The 2014 International Conference on Marine and Freshwater Environments

Joint Conferences:
The 2014 Annual Conference of the International Society for Environmental Information Sciences (ISEIS)
The 2014 Atlantic Symposium of the Canadian Association on Water Quality (CAWQ)
The 2014 Annual General Meeting and 30th Anniversary Celebration of the Canadian Society for Civil Engineering Newfoundland and Labrador Section (CSCE-NL)
The 2nd International Conference of Coastal Biotechnology (ICCB) of the Chinese Society of Marine Biotechnology and Chinese Academy of Sciences (CAS)

# Evaluation of the Causal Relationship between Variables Using a Probabilistic Approach for Water Quality Management

Jianxun He[1,*]

[1]*Department of Civil Engineering, Schulich School of Engineering, University of Calgary, 2500 University Drive, T2N 1N4 Calgary, Alberta, Canada*

**Corresponding author's e-mail:* jianhe@ucalgary.ca

## Abstract

In aquatic environments, complex interplay exists among physical, chemical, and biological water quality parameters, which are further influenced by exogenous factors such as hydrological, meteorological and geological conditions. To understand the spatial and temporal variations of water quality, and furthermore, the relationships between the variables of interest is hence a challenging task. Given the large data matrix, one category of methods frequently employed in the literature is multivariate analysis such as cluster analysis, principal component analysis, and factor analysis. These techniques are straightforward and intuitive to identify the qualitative associations among variables. 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 will illustrate 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 developed expectation maximization algorithm, which is capable of dealing with multiple variables and censored data. The joint distribution of two variables of interest and the conditional distributions were used to describe the complete statistical distribution of water quality parameters, which are subject to the effects of hydro-meteorological conditions. The method was demonstrated using data from the Bow River in Alberta. 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

## Introduction

In water resources management, water quality has continuously attracted attention at national, provincial, and municipal levels in the recent decades. Water quality monitoring programs have been established in many aquatic environments, for example in Canada, to capture the water quality level in the natural water bodies and spatial and temporal characterization of water quality for management purposes. Knowledge obtained from water quality monitoring and following assessment/or characterization and modelling has often been employed to formulate management strategies, for instance for reducing pollutant loads into water bodies, to protect water sources and ecosystems.

It has been acknowledged that the complexity of water quality processes, including physical, chemical and biological processes, occurring in natural water bodies challenges our understanding of water quality issues and consequently presents major obstacles to the

development and implementation of water quality models. Therefore, to date, water quality monitoring and assessment have been playing very important roles in identifying and resolving water quality issues. With the gradual increase of the water quality data, water quality managers and researchers have searched for effective approaches to statistically present the data and subsequently to formulate feasible water quality management objectives/or targets, upon which to base more effective management decisions. Water quality phenomena are multidimensional. Water quality processes intertwine with each other and exogenous factors including geological, meteorological, and hydrological conditions, which affect the spatial and temporal variation of water quality. Recently, multivariate statistical techniques such as cluster analysis, principal component analysis, factor analysis, and discriminant analysis have been employed to interpret complex data sets of surface waters (Panda et al., 2006; Shrestha and Kazama, 2007; Singh et al., 2004). Despite the effectiveness of these methods in qualitatively identifying critical influential factors and spatial and temporal variations, they however do not quantitatively present the linkage between two or more variables of interest. In surface water bodies, hydro-meteorological response of water quality could largely vary under different conditions. For example, the variation of dissolved oxygen (DO) levels in a river could be largely explained by flow and water temperature, while their roles varies under different conditions, such as high, medium, and low flows (He et al., 2011). The influence of suspended solids on water quality and aquatic biota varies under different hydrological regimes (such as flood conditions and base-flow conditions) (Bilotta and Brazier, 2008). In addition, the dependence of water quality on the natural conditions may confound the identification of the water quality limits for management and the cause of water quality degradation (human activities or changes in natural conditions) (Poole et al., 2004), and the assessment of management effectiveness (Stow and Borsuk, 2003). The aforementioned 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).

Due to the recognition of the importance of the associations between water quality and hydro-meteorological conditions, water quality management attempts to take them into consideration. For example, water quality data are stratified according to different seasons and/or flow conditions and subsequently employed to identify different management targets for different conditions (CCME, 2003; Government of Alberta, 2012). However in most cases, the division of the flow conditions has been usually arbitrary and lack of statistical justification. All these suggest the needs of the probabilistic characterization of water quality conducted in a multivariate context such that the associations between two or more variables can be taken into account. The multivariate distribution analysis, from which the probability distribution of one variable conditioned upon the values of the remaining variables can be readily derived, has showed its potentials for this purpose. Most recently, Hoffman and Johnson (2011) and Wang et al. (2012) employed the multivariate distribution analysis to assess the overall contamination level of several dissolved trace metals and to investigate the complicated linkages between chlorophyll *a* and ambient water quality, respectively. Therefore, the primary objective of this paper is to develop an efficient multivariate distribution analysis, which can statistically characterize data while accounting for the dependence between two or more variables. Similar to univariate distribution analysis, two issues: i) underlying distribution of data, and ii) observations below detection limits (DLs), also need to be tackled to properly represent the statistical characteristics of water quality data. The proposed methodology by He (2013) employed

Gaussian mixture model (GMM) and the expectation maximization (EM) algorithm to describe data distribution and to estimate distribution parameters, respectively. This paper extends the above methodology to derive joint and conditional distributions and further illustrates the applications of the methodology using data collected from the Bow River in southern Alberta, Canada.

## Materials and Methods

### The study area

The Bow River originates from the Rocky Mountains in Alberta, Canada, and flows towards east. In the upper watershed located within the Banff National Park, the river flows through largely undeveloped and low intensity agricultural land; the water quality is fairly good. 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 besides providing drinking water to over half Calgary's population. During December and March, the river is usually covered or partially covered by ice; while open water generally begins in April. Flow peaks around June or July annually, driven by both rain events and snowmelt.

### 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., urbanization). There are 5 long-term water quality monitoring stations on the river. The water quality data collected on a monthly basis between 1988 and 2009 from two long-term monitoring stations, one located in the upstream of the river about 4.5 km above Canmore and the other located just the upstream of the confluence with the South Saskatchewan River (near Ronalane Bridge), were used in this paper. These two stations are called the upstream and downstream stations, respectively, throughout this paper. Since this paper did not target any specific water quality parameters, data including DO, water temperature, turbidity, and dissolved total phosphorus (TP), were selected. The data sizes range from 258 to 278. All data of DO, water temperature and turbidity are above DL; while 20% of dissolved TP data is censored.

Daily flow data collected by the Water Survey of Canada on the Bow River at Banff and the Bow River Near the Mouth, which are in close proximity to the upstream and downstream water quality monitoring stations, respectively, were used. The flow data corresponding to the water quality sampling dates were extracted from the daily flow data sets for the analysis. Water temperature will be viewed as a meteorological factor in this paper, since it is usually strongly associated air temperature and largely affects chemical and biological reactions.

### Methodologies

*Gaussian mixture model*. The GMM is a desirable parametric model, which can approximate the unknown probabilistic distribution of many water quality parameters (He, 2013). The typical finite GMM, which is an additive model, is given by

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

where **x** is an *M*-dimensional data vector distributed according to $p(\mathbf{x}|\Theta)$ parameterized by $\Theta$; $\alpha_k$ is the nonnegative weight for the *k*-th component of the GMM; $p_k(\mathbf{x}|\mu_k,\Sigma_k)$ is the *k*-th Gaussian component with mean $\mu_k$ and covariance matrix $\Sigma_k$; i.e., $p_k(\mathrm{x}|\mu_k,\Sigma_k)=(2\pi)^{-M/2}|\Sigma_k|^{-1/2}\exp\{-\frac{1}{2}(\mathrm{x}-\mu_k)^{\mathrm{T}}\Sigma_k^{-1}(\mathrm{x}-\mu_k)\}$, where the superscript T and |稹denote the transpose and determinant of a matrix, respectively. Furthermore, the summation of all weights must be equal to 1. Therefore, the parameter set $\Theta$ is the set of all parameters appeared in the right side of (1) and can be represented by $\Theta=\{\alpha_k,\ \mu_k,\Sigma_k\,|\,k=1,\cdots,K\}$.

In addition, 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 **x** is a mixture of *K* multivariate normal distributions with the same weights (Kotz et al., 2000; Titterington et al., 1985).

*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 Titterington et al. (1985) and He (2013).

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

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

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

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 multiply 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 **x** of the length of $N_0$. The estimation algorithm derived by He (2013) are summarized as follows:

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

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

where $\Theta^p$ 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

$$C = \sum_{i=1}^{N_0} p(k \mid x_i, \Theta^p) + \sum_{i=1}^{N_d} p(k \mid z_i^d, \Theta^p)$$

$$D = \sum_{i=1}^{N_0} x_i p(k \mid x_i, \Theta^p) + \sum_{i=1}^{N_d} p(k \mid z_i^d, \Theta^p) \int_{\Omega_i} x_i p(x_i \mid z_i^d, k, \Theta^p) dx_i$$

$$G = \sum_{i=1}^{N_0} p(k \mid x_i, \Theta^p) \hat{\Lambda}_{k,i} + \sum_{i=1}^{N_d} p(k \mid z_i^d, \Theta^p) \int_{\Omega_i} \Lambda_{k,i} p(x_i \mid z_i^d, k, \Theta^p) dx_i$$

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

*Conditional distribution.* 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 $x \sim \mathrm{N}(\mu_x, \sigma_x^2)$ and $y \sim \mathrm{N}(\mu_y, \sigma_y^2)$, respectively, the conditional distribution of *y* given *x* is

$$y \mid x \sim \mathrm{N}\left(\mu_y + \frac{\sigma_y}{\sigma_x}\rho(x - \mu_x), (1-\rho^2)\sigma_y^2\right) \tag{7}$$

where $\rho$ is the correlation coefficient between *x* and *y*. For data reasonably described by the GMM, the conditional probability distribution can be computed by (8).

$$p(y \mid x) = \sum_k^K \alpha_k \frac{p_k(x, y \mid \Theta)}{p_k(x \mid \Theta)} = \sum_k^K \alpha_k p_k(y \mid x, \Theta) \tag{8}$$

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 assessed.

## Results

This section demonstrated the applications of above algorithms to derive the joint distribution and the conditional distributions using real observations. To illustrate the applicability of this proposed approach for both the uncensored and censored data sets, the results from these two scenarios along with variations of hydrological and water quality data were presented herein.

### Variations of hydrological and water quality variables and their dependence

In the river, both flow and water quality, in general, present seasonal variations. Figure 1 displays the boxplots of flow and DO, respectively, at the downstream station as examples. It is obvious that lower DOs generally 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 flow conditions. As demonstrated in these figures,

the dependence of the water quality parameters on flow appears 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 (Figure 2(a)). Similar to DO at the upstream station, the dependences of both turbidity and dissolved TP on flow appear different under low and high flow conditions at the downstream station (Figures 2(b) and 2(c)).

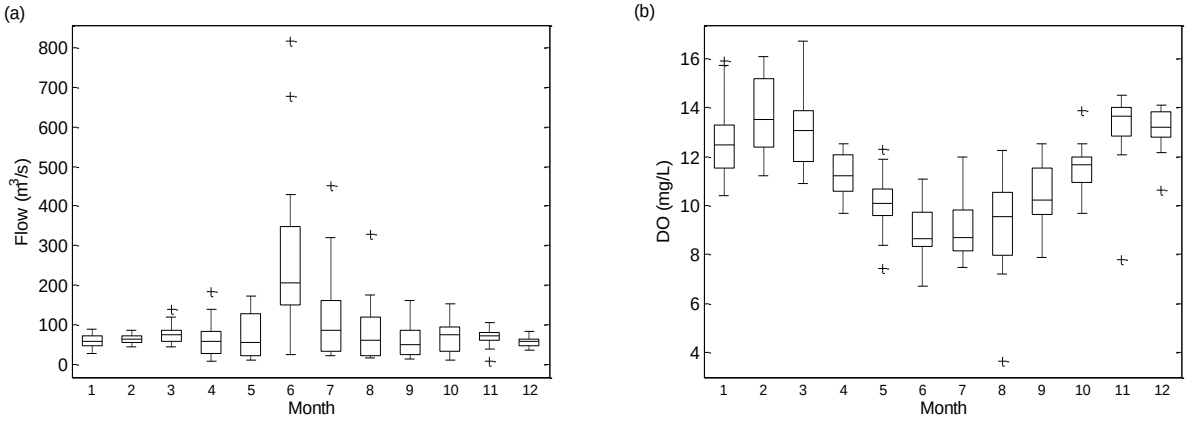


**Figure 1**. Seasonal variations of (a) flow and (b) DO at the downstream station

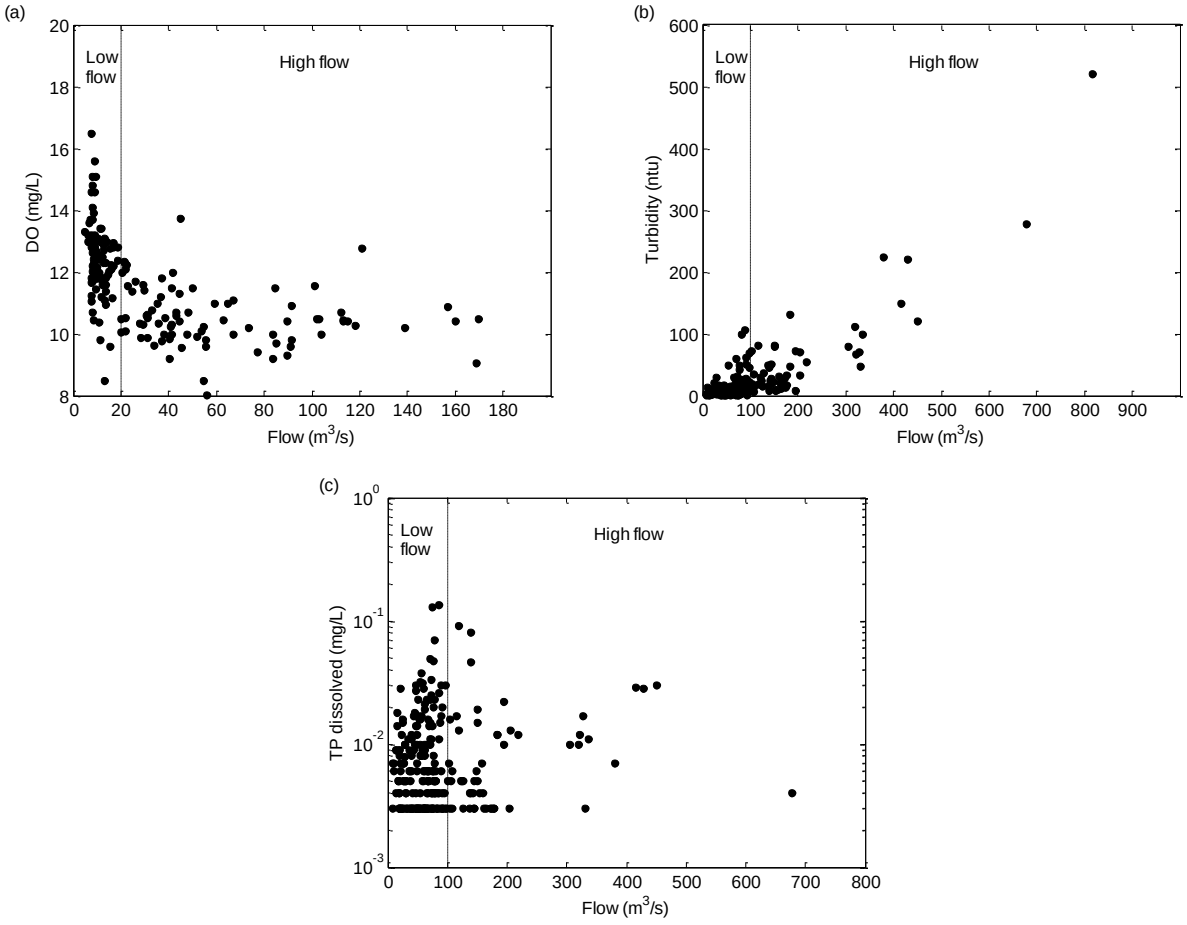

**Figure 2**. Scatter plots of (a) flow and DO at the upstream station, (b) flow and turbidity at the downstream station, and (c) flow and dissolved TP at the downstream station.

Applications to uncensored and censored data

For the uncensored case, the data of flow and DO observed at the upstream station, both of which are not censored, were used as an example. As shown in Figure 3(a), different distributions 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 3(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 4 - 6 illustrate the results for flow and DO at the downstream station, 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 (Figures 3 and 4).

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 this data set, dissolved TP data are censored (DL=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 displayed in Figure 7(a) and 7(b), respectively. Differences in the conditional cumulative probability distributions under different flow conditions can be observed, although the differences are not as prominent as those in Figures 3(b), 4(b) and 5(b).

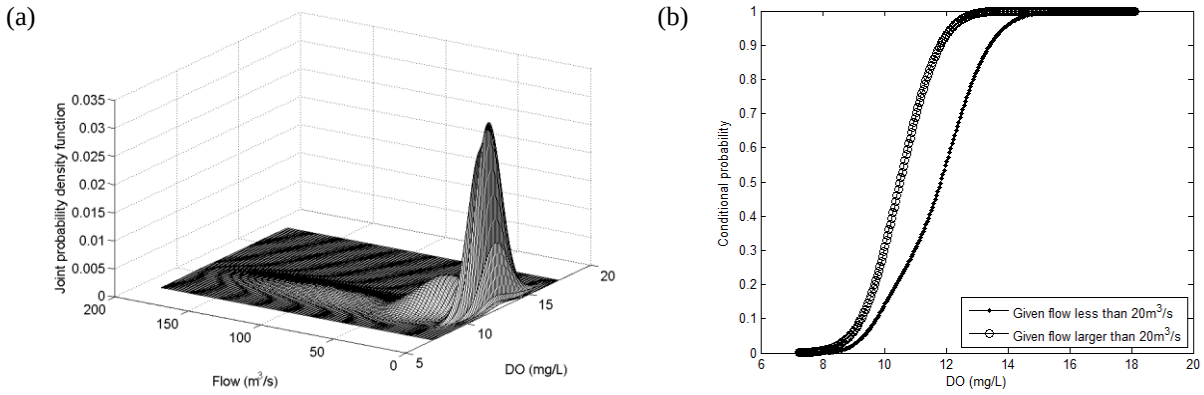


**Figure 3**. Results of (a) the joint probability density function between flow and DO and (b) the conditional cumulative probability distributions of DO given flows at the upstream station

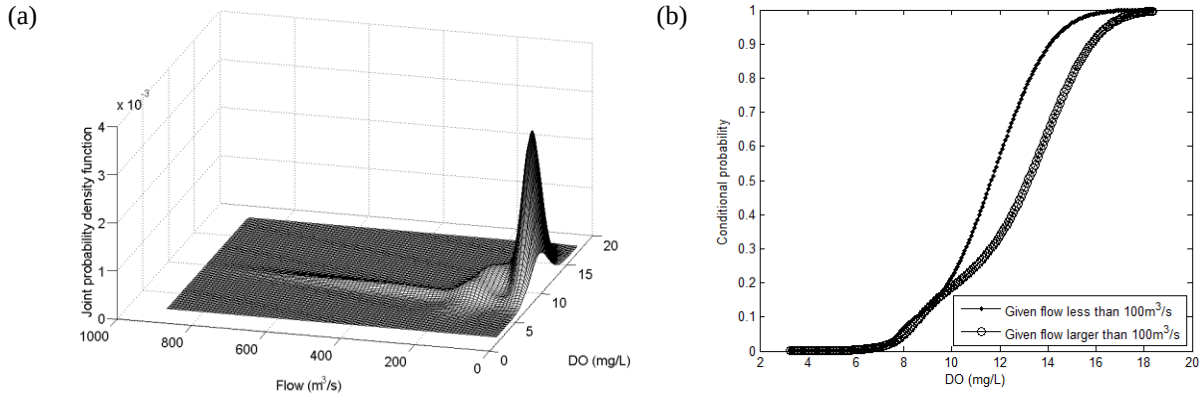

**Figure 4**. Results of (a) the joint probability density function between flow and DO and (b) the conditional cumulative probability distributions of DO given flows at the downstream station

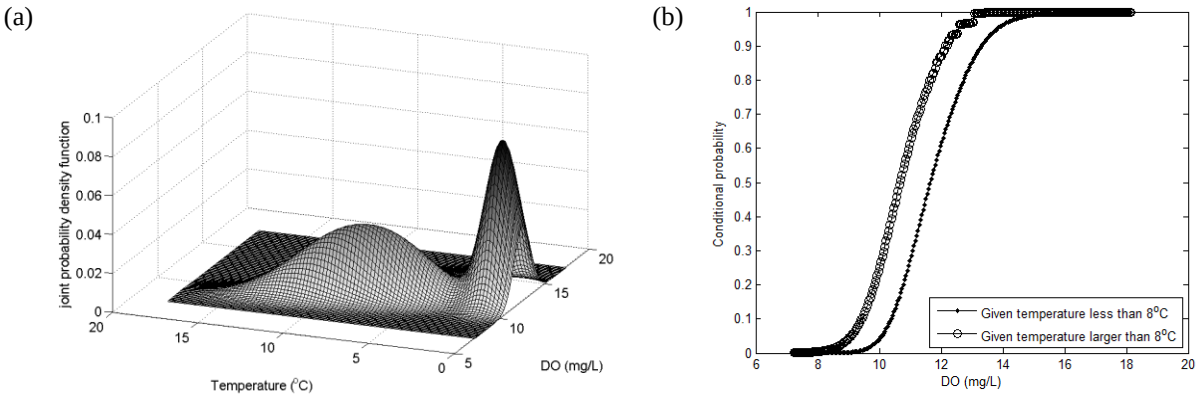


**Figure 5**. Results of (a) the joint probability density function between water temperature and DO and (b) the conditional cumulative probability distributions of DO given temperatures at the upstream station

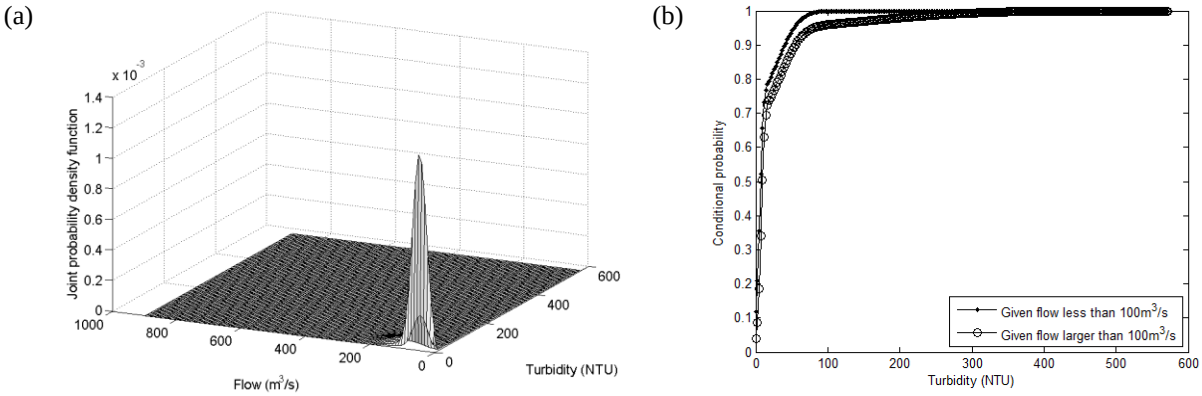


**Figure 6**. Results of (a) the joint probability density function between flow and turbidity and (b) the conditional cumulative probability distributions of turbidity given flows at the downstream station

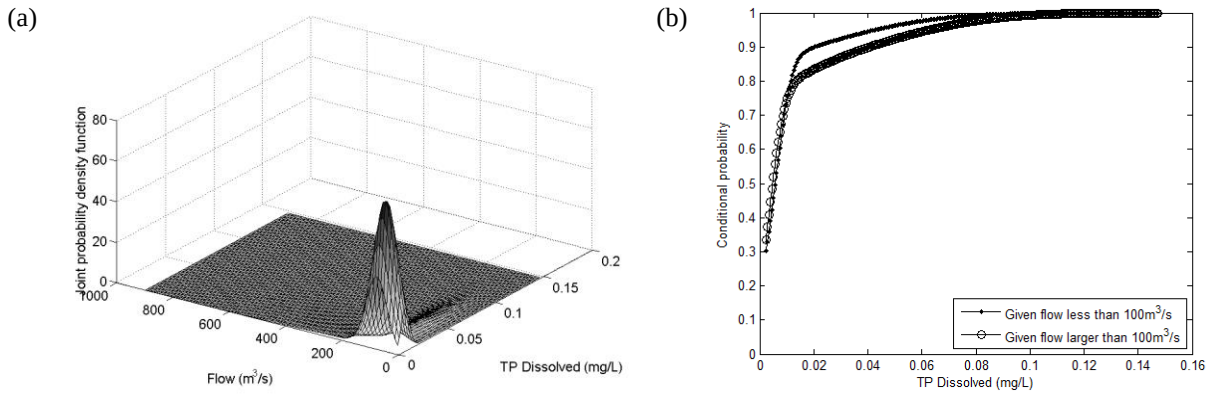


**Figure 7**. Results of (a) the joint probability density function between dissolved TP and flow and (b) the conditional cumulative probability distributions of dissolved TP given flows at the downstream station

## Discussions

In aquatic environments, the causal relationships between water quality parameters and hydro-meteorological variables are very difficult to be represented by physically-based models or simple empirical models obtained from statistical analysis, due to the fact that the governing mechanisms of water quality are complicated. The governing mechanisms of transport and/or the sources of a pollutant can be different under different hydro-meteorological conditions. For example under high flows, both flow and temperature might play significant roles on DO levels; while the biological processes, photosynthesis and respiration of periphyton and macrophytes, could override the roles of hydro-meteorological factors under low flows. Thus, it is practical to conduct water quality management considering different hydro-meteorological conditions, which shift annually and would lead to the shift of their roles on water quality in an intra-annual scale. As the results demonstrated in this paper, the distribution of water quality parameters can be linked to hydro-metrological conditions using this proposed approach.

The aim of water quality management is to either maintain the existing conditions or to improve degraded water quality. The required management actions are often determined based on the identified thresholds or targets from water quality assessments, such as using 90 percentiles of pollutant concentrations. However, regardless of the understanding of the roles of hydro-meteorological factors, current water quality management usually ignore the intra-annual variation of water quality posed by the shift of hydro-meteorological conditions. As demonstrated in this paper, different management thresholds can be obtained under given different hydro-meteorological conditions (e.g., low and high flows). This suggests that the determination of the thresholds for management ignored the complicate causal relationship is likely to lead to inefficient management. For instance, the water quality violation may not be detected. In addition, assessing the effectiveness of pollutant management actions in reducing pollutant loads may result in unexpected results if without taking into account the effects of flow on water quality (Stow and Borsuk, 2003). Therefore, the causal relationships between water quality and hydro-meteorological variables should be addressed in water quality assessment for developing efficient management. The proposed approach in this paper, in which the dependence of two variables is considered when conducting the statistical analysis using multivariate analysis techniques, provides the distribution of a variable derived upon a quantitative description of its relationship with the other variable(s). Therefore, it provides a complete statistical explanation of the variable's variation, which would benefit in developing efficient management strategies.

The results obtained from the real applications demonstrated the applicability and the potential of the proposed approach for enhancing water quality management. This approach can provide the decision-makers and water quality managers probabilistic distributions of a water quality parameter under a given hydrological conditions or the level of another water quality parameter. This approach can also assist in distinguishing either natural (hydrological and meteorological) or anthropogenic causes of changes in water quality, if the probabilistic distribution is given conditioned on the natural conditions; namely the effects of the natural conditions can be removed in the water quality assessment. This paper only demonstrated the applicability of this approach in the bivariate context; while higher dimensions might be required since a water quality parameter can be affected by both hydrological and climatological variables and other ambient environmental conditions. On the other hand, copula has been recently commonly used for multivariate probabilistic analysis, for instance, multivariable hydrological frequency analysis. Copula has advantages in its applications due to the fact that it is relatively easy to implement and has no limitations on the distribution of each variable in the analysis. For water

quality data analysis, multivariate analysis has been presented to be promising, however further research on some issues, such as different distribution family (not only Gaussian distribution) and high dimensions (trivariate or more) are recommended.

## Conclusions

This paper proposed a multivariate probabilistic analysis approach to take the dependences of water quality and hydro-meteorological variables into account. Through applying this proposed approach to real water quality data collected on the Bow River, the applicability of this approach was demonstrated and the potential for improving water quality management were discussed. As demonstrated in the results, this approach is capable of bridging the water quality and the effects of its influential factor(s) in a probabilistic framework, thus providing more efficient way to identify water quality thresholds/targets and problems, for example, the cause of water quality violation, for management purposes.

## Acknowledgement

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

## References

Bilotta G.S. and Brazier R.E. (2008). Understanding the influence of suspended solids on water quality and aquatic biota. *Water Research*, 42, 2849-2861.

CCME. (2003). *Canadian water quality guidelines for the protection of aquatic life*. Canadian Council of Ministers for the Environment.

Frey H.C. and Rhodes D.S. (1998). Characterization and simulation of uncertain frequency distributions: effects of distribution choice, variability, uncertainty, and parameter dependence. *Human and Ecological Risk Assessment*, 4, 423-468.

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. *Advances in Water Resources*, 59, 15-24.

He J., Chu A., Ryan M.C., Valeo C., and Zaitlin B. (2011). Abiotic influences on dissolved oxygen in a riverine environment. *Ecological Engineering*, 37, 1804-1814.

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

Kotz S., Balakrishnan N., and Johnson N.L. (2000). *Continuous multivariate distributions*, vol. 1: models and applications, 2ed. John Wiley & Sons, New York

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: Mahanadi River (India). *Journal of Hydrology*, 331, 434-445.

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. *BosScience*, 54(2), 155-161.

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 Research*, 38, 3980-3992.

Shrestha S. and Kazama F. (2007). Assessment of surface water quality using multivariate statistical techniques: A case study of the Fuji river basin, Japan. *Environmental Modelling and Software*, 22(4), 464-475.

Stow, C.A. and Borsuk, M.E. (2003). Assessing TMDL effectiveness using flow-adjusted concentrations: A case study of the Neuse River, North Carolina. Environmental Science and Technology, 37, 2043-2050.

Titterington D.M., Smith A.F.M., and Makov U.E. (1985). *Statistical analysis of finite mixture distribution*. Wiley, New York.

Wang Y., Ma H., Sheng D., and Wang D. (2012). Assessing the interactions between chlorophyll a and environmental variables using copula method. *Journal of Hydrological Engineering*, 17(4), 495-506.