J. Limnol., 2016; 75(s1): 242-252
DOI: 10.4081/jlimnol.2016.1374

ORIGINAL ARTICLE

A non-deterministic approach to forecasting the trophic evolution of lakes

Roberto BERTONI,[1*] Martino BERTONI,[2] Giuseppe MORABITO,[1] Michela ROGORA,[1] Cristiana CALLIERI[1]

[1]Institute of Ecosystem Study - CNR, Largo Tonolli 50, 28922 Verbania, Italy; [2]Biozentrum, University of Basel, Klingelbergstrasse 50/70, CH-4056 Basel
*Corresponding author: r.bertoni@ise.cnr.it

ABSTRACT

Limnologists have long recognized that one of the goals of their discipline is to increase its predictive capability. In recent years, the role of prediction in applied ecology escalated, mainly due to man's increased ability to change the biosphere. Such alterations often came with unplanned and noticeably negative side effects mushrooming from lack of proper attention to long-term consequences. Regression analysis of common limnological parameters has been successfully applied to develop predictive models relating the variability of limnological parameters to specific key causes. These approaches, though, are biased by the requirement of a priori cause-relation assumption, oftentimes difficult to find in the complex, nonlinear relationships entangling ecological data. A set of quantitative tools that can help addressing current environmental challenges avoiding such restrictions is currently being researched and developed within the framework of ecological informatics. One of these approaches attempting to model the relationship between a set of inputs and known outputs, is based on Genetic Algorithms (GA) and Genetic Programming (GP). This stochastic optimization tool is based on the process of evolution in natural systems and was inspired by a direct analogy to sexual reproduction and Charles Darwin's principle of natural selection. GP is an evolutionary algorithm that uses selection and recombination operators to generate a population of equations. Thanks to a 25-year long time-series of regular limnological data, the deep, large, oligotrophic Lake Maggiore (Northern Italy) is the ideal case study to test the predictive ability of GP. Testing of GP on the multi-year data series of this lake has allowed us to verify the forecasting efficacy of the models emerging from GP application. In addition, this non-deterministic approach leads to the discovery of non-obvious relationships between variables and enabled the formulation of new stochastic models.

Key words: Deep lakes; trophic evolution; ecological modelling; genetic programming; Lake Maggiore.

Received: April 2015. Accepted: May 2015.

INTRODUCTION

Limnologists have long recognized that one of the goals of their discipline is to increase its predictive capability (Peters, 1986, 1991). In recent years, the role of prediction in applied ecology grew up, mainly due to man's increased ability to change the biosphere. The man-induced alterations often came with unexpected and remarkably negative side effects arising from the lack of adequate attention to long-term consequences. The current threat to biodiversity and global climate change are paradigmatic examples of such negative effects. Their mitigation requires actions based on efficient models for ecological forecasting (Clark *et al*., 2001). Past applications of predictive limnology proved fundamental, for example, to eutrophication control. Vollenweider (1968), searching for effective responses to the eutrophication problem, formulated successful deterministic models for lake management that predict lake total phosphorus concentrations as a function of lake morphometric/hydraulic characteristics. Dillon and Rigler (1974) among others developed regression models to predict the change in phytoplankton standing crop with a given change in nutrient loading.

Following this path, the regression analysis has been successfully applied to identify the variables most suitable to predict the evolution of the specific components of lake ecosystems. Recent examples are, for instance, the prediction of the cyanobacterial biomass in relation to climate change (Beaulieu *et al.,* 2013) and the modeling of the main driving force of zooplankton dynamics (Perhar *et al.,* 2013), a task intrinsically more difficult than modeling an assemblage of unicellular algal species. These approaches, though, are biased by the requirement of a priori cause-relation assumption, oftentimes difficult to find in the complex, nonlinear relationships entangling ecological data. In addition, it is often difficult to satisfy the restrictive assumptions required by conventional parametric approaches.

One promising set of quantitative tools that can help addressing current environmental challenges avoiding such restrictions is currently being studied and developed within the framework of ecological informatics. This is an interdisciplinary framework promoting the use of advanced computational technology to reveal ecological processes and patterns across levels of ecosystem complexity. Machine Learning (ML) is a rapidly growing area of eco-informatics that is concerned with identifying structure in complex, often nonlinear data and generating accurate pre-

dictive models. Supervised learning is a form of machine learning that aims to model the relationship between a set of inputs given the known outputs. This approach finds its implementation in a huge variety of algorithms. Among these we can find evolutionary algorithms such as Genetic Algorithms (GA) and Genetic Programming (GP). Evolutionary algorithms are stochastic optimization heuristics based on mimicking the process of evolution in natural systems and are inspired by a direct analogy to sexual reproduction and Charles Darwin's principle of natural selection. GP and GA use selection and recombination operators to generate a population of solutions to a problem's instance. These evolve over generations where each individual has a chance to survive or reproduce proportional to its fitness, *i.e.* how well it satisfies the problem. Given enough generation, these algorithms converge to an optimal solution (Poli, 2001; Recknagel, 2001). The best solution, *i.e.* the best predicting equation, can be tested on a subset of data from the time series used to construct the model. The full parallelism between Nature and Computer is summarized in Tab. 1 (from Cagnoni and Poli, 2006). The growing use of these methods in recent years is the direct result of their ability to model complex, nonlinear relationships in ecological data without having to satisfy the restrictive assumptions required by conventional, parametric approaches (Guisan and Zimmermann, 2000; Olden and Jackson, 2002; Elith *et al.,* 2006). As a result, eco-informatics techniques, and in particular GA and GP, have been often applied in limnology to unravel connections between variables controlling the algal population dynamics and to forecast their short and long term evolution (Recknagel *et al.,* 2006; Kim *et al.,* 2012; Recknagel *et al.,* 2013). GP applications to trophic levels higher than phytoplankton are less abundant and more frequently utilized in marine environment (Perhar *et al.,* 2013; Marini and Conversi, 2012).

This paper contains a brief introduction to (GP) and an evaluation of its use in forecasting time series of organic carbon production in a lake for which long time-series of measurements of limnological, hydrological and climate variables are available. A further purpose is to identify, with no a priori deterministic assumptions, the variables with greater predictive power in the complex ecological relationships between plankton populations, physical and chemical water properties as well as climate and environmental changes over time.

Thanks to a 25-year long time-series of regular limnological data, the deep, oligotrophic and large Lake Maggiore (Northern Italy) is the ideal case study to test the predictive ability of GP.

Introduction to Genetic Programming

Genetic Programming is an evolutionary algorithm optimization technique that generates computer programs as solution to a problem. GP derived from Genetic Algoritm (GA), both works in analogy of organisms' reproduction and evolution through selection. However, while GAs are working on solutions of fixed size, GP can generate solutions of variable and increasing size, thus resulting more suitable for ecological modeling. Another advantage of GP over other optimization techniques is its ability to perform automatic feature selection, automatically disregarding those features (*i.e.* variables) not relevant for the solution of the problem. In the context of ecology, feature reduction is extremely desirable, given the huge number of possible variables influencing a system.

GP has been utilized in a variety of benchmark problems (White *et al.,* 2013) and it is applicable in a high number of different contexts (Koza, 1992). In this study we will focus on symbolic regression, as we want to assemble equations effective in forecasting a variable of interest. The general workflow of GP is similar to GA: starting from a randomly initialized population of models/equations, the fitness of each model is evaluated and the process of natural selection is simulated. After the selection of the variable to be predicted, the generation of the best predictive model is accomplished in 5 steps:

i) Generating a random **population** of *individuals*. Each of them is an equation designed to provide the value of a variable of interest as a function of some other measured variables. A common form of encoding is the parse tree, *i.e.* a way of writing an equation compatible with the programming language used in this kind of software (example in Fig. 1).
ii) Evaluating the **fitness** of each equation, as its ability

Tab. 1. Parallelism between Nature and Computer in evolutionary algorithms.

Nature	Computer
Individual	Solution to a problem
Population	Set of solutions
Fitness	Quality of the solution
Chromosome	Representation for a solution (*e.g.*, set of parameters)
Gene	Part of the representation of a solution
Crossover, mutation	Search operators
Natural selection	Promoting the reuse of good (sub-)solutions

to provide estimated data as close as possible to the observed data. This evaluation is performed measuring the error between predicted and actual values.

iii) **Selecting** phase the best performing equations (those with lower error).

iv) To increase the overall population fitness, mainly the 'promising' solutions undergo to **offspring generation**. The process is accomplished through search operators like crossover and mutation, which generate new equations.

v) The **termination** criteria for these algorithms can be an evaluation of the newly generated offspring. If they meet a certain quality the solution is accepted, otherwise the offspring forms a new generation and the process is iterated for several generations.

These steps are illustrated in Fig. 1 using as example a population of equations predicting phytoplankton chlorophyll from nitrogen (N), phosphorus (P) and solar radiation (rad).

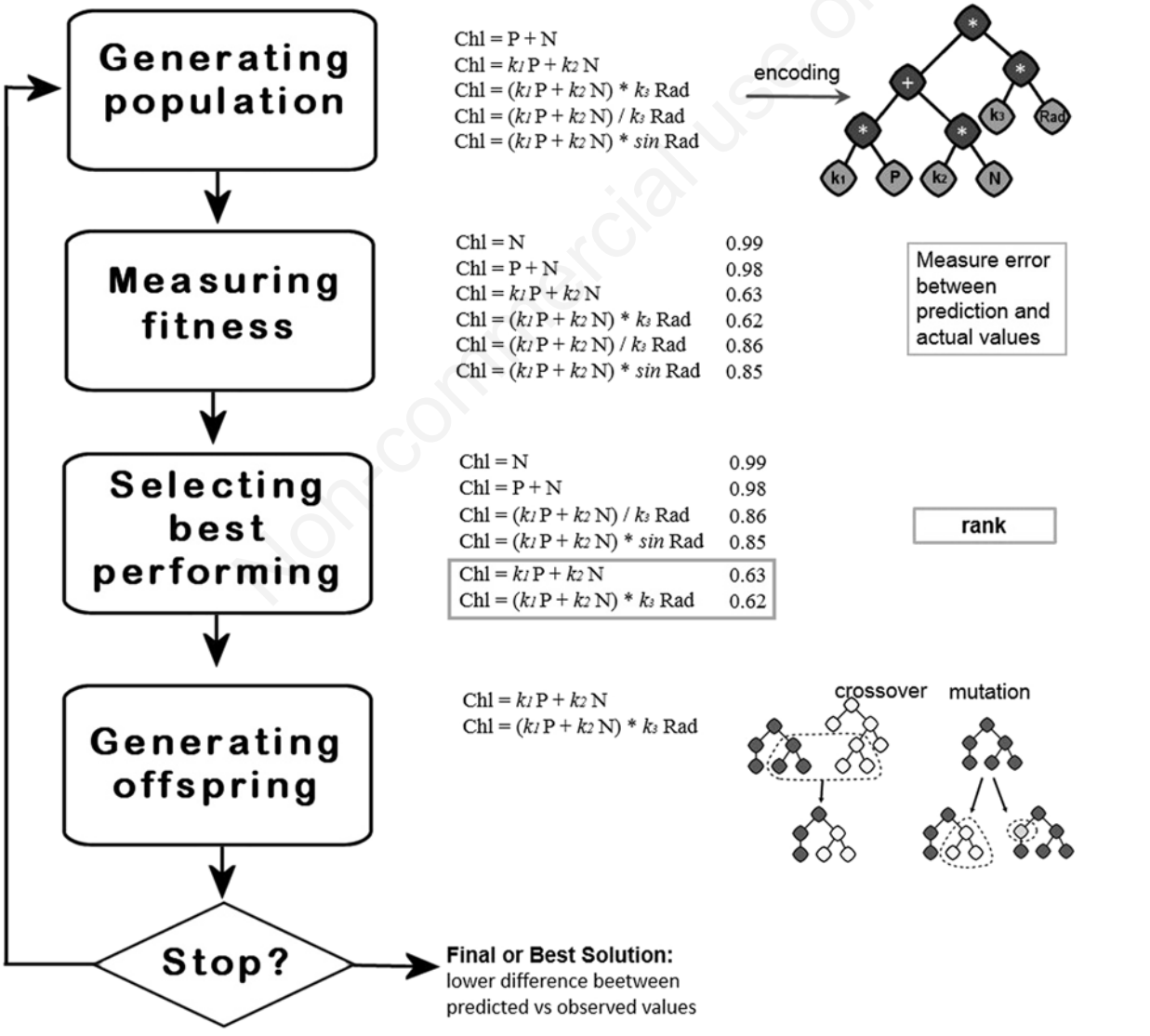

Fig. 1. Flow chart of Genetic Programming data processing.

METHODS

Study site and sampling

Lake Maggiore is a large, deep, subalpine lake (lake area 212 km^2, Z_{max} 372 m) in Northern Italy, included in the Southern Alpine Lakes LTER site. This lake is classified as holo-oligomictic since complete overturn takes place only during periods of strong wind and low air temperatures (Ambrosetti *et al.,* 2003). The total P concentration decreased from 1977 to 1995 by a factor of 4.6 and the lake is now oligotrophic, with TP around 10 µg L^{-1} (Gallina *et al.,* 2013). In this lake, the trend over time of physical, chemical (Salmaso and Mosello, 2011; Salmaso *et al.,* 2013) and biological variables in the size range of the microbial food chain are well studied (Bertoni *et al.,* 2010).

Since 1980, the lake is included in a monitoring program of biological, chemical and physical variables, with monthly/fortnightly samplings along the whole water column, funded by the International Commission for Protection of Italian Swiss Waters (CIPAIS). The data used in this paper, with the relevant analytical details, are available in the annual reports of the Commission downloadable from www.cipais.org.

Sampling took place at the deepest point of the lake. The samples for the biological variables were taken monthly (winter and autumn) and fortnightly (spring and summer). Water samples from the 0-20 m layer, corresponding to epilimnion when the lake is stratified, were collected using a sampler (Bertoni, pat. 96/A 000121) that collects a 5 L integrated sample in a single operation from the surface to 20 m. For the 20-370 m layer, corresponding to hypolimnion when the lake is stratified, samples were taken at 20 m depth and at 50 m, and then at 50-meter intervals down to the bottom Volumes of samples proportional to the thickness of the layer were pooled to obtain an integrated sample. The samples for chemical analysis were collected monthly at 12 depths along the water column. The 0-20 m and 20-370 m integrated values are depth-weighted mean of values from each depth. Samples were prefiltered through a 126 µm plankton net to eliminate larger zooplankton.

Climatic variables

Lake water temperature profiles were obtained at any sampling with reversing thermometers and continuous recording probe (Ocean 316, Idronaut) and used to calculate the monthly average temperature of 0-20 m and 20-370 m layers. Continuous recording of solar radiation was provided by pyranometers placed at the meteorological station of the CNR ISE and used to calculate the monthly average solar radiation (rad). As proxy of the precipitations in the drainage basin we used the monthly average discharge (Q) of the three most important tributaries of Lake Maggiore, whose catchments account for almost 70% of the entire lake watershed.

Water chemistry

Details about the analytical methods and quality assurance/quality controls (QA/QC) procedures used in the hydrochemical laboratory can be found in Mosello *et al.* (2001). For the purposes of this research, monthly data of pH, conductivity (cond), alkalinity (alk), dissolved oxygen concentration (O_2), inorganic carbon (IC), total and reactive phosphorus (TP and RP), nitric and total nitrogen (NO_3 and TN), and reactive silica (Si) were used to calculate their average values in the layers 0-20 and 20-370 m.

Organic carbon and Chlorophyll *a*

Total organic carbon (TOC) concentration was measured with a total organic carbon analyzer (Shimadzu, 5000A). Duplicate subsamples where filtered through GF/C filters (Whatman) previously combusted at 450°C for 3 h. The filters where used to perform Particulate Organic Carbon (POC) analyses by CHN Elemental Analyzer (Carlo Erba, ANA 1500) (Bertoni, 1978).

Chlorophyll *a* (Chl) was determined fluorometrically by Perkin-Elmer LS-2 Filter Fluorometer in duplicate subsamples, filtered through GF/C (Whatman), after pigment extraction with methanol (Holm-Hansen and Riemann, 1978). Chlorophyll to carbon conversion was calculated according to Riemann *et al.* (1989).

Bacteria counting and biovolume measuring

Samples for cell counting were immediately fixed with 0.2 µm filtered formaldehyde (final concentration 2% vol/vol). After 4', 6-diamidin-2-phenylindole (DAPI) staining (final concentration, 0.1 µg mL^{-1}), the samples were filtered onto 0.2 µm pore-size polycarbonate membranes, and counted by epifluorescence microscopy (Zeiss Axioplan microscope equipped with an HBO 100 W lamp, a Neofluar 100 x objective 1.25 x additional magnification and filter sets for UV: BP365, FT 395, LP420). 400 bacterial cells were counted on at least 10 fields in individual filters. Cellular size was measured using image analysis software (Image-Pro Plus 5.1, Media Cybernetics) after manual thresholding and the volumes were estimated according to simple geometric shapes. Bacterial carbon (Bact C) was computed from mean cell volume according Loferer-Krößbacher *et al.* (1998).

Phytoplankton

Phytoplankton determinations were performed on subsamples preserved in acetic Lugol's solution. Counting and measuring were carried out in sedimentation chambers (Vollenweider, 1974) using an inverted microscope.

Cell dimensions were measured using an electronic caliper and a mean biovolume of each species was estimated from linear measures and using the closest geometrical shape (Hillebrand *et al.,* 1999). The organic carbon content of phytoplankton was estimated from mean cell volume using an equation of the form $y=ax^b$, where y is the carbon content, x is the cell volume, the coefficient a and b are equal to 0.12 and 1.015 respectively, according to the equation reported in Montagnes *et al.* (1994). The same coefficients were used for phytoplankton as whole, as well as for diatoms and cyanobacteria.

Calculation and statistic

Trends in the long-term data series were tested using the Mann-Kendall trend test. P values of the statistical tests were related to a significance value of $\alpha=0.05$. In cases with p values <0.05, the null hypothesis (no trend in the data series) was rejected. All analyses were carried out with the Microsoft EXCEL add-in program XLSTAT 2014-TIME.

The Genetic Programming computations were made using Eureqa 1.10 academic version (Schmidt and Lipson, 2009), after statistical standardization to improve data scaling: $y=(Y-\mu)/\sigma$. We wanted to predict the time evolution of TOC, POC, phytoplanktonic carbon and bacterial carbon as function of physical, chemical and climatic variables. These variables to predict are proxy of the lake's trophic conditions and can provide information on the variation over time of the various components of the organic carbon cycle. The GP software was used to search for a formula $y=f(x_1,...,x_n)$, where y is a form of Organic Carbon (OC) and $x_1,...,x_n$ are the 0-20 m and 20-370 m data for each variable:

$$(OC)=f((O_2), (pH), (Cond), (Alk), (NO_3), (RP), (TP), (TN), (Si), (IC), (temp), (S.d), (Q), (rad))$$

The pool of functions which GP uses to map the relation between variables included arithmetic operators (constant, input variable, addition, subtraction, multiplication, division), and exponential operators (exponential, natural logarithm, power, square root) as well as functions accounting for the previous history of the variables (delayed variable, simple moving average). The predictive ability, or fitness, of the equations generated by the GP software was measured as Mean Absolute Error between predicted and actual values. We used the first 21 years of our time series as train-test set and retained the last 4 years for validation purposes. We further divided the train-test set: we used 75% of data for training the GP and 25% for testing and model selection. By default, Eureqa ranks the equations found according to a ratio of complexity (size of the equation) and accuracy (lower mean absolute error); solutions that are accurate but not too complex are accepted and selected for further validation on the last 4-year dataset.

For comparison purposes, using the Python module - Scikit-learn (Pedregosa *et al.,* 2011), we applied a Multiple Linear Regression (MLR) model assuming as dependent variables the organic carbon parameters listed above and as independent variables the physical, chemical and climatic variables appearing in the above equation. The same statistical standardization of data as for Genetic Programming computation was applied.

COMPARING GENETIC PROGRAMMING AND REGRESSION MODELS

Trend of time series

The results of trend test for all the variables considered in the two layers 0-20 m and 20-370 m are presented in Tab. 2. The minimum and maximum values of the water column for the period 1988-2012 are also shown.

Observing the climatic variables it is evident that in the past 25 years the temperature of the surface and deep layers of the lake has increased as the solar radiation, also showing a trend towards the increase. The water supply from the watershed has instead shown no significant trend although there are evidences of a change in the seasonal distribution, frequency and intensity of the rainfall events (Saidi *et al.,* 2015).

The water chemistry data reveal a significant decrease in time of the oxygen concentration, which, however, remained always high enough to avoid the hypolimnetic hypoxia (minimum measured 6.3 mg L^{-1}). This is possible because the oxygen supply in this lake occurs also through mechanisms different from the convective circulation, such as riverine water intrusion and deepening of colder littoral waters (Ambrosetti *et al.,* 2003). As consequence, the oxygen supply of the lake takes place through a complex multi-year cycle, undetectable in a 25 years trend estimate.

The ionic content of lake water also increased, as shown by conductivity trend. The last 25 years have also been characterized by a significant rise in the concentration of inorganic carbon (IC) and of nitrate (NO_3) and total nitrogen (TN) in the hypolimnetic layers. On the contrary, the concentration of total phosphorus (TP) decreased significantly while the inorganic fraction, the reactive phosphorus (RP), showed no significant variation in time.

All the variables related to different forms of organic carbon we wanted to predict showed a significant concentration decrease in the 25 years both in the epilimnetic and the hypolimnetic layers. This holds true also for phytoplanktonic carbon data, which obviously refers only to the photic zone. The bacterial carbon is an exception since significantly increased in the whole water column.

Genetic Programming models

The results obtained with GP are summarized in Tab. 3, where are presented the best performing equations, *i.e.* those

with lower complexity (size of the equation) and higher accuracy (lower Mean Absolute Error), with respect to the test data. The equations of high complexity (size >25) were discarded since the increased complexity added little accuracy increasing the risk of overfitting. The equations with high Mean Absolute Error were also discarded because of their poor forecasting accuracy. It is worth noting that some variables have a relevant predictive value appearing often in the models. Considering the climatic variables, temperature and radiation are often present in 0-20 m layer equations. The predictive role of the river discharge, a proxy of the precipitations in the entire watershed, seems poor since this variable appears only in 20-370 m in the POC model. This result is likely since in Lake Maggiore the riverine inflowing water, often with temperature lower than that of lake's superficial layers, tend to sink rapidly into the deeper layers. The frequent presence of nitrogen in the equations leads to reconsider the role of this element, reputed non-limiting due to its high concentration in Lake Maggiore waters (Morabito *et al.,* 2003). Nitrate also ap-

Tab. 2. Trends of time series for different variables during the period 1988-2012 evaluated with the Mann-Kendall Test using monthly average data.

Variable	Short	Unit	0-20 m	20-370 m	Range, min-max
Water temperature	Temp	°C	↑	↑	6.2-21.4
Main tributaries discharge	Q	m^3 sec^{-1}	-		40.7-738.8
Solar radiation	Rad	MJ m^{-2}	↑		2182-23650
Dissolved oxygen	O_2	mg L^{-1}	↓	↓	6.3-12.7
pH	pH		-	↑	7.1-8.9
Conductivity at 20°C	Cond	µS cm^{-1}	↑	↑	117.8-156.4
Alkalinity	Alk	meq L^{-1}	↑	↑	0.640-0.888
Nitrate	NO_3	µg N L^{-1}	↑	↑	531-904
Total nitrogen	TN	µg N L^{-1}	-	↑	693-1124
Reactive Phosphorus	RP	µg P L^{-1}	-	-	1.0-17
Total Phosphorus	TP	µg P L^{-1}	↓	↓	3.3-21
Reactive silica	Si	mg Si L^{-1}	-	↑	0.1-1.9
Inorganic carbon	IC	mg C L^{-1}	↑	↑	8.1-12.1
Total organic Carbon	TOC	µg C L^{-1}	↓	↓	223-2362
Particulate organic Carbon	POC	µg C L^{-1}	↓	↓	32-721
Bacterial Carbon	Bact C	µg L^{-1}	↑	↑	3-251
Phytoplankton carbon (*from Chlorophyl*)	Phy (Chl)	µg L^{-1}	↓		3.3-436.7
Phytoplankton carbon (*from biovolume*)	Phy (biov)	µg L^{-1}	↓		3-619
Diatoms carbon (*from biovolume*)	Dia (biov)	µg L^{-1}	-		0.4-508
Cyano carbon (*from biovolume*)	Cy (biov)	µg L^{-1}	↓		0.1-532
Secchi disk	S.d.	m	-		2.5-16.4

↓, significant decrease; ↑, significant increase; -, no significant trend.

Tab. 3. Best performing equations for the 0-20m and 20-370m layers with their complexity (size) and accuracy (fit) found with GP using data from the period 1988- 2008 (acronyms in Tab. 1). The operator sma_{12} (simple moving average of the last 12 values of a variable) appears in the equation if the previous history of a variable has a relevant predictive value.

0-20 m	Size
TOC=0.109*TN+sma_{12} pH-0.594*IC	12
POC=0.478*pH+0.143* rad-0.286*NO_3-0.343*S.d-0.419*temp	19
Bact C=0.369*temp+0.237*NO_3-0.134-0.255*Si	13
Phy (biov)=0.289*rad+0.285*pH-0.222*NO_3-0.257*Si-0.361*S.d.-0.683*temp	23
Phy (Chl)=0.438*O_2+0.319* TP-0.077-0.132*RP-0.438 *NO_3	17
Dia (biov)=0.401*rad-0.313*S.d.-0.604*Si-0.812*temp	15
Cy (biov)=(0.292+0.042*RP)/exp(IC)-0.642	14
20-370 m	
TOC=0.004-0.082*TN-0.164*O_2-0.275*pH	13
POC=0.122*Q+0.140* O_2*alk-0.079-0.419*IC	15
Bact C=0.611*NO_3-0.345*Cond	7

pears in conjunction with the total and inorganic phosphorus as predictor of phytoplanktonic carbon estimated through Chlorophyll *a* concentration.

Multiple Linear Regression models

An application of MLR consistent with the statistical theory would involve a selection of predictor variables, excluding from the model the variables which are not independent and those without a known causal link with the dependent variable. Nevertheless, we decided to keep all the available variables to ensure maximum comparability of the two approaches. For the same reason we tested the MLR model retaining all variables regardless of whether a variable reaches statistical significance. Appropriate data transformations were applied when necessary to satisfy the requirements of parametric statistic. The equations obtained with the MLR model are presented in Tab. 4. Also the MLR equations were computed from data of the period 1988-2008. The remaining four years (2009-2012) were used to validate the models.

In Fig. 2 the forecasted trends of TOC, POC and Bact C for the period 2009-2012 computed from GP and MLR models are presented. For each variable, the trend of the actual data is also reported for sake of comparison. The bar chart below each trend plot reports the absolute error of GP and MLR based forecast. The overall error of GP and MLR on each variable is illustrated in the side boxplot showing the median, the first and third quartiles, the interquartile range, and the outliers.

Looking to the variables related to the entire water column the predictive accuracy of GP and MLR models for TOC, POC and Bact C is similar both in 0-20 m and 20-370 m layer. However the overall error of GP is lower than that of MLR, with the exception of Bact C in the more superficial zone. The difference between predicted and actual values (*i.e.*, the absolute error) is higher when the observed values peak. This suggests that some trigger variable or some synergistic effect is not included in the model. Note that in the bar chart of absolute error of TOC, 0-20 m layer, the data for the first year are missing. This is due to the presence in the relevant equation (TOC=0.109*TN+sma_{12} pH-0.594 * IC) of a *history term* (sma_{12} pH) including the simple mobile average of the pH values of previous year as relevant forecasting variable. This suggest a lasting effect of pH on the IC concentration, which is also a relevant predictor of TOC concentration. Because of the *history term* the Absolute Error of GP forecasting can not be computed for the first year. The GP and MLR comparison in the boxplot refers to years 2010-2012 only.

Fig. 3 presents the forecasted trend of the phytoplanktonic organic carbon data, computed from chlorophyll and algal biovolume data for the period 2009-2012. The forecast obtained from GP and MLR models for each variable are plotted together with the actual data plot. The bar chart below each trend plot report the absolute error of GP and MLR based forecast. The overall error of GP and MLR, for each variable, is compared in the side boxplot showing the median, the first and third quartiles, the interquartile range and the outliers.

Considering the variables accounting for phytoplank-

Tab. 4. Coefficients of the MLR equations for the 0-20m and 20-370m layers obtained using data from the period 1988-2008

	0-20 m							20-370 m		
dep / *indep*	TOC	POC	Bact C	Phy (Chl)	Phy (biov)	Dia (biov)	Cy (biov)	TOC	POC	Bact C
Intercept	-0.016	0.049	-0.029	0.045	0.018	0.010	0.035	-0.079	0.044	0.033
O_2	0.122	-0.029	-0.086	0.159	-0.026	0.022	-0.059	-0.112	-0.092	0.124
pH	0.182	0.501	-0.636	0.240	0.556	0.090	0.514	-0.517	0.229	-0.542
Cond	0.218	0.159	-0.215	0.236	0.182	-0.003	0.055	0.126	0.153	-0.340
Alk	-0.555	-0.326	1.041	-0.682	-0.645	-0.084	-0.344	0.245	-0.148	0.957
NO_3	-0.211	-0.236	0.292	-0.298	-0.188	-0.045	-0.396	-0.062	-0.032	0.424
RP	0.101	-0.090	0.076	-0.072	-0.086	-0.042	-0.014	-0.008	-0.005	0.014
TP	0.058	0.122	-0.083	0.260	0.035	0.010	-0.008	-0.052	0.017	0.014
TN	0.143	0.061	-0.172	-0.008	0.038	-0.052	0.098	0.007	0.029	0.090
Si	-0.062	-0.028	-0.342	-0.296	-0.197	-0.589	0.395	-0.105	-0.337	0.271
IC	-0.081	0.144	-1.042	0.406	0.464	0.265	-0.029	-0.635	0.072	-1.005
Temp	-0.002	-0.335	0.216	-0.529	-0.719	-0.833	-0.283	0.043	-0.123	-0.137
Q	-0.081	-0.061	-0.071	-0.084	-0.069	-0.031	0.022	-0.068	0.099	0.090
S.d.	-0.115	-0.360	-0.131	-0.360	-0.423	-0.379	-0.106			
Rad	-0.085	0.201	0.299	0.048	0.349	0.383	0.021			

dep, dependent variables; indep, independent variables; other acronyms in Tab. 2.

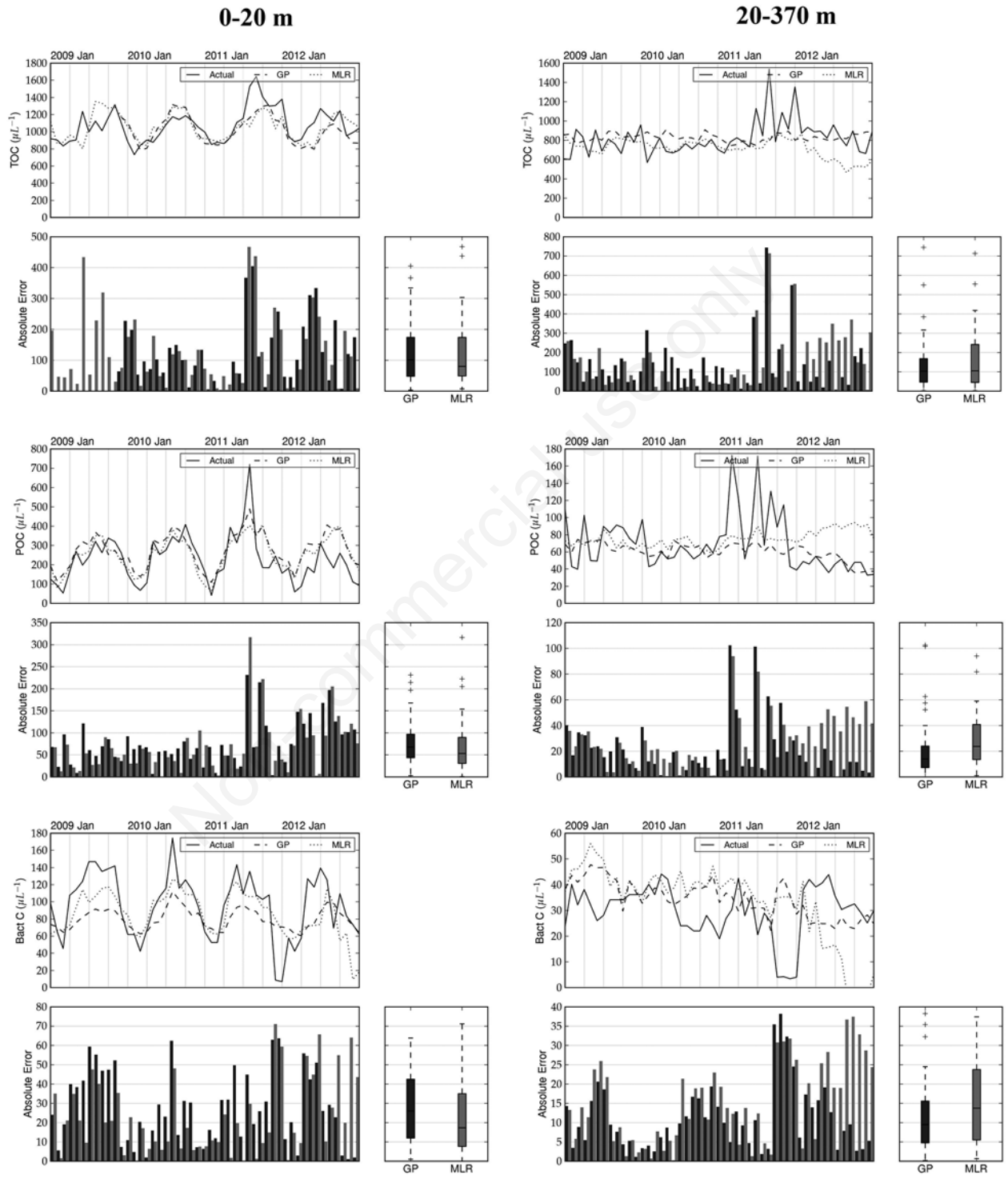

Fig. 2. GP and MLR based forecast, absolute error (bar chart) and mean absolute error (box plot), compared with the actual trend of TOC, POC and Bact C for the period 2009-2012 in the superficial (0-20 m) and deep (20-370 m) layers of Lake Maggiore.

tonic carbon production, the GP and MLR perform in a comparable way providing data forecast affected by errors of comparable size. Also with phytoplankton related variables, the difference between predicted and actual values is larger than the measured values, suggesting the exclusion from the model of some relevant driving variable.

If we compare GP and MLR results (Figs. 2 and 3), the predictive capability of the two approaches result similar. However GP provides effective forecasting models automatically selecting the variables with the highest predictive power. This result allows a drastic reduction in the number of variables to be monitored, with a consequent reduction of monitoring costs and time. In addition the GP approach highlights non-obvious and unforeseen relationships between the variables considered. In particular, in 0-20 m layer TN and NO_3 concentration are present in the equations predicting TOC, POC, Bacterial C and phytoplanktonic C evaluated from both microscopic determination and cell Chlorophyll content. This result is somewhat unexpected since the high concentration of N in Lake Maggiore, and the high N:P ratio, had suggested this nutrient to play a minor role respect to phosphorus as a controlling variable. TP and RP are relevant predictive variables only forecasting phytoplanktonic carbon estimated through chlorophyll concentration.

TN and NO_3 appear in 20-370 m equations forecasting TOC and Bact C, supporting that hypolimnetic layer is a site of intense activity in the N cycle, as suggested by Callieri *et al.* (2014).

Another variable that appears to have a significant role in the GP models is, in the layer 0-20 m, the temperature, present in the equations of POC, Bact C, Phy (biov) and Dia (biov). Silica is a significant predictor of this latter variable, as might be expected; Si also appears significant in predicting the Bact C, suggesting the important role of diatoms in producing the substrate for bacterial populations.

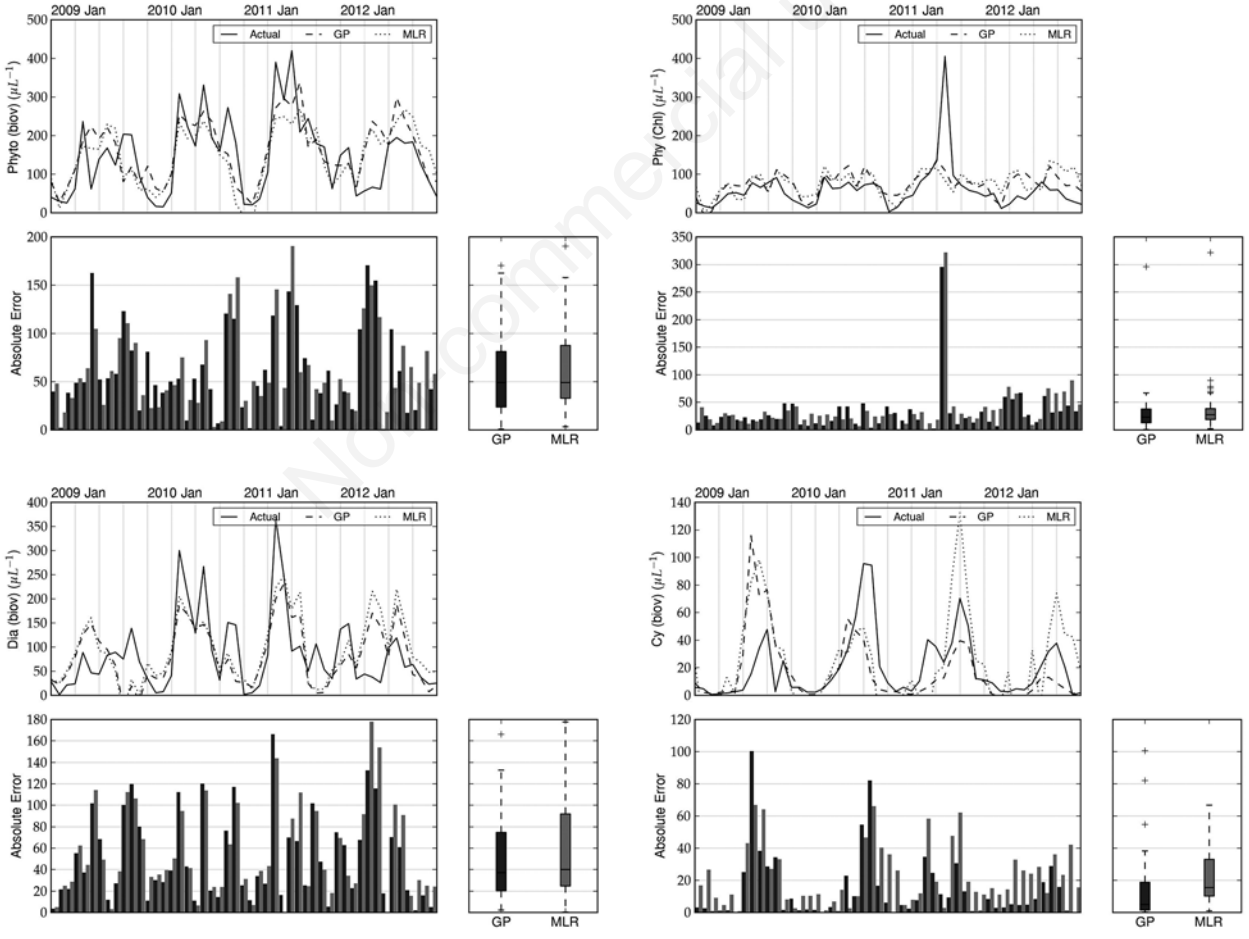

Fig. 3. GP and MLR based forecast, absolute error (bar chart) and mean absolute error (box plot), compared with the actual trend of the organic carbon content of phytoplankton, evaluated from biovolume [Phy(biov)] and Chlorophyll [Phy (Chl)], and of Diatoms [Dia (biov)] and Cyanobacteria [Cy (biov)] for the period 2009-2012 in the superficial (0-20 m) layers of Lake Maggiore.

In the past, the comparison was made of different evolutionary computation techniques such as Artificial Neural Networks (ANN) and GP (Muttil *et al.,* 2006). Furthermore, many examples exist of use of GP to build predictive models of the time-series dynamics of phytoplankton in lakes (Dong-Kyun *et al.,* 2007). However, to our knowledge in this paper for the first time regressive and evolutionary computation techniques are compared on a twenty-five years data series to forecast a compartment of the ecosystem wider than the phytoplankton.

CONCLUSIONS

Continuing *in situ* measurements of limnological, hydrological and climate variables of lakes and rivers retain most momentous information about complex ecological relationships between plankton populations, physical and chemical water properties, as well as climate and environmental changes over time. The extraction of valuable ecological information from long-term time series data allows the generation of predictive models useful for the management of freshwater ecosystems and the advance of theories on their evolution and functioning. GP is an excellent tool to achieve these objectives, since it is capable of producing effective predictive stochastic models, without the constraints imposed by parametric statistic and deterministic models. In addition, the GP approach is able to highlight non-obvious relationships between variables, which can help understanding the long-term ecological variability.

REFERENCES

Ambrosetti W, Barbanti L, Sala N, 2003. Residence time and physical processes in lakes. J. Limnol. 62(1s):1-15.

Beaulieu M, Pick F, and Gregory-Eaves I, 2013. Nutrients and water temperature are significant predictors of cyanobacterial biomass in a 1147 lakes data set. Limnol. Oceanogr. 58: 736-1746.

Bertoni R, 1978. Automatic determination of carbon and nitrogen in suspended matter of natural water with Carlo Erba 1106 CHN elemental analyzer. Mem. Ist. Ital. Idrobiol. 36:297-301.

Bertoni R, Callieri C, Corno G, Rasconi S, Caravati E, Contesini M, 2010. Long-term trends of epilimnetic and hypolimnetic bacteria and organic carbon in a deep holo-oligomictic lake. Hydrobiologia 644:279-287.

Cagnoni S, Poli R, 2006. Genetic and evolutionary computation. Intelligenza Artificiale 3:94-101.

Callieri C, Coci M, Eckert EM, Salcher MM, Bertoni R, 2014. Archaea and Bacteria in deep lake hypolimnion: *in situ* dark inorganic carbon uptake. J. Limnol. 73:47-54.

Clark JS, Carpenter SR, Barber M, Collins S, Dobson A, Foley JA , Lodge DM, Pascual M, Pielke R Jr., Pizer W, Pringle C, Reid WV, Rose KA, Sala O, Schlesinger WH, Wall DH, We D, 2001. Ecological forecasts: an emerging imperative. Science 29:657-660.

Dillon PJ, Rigler FH, 1974. The phosphorus-chlorophyll relationship in lakes. Limnol. Oceanogr. 19:767-773.

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

Elith J, Graham CH, Anderson RP, Dudík M, Ferrier S, Guisan A, Hijmans RJ, Huettmann F, Leathwick JR, Lehmann A, Li J, Lohmann LG, Loiselle BA, Manion G, Moritz C, Nakamura M, Nakazawa Y, Overton JMcC, Townsend Peterson A, Phillips SJ, Richardson K, Scachetti-Pereira R, Schapire RE, Soberón J, Williams S, Wisz MS, Zimmermann NE, 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29:129-151.

Gallina N, Salmaso N, Morabito G, Beniston N, 2013. Phytoplankton configuration in six deep lakes in the peri-Alpine region: are the key drivers related to eutrophication and climate? Aquat. Ecol. 47:177-193.

Guisan A, Zimmermann NE, 2000. Predictive habitat distribution models in ecology. Ecol. Model. 135:147-186.

Holm-Hansen O, Riemann B, 1978. Chlorophyll *a* determination: improvements in methodology. Oikos 30:438-447.

Hillebrand H, Dürselen CD, Kirschtel D, Pollingher U, Zohary T, 1999. Biovolume calculation for pelagic and benthic microalgae. J. Phycol. 35:403-424.

Kim DK, Jeong KS, McKay RIB, Chon TS, Joo GJ, 2012. Machine learning for predictive management: short and long term prediction of phytoplankton biomass using genetic algorithm based recurrent neural networks. Int. J. Environ. Res. 6:95-108.

Koza JR, 1992. Genetic Programming: on the programming of computers by means of natural selection. MIT Press: 699 pp.

Loferer-Krößbacher M, Klima J, Psenner R, 1998. Determination of bacterial cell dry mass by transmission electron microscopy and densitometric image analysis. Appl. Environ. Microbiol. 64:688-694.

Marini S, Conversi A, 2012. Understanding zooplankton long term variability through genetic programming, p. 50-61. In: E. Marchiori, J.H. Moore and J.C. Rajapakse (eds.), Evolutionary computation, machine learning and data mining in bioinformatics. Springer.

Montagnes DJS, Berges JA, Harrison PJ, Taylor FJR, 1994. Estimating carbon, nitrogen, protein and Clorophyll *a* from volume in marine phytoplankton. Limnol. Oceanogr. 39:1044-1060.

Morabito G, Oggioni A, Panzani P, 2003. Phytoplankton assemblage at equilibrium in large and deep subalpine lakes: a case study from Lago Maggiore (N. Italy), p. 37-48. In: L. Naselli-Flores, J. Padisák and M.F. Bach (eds.), Phytoplankton and equilibrium concept: the ecology of steady-state assemblages. Springer.

Mosello R, Barbieri A, Brizzio MC, Calderoni A, Marchetto A, Passera S, Rogora M, Tartari GA, 2001. Nitrogen budget of Lago Maggiore: the relative importance of atmospheric deposition and catchment sources. J. Limnol. 60: 27-40.

Muttil N, Chau KW, 2006. Neural network and genetic programming for modelling coastal algal blooms. Int. J. Environ. Pollut. 28:223-238.

Olden JD, Jackson DA, 2002. A comparison of statistical approaches for modelling fish species distributions. Freshwater Biol. 47:1976-1995.

Pedregosa, F, Varoquaux G, Gramfort A, Michel V, Thirion B,

Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E, 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12:2825-2830.

Perhar G, Arhonditsis GB, Brett M, 2013. Modeling zooplankton growth in Lake Washington: a mechanistic approach to physiology in a eutrophication model. Ecol. Model. 258:101-121

Peters RH, 1986. The role of prediction in limnology. Limnol. Oceanogr. 31: 1143-1159.

Peters RH, 1991. A critique for ecology. Cambridge Univeristy Press, Cambridge: 366 pp.

Poli R, 2001. Exact schema theory for genetic programming and variable-length genetic algorithms with one-point crossover. Genet. Program. Evol. M. 2:123-163.

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

Recknagel F, Cao H, Kim, Takamura N, Welk A, 2006. Unravelling and forecasting algal population dynamics in two lakes different in morphometry and eutrophication by neural and evolutionary computation. Ecol. Inform. 1: 133-151.

Recknagel F, Ostrovsky I, Cao H, Zohary T, Zhangc X, 2013. Ecological relationships, thresholds and time-lags determining phytoplankton community dynamics of Lake Kinneret, Israel elucidated by evolutionary computation and wavelets. Ecol. Model. 255:70-86.

Riemann B, Simonsen P, Stensgaard L, 1989. The carbon and Clorophyll content of phytoplankton from various nutrient regimes J. Plankton Res. 11:1037-1045.

Schmidt M, Lipson H, 2009. Distilling free-form natural laws from experimental data. Science 324:81-85.

Saidi, H, Ciampittiello M, Dresti C, Ghiglieri G, 2015. Assessment of trends in extreme precipitation events: a case study in Piedmont (North-West Italy). Water Resour. Manag. 29:63-80.

Salmaso N, Mosello R, 2011. Limnological research in the deep southern subalpine lakes: synthesis, directions and perspectives. Adv. Oceanogr. Limnol. 1:29-66.

Salmaso N, Buzzi F, Cerasino L, Garibaldi L, Leoni B, Morabito G, Rogora M, Simona M, 2013. Influence of atmospheric modes of variability on the limnological characteristics of large lakes south of the Alps: a new emerging paradigm. Hydrobiologia 731:31-48.

White DR, Luke S, Manzoni L, Castelli M, Vanneschi L, Jaskowski W, Krawiec K, Harper R, De Jong K, O'Reilly UM, 2013. Genetic programming needs better benchmarks, p. 791-798. Proc. 14th Annual Conf. on Genetic and evolutionary computation. ACM New York, NY, USA.

Vollenweider RA, 1968. Scientific fundamentals of the eutrophication of lakes and flowing waters, with particular reference to nitrogen and phosphorus as factors in eutrophication. Organization for Economic Cooperation and Development, Paris, France: 192 pp.

Vollenweider RA, 1974. A manual on methods for measuring primary production in aquatic environments. 2nd ed. Blackwell Scientific Publ., Oxford: 213 pp.