ELSEVIER

Ecological Modelling 146 (2001) 275–287

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecolmodel

An inductive approach to ecological time series modelling by evolutionary computation

P.A. Whigham [a,*], Friedrich Recknagel [b]

[a] *Department of Information Science, University of Otago, P.O. Box 56, Dunedin, New Zealand*
[b] *Department of Soil and Water, Adelaide University, 5064 Glen Osmond, Australia*

Abstract

Building time series models for ecological systems that can be physically interpreted is important both for understanding the dynamics of these natural systems and the development of decision support systems. This work describes the application of an evolutionary computation framework for the discovery of predictive equations and rules for phytoplankton abundance in freshwater lakes from time series data. The suggested framework evolves several different equations and rules, based on limnological and climate variables. The results demonstrate that non-linear processes in natural systems may be successfully modelled through the use of evolutionary computation techniques. Further, it shows that a grammar based genetic programming system may be used as a tool for exploring the driving processes underlying freshwater system dynamics.

Keywords: Time series modelling; Machine learning; Genetic programming; Chlorophyll-*a* modelling

1. Introduction

This paper describes evolutionary computation techniques for inferring predictive models based on ecological time series data. The approach is applied in order to discover equations and rules for the prediction of chlorophyll-*a* in a freshwater lake. Previous work on developing predictive models for time series ecological data has generally focussed on the use of neural networks (Recknagel et al., 1997, 1998; Liu and Yao, 1999), hybrid systems (Bobbin and Recknagel, 1999) or process-based models (Recknagel and Benndorf, 1982; Elliot et al., 2000). Although neural networks have been very successful in developing models, it is often difficult to deduce any underlying process understanding based on the model, since it is normally represented as a 'black box'. Process-based models are promising, and represent the ideal situation, however they are often difficult to calibrate, are usually based around certain units or assumptions, and are designed based on specific independent variables that may not always be available for the system being modelled. The approach presented in this paper attempts to address these issues by allowing a flexible language for describing models without any assumptions regarding the form of input, and producing models that are symbolic and therefore

* Corresponding author. Tel.: +64-3-4797391; fax: +64-3-4798311.
E-mail address: pwhigham@infoscience.otago.ac.nz (P.A. Whigham).

0304-3800/01/$ - see front matter
PII: S0304-3800(01)00313-1

open to interpretation. Similar work in equation discovery (Todorovski et al., 1998) has been previously applied to phytoplankton time-series prediction. This work used a grammar to define differential equations to describe a model, however the system (Lagramge) had to enumerate all possible equations. This restriction is overcome by the approach described in this paper by using an evolutionary algorithm to select good models from a search space defined by a largely unconstrained grammar.

1.1. Data characteristics of Lake Kasumigaura

Lake Kasumigaura is situated in the south-eastern part of Japan. It is a large, shallow water body where no thermal stratification occurs. Water temperatures vary widely, from 4°C in the winter to 30°C in summer. The lake has high external and internal nutrient loading and therefore primary productivity is high. Algal succession changes species abundance year by year, therefore making it very difficult to predict algal blooms or develop causal models of algal behaviour. Kasumigaura is dominated by harmful blue-green algal species such as *Microcystis* spp, *Oscillatoria* and *Anabaena flos aquae*. The development of models that will support understanding of the dynamics of this system may allow better predictions of future bloom behaviour and allow better management of the water as a resource. This paper describes two approaches to developing predictive models for chlorophyll-*a* concentration, based on a mathematical equation and a set of rules.

1.2. The ecology of freshwater phytoplankton

Phytoplankton include representatives of several groups of algae and cyanobacteria. They are usually distinguished by being freely floating and dependent on water movement for maintenance and transport (Reynolds, 1984). Algal species rely on light as a basic input for photosynthesis and require nutrients such as nitrogen and phosphorus for growth and reproduction. Factors such as water temperature, turbidity, mixing, competition and grazing are also relevant to the population dynamics of algae. Even though much work has been done on phytoplankton, there are still difficulties with developing reliable predictive models for algal growth due to the highly non linear behaviour of the population as a whole.

2. Evolutionary computation

Evolutionary computation techniques are a search algorithm using the concepts of evolutionary pressure to search for fit solutions to problems. The basic concepts behind evolutionary algorithms are shown in Fig. 1. All evolutionary systems require some form of population, representing possible solutions to the problem, a mechanism for selecting individuals based on their fitness, to form the next population, a method for mixing parents solutions to form new solutions, and a termination criteria.

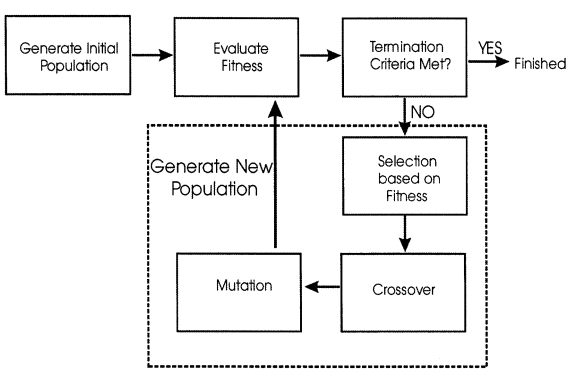

Fig. 1. A simple framework for evolutionary algorithms.

3. Genetic Programming

Genetic Programming (GP) is an evolutionary algorithm that represents population individuals as functional programs (Koza, 1992). GP extended the fixed-length approach of Genetic Algorithms (Holland, 1992) to allow basic computer programs to be evolved in the form of functional LISP expressions. This extended the GA concepts by allowing the size and shape of the evolved solutions to change and allowed the language

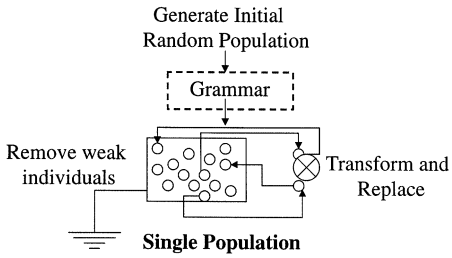

Fig. 2. A steady-state evolutionary population.

describing the problem to be varied and expressive. GP has been applied successfully to many problems (Koza, 1990; Roston and Sturges 1995; McKay et al., 1997) and has been previously shown to be useful in developing time series expressions (Whigham and Crapper, 1999). A variant of GP, using a context-free grammar, has been developed to allow the user to explicitly define a language bias when searching for solutions. This system, entitled CFG-GP (Whigham, 1995), has been applied successfully to produce time series models and habitat density models (Whigham, 2000).

This paper will apply CFG-GP to explore two structures for representing time-series data models; as a multivariate mathematical function, and as a set of rules. The use of a grammar to define the language will be shown to be a simple and useful method for expressing the form of potential solutions. It appears to be relevant to many applications in ecological modelling.

3.1. An introduction to CFG-GP

The CFG-GP system uses a population of expressions, generated from a context-free grammar, to represent possible solutions. Previous work using grammars for representing equations in an ecological context have searched the space of possible solutions in an exhaustive (breadth or depth first) manner (Todorovski and Dzeroski, 1997; Todorovski et al., 1998). Although this work was extremely successful, the approach was limited by the exponential growth in possible solutions as the depth of solutions expressed by the grammar was increased. Evolutionary algorithms are designed to efficiently handle large search spaces and therefore the combination of a grammar to represent the language and an evolutionary search shows great promise. The CFG-GP system described in this paper uses a steady-state population, as shown in Fig. 2. The basic steps involved with CFG-GP are as follows:

1. Create an initial random population based on the grammar.
2. Evaluate the fitness of the population members (based on the problem).
3. While termination criteria not met do.
   - 3.1. Select two members of the population, based on fitness, for breeding.
   - 3.2. Probabilistically apply crossover to swap sub-trees of the parents to create two children.
   - 3.3. Probabilistically apply mutation to modify a random sub-tree of each child.
   - 3.4. Evaluate the fitness of each child, and insert them into the population.
   - 3.5. Remove two weak individuals from the population based on fitness.

3.2. Introduction to context-free grammars

A context-free grammar is a production system for generating strings in a language. Productions define how non-terminal symbols may be rewritten into strings containing non-terminal and terminal symbols. This process is continued until no non-terminal symbols remain. The final terminal symbols represent the string in the language. More formally, a context-free grammar is a four-tuple $G\{S,P,N,\Sigma\}$ where S is the designated start symbol, P is a set of productions of the form $A \rightarrow \alpha$, where $A \in N$ and $\alpha \in \{N, \Sigma\}^*$, N is a set of non-terminals and Σ is a set of terminal symbols. Commencing with S the non-terminal symbols are rewritten using the productions to create strings of the language L(G). Since the productions constrain the legal combinations of strings they effectively form a language bias for the search space.

3.3. Population representation

Population members are stored as a tree structure representing the derivation steps of the grammar that was used to create the member. For

example, based on the following grammar, $G_{equation}$, the program derivation trees representing the equations $(x + y - 34.5)$ and $(\exp(x + 8))$ are shown in Fig. 3. Note that the terminal 'real[0.0:100.0]' represents a random real number between 0.0 and 100.0, which is generated when the initial population is created. This is represented as the terminal symbol $\Re$.

$G_{equation} = \{S,$
$N = \{E\},$
$P = \{$
$S \rightarrow E$
$E \rightarrow + E\ E | - E\ E | * E\ E | \backslash E\ E | \exp\ E$
$E \rightarrow x | y | z |$ real[0.0:100.0]
$\},$
$\Sigma = \{x,y,z,+,-,*,\backslash,\exp,\Re\}$
$\}$

Since the non-terminals (left-hand sides of the grammar rules) are stored as part of the program, the implicit structure of the language defined by the grammar is maintained. Crossover and mutation act on these non-terminals, and therefore any mixing of individuals maintains the language defined by the grammar. This allows complex structures to be described as part of the language, and allows the user to control what areas of the language are searched.

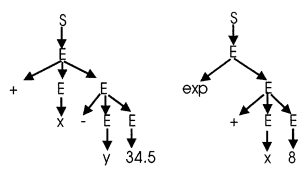

Fig. 3. Equations represented as derivation trees.

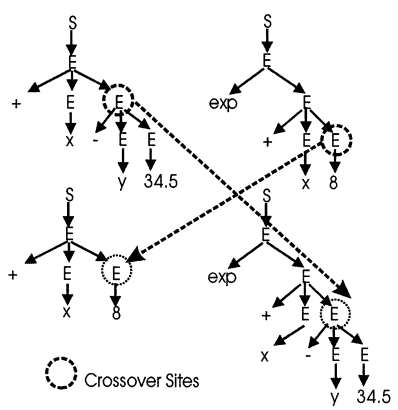

Fig. 4. Crossover based on derivation trees.

3.4. Crossover and mutation

Crossover is performed by randomly swapping components from two programs, based on a non-terminal site, as shown in Fig. 4. Here, the programs $(x + y - 34.5)$ and $(\exp(x + 8))$ are crossed to give the new programs $(\exp(x + y - 34.5))$ and $(x + 8)$. Crossover is designed to allow useful components of a fit partial solution to propagate throughout the population.

Mutation is performed by randomly deleting a sub-tree rooted in a non-terminal, and generating a new, random, sub-tree based on the grammar, in the sa–me manner as the initial creation of the population members.

Note that a set maximum tree depth limits the initial creation of the population, and subsequent population members formed from crossover and mutation. This limits the total search space, although the depth can be selected so that novel and complex programs can be created.

3.5. Hill climbing mutation for real numbers

Random real numbers are used as constants to allow the evolving programs to adjust their scale and magnitude. These numbers are generated at random at the commencement of the evolution and are not tuned in any way with the final solution, although they are able to mutate. To allow a fine tuning of an evolved program, a hill climbing mutation for the real numbers is used. This operation can be applied to the current fittest solution during the evolution for a solution at any time by user control. It is typically used to tune the constant values when the evolution is complete.

Table 1
CFG-GP population setup

Parameter	Value
Population size	1000
Maximum depth of program tree	8
Crossover probability	90%
Mutation probability	5%
Number of evaluations	50 000

Table 2
Factors measured with the daily time series data

Measured factor	Av ± S.D.	Units
Ortho-phosphate (p)	14.14 ± 25.71	Mg/l
Nitrate (n)	520.56 ± 503.4	µg/l
Secchi depth (sd)	85.43 ± 44.57	Cm
Dissolved oxygen (do)	11.2 ± 2.14	Mg/l
PH (ph)	8.74 ± 0.59	
Water temperature (t)	16.36 ± 7.79	°C
Rotifera (r)	229.2 ± 293.4	Ind/l
Cladocera (clad)	169.9 ± 221.7	ind/l
Copepoda (cop)	156.4 ± 83.7	ind/l
Chlorophyll-*a* (chla)	74.43 ± 42.51	µg/l

3.6. CFG-GP setup

Table 1 defines the setup up for all of the experiments described in this paper. The maximum depth of a program tree was limited to eight so that relatively simple and generalised expressions were created. Although larger programs may have lower errors (especially on the training data), the purpose of this paper is to demonstrate that good, simple expressions that may have a physical interpretation can be discovered. The probability of crossover and mutation indicates the likelihood of these operators being applied each time two parents are selected to produce two new children programs. All non-terminal sites were set as legal crossover and mutation sites. Specific issues related to each approach will be described in the relevant sections when the setup and results are described.

4. Training and test data setup

Table 2 shows the measured variables used for developing the models. For all experiments, 8 years of daily data ('84, '85, '87, '88, '89, '90, '91, '92) were used for training and 2 years of daily data ('86 and '93) for testing the CFG-GP system. The root mean square error (RMSE) was used as the fitness function for the training data and as a measure of accuracy for the test data. A lower RMSE was taken to indicate a better prediction of the test data. When comparing two different learning techniques a lower RMSE for the unseen (test) data implied that the learning system had better generalised the patterns found in the training data. RMSE was selected as the fitness measure since it tends to bias towards larger values in a series. Since bloom behaviour is an important aspect of chlorophyll-*a* prediction this seemed a reasonable choice for measuring performance.

5. Predicting chlorophyll-*a*

Chlorophyll-*a* is used as a sampling technique for estimating the total biomass of the phytoplankton community in a waterbody. Hence the driving factors for chlorophyll-*a* tend to represent the overall behaviour of the plankton community. The daily time series data for chloroplyll-*a* is shown in Fig. 5. Note that the validation (test) year 1986 has a far larger concentration measure than any of the training years. The other validation year (1993) is more typical of the training years.

5.1. Exploring mathematical expressions for chlorophyll-a prediction

This section will describe two mathematical approaches to developing predictive models for chlorophyll-*a*. The two approaches both allow mathematical expressions to be created, the second structure allowing past values of the independent variables to be used as input variables to the equation. The grammar, G_{chla}, defines mathematical expressions using the current values of the independent variables for prediction. The expo-

nential, natural logarithm and power functions are provided along with the standard arithmetic operators.

G_{chla} = {S,
N = {T},
P = {
S → T
T → + T T|−T T|*T T|/T T|exp T
T → ln T|^T real[0.0:1.0]
T → p|n|sd|do|ph|t|r
T → clad|cop| real[−100.0:100.0]
},

Σ = {+, −,*,/,exp,ln,^,p,n,sd,do,ph,t,r,clad,cop,ℜ}
}

The second form of equation allows past values of the independent variables to be used as part of the equation. The function 'past ⟨var⟩ ⟨num⟩' represents the value of ⟨var⟩ for the day ⟨num⟩ before the current day being evaluated. For example, 'past p 3' represents the value of *ortho*-phosphate 3 days before the current day. Based on the k-autocorrelation function for chlorophyll-*a*, as shown in Fig. 6, a 95% relationship between the current and past values occurs until the 7th day. Hence past values up to 7 days were allowed in the expressions. The grammar, $G_{chlapast}$, defines the language for this type of equation.

$G_{chlapast}$ = {S,
N = {T, V, PV},
P = {
S → T
T → + T T|−T T|*T T|/T T|exp T
T → ln T|^T real[0.0:1.0]| past V PV
T → p|n|sd|do|ph|t|r
T → clad|cop| real[−100.0:100.0]
V → p|n|sd|do|ph|t|r|clad|cop
PV → int[1:7]
},
Σ = {+, −, *,/,exp,ln,^,
p,n,sd,do,ph,t,r,clad,cop,past,ℜ}
}

5.2. Developing a rule set for prediction

An alternative structure for representing knowledge is a set of rules. By defining a different grammar the CFG-GP system can evolve a rule set to

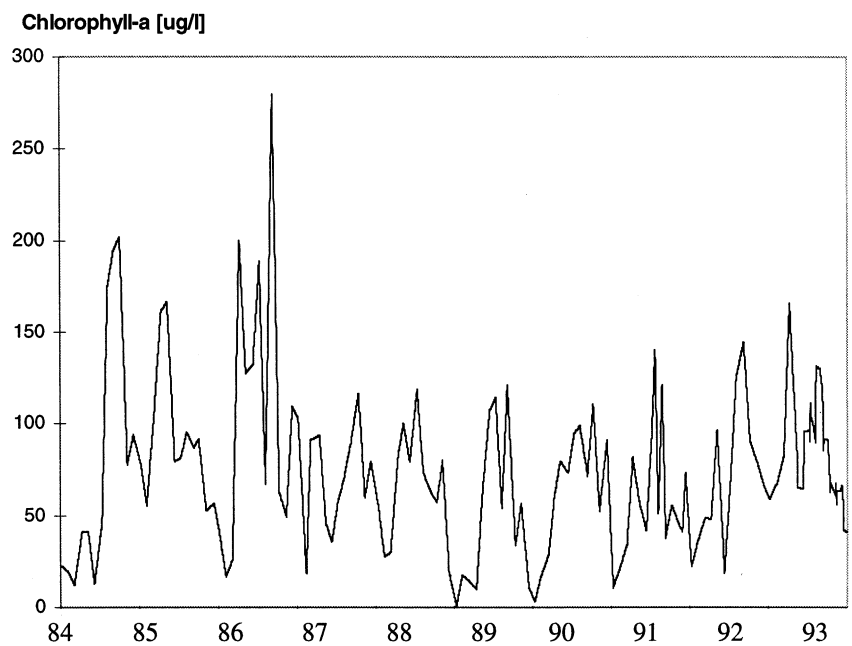

Fig. 5. Daily time series data for chlorophyll-*a* for all years.

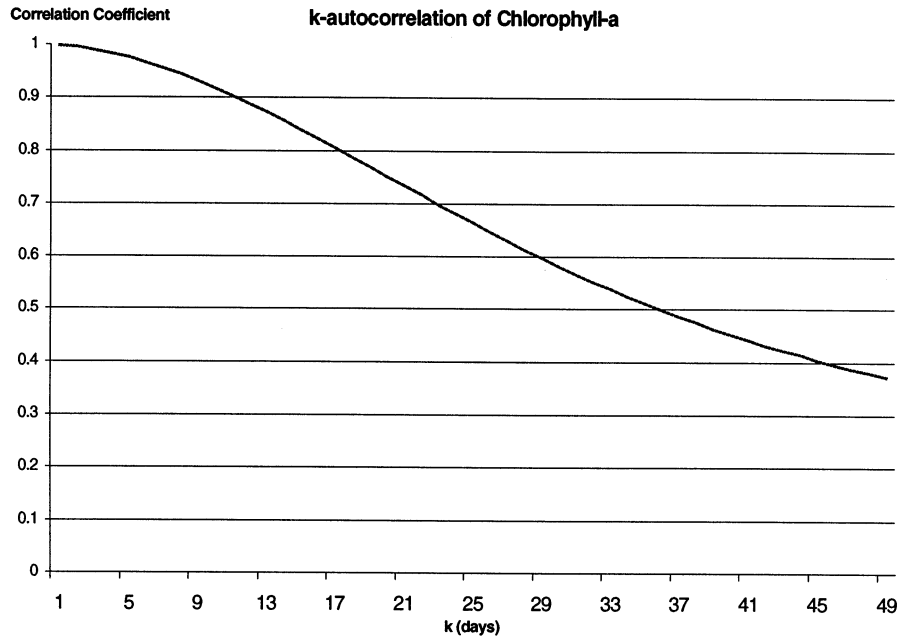

Fig. 6. The k-autocorrelation function for chlorophyll-a.

represent a model for predicting chlorophyll-a. The grammar, G_{rules}, defines a set of rules allowing a combination of boolean conditions, based on values of the independent variables, to predict chlorophyll-a concentration.

$G_{rules} = \{S,$
$N = \{R, COND, RESULT, T, V\},$
$P = \{$
$S \rightarrow R$
$R \rightarrow$ if COND R RESULT | RESULT
COND $\rightarrow$ and COND COND

COND $\rightarrow$ or COND COND | $<$ T V | $>$ T V
$T \rightarrow p|n|sd|do|ph|t|r|clad|cop$
$V \rightarrow$ real[0.0:1500.0]
RESULT $\rightarrow$ real[0.0:500.0]
$\},$

$\Sigma = \{$if,and,or,p,n,sd,do,ph,t,r,clad,cop, $<$, $>$,$\Re\}$
$\}$

The grammar G_{rules} defines a language expressing a list of if-then-else expressions, based on conditions of the independent variables, with a resulting value ranging between 0.0 and 500. The range of the RESULT value was selected since chlorophyll-a concentration is always zero or above, and 500.0 is well above all of the observed values for chlorophyll-a.

An extension to this simple rule set is possible by allowing the RESULT expression to be a mathematical equation, rather than a single value. The grammar $G_{ruleseqn}$ allows rules to be constructed that have mathematical equations as the result of a conditional statement. This language extends the possible mathematical expressions described by G_{chla} by allowing different equations to be executed based on the condition of the independent variables, thus allowing different equations to model different states of the system.

$G_{ruleseqn}$ = {S,
N = {R, COND, E, T, V},
P = {
S → R
R → if COND R E | E

COND → and COND COND | or COND
COND | < T V | > T V
T → p|n|sd|do|ph|t|r|clad|cop
V → real[0.0:1500.0]
E → + E E|−E E|*E E|/E E|
E → exp E|In E|^ E real[0.0:1.0]
E → p|n|sd|do|ph|t|real[−100.0:100.0]
},
Σ = {if,and,or, +, −, *, /,exp,In, ^,p,n,sd,do, ph,t,r,clad,cop, <, >,ℜ}
}

5.3. Results using G_{chla}

The best Eq. (1) discovered using G_{chla}, based on the training data, had a RMSE for the test years of 41.35, which is comparable with previous studies based on this data (Recknagel et al., 1998; Whigham and Recknagel 1999; Recknagel et al., 2000). The resulting prediction is shown in Fig. 7. Note that this equation is relatively simple due to the constraint on the maximum depth of any created program. Although better equations (in terms of error on both the training and test set) were created, they were too large to be easily interpreted. Eq. (1) shows that the chlorophyll-*a* concentration is related to the pH and phosphorus concentrations, and to a lesser extent the water temperature and dissolved oxygen concentration. Note also that there is a limiting effect due to nitrogen. Eq. (1) is also a good predictor of the peak concentration measured towards the end of 1986.

$$\mathrm{chla}_t = \mathrm{ph}^2 + p + \frac{t \times \mathrm{do}}{(n + \mathrm{ph} + 28.29)^{0.51}}. \quad (1)$$

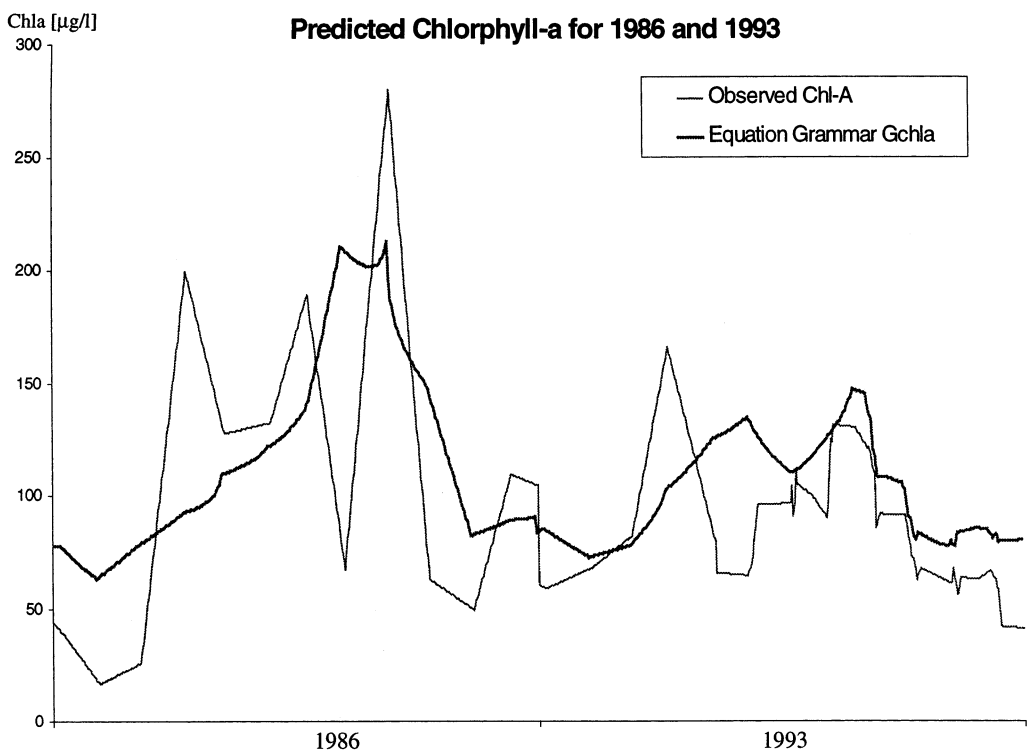

Fig. 7. Prediction derived using G_{chla}.

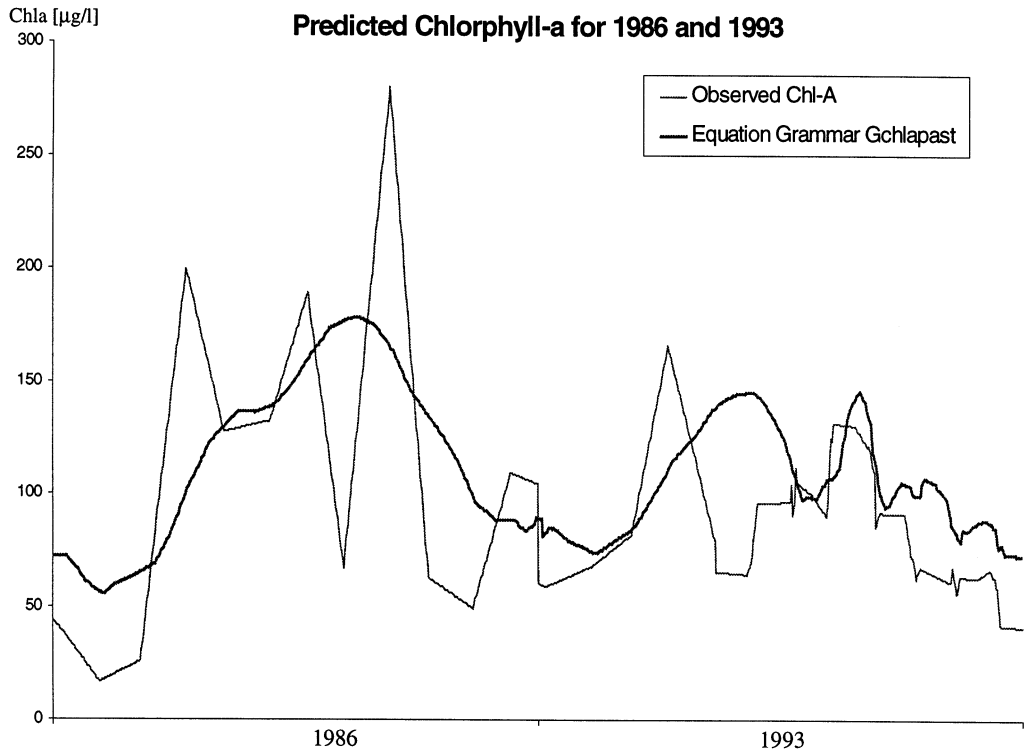

Fig. 8. Prediction using $G_{chlapast}$.

5.4. Results using $G_{chlapast}$

The best Eq. (2) discovered using $G_{chlapast}$ had an RMSE or 40.44, which is comparable with the results for Eq. (1). Although the magnitude of the peak value was not predicted, as shown in Fig. 8, the overall response of the system (i.e. the shape of the curve) was reproduced reasonably well.

$$\text{chla}_t = \text{ph}_{t-5}^2 + 2t_{t-1} - \frac{\text{cop}_{t-13}}{\text{ph}_1} \quad (2)$$

There are several comments to be made regarding Eq. (2). Past values of the variables have been used, rather then current values, for all parts of the equation other than the ph value in the reducing term. Note that although the grammar only allowed past values up to 7 days, the hill-climbing mutation has created a value back 13 days for copepoda. The equation can be interpreted as stating that the driving force for chlorophyll-*a* concentration is the ph concentration, with a lag of approximately 5 days, and the water temperature during the previous day. The concentration is limited by the grazing of copepoda, which has an influence from up to 13 days in the past, however this grazing is reduced as ph increases. It is interesting to observe that the ph^2 term is used in both Eq. (1) and Eq. (2), indicating that this is probably a fundamental measure when predicting chlorophyll-*a* in a hypereutrophic lake dominated in summer blue-green algae. This result compares very well with findings of Shapiro (1973) and Reynolds (1984) that Microcystis and Anabaena can continue to grow even above ph values of 9.5 while green-algae and diatoms become inhibited and cease growth at pH values > 8.5. As outlined by Harris (1986) the ability of blue-green algae to grow at high pH gives them an advantage over other algal groups and appears to be correlated with their ability to produce late summer blooms

in eutrophic waters. Even though it is very difficult to prove these explanations by field experiments (Harris, 1986) the emphasis of ph as key driving variable by means of the presented CFG-GP model clearly supports these findings. Hence, for Lake Kasumigaura a simple indicator for warning of an impending bloom condition could be to measure the ph concentration when ph and water temperature start to rise it is likely that within the next week bloom conditions may arise. The next step would be to analyse factors which lead to an increase in ph such as exhaustion of freely available dissolved CO_2. Although an inductive approach could be used to start forming predictive models of ph this is beyond the scope of this work.

5.5. Results using G_{rules}

Fig. 9 shows the results of applying one rule set discovered using G_{rules}, which has a RMSE of 40.02. Although this was not the best-predicted set of rules (several rules sets had slightly lower RMSE, however they did not predict the peak shape as well) the peak chlorophyll-*a* concentration in 1986 is predicted and therefore the results are of interest. The simplified version of this rule set is as follows:

IF (sd > 37.9 cm) OR (p > 222.0 mg/l) THEN
  IF (n > 177.8 μg/l) THEN chla = 77.9 μg/l
  ELSE chla = 148.3 μg/l
ELSE chla = 220.1 μg/l

Based on these rules, conditions for an algal bloom occur when the Secchi depth is less than 38 cm and the *ortho*-phosphate concentration is less than 222.0 mg/l. Since the phosphate concentration is less than 222.0 mg/l for all except 2 days out of the 10 years of measured data, it is clear that the Secchi-depth (i.e. the turbidity) is a major driver of algal production. Based on this simple set of conditions the timing of two out of the three largest peaks with chlorophyll-*a* for the 10 years of data are predicted. This set of rules also

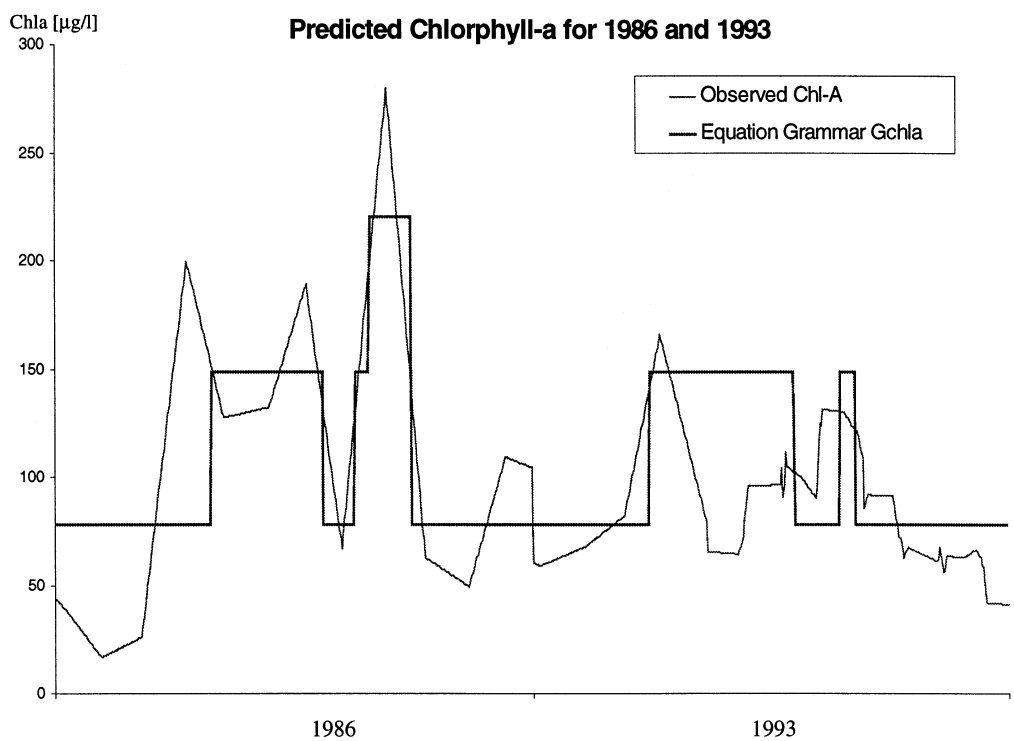

Fig. 9. Prediction using G_{rules}.

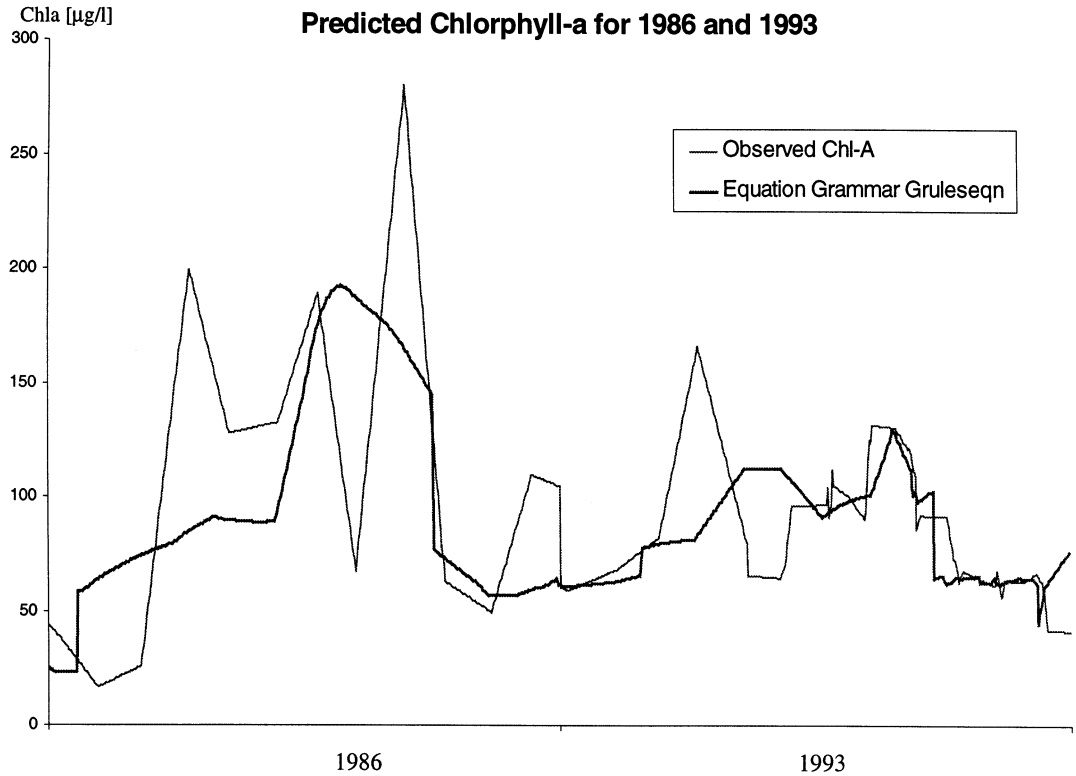

Fig. 10. Prediction using $G_{ruleseqn}$.

follows a similar pattern to previous work (Bobbin and Recknagel, 1999), where a low Secchi-depth and high nitrogen concentration were related to high concentrations of chlorophyll-*a*.

5.6. Results using $G_{ruleseqn}$

The main setup change for this example was that the maximum depth of programs was increased to 12, allowing for the fact that the grammar $G_{ruleseqn}$ was more complex then previous grammars. When the system was initially run with a maximum depth of 10 the complexity of the equations was very limited. Note, however, that as the complexity of programs increases the execution time increases. Fig. 10 shows the best discovered result after 20 runs, based on the 8 years of training data. The RMSE for the test data is 39.7, which is slightly lower then the previous results. Although the peak in 1986 is underestimated, the overall shape is reasonable. The result does not predict the early peak in 1993, nor the first peak in 1986. Although this prediction does not look as similar to the measured data as that shown in Fig. 9 or Fig. 8, the RMSE is lower. This is an indication of one problem associated with using a simple metric, such as RMSE, when dealing with time series data. Since RMSE does not take into account the shape of the data (treating each point independently) is can produce anomalies in terms of how we would measure similarity based on visual appearance. Issues related to the use of RMSE as a metric will be referred to in Section 6. The discovered rule equation is as follows:

IF $n > 640.9$ µg/l THEN
  IF sd > 109.2 cm THEN
    $chla = (do*(3p)^{0.59})$ µg/l
  ELSE
    $chla = (do*(t + t^{-2.7} + do)^{0.52})$ µg/l
ELSE
  $chla = (do*(p + do + 2t)^{0.52})$ µg/l

There are several comments to be made about this rule equation. Firstly, since the maximum depth of program was limited to 12 the rule equations were relatively simple. The use of nitrogen and Secchi depth as the main variables driving the state of the system (i.e. the selection of equations) has been independently discovered using $G_{ruleseqn}$ and G_{rules}, which would suggest that these variables are fundamental to the dynamics of Lake Kasumigaura.

6. Discussion and conclusion

The previous studies have demonstrated that models can be developed for the non-linear dynamics of phytoplankton, both as a set of rules and as mathematical equations. The flexibility of the CFG-GP approach has been shown by creating simple mathematical equations, equations that use past values for variables, a set of rules and a set of rules to select different equations. This approach should be considered as a further tool for symbolic model discovery, where the emphasis is on selecting and searching through an appropriate language that describes the problem. Although the resulting errors for the predicted model were inferior to those produced by a neural network on the same data (Liu and Yao, 1999), the study of Liu and Yao divided the data into several ensembles, each representing a different state of the system, before training commenced. This work also used a normalised version of the data, which has been shown previously to often improve predictions, especially for extreme values (Whigham and Recknagel, 1999). It is therefore difficult to determine whether the approach described in this paper would be able to produce a model that was comparable in accuracy to the described neural network.

The fitness function for all of these time series problems used the standard RMSE, as a measure of the similarity between the predicted and actual time series data. Previous work in the field of data mining has indicated that RMSE is not always an appropriate measurement for similarity between time series (Keogh and Pazzini, 1999). Further work is required to determine whether more appropriate measures, such as shape indicators like fourier transforms and differencing, would improve the evolved equations. A second approach would be to compare the evolved solutions when different metrics are used for similarity. A simple hypothesis would be to test whether different metrics discover different dominant factors driving the system, with the likely outcome being that certain basic combinations of terms would consistently appear.

Phytoplankton dynamics are complex and often have different phases based on climate and other population histories. This study has considered one mechanism for detecting these different phases, based on using simple rules and rules combined with equations. Since these approaches allow different phases of growth and decay to be recognised and use different equations for each phase it was expected that the models would be significantly better then single equation representations. However, based on the results shown here it is inconclusive as to whether a more complex, state-based approach will have any improvement over a single equation.

The current study attempted to exploit the time series nature of the data by allowing past values as input to the evolved equations. Equations for population growth, which include the current population as part of the next time step prediction, have been previously explored using this approach (Whigham and Reckangel, 2000). Although the results were positive, the models were too complex to be easily interpreted. Further work is required to determine whether process-based models, such as difference equations, combined with rule structure, may allow further knowledge of the phytoplankton dynamics for Lake Kasumigaura to be established.

Acknowledgements

The authors would like to thank the anonymous referees for their contribution towards improvements in this paper, and Takehiko Fukushima, Noriko Takamura and Takayuki Hanazato of the National Institute of Environmental Studies in Tsukuba, Japan, for providing data of the Lake Kasumigaura.

References

Bobbin, J., Recknagel, F., 1999. Min Ing water quality time series for predictive rules of algal blooms by genetic algorithm. MODSIM'99 International Congress on Modelling and Simulation, In: Oxley, L., Scrimgeour, F., (Eds.), Hamilton, New Zealand, Modelling and Simulation Society of Australia and New Zealand Inc.

Elliot, J.A., Irish, A.E., Reynolds, C.S., Tett, P., 2000. Modelling freshwater phytoplankton communities: an exercise in validation. Ecological Modelling 128, 19–26.

Harris, G.P., 1986. Plankton Ecology – Structure Function and Fluctuation. Chapman and Hall, NY, p. 384.

Holland, J., 1992. Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence. Second Edition, MIT Press/Bradford Books, Cambridge, MA

Keogh, E.J., Pazzini, M.J., 1999. Relevance Feedback Retrieval of Time Series Data. The 22nd Annual international ACM-SIGIR conference on research and development of information retrieval.

Koza, J.R., 1990. Concept formation and decision tree induction using the genetic programming paradigm. In: Schwefel, H.P., Manner, R. (Eds.), Parallel Problem Solving from Nature. Springer-Verlag, pp. 124–129.

Koza, J.R., 1992. Genetic Programming: On the Programming of Computers by means of Natural Selection. MIT Press, Cambridge, MA.

Liu, Y., Yao, X., 1999. Time Series Prediction by Using Negatively Correlated Neural Networks. Lecture Notes in Artificial Intelligence, vol. 1585. Springer-Verlag, Berlin, pp. 325–332.

McKay, R.I., Pearson, R.A., Whigham, P.A., 1997. Learning spatial relationships: some approaches. In: Pascoe, R.T. (Ed.), GeoComputation'97. University of Otago, Dunedin, New Zealand.

Recknagel, F., Benndorf, J., 1982. Validation of the ecological simulation model SALMO. Int. Revue ges. Hydrobiol. 67 (1), 113–125.

Recknagel, F., Bobbin, J., Whigham, P.A., Wilson, H., 2000. Multivariate time series modelling of algal blooms in freshwater lakes by machine learning. Proceedings of the 5th International Symposium WATERMATEX on Systems Analysis and Computing in Water Quality Management, Gent, Belgium.

Recknagel, F., French, M., Harkonen, P., Yabunaka, K., 1997. Aritificial neural network approach for modelling and prediction of algal blooms. Ecological Modelling 96 (1-3), 11–28.

Recknagel, F., Fukushima, T., Hanazato, T., Takamura, N., Wilson, H., 1998. Modelling and prediction of Phyto- and Zooplankton dynamics in Lake Kasumigaura by artificial neural networks. Lakes and Reservoirs: Research and Management 3, 123–133.

Reynolds, C.S., 1984. The Ecology of Freshwater Phytoplankton. Press Syndicate of the University of Cambridge, New York.

Roston, G., Sturges, R., 1995. A genetic design methodology for stucture configuration. ASME Advances in Design Automation DE82, 73–90.

Shapiro, J., 1973. Blue-green algae: why they become dominant. Science 179, 382–384.

Todorovski, L., Dzeroski, S., 1997. Declarative bias in equation discovery, Proceedings Fourteenth International Conference on Machine Learning, Morgan Kaufmann, San Mateo, CA.

Todorovski, L., Dzeroski, S., Kompare, B., 1998. Modelling and prediction of phytoplankton growth with equation discovery. Ecological Modelling 113, 71–81.

Whigham, P.A., 1995. Inductive Bias and Genetic Programming, First International Conference on Genetic Algorithms in Engineering Systems: Innovations and Applications, The Institute of Electrical Engineers.

Whigham, P.A., 2000. Induction of a marsupial density model using genetic programming and spatial relationships. Ecological Modelling 131 (2–3), 299–317.

Whigham, P.A., Crapper, P.F., 1999. Time series modelling using genetic programming: an application to rainfall-runoff models. In: Spector, L., Langdon, W.B., O'Reilly, U., Angeline, P.J. (Eds.), Advances in Genetic Programming 3, vol. 5. MIT Press, Cambridge, MA, USA, pp. 89–104.

Whigham, P.A., Reckangel, F., 2000. Evolving difference equations to model freshwater phytoplankton. In: 2000 Congress on Evolutionary Computation, San Diego, USA. IEEE, Piscataway, NJ.

Whigham, P.A., Recknagel, F., 1999. Predictive modelling of plankton dynamics in freshwater lakes using genetic programming. In: Oxley, L., Scrimgeour, F., (eds.), MODSIM'99 International Congress on Modelling and Simulation, The Modelling and Simulation Society of Australia and New Zealand Inc., Hamilton, New Zealand.