for river systems. *J. Environ. Manag.* **2009**, *90*, 2987–2998. [CrossRef] [PubMed]