Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2014, Article ID 879736, 14 pages
http://dx.doi.org/10.1155/2014/879736

Research Article

Time Series Outlier Detection Based on Sliding Window Prediction

Yufeng Yu, Yuelong Zhu, Shijin Li, and Dingsheng Wan

College of Computer & Information, Hohai University, Nanjing 210098, China

Correspondence should be addressed to Yufeng Yu; hhuheiyun@126.com

Received 18 July 2014; Accepted 15 September 2014; Published 30 October 2014

Academic Editor: Jun Jiang

In order to detect outliers in hydrological time series data for improving data quality and decision-making quality related to design, operation, and management of water resources, this research develops a time series outlier detection method for hydrologic data that can be used to identify data that deviate from historical patterns. The method first built a forecasting model on the history data and then used it to predict future values. Anomalies are assumed to take place if the observed values fall outside a given prediction confidence interval (*PCI*), which can be calculated by the predicted value and confidence coefficient. The use of *PCI* as threshold is mainly on the fact that it considers the uncertainty in the data series parameters in the forecasting model to address the suitable threshold selection problem. The method performs fast, incremental evaluation of data as it becomes available, scales to large quantities of data, and requires no preclassification of anomalies. Experiments with different hydrologic real-world time series showed that the proposed methods are fast and correctly identify abnormal data and can be used for hydrologic time series analysis.

1. Introduction

As the fundamental resources for water resources management and planning, long-term hydrological data are sets of discrete record values of hydrological elements that are collected with time and have been frequently analyzed in the field such as flood and drought control, water resources management, and water environment protection. With the development of data acquisition technology and data transmission technology, hydrological departments collected ever-increasing amounts of time series data from automatic monitoring systems via loggers and telemetry systems. Within these datasets, hydrologic time series analysis becomes workable and credible for building mathematical model to generate synthetic hydrologic records, to forecast hydrologic events, to detect trends and shifts in hydrologic records, and to fill in missing data and extend records [1]. However, a hydrologic time series is generally composed of a stochastic component superimposed on a deterministic component and usually shows stochastic, fuzzy, nonlinear, nonstationary, and multitemporal scale characteristics [2]. So, it is a challenging task to process and interpret the original hydrological time series due to the following:

(i) the large volumes of data,

(ii) the parameter pattern being specific and changing to different hydrology acquisition system due to multitemporal scale characteristic,

(iii) abnormal events or disturbances that create spurious effects in the data series and result in unexpected patterns,

(iv) inaccuracies in hydrological models due to imprecise and outdated information, logger and communications failures, poor calibration, and lack of system feedback.

Consequences of such situations in hydrological information systems may result in the DRQP (data rich, but quality poor) phenomenon. Consequently, the original monitoring data (i.e., precipitation, discharge, and water levels) should undergo a preprocessing step to eliminate the negative influence caused by incorrect or abnormal data due to

instrumentation faults, data inherent change, operation error, or other possible influencing factors [3]. Therefore, outlier detection usually becomes a vital step for hydrologic time series analysis based on the monitoring data. Regarding small monitoring datasets, data managers can detect and deal with outliers directly with a simple graphical or manual process. However, for massive datasets or data stream, an automatic and objective technique for effectively detecting and treating outliers is necessary.

This study develops a real-time outlier detection method that employs a window-based forecasting model for hydrologic time series collected from automatic monitoring systems. The method builds a forecasting model from a sequence of historical point values with a given window to predict future values. If the observed value differs from the predicted value beyond a certain threshold, an outlier would be indicated. The method uses prediction confidence interval (*PCI*) as threshold in consideration of uncertainty in the data series parameters in the forecasting model. Data are classified as anomalous/nonanomalous based on whether or not they fall outside a given *PCI*. Thus, the method provides a principled framework for selecting a threshold. This method does not require any preclassified examples of data, scales well to large volumes of data, and allows for fast incremental evaluation of data as it becomes available.

In order to evaluate the proposed method, it was applied to two different hydrological variables, water level and daily flow, from *Huayuankou* (hereafter *HYK*, 34.76°N, 113.58°E) and *Lanzhou* (hereafter *LZ*, 36.04°N, 103.49°E) stations obtained from national hydrology database of MWR, China. The results show that the proposed method can exactly detect the outliers in the hydrological time series with near negligible false positive rate. Furthermore, the algorithm's efficiency is analyzed based on the detection results.

The rest of the paper is organized as follows. In the next section (Section 2) we present the related work to this area of research. In Section 3 we present details about the proposed algorithm for outlier detection in time series based on prediction confidence interval. A number of experiments with the proposed method using real-world hydrological time series are reported in Section 4. Finally, Section 5 gives conclusions and suggestions for further research.

2. Related Work

2.1. Time Series Analysis. A time series (TS) $X = \{x(t) \mid 1 \leqslant t \leqslant m\}$ is a sequence of d-dimensional observations vector $x(t) = (x_1(t), x_2(t), \ldots x_d(t))$ ordered in time. Mostly these observations are collected at equally spaced, discrete time intervals. It is called a univariate (or single) time series when d is equal to 1 and a multivariate time series when d is equal to or greater than 2 [4]. Generally, a time series can be regarded as a sample realization from an infinite population of such time series generated by a stochastic process, which can be stationary or nonstationary. In addition, the understanding of the structure and dependence of time series is achieved through time series analysis.

Time series analysis is the investigation of a temporally distributed sequence of data or the synthesis of a model for prediction wherein time is an independent variable; as a consequence, the information obtained from time series analysis can be applied to forecasting, process control, outlier detection, and other applications [5]. A basic assumption in time series analysis is that some aspects of the past pattern will continue to remain in the future. Also under this setup, often the time series process is assumed to be based on past values of the main variable but not on explanatory variables which may affect the variable/system.

In hydrology, time series analysis is one of frontier scientific issues because it can detect and describe quantitatively each of the hydrologic processes underlying a given sequence of observations. Moreover, hydrologic time series analysis can also be used for building mathematical models to generate synthetic hydrologic records, to forecast hydrologic events, to detect trends and shifts in hydrologic records, and to fill in missing data and extend records. Consequently, time series analysis has become a vital tool in hydrological sciences and its importance has been dramatically enhanced in the recent past due to ever-increasing interest in the scientific understanding of climate change [6].

In the time series analysis, it is assumed that the data (observations) consist of a systematic pattern and stochastic component; the former is deterministic in nature, whereas the latter accounts for the random error and usually makes the pattern difficult to be identified. Previous research usually equates stochastic component to system error and then simply discards it so as to not complicate the statistical analyses. However, the stochastic component potentially includes interesting and meaningful information; it must be treated with caution. It is for this reason that outlier detection becomes a hotspot research issue in recent years.

2.2. Outlier Detection. An outlier can be defined as "observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism" [7], or "patterns in data that do not conform to a well-defined notion of normal behavior" [8]. Generally, an outlier may be incorrect and/or the circumstances around the measurement may have changed over time. Thus, the identification and treatment of outliers constitute an important component of the time series analysis before modeling because outliers can have negative impacts on the selection of the appropriate model as well as on the estimation of the associated parameters.

Outlier detection, also known as anomaly detection in some literatures, is an important long-standing research problem in the domains of data mining and statistics. The major objective of outlier detection is to identify data objects that are markedly different from, or inconsistent with, the remaining set of data [9, 10]. In recent decades, this research problem is attracting significant research attentions in the fields of statistical analysis, machine learning, and artificial intelligence due to its important applications in a wide range of areas in business, security, insurance, health care, and engineering, to name a few. Chandola et al. [8] provide a comprehensive classification of the outlier detection techniques

and classify the techniques into six categories: classification-based [11], nearest-neighbor-based [12], clustering-based [13], statistical [14], information theory-based [15], and spectral theory-based [16]. In addition to the abovementioned algorithms, other approaches have been adopted to solve the outlier detection problem. For instance, signal processing techniques such as wavelet transform [17] and Fourier transform [18] have been used to detect outlier regions in meteorological data.

Outlier detection is a very broad field and has been studied in the context of a large number of application domains where many detection methods have been proposed according to the different data characteristics. Recently, there has been significant interest in detecting outliers in time series. Generally, methods for time series outlier detection should consider the sequence nature of data and operate either on a single time series or on a time series database. The goal of outlier detection on a single time series is to find an anomalous subregion, while the goal of the latter is to identify a few sequences as outliers or to identify a subsequence in a test sequence as an outlier. In some cases, a single time series is converted to a time series database through the use of a sliding window [19].

Given a single time series, one can find particular elements (or time points) within the time series as outliers or find subsequence outliers. Fox defines two types of outliers (type I/additive and type II/innovative) based on the data associated with an individual object across time, ignoring the community aspect completely [20]. Since then, a large number of t prediction models and profile based models have been proposed to find point outlier within a time series. A straightforward method for outlier detection in time series is based on forecasting [21–23]. This approach first builds a prediction model from the historical values and is then used to predict future values. If a predicted value differs from the observed value beyond a certain threshold, an outlier would be indicated. The definition of the value of the threshold to be used for detecting outliers is the main problem of outlier detection based on prediction. The difficulties of forecasting-based outlier detection have motivated the proposal of anomaly subsequences detection techniques, based on outlier score calculated with respect to similar measure between subsequences [24–27]. Keogh et al. [24] proposed using one nearest-neighbor approach to detect maximal different subsequences within a longer sequence (called discords). Subsequence comparisons can be smartly ordered for effective pruning using various methods like heuristic reordering of candidate subsequences [25], locality sensitive hashing [26], Haar wavelet [27], and SAX with augmented tries [28].

The problem of outlier detection in time series database mainly focuses on how to find all anomalous time series. It is assumed that most of the time series in the database are normal while a few are anomalous. Similar to the traditional outlier detection, the usual recipe of solving such problems is to first learn a model based on all the time series sequences in the database and then compute an outlier score for each sequence with respect to the model [29]. Statistics-based methods are still conducted to detect outliers in time series database because of their efficient structure. In 2012, Zhang developed an average-based methodology based on time series analysis and geostatistics, which achieved satisfactory detection results in short snapshots [30]. However, it will fail if the outliers gather together closely in the same short time slot. Hence, recent researches have mainly focused on nonparametric outlier detection methods such as Bayesian method and discrete wavelet transform (DWT). Frieda proposed a Bayesian approach to outlier modeling, approximating the posterior distribution of the model parameters by application of a componentwise Metropolis-Hastings algorithm [31]. Apart from the Bayesian method, Grané and Veiga [32] identified the outliers as those observations in the original series whose detail coefficients are greater than a certain threshold based on wavelet transform. They iterated the process of DWT and outlier correction until all detail coefficients are lower than the threshold. Based on real-world financial time series, their method achieved a lower average number of false outliers than Bilen and Huzurbazar's [33]. However, in these works, thresholds are mainly set subjectively, which makes these methods inefficient and insensitive when there are several different kinds of outliers appearing in the same time series.

2.3. Outlier Detection in Hydrological Time Series. Hydrologic systems involving outliers invariably represent complex dynamical systems. The current state and future evolutions of such dynamical systems depend on countless properties and interactions involving numerous highly variable physical elements. The representation of such dynamical systems in their corresponding models is complicated because certain relationships can only be developed through analyses.

Outlier detection in hydrologic data is a common problem which has received considerable attention in the univariate framework. In the multivariate setting, the problem is well established in statistics. However, in the hydrologic field, the concepts are much less established. A pioneering work in this direction was recently presented by Chebana and Ouarda [34]. Moreover, many outlier detection techniques, such as Chauvenet's method, Dixon-Thompson outlier test, and Rosner's test [35], are statistics based on the principle of hypothesis testing with the underlying assumption of log-Pearson type III (LP3) distribution [36], which may not always be readily available for hydrological time series. Hyndman and Shang's [37] methods, on the basis of real data, are graphical and consist first in visualizing functional data through the rainbow plot and then in identifying functional hydrological outliers using the functional bag-plot and the functional highest-density region box-plot. The methods can detect outlier curves that may lie outside the range of the majority of the data or may be within the range of the data but have a very different shape. However, as indicated by Chebana and Ouarda [34], the points outside the fence of the bag-plot or box-plot are considered as extremes rather than outliers. On this basis, Chebana et al. proposed a nongraphical outlier detection method based on the bivariate score points, which were obtained from the first two principal components score vectors generated by functional principal component analysis, to detect outliers in flood frequency analysis [38].

Ng et al. [39] found that chaotic approach can determine the level of complexity of a system after they applied the chaotic analytical techniques to daily hydrologic series comprising of outlier. It adopts box-plot [40] as the outlier detection tools based on its (1) statistical popularity, (2) ability to handle and detect multiple outliers simultaneously without preceding to an iterative detection, and (3) ability to detect outliers by block detection where no masking effects are involved before conducting the chaotic analysis. The method can effectively detect numerous outliers in the original daily discharge dataset but may understate the actual number of outliers because the data is usually clustered around the zero mean and consequently results in less skewness in the data.

Although many outlier detection methods exist in the literature, there is a lack of discussion on the selection of a proper detection method for hydrological outliers. It is mainly because of the fact that most of outlier detection methods belong to statistical approaches and demanded that the data must follow some distributions, and the selection of a suitable outlier detection method is critically determined by the intent of analyst and the intended use of the results. Analysts have to consider several technical aspects in its decisions-making such as the tradeoff between accurate and efficient, the evaluation of consequences subject (i.e., masking and swamping), the design assumptions and the limitation of different methods, and the preference on parametric or nonparametric approach. Without a thorough understanding of outlier phenomena, it is difficult to determine a suitable outlier detection method.

Faced with such challenges, this work proposes a new method to detect outliers that splits given historical hydrological time series into subsequences by a sliding-window and then an autoregressive (*AR*) prediction model of time series and prediction confidence interval (*PCI*) calculated from nearest-neighbor historical data to identify time series anomalies. The method used an autoregressive prediction model, which belongs to data-driven time series model essentially, rather than a physics-based time series model, due to the fact that it is simpler to develop and can rapidly produce accurate short forecast horizon predictions. Data are classified as anomalous/nonanomalous based on whether or not they fall outside a given *PCI*. The *PCI* which is employed in this study not only accounts for the fact that the correlations between adjacent data points in the time series are higher than those farther away but also avoids falling into inefficient and insensitive dilemma caused by a subjective-setting threshold. Moreover, the *PCI* can be calculated dynamically according to different nearest-neighbor windows size and confidence coefficient of difference users, which make it suitable for different variables of hydrologic time series outlier detection for different user's demand. In conclusion, this method does not require any preclassified examples of data, scales well to large volumes of data, and allows for fast incremental evaluation of data as it becomes available.

3. Window-Based Outlier Detection

In this section, we formulate the outlier detection problem and give a formal definition of some concepts which are used in the proposed algorithm. And then we will introduce the algorithm detecting outliers in time series based on the sliding-window prediction model. In addition, we also mention efficient strategies to choose the optimal parameters to meet the users' requirements. Next is the formal formulation of the contextual outlier detection algorithm.

3.1. Problem Definition. Hydrology is a time-varying phenomenon, the change of which is referred to hydrological processes. As important scientific data resources, hydrological data are the discrete records of hydrological processes and could be divided into flow, water level, rainfall, evaporation, and other hydrologic time series according to the physical quantities of its representation.

Definition 1 (hydrological time series). A hydrological time series T is a set of real-valued data in successive order, occurring uniform time interval. In this work, $T = \langle d_1 = (v_1, t_1),\ d_2 = (v_2, t_2), \ldots d_m = (v_m, t_m)\rangle$, where m is the length of the time series and point $d_i = (v_i, t_i)$ stands for the observation v_i at the moment t_i; moreover, t_i is strictly increased.

For outlier detection purposes, we are typically not interested in any of the global properties of a time series; rather, we are interested in local subsections of the time series, which are called subsequences.

Definition 2 (subsequence). Given a time series T of length m, a subsequence C of T is a sampling of length $n \leq m$ of contiguous position from p; that is, $C = \langle d_p = (v_p, t_p), d_{p+1} = (v_{p+1}, t_{p+1}), \ldots d_{p+n-1} = (v_{p+n-1}, t_{p+n-1})\rangle$ for $1 \leq p \leq m - n + 1$. Particularly, subsequence C may contain only one data point on the condition that $n = m$.

Since all subsequences may potentially be abnormal, any algorithm will eventually have to extract all of them; this can be achieved by use of a sliding window.

Definition 3 (sliding window). Given a time series T of length m and a user-defined subsequence length of n, all possible subsequences can be extracted by sliding window of size n across T and considering each subsequence C_p.

Generally, the first problem of time series outlier detection is to define what kind of data in a given dataset is abnormal. That is, the definition of outlier determines the outlier detections' goals. In hydrologic time series, time sequences which are composed of different physical quantities show great difference anomaly characteristics; therefore, it is difficult to give a uniform definition of abnormality. In this paper, we identify a subsequence to be outlier based on its nearest-neighbor.

Definition 4 (k-nearest-neighbor). Given a time series T of length m and a data point d_i ($i < m$), the k-nearest-neighbor $\eta_i^{(k)}$ of d_i is a sampling of length $2k$ of contiguous position

from $i - k$ to $i + k$ and does not contain i; that is, $\eta_i^{(k)} = \{d_{i-k}, \ldots d_{i-1}, d_{i+1}, \ldots d_{i+k}\}$ for $i + 1 \leq k \leq m - i$.

Definition 5 (time series outlier). Given a time series T of length m and the k-nearest-neighbor $\eta_i^{(k)}$ of d_i, d_i will be identified as an outlier if the observed value of d_i falls outside a PCI calculated by confidence coefficient p and predicted value according to $\eta_i^{(k)}$.

From the above definition, one can see that the nearest-neighbors window size k and confidence coefficient p become key parameters of outlier detection. Therefore, it can dynamically adjust k and p for different users and different hydrological elements to achieve the optimal detection result.

3.2. Algorithm Description. After studying the current situation and challenge of hydrological time series and its outliers, this study proposes a new outlier detection method that uses a sliding window of hydrological time series $T^m = \langle d_1 = (v_1, t_1), d_2 = (v_2, t_2), \ldots d_m = (v_m, t_m)\rangle$, where point $d_i = (v_i, t_i)$ stands for the measurement v_i at the moment t_i, to classify a particular data point as anomalous or not. A data point d_i is classified as anomalous if its measurement deviates significantly from its k-nearest-neighbor (KNN) prediction value calculated using its neighboring point set $\eta_i^{(k)}$ as input. Upon initialization, the method fills the window with the most recent measurements and commences classification with the next measurement taken by the time series.

In brief, the method consists of the following steps beginning at time i within a given hydrological time series T^i.

Step 1. Define k-nearest-neighbor window $\eta_i^{(k)}$ for the point d_i, based on the data source, from where outliers are to be detected (history data including complete sequence or only past data to forecast new incomings); the $\eta_i^{(k)}$ can be divided into one-sided-windows and two-sided-windows types.

Step 2. Build a nearest-neighbor-window prediction model that takes $\eta_i^{(k)}$ as input to predict $\overline{v}_{i+1}$, the expected value of the point d_{i+1}; in addition, calculate the upper and lower bounds of the range within which the measurement should lie (i.e., the prediction confidence interval, PCI).

Step 3. Compare the actual measurement at time $i + 1$ with the range calculated in Step 2 and classify it as anomalous if it falls outside the range; otherwise, classify it as nonanomalous.

Step 4. Modify T^i by removing d_{i-k+1} from the back of the window and adding d_{i+1} which holds the value $\overline{v}_{i+1}$ to the front of the window to create T^{i+1}; slide one step and modify the $\eta_i^{(k)}$ to create $\eta_{i+1}^{(k)}$ if the measurement is classified as anomalous; else, modify T^i by removing d_{i-k+1} from the back of the window and adding d_{i+1} which takes its original value to the front of the window to create T^{i+1}; slide one step and modify the $\eta_i^{(k)}$ to create $\eta_{i+1}^{(k)}$.

Step 5. Repeat Steps 1–4, until all sequence has been detected.

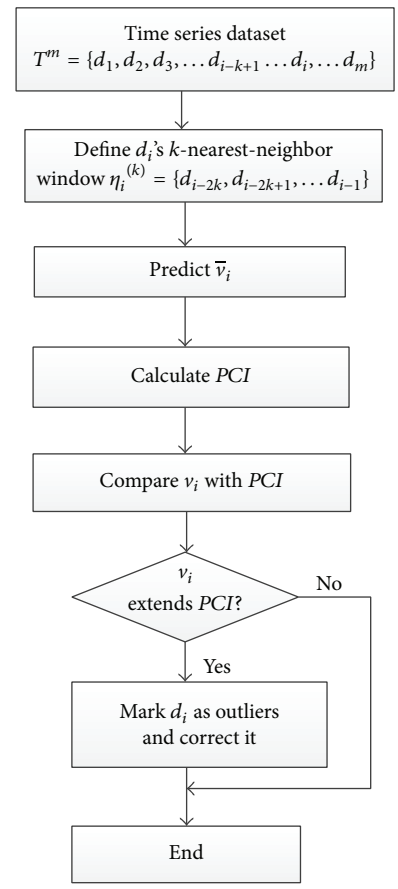


Figure 1: Diagram of the proposed outlier detection method.

The detection process of data point d_i is illustrated with a flow chart in Figure 1. The remainder of this section describes these steps in detail.

3.2.1. Window Definition. The first step of this outlier detection process, the KNN window of the test point d_i in time series data, is defined to illustrate the relations between the data point and its nearest-neighbor. And then, the prediction model can use only the test point's KNN window to predict the measurement of d_i for the purpose of simplifying the computational complexity.

Generally, there are two types of hydrological data from where outliers are to be detected: history time series data or real-time data. The difference between them is primarily based on the fact that the former uses the previous and subsequent neighbor window as input parameters to detect the outlier while the latter only uses the previous neighbor window as input parameters. Then, the neighbor window can be divided into one-sided and two-sided types.

(1) Two-Sided Neighbor Windows. The two-sided-windows outlier detection method uses a data points' previous (left) and subsequent (right) neighbor's data point to determine whether this data point is outlier or not. Given a water

level time series $T^m = \langle d_1 = (v_1, t_1), d_2 = (v_2, t_2), \ldots d_m = (v_m, t_m)\rangle$, the two-sided-windows neighboring point sets $\eta_i^{(k)}$ for the point d_i can be defined as follows:

$$\eta_i^{(k)} = \{d_{i-k}, \ldots d_{i-1}, d_{i+1}, \ldots d_{i+k}\}. \quad (1)$$

Note that $2k$ is the size of the neighborhood window, starting at $i-k$ and ending at $i+k$ (not including i).

(2) One-Sided Neighbor Windows. As the right neighbors may contain outliers that had not been detected, therefore, in the real-time detection application, only left neighbor can be available. So, one-sided-windows outlier detection method only chooses left neighbors which removed (or revised) the outliers had been detected will make the detection results more meaningful. Other than the two-sided-windows method, one-sided-windows method outlier detection algorithms define neighborhood point set $\eta_i^{(k)}$ for the point d_i, as follows:

$$\eta_i^{(k)} = \{d_{i-2k}, d_{i-2k+1}, \ldots d_{i-1}\}. \quad (2)$$

Here, $2k$ is the size of the neighborhood window for the point d_i, starting at $i-2k$ and ending at $i-1$.

3.2.2. Model Selection. In this step, an autoregressive (AR) prediction model is built to forecast the particular point's value using its neighborhood point sets as input. The AR model forecasts future measurements in time series datasets using only a specified set of observations, that is, $\eta_i^{(k)}$, from the same discharge site; they are used because they avoid complications caused by different sampling frequencies that can arise if a heterogeneous set of time series data was used; moreover, the use of AR model reduces the number of predictions that cannot be made due to insufficient data caused by the embedded telemetry equipment that went offline.

The neighborhood point sets $\eta_i^{(k)}$ are used as input to the AR model of the time series to predict the following observation:

$$\overline{d}_i = M\left(\eta_i^{(k)}\right), \quad (3)$$

where $M(\,)$ is the model. This method assumes that the behavior of the processes at time $t+1$ can be described by a finite set of k previous measurements; thus it implicitly assumes that the time series is an order k Markov process. Literature [22] compares native, nearest cluster (NC), and single-layer linear network (LN) and multilayer perceptron (MLP) on different dataset and concludes that LN and MLP would obtain better detection result than other models. Based on this work, we use LN as prediction model and assume that observation v_i is a linear combination of two-sided neighbors windows data point:

$$\overline{v}_i = \frac{\left(\sum_{j=1}^{k} w_{i-j} v_{i-j} + \sum_{j=1}^{k} w_{i+j} v_{i+j}\right)}{\left(\sum_{j=1}^{k} w_{i-j} + \sum_{j=1}^{k} w_{i+j}\right)}, \quad (4)$$

where $\langle w_{i-k}, \ldots, w_{i-1}, w_{i+1}, w_{i+k}\rangle$ stands for the weight of neighborhood, which defines the relationship between the two-sided-windows neighborhood $\{d_{i-k}, \ldots d_{i-1}, d_{i+1}, \ldots d_{i+k}\}$ and the expected value of d_i. Generally, the weight of the neighboring point d_j is inversely proportional to the distance between the point d_i and d_j; that is, the larger the distance is, the small the weight w is. For simplicity, it assigns the weight vector $\langle w_{i-k}, \ldots, w_{i-1}, w_{i+1}, w_{i+k}\rangle$ with the values $\langle 1, 2, \ldots k, k, \ldots 2, 1\rangle$.

Generally, two-sided neighbor windows need points' previous and subsequent neighbors; however, the right neighbors may contain outliers that had not been detected, which may affect detection results subsequently. So, in some application fields, only previous (left) neighbor-window data can be used to predict and identify forthcoming outliers. Therefore, it often uses a simple modification of the two-sided-windows model to predict the measurement at time i:

$$\overline{v}_i = \frac{\sum_{j=1}^{2k} w_{i-j} v_{i-j}}{\sum_{j=1}^{2k} w_{i-j}}. \quad (5)$$

Similar to two-sided neighbors windows, the weight vector $\langle w_{i-2k}, w_{i-2k+1}, \ldots w_{i-1}\rangle$ stands for the weight of neighborhood and is assigned with the values $\langle 1, 2, \ldots 2k\rangle$.

3.2.3. Outliers Identification. Given the model prediction from Section 3.2.2 based on test point's neighbor, the data point can then be classified as anomalous using a confidence boundary PCI calculated via predicted value and confidence coefficient. The PCI gives the range of plausible values that the test measurement can take; the confidence coefficient ($p = 100(1-\alpha)$) indicates the expected frequency with which measurements will actually fall in this range. If it is assumed that the model residuals have a zero-mean Gaussian distribution, the $p\%$ PCI can be calculated as follows [22]:

$$PCI = \overline{v}_{t+1} \pm t_{\alpha/2,2k-1} \times s\sqrt{1 + \frac{1}{2k}}, \quad (6)$$

where $\overline{v}_{t+1}$ is the prediction value of the test point, $t_{\alpha/2,n-1}$ is the pth percentile of a Student's t-distribution with $2k-1$ degrees of freedom, s is the standard deviation of the model residual, and k is the window size used to calculate s. This type of PCI is a type of t-interval because it relies on Student's t-distribution. If the test point's actual measurement falls within the bounds of the PCI, then the point is classified as nonanomalous; otherwise, it is classified as anomalous. Thus, the PCI represents a threshold for acceptance or rejection of a data point. The benefit of using the PCI instead of an arbitrary threshold is that the prediction level guides the selection of the interval width.

The two-sided-windows approach for outlier detecting is illustrated with a simple example in Figure 2 with a neighborhood window width of $k = 4$. It should be noted that the area between the two green lines in Figure 2 covering the neighborhood of data points indicates the confidences bound (PCI). In this example, d_7 is an outlier in the current neighborhood of points and proposes to be replaced by the

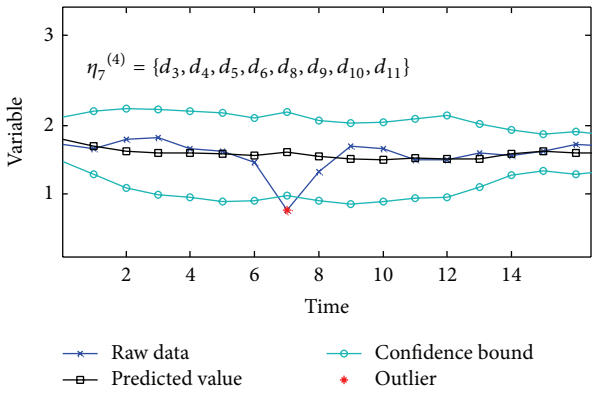


FIGURE 2: Outlier detection example, where the raw data, predicted value, confidence bound, and outlier are identified by different symbols. In this example, d_7 is an outlier in the current neighborhood of points and has been detected correctly.

predicted value $\bar{v}_7$ for analysis and modeling of the time series.

3.3. Parameter Optimization. In order to detect the outliers in the time series, the proposed methods should calculate the plausible values range *PCI* via predicted value and confidence coefficient based on the test point's k-nearest-neighbor. Hence, the values of the two parameters, k and p, become the key issues for improving the outlier detection methods mentioned in the previous section. In this case, we can use the following experiments to choose a proper value of those two parameters.

(1) The window width (k) of the methods controls how many neighboring points are included in the calculation of the prediction value. The larger the window width is, the more points are included to compute the prediction value. It varies k from 3 to 15 in increments of one; that is, $k = \{3, 4, \ldots, 15\}$.

(2) The confidence coefficient p calculates the plausible values range *PCI* to classify a data point as an outlier. The larger the confidence coefficient is, the more plausible the prediction value is. Therefore, it varies p from 85% to 99% in increments of one percent.

To tune the best combination of algorithm parameters that maximizes the ratio of detection, the cross-validation scheme is applied. The complete sample is split into two segments: a training dataset and a testing dataset. This is a way of cross-validating whether the parameters found during the first period, the training phase, are consistent and still valid in a different period, the testing phase. The principle of cross-validation is a generic resource to validate statistical procedures and it has been applied in different contexts. Furthermore, in training phase, the training set is divided into 10 nonintersecting subsets of equal size, chosen by random sampling. The model is then trained 10 times, each time reserving one of the subsets as a validation set on which the model error is evaluated while fitting the model parameters using the remaining nine subsets. The model parameters with the lowest mean squared error among the 10 training models are then selected for the final model [41].

4. Experiments and Analysis

To demonstrate the efficacy of the outlier detection methods developed in this study for data QA/QC, it will be applied to hydrological data series from national hydrology database of MWR, China. In the following, the real data are described and are functional and results are presented and discussed. More precisely, it first uses the previously presented approaches to identify outliers; then, some performance evaluation will be discussed and interpreted on the basis of hydrological data; and some results using multivariate approaches for comparison purposes will be provided at last.

4.1. Study Area and Data. In this subsection we report on a number of experiments using two different hydrologic elements, water level (m) and daily flow ($m^3\ s^{-1}$), from *LZ* and *HYK* stations with reference numbers 40101200 and 40105150, respectively, obtained from national hydrology database of MWR, China. The *LZ* and *HYK* stations are the important flood prevention and control stations and are generally known as typical hydrological stations of upper and middle reaches of the Yellow River. They play an important role in downstream of Yellow River in the fields such as hydrological data collection and forecasting, flood control scheduling, river control experiments, and water resources development. Figure 3 indicates the geographical location of *LZ* and *HYK* stations.

The data downloaded from the national hydrology database of MWR, China, were available as raw data. In addition, we followed the procedures described in the previous section. Figures 4 and 5 illustrate the whole dataset within a given time series; it shows that the dataset is nearly periodic and has some suspicious data point obviously.

4.2. Experiment Result. Since the data used in this study were subjected to manual quality control before being archived to the national hydrology database, it was expected that the detectors would not identify many data outliers in the archive. However, we can easily see that some data points deviate from their neighbor. And then, we apply our methods to detect the outliers in the given hydrological time series with the window size $k = 6$ and probability $p = 95\%$. And the detection results for the proposed methods are shown in Figures 6, 7, 8, and 9.

Figures 6–9 depict real and predicted values with the proposed outlier detected methods in the daily flow and water level time series of *LZ* and *HYK* stations, respectively. In each of these graphics the *PCI* for the predictions is also depicted. These graphics show that most of the real values are very near the respective predicted value, while a spot of them lies outside the *PCI* boundaries for the predictions. An outlier detection method based on *PCI* would indicate an outlier if the real value lies outside the confidence bounds, as described in Section 3.2. The experiments reported here showed that these bounds, built from cross-validation scheme on training and validation sets, can correctly bind the region of normality. Furthermore, these intervals could be used to indicate the level of suspicion associated with a given point in the future.

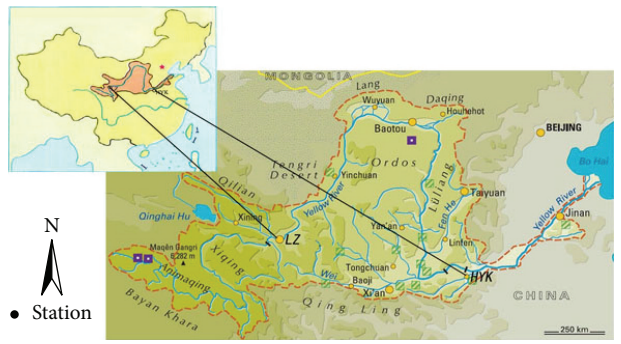
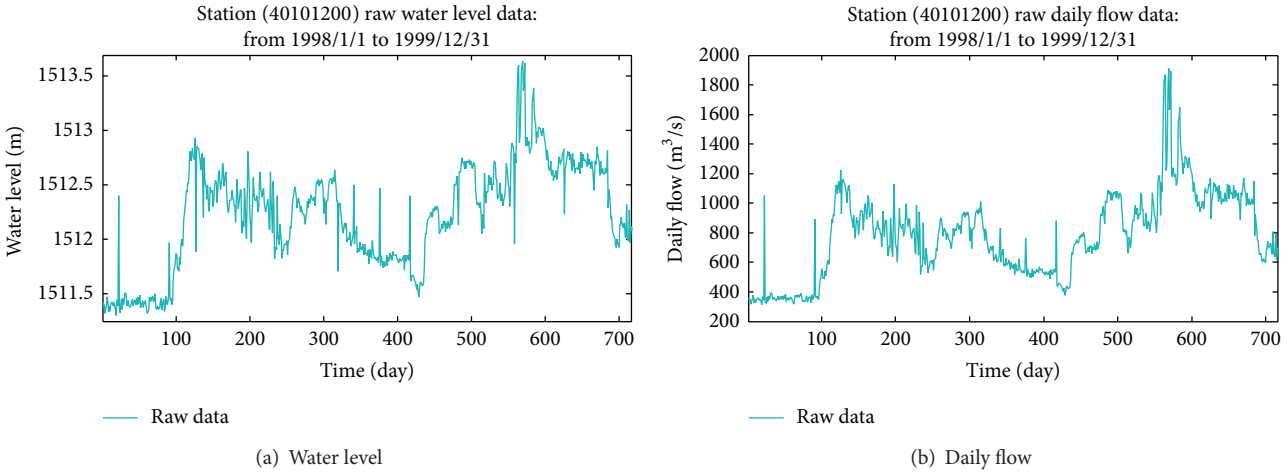


Figure 3: Geographical location of *LZ* and *HYK* stations in Yellow River.



(a) Water level

(b) Daily flow

Figure 4: Raw hydrological time series of *LZ* station.

4.3. Experiment Evaluation

4.3.1. Parameters of Quality. The detection results for the proposed methods with the two different hydrologic elements from *LZ* and *HYK* stations illustrate that the outlier detection methods are promising and can successfully detect a considerable amount of outliers from the hydrologic dataset. However, we also note that there are some valid data points which had been detected as outliers and imputed in error.

The objective of the evaluation analysis described in this section is to assess the effectiveness of method; we can classify the results from our experiment into four categories (see Table 1).

The categories in Table 1 correspond to the four possible outcomes of one experimental run, which consists of using both methods with a particular combination of window width (k) and probability (p). Categories A and D are ideal situations in which a point can be detected correctly, while categories B and C are undesired, because the methods are not able to distinguish between outliers and exactness.

According to these definitions, the *Sensitivity* is the probability that the proposed methods discovered a real outlier. Its formula is defined as follows:

$$\text{Sensitivity} = \frac{TP}{(TP + FN)}. \tag{7}$$

Table 1: Assessment of both methods.

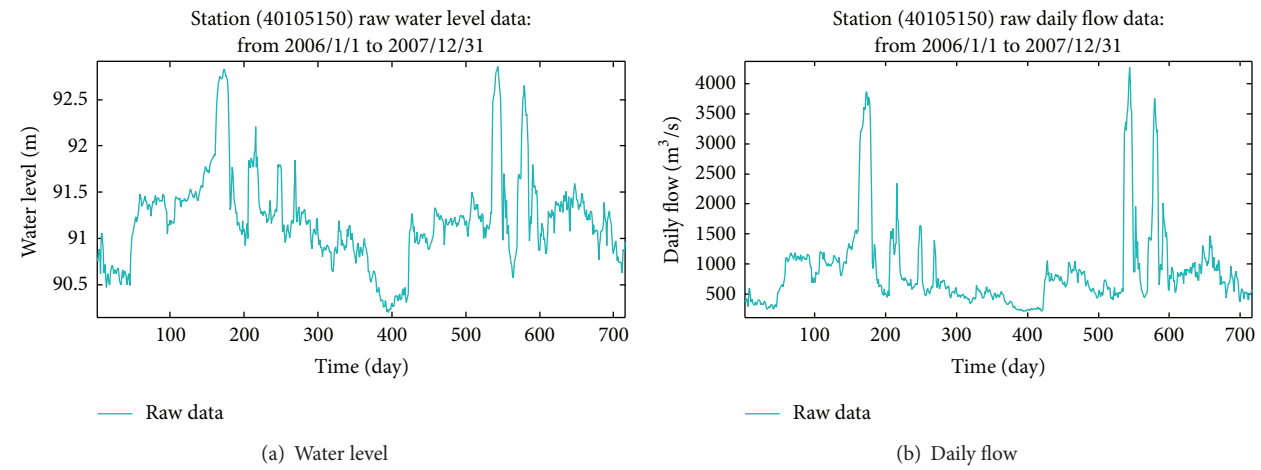
Truth	Detection	
	Outlier	Not an outlier
Outlier	True positives or TP (A) data points that are outliers and identify as outliers	False negatives or FN (C) data points that are outliers but identify as normal
Not an outlier	False positives or FP (B) data points that are normal but identify as outliers	True negatives or TN (D) data points that are normal and identify as normal

Another relevant parameter is the *Specificity*, the proportion of negative test results among the normal; the mathematical expression is as follows:

$$\text{Specificity} = \frac{TN}{(TN + FP)}. \tag{8}$$

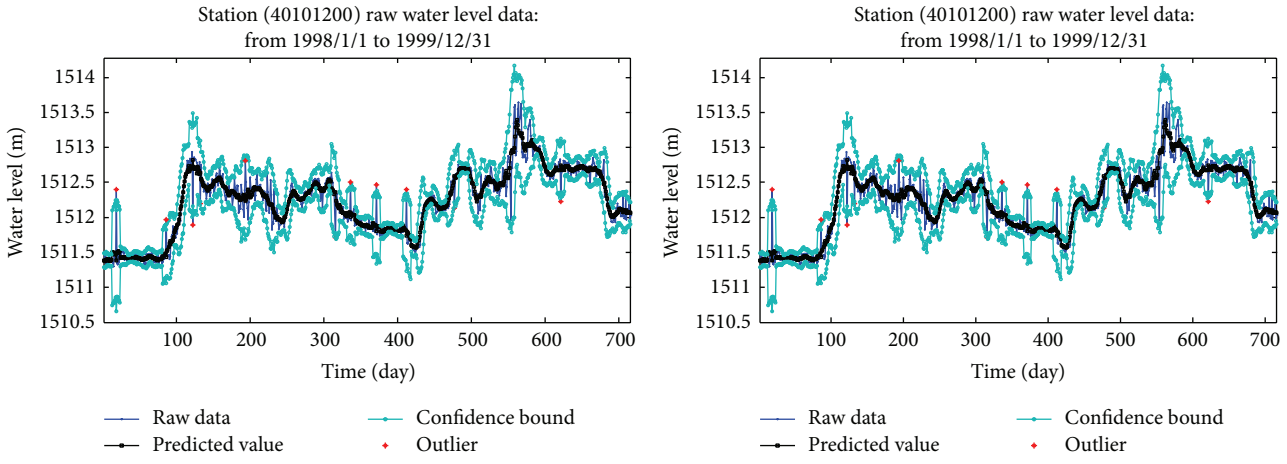
The positive predictive value (PPV) is the probability that a detected outlier is indeed a real one. Its formula is as follows:

$$PPV = \frac{TP}{(TP + FP)}. \tag{9}$$



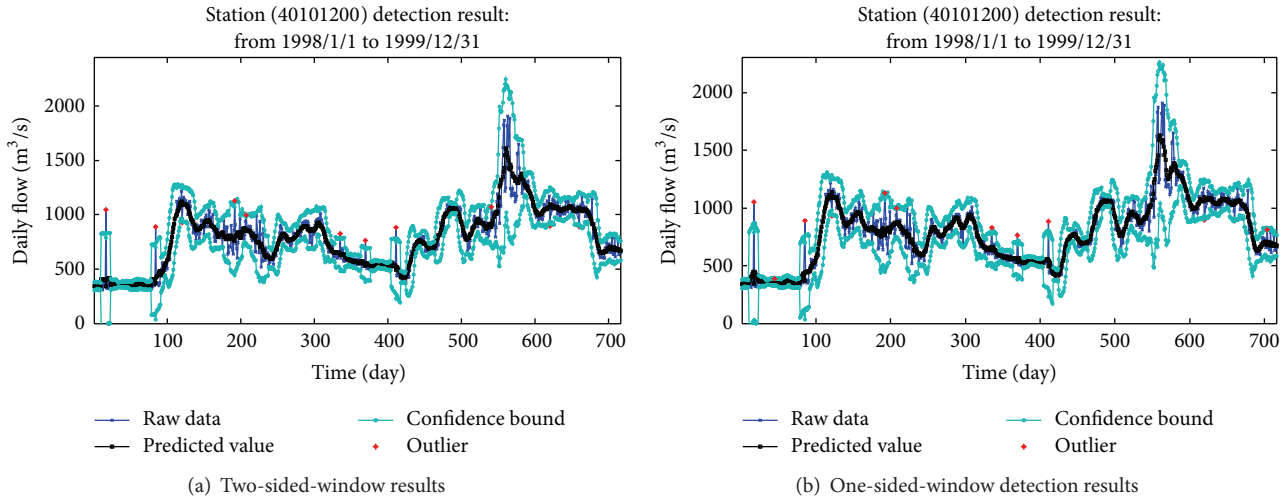
(a) Water level (b) Daily flow

Figure 5: Raw hydrological time series of *HYK* station.



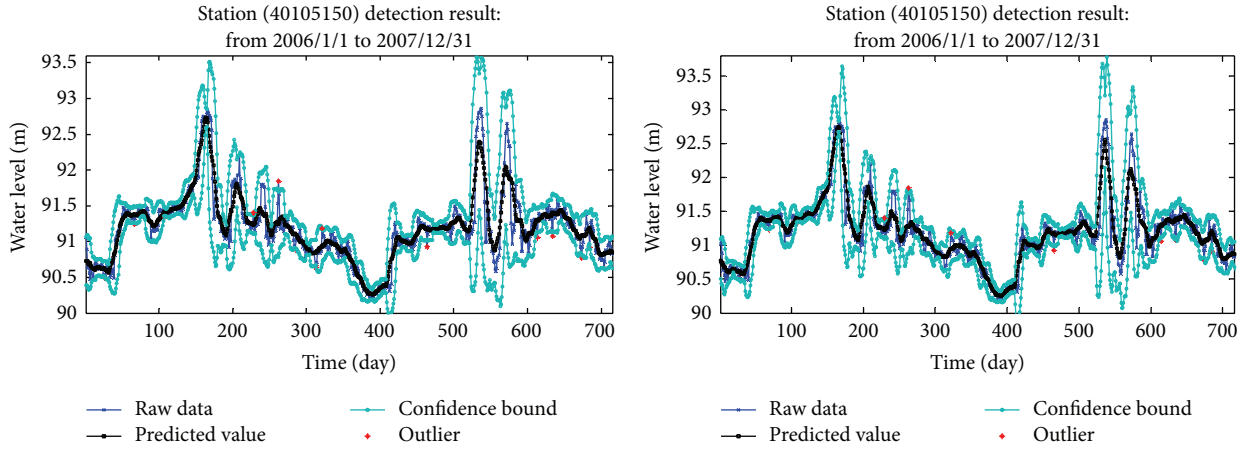
(a) Two-sided window detection results (b) One-sided window detection results

Figure 6: Detection results over *LZ* water level.



(a) Two-sided-window results (b) One-sided-window detection results

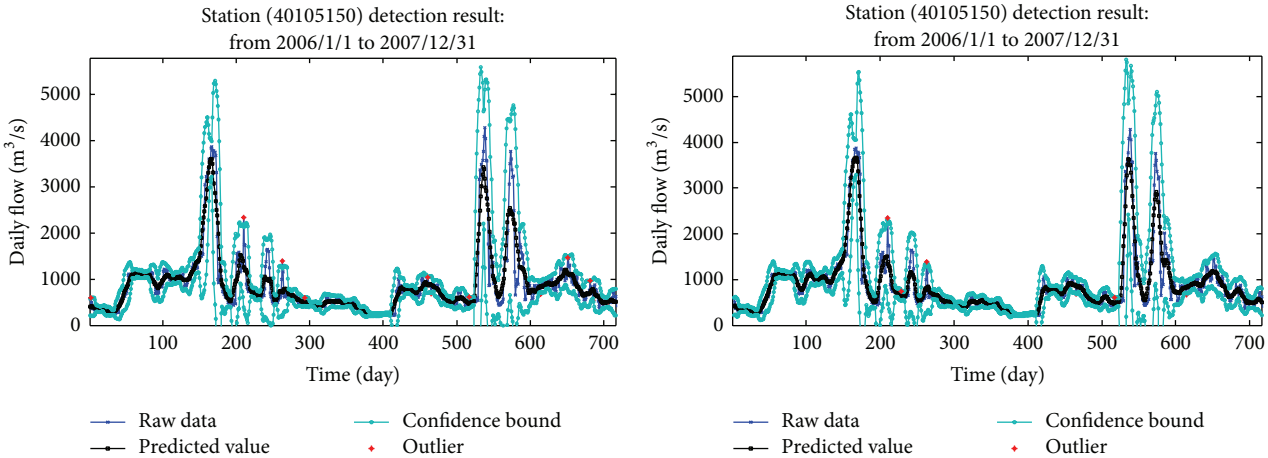
Figure 7: Detection results over *LZ* daily flow.



(a) Two-sided-window detection results

(b) One-sided-window detection results

Figure 8: Detection results over *HYK* water level.



(a) Two-sided-window detection results

(b) One-sided-window detection results

Figure 9: Detection results over *HYK* daily flow.

Finally, the negative predictive value (*NPV*) is the proportion of nonoutliers among subjects with a negative test result. Its formula is as follows:

$$NPV = \frac{TN}{(TN + FN)}. \tag{10}$$

4.3.2. Quantifying the Outlier Occurrences. A statistical measure of the accuracy is provided in this section. The parameters used are the ones described in Section 4.3.1 and summarized in Tables 2 and 3. Note that all parameters refer to the whole year from 2006 to 2007 of *HYK* station and 1998 to 1999 of *LZ* station. And just for simplicity, only the detection result with three pairs of parameters and the explanation for the water level where (k, p) takes of the value (6, 0.95) is provided, since the reasoning for the remaining hydrologic elements and parameters is analogous. Note that the optimal values for the parameters (k, p) would be (6, 0.95) on water level time series of *HYK* station according to parameter optimization method described in Section 3.3.

As it is shown in Table 2, during the one-sided methods detecting process on water level time series of *HYK* station with (k, p) taking the value of (6, 0.95), there were 14 outliers properly detected; that is, $TP = 14$. On the other hand, there were two occasions which originally represent normal event but are to be considered an outlier by the methods; therefore, $FP = 2$. Furthermore, there was one day which was outlier and was not detected by the methods; that is, $FN = 1$. And the remaining 713 days were correctly judged as normal event. Thus, $TN = 713$ (14 + 2 + 1 + 713 = 730 forecasted days from 2006 to 2007).

The results show great accuracy for both features. Particularly remarkable are the values reached by the specificity. In particular, it exceeds 99% in all situations. These values mean that when the approach classifies the day to be predicted as normal, it does it with high reliability.

Table 2: Statistical analysis of one-sided methods with different parameters of HYK station.

Parameters	Water level				Daily flow			
	(5, 0.95)	(6, 0.95)	(6, 0.96)	(7, 0.95)	(5, 0.95)	(5, 0.96)	(6, 0.96)	(5, 0.97)
TP	12	14	12	11	15	16	14	13
TN	713	713	711	712	707	710	708	709
FP	2	2	4	3	5	2	4	3
FN	3	1	3	4	3	2	4	5
Sensitivity	80.00%	93.33%	80.00%	73.33%	83.33%	88.89%	77.78%	72.22%
Specificity	99.72%	99.72%	99.44%	99.58%	99.30%	99.72%	99.44%	99.58%
PPV	85.71%	87.50%	75.00%	78.57%	75.00%	88.89%	77.78%	81.25%
NPV	99.58%	99.86%	99.58%	99.44%	99.58%	99.72%	99.44%	99.30%

Table 3: Statistical analysis of both methods with optimal parameters of given dataset.

Parameters	LZ station				HYK station			
	Water level		Daily flow		Water level		Daily flow	
	One-sided (6, 0.96)	Two-sided (6, 0.95)	One-sided (7, 0.96)	Two-sided (6, 0.95)	One-sided (6, 0.95)	Two-sided (6, 0.96)	One-sided (5, 0.96)	Two-sided (6, 0.95)
TP	20	18	18	19	14	13	17	17
TN	704	704	706	704	713	710	710	708
FP	4	4	3	5	2	5	2	4
FN	2	4	3	2	1	2	1	1
Sensitivity	90.91%	81.82%	85.71%	90.48%	93.33%	86.67%	94.44%	94.44%
Specificity	99.44%	99.44%	99.58%	99.29%	99.72%	99.30%	99.72%	99.44%
PPV	83.33%	81.82%	85.71%	79.17%	87.50%	72.22%	89.47%	80.95%
NPV	99.72%	99.44%	99.58%	99.72%	99.86%	99.72%	99.86%	99.86%

As for the sensitivity, all the results reached values greater than 73% (except for daily flow with the situation (k, p) taking the value of (5, 0.97), in which it reaches 72.22%) and obtains 81.6% on average.

The *PPV* reached values similar to those of the sensitivity, in particular, slightly greater (82.8% on average). Therefore, it can be stated that when the proposed method determines that there is an upcoming outlier, it is highly reliable.

Finally, *NPV* provides similar values for all the situations, reaching 99.58% on average; that is, the rate of real outliers not found by the approach cannot be considered significant.

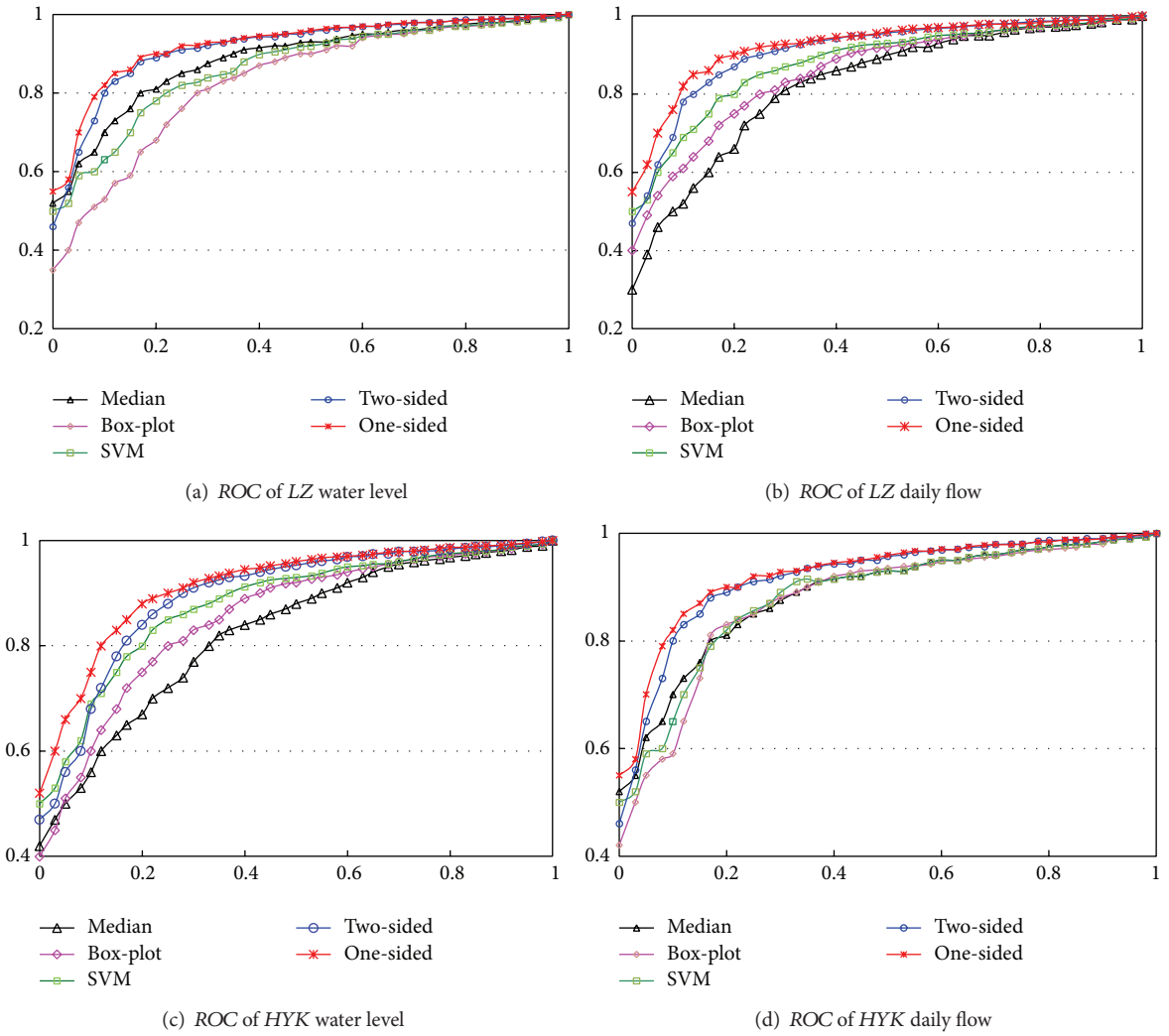
As for Table 3, it can be easy to draw a conclusion that different hydrologic elements of the same station and the same hydrologic elements of different station may take different optimal parameters values for (k, p) from experiments described in Section 3.3. Moreover, we can see that the detection accuracy of one-sided method is better than two-sided one as a whole. For example, there were 15 outliers in water level dataset of *HYK* station; one-sided method can correctly detect 14 abnormal and 713 normal data points; simultaneously, it masks one outlier as normal and swamps two normal samples as outliers. As a comparison, two-sided methods can correctly detect 13 abnormal and 710 normal data points; correspondingly, the masking and swamping effects reach to 2 and 5, respectively. The reason for this difference may lie in that the latter method needs both sides' neighbors which may contain outliers, which may lead to the fact that masking and swamping event occurs.

4.4. Analysis and Discussion. We compared our methods with other methods such as *SVM* (support vector machine) [42], box-plot techniques [40], and median method [23] on the same test datasets. The comparison results will display in the receiver operating characteristic (*ROC*) [43] curves, which will be shown later. By convention, the *ROC* curve displays sensitivity (*TPR*) on the vertical axis against the complement of specificity (1-specificity or *FPR*) on the horizontal axis. The *ROC* curve then demonstrates the characteristic reciprocal relationship between sensitivity and specificity, expressed as a tradeoff between the *TPR* and *FPR*. This configuration of the curve also facilitates calculation of the area beneath it as a summary index of overall test performance. Therefore, the larger the area under the *ROC* curve, the better the performance of the technique.

Figure 10 reveals the *ROC* curves obtained by proposed methods and other techniques. For these datasets, the performances of proposed methods are satisfactory and stable. For the water level dataset of *LZ*, our methods obtain similar results; their *ROC* curves show better performance than those of *SVM* and box-plot methods, while median method shows second-better *ROC* curves. For the daily flow dataset of *LZ*, our methods and *SVM* show better performance than median method and box-plot method; moreover, the performance of one-sided-window method is slightly better than that of two-sided-window. For the water level of *HYK*, our methods show preferable results, while *SVM* method obtains relatively better results than box-plot and median methods. For the

Table 4: Comparisons of area under ROC curves.

AUCs	Median	Box-plot	SVM	Two-sided	One-sided
LZ water level	0.922	0.894	0. 871	0.935	0.957
LZ daily flow	0. 843	0.852	0.895	0.92	0.933
HYK water level	0.819	0.836	0.865	0.903	0.921
HYK daily flow	0.934	0.928	0.93	0.942	0.955



(a) *ROC* of *LZ* water level

(b) *ROC* of *LZ* daily flow

(c) *ROC* of *HYK* water level

(d) *ROC* of *HYK* daily flow

Figure 10: *ROC* curves for different datasets. Five methods are compared: median method, box-plot method, *SVM* method, and our method, which are described using different colored lines.

daily flow of *HYK*, the *ROC* curves interlace with each other, so it is difficult to evaluate the performance only according to the observation. We have also calculated the *AUCs* for the *ROC* curves to compare the results (see Table 4). Note that there is a remarkable improvement in the detection performance. The average *AUCs* of our method are 0.925 and 0.942, which are the highest two among these five methods. The average *AUCs* of median, box-plot, and SVM methods are 0.892, 0.878, and 0.897, respectively.

For our method, it can be inferred that the anomalies can be effectively detected by the window-based forecasting model, which is constructed using *AR* prediction model and dynamically boundary movement strategies. The results suggest that our methods improve the robustness of the overall decision, while the other methods show different performances on different datasets. The other three techniques have the similar characteristic; they achieve better or worse results on different datasets. Results indicate that the proposed framework can find robust *PCI* as a means for defining the thresholds for detecting novelties in time series and can therefore improve performance of forecasting-based time series novelty detection. Moreover, our method is robust

to detect anomalies in different datasets, which means that it is more data independent.

It is important to mention that we have also validated our method on different datasets. The results were similar to those reported in this section. Furthermore, this paper used *PCI* rather than an arbitrary, user-defined threshold, which is mainly because of the fact that *PCI* can provide guidance (in the form of the confidence coefficient) for the calculation of the normal boundary region without requiring any knowledge of the process variables being measured. The $p\%$ *PCI* indicates the region likely to contain at least $p\%$ of the possible sensor measurements, and we expect that approximately $(1 - p)\%$ of the nonanomalous data will be misclassified as anomalous. This indicates that the normal assumption implicit in the *PCI* calculation is reasonable for the hydrological time series data. One limitation of our methods is that we identify a data point to be an outlier or not by comparing new observed value to *PCI* calculated dynamically according to different nearest-neighbor-windows size and confidence coefficient of difference users, which may cost a certain amount of time complexity to calculate optimization parameter for best detection results. Notwithstanding that limitation, this study does suggest that the proposed method can improve the performance of outlier detection. Results confirm that our methods can achieve a more robust performance than other outlier detection techniques.

5. Conclusions

Outlier detection, one of the classical topics of data mining, has generated a great deal of research in recent years owing to the new challenges posed by large high-dimensional data. In the meantime, outliers in hydrological time series have many practical applications, such as data QA/QC, adaptive sampling, and anomalous event detection. This research developed a time series outlier detection method that employs a window-based forecasting mode in conjunction with *PCI* to detect novelties in hydrological time series. The method first splits given historical hydrological time series into subsequences by a sliding-window, and then an autoregressive prediction model of time series was built from its nearest-neighbor-window to predict future values. Anomalies are assumed to take place if the observed values fall outside a given *PCI*, which can be calculated dynamically according to different nearest-neighbor-windows size and confidence coefficient of different user. Moreover, experiments with two different hydrological variables, water level and daily flow, from *LZ* and *HYK* stations obtained from national hydrology database of MWR, China, were performed to examine the effects of the method. In the experiments, single-layer linear networks were used for forecasting model; confidence coefficient and window size are assigned to 95% and 6 respectively in the initialization phase. Moreover, the best combination of parameters confidence coefficient and window size can be tuned according to different user's requirements and time series' features.

The case study results suggest that the proposed outlier detection methods developed in this study are useful tools for identifying anomalies in hydrological time series. Since these methods only require a time-series model of the time series, they can be easily applied to many real-time hydrological time series. However, it should be noted that, while the *LN* produced the best model for the water level and daily flow time series considered in the case study, this model may not be the most appropriate choice for other types of hydrological data. Furthermore, the optimization combination of parameters confidence coefficient and window size may cost a certain amount of time complexity for different hydrological element and different user's requirements. In spite of this, these methods do not require any preclassified examples of data, scale well to large volumes of data, and allow for fast incremental evaluation of data, which make them an ideal choice for correctly and efficiently hydrological time series outlier detection, especially for cases in which there is little information about how to set the threshold value.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work is supported by the Natural Science Foundation of China (nos. 51079040, 61170200, and 61370091) and the National Science and Technology Infrastructure of China (no. 2005DKA32000).

References

[1] J. D. Salas, "Analysis and modeling of hydrologic time series," in *Handbook of Hydrology*, vol. 19, pp. 1–72, 1993.

[2] W. Gujer, *Systems Analysis for Water Technology*, Springer, Berlin, Germany, 2008.

[3] N. Lauzon, *Water resources data quality assessment and description of natural processes using artificial intelligence techniques [Ph.D. thesis]*, University of British Columbia, 2003.

[4] K. Yang and C. Shahabi, "A PCA-based similarity measure for multivariate time series," in *Proceedings of the 2nd ACM International Workshop on Multimedia Databases*, pp. 65–74, ACM, November 2004.

[5] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis: Forecasting and Control*, John Wiley & Sons, New York, NY, USA, 2013.

[6] D. Machiwal and M. K. Jha, *Hydrologic Time Series Analysis: Theory and Practice*, Springer, New York, NY, USA, 2012.

[7] D. M. Hawkins, *Identification of Outliers*, vol. 11, Chapman &Hall, London, UK, 1980.

[8] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: a survey," *ACM Computing Surveys*, vol. 41, no. 3, article 15, 2009.

[9] M. Gupta, J. Gao, C. Aggarwal, and J. Han, *Outlier Detection for Temporal Data*, Synthesis Lectures on Data Mining and Knowledge Discovery, Morgan & Claypool, 2014.

[10] V. J. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artificial Intelligence Review*, vol. 22, no. 2, pp. 85–126, 2004.

[11] K. Das and J. Schneider, "Detecting anomalous records in categorical datasets," in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 220–229, ACM, August 2007.

[12] M. M. Breuniq, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: identifying density-based local outliers," *ACM Sigmod Record*, vol. 29, no. 2, pp. 93–104, 2000.

[13] Z. He, X. Xu, and S. Deng, "Discovering cluster-based local outliers," *Pattern Recognition Letters*, vol. 24, no. 9-10, pp. 1641–1650, 2003.

[14] C. C. Aggarwal and P. S. Yu, "Outlier detection with uncertain data," in *Proceedings of the 8th SIAM International Conference on Data Mining*, pp. 483–493, April 2008.

[15] S. Ando, "Clustering needles in a haystack: An information theoretic analysis of minority and outlier detection," in *Proceedings of the 7th IEEE International Conference on Data Mining (ICDM '07)*, pp. 13–22, October 2007.

[16] A. Agovic, B. Arindam, G. Auroop, and P. Vladimir, "Anomaly detection using manifold embedding and its applications in transportation corridors," *Intelligent Data Analysis*, vol. 13, no. 3, pp. 435–455, 2009.

[17] S. Barua and R. Alhajj, "A parallel multi-scale region outlier mining algorithm for meteorological data," in *Proceedings of the 15th ACM International Symposium on Advances in Geographic Information Systems (GIS '07)*, pp. 352–355, ACM, November 2007.

[18] F. Rasheed, P. Peng, R. Alhajj, and J. Rokne, "Fourier transform based spatial outlier mining," in *Intelligent Data Engineering and Automated Learning—IDEAL 2009*, vol. 5788 of *Lecture Notes in Computer Science*, pp. 317–324, Springer, Berlin, Germany, 2009.

[19] U. Rebbapragada, P. Protopapas, C. E. Brodley, and C. Alcock, "Finding anomalous periodic time series : an application to catalogs of periodic variable stars," *Machine Learning*, vol. 74, no. 3, pp. 281–313, 2009.

[20] A. J. Fox, "Outliers in time series," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 34, pp. 350–363, 1972.

[21] J. Ma and S. Perkins, "Online novelty detection on temporal sequences," in *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '03)*, pp. 613–618, ACM, August 2003.

[22] D. J. Hill and B. S. Minsker, "Anomaly detection in streaming environmental sensor data: a data-driven modeling approach," *Environmental Modelling and Software*, vol. 25, no. 9, pp. 1014–1022, 2010.

[23] A. L. I. Oliveira and S. R. L. Meira, "Detecting novelties in time series through neural networks forecasting with robust confidence intervals," *Neurocomputing*, vol. 70, no. 1–3, pp. 79–92, 2006.

[24] E. Keogh, J. Lin, A. W. Fu, and H. Van Herle, "Finding unusual medical time-series subsequences: algorithms and applications," *IEEE Transactions on Information Technology in Biomedicine*, vol. 10, no. 3, pp. 429–439, 2006.

[25] E. Keogh, J. Lin, and A. Fu, "HOT SAX: efficiently finding the most unusual time series subsequence," in *Proceedings of the 5th IEEE International Conference on Data Mining (ICDM '05)*, pp. 226–233, Houston, Tex, USA, November 2005.

[26] L. Wei, E. Keogh, and X. Xi, "SAXually explicit images: finding unusual shapes," in *Proceedings of the 6th International Conference on Data Mining (ICDM '06)*, pp. 711–720, Hong Kong, December 2006.

[27] Y. Bu, T.-W. Leung, A. W.-C. Fu, E. J. Keogh, J. Pei, and S. Meshkin, "WAT: finding top-K discords in time series database," in *Proceedings of the 7th SIAM International Conference on Data Mining (SDM '07)*, pp. 449–454, April 2007.

[28] J. Lin, E. Keogh, A. Fu, and H. van Herle, "Approximations to magic: finding unusual medical time series," in *Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems*, pp. 329–334, June 2005.

[29] M. Gupta, J. Gao, C. C. Aggarwal, and J. Han, "Outlier detection for temporal data: a survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 9, p. 1, 2014.

[30] Y. Zhang, N. A. S. Hamm, N. Meratnia, A. Stein, M. van de Voort, and P. J. M. Havinga, "Statistics-based outlier detection for wireless sensor networks," *International Journal of Geographical Information Science*, vol. 26, no. 8, pp. 1373–1392, 2012.

[31] R. Frieda, I. Agueusopa, B. Bornkampb et al., "Bayesian outlier detection in INGARCH time series," Sonderforschungsbereich (SFB) 823, 2012.

[32] A. Grané and H. Veiga, "Wavelet-based detection of outliers in financial time series," *Computational Statistics & Data Analysis*, vol. 54, no. 11, pp. 2580–2593, 2010.

[33] C. Bilen and S. Huzurbazar, "Wavelet-based detection of outliers in time series," *Journal of Computational and Graphical Statistics*, vol. 11, no. 2, pp. 311–327, 2002.

[34] F. Chebana and T. B. M. J. Ouarda, "Depth-based multivariate descriptive statistics with hydrological applications," *Journal of Geophysical Research: Atmospheres*, vol. 116, no. D10, 2011.

[35] R. H. McCuen, *Modeling Hydrologic Change: Statistical Methods*, CRC Press, New York, NY, USA, 2002.

[36] Interagency Advisory Committee on Water Data, *Guidelines for Determining Flood Flow Frequency: Bulletin 17B*, U.S. Geological Survey, Office of Water Data Coordination, Reston, Va, USA, 1982.

[37] R. J. Hyndman and H. L. Shang, "Rainbow plots, bagplots, and boxplots for functional data," *Journal of Computational and Graphical Statistics*, vol. 19, no. 1, pp. 29–45, 2010.

[38] F. Chebana, S. Dabo-Niang, and T. B. M. J. Ouarda, "Exploratory functional flood frequency analysis and outlier detection," *Water Resources Research*, vol. 48, no. 4, Article ID W04514, 2012.

[39] W. W. Ng, U. S. Panu, and W. C. Lennox, "Chaos based Analytical techniques for daily extreme hydrological observations," *Journal of Hydrology*, vol. 342, no. 1-2, pp. 17–41, 2007.

[40] J. M. Chambers, W. S. Cleveland, B. Kleiner, and P. A. Tukey, *Graphical Methods for Data Analysis*, Wadsworth, Belmont, Calif, USA, 1983.

[41] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, San Francisco, Calif, USA, 2001.

[42] J. Ma and S. Perkins, "Time-series novelty detection using one-class support vector machines," in *Proceedings of the International Joint Conference on Neural Networks*, vol. 3, pp. 1741–1745, July 2003.

[43] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.