ISEIS

Journal of Environmental Informatics 28(2) 110-119 (2016)

Journal of Environmental Informatics

www.iseis.org/jei

Probabilistic Evaluation of Causal Relationship between Variables for Water Quality Management

J. He*

Department of Civil Engineering, Schulich School of Engineering, University of Calgary, 2500 University Drive, Calgary, T2N 1N4, Canada

Received 2 April 2015; revised 15 July 2105; accepted 30 August 2015; published online December 30 2016

ABSTRACT. In aquatic environments, a complex interplay exists among physical, chemical, and biological water quality characteristics, which are constantly influenced by exogenous factors such as hydrological, meteorological and geological conditions. Due to the spatial and temporal variations of exogenous factors, the relationship between the water quality parameters and these factors hence becomes complicated and challenging. Given the large data matrix, one type of methods frequently seen in the literature belongs to the multivariate analysis which generates a qualitative measure of the relationships among variables in a geometrically intuitive way. However, a quantitative evaluation from a probabilistic perspective is favorable since it defines a measurable causality among variables so that more efficient water management strategies can be formulated. This paper illustrates a new way to discover the relationship between two variables by estimating their joint distribution which fully interprets the statistical dependence. A multivariate Gaussian mixture model was employed to describe the data. The model parameters were determined using the previously developed estimation approach, which is capable of dealing with both multivariate variables and censored data. The joint distribution and the conditional distribution were computed and used to describe the statistical distribution of water quality parameters, which are subject to the effects of hydro-meteorological conditions. The method was demonstrated by a case study on the Bow River in Alberta, Canada. The results shed light on how one variable affects the distribution of the other variable under complex environments in a probabilistic context.

Keywords: joint distribution, conditional probability, water quality management, mixture model, censored water quality data

1. Introduction

Water quality protection and restoration has received increasing attention over the past several decades because of the concern of water quality degradation in natural water bodies including streams, lakes, and estuaries. The degradation of water quality, which is ascribed to both anthropogenic activities (e.g., urbanization and industrialization) and changes in hydro-meteorological conditions, could lead to the impairment of water quality standards and thus hinder various beneficial uses of water. It was documented that, in the U.S., about 44% of stream miles, 64% of lake acres, and 30% of assessed bay and estuarine square miles in the less than 30% waters investigated, do not meet the requirements for water uses such as fishing and swimming (US EPA, 2004). In Canada, the water quality problems exist in many regions as stated in a recent report (Environment Canada, 2011).

In water quality management, water quality standards describe the conditions of the water quality variables and play a fundamental role in protecting the quality of water bodies. Having the quantitative baselines of these variables in place facilitates establishing treatment controls, conducting watershed planning, as well as protecting and restoring aquatic environments. However, these standards, determined for a specific type of water beneficial use, have rarely been elaborated to address complex issues. For example, the issue of site-specific attainability and impacts of flow (US EPA, 2003), which vary at intra-annual scale, is largely responsible for the variation of water quality in aquatic environments in addition to anthropogenic activities. From this viewpoint, site-specific thresholds/targets as the benchmarks for identifying whether a water body is impaired and further taking management actions are required. These thresholds are normally derived from the current situation or baseline conditions. If the thresholds do not reflect the practical changes, all subsequent procedures and management actions will be affected. Similar to water quality standards, the processes to determine the thresholds, in general, do not take the impacts of exogenous factors such as hydro-meteorological variables into consideration, regardless of the recognition of their roles in affecting water quality.

Water quality phenomena are naturally multidimensional since water quality processes intertwine with each other, and in addition to that, exogenous factors such as geological, meteorological, and hydrological conditions largely contribute to the spatial and temporal variation of water quality in aquatic environments. While there are many different approaches to explore the relationship of variables, multivariate statistical tech-

* Corresponding author. Tel.: +1 403 2204112; Fax: +1 403 2827026.
E-mail address: jianhe@ucalgary.ca (J. He)

ISSN: 1726-2135 print/1684-8799 online
 doi: 10.3808/jei.201600353

niques including cluster analysis, principle component analysis, factor analysis, discriminant analysis, and self-organizing map are most frequently adopted to qualitatively identify critical influential factors on water quality and spatial and temporal variations from complex data sets (Singh et al., 2004; Panda et al., 2006; Shrestha and Kazama, 2007; Li et al., 2015). In particular, these methods have the merit of computational simplicity and provide a geometrically intuitive interpretation due to the data matrix structure, for instance, in the principal component analysis. However, they are not capable of quantitatively measuring the relationship between two or more variables of interest and therefore are hard to establish the linkage of variables for model formulation. On the other hand in surface water bodies, the hydro-meteorological response of water quality can magnificently vary under different conditions. For example in a river, flow and water temperature can predominantly explain the variation of dissolved oxygen (DO) levels, while their roles vary with hydrological conditions, such as high, medium, and low flows (He et al., 2011). Suspended solids affect aquatic biota differently under different hydrological regimes (such as flood conditions and base-flow conditions) (Bilotta and Brazier, 2008). Furthermore, the dependence of water quality on the hydro-meteorological conditions may complicate the task of identifying the water quality targets for management, determining the cause of water quality degradation (human activities or changes in natural conditions) (Poole et al., 2004), and assessing management effectiveness (Stow and Borsuk, 2003). Different water quality levels have often been observed during dry and wet seasons in various water bodies, such as in the Danjiangkou Reservoir in China (Tan et al., 2015). In addition, Xia et al. (2015) stated that climate change, which would alter hydro-meteorological conditions, potentially affects water quality in different types of water bodies in different ways. These facts argue that the hydro-meteorological dependence of water quality is very common and should be properly represented in the statistical characteristics of water quality data (Frey and Rhodes, 1998). The multidimensional nature of water quality challenges researchers and practitioners to statistically explore the data and subsequently to derive feasible water quality management objectives/or targets, upon which to base more effective management decisions.

Recently, water quality management has attempted to address the causal relationships between water quality and hydro-meteorological conditions in the practices. For instance, different management targets are determined based on stratified water quality data according to the seasons and/or flow conditions (CCME, 2003; Government of Alberta, 2012). However, so far it appears to be quite arbitrary and lack statistical justification when grouping data under different conditions for the subsequent water quality assessment and statistical analysis. This can be overcome by conducting the probabilistic characterization of water quality in a multivariate context such that the causal relationship between two or more variables can be represented and reflected into the obtained characterization of water quality. The multivariate distribution describes the correlated random variables in terms of joint distribution, from which the probability distribution of one variable conditioned upon the remaining variables can be readily derived. Therefore, the advantages of the multivariate distribution analysis can be employed to fulfill the aforementioned needs for water quality characterization. Most recently, Hoffman and Johnson (2011) employed the multivariate distribution analysis to assess the overall contamination level of several dissolved trace metals including copper, lead, and zinc whose toxicity need to be corrected based on water hardness. Wang et al. (2012) used the multivariate distribution analysis to investigate the complicated linkages of chlorophyll a and ambient water quality. Both studies supported that the multivariate distribution is more effective for understanding the interaction of water quality and environmental variables.

This paper attempts to explore the probabilistic causal relationship of water quality variables by applying an efficient multivariate distribution approach that can statistically characterize data while accounting for the dependence between two or more variables. Similar to univariate distribution analysis for environmental data, two practical problems, i.e., unknown underlying distribution of data and censored observations below detection limits (DLs) were dealt with for properly representing the statistical characteristics of water quality data. Conventionally, the substitution method, which replaces the data points below DLs with zeros, DLs, or half of the DLs, is used due to its simplicity. However, it is well acknowledged that the substitution method lacks statistical justification and often yields biased results (Singh and Nocerino, 2004; Helsel, 2010). In the proposed methodology, the multivariate variables are modeled with Gaussian mixtures which can flexibly approximate the statistical characteristics of the complex data set. This paper further applied the developed expectation maximization (EM) algorithm (He, 2013) to estimate the distribution parameters in the presence of both uncensored and multiple censored observations. The joint and conditional probability distributions of variables of interest were hence derived from the estimated multivariate distribution. This paper incorporates the covariance of variables for taking their dependence into the analysis. In addition, applications of this method are illustrated with a case study of the Bow River in southern Alberta, Canada and potential in enhancing water quality management is discussed.

2. Materials and Methods

2.1. Study Area

The Bow River as the largest tributary of the South Saskatchewan River originates from the Rocky Mountains in Alberta, Canada and flows towards east through the mountains, the Foothills, and the plains. In the upper watershed located within the Banff National Park, the river flows through largely undeveloped and low intensity agricultural land with good water quality. Before entering the downstream watershed largely consisting of agriculture land, the river flows through the City of Calgary, the most populated community along the river. The river supports a blue ribbon fishery and provides drinking water to over half of Calgary's population. The river is usually covered or partially covered by ice between December and March with open water generally begins in April. Flow peaks occur around June or July annually, driven by combined rainfall and snowmelt.

2.2. Water Quality and Hydrological Data

Water quality varies considerably along the Bow River due to the variation in both natural conditions (e.g., hydrology and geology) and anthropogenic activities (e.g., point- and non-point sources pollution due to urbanization). Due to the significant spatial variation in water quality, five long-term water quality monitoring stations have been deployed on the river to comprehensively capture the water quality variations. The water quality data have been collected on a monthly basis. This paper uses the data collected from 1988 to 2009 at two of long-term monitoring stations; one is located in the upstream of the river about 4.5 km above Canmore and the other is situated just upstream of the confluence with the South Saskatchewan River (near Ronalane Bridge). These two stations are called the upstream and downstream stations, respectively, throughout this paper. It should be noted that this paper does not target any specific water quality parameters but rather aiming to develop methods for the data analysis. In this paper, water quality parameters including DO, water temperature, turbidity, specific conductance, dissolved total phosphorus (TP), and TP are selected for the analysis. The data sizes range from 258 to 278. All data of DO, water temperature, turbidity, specific conductance, TP are above DL; while 20% of dissolved TP data is censored.

Daily flow data collected by the Water Survey of Environment Canada at the Bow River at Banff (station: 05BB001) and the Bow River Near the Mouth (station: 05BN012) were used. These two hydrometric stations are in close proximity to the upstream and downstream water quality monitoring stations, respectively. The flow data corresponding to the water quality sampling dates were extracted from the daily flow data sets of these two hydrometric stations. It should be noted that the water temperatures measured at the upstream and downstream water quality monitoring stations are considered to represent the meteorological air temperatures since water temperatures are usually strongly associated with air temperatures and largely affects chemical and biological reactions occurring in water column.

2.3. Methodology

2.3.1. Gaussian Mixture Model (GMM)

The GMM is a desirable parametric model which can closely approximate the unknown probabilistic distribution of many water quality parameters (He, 2013). In particular, the GMM is theoretically proven to be a universal approximator which means that it can approximate any continuous distribution to any degree given a sufficient number of components (Titterington et al., 1985). The typical finite GMM is a linear additive model which is given by:

$$p(\mathbf{x}\mid\Theta)=\sum_{k=1}^{K}\alpha_k p_k(\mathbf{x}\mid\mu_k,\Sigma_k) \tag{1}$$

where $\mathbf{x}$ is an M-dimensional data vector distributed accrding to $p(\mathbf{x}\mid\Theta)$ parameterized by Θ; α_k is the nonnegative weight of the k-th component of the GMM; $p_k(\mathbf{x}\mid\mu_k,\Sigma_k)$ is the k-th Gaussian component with mean μ_k and covariance matrix Σ_k; i.e.,

$$p_k(x\mid\mu_k,\Sigma_k)=(2\pi)^{-M/2}\mid\Sigma_k\mid^{-1/2}\exp\{-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)\}$$

where the superscript T and $|\cdot|$ denote the transpose and determinant of a matrix, respectively. Furthermore, the summation of all weights must be equal to 1. The parameter set Θ is the set of all component-specific parameters appeared in the right side of (1) and can be represented by:

$$\Theta=\{\alpha_k,\ \mu_k,\Sigma_k\mid k=1,\cdots,K\}$$

The GMM has excellent properties for analysis. For instance, if the joint distribution of $x_1,\cdots,x_M$ is a mixture of K multivariate normal distributions with weights $\{\alpha_1,\cdots,\alpha_K\}$, then the joint distribution of any subset of $\mathbf{x}$ is a mixture of K multivariate normal distributions with the same weights (Titterington et al., 1985; Kotz et al., 2000).

2.3.2. Estimation Methods

Given uncensored data, the standard EM algorithm described as follows is employed to recursively estimate the parameters and weights until a local maximum of likelihood function is reached. More details on the standard EM algorithm can be found in He (2013).

$$\hat{\alpha}_k=\frac{1}{N}\sum_{i=1}^{N}p(k\mid x_i,\Theta^p) \tag{2}$$

$$\hat{\mu}_k=\frac{\sum_{i=1}^{N}\mathrm{x}_i p(k\mid \mathrm{x}_i,\Theta^p)}{\sum_{i=1}^{N}p(k\mid \mathrm{x}_i,\Theta^p)} \tag{3}$$

$$\hat{\Sigma}_k=\frac{\sum_{i=1}^{N}p(k\mid \mathrm{x}_i,\Theta^p)(\mathrm{x}_i-\hat{\mu}_k)(\mathrm{x}_i-\hat{\mu}_k)^{\mathrm{T}}}{\sum_{i=1}^{N}p(k\mid \mathrm{x}_i,\Theta^p)} \tag{4}$$

In more general cases, some water quality data are often asynchronously censored, which means that individual variables in the multivariate data are subject to different DLs at different time instant. To deal with the multiple censored data, the complete data defined in the EM method include the missing and/or censored data points and the measurements above the DLs or uncensored data, both of which subsequently form two conditional expectations respectively corresponding to the censored data $\mathbf{z_d}$ of the length of N_d and the uncensored data $\mathbf{x}$ of the length of N_0. The estimation algorithm derived by He (2013) is summarized as follows:

$$\hat{\alpha}_k=\frac{1}{N}\Big[\sum_{i=1}^{N_0}p(k\mid \mathrm{x}_i,\Theta^p)+\sum_{i=1}^{N_d}p(k\mid \mathrm{z}_d^i,\Theta^p)\Big] \tag{5}$$

$$\hat{\mu}_k=\mathcal{D}/\mathcal{C},\quad \hat{\Sigma}_k=\mathcal{G}/\mathcal{C} \tag{6}$$

where Θ^p denotes the set of parameters as defined in previous section; N is the length of the data; and C, D, and G are calculated by:

$$\mathcal{C} = \sum_{i=1}^{N_0} p(k \mid \mathrm{x}_i, \Theta^p) + \sum_{i=1}^{N_d} p(k \mid \mathrm{z}_i^d, \Theta^p) \tag{7}$$

$$\mathcal{D} = \sum_{i=1}^{N_0} \mathrm{x}_i p(k \mid \mathrm{x}_i, \Theta^p) + \sum_{i=1}^{N_d} p(k \mid \mathrm{z}_i^d, \Theta^p) \int_{\Omega_i} \mathrm{x}_i p(\mathrm{x}_i \mid \mathrm{z}_i^d, k, \Theta^p) d\mathrm{x}_i \tag{8}$$

$$\mathcal{G} = \sum_{i=1}^{N_0} p(k \mid \mathrm{x}_i, \Theta^p) \hat{\Lambda}_{k,i} + \sum_{i=1}^{N_d} p(k \mid \mathrm{z}_i^d, \Theta^p) \int_{\Omega_i} \hat{\Lambda}_{k,i} p(\mathrm{x}_i \mid \mathrm{z}_i^d, k, \Theta^p) d\mathrm{x}_i \tag{9}$$

with $\hat{\Lambda}_{k,i}$ being the estimated covariance matrix between $\mathbf{x}_i$ and μ_k.

2.3.3. Joint and Conditional Distributions

After obtaining the joint distribution of multivariate data, the conditional probability, namely the distribution of *y* given a specific *x* or the range of *x*, can be derived. For the convenience of illustration, the computation of the conditional probability in a bivariate context is given below. This however can be easily extended to more than two variables.

For two Gaussian random variables *x* and *y* with $x \sim \mathcal{N}(\mu_x, \sigma_x^2)$ and $y \sim \mathcal{N}(\mu_y, \sigma_y^2)$, respectively, the conditional distribution of *y* given *x* is:

$$y \mid x \sim \mathcal{N}(\mu_y + \frac{\sigma_y}{\sigma_x}\rho(x - \mu_x), (1-\rho^2)\sigma_y^2) \tag{10}$$

where ρ is the correlation coefficient between *x* and *y*. For data reasonably described by the GMM, the conditional probability distribution can be computed by the following equation:

$$p(y \mid x) = \sum_{k=1}^{K} \alpha_k \frac{p_k(x, y \mid \Theta)}{p_k(x \mid \Theta)} = \sum_{k=1}^{K} \alpha_k p_k(y \mid x, \Theta) \tag{11}$$

where $p_k(y \mid x, \Theta)$ is obtained from (1). With the computed $p(y \mid x)$, it is convenient to assess the likelihood of two random variables. Furthermore, the likelihood of one variable given the other variable falling within any specified range of interest can also be evaluated.

3. Results

This paper demonstrates the potential application of the proposed approach with the real water quality data instead of targeting specific water quality parameters. To illustrate the generality of the proposed approach for different scenarios of water quality parameters, the proposed approach was applied to both the uncensored and censored data sets.

3.1. Variations of Hydro-meteorological and Water Quality Variables and Their Dependence

In the riverine environment, both flow and water quality generally exhibit certain seasonal variations. Taking downstream station for example, Figure 1 shows the boxplots of flow, DO, dissolved TP and TP, respectively. It is obvious that DOs vary inversely as flow. In particular, the lower DOs correspond to higher flows. In addition, the water quality response to the hydrological conditions can be further observed from Figure 2, in which the flow is divided into low flow (base flow) and high

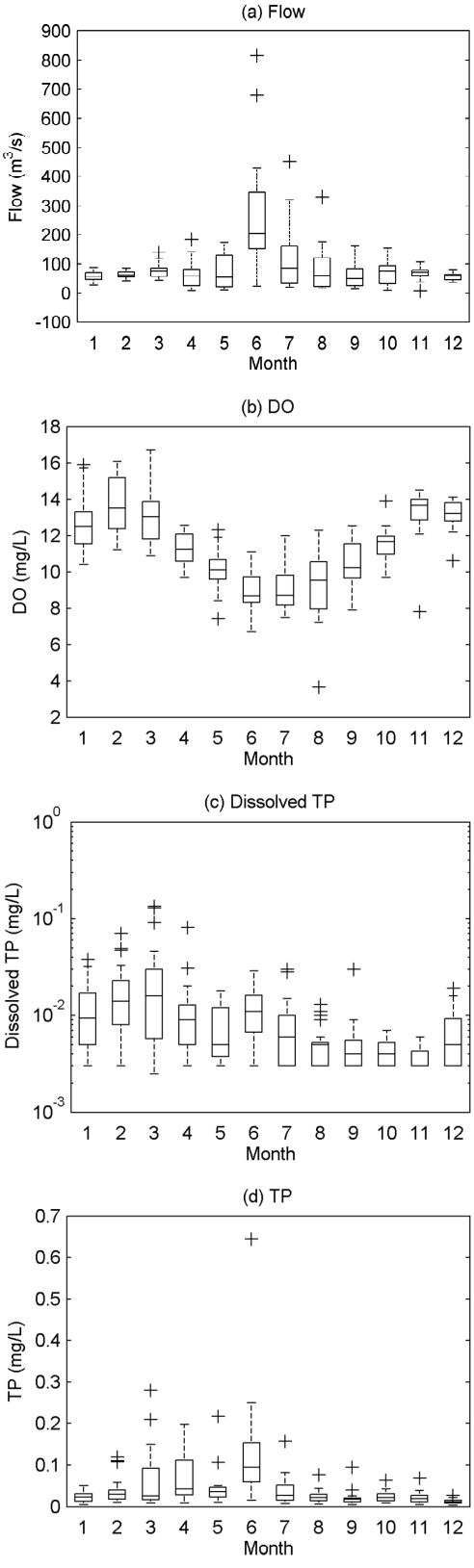

Figure 1. Box plots of (a) flow, (b) DO, (c) dissolved TP and (d) TP at the downstream station.

flow (above base flow) conditions. As demonstrated in these figures, the dependence of the water quality parameters on flow tends to vary with the hydrological conditions. For example, DO's variation under low flows appears stochastic; whereas the hydrological dependence of DO can be seen under high flows in Figure 2(a). Similar to DO at the upstream station, the dependences of both turbidity and dissolved TP on flow show different trends under low and high flow conditions at the downstream station as illustrated in Figures 2(b) and 2(c), respectively. By closely examining Figure 2(d), flow response of specific conductance can be spotted, whereas its response to flow is different from the other water quality parameters which also present their different responses to the low and high flows. In addition, this figure shows that the dependence of specific conductance on flow is consistent in the range of flow from 0 to 60 m^3/s and the specific conductance is more or less constant when flow is above 60 m^3/s. Similarly, the dependence of TP on flow and its variation under different flow ranges at the downstream station can be observed in Figure 2(f). The scatter plots, on the other hand, can assist in dividing regimes in most cases as seen in Figure 2. A rough division of flow can be observed for each variable while this appears to be ambiguous for dissolved TP. In addition to flow, the dependence of DO on water temperature at the downstream station is illustrated in Figure 2(e).

The histograms of DO and dissolved TP at the downstream station and specific conductance at the upstream station are displayed in Figure 3 as examples. The shapes of these histograms imply that a single distribution will not be able to closely fit the data, especially shown in Figures 3(a) and 3(c). It appears that different distributions would be needed for differrent ranges of the water quality parameters. This is expected since different governing mechanisms play determining role in different hydro-meteorological conditions, which correspondingly result in complex distribution of the water quality parameters.

3.2. Uncensored Multivariate Distribution Analysis

For the uncensored case, the data of flow and DO observed at the upstream station were used here as an example. As shown in Figure 4(a), individual Gaussian components in the GMM were determined for describing the variables under different flow regions as the shape of the joint distribution appears to be separated by flows (high and low flows). Figure 4(b) presents the conditional cumulative probability distributions of DO given the ranges of flow, which show the different hydrological response of DO under different flows. Figures 5 ~ 7 illustrate the results for flow and DO at the downstream station, water temperature and DO at the upstream station, and flow and turbidity at the downstream station, respectively. These figures indicate that the dependence of water quality on hydro-meteorological variables varies at different conditions and also suggest that their dependence varies spatially, as observed from Figures 4 and 5.

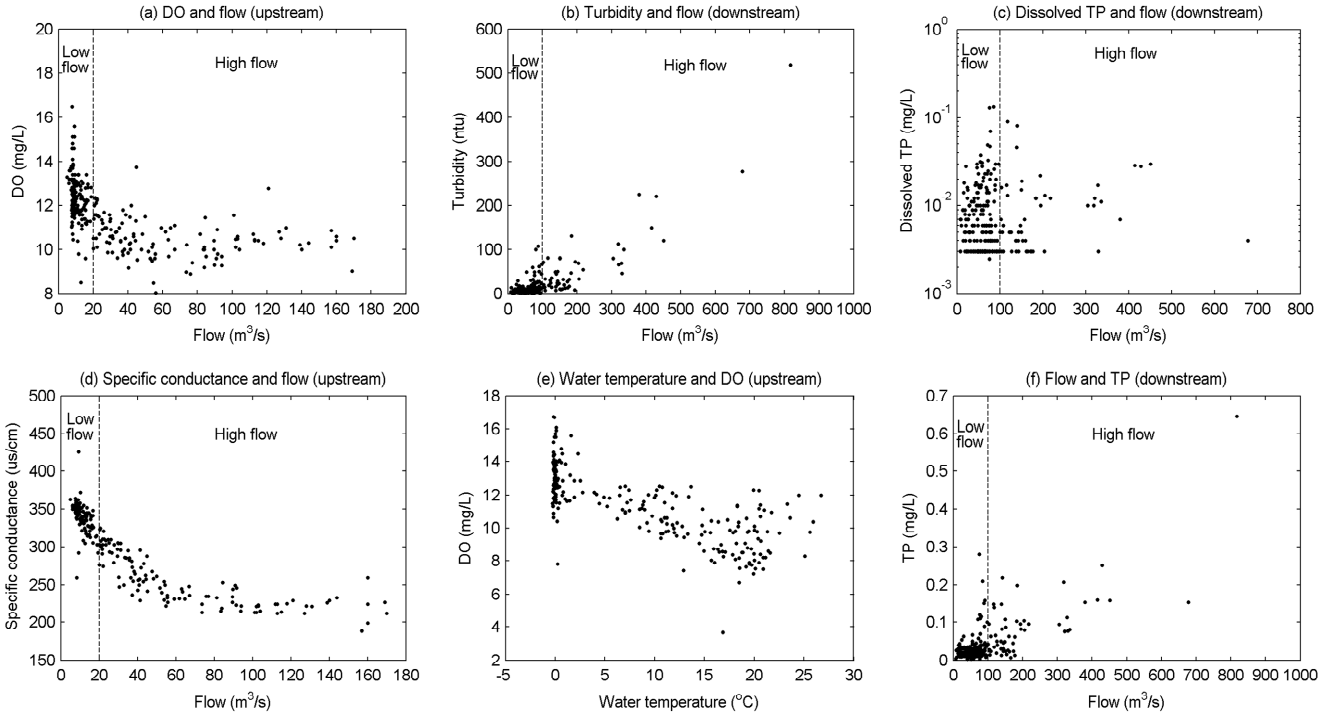

Figure 2. Scatter plots of (a) DO and flow at the upstream station, (b) turbidity and flow at the downstream station, (c) dissolved TP and flow at the downstream station, (d) specific conductance and flow at the upstream station, (e) water temperature and DO at the upstream station, and (f) flow and TP at the downstream station.

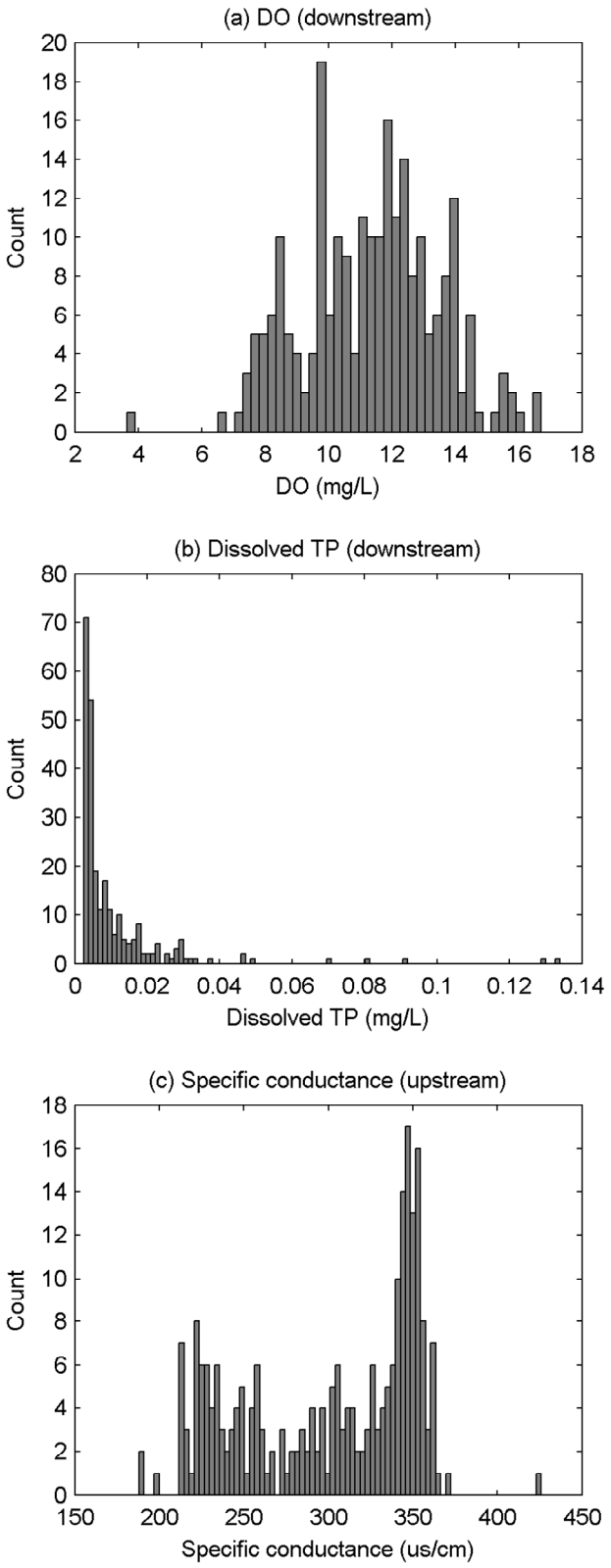

Figure 3. Histograms of (a) DO and (b) dissolved TP at the downstream station and (c) specific conductance at the upstream station.

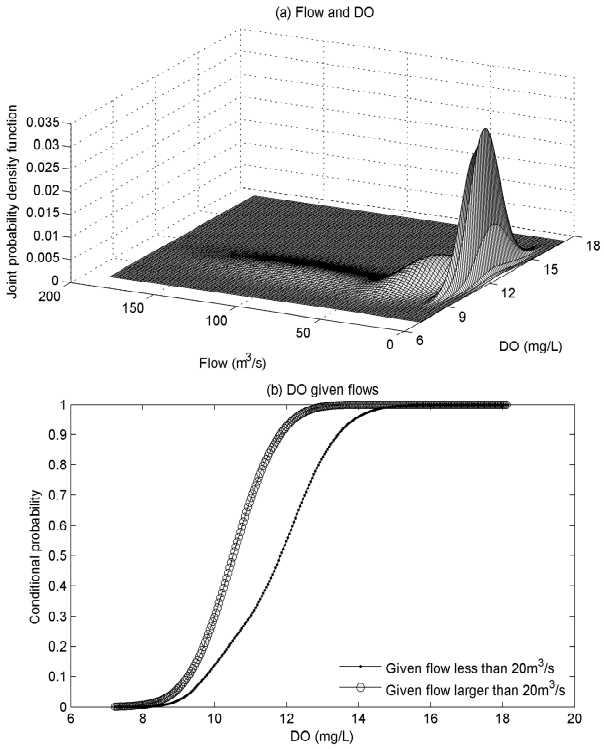

Figure 4. Results of (a) the joint probability density function between flow and DO and (b) the conditional cumulative probability distributions of DO given flows at the upstream station.

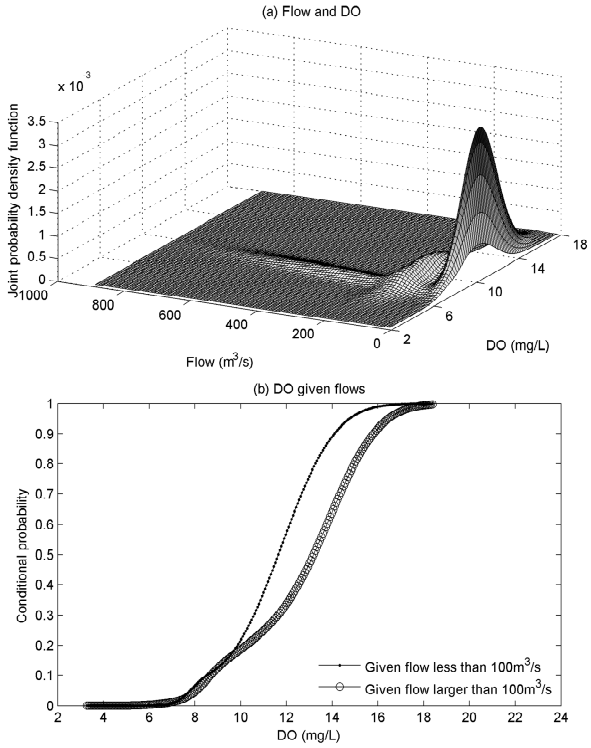

Figure 5. Results of (a) the joint probability density function between flow and DO and (b) the conditional cumulative probability distributions of DO given flows at the downstream

station.

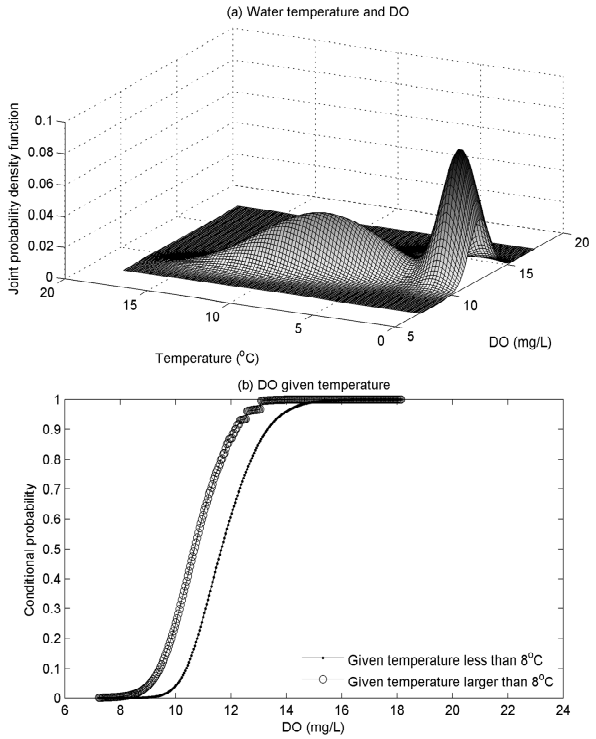

Figure 6. Results of (a) the joint probability density function between water temperature and DO and (b) the conditional cumulative probability distributions of DO given temperatures at the upstream station.

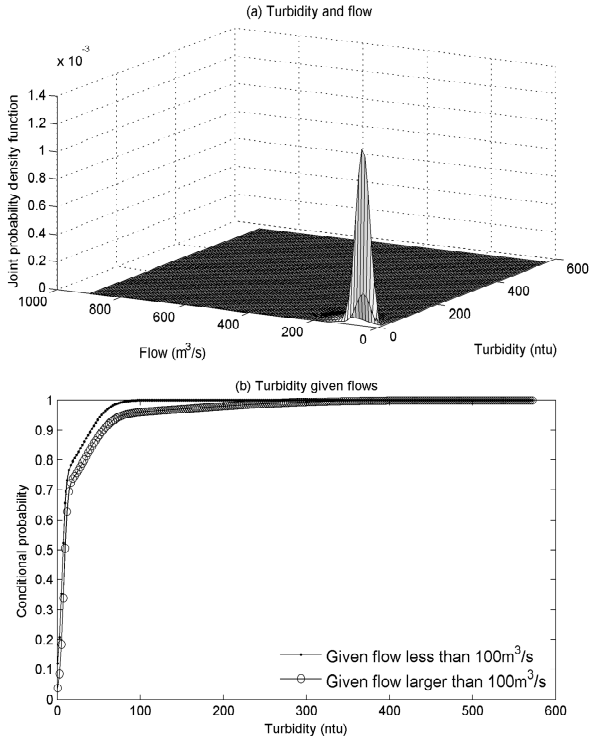

Figure 7. Results of (a) the joint probability density function between flow and turbidity and (b) the conditional cumulative probability distributions of turbidity given flows at the downstream station.

3.3. Censored Multivariate Distribution Analysis

The data sets of flow and dissolved TP observed at the downstream station were used as an example to demonstrate the applicability of the proposed approach to censored data sets. In the data sets, dissolved TP data are subject to detection limit of 0.003mg/L; while there is no censoring in the flow data. The derived joint distribution and conditional cumulative probability distributions of dissolved TP given flows are presented in Figures 8(a) and 8(b), respectively. It is observed that the conditional cumulative probability distributions behave quite similar for small and large values of dissolved TP under both flow conditions while the conditional probability on high flows is relatively smaller than that on lower flows when the dissolved TP has a moderate value, typically ranging from 0.01 to 0.08 mg/L.

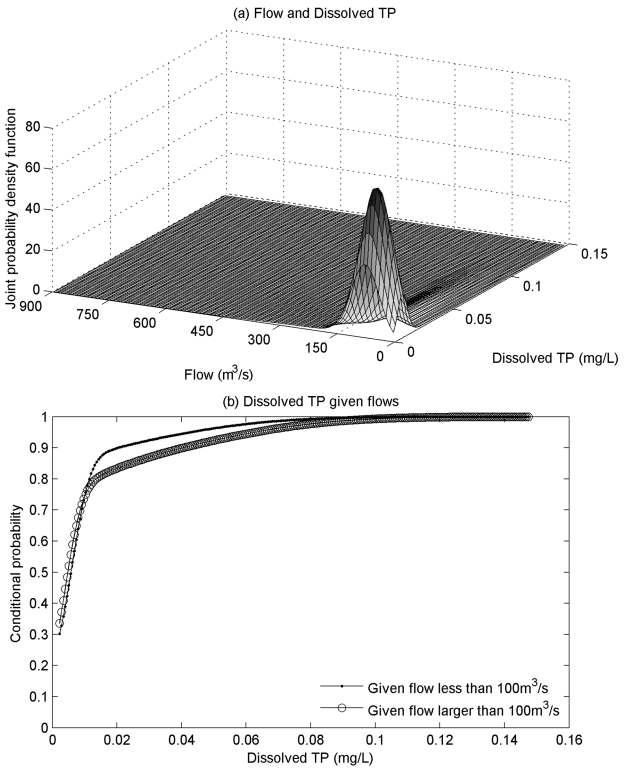

Figure 8. Results of (a) the joint probability density function between dissolved TP and flow and (b) the conditional cumulative probability distributions of dissolved TP given flows at the downstream station.

4. Discussions

4.1. Hydro-meteorological Regimes and Water Quality Management

In water quality management, in particular for ecosystem protection, regime based water quality management has been increasingly advocated by many researchers (e.g. Poff et al., 1997; Poole et al., 2004). This implies that the information on natural factors and their regimes should be emphasized in the practices by the means of their distributions across space and

time. Among the natural or environmental factors, flow regularly manifests prominent seasonal variation of water quality and has been demonstrated to be primary cause of seasonal variations of many physical, chemical and biological water quality parameters (Interlandi and Crockett, 2003). In addition, water temperature which is strongly associated with meteorological condition, especially air temperature, certainly affects chemical and biological processes. From the perspective of ecosystem protection, Poff et al. (1997) argued to consider flow regimes in terms of the magnitude, frequency, timing, duration, and rate of change in flow, which describe the aspects of a flow regime both temporally and spatially. Regarding flow, this proposed distribution analysis approach is capable of taking the effects of two out the five aspects aforementioned, namely magnitude and frequency, into consideration. The effects of other three aspects of flow regimes can be included if having sufficiently long-term data series to quantify them statistically; it is however not discussed in this paper.

On the other hand, in aquatic environments, the causal-effect relationship between water quality parameters and hydro-meteorological variables is very difficult to formulate using physically-based models or simple empirical models obtained from statistical analysis. This is especially true considering the fact that the governing mechanisms of water quality are complicated, as implicitly suggested in Figures 2 and 3. The governing mechanisms of pollutant transport and/or the sources of pollutants can be distinct under different hydro-meteorological conditions. For example, under high flows both flow and temperature might play significant roles on DO levels; whereas the biological processes, photosynthesis and respiration of periphyton and macrophytes, may override the roles of hydro-meteorological factors under low flows in rivers. As a result it is logical to conduct water quality management considering different hydro-meteorological conditions, which shift and in turn would lead to the shift of their roles on water quality in an intra-annual scale. Therefore, the determination of thresholds/targets of water quality without considering the complicated causal-effect relationship is very likely to cause inefficient water quality management. For instance, the water quality violation may not be detected. In addition, the assessment of the effectiveness of pollutant management actions in reducing pollutant loads may yield unexpected results if without taking the effects of flow on water quality into account (Stow and Borsuk, 2003).

The results obtained from the real data analysis demonstrate that the proposed approach is applicable to multivariate water quality data with different distributions. It also has the potential to quantitatively formulate the statistical relationship so as to incorporate it into water quality management. The approach can also provide the decision-makers and water quality managers with the probabilistic distributions of a water quality parameter under a given hydrological conditions or the level of another water quality parameter. If the probabilistic distribution is given conditioned on the natural conditions, the proposed approach is able to assist in distinguishing either natural (hydrological and meteorological) or anthropogenic causes of changes in water quality. Therefore, the effects of the natural conditions can be removed in the subsequent water quality assessment.

4.2. Regime Division

As mentioned in the previous section, the regime based water quality management has the potential to enhance water quality management, as it is capable of assisting in identifying water quality impairment which is not caused by changes of natural conditions, such as hydro-meteorological conditions. In fact, controlling these natural conditions is not feasible to improve water quality. Thus capturing the statistical linkage between water quality and different natural conditions is needed. From this perspective, regime based management would be more efficient compared to the conventional management, which neglects distinguishing water quality changes due to the shift of natural conditions and anthropogenic activities.

The aim of water quality management is to either maintain the existing conditions or to improve degraded water quality. The required management actions for reducing impacts from human activities and consequently the pollutant loadings are often determined based on whether the identified thresholds or targets from water quality assessments are violated or surpassed. In particular, trends of water quality would be evaluated in a way through observing both the average (50% percentiles) and the extreme (95% percentiles) water quality conditions over time for ensuring no further degradation in water quality (NSWA, 2010). Unfortunately, some water quality management strategies neglect the intra-annual variation of water quality posed by the shift of hydro-meteorological conditions, regardless of the understanding of the roles of hydro-meteorological factors placing on water quality. Most recently, the movement towards the use of separate statistics derived from stratified data according to the selected flow ranges has been initiated, for instance, the site-specific water quality objectives proposed for the North Saskatchewan River (NSWA, 2010).

As demonstrated in the results of this paper, different divisions of flow regimes might be required for different water quality parameters, for example, DO and specific conductance at the upstream station illustrated in Figures 2(a) and 2(d). The relationship of DO and flow appears to be different between low flows (< 20 m^3/s) and high flows (> 20 m^3/s); whereas the relationship of specific conductance and flow is different between flow below 60 m^3/s and flow above 60 m^3/s. As a result, it is not very intuitive to divide the flow regimes based on the data shown in Figure 2. In addition for high dimensional data, for example DO, water temperature and flow, among which water temperature is also correlated with flow, the subjective data division is hence a challenging task. However in the proposed approach, the statistical distribution of water quality data given any conditions can be derived in terms of a conditional distribution. This can overcome the challenge in stratifying data, which is largely dependent on subjective judgment, for water quality characterization.

The data stratification approach based on the pre-selected flow ranges can, to some extent, take the flow into consideration. It, however, should be noted that various statistics can be

obtained using different ranges of influential factors. As such, the derived numerical numbers might be biased. This implies that the selection of the flow ranges tends to lack statistical justification if without quantifying the causal-effect relationship between variables. As discussed previously, the division of regimes can be different among water quality parameters as their relationship with the influential factors may vary. In this respect, how to stratify the data would indeed challenge the implementation of the data stratification approach. Therefore, the causal-effect relationship between variables, such as that between water quality and hydro-meteorological variables, should be assessed in water quality assessment process. In the proposed approach, the casual-effect relationships between variables, represented in the form of the covariance, are incorporated into the analysis processes and thus they are reflected into the results. Therefore this approach demonstrates its capability in providing a complete statistical explanation of the variable's variation, which would benefit in developing efficient management strategies.

4.3. Future Work Recommendations

This paper discussed the applicability of the proposed approach in the bivariate context, however in theory the proposed approach can be employed for analyzing higher dimensional data. The application to higher dimensional data also appears to be of practical significance since a water quality parameter can be affected by multiple hydro-meteorological variables and other ambient environmental parameters. On the other hand, copula method has recently been employed for multivariate probabilistic analysis, for instance, multivariable hydrological frequency analysis. It has some advantages in relaxing the distribution limitations of the underlying data so that it can treat various data with different statistical distribution; however, it is not applicable to censored multivariate data. In water quality data analysis, multivariate analysis has been presented to be promising, however further research on some issues such as statistical evaluation of the proposed method, optimal mixture model and application to higher dimensional data are required. In addition, future research on the linkage between the statistical analysis and the physical processes is recommended, as it is crucial to formulate more efficient water quality management strategies. As the dependence of two variables can be different under different regimes of the independent variable, regime division (such as flow regime division) is required in order to link the statistical analysis to the physical process. As illustrated in the results from this paper, it is very promising to use the derived joint distribution to divide different regimes and further research on this topic is needed.

5. Conclusions

This paper proposed a multivariate probabilistic analysis approach to examine the quantitative probabilistic causal relationship between water quality and hydro-meteorological variables. The applicability of the proposed approach was demonstrated by the numerical study of real water quality data on the Bow River, Alberta, Canada. The potential for improving water quality management was also discussed. As indicated in the results, the approach is capable of bridging the water quality and the effects of its influential factors in a probabilistic framework, thus providing more efficient way to identify water quality thresholds/targets and to solve relevant problems, for example, the cause of water quality violation complicating water quality management.

Acknowledgement. This research was financially supported by the NSERC Discovery Grant and the start-up funding from the University of Calgary, Canada. The author thanks Alberta Environment and Sustainable Resource Development as well as the Water Survey of Canada for the data source.

References

Bilotta, G.S., and Brazier, R.E. (2008). Understanding the influence of suspended solids on water quality and aquatic biota. Water Res., 42(12), 2849-2861. http://dx.doi.org/10.1016/j.watres.2008.03.018

CCME. (2003). Canadian Water Quality Guidelines for The Protection of Aquatic Life, Canadian Council of Ministers for the Environment.

Environment Canada. (2011). Water Quality Status and Trends of Nutrients in Major Drainage Areas of Canada, Cat, No.: En154-63/2011E-PDF.

Frey, H.C., and Rhodes, D.S. (1998). Characterization and simulation of uncertain frequency distributions: effects of distribution choice, variability, uncertainty, and parameter dependence. Hum. Ecol. Risk Assess., 4, 423-468. http://dx.doi.org/10.1080/10807039891284406

Government of Alberta. (2012). Guidance for Deriving Site-Specific Water Quality Objectives for Alberta Rivers. Policy Division, Alberta Environment and Water.

He, J. (2013). Mixture model based multivariate statistical analysis of multiply censored environmental data. Adv. Water Res., 59, 15-24. http://dx.doi.org/10.1016/j.advwatres.2013.05.001

He, J., Chu, A., Ryan, M.C., Valeo, C., and Zaitlin, B. (2011). Abiotic influences on dissolved oxygen in a riverine environment. Ecol. Eng., 37(11), 1804-1814. http://dx.doi.org/10.1016/j.ecoleng.2011.06.022

Helsel, D. (2010). Much ado about next to nothing: incorporating nondetects in science. Ann. Occup. Hyg., 54(3), 257-262. http://dx.doi.org/10.1093/annhyg/mep092

Hoffman, H.J., and Johnson, R.E. (2011). Estimation of multiple trace metal water contaminants in the presence of left-censored and missing data. J. Environ. Stat., 2(2), 1-16.

Interlandi, S.J., and Crockett, C.S. (2003). Recent water quality trends in the Schuylkill River, Pennsylvania, USA: a preliminary assessment of the relative influences of climate, river discharge and suburban development. Water Res., 37(8), 1737-1748. http://dx.doi.org/10.1016/S0043-1354(02)00574-2

Kotz, S., Balakrishnan, N., and Johnson, N.L. (2000). Continuous Multivariate Distributions, Vol. 1: Models and Applications, 2ed, John Wiley & Sons, New York. http://dx.doi.org/10.1002/0471722065

Li, W., Zhang, H.T., Zhu, Y., Liang, Z.W., He, B., Hashmi, M.Z., Chen, Z.L., and Wang, Y.S. (2015). Spatiotemporal classification analysis of long-term environmental monitoring data in the northern part of lake Taihu, China by using a self-organizing map. J. Environ. Inf., 26(1), 71-79.

NSWA (North Saskatchewan Water Alliance). (2010). Proposed reach-specific water quality objectives for the mainstem of the North Saskatchewan River. http://www.nswa.ab.ca/userfiles/NSWA2010-proposed-site-specific-water-quality-objectives.pdf.

Panda, U.C., Sundaray, S.K., Rath, P, Nayak, B.B., and Bhatta D. (2006). Application of factor and cluster analysis for characterization of river and estuarine water systems - A case study: Ma-

hanadi River (India). J. Hydrol., 331(3-4), 434-445. http://dx.doi.org/10.1016/j.jhydrol.2006.05.029

Poff, N.L., Allan, J.D., Bain, M.B., Karr, J.R., Prestegaard, K.L., Richter, B.D., Sparks. R.F., and Stromberg, J.C. (1997). The natural flow regime. Bioscience, 47(11), 769-784. http://dx.doi.org/10.2307/1313099

Poole, G.C., Dunham, J.B., Keenan, D.M., Sauter, S.T., McDullough, D.A., Mebane, C., Lockwood, J.C., Essig, D.A., Hicks, M.P., Sturdevant, D.J., Materna, E.J., Spalding, S.A., Risley, J., and Deppman, A. (2004). The case for regime-based water quality standards. Bioscience, 54(2), 155-161. http://dx.doi.org/10.1641/0006-3568(2004)054[0155:TCFRWQ]2.0.CO;2

Singh, K.P., Malik, A., Mohan, D., and 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(18), 3980-3992. http://dx.doi.org/10.1016/j.watres.2004.06.011

Singh, A., and Nocerino, J. (2004). Robust estimation of mean and variance using environmental data sets with below detection limit observations, Chemometr Intell. Lab. Syst., 60(1), 69-86.

Shrestha, S., and Kazama, F. (2007). Assessment of surface water quality using multivariate statistical techniques: A case study of the Fuji river basin, Japan. Environ. Model. Soft., 22(4), 464-475. http://dx.doi.org/10.1016/j.envsoft.2006.02.001

Stow, C.A., and Borsuk, M.E. (2003). Assessing TMDL effectiveness using flow-adjusted concentrations: A case study of the Neuse River, North Carolina. Environ. Sci. Technol., 37(10), 2043-2050. http://dx.doi.org/10.1021/es020802p

Tan, X., Xia, X.L., Li, S.Y., and Zhang, Q.F. (2015). Water quality characteristics and integrated assessment based on multistep correlation analysis in the sanjiangkou reservoir, China. J. Environ. Inf., 25(1), 60-70. http://dx.doi.org/10.3808/jei.201500296

Titterington, D.M., Smith, A.F.M., and Makov, U.E. (1985). Statistical Analysis of Finite Mixture Distribution, Wiley, New York.

US EPA. (2004). The National Water Quality Inventory: Report to Congress for the 2004 Reporting Cycle, Report, No. EPA 841-R-08-001.

US EPA. (2003). Strategy for Water Quality Standards and Criteria, Office of Science and Technology.

Wang, Y., Ma, H., Sheng, D., and Wang D. (2012). Assessing the interactions between chlorophyll a and environmental variables using copula method. J. Hydrol. Eng., 17(4), 495-506. http://dx.doi.org/10.1061/(ASCE)HE.1943-5584.0000387

Xia, X.H., Wu, Q., Mou, X.L., and Lai, Y.J. (2015). Potential impacts of climate change on the water quality of different water bodies. J. Environ. Inf., 25(2), 85-98. http://dx.doi.org/10.3808/jei.201400263