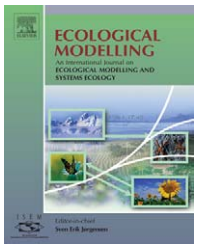
available at www.sciencedirect.com

SCIENCE DIRECT®

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



Automated modelling of a food web in lake Bled using measured data and a library of domain knowledge

Nataša Atanasova[a,*], *Ljupčo Todorovski*[b], *Sašo Džeroski*[b], *Špela Rekar Remec*[c], *Friedrich Recknagel*[d], *Boris Kompare*[a]

[a] *Faculty of Civil and Geodetic Engineering, University of Ljubljana, Slovenia*
[b] *Jožef Stefan Institute, Slovenia*
[c] *Environmental Agency of the Republic of Slovenia, Slovenia*
[d] *University of Adelaide, Australia*

ARTICLE INFO

Article history:
Available online 20 December 2005

Keywords:
Aquatic ecosystems
Food web modelling
Dynamic systems
Automated modelling
Computational scientific discovery

ABSTRACT

In this paper, we applied automated modelling (computer model construction) method to the task of modelling a complex lake ecosystem. The method (Lagramge) integrates domain expert knowledge in the process of automated model induction from given data set. The data set comprises long-term measurements (from 1995 to 2002) of physical, chemical and biological data in lake Bled, Slovenia. Given expert knowledge in terms of a simple food web concept and rules for modelling thereof, we first induced a model for long-term dynamics of the phytoplankton in the lake. Failing to obtain a good fit, we also induced models of phytoplankton dynamics for each year separately. The differences between these models indicate structural dynamics of the food web in lake Bled, i.e., indicate that the behaviour of the lake is changing from year to year. Additionally, we successfully induced a three-equation model (nutrient–phytoplankton–zooplankton) on the data from year 1996.

1. Introduction

Lake Bled has been a subject of exploration since late 1950s when first indices of eutrophication became obvious (Sketelj and Rejic, 1958). At early stages the research focused on measurements and observations, which amongst other showed very high dynamicity of the lake behaviour and rank it among complex ecosystems. Rismal (1980) set the first model of the lake using a stationary version of the Imboden's model (1974), which he improved to obtain inflow and outflow from each layer in order to simulate various proposed restoration measures, i.e., bringing additional fresh water into the hypolimnion, construction of a hypolimnetic siphon that takes the most nutrients' rich water from the bottom layers and reducing the nutrients' input to the lake. The benefits of the proposed siphoning were presented by a 2D and three-dimensional (3D) hydrodynamic model (Rismal et al., 1997). Later Kompare (1995) and Kompare et al. (1997) used machine learning techniques to model the lake's behaviour and to discover some additional knowledge from measured data. His models showed a typical (3D) behaviour. Thus, the lake cannot be modelled properly with 0D, 1-box models such as Vollenweider's model (1968), or 2-box model (Imboden, 1974; Rismal, 1980). According to this fact, physical segmentation of the system is required, i.e. at least 3-box (epi-, meta-, hypolimnion) for each of the two basins (eastern and western). This research showed that we need a very complex mathematical model to adequately describe this system.

On the other hand, regardless to their complexity, models represent not more than a simplified perception and under-

* *Corresponding author.*
E-mail address: natanaso@fgg.uni-lj.si (N. Atanasova).

0304-3800/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2005.10.029

standing of the natural processes. Even if we know the concept of the system very well, we still have to solve a number of equations containing constants and parameters, which need to be estimated. Usually this leads to some numerical problems. Thus, we have to balance between too sophisticated models with many parameters, difficult to be estimated and too simple models with limited use.

Several questions emerge from this dilemma: (1) do we really need and can cope with such complex models, (2) is it possible to find a model structure that will properly cover the lake dynamics under all external conditions and for long period of time and (3) is the lake's system structure too dynamic for one model to cover the full long-term behaviour? We offer some answers to these questions using an advanced machine learning technique Lagramge (Džeroski and Todorovski, 2003; Todorovski, 2003). Lagramge joins two fields in automated modelling, i.e. compositional modelling and a machine learning method, i.e. model induction from data. Compositional modelling builds models by assembling model fragments, typically from a library of model fragments, into an adequate model. In contrast, induction methods usually tackle the same task without incorporating domain expert knowledge in the procedure for model construction. The method used in the paper integrates the domain expert knowledge, gathered in a knowledge library (Atanasova et al., 2006), in the process of induction, performed by machine learning tools. This integration provides us with a guarantee that the constructed models will follow the basic principles from the domain of interest.

In the early days of the development of these methods (Todorovski and Džeroski, 1997), the knowledge had to be provided as an explicit specification of the space of candidate models. Now, Lagramge allow the user to provide higher level (generic) knowledge about building mathematical models of complex real-world systems in the domain of interest (Todorovski, 2003). Given such library of knowledge and a specification of the modelling task, Lagramge first builds a specification of the space of candidate models and then, following the specification, searches for the model that follows the specification and fits measurement data best. Note that Lagramge searches for both optimal structure of the model as well as the optimal values of the model parameters.

2. The method: automated modelling framework

The machine learning method, used in this paper, supports introduction of the background modelling knowledge in the procedure of model induction from data. The knowledge provides recipe for building models in the domain of interest—it provides: (1) taxonomy of basic process classes in the domain, (2) commonly used modelling alternatives for the processes in these classes, as well as (3) rules for combining the models of individual processes into the model of the whole observed system. Process classes represent a set of similar processes, for example, a process class "primary producer growth" represents different types of growth processes including unlimited (exponential) growth, logistic (limited) growth, nutrient limited growth, etc. The knowledge library used here provides knowledge for modelling of food webs in lakes, following the mass conservation principle. The models are based on ordinary differential equations. For further details see (Atanasova et al., 2006).

In order to apply the modelling framework to a particular task of modelling a specific ecosystem, we have to provide modelling task specification, i.e., specification of the observed system variables and processes. Given a specification of modelling task at hand, Lagramge's pre-processor can transform the high-level knowledge from the library into an operational form of a grammar that specifies the space of candidate models of the observed system. Once we have the grammar, we can use equation discovery method Lagramge to heuristically search through the space of candidate models and match each of them to submitted data by fitting the values of the constant parameters. These models can be evaluated by two heuristic functions. One is mean square error (MSE) – it measures the discrepancy between measured data and data obtained by simulating the model. The other is minimum description length (MDL) function that takes into account model complexity and introduces preference towards simpler models.

As described above, the space of candidate models depends on knowledge library and modelling task specification. User can control the space of candidate models by providing different levels of detail about the model in the task specification. The detail level of the model definition can vary according to the expert knowledge about the observed system—the more structure is defined (or fixed) in the task specification, the smaller is the space of candidate models. This space is largest, if user only defines the state variables and does not specify any processes. In this case, Lagramge would search for models that are based on arbitrary combination of possible basic processes. If we know the relevant process classes for the observed system (or the particular system variable), we can further limit the space of candidate models to those that include these processes from those classes. Now, Lagramge will search for suitable process formulation within the specified process classes. Further limitation would include specification of the process formulation within the process class. In this case, the structure of the model is completely defined by the user and Lagramge performs only parameter calibration according to the given data set. Theoretically, we could even determine the parameters values and contract the search space to a single model, which would be a null task for Lagramge. This can be beneficial when we want to fix equations for only some of the system variables and let Lagramge to look for appropriate structure and parameters for the rest. Thus, the modelling formalism has the ability to complete a partially specified model.

Further details about the modelling framework can be found in (Todorovski, 2003; Atanasova et al., 2006).

3. Lake Bled data set

Lake Bled is a typical dimictic, subalpine lake of glacial-tectonic origin, situated in the NW Slovenia (14°7′N and 46°23′E), Europe. It occupies an area of 1.4 km^2 with a maximum depth of 30.1 m and an average depth of 17.9 m (Sketelj and Rejic, 1958). A sunken reef in the north–south direction at the position of the Bled island divides the lake into two

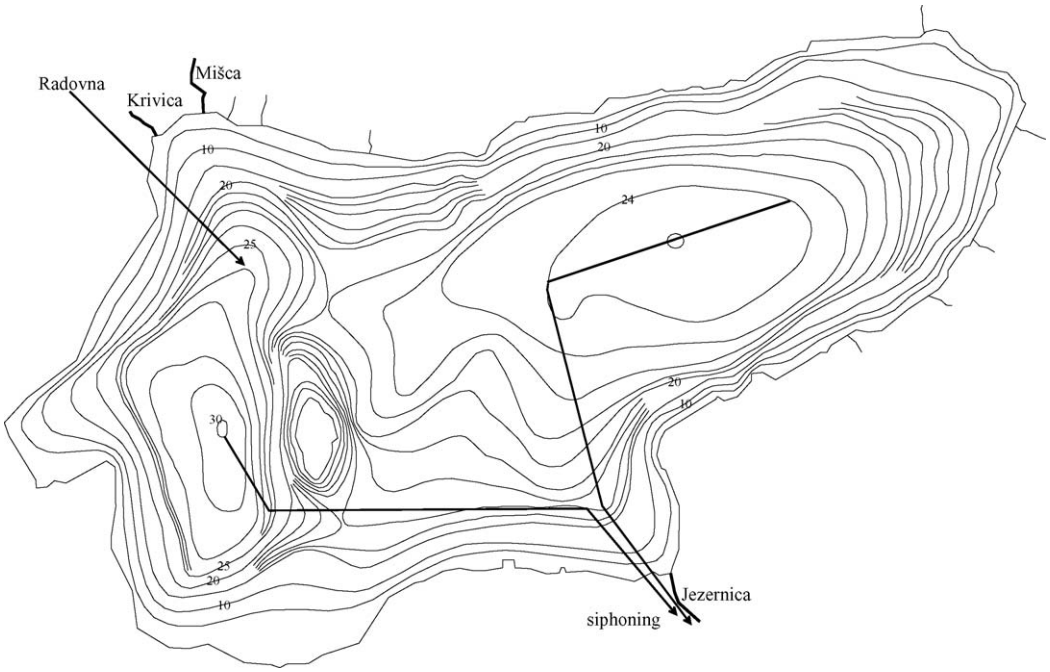


Fig. 1 – Lake Bled, adopted by (Sketelj and Rejic, 1958) and (Rismal, 1980). The dots on each side of the lake represent the sampling points.

basins – eastern and western (see Fig. 1). The morphological characteristics of the lake are shown in Table 1. The monitoring of lake Bled has been a part of the Slovene National Water-Quality Monitoring Programme since 1975. The data, obtained from the Ministry for Environment, Spatial Planning and Energy, Environmental Agency of The Republic of Slovenia, comprise long-term (from 1987 to 2002) measurements of physical, chemical and biological parameters, but only the data from 1995 to 2002 are consistent and suitable for model induction. Samples are taken at two deepest locations in western and eastern basin, at every 2 m from the surface to the bottom (Fig. 1). During the periods when the lake surface was covered with ice, sampling was not performed. That is the reason why some data, especially data at the beginning or the end of the year, is missing. In 1995 and 1996 the sampling was performed all year around, since the lake was not ice covered, while in other years sampling usually started in March (or even April in 2001).

3.1. *Physical data*

The lake receives three major streams, i.e. the river Radovna, small torrent Krivica and the creek Mišca. There are also some minor inflows but for modelling purposes their influence was neglected. Flow rates of the inflows are measured daily, whereas the quality parameters of the streams are measured monthly. The lake has one natural outflow, the river Jezernica and a siphoning outflow. The flow rates of both outflows are measured daily. Light is measured half-hourly, from 1993 as global radiation in W/m^2 at a location near the lake. Water temperature and water transparency are measured monthly.

3.2. *Chemical and biological data about the lake*

Samples for physical, chemical and biological analyses were taken in the period from 1995 to 2002 in the eastern and western lake basins monthly, at 2-m intervals through the water column from the surface to 30 m at the western, and from the surface to 24 m at the eastern lake basins. Sampling from the depths was carried out by a Van Dorn bottle. The chemical analyses were carried out at the Environmental Agency of The Republic of Slovenia, following the standard methods.

The chemical data include measurements of inorganic nutrients important for algae modelling. These include concentrations of all forms of phosphorus (dissolved inorganic and total phosphorus), nitrogen (nitrate, nitrite and total nitrogen) and silica measured in (mass/volume).

The biological data include measurements of six taxonomical groups of phytoplankton and seven species of zooplankton. Their concentrations are measured in number of individu-

Table 1 – Morphological characteristics of lake Bled

	Eastern basin	Western basin	Entire lake
Volume (m^3, $\times 10^6$)	17.5	8.2	25.7
Area (m^2, $\times 10^6$)	0.98	0.49	1.47
Depth, max	24	30	30

Table 2 – Measured data (variables) in lake Bled used for model induction

Variable name	Description	Frequency
q_krivica (m^3/day)	Inflow to the lake	Daily
q_misca (m^3/day)	Inflow to the lake	Daily
q_radovna (m^3/day)	Inflow to the lake	Daily
q_jezernica (m^3/day)	Outflow (at surface)	Daily
q_natega (m^3/day)	Outflow (syphon)	Daily
ps_krivica, ps_misca, ps_radovna (mg/l)	Nutrient (orthophosphate) concentration in the inflows	Monthly
temp (°C)	Water temperature of the streams and lake	Monthly
light (J/(cm^2 day))	Calculated underwater light	Monthly
ps, no, silica (mg/l)	Inorganic nutrients' concentration in the lake (ps is soluble phosphorus and NO is nitrate)	Monthly
phyto (mg DW/l)	Phytoplankton biomass concentration in the lake	Monthly
daph (No. ind/ml or mg DW/l (see text))	Zooplankton (*Daphnia hyalina*) biomass concentration in the lake	Monthly

als per volume unit (No. ind/ml). In order to get compatible measurement units (mass/volume), we have to transform the measurement units to mg of dry weight (DW) per volume unit. While this transformation was already done for the total concentration of phytoplankton (Remec-Rekar, 1995), we used available information from literature and expert estimate to transform the measurement units of zooplankton. Of all the observed zooplankton species, only *Daphnia hyalina* (as most representative zooplankton species) was converted in [mass/volume] units. We estimated the average body length to be 2 mm and calculated the dry weight using the equation suggested by (Dumont et al., 1975). The list of all variables used for modelling is presented in Table 2. Note however that for purposes of modelling phytoplankton change only (i.e., considering zooplankton to be an independent variable), this (approximate) transformation is not really necessary. To avoid it, we used zooplankton as measured (in No. ind/ml), where possible.

3.3. *Data preparation*

3.3.1. *Light, euphotic zone and temperature*

Light was used as averaged daily value for underwater light in the euphotic (illuminated) zone. The depth of the euphotic zone was calculated from the measured transparency in the lake:

$$z_{eu} = 1.7 \cdot \text{transparency} \quad (1)$$

The light extinction factor (k_e in m^{-1}) and the underwater light in the euphotic zone were calculated from the averaged daily global radiation (I), as follows:

$$k_e = \frac{4.6}{z_{eu}} \quad (2)$$

$$\text{PAR} = 0.5 \cdot I \quad (3)$$

$$\text{PAR}(z) = \text{PAR} \cdot e^{-k_e \cdot z} \quad (4)$$

$$\text{light} = \text{avg}(\text{PAR}(z)) \quad (5)$$

where PAR is photosynthetically available radiation, z the water depth and light is depth averaged value for the underwater light in (J/cm^2 day) in the illuminated zone.

Daily water temperature data were obtained by a cubic spline interpolation over the monthly measured data.

3.3.2. *Other data*

As the majority of the measurements were performed on monthly basis we interpolated the daily data by cubic spline interpolation to get a convenient data set of "daily" measurements for induction of differential equations with Lagramge.

4. Experiments

The lake is naturally divided into an eastern and a western basin. According to the measurements the two basins have quite different characteristics and dynamics, which should be considered in the modelling procedure. Our modelling experiments refer to the eastern (bigger) basin and to the upper 10-m zone. No communication between the basins and between the upper and lower (hypolimnion) zone was taken into account.

4.1. *Introducing the expert knowledge to Lagramge*

We introduced modelling knowledge in the process of model discovery at two different levels. At the higher level is the general modelling knowledge about aquatic ecosystems (knowledge library) as described in (Atanasova et al., 2006). The lower level consists of a task specification that includes a list of variables and processes relevant for the modelling of lake Bled. The modelling task for lake Bled was introduced in a form of simple food web concept shown in Fig. 2. It includes three state variables, i.e. inorganic dissolved phosphorus, phytoplankton and zooplankton (*D. hyalina*) and the following processes: inflow/outflow of phosphorus, primary producer growth (PP_growth), predatory loss of phytoplankton (which is equal to the growth of daphnia (Feeds_on)), non-predatory loss of phytoplankton (Respiration_PP, Settling), non-predatory loss of daphnia (Respiration_A) and mortality of daphnia (Mortality_A), which also accounts for daphnia predatory loss.

The knowledge library includes several formulations for each of above listed process classes (Atanasova et al., 2006). For example the process class PP_growth contains five different models for primary producer growth, i.e. exponential, logistic, growth limited by temperature, light and nutrients, growth limited model that accounts for variable optimal temperature, as well as growth limited model that couples the effects of light

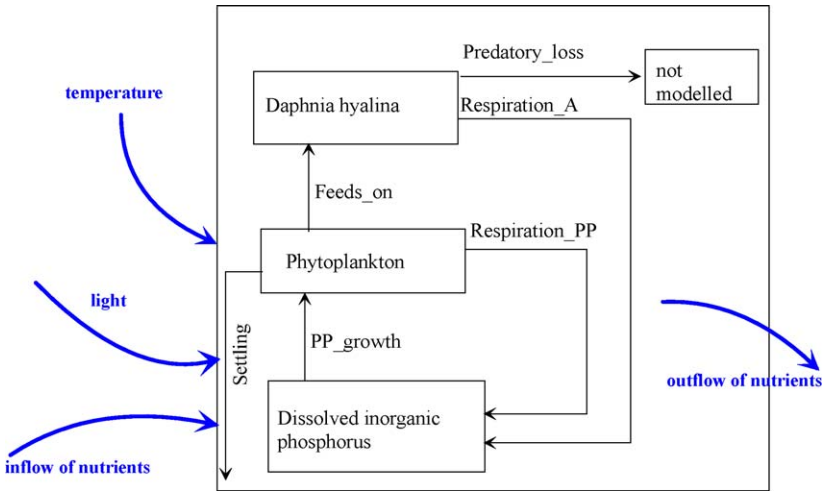


Fig. 2 – Simple conceptual model for lake Bled.

and temperature. Furthermore light, temperature and nutrients limitations are defined as function classes that include several different formulations for each. Thus, we have more than 50 possible formulations for the PP_growth process. Similarly, we have a number of candidate formulations for the rest of the process classes in this lake.

The library of modelling knowledge also specifies how to combine the processes into a corresponding model of the whole system (Džeroski and Todorovski, 2003; Todorovski, 2003; Atanasova et al., 2006). The combining rules in the library support modelling with ordinary differential equations by following the mass conservation principle. More details about this kind of modelling can be found in (e.g. Jørgensen and Bendoricchio, 2001) or (Chapra, 1997)). According to the combining rules from the library, the processes defined in the task specification, will be composed in the following model based on three differential Eqs. (6–8):

$$\frac{d(nut)}{dt} = inflow - outflow + const \cdot Respiration\_PP + const \cdot Respiration\_A - const \cdot PP\_growth \quad (6)$$

$$\frac{d(phyto)}{dt} = PP\_growth - Respiration\_PP - Sedimentation - Feeds\_on \quad (7)$$

$$\frac{d(daph)}{dt} = const \cdot Feeds\_on - Respiration\_A - Mortality\_A \quad (8)$$

4.2. *Discovering models*

First, we made an attempt to discover a model that would describe the long-term behaviour of the lake. For this, we used the task specification, described in the previous section with the only difference being that phyto is the only system variable, while daphnia and phosphorus were considered to be exogenous variables (i.e., forcing functions). Lagramge was then used to discover a specific model following Eq. (7) from the data for years 1995–2001.

Failing to get a very good fit to the long-term data, we conjectured that the lake dynamics changes from year to year. In our second experiment, we aimed at testing this hypothesis, so we applied Lagramge to build separate models for each year data.

In the final experiment, we aimed at discovering a model that includes three system variables (phosphorus, phytoplankton and zooplankton) from 1 year's data (1996). Due to the complexity of space of candidate models and limited computational resources[1] we decided to induce equation for each of the system variables at a time, following the food web hierarchy (phosphorus–phytoplankton–zooplankton). According to the mass balance for inorganic nutrient (see Eq. (6)) following processes defined in the expert task definition were included: inflow of inorganic phosphorus, outflow, release of nutrients due to phytoplankton and zooplankton respiration and loss due to phytoplankton growth.[2] The two processes that influence both the phosphorus and phytoplankton (Eq. (7)) equation are PP_growth and Respiration_pp. Since the discovered equation for phosphorus already fixed the formulation for these processes, we used the same fixed formulation for discovering the phytoplankton equation and only search for appropriate formulation of the other processes involved there (i.e., sedimentation and predatory loss, Feeds_on). Similarly, when the phytoplankton equation is discovered, we used the already discovered formulation of the processes Respiration_A in the phosphorus equation and Feeds_on, in the phytoplankton equation, and let Lagramge find an appropriate formulation for the mortality of daphnia. Note finally that the models induced in the last experiment involve zooplankton measured in [mg DW/l]. Experimental setup is summarized in Table 3.

[1] Note that induction of a single equation with Lagramge takes tens of hours of CPU time on the equipment (Pentium based Linux Platform with 2 GHz processor and 1 GB of RAM) we used for the experiments.

[2] We should point here that the recycling of nutrients goes through more stages (e.g. decomposition of dead organic matter, detritus), which were skipped here in favour of model simplicity.

Table 3 – The experimental setup in lake Bled

	Experiment 1	Experiment 2	Experiment 3
System (target) variables	phyto	phyto	ps, phyto, daph[a]
Independent variables (forcing functions)	temp, light, ps, no, silica, daph	temp, light, ps, no, silica, daph	q_radovna, q_krivica, q_misca, q_jezernica, q_natega, ps_radovna, ps_krivica, ps_misca, temp, light
Training data set(s)	1995–2001	1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002	1996

[a] The model of three differential equations was not discovered simultaneously (see text).

5. Results

5.1. *Long-term phytoplankton model*

Using the task specification from Section 3 and the data series from 1995 to 2001, Lagramge discovered the following model for phytoplankton (Eq. (9)):

$$\frac{d(\text{phyto})}{dt} = \text{phyto}\cdot 0.145\cdot\frac{\text{ps}}{\text{ps}+0.0006}\cdot\frac{\text{silica}}{\text{silica}+0}\cdot\frac{\text{no}}{\text{no}+0.01}\cdot 1.2^{\text{temp}-20}\cdot\frac{\text{light}}{200}\cdot e^{\left(1-\frac{\text{light}}{132}\right)} - \text{phyto}^2\cdot 0.002\cdot\frac{\text{temp}-0.5}{20} - \text{phyto}\cdot\frac{0.28}{10} - \text{daph}\cdot 0.96\cdot\frac{\text{temp}}{6.5}\cdot\frac{\text{phyto}}{\text{phyto}+19}\cdot 0 \quad (9)$$

The first term in Eq. (9) represents the growth process of phytoplankton, which is formulated as nutrient, temperature and light limited. Nutrient limitation is modelled with the Monod term, where phosphorus and nitrate are found as limiting nutrients. Temperature influence on growth is modelled using the exponential adjustment curve, while light limitation on growth is modelled with the photoinhibition curve (Steele, 1965). Respiration of phytoplankton (the second term) is modelled with second-order kinetics, where temperature influence is formulated with the linear response curve. Finally, the last two additive terms in the phytoplankton equation represent the settling process and the process of grazing by zooplankton (daph). According to the model the grazing term equals zero, i.e., grazing has no influence on the phytoplankton dynamics.

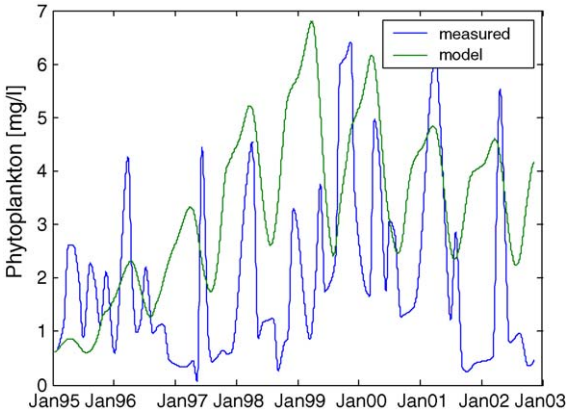


Fig. 3 – Long-term simulation of the phytoplankton model (Eq. (9)).

The comparison of measured and simulated phytoplankton concentration shows a poor fit (Fig. 3) to the training as well as to the testing data set (data from year 2002). There are several possible reasons for this. First, we might need more complex model structure including several alga species and perhaps also the diet preferences of zooplankton. However, due to the limitations of measured data, this hypothesis cannot be properly tested. Second reason might be that the lake dynamics changes through the time.

We can easily test this hypothesis by inducing models from data on individual lake cycles, i.e., calendar years. The test of this hypothesis is the objective of the second experiment.

5.2. *Discovering a phytoplankton model for different years*

In this experiment Lagramge was set to discover phytoplankton models with same basic structure (see Eq. (7)), but for each year data from 1995 to 2002 separately.

Note that the induced models have different formulations of the processes and different parameter values. The growth term in all models is formulated as nutrient, temperature and light influenced process. Nutrient limitation functions for ps, no3 and silica is formulated either with the two variations of the Monod term (i.e., $x/(x+\text{const})$ and $x^2/(x^2+\text{const})$) or with the exponential limiting term (i.e., $1-e^{-\text{const}\times x}$). Note that the smaller values of the constant (also called saturation coefficient) in the Monod terms indicate smaller influence by (nutrient) x on phytoplankton growth. In the limit, a term with saturation coefficient zero, (i.e., $x/(x+0)$), the influence equals one, which means that phytoplankton growth is not limited by x. In contrast with Monod terms, exponential term behaviour is the opposite – larger constant parameter values correspond to smaller influence by x is (the term's value is closer to 1). Following these simple rules, we can interpret the discovered models in terms of the nutrients' influence on the total phytoplankton growth and analyze how this influence changes from year to year.

Table 4 summarizes the types of influences in the induced Eqs. ((15–22) in Appendix A).

Analysis of the nutrient, light and temperature influence on phytoplankton growth shows the following. In 1995, silica was found as the only limiting nutrient, in 1996 and 2001, ps and silica, in 2002 silica and nitrate, while in the period from 1997 to 2000 all of the nutrients were important (limiting) for the phytoplankton growth. It is interesting that in 1997 and 1998

Table 4 – Description of the variables influences on the phytoplankton dynamics equations induced on 1-year data sets (1995–2002)

	Process/term							
	Growth					Respiration temp[a]	Settling temp[a]	Grazing
	Temp[a]	Light[a]	ps[a]	Silica[a]	no[a]			
1995	exp	mon	no	mon	no	lin	lin	No
1996	exp	inh	mon	no	mon	no	no	No
1997	lin	no	mon	mon	mon	exp	no	No
1998	lin	no	exp	mon	exp	exp	no	Yes
1999	exp	inh	mon	mon2	mon	lin	exp	Yes
2000	exp	inh	mon	exp	mon	lin	no	Yes
2001	exp	inh	mon2	exp	no	lin	lin	No
2002	lin	mon	no	mon2	mon2	lin	no	Yes

The influence is described using the following labels: no denotes no influence, yes denotes presence of influence and other labels (exp, lin, mon, mon2) denote the specific influence model (exponential, linear, Monod, second-order Monod, respectively).

[a] Variable.

light was found not to influence the phytoplankton growth. In the models for 1995 and 2002, light influence on growth was modelled with the Monod term (saturation curve), while in 1999–2001 by the photoinhibition curve (Steele, 1965). Temperature influence on growth is modelled either using the linear model (1997, 1998 and 2002) or the exponential one (1995, 1996 and 1999–2001).

Next, we analyzed the influence of respiration on the phytoplankton dynamics. It is modelled with first (1995, 1996, 1998 and 1999) or with second (in 1997 and 2000–2002) order kinetics. The respiration is temperature influenced in all years except for 1996.

Finally, the last two additive terms in the phytoplankton equation represent the settling process and the process of grazing by zooplankton (daph). According to the models in 1995, 1996, 1997 and 2001 the grazing term equals zero, i.e., grazing has no influence on the phytoplankton dynamics.

The simulation of the models is compared with the measurement data in Fig. 4. Note the different starting time of the simulations due to missing data in winter periods (see Section 3). The models perform much better than the one induced on the data from the full time span (see Fig. 3). The goodness of fit is evaluated by the root mean square error (RMSE). The best fit (lowest RMSE) is obtained on the 2002 data, and the worst one on the 1996. A possible reason for poor fit is that we took the nutrient and zooplankton data for granted, instead of treating these two variables as system variables. So, in the last experiment, we aimed at building a complete model of food web in the lake from the 1996 measurements.

In addition, we validated each of the models discovered on specific year on unseen data measured in other years. The validation of almost all models revealed that there is a big discrepancy between the simulated and measured data, which indicates that we deal with a very complex system without yearly repeating patterns. Yet, the model induced on 2002 data shows fairly good performance on the other years (Fig. 5). The model correctly follows the trend of phytoplankton dynamics in all years, except for year 1999. Note also that the model systematically overestimates the spring phytoplankton peaks.

5.3. *Inducing basic food web model*

As already explained in Section 4 this model was discovered gradually, one equation at a time, starting with phosphorus equation, continuing with the discovery of phytoplankton, and daphnia equation.

5.3.1. *Phosphorus equation*

When discovering this model phytoplankton and daphnia are taken as forcing functions (data). Using the task specification from Section 3 and the 1996 daily data Lagramge discovered several good phosphorus models having the form shown in Eq. (6). The one with the lowest error is shown below (10).

$$\begin{aligned}\frac{\mathrm{d(ps)}}{\mathrm{dt}} &= \mathrm{ps\_krivica}\cdot\frac{\mathrm{q\_krivica}}{7\cdot10^6}+\mathrm{ps\_misca}\cdot\frac{\mathrm{q\_misca}}{7\cdot10^6}\\ &+\mathrm{ps\_radovna}\cdot\frac{\mathrm{q\_radovna}}{7\cdot10^6}-\mathrm{ps}\cdot\frac{\mathrm{q\_jezernica}}{7\cdot10^6}\\ &-\mathrm{ps}\cdot\frac{\mathrm{q\_natega}}{7\cdot10^6}+0.0022\cdot\mathrm{phyto}^2\cdot0.072\cdot\frac{\mathrm{temp}-2.7}{20.4-2.7}\\ &+0.07\cdot\mathrm{daph}\cdot0.0026\cdot\frac{\mathrm{temp}}{12.3}-0.0023\cdot\mathrm{phyto}\cdot0.21\\ &\cdot\frac{\mathrm{ps}}{\mathrm{ps}+0.00042}\cdot\frac{\mathrm{temp}}{16.7}\cdot\frac{\mathrm{light}}{170}\cdot\mathrm{e}^{\left(1-\frac{\mathrm{light}}{170}\right)}\end{aligned}\tag{10}$$

The first five terms in the equation represent the inflow and outflow of inorganic phosphorus. Next term, phytoplankton respiration (i.e. release of phosphorus due to phytoplankton respiration), is formulated as second-order kinetics, while daphnia respiration as first-order kinetics. Both processes are temperature influenced. Growth of phytoplankton, i.e., consumption of phosphorus by phytoplankton is modelled as temperature, light and nutrient limited growth. Nutrient limitation is modelled with Monod expression, light limitation with the photoinhibition curve (Steele, 1965) and temperature influence with linear response curve.

5.3.2. *Phytoplankton equation*

In the phytoplankton mass balance Eq. (7) following processes (Eqs. (11 and 12)) are already discovered in the phosphorus

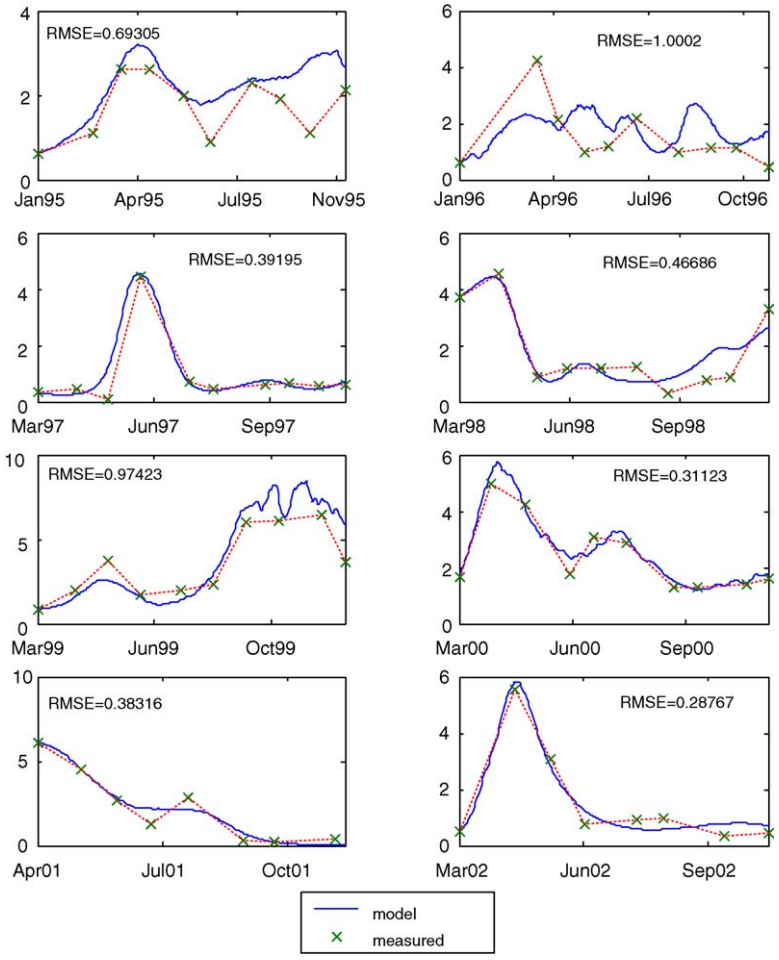


Fig. 4 – Performance of the phytoplankton models, discovered separately for each year data. The goodness of fit of each model is evaluated by the root mean square error (RMSE).

equation:

$$\text{PP\_growth} = \text{phyto} \cdot 0.21 \cdot \frac{\text{ps}}{\text{ps} + 0.00042} \cdot \frac{\text{temp}}{16.7} \cdot \frac{\text{light}}{170} \cdot \text{e}^{\left(1 - \frac{\text{light}}{170}\right)} \tag{11}$$

$$\text{Respiration\_PP} = \text{phyto}^2 \cdot 0.072 \cdot \frac{\text{temp} - 2.7}{19.7 - 2} \tag{12}$$

According to this we set Lagramge to discover the rest of the processes in the phytoplankton equation, i.e., sedimentation and Feeds_on (or grazing by daphnia). The best phytoplankton model using the growth and respiration terms from the phosphorus model is shown below (13).

$$\begin{aligned}\frac{\text{d(phyto)}}{\text{dt}} = {} & \text{phyto} \cdot 0.21 \cdot \frac{\text{ps}}{\text{ps} + 0.00042} \cdot \frac{\text{temp}}{16.7} \cdot \frac{\text{light}}{170} \\ & \cdot \text{e}^{\left(1 - \frac{\text{light}}{170}\right)} - \text{phyto}^2 \cdot 0.072 \frac{\text{temp} - 2.7}{19.7 - 2} \\ & - \text{phyto} \cdot \frac{0.5}{10} \cdot \frac{\text{temp} - 2}{18 - 4} - \text{daph} \cdot 0.5 \cdot \frac{\text{temp} - 2.6}{18 - 4} \\ & \cdot (1 - \exp(-0.58 \cdot \text{phyto})) \cdot 0.56 \cdot \text{phyto}\end{aligned} \tag{13}$$

Sedimentation is formulated as temperature influenced, with sedimentation rate 0.5 m/day. The grazing term is formulated using the filtration coefficient (0.5 l/(mg day)), linear temperature response curve and exponential term for food limitation on daphnia growth.

Comparison of the phytoplankton equation with the one from the previous section (Eq. (16) in Appendix A) shows the differences in practically all process' formulations. The phytoplankton growth is limited by phosphorus concentration and temperature, while the growth in Eq. (16) is limited by two nutrients (phosphorus and silica). Note also the difference in the temperature influence terms. Also, respiration is formulated with second-order kinetics (first-order in Eq. (16)) and sedimentation (third term) is temperature influenced (temperature independent in Eq. (16)). Finally, the most important difference between two models is that grazing influence is important for the phytoplankton dynamics, unlike the previous experiment where the grazing term equals zero.

Considering that this model has better performance (i.e., lower RMSE, see Figs. 6 and 4), it is a bit of surprise that Lagramge could not find a suitable set of parameters in the previous experiment. Obviously, the change of daphnia units pushed the parameters' values in a range where the optimization method is unsuccessful. Note that in this experiment

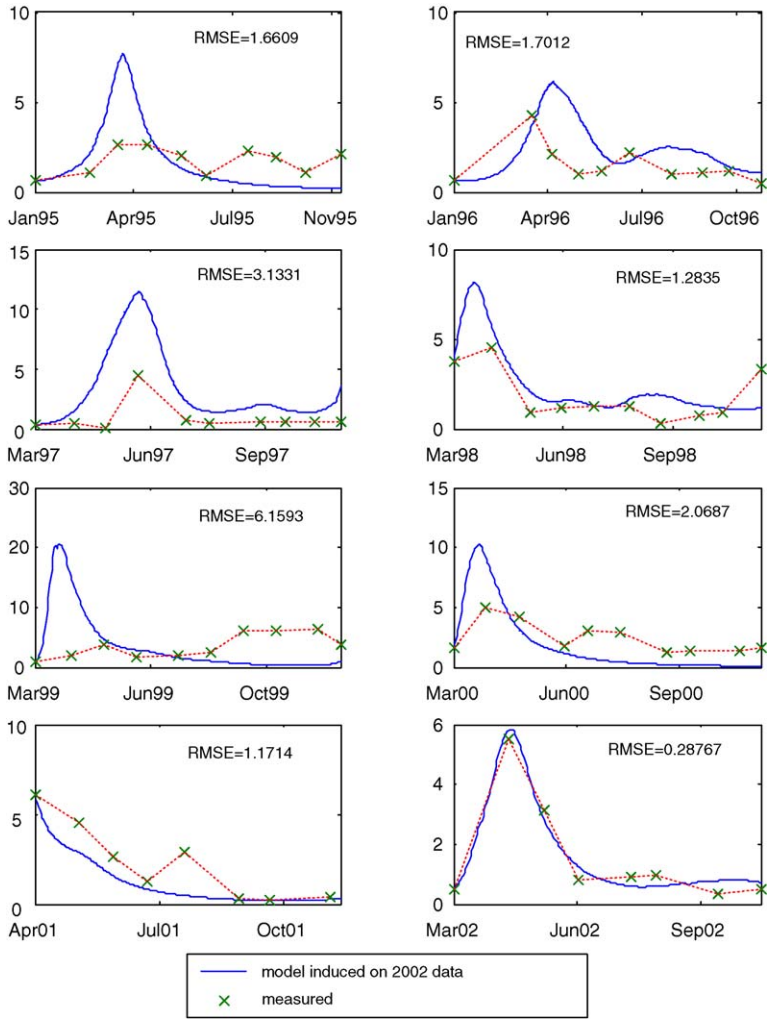


Fig. 5 – Validation of the phytoplankton model induced from 2002 data on the rest of the simulation period.

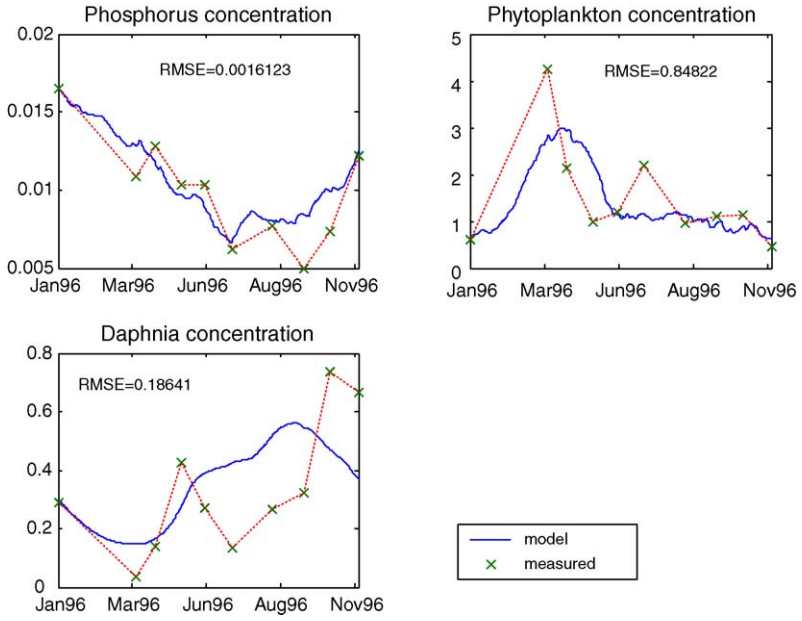


Fig. 6 – Performance of the food web model for phosphorus, phytoplankton and daphnia.

daphnia is expressed in [mg DW/l], while in the previous in [No ind/ml].

5.3.3. *Zooplankton (daphnia) equation*

Zooplankton equation contains the following processes: Feeds_on, Respiration_A and Mortality_A. Feeds_on correspond to grazing of phytoplankton and therefore was already discovered in the phytoplankton equation, while Respiration_A is already discovered in the phosphorus equation. The task here is to find suitable mortality process for daphnia, which is a closure term for the model. The equation with lowest error is presented in (14). Lagramge found the hyperbolic term as the most suitable closure term for the model.

$$\frac{\mathrm{d(daph)}}{\mathrm{dt}} = 0.14 \cdot \mathrm{daph} \cdot 0.5 \cdot \frac{\mathrm{temp} - 2.6}{18 - 4} \cdot (1 - \exp(-0.58 \cdot \mathrm{phyto})) \cdot 0.56 \cdot \mathrm{phyto} - \mathrm{daph} \cdot 0.026 \cdot \frac{\mathrm{temp}}{12.3} - 0.01 \cdot \frac{\mathrm{daph}^2}{0.001 + \mathrm{daph}} \quad (14)$$

Thus, the complete basic food web model for phosphorus, phytoplankton and daphnia consists of Eqs. (10), (13) and (14). Simulation of the model is shown in Fig. 6.

Though the model was successfully calibrated on 1996 data it shows problems when validating it on the other years' data sets. This behaviour yet again confirms our hypothesis that the lake dynamics changes yearly.

6. Conclusions and further work

Discovering a model based on ordinary differential equations that covers long-term behaviour of such a complex system is very difficult task, mostly due to the system complexity and constantly changing patterns of the real system behaviour through time. Another issue is the complexity of the computational method itself, which is strongly limited by the present computational resources used in this research. Therefore, we limited our task on either discovering equation for one state variable instead of discovering the complete model simultaneously, or discovering a complete model with strong limitations on the search space of candidate models.

The aim of discovering equation for one state variable instead of complete model of the system is above all to search connections and patterns among data. In our case the relationships were prescribed by domain expert – therefore they are transparent and understandable to experts. According to our expectations, we could not find a single model that will be suitable for such complex system over the whole time period. Comparison of the measurement data from different years shows that lake Bled usually has two peaks of algae bloom each year (Spring and Autumn) caused by different algae species. However, the situation in certain years can also be different: in 1995 we can notice three algal blooms, while in 1997 and 2002, we observe only one. On the other hand, modelling 1-year behaviour is a reasonable task, as it is evident from Fig. 4.

For 1996, Lagramge successfully discovered a complete three-equation food web model, though the search space of candidate models was strictly limited and controlled. The chosen induction order that follows the food web hierarchy is only one possible and plausible order used here as a heuristic. However, further experiments are required to evaluate the influence of induction ordering on the obtained model.

Phosphorus and phytoplankton equations were discovered very successfully as evident from the simulations in Fig. 6. Discovering daphnia equation was more difficult task due to following reasons: (1) the conversion of data from number of individuals to biomass was approximated with literature data and (2) daphnia predation by fish was not modelled. Daphnia mortality was the closure term in the model. Of the four animal loss terms defined in the library, i.e. first-order kinetics, second-order kinetics, hyperbolic and sigmoid form, Lagramge found the hyperbolic form as the most suitable one. Simulation of the complete model indicates good fit with the measurements for phosphorus and phytoplankton, while for daphnia, the model only captures the trend and not the daphnia dynamics.

In order to find a complete model that will cover the long-term behaviour of the lake, several investigations still need to be done, to confirm some assumptions that emerged during this research. The first assumption is that the lake has dynamic structure and therefore we cannot model it with a single model with constant parameters. This assumption was partly confirmed in the second experiment when Lagramge successfully discovered phytoplankton models, which were trained for each year separately (Eqs. (15–22) in Appendix A). The structure of all models was defined by expert (in the task specification) to have four processes, i.e. growth, respiration, sedimentation and grazing. To our expectations discovered models differ in the processes' formulations and in their parameters values, which indicates the structural dynamicity of the lake, i.e. the system's structure is different from year to year.

Our second assumption is that the concept that we used for discovering models is too simple to cover long-term behaviour of the lake. It is necessary to increase the complexity of the concept by introducing the algal functional groups as state variables. For this we need more detailed insight into the lake food web and some more expert knowledge. In order to accomplish such demanding task we also need faster computational resources.

Finally, our last assumption is that improvement of the optimisation method for simultaneous multiple parameter estimation would result in better models even by using the current simple concept.

Acknowledgements

The data for this research were provided by the Ministry for Environment, Spatial Planning and Energy, Environmental Agency of The Republic of Slovenia. First author would like to acknowledge the Ministry of Education, Science and Sport for 2-year junior researcher grant No. 3311-02-831/433. We also acknowledge the support of the ECOGRID project, funded by the Slovenian Research Agency, under contract No. 3311-04-828125. Finally, thanks to the reviewers for their constructive comments on an earlier version of the manuscript.

Appendix A

1995:

$$\frac{d(phyto)}{dt} = phyto \cdot 0.25 \cdot \frac{ps}{ps+0} \cdot \frac{silica}{silica+0.34} \cdot \frac{no}{no+0} \cdot 1.12^{(temp-20)} \cdot \frac{light}{light+45.2} - phyto \cdot 0.07 \cdot \frac{temp-4.2}{15.5-4.4} - phyto \cdot \frac{0.0002}{10} \cdot \frac{temp-0}{15-5} - daph \cdot 10^{-5} \cdot \frac{temp}{18} \cdot \frac{phyto^2}{phyto^2+19.8} \cdot phyto \cdot 0 \quad (15)$$

1996:

$$\frac{d(phyto)}{dt} = phyto \cdot 5 \cdot \frac{ps}{ps+0.0024} \cdot \frac{silica}{silica+0.19} \cdot \frac{no}{no+0} \cdot 1.11^{(temp-15)} \cdot \frac{light}{100} \cdot e^{\left(1-\frac{light}{100}\right)} - phyto \cdot 0.032 - phyto \cdot \frac{0.31}{10} - daph \cdot 10^{-5} \cdot \frac{temp}{7.8} \cdot \frac{phyto^2}{phyto^2+4.8} \cdot phyto \cdot 0 \quad (16)$$

1997:

$$\frac{d(phyto)}{dt} = phyto \cdot 5.04 \cdot \frac{ps}{ps+0.0005} \cdot \frac{silica^2}{silica^2+0.688} \cdot \frac{no^2}{no^2+15} \cdot \frac{temp-5}{16.4-4.3} \cdot \frac{light}{light+0} - phyto^2 \cdot 0.058 \cdot 1.13^{(temp-16)} - phyto \cdot \frac{0.43}{10} - daph \cdot 0.0063 \cdot \frac{temp}{18} \cdot \frac{phyto^2}{phyto^2+11.7} \cdot phyto \cdot 0 \quad (17)$$

1998:

$$\frac{d(phyto)}{dt} = phyto \cdot 4.66 \cdot (1-e^{-4.5 \cdot ps}) \cdot \frac{silica}{silica+0.15} \cdot (1-e^{-3.3 \cdot no}) \cdot \frac{temp}{10.2} \cdot \frac{light}{light+0} - phyto \cdot 0.32 \cdot 1.12^{(temp-17.7)} - phyto \cdot \frac{0.5}{10} - daph \cdot 15 \cdot \frac{temp}{15-5} \cdot \frac{phyto^2}{phyto^2+20} \cdot phyto \cdot 0.79 \quad (18)$$

1999:

$$\frac{d(phyto)}{dt} = phyto \cdot 0.55 \cdot \frac{ps}{ps+0.033} \cdot \frac{silica^2}{silica^2+0.02} \cdot \frac{no}{no+0.073} \cdot 1.11^{(temp-18.5)} \cdot \frac{light}{140} \cdot e^{\left(1-\frac{light}{174.9}\right)} - phyto \cdot 0.092 \cdot \frac{temp-0.4}{19.4-1.5} - phyto \cdot \frac{0.09}{10} \cdot 1.11^{(temp-15.3)} - daph \cdot 10^{-5} \cdot \frac{temp}{15-3.4} \cdot \frac{phyto}{phyto+2.1} \cdot phyto \cdot 0.37 \quad (19)$$

2000:

$$\frac{d(phyto)}{dt} = phyto \cdot 1.9 \cdot \frac{ps}{ps+0.0017} \cdot (1-e^{-0.11 \cdot silica}) \cdot \frac{no^2}{no^2+0.012} \cdot 1.13^{(temp-15)} \cdot \frac{light}{100} \cdot e^{\left(1-\frac{light}{116.3}\right)} - phyto^2 \cdot 0.004 \cdot \frac{temp}{5.2} - phyto \cdot \frac{0.23}{10} - daph \cdot 0.52 \cdot \frac{temp}{9.1} \cdot \frac{phyto}{phyto+3.9} \cdot phyto \cdot 0.72 \quad (20)$$

2001:

$$\frac{d(phyto)}{dt} = phyto \cdot 9.3 \cdot \frac{ps^2}{ps^2+0.09} \cdot (1-e^{-5.15 \cdot silica}) \cdot \frac{no}{no+0} \cdot 1.13^{(temp-15)} \cdot \frac{light}{114.4} \cdot \exp\left(1-\frac{light}{197.9}\right) - phyto^2 \cdot 0.0005 \cdot \frac{temp}{2.4} - phyto \cdot \frac{0.5}{10} \cdot \frac{temp-4.7}{15-5} - daph \cdot 0.33 \cdot \frac{temp}{15-5} \cdot \frac{phyto}{phyto+0.26} \cdot phyto \cdot 0 \quad (21)$$

2002:

$$\frac{d(phyto)}{dt} = phyto \cdot 9.4 \cdot \frac{ps}{ps+0} \cdot \frac{silica^2}{silica^2+15} \cdot \frac{no^2}{no^2+10} \cdot \frac{temp}{5.7} \cdot \frac{light}{light+41} - phyto^2 \cdot 0.0054 \cdot \frac{temp}{5.1} - phyto \cdot \frac{0.05}{10} - daph \cdot 10^{-5} \cdot \frac{temp-2}{19.8-2.6} \cdot \frac{phyto}{phyto+17.3} \cdot phyto \cdot 0.15 \quad (22)$$

REFERENCES

Atanasova, N., Todorovski, L., Džeroski, S., Kompare, B., 2006. Constructing a library of domain knowledge for automated modelling of aquatic ecosystems. Ecol. Model. 194, 14–36.

Chapra, S.C., 1997. Surface Water-Quality Modeling. McGraw-Hill (0-07-011364-5).

Dumont, H.J., Van de Velde, I., Dumont, S., 1975. The dry weight estimate of biomass in a selection of cladocera, copepoda and rotifera from the plankton, periphyton and benthos of continental waters. Oecol. (Berl.) 19, 75–97.

Džeroski, S., Todorovski, L., 2003. Learning population dynamics models from data and domain knowledge. Ecol. Model. 170 (2–3), 129–140.

Imboden, D., 1974. Phosphorus model of lake eutrophication. Limnol. Oceanogr. 19, 297–304.

Jørgensen, S.E., Bendoricchio, G., 2001. Fundamentals of Ecological Modelling. Elsevier, ISBN 0-080-44028-2.

Kompare, B., 1995. The Use of Artificial Intelligence in Ecological Modelling. University of Ljubljana, Faculty of Civil and Geodetic Engineering, Royal Danish School of Pharmacy, Ljubljana, Copenhagen.

Kompare, B., Džeroski, S., Karalic, A., 1997. Identification of the Lake Bled ecosystem with the artificial intelligence tools M5 and FORS. In: Fourth International Conference on Water Pollution, p. 798.

Remec-Rekar, S., 1995. Življenska Strategija in Absorbcija Fosforja pri Nekaterih Fitoplanktonskih vrstah Blejskega jezera-123. University of Ljubljana, Ljubljana.

Rismal, M., 1980. Presoja posameznih metod za sanacijo Blejskega jezera. Gradbeni Vestnik 29 (2–3), 34–46.

Rismal, M., Kompare, B., Rajar, R., 1997. Contribution of hydrodynamic and limnological modelling to the sanitation of Lake Bled. In: Fourth International Conference on Water Pollution, p. 139.

Sketelj, J., Rejic, M., 1958. Preliminary account on the examination of Lake Bled. Gradbeni Vestnik, 61–64.

Steele, J., 1965. Notes on some theoretical problems in production ecology. In: Goldman, C. (Ed.), Primary Production in Aquatic Environments. University of California Press, Berkeley, California, pp. 393–398.

Todorovski, L., 2003. Using domain knowledge for automated modeling of dynamic systems with equation discovery. In: Fakulteta Zaraèunalništvo in Informatiko. University of Ljubljana, Ljubljana, Slovenia.

Todorovski, L., Džeroski, S., 1997. Declarative bias in equation discovery. In: 14th International Conference on Machine Learning, p. 384.

Vollenweider, R.A., 1968. The Scientific Basis of Lake and Stream Eutrophication with Particular Reference to Phosphorus and Nitrogen as Eutrophication Factors. Organisation for Economic Cooperation and Development, Paris.