International Environmental Modelling and Software Society (iEMSs)
2012 International Congress on Environmental Modelling and Software
Managing Resources of a Limited Planet, Sixth Biennial Meeting, Leipzig, Germany
R. Seppelt, A.A. Voinov, S. Lange, D. Bankamp (Eds.)
http://www.iemss.org/society/index.php/iemss-2012-proceedings

Model-Driven Approach to Optimization of Monitoring Designs for Multiple Water Quality Parameters

Marina G. Erechtchoukova[a], Peter A. Khaiter[a]
[a]*York University, Toronto, Canada*
marina@yorku.ca, pkhaiter@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. 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 which minimizes the total number of water samples collected over an investigated period of time under the condition that the uncertainty of estimates derived from the monitoring data is kept below an acceptable level. Since concentrations of water constituents exhibit different variability, the numbers of observations required to achieve the same uncertainty level in their estimates vary significantly. In order to make a practically meaningful recommendation on the frequencies of observations, it is necessary to compromise temporal monitoring designs for all water quality parameters whose concentrations are derived from the same grab water sample. Given that concentrations of these parameters are formed 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 develop temporal monitoring designs common for water quality parameters determined from the same water sample and can significantly reduce the total number of observations required for water quality assessment. The proposed approach has been tested on observation data collected on a section of a small river in a highly urbanized area. The proposed approach may help to develop efficient monitoring designs with the reasonable cost of sampling programs by considering subsets of the water quality parameters.

***Keywords**:* Monitoring design; water quality parameter; constrained optimization model

1 INTRODUCTION

Simulation models and mathematical techniques have been suggested for water resource management for many years as a tool supporting decision making. Models offer a means to transform data into information and to complement observation data in a cost-effective way. Due to the ability to simulate the behaviour of the real world ecosystems without their actual perturbation, models can be used for evaluation of impacts of various management scenarios that could not be implemented otherwise. Many models are accompanied by user-friendly interfaces and data pre-processing and post-processing facilities, e.g. WASP7 [Ambrose et al. 2006]. Nevertheless, models are not acclaimed by decision and policy makers, and model application to assessment of human impact on water resources and their sustainable management is yet to be expanded. There are

several factors limiting usage of models in water related decision making process. The necessity to obtain a representative data set for simulations is one of them.

1.1 Data-model paradigm

Any modelling methodology includes such important steps as site specific identification of model parameters, model validation for a given case study, and model stability and uncertainty analyses [Jakeman et al. 2006]. All these steps require rather detailed data sets describing the investigated aquatic environment. Where a process-based approach is employed, environmental indicators relevant to an investigated case study are selected, and their spatial and/or temporal dynamics are imitated by describing natural processes affecting the indicators based on mathematical formulae. The set of indicators determines the number of model state variables and processes which must be taken into account. The processes which significantly contribute to indicators' variability are included into a model using balance equations. It may call for additional processes to be added to the model and each process to be described by a particular mathematical term. Sets of data obtained though observations and measurements should support computational procedures for model identification and investigation and are to satisfy statistical conditions specific for the given case study and the selected model. In many cases, available data limit the set of model state variables and the corresponding set of natural processes which the model takes into account. It results in replacing of detailed indicators by aggregate ones and reducing model ability to detailed description of natural processes. The model to be applied imposes additional requirements on the structure and the size of the data set necessary for simulations. Data-model interplay had been investigated in [e.g., Richardson and Berish 2003, Erechtchoukova and Khaiter 2007].

1.2 Data sources

Ideally, data collection is implemented for a particular modelling exercise and follows modelling methodology requirements. It helps to identify sampling sites, set of water quality and quantity parameters to be observed, duration of sample collection and frequencies of sampling. Even in this the most favourable case, preliminary analysis of variability of the investigated water quality parameters is necessary, and data collected on the waterbody for different purposes are an important source of information. In many cases, data collected for a totally different purpose are used to calculate preliminary estimates. In general, routine monitoring systems become an important source of data on the status of the aquatic environment used in the modelling exercise. These systems implement observations and measurements of important physical and chemical parameters of a waterbody on a network of sampling sites in the most systematic way. US EPA [2003] investigated various topologies for sampling networks and approaches to the development of water quality monitoring programs.

Many monitoring systems operate based on the tiered approach. The approach requires identification of a core set of water quality parameters which reflect designated uses and can be monitored routinely to assess attainment of applicable water quality standards. In addition to the core set of indicators of the aquatic environment, supplemental indicators dictated by the site or project specific needs are to be determined and monitored with the lesser frequency.

Most of the data supplied by routine monitoring systems are obtained from grab samples collected routinely at certain time intervals at different sampling sites of a waterbody or by using automatic samplers. Although automatic samplers are able to determine values of water quality parameters 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 for a limited set of water quality parameters. Second, limited budgets of monitoring systems cannot afford a large number of such devices, thus leaving many monitoring sites to operate under routing sampling programs.

For an efficient assessment of water quality, US EPA [2003] recommends an integrated approach incorporating several techniques including sampling programs supplying data for a statistically valid assessment along with watershed and water quality models used to transform collected data into information. It implies the necessity to determine monitoring design providing data sufficient to apply models to a given case study.

1.3 Modelling support

It is worth noting, that routine monitoring systems supply data for 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. Monitoring objectives dictate important characteristics of selected water quality indicators required for decision making and constraints imposed on sampling programs. For example, implementing the total maximum daily load process, it is important to know not only concentrations of water constituents and violations of water quality standards, but also the duration and magnitude of the violation [Shabman and Smith 2003]. Trend detection requires sampling of selected water quality indicators with a fixed frequency at the same location and at the reference site for a long period of time [Lettenmaier 1978]. Attainment of water quality standards can be confirmed using different schemes, including fixed frequency, sequential or Markov sampling where the number of observations is determined based on the outcome of previous observations [Whitfield 1988]. For mass transport estimation, simple random sampling or stratified random sampling are preferable [Robertson and Roerish 1999; Erechtchoukova and Khaiter 2009]. The designs supporting environmental assessment are project specific. They must be compliant with the type of analysis of the effects in an investigated project. In addition to that, models require data sets with significantly different properties. To compromise all the requirements and to create data sets suitable for a wide range of objectives and models is hardly possible, because it requires data sets representing very detailed dynamics of water quality and quantity parameters. The dynamics of water quality parameters can be described by relatively simple models which can restore the missing values of concentrations of water constituents or physical characteristics of a waterbody using known relationships between water quality and quantity parameters. These models can complement observation data for application of watershed and water quality models. They can also be used to develop monitoring designs supplying data sufficient to generate series of values of water quality parameters for subsequent modelling exercises. 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. Simple random designs depend on the variability of water constituents and vary significantly for different water quality parameters [e.g., Erechtchoukova and Khaiter 2009]. At the same time all water constituents in a water column are affected by a set of common natural and anthropogenic processes. Therefore, it is reasonable to expect that series of concentrations of these constituents are somehow related. If such relationships are discovered, they can be used to restore the values of one water quality parameter based on another water quality parameter.

1.4 Aim and scope

The paper presents an approach to the development of efficient temporal monitoring designs for a water quality monitoring system with a network of fixed stations where grab samples are collected periodically. The designs are developed as solutions of the operation research model formulated based on the cost-

effectiveness analysis [Erechtchoukova and Khaiter 2009]. The uncertainty of the estimates obtained from water quality data collected in accordance with the monitoring design is used as the effectiveness measure. Since several water quality parameters, with different temporal variability, are determined from the same water sample, monitoring designs for these parameters must be compromised in a way to ensure a required level of effectiveness for all parameters being detected. The approach takes into account variability and relationships between these water quality parameters and generates temporal designs common for these parameters [Erechtchoukova and Khaiter 2011]. 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 parameters.

2 METHODS

Environmental monitoring systems operate under financial constraints. Available budget must be allocated in a way supporting data collection sufficient to fulfil monitoring objectives. Existing frameworks for development of monitoring programs suggest cost-effectiveness analysis as the main approach to determine the efficient sets of sampling sites and monitoring designs at these sites. If the cost of a proposed monitoring design is known and the effectiveness of the same design is expressed in monetary form, the efficient monitoring design can be easily determined by a direct comparison. In many cases, neither of the estimates is available. However, if both the cost and effectiveness of a design can be expressed as functions of the same set of variables, the efficient design can be obtained as a solution of the operation research model. The constrained optimization approach supports two possible articulations of the problem of the development of efficient monitoring designs: (1) to maximize the effectiveness of the design under limited budget and (2) to minimize the cost of the design within an acceptable level of effectiveness. In both cases it is necessary to provide quantitative estimates of the cost and the effectiveness of a monitoring design [Erechtchoukova and Khaiter 2010]. In this paper, articulation (2) is considered.

2.1 Cost estimate of a monitoring design

In general the cost of sample collection and processing comprises direct and indirect components. Loftis and Ward [1980] considered only the direct cost of sample collection, 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 monotonically increasing function of the number of collected and processed samples. This assumption allows to avoid the necessity to calculate the actual cost of a monitoring design and 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.

2.2 Effectiveness estimate of a monitoring design

Formal definition of the effectiveness of a monitoring design should quantitatively express the extent, to which monitoring results meet the objectives. Since one of the main goals of a monitoring system is to provide information for decision making, the effectiveness may take into account the reliability of estimates obtained from the collected data. In general, the estimates derived from monitoring data are considered reliable when their uncertainty does not exceed an established level. Traditionally, the uncertainty of an estimate is evaluated via the variance of its estimator, reflecting the fact that the larger the data set, the lesser the uncertainty. Although measurement uncertainty of observation data is an important source of uncertainty in the estimates, this type of uncertainty was not included into consideration. In order to define uncertainty formally, a water quality indicator of

interest is to be selected. Since the presence and amount of various constituents in a water column are determined based on their concentrations, the current study used the average concentration of a substance over an investigated period of time as a primary indicator of water quality. The variance of this estimator can be directly obtained from the variance of concentrations of selected water constituents. The effectiveness of a monitoring design monotonnically decreases with the increasing uncertainty of the estimate. It is necessary to consider the variance of all water quality constituents determined from the same water sample.

2.2 The operation research model

Assuming that the maximum effectiveness is achieved when the accurate estimate is available and that the accurate estimates are calculated based on daily measurements, the operation research model can be formulated as:

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

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

where $I_k(n)$ is the estimator of the k-th water constituent on a set of n observations, $D^2(I)$ is the variance of the estimator I_k, V_k is the acceptable level of uncertainty in this estimate, and K is the total number of water constituents of interest. Since concentrations of these water quality parameters form under common environmental conditions, it is reasonable to assume that series of concentrations are somehow related. If such relationships are registered and their approximations are defined, they can be used to reduce the number of water samples required to achieve the established level of uncertainty in the estimates through the substitution of the following formulae into the constraint function of model (1):

$$C = f(C_{CMV}), \tag{2}$$
$$D(I_C) = D(I(f(C_{CMV}))),$$

where C_{CMV} is the concentration of the base water constituent, f is a regression function identified based on the least squares fitting.

3 CASE STUDY

The model (1) with linear regression function 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 part of the main section of the Humber River (Ontario, Canada). The main branch of the Humber River 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 annual water discharge of about 0.20 – 0.32 km^3/year (Figure 1). The river flows in the Southern Ontario from the Georgian Bay to the Lake Ontario through the Greater Toronto Area, the most urbanized centre in Canada. Its waters experience significant anthropogenic impact. The major ions, namely, the total calcium (Ca), the dissolved inorganic carbon (C), and the total magnesium (Mg), 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.

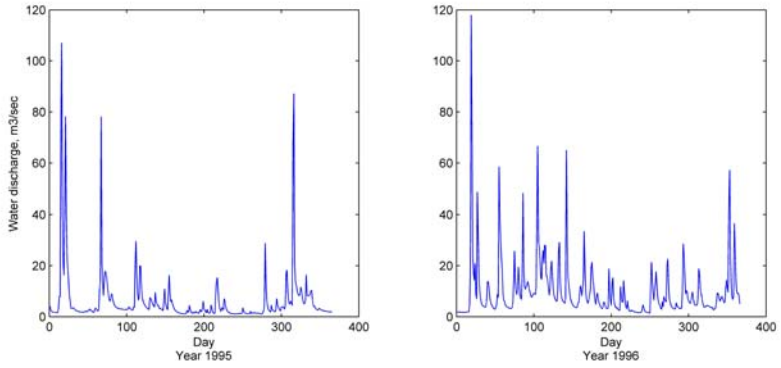

Figure 1. Water hydrograph at Old Mill Rd. station, Humber River (Ontario, Canada)

It is necessary to specify a water quality indicator which is used in a decision making. In this study the average concentration over a year period was chosen as an indicator of interest. Simple random designs with the minimum total number of observations are presented in Figure 2. They were developed for each investigated water quality parameter and different levels of uncertainty as solutions of the constrained optimization model (1).

Table 1. Basic statistics of the investigated water quality parameters

Water constituent	Year 1995			Year 1996		
	Mean	Variance	Coefficient of variance	Mean	Variance	Coefficient of variance
Ca	80.04	520.42	0.28	77.81	227.69	0.19
C	4.37	1.62	0.29	4.79	1.33	0.24
Mg	15.59	19.38	0.28	14.98	41.38	0.43

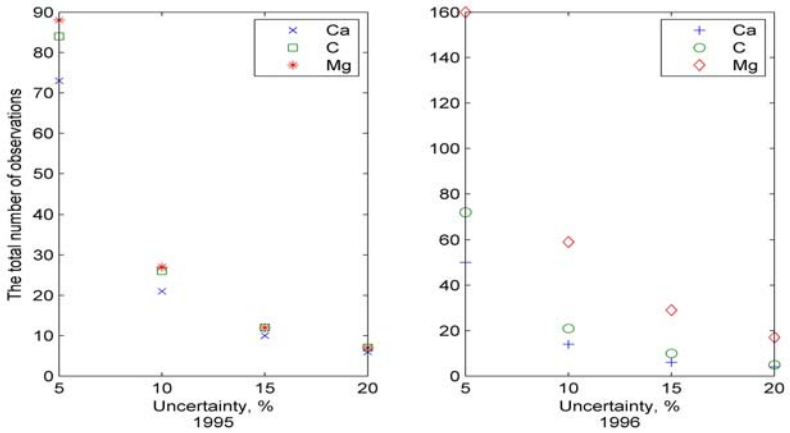

Figure 2. The required number of observations vs. uncertainty in the estimates

To reduce the total number of required observations, the base water quality parameter had to be chosen and regression functions to be determined. The concentration of the total Ca ions had been used as the base water quality parameter. The relationships between the investigated water quality parameters were analyzed using Basic Fitting tool of MATLAB. The analysis had shown that linear functions describe relationships between concentrations of the total calcium ions and two other water quality parameters very well. The polynomials of higher

orders produced almost the same results and the coefficients for the second and higher order terms were significantly smaller than 1.0. Linear regression functions have an obvious advantage because analytical expressions for formulae (2) can be easily derived and substituted in the model (1). The regression coefficients determined for years 1995 and 1996 are presented in Table 2. Monitoring designs common for all three investigated water quality parameters supporting evaluation of the annual average concentration of the water constituents with different levels of uncertainty are presented in Table 3.

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

Year	Water constituent	First order coefficient	Constant
1995	C	-0.0166	5.684
	Mg	0.1011	7.422
1996	C	-0.0217	6.488
	Mg	0.2199	-2.1385

4 DISCUSSION AND CONCLUSIONS

The comparison of the designs developed for the investigated water quality parameters individually with the designs common for all these parameters in a given year clearly speaks in favour of application of regression models (2) in the operation research model (1). It helps to reduce the required number of observations for the most variable investigated water constituent, Mg, by 57% for the price of the 31% increase in observations for the base water quality parameter in the study - the total calcium. Given that all three water quality parameters are determined from the same water sample, the compromise looks very reasonable and the derived common designs reduce the cost of data collection and processing. The application of the models (1) and (2) to the case study suggests some recommendations on the approach refinements. The concentration of the total calcium ions was chosen as a base parameter, because it has the highest mean value and the least coefficient of variation. These facts imply that this water constituent requires the least total number of observations to achieve a desired level of uncertainty in the estimates and regression coefficients of the first order terms in linear regression models for other water constituents are less than one. Otherwise, the improvement of the designs is not possible. The approach intensively uses regression models. That is why the proposed designs must be sufficient to determine statistically valid regression coefficients. In this study, the statistical significance of the regression models (2) has been checked afterward. The way to incorporate this requirement into a framework for efficient temporal monitoring designs is the subject of further investigation. 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 parameters 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

Table 3. Efficient monitoring designs common for the investigated water quality parameters

Year	Uncertainty, %			
	5	10	15	20
1995	73	21	10	6
1996	69	20	9	5

concentrations of subsets of the water quality indicators. Model (1) does not require the same level of uncertainty in the estimates to be specified for all monitored water quality parameters. The level of uncertainty can be constituent-specific. It makes model (1) useful to determine efficient monitoring designs for tiered monitoring systems. Constituent-specific levels of uncertainty may also help to reduce the total number of required observations according to project-specific needs.

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

Ambrose, R.B., Martin, J.L., and T.A. Wool, WASP7 Benthic algae-model theory and user's guide. Supplement to Water Quality Analysis Simulation Program (WASP) user documentation. EPA 600/R-06/106, Athens, GA, 2006

Erechtchoukova, M.G. and P.A. Khaiter, Data-model issues in environmental impact assessment, *The International Journal of Environmental, Cultural, Economic and Social Sustainability*, 3(6), 149–156, 2007.

Erechtchoukova, M.G. and P.A. Khaiter, 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*,* 2009

Erechtchoukova, M.G. and P.A. Khaiter, 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, 2010

Erechtchoukova, M.G. and P.A. Khaiter, A model-driven approach to uncertainty reduction in environmental data. In Golinska, P., Fertsch, M. and J.Marx-Gomez (eds) *Information Technology in Environmental Engineering. New Trends and Challenges*. Springer-Verlag, Berlin, 107–122, 2011

Jakeman, A.J., Letcher, R.A., and J.P. Norton, Ten iterative steps in development and evaluation of environmental models, *Environmental Modelling and Software*, 21: 602–614, 2006.

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

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

Richardson, J. R., and C.W. Berish, Data and information issues in modelling for resource management decision making: communication is the key. In Dale, V.H.(ed) Ecological modelling for resource management, Springer, New York, 167–179, 2003.

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

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

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

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