Importance of Genetic Algorithm Operators in River Water Quality Model Parameter Optimisation

A.W.M. NG and B.J.C. PERERA

School of the Built Environment, Victoria University of Technology,
PO Box 14428 MCMC, Melbourne, Victoria 8001, Australia (Anne.Ng@research.vu.edu.au)

Abstract: Well-calibrated river water quality models are required to assess the effectiveness of various management strategies, which are aimed at improving river water quality. Model calibration (or parameter estimation) is an important part of overall model development. A river water quality model was developed for Yarra River in Victoria (Australia) and was calibrated using a genetic algorithm (GA). In general, the efficiency of GA depends on the proper selection of GA operators, which prompted an investigation of these operators in achieving the 'optimum' model parameter set for the Yarra River water quality model. This was conducted by considering a hypothetical river network water quality model with both insensitive and sensitive reaction parameters and later verified by the Yarra River water quality model. Based on limited numerical experiments, it was found that GA with a reasonable operator set obtained from literature was capable of achieving a near-optimum model parameter set in river water quality models. However, it is recommended that further studies be conducted to verify the above findings.

Keywords: Water Quality Modelling; Genetic Algorithm; Parameter Optimisation; QUAL2E; Calibration

1. INTRODUCTION

Successful management of river water quality requires the development and use of river water quality models. Model calibration (or parameter estimation) is an important part of overall model development. Some model parameters can be physically measured, while the remaining model parameters should be estimated through model calibration. Model calibration is generally done through a trial and error iterative process by comparing model predictions with observations. This method is time consuming, and can also miss the 'optimum' model parameter set. Recently, genetic algorithm (GA) optimisation has proven to be successful and efficient in identifying the 'optimum' parameter set for river water quality models [Mulligan and Brown, 1998]. GA is a global optimisation technique that is based on the concept of natural selection and genetics [Goldberg, 1989].

In general, the efficiency of GA depends on the proper selection of GA operators, which are essentially the components that make up the overall GA process. The GA operators deal with parameter coding, population initialisation, selection of subsequent populations, crossover and mutation. The reader is referred to Goldberg [1989] for definitions of these terms.

The significance of GA operators on the optimised model parameter set has been studied in a number of water resource applications. The most comprehensive study was on a rainfall and runoff application by Franchini and Geleati [1997]. They found that the GA operators did not have any significant impact on the optimum model parameter set, and therefore stated that a robust GA operator range was adequate. On the other hand, Davis [1991] commented that the optimal GA operator set varies from problem to problem, but a reasonable robust GA operator range can provide an efficient solution. The reasonable robust ranges for various GA operators (for crossover and mutation rates) are given in Goldberg [1989]. Mulligan and Brown [1998], in their water quality modelling application, explored the effect of constant mutation rate throughout the run against varying rates, and found that the constant rate was efficient. As seen from these studies, the importance of the GA operators in

achieving the 'optimum' model parameter set was inconclusive. Therefore, a detailed study was conducted to investigate the effect of GA operators on the model parameter optimisation of Yarra River Water Quality Model (YRWQM). The paper begins by introducing the GA optimisation technique. The aims and methodology are then discussed. A detailed study of GA operators on a hypothetical river system is discussed, followed by a similar study on YRWQM. Finally, the conclusions drawn from these studies are presented.

2. GENETIC ALGORITHM

GA is a powerful optimisation technique that has been applied successfully in many disciplines [Paz, 1998]. It is based on concepts of natural selection and genetics, and therefore, the terminology used in GA is borrowed from genetics. The GA process as applicable to model parameter optimisation of river water quality models is described below.

Every model has its own model parameters. According to the genetics terminology, each model parameter is a gene, while a complete set of model parameters is a chromosome. Each parameter in GA is encoded using binary, gray and (recently) real coding systems [Wardlaw and Sharif, 1999]. Each GA run consists of a number of generations with constant population size. The process of GA begins with an initial population of a user-defined number of model parameter sets, which are chosen at random or some pre-defined rule, within a specified parameter range (search space). Each model parameter set is then evaluated by an objective function (e.g. simple least squares) to yield its fitness value [Sorooshian and Gupta, 1995].

The second and subsequent populations are generated by combining model parameter sets with high fitness value from the previous population (parent) through selection, mutation and crossover operations to produce successively fitter model parameter sets (offsprings). The selection GA operator favours those parent parameter sets with high fitness value to those of lower fitness value in producing offsprings. The mutation operator adds variability to randomly selected model parameter sets by altering some of the values arbitrarily. The crossover operator exchanges model parameter values from two selected parent model parameter sets. Several generations are considered in one GA run, until no further improvement (within a certain tolerance) is achieved in the objective functions.

3. AIMS AND METHODOLOGY

The main aim of this study is to investigate the effect of GA operators on YRWQM model parameter optimisation. In this investigation, an initial hypothesis was made that the GA operators have an effect on the 'optimised' model parameters. The following river networks were considered to study this hypothesis.

- A hypothetical river system with known insensitive and sensitive model parameter sets.
- YRWQM river network.

The QUAL2E [Brown and Barnwell, 1987] models of the two river networks were linked (via input and output files) with GENESIS [Grefenstette, 1995] - a standard GA software package. The hypothetical river network model was initially assembled using the data of a hypothetical river network (with no modifications) of Chapra [1998], which deals with modelling of Dissolved Oxygen (DO). An uncertainty and sensitivity analysis of model parameters was then conducted using Monte Carlo simulation (MCS). The criterion used to determine the sensitivity of model parameters to output water quality was based on the relative deviation ratio (RDR). A critical value of RDR of 1 was considered in this study, as in Hamby [1994]. Any model parameter with RDR greater than 1 was considered to be sensitive and vice versa. The details of uncertainty and sensitivity analysis in relation to water quality model parameters are described in Ng and Perera [2001]. The model parameters assembled using data of Chapra [1998] was found to be insensitive to DO response. After modifying the effluent data of Chapra [1998], the model parameters were shown to be sensitive. Therefore, two models were developed using the hypothetical river network, one with insensitive model parameters and the other with sensitive parameters.

The hypothesis was initially tested with the hypothetical network models using GA operators obtained from the literature, subject to capabilities of GENESIS (which is called 'LIT' set in this paper). Although gray and binary coding systems were available in encoding model parameters in GENESIS, the former was used since it is an improvement of the latter [Goldberg, 1989]. Model parameter set initialisation can be performed randomly or heuristically, however, the random method was used in this study. Linear ranking together with stochastic universal sampling were used for selection of the 'fitter' model parameter sets for the next generation. A

population size of 125 and 40 generations (equivalent of 5000 model parameter sets) were selected based on the study by Franchini and Galeati [1997]. The mutation rate of 0.03 and crossover rate of 0.6 were obtained from Mulligan and Brown [1998]. The commonly used simple least squares objective function was used in this study. Based on the 'LIT' GA operators, the model parameters (i.e. reaction rates) were optimised for both insensitive and sensitive models. Since objective function value may not vary greatly for a number of 'best' model parameter sets, it was also necessary to select a certain number of model parameter sets from the last generation as the 'optimum' sets. However, no guidance was found from the literature on this issue, and therefore it was studied first.

Depending on the outcome of using the 'LIT' set in achieving convergence to the 'optimum' model parameter set, a second stage of the investigation was conducted for both insensitive and sensitive models, provided the models did not converge. The second stage involves a systematic optimisation of GA operators. For each combination of GA operators within the second stage, a GA optimisation of model parameters is conducted. The GA operators considered in the second stage optimisation were population size, and mutation and crossover rates, because the other operators (parameter coding, population initialisation and selection) were restricted by the capabilities of GENESIS.

The findings obtained from the hypothetical models were then tested on the YRWQM parameters, using the same method as for the hypothetical models. In both models (i.e. hypothetical and YRWQM), only the reaction parameters were considered.

4. HYPOTHETICAL RIVER WATER QUALITY MODEL

The river network system and the data used for this part of the study was extracted from an example given in Chapra [1998]. The river network was modelled using 6 reaches. Each reach was sub-divided into a number of computational elements of 1-km length, which provides sufficient resolution for water quality modelling. The reaction rates of CBOD, $CBOD_s$ and SOD of both insensitive and sensitive models and other input data were used in QUAL2E to produce the output DO concentration. This output DO was then considered as the observed concentrations for the GA optimisation. The reaction rates were then treated as unknown, and were optimised through GA. The search space for respective reaction rates used in GA optimisation are shown in Table 1, together with the actual reaction rates. The use of this relatively large search space allows the investigation of use of GA in achieving the 'optimum' model parameter set.

Table 1. Search space and actual reaction rate for hypothetical example.

Reaction Rates	Search space	Actual reaction rate
CBOD decay (CBOD)	0.0042-0.7	0.5
CBOD Settling ($CBOD_s$)	0.03-1.53	0.25
Sediment Oxygen demand (SOD)	0.05-7	5.0

*Modified from Bowie et al. [1985]

4.1 Parameter Sets from Final Generation

An experiment was first conducted to select a certain number of model parameter sets from the last generation of the run as the 'optimum' parameter sets. The effect of population size was hypothesised as a factor affecting these 'optimum' model parameter sets. Therefore, four population sizes of 125 (i.e.'LIT' set), 250, 500 and 1000 were investigated.

The results (i.e. objective functions corresponding to number of parameter sets) from the last generation of each run were extracted and ranked. The ranked objective functions were plotted as in Figure 1 (for population size of 125), showing the rate of change of the objective function value $(mg/L)^2$ with respect to number of model parameter sets in the last generation. This plot indicates the slope changes in the objective function value at increasing rates, as more parameters are considered. Although the first slope changes at around 3 model parameter sets, it was more appropriate to adopt the second slope change (at 8 model parameter sets), since the objective function value is fairly low with 8 model parameter sets. Similar results were seen for the other population sizes, with rapid increase in slope after 10 parameter sets. Therefore, the mean of the 10 best model parameter sets, was considered as the final model parameter set.

4.2 Model with Insensitive Parameters

After conducting the GA optimisation for the insensitive model parameters, the 'optimum' reaction rates was compared with the actual reaction rates (Table 1) and found that the maximum difference was around 9%. Therefore, it was considered that the model parameters did not converged. The 'optimum' model parameter

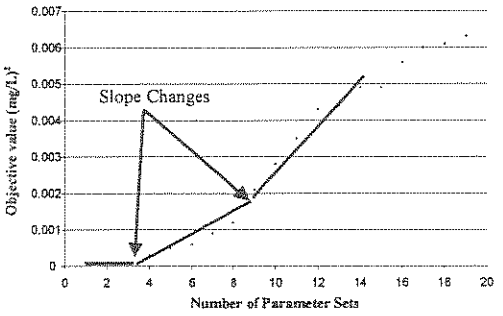

Figure 1. Objective function Vs number of parameter set.

set obtained was then used in QUAL2E to predict the DO concentration. Although the model parameters did not converge, the difference of DO response from both sets of model parameters was insignificant (within 0.5%). This was because the reaction rates were insensitive to output DO. Since the convergence was not achieved in the insensitive model, a systematic optimisation of GA operators was conducted to investigate whether the GA operators play a role in achieving convergence in an insensitive model.

The first GA operator considered in this 'optimisation' investigation was the effect of population size, which may have some influence on the 'optimum' model parameter set. Four different population sizes of 125 (i.e. 'LIT' set), 250, 500 and 1000 were investigated. The number of simulations (or model parameter sets) used for all these population sizes were constant at 32,000, which was the maximum limit in GENESIS. It was found that population sizes of 125 and 250 facilitated the convergence of the model parameters, while the model parameters did not converge with population sizes of 500 and 1000. This result was consistent with Franchini and Galeati [1997], where they stated that large population sizes require large number of simulations to reach convergence. Since the population size of 250 required more QUAL2E simulations to reach the 'optimum' set compared to the population size of 125, it was not considered any further. Therefore, the population size of 125 with 190 generations (equivalent of 23,750 QUAL2E simulations) was adopted as the 'optimum'.

The next GA operator optimisation was on mutation and crossover rates. Since these two rates simultaneously determine the rate of convergence of model parameters, they should be studied together. However, initially the range for each of these rates was optimised independently to narrow down the range for these rates.

Twenty-four different mutation rates were explored within the range of 0.001-1.0 at varying increments considering more values for lower rates, while 16 crossover rates were considered within the range of 0.25-1, at equal increments of 0.05. It was found that the model parameters did converge within the mutation and crossover rate ranges of 0.001-0.03 and 0.45-0.85 respectively. Based on these ranges, 63 different combinations of mutation and crossover rates were considered in GA optimisation of model parameters through variable increments for mutation and constant increment of 0.05 for crossover rate. The mean and the coefficient of variation (CV) of the best 10 parameter sets from each of the 63 runs after convergence of model parameters were determined. Contour plots of mean and CV were produced with respect to mutation and crossover rates for all three reaction parameters. One such plot (i.e. mean value for $CBOD_s$) is shown in Figure 2. As shown with the arrows in Figure 2, several combinations of mutation and crossover rates can yield the 'optimum' solution of 0.25 for $CBOD_s$. Similar contour plot of CV was produced for $CBOD_s$. The 'optimum' $CBOD_s$ was found from the contour plots of mean (Figure 2) and CV considering the value of 0.25 (or closer) and 0 (or closer) respectively, but for the same region of mutation and crossover rates.

After considering the contour plots for both mean and CV for all three model parameters, it was found that mutation and crossover rates of 0.003-0.007 and 0.66-0.72 respectively can simultaneously converge all three model parameters to their actual values. However, the middle values of the above ranges (i.e. mutation rate of 0.005 and crossover rate of 0.69) were adopted as the 'optimised' rates.

A final GA optimisation run was then conducted using the GA operators obtained from the 'optimisation' as described above, and found the 'optimum' reaction rates to be within 1% of the actual reaction rates, compared to 9% previously found with 'LIT' GA operator set. However, when these reaction rates were used in QUAL2E simulation, the output DO concentration did not show significant changes to the concentration corresponding to reaction rates obtained from the 'LIT' set. This experiment shows that it is not necessary to optimise GA operators for an insensitive model, and that reasonable values for GA operators obtained from literature can be used to optimise the reaction rates.

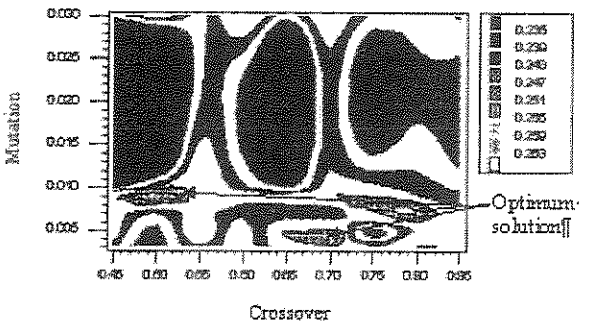

Figure 2. Mean contour plot on mutation and crossover rates for $CBOD_s$.

4.3 Model with Sensitive Parameters

A similar procedure was used for the sensitive model, as for the insensitive model. The 'optimum' reaction rates were then compared with the actual and it was found that the maximum difference was less than 2%, which can be considered as 'converged'. This model parameter set was then used in the QUAL2E simulation to determine the predicted DO concentration. The percentage difference between measured and predicted DO concentration was around 1%, which was considered sufficient in a water quality modelling study. Since a standard 'LIT' GA operator set was able to optimise reaction rates in a sensitive model and reach convergence, it can be said that the GA operators do not play a role in a sensitive model.

5. YRWQM MODEL

The Yarra River was first discretised, into a number of reaches that have uniform pollution loading, hydraulic and hydrological characteristics based on the locations of Sewage Treatment Plants (STPs), the confluences of tributaries and the water quality sampling stations. When the reaches defined based on above criteria were long, they were further sub-divided. In total, 29 reaches were considered. Each reach was then sub-divided into 1-km computational elements. This information together with flow and effluent data were used in building the YRWQM. The details are given in Ng and Perera [2001], which also stated that the YRWQM was an insensitive model, based on a detailed uncertainty and sensitivity analysis of model parameters. Therefore, the procedure used for the insensitive hypothetical model (Section 4.2) was used for YRWQM. Total Kjeldahl Nitrogen (TKN), Total Nitrogen (TN), Total Phosphorus (TP) and DO were considered in YRWQM and their respective reaction rates in GA optimisation. Similar to the hypothetical example, the parameter search space used in the YRWQM was large.

The reaction rates obtained from the GA optimisation using both GA operator sets were different, but of a similar order of magnitude. The percentage difference between the 2 'optimum' model parameter sets was from 6% to 30% on average for reaction parameters of TKN, TN, TP and DO considered in YRWQM. These 2 reaction rates sets were then used in YRWQM and the differences in the output responses of TKN, TN, TP and DO were compared. The comparison on TN (as an example) is shown in Figure 3. As can be seen from this figure, the difference in TN is not great considering that there are differences in the reaction rates values. Similar results were found for TKN, TP and DO.

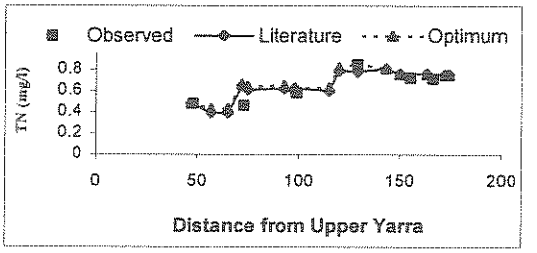

Figure 3. TN profile using 'LIT' and 'optimised' GA operator sets.

This confirmed the findings of the hypothetical insensitive model that the effect of GA operators in an insensitive model is insignificant in optimising model parameters, and that reasonable values of GA operators obtained from the literature can be used. However, this finding requires further verification Therefore, the 'LIT' set was used in the YRWQM-GA calibration.

6. SUMMARY AND CONCLUSIONS

Although the importance of GA operators on model parameter optimisation has been studied in the past, the findings were inconclusive. Therefore, a comprehensive study on the significance of GA operators was conducted using a hypothetical river network model with insensitive and sensitive model parameters, and verified using the Yarra River Water Quality Model (YRWQM) of which model parameters were found to be insensitive. The reaction rates in both models were considered in parameter optimisation, and the model parameter search space was large enough to test the capability of GA in finding the 'optimum' model parameter set.

In both models, QUAL2E software was used to model river water quality. QUAL2E uses standard river water quality advection-dispersion mass transport equation, which are also used in other river water quality software tools. Therefore, it can be said that the model structure is the same in all river water quality software.

A set of literature ('LIT') GA operators was initially used in the hypothetical model with both insensitive and sensitive model parameters. It was found that the 'LIT' set was able to achieve the 'optimum' model parameter set for the sensitive model, but not for the insensitive model. A systematic 'optimisation' on the GA operators was then undertaken to optimise the set of GA operators in the insensitive model to reach convergence of the model parameters. Using both 'LIT' and 'optimised' set of GA operators in QUAL2E model, no differences were observed in DO response due to insensitivity of the model parameters. This study showed that although the GA operators were significant in an insensitive model in reaching the 'optimum' model parameter set, its overall effect in predicting water quality was insignificant. Similar results were found with YRWQM with water quality responses of TKN, TN, TP and DO.

In conclusion, based on limited numerical experiments, it was found that the use of GA in optimising reaction rates of river water quality models can be done efficiently by selecting robust GA operators from the literature. Although the 'optimisation' GA operator sets can provide the 'optimum' reaction rates set, it is necessary to consider the amount of effort required in achieving such accuracy, which does not contribute a great difference in overall water quality prediction. It should be noted that these conclusions are based on limited numerical experiments and therefore, it is recommended that further studies should be conducted using different river settings to substantiate the findings of this study.

7. ACKNOWLEDGEMENT

The authors gratefully acknowledge the support of Environmental Protection Authority of Victoria, Melbourne Water and Yarra Valley Water for supplying data for this study.

8. REFERENCES

Bowie, G., W. Mills, D. Porcella, C. Campbell, J. Pagenkopf, G. Rupp, K. Johnson, P. Chan, S. Gherini, and C. Chamberlin, rates, constants and kinetic formulations in surface water quality modeling, (EPA/600/3-85/040). U.S. EPA, 1985.

Brown, L. C. and J. O. Barnwell, The enhanced stream water quality models QUAL2E and QUAL2E -UNCAS: documentation and user manual, EPA, Athens, 1987.

Chapra, S.C., *Surface water quality modeling*, McGraw-Hill, 1998.

Davis, L., *Handbook of genetic algorithms*, Van Nostrand Reinhold, 1991.

Franchini, M. and G. Galeati, Comparing several genetic algorithm schemes for the calibration of conceptual rainfall-runoff models, *Hydrological Sciences*, 42(3), 357-379, 1997.

Goldberg, D., *Genetic algorithms in search, optimisation and machine learning*, Addison-Wesley, 1989.

Grefenstette, J., A user's guide to the genetic search implementation system (GENESIS), ftp://www.aic.nrl.navy.mil/pub/galist/src/genesis.tar.Z', 1995.

Hamby, D., A review of techniques for parameter sensitivity analysis of environmental models. *Environmental Monitoring and Assessment*, 32, 135-154, 1994.

Mulligan, A. and L. Brown, Genetic algorithm for calibrating water quality models, *ASCE Journal of Environmental Engineering*, 124(3), 202-211, 1998.

Ng, A.W.M. and B. J. C. Perera, Uncertainty and sensitivity analysis of river water quality model parameters, Water Pollution - 2001, Sixth International Conference on Modelling, Measuring and Prediction of Water Pollution, Rhodes, Greece, 2001.

Paz, E., A survey of parallel GA, *Calculateurs Paralleles*, 10(2), 141-171, 1998.

Sorooshian, V. and K. Gupta, Model calibration, *In Computer models of watershed hydrology*, V. Singh (ed.), Water Resources Publications, 26-68, 1995.

Wardlaw, R. and M. Sharif, Evaluation of genetic algorithms for optimal reservoir system operation, *ASCE Journal of Water Resources Planning and Management*, 125(1), 25-33, 1999.