Science of the Total Environment 610–611 (2018) 1390–1399

Contents lists available at ScienceDirect

Science of the Total Environment

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

Applying high-frequency surrogate measurements and a wavelet-ANN model to provide early warnings of rapid surface water quality anomalies

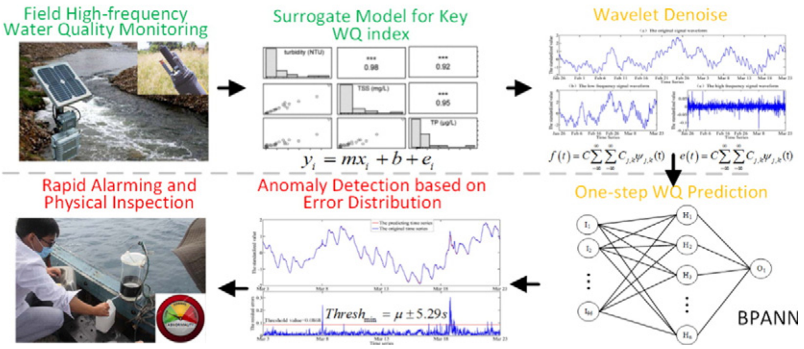
Bin Shi [a], Peng Wang [a,b], Jiping Jiang [a,c,*], Rentao Liu [a]

[a] School of Environment, Harbin Institute of Technology, Harbin 150090, China
[b] State Key Laboratory of Urban Water Resource and Environment, Harbin Institute of Technology, Harbin 150090, China
[c] School of Environmental Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China

HIGHLIGHTS

- Realized rapid surface water quality anomaly warnings using a data-driven approach
- Wavelet, ANN, high-frequency monitoring, and surrogate methods were combined.
- Anomaly thresholds for logistically distributed prediction residual error were found.
- Verified via anomaly events from the Potomac River monitoring program
- The method can be applied for urban aquatic (sponge city) and watershed management.

GRAPHICAL ABSTRACT


Field High-frequency Water Quality Monitoring
Surrogate Model for Key WQ index
turbidity (NTU)
0.98
0.92
TSS (mg/L)
0.95
TP (ug/L)
$y_i = mx_i + b + e_i$
Wavelet Denoise
$f(t) = C\sum\sum C_{j,k}\psi_{j,k}(t)$
$e(t) = C\sum\sum C_{j,k}\psi_{j,k}(t)$
Rapid Alarming and Physical Inspection
Anomaly Detection based on Error Distribution
$Thresh_{min} = \mu \pm 5.29s$
One-step WQ Prediction
BPANN

ARTICLE INFO

Article history:
Received 9 June 2017
Received in revised form 16 August 2017
Accepted 23 August 2017
Available online 30 August 2017

Editor: D. Barcelo

Keywords:
Water quality
Surrogate parameters
Anomaly detection
Wavelet denoising
Back-propagation neural networks

ABSTRACT

It is critical for surface water management systems to provide early warnings of abrupt, large variations in water quality, which likely indicate the occurrence of spill incidents. In this study, a combined approach integrating a wavelet artificial neural network (wavelet-ANN) model and high-frequency surrogate measurements is proposed as a method of water quality anomaly detection and warning provision. High-frequency time series of major water quality indexes (TN, TP, COD, etc.) were produced via a regression-based surrogate model. After wavelet decomposition and denoising, a low-frequency signal was imported into a back-propagation neural network for one-step prediction to identify the major features of water quality variations. The precisely trained site-specific wavelet-ANN outputs the time series of residual errors. A warning is triggered when the actual residual error exceeds a given threshold, i.e., baseline pattern, estimated based on long-term water quality variations. A case study based on the monitoring program applied to the Potomac River Basin in Virginia, USA, was conducted. The integrated approach successfully identified two anomaly events of TP variations at a 15-minute scale from high-frequency online sensors. A storm event and point source inputs likely accounted for these events. The results show that the wavelet-ANN model is slightly more accurate than the ANN for high-frequency surface water quality prediction, and it meets the requirements of anomaly detection. Analyses of the performance at different stations and over different periods illustrated the stability of the proposed method. By combining monitoring instruments and surrogate measures, the presented approach can support timely anomaly identification and be applied to urban aquatic environments for watershed management.

* Corresponding author at: School of Environment, Harbin Institute of Technology, Harbin, 150090, China.
E-mail address: jiangjp@sustc.edu.cn (J. Jiang).

http://dx.doi.org/10.1016/j.scitotenv.2017.08.232
0048-9697/© 2017 Elsevier B.V. All rights reserved.

1. Introduction

In recent years, the hydrosphere has been affected by human activities. Public drinking water systems, especially surface water, which serves most human populations, may be vulnerable in some localities. Anomalous water quality levels may be due to a variety of factors, such as natural accidents, uncertain point sources, and the intentional injection of contaminants (Liu et al., 2015; Hall et al., 2007). Although water quality monitoring systems generate data continuously using automatic sensors, the full benefits of online monitoring data cannot be obtained without real-time analysis. Smart anomaly detection systems (ADSs) based on real-time monitoring data are in high demand (Liu et al., 2015; Jiang et al., 2012).

The detection and identification of specific substances in water involve the use of traditional wet chemistry and emerging data-based technologies. Wet chemistry methods (e.g., titrimetry, spectral analysis and mass spectrometry analysis) are generally based on the use of sophisticated analytical instruments, such as spectrophotometers, chromatographs and mass spectrometers (Hou et al., 2013; Xi et al., 2004). Given that these sophisticated instruments are expensive, they are infrequently used for continuous monitoring. In addition, the analysis periods of these instruments are lengthy and may not be compatible with real-time monitoring objectives.

Sensor and time series-based mathematical methods involve event cluster analysis and threshold value judgment. Since the 9/11 terrorist attacks on the United States in 2001, the water security network of the U.S. Environmental Protection Agency (EPA) has developed a contamination warning system (CWS) for the real-time monitoring of drinking water distribution systems (US EPA, 2013). Baseline and simulated contamination events are input into the system, and abnormal and alert status levels are assigned. Concentration warning systems are generally based on real-time detection components (e.g., online water quality monitoring, enhanced security monitoring, customer complaint surveillance, and public health surveillance). Such data are often limited by monitoring technologies or due to issues related to information disclosure. Recently, with the development of big data applications, threshold values based on time series analysis have been used for anomaly detection. For example, Mckenna and Klise (2006) developed a multivariate distance measure based on water quality measurements and the closest observation for detection applications in water distribution systems. Yang et al. (2009) applied the adaptive transformation method by enhancing contaminant signals while reducing background noise to detect contamination events in a drinking water pipe. Miguntanna et al. (2010) used surrogate parameters for the rapid identification of urban stormwater quality levels.

Over the last several years, technologies have transformed watershed sciences by allowing for the direct measurement of water quality levels (Rode et al., 2016). With the development of information technologies, wireless sensor networks have become more popular as tools for watershed management. Water quality variables such as water temperature, dissolved oxygen, turbidity, pH, specific conductance and nitrite plus nitrate nitrogen can now be monitored continuously using probes. However, several water quality variables, such as total suspended solids, alkalinity, total nitrogen, total phosphorus, sodium, chloride, fluoride, sulfate, fecal *Escherichia coli*, polycyclic aromatic hydrocarbons, and mercury, cannot be monitored online at a high frequency (e.g., 15 min) due to the associated reaction mechanisms and measurement time requirements (Jones et al., 2011). The surrogate measurements can enhance the rapid generation of water quality data based on on-site measurements and thereby simplify resource-intensive laboratory analysis methods.

Many key water quality indexes, such as total nitrogen (TN), total phosphorus (TP), total suspended solids (TSS), etc., cannot be monitored at a high temporal resolution, and some may not be detected by a sensor in real time (Bieroza and Heathwaite, 2015). It is necessary to express these indexes in terms of online monitoring data using simple linear regression equations (Christensen, 2001; Horsburgh et al., 2010). Surrogate relationships can be used to estimate water quality concentrations at a much higher temporal resolution. Jones et al. (2011) studied the relationship between turbidity and TP that between turbidity and TSS to provide high-frequency estimates for TP and TSS concentrations, which cannot be determined by traditional monitoring approaches. As a form of surrogate measurement, these relationships have been used to measure the pollutant flux in water (Kirchner et al., 2011; Rügner et al., 2013), provide warnings of algal blooms (Bowes et al., 2016) and test hydrological models (Rode et al., 2016).

Surface water is part of a complex system, as water quality variables are influenced by many factors. It is difficult to use traditional mechanism models to describe high-frequency water quality variations. However, data-driven models such as artificial neural networks (ANN) and wavelet artificial neural networks (wavelet-ANN) can be used to describe these variations at a high resolution. Wavelets are treated as a "lens" that enables the researcher to explore long-term water quality anomalies. This approach can eliminate the effects of noise on the original time series during water quality prediction. ANN-based methods are a popular form of non-linear modeling, and such models can be used to identify correlated patterns between input and objective values via learning, memory, self-adaption and intelligent processes (Jorjani et al., 2009). Sun et al. (2010) applied time delay neural networks to predict hydrological model errors and provide theoretical guidance for the establishment of neural networks. For example, the results can be used to adjust the network synaptic weights, the number of hidden neurons and the number of epochs, as well as evaluate the error prediction efficiency.

Wavelet-ANNs and ANNs have been widely used for long-term time series analysis in water quality management. ANN models have been used to forecast pollutant concentrations in surface water based on monitoring data, and the corresponding results show that such models serve as easy-to-use modeling tools for engineers and water resource managers (Huang and Foo, 2002; Kuo et al., 2006). In subsequent studies, the wavelet was introduced. Recently, a wavelet-ANN model was used to predict water quality parameters, and the prediction accuracy was significantly better than those of ANN, multiple linear regression, and conventional sediment rating curve models (Rajaee, 2011). However, Du et al. (2017) raised an important issue regarding wavelet-ANN models. Because singular spectral analysis (SSA) and discrete wavelet transform (DWT) can potentially adopt 'future' values to perform calculations, the series generated by SSA reconstruction or DWT decomposition contain information associated with 'future' values. Thus, it is incorrect to train and test the predictor using the entire time series. This issue should be given particular attention in wavelet-ANN modeling. Other auxiliary methods, such as fuzzy inference, bootstrap simulation and surrogate measurement methods, have been introduced into ANN models to forecast water quality parameters under different scenarios (Csábrági et al., 2017; Erkyihun et al., 2016; Parmar and Bhardwaj, 2015). All of these studies serve as valuable references for the prediction of water quality time series.

Anomalies in surface water quality can be divided into three typical scenarios based on time scale. (1) *Average water quality value deviations from the baseline of more than a few days or longer*: this case always results from a change in the environment or in hydraulic conditions. (2) *Short-term outliers of less than an hour*: these are always caused by equipment failures or recording errors. They commonly occur during anomaly detection and are often eliminated through data processing and quality control. (3) *Average water quality deviations from the baseline over several hours or days* (US EPA, 2013): this case is always caused by contaminants from point or non-point sources. This anomaly scenario is the focus of the present study.

Our goal is to establish an efficient method of water quality anomaly detection based on online high-frequency water quality variables. To overcome the challenges related to data availability and detection efficiency during anomaly event detection, we present a combined method of data collection and surrogate measurement. Key water quality variables, such as TN, TP and TSS, are measured at a high resolution based on surrogate relationship analyses (Bieroza and Heathwaite, 2015; Christensen, 2001; Horsburgh et al., 2010). A wavelet-denoising algorithm is introduced to remove noise from an original water quality time series. A back-propagation neural network (BPNN) model and a goodness of fit test were used for anomaly detection. A case of abnormal water quality in the Potomac River Basin was used to verify the feasibility of the proposed method. Limitations and avenues for future research are discussed at the end of the paper.

2. Methodology and study area

2.1. A data-driven framework for the provision of water quality anomaly event warnings

The proposed method of water quality anomaly detection is depicted in Fig. 1. The method involves three steps.

Step 1: Calculate the anomaly threshold range from baseline observations. The standardized time series of historical water quality is treated by wavelet transformation to remove noise. The low-frequency component is used for BPNN, modeling while the high-frequency part is zeroed. Then, the two predicted parts are reconstructed and combined to form a complete water quality time series from corresponding wavelet functions. The threshold value interval of anomaly detection is calculated based on a residual error series between the predicted and actual water quality time series.

Step 2: Upcoming water quality values are predicted from the trained wavelet-ANN model via online monitoring. As soon as the latest monitoring data are available, the next time step of water quality values is predicted from the trained ANN model. The residual error is calculated for as long as monitoring data are collected.

Step 3: The residual error is compared to the baseline threshold. When the residual error exceeds the threshold interval and when the duration exceeds a specified period, water quality levels may be experiencing an anomaly event requiring a warning, physical inspection or emergency response.

2.2. The high-frequency monitoring and surrogate measurement method

Surrogate measurements are appropriate for water quality variables that cannot be measured directly in situ or that are difficult to measure from high-resolution online observations (TP, TN, COD, etc.). High-frequency, in situ water quality monitoring data can capture characteristic trends and periods overlooked by traditional periodic sampling and is an emerging area of research (Kunz et al., 2017; Rode et al., 2016).

As an example, suspending solids are a major contributor to unfiltered TN and TP in natural surface waters (Jones et al., 2011). Turbidity, the degree of the relatively clarity of a liquid, which is influenced by suspended solids, has been used as a surrogate measure of the concentration of suspended solids in addition to constituents such as TP and TN, which may be associated with suspended matter (Kuo et al., 2006; Viviano et al., 2014). Notably, most TN and TP particles are attached to fine suspended materials during migration in water. Specific conductance is related to concentrations of TN and TP, and high specific conductance values can affect the precipitation and adsorption of nitrogen and phosphorus. Additionally, dissolved nitrogen and phosphorus are closely related to pH and water temperature; thus, TN and TP concentrations are related to water temperature (Christensen, 2001).

Water quality data collected using automatic monitoring sensors can be explored as surrogate information to develop regression equations for rapid anomaly detection (Helsel, 1992; Parmar and Bhardwaj, 2014). The simplest regression equation can be expressed as follows:

$$y_i = kx_i + b + e_i \quad i = 1, 2, \ldots, n \tag{1}$$

where y_i is the ith observation of the response variable, x_i is the ith observation of the explanatory (independent) variable, k is the slope, e_i is the random error of the ith observation, b is the intercept, and n is the number of samples.

In addition, certain surrogate variables have significant relationships with the response variable. However, there should be a reasonable physical basis or explanation for their inclusion in the regression equation. Relationships between surrogate measurements and water quality constituent concentrations are also site specific and must be developed

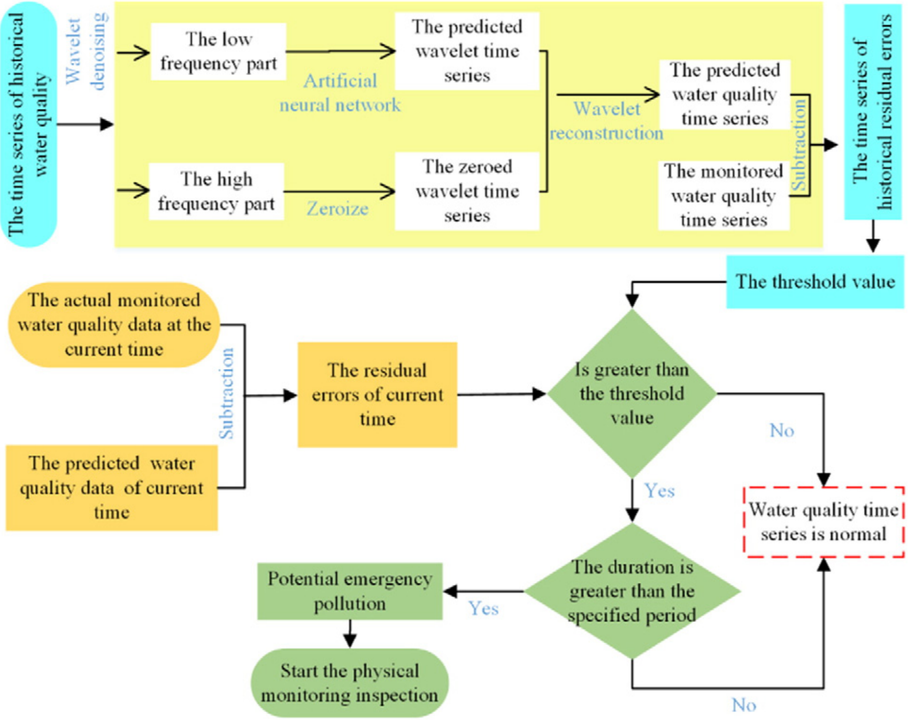

The time series of historical water quality
Wavelet denoising
The low frequency part
Artificial neural network
The predicted wavelet time series
The high frequency part
Zeroize
The zeroed wavelet time series
Wavelet reconstruction
The predicted water quality time series
The monitored water quality time series
Subtraction
The time series of historical residual errors
The threshold value
The actual monitored water quality data at the current time
Subtraction
The predicted water quality data of current time
The residual errors of current time
Is greater than the threshold value
No
Yes
Water quality time series is normal
The duration is greater than the specified period
Yes
Potential emergency pollution
Start the physical monitoring inspection
No

Fig. 1. Flow chart of the proposed combined approach of water quality anomaly detection.

for each sensor node. Therefore, these relationships vary in different watersheds or locations.

2.3. Wavelet denoising of the original time series

Wavelet technologies have been widely used for data-driven hydrological and environmental modeling (Erkyihun et al., 2016; Parmar and Bhardwaj, 2015; Rajaee, 2011). The wavelet transform function is an integral function that transforms a time series into a two-dimensional plane to obtain essential features of the original time series. Wavelet analysis can be used to transform an original time series into components (major trend (low-frequency parts) and noise (high-frequency parts)). In this study, the low-frequency part contains the sub-hourly variations in surrogate water quality time series (Shupe, 2017). Such series reflect the complex temporal dynamics that are obscured by traditional sampling frequencies and provide new insight into the inner-workings of watersheds and streams. The high-frequency part contains noise caused by measurement errors or the surrounding environment. The original product of water quality monitoring $S(\mathrm{i})$ is a one-dimensional discrete time series expressed as follows:

$$S(\mathrm{i}) = f(t) + \sigma \times e(t),\ \ t = 0, \ldots, \mathrm{n}-1 \tag{2}$$

where $f(t)$ is the real time series, $e(t)$ is the time series of noise, i is the sampling time, and σ is the coefficient of noise (standard deviation).

During wavelet denoising, the standard deviation can be replaced with 0. The corresponding discrete wavelet function can be expressed as follows:

$$\psi_{j,k}(\mathrm{t}) = a_0^{-\frac{j}{2}}\psi\left(\mathrm{a}_0^{-j}t - kb_0\right) \tag{3}$$

where a_0^j is the scaling factor, b_0 is the time factor, t is time, and j is an integer that represents the element number of the time series.

The wavelet coefficients of discretization can be expressed as follows.

$$C_{j,k} = \int_{-\infty}^{\infty} f(t)\psi^*_{j,k}(t)dt \tag{4}$$

Additionally, the wavelet reconstruction formula can be expressed in the following form:

$$f(t) = C\sum_{-\infty}^{\infty}\sum_{-\infty}^{\infty} C_{j,k}\psi_{j,k}(\mathrm{t}) \tag{5}$$

where C is a constant. Moreover, $e(t) = C\sum_{-\infty}^{\infty}\sum_{-\infty}^{\infty} C_{j,k}\psi_{j,k}(\mathrm{t})$.

2.4. Water quality prediction using the BPNN and a baseline set

A BPNN is a multilayered feedforward network trained according to an error back-propagation algorithm (Bade et al., 2015). It can learn and store numerous input-output mapping relations without requiring mathematical equations. The original time series were transformed into different frequency components via wavelet transform. The number of input parameters for the neural network was determined based on the periods of different frequency components. For example, if there are n data sets in a given period, the number of input parameters in the neural network is n.

Here, the data set was divided into two subsets: the training set (January 26–March 2) and test set (March 3–March 23). The training data set contained training data (January 26–February 14) and cross-validation data (February 15–March 2). The training data were used for training the network and adjusting the synaptic weights. Cross-validation data were used to avoid overfitting. When the error between the predicted values and desired values in the cross-validation data set increases, the training stops, and the result is considered the optimal generalization (Babovic, 2005; Sannasiraj et al., 2004; Sun et al., 2010). As Du et al. (2017) suggested, in this forecasting method, the data in the 'next time step' are not included in the DWT pretreatment, and the validation data sets are transformed separately to prevent the 'fake' good performance of the hybrid data-driven model.

The neural network contained two hidden layers. The log-sigmoid function (the input value is unlimited; the output value ranges between 0 and 1) and the purelin function (the input and output values are unlimited) were selected as the activation functions to improve the performance of the BPNN model (Fig. 2). When the neural network has been trained and tested, a new time series can be predicted. The new predicted time series was reconstructed from a wavelet reconstruction function and was compared with the actual monitoring data to evaluate its accuracy.

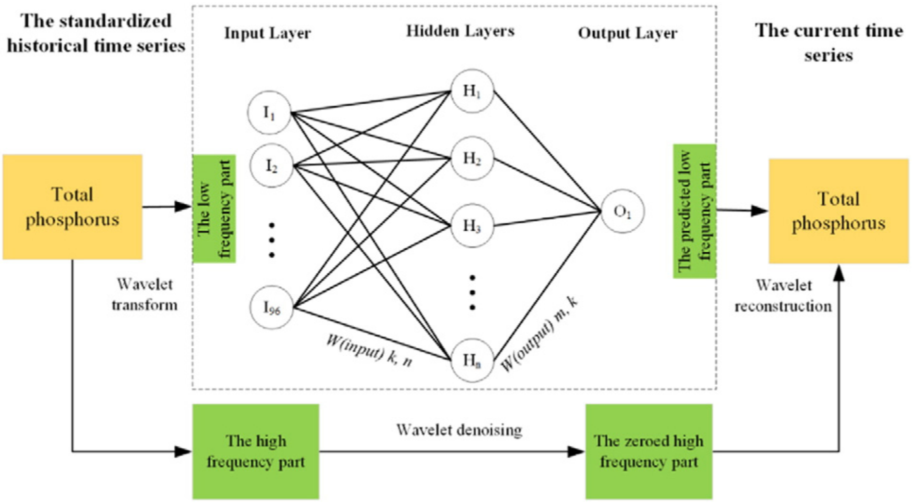


Fig. 2. The notional structure of the neural network prediction method based on wavelet analysis.

The performance of the forecasting model can be assessed based on the residual error (RE) between the predicted and original monitoring values. RE hypothetically follows a certain distribution (normal, gamma, beta, logistic, etc.), and it can be tested via the goodness of fit method. Moreover, the mean, medium and square deviation can be estimated. Hypothesis testing is used to calculate the approximate probability distribution model. Threshold values must be used to classify residuals that are indicative of background or outlier water quality (Mckenna et al., 2011). In this study, the accumulation of 99% of RE was set as normal. During new monitoring data detection, the new predicted RE was compared with the threshold value. When the new RE falls within the range of threshold values, the water quality level is normal. By contrast, when the new RE falls outside of the threshold value range, the water quality level represents an anomaly.

2.5. Study area and monitoring data

The monitoring program applied to the Potomac River by the United State Geological Survey (USGS) was adopted for our case study. The Potomac River is located along the mid-Atlantic coast of the United States and flows into the Chesapeake Bay (Fig. 3). The river has north and south branches that originate from sources in the Appalachian Mountains of West Virginia. The river (including both branches) is approximately 405 miles (652 km) long with a drainage area of approximately 14,700 mile^2 (38,000 km^2). It is the fourth largest river along the Atlantic coast of the United States. Approximately 5 million people live in this basin. Moreover, the average river flow is approximately 10,800 ft^3/s (306 m^3/s) (US EPA, 2017).

As previously noted, the river has two sources. The source of the northern branch is located at the Fairfax Stone, which is positioned at the junction of Grant, Tucker, and Preston counties in West Virginia. The source of the southern branch is located close near Hightown in northern Highland County, Virginia. There are three major tributaries to the river (the Shenandoah River at Harpers Ferry, the Monocacy River in the Piedmont region of Virginia, and the Anacostia River in Washington, D.C.). Prior to 2007, the Potomac River and Chesapeake Bay experienced serious environmental problems due to eutrophication. On November 13, 2007, the Potomac Conservancy instituted a "D-plus" program, and nitrogen and phosphorus loading have been strictly controlled since then (Byrand, 2010).

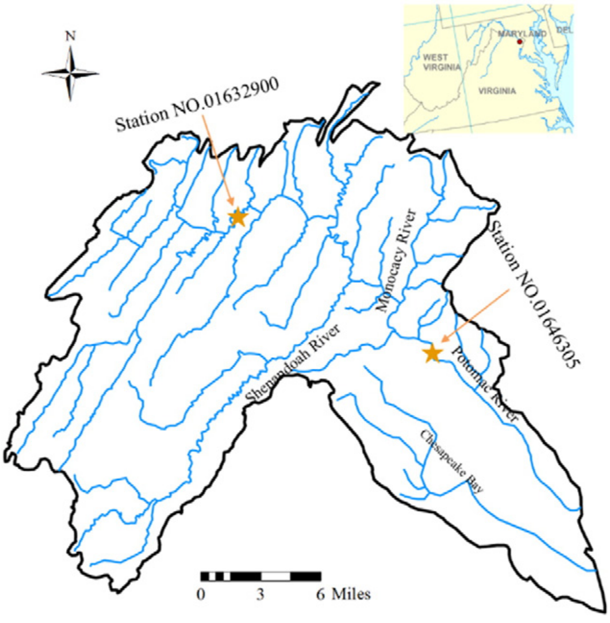


Fig. 3. Locations of the Potomac River and monitoring stations in this study.

The USGS (https://waterdata.usgs.gov/nwis/sw) collects surface water data, including water levels, streamflow (discharge), surface water quality, rainfall, etc., at >850,000 stations. These data are mainly detected by sensors and via manual field measurements. The surface water quality variables include water temperature (WT), specific conductance (SC), dissolved oxygen (DO), pH, turbidity (TURB), and nitrate + nitrite nitrogen (NOx). Measurements are commonly recorded at a fixed interval of 15 to 60 min and transmitted to the USGS every hour.

In relevant hydrological studies, different types of stations and different temporal periods should be investigated. Monitoring station no. 01632900, which is located at Smith Creek, an upstream tributary close to New Market, Virginia, was selected as a major station for validating the proposed anomaly detection approach. In addition, station no. 01646305, which is located on the lower main stream of the Potomac River near McLean, was selected for further comparison. The observation period of station no. 01646305, which spanned from January 26, 2017, to March 23, 2017 (55 days), was selected as the major period of investigation. At station no. 01632900, two different periods (10 days and 1 year) were selected for comparative study. Stations and periods were not specially selected. Any period that meets the basic requirements of an analysis can be used for validation as long as it contains abnormal events and has high data quality (for example, records at some stations are missing for long periods due to various unforeseen reasons).

Intermittent data losses are unavoidable due to sensor malfunctions despite weekly station maintenance (Cassidy and Jordan, 2011). Low-quality data (incomplete, noisy, inconsistent, redundant, etc.) in an original sampling time series must be filtered by adding missing data, disregarding redundant data, and transforming noisy and inconsistent data. In this study, missing data were interpolated using adjacent data. Singularities in the data were corrected with adjacent data or by adding an appropriate and constant offset to the raw data (Wagner et al., 2006).

Original data set $\mathrm{d_i}$ was standardized via *Z*-score transform with an average of 0 and a variance of 1:

$$D_i = \frac{d_i - \overline{d}}{\sigma} \tag{6}$$

where $\overline{d}$ is the mean of d_i and σ is the variance of d_i.

The standardized time series of water quality parameters are shown in Fig. S1. These parameters fluctuate periodically (and especially pH, DO and specific conductance levels) during the study periods. Water quality variables such as turbidity and specific conductance show no obvious periodic characteristics. These two variables are sensitive to surrounding environmental conditions, and slight changes in the environment can change these two variables considerably. Therefore, considerable noise can be observed in these two time series, making it difficult to use the time series for prediction. Thus, noise should be removed to provide accurate predictions of water quality variables.

3. Results and discussion

3.1. TP surrogate measurements and anomaly events in the study area

The concentrations of TP in the Potomac River Basin are strongly correlated with WT ($r = 0.92$, $p < 0.001$) and have significant but weaker relationships with TURB and SC ($r = 0.64$ and 0.48, respectively) (Viviano et al., 2014). Water quality surrogate relationships are dynamic and fluctuate seasonally and over longer time periods, as land uses and sources of constituent loading change and should be periodically refined through continuous data collection. Eq. (7) shows the surrogate relationships of TP (Jastram, 2014) used here to construct high-frequency TP time series (although we do not calibrate them).

$$TP = 0.00103\mathrm{TURB} + 0.00570\mathrm{WT} - 0.227\log_{10}SC + 0.776 \tag{7}$$

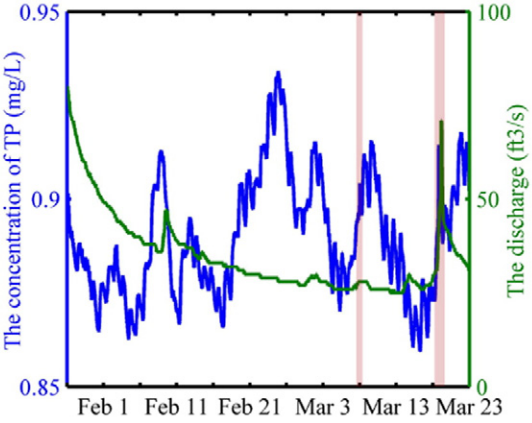


Fig. 4. Concentration time series of TP and discharge.

The concentrations of TP and corresponding discharge levels at the station are shown in Fig. 4. A sharp increase in TP concentrations occurs on March 19 and is consistent with increasing water discharge. According to historical observations of discharge and meteorological conditions, one obvious anomaly event for TP concentrations due to rainfall occurs on March 18–19. The average discharge on March 19 is 49 ft^3/s (1.38 m^3/s), which is higher than that of the predictive period (29 ft^3/s (0.82 m^3/s)). This phenomenon may have occurred when phosphorus organic matter on land was washed into the river by rain. Another sharp increase in the TP concentration occurred on March 8, although discharge levels did not change considerably in this period; thus, this increase was potentially due to point source spills. This event can be considered a water quality anomaly. In the following section, we use the developed method to detect these two anomalous events.

3.2. Wavelet denoising from the original TP time series

The results of the one-dimensional DWT of the TP time series are shown in Fig. 5. The original time series were transformed using the multilevel wavelet reconstruction function. In this study, Daubechies-4 (db4) was selected as the mother-wavelet function. Named after its inventor, Ingrid Daubechies, the Daubechies wavelet transform function includes a family of orthogonal wavelets and is characterized by a maximal number of vanishing moments for a given wavelet. For each wavelet type, a scaling function is used to generate an orthogonal multiresolution analysis. In the dbN (N is the order number) method, the number of vanishing moments increases as N increases, and the vanishing moments become smoother. Fig. 5(b) presents the low-frequency component and shows the major variations in TP. Fig. 5(a) shows that the time scale of each minor fluctuation is close to one day. Thus, we can determine the sizes of input signals via artificial network prediction. Fig. 5(c) shows the high-frequency component, which contains noise. This component is shaped by intermittent changes in the surrounding environment and in sensor drift. Such issues may compromise the predictive performance of the neural network and affect anomaly detection. Therefore, 'noise' is removed during the following baseline construction process.

3.3. Baseline construction for water quality prediction

In this case, data for the previous 35 days were set as a baseline for anomaly detection and were used for BPNN training and cross-validation. The latter 20 days, spanning from March 3 to March 23, was set as the period for the application of rapid surface water quality anomaly warnings. The standardized TP time series is illustrated in Fig. 6.

The neuron number of the input layer is determined from the potential period of the original time series. The temperature changes via a diurnal cycle (Zhai et al., 2014), and water quality is normally associated with instream photosynthesis/respiration dynamics (Halliday et al., 2015). Thus, the diurnal cycles of surface water quality parameters are considered in the baseline condition. In this study, the number of neurons in the input layer is 96, and the 15-min intervals of these 96 neurons are transformed to one day, or 24 h. Since there is no specific algorithm available to determine the number of neurons in the hidden layer, we tested 2 to 16 neurons and selected the optimal number according to the algorithm performance. The output layer includes one neuron, which denotes the next time step of water quality observation for one-step prediction.

The neural network was trained using a standard gradient descent algorithm to minimize the least-squares objective function. In BPNN modeling, the network structure and training iteration number are two important factors that should be considered to improve model efficiency and prevent the BPNN model from becoming over-trained. From this study, we found that 1000 epochs satisfy the training network based on a performance goal of 10^{-4}.

The time series of the two components were reconstructed using the wavelet reconstruction function. Fig. 7 shows a histogram of the REs the baseline. Normal and logistic functions were used to fit the error distribution. The random error clearly does not obey the normal distribution in this case. A logistic distribution (Eq. (8)) is symmetric, and the 0.5% and 99.5% quantiles of the cumulative distribution are used to identify

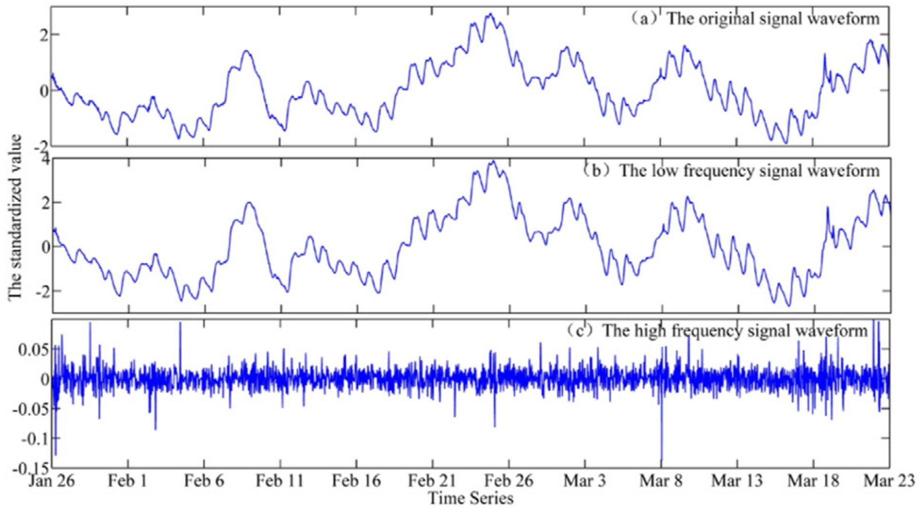


Fig. 5. Wavelet transform results for the original time series of surrogate TP.

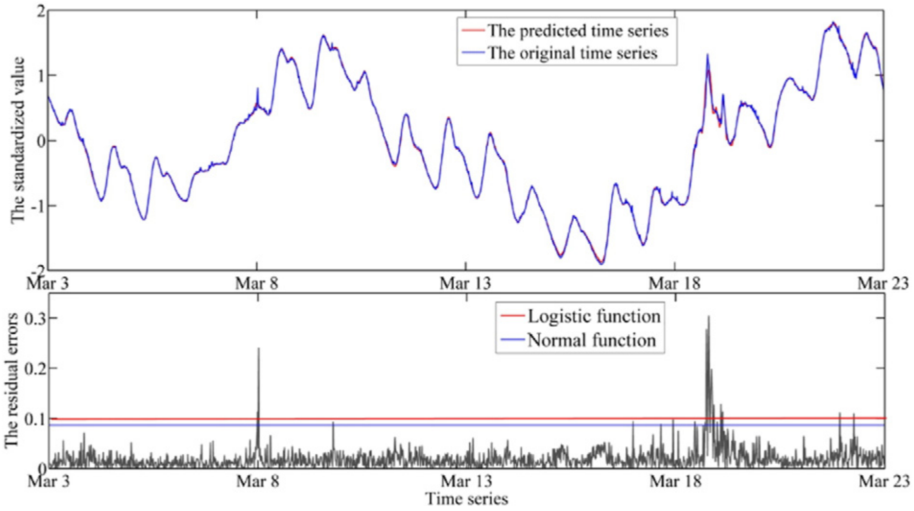


Fig. 6. Standardized TP time series and the prediction of residual errors during the test period (20 days).

outliers (Section 2.4). Eq. (9) shows the upper and lower threshold values of the baseline scenario.

$$f(x;\mu,s)=\frac{e^{-\frac{x-\mu}{s}}}{s\left(1+e^{-\frac{x-\mu}{s}}\right)} \tag{8}$$

$$\begin{aligned} Thresh_{\text{inf}} &= \mu+\ \ln\left(\frac{p_{\text{inf}}}{1-p_{\text{inf}}}\right)s=\mu-5.29s \\ Thresh_{\text{sup}} &= \mu+\ \ln\left(\frac{p_{\text{sup}}}{1-p_{\text{sup}}}\right)s=\mu+5.29s \end{aligned} \tag{9}$$

where μ and s are the mean and scale parameter, respectively.

In this case, the logistic error distribution has a mean value of $\mu=0.001$ and a scale parameter of $s=0.019$. Therefore, we find that $Thresh_{inf}=-0.098$ and $Thresh_{inf}=0.100$.

Fig. 7 illustrates the RE time series over the last 20 days of the test period. Compared to the logistic distribution based on threshold intervals (red line in Fig. 7), the normal distribution based on threshold intervals (blue line) is narrower (0.099 vs. 0.087), and the logistic distribution is more centered around the mean. Thus, more events involving significant variations will be considered abnormal events according to the assumptions of the normal distribution.

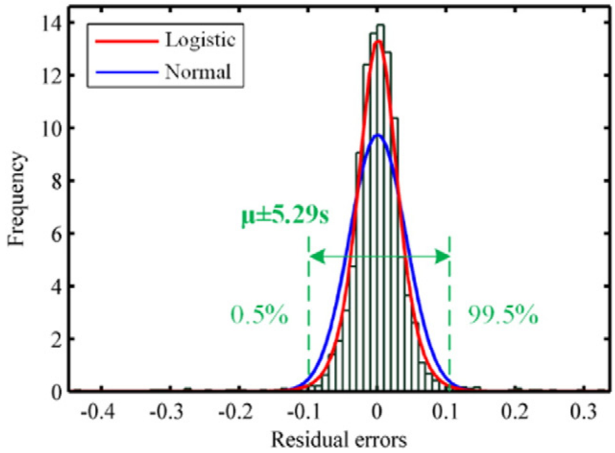


Fig. 7. Distribution of the residual errors of prediction for the baseline (station no.01632900).

Notably, the RE is greater than the threshold value on March 8 and March 19. The anomaly event on March 19 confirms the hypothesis. The event occurred when phosphorus organic matter on land was washed into the river by rain. However, the anomaly event on March 8 may have been caused by point sources. An emergency monitoring network should be initiated to confirm this theory.

3.4. Reliability analysis of anomaly detection results

The receiver operating characteristic (ROC) curve from electronic signal observation theory has been widely used in water quality event-detection algorithms (He et al., 2013). It can be used to assess the quality of decision making and the accuracy of detection information.

The ROC curve was plotted with the false alarm rate (FAR) as an abscissa and with the probability of detection (PD) as an ordinate. The PD is a function of true positives and false negatives. True positives are actual anomaly observations identified as abnormities through this method. False negatives are actual anomaly observations estimated as normal conditions through this method. The FAR is a function of true negatives and false positives. True negatives are normal conditions estimated as normal through this method. False positives are normal conditions identified as abnormal through this method.

The area between the curve and x-axis is used to evaluate the detection accuracy. To test these two anomaly detection methods, two hypothetic scenarios were established. We established an abnormal event that occurred each day from March 3 to March 23 at 10:00 am and 12:00 am. In scenario A, the abnormal concentration is 2 times that of the original time series, and in scenario B, the concentration is 3 times that of the original time series. The wavelet-ANN and ANN methods were employed for water quality prediction. The detection accuracies of these two methods are shown in Fig. 8. Generally, all of the ROC curves are positioned over the line of $y=x$. Both predication models perform well in terms of anomaly detection. The area under the ROC curve of the wavelet-ANN method (0.981 and 0.993) is higher than that of the ANN method (0.820 and 0.936), suggesting that the wavelet-ANN is slightly more accurate than the ANN in this case. In both scenarios, the detection accuracy is positively related to an increase in the 'abnormal concentration' of TP.

Furthermore, the prediction performances of the wavelet-ANN and ANN are compared in Table S2 and Fig. 9. Table S2 compares their prediction errors. The five statistical indicators of RE show that the predictive performances of the ANN and wavelet-ANN are similar. However, the ANN is slightly more accurate than the wavelet-ANN for baseline calibration and validation (network training period), while the

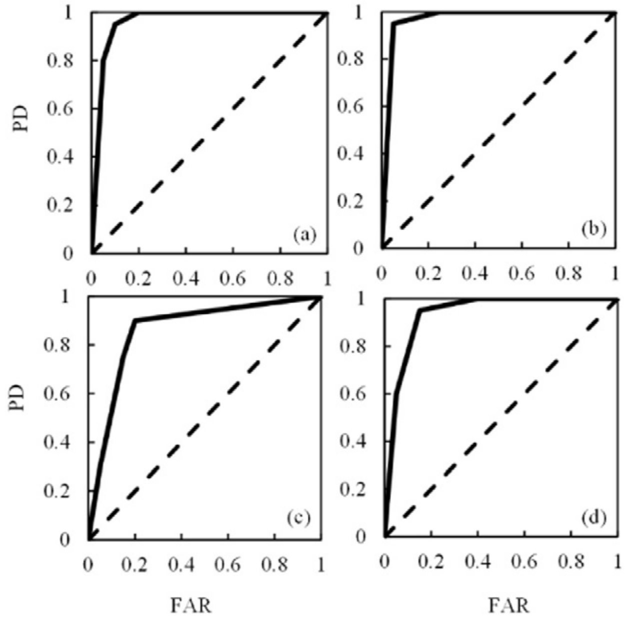


Fig. 8. ROC curves for Scenario A according to the wavelet-ANN (a) and ANN (c) methods and for Scenario B according to the wavelet-ANN (b) and ANN (d) methods.

wavelet-ANN is superior to the ANN in terms of verification. The wavelet-removed noise associated with large variations may account for this phenomenon. Fig. 9 compares the local time series (one day) and corresponding REs of the two models. In Fig. 9(a), compared with the ANN curve, the wavelet-ANN curve is more stable amidst sharp changes in the original time series because the high-frequency noise was removed.

These results are embodied in Fig. 9(b). As is shown in the yellow circle, the predicted RE of the wavelet-ANN method is 0.198 (great than the threshold value), while the residual error of the ANN is 0.040 (less than the threshold value). Thus, the wavelet-ANN can detect an abnormal event more accurately. However, as is shown in the green circle, when the original monitoring data are typical, the REs of the ANN are 0.454, 0.768 and 0.304, which exceed the threshold value. By contrast, the REs of the wavelet-ANN method are 0.036, 0.069 and 0.013, and all of these values are less than the threshold. These results suggest that the wavelet-ANN can improve the accuracy of anomaly detection by increasing the PD and decreasing the FAR.

To further validate the anomaly detection performance at other locations, station no. 01646305 (Fig. 3) in the downstream portion of the basin was analyzed. Fig. S2 shows that the REs of prediction on March 8 and March 19, 2017, at station no. 01646305 are greater than the threshold values, and anomaly events can be identified. This finding is consistent with the results at station no. 01632900 (Fig. 6), which is expected because point sources in the upper portion of the watershed and storm water affect the entire basin. Additional comparisons were not made because the baseline pattern and surrogate model are site specific.

Comparisons of the wavelet-ANN prediction performance in different observation period were performed, as reported in Table S1. The wavelet-ANN model exhibited good performance over three different time spans. However, an insufficiently short period of data records in a data-driven model may lead to poor learning of the water quality pattern, low accuracy, and mistakes in identifying anomaly events. A two-month to one-year period may be suitable for training a wavelet-ANN for anomaly detection. However, this approach is site specific, and the length of records used for constructing a baseline will vary from site to site.

3.5. Discussion

The presented approach can be applied to many applications, including watershed management, stormwater management, and city maintenance (Scott and Frost, 2017). With respect to non-point source pollution, the real-time detection of extreme water quality events based on water quality data is more direct but more challenging than the event-based hydrologic monitoring of streamflow. In practice, quantitative analyses of the following issues should be carefully conducted in the future.

1) How does the type of data-based prediction model affect the anomaly detection results? The accuracy of prediction has obvious effects on the results of anomaly detection. Although further verification is required, the ANN results were acceptable and competent in this

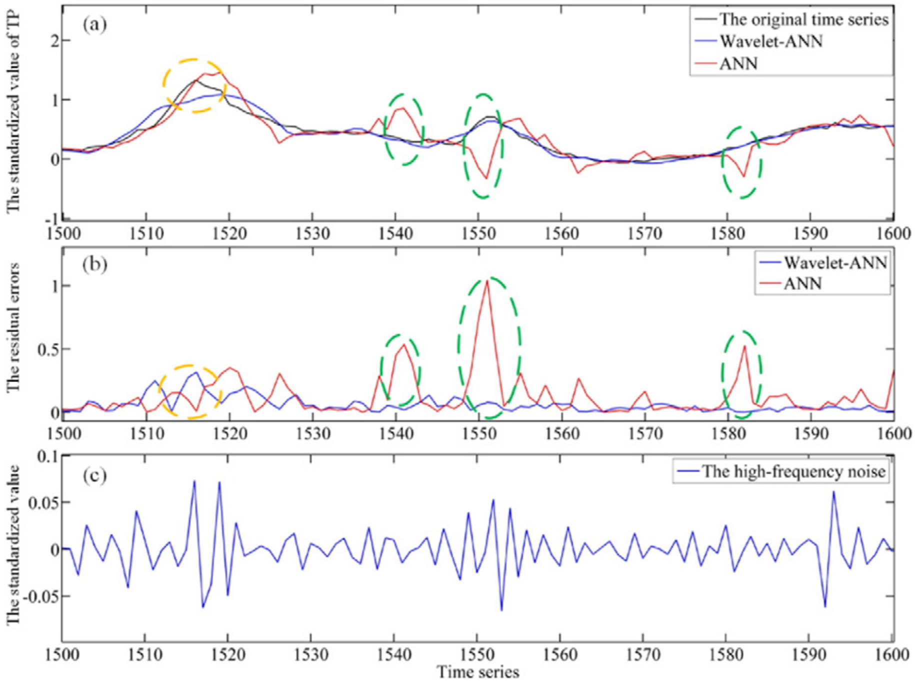


Fig. 9. Comparison of the predictive performance of the wavelet-ANN and ANN on March 19.

study. 2) What is the most appropriate water quality variation baseline? Qualitatively, a good baseline should be as long as possible, with acceptable fluctuations. Additionally, the baseline pattern may vary in different seasons. 3) Is it sufficient to consider one hydrological year? After a preliminary comparative investigation of different lengths of time series, including 10 days, 55 days and 1 year, it was found that the length of the time series had no obvious effect on the anomaly detection results (Table S1). However, as the amount of data used in the ANN model increased, the anomaly detection results stabilized. 4) How does uncertainty propagate throughout the anomaly detection process? For example, the output of the wavelet-ANN model cannot achieve 100% accuracy, and a confidence interval exists for REs. These issues can lead to anomaly detection uncertainties and contribute to the FAR. Factors that influence wavelet-ANN model prediction and the distribution of REs will lead to uncertainty in the anomaly detection results.

High-frequency sensors are used in the proposed approach. Surrogate relationships play an important role in this method in conjunction with high-frequency monitoring, thereby extending the use of this method to other water quality parameters. For example, surrogate models of organic chemicals or heavy metal concentrations can be established, and the presented approach can serve as a powerful tool for predicting the occurrence of chemical spills.

Compared to traditional methods of anomaly detection, i.e., online water quality monitoring, consumer complaint surveillance, public health surveillance, enhanced security monitoring and routine sampling and analysis, the average detection time of this method is short. Thus, the emergency response time required addressing sudden water pollution events and the response time for emergency monitoring and disposal should decrease (Shi et al., 2017).

4. Conclusions

This paper presents a novel approach to anomaly detection for surface water quality management. State-of-the-art high-frequency automatic monitoring methods, surrogate models and prediction error-based outlier detection methods were combined. Wavelets were used to remove insignificant noise in ANN prediction based on normal variations in water quality. The threshold range was defined according to the distribution of the RE time series. The dynamic data-driven process proposed can be adapted to longer normal time series.

Our method was verified based on the Potomac River monitoring program applied in Virginia, USA. We found that our method accurately identified anomaly events and sudden changes that occurred over short intervals. Compared to the ANN prediction method, the wavelet-ANN method was more sensitive to sudden water quality anomaly events and avoided the effects of false positive events in many cases. ROC tests based on two hypothetic scenarios yielded a detection accuracy of up to 0.98.

We live in the age of big data, and intelligent and precise environmental management is essential. The data-driven method proposed can serve as a state-of-the-art alternative for safeguarding water quality, improving urban and aquatic environmental management, controlling non-point and point source pollution, and enhancing watershed management.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No.51509061), the China Postdoctoral Science Foundation (Grant No. 2014M551249) and the HIT Environment and Ecology Innovation Special Funds (Grant No. HSCJ201607). Additional support was provided by the Southern University of Science and Technology (No. G01296001). We thank Professor Bellie Sivakumar for offering suggestions and comments on the statistical applications. We are also grateful for Editor Dr. Ashantha Goonetilleke and the anonymous reviewers for their generous feedback that improved the quality of the paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.scitotenv.2017.08.232.

References

Babovic, V., 2005. Data mining in hydrology. Hydrol. Process. 19, 1511–1515.

Bade, R., Bijlsma, L., Miller, T.H., Barron, L.P., Sancho, J.V., Hernández, F., 2015. Suspect screening of large numbers of emerging contaminants in environmental waters using artificial neural networks for chromatographic retention time prediction and high resolution mass spectrometry data analysis. Sci. Total Environ. 538, 934–941.

Bieroza, M.Z., Heathwaite, A.L., 2015. Seasonal variation in phosphorus concentration–discharge hysteresis inferred from high-frequency in situ monitoring. J. Hydrol. 524, 333–347.

Bowes, M.J., Loewenthal, M., Read, D.S., Hutchins, M.G., Prudhomme, C., Armstrong, L.K., Harman, S.A., Wickham, H.D., Gozzard, E., Carvalho, L., 2016. Identifying multiple stressor controls on phytoplankton dynamics in the river Thames (UK) using high-frequency water quality data. Sci. Total Environ. 569-570, 1489–1499.

Byrand, K., 2010. Nature and history in the Potomac country: from Hunter–Gatherers to the age of Jefferson. Ethnohistory 36 (3), 233–234.

Cassidy, R., Jordan, P., 2011. Limitations of instantaneous water quality sampling in surface-water catchments: comparison with near-continuous phosphorus time-series data. J. Hydrol. 405 (1–2), 182–193.

Christensen, V.G., 2001. Characterization of Surface-water Quality Based on Real-time Monitoring and Regression Analysis, Quivira National Wildlife Refuge, South-central Kansas, December 1998 Through June 2001. Center for Integrated Data Analytics Wisconsin Science Center.

Csábrági, A., Molnár, S., Tanos, P., Kovács, J., 2017. Application of artificial neural networks to the forecasting of dissolved oxygen content in the Hungarian section of the river Danube. Ecol. Eng. 100, 63–72.

Du, K., Zhao, Y., Lei, J., 2017. The incorrect usage of singular spectral analysis and discrete wavelet transform in hybrid models to predict hydrological time series. J. Hydrol. 552, 44–51.

Erkyihun, S.T., Rajagopalan, B., Zagona, E., Lall, U., Nowak, K., 2016. Wavelet-based time series bootstrap model for multidecadal streamflow simulation using climate indicators. Water Resour. Res. 52 (5), 4061–4077.

Hall, J., Zaffiro, A.D., Marx, R.B., Kefauver, P.C., Krishnan, E.R., Haught, R.C., Herrmann, J.G., 2007. On-line water quality parameters as indicators of distribution system contamination. J. Am. Water Works Assoc. 99 (1), 66–77.

Halliday, S.J., Skeffington, R.A., Wade, A.J., Bowes, M.J., Gozzard, E., Newman, J.R., 2015. High-frequency water quality monitoring in an urban catchment: hydrochemical dynamics, primary production and implications for the water framework directive. Hydrol. Process. 29, 3388–3407.

He, H.M., Hou, D.B., Zhao, H.F., Huang, P.J., Zhang, G.X., 2013. Multi-parameters fusion algorithm for detecting anomalous water quality. J. Zhejiang Univ. (Eng. Sci.) 47 (4), 735–740.

Helsel, D.R., 1992. In: Helsel, D.R., Hirsch, R.M. (Eds.), Statistical Methods in Water Resources. Elsevier, Amsterdam; New York.

Horsburgh, J.S., Jones, A.S., Stevens, D.K., Tarboton, D.G., Mesner, N.O., 2010. A sensor network for high frequency estimation of water quality constituent fluxes using surrogates. Environ. Model. Softw. 25 (9), 1031–1044.

Hou, D., Song, X., Zhang, G., Zhang, H., Loaiciga, H., 2013. An early warning and control system for urban, drinking water quality protection: China's experience. Environ. Sci. Pollut. Res. Int. 20 (7), 4496–4508.

Huang, W., Foo, S., 2002. Neural network modeling of salinity variation in Apalachicola river. Water Res. 36 (36), 356–362.

Jastram, J., 2014. Streamflow, water quality, and aquatic macroinvertebrates of selected streams in Fairfax county, Virginia, 2007–12. Scientific Investigations Report, pp. 2014–5073.

Jiang, J., Wang, P., Lung, W.S., Guo, L., Li, M., 2012. A GIS-based generic real-time risk assessment framework and decision tools for chemical spills in the river basin. J. Hazard. Mater. 227, 280–291.

Jones, A.S., Stevens, D.K., Horsburgh, J.S., Mesner, N.O., 2011. Surrogate measures for providing high frequency estimates of total suspended solids and total phosphorus concentrations 1. J. Am. Water Resour. Assoc. 47 (2), 239–253.

Jorjani, E., Asadollahi Poorali, H., Sam, A., Chehreh Chelgani, S., Mesroghli, S., Shayestehfar, M.R., 2009. Prediction of coal response to froth flotation based on coal analysis using regression and artificial neural network. Miner. Eng. 22 (11), 970–976.

Kirchner, J.W., Austin, C.M., Myers, A., Whyte, D.C., 2011. Quantifying remediation effectiveness under variable external forcing using contaminant rating curves. Environ. Sci. Technol. 45 (18), 7874–7881.

Kunz, J.V., Hensley, R., Brase, L., Borchardt, D., Rode, M., 2017. High frequency measurements of reach scale nitrogen uptake in a fourth order river with contrasting hydromorphology and variable water chemistry (Weibe Elster, Germany). Water Resour. Res. 53 (1), 328–343.

Kuo, J.T., Wang, Y.Y., Lung, W.S., 2006. A hybrid neural–genetic algorithm for reservoir water quality management. Water Res. 40 (7), 1367–1376.

Liu, J., Guo, L., Jiang, J., Hao, L., Liu, R., Wang, P., 2015. Evaluation and selection of emergency treatment technology based on dynamic fuzzy GRA method for chemical contingency spills. J. Hazard. Mater. 299, 306–315.

Mckenna, S.A., Klise, K.A., 2006. Multivariate applications for detecting anomalous water quality. Proceedings of the 8th Annual Water Distribution Systems Analysis Symposium. WDSA, Cincinnati, Ohio, USA:pp. 1–11 http://dx.doi.org/10.1061/40941(247)130 (accessed May 20, 2017).

Mckenna, S.A., Hart, D.B., Murray, R., Haxton, T., 2011. Testing and evaluation of water quality event detection algorithms. Handbook of Water and Wastewater Systems Protection, pp. 369–396.

Miguntanna, N.S., Egodawatta, P., Kokot, S., Goonetilleke, A., 2010. Determination of a set of surrogate parameters to assess urban stormwater quality. Sci. Total Environ. 408 (24), 6251–6259.

Parmar, K.S., Bhardwaj, R., 2014. Water quality management using statistical analysis and time-series prediction model. Appl Water Sci 4 (4), 425–434.

Parmar, K.S., Bhardwaj, R., 2015. River water prediction modeling using neural networks, fuzzy and wavelet coupled model. Water Resour. Manag. 29 (1), 17–33.

Rajaee, T., 2011. Wavelet and ANN combination model for prediction of daily suspended sediment load in rivers. Sci. Total Environ. 409 (15), 2917–2928.

Rode, M., Wade, A.J., Cohen, M.J., Hensley, R.T., Bowes, M.J., Kirchner, J.W., Arhonditsis, G.B., Jordan, P., Kronvang, B., Halliday, S.J., Skeffington, R.A., Rozemeijer, J.C., Aubert, A.H., Rinke, K., Jomaa, S., 2016. Sensors in the stream: the high-frequency wave of the present. Environ. Sci. Technol. 50 (19), 10297–10307.

Rügner, H., Schwientek, M., Beckingham, B., Kuch, B., Grathwohl, P., 2013. Turbidity as a proxy for total suspended solids (TSS) and particle facilitated pollutant transport in catchments. Environ. Earth Sci. 69 (2), 373–380.

Sannasiraj, S.A., Zhang, H., Babovic, V., Chan, E.S., 2004. Enhancing tidal prediction accuracy in a deterministic model using chaos theory. Adv. Water Resour. 27, 761–772.

Scott, A.B., Frost, P.C., 2017. Monitoring water quality in Toronto's urban stormwater ponds: assessing participation rates and data quality of water sampling by citizen scientists in the fresh water watch. Sci. Total Environ. 592, 738–744.

Shi, B., Jiang, J., Liu, R., Khan, A.U., Wang, P., 2017. Engineering risk assessment for emergency disposal projects of sudden water pollution incidents. Environ. Sci. Pollut. Res. 1–15.

Shupe, S.M., 2017. High resolution stream water quality assessment in the Vancouver, British Columbia region: a citizen science study. Sci. Total Environ. 603-604, 745–759.

Sun, Y., Babovic, V., Chan, E.S., 2010. Multi-step-ahead model error prediction using time-delay neural networks combined with chaos theory. J. Hydrol. 395, 109–116.

US EPA, 2013. Water Quality Event Detection System Challenge: Methodology and Findings. Office of Water ((MC-140) EPA 817-R-13-002).

US EPA, 2017. National hydrography dataset high-resolution flowline data. The national map. https://www.data.gov/, Accessed date: 20 May 2017.

Viviano, G., Salerno, F., Manfredi, E.C., Polesello, S., Valsecchi, S., Tartari, G., 2014. Surrogate measures for providing high frequency estimates of total phosphorus concentrations in urban watersheds. Water Res. 64, 265–277.

Wagner, R.J., Boulger, R.W., Oblinger, C.J., Smith, B.A., 2006. Guidelines and standard procedures for continuous water-quality monitors: station operation, record computation, and data reporting. J. Veg. Sci. 2 (3), 377–384.

Xi, D., Sun, Y., Liu, X., 2004. Environment Monitoring. Higher Education Press (in Chinese).

Yang, Jeffrey Y., Haught, R.C., Goodrich, J.A., 2009. Real-time contaminant detection and classification in a drinking water pipe using conventional water quality sensors: techniques and experimental results. J. Environ. Manag. 90 (8), 2494–2506.

Zhai, X., Xia, J., Zhang, Y., 2014. Water quality variation in the highly disturbed Huai river basin, China from 1994 to 2005 by multi-statistical analyses. Sci. Total Environ. 496, 594–606.