Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Environmental Modelling & Software 21 (2006) 419–425
Environmental
Modelling & Software
www.elsevier.com/locate/envsoft

Short communication

QUAL2Kw – A framework for modeling water quality in streams and rivers using a genetic algorithm for calibration

Gregory J. Pelletier [a,*], Steven C. Chapra [b], Hua Tao [b]

[a] Washington State Department of Ecology, P.O. Box 47600, Olympia, WA 98504-7600, USA
[b] Department of Civil Environmental Engineering, Tufts University, Medford, MA, USA

Received 7 April 2005; received in revised form 11 July 2005; accepted 21 July 2005

Abstract

QUAL2Kw is a framework for the simulation of water quality in streams and rivers. Dynamic diel heat budget and water quality kinetics are calculated for one-dimensional steady-flow systems. The framework includes a genetic algorithm to facilitate the calibration of the model in application to particular waterbodies. The genetic algorithm is used to find the combination of kinetic rate parameters and constants that results in a best fit for a model application compared with observed data. The user has the flexibility to select any combination of parameters for the optimization and specify any appropriate function for goodness-of-fit.

Keywords: Genetic algorithm; Calibration; Water quality modeling; Streams and rivers; Conventional pollutants

Software availability

Name of software: QUAL2Kw
Developers: Greg Pelletier, Steve Chapra, and Hua Tao
Contact address: Greg Pelletier, Department of Ecology, P.O. Box 47710, Olympia, WA 98504-7710, USA
Tel.: +1 360 407 6485
Fax: +1 360 407 6884
Email: gpel461@ecy.wa.gov
Software requirements: Microsoft Windows and Excel
Program language: Microsoft Excel VBA, Fortran 95
Availability: Free download of Excel/VBA source code and Fortran executable at http://www.ecy.wa.gov/programs/eap/models/

1. Introduction

Computer models are used extensively for water-quality management of rivers and streams (see Thomann and Mueller, 1987; Chapra, 1997 for reviews). These models must typically be calibrated by adjusting a large number of parameters to attain optimal agreement between model output and field measurements. Such calibration is often performed by the time-consuming process of manual trial-and-error. The present paper describes a new river model that includes automatic calibration.

The advantages of global optimization algorithms for calibration of water quality models have been noted by many authors (e.g. Zou and Lung, 2004; Mulligan and Brown, 1998). Several alternative tools are available for automatic calibration of models (e.g. UCODE by Poeter and Hill, 1998, or PEST by Scientific Software Group). However, the present model is not compatible with the requirements of tools such as UCODE and PEST, and a customized function optimization algorithm was required.

---

* Corresponding author. Tel.: +1 360 407 6485; fax: +1 360 407 6884.
E-mail address: gpel461@ecy.wa.gov (G.J. Pelletier).

1364-8152/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.envsoft.2005.07.002

Genetic algorithms (hereafter GAs) are a class of search techniques analogous to the process of natural selection during evolution (see Goldberg, 1989 for reviews). GAs have been used in many types of models (e.g. Gupta et al., 1999; Wang, 1997). The GA used in QUAL2Kw is the PIKAIA algorithm (Charbonneau and Knapp, 1995). PIKAIA has been applied successfully in other modeling applications (Metcalfe, 2001).

2. Overview of the water quality model in QUAL2Kw

QUAL2Kw is a modeling framework that is intended to represent a modernized version of the U.S. Environmental Protection Agency's standard river water-quality model: QUAL2E (Brown and Barnwell, 1987). In addition to incorporating more current science, the framework also includes several new features that allow it to be applied to shallow, upland streams.

As with QUAL2E, QUAL2Kw simulates the transport and fate of conventional (i.e., non-toxic) pollutants. The framework represents the river as a one-dimensional channel with non-uniform, steady flow, and simulates the impact of both point and non-point pollutant loadings. The model simulates changes within the daily cycle with a user-selected time step of less than 1 h.

The model simulates the transport and fate of a number of constituents such as temperature, carbonaceous biochemical oxygen demand, dissolved oxygen, phytoplankton and several forms of the nutrients phosphorus and nitrogen (Table 1). It also simulates several other constituents that are not typically included in generally-available software. In particular, the model simulates pH, alkalinity, inorganic suspended solids, pathogenic bacteria, and bottom algae. The inclusion of bottom algae is essential for simulating shallow streams. These algae have the novel feature of variable stoichiometry of the nutrients nitrogen and phosphorus.

Table 1
State variables in QUAL2Kw

Variable	Units[a]
Temperature	°C
Conductivity	µmhos
Inorganic suspended solids	mg D/L
Dissolved oxygen	mg O_2/L
Slowly reacting CBOD	mg O_2/L
Fast reacting CBOD	mg O_2/L
Organic nitrogen	µg N/L
Ammonia nitrogen	µg N/L
Nitrate nitrogen	µg N/L
Organic phosphorus	µg P/L
Inorganic phosphorus	µg P/L
Phytoplankton	µg A/L
Detritus	mg D/L
Pathogen	cfu/100 mL
Alkalinity	mg $CaCO_3$/L
Total inorganic carbon	mole/L
Bottom algae biomass	g D/m^2
Bottom algae nitrogen	mg N/m^2
Bottom algae phosphorus	mg P/m^2

[a] mg/L ≡ g/m^3, D = dry weight, A = chlorophyll *a*.

The model has two other features that distinguish it from other frameworks. First, sediment–water fluxes of dissolved oxygen and nutrients are simulated internally rather than being prescribed. That is, oxygen and nutrient fluxes are computed as a function of settling particulate organic matter, reactions within the sediments, and the concentrations of soluble forms in the overlying waters. Second, the hyporheic zone is modeled. This is the area below the streambed where water percolates through spaces between the rocks and cobbles. This is another feature that is necessary in order to simulate shallow streams.

QUAL2Kw is implemented within Microsoft Excel. It is programmed in Visual Basic for Applications (VBA). Excel is used as the graphical user interface for input, running the model, and viewing of output. The numerical integration during a model run is performed by a compiled Fortran 95 program that is run by the Excel VBA program.

A general mass balance for a constituent concentration (c_i) in the water column of a reach (excluding hyporheic exchange) is written as (Fig. 1):

$$\frac{\mathrm{d}c_i}{\mathrm{d}t}=\frac{Q_{i-1}}{V_i}c_{i-1}-\frac{Q_i}{V_i}c_i-\frac{Q_{\mathrm{ab},i}}{V_i}c_i+\frac{E'_{i-1}}{V_i}(c_{i-1}-c_i)+\frac{E'_i}{V_i}(c_{i+1}-c_i)+\frac{W_i}{V_i}+S_i \tag{1}$$

where Q_i = flow [m^3/d, ab = abstraction], V_i = volume (m^3), E'_i = the bulk dispersion coefficient between reaches i and $i+1$ [m^3/d], W_i = the external loading of the constituent to reach i [g/d or mg/d], and S_i = sources and sinks of the constituent due to reactions and mass transfer mechanisms [g/m^3/d or mg/m^3/d]. For bottom algae in the water column the transport and loading terms are omitted from the mass balance differential equations.

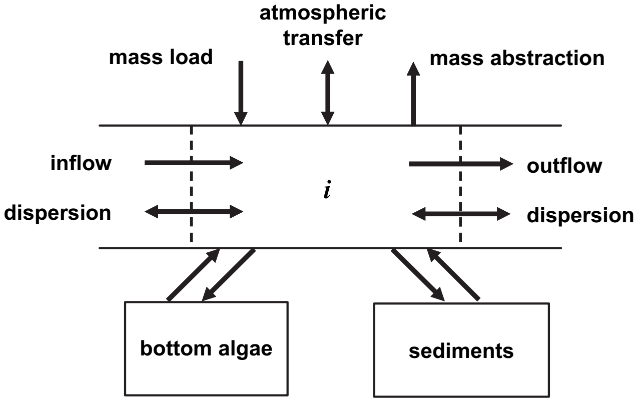

Fig. 1. Mass balance for constituents in a reach segment "*i*".

The source/sink term (S_i in Eq. (1)) requires specification of a large number of parameters for each state variable (e.g. maximum growth rate of bottom algae). The user may select which parameters are held at constant values and which are to be optimized by the GA.

A detailed description of the model is provided elsewhere (http://www.ecy.wa.gov/programs/eap/models/).

3. The genetic algorithm for the calibration of QUAL2Kw

A flowchart of the PIKAIA GA used in QUAL2Kw is shown in Fig. 2. The GA maximizes the goodness-of-fit of the model results compared with measured data. The GA carries out its maximization task on a user-selected number of model runs to define a population. The population size remains constant throughout the evolutionary process. Rather than evolving the population until some preset tolerance criterion is satisfied, the GA carries the evolution forward over a user-specified number of generations.

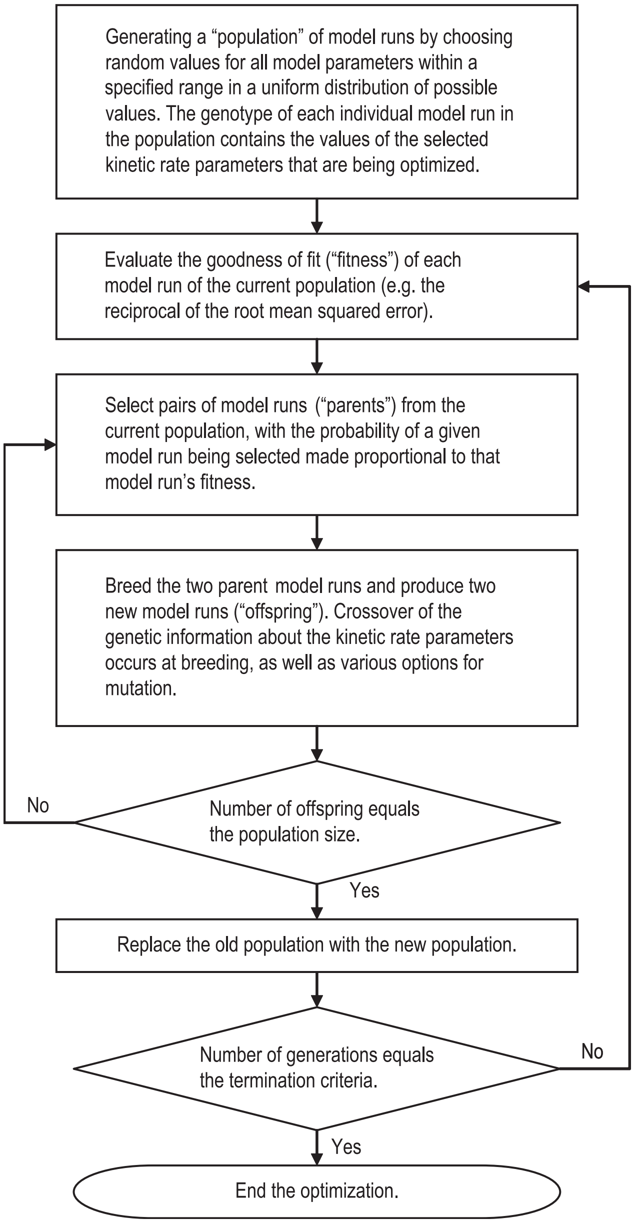

Fig. 2. Flowchart for the genetic algorithm.

Charbonneau and Knapp (1995) provide a tutorial for GA concepts and comprehensive documentation of the PIKAIA GA. The PIKAIA GA is adaptable for use in a wide variety of modeling applications. The original Fortran 77 code for PIKAIA was translated to Excel VBA for use in QUAL2Kw.

The VBA code accounts for only a small fraction of the computational time when the GA is run. Most of the computational time is spent by the compiled Fortran program that is driven by the GA to perform the numerical integration of the water quality model each time the fitness function is evaluated.

The user may select any combination of kinetic rate parameters to include in the optimization. The user also specifies the minimum and maximum values for any kinetic rate parameters that are being optimized.

The GA maximizes the function $f(x)$ in a bounded n-dimensional space, for

$$x \equiv (x_1, x_2, \ldots, x_n) \quad x_k \in [0.0, 1.0] \tag{2}$$

where n is the number of parameters that are being optimized. The parameters (x) are bounded in the range of 0.0 to 1.0 in the GA. The kinetic rate parameters for the model are scaled from the values of x according to a linear interpolation between the specified minimum and maximum value of each kinetic rate parameter that is being optimized.

The value of the function $f(x)$ corresponds to the fitness of a particular model that is run with the set of kinetic rate parameters that are scaled from x. For example, the fitness may be determined as the reciprocal of the root mean squared error (RMSE) of the difference between the model predictions and the observed data for water quality constituents. The reciprocal of the RMSE is a better indicator of fitness than the RMSE because the GA maximizes the function $f(x)$. Specification of the fitness function and inclusion of appropriate variables and data are crucial for the successful performance of the GA.

A robust fitness function should represent all of the state variables of the model (Table 1). An example of a possible fitness function for multiple state variables is the reciprocal of a weighted average of the normalized RMSE, which can be estimated as follows,

$$f(x)=\left[\sum_{i=1}^{q} w_i\right]\left[\sum_{i=1}^{q}\frac{1}{w_i}\left[\frac{\frac{\sum_{j=1}^{m} O_{i,j}}{m}}{\left[\frac{\sum_{j=1}^{m}\left(P_{i,j}-O_{i,j}\right)^2}{m}\right]^{1/2}}\right]\right] \qquad (3)$$

where $O_{i,j}$ = observed value, $P_{i,j}$ = predicted value, m = number of pairs of predicted and observed values, w_i = weighting factor, and q = number of different state variables (e.g. dissolved oxygen, pH, nutrient concentrations, etc.) included in the reciprocal of the weighted normalized RMSE. The Excel framework provides flexibility to construct any fitness function with any combination and weighting of water quality constituents to control the calibration results.

At the beginning of an evolutionary run, the initial values for x for each individual model run in the population are selected from a uniform random distribution between 0 and 1. Most of the individual model runs in the initial population have very poor fitness values. However, some individuals have better fitness than others, and the natural selection process during the evolution favors those individuals.

A "roulette wheel" algorithm is used to select both parents during the reproductive cycle. The relative fitness based on rank is used as a measure of the selection probability. The user specifies the fitness differential that is used to translate the rank into the relative fitness.

A "chromosome" is created for each parent from the n parameters in x. The GA encodes the values in x with 1-digit base 10 integers (0–9). Each digit represents a "gene" in the "chromosome". The user specifies the number of digits for the encoding. For example, if $n = 2$, $x_1 = 0.25034275$, $x_2 = 0.6718247$, and 5 digits are used for encoding, then the encoded chromosome would have a value of 2503467182. The GA also incorporates a mutation operator that may vary dynamically over the course of the evolution to potentially alter the values of each "gene".

Optional crossover modes are provided: one-point, two-point, uniform, and arithmetic. The crossover operator acts on a pair of parent chromosomes to produce a pair of offspring chromosomes. For one-point and two-point crossover, the break points are randomly selected along the length of the chromosome. For example, if the two parents have the encoded chromosome values 2503467182 and 4276986439, and one-point crossover occurs at the 4th digit, then the offspring chromosomes are 2503986439 and 4276467182. Crossover occurs if a user-specified number is exceeded by a uniform random number between 0 and 1, otherwise the offspring are exact copies of the parents.

Two-point crossover is similar to one-point except that the chromosome is randomly split into three segments and the middle segment is crossed. Uniform crossover randomly crosses each gene. Arithmetic crossover decodes the parent chromosomes, randomly interpolates the real values in vector x between the parents, and then encodes the interpolated vectors to create the offspring chromosomes. Hybrid modes are also available for combinations of one-point, two-point, uniform, and arithmetic crossover.

The offspring chromosomes are decoded back to real values of x between 0.0 and 1.0, scaled between the

Fig. 3. An example calibration for dissolved oxygen in a small stream.

specified minimum and maximum values for the kinetic rate parameters, and the model is run with the new decoded and scaled kinetic rate parameters. The fitness value of these new offspring model runs is then determined and this process is repeated until the number of offspring equals the number of model runs in the population.

Several optional reproduction plans are available to control how the algorithm incorporates the offspring into the population during the evolution. These include (1) full generational replacement, (2) steady-state-delete-random, and (3) steady-state-delete-worst. Under the first plan the entire population is replaced by the offspring as soon as the number of offspring equals the size of the population. Under the second and third, the offspring are incorporated into the population as they are produced. The second and third plans differ in terms of which members of the population are deleted from the population when the offspring are incorporated.

The GA incorporates an optional strategy of elitism which allows the user to specify whether the fittest individual in a population will be guaranteed to be passed on to the next generation.

4. Example application

Application of a model generally includes calibration and confirmation (e.g. Chapra, 2003). QUAL2Kw

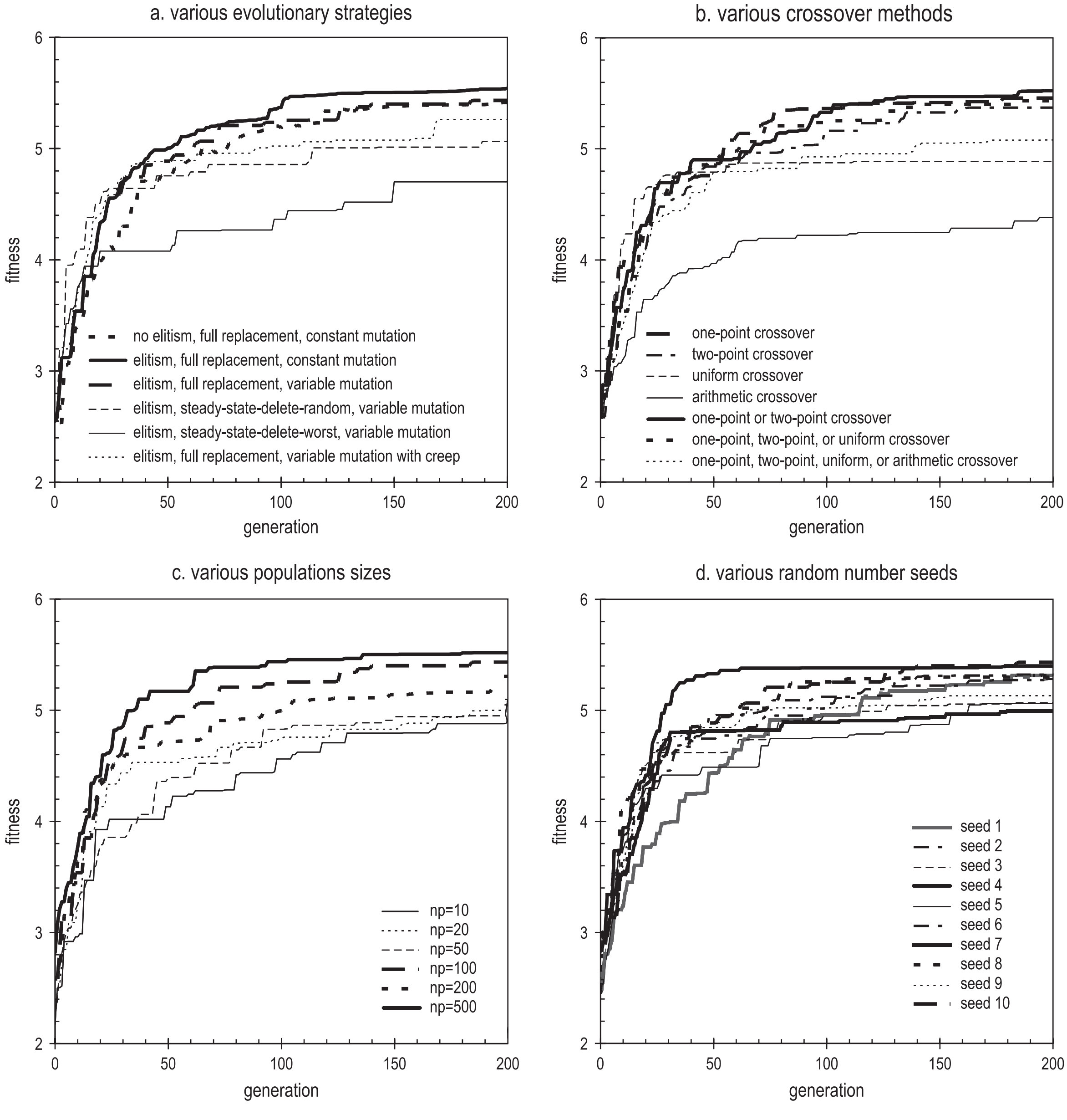

Fig. 4. Results of various evolutionary strategies and options. A population size of 100 was used for a, b, and d. The same random number seed was used for a, b, and c.

Table 2
Variability of selected optimal parameters for bottom algae kinetics from 10 GA simulation runs (at final fitness shown in Fig. 4d)

Parameter	Units[a]	Mean	Standard deviation	Range of GA constraints
Maximum unlimited growth rate	mg A/m^2/d	472	23	(0–500)
Respiration	day^{-1} at 20 °C	0.39	0.04	(0–0.5)
Excretion of N and P	day^{-1} at 20 °C	0.39	0.07	(0–0.5)
Death	day^{-1} at 20 °C	0.29	0.08	(0–0.5)
External N half saturation	μg N/L	158	66	(0–300)
External P half saturation	μg P/L	46	38	(0–100)
Subsistence quota for N	mg N/mg A	0.30	0.24	(0.0072–7.2)
Subsistence quota for P	mg P/mg A	0.031	0.029	(0.001–1)
Maximum uptake rate for N	mg N/mg A/d	94	78	(1–500)
Maximum uptake rate for P	mg P/mg A/d	17	12.3	(1–500)

[a] A = chlorophyll *a*, assumed chlorophyll stoichiometry of 1% of dry weight.

allows for separate calibration and confirmation evaluations, or simultaneous calibration of up to three data sets. The following example of calibration would generally be followed by confirmation with independent data.

An example showing the results of an application of the GA for calibration of a water quality model for dissolved oxygen in a small effluent-dominated stream is presented in Fig. 3. The GA was able to accurately calibrate the model over a very wide diel range for dissolved oxygen (Fig. 3). Fitness in this example was defined as the reciprocal of a weighted average of the normalized root mean squared errors of the differences between observed and predicted concentrations of various water quality constituents (Eq. (3)). Since the fitness function included most of the state variables, the same calibration run also resulted in accurate simultaneous calibration of the other water quality constituents, including pH and nutrient concentrations. Model run time for a population of 100 with 100 generations was approximately 6 h using a 3.2 GHz Pentium 4 processor.

Fig. 4 shows the improvement in fitness over the course of the evolution for various (a) evolutionary strategies, (b) crossover methods, (c) populations sizes, and (d) random number seeds. Performance was diminished by using steady state reproduction (Fig. 4a) and arithmetic crossover (Fig. 4b). The best strategy appears to include a combination of elitism and full generational replacement (Fig. 4a). Adjustable mutation and constant mutation performed similarly in this example, although adjustable mutation could be better in case convergence occurs quickly. The best crossover method appears to be a hybrid with equal probability of one-point or two-point, possibly also including the uniform method (Fig. 4b). A population size of 100 performs better than smaller numbers and nearly as well as a population of 500 (Fig. 4c).

The random number seed determines the sequence of random numbers that are generated during the optimization to create the initial population and control the genetic operators. Each of the 10 GA optimizations in Fig. 4d uses a different random number seed but are otherwise identical, using a strategy of elitism, full generational replacement, adjustable mutation, and equal probability of crossover with one-point, two-point, or uniform operators. Each of the 10 GA optimizations in Fig. 4d could be acceptable for calibration even though each has a different set of optimum values for the kinetic rates and constants. Results of the GA allow exploration of the variability of the optimum kinetic rate parameters (Table 2).

5. Conclusions and recommendations

The Excel framework performs well and allows a great deal of flexibility with reasonable computational speed. A hybrid method using one-point, two-point, and uniform crossover combined with a full generational replacement strategy with adjustable mutation and elitism is recommended. Future enhancements could include additional GA methods for selection, crossover, and mutation. Reasonable ranges for parameters and strategies for selecting which parameters to include in the optimization should also be explored. Additional research is also suggested for methods of calculating goodness-of-fit and for the interpretation of the variability of the optimum parameter set.

References

Brown, L.C., Barnwell, T.O., 1987. The Enhanced Stream Water Quality Models QUAL2E and QUAL2E-UNCAS (EPA/600/3-87-007). U.S. Environmental Protection Agency, Athens, GA, 189 pp.

Chapra, S.C., 1997. Surface Water Quality Modeling. McGraw-Hill, New York.

Chapra, S.C., 2003. Engineering water quality models and TMDLs. Journal of Water Resources Planning and Management 129 (4), 247–256.

Charbonneau, P., Knapp, B., 1995. A user's guide to PIKAIA 1.0. NCAR Technical Note 418+IA. National Center for Atmospheric Research, Boulder.

Goldberg, D.E., 1989. Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley.

Gupta, I., Gupta, A., Khanna, P., 1999. Genetic algorithm for optimisation of water distribution systems. Environmental Modelling and Software 14, 437–446.

Metcalfe, T.S., 2001. Computational asteroseismology. Dissertation, presented to the faculty of the graduate school of the University of Texas at Austin, in partial fulfillment of the requirements for the degree of Doctor of Philosophy, University of Texas at Austin, August 2001.

Mulligan, A.E., Brown, L.C., March 1998. Genetic algorithms for calibrating water quality models. ASCE Journal of Environmental Engineering 124 (3), 202–211.

Poeter, E.P., Hill, M.C., 1998. Documentation of UCODE, a computer code for universal inverse modeling. U.S. Geological Survey Water-resources Investigations Report 98-4080, 116 p.

Thomann, R.V., Mueller, J.A., 1987. Principles of Surface Water Quality Modeling and Control. Harper-Collins, New York.

Wang, Q.J., 1997. Using genetic algorithms to optimise model parameters. Environmental Modelling and Software 12 (1), 27–34.

Zou, R., Lung, W.S., November/December 2004. Robust water quality model calibration using an alternating fitness genetic algorithm. ASCE Journal of Water Resources Planning and Management 130 (6), 471–479.