Ecological Modelling 291 (2014) 205–217

Contents lists available at ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel

Improvement of complex and refractory ecological models: Riverine water quality modelling using evolutionary computation

MinHyeok Kim [a], Namyong Park [b], R.I. (Bob) McKay [b], Haisoo Shin [b], Yun-Geun Lee [b], Kwang-Seuk Jeong [c,d], Dong-Kyun Kim [e,*]

[a] Institute of Computer Technology, Seoul National University, Seoul 151-721, Republic of Korea
[b] Department of Computer Science & Engineering, Seoul National University, Seoul 151-721, Republic of Korea
[c] Department of Biological Sciences, Pusan National University, Busan 609-735, Republic of Korea
[d] Institute of Environmental Technology & Industry, Pusan National University, Busan 609-735, Republic of Korea
[e] Department of Physical & Environmental Sciences, University of Toronto, Toronto, Ontario M1C 1A4, Canada

ARTICLE INFO

Article history:
Received 31 March 2014
Received in revised form 24 July 2014
Accepted 25 July 2014
Available online 28 August 2014

Keywords:
Ecological model
River process
Parameterisation
Optimisation
Evolutionary algorithms
Model improvement

ABSTRACT

Complex environmental models have frequently suffered from large discrepancies between prediction and reality, inaccurate quantification of multivariate parameters, and difficulties in dealing with non-linearities. We introduce an interdisciplinary project combining an ecological river-process model and evolutionary optimisation of model parameters, resulting in tools for more effective water resource management. The aim is to more tightly integrate the expert's knowledge and the evolutionary system through an iterated cycle of knowledge refinement and evolutionary search. This requires new methods to specify the expert knowledge in ways that can be integrated into the search. We used an evolutionary algorithm to optimise the multivariate values of the model parameters while retaining their acceptability, verifying that their ranges and values were consistent with ecological knowledge and constraints. The best model had a significantly lower predictive error than the initial process model parameterised from literature estimates. Its error was also over 50% less than those of the purely empirical modelling methods of linear regression and neural network learning. We conclude that combining process knowledge with evolutionary learning can play an important role in ecological modelling.

1. Introduction

Ecological modelling problems frequently combine coarse datasets and weak domain theories (Shan et al., 2006). Ecosystem processes embody enormous complexity in both the structure and functional mechanisms of the system. Striving to model these complex and nonlinear systems, we inevitably introduce simplifications and approximations. Modelling efforts have often aimed to do so without compromising the predictive or explanatory power of the model.

In general, ecological modelling encompasses three types of mathematical or computational methods. The first is process (or mechanistic) modelling. A model is built up from known processes, and parameter values are determined based on the best available knowledge, with perhaps some judicious parameter adjustment at the end of the modelling process to obtain better fit. In this method, the modeller's knowledge and expertise are paramount, and available data is mainly used to validate the model (Brown and Barnwell, 1987; Pei and Ma, 2002). It can work well for simpler model structures that embody a relatively small number of model parameters. As the number of parameters increases, it becomes increasingly likely that some of the expertise-derived parameter values are sub-optimal, so that validating the model becomes more difficult.

The second is heuristic modelling such as machine learning. Some form of machine learning – for example neural network learning (Recknagel, 2001; Yao and Liu, 2001) or genetic programming (Whigham and Recknagel, 2001a; Peterson et al., 2002; Jeong et al., 2003; Kim et al., 2007b) – is used to generate a model from data. The data alone determines the model, which is highly data-driven. These approaches have an important limitation: the amount of data required to learn a model of a given complexity. Since fairly limited data is typically available in ecosystems modelling, this imposes stringent restrictions on the complexity of the models that can be learnt.

The third method is a variant of process modelling, in which the model itself is built from the best available knowledge, but

* Corresponding author. Tel.: +1 4162084878.
E-mail address: dkkim1004@gmail.com (D.-K. Kim).

http://dx.doi.org/10.1016/j.ecolmodel.2014.07.021
0304-3800/

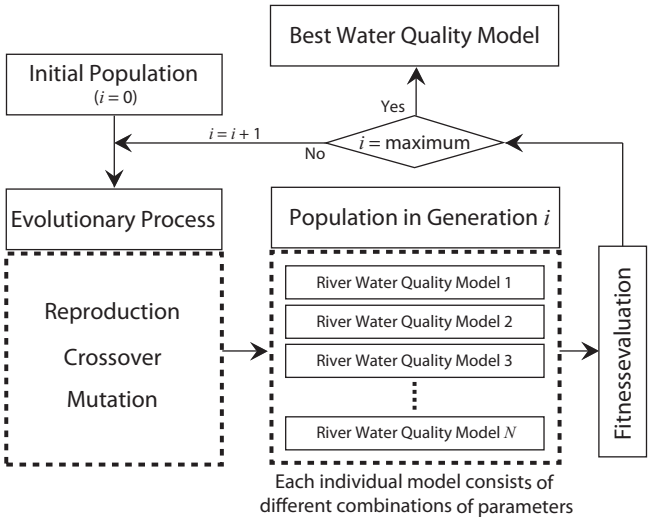

Fig. 1. Conceptual diagram of the use of evolutionary parameter optimisation in river water quality modelling.

the parameter values, instead of being estimated from knowledge, are optimised to give the best fit to the available data (Whigham and Recknagel, 2001b; Cho et al., 2004; Cao et al., 2008). While the second method may sometimes give greater predictive power, it cannot always guarantee greater explanatory power. This third method balances the influences between expertise about the model structure and data observation of the real ecosystem – the model is determined from expertise but parameter values are estimated from observed data.

Although these three methods are very different, they share one important characteristic: limited interaction between the expertise of the ecosystems modeller and that of the computer system developer. An increased use of informed expertise could play an important role in determining both predictability and interpretability of the model. In this context, some modelling studies using Bayesian approaches have generated results allocating different confidences to predicted values based on prior distributions specified for model parameters (Gelman, 2006; Wellen et al., 2014). However their emphasis is on finding relevant parametric bounds within assessed uncertainties, thus differing from our aim of better prediction.

The overarching goal of our research is to obtain an ecologically sound model through cooperation between an expert and a computationally robust system. The knowledge of the ecosystems modeller can influence the behaviour of the optimisation/learning system in intricate ways, and conversely, it is possible to increase the level of self-tuning of the system. This deeper interaction can generate more detailed, and hopefully more accurate, models. In this paper, we investigate how this can benefit modelling methods of the third type noted above.

This paper focuses on the combination of process-based modelling and data-driven parameter estimation. It presents model improvement by use of evolutionary algorithms in parameter adjustment of the process model (i.e. the third type of modelling described above, Fig. 1). We specifically address a water quality modelling program implemented in the lower Nakdong River, Korea. Our study aims to generate better models to predict plankton dynamics in a river ecosystem using evolutionary methods. We strive to improve an existing process-based model through adaptive implementation via evolutionary algorithms. The model we develop is based on generic limnological knowledge of a freshwater ecosystem. The methodological techniques are general, and can be readily extended to other problems. We emphasise that the underlying philosophy is adaptable to a much wider range of ecological modelling problems.

2. Background of research

2.1. Eutrophication and water resource management in fresh waters

Large open freshwater ecosystems contain numerous internal components, but are also affected by unpredictable external forces (Moss, 1998), both natural (e.g. weather variations) and anthropogenic (land use changes, dams and barrages etc.). A river ecosystem is generally seen as a very ambitious domain for modelling. As a consequence of eutrophication of freshwater ecosystems, algal blooms have become ubiquitous in favourable conditions. Rivers around the world are subject to increasing development, and subsequent algal proliferations have become a major concern in many countries. To resolve these problems, establishing guidelines and assisting decision-making through modelling is one of the most promising options for water resources management (Chapra, 1997).

Effective ecosystems management requires robust and reliable predictive models of ecological phenomena. It is almost impossible to manage a river ecosystem effectively without understanding the potential effects of management decisions (Calow and Petts, 1992). In the case of algal blooms, we need to model the effects both of regulatory management – e.g. decisions on water discharge from dams, or controls on nutrient export – and of shifts in Nature – e.g. changes in precipitation levels and timing as a result of climate change. While currently available algal bloom models have often helped to elucidate the ecosystem properties and dynamics, they may be unsatisfactory in terms of structural complexity. The complexity, even of the known system processes, is far beyond what we can hope to model. Thus models can only approximate what we think are the most important influences; but this leaves us hostage to fortune: if we are even slightly wrong in what we choose to model, our models will not perform accurately, as indeed we often find.

Ecological models provide the capability to explain and predict ecosystem dynamics, ranging from specific components to system structure. Both quantitative and qualitative properties of the data sets are crucial in determining the performance and robustness of the ecological models. In building the process structure of the ecological model, the most straightforward method is to base it directly on expert knowledge. However a knowledge based approach does not guarantee an effective model. The data may be too coarse in quality and/or quantity to be used without treatment, and the knowledge used to generate the process may be inaccurate – or more commonly, may not use the most suitable abstractions (Shan et al., 2006).

2.2. Necessity of water quality prediction and evolution of the predictive models in a complex ecosystem

Freshwater ecosystems fall naturally into two groups, lentic and lotic, based on the flow rate of the water body. Lakes, the extreme form of lentic systems, are one of the most popular domains for water quality process modelling. Notably, biogeochemical models have played a key role in lake ecosystem research, and have been used to elucidate ecological patterns from the perspective of system dynamics (Mieleitner and Reichert, 2006). The classical models of lake eutrophication started from empirical models (e.g. statistical regression analysis) and have developed into the mass balance approach (Chapra, 1997; Mooij et al., 2010). To date, a wide variety of lake ecosystem models have been proposed and developed.

For example, Simulation of an Analytical Lake MOdel (SALMO) simulates the major plankton dynamics of lakes and reservoirs (Recknagel and Benndorf, 1982), while the Lake Web model quantifies lake food web interactions (Håkanson and Boulion, 2003).

On the other hand, a riverine ecosystem introduces much greater complexity in modelling water quality, because the determinants of water quality are heavily affected by flow. USEPA's QUAL2E model (Brown and Barnwell, 1987) has been the most widely used river model. Taking our example of algal growth, both the algae themselves, and the nutrients on which they depend, are carried with the flow. But we cannot concentrate purely on the flow, because the algae grow as they are transported. Park and Lee (2002) demonstrated application of the QUAL2E and QUAL2K (improved version of the QUAL2E) to the Nakdong River, Korea. However their study developed an abstracted river model under the simplifying assumption that the river flow was constant. A major drawback of these models was the requirement to calibrate and validate the process under conditions of steady-state flow. Good fit under steady flow does not necessarily guarantee good fit in other conditions. Specifically, the models exhibited limitations in their predictions of water quality, especially during lower (Winter) and higher flow (Summer) periods.

These difficulties influenced ecological modellers to seek other methodologies, one being machine learning. Ecological applications of machine learning algorithms (e.g. artificial neural networks and evolutionary algorithms) have been used to predict water quality and plankton dynamics (Recknagel, 2001). While many water resource agencies still rely on manually-constructed process models, machine learning methods have demonstrated somewhat enhanced performance. Recknagel et al. (2002) demonstrated artificial neural networks (ANN) as a new and promising approach for modelling and prediction of algal blooms, while Kim et al. (2007b) applied evolutionary algorithms in developing a forecasting model of diatom blooms in a river ecosystem. While both studies had acceptable predictive power, they did not provide explicit process representation that could directly elucidate ecological causality. In such cases, sensitivity analysis is generally used to clarify the relationship between explanatory variables and response.

3. Materials and methods

3.1. Description of the study sites and data collection

The Nakdong River system in South Korea is one of the major regulated river systems of East Asia. As Korea's development has progressed, the river has become highly regulated by upstream dams and an estuarine barrage. As a result, its characteristics regularly shift between those of a reservoir and of a river. Regional climatic conditions govern the hydrological regime: over 60% of precipitation falls during the Summer period (June-September) (Jeong et al., 2007). High demand and intensive use of water resources (≈10 million people reside in the Nakdong River Basin) lead to conflicting requirements, a key issue being the occurrence of algal blooms, fueled by the nutrients injected upstream, which periodically blight the river near Busan (≈5 million people) in the lower part of the river. The lower Nakdong River experiences recurrent algal blooms of Summer cyanobacteria and Winter diatoms (Ha et al., 1999, 2003). Mitigating these algal blooms is a key economic and social issue. A wide variety of limnological research in terms of water quality (Kim et al., 1998, 2007a) and plankton dynamics (Ha et al., 1999; Kim and Joo, 2000) has been conducted. So important is the management of this river that the Korean government invested approximately USD 19 billion (for four major rivers, of which this is the largest) in a scheme to improve its water management, and

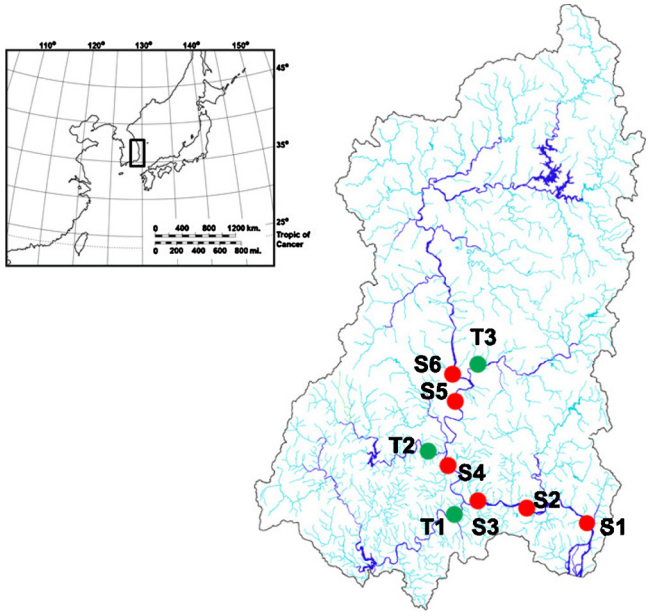

Fig. 2. Map of the Nakdong River basin. The circles indicate measuring stations (study sites); S_1, S_2, S_3, S_4, S_5 and S_6 are on the main channel, and T_1, T_2 and T_3 on major tributaries.

an intensive monitoring programme known as the National/Korean Long-Term Ecological Research (KLTER) has been carried out over the past decade (Kim and Kim, 2011).

In building the models presented in these papers, geographical, hydrological, meteorological, physicochemical and biological datasets were compiled. We also used 13 years (1996–2008) of observation data (e.g. water temperature, irradiance, precipitation, flow rates, nutrient concentrations, and chlorophyll *a*). Most were collected on a daily basis, but nutrient concentrations and chlorophyll *a* concentration were measured weekly at the primary study site S_1 (near the river mouth) and bi-weekly at the upstream study sites. Spatially explicit data were used at nine study sites that were located on the main channel from upstream S_6 to downstream S_1 and tributaries at T_1, T_2 and T_3 (Fig. 2). Hydrological (e.g. flow rate) and meteorological (e.g. irradiance and precipitation) data were provided by the Korean WAter Management Information System (WAMIS, Korean Ministry of Land, Infrastructure and Transport) and the Korea Meteorological Administration (KMA).

3.2. Model description

The river process model used in this paper was first introduced in Kim et al. (2010). We adapted a simple process model to test parameter optimisation, rather than aiming to represent the real ecosystem in full (and probably inaccurate) detail. In this study, we slightly modified the nutrient and temperature equations to more closely reflect commonly-used forms (Cole and Buchak, 1995; Arhonditsis and Brett, 2005a) by incorporating maximum and minimum values of parameters. We have also added a new state variable representing zooplankton, which play a key role in limiting phytoplankton through grazing (Kim and Joo, 2000). The model was modulated by two contemporaneous processes describing the hydrological (flow of water bodies) and biological (plankton dynamics) mechanisms.

Firstly, the flow model uses a simple flow mass balance between stations, and provides flow-at-time information to the biological process model. Eq. (1) shows the basic flow model, which has three

parts; inflow from upstream A, flow retention downstream B, and run-off R by precipitation):

$$F_{B,t+d} = (1 - r_A) \cdot F_{A,t} + r_B \cdot F_{B,t} + R_{B,t+d} \quad (1)$$

where $F_{X,t}$ denotes the flow at station X at time t, d is the time it takes water to flow from station A to station B, and r_X is the fraction of the water that is retained at station X. The pool ratio r_X was obtained from empirical regression against the water velocity. Thus $(1 - r_A) \cdot F_{A,t}$ is the outflow from station A (which should be identical, apart from other inflows, to the inflow at the next station B). The quantity $r_B \cdot F_{B,t}$ is the proportion of flow retained at station B (the pool body) due to non-laminar flow. $R_{B,t+d}$ indicates the inflow arising from run-off of precipitation occurring in the catchment of station B at time $t + d$. A simple additive model is used for the confluence of two streams. Concerning the pool ratio r_X, the water velocity data were provided by the Nakdong River Environment Research Center of the National Institute of Environmental Research (personal communication); they were measured at intermittent intervals during periods of high, normal and low flow. We downscaled the velocities based on site-specific regression using the corresponding flow rates. All the coefficients of determination (r^2) were greater than 0.99.

Second, the biological process interacts bidirectionally with the hydrological process. It determines the temporal dynamics of the phytoplankton biomass (B_{Phyt}), a proxy for the trophic state of water body (i.e. the level of eutrophication). The biological process model mediates the growth or decline of phytoplankton in a flowing body of water over time – specifically, the transit time between stations. This transit time is in turn determined by the distance and corresponding water velocity between successive stations.

$$\begin{aligned}
\frac{dB_{Phyt}}{dt} &= B_{Phyt} \cdot (\mu_{Phyt} - \gamma_{Phyt}) - B_{Zoop} \cdot \varphi \\
\mu_{Phyt} &= C_{UA} \cdot f(V_{lgt}) \cdot g(V_n, V_p, V_{si}) \cdot h(V_{tmp}) \\
\gamma_{Phyt} &= C_{BRA} \cdot e^{Q_{10a}(V_{tmp} - C_{ref})} \\
\varphi &= C_{MFR} \cdot \frac{B_{Phyt} - C_{Fmin}}{K_{FS} + B_{Phyt} - C_{Fmin}} \cdot e^{-C_{ZT}(V_{tmp} - C_{ref})^2} \\
f(V_{lgt}) &= \frac{V_{lgt}}{C_{bl}} \cdot e^{1-(V_{lgt}/C_{bl})} \\
g(V_n, V_p, V_{si}) &= MIN\left(\frac{V_n}{K_n + V_n}, \frac{V_p}{K_p + V_p}, \frac{V_{si}}{K_{si} + V_{si}}\right) \\
h(V_{tmp}) &= MAX(e^{-C_{PT}(V_{tmp} - C_{btp1})^2}, e^{-C_{PT}(V_{tmp} - C_{btp2})^2})
\end{aligned} \quad (2)$$

$$\begin{aligned}
\frac{dB_{Zoop}}{dt} &= B_{Zoop} \cdot (\mu_{Zoop} - \gamma_{Zoop} - \delta_{Zoop}) \\
\mu_{Zoop} &= C_{UZ} \cdot \frac{B_{Phyt} - C_{Fmin}}{K_{FS} + B_{Phyt} - C_{Fmin}} \cdot e^{-C_{ZT}(V_{tmp} - C_{ref})^2} \\
\gamma_{Zoop} &= C_{BRZ} \cdot e^{Q_{10b}(V_{tmp} - C_{ref})} + C_{BMT} \cdot \varphi \\
\delta_{Zoop} &= C_{DZ} \cdot \alpha^{(V_{tmp} - C_{ref})}
\end{aligned} \quad (3)$$

The primary equations for algal biomass were a simplified form incorporating photosynthetic production (μ_{Phyt}), metabolic degradation (γ_{Phyt}) and herbivorous zooplankton grazing activity (φ) (Eq. (2)). Algal growth was subject to multiplicative influences from external forces such as solar radiation (V_{lgt}), water temperature (V_{tmp}) and nutrient concentrations (nitrate V_n, phosphate V_p and silica V_{si}). These limiting functions were partially adapted from the studies of Cho and Shin (1998), Pei and Ma (2002), and Arhonditsis and Brett (2005a). We modified them to use two optimal temperatures for phytoplankton growth, since this river has been dominated by Summer cynanobacteria (Ha et al., 1999) and Winter diatom (Ha et al., 2003) blooms. The optimal values were based on the experimental literature (Cho and Shin, 1998; Reynolds, 2006). Zooplankton abundance (B_{Zoop}) plays a key role in limiting phytoplankton biomass due to grazing pressure (Eq. (3)). The governing equations of zooplankton metabolism and grazing activity were derived from Pei and Ma's study (Pei and Ma, 2002). Specifically we added a temperature-dependent factor for phyto- and zooplankton respiration rates, and both grazing and mortality of zooplankton.

For the process model's parameterisation, we explored and defined their ranges through a rigorous review of related literature (Cho and Shin, 1998; Everbecq et al., 2001; Pei and Ma, 2002; Arhonditsis and Brett, 2005a; Reynolds, 2006). We used the parameter values (reference value, Table 1) as a comparative baseline without model calibration, because our focus is on automated parameter adjustment and manual adjustment would disturb this focus. Numerical bounds of each parameter were set, indicating our level of confidence in ecological knowledge. We set their ranges very loosely so that the search would be certain to include all feasible values. Table 1 describes these parameters in detail.

To recap, the model consists of two top-level processes. The river flow process manages the interaction between stations, while the algal growth process calculates the change of status at each station. All measured data from the four upstream stations, tributary stations T_1, T_2, T_3 and main channel station S_6, were used as sources. The model inputs the upstream data to predict values of downstream stations. We treated confluences between measuring sites as virtual stations. The values of all variables describing a body of water were recalculated based on flow-weighted averages, and then returned to the next downside reach.

3.3. Evolutionary algorithms with Adaptive Operator Selection

The primary focus of this work is to explore parameter ranges and optimise the model accuracy using the well-known genetic algorithm (GA, Holland, 1975; Goldberg, 1989). A GA abstracts natural evolutionary objects such as genes, and the processes that operate on them (reproduction, mutation, crossover and selection), and has been widely applied to solving real world problems. A key feature of such stochastic population-based search algorithms is their low dependence on assumptions about the form of solutions. In particular, they are well suited to intricate search problems that are not easily resolved by ordinary methods. All the parameters are represented by genes, and evolve into new combinations of values by crossover and mutation. Through iterating this process, the GA searches for the best fit of the model by optimising those parameters.

One issue in applying evolutionary algorithms to difficult problems is the match of the genetic operators to the problem; a poor choice of operators can lead to worse performance. A conventional genetic algorithm combines two operators: single-point crossover and mutation. However the effectiveness of operators can vary depending on the complexity of the search space, especially if the space is (as in this case) very large, covering a vast range of combinations of real values. Other operators may be more effective. In this work, we used a range of genetic operators, incorporating them in an adaptive algorithm to select the most effective.

Adaptive Operator Selection (AOS) is an online strategy for selecting the most appropriate variation operators (Goldberg, 1990; Thierens, 2005). It uses historical information from the evolutionary process to optimise model parameters. In our study, the probability of use of each operator is adjusted based on how often its application contributed to improving the best individuals in the genetic population. AOS is thus capable of autonomously identifying the most effective operators and generating better parameter values.

Table 1
Model parameters/variables and exploration bounds.

Parameter/variable	Description	Description	Reference value	Bounds
B_{Phyt}	Phytoplankton biomass (referred to as chl *a*)	d^{-1}		
B_{Zoop}	Zooplankton biomass	$\mu g\,L^{-1}$		
C_{BMT}	Respiration multiplier on grazing	N/A	0.04	0.01–0.07
C_{BRA}	Phytoplankton respiration rate	d^{-1}	0.021	0.0–0.17
C_{BRZ}	Zooplankton respiration rate	d^{-1}	0.05	0.0–0.20
C_{bl}	Best light for phytoplankton	$MJ\,m^{-2}\,d^{-1}$	26.78	24.0–30.0
C_{btp1}	Blue-green optimal temperature	°C	27.0	20.0–34.0
C_{btp2}	Diatom optimal temperature	°C	5.0	1.0–20.0
C_{DZ}	Zooplankton mortality	d^{-1}	0.04	0.01–0.10
C_{Fmin}	Minimum food concentration	$\mu g\,L^{-1}$	1.0	0.1–1.9
C_{MFR}	Maximum feeding rate	d^{-1}	0.19	0.01–0.8
C_{PT}	Temp coefficient for phytoplankton growth	$°C^{-2}$	0.005	0.003–0.2
C_R	Choosing coefficient for feeding	N/A	0.88	0.2–1.0
C_{ref}	Standard temperature	°C	20	
C_{UA}	Maximum growth rate of phytoplankton	d^{-1}	1.89	0.1–4.0
C_{UZ}	Maximum growth rate of zooplankton	d^{-1}	0.15	0.0–0.3
C_{ZT}	Temp coefficient for zooplankton growth	$°C^{-2}$	0.005	0.003–0.2
K_{FS}	Half-saturation constant of food	$\mu g\,L^{-1}$	5.0	4.0–6.0
K_n	Half-saturation constant of nitrogen	$mg\,L^{-1}$	0.0351	0.02–0.05
K_p	Half-saturation constant of phosphorus	$mg\,L^{-1}$	0.00167	0.001–0.020
K_{si}	Half-saturation constant of silica	$mg\,L^{-1}$	0.00467	0.001–0.2
Q_{10A}	Q_{10} coefficient (for BA)	$°C^{-1}$	0.069	0.01–0.13
Q_{10Z}	Q_{10} coefficient (for BZ)	$°C^{-1}$	0.05	0.01–0.09
V_{lgt}	Irradiance	$MJ\,m^{-2}\,d^{-1}$		
V_n	Nitrogen concentration	$mg\,L^{-1}$		
V_p	Phosphate concentration	$mg\,L^{-1}$		
V_{si}	Silica concentration	$mg\,L^{-1}$		
V_{tmp}	Water temperature	°C		
α	Temp coefficient for zooplankton mortality		0.9	
$\delta_{Phyt,Zoop}$	Phyto- and zooplankton mortality	d^{-1}		
$\gamma_{Phyt,Zoop}$	Phyto- and zooplankton respiration rate	d^{-1}		
$\mu_{Phyt,Zoop}$	Phyto- and zooplankton respiration rate	d^{-1}		
φ	Grazing rate of zooplankton	d^{-1}		

In this paper, we used Adaptive Probability Matching (APM, Kim et al., 2012), which combines two earlier AOSs; Adaptive Pursuit (AP, Thierens, 2005) and Probability Matching (PM, Goldberg, 1990). Its dominant feature is to emphasise the most effective operator at each generation (as in AP), but to distribute the remaining operator rate to other operators based on differences between their impacts on performance (as in PM). The detailed algorithm of AOS is shown in Table A.2.

3.4. Experiment configuration

A total of five model-building systems were compared:

1. The reference process directly uses the mathematical model illustrated in Section 3.2, coming under the first methodological approach described in the introduction. All the parameter values were directly derived from the related literature (Table 1). This model serves as the reference for comparison with other approaches. We did not calibrate the model, so that poor performance could be anticipated.
2. Linear regression (LR) is a simple data-driven approach based on linear statistical prediction, and so comes under the second approach described in Section 1.
3. An artificial neural network (ANN) is another data-driven model, well-known as one of the most powerful heuristic algorithms. A feedforward multilayer perceptron was constructed with a single hidden layer half the size of the input layer.
4. Bound-free GA (GA-BF) uses a GA with conventional genetic operations. While the model itself derives from ecological expertise, no expertise was used in controlling the GA – in particular on parameter bounds. There is nothing to present the system learning parameter values outside their realistic ranges.
5. Adaptive Operator Selection GA (GA-AOS) also evolves the model parameters based on the reference process. However we imposed constraints on realistic ranges for the model parameters based on ecological expertise (see Table 2). The ranges were subjectively determined based on a wide literature review (Cole and Buchak, 1995; Cho and Shin, 1998; Everbecq et al., 2001; Park and Lee, 2002; Pei and Ma, 2002; Arhonditsis and Brett, 2005a). Adaptive Operator Selection was used in the GA. We also compared the model performances between different operators (Table A.1). We used the ratio of fitness values between a child and the mean of its parents for evaluating the impact of the operator that created it. To avoid arithmetic overflows that can result from extreme fitness values from unfit individuals, and to focus evolutionary attention on the best individuals, we used only the best 30% of individuals for impact evaluation.

The systems based on the process model used the whole data for prediction; for these, all values for stations other than S_1, and the exogenous values for station S_1, were used to predict the state variables. The measured value for plankton biomass at S_1 was used as the ground truth, from which the root mean squared error (RMSE) was computed. In addition, we also assessed the model performance based on the relative error (RE), another error metric (Arhonditsis and Brett, 2005b). The RE may be a good complement to the RMSE, since the RE-based accuracy can be interpreted independent of the range of values in the response. However, please note that the model still predicts plankton biomass at upper stations (S_2, S_3, S_4 and S_5) based on external forces (including variables at S_6, T_1, T_2 and T_3). The station S_1 is the furthest downstream, and the model's predictive errors are serially correlated along the river from upstream to downstream. Errors may also be generated by mismatched flow mass balance between the stations. The flow mass balance of our model relies on measured data from meteorological

Table 2
Parameter configuration of evolutionary systems.

Parameters		GA-BF	GA-AOS
GA type		Real coded	
Fitness value		RMSE at station S_1	
Number of tuns		500	
Max generation	G_{max}	100	
Population size		100	
Elite size		1	
Tournament size		4	
Operators		1-point crossover, uniform mutation P_{initX} = 0.7, P_{initM} = 0.1	All 8 operators P_{init} = 0.125
Parameter ranges		Fixed	Expert-derived
Operator adaptation			APM
Minimum rate	P_{min}		$\frac{1}{10K}$
Maximum rate	P_{max}		$1-(K-1)\cdot P_{min}$
Adaptation rate	α		0.8
Learning rate	β		0.8

stations. These measurements are the best available, but they have known limitations in accuracy. The evolutionary method will nevertheless reduce discrepancy in the model predictions. Hence it is possible that the parameter values obtained by the GA may reflect not only the actual processes, but also some compensation of the known inaccuracies in the flow measurements. Unfortunately there is no realistic way to disentangle these.

For all the experiments, we had available 13 years of data, from 1996 to 2008.[1] The data was divided into two periods: data from 1996 to 2005 was used for training, and that from 2006 to 2008 was used for testing.

3.5. Sensitivity estimation of the model parameters

From a practical perspective, the accuracy of the model predictions is the most significant measure for policy makers. However the model's structure and function may also be important for scientific elucidation, so simple measurements of model accuracy need to be supported by a comprehensive description of the model's responses, ideally supported by a causal understanding of the reciprocal relationship between input and output variables. Sensitivity to the model parameters is one of the easiest ways to understand the model's causality and assess its reliability. Thus we performed a global sensitivity analysis of the most accurate model arising from these experiments.

The effects of the model parameters on the state variables were evaluated by inducing perturbations of each of the model parameters. Each model parameter varied along a range $\pm\sigma$, i.e. one standard deviation (S.D.) from the value specified by the best model. For the parametric perturbation we used Latin hypercubic sampling (McKay and Joo, 1979), a statistical method for generating a sample of plausible collections of model parameters from a multidimensional distribution. We supplied all 1000 cases of these simulation results to a multiple linear regression. Our aim was to detect significant interactions between the values of the 20 model parameters and the two state variables, *BP* and *BZ* (phytoplankton and zooplankton respectively). The sensitivity was defined and ranked based on semi-partial correlation coefficients (r_{sp}^2) of the regression.

[1] The riverbed profile was subsequently drastically changed by the four rivers project, hence more recent data is not comparable or usable for modelling, though the same methods will be usable in future once sufficient post-reconstruction data has accumulated. The previous modelling effort is far from wasted, since the ability to compare what would have happened under the previous flow regime with what happens now is potentially of very high value.

4. Results and discussion

4.1. Reference process versus data-learning models

The comparative results between the reference process and the data-learning models (LR and ANN) are shown in Fig. 3 and Table 3. The non-calibrated reference process exhibits a poor level of chlorophyll *a* prediction. The prediction was generally greatly overestimated. This pattern was anticipated, since the model parameters were not carefully calibrated by trial and error. However our previous work (Kim et al., 2010) provided a general feel for what could be achieved by reasonable manual calibration (38.76 RMSE). The results in this case would be slightly different, as there have been the small changes in the biological process model, as mentioned in the method section.

LR and ANN generated substantially better predictions – in the ballpark of, but probably slightly better than, what we could expect to obtain by a manual calibration of the model parameters in the reference process. Over the whole period 1996–2008 (both training and test), LR and ANN generated RMSEs of 36.03 and 33.46, and REs of 54% and 50%. The two models' RMSEs were better than the baseline model's predictive capacity (≈38 RMSE). However the models were unsatisfactory for practical application, since both underestimated biomass overall, and neither modelled the annual variability of chlorophyll *a*. Interestingly, while the RMSEs of the two models were lower in training (1996–2005) than test (2006–2008), the REs were higher in the test period for both models. This demonstrates the value of RE as a complement to RMSE, especially when the training and test sets cannot be randomly sampled from the same distribution, as in this case. The two bottom panels of Fig. 3 depict the temporal dynamics of phytoplankton biomass (as chlorophyll *a* concentration), and allow us to compare model performance between the methods. Like the reference model, LR overestimated chlorophyll *a* in 2008. This may reflect change in data attributes (e.g. in the model's external driving forces such as climatic conditions, nutrient export, and hydrological regime), indicative of a key drawback of linear statistical methods. Overall, ANN gave better prediction of chlorophyll *a*, with the timing of events being well-fitted, as would be expected of a widely accepted learning method, although there are some under- or over-estimates when predicting high biomass. However the model's lack of transparency (i.e. black-box nature) limits its value for scientific understanding.

Thus it seems that, if the alternative is an un-fitted or manually-fitted process model, linear regression is preferable. If we are prepared to accept an ANN model, then its nonlinearity can generate further improvements in predictive accuracy, particularly in test set accuracy (i.e. the ability of the model to generalise to unseen data). On the other hand, this may have been partially the result of

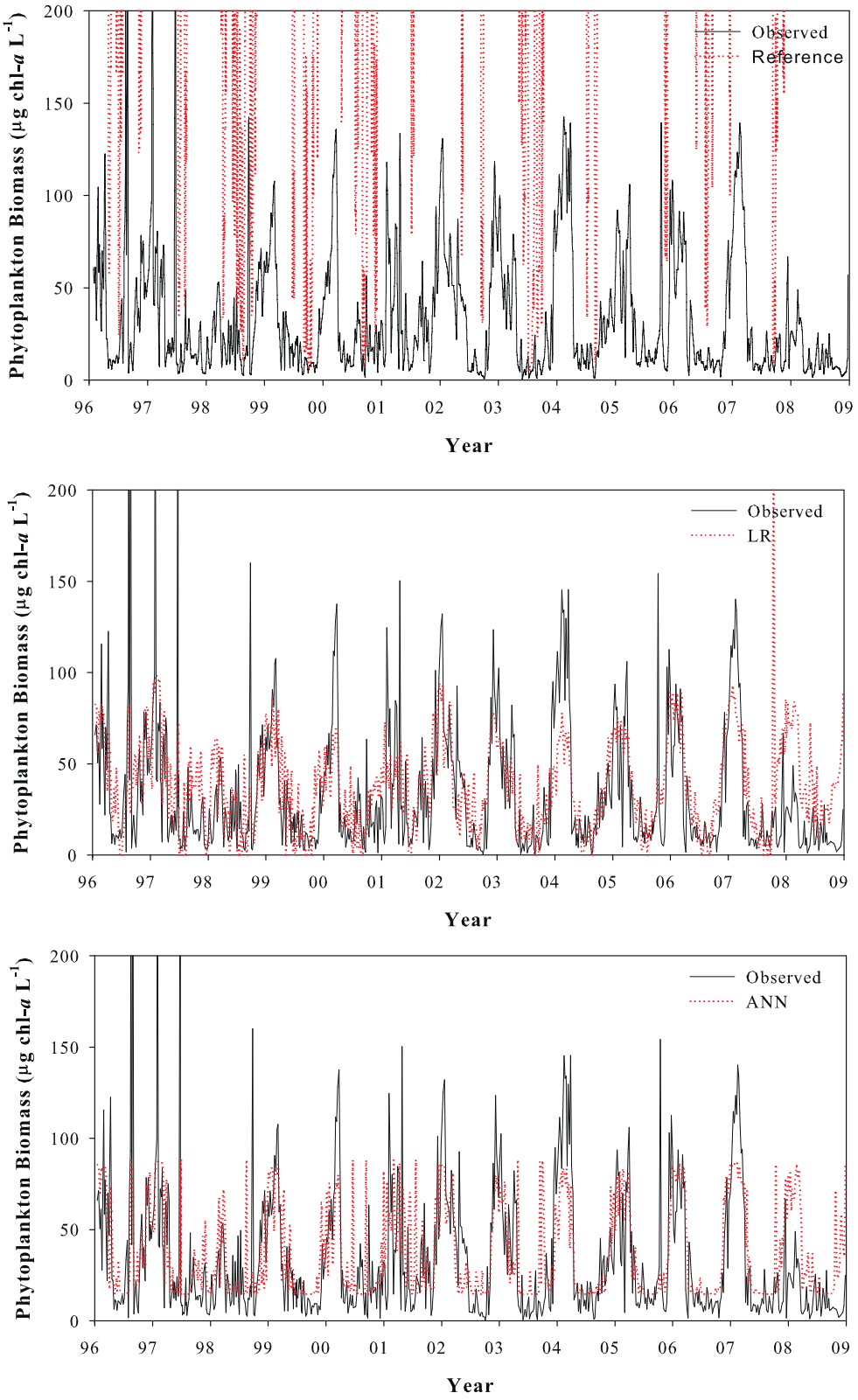

Fig. 3. Phytoplankton prediction at S_1 by time for the simple process model (top), linear regression (middle) and artificial neural network (bottom).

general under-prediction by the ANN model, which served it well in the test period because two of the test years (2008 and 2009) saw relatively low values. Nevertheless neither model is really satisfactory for practical application; they both underestimated the variability of chlorophyll *a*, and neither was particularly good at predicting the peak concentrations, with the ANN in particular predicting many false peaks, while LR failed to fit the trends in the data, drastically overestimating algal production in 2008.

Table 3
Performance comparison between different models. The primary focus of this paper is the test (2006–2008) indicator.

				GA-BF		GA-AOS	
	Reference	LR	ANN	Best	Mean ± S.D.	Best	Mean ± S.D.
RMSE task							
Training (1996–2005)	1.91e+09	36.21	35.56	26.14	27.15 ± 3.64	24.62	24.75 ± 0.11
Test (2006–2008)	1.67e+06	35.40	24.88	13.45	275.71 ± 3023.68	12.03	12.43 ± 0.62
Overall (1996–2008)	1.67e+09	36.03	33.46	23.72	1152.17 ± 9817.94	22.31	22.47 ± 0.12
RE task							
Training (1996–2005)	2.693e+06	0.4816	0.4463	0.3562	0.3521 ± 0.0059	0.3375	0.3470 ± 0.0045
Test (2006–2008)	2.082e+04	0.8400	0.7701	0.2324	0.2539 ± 0.0584	0.2389	0.2497 ± 0.0084
Overall (1996–2008)	2.216e+06	0.5457	0.5042	0.3341	0.3424 ± 0.0679	0.3198	0.3295 ± 0.0041

4.2. Parameter-evolved process models

Before discussing the results from parameter optimisation, we need to address some important issues regarding chlorophyll *a* concentration. The peaks are of most interest, since low algal levels pose no problem. There is clear evidence of seasonality and recurrent variation in time scale, but the magnitude and timing of onset of algal blooms have been irregular (Figs. 2–4). Over the long term, algal blooms were more severe during the 1990s, with peaks moderating since 2002. Algal blooms were formed by diatom species of *Stephanodiscus hantschii* most commonly in Winter, but on the few occasions when they did occur in Summer along with cyanobacteria of *Microcystis aeruginosa*, peak concentrations were much higher ($\approx$570 μg chl-*a* L^{-1} in June 1997).

Table 3 shows the overall performance figures for all techniques (for the GA runs, the means and standard deviations were computed from 500 runs with different random seeds, while the 'best' runs were selected on the basis of training set performance). The first point to note is that the parameter-optimised process models' training set performance was far better than that of both the un-optimised process models and the two pure machine learning techniques. The second obvious point is that expert setting of the parameter ranges is essential. While inexpertly set parameter ranges gave good training performance relative to the non-GA treatments (and the best model was only slightly behind the other GA treatments), the generalisation performance was poor: the corresponding models predicted the 2006–2008 period inaccurately.

When we come to the GA treatments with expert-defined (i.e. knowledge or literature-based) parameter ranges, we note that the error was substantially reduced overall, but it is particularly notable that it was more than halved over the test period: the parameter-evolved process models were able to generalise to the lower algal levels of the test years, and the best models gave less than half the test-set RMSE of the LR and ANN methods. Moreover, the test-set REs consistently decreased in both GA-BF and GA-AOS models, which contrasts with the RE pattern in LR and ANN models (see Table 3). Within these treatments, when we compare between the different operator treatments, we see that in terms of best performance out of 500 runs, the eight-operator version gives worse 'best' performance than the two-operator. The variance is much higher, meaning that the risk of using the eight-operator version is substantially higher: choosing a good pair of operators works better than choosing a wide range of operators. On the other hand, Adaptive Operator Selection reverses this conclusion: the 'best' results are slightly improved relative to the two-operator version, while the average results are mixed, and statistically indistinguishable. Thus choosing GA-AOS may be the safest option overall.

As well as checking the predictive accuracy at S_1, we also evaluated the ability of the best model to predict chlorophyll *a* concentrations along the river, by taking the best model and measuring the accuracy of its predictions of each station from the data for its upstream stations. We saw overall RMSEs of 22.31 (RE = 32.8%, S_1), 25.09 (48.7%, S_2), 15.27 (36.8%, S_3), 18.29 (46.8%, S_4) and 12.97 (37.3%, S_5, from upstream). Thus the model fitted well not only over time, but over space, between different measuring stations. Fig. 4 shows the overall performance of the best models over time. The much better fit, not only to the scale, but also to the timing, of phytoplankton events is readily apparent.

Machine learning has been applied across a range of ecological research demonstrating its explanatory and predictive power (Recknagel, 2001; Peterson et al., 2002; Ahmed and Sarma, 2005; McNyset, 2005; Kim et al., 2007b). However there are also reasons for scepticism: some learning systems, such as neural networks, generate black-box predictions, that are not fully available for inspection; and even white-box systems may not aid in understanding because there is no guarantee that the white-box structures usefully correspond to ecological understanding. On the other hand, starting from an ecologically-informed process model provides increased likelihood that the resulting model will be ecologically comprehensible.

One important issue relates to test set validation. It is relatively common practice to use a leave-one-out validation mechanism, choosing each year as test year, training from all the others, and predicting the test year from the trained model – and repeating this for all years in turn. This is based on standard practice in machine learning, where leave-one-out is one of the commonest methods of validation. However in this context, there are two important issues with this approach, one practical, the other theoretical (but equally important).

The practical issue relates to the start-up procedure for the process model. When we start to run the model, we do not have any available information about the water in the stream system. Hence we run the flow model effectively with initially empty streams, allowing the water to flow through until the complete system is filled with water. We only start to measure model error from that point. When the model error is computed over a few years, this initial startup error is not significant. However when the error is divided into individual years, it can be significant – especially because it is somewhat biased: only the start of the year, an important but atypical period, is lost.

The theoretical issue arises from the underpinnings of the leave-one-out strategy. It relies on the assumption that the instances (in this case, the years) are independent. However this is clearly not true of our data. There is a regular change in the algal behaviour over the time period of the dataset. In particular, there appear to be clear trends in the data – perhaps from climate change, from continuing cleanup of pollution sources in the catchment, or from improved flow management. A leave-one-out strategy would penalise (unfairly) models that are able to handle the trend; instead it would favour (unfairly) models that cancel out the trend, predicting average values over the period, because in a leave-one-out strategy, that is what the test years will be: typical years from the data. Thus it would select models that are likely to predict poorly in the future (precisely because we do not expect the future to be exactly like the past, on the basis of the trends in the data).

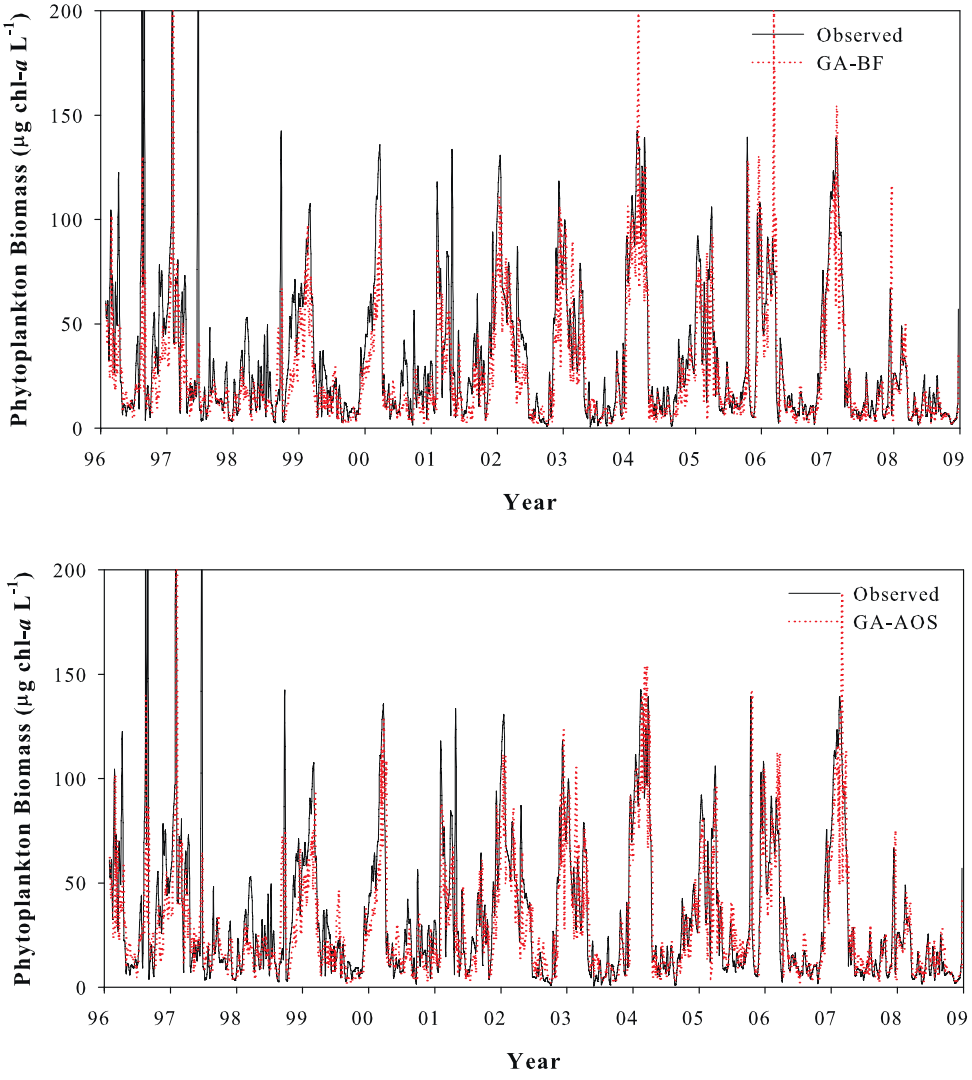

Fig. 4. Phytoplankton prediction at S_1 by time for GA-BF (top) and GA-AOS (bottom).

For these reasons, we based our primary results on prospective prediction for the three years 2006–2008. Tables 3 and A.3 show the underlying basis for this. It shows the RMSE values for the three models (Reference, GA-BF and GA-AOS) for each year from 1996 to 2008. The first point to note is that the training RMSEs were higher than the test ones (Table 3). This, of course, is unusual for machine learning. It reflects unusual characteristics of the data – the later years were more predictable than the earlier (Table A.3), perhaps because of improved measurement, or reduced external noise in the system. We saw a similar pattern, with the same data, in our previous study (Kim et al., 2010). We emphasise that this is a characteristic of the data, not of the methods. In applications of these methods to other data, we are likely to see the more usual behaviour, of test RMSE error being higher than training.

It is worth noting that the expert-bounded-range evolved process models are able to predict well even in unusual years (e.g. 1996–1997 and 2008). Chlorophyll *a* was exceptionally high in 1996–1997 and low in 2008. In Fig. 3, both LR and ANN underestimated the chlorophyll *a* in 1996–1997, and overestimated it in 2008. However, the evolved process models (i.e. GA-BF and GA-AOS) conform to the observations far better than LR and ANN during the unusual periods in 1996–1997 and 2008 (Fig. 4). Interestingly, the fixed-range evolved process models were quite good in most years. On the other hand, they were spectacularly bad (on average) at fitting 1998 in the training period, and 2006 in the test period.

Based on the aforementioned results, it is clear that the reference process model (Reference) was substantially improved by evolutionary parameter fitting. Conventional learning methods (LR and ANN) showed better performances than the Reference model, but were not competitive with GA-BF and GA-AOS.

4.3. Evaluation of the best model

Table 4 summarises the optimised parameters of the evolved process models, comparing uninformed (i.e. with an arbitrary bound) and informed (with expert-specified bound) evolutionary runs. We examined the exploratory power of GA – its ability to search for reasonable and relevant parameters. The uninformed GA model generated unrealistic parameter values, such as a negative growth rate for zooplankton (C_{UZ}) and an optimal temperature of −17 °C for algal growth (C_{btp1}), even though the model performance was quite good apart from 2006 (Tables A.3 and 4). By contrast, the parameter values of the best model obtained with an informed GA were more similar to the reference values, and mostly within realistic ranges. However they still differed from the reference values, with the reference value lying outside one standard deviation of the GA-AOS values for 10 of the 20 parameters (C_{UA},C_{BRA},C_{BRZ},Q_{10A},C_R,C_{btp2},C_{MFR},K_n,C_{PT},C_{ZT}). Some parameter values were more surprising, even though they were within the 'acceptable' range. Fig. 5 shows the statistical distribution of each

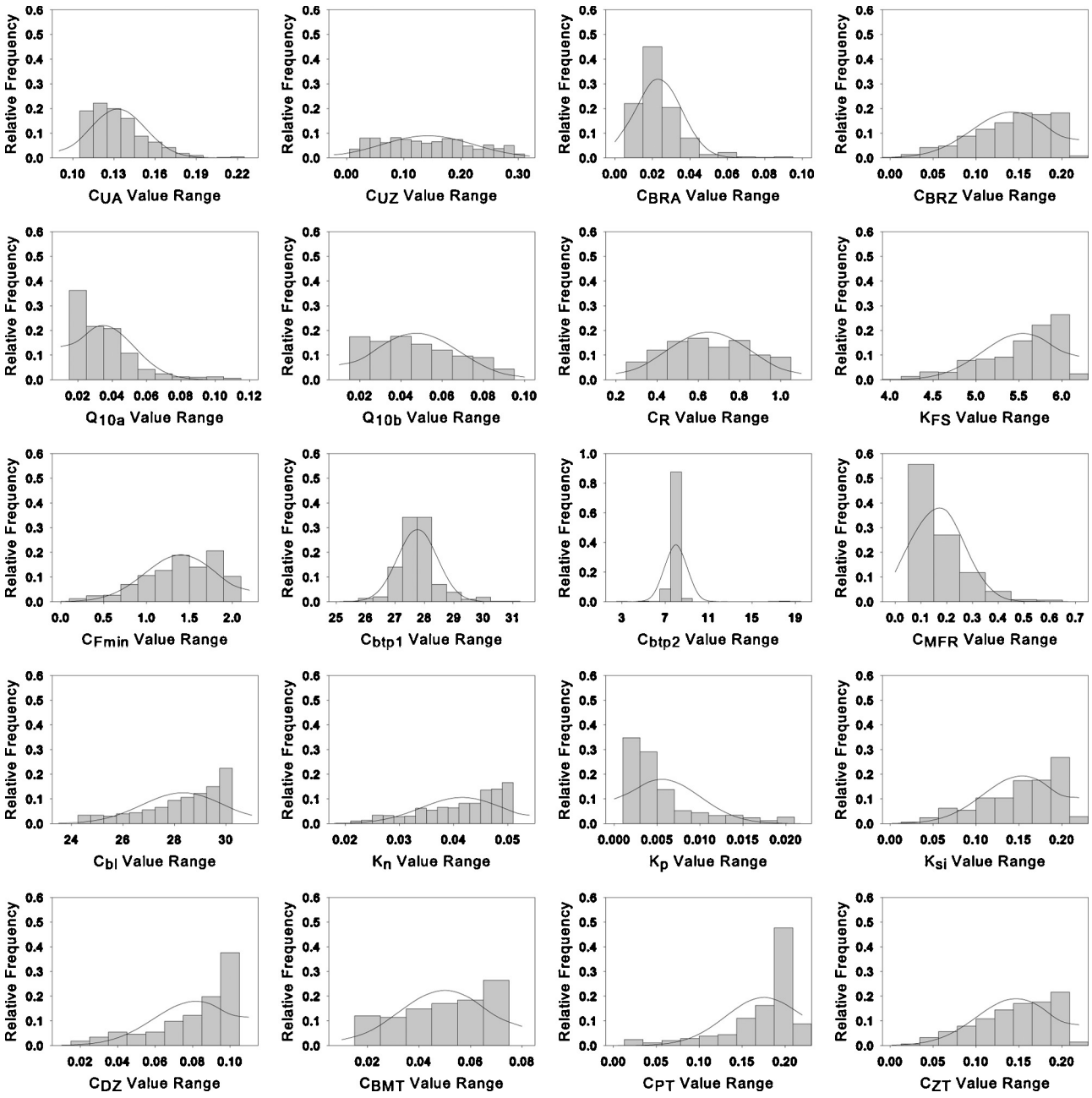

Fig. 5. Statistical distributions for parameter values from best evolved process model (GA-AOS).

parameter obtained from the 500 evolved process models (i.e. candidate solutions).

The growth rates of phyto- (C_{UA}) and zooplankton (C_{UZ}) are of most interest. The C_{UA} parameter was generally modelled as having a quite low value – within the expert's extreme limits of acceptability, but nevertheless highly questionable. We believe that it probably stems from the strong interaction between two parameters (growth rates between phyto- and zooplankton, or possibly growth and respiration rates within phytoplankton) which can lead to high correlation in the model; thus large paired changes in the parameter values can lead to small changes in the overall error – and the system has no way to know that it may be over-fitting these. Thus it seems that asking the ecological expert to specify a feasible range is not sufficient. The expert has much deeper knowledge about the prior likelihood of particular combinations of parameter values. We need to capture this knowledge as a probability distribution – and then incorporate it into the fitness function so that the evolutionary algorithm is aware of it. This will be an important issue in our future work.

Among the major driving factors, we find that water temperature plays a key role in the growth of phyto- and zooplankton. We specified two optimal temperatures (C_{btp1} and C_{btp2}) for the different algal groups present in Winter and Summer. We confined both within a narrow range. The value of C_{btp2} that we assumed for Winter diatoms (e.g. *S. hantzschii*) is somewhat controversial. van Donk and Kilham (1990) stated that the maximum growth of *S. hantzschii* was observed at 20 °C in experiments, while Jung et al. (2011) recently reported that the maximum growth was found at 10 °C. Nonetheless, the abundance of *S. hantzschii* is highest in the 3–6 °C range in the lower Nakdong River (Ha et al., 2003). In our parameter optimisations, C_{btp2} was selected as 17.6 °C in the best model, but in most good models, C_{btp2}s was set around 7.5 °C. Thus

Table 4
Fitted parameter values for the process model. Values which lie outside reasonable (expert-specified) bounds are in italic.

		GA-BF	GA-AOS	
Parameter	Reference	Best	Best	Mean ± S.D.
C_{UA}	1.89	0.410	0.104	0.128 ± 0.020
C_{UZ}	0.15	*−0.156*	0.170	0.133 ± 0.080
C_{BRA}	0.021	0.057	0.008	0.008 ± 0.012
C_{BRZ}	0.05	0.046	0.127	0.131 ± 0.046
Q_{10A}	0.069	0.067	0.011	0.011 ± 0.018
Q_{10Z}	0.05	0.079	0.063	0.064 ± 0.021
C_R	0.88	*1.118*	0.319	0.319 ± 0.204
K_{FS}	5.0	*0.863*	4.432	5.432 ± 0.459
C_{Fmin}	1.0	*10.585*	1.498	1.300 ± 0.415
C_{btp1}	27.0	*−1.197*	27.281	27.511 ± 0.661
C_{btp2}	5.0	5.101	7.464	7.475 ± 0.993
C_{MFR}	0.19	*−0.129*	0.012	0.012 ± 0.098
C_{bl}	26.78	*45.964*	26.859	28.081 ± 1.588
K_n	0.0351	*0.142*	0.044	0.041 ± 0.008
K_p	0.00167	*0.060155*	0.00107	0.004575 ± 0.004408
K_{si}	0.00467	0.141	0.090	0.187 ± 0.045
C_{DZ}	0.04	*0.278*	0.037	0.077 ± 0.022
C_{BMT}	0.04	*0.428*	0.049	0.049 ± 0.018
C_{PT}	0.005	*0.595*	0.198	0.198 ± 0.044
C_{ZT}	0.005	*0.327*	0.078	0.136 ± 0.045

it is clear that more detailed experimentation is required on the true optimal growth temperatures for Winter diatoms.

Other major driving factors such as solar radiation (C_{bl}) and half-saturated coefficients of nutrients (K_n, K_p, and K_{si}) were optimised to easily comprehensible values. 29.7 MJ m^{-2} d^{-1} as the optimal solar radiation was found in Summer, and the nutrient coefficients suggest that the current typical levels of nutrient concentrations (≈2.6 mg N L^{-1}, ≈0.06 mg P L^{-1}, and ≈5.6 mg Si L^{-1}) do not operate as limiting factors in the Nakdong River.

Other parameters were similar to the reference values, but much higher values of C_{PT} and C_{ZT} indicated that the ecosystem process of the Nakdong River might also be significantly driven by temperature-dependent interplay between process components. In this context, Fig. 6 can support this speculative conjecture with explicit evidence indicated by the zooplankton pattern. Zooplankton biomass is significantly different between the two models (Ref and GA-AOS) during the Winter period. The evolved process model experiences dramatic declines of zooplankton biomass, while the reference model produces high densities of zooplankton . This pattern of the GA-AOS model is consistent with ecological understanding of dormancy among zooplankton in Winter.

4.4. Sensitivity analysis and ecological understanding through the model

The sensitivity analysis of the evolved process model involved 20 parameters along with the corresponding phyto- and zooplankton biomass of the Nakdong River (Table 5). We interpreted sensitivity of the model parameters based on the squared semi-partial correlation coefficients of multiple linear regression, and identified the top 7 most influential parameters along the river (S_1 to S_5). For phytoplankton biomass, its growth rate (C_{UA}) exerted 20.5–25.3% influence at S_2 to S_5, while the influence of C_{UA} was reduced almost by half (11.5%) in the lower Nakdong River (S_1). Instead, zooplankton factors (C_{ZT} and C_{MFR}) were identified as the major drivers of phytoplankton dynamics. Our elucidation is supported by Kim et al.'s study (Kim and Joo, 2000), in which zooplankton biomass was larger in the lower Nakdong River, and often induced a clear water phase in Spring and Autumn. In a similar manner, we examined sensitivities of the model parameters on zooplankton biomass. Overall, the influence ranks of the parameters were similar in all sites, and zooplankton biomass seems to be driven by self-regulatory factors such as respiration rate (C_{BRZ}) and mortality (C_{ZT}) rather than by phytoplankton dynamics.

The result of the sensitivity analysis suggests several underlying patterns of phyto- and zooplankton biomass along the Nakdong River:

1. The major driving factors differ with longitudinal distribution.
2. Zooplankton seems to be primarily affected by self-metabolic rates that can be affected by temperature, while phytoplankton can be significantly influenced by zooplankton activity.
3. The effect of zooplankton grazing on phytoplankton is approximately twice as strong in the lower part of the river (based on feeding rates from S_1 to S_5).

5. Synopsis

5.1. Summary

The experimental results have validated the process model with evolved parameters as generally correct, and able to fit (and

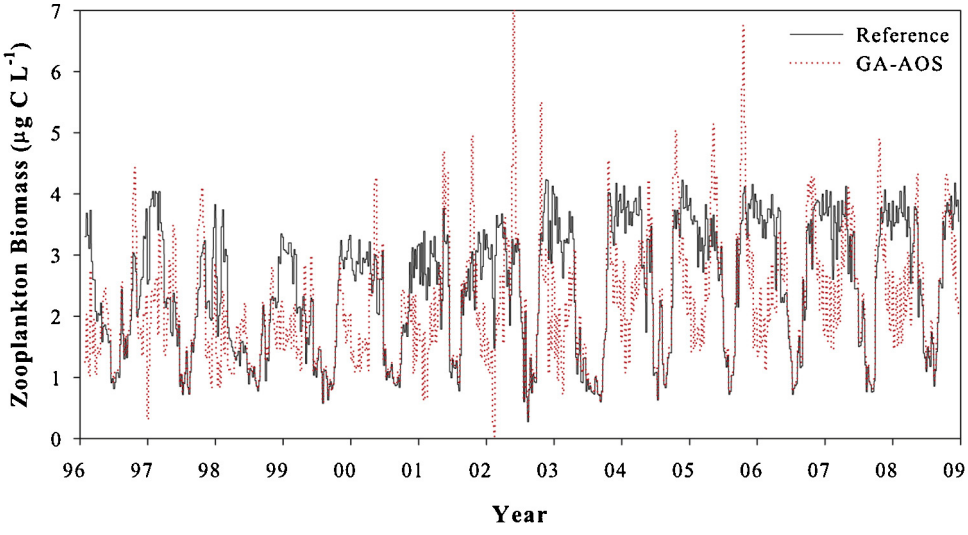

Fig. 6. Comparative zooplankton dynamics at S_1 between reference and best evolved process model (GA-AOS).

Table 5
Sensitivity analysis for the 20 parameters of the best evolved process model. The bracketed values indicate multiple regression r^2 values.

	S_1	r^2_{sp}	S_2	r^2_{sp}	S_3	r^2_{sp}	S_4	r^2_{sp}	S_5	r^2_{sp}
BP	(0.968)		(0.977)		(0.979)		(0.981)		(0.978)	
	C_{ZT}	0.188	C_{UA}	0.205	C_{UA}	0.230	C_{UA}	0.253	C_{UA}	0.243
	C_{MFR}	0.135	C_{ZT}	0.120	C_{ZT}	0.105	C_{BRA}	0.105	C_{BRA}	0.102
	C_{UA}	0.115	C_{BRA}	0.099	C_{BRA}	0.150	C_{ZT}	0.090	C_{ZT}	0.100
	C_{BRA}	0.110	C_{MFR}	0.074	C_{MFR}	0.074	C_{MFR}	0.062	C_{MFR}	0.076
	C_{BRZ}	0.044	C_{BRZ}	0.048	C_{BRZ}	0.045	C_{BRZ}	0.035	C_{BRZ}	0.038
	K_p	0.040	K_p	0.043	K_p	0.042	K_p	0.031	C_{PT}	0.024
	C_{PT}	0.013	C_{PT}	0.018	C_{PT}	0.020	C_{PT}	0.025	K_p	0.023
BZ	(0.944)		(0.898)		(0.906)		(0.917)		(0.931)	
	C_{BRZ}	0.341	C_{BRZ}	0.326	C_{BRZ}	0.334	C_{BRZ}	0.349	C_{BRZ}	0.351
	C_{DZ}	0.331	C_{DZ}	0.257	C_{DZ}	0.261	C_{DZ}	0.262	C_{DZ}	0.285
	C_{ZT}	0.028	C_{ZT}	0.037	C_{ZT}	0.036	C_{ZT}	0.036	C_{ZT}	0.032
	C_{UZ}	0.012	C_{UZ}	0.019	C_{UZ}	0.020	C_{UZ}	0.019	C_{UZ}	0.020
	C_{bl}	0.010	C_{bl}	0.018	C_{bl}	0.017	C_{bl}	0.015	C_{bl}	0.013
	C_{MFR}	0.004	C_{MFR}	0.007	Q_{10Z}	0.007	Q_{10Z}	0.005	C_{MFR}	0.005
	Q_{10Z}	0.003	Q_{10Z}	0.006	C_{MFR}	0.006	C_{PT}	0.004	Q_{10Z}	0.004

predict) the algal status of the river system accurately. Combined with reasonably accurate stream flow estimation, the model should be sufficiently accurate to guide management decisions for the river system, because our hydrological process is modulated by flow-rate based mass balance exchange between stations.

The performance is substantially better than regression and neural network models (that perforce treat the system as if it were a lake), and very much better than the un-optimised process model. It is also clear that expert input into the parameter bounds is essential: without this, the model fit was poor, and the predictive accuracy much lower. Moreover without this expertise, the system evolved parameter values that were physically impossible. Physical implausibility undermines the value of a process-based model, since counter-intuitive values mean that ecologists will not accept the model as explanatory, no matter how well it predicts.

The evolutionary experiments also examined adaptive operator selection techniques, aimed at reducing the need for algorithm expertise in selecting suitable operators (and thus restricting the expertise requirements to the ecological subject matter). They demonstrated that it was possible to remove the need for this algorithm expertise, generating results at least as good as those given by careful expert selection of operators.

5.2. Future work

The research reported here aimed only at adapting parameter values of the process model. A parallel, and somewhat more ambitious, arm of this project adapts the process model itself. The aim is to allow for the possibility that the process model is oversimplified, either abstracting out processes that are important in the real world, or perhaps incorrectly representing the form of the equations. This work uses genetic programming to generate new versions of the model, better fitted to the data. We hope to submit results of this work soon.

A related issue is how to fully utilise the expert's knowledge about likely parameter values. As we have seen, the form of expertise we used here (simple univariate parameter bounds) generated more believable parameter values and better model fit. On the other hand, it under-utilises the available expert knowledge: the expert knows not merely that some values are impossible, but also that other values, while possible, are extremely unlikely – they should only be accepted if there is no other way to assure low error on the data. This requires incorporating the expert's knowledge into the system as a probabilistic prior, and using probabilistic reasoning to determine the best solution. Work along these lines is currently in progress.

Acknowledgments

Water quality data were provided by National Long-Term Ecological Research Programme of Korea. Seoul National University Institute for Computer Technology provided research facilities for this study, which was supported by the Basic Science Research Program of the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (Project No. 2010-0012546), and the BK21-IT program of MEST.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ecolmodel.2014.07.021.

References

Ahmed, J.A., Sarma, A.K., 2005. Genetic algorithm for optimal operating policy of a multipurpose reservoir. Water Resour. Manage. 19 (2), 145–161.
Arhonditsis, G.B., Brett, M.T., 2005a. Eutrophication model for Lake Washington (USA): Part I. Model description and sensitivity analysis. Ecol. Modell. 187 (2–3), 140–178.
Arhonditsis, G.B., Brett, M.T., 2005b. Eutrophication model for Lake Washington (USA): Part II. Model calibration and system dynamics analysis. Ecol. Modell. 187 (2–3), 179–200.
Brown, L.C., Barnwell, T.O., 1987. The Enhanced Stream Water Quality Models QUAL2E and QUAL2E-UNCAS: Documentation and User Manual. US Environmental Protection Agency, Office of Research and Development, Environmental Research Laboratory.
Calow, P., Petts, G.E., 1992. Rivers Handbook: Hydrological and Ecological Process, vol. 1. Wiley-Blackwell, Oxford.
Cao, H., Recknagel, F., Cetin, L., Zhang, B., 2008. Process-based simulation library SALMO-OO for lake ecosystems: Part 2. Multi-objective parameter optimization by evolutionary algorithms. Ecol. Inform. 3, 181–190.
Chapra, S.C., 1997. Surface Water-Quality Modeling. McGraw-Hill Series in Water Resources and Environmental Engineering. McGraw-Hill, New York.
Cho, J.H., Seok Sung, K., Ryong Ha, S., 2004. A river water quality management model for optimising regional wastewater treatment using a genetic algorithm. J. Environ. Manage. 73 (3), 229–242.
Cho, K.-J., Shin, J.-K., 1998. Growth and nutrient kinetics of some algal species isolated from the Naktong River. Algae 13 (2), 235–240.
Cole, T.M., Buchak, E.M., 1995. CE-QUAL-W2: A Two-Dimensional, Laterally Averaged, Hydrodynamic and Water Quality Model, Version 2.0. User Manual. Instruction Report EL-95-1. Tech. Rep., DTIC Document.
Everbecq, E., Gosselain, V., Viroux, L., Descy, J.-P., 2001. POTAMON: a dynamic model for predicting phytoplankton composition and biomass in lowland rivers. Water Res. 35 (4), 901–912.
Gelman, A., 2006. Prior distributions for variance parameters in hierarchical models. Bayes. Anal. 1 (3), 151–534.
Goldberg, D., 1989. Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley Longman Publishing Co. Inc., Boston, MA, USA.
Goldberg, D.E., 1990. Probability matching, the magnitude of reinforcement, and classifier system bidding. Mach. Learning 5 (4), 407–425.
Ha, K., Cho, E.-A., Kim, H.-W., Joo, G.-J., 1999. Microcystis bloom formation in the lower Nakdong River, South Korea: importance of hydrodynamics and nutrient loading. Mar. Freshw. Res. 50, 89–94.

Ha, K., Jang, M.-H., Joo, G.-J., 2003. Winter *Stephanodiscus* bloom development in the Nakdong River regulated by an estuary dam and tributaries. Hydrobiologia 506/509, 221–227.

Håkanson, L., Boulion, V.V., 2003. A general dynamic model to predict biomass and production of phytoplankton in lakes. Ecol. Modell. 165, 285–301.

Holland, J., 1975. Adaptation in Natural and Artificial Systems, vol. 1. University of Michigan Press, Ann Arbor, MI, pp. 5.

Jeong, K.-S., Kim, D.-K., Joo, G.-J., 2007. Delayed influence of dam storage and discharge on the determination of seasonal proliferations of *Microcystis aeruginosa* and *Stephanodiscus hantzschii* in a regulated river system of the lower Nakdong River (South Korea). Water Res. 41 (6), 1269–1279.

Jeong, K.-S., Kim, D.-K., Whigham, P., Joo, G.-J., 2003. Modelling *Microcystis aeruginosa* bloom dynamics in the Nakdong River by means of evolutionary computation and statistical approach. Ecol. Modell. 161 (1), 67–78.

Jung, S.W., Min Joo, H., Kim, Y.-O., Hwan Lee, J., Han, M.-S., 2011. Effects of temperature and nutrient depletion and reintroduction on growth of *Stephanodiscus hantzschii* (bacillariophyceae): implications for the blooming mechanism. J. Freshw. Ecol. 26 (1), 115–121.

Kim, D.-K., Cao, H., Jeong, K.-S., Recknagel, F., Joo, G.-J., 2007a. Predictive function and rules for population dynamics of *Microcystis aeruginosa* in the regulated Nakdong River (South Korea), discovered by evolutionary algorithms. Ecol. Modell. 203 (1–2), 147–156.

Kim, D.-K., Jeong, K.-S., Whigham, P.A., Joo, G.-J., 2007b. Winter diatom blooms in a regulated river in South Korea: explanations based on evolutionary computation. Freshw. Biol. 52 (10), 2021–2041.

Kim, D.-K., McKay, B., Shin, H., Lee, Y.-G., Nguyen, X.H.,2010. Ecological application of evolutionary computation: improving water quality forecasts for the Nakdong River, Korea. In: World Congress on Computational Intelligence. IEEE Computational Intelligence Society, IEEE Press, Barcelona, pp. 2005–2012.

Kim, E.-S., Kim, Y.-S., 2011. Current status of Korea long-term ecological research (KLTER) network activities compared with the framework activities of the long-term ecological research (LTER) networks of the United States and China. J. Ecol. Field Biol. 34 (1), 19–29.

Kim, H.-W., Ha, K., Joo, G.-J., 1998. Eutrophication of the lower Nakdong River after the construction of an estuarine dam in 1987. Int. Rev. Hydrobiol. 83, 65–72.

Kim, H.-W., Joo, G.-J., 2000. The longitudinal distribution and community dynamics of zooplankton in a regulated large river: a case study of the Nakdong River (Korea). Hydrobiologia 438 (1), 171–184.

Kim, M.-H., McKay, R.I.B., Kim, D.-K., Nguyen, X.H., 2012. Evolutionary operator self-adaptation with diverse operators. In: Moraglio, A., Silva, S., Krawiec, K., Machado, P., Cotta, C. (Eds.), Genetic Programming. Vol. 7244 of Lecture Notes in Computer Science. Springer-Verlag, Heidelberg, Germany, pp. 230–241.

Korea Meteorological Administration. Korea Meteorological Administration Website. http://www.kma.go.kr/

Korean Ministry of Land, Infrastructure and Transport. Korean Water Management Information System. http://www.wamis.go.kr

McKay, M.D.J.B.R., Conover, W.J., 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 21 (2), 239–245.

McNyset, K.M., 2005. Use of ecological niche modelling to predict distributions of freshwater fish species in kansas. Ecol. Freshw. Fish 14 (3), 243–255.

Mieleitner, J., Reichert, P., 2006. Analysis of the transferability of a biogeochemical lake model to lakes of different trophic state. Ecol. Modell. 194, 49–61.

Mooij, W., Trolle, D., Jeppesen, E., Arhonditsis, G., Belolipetsky, P., Chitamwebwa, D., Degermendzhy, A., DeAngelis, D., De Senerpont Domis, L., Downing, A., Elliott, J., Fragoso, C., Gaedke, U., Genova, S., Gulati, R., Håkanson, L., Hamilton, D., Hipsey, M., 't Hoen, J., Hülsmann, S., Los, F., Makler-Pick, V., Petzoldt, T., Prokopkin, I., Rinke, K., Schep, S., Tominaga, K., Van Dam, A., Van Nes, E., Wells, S., Janse, J., 2010. Challenges and opportunities for integrating lake ecosystem modelling approaches. Aquat. Ecol. 44 (3), 633–667.

Moss, B., 1998. Ecology of Fresh Waters: Man and Medium, Past to Future. Blackwell Science Ltd., Osney Mead.

Park, S.S., Lee, Y.S., 2002. A water quality modeling study of the Nakdong River, Korea. Ecol. Modell. 152 (1), 65–75.

Pei, H., Ma, J., 2002. Study on the algal dynamic model for West Lake, Hangzhou. Ecol. Modell. 148 (1), 67–77.

Peterson, A.T., Ball, L.G., Cohoon, K.P., 2002. Predicting distributions of Mexican birds using ecological niche modelling methods. Ecol. Modell. 144 (1), E27–E32.

Recknagel, F., 2001. Applications of machine learning to ecological modelling. Ecol. Modell. 146 (1–3), 303–310.

Recknagel, F., Benndorf, J., 1982. Validation of the ecological simulation model "SALMO". Internationale Revue der Gesamten Hydrobiologie 67 (1), 113–125.

Recknagel, F., Bobbin, J., Whigham, P., Wilson, H., 2002. Comparative application of artificial neural networks and genetic algorithms for multivariate time-series modelling of algal blooms in freshwater lakes. J. Hydroinform. 4 (2), 125–133.

Reynolds, C.S., 2006. The Ecology of Phytoplankton. Cambridge University Press, Cambridge, United Kingdom.

Shan, Y., Paull, D., McKay, R.I., 2006. Machine learning of poorly predictable ecological data. Ecol. Modell. 195, 129–138.

Thierens, D.,2005. An adaptive pursuit strategy for allocating operator probabilities. In: Proceedings of the 7th Annual Conference on genetic and Evolutionary Computation (GECCO 2005). ACM, pp. 1539–1546.

van Donk, E., Kilham, S.S., 1990. Temperature effects on silicon- and phosphorus-limited growth and competitive interactions among three diatoms. J. Phycol. 26, 40–50.

Wellen, C., Arhonditsis, G.B., Labencki, T., Boyd, D., 2014. Application of the sparrow model in watersheds with limited information: a Bayesian assessment of the model uncertainty and the value of additional monitoring. Hydrol. Process. 28, 1260–1283.

Whigham, P.A., Recknagel, F., 2001a. An inductive approach to ecological time series modelling by evolutionary computation. Ecol. Modell. 146 (1–3), 275–287.

Whigham, P.A., Recknagel, F., 2001b. Predicting chlorophyll-a in freshwater lakes by hybridising process-based models and genetic algorithms. Ecol. Modell. 146 (1–3), 243–251.

Yao, X., Liu, Y., 2001. Evolving neural networks for chlorophyll-a prediction. In: Fourth International Conference on Computational Intelligence and Multimedia Applications, 2001. ICCIMA 2001. Proceedings, pp. 185–189.