Rule-based agents for forecasting algal population dynamics in freshwater lakes discovered by hybrid evolutionary algorithms

Amber Welk, Friedrich Recknagel, Hongqing Cao, Wai-Sum Chan, Anita Talib

School of Earth and Environmental Sciences, University of Adelaide, Australia
Cooperative Research Centre for Water Quality and Treatment, Australia
University Saints Malaysia, Penang, Malaysia

ARTICLE INFO

Article history:
Received 29 December 2006
Received in revised form
11 November 2007
Accepted 4 December 2007

Keywords:
Rule-based agents
Hybrid evolutionary algorithm
k-fold cross-validation
Generic rules
Lake categories
Forecasting
Chlorophyll-a
Microcystis
Oscillatoria

ABSTRACT

In the context of this study two concepts were applied for the development of rule-based agents of algal populations: (1) rule discovery by means of a hybrid evolutionary algorithms (HEA) and rigorous k-fold cross-validation, and (2) rule generalisation by means of merged time-series data of lakes belonging to the same lake category. The rule-based agents developed during this study proved to be both explanatory and predictive. It has been demonstrated that the interpretation of the rules can be brought into the context of empirical and causal knowledge on chlorophyll-a dynamics as well as population dynamics of Microcystis and Oscillatoria under specific water quality conditions. The k-fold cross-validation of the agents based on measured data of each year of similar lakes revealed good forecasting accuracy resulting in \( r^2 \) values ranging between 0.39 and 0.63.

© 2007 Elsevier B.V. All rights reserved.

1. Introduction

Predictive agents for specific algal populations can be powerful tools for early warning and operational control of harmful algal blooms in lakes and drinking water reservoirs. One way of developing predictive agents for algal populations is the extraction of generic rules from ecological time-series data by means of evolutionary algorithms as suggested by Recknagel (2003).

This study demonstrates the development of rule-based agents from merged time-series data of lakes belonging to the same lake category by means of a hybrid evolutionary algorithm (HEA) and a rigorous k-fold cross-validation framework.

Three rule-based agents will be discussed and validated which are applicable for forecasting 5 to 7-days-ahead the concentration of chlorophyll-a in the warm monomictic and mesotrophic reservoirs Myponga and Happy Valley (Australia), the abundance of Microcystis in the shallow polymictic hypertrophic lakes Kasumigaura and Suwa (Japan) as well as the abundance of Oscillatoria in the shallow polymictic and hypertrophic lakes Veluwemeer and Wolderwijd (The Netherlands).

The resulting rule-based agents proved to be both explanatory and predictive. It has been demonstrated that the interpretation of the rules can be brought into the context of empirical and causal understanding of chlorophyll-a dynamics as well as population dynamics of Microcystis and Oscillatoria.
under specific water quality conditions. The k-fold cross-validation of the agents based on measured data of each year of two similar lakes revealed good forecasting accuracy resulting in \( r^2 \) values ranging between 0.39 and 0.63.

### 2. Methods and materials

#### 2.1. Study sites and data

Myponga and Happy Valley reservoirs are located in South Australia and used for drinking water storage. Both reservoirs are classified as warm monomictic and mesotrophic, which experience in summer blue-green algal blooms. Efforts to control algal bloom events in both reservoirs include implementation of artificial mixing and aeration in summer, and operational application of copper sulphate (CuSO₄).

The Japanese lakes Kasumigaura and Suwa are classified as shallow polymictic and hypertrophic. High internal nutrient loadings because of shallowness as well as high external nutrient loadings because of agricultural use and urbanisation of their catchments are typical for both lakes and cause recurrent summer blooms of *Microcystis* (Takamura et al., 1992; Park et al., 1998).

The Dutch lakes Veluwemeer and Wolderwijd are shallow and adjacent where Wolderwijd receives outflowing water from Veluwemeer. For the study period from 1976 to 1993 they were classified as shallow polymictic and hypertrophic, and experienced summer blooms of *Oscillatoria agardhii* (Reeders et al., 1998). Since the 1980s integrated concepts of eutrophication management are being implemented including external nutrient control and food web manipulation (Jagtman et al., 1992; Meijer and Hosper 1997).

As the measurement intervals of the raw data from all water bodies were highly irregular and sampling dates different for physical, chemical and biological variables, the data was interpolated to create daily values as required for the forecasting of daily values.

Table 1 summarises the characteristics of water quality variables from all lakes that were considered in the present study.

<table>
<thead>
<tr>
<th>Table 1 – General characteristics and limnological variables of the lakes Myponga and Happy Valley (Australia), Kasumigaura and Suwa (Japan), and Veluwemeer and Wolderwijd (The Netherlands)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surface area (km²)</td>
</tr>
<tr>
<td>Mean depth (m)</td>
</tr>
<tr>
<td>Circulation type</td>
</tr>
<tr>
<td>Limnological variable</td>
</tr>
<tr>
<td>Turbidity TURB (NTU)</td>
</tr>
<tr>
<td>Conductivity COND (μS/cm)</td>
</tr>
<tr>
<td>Diss. oxygen DO (mg/L)</td>
</tr>
<tr>
<td>Secchi depth SD (m)</td>
</tr>
<tr>
<td>pH</td>
</tr>
<tr>
<td>Nitrate NO₃-N (mg/L)</td>
</tr>
<tr>
<td>Phosphate PO₄-P (mg/L)</td>
</tr>
<tr>
<td>Silica SO₂ (mg/L)</td>
</tr>
<tr>
<td>Chlorophyll-a Chl-a (μg/L)</td>
</tr>
</tbody>
</table>

2.2. Methods

#### 2.2.1. Hybrid evolutionary algorithms

Evolutionary algorithms (EA) are adaptive methods, which mimic processes of biological evolution, natural selection and genetic variation. They search for suitable representations of models by means of genetic operators and the principle of “survival of the fittest”. Owing to their qualities of self-organisation, self-learning, intrinsic parallelism and generality, EAs have been successfully applied to pattern recognition, optimum control and parallel processing (Goldberg, 1989; Bäck et al., 1997).

Fig. 1 represents the flowchart of hybrid evolutionary algorithm (HEA) and Fig. 2 shows the principal approach of using HEA for the discovery of predictive rule sets in ecological time-series data according to Cao et al. (2006). HEA uses both genetic programming (GP) (Koza, 1992, 1994; Banzhaf et al., 1997) to generate and optimise the structure of
rule sets, and genetic algorithms (GA) (Holland, 1975; Mitchell 1996) to optimise the parameters of rule sets. GP is an extension of GA, in which the genetic population consists of computer programs of various shapes and sizes. These programs are subsequently evaluated by means of “fitness cases”. Fitter programs are selected for recombination to
create the next generation by using genetic operators such as crossover and mutation. This step is iterated for consecutive generations until the termination criterion of the run has been satisfied.

### 2.2.2. Framework for agent development based on generic rule sets

Fig. 3 shows the framework for using HEA to develop rule-based agents from merged time-series data of same category lakes by means of rigorous k-fold cross-validation where k-fold cross-validation according to Kohavi (1995) involves k-fold data partitioning and the consecutive use of each part of the data for both training and validation. The framework includes the following 6 steps: (1) determine best performing predictive rule for lakes by iterative k-fold training and validation using merged time-series data, (2) freeze the structure of best performing rule and determine best performing parameter values within the rule by iterative k-fold training and validation, (3) determine the mean and range of parameter values, (4) determine best performing water temperature functions within the established range by iterative k-fold training and validation, (5) substitute constant parameter values in rule with best performing water temperature functions, and (6) freeze generic rule to be validated for each year and lake. It has successfully been applied by Recknagel et al. (2008) for developing a rule-based Microcystis agent for warm monomictic hypertrophic lakes in South Africa.

### 2.2.3. Lake ecosystem categories

Rule-based agents for algal populations need to be generic for more than one freshwater lakes. As lake ecosystems appear in a great variety of structures and functioning determined by climate, morphometry, trophic state, we classified the lakes of which data were used in this study into lake categories as suggested by Recknagel et al. (2008). The lake categories were defined by trophic states and circulation types assuming that the circulation type reflects to some extent climate conditions and morphometry whilst the trophic state indicates habitat properties and community structures. Therefore lakes belonging to the same category are expected to encounter similar ecological behaviours, health issues and management options. After successful validation of specific rule-based agents on several lakes of the same category it needs to be tested if the agents can also be applied to other lakes within that category. Table 2 classifies the six lakes used in this research into the two lake categories warm monomictic and mesotrophic, and shallow polymictic and hypertrophic.

### 3. Results and discussion

As a result of applying HEA within the framework of rigorous k-fold cross-validation according to Figs. 1-3 to time-series data of three pairs of lakes belonging to the same lake category, three rule-based agents have been developed. A rule-based chlorophyll-a agent was discovered for the two warm monomictic and mesotrophic reservoirs Myponga and Happy Valley. A rule-based Microcystis agent was developed for the two shallow polymictic and hypertrophic lakes Kasumigaura and Suwa. A rule-based Oscillatoria agent was developed for the two shallow polymictic and hypertrophic lakes Veluwemeer and Wolderwijd. The agents were determined by the best performing IF-THEN-ELSE rule sets with the best performing temperature function for parameters $p_i$ evaluated by the lowest root mean square error (RMSE), the highest $r^2$, and the visual closeness between measured and predicted data.

#### 3.1. Rule-based chlorophyll-a agent

A total of 10 years of merged and daily interpolated data of the electronically measurable limnological variables water temperature WT, turbidity TURB, conductivity COND and dissolved oxygen DO of the Myponga and Happy Valley reservoirs (see Table 1) were used to develop a rule-based chlorophyll-a

<table>
<thead>
<tr>
<th>Table 2 – Lake ecosystem categories and related rule-based agents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lake categories and ecosystem categories</td>
</tr>
<tr>
<td>Warm monomictic and mesotrophic</td>
</tr>
<tr>
<td>Shallow polymictic and hypertrophic</td>
</tr>
<tr>
<td>Lake examples</td>
</tr>
<tr>
<td>Myponga (Australia)</td>
</tr>
<tr>
<td>Barossa (Australia)</td>
</tr>
<tr>
<td>Kasumigaura (Japan)</td>
</tr>
<tr>
<td>Suwa (Japan)</td>
</tr>
<tr>
<td>Wolderwijd (The Netherlands)</td>
</tr>
<tr>
<td>Veluwemeer (The Netherlands)</td>
</tr>
<tr>
<td>Rule-based agents</td>
</tr>
<tr>
<td>Chlorophyll-a</td>
</tr>
<tr>
<td>Microcystis</td>
</tr>
<tr>
<td>Oscillatoria</td>
</tr>
</tbody>
</table>
In order to apply the agent for 7-days-ahead forecasting, a time lag of 7 days was imposed between input and output data. As a result, we discovered the rule set as documented in Fig. 4. The IF condition of the rule determines a threshold value for water temperature (WT). The rule indicates that IF WT is less than 17.88 °C THEN the Chl-a concentration is calculated by the equation \( \text{Chl-a} = \exp\left(\frac{\text{WT}}{P_1/\text{DO} + P_2}\right) \) ELSE by the equation \( \text{Chl-a} = \exp\left(\frac{\text{WT}}{P_3/\text{DO} + \text{Conductivity}/P_4}\right) \). Obviously, the rule distinguishes between conditions of low growth with cooler water temperatures and higher DO, and conditions of high algal growth reflected by higher water temperatures and lower DO levels. These relationships can also be seen in the input sensitivity plots for the then- and the else-branches of the rule in Fig. 4 and indicate that warmer
water temperatures stimulate algal growth (Fig. 4a and b) and high algal biomass increases BOD resulting in decreasing DO (Fig. 4b).

Fig. 5 shows the validation of the rule-based agent for 7-days-ahead forecasting of Chl-a concentrations in Myponga and Happy Valley reservoirs for the years 1999–2003. Generally the model predicts the timing of peak events very well for both reservoirs but sometimes overestimates magnitudes of years with relative low peaks and underestimates magnitudes of years with relative high peaks. For Myponga reservoir, the overall $r^2$ value of the linear regression between measured and calculated data of all 5 years amounts to 0.42. For Happy Valley

Fig. 6 – Structure of the rule-based Microcystis agent for the lakes Kasumigaura and Suwa. a) Input sensitivity of the then-branch of the rule, b) input sensitivity of the else-branch of the rule.

Fig. 7 – k-fold cross-validation of the rule-based agent for 7-days-ahead forecasting of Microcystis cell concentrations in lakes Kasumigaura and Suwa.
reservoir, the overall $r^2$ value of the linear regression between measured and calculated data of all 5 years amounts to 0.38. These results can be considered quite successful, particularly considering the stochastic nature of the lake conditions as reflected by the data. The ecology of both lakes is temporarily affected by artificial mixing and CuSO4-treatment.

### 3.2. Rule-based Microcystis agent

A total of 20 years of merged and daily interpolated data of the limnological variables water temperature WT, dissolved oxygen DO, Secchi depth SD, pH, phosphate PO4–P, nitrate NO3–N, N:P ratio and chlorophyll-a Chl-a of the lakes Kasumigaura and Suwa (see Table 1) were used to develop a rule-based Microcystis agent. In order to apply the agent for 7-days-ahead forecasting a time lag of 7 days was imposed between input and output data. As a result we discovered the rule set as documented in Fig. 6. The IF condition (Chl-a – WT) > $P_1$ AND ($P < 15.92$) is satisfied, the THEN branch Microcystis = WT * SD is used to calculate the Microcystis cell concentrations. Otherwise, the ELSE branch Microcystis = Chl-a * $P_2$ is used.

The rule set suggested that Chl-a, WT, P and SD were important criteria to stimulate the growth of Microcystis. The causal relationships between Microcystis and these variables were further revealed by their sensitivity curves shown in Fig. 6a and b. The input sensitivity for the THEN branch (Fig. 6a) indicates that both WT and SD are positively related to high abundances of Microcystis (up to 1,200,000 cells/mL). The increase in SD suggested more available photosynthetic light in the water column, which promoted the growth of Microcystis cells. On the other hand, the input sensitivity for the ELSE branch (Fig. 6b) shows positive relationships between Chl-a and WT and the lower abundances of Microcystis (below 200,000 cells/mL).

Although the development of this rule set was limited by the incomplete and inconsistent annual data of two lakes, the rule set was still capable of predicting the Microcystis concentrations 7-days-ahead by feeding the corresponding input data. Fig. 7 illustrates the application of the rule set in every year of both lakes Kasumigaura and Suwa. In general, the rule set was able to capture the inter-annual dynamics of Microcystis with good timing and magnitude represented in the $r^2$ values of 0.39 and of 0.48 for lakes Kasumigaura and Suwa, respectively. Even though the rule set overestimated the low Microcystis events in lake Kasumigaura, it successfully predicted the extreme peak events of Microcystis blooms in lake Suwa. The ability to forecast the bloom at right timing and magnitude is far more important in term of bloom-control practices.

### 3.3. Rule-based Oscillatoria agent

A total of 28 years of merged and daily interpolated data of the limnological variables water temperature WT, dissolved oxygen DO, Secchi depth SD, pH, phosphate PO4–P, nitrate NO3–N, and silica SiO2 of the lakes Veluwemeer and Wolderwijd (see Table 1) were used to develop a rule-based Oscillatoria agent. In order to apply the agent for 5-days-ahead forecasting a time lag of 5 days was imposed between input and output data. As a result we discovered the rule set as documented in Fig. 8. The IF condition (Chl-a – WT) > $P_1$ AND ($P < 15.92$) is satisfied, the THEN branch Oscillatoria = WT * SD is used to calculate the Oscillatoria cell concentrations. Otherwise, the ELSE branch Oscillatoria = Chl-a * $P_2$ is used.

The rule set suggested that Chl-a, WT, P and SD were important criteria to stimulate the growth of Oscillatoria. The causal relationships between Oscillatoria and these variables were further revealed by their sensitivity curves shown in Fig. 8a and b. The input sensitivity for the THEN branch (Fig. 8a) indicates that both WT and SD are positively related to high abundances of Oscillatoria (up to 1,200,000 cells/mL). The increase in SD suggested more available photosynthetic light in the water column, which promoted the growth of Oscillatoria cells. On the other hand, the input sensitivity for the ELSE branch (Fig. 8b) shows positive relationships between Chl-a and WT and the lower abundances of Microcystis (below 200,000 cells/mL).
The then-branch can be interpreted as being relevant for turbid water conditions caused by and favoring high Oscillatoria abundances whilst the else-branch relates to clear-water conditions with relatively low Oscillatoria abundances.

According to the sensitivity curves in Fig. 8a both WT and SiO$_2$ were positively related to the high concentrations of Oscillatoria. The decrease in SD suggests that increasing abundance of Oscillatoria causes more shading. The sensitivity curves in Fig. 8b showed same trends in the relationships between SD and WT, and Oscillatoria abundance as observed in Fig. 8a.

Fig. 9 shows the validation of the rule-based Oscillatoria agent for the lakes Veluwemeer and Wolderwijd. It can be concluded that in general the model predicts the timing of peak events very well for both lakes. However occasionally the agent overestimates magnitudes of years with relative low peaks during post-management years of 1985 and 1993 and underestimates magnitudes of years with relative high peaks in the period prior to 1979 i.e. before the implementation of lake management measures. The prediction results for lake Veluwemeer achieve a $r^2$ value of 0.63 whereas for Lake Wolderwijd the $r^2$ value is 0.62. The interpretation of the rule sets discovered in this study corresponds to some extent with observations on alternative turbid and clear-water states in shallow lakes as reported by Scheffer (1998).

5. Conclusions

In the context of this study two concepts have been applied for the development of rule-based agents of algal populations: (1) rule discovery by means of a hybrid evolutionary algorithms (HEA) and rigorous k-fold cross-validation, and (2) rule generalisation by means of merged time-series data of lakes belonging to the same lake category.

The rule-based agents that have been discovered as an outcome of this study proved to be both explanatory and predictive. It has been demonstrated that the interpretation of the rules can be brought into the context of empirical and causal knowledge on chlorophyll-a dynamics as well as population dynamics of Microcystis and Oscillatoria under specific water quality conditions. The k-fold cross-validation of the agents based on measured data of each year of two similar lakes revealed good forecasting accuracy resulting in $r^2$ values between 0.39 and 0.63.

Future research will focus on developing agent libraries for specific algal populations and lake categories applicable for early warning and operational control of algal blooms. Special attention will be given to the use of electronically measurable water quality data as input variables in order to allow the use of rule-based agents for real-time forecasting of algal blooms.

REFERENCES


