19th International Congress on Modelling and Simulation, Perth, Australia, 12–16 December 2011
http://mssanz.org.au/modsim2011

Application of optimization techniques to water quality monitoring designs

M.G. Erechtchoukova [a] and P.A. Khaiter [a]

[a] *School of Information Technology, Faculty of Liberal Arts and Professional Studies, York University, Canada*
Email: marina@yorku.ca

Abstract: The issues of possible improvements, increased efficiency and/or optimization of monitoring systems in general, and monitoring designs, in particular, attract the attention of researchers for years. Traditionally, these issues are addressed using expert knowledge and heuristic approaches to monitoring system development under the budgetary constraints. Application of formal techniques for these purposes looks appealing since it may validate suggested procedures or justify expenses required for data collection. The paper describes an approach to the development of sampling programs as a solution of the operation research model articulated in terms of the cost-effectiveness analysis. The effectiveness of a sampling program is described through the uncertainty of the estimates obtained from water quality data collected in accordance with the sampling program. Since several water quality indicators, with different temporal and/or spatial variability, are determined from the same water sample, monitoring designs for these indicators must be compromised in a way to ensure a required level of efficiency for all indicators being detected.

The proposed approach is based on the operation research model which minimizes the total number of water samples been collected over an investigated period of time under the condition that the uncertainty of an estimate derived from the monitoring data is kept below an acceptable level. The efficient monitoring designs are determined as the solutions of this model. Since concentrations of water constituents exhibit different variability, the numbers of observations required to achieve the same uncertainty level in their estimates vary significantly, even for those water constituents whose concentrations are derived from the same grab water sample. In order to make a practically meaningful recommendation on the frequencies of observations, it is necessary to comprise temporal monitoring designs for all water quality indicators from the same water sample. Given that concentrations of these parameters form under common hydrological and climatic conditions, it is reasonable to assume that series of concentrations are somehow related. It had been shown that if such dependencies are detected, they can be used to significantly reduce the total number of observations required for water quality assessment.

The proposed approach has been tested on observation data collected on the Humber River (Ontario, Canada). The major ions, namely, calcium, carbon, magnesium, and potassium have been selected for the study. Since monitoring data can be used for various purposes, simple random designs supporting the evaluation of basic statistics of the investigated water quality indicators are preferable. The relationships between concentrations of the investigated water constituents were described by linear regression models. These models were used in the proposed operation research model to obtain efficient monitoring designs supporting estimation of water quality indicators with a given level of uncertainty. These designs are common for all investigated water quality indicators detected from the same water sample at the Old Mill Road cross-section of the Humber River.

The proposed operation research model can be applied to tiered monitoring systems where water quality indicators of interest are split in two sets: core and supplemental, according to their importance for a given site with different accuracy requirements. The designs common for all water quality indicators measured at the given site may result in a higher number of observations. Depending on the desired level of accuracy, it may lead to daily sample collection. The proposed approach may help to develop efficient monitoring designs with the reasonable cost of sampling by considering subsets of the water quality indicators.

Keywords: *Water quality monitoring, temporal monitoring design, operation research model*

1. INTRODUCTION

Water quality monitoring systems are an integral part of such activities as the environmental impact assessment and decision-making related to water resources. The systems are also the main source of data on the status of the aquatic environment which is described by a set of physical characteristics, concentrations of various substances in natural water, and indicators of water ecosystem states. These data are collected to support general goals including formulation of water quality standards, attainment of the standards, identification of impaired waters, as well as causes and sources of water quality impairments and detection of long-term trends (US EPA, 2003) and certain site-specific or project-specific needs. Normal functioning of monitoring systems requires such operations as collection and analysis of physical, chemical, and biological data, as well as quality assurance and control programs to ensure that the data are scientifically valid. For this purpose, the Canada-wide framework for water quality monitoring identifies the following important phases of monitoring which must be included in any system: (1) formulation of monitoring program objectives, (2) monitoring program design, (3) field sampling program, (4) laboratory analysis and procedures, (5) data analysis and interpretation, (6) reporting and follow-up (WQTG, 2006). While all these phases are mandatory for monitoring activities, the study deals only with the second phase, namely developing monitoring designs.

A network of sampling sites where sample collection and observations are implemented is a key component of a monitoring system. There are several topologies for networks of monitoring sites. Their overview is presented in (US EPA, 2003). Observations and measurements at a sampling site can be done routinely by collecting a grab sample at certain time intervals or by using automatic samplers. Although automatic samplers are able to determine values of concentrations of water constituents with high frequencies, they cannot replace routine water sample collection followed by a laboratory analysis due to the following reasons. First, these automatic tools can determine values only for a limited set of water constituents. Second, limited budgets of monitoring systems cannot afford a large number of such devices, thus leaving many monitoring sites to operate under routine sampling programs. Financial constraints force researchers and monitoring data users to seek more efficient utilization of available funds. That is why the issues of possible improvements, increased efficiency and/or optimization of monitoring systems in general, and monitoring designs, in particular, attract the attention of researchers for years.

The development of a network of sampling sites and temporal monitoring designs at the sites is guided by monitoring objectives to a great extent. Lettenmaier (1978) mapped different types of monitoring designs to main water quality monitoring objectives. A systematic analysis of monitoring objectives can be found in (Whitfield, 1988). Groot and Schilperoort (1983) proposed a framework for developing an efficient monitoring network based on cost-effectiveness analysis. This framework is very generic and can be applied to create an entire monitoring program or its part provided that monetary estimates of all the system components and the extent of attainment of monitoring goals are available. The multidisciplinary nature of a monitoring system explains different and sometimes contradicting aspects of the system which must be taken into account during its optimization. It justifies various approaches which are aimed to replace intuitive improvements of monitoring systems in order to increase their efficiency. Thus, the problem of monitoring optimization can be solved as a multi-objective mixed integer programming model with constraints (Ning and Chang, 2002), or as a constrained optimization using generic algorithms (Icaga, 2005; Cieniawski et al., 1995). Fuzzy optimization approach is also used in optimization of a water quality monitoring network (Ning and Chang, 2004).

The influence of various monitoring designs on the quality of estimates generated from the monitoring data has been studied by Shabman and Smith (2003) and Robertson and Roerish (1999). Formal approaches to the development of efficient temporal monitoring designs for a single water constituent have been already investigated elsewhere (Erechtchoukova and Khaiter, 2009; Erechtchoukova et al., 2009). A particular attention has been given to models employed for subsequent data analysis, since these models may require a data set with specific statistical properties. Although model selection can reduce the required number of observations for achieving a desired level of accuracy, these numbers are water constituent-specific and reflect their variability.

The current paper presents an approach to the development of efficient temporal monitoring designs at sites of routine water quality monitoring systems operating based on fixed stations where samples are collected in accordance with tiered monitoring programs. The approach takes into account variability and relationships between water quality indicators derived from the same grab sample, thus generating frequencies of observations common for these indicators. The approach has been tested on a case study presented in the paper. The results led to further recommendations on its applicability to different natural streams and water quality indicators.

2. PROBLEM ARTICULATION BASED ON THE OPERATION RESEARCH MODEL

Traditionally, optimization of monitoring designs is motivated by numerous complaints from researchers and practitioners about scarce data sets which are not sufficient for water related assessments. Admitting the limited funds allocated to water quality monitoring, the problem of minimizing the cost of observations in order to get required information on the status of the aquatic environment remains urgent nowadays. Contrary to the expectation that evaluation of the cost of a monitoring program is a simple technical exercise, the cost of the components comprising the total cost is not readily available. Thus, Loftis and Ward (1980) considered only the direct cost of sample collections, and even such estimates, in many cases, are not known. However, it is reasonable to assume that the cost of a monitoring program is a monotonously increasing function of the number of collected samples. This assumption implies that the goal of the optimization problem can be declared as to minimize the total number of observations, instead of minimizing the cost of the program. At the same time, data sets generated by monitoring systems must suffice scientifically valid conclusions and meet expectations of decision makers with respect to the tolerable level of uncertainty in information supporting their decisions. Since information is derived from the monitoring data based on models and statistical estimators, their uncertainty evaluated from data sets collected in accordance with a monitoring program can be used to reflect the level of efficiency of this program. As a result the operation research model which minimizes the total number of required observations under conditions that the uncertainty of the estimates does not exceed a specified level can be formulated (Erechtchoukova and Khaiter, 2010):

$$\min n \text{ subject to} \tag{1}$$

$$\left|\frac{D(I(n))}{I(n)}\right| \cdot 100\% \le V, \tag{2}$$

where I is the selected estimator, $I(n)$ is its estimate on a set of n observations, $D^2(I)$ is the variance of the estimator I, and V is the acceptable level of the uncertainty in I.

Monitoring designs providing data sufficient for keeping the uncertainty of the estimates below the required level can be obtained as solutions of the operation research model (1)-(2). These designs depend on the water quality indicators used for the estimate. Since concentrations of water constituents exhibit different variability, the numbers of observations required to achieve the same uncertainty level in their estimates vary significantly even for those water constituents whose concentrations are derived from the same grab water sample. In order to make a practically meaningful recommendation on the frequencies of observations, it is necessary to comprise temporal monitoring designs for all water quality indicators from the same water sample. The development of an efficient temporal monitoring design at a single observation site common for the water quality indicators derived from the same water sample is considered further in the paper. For this purpose, the constraint function (2) must be replaced by a set of constraints for each investigated parameter:

$$\left|\frac{D(I_k(n))}{I_k(n)}\right| \cdot 100\% \le V_k, k = 1,...,K, \tag{3}$$

where I_k is the estimator of the k-th constituent, V_k is the acceptable level of uncertainty in this estimate, and K is the total number of constituents of interest.

Since the objective function is linear, the optimal solutions are expected to be found on the border of the domain determined by the constraint functions. Hence, the solution of the model (1) and (3) recommends the number of observations which is the maximum of the numbers for all K water constituents. It may result in oversampling for many water constituents. Given that concentrations of these parameters form under common hydrological and climatic conditions, it is reasonable to assume that series of concentrations are somehow related. If such relationships are registered, they can be used to reduce the number of water samples required to achieve the established level of uncertainty in the estimates.

Considering linear regression models, the concentration C of a water constituent is estimated as:

$$C = a \cdot C_{CMV} + b, \tag{4}$$

where C_{CMV} is the concentration of the base water constituent, a and b are regression coefficients which are identified based on the least squares fitting. Then, the variance of the estimator I_C can be evaluated from the series of the constituent with the minimal variability in the following way:

$$D(I_C) = a^2 \cdot D(I_{CMV}). \quad (5)$$

Formulae (4) and (5) can be used in the constraint expressions (3) to obtain monitoring designs sufficient to estimate the average concentrations of all water constituents from the same sample with a given level of uncertainty. It is worth noting, that the higher the uncertainty of an estimate, the more observations are required to keep the uncertainty below an accepted level. That is why the formulae (4) and (5) can improve monitoring designs only if the regression coefficient a is less than 1.0. This condition provides an insight on the selection of the base water constituent and allows for a conclusion whether the improvement is possible.

An obvious advantage of the models in the form of either (1) and (3) or (1), (3)-(5) is that they do not require site-specific parameterization. The constraint functions (3)-(5) are evaluated based on the time series of concentrations of the selected water quality indicators collected at the investigated sites. The operation research model (1), (3)-(5) has been tested on observation data collected on the Humber River (Ontario, Canada).

3. CASE STUDY

The model (1), (3)-(5) had been applied to develop monitoring designs at Old Mill Rd. station of the Toronto and Region Conservation Authority monitoring network. The observation site is located at the lower main section of the Humber River (Ontario, Canada). The main branch of the Humber Rive travels more than 120 km through 908 km^2 watershed covering Niagara Escarpment, the rolling hills and kettle lakes of the Oak Ridges Moraine, the high-quality agricultural lands of the South Slope and Peel Plain, and the ancient Lake Iroquois shoreline. The Humber River is classified as a small river with the average water discharge of about 0.24 km^3/year (Figure 1). The river flows in Southern Ontario from Georgian Bay to Lake Ontario through the Greater Toronto Area, the most urbanized centre in Canada. Its waters experience significant anthropogenic impact. The major ions, namely, calcium (Ca), carbon (C), magnesium (Mg), and potassium (K) have been selected for the study. The choice can be explained by the availability of the relatively long series of concentrations. Table 1 presents basic statistics of the selected water constituents.

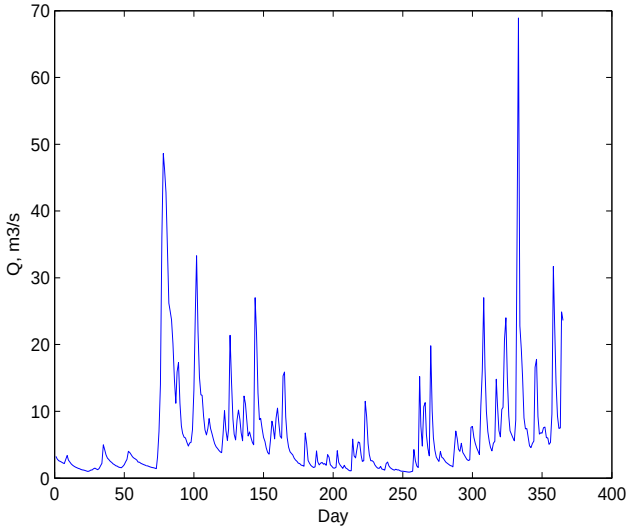

Figure 1. Water Discharge, the Humber River (Ontario, Canada)

Table 1. The investigated water constituents

Water constituent	Mean, mg/L	Variance	Coefficient of variance
Ca	80.11	297.43	0.2153
C	5.08	2.74	0.3258
Mg	16.85	19.88	0.26456
K	3.78	0.66	0.2157

Since monitoring data can be used for various purposes, simple random designs supporting the evaluation of basic statistics of the investigated water quality indicators are preferable (Shabman and Smith, 2003). Such designs have observations randomly distributed within an investigated period of time and their main characteristic is the total number of observations for the period.

Models (1)-(2) and (1), (3)-(5) can be solved using various computational algorithms. Their choice is conditioned by the estimators employed for assessment since the goal function and constraint functions must satisfy basic mathematical assumptions. In the present study, non-gradient methods are deemed to be preferable. The proposed models have been solved using the non-gradient constrained optimization method implemented in MATLAB 7.1 (Conn et al., 1997).

First, model (1)-(2) has been used to develop simple random monitoring designs for each water constituent over a year period (Figure 2). The suggested numbers of observations varied significantly for different water quality indicators. After that, model (1) and (3) with the same levels of uncertainty for all four water constituents has been used to generate monitoring designs. These designs require the same number of observations as the designs developed for the most variable investigated water constituent, i.e. carbon. Such recommendations result in oversampling for all other water quality indicators: from 53% for magnesium to 87% for calcium and potassium. In order to further synchronize monitoring designs for all four water constituents, statistical models have been employed.

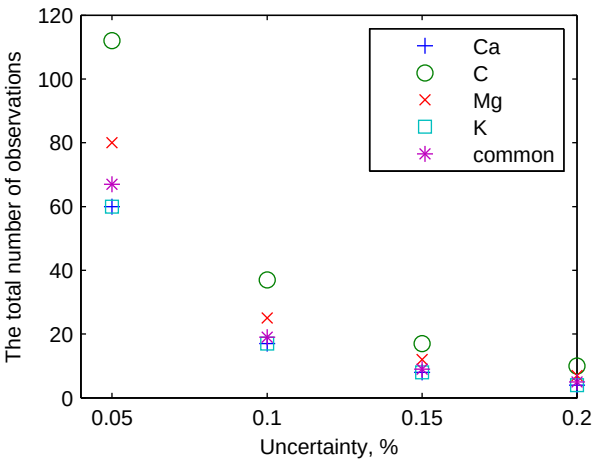

Figure 2. The monitoring designs vs. the uncertainty level

Detection of the relationships between concentrations of these water constituents has been done using Basic Fitting tools in MATLAB where the parameters for the best fitting linear regression model have been determined. An important question in discovering these relationships is which water constituent to choose as the base water constituent. Since the first order parameter of a linear model must be less than 1.0, the base water constituent must have the mean value of concentrations higher than mean concentrations of others. In order to reduce the total number of required observations, the variability of the base water constituent must be less than variability of other water constituents of interest. In the given case study, the total calcium has been chosen as the base water constituent and for other water constituents linear models have been identified (Table 2). These models have been used in (4)-(5) to determine monitoring designs efficient for all four investigated water constituents. The suggested designs for different level of uncertainty in the estimates are shown in Figure 2.

Table 2. Parameter values for the regression models used in the case study

Water constituent	*a*	*b*	Deviation
C	-0.0290	7.3052	2.48
Mg	0.1811	2.2837	8.5198
K	0.0275	1.6374	0.4186

4. DISCUSSION

The comparison of the designs is presented in Table 3. At Old Mill Rd observation site, the suggested common monitoring designs allow for reduction of water samples up to 50% over a year period for the most variable water constituent in the study for the cost of an increase in samples for total calcium, the least variable water constituent. The application of model (1) and (3)-(5) is justified by detected relationships between the investigated water constituents. Relatively high correlation coefficients suggested application of the linear model.

Table 3. Changes in efficient monitoring designs when the designs common for all water constituents are used, %

Water constituent	Uncertainty, %			
	5	10	15	20
Ca	11.6	11.7	12.5	12.5
C	-40.2	-48.6	-47.1	-50.0
Mg	-16.3	-24.0	-25.0	-28.7
K	11.6	11.7	12.5	12.5

Strictly speaking, the choice of a regression function must be validated. Even if relationships between water quality indicators determined from the same water sample exist, they can be better described by non-linear functions. In that case, equation (5) must be adjusted and the explicit expression for the constraint function may not be available. This issue is a subject of further investigation. In the present case study, polynomial functions have been used for the regression analysis. The polynomials of a degree higher than one only slightly improved fitting curves, and their regression coefficients of the highest degree were very small. Figure 3 presents the comparison of two polynomials describing relationships between calcium ions and magnesium ions. That is why the linear functions were considered as a good fit. Relatively good approximation of the relationships between water quality indicators and relatively simple analytical expressions for estimator variance determined the choice of the linear models for the development of efficient monitoring designs.

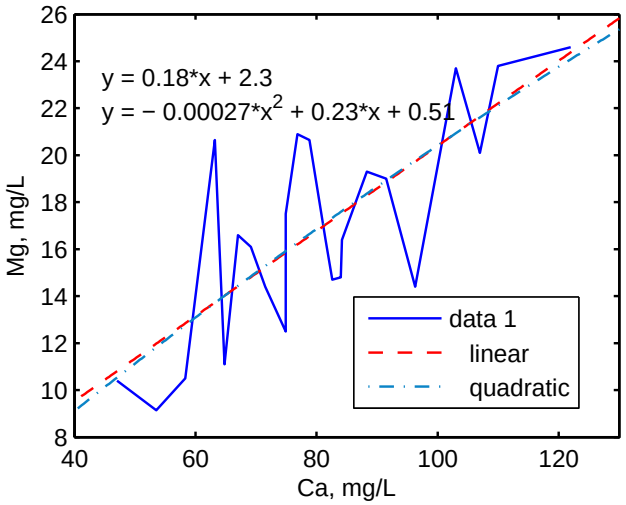

Figure 3. Regression analysis of the relationships between the investigated water constituents

The proposed approach is based on two key points. First, that there is a water quality indicator with the average concentration higher than average concentrations of all other water constituents determined from the same water sample and its variability is less than variability of others. Second, there are relationships between these water constituents described by relatively simple models.

Admitting that these two conditions may not always hold, it is necessary to point out that there exist waterbodies and observation sites which satisfy these conditions. In many monitoring programs, aggregate water quality indicators are taken into account along with concentration of their chemical compounds. Total Dissolved Solids (TDS) is one of the examples. The concentration of this water quality indicator is a sum of major ions, thus, higher than the concentrations of its components with most likely lesser variability. In general, concentrations of constituents in a water column are conditioned by natural and anthropogenic factors typical for a site and common for all water constituents at a given cross section. The effect of these factors on the dynamics of concentrations of water quality indicators can be considered as a common pattern and as a rationale for dependencies in concentrations of subsets of the water quality indicators.

Model (1) and (3) does not require the same level of uncertainty in the estimates to be specified for all monitored water quality indicators. The level of uncertainty can be constituent-specific. This may help to reduce the total number of required observations according to project-specific needs.

The nature of the problem of optimization of monitoring designs is in developing recommendations for future sampling based on the results from the past. In other words, the developed recommendations are correct if they are made based on representative data sets. Although the designs have been obtained as optimal solutions of the model (1) and (3)-(5), they are satisficing rather than optimal. At the same time, they can be used as benchmarks to better understand the quality of information generated from such data sets.

5. CONCLUSIONS

The application of simple statistical models to optimization of temporal monitoring design at a given observation site leads to conclusions and provides an insight for further research. It had been shown that if the dependencies between water quality indicators whose concentrations are derived from the same water sample are detected, they can be used to significantly reduce the total number of observations required for water quality assessment.

The utilization of non-linear regression functions instead of equations (4) and (5) is a subject for further investigation. The operation research model (1) and (3) can be applied to tiered monitoring systems when water quality indicators of interest are split in two sets: core and supplemental according to their importance

for a given site with different accuracy requirements. Testing the approach on the waterbodies with different hydrological and hydrochemical characteristics will create a basis for articulation of formal criteria for applicability of the approach to a waterbody of interest.

The data sets used in the study support estimates of the selected water quality indicators at the selected cross-section of the Humber River with 90% of accuracy. To upgrade the estimates to 95% accuracy level, frequencies of observations must be increased by more than three times. The designs common for all water quality indicators measured at the given site may result in higher numbers of observations. Depending on the desired level of accuracy, it may lead to daily sample programs. The proposed approach may help to develop efficient monitoring designs with the reasonable cost of sampling by considering subsets of the water quality indicators.

ACKNOWLEDGMENTS

The research has been implemented using data sets provided by the Toronto and Region Conservation Authority (Ontario, Canada). The authors are grateful to Angela Wallace for her work on data files and valuable comments on data. The authors are thankful to anonymous reviewers for their thoughtful comments and suggestions on the improvement of the manuscript.

REFERENCES

Cieniawski, S., Ehart, J., Ranjithan, S. (1995). Using genetic algorithms to solve a multiobjective groundwater monitoring problem. *Water Resources Research* 31(2), 399-409

Conn, A.R., Gould, N.I.M., Toint, P.L. (1997). A globally convergent augmented Lagrangian algorithm for optimization with general constraints and simple bounds. *Mathematics of Computation* 66(217), 261-288.

Erechtchoukova, M.G. and Khaiter, P.A. (2009). Investigation of Monitoring Designs for Water Quality Assessment In: Anderssen, B. et al. (Eds.),18th IMACS World Congress - MODSIM09 International Congress on Modelling and Simulation, 13-17 July 2009, Cairns, Australia, 3612-3618.

Erechtchoukova, M.G., Chen, S.Y. and Khaiter, P.A. (2009). Application of Optimization Algorithms for the Improvement of Water Quality Monitoring Systems. In: Athanasiadis, I.N., Mitkas, P.A., Rizzoli, A.E., Marx Gomez J. (Eds.) Information Technologies in Environmental Engineering Proc. of the 4th International ICSC Symposium Thessaloniki, Greece, May 28-29, 2009. Springer, 176-188.

Erechtchoukova, M.G. and Khaiter, P.A. (2010). Efficiency Criteria for Water Quality Monitoring. In: Swayne, D.A., Yang, W., Voinov, A.A., Rizzoli, A.E., Filatova, T. (Eds.) 2010 International Congress on Environmental Modelling and Software Modelling for Environment's Sake, Fifth Biennial Meeting, 5-8 July 2010, Ottawa, Canada, 272-279.

Groot, S., and Schilperoort, T. (1983). Optimization of water quality monitoring networks, *Water Science and Technology*, 16, 275–287.

Icaga, Y. (2005). Genetic algorithm usage in water quality monitoring networks optimization in Gediz (Turkey) river basin, *Environmental Monitoring and Assessment,* 108, 261–277.

Lettenmaier, D.P. (1978). Design considerations for ambient stream quality monitoring, *Water Resources Bulletin*, 14, 884–902.

Loftis, J.C. and Ward, R.C. (1980). Cost-effective selection of sampling frequencies for regulatory water quality monitoring. *Environment International,* 3, 297-302.

Ning, S.K. and Chang, N.-B. (2002). Multi-objective, decision-based assessment of a water quality monitoring network in a river system, *Journal of Environmental Monitoring*, 4, 121–126.

Ning, S.K. and Chang, N.-B. (2004). Optimal expansion of water quality monitoring network by fuzzy optimization approach, *Environmental Monitoring and Assessment*, 91, 145–170.

Robertson, D.M. and Roerish, E.D. (1999). Influence of various water quality sampling strategies on load estimates for small streams, *Water Resources Research*, 35(12), 3747–3759.

Shabman, L. and Smith, E. (2003). Implications of applying statistically based procedures for water quality assessment, *Journal of Water Resources Planning and Management*, 129(4), 330–336.

Whitfield, P.H. (1988). Goals and data collection designs for water quality monitoring, *Water Resources Bulletin*, 24(4), 775–780.

WQTG (2006). A Canada-wide framework for water quality monitoring. PN 1369, online URL http://www.ccme.ca/assets/pdf/wqm_framework_1.0_e_web.pdf.

US Environmental Protection Agency (2003). Elements of a state water monitoring and assessment program (EPA 841-B-03-003), Online URL http://www.epa.gov/owow/monitoring/elements/index.html