Environmental Modelling & Software xxx (2017) 1–14

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

A generic framework to analyse the spatiotemporal variations of water quality data on a catchment scale

Qinli Yang [a, b], Miklas Scholz [c, d, e], Junming Shao [b, f, *], Guoqing Wang [g], Xiaofang Liu [h]

[a] School of Resources and Environment, University of Electronic Science and Technology of China, No. 2006, Xiyuan Avenue, West Hi-Tech Zone, Chengdu 611731, China
[b] Big Data Research Center, University of Electronic Science and Technology of China, No. 2006, Xiyuan Avenue, West Hi-Tech Zone, Chengdu 611731, China
[c] Division of Water Resources Engineering, Faculty of Engineering, Lund University, P.O. Box 118, 22100 Lund, Sweden
[d] Civil Engineering Research Group, School of Computing, Science and Engineering, The University of Salford, Newton Building, Greater Manchester M5 4WT, UK
[e] Department of Civil Engineering Science, School of Civil Engineering and the Built Environment, University of Johannesburg, Kingsway Campus, PO Box 524, Aukland Park 2006, Johannesburg, South Africa
[f] School of Computer Science and Engineering, University of Electronic Science and Technology of China, No. 2006, Xiyuan Avenue, West Hi-Tech Zone, Chengdu 611731, China
[g] State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, Nanjing Hydraulic Research Institute, Nanjing 210029, China
[h] Sichuan Provincial Academician (Expert) Workstation, Sichuan University of Science and Engineering, Zigong 643000, China

ARTICLE INFO

Article history:
Received 21 September 2016
Received in revised form
20 June 2017
Accepted 8 November 2017
Available online xxx

Keywords:
Spatiotemporal analysis
Environmental data
Cluster analysis
Dynamic time warping

ABSTRACT

Most spatiotemporal studies treat spatial and temporal analysis separately. However, spatial and temporal changes occur simultaneously and are correlated. In this study, we propose a generic framework to simultaneously analyse the spatial and temporal variations of water quality on a catchment scale. Specifically, we analyse the heterogeneity of temporal evolution of water quality data among different sampling sites, and the heterogeneity of spatial distribution of water quality data over different sampling times, respectively, by integrating the techniques of normalized mutual information, dynamic time wrapping and cluster analysis. To bring deep insight into the spatiotemporal variations, inter-change and intra-change are further defined and distinguished, respectively. Taking the Fuxi River catchment as a case study, results indicate that the proposed framework is intuitive and efficient. Beyond this, the generic framework can be expanded for other catchments and various environmental data.

Software availability

Name of software: Spatiotemporal Analysis on Environmental Data (SAE)
Developers: Junming Shao, Qinli Yang
Contact: Junming Shao; Address: No. 2006, Xiyuan Avenue, West Hi-Tech Zone, Chengdu 611731, China; Email: junmshao@uestc.edu.cn
Year first available: 2016
Required hardware and software: SAE works on Matlab on Windows, Linux based computers
Cost: Free. SAE software package is public available at http://dm.uestc.edu.cn/sae

1. Introduction

1.1. Challenges of water quality data variation analysis

The water quality of a river can be attributed to both natural processes and anthropogenic activities including surface runoff (Abbaspour et al., 2007), climate change (Whitehead et al., 2009), geological structure (Bilgin and Konanç, 2016), land use (Sliva and Williams, 2001; Ding et al., 2015) and sewage discharge (Zhen and Zhu, 2016). Different atmospheric inputs, climatic conditions and anthropogenic inputs may result in spatial variation of water quality in rivers (Bricker and Jones, 1995). On the other hand,

* Corresponding author. Big Data Research Center, University of Electronic Science and Technology of China, No. 2006, Xiyuan Avenue, West Hi-Tech Zone, Chengdu 611731, China.
E-mail addresses: qinli.yang@uestc.edu.cn (Q. Yang), miklas.scholz@tvrl.lth.se, m.scholz@salford.ac.uk (M. Scholz), Junmshao@uestc.edu.cn (J. Shao), gqwang@nhri.cn (G. Wang), lxf1969@163.com (X. Liu).

https://doi.org/10.1016/j.envsoft.2017.11.003
1364-8152/

seasonal variation in precipitation may cause river discharge variations and subsequently affect the concentration of pollutants in river water (Vega et al., 1998). Therefore, a good understanding of temporal and spatial variations of water quality on a catchment scale is of great importance for water pollution control, aquatic ecosystem restoration and water management by regional communities (Bu et al., 2010; Iscen et al., 2008).

The methods for temporal and/or spatial analysis on water quality data have developed from univariate analysis to multivariate analysis. In the early stage, spatial and/or temporal analysis of water quality data mainly focused on a single argument (Laznik et al., 1999; Niu et al., 2004). This kind of analysis method cannot fit well to more and more complex water challenges. With the development of computer science, multivariate statistical analysis methods have grown, mainly including principal component analysis (PCA), principal factor analysis (PFA), factor analysis (FA), cluster analysis (CA) and discriminant analysis (DA) (Singh et al., 2004; Smeti and Golfinopoulos, 2016). These multivariate statistical techniques have been well-accepted and widely applied in water quality assessments (Shrestha and Kazama, 2007; Iscen et al., 2008; Wang et al., 2013), particularly for key parameter extraction and main pollution source identification.

For instance, applying CA and FA, Bu et al. (2010) grouped 12 sampling sites into three pollution level clusters (no pollution, moderate pollution and high pollution) and identified five factors of pollution sources for the Jinshui River of the South Qinling Mountains in China. Wang et al. (2013) applied CA and PCA/FA to evaluate temporal/spatial variations in water quality and identify latent sources of water pollution in the Songhua River Harbin region, China. Smeti and Golfinopoulos (2016) applied DA on surface water quality data of Yliki Lake to determine which variables were the most efficient in discriminating between clusters.

However, most existing spatiotemporal variation analytical works treat the temporal and spatial analysis of water quality data separately. For instance, Ouyang et al. (2006) assessed the seasonal variations in surface water quality in the lower St. Johns River. Wang et al. (2013) performed the temporal and spatial cluster analysis of water quality in the Songhua River Harbin region separately, classifying monitoring time into three periods and classifying monitoring stations into three groups. Chang (2008) conducted spatial analysis of water quality trends in the Han River basin, South Korea. However, the co-occurrence and correlation of temporal and spatial variations were not thoroughly considered during the previous spatiotemporal variation assessment.

Therefore, an approach, which allows to comprehensively analyse the temporal and spatial variations of water quality data simultaneously, is highly needed. Especially, in the context of a changing environment, it is essential to study the spatiotemporal variations of water quality data, so as to adapt to climate change and environment deterioration.

1.2. Aims and objectives

This paper aims to propose a generic framework to simultaneously analyse the spatiotemporal variations of water quality data on a catchment scale. The importance of this paper lies in teaching environmental engineers and scientists without a computational background to systemically analyse temporal and spatial variations of water quality data in a more effective, intuitive and easier way.

The objectives of this paper are as follows.

1) to analyse how the spatial distributions of water quality data change over time;
2) to analyse the spatial heterogeneity of temporal evolution of water quality data; and
3) to demonstrate the procedure and verify the effectiveness of the proposed framework, taking a water quality dataset in Fuxi River catchment as a typical example.

2. Study area and data acquisition

2.1. Study area

To demonstrate the proposed framework, any catchment could have been chosen for illustration. Here, the Fuxi River catchment is used as a case study as the authors are familiar with this example catchment and reliable data have been collected by trusted sources.

The Fuxi River catchment (28°58′-29°46′ N, 103°43′-105°36′ E, Fig. 1) covers 3490 km^2 of drainage area with a mainstream (Fuxi River) length of 73.2 km, a bending coefficient of 2.21 and an average gradient of 0.27‰. According to the recorded hydrological data, the natural multiple annual average runoff in the Fuxi River catchment is 42.25 m^3/s and the measured annual average runoff is 19.26 m^3/s. The meteorological data indicate that the average annual precipitation is 1023 mm (58% in July and August), and the mean temperature is 17.8 °C. The altitude ranges from 250 m to 500 m. The land use in the catchment is dominated by farmland (about 70%), which contributes to non-point pollution of Fuxi River.

The Fuxi River is one of the first-level tributaries of the Tuojiang River, which is located in the southwest of China. It originates from the Xushui River (length of 118 km; source in the west) and the Weiyuan River (length of 123 km, source in the north), joining as the mainstream at Shuanghekou (Fig. 1, Site 4). Further downstream, the Tieqian River and the Zhenxi River join the Fuxi River, which finally flows into the Tuojiang River. The locations of seven sampling sites are presented in Table 1. Site 1 is located on Weiyuan River, sites 2 and 3 are by the Xushui River, and the other sites are selected on the mainstream of Fuxi River (Fig. 1).

The Fuxi River is the only watercourse flowing across Zigong city, and is regarded as the ‘mother river’ by local citizens. Zigong city is one of 50 cities facing serious water shortages in China, with a population of 3.30 million in 2013 and per capita water resources of 585 m^3, accounting for 18.84% of the provincial level (Sichuan province). In recent years, with the rapid economic development and population growth, the water quality in the Fuxi River has degraded considerably. The degradation grades differ at different times (e.g., eutrophication often shows up during the dry season) and different locations (e.g., water quality is worse downstream). Water quality in the catchment exhibits temporal and spatial variations, and makes water pollution control challenging.

2.2. Data acquisition

Water quality data in the study area were provided by the Environmental Protection Agency of Zigong City, China. The dataset consists of ten water quality variables at seven sampling stations with monthly records from January 2012 to December 2014. A basic summary of the water quality data is presented in Table 2. Here, the authors use it as a representative dataset of water quality measurements on a catchment scale to illustrate the proposed new method. For extended applications, the studied variables, the number of variables and the number of sampling stations could be different depends on specific practices.

Please cite this article in press as: Yang, Q., et al., A generic framework to analyse the spatiotemporal variations of water quality data on a catchment scale, Environmental Modelling & Software (2017), https://doi.org/10.1016/j.envsoft.2017.11.003

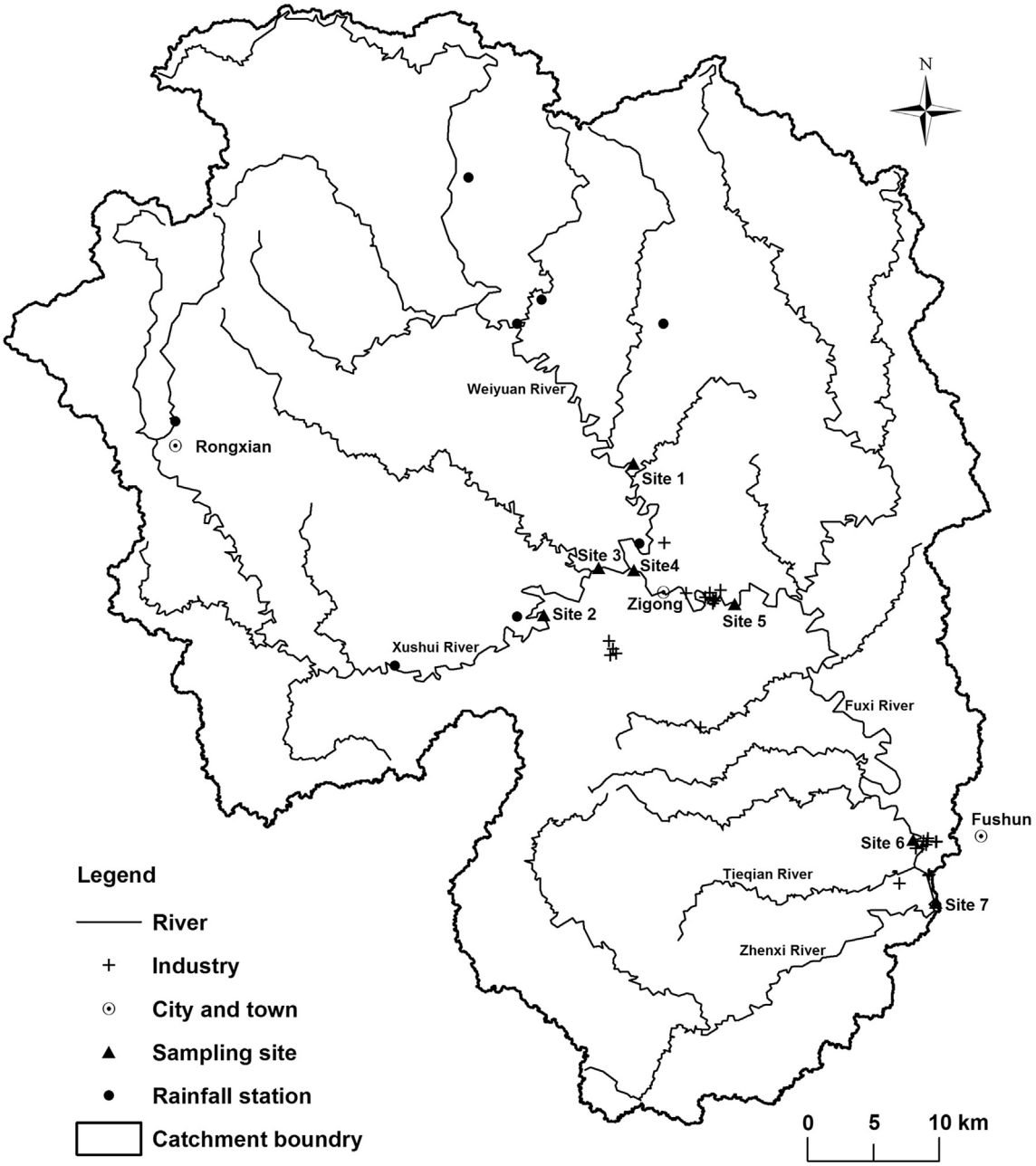


Fig. 1. The Fuxi River catchment (located in Sichuan province, China) and locations of seven sampling sites, rainfall stations and main industries.

Table 1
Basic information regarding the sampling sites in the Fuxi River catchment.

Site number	Site Name	Latitude (degree)	Longitude (degree)	Location
Site 1	Liaojiayan	29.438	104.746	Cross-section of Neijiang city and Zigong city
Site 2	Changtuhe	29.334	104.685	Reserve point for Changtu drinking water factory
Site 3	Leigongtan	29.367	104.722	1000 m up the junction of Weiyuan River to Fuxi River
Site 4	Shuanghekou	29.365	104.746	1000 m down the junction of Weiyuan River and Xushui River
Site 5	Tanyansuo	29.342	104.815	1000 m down the Fuxi River flowing out of Zigong city
Site 6	Dengguan	29.180	104.947	5000 m up the junction of Fuxi River to Tuojiang River
Site 7	Rutuobakou	29.138	104.953	500 m up the junction of Fuxi River to Tuojiang River

3. Framework and the fundamental techniques

3.1. Overview of the framework

An overview of the proposed framework to analyse the spatiotemporal variation of water quality data is presented in Fig. 2. In general, the framework consists of two routes. The first route is to analyse the heterogeneity of temporal evolution of water quality among different sampling sites. Namely, how the water quality of each sampling site evolves over time, and is the temporal evolution

Please cite this article in press as: Yang, Q., et al., A generic framework to analyse the spatiotemporal variations of water quality data on a catchment scale, Environmental Modelling & Software (2017), https://doi.org/10.1016/j.envsoft.2017.11.003

Table 2
Basic statistics of the water quality data (2012–2014) in the Fuxi River catchment.

Variable	Abbreviation	Unit	Min	Mean ± Stdev	Max
Temperature	Temp	°C	5.80	19.59 ± 6.43	31.00
pH	pH	–	6.84	7.82 ± 0.30	8.89
Dissolved Oxygen	DO	mg/L	0.20	5.36 ± 1.77	10.20
Chemical Oxygen Demand	COD_{Mn}	mg/L	2.70	6.74 ± 1.90	13.10
Biological Oxygen Demand	BOD_5	mg/L	0.80	4.99 ± 3.14	21.40
Ammonia-nitrogen	NH_4-N	mg/L	0.20	3.26 ± 4.22	37.20
Petroleum	Petrol	mg/L	0.01	0.05 ± 0.02	0.22
Volatile phenol	VP	mg/L	0.00	0.00 ± 0.00	0.01
Fluoride	F	mg/L	0.25	1.36 ± 1.29	7.83
Electrical conductivity	EC	μS/cm	29.90	101.61 ± 57.2	306.00

Sampling number: 252.

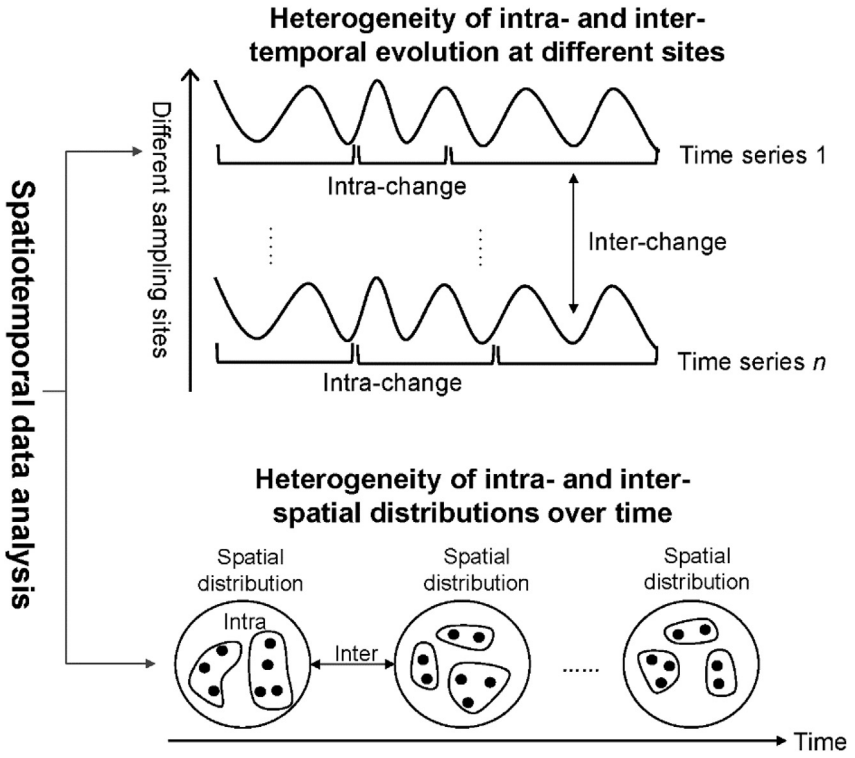


Fig. 2. The generic framework for spatiotemporal variation analysis of water quality data at a catchment level.

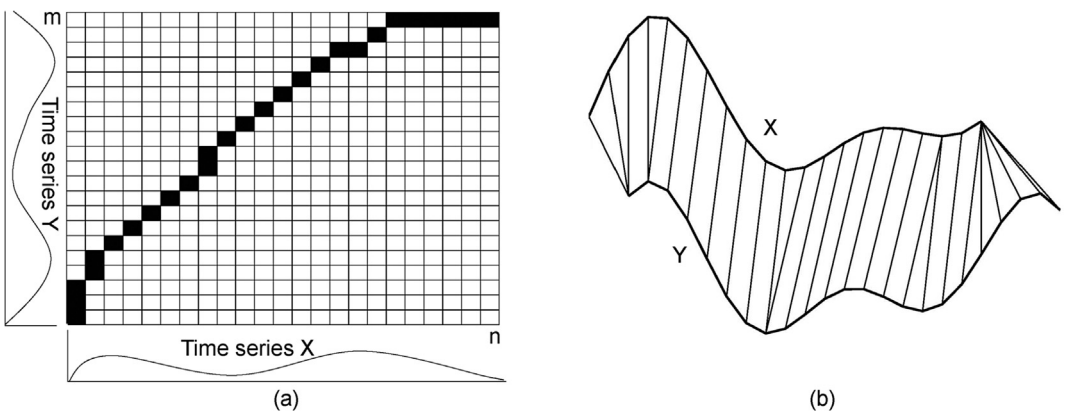


Fig. 3. Illustration of dynamic time warping. (a) Cost matrix, (b) Optimal warping path for time series.

Please cite this article in press as: Yang, Q., et al., A generic framework to analyse the spatiotemporal variations of water quality data on a catchment scale, Environmental Modelling & Software (2017), https://doi.org/10.1016/j.envsoft.2017.11.003

of water quality at different sampling sites similar or not? The second route focuses on the heterogeneity of spatial distribution of water quality over time. As the water quality is usually spatially distributed (i.e., different sampling sites show different levels of water quality), the second route aims to answer the question: Does the spatial distribution of water quality evolves over time and how does this happen?

Beyond this, to systematically explore the spatiotemporal variations, inter-change and intra-change of each route are further defined, respectively. Specifically, inter-change of the first route focuses on the temporal variation of water quality among different sampling sites. For intra-change, it characterizes the differences of inner temporal clusters generated from each sampling site. Similarly, the inter-change of the second route refers to the difference of spatial distribution of water quality among different sampling times, while the intra-change means the difference among inner spatial clusters generated at each sampling time.

To capture the similarities of the inter- and intra-changes of water quality data, normalized mutual information (NMI) and dynamic time warping (DTW) were integrated. More detailed descriptions of how to use the framework can be found in the following subsections 3.2 to 3.4.

3.2. Analysis on heterogeneity of temporal evolution at different sites

For a catchment, water quality often varies with time due to climatic or/and anthropic factors. Furthermore, different sites generally show different temporal variations of water quality. To study the heterogeneity of temporal evolution of water quality at different sites, two perspectives are considered: One is from the inter-change view, which aims to investigate which sampling sites have similar temporal evolution. Another is from the intra-change view, which aims to study which sampling sites exhibit similar change of inner temporal pattern.

3.2.1. Inter-change analysis

To investigate the heterogeneity of temporal evolution of water quality among different sampling sites at the inter-change level, the following steps are involved:

(1) Dynamic time warping (DTW) is first extended to calculate the pair-wise dissimilarity of the temporal evolution of water quality among different sampling sites. Specifically, the water quality data of each sampling site is characterized as a multi-variable time series.

For instance, assume that a given water quality data set consists of k variables with w sampling sites ($k = 10$ and $w = 7$ in our case study). Since the data of each variable are collected over time for a given sampling site, we can use k time series to represent its data, where each time series captures the values of one variable over time. With the time series representation, the dissimilarity between any two sampling sites for each water quality variable is calculated based on dynamic time warping, and then the derived dissimilarities of all variables are averaged to quantify the dissimilarity of water quality between any two sampling sites. The description of DTW will be introduced in Section 3.4.1.

(2) Based on the dissimilarity matrix, cluster analysis is employed to partition of all sampling sites into different groups, where the sites with similar temporal evolution characteristics tend to stay in the same group.

3.2.2. Intra-change analysis

To investigate the heterogeneity of temporal evolution of water quality among different sampling sites at the intra-change level, three steps are needed:

(1) First, we investigate the temporal cluster on each sampling site via cluster analysis. Namely, since the water quality of each sampling site evolves over time, the time steps of similar water quality are identified. For example, given the water quality of a sampling site i in a certain period (e.g., one year), the cluster analysis is applied to investigate in which months the water quality is similar.
(2) Next, normalized mutual information is used to quantify the similarities of the derived temporal clusters at different sampling sites; and
(3) Finally, based on the similarity, cluster analysis is used again to group the sampling sites into different clusters. The sites having similar inner temporal pattern will fall into the same cluster.

3.3. Analysis on heterogeneity of spatial distributions over time

The spatial distribution of water quality within a catchment may differ over time, therefore it is essential to investigate what changes have happened. In order to analyse the heterogeneity of temporal evolution at different sites (section 3.2) and to study the heterogeneity of spatial destructions over time, two aspects are also investigated: inter-change and intra-change. For inter-change analysis, the objective is to investigate the change of spatial distributions at different sampling times. As for intra-change, the aim is to study the change of inner spatial patterns at different sampling times.

3.3.1. Inter-change analysis

To investigate the heterogeneity of spatial distributions at different sampling times at the inter-change level, the following two steps are performed:

(1) We regard the water quality data at each sampling time as a multi-instance, and then use Euclidean distance to calculate the dissimilarity of spatial distribution of water quality among different sampling times.

For instance, given water quality data with k variables sampling at w sites over m time points ($k = 10$, $m = 36$ and $w = 7$ in our case study), the multi-instance consists of w instances, and each instance is a k-dimensional feature vector. The dissimilarity of spatial distribution of water quality between any two sampling times are thus transformed to compute the distance between the corresponding two multi-instances. Specifically, the Euclidean distance for each instance is calculated and then distances for all w instances are averaged to obtain the dissimilarity between two multi-instances.

(2) Building upon the dissimilarity matrix, we cluster all sampling times into different groups. The sampling times, which have similar spatial distributions of water quality, tend to group together.

Please cite this article in press as: Yang, Q., et al., A generic framework to analyse the spatiotemporal variations of water quality data on a catchment scale, Environmental Modelling & Software (2017), https://doi.org/10.1016/j.envsoft.2017.11.003

3.3.2. Intra-change analysis

To investigate the heterogeneity of temporal evolution of water quality among different sampling sites at the intra-change level, three steps are needed:

(1) First, we investigate the temporal cluster on each sampling site via cluster analysis. Namely, since the water quality of each sampling site evolves over time, the time steps of similar water quality are identified. For example, given the water quality of a sampling site *i* in a certain period (e.g., one year), the cluster analysis is applied to investigate in which months the water quality is similar.
(2) Normalized mutual information is used to quantify the similarities of the derived temporal clusters at different sampling sites.
(3) Finally, based on the similarity matrix, cluster analysis is used again to group the sampling sites into different clusters. Therefore, the sampling sites having similar inner temporal pattern will fall into the same cluster.

3.4. Fundamental techniques

3.4.1. Dynamic time warping (DTW)

Dynamic time warping is a technique that identifies the optimal alignment of two time series. To achieve this goal, the time series are "warped" together non-linearly by stretching or shrinking them along the time axes (Salvador and Chan, 2007). The DTW technique provides a good solution known for time series challenges in many domains, including environment science (Jha and Datta, 2014), medicine (Pan and Li, 2016; Shao et al., 2010a; 2010b), engineering (Kim et al., 2016) and entertainment (Kostoulas et al., 2015).

Suppose we have two time series X and Y (equations (1) and (2)), of lengths m and n, respectively (Here $m = n = 36$ in our case study, Fig. 3), where

$$X = (x_1, x_2, \ldots, x_i \ldots, x_m) \tag{1}$$

$$Y = \left(y_1, y_2, \ldots, y_j \ldots, y_n\right) \tag{2}$$

The objective is to optimise a warping path W (equation (3)):

$$W = (w_1, w_2, \ldots, w_i, \ldots, w_K) \tag{3}$$

where K is the length of W, with $max(m, n) < K < m + n - 1$. The kth element of W is a pair of indices indicating a connection of time points in X and Y, and is written as $w_k = (i,j)$. A warping path follows these constraints (Keogh and Ratanamahatana, 2005):

1) Boundary conditions: $w_1 = (1, 1)$ and $w_K = (m, n)$. This requires the warping path to start and finish in the first and last points of the series, respectively;
2) Monotony: Given $w_k = (i, j)$, then $w_{k+1} = (i', j')$, with $i' - i >= 0$ and $j' - j >= 0$. This forces the points in W to be monotonically spaced in time;
3) Continuity: Given $w_k = (i, j)$, then $w_{k+1} = (i', j')$, with $i' - i \leq 1$ and $j' - j \leq 1$. This restricts the admissible steps in the warping path to adjacent points of the series.

There are many warping paths satisfying the above conditions. In order to find a best match between two time series, it looks for that path which minimizes the cumulative distance between them. The distance DTW for this optimum path is defined in equation (4).

$$DTW(X, Y) = \min\left(\sum_{i=1}^{K} d(w_i)\right) \tag{4}$$

where $d(\cdot)$ is a distance function. We use the Euclidean distance in this study.

The optimum warping path for *DTW(X,Y)* can be obtained through a dynamical programming approach (Itakura, 1975). It proceeds as follows: First, a m by n cost matrix D is constructed (Fig. 3 (a)). A component $D(i, j)$ is defined recursively as sum of the distance $d(i, j)$ and the minimum of the cumulative distances in the adjacent elements (equation (5)).

$$D(i, j) = d(i, j) + min\{D(i-1, j-1), D(i-1, j), D(i, j-1)\} \tag{5}$$

3.4.2. Cluster analysis

Cluster analysis is an unsupervised multi-variate technique used to uncover the grouping structure of a given data set (Vega et al., 1998). Among the clustering approaches, agglomerative hierarchical clustering gains wide popularity, since it is simple, intuitive and easy to interpret (Jain et al., 1999). Therefore, we chose it to analyse the water quality data in this example case study.

Agglomerative hierarchical clustering forms clusters sequentially by starting with the most similar pair of objects and merge higher clusters step-by-step (Steinbach et al., 2000). To determine which clusters should be merged, it requires a measure of similarity between sets of observations and the linkage criterion, which is a function of the pair-wise distances of observations. The similarity of the clusters is determined by the corresponding measure of distance (e.g., Euclidean distance) between pairs of observations. For the merging criterion, several alternatives have been proposed, the most well-known of which are Single Link, Average Link, Complete Link and Ward's Link (Olson, 1995). Various partitioning clusters can be obtained by cutting the hierarchy at various desired levels. The higher the level, the coarser the clustering is, and the smaller is the number of clusters.

Ward's link (Ward, 1963) has been applied in this study. The method works in a manner that the pair of clusters whose combination results in minimum increase in "information loss" are merged. This happens on every possible pair of clusters at each stage of the hierarchical clustering. "Information loss" is defined by Ward's link in terms of an error sum of squares (ESS) criterion (equation (6)).

$$\mathrm{ESS} = \frac{N_A N_B}{N_A + N_B} \overline{X}_A - \overline{X}_B{}^2 \tag{6}$$

Where *EES* is the error sum of squares, $\overline{X}_A$ and $\overline{X}_B$ are the centroids of clusters A and B, $\overline{X}_A - \overline{X}_B$ is the Euclidean distance between the clusters A and B, and N_A and N_B are the number of objects in clusters A and B.

3.4.3. Normalized mutual information (NMI)

Mutual information, which is a symmetric measure to quantify the statistical information shared between two distributions (Cover and Thomas, 1991), provides a sound indication of the shared information between a pair of clustering.

The mutual information of two discrete random variables X and Y is defined as

Please cite this article in press as: Yang, Q., et al., A generic framework to analyse the spatiotemporal variations of water quality data on a catchment scale, Environmental Modelling & Software (2017), https://doi.org/10.1016/j.envsoft.2017.11.003

$$I(X,Y) = \sum_{y\in Y}\sum_{x\in X} p(x,y)log_2\left(\frac{p(x,y)}{p(x)p(y)}\right) \quad (7)$$

where $p(x)$ and $p(y)$ are the probability functions of X and Y; and $p(x,y)$ is the joint probability function of variables of X and Y.

As $I(X,Y)$ has no upper bound, for easier interpretation and comparisons, a normalized mutual information ranging from 0 to 1 is defined in equations (8)–(10) (Witten and Frank, 2005).

$$NMI(X,Y) = \frac{2I(X,Y)}{H(X)+H(Y)} \quad (8)$$

$$H(X) = -\sum_{x\in X}\int p(x)log_2 p(x) \quad (9)$$

$$H(Y) = -\sum_{y\in Y}\int p(y)log_2 p(y) \quad (10)$$

$H(X)$ and $H(Y)$ indicate the information entropy of X and Y, respectively.

4. Results and discussion

4.1. Overview

This section presents the spatiotemporal analysis results of water quality by applying the proposed generic framework on Fuxi River catchment. The results include two parts: the results of spatial heterogeneity of temporal evaluation of water quality and the results of distinct changes of spatial distribution over time. In addition, potential reasons behind the results have been discussed by integrating the meteorological environment, agriculture and human activities in the studied catchment.

4.2. Spatial heterogeneity of temporal evolution of water quality

4.2.1. Inter-change assessment

Regarding the data of each water quality variable at one sampling site as a time series, extended DTW is used to quantify the similarities of multi-variable time series among different sampling sites. Fig. 4 (a) shows the dissimilarities of seven sampling sites in terms of the temporal evolution of water quality based on DTW. From the dissimilarity matrix, we can observe that the sites 1 and 4 are most similar (with the lowest distance). The rational is that site 1 and site 4 are nearly located on the same branch of river. Furthermore, the industrial water pollution sources are rare around site 1 and site 4.

Fig. 4 (b) shows the dendrogram of cluster analysis based on the calculated similarities. With a given split indicated by the dashed line, the 7 sites can be grouped into 2 clusters, where the water quality at sites 1, 2, 4 and 7 (Cluster 1) have similar temporal evolution patterns, while the water quality at the other sites 3, 5 and 6 (Cluster 2) exhibit similar temporal variation characteristics. Cluster 2 is consistent with the distribution of those industries which are intensively monitored (see Fig. 1). This can be explained by industrial factories located near sites 3, 5 and 6 discharging water pollutants in a certain pattern (e.g., time, amount) into the river and thus worsening the water quality. In the long term, to improve the water quality of Fuxi river, Zigong government may consider industrial discharge control.

To further validate our results, the spatial variations of 10 water quality variables in the Fuxi River catchment are presented in Fig. 5 by box-whiskers plots showing the median, 25th and 75th percentiles. The bottom and top of the box are the first and third quartiles, and the band inside the box is the median. The "whisker" above and below the box represents the maximum and minimum of all data. The cross ("+") represents outliers with three times the interquartile range (3IQR). The box-whiskers provide a statistical summary of each water quality variable of each individual sampling site, where the median indicates the averaging value, and the bottom and top of the box characterize the variation. By first examining the medians of water quality variables, we can see that the sites 3, 5 and 6 have higher concentrations of BOD_5 than the others. Similarly, the sites 3 and 5 show highest concentrations of COD, while sites 5 and 6 also exhibit higher concentrations of NH_4-N compared with other sites. Therefore, water quality at sites 3, 5 and 6 are similar (from statistic perspective) and worse than those of sites 1, 2, 4 and 7. That is because more industries discharge pollutants into the river, causing the water quality to deteriorate. In addition, by examining the variance of the water quality variables, we can observe that the sampling sites 3, 5 and 6 are similar in most cases. The similarity of median and variance gives a hint that the three sites may show a similar temporal evolution of water quality.

In summary, in the Fuxi River catchment, sites 3, 5 and 6 show similar temporal variation of water quality, mainly due to industrial wastewater discharge. In contrast, sites 1, 2, 4 and 7 are grouped together as they indicate better water quality.

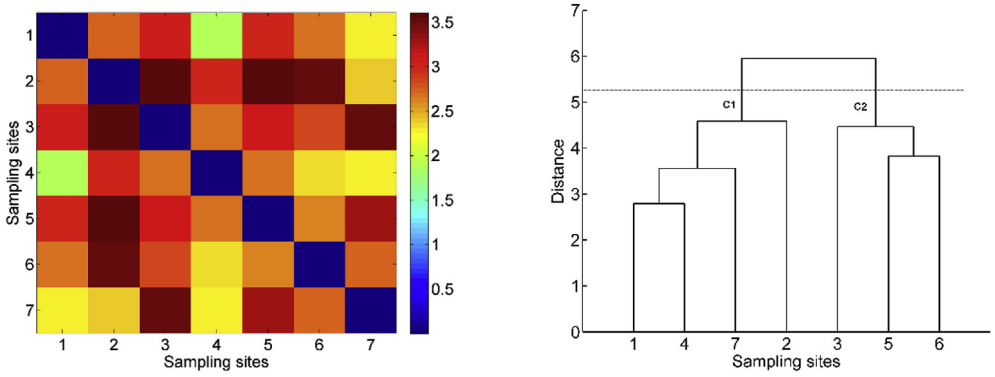


Fig. 4. The inter-change of temporal evolution of water quality among different sampling sites. (a) The derived distance matrix among different sampling sites based on multi-variable DTW; (b) The visualized cluster structure with dendrogram, where the dashed line is the user specified split.

Please cite this article in press as: Yang, Q., et al., A generic framework to analyse the spatiotemporal variations of water quality data on a catchment scale, Environmental Modelling & Software (2017), https://doi.org/10.1016/j.envsoft.2017.11.003

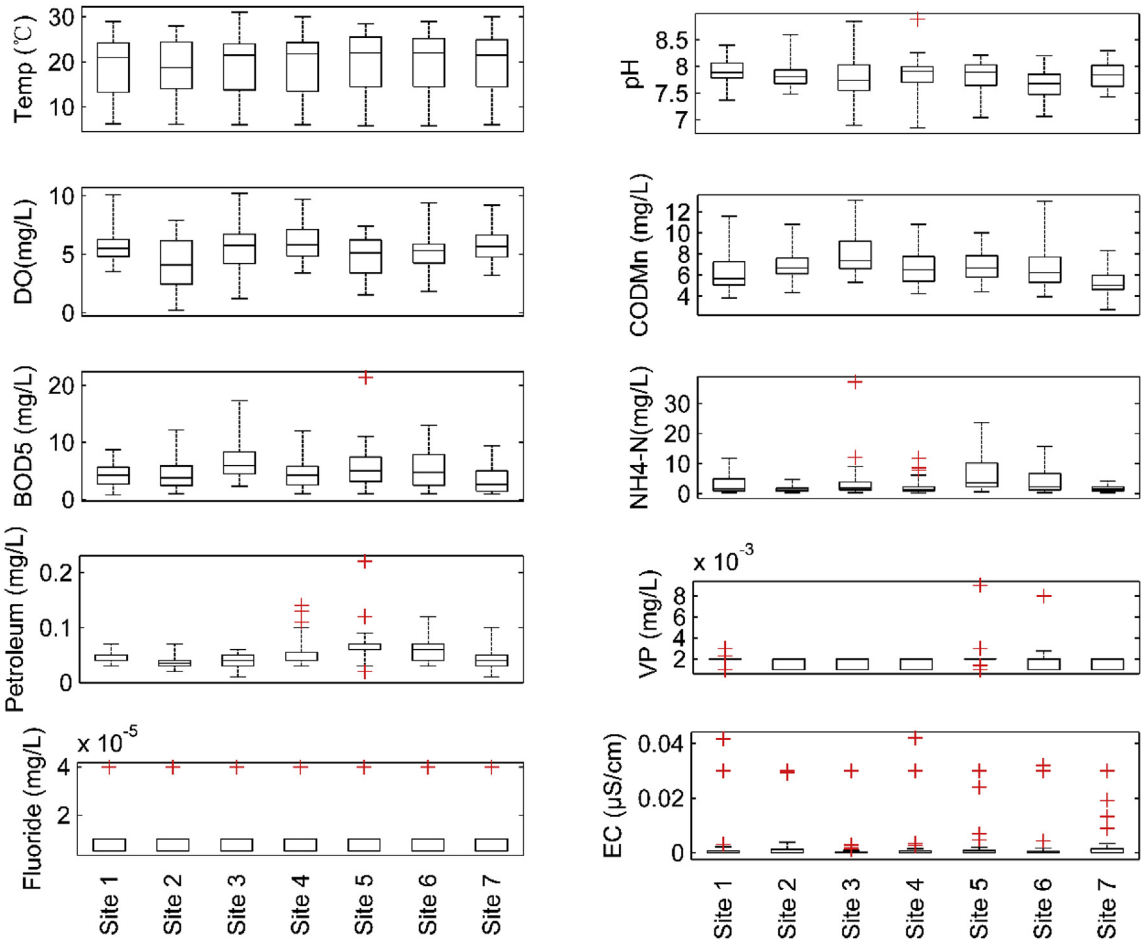


Fig. 5. Spatial variation of 10 water quality variables in Fuxi River catchment.

4.2.2. Intra-change assessment

To investigate the intra-change of water quality's temporal variations, cluster analysis was applied on the time series data of each site. Fig. 6 plots the corresponding temporal clusters over the period between 2012 and 2014, and over each year. Specifically, the sub-figures on the left column of Fig. 6 show the temporal clusters over the whole period of 2012–2014, and the subfigures on the right three columns demonstrate the specific temporal clusters in each individual years. Not surprisingly, different sampling sites exhibit different temporal clustering results for the annual scale and three-year period; i.e. the spatial heterogeneity of temporal distribution patterns.

On an annual scale, water quality at most sites are clustered into two groups, which roughly correspond to wet seasons (generally from December to next May) and dry seasons (generally from June to November). This is in line with the uneven distributions of precipitation and runoff in the catchment, which implies that water quality's temporal variations are largely impacted by seasons. The rationale is that in wet seasons, more precipitation and runoff would dilute pollutants and thus improve water quality.

In contrast, during dry seasons, less rainfall and runoff will increase the concentration of the contaminants and thus lead to worse water quality. Specifically, sites 1 and 3 obtained the same temporal cluster results (Jan.–May, Dec.; Jun.–Nov.) in 2012. Sites 1, 3 and 5 obtained the same cluster results (Jan.–May; Jun.–Dec.) in 2013. Sites 3 and 6, sites 2 and 5, and sites 1 and 7 formed the same cluster results, respectively, in 2014. Here, sites which achieve the same temporal cluster results mean they have similar temporal variation patterns in the specific year. The differences of temporal clusters among different sampling sites might be due to the fact that the runoff is interfered with by humans, such as setting-up sluices on the river for irrigation purposes.

More broadly, looking at the circumstances from the three-year period, different sites show various temporal clusters of water quality. In general, water quality at each sampling site can be roughly grouped into 2 clusters (from Jun. to Oct.; from Nov. to May), corresponding to two hydrologic levels (high flow period and low flow period). Time steps with similar precipitation and runoff tend to cluster together. This can be further verified by the changes of precipitation and runoff during 2012 and 2014 (Fig. 7).

To quantify the similarity of these temporal clusters generated from each sampling site, NMI is introduced. The similarities of temporal clusters among different sampling sites are shown in Fig. 8 (a). Afterwards, based on the calculated NMI, cluster analysis was reapplied, and the dendrogram is presented in Fig. 8 (b). The dendrogram can be interpreted on two levels: On a coarse level, seven sites are grouped into two clusters (C1 and C2), where site 7 is isolated from all the other sites, which can be seen in the similarity matrix (Fig. 8 (a)). This implies that the water quality at site 7 has a different temporal structural pattern relative to those of the other sites. The reason might be that site 7 is located at the outlet of the catchment, the quality and quantity of water at other places would lead to water quality changes at site 7. Under the complicated impacts of both climate and human activities in the

Please cite this article in press as: Yang, Q., et al., A generic framework to analyse the spatiotemporal variations of water quality data on a catchment scale, Environmental Modelling & Software (2017), https://doi.org/10.1016/j.envsoft.2017.11.003

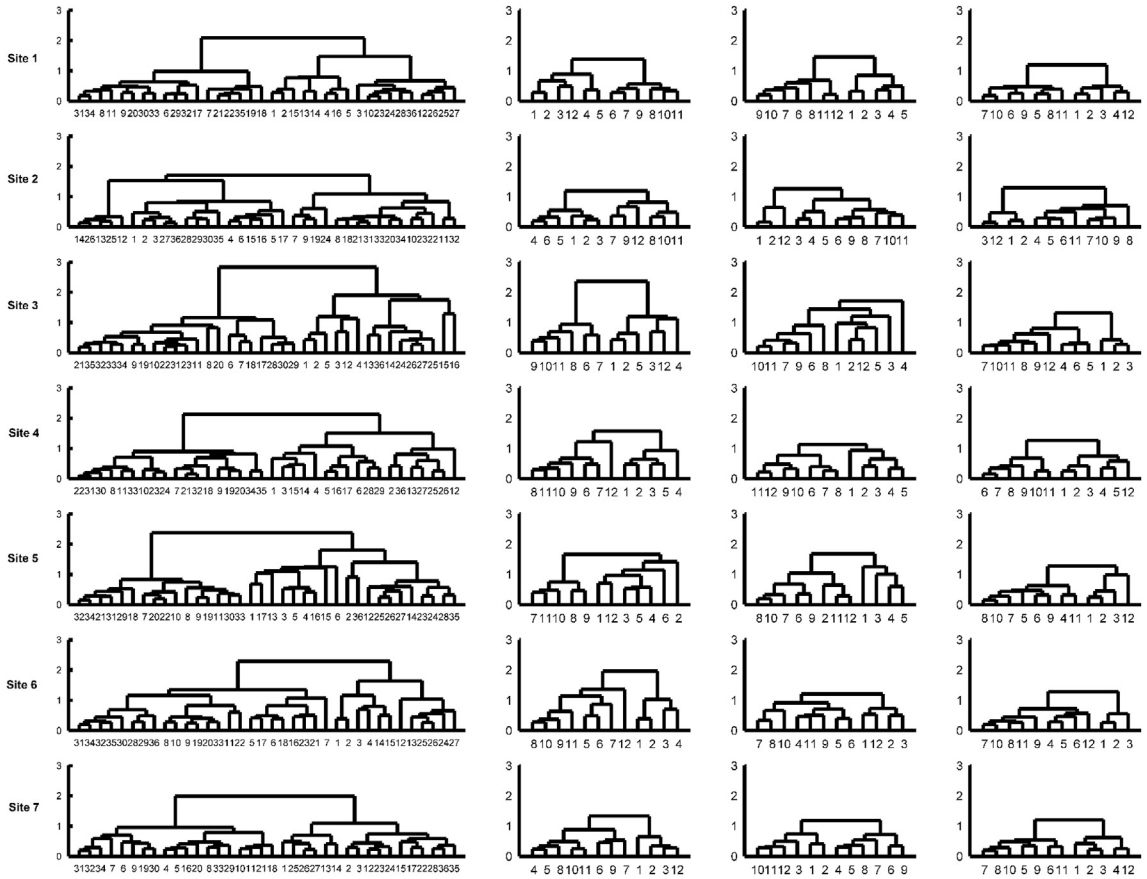


Fig. 6. Dendrograms of temporal clusters of water quality at different sampling sites for the period 2012 to 2014 (months from Jan. 2012 to Dec. 2014 were numbered from 1 to 36), and over each year, respectively.

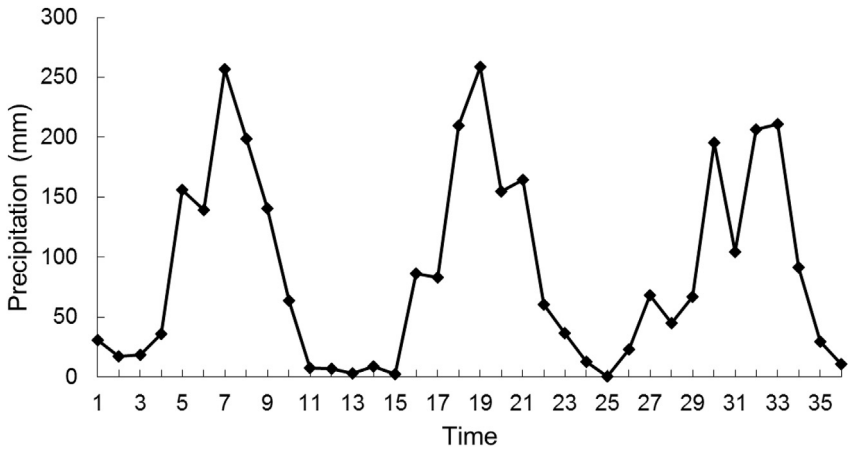


Fig. 7. Monthly precipitation in the Fuxi River catchment during 2012–12014 (months from Jan. 2012 to Dec. 2014 were numbered from 1 to 36).

catchment, temporal change of water quality at site 7 is different from others. On a fine-grained lever, C2 was further separated and seven sites can be divided into three clusters (C1: 7; C3: 2, 4; C4: 1, 3, 5, 6), which means that the water quality characteristics at sites 2 and 4 have more similar inner temporal change patterns and differ from those at sites 1, 3, 5 and 6. Specifically, sites 1, 3, 5 and 6 show similar temporal clusters in line with seasonal change, which implies that water quality at these sites are mainly affected by precipitation. In contrast, water at the upstream of site 2 is mainly reserved for Changhu reservoir used to supply drinking water, while water between Liaojiayan and Shuanghekou (site 4) is for the purpose of agricultural irrigation. Therefore, the temporal changes of water quality at sites 2 and 4 are strongly interfered by humans and different from that of sites 1, 3, 5 and 6. In the long run, climate change will affect the water quality temporal evolution and people hav eto take appropriate actions (such as water storage, water use) to adapt to climate change.

Water quality samples at sites 1, 3, 5 and 6 exhibit similar

Please cite this article in press as: Yang, Q., et al., A generic framework to analyse the spatiotemporal variations of water quality data on a catchment scale, Environmental Modelling & Software (2017), https://doi.org/10.1016/j.envsoft.2017.11.003

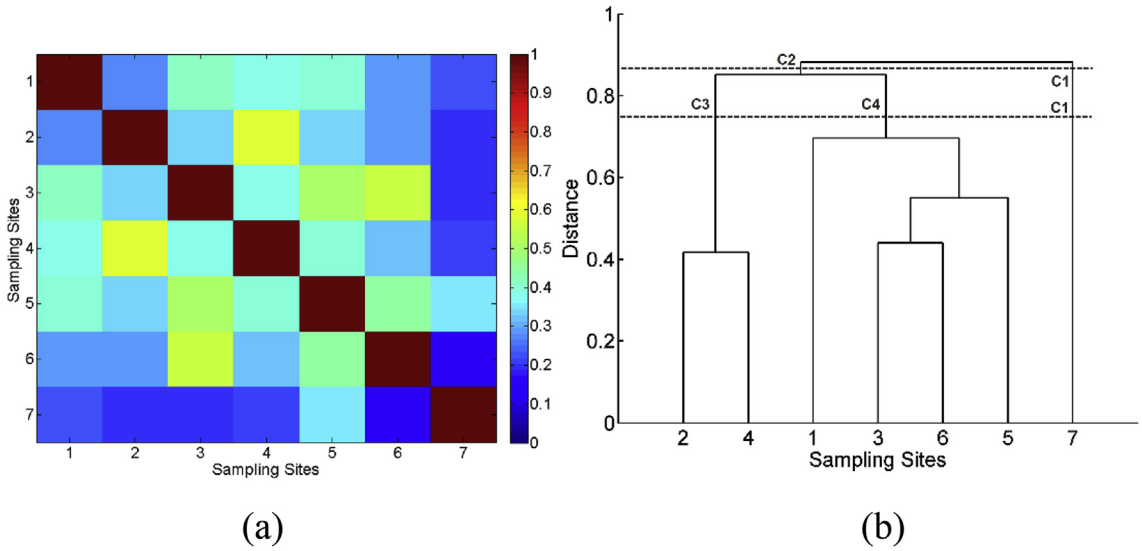


Fig. 8. (a) Similarity of temporal clusters at different sampling sites based on normalized mutual information; and (b) the corresponding dendrogram of cluster analysis.

temporal variation, which might be mainly because of precipitation change. Differently, sites 2 and 4 show another temporal pattern of variation, where human activities might be the main contributor. Water quality at site 7 shows a different temporal change pattern, and the derived reason might be that both climate and human activities impact on that. Climate conditions (e.g. precipitation) and human activities (e.g. drinking water storage, irrigation) are two main contributors to the spatial heterogeneity of temporal variation of water quality in the Fuxi River catchment.

4.3. Distinct changes of spatial distribution over time

4.3.1. Inter-change assessment

In this part, regarding the spatial distribution of water quality at each sampling time as an instance, the dissimilarity of these instances have been investigated based on Euclidean distances. Fig. 9 (a) shows the Euclidean distances of water quality distribution among different sampling times. From the distance matrix, we can observe that time steps from June to November for each year are relatively closer. It indicates that water quality parameters at these sampling times have similar spatial distributions. The potential reasons are that precipitation during June and November (wet season) is relative high and thus it can mitigate water pollution at each site by diluting pollutants, which finally reduce the spatial heterogeneity of water quality. While during the dry season, water quality is dominated by local conditions (e.g., different pollutants, different amount of discharge and water capacity), and thus tend identification to reveal reasons for spatial heterogeneity is challenging. Therefore, in the future, people should be cautious concerning local water quality degradation, especially during dry seasons.

Based on the Euclidean distance, cluster analysis was utilised and the corresponding result is illustrated in Fig. 9 (b). The dendrogram can be divided into three clusters, with the first cluster (C1) mainly ranging from June to November (summer and autumn), and the second (C2) and third cluster (C3) both ranging from December to May (winter and spring). The difference between C2 and C3 is noticeable. C2 mainly includes time steps of winter and

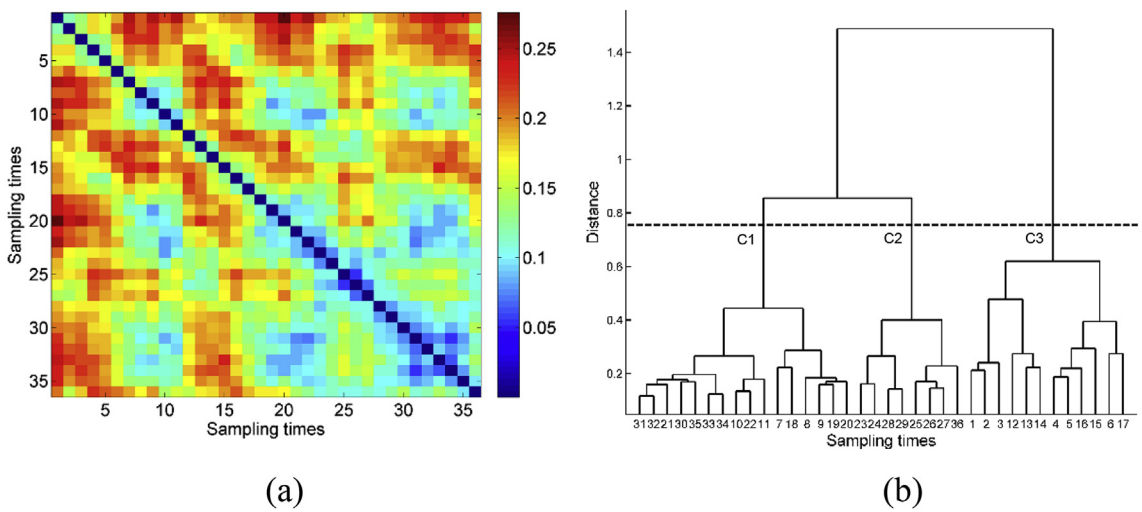


Fig. 9. (a) The Euclidean distances of water quality at different sampling times; and (b) the corresponding clustering result (months from Jan. 2012 to Dec. 2014 were numbered from 1 to 36).

Please cite this article in press as: Yang, Q., et al., A generic framework to analyse the spatiotemporal variations of water quality data on a catchment scale, Environmental Modelling & Software (2017), https://doi.org/10.1016/j.envsoft.2017.11.003

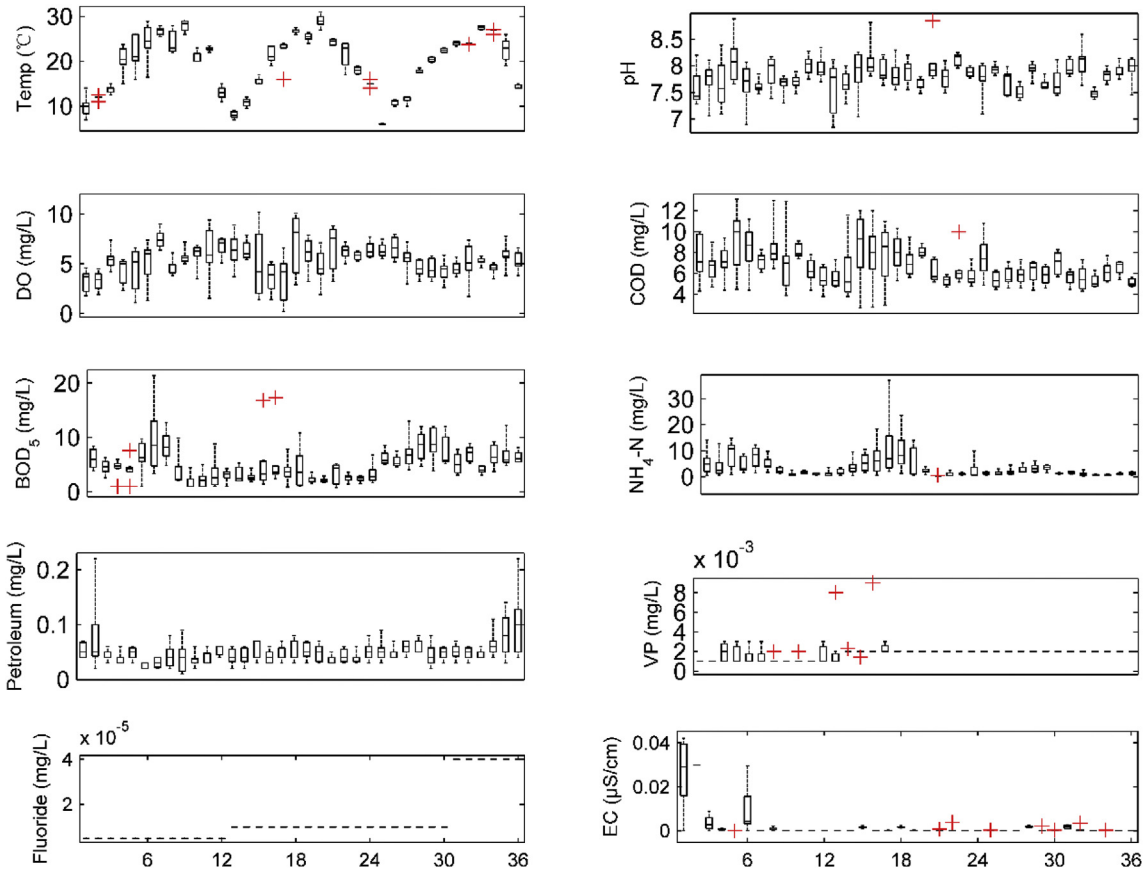


Fig. 10. Temporal variation of ten water quality variables in the Fuxi River catchment, China, from 2012 to 2014 (red cross "+" denotes outliers with 3IQR; months from Jan. 2012 to Dec. 2014 were numbered from 1 to 36). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

spring in the year of 2014, while C3 primarily covers winter and spring time steps over the years of 2012 and 2013. These results reveal that the spatial distribution of water quality exhibits both seasonal and annual variations.

To verify the effectiveness of our results, the temporal variations of ten water quality variables in the Fuxi River catchment are displayed by box-whiskers plots in Fig. 10. Seen from the medians (bands inside the boxes) in Fig. 10, among ten water quality variables, temperature shows the most obvious seasonal variation patterns for the years 2012–2014 followed by DO, COD and BOD_5. Observed from the box variations (from top to bottom) in Fig. 8, a relatively high spatial variation of temperature appeared between April and June of 2012 and the high spatial variations of COD, BOD_5 and NH_4-N often happened in spring and early summer. These findings are consistent with the cluster results in Fig. 9 (b). Furthermore, the concentrations of COD and NH_4 in 2014 were stable (small variations) and kept at a low level, which implies that water quality improved in 2012 and 2013. This also explain our findings that time steps in winter and spring for 2014 fell in a different cluster compared to those of 2012 and 2013.

In summary, water quality at time steps during June and November show similar spatial distributions, since precipitation is intensive during this period and thus can dilute pollutants and reduce the spatial heterogeneity of water quality. In contrast, water quality in winter and spring share similar spatial distributions, exhibiting larger spatial heterogeneity, which might be due to the water quality during dry seasons being more impacted by local conditions.

4.3.2. Intra-change assessment

From the cluster perspective, the variation of spatial distribution of water quality is studied. Cluster analysis using the "Ward" distance is applied on water quality data at each sampling time, and all the corresponding spatial cluster results are displayed in Fig. 9. Apparently, spatial distribution of water quality differs among different sampling times. In general, Fig. 11 indicates that the Ward distances among different sampling sites become smaller and smaller on average from 2012 to 2014, especially for the period from January to August. This implies that water quality at these sampling sites becomes more similar. Specifically, Ward distances among 7 sites larger than 1.5 appeared in March and April of 2012, and only in April of 2013. All Ward distances of 7 sampling sites in 2014 were less than 1.

To compare the differences of these generated spatial clusters, from a perspective of entropy, NMI is firstly applied to represent their similarities (as illustrated in Fig. 12 (a)) and cluster analysis is then performed based on the similarities. Fig. 12 (b) presents the cluster results, where the dendrogram can be split into three clusters by the user-defined dash line. The time steps represent numbered months. Jan. 2012 to Dec. 2014 were numbered from 1 to 36. The first cluster (C1) includes the numbered months 7 (Jul. 2012), 22 (Oct. 2013), 23 (Nov. 2013), 27 (Mar. 2014), and 28 (Apr. 2014). The third cluster (C3) consists of time steps 25 (Jan. 2014), 33 (Sep. 2014), 34 (Oct. 2014) and 31 (Jul. 2014), and the other sampling times are grouped together (C2). The time steps, which fall into one group, indicate that water quality variables at these sampling times have similar spatial distribution.

Please cite this article in press as: Yang, Q., et al., A generic framework to analyse the spatiotemporal variations of water quality data on a catchment scale, Environmental Modelling & Software (2017), https://doi.org/10.1016/j.envsoft.2017.11.003

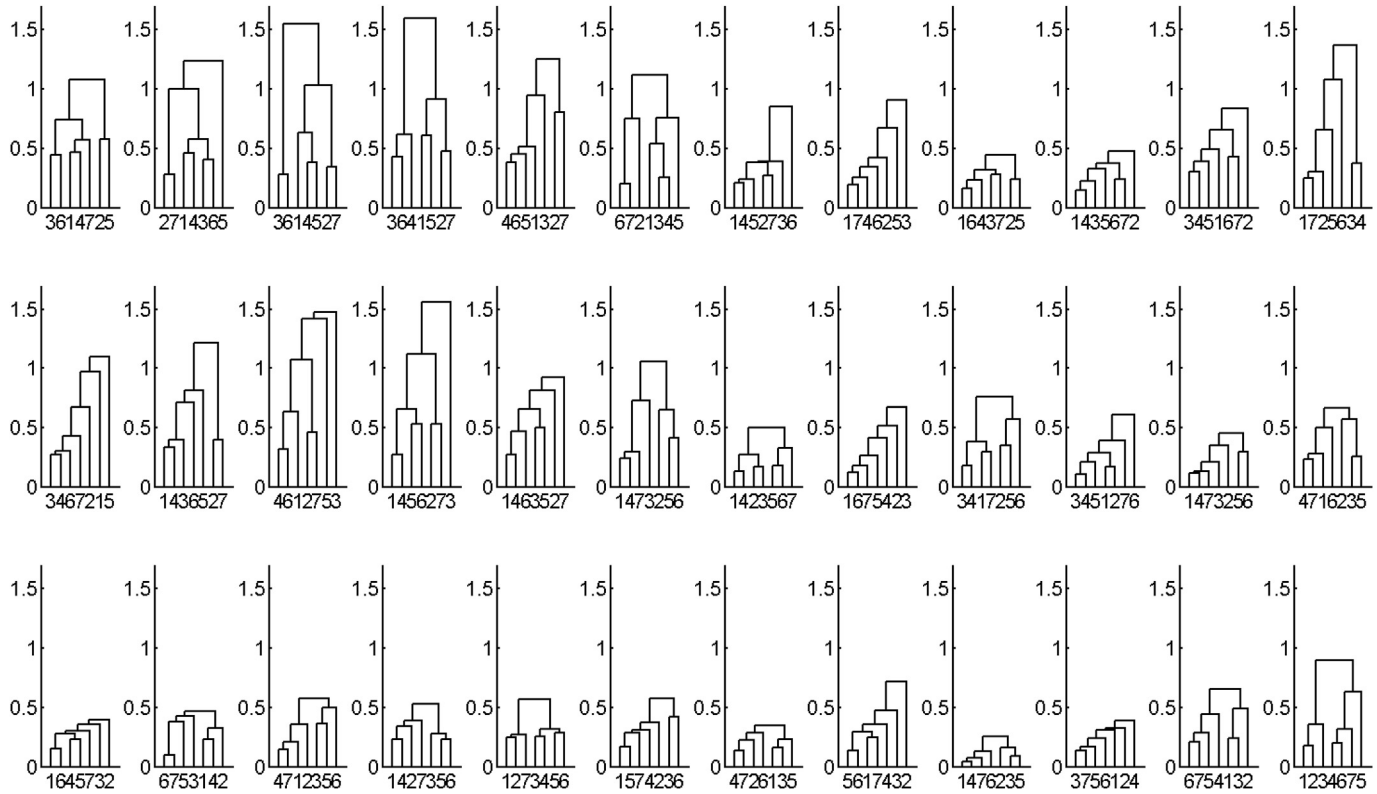
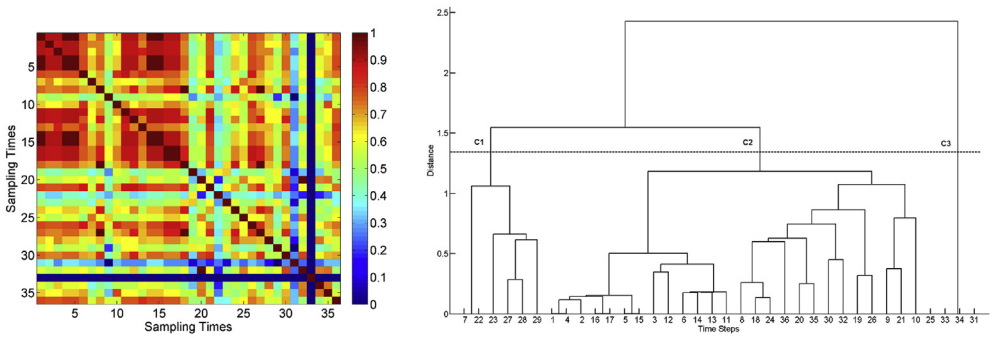


Fig. 11. Dendrograms of spatial clusters of water quality at different sampling times (months from Jan. 2012 to Dec. 2014).



(a) (b)

Fig. 12. (a) Spatial distribution similarities among different sampling times; and (b) the corresponding clustering result.

4.4. Application and extension of the framework and limitations

The proposed framework can be applied to other catchments where there are continuous monitoring data of water quality from certain sampling points (more than 2 points) during a certain period (suggest no less than 3 years). When the computational results come out of the proposed framework, we suggest users interpret the results and reasons behind the results by thoroughly identifying all kinds of natural and human factors in the studied catchment for better understanding. Furthermore, the proposed framework can be extended to other environmental phenomenon studies on catchment scale (such as air pollution, noise pollution and climate change), whereas there are different sampling sites in the study area and each site has a series of observations over a certain period. The number of the variables are flexible, depending on the research problems in practice and the data availability.

After finding the research problem and available data, how can the framework be applied? Fig. 13 gives a flowchart to guide users to apply the generic framework to other applications. Before spatiotemporal analysis, data pre-processing is usually needed, such as normalization or noisy filtering if necessary. The analysis mainly consists of two tasks. One task is the analysis on heterogeneity of temporal evolution at different sites while another task is the analysis of change concerning spatial distributions over time.

The application and extension of the proposed framework is limited by data availability and data quality. Regarding data availability, one aspect is that for long-term continuous water quality monitoring data are generally owned by environment agencies and there is a lack of sharing detailed data internationally. Another aspect is that the number of data monitoring points are limited,

Please cite this article in press as: Yang, Q., et al., A generic framework to analyse the spatiotemporal variations of water quality data on a catchment scale, Environmental Modelling & Software (2017), https://doi.org/10.1016/j.envsoft.2017.11.003

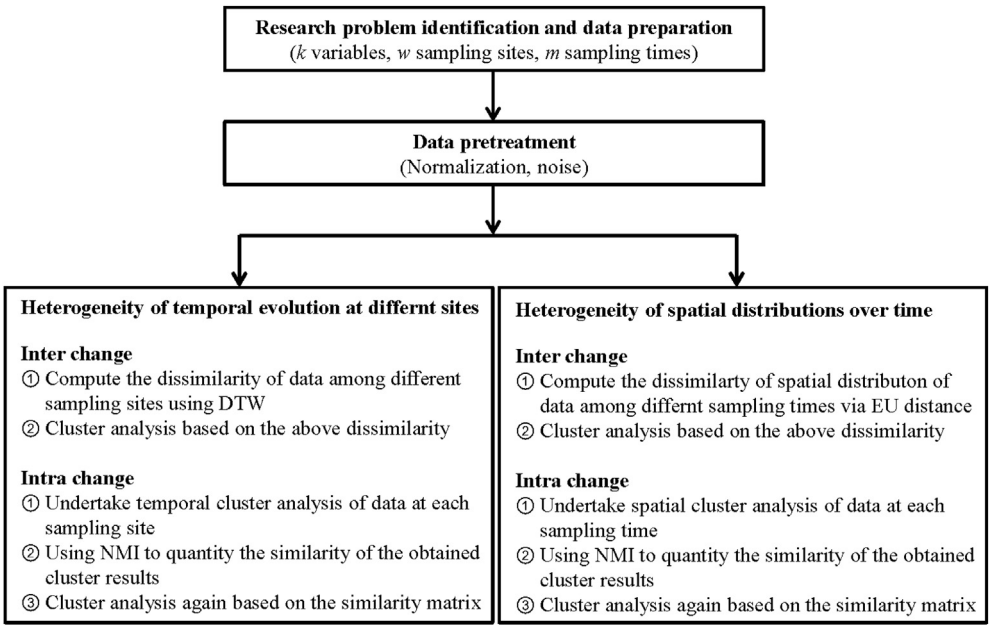


Fig. 13. Flowchart for using the generic framework for other applications. NMI: normalized mutual information, DTW: dynamic time warping, EU: Euclidean.

especially in remote areas due to economic and human resources costs. There might be a missing data challenge. Missing data often result in uncertainty and imprecision of the analysis results. Therefore, high data availability and good data quality are important prerequisites for using the proposed framework.

5. Conclusions and recommendations

A generic framework for the spatiotemporal variations of water quality on the catchment scale has been proposed. This framework includes two tasks of water quality analysis: spatial heterogeneity of temporal evolution and changes of spatial distribution over time. The results with respect to the specific case study indicate that our framework allows revealing the inter- and intra-change of water quality systematically and report the spatial and temporal changes of water quality simultaneously. The proposed framework will provide environment scientists and managers with an intuitive and effectively tool for spatiotemporal analysis of water quality data, allowing for better decision making on water pollution control and ecological restoration on a catchment scale. In the future, more research should be undertaken on extending the proposed framework to other environmental data.

Acknowledgments

Author contributions: J. S. and Q. Y. designed the research; G. W. and M. S. made valuable suggestions and comments to the research design; Q.Y., and J. S. analysed the data; G. W. and X. L. provided the data, Q. Y. wrote the first manuscript draft; all authors read and commented on the paper, and M. S. edited the final paper.

This work has been financially supported by the National Natural Science Foundation of China (grant numbers 61403062 and 41601025), The National Key Research and Development Program of China (grant numbers 2016YFA0601501, 2016YFA0601601 and 2016YFB0502303), State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering (grant number 2015490811), the Postdoctoral Science Foundation of China (grant numbers 2014M552344 and 2015T80973), Science-Technology Foundation for Young Scientist of Sichuan Province (grant number 2016JQ0007), and Fund Project of Sichuan Provincial Academician (Expert) Workstation (grant number 2014YSGZZ02).

References

Abbaspour, K.C., Yang, J., Maximov, I., Siber, R., Bogner, K., Mieleitner, J., Srinivasan, R., 2007. Modelling hydrology and water quality in the pre-alpine/alpine Thur watershed using SWAT. J. Hydrology 333 (2), 413–430.

Bilgin, A., Konanç, M.U., 2016. Evaluation of surface water quality and heavy metal pollution of Coruh River Basin (Turkey) by multivariate statistical methods. Environ. Earth Sci. 75 (12), 1–18.

Bricker, O.P., Jones, B.F., 1995. Main Factors Affecting the Composition of Natural Waters. Trace Elements in Natural Waters. CRC Press, Boca Raton, FL, pp. 1–5.

Bu, H., Tan, X., Li, S., Zhang, Q., 2010. Temporal and spatial variations of water quality in the Jinshui River of the south Qinling Mts., China. Ecotoxicol. Environ. Saf. 73 (5), 907–913.

Chang, H., 2008. Spatial analysis of water quality trends in the Han River basin, South Korea. Water Res. 42 (13), 3285–3304.

Cover, T.M., Thomas, J.A., 1991. Entropy, relative entropy and mutual information. Elem. Inf. Theory 2, 1–55.

Ding, J., Jiang, Y., Fu, L., Liu, Q., Peng, Q., Kang, M., 2015. Impacts of land use on surface water quality in a subtropical River basin: a case study of the dongjiang River basin, southeastern China. Water 7 (8), 4427–4445.

Iscen, C.F., Emiroglu, Ö., Ilhan, S., Arslan, N., Yilmaz, V., Ahiska, S., 2008. Application of multivariate statistical techniques in the assessment of surface water quality in Uluabat Lake, Turkey. Environ. Monit. Assess. 144 (1–3), 269–276.

Itakura, F., 1975. Minimum prediction residual principle applied to speech recognition. IEEE Trans. Acoust. Speech, Signal Process. 23 (1), 67–72.

Jain, A.K., Murty, M.N., Flynn, P.J., 1999. Data clustering: a review. ACM Comput. Surv. (CSUR) 31 (3), 264–323.

Jha, M.K., Datta, B., 2014. Linked simulation-optimization based dedicated monitoring network design for unknown pollutant source identification using dynamic time warping distance. Water Resour. Manag. 28 (12), 4161–4182.

Keogh, E., Ratanamahatana, C.A., 2005. Exact indexing of dynamic time warping. Knowl. Inf. Syst. 7 (3), 358–386.

Kim, H., Sa, J., Chung, Y., Park, D., Yoon, S., 2016. Fault diagnosis of railway point machines using dynamic time warping. Electron. Lett. 52 (10), 818–819.

Kostoulas, T., Chanel, G., Muszynski, M., Lombardo, P., Pun, T., 2015. November). Dynamic time warping of Multimodal signals for detecting highlights in Movies. In: Proceedings of the 1st Workshop on Modeling INTERPERsonal SynchrONy and InfLuence. ACM, pp. 35–40.

Laznik, M., Stålnacke, P., Grimvall, A., Wittgren, H.B., 1999. Riverine input of nutrients to the Gulf of Riga—temporal and spatial variation. J. Mar. Syst. 23 (1), 11–25.

Niu, X., Geng, J., Wang, X., Wang, C., Gu, X., Edwards, M., Glindemann, D., 2004. Temporal and spatial distributions of phosphine in Taihu Lake, China. Sci. Total Environ. 323 (1), 169–178.

Olson, C.F., 1995. Parallel algorithms for hierarchical clustering. Parallel Comput. 21 (8), 1313–1325.

Ouyang, Y., Nkedi-Kizza, P., Wu, Q.T., Shinde, D., Huang, C.H., 2006. Assessment of seasonal variations in surface water quality. Water Res. 40 (20), 3800–3810.
Pan, H., Li, J., 2016. Online human action recognition based on improved dynamic time warping. In: In 2016 IEEE International Conference on Big Data Analysis (ICBDA). IEEE, pp. 1–5.
Salvador, S., Chan, P., 2007. Toward accurate dynamic time warping in linear time and space. Intell. Data Anal. 11 (5), 561–580.
Shao, J., Hahn, K., Yang, Q., Bohm, C., Wohlschlager, A., Myers, N., Plant, C., 2010a. Combining time series similarity with density-based clustering to identify fiber bundles in the human brain. In: In 2010 IEEE International Conference on Data Mining Workshops. IEEE, pp. 747–754.
Shao, J., Hahn, K., Yang, Q., Wohlschlager, A., Bohm, C., Myers, N., Plant, C., 2010b. Hierarchical density-based clustering of white Matter tracts in the human brain. Int. J. Knowl. Discov. Bioinforma. 1 (4), 1–26.
Shrestha, S., Kazama, F., 2007. Assessment of surface water quality using multivariate statistical techniques: a case study of the Fuji river basin, Japan. Environ. Model. Softw. 22 (4), 464–475.
Singh, K.P., Malik, A., Mohan, D., Sinha, S., 2004. Multivariate statistical techniques for the evaluation of spatial and temporal variations in water quality of Gomti River (India)-a case study. Water Res. 38, 3980–3992.
Sliva, L., Williams, D.D., 2001. Buffer zone versus whole catchment approaches to studying land use impact on river water quality. Water Res. 35 (14), 3462–3472.
Smeti, E.M., Golfinopoulos, S.K., 2016. Characterization of the quality of a surface water resource by multivariate statistical analysis. Anal. Lett. 49 (7), 1032–1039.
Steinbach, M., Karypis, G., Kumar, V., 2000, August. A comparison of document clustering techniques. In: In KDD Workshop on Text Mining, vol. 400, pp. 525–526, 1.
Vega, M., Pardo, R., Barrado, E., Debán, L., 1998. Assessment of seasonal and polluting effects on the quality of river water by exploratory data analysis. Water Res. 32 (12), 3581–3592.
Wang, Y., Wang, P., Bai, Y., Tian, Z., Li, J., Shao, X., Li, B.L., 2013. Assessment of surface water quality via multivariate statistical techniques: a case study of the Songhua River Harbin region, China. J. Hydro-Environment Res. 7 (1), 30–40.
Ward Jr., J.H., 1963. Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. 58 (301), 236–244.
Whitehead, P.G., Wilby, R.L., Battarbee, R.W., Kernan, M., Wade, A.J., 2009. A review of the potential impacts of climate change on surface water quality. Hydrological Sci. J. 54 (1), 101–123.
Witten, Ian H., Frank, Eibe, 2005. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, Amsterdam.
Zhen, S., Zhu, W., 2016. Analysis of isotope tracing of domestic sewage sources in Taihu Lake — a case study of Meiliang Bay and Gonghu Bay. Ecol. Indic. 66, 113–120.

Please cite this article in press as: Yang, Q., et al., A generic framework to analyse the spatiotemporal variations of water quality data on a catchment scale, Environmental Modelling & Software (2017), https://doi.org/10.1016/j.envsoft.2017.11.003