Review Paper

J. Ecol. Field Biol. 33(4): 275-288, 2010
DOI: 10.5141/JEFB.2010.33.4.275

JEFB Journal of Ecology and Field Biology

Model development in freshwater ecology with a case study using evolutionary computation

Dong-Kyun Kim[1], Kwang-Seuk Jeong[2], Robert Ian (Bob) McKay[1], Tae-Soo Chon[2], Hyun-Woo Kim[3] and Gea-Jae Joo[2,*]

[1]School of Computer Science & Engineering, Seoul National University, Seoul 151-744, Korea
[2]Department of Biology, Pusan National University, Busan 609-735, Korea
[3]Department of Environmental Education, Sunchon National University, Suncheon 540-742, Korea

Ecological modeling faces some unique problems in dealing with complex environment-organism relationships, making it one of the toughest domains that might be encountered by a modeler. Newer technologies and ecosystem modeling paradigms have recently been proposed, all as part of a broader effort to reduce the uncertainty in models arising from qualitative and quantitative imperfections in the ecological data. In this paper, evolutionary computation modeling approaches are introduced and proposed as useful modeling tools for ecosystems. The results of our case study support the applicability of an algal predictive model constructed via genetic programming. In conclusion, we propose that evolutionary computation may constitute a powerful tool for the modeling of highly complex objects, such as river ecosystems.

Key words: complex river ecosystem, data learning process, ecological modeling, evolutionary computation, phytoplankton proliferation, time-series prediction

INTRODUCTION

Diverse ecosystem phenomena arising from combinations of living organisms and their interactions with the physical environment are highly nonlinear, very complex, and frequently chaotic (Fielding 1999). In a Newtonian physical simulation of a thrown ball, for example, it is necessary to incorporate factors such as the effects of gravity, the mass of the ball, etc. In many circumstances, we can ignore aspects such as air density, wind, etc. In other circumstances (golf, baseball), however, these factors may prove important. Generally, we know with a fair degree of accuracy what must be included and what must be omitted to construct a model with the required level of accuracy, according to the degree of relevance to the issue. By way of contrast, we frequently possess little of this type of knowledge in the study of ecology, which is not the case in physics or chemistry.

Hence, it can prove quite difficult to forecast and explain the broad variety of environmental aspects and their emergent behaviors in ecosystems, especially as compared to other existing scientific systems. Ecological modeling faces some unique problems in dealing with complex environment-organism relationships, and is one of the toughest domains that might be encountered by a modeler. The relevant difficulties derive both from the complexity of the systems being modeled and the quality and quantity of data available for model development (Shan et al. 2006).

However, this is not the only reason that ecological

Received 07 July 2010, **Accepted** 20 September 2010

*Corresponding Author
E-mail: gjjoo@pusan.ac.kr
Tel: +82-51-510-2258

modeling is difficult. The data, too, may introduce some difficulty. Ecological data is frequently both rough and noisy, particularly when it is sampled from the field. Field sampling is generally expensive, since it is often collected by hand. The data is frequently sparse (missing) and/or collected in an irregular fashion, owing to exceptional conditions including illness, equipment failure, or holidays. As reported previously by Lek (2007), ecological data frequently contains sampling errors and measurement and intermittent estimation mistakes, thereby introducing uncertainty into the resultant models. Moreover, the sources of errors themselves may be biased, thus creating errors that are correlated with the measurements.

Although the development of sampling and measuring technologies for data collection has ameliorated these problems to some degree, many ecological datasets have been collected over periods of multiple years, and these changes have had limited impact thus far. The issues relevant to coping with imperfect data remain very important in the field of ecological modeling. Moreover, additional difficulties arise from the large numbers of variables relative to the number of instances in ecological datasets – because we do not generally know which relationships are important, we tend to include all available measurements, rather than risk omitting an important one. Consequently, the redundant variables may create additional difficulties in the development of automated modeling methods.

A broad variety of methods have been employed in the development of such models, ranging from classical mathematical modeling (Recknagel and Benndorf 1982, Chapra and Reckhow 1983) to evolutionary computation (EC) (Kim et al. 2007b, Cao et al. 2008). EC can be used to create automatic functions or models, producing diverse candidates with a nonlinear computational structure; EC, as well as artificial neural networks (ANN), has yielded promising results in terms of the prediction certain environmental phenomena in ecological research (Recknagel et al. 2002, Cho and Sung 2004, Park et al. 2006a). In this paper, we discuss the relevance of EC to ecological modeling, illustrating it with an application to water quality modeling, and specifically to plankton population dynamics.

The remainder of this paper is structured as follows. First, we detail the relevant background of ecological modeling, describing the wide range of techniques that have been used thus far. We then attempted to identify the appropriate situations for the use of ecological modeling. We described some nature-inspired computational methods (of which EC constitutes a sub-class). We then investigated the important considerations to be taken into consideration in the development of an ecological model. We illustrate this via specific applications to water quality, and then conclude with a discussion of the applicability of EC to ecological modeling.

NECESSITY OF ECOLOGICAL MODELING

Why is ecological modeling special?

The model can be broadly defined as a specific representation of a system, in which each component involves a combination of relationships and interactions. In some cases, the models do not reflect the full mechanisms of the dynamic and integrated systems – relatively simple model approaches such as regression, logistic-type models and predator-prey models may be employed in order to gain insight into general principles and probabilities (Lotka 1925, Volterra 1926, Schaefer 1968, Boerema and Gulland 1973, Cloern 1996). However, the ultimate objective of almost all ecological model construction is the construction of a system that can reproduce and simulate patterns of outcomes. Thus, the constructed models must be sufficiently sophisticated to accurately represent the target system, with the additional assumption that all of the knowledge is suited to the representation. Such models can be employed in the interpretation of general possibilities or the prediction of outcomes for particular populations, communities, or ecosystems.

Initially, ecosystems researchers engaged in great debates as to whether *in vitro* or *in vivo* investigations were more appropriate for ecosystems research. Although both experimental approaches require more time and effort than mathematical or other theoretical approaches, such experiments do not guarantee a high probability that the system's performance will be satisfactory. This has compelled researchers to search for methods capable of representing the target system in ways suitable to the principal objectives of ecological modeling.

Many approaches to ecosystems modeling have been developed that reproduce a system and reveal interactions and relationships, particularly when other experimental approaches prove impossible or impractical. Since Eugene Odum introduced theoretical modeling methods for use in systems ecology (Odum 1983), a number of models have been constructed in efforts to elucidate ecological processes more accurately. Jørgensen (1992) previously proposed the concept of exergy, as well as methods for computing ecosystem quality, to better un-

derstand the information level and interactions between ecological theory and the models. Deaton and Winebrake (2000) previously surveyed a variety of dynamic models that could be applied to environmental systems to model growth patterns, coupled predator-prey populations, water pollution, global warming, and so forth.

Ecological issues for freshwater systems in South Korea

These recent developments in modeling techniques have been previously applied to a case study examining algal communities in freshwater ecosystems. In Korea, modeling has been more frequently applied to the fields of hydrology and hydraulics than to limnology and freshwater ecology (Park and Lee 2002, Cho and Sung 2004). In this paper, we demonstrate the application of EC to ecological analysis and modeling in the context of Korean freshwater ecosystems.

The majority of freshwater ecosystems in South Korea no longer bear any resemblance to natural streams or lakes. They have generally been heavily modified by physical alterations, including dam construction and estuarine barrages (Kim et al. 1998, Kim et al. 2004). Trophic states are largely nutrient-enriched due to the approximately forty million people residing within this relatively small area (Joo et al. 1997). Additionally, climate characteristics, particularly the biased rainfall pattern (rainy summer and dry winter), are known to accentuate the effects of this freshwater eutrophication. Korean freshwater ecosystems, therefore, differ profoundly from, and perhaps are more complex to model than some other modified ecosystems.

MAJOR APPROACHES TO ECOLOGICAL MODELING

Conventional modeling

Statistical methods have been extensively employed for the analysis of datasets across different scientific regimes. In the field of ecological research, statistical analysis has given rise to the increasingly important field of biostatistics (Zar 1999). In the infancy of this discipline, readily applicable linear and statistical approaches were employed to isolate and identify significant ecosystem properties. In particular, many ecologists have analyzed their experimental data primarily via multivariate analyses such as principal component analysis (PCA) and canonical correspondence analysis. These ordination methods have commonly been employed in efforts to simplify the aquatic ecology data (Magadza 1980, Matta and Marshall 1984, van Tongeren et al. 1992, ter Braak and Verdonschot 1995, Romo et al. 1996). The limitations of these methods have been well established (e.g., horseshoe and arch effects). However, we do not discuss this in depth herein, since EC seldom deals with ordination methods, especially in ecological areas.

Second, a variety of time-series analyses have also been employed. In statistical approaches, multivariate linear regression (MLR) methods are probably the most popular. However, they are limited in several ways, including the presence of strong distortion deriving from nonlinear relations attributable to outliers, heteroscedasticity, and colinearity (Zuur et al. 2009). Among more advanced linear methods, an autoregressive model is a type of random process employed in the prediction of certain types of values and phenomena. AutoRegressive (integrated) moving averages (ARMA/ARIMA) are representatives, which are used for the prediction of continuous values, particularly in time-series analyses. Harding and Perry (1997) predicted a long-term increase in phytoplankton biomass using ARMA, and Mishra and Desai (2006) conducted comparative experiments between linear statistical models and neural networks to forecast droughts on the basis of the precipitation index of the river basin. Recently, Jeong et al. (2008) also compared forecasting performances between ARIMA and autoregressive ANN in predicting chlorophyll *a*. Generally, these approaches appear to have a somewhat limited ability to capture non-stationary and nonlinear peaks in ecological data. Consequently, ecologists searching for better prediction methods have become increasingly interested in artificial intelligence methods, which are able to deal with data in highly nonlinear structures.

In addition to linear statistical approaches, mathematical and numerical modeling techniques provide some of the most common tools used for the quantitative description of a system, frequently relying on mass balance equations. In these models, all components employed to represent and evaluate the system are described in the initial stages of model construction. Each component of the system interconnects and interacts with others in the model, based on known causal relationships; the succession of the resultant values generates the results. The majority of such models are deterministic models, which are represented as individual-based and object-oriented processes. Commonly, such models consist of a set of ordinary differential equations that model the dynamic system. For example, Odum (1983) previously introduced

and exemplified many types of deterministic models to represent virtual ecosystems. In freshwater systems, a plethora of water quality models have already been designed and developed. Håkanson and Boulion (2003) presented a general dynamic model to predict phytoplankton biomass and production, and Arhonditsis and Brett (2005) developed a more complex model that incorporated phyto- and zooplankton in Lake Washington. For assessments of streams and rivers, QUAL2E is one of the most popular water quality models (Brown and Barnwell 1987). However, this technique has had some difficulties in cases in which the errors between predicted and observed values have been too large for direct application to target river systems. Hence, Park and Lee (2002) added some tuning parameters, such as autochthonous sources, in order to improve their model predictions. Nonetheless, this technique is still limited in terms of its ability of predict specific values (e.g. Biochemical Oxygen Demand and chlorophyll *a*) relevant to water quality, particularly in regulated river systems (Choi et al. 2008). In addition to these QUAL-based models, POTAMON is a unidimensional, non-stationary model that was designed to simulate potamoplankton. This is a more biologically friendly technique than QUAL2E, but does not reduce the errors inherent to the prediction of real observed values (Everbecq et al. 2001).

Empirical modeling

The rapid advance of computer science has ushered in a host of new technologies relevant to a broad range of sciences since the 1990s. Newer technologies and paradigms of ecosystem modeling have been proposed, aiming to reduce the uncertainty in models arising from qualitative and quantitative imperfections in the ecological data (Lek 2007). With the advent of computer-based modeling, data-collecting systems have also been developed and larger quantities of data have become available. This phenomenon has grown to encompass and delineate a wholly novel research field, referred to as ecological informatics (Recknagel 2006).

Computational algorithms take advantage of quick iterative calculations conducted with large volumes of data. Generally, empirical computational ecosystem models are designed to derive the best-fitting representation for an ecological dataset via a training and validation process (Fielding 1999). As many empirical computational models are constructed via data learning, they also fall under the rubrics of 'machine learning', 'inductive model' or 'data-driven model' (Recknagel 2006). Some representative examples include ANN, EC, decision tree models, fuzzy logic, etc. (Silvert 1997, Whigham and Recknagel 2001a, Goethals et al. 2003, Shan et al. 2006).

Among these, ANN and EC may be classified as biologically inspired methods, and ecological scientists have begun to take increasing interest in applying them to ecosystem modeling. Recknagel (2001) previously demonstrated some useful empirical models for ecological time-series modeling, emphasizing the limitation in the complexity of deductive ecological models with their rigid structures. Jeong et al. (2003) described an empirical predictive model in a comparison between statistical linear models and evolutionary computation. Kim et al. (2007a) also interpreted ecological significance on the basis of an empirical predictive model.

EVOLUTIONARY COMPUTATIONS AND RELATED RESEARCH

Genetic algorithms (GA) are a mechanism originally inspired by natural evolution (Holland 1975), which operate on strings of bits that are analogous to chromosomes. One unique attribute of the GA is that it adopts the evolutionary mechanisms of heritable variation and selection. Crossover and mutation processes in the GA cause variations in the population (chromosomes) over time. The individuals with poor fitness are excluded in the selection of the next generation's parents. A near-optimal solution eventually results from the iterated application of these mechanisms.

Genetic programming (GP) is an extension of the GA concept, in which the individuals exhibit a more complex (labeled tree) structure, thereby allowing them to reflect more complex target solutions (Koza 1992), comparable with ANN. This may ease the process of creating new offspring populations from the two parents. New populations are generated by removing a branch from one tree and inserting it into another, or replacing it with a whole new branch, by analogy with genetic operators such as crossover and mutation (Fig. 1). The overall procedure of the GP is described in Fig. 2. Population size, $P(t)$, refers to the initial number of candidate tree models at time t. Better individuals in the population are selected via reproduction and genetic operations. A single cycle of this process is referred to as a generation, with the computation eventually halting when a predetermined maximum number of generations is reached.

At the termination of the computation, GP supplies labeled tree structures that can, in principle, be under-

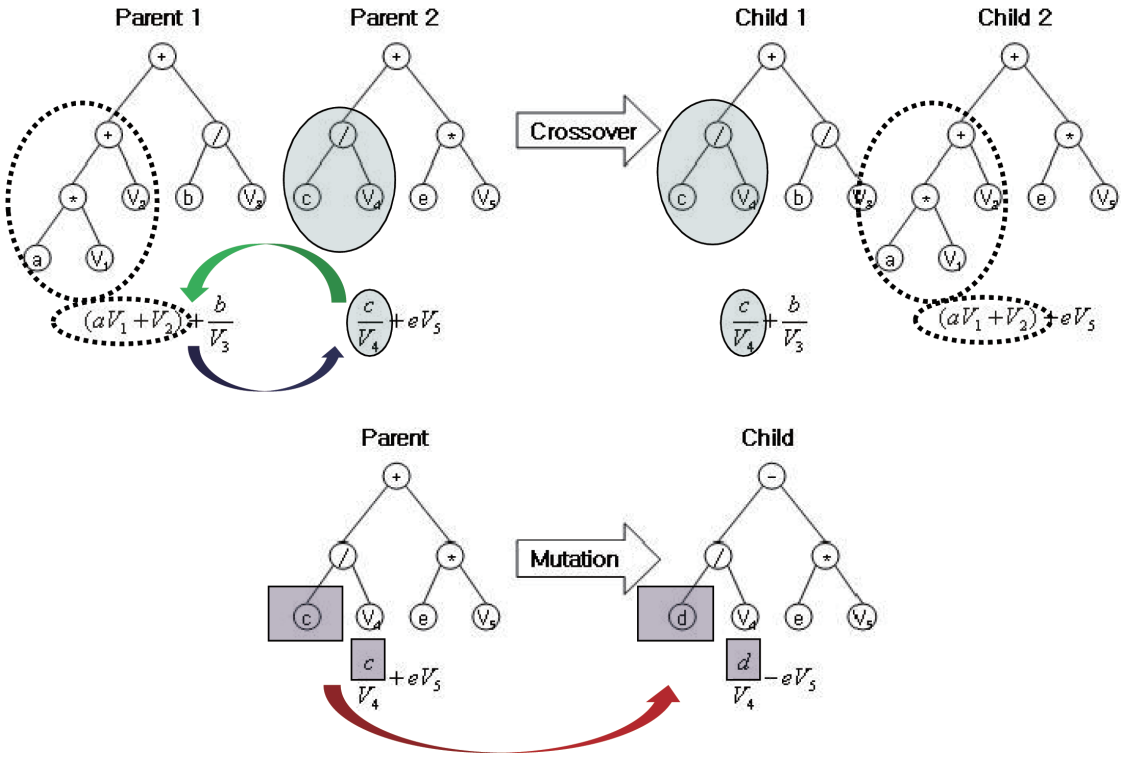

Fig. 1. Basic structure and evolutionary principle of genetic programming (letters from *a* to *e* imply constant parameters, and V_i means variable parameter for inputs).

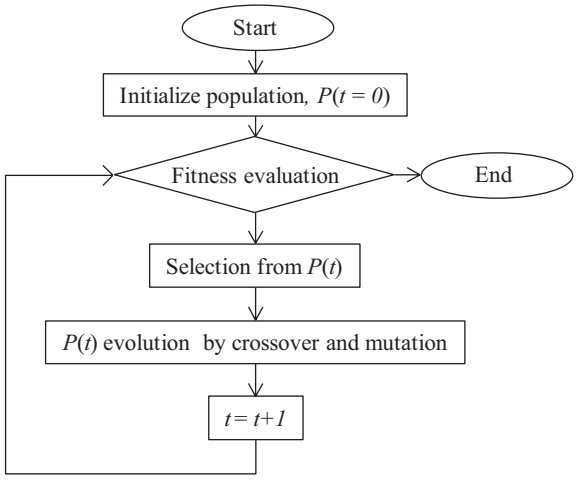

Fig. 2. Computational procedure of genetic programming.

stood by the user. This is an advantage of GP in terms of the readability of the model, whereas ANNs are a black-box model (their meaning is not readily comprehensible to humans). Nonetheless, ANNs have been utilized more extensively in ecological research (Lek et al. 1996, Chon et al. 2001, Park et al. 2006b, Goethals et al. 2007), and relatively few ecologists have presented the results of predictive modeling using GP (Savic et al. 1999, Whigham and Recknagel 2001a).

Table 1 presents some environmental and ecological research related to the applications of EC. Internationally, EC has been fairly broadly employed in environmental research. In particular, GA has been generally perceived as a favorable tool for parameter optimization in the engineering field, and has consequently come into common use for the constant fitting of complex structured models such as QUAL2K (Pelletier et al. 2006, Cho and Lee 2009). This methodology has been recently adopted for model optimization in Korea (Cho et al. 2004) and utilized for operational purposes in management policy (Lee and Chung 2004, Park et al. 2006a). Nonetheless, the applications of this technique in biological research are far fewer than those possible at an international level. Moreover, it appears that GA is more familiar to domestic researchers than is GP. GP has been used only rarely in the environmental engineering field, although its solutions are more transparent and extensible than GA. In rainfall-runoff modeling, GA-optimized tank structured (Paik et al. 2005) and GP-based self-automated models (Khu et al. 2001, Rabuñal et al. 2007) have been used in both domestic and international research.

Table 1. Comprehensive environmental and ecological research in relation to evolutionary computations

		Issues and topics	Applications of evolutionary computations/algorithms
Environmental issue	Water quality	Time-series forecasting of water quality Parameter calibration	Cao et al. (2006) Cho and Lee (2009) Cho and Sung (2004) Kim et al. (2007b) McKay et al. (2006) Pelletier et al. (2006) Whigham and Recknagel (2001b)
	Flow/Runoff	Real time runoff forecasting Modeling the rain effect on flow rate Parameter optimization of a given model	Khu et al. (2001) Dorado et al. (2002) Paik et al. (2005) Rabuñal et al. (2007) Savic et al. (1999)
	Policy	Finding an optimal operating policy in reservoirs Optimization of water quality monitoring networks Optimizing water consumption and wastewater networks / Evaluation of riparian zones with satellite sense images	Ahmed and Sarma (2005) Icaga (2005) Lavric et al. (2005) Lee and Chung (2004) Makkeasorn et al. (2009) Park et al. (2006a)
Ecological issue	Plant	Predicting patters of non-native plant invasions	Underwood et al. (2004)
	Plankton	Algal bloom forecasting Prediction of cyanobacterial dynamics Ecological explanation based on EC modeling	Bobbin and Recknagel (2001) Cao et al. (2008) Jeong et al. (2003) Kim et al. (2007a) Recknagel et al. (2002) Recknagel et al. (2008) Welk et al. (2008)
	Fish	Predicting distributions of fish species	McNyset (2005)
	Invertebrate	Prediction of spatial distributions	Stockman et al. (2006)
	Bird	Prediction of probability of presence	Peterson et al. (2002)
	Others	Assessment of machine learning with poorly predictable ecological data Introduction of inductive modeling methods	Shan et al. (2006) Whigham (2000)

COMPARATIVE ADVANTAGES OF DIFFERENT MODELING APPROACHES

In assessing specific phenomena and ecological events, we must first gain insight into the properties of the different potential modeling methods. In this section, we compare the characteristics of each modeling method, delineating the advantages and disadvantages of the methods.

Statistical models and analyses are the most commonly used tools in many scientific disciplines. They are predicated on simple statistical relationships (generally correlations) between important parameters – most often linear, commonly also polynomial or logarithmic, but always in a pre-defined simple form. MLR models have been broadly employed for the prediction of responses to independent effects. However, ecological datasets frequently contain many variables, particularly relative to the total number of instances; however, too many variables can conceal causal relationships, confusing attempts to extract them via automated methods. Thus, it has been known for some time that the limitation of classical statistical models to the extraction of linear relationships meant that these models might miss important nonlinear relationships in ecosystems (Lek et al. 1996, Jeong et al. 2003).

Mathematical mechanistic models are used to construct a representation of the ecosystem on the basis of known physical principles, most commonly the mass balance between various components within the ecosystem boundaries. In mechanistic models, it is important to model all relevant components within the system (otherwise, the assumption of mass balance may be invalid). Such models have been particularly favored for decision-making by managers and administrators in the field of water resource operations, owing primarily to the completeness of the models; this means that very flexible operation, extrapolating beyond the range of previous data, might prove possible. However, they commonly evidence very complicated architectures. As with statistical

Table 2. Comparison between conventional models and evolutionary computation

Attribute	Conventional models	Evolutionary computation
Flexibility of model structure	Predetermined based on prior knowledge or statistically linear relationships	Flexible structures learnt from given datasets
Time-series analysis	Past data use Using one point of past data, by means of moving average or some time-series statistics Future data prediction A series of future prediction (e.g. *n*-consecutive days prediction; *n* is the number of days interested in future) is possible based on the process built into the model	Past data use Possible using time-lagged metadata rearrangement Future data prediction One point prediction is possible (one-week- or month-ahead)
Capacity to handle unanticipated structure in the data	The model will only work in representing the real world when the process relationships in the model fit exactly to the actual processes	Genetic programming is often able to find a good model by efficiently searching a vast space of possible models
Amount of data required	The process is already represented in the model, only validation data is needed	Requires much larger volumes of data in order to obtain solutions that generalize to new data
Ease of future model application	Consecutive changes of an interested factor can be obtained only with small number of data	Simple and easy-to-understand structure of final outcome allowed users to apply the model to the world easily
Ecological/Environmental explanation	The phenomena that the process does not represent cannot be explained	Various sensitivity analyses can discover previously unknown relationships between input and output parameters
Research time required for model development	Longer time to determine the model structures and calibrations	Relatively short time due to self-automation and adaptation

models, prediction uncertainty is apt to be large, owing to a lack of knowledge regarding the non-mass-balance components of the ecosystem. Generally, the prediction accuracy is not sufficiently high for practical applications. Thus, determining how to incorporate the benefits of mechanistic models, while dealing with the uncertainty and nonlinearity of ecological data, is one of the most important issues in the field of ecological modeling.

By way of contrast with the above methods, empirical computational models can be employed in constructing a representation of an ecosystem on the basis of the observed data. Their primary objective is usually to find the optimal model structure for the target ecosystem ('best' is usually taken to mean 'lowest predictive error') based on computations and reasoning from large quantities of data. The higher level of automation makes it feasible for end users to select and apply the most appropriate methods. In this regard, machine learning (ML) techniques are employed in order to extract information regarding the relevant interactions and relationships between environmental entities, through the optimization of a model to fit the target ecosystem. A major premise in this regard is that data is inherently noisy, and thus this noise may mask weaker relationships within the data, thus making the development of a perfect and complete ecosystem model impossible; these methods are premised on finding the best model justified by the specific data available. These methods are also thought to be particularly useful when the important relationships within the target ecosystem are not fully known, or are too complicated to represent in a model, or when the quantity and quality of the data are insufficient for the construction of a complete representation of the system (Table 2).

CASE STUDY: WATER QUALITY PREDICTION IN THE LOWER NAKDONG RIVER

Site description and methods

The study site (Mulgeum) was located within the lower part of the Nakdong River, the longest (ca. 525 km) river in South Korea (Fig. 3). The trophic state of the river is a persistent eutrophic level (chlorophyll *a*: 40 μg/L) throughout the year, except during the summer heavy rainfall season. Algal proliferations comprise two severe problems: 1) summer cyanobacterial blooms and 2) winter diatom blooms (Ha et al. 1999, Ha et al. 2003). Large populations of people also reside in this area, and thus demand for water resources availability is relatively high.

A total of 17 input variables were used to generate a one-week-ahead predictive GP model to forecast algal

Fig. 3. Site location in the Nakdong River.

abundance. Hydrological and meteorological data (flow rate, 4 dam discharges and rainfall) were acquired from the Korean Water Management Information System, and other data (water temperature, dissolved oxygen, pH, Secchi disc depth, conductivity, alkalinity, turbidity, nitrate, phosphate, silica and nitrogen:phosphorus ratio) were collected and measured via field sampling (Table 3). The concentration of chlorophyll *a* was employed as a proxy for algal abundance as the output measure. Data from 1994 to 2008 were used for model construction (N = 782).

In this study, we employed a GP program in the C++ language, which was originally designed by Cao et al. (2006). One key issue in this type of time-series prediction is how to allocate data to training and the testing of the model (the two need to be kept separate for fair validation). We employed 702 data instances for training, with the remaining 80 reserved for testing; the partition of the data was conducted using the bootstrapping method (Adams et al. 1997) per trial (200 runs) to avoid tedious k-fold cross-validation. The initial population size was fixed at 5,000, and the maximum tree depth (length of the model structure, i.e., limit on model complexity) was 5. The GP system was allowed to construct solutions however it liked using the standard arithmetic operators (+, -, *, \) along with the exponential and logarithmic functions, arithmetic relations (>, =, <) and the Boolean if then else construct. Each GP run continued for a total of 100 generations, for a total of 200 runs overall. The root mean squared error (RMSE) was used as the fitness function in this experiment.

Table 3. Data used in evolutionary computation modeling (N = 782)

Variable (input/output)	Acronym	Unit	Mean	Standard error
Flow rate	FL	m^3/s	692	28.8
Andong dam discharge	AD	m^3/s	34.4	1.0
Imha dam discharge	IH	m^3/s	23.2	1.3
Namgang dam discharge	NG	m^3/s	41.5	1.6
Hapcheon dam discharge	HC	m^3/s	22.1	0.7
Rainfall	Ra	mm/d	3.5	0.2
Water temperature	WT	°C	16.3	0.3
Dissolved oxygen	DO	mg/L	106.6	0.9
pH	pH		8.25	0.02
Secchi depth	Se	cm	80.0	1.0
Conductivity	Co	μs/cm	299.7	4.0
Alkalinity	Al	mg/L	53.1	0.6
Turbidity	Tu	NTU	15.4	1.4
Nitrate	NO	mg/L	2.58	0.031
Silica	Si	mg/L	5.60	0.134
Phosphate	PO	mg/L	0.056	0.001
Nitrogen:Phosphorus ratio	NP		114.9	8.5
Chlorophyll *a*	chl.a	μg/L	38.3	1.5

RESULTS AND DISCUSSION

The best predictive model was generated via selection by both RMSE and the determination coefficient (r^2). The optimal model contained eight input variables, and was as follows:

If WT ≥ 33.4

Then $chl.a = DO + FL$ (1)

Else $$chl.a = \frac{NO + pH - 3Se + 407.4}{WT + 2\ln|Si|} + \frac{40.6 \cdot DO}{Se + AD + \frac{FL}{pH}}$$

Where, WT: water temperature
chl.a: chlorophyll *a*
DO: dissolved oxygen
FL: flow rate
AD: Andong dam discharge
Se: Secchi disc depth

Among our conditional criteria, water temperature (WT) was selected, as the pattern of chlorophyll *a* concentration is affected profoundly by temperature. At high temperatures, rule-based expression was rather simple, whereas more complicated expression patterns were produced by GP for normal and lower temperature ranges.

The overall prediction error was 31.32 (RMSE) with r^2 = 0.45. Note that random data partitioning between the training and test was used. Fig. 4 shows the comparison between the observed and predicted values for chlorophyll *a* concentration. Although the predicted peak values were generally slightly underestimated, the model does accurately depict the dynamic pattern of chlorophyll *a* and also accurately matches the timing. If we regard 40 μg/L and a high eutrophic level and as the critical indicators for water quality deterioration at the study site, the predictive model performs with an accuracy of 82.5% (212 of 257 cases) when employed as an early warning system for the management of the river ecosystems. Additionally, the stability of the model predictions should be taken into consideration when assessing the application of the models. Error ranges for the predictive models generated by GP were 32.2 ± 1.5 (mean ± standard deviation) for training and 37.1 ± 12.9 for test (N = 200).

From the GP predictive models, we can observe that input variable selection provides important information. The frequency with which GP selects a variable in model construction is a good proxy for the degree of influence it exerts (Kim et al. 2007a). The selection frequencies for the different variables are quite diverse; their distribution is presented in Fig. 5. WT was most frequently selected, whereas pH and Secchi depth were also included in more than 50% of the models. In fresh water, these factors are highly influential on chlorophyll *a* concentrations, because algal growth is regulated markedly by temperature and light intensity (Jeong et al. 2001). Additionally, the more frequent selection of silica than nitrogen and phosphorus observed in this study is consistent with ecological knowledge regarding the reoccurrence of winter diatom blooms in the lower Nakdong River (Ha and Joo 2000). Besides, in reference to equation 1, lower silica

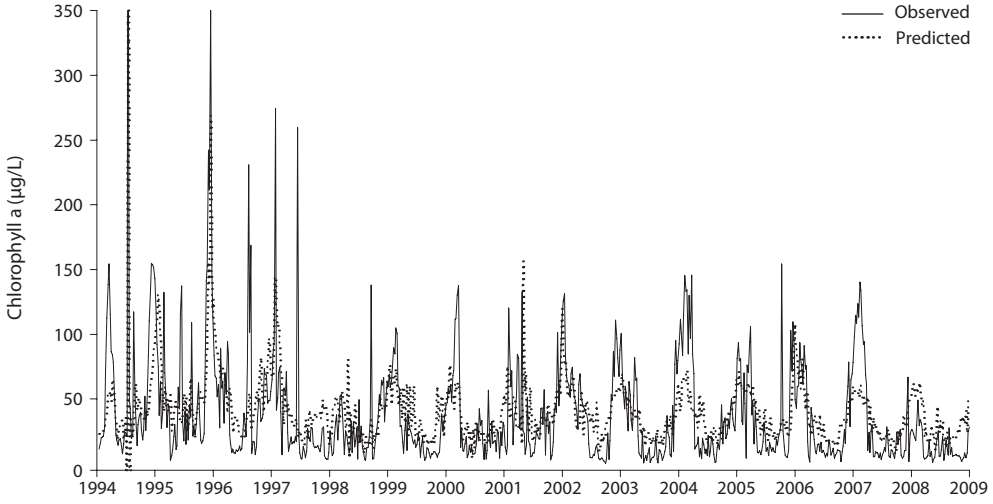

Fig. 4. Comparative result between observed and predicted data for algal dynamics.

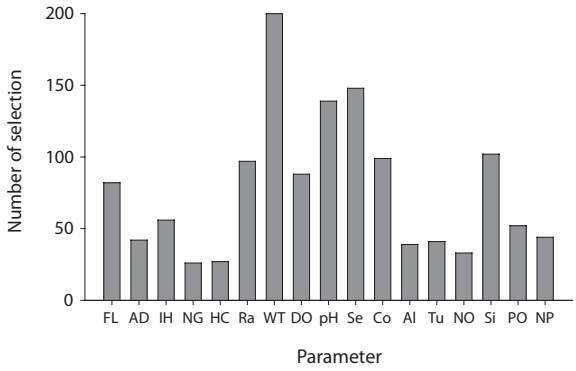

Fig. 5. Selectivity of input variables in the genetic programming predictive models. FL, flow rate; AD, Andong dam discharge; IH, Imha dam discharge; NG, Namgang dam discharge; HC, Hapcheon dam discharge; Ra, rainfall; WT, water temperature; DO, dissolved oxygen; Se, Secchi depth; Co, conductivity; Al, alkalinity; Tu, turbidity; NO, nitrate; Si, silica; PO, phosphate; NP, nitrogen:phosphorus ratio.

concentrations result in increased algal biomass. Kilham et al. (1986) previously stressed that *Stephanodiscus* species – a predominant diatom in the Nakdong River – required a high supply rate of phosphorus, but could grow successfully under low silica and light conditions, although diatom species employ silica to build their shells (frustule). However, it is also reasonable to assume that high silica consumption induces increasing algal concentrations, particularly winter diatom species, as there is a time lag for algal growth via nutrient absorption (Kim et al. 2007a). Although we can understand and explain this effect, the importance of silica was somewhat counter to our expectations. This highlights one important advantage of GP: it can be used to extract unexpected information via learning in data-driven modeling.

With regard to predictability, the most significant issue is how to acquire larger quantities of higher quality data. The data quality issue is related directly to how we can obtain data from stable analytical methodologies (i.e., high consistency in monitoring and measuring). In addition to the qualitative issue, empirical models such as EC require large quantities of data for data learning/training – perhaps larger quantities than are required for other methods. A great deal of time may be required to gain sufficient data using traditional methods, but we anticipate that the rapid development of ecological monitoring and analysis systems will help to remedy this problem before too long. Data cleaning is a favorable option not only for the extraction of potentially useful information, but also for the removal of outliers and noise from data. Consequently, it should prove possible to reduce predictive errors through the appropriate data cleaning techniques.

APPLICABILITY OF INTEGRATED MODELS IN FUTURE ECOSYSTEMS RESEARCH

In ecological research, data accumulation is accelerating precipitously, as the measuring equipment used for ecosystems is under rapid and continuous development. A broad variety of tools and techniques for the analysis and assessment of ecological properties are continuously being created and deployed. Although we introduced a variety of analytical methodologies and categorized them, we are currently unable to pre-determine a specific framework of modeling approaches for a particular range of ecosystems. Each modeling approach has some useful properties for the analysis of a target ecosystem, which may prove valuable in the interpretation and understanding of that ecosystem. For instance, in a comparison between linear (PCA and correspondence analysis) and nonlinear methods (self-organizing map, SOM), it may prove desirable to employ nonlinear methods in ecological patterns to prevent horseshoe (PCA) and arch effects (CA), but alternatives such as SOM do not allow for the control of gradient directions (Giraudel and Lek 2001). Consequently, a combination or fusion of analytical techniques is desirable, particularly in the patterning and clustering of the structures of ecosystem populations and communities.

The applicability of different modeling methodologies is a matter under continual discussion, regardless of whether deductive or inductive approaches are employed. Previously, conventional modeling techniques involved a variety of standardized mathematical and stochastic methods, such as differential equations, multivariate statistics, and regression models, whereas recent modeling approaches have been biased toward heavily computational models based on data warehousing and biologically inspired algorithms (Dolk 2000, Recknagel 2006). Additionally, a few ecological scientists have reported some promising results via hybrid approaches. Hybrid evolutionary algorithms, in which rule sets and algebraic equations define the model architecture but the content is selected via evolution, have been employed in the prediction of chlorophyll *a* concentrations in rivers and lakes (Kim et al. 2007a, Welk et al. 2008). Atanasova et al. (2006) reported good simulation results for chlorophyll *a* in Lake Kasumigaura using an assembly of ODEs. Additionally, some of the generic lake models (SALMO and Lake Washington model) have been upgraded and updated via GP techniques (Cetin et al. 2005, Cao et al. 2008).

In studies of South Korean freshwater ecosystems, eco-

logical scientists have undertaken only a limited amount of modeling via comparison with the data of hydrological engineers. Thus far, the majority of such research has been biased toward specific analysis methods, particularly statistically based approaches (Yoo 2002, An et al. 2006, Kim et al. 2007c). Mechanistic models have been employed in a few applications, and these have focused principally on pollutant transportation (Shim et al. 1995, Park and Lee 2002). However, these models regarded the physicochemical impacts as more important than the biological influences. However, in the lakes and regulated rivers of South Korea, grazing activity by zooplankton is a critical component in determining water quality during the dry winter period (Kim et al. 2000). Although the modified QUAL2E (QUAL-NIER) incorporated 31 variables in the model, zooplankton activity is not one of them (Choi et al. 2008). Comparatively, in regard to the use of empirical modeling approaches, only a few ML techniques have been applied thus far to the prediction of population and community dynamics in stream and river ecosystems (Chon et al. 2000, Jeong et al. 2006); the numbers of such studies are relatively small compared to other countries.

This imbalance in model application may limit future scientific research. Thus, interdisciplinary collaborations may prove an effective solution for understanding and improving ecological modeling. In turn, the development of better ecological models is expected to allow for the development of effective and efficient strategies for water resource management.

ACKNOWLEDGMENTS

This work was supported by a research project (355-2008-1-C00047) of the Korea Research Foundation during Dr. Kim, Dong-Kyun's postdoctoral fellowship. This study was also partially supported by the LTER Programme of the Ministry of Environment in the Nakdong River. The authors are grateful to Dr. Cao, Hongqing for sharing the GP programming source code and to Mr. Oh, Insoo for revising and updating the code. The Seoul National University Institute for Computer Science and Technology provided facilities for the research.

LITERATURE CITED

Adams DC, Gurevitch J, Rosenberg MS. 1997. Resampling tests for meta-analysis of ecological data. Ecology 78: 1277-1283.

Ahmed JA, Sarma AK. 2005. Genetic algorithm for optimal operating policy of a multipurpose reservoir. Water Resour Manag 19: 145-161.

An KG, Park SJ, Choi SM, Park JS. 2006. Comparative analysis of long-term water quality data monitored in Andong and Imha Reservoirs. Korean J Limnol 39: 21-31.

Arhonditsis GB, Brett MT. 2005. Eutrophication model for Lake Washington (USA): part I. model description and sensitivity analysis. Ecol Model 187: 140-178.

Atanasova N, Recknagel F, Todorovski L, Dzeroski S, Kompare B. 2006. Computational assemblage of ordinary differential equations for chlorophyll-a using a lake process equation library and measured data of Lake Kasumigaura. In: Ecological Informatics: Scope, Techniques and Applications (Recknagel F, ed). Springer-Verlag, Berlin, pp 409-428.

Bobbin J, Recknagel F. 2001. Knowledge discovery for prediction and explanation of blue-green algal dynamics in lakes by evolutionary algorithms. Ecol Model 146: 253-262.

Boerema LK, Gulland JA. 1973. Stock assessment of the peruvian anchovy (*Engraulis ringens*) and management of the fishery. J Fish Res Board Can 30: 2226-2235.

Brown LC, Barnwell TO Jr. 1987. The Enhanced Stream Water Quality Models QUAL2E and QUAL2E-UNCAS: Documentation and User Manual. EPA/600/3-87/007. U.S. Environmental Protection Agency, Athens, GA.

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.

Cao H, Recknagel F, Joo GJ, Kim DK. 2006. Discovery of predictive rule sets for chlorophyll-a dynamics in the Nakdong River (Korea) by means of the hybrid evolutionary algorithm HEA. Ecol Inform 1: 43-53.

Cetin L, Zhang B, Recknagel F. 2005. Process-based simulation library SALMO-OO for lake ecosystems. International Congress on Modelling and Simulation, 2005 Dec 12-15, Melbourne, pp 318-324.

Chapra SC, Reckhow KH. 1983. Engineering Approaches for Lake Management. Vol. II: Mechanistic Modeling. Butterworth Publishers, Boston, MA.

Cho JH, Lee CH. 2009. Parameter optimization of QUAL2K using influence coefficient algorithm and genetic algorithm. J Environ Impact Assess 18: 99-109.

Cho JH, Sung KS. 2004. A study on the river water quality management model using genetic algorithm. J Korean Soc Water Wastewater 18: 453-460.

Cho JH, Sung KS, Ha SR. 2004. A river water quality manage-

ment model for optimising regional wastewater treatment using a genetic algorithm. J Environ Manag 73: 229-242.

Choi JK, Chung S, Ryoo JI. 2008. Comparative evaluation of QUAL2E and QUAL-NIER models for water quality prediction in eutrophic river. J Korean Soc Water Qual 24: 54-62.

Chon TS, Park YS, Cha EY. 2000. Patterning of community changes in bentic macroinvertebrates collected from urbanized streams for the short term prediction by temporal artificial neuronal networks. In: Artificial Neuronal Networks: Application to Ecology and Evolution (Lek S, Guegan JF, eds). Springer, Berlin.

Chon TS, Kwak IS, Park YS, Kim TH, Kim Y. 2001. Patterning and short-term predictions of benthic macroinvertebrate community dynamics by using a recurrent artificial neural network. Ecol Model 146: 181-193.

Cloern JE. 1996. Phytoplankton bloom dynamics in coastal ecosystems: a review with some general lessons from sustained investigation of San Francisco Bay, California. Rev Geophys 34: 127-168.

Deaton ML, Winebrake JJ. 2000. Dynamic Modeling of Environmental Systems. Springer-Verlag, New York, NY.

Dolk DR. 2000. Integrated model management in the data warehouse era. Eur J Oper Res 122: 199-218.

Dorado J, Rabuñal J, Puertas J, Santos A, Rivero D. 2002. Prediction and modelling of the flow of a typical urban basin through genetic programming. In: Applications of Evolutionary Computing (Cagnoni S, Gottlieb J, Hart E, Middendorf M, Raidl G, eds). Springer, Berlin, pp 190-201.

Everbecq E, Gosselain V, Viroux L, Descy JP. 2001. Potamon: a dynamic model for predicting phytoplankton composition and biomass in lowland rivers. Water Res 35: 901-912.

Fielding AH. 1999. An introduction to machine learning methods. In: Machine Learning Methods for Ecological Applications (Fielding AH, ed). Kluewer Academic Publishers, Norwell, MA.

Giraudel JL, Lek S. 2001. A comparison of self-organizing map algorithm and some conventional statistical methods for ecological community ordination. Ecol Model 146: 329-339.

Goethals P, Dedecker A, Gabriels W, De Pauw N. 2003. Development and application of predictive river ecosystem models based on classification trees and artificial neural networks. In: Ecological Informatics (Recknagel F, ed). Springer-Verlag, New York, NY, pp 91-107.

Goethals PLM, Dedecker AP, Gabriels W, Lek S, De Paw N. 2007. Applications of artificial neural networks predicting macroinvertebrates in freshwaters. Aquat Ecol 41: 491-508.

Ha K, Joo GJ. 2000. Role of silica in phytoplankton succession: an enclosure experiment in the downstream Nakdong River (Mulgum). Korean J Ecol 23: 299-307.

Ha K, Jang MH, Joo GJ. 2003. Winter *Stephanodiscus* bloom development in the Nakdong River regulated by an estuary dam and tributaries. Hydrobiologia 506: 221-227.

Ha K, Cho EA, Kim HW, Joo GJ. 1999. *Microcystis* bloom formation in the lower Nakdong River, South Korea: importance of hydrodynamics and nutrient loading. Mar Freshw Res 50: 89-94.

Håkanson L, Boulion VV. 2003. A general dynamic model to predict biomass and production of phytoplankton in lakes. Ecol Model 165: 285-301.

Harding LW, Perry ES. 1997. Long-term increase of phytoplankton biomass in Chesapeake Bay, 1950-1994. Mar Ecol Prog Ser 157: 39-52.

Holland JH. 1975. Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence. University of Michigan Press, Ann Arbor, MI.

Icaga Y. 2005. Genetic algorithm usage in water quality monitoring networks optimization in Gediz (Turkey) river basin. Environ Monit Assess 108: 261-277.

Jeong KS, Recknagel F, Joo GJ. 2006. Prediction and elucidation of population dynamics of a blue-green algae (*Microcystis aeruginosa*) and diatom (*Stephanodiscus hantzschii*) in the Nakdong River-Reservoir System (South Korea) by artificial neural networks. In: Ecological Informatics: Scope, Techniques and Applications (Recknagel F, ed). Springer, Berlin, pp 255-273.

Jeong KS, Kim DK, Whigham P, Joo GJ. 2003. Modelling *Microcystis aeruginosa* bloom dynamics in the Nakdong River by means of evolutionary computation and statistical approach. Ecol Model 161: 67-78.

Jeong KS, Joo GJ, Kim HW, Ha K, Recknagel F. 2001. Prediction and elucidation of phytoplankton dynamics in the Nakdong River (Korea) by means of a recurrent artificial neural network. Ecol Model 146: 115-129.

Jeong KS, Kim DK, Jung JM, Kim MC, Joo GJ. 2008. Nonlinear autoregressive modelling by Temporal Recurrent Neural Networks for the prediction of freshwater phytoplankton dynamics. Ecol Model 211: 292-300.

Joo GJ, Kim HW, Ha K, Kim JK. 1997. Long-term trend of the eutrophication of the lower Nakdong River. Korean J Limnol 30 Suppl: 472-480.

Jørgensen SE. 1992. Integration of Ecosystem Theories: A Pattern. Kluwer Academic Publishers, Dordrecht.

Khu ST, Liong SY, Babovic V, Madsen H, Muttil N. 2001. Ge-

netic programming and its application in real-time runoff forecasting. J Am Water Resour Assoc 37: 439-451.

Kilham P, Kilham SS, Hecky RE. 1986. Hypothesized resource relationships among African planktonic diatoms. Limnol Oceanogr 31: 1169-1181.

Kim DK, Jeong KS, Whigham PA, Joo GJ. 2007a. Winter diatom blooms in a regulated river in South Korea: explanations based on evolutionary computation. Freshw Biol 52: 2021-2041.

Kim DK, Cao H, Jeong KS, Recknagel F, Joo GJ. 2007b. Predictive function and rules for population dynamics of *Microcystis aeruginosa* in the regulated Nakdong River (South Korea), discovered by evolutionary algorithms. Ecol Model 203: 147-156.

Kim G, Kim Y, Song M, Ji K, Yu P, Kim C. 2007c. Evaluation of water quality charateristics in the Nakdong River using multivariate analysis. J Korean Soc Water Qual 23: 814-821.

Kim HW, Ha K, Joo GJ. 1998. Eutrophication of the lower Nakdong River after the construction of an estuarine dam in 1987. Int Rev Hydrobiol 83: 65-72.

Kim HW, Hwang SJ, Joo GJ. 2000. Zooplankton grazing on bacteria and phytoplankton in a regulated large river (Nakdong River, Korea). J Plankton Res 22: 1559-1577.

Kim LH, Choi E, Gil KI, Stenstrom MK. 2004. Phosphorus release rates from sediments and pollutant characteristics in Han River, Seoul, Korea. Sci Total Environ 321: 115-125.

Koza JR. 1992. Genetic Programming: On the Programming of Computers by Means of Natural Selection. The MIT Press, New York, NY.

Lavric V, Iancu P, Pleşu V. 2005. Genetic algorithm optimisation of water consumption and wastewater network topology. J Clean Prod 13: 1405-1415.

Lee KS, Chung ES. 2004. Optimal operation rules for multireservoir systems using genetic algorithm. J Korean Soc Civil Eng 24: 9-17.

Lek S. 2007. Uncertainty in ecological models. Ecol Model 207: 1-2.

Lek S, Delacoste M, Baran P, Dimopoulos I, Lauga J, Aulagnier S. 1996. Application of neural networks to modelling nonlinear relationships in ecology. Ecol Model 90: 39-52.

Lotka AJ. 1925. Elements of Physical Biology. Dover Publications, New York, NY.

Magadza CHD. 1980. The distribution of zooplankton in the Sanyati Bay, Lake Kariba: a multivariate analysis. Hydrobiologia 70: 57-67.

Makkeasorn A, Chang NB, Li J. 2009. Seasonal change detection of riparian zones with remote sensing images and genetic programming in a semi-arid watershed. J Environ Manag 90: 1069-1080.

Matta JF, Marshall HG. 1984. A multivariate analysis of phytoplankton assemblages in the western North Atlantic. J Plankton Res 6: 663-675.

McKay RIB, Hao HT, Mori N, Hoai NX, Essam D. 2006. Model-building with interpolated temporal data. Ecol Inform 1: 259-268.

McNyset KM. 2005. Use of ecological niche modelling to predict distributions of freshwater fish species in Kansas. Ecol Freshw Fish 14: 243-255.

Mishra AK, Desai VR. 2006. Drought forecasting using feed-forward recursive neural network. Ecol Model 198: 127-138.

Odum HT. 1983. Ecological and General Systems: An Introduction to Systems Ecology. University Press of Colorado, Niwot, CO.

Paik K, Kim JH, Kim HS, Lee DR. 2005. A conceptual rainfall-runoff model considering seasonal variation. Hydrol Process 19: 3837-3850.

Park SS, Lee YS. 2002. A water quality modeling study of the Nakdong River, Korea. Ecol Model 152: 65-75.

Park SY, Choi JH, Wang S, Park SS. 2006a. Design of a water quality monitoring network in a large river system using the genetic algorithm. Ecol Model 199: 289-297.

Park YS, Tison J, Lek S, Giraudel JL, Coste M, Delmas F. 2006b. Application of a self-organizing map to select representative species in multivariate analysis: a case study determining diatom distribution patterns across France. Ecol Inform 1: 247-257.

Pelletier GJ, Chapra SC, Tao H. 2006. QUAL2Kw: a framework for modeling water quality in streams and rivers using a genetic algorithm for calibration. Environ Model Software 21: 419-425.

Peterson AT, Ball LG, Cohoon KP. 2002. Predicting distributions of Mexican birds using ecological niche modelling methods. Ibis 144: E27-E32.

Rabuñal JR, Puertas J, Suárez J, Rivero D. 2007. Determination of the unit hydrograph of a typical urban basin using genetic programming and artificial neural networks. Hydrol Process 21: 476-485.

Recknagel F. 2001. Applications of machine learning to ecological modelling. Ecol Model 146: 303-310.

Recknagel F. 2006. Ecological Informatics: Scope, Techniques and Applications. Springer-Verlag, Berlin.

Recknagel F, Benndorf J. 1982. Validation of the ecological simulation model "SALMO". Int Rev Gesamten Hydrobiol 67: 113-125.

Recknagel F, Bobbin J, Whigham P, Wilson H. 2002. Comparative application of artificial neural networks and ge-

netic algorithms for multivariate time-series modelling of algal blooms in freshwater lakes. J Hydroinformatics 4: 125-133.

Recknagel F, van Ginkel C, Cao H, Cetin L, Zhang B. 2008. Generic limnological models on the touchstone: testing the lake simulation library SALMO-OO and the rule-based *Microcystis* agent for warm-monomictic hypertrophic lakes in South Africa. Ecol Model 215: 144-158.

Romo S, Van Donk E, Gylstra R, Gulati R. 1996. A multivariate analysis of phytoplankton and food web changes in a shallow biomanipulated lake. Freshw Biol 36: 683-696.

Savic DA, Walters GA, Davidson JW. 1999. A genetic programming approach to rainfall-runoff modelling. Water Resour Manag 13: 219-231.

Schaefer MB. 1968. Methods of estimating effects of fishing on fish populations. Trans Am Fish Soc 97: 231-241.

Shan Y, Paull D, McKay RI. 2006. Machine learning of poorly predictable ecological data. Ecol Model 195: 129-138.

Shim SB, Oh YS, Lee YS, Koh DK. 1995. Eutrophication forecasting of Daechong Reservoir using WASP5 water quality model. J Inst Constr Technol 14: 41-53.

Silvert W. 1997. Ecological impact classification with fuzzy sets. Ecol Model 96: 1-10.

Stockman AK, Beamer DA, Bond JE. 2006. An evaluation of a GARP model as an approach to predicting the spatial distribution of non-vagile invertebrate species. Divers Distrib 12: 81-89.

ter Braak CJF, Verdonschot PFM. 1995. Canonical correspondence analysis and related multivariate methods in aquatic ecology. Aquat Sci 57: 255-289.

Underwood EC, Klinger R, Moore PE. 2004. Predicting patterns of non-native plant invasions in Yosemite National Park, California, USA. Divers Distrib 10: 447-459.

van Tongeren OFR, van Liere L, Gulati RD, Postema G, Boesewinkel-De Bruyn PJ. 1992. Multivariate analysis of the plankton communities in the Loosdrecht lakes: relationship with the chemical and physical environment. Hydrobiologia 233: 105-117.

Volterra V. 1926. Fluctuations in the abundance of a species considered mathematically. Nature 118: 558-560.

Welk A, Recknagel F, Cao H, Chan WS, Talib A. 2008. Rule-based agents for forecasting algal population dynamics in freshwater lakes discovered by hybrid, evolutionary algorithms. Ecol Inform 3: 46-54.

Whigham PA. 2000. Induction of a marsupial density model using genetic programming and spatial relationships. Ecol Model 131: 299-317.

Whigham PA, Recknagel F. 2001a. An inductive approach to ecological time series modelling by evolutionary computation. Ecol Model 146: 275-287.

Whigham PA, Recknagel F. 2001b. Predicting chlorophyll-a in freshwater lakes by hybridising process-based models and genetic algorithms. Ecol Model 146: 243-251.

Yoo HS. 2002. Statistical analysis of factors affecting the Han River water quality. J Korean Soc Environ Engin 24: 2139-2150.

Zar JH. 1999. Biostatistical Analysis. Prentice-Hall, Upper Saddle River, NJ.

Zuur AF, Ieno EN, Walker NJ, Saveliev AA, Smith GM. 2009. Limitations of linear regression applied on ecological data. In: Mixed Effects Models and Extensions in Ecology with R (Zurr AF, Ieno EN, Walker NJ, Saveliev AA, Smith GM, eds). Springer, New York, NY, pp 11-33.