Journal of Cleaner Production 251 (2020) 119145

Contents lists available at ScienceDirect

Journal of Cleaner Production

journal homepage: www.elsevier.com/locate/jclepro

# An integrated data-driven framework for surface water quality anomaly detection and early warning

Jie Liu [a], Peng Wang [a, *], Dexun Jiang [b], Jun Nan [a, **], Weiyu Zhu [a]

[a] School of Environment, State Key Laboratory of Urban Water Resource and Environment, Harbin Institute of Technology, Harbin, 150090, China
[b] School of Information Engineering, Harbin University, Harbin, 150086, China

ARTICLE INFO

*Article history:*
Received 3 April 2019
Received in revised form
9 October 2019
Accepted 2 November 2019
Available online 25 November 2019

Handling editor: Prof. S Alwi

*Keywords:*
Surface water quality
Anomaly detection
Early warning
Bayesian autoregressive model
Isolation forest algorithm

ABSTRACT

The surface water quality anomaly detection for rapid early warning is essential to prevent potential harmful compounds, resulting from river environmental spills or intentional releases, from dispersing in large scale. In this study, an effective data-driven framework for surface water quality anomaly detection is developed to provide early warnings for dealing with river environmental pollution in advance. The developed framework is constructed by an integration of Bayesian autoregressive (BAR) model for water quality variation prediction and Isolation Forest (IF) algorithm for water quality anomaly detection. First, an autoregressive method based on Bayesian inference is used to forecast the tendencies of water quality variations. Second, an IF algorithm is applied to identify the features of water quality anomalies using the prediction residuals obtained in the previous stage. The integration framework is then applied to analyze and detect the surface water quality variations and anomalies in Potomac River of West Virginia, USA, comparing with prediction-based anomaly detection method, classification-based anomaly detection method, and different scenarios. The results demonstrate that the developed integration framework not only could enhance water quality anomaly detection accuracy, but also effectively provide early warning for emergency operations in a quick response.



## 1. Introduction

In recent years, with an accelerated development of industrialization and urbanization, the surface water quality is suffering from increasingly serious deterioration due to the intense human activities, frequent river environmental pollution, and extreme weather condition, especially in China and other developing countries (Bertone et al., 2016; Wang et al., 2018; Li et al., 2018). In order to improve emergency response capacity and protect water environmental quality from potential harmful compounds resulting from river environmental spills or intentional releases, it is imperative to develop an effective detection method in order to identify water quality variations and anomalies, and provide rapid early warning for surface water environmental emergency management (Shi et al., 2018; Leigh et al., 2019).

Anomaly detection, or outlier detection, is a procedure for finding the patterns in dataset which do not conform to the expected pattern or deviate from the expected behavior greatly. Recently, anomaly detection methods have been widely used in various fields for water quality anomaly detection, such as water distribution system analysis (Housh and Ostfeld, 2015), surface water quality monitoring (Hill and Minsker, 2010), groundwater quality management (Jeong et al., 2017a). In general, anomaly detection methods can be categorized as statistic-based algorithm (Hou et al., 2015), distance-based algorithm (Knorr et al., 2000), density-based algorithm (Breunig et al., 2000), clustering-based algorithm (He et al., 2003), prediction-based algorithm (Arad et al., 2013) and isolation-based algorithm (Liu et al., 2008). In many water quality anomaly detection cases, traditional statistic-based anomaly detection algorithms are adopted to identify water quality dataset anomalies, assuming that normal data instances occur in high probability regions of a stochastic model, while anomalies occur in low probability regions of the stochastic model, such as Gaussian distribution and Poisson distribution (Jeong et al., 2017b). However, most of the statistic-based algorithms rest on the premise of a probability distribution model and are not suitable for high-dimensional data which may subject to different distributions. In terms of clustering-based anomaly detection algorithms, Euclidean distance or density-attractor are employed to evaluate distance between data instances and sample center or look for the dense areas of data instances for water quality anomaly detection and pattern discovery (Liu et al., 2015; Deng and

* Corresponding author.
** Corresponding author.
*E-mail addresses:* pwang73@hit.edu.com (P. Wang), nanjun11@163.com (J. Nan).

https://doi.org/10.1016/j.jclepro.2019.119145
0959-6526/

Wang, 2017). Since outliers are supposed to be far away from the most normal samples or exist in area with low-density normal samples, it is difficult for clustering-based algorithms to determine valid parameters and global threshold if the anomalous data have high local density. For classification-based anomaly detection algorithms, classification models are established using the normal-labeled samples to identify the deviate degree of abnormal data from normal data instances (Koch and Mckenna, 2011). Normal-unlabeled samples may have large influences on the classification accuracy and the generalized ability.

Recently, prediction-based anomaly detection algorithms are becoming an essential tool for water quality anomaly detection and early warning as an emerging technique without consideration of various related model parameters in contrast to traditional mechanism prediction model which is always restricted by limited information and high uncertainties. In general, autoregressive (AR) model, support vector machine (SVM) and artificial neural networks (ANNs), as prediction-based algorithms, which are inherently more efficient, have been used to forecast the water quality variation tendencies and detect water quality anomalies using the abnormal thresholds of prediction residuals (Perelman et al., 2012; Hou et al., 2013; Jin et al., 2019). However, most of the prediction-based anomaly detection algorithms are usually analyzed according to variation features of a single water quality parameter (Shi et al., 2018). The fluctuations and anomalies of prediction residuals, which are identified by prediction-based algorithms using a single water quality parameter, may sometimes be caused by sensor fault or equipment noise, but not by contaminants injected into the surface water system (SWS). Hence, it is indispensable that the combination of multiple water quality parameters should be detected to enhance the accuracy of water quality anomaly detection results.

Isolation Forest (IF) algorithm, as an isolation-based anomaly detection algorithm, can efficiently deal with anomaly data in the massive dataset with multiple dimensions, has lower computation complexity compared to other existing methods, and is of great significance in many applications, such as cloud information (Calheiros et al., 2017), groundwater level monitoring (Azimi et al., 2018), semiconductor manufacturing management (Puggini and McLoone, 2018), and mineral exploration analysis (Chen and Wu, 2018). IF algorithm can provide a synchronized decision support for water quality anomaly detection by data fusion using multiple quality parameters and employ only a tiny proportion of training data to build effective models with a linear time complexity and low memory requirement which is ideal for high volume data sets (Liu et al., 2008). Since the binary trees are generated and regarded as independent of each other, the IF algorithm can be deployed on the large-scale distributed parallel computing system to accelerate the computation. An integration method of prediction-based and isolation-based method can make a difference in water quality anomaly detection and early warning.

Therefore, the objective of this study is to develop an effective data-driven framework by an integration of prediction-based and isolation-based methods for surface water quality anomaly detection according to water quality variation prediction and help decision makers (DMs) provide early warnings for dealing with river environmental pollution in advance. An autoregressive method based on Bayesian inference is used to forecast the tendencies of water quality variations in the first stage. Then an IF algorithm is applied to identify the features of water quality anomalies based on the prediction residuals in the second stage. The obtained results are beneficial for DMs to guarantee the efficiency of water quality anomaly detection, and enhance the emergency response ability in dealing with river environmental pollution. The paper is organized as follows: Section 2 presents the development of a data-driven framework, including Bayesian autoregressive (BAR) model for water quality variation prediction and IF algorithm for water quality anomaly detection. Section 3 describes the application in Potomac River of West Virginia based on the developed integration framework, and demonstrates the related results and discussions comparing with different methods and scenarios, where the performance analysis and comparison of water quality anomaly detection are analyzed by Receiver Operating Characteristic (ROC) curves and Area Under roc Curve (AUC) values. Section 4 gives some conclusions.

## 2. Methods

In this study, an effective data-driven framework integrating BAR model with IF algorithm is developed in order to forecast the tendencies of water quality variations, identify the features of water quality anomalies and provide emergency warning in advance. The developed integration framework is composed of three main phases: (1) water quality variation prediction, (2) water quality anomaly detection, and (3) performance analysis. In water quality variation prediction phase, the tendencies of water quality variations are forecasted by autoregressive model based on Bayesian inference. In water quality anomaly detection phase, the features of water quality anomalies are identified by IF algorithm using the prediction residuals obtained in the previous stage. In performance analysis phase, two metric measurements are used to numerically quantify the anomaly detection performance and analyze the accuracy and efficiency of the developed anomaly detection framework.

### 2.1. BAR model for water quality prediction

In this study, A BAR model is developed and applied to multivariate time series model for water quality time-varying forecast in order to simulate the major features of water quality variations. A set of time series observation $Y_t$ is built to analyze the time persistence of water quality parameters, such as turbidity (TURB), specific conductance (SC), dissolved oxygen (DO). In AR model, the observation in the next time point can be expressed by the linear combination and random error of the observations in the previous $p$ time points of the time series. A $p$th-order AR model is also called an AR model with p lags which can be expressed as (Enders and Sandler, 1993):

$$Y_t = A_1 Y_{t-1} + A_2 Y_{t-2} + \cdots + A_p Y_{t-p} + BX_t + \in_t \quad (1)$$

where $p$ is the number of time points (order of the AR model); $Y_{t-1}$, $Y_{t-2}$, …, $Y_{t-p}$ are the historical observations; $A_1$, $A_2$, …,$A_p$ are the coefficients of the historical observations; the vector $BX_t$ is a set of constant variables; and the vector $\in_t$ is assumed to be white noise. Stationarity of time series is examined by Augmented Dickey Fuller (ADF) Test. Bayesian information criterion (BIC) is used to determine the order $p$ of the AR model which can be calculated as (Wit et al., 2012):

$$BIC = \ln(n)k - 2\ln(\widehat{L}) \quad (2)$$

$$\widehat{L} = P(x|\widehat{\theta}, M) \quad (3)$$

where $\widehat{L}$ is the maximized value of the likelihood function for the estimated model $M$; $\widehat{\theta}$ are the parameter values that maximize the likelihood function; $x$ are the observation data; $n$ is the number of data points in $x$, or equivalently, the sample size; $k$ is the number of parameters to be estimated. If the estimated model is a linear regression, $k$ is the number of regressors, including the constant. In general, $p$ can be supposed to be the best order of the AR model when BIC reaches the minimized value. BAR model can be established with order $p$, and the current water quality data then can be

forecasted based on the historical observation data.

### 2.2. IF algorithm for anomaly detection

In this study, a two-stage IF algorithm is developed based on Liu et al. (2008) and applied to water quality anomaly detection for surface water pollution early warning in order to deal with potential environmental accidents in advance. More specifically, isolation trees (iTrees) are constructed as an isolation forest (iForest) using sub-samples of the training data set in the first stage for data training process. Then, an anomaly score for each instance is obtained from the test instances based on the iTrees in the second stage for data testing process. The base IF algorithm used in this study is described in the following subsections.

#### 2.2.1. Model training process

Given a sample of training data set $X = \{x_1, x_2, \ldots, x_n\}$ with attribute dimension $D$, an attribute $d$ and a split value $p$ are randomly selected in order to build an iTree. In the data training process, the training data set is recursively partitioned and iTrees are finally established until one of the following conditions is arbitrarily satisfied: (1) the given instances are isolated; (2) the attributes of given instances are same with each other; or (3) a specific iTree height reaches at the depth limit. The iTree height limit $l$, which is equivalent to the average iTree height, can be automatically set by the sub-sampling size $\psi$ using Eq. (4):

$$l = ceiling(\log_2 \psi) \tag{4}$$

An iForest is constructed based on Algorithm 1 adapted from Liu et al. (2008). In general, the sample instances which have shorter-than-average path lengths are likely to be isolated as to be anomaly instances in the process of growing iTree up to the average tree height. In this stage, two input parameters are applied to the IF algorithm, including number of iTrees $t$, and sub-sampling size $\psi$. More specifically, sub-sampling size $\psi$ determines the training data size, which are of important influence on processing time and memory size in anomaly detection performance. And number of iTrees $t$ determines the ensemble size, which can enhance the anomaly detection accuracy. Path lengths usually converge well before $t = 100$ (Liu et al., 2008). In this study, $\psi = 256$ and $t = 100$.

Supposing that $N$ is a node of an iTree, it may be a leaf node with no sub-node, or a non-leaf node with two sub-nodes (left child node and right child $N_r$). Recursively partitioning process is executed by randomly selecting attribute $d$ and a split value $p$. The test instances would be distributed to $N_l$ if the attribute values of the test instances $x_d < p$, while the test instances would be distributed to $N_r$ if the attribute values of the test instances $x_d \geq p$. An iTree is generated based on Algorithm 2 adapted from Liu et al. (2008).

**Algorithm 1:** *iForest* $(X, t, \psi)$,

**Inputs:** Sampling instances $X$; Number of iTrees $t$; Sub-sampling size $\psi$;
**Outputs:** A set of $t$ iTrees
**Procedure:**

```
begin
  initialize Forest
  set iTree height limit l according to Eq.(4)
  for i = (1, 2,...t) do
    X' ← sub-sampling (X, ψ)
    Forest ← Forest ∪ iTree (X', 0, l)
  end for
  return Forest
end
```

**Algorithm 2:** *iTree* $(X', e, l)$,

**Inputs:** Sub-sampling instances $X'$; Current tree height $e$; iTree height limit $l$;
**Outputs:** An iTree
**Procedure:**

```
begin
  if e ≥ l or |X'| ≤ 1 then
    return leaf node
  else
    randomly select an attribute d ∈ D of X'
    randomly select a split value p ∈ (min(x_d), max(x_d))
    X_l ← filter(X', x_d < p)
    X_r ← filter(X', x_d ≥ p)
    return non-leaf node {left child ← iTree (X'_l, e+1, l),
                          right child ← iTree (X'_r, e+1, l),
                          split attribute ← d,
                          split value ← p }
  end if
end
```

*2.2.2. Model testing process*

In the first stage, $T$ iTrees are built by recursively partitioning. An iForest is then defined by $T$ iTrees.

$$iForest = \{t_1,\ t_2,\ \ldots,\ t_T\} \tag{5}$$

Then, abnormal levels are evaluated according to the path lengths of the sample instances for each iTree $t$ in the second stage. For a sample instance $x$, the path length $h_t(x)$ from leaf node to root node is the number of iterations when the sample instance $x$ can be isolated in each iTree $t$. Expected value $E\,(h(x))$ of all path lengths among $T$ iTrees can be obtained as:

$$E(h(x)) = \frac{1}{T}\sum_{t=1}^{T} h_t(x) \tag{6}$$

The abnormal levels of sample instance $x$ can be calculated by anomaly score $s\ (x,\ \psi)$. In general, an anomaly instance can be isolated only in a few iteration steps which are influenced by the sub-sampling size. Based on the Python implementation of IF algorithm available in the scikit-learn library, a normalized anomaly score $s\ (x,\ \psi)$ can be defined as (Stripling et al., 2018):

$$s(x,\ \psi) = 0.5 - 2^{-\frac{E(h(x))}{c(\psi)}} \tag{7}$$

where $\psi$ is the sub-sampling size; $c(\psi)$ is the average path length of Binany Searching Tree (BST), which can be used to normalize the $h_t(x)$. And $c(\psi)$ can be expressed as:

$$c(\psi) = \begin{cases} 2H(\psi-1) - (2(\psi-1)/\psi) & if\ \ \psi > 2 \\ 1 & if\ \ \psi = 2 \\ 0 & otherwise \end{cases} \tag{8}$$

where $H(i)$ is the harmonic number of sub-sampling size $i$, and can be estimated by Euler-Mascheroni constant:

$$H(i) \approx \ln(i) + 0.5772156649 \tag{9}$$

For a sample instance $x$, the path length $h_t(x)$ can be obtained based on Algorithm 3 adapted from Liu et al. (2008). The path length $h_t(x)$ of a sample instance $x$ is revised according to Eq. (6) based on the number of instances $n$ at the leaf node. And $h_t(x) = e + c(n)$.

According to anomaly score $s\ (x,\ \psi)$, three conditions can be obtained as follows:

(1) When $E(h(x)) \to c(\psi),\ \ s \to 0$, the average path length of a sample instance $x$ is approximately equal to the average path length of the iTree. The sample instance $x$ cannot be judged as an anomaly instance;
(2) When $E(h(x)) \to 0,\ \ s \to -0.5$, the anomaly score of a sample instance $x$ is approximately equal to $-0.5$. The sample instance $x$ can be judged as an anomaly instance;
(3) When $E(h(x)) \to n-1,\ \ s \to 0.5$, the anomaly score of a sample instance $x$ is approximately equal to 0.5. The sample instance $x$ can be judged as a normal instance.

*2.3. Performance analysis for anomaly detection*

In order to compare the performance of the anomaly detection algorithm, it is important to assess the accuracy and efficiency of the developed method. In this study, two metric measurements are conventionally used to numerically quantify the efficiency of anomaly detection performance, including the one-dimensional and two-dimensional measures.

*2.3.1. True positive rate and false positive rate*

In this study, two single measures true positive rate (TPR) and false positive rate (FPR) as one-dimensional metric method, are used to demonstrate anomaly detection rate and false alarm/fall-out rate (Perelman et al., 2012). True positive represents that actual anomaly observations are identified as anomaly conditions through anomaly detection method. And false positive represents that actual normal observations are identified as anomaly conditions through anomaly detection method. TPR and FPR can be calculated as (Perelman et al., 2012):

$$TPR = \frac{TP}{TP + FN} = Sensitivity \tag{10}$$

$$FPR = \frac{FP}{TN + FP} = 1 - Specificity \tag{11}$$

where $TP$ represents true positive, which means that the actual water quality is abnormal, and anomaly detection result is also

**Algorithm 3:** *PathLength* $(x, N, e)$

**Inputs:** A sub-sampling instance $x$; An iTree node $N$; Current tree height $e$;
**Outputs:** Path length of $x$ $\{h(x)\}$
**Procedure:**

```
begin
  if N is a leaf node and contains n instances {one instance/ n same instances/ n
  different instances constraint with e<l} then
    return e+c( n_N ) according to Eq.(8)
  else { N is a non-leaf node }
    a ← Split attribute q at node N
    if x_a < Split value p then
      return PathLength (x, N_l, e+1)
    else { x_a ≥ Split value p }
      return PathLength (x, N_r, e+1)
    end if
  end if
end
```

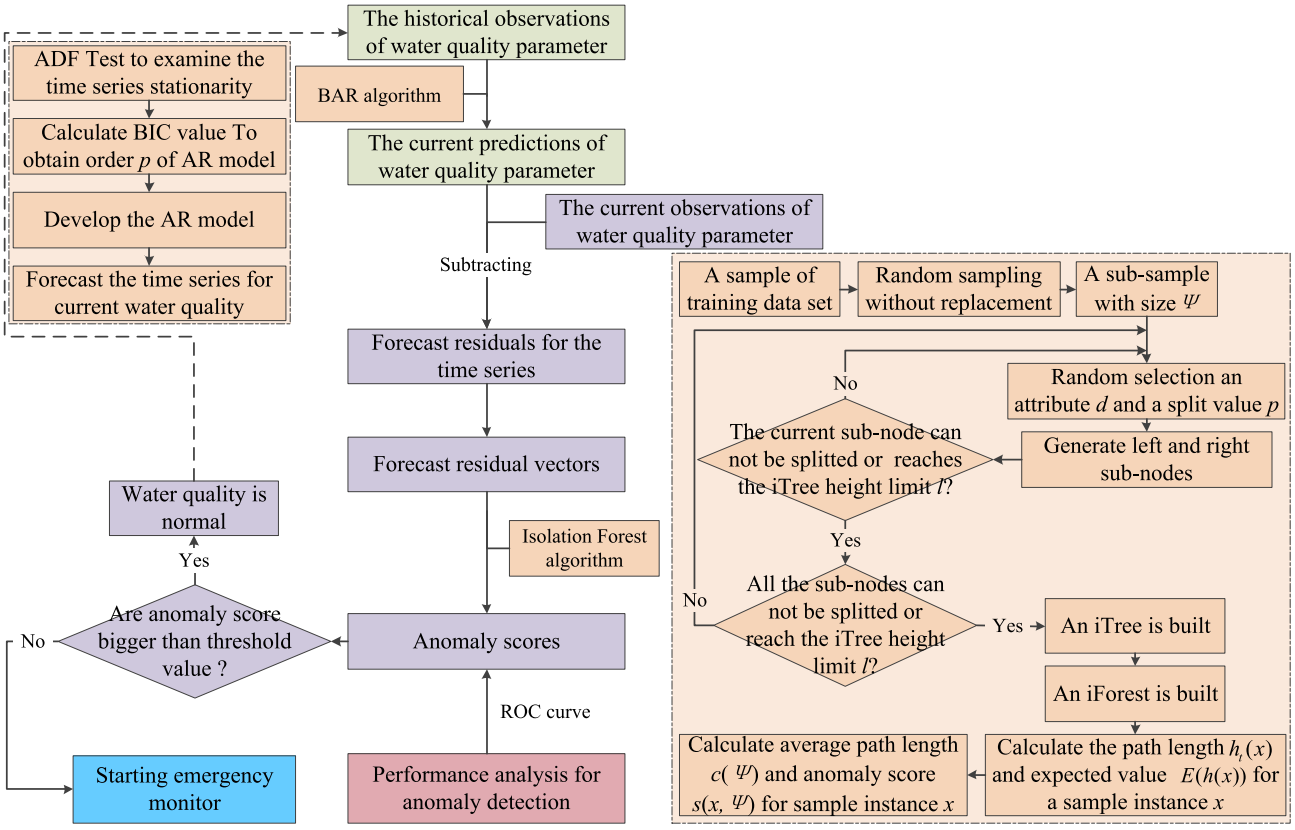


**Fig. 1.** Flow chart of the data-driven framework.

abnormal; *FN* represents false negative, which means that the actual water quality is abnormal, while anomaly detection result is normal; *FP* represents false positive, which means that the actual water quality is normal, while anomaly detection result is abnormal; *TN* represents true negative, which means that the actual water quality is normal, and anomaly detection result is also normal.

### 2.3.2. ROC curve and AUC value

In this study, ROC curve, as the two-dimensional metric method, is used for performance analysis and quantitative evaluation for the anomaly detection information (Housh and Ohar, 2017). It is a graphical plot which illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. In ROC curve, the abscissa denotes FPR value, reflecting the specificity of the anomaly detection or probability of false alarm. And the ordinate denotes TPR value, reflecting the sensitivity of the anomaly detection or probability of detection. AUC value, the areas under ROC curve, is used to evaluate the accuracy of the anomaly detection. AUC values can be calculated according to Fawcett (2006). In general, the points in the top left of the ROC curve, have higher critical values of specificity and sensitivity than others. And the bigger AUC values are, the higher reliability and accuracy of the anomaly detection algorithm is. AUC values are expected to be distributed between 0.5 and 1. AUC = 1 indicates a perfect detector, which detects all anomalies without any false alarms, while AUC = 0.5 indicates a random detector (Khreich et al., 2017). The Youden index is generally applied to optimize the threshold point of the ROC curve for anomaly detection method, and is defined as the difference between TPR and FPR (Maso and Montecchio, 2014).

$$\Upsilon = Sensitiviey + Specificity - 1 = TPR - FPR \quad (12)$$

The maximum of the Youden index is used to select the optimal point on the ROC curve with the farthest vertical distance from the diagonal line (Bantis et al., 2014; Pang et al., 2018). A graphical view of the developed framework is shown in Fig. 1.

## 3. Case study for water quality anomaly detection

In this study, the application of the developed data-driven framework is described for Potomac River in West Virginia to illustrate the related results for water quality variation prediction and anomaly detection, and demonstrate the efficacy of the developed method comparing with different methods and scenarios.

### 3.1. Study area and monitoring data

Potomac River, which is located in West Virginia, is one of the most important rivers of the east-central United States with the river basin area of 37,000 km$^2$ (Fig. 2). It originates from the western foothills of the Appalachian Mountains and is formed by the confluence of North Blanche River and South Blanche River, finally flowing into the Chesapeake Bay. Potomac River Basin is the main water resource providing area for the cities with high population density (approximately 5 million people). However, the river is subject to the impacts of severe environmental pollution due to intense human activities and water eutrophication (Byrand, 2010). The monitoring program is applied to the Potomac River by the United State Geological Survey (USGS) and sufficient water quality monitoring sites are located for high-frequency time series of major water quality parameters which provides enough basic data for this study.

In this study, the surface water quality data for Potomac River

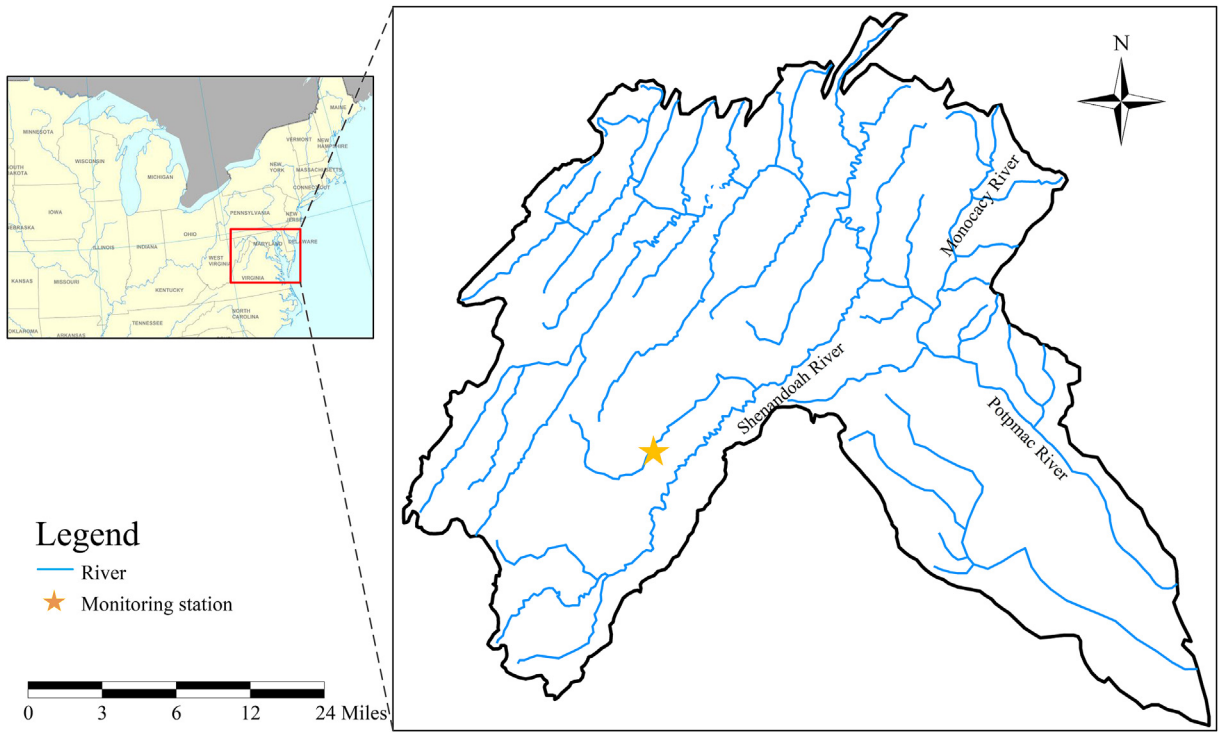


**Fig. 2.** Location and spatial distribution of Potomac River.

are obtained from USGS (https://waterdata.usgs.gov/nwis/sw), which are mainly detected by sensors and via manual field measurements. Monitoring site No.01632900, which is located at upstream tributary of Potomac River, is selected as a major station for validating the proposed anomaly detection approach. And three surface water quality parameters are selected to determine whether surface water quality are abnormal and applied to the developed integration framework to forecast the tendencies of water quality variations and identify the features of water quality anomalies, including water quality parameters of TURB, SC and DO. The time series observation data are commonly recorded at a fixed interval of 15 min in order to obtain high-frequency time series data. In this study, the observation period, which spans from January 27, 2017 (0:00) to March 21, 2017 (0:00), is selected as the major study period of investigation, accounting for 5089 groups observation data. More specifically, the time series observation data from March 18 to March 21 showed surface water quality anomalies due to intensity rainfall in short duration. This event can be considered as a water quality anomaly. Table 1 shows the raw data statistics of the water quality parameters at monitoring site No.01632900. In this study, the original observation data set $Y_i$ is standardized via Z-score transform with an average of 0 and a variance of 1:

$$D_i = \frac{Y_i - \overline{Y}}{\sigma} \tag{13}$$

where $\overline{Y}$ is the mean of $Y_i$ and $\sigma$ is the variance of $Y_i$. The standardized time series of water quality parameters (observation values) are shown in Fig. 4(a1, b1 and c1). The water quality parameter of DO fluctuates periodically during the study periods. The water quality parameters of SC and TURB show no obvious periodic characteristics. These two variables are sensitive to surrounding environmental conditions, and slight changes in the water environment can change these two variables considerably.

**Table 1**
Raw data statistics of the water quality parameters in monitoring site.

| Parameters | TURB (NTU) | SC (μS/cm) | DO (mg/L) |
|---|---|---|---|
| Maximum | 23.00 | 520 | 14.80 |
| Minimum | 0.00 | 429 | 7.90 |
| Mean | 1.54 | 492.36 | 11.54 |
| Variance | 1.65 | 15.99 | 1.33 |

### 3.2. Results and discussion

In this study, an effective data-driven framework is developed, including three main phases: (1) water quality variation prediction, (2) water quality anomaly detection and (3) performance analysis. The time series of water quality parameters including TURB, SC and DO are used to analyze and verify the efficiency of the developed anomaly detection method.

#### 3.2.1. Water quality prediction by BAR model

In the water quality variation prediction phase, the standardized historical observations as the training set (1920 groups), which span from January 27, 2017 to February 15, 2017, are used to train the parameters of BAR model. The standardized historical observations as the testing set (3169 groups), which span from February 16, 2017 to March 21, 2017, are used to verify the prediction performance of BAR model, comparing with the predictions. In this study, a one-step-ahead prediction with a sliding window pattern is adopted using the current historical observations to forecast the next 15-min water quality prediction value. In BAR model training process, the BIC value and order $p$ of BAR model are two important factors that should be considered to improve water quality prediction efficiency and prevent the BAR model from becoming over-trained. Fig. 3 shows the relationship between the BIC values and orders of BAR models for TURB, SC and DO. According to the BIC values, the optimal number orders of BAR models for TURB, SC and DO can be selected as

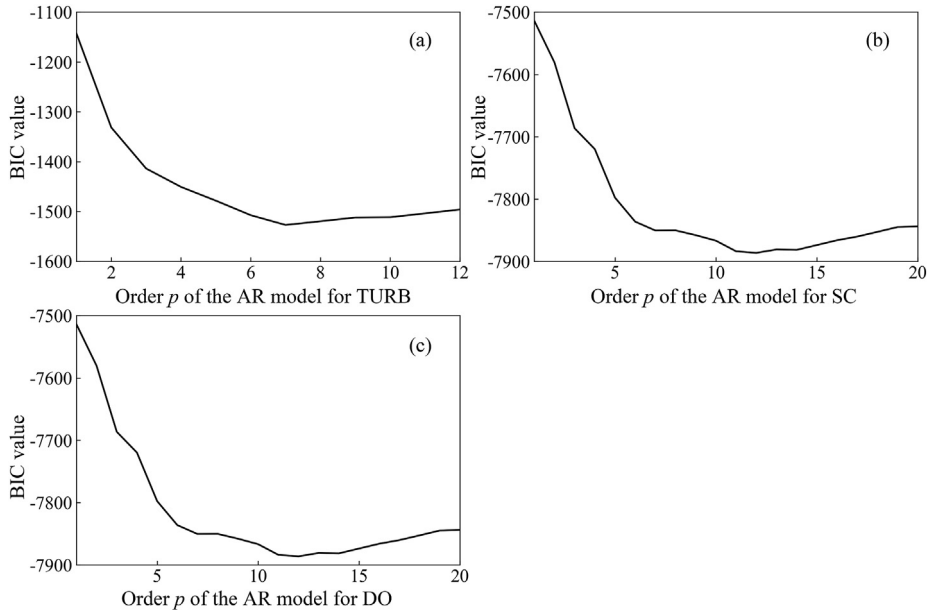


**Fig. 3.** The relationship between the BIC values and orders of BAR model.

7, 12 and 7. The current surface water quality parameters can be calculated by the linear combination of the last number of 7, 12 and 7 historical observations. In BAR model testing process, Fig. 4 shows the prediction results and the prediction residuals of TURB, SC and DO comparing with the standardized historical observations. It is indicated that the prediction results can effectively reflect the current water quality variation conditions during the steady period of water quality variation from January 16, 2017 to March 12, 2017, while the prediction results of TURB and SC show great differences from current observations during the anomaly period of water quality variations from March 18, 2017 to March 21, 2017. The maximum absolute values of prediction residuals of TURB and SC can reach to 11.35 and 2.11. The observations of DO are not affected by the emergency environmental influence with the maximum absolute value of prediction residuals at 0.07. In this study, a one-step-ahead water quality prediction BPNN model (Jin et al., 2019) is used to identify the reliability of the prediction results of BAR model. To maintain a reasonable neural network construction for the BPNN model, 7, 12, 7 neurons are used in the input layer according to the BIC values above, 6 neurons are used in the hidden layer, and 1 neuron is used in the output layer for TURB, SC and DO prediction. The standardized historical observations are discretized into an input variable as a sliding window pattern. A comparison result of error analysis (from January 16, 2017 to March 12, 2017) between the BAR model and BPNN model are shown in Table 2. Although the prediction errors are slightly different, water quality prediction values obtained by the two models are very similar in a long time range, indicating that both the BAR model and BPNN model can identify water quality variation tendencies and provide water quality prediction values within an acceptable error range. Hence, the BAR model used in this study can reflect the current water quality variation conditions and provide a reasonable and reliable prediction values for water quality anomaly detection phase.

### 3.2.2. Water quality anomaly detection by IF algorithm

In the water quality variation prediction phase, the predictions (3169 groups), which span from January 16, 2017 to March 21, 2017, are forecasted by BAR model. And the prediction residual values can be obtained by subtraction operations between observations and predictions. In the water quality anomaly detection phase, the prediction residual values as the training set (2400 groups), which span from January 16, 2017 to March 12, 2017, are used to train iForest consisting of multiple iTrees. More specifically, each iTree is constructed by randomly selecting 256 sub-samples from the training set with the constraint of height limit $l = 8$. Then, 100 iTrees are built by recursively partitioning in order to train iForest and obtain the anomaly scores of the training set. Then, the prediction residual values as the testing set (769 groups), which span from March 13, 2017 to March 21, 2017, are used to calculate the anomaly scores of the testing set and evaluate anomaly degree for surface water quality anomaly detection. In this study, pair-wise training and testing pattern is adopted to detect surface water quality anomaly considering the prediction residuals of TURB, SC and DO together. And a threshold setting $S$ of anomaly score is set at $-0.0334$ based on maximum value of the Youden index and it is supposed to be water quality anomaly conditions if the anomaly scores of the prediction residuals are less than $-0.0334$. The water quality anomaly detection results are obtained in Fig. 5. It is shown that most of the outlier sample instances can be verified and detected as water quality anomalies and the whole water quality anomaly detection results are reasonable to provide early warnings for environmental emergency management. However, Fig. 5(b) shows that some sample instances cannot be detected as the water quality anomalies using TURB and DO attributes due to the lack of anomaly information of SC attribute. In terms of 5(a) and 5(c), it is shown that a portion of sample instances in $-0.24$-contour line region can be effectively detected due to the obvious anomalies of prediction residual values with TURB attribute. However, some of sample instances in $-0.18$-contour line region are identified as normal conditions, since these sample instances have no TURB anomaly features and conspicuous DO anomaly features comparing with other anomaly instances. Once the developed detection framework can identify the water quality anomalies with continuous time series for a river environmental pollution, an emergency environmental monitoring should be started for early warning.

### 3.2.3. Performance analysis and comparison by ROC curve

In this study, two measures TPR and FPR are used to

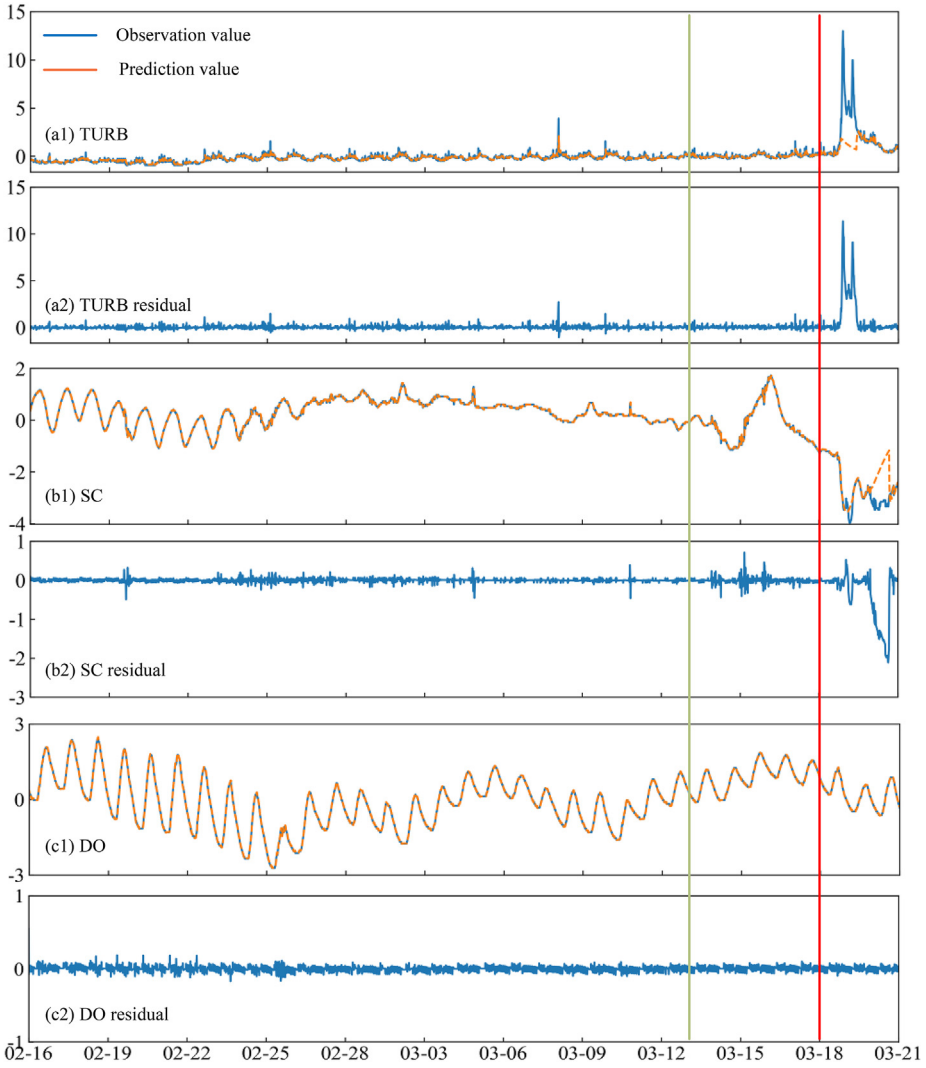


**Fig. 4.** The prediction results of BAR model for case study.

demonstrate the anomaly detection rate and false alarm rate. ROC curves are obtained with FPR values as the abscissa and TPR values as the ordinate, depicting the relative tradeoffs between benefits (true positives) and costs (false positives). The areas between the ROC curves and x-axis are calculated as AUC values to evaluate the accuracy of the anomaly detection method. For performance analysis of anomaly detection, a scoring classifier can be used with a threshold to generate a discrete (binary) classifier. The conditions of surface water quality from March 18 to March 21, which showed water quality anomalies due to intensity rainfall in short duration, are supposed be 1 as anomaly label. Other conditions of surface water quality are supposed be 0 as normal label. A water quality anomaly classification based on real abnormal conditions would be obtained. If a threshold setting of the anomaly detection method is selected, a water quality anomaly classification based on the anomaly detection method would be obtained. Comparing with the two classifications obtained by true event and the anomaly detection method, TPR and FPR values can be calculated based on the selected threshold setting.

(1) Performance analysis for case study

In this study, the developed anomaly detection method is integrated by prediction-based and isolation-based methods. In order

**Table 2**
Performance analysis for prediction results of the BAR and BPNN model.

| | Evaluation index | TURB | SC | DO |
|---|---|---|---|---|
| BAR model | Mean absolute error (MAE) | 0.1086 | 0.0453 | 0.0282 |
| | Mean square error (MSE) | 0.0287 | 0.0069 | 0.0011 |
| | Root mean squared error (RMSE) | 0.1694 | 0.0831 | 0.0332 |
| BPNN model (Jin et al., 2019) | Mean absolute error (MAE) | 0.1040 | 0.0275 | 0.0294 |
| | Mean square error (MSE) | 0.0289 | 0.0020 | 0.0018 |
| | Root mean squared error (RMSE) | 0.1700 | 0.0450 | 0.0430 |

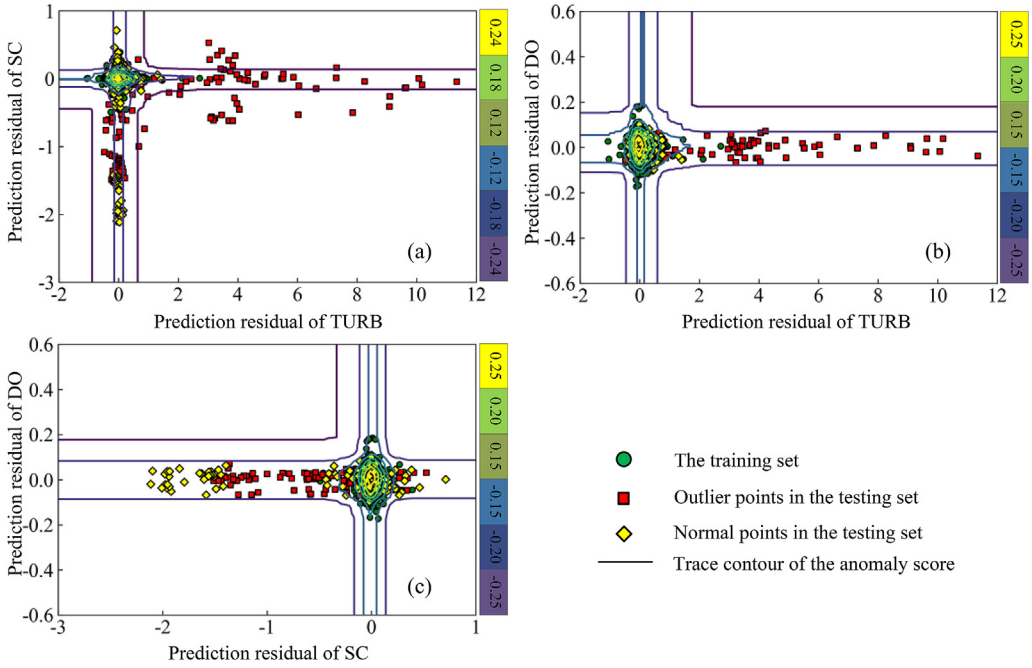


**Fig. 5.** The anomaly detection results using IF algorithm for case study.

to identify the efficiency of the anomaly detection method, the developed integration method is applied to the case study of Potomac River. BAR model is used for water quality variation prediction and IF algorithm is applied for surface water quality anomaly detection based on the prediction residuals obtained by BAR model. The water quality anomaly detection using two attributes is adopted by the developed integration method. In terms of anomaly scores ($s \in (-0.5,\ 0.5)$) obtained by IF algorithm, if a threshold setting $S$ is selected from the anomaly scores $s$, the anomaly scores of prediction residuals less than $S$ would be sorted into 1, while the anomaly scores of prediction residuals larger than $S$ would be sorted into 0. The ROC curves are plotted with TPR and FPR values shown in Fig. 6(a) by traversing all the threshold settings $s \in (-0.5,\ 0.5)$. It is shown that ROC curve obtains the maximum AUC value (0.919) for water quality anomaly detection using TURB and SC attributes. With the threshold setting $S$ of anomaly score setting at −0.0334, the maximum value of the Youden index $\Upsilon$ can be obtained with the TPR value 80% and the FPR value 9.7%. The AUC values are 0.797 and 0.805 for the water quality anomaly detection using TURB-DO attributes and SC-DO attributes. And the FPR values would reach up to 45.4% and 33.9% when TPR value is required at 80%. Therefore, the water quality anomaly detection using TURB and SC attributes can better identify and analyze water quality anomalies than other two attribute combinations.

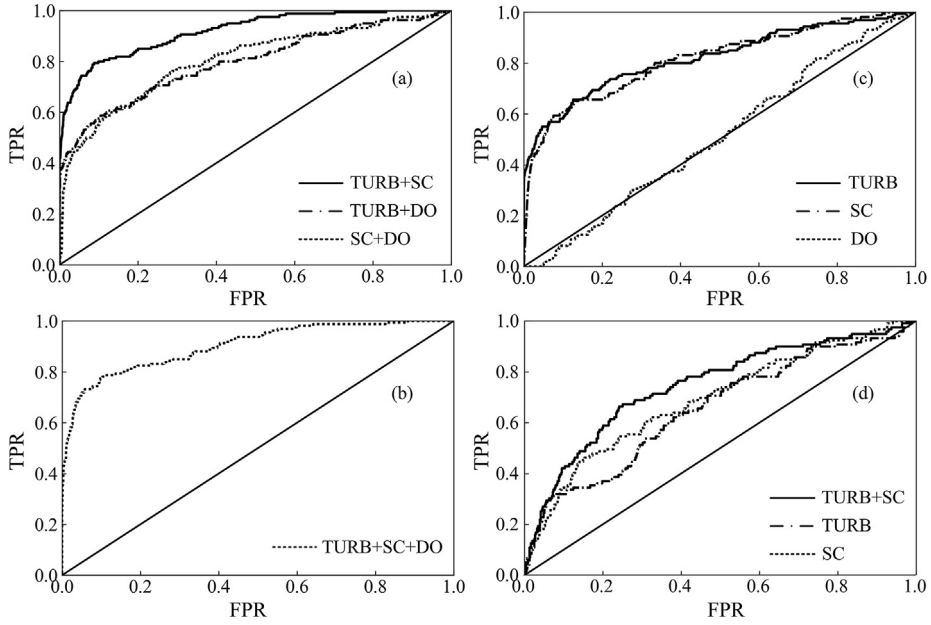


**Fig. 6.** The ROC curves for case study using IF algorithm (a and b) and BAR model (c), and for hypothetical scenario using IF algorithm and BAR model (d).

The water quality anomaly detection using three attributes together is also adopted by the developed integration method. For anomaly detection using three attributes together, the ROC curve is constructed by plotting the TPR values against the FPR values according to various threshold settings which are selected among anomaly scores obtained by IF algorithm shown in Fig. 6(b). And the AUC values can reach 0.902 using three attributes together by the developed integration method. It is shown that the water quality anomaly detection efficiency using three attributes together is slightly lower than the one using pair-wise TURB and SC attributes and is obviously better than the ones using TURB-DO attributes and SC-DO attributes. In fact, IF algorithm can efficiently deal with anomaly data in the massive dataset with multiple dimensions. However, it is not suitable for high-dimension attribute data due to the influences of noise attributes and irrelevant attributes. Therefore, for water quality anomaly detection and early warning, ROC curve can provide an effective tool to select possibly optimal water quality attribute combination/dimension and discard suboptimal ones independently by a cost/benefit analysis of anomaly detection and diagnostic decision making.

(2) Performance comparison with prediction-based and classification-based methods

A performance comparison for the water quality anomaly detection is proposed to assess the accuracy and efficiency among the developed integration method, prediction-based method (BAR model) and classification-based method (One-Class Support Vector Machine, OC-SVM) based on the case study of Potomac River. For prediction-based method, the water quality anomaly detection using a single attribute is adopted only by BAR model. In terms of the prediction residuals obtained by BAR model, if a threshold setting $S$ is selected from the absolute values of prediction residuals $pr$, the absolute values of prediction residuals larger than $S$ would be sorted into 1, while the absolute values of prediction residuals less than $S$ would be sorted into 0. The ROC curves are constructed by plotting the TPR values against the FPR values according to various threshold settings which are directly selected among the absolute values of prediction residuals obtained by BAR models shown in Fig. 6(c). The AUC values are 0.821, 0.816 and 0.509 using TURB, SC and DO attributes, respectively. It is shown that the water quality anomaly detection efficiency by the integration of BAR model and IF algorithm using pair-wise TURB and SC attributes or three attributes together is obviously better than the ones by prediction-based method, directly analyzing the prediction residuals of BAR model using a single attribute. Therefore, anomaly detection with appropriate multiple water quality parameters should be taken into account in order to ensure the accuracy and reliability of the anomaly detection results.

For classification-based method, the water quality anomaly detection using two attributes is also adopted by the integration of BAR model and OC-SVM method (Schölkopf et al., 2001). OC-SVM method, instead of IF algorithm, is applied for surface water quality anomaly detection based on the prediction residuals obtained by BAR model. In this study, Gaussian kernel is used to train the performance of OC-SVM. OC-SVM algorithm maps the input data into a high dimensional feature space using kernel functions, and finds the smallest hypersphere to make a clearer separation between normal and abnormal data (Xiao et al., 2016; Khreich et al., 2017). The training error $\nu \in (0,\ 1]$, as a predefined regularization parameter of OC-SVM method, controls the trade-off between the size of the hypersphere and the fraction of data points falling outside the hypersphere (abnormal data) (Erfani et al., 2016). In terms of training errors ($\nu \in (0,\ 1]$) predefined by OC-SVM method, if a threshold setting $S$ is selected from the training errors $\nu$, the decision function returns 1 in a small region (a normal region), and returns $-1$ elsewhere (an abnormal region). The ROC curves are plotted with TPR and FPR values shown in Fig. 7(b) by traversing all the threshold settings $\nu \in (0,\ 1]$. It is shown that ROC curve obtains the maximum AUC value (0.790) for water quality anomaly detection using TURB and SC attributes. With the threshold setting $S$ of training error setting at 0.01, the maximum value of the Youden index $\gamma$ can be obtained with the TPR value 57.8% and the FPR value 9.8%. The partial enlarged details of water quality anomaly detection result with an estimated training error of 0.01 are shown in Fig. 7(a). It is obvious that this boundary is negatively affected by the outliers, and unable to properly describe the "Ellipse". OC-SVM method allows some training samples to be located outside of its decision boundary in order to obtain a more general model and tolerate a certain fraction of outliers. According to the performance analysis and comparison among the developed integration method, prediction-based method and classification-based method, the developed method integrating BAR model with IF algorithm can exhibit excellent performance in accuracy and efficiency for water quality anomaly detection.

(3) Performance comparison with hypothetical scenario

In order to further identify the efficiency of the developed anomaly detection method, the developed integration method is applied to a hypothetical scenario with low-level anomaly condition. In this study, the historical observations of TURB and SC at 0:00 to 4:00 every day, which span from March 1, 2017 to March 7, 2017, are doubled and supposed to be 1 as anomaly label. The new time series of TURB and SC are shown in Fig. 8(a) and (b), and the abnormal data are distributed in the red regions. In the prediction

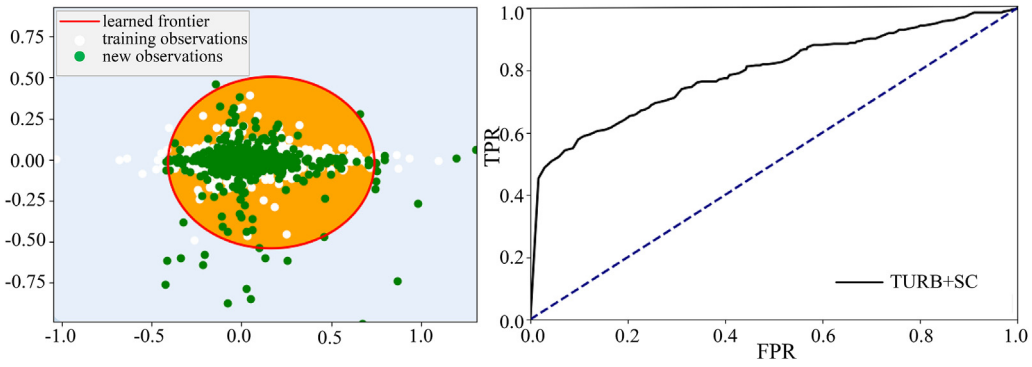


**Fig. 7.** The anomaly detection result and ROC curve using OC-SVM method for case study.

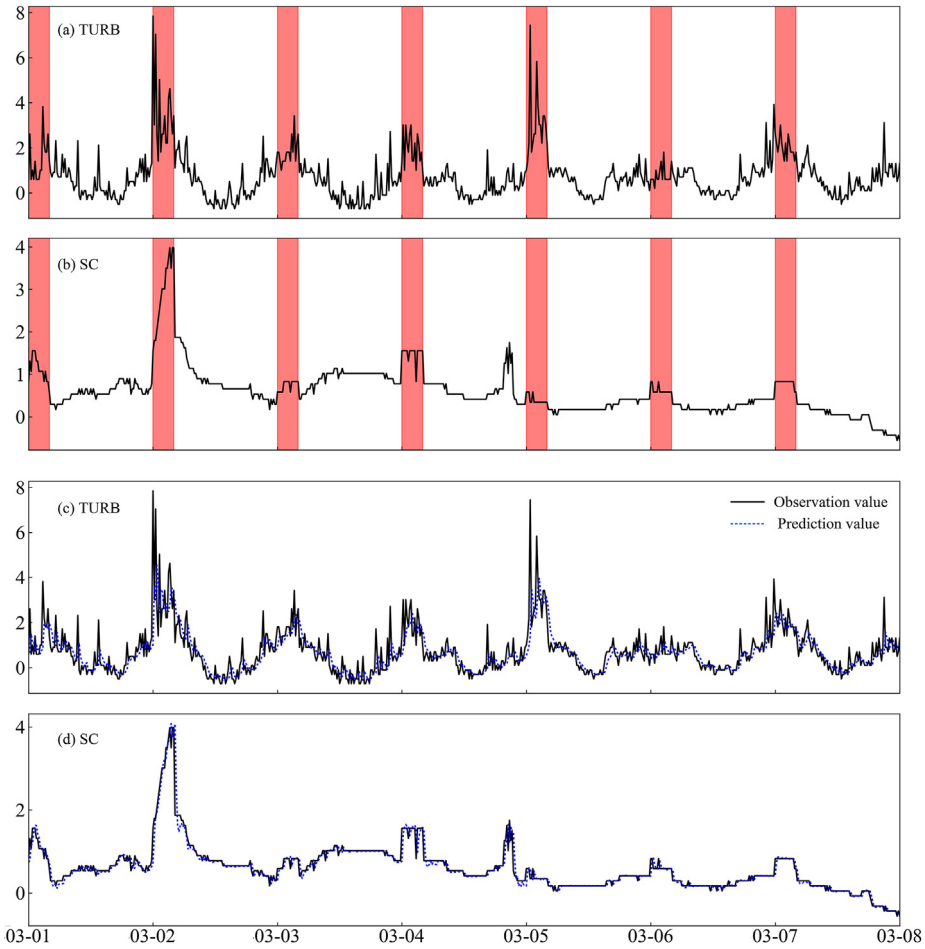


**Fig. 8.** The prediction results of the BAR models for hypothetical scenario.

phase, the historical observations as the training set, which also span from January 27, 2017 to February 15, 2017, are used to train the parameters of BAR model. And the optimal number orders of BAR models of TURB and SC are obtained as 7 and 12. Then, the developed BAR models of TURB and SC are used to dynamically forecast the current water quality parameters by constantly introducing and analyzing the observation data. In this low-level anomaly scenario, the water quality prediction results, which span from February 16, 2017 to March 7, 2017, can be obtained by the BAR models of TURB and SC. The prediction results of TURB and SC from March 1, 2017 to March 7, 2017 are shown in Fig. 8(c) and (d), and the related water quality prediction residuals can be obtained. In the anomaly detection phase, similarly, 100 iTrees are built by randomly selecting 256 sub-samples from the training set. The prediction residual values as the training set, which span from February 16, 2017 to February 28, 2017, are used to train iForest consisting of multiple iTrees and obtain the anomaly scores of the training set. Then, the prediction residual values as the testing set, which span from March 1, 2017 to March 7, 2017, are evaluated by calculating the anomaly scores for surface water quality anomaly detection. In the performance analysis phase, the ROC curves are plotted to demonstrate the performance tradeoff between FPR and TPR shown in Fig. 6(d). It is shown that the water quality anomaly detection efficiency by IF algorithm using pair-wise TURB and SC attributes (AUC = 0.745) is also obviously better than the ones by directly analyzing the prediction residuals of BAR model using a single TURB attribute (AUC = 0.691) and SC attribute (AUC = 0.654). The water quality anomaly detection efficiency by IF algorithm using pair-wise TURB and SC attributes in the low-level anomaly scenario is lower than the one in the case study of Potomac River above (AUC = 0.919). The main reason is that anomaly detection efficiency by IF algorithm is influenced by the distribution of the prediction residuals and the more obvious anomaly feature differences are, the easier the anomaly instances can be isolated.

## 4. Conclusions

In this study, an effective data-driven framework is developed by an integration of BAR model and IF algorithm for surface water quality anomaly detection and rapid early warning in response to river environmental pollution. The developed integration framework based on the combination of BAR model and IF algorithm is then applied to the case study of Potomac River in West Virginia comparing with prediction-based method (BAR model), classification-based method (OC-SVM method), and different scenarios. The results demonstrate that the developed integration method using TURB and SC attributes (AUC = 0.919) can better identify and analyze water quality anomalies than other two attribute combinations (TURB-DO and SC-DO) do (AUC = 0.797 and 0.805). And the TPR value of anomaly detection can reach at 80% and the FPR value is only 9.7%. Compared to prediction-based method (BAR model) and classification-based method (OC-SVM method), the IF algorithm using multiple attributes is more sensitive to surface water quality anomaly detection for river

environmental pollution. At meanwhile, the IF algorithm can also be applied well to anomaly detection at a low-level anomaly condition. Performance analysis of ROC curve for a hypothetical scenario with low-level anomaly condition yields an anomaly detection accuracy reaching at 0.745 using pair-wise TURB and SC attributes.

The results suggest that the developed integration framework for surface water quality anomaly detection is effective in providing rapid early warnings, and demonstrate that (a) the developed data-driven framework integrates BAR model and IF algorithm, and can prove to be effective in detecting water quality anomalies using multiple water quality attributes; (b) the developed data-driven framework can compensate the lack of anomaly detection method using single attribute, and obtain optimal anomaly detection results by ROC curve analysis and (c) the developed data-driven framework is applied to the case study of Potomac River in West Virginia in order to provide a rapid early warning, maximize emergency response capability, and prevent the negative influence from dispersing in large scale caused by river environmental pollution. DMs can start emergency monitoring according to surface water quality anomaly detection results, and carry out emergency response scheme in advance in order to deal with river environmental pollution.

## Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements.

This research was supported by the National Natural Science Foundation of China [Grant No.51779066], the China Postdoctoral Science Foundation [Grant No. 2018M631935] and Youth Doctoral Scientific Research Staring Foundation of Harbin University [Grant No. HUDF2017102]. The authors are extremely grateful to the editors and anonymous reviewers for their insightful comments and suggestions.

## References

Arad, J., Housh, M., Perelman, L., Ostfeld, A., 2013. A dynamic thresholds scheme for contaminant event detection in water distribution systems. Water Res. 47 (5), 1899–1908.
Azimi, S., Moghaddam, M.A., Monfared, S.A.H., 2018. Anomaly detection and reliability analysis of groundwater by crude Monte Carlo and importance sampling approaches. Water Resour. Manag. 32, 4447–4467.
Bantis, L.E., Nakas, C.T., Reiser, B., 2014. Construction of confidence regions in the roc space after the estimation of the optimal Youden index-based cut-off point. Biometrics 70 (1), 212–223.
Bertone, E., Sahin, O., Richards, R., Roiko, A., 2016. Extreme events, water quality and health: a participatory Bayesian risk assessment tool for managers of reservoirs. J. Clean. Prod. 135, 657–667.
Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J., 2000. LOF: identifying density-based local outliers. ACM SIGMOD International Conference on Management of Data 29 (2), 93–104.
Byrand, K., 2010. Nature and history in the potomac country: from hunter–gatherers tothe age of jefferson. J. Hist. Geogr. 36 (2), 233–234.
Calheiros, R.N., Ramamohanarao, K., Buyya, R., Leckie, C., Versteeg, S., 2017. On the effectiveness of isolation - based anomaly detection in cloud data centers. Concurrency Comput. Pract. Ex. 29 (18), e4169.
Chen, Y.L., Wu, W., 2018. Isolation forest as an alternative data-driven mineral prospectivity mapping method with a higher data-processing efficiency. Nat. Resour. Res. https://doi.org/10.1007/s11053-018-9375-6.
Deng, W.H., Wang, G.Y., 2017. A novel water quality data analysis framework based on time-series data mining. J. Environ. Manag. 196, 365–375.
Enders, W., Sandler, T., 1993. The effectiveness of antiterrorism policies: a vector-autoregression-intervention analysis. Am. Political Sci. Rev. 87 (4), 829–844.
Erfani, S.M., Rajasegarar, S., Karunasekera, S., Leckie, C., 2016. High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning. Pattern Recognit. 58, 121–134.
Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognit. Lett. 27 (8), 861–874.
He, Z.Y., Xu, X.F., Deng, S.C., 2003. Discovering cluster-based local outliers. Pattern Recognit. Lett. 24 (9), 1641–1650.
Hill, D.J., Minsker, B.S., 2010. Anomaly detection in streaming environmental sensor data: a data-driven modeling approach. Environ. Model. Softw 25, 1014–1022.
Hou, D.B., He, H.M., Huang, P.J., Zhang, G.X., Loaiciga, H., 2013. Detection of water-quality contamination events based on multi-sensor fusion using an extented Dempster-Shafer method. Meas. Sci. Technol. 24 (5), 055801.
Hou, D.B., Zhang, J., Yang, Z.L., Liu, S., Huang, P.J., Zhang, G.X., 2015. Distribution water quality anomaly detection from UV optical sensor monitoring data by integrating principal component analysis with chi-square distribution. Opt. Express 23 (13), 17487–17510.
Housh, M., Ohar, Z., 2017. Integrating physically based simulators with Event Detection Systems: multi-site detection approach. Water Res. 110, 180–191.
Housh, M., Ostfeld, A., 2015. An integrated logit model for contamination event detection in water distribution systems. Water Res. 75, 210–223.
Jeong, J., Park, E., Han, W.S., Kim, K.Y., 2017a. A subagging regression method for estimating the qualitative and quantitative state of groundwater. Hydrogeol. J. 25, 1491–1500.
Jeong, J., Park, E., Han, W.S., Kim, K.Y., Choung, S.W., Chung, I.M., 2017b. Identifying outliers of non-Gaussian groundwater state data based on ensemble estimation for long-term trends. J. Hydrol 548, 135–144.
Jin, T., Cai, S.B., Jiang, D.X., Liu, J., 2019. A data-driven model for real-time water quality prediction and early warning by an integration method. Environ. Sci. Pollut. Res. https://doi.org/10.1007/s11356-019-06049-2.
Khreich, W., Khosravifar, B., Hamou-Lhadj, A., Talhi, C., 2017. An anomaly detection system based on variable N-gram features and one-class SVM. Inf. Softw. Technol. 91, 186–197.
Knorr, E.M., Ng, R.T., Tucakov, V., 2000. Distance-based outliers: algorithms and applications. VLDB J 8 (3–4), 237–253.
Koch, M.W., Mckenna, S.A., 2011. Distributed sensor fusion in water quality event detection. J. Water Resour. Plan. Manag. 137 (1), 10–19.
Leigh, C., Alsibai, O., Hyndman, R.J., Kandanaarachchi, S., King, O.C., McGree, J.M., Neelamraju, C., Strauss, J., Talagala, P.D., Turner, R.D.R., Mengersen, K., Peterson, E.E., 2019. A framework for automated anomaly detection in high frequency water-quality data from in situ sensors. Sci. Total Environ. 664, 885–898.
Li, T.Y., Li, S.Y., Liang, C., Bush, R.T., Xiong, L.H., Jiang, Y.J., 2018. A comparative assessment of Australia's Lower Lakes water quality under extreme drought and post-drought conditions using multivariate statistical techniques. J. Clean. Prod. 190, 1–11.
Liu, F.T., Ting, K.M., Zhou, Z.H., 2008. Isolation forest. Eighth IEEE International Conference on Data Mining 413–422.
Liu, S.M., Smith, K., Che, H., 2015. A multivariate based event detection method and performance comparison with two baseline methods. Water Res. 80, 109–118.
Maso, E.D., Montecchio, L., 2014. Risk of natural spread of hymenoscyphus fraxineus with environmental niche modelling and ensemble forecasting technique. For. Res. 3 (4), 1000131.
Pang, J.Y., Liu, D.T., Peng, Y., Peng, X.Y., 2018. Optimize the coverage probability of prediction interval for anomaly detection of sensor-based monitoring series. Sensors 18 (4), 967.
Perelman, L., Arad, J., Housh, M., Ostfeld, A., 2012. Event detection in water distribution systems from multivariate water quality time series. Environ. Sci. Technol. 46 (15), 7927–8522.
Puggini, L., McLoone, S., 2018. An enhanced variable selection and Isolation Forest based methodology for anomaly detection with OES data. Eng. Appl. Artif. Intell. 67, 126–135.
Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., Williamson, R.C., 2001. Estimating the support of a high-dimensional distribution. Neural Comput. 13 (7), 1443–1471.
Shi, B., Wang, P., Jiang, J.P., Liu, R.T., 2018. Applying high-frequency surrogate measurements and a wavelet-ANN model to provide early warnings of rapid surface water quality anomalies. Sci. Total Environ. 610–611, 1390–1399.
Stripling, E., Baesens, B., Chizi, B., Broucke, S.V., 2018. Isolation-based conditional anomaly detection on mixed-attribute data tooncover workers' compensation fraud. Decis. Support Syst. 111, 13–26.
Wang, Y.G., Engel, B.A., Huang, P.P., Peng, H., Zhang, X., Cheng, M.L., Zhang, W.S., 2018. Accurately early warning to water quality pollutant risk by mobilemodel system with optimization technology. J. Environ. Manag. 208, 122–133.
Wit, E., Heuvel, E.V.D., Romeyn, J.W., 2012. 'All models are wrong...': an introduction to model uncertainty. Stat. Neerl. 66 (3), 217–236.
Xiao, Y.C., Wang, H.G., Xu, W.L., Zhou, J.W., 2016. Robust one-class SVM for fault detection. Chemometr. Intell. Lab. 151, 15–25.