Environmental Modelling & Software xxx (2010) 1–11

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

Automated discovery of a model for dinoflagellate dynamics

Nataša Atanasova [a,*], Sašo Džeroski [b], Boris Kompare [a], Ljupčo Todorovski [c], Gideon Gal [d]

[a] Faculty of Civil and Geodetic Engineering, University of Ljubljana, Hajdrihova 28, SI-1000 Ljubljana, Slovenia
[b] Department of Knowledge Technologies, Jožef Stefan Institute, Jamova 39, SI-1000 Ljubljana, Slovenia
[c] Faculty of Administration, University of Ljubljana, Gosarjeva 5, SI-1000 Ljubljana, Slovenia
[d] Yigal Alon Kinneret Limnological Laboratory, IOLR, PO Box 447, Migdal 14950, Israel

ARTICLE INFO

Article history:
Received 2 April 2010
Received in revised form
5 November 2010
Accepted 12 November 2010
Available online xxx

Keywords:
Dinoflagellate
Lake Kinneret
Dynamic modelling
Machine learning
Automated modelling

ABSTRACT

The aim of this paper is to discover a model equation for predicting the concentration of the algal species *Peridinium gatunense* (Dinoflagellate) in Lake Kinneret. This is a rather difficult task, due to the sudden ecosystem changes that occurred in the mid-1990s. Namely, the stable ecosystem (with regular *Peridinium* blooms until 1993) underwent changes and has transformed into an unstable system, with cyanobacterial blooms now occurring regularly. This shift in the algal succession is expected to influence attempts to model the lake ecosystem. Namely, the model structure before and after the change is likely to be different. Our modelling experiments were directed to discover a single model equation that can simulate dinoflagellate dynamics in both periods. We apply an automated modelling tool (Lagramge), which integrates the knowledge- and the data-driven modelling approach. In addition we include an expert visual estimation of the models discovered by Lagramge to assist in the selection of the optimal model. The dataset used included time-series measurements of typical data from the periods 1988 to 1992 and 1997 to 1999. Using the data and expert knowledge coded in a modelling knowledge library, Lagramge successfully discovered several suitable mathematical models for *Peridinium*. After the expert's visual estimation and validation of the models, we propose one optimal model capable of long-term predictions.

1. Introduction

Lake Kinneret is the only natural freshwater lake in Israel, not only providing some 30% of the country's drinking water but also serving as a key recreational site. Routine monitoring of the lake ecosystem has been conducted since 1969 and until 1993 the ecosystem exhibited noticeable stability in its key characteristics (Berman et al., 1995). One key element was the annual spring bloom of the large dinoflagellate, *Peridinium gatunense* (Zohary, 2004). *P. gatunense*, through to the early 1990s was the dominant algal species in the lake often reaching over 90% of the algal biomass. Since the mid-1990s, however, a change has occurred in the lake leading to the appearance of cyanobacterial blooms and years in which no *Peridinium* bloom was detected (Zohary, 2004). Hence, unlike the period prior to 1994, it is no longer possible to predict which species will bloom or when. The reasons for the shift in algal succession are still unclear as are the reasons that have led to observed large inter-annual variation in *Peridinium* biomass (Roelke et al., 2007).

Modelling of such a system can be very useful but also a very difficult task. Mechanistic ecological model, such as DYRESM-CAEDYM have been applied to simulate the seasonal dynamics of nutrients, and multiple phytoplankton and zooplankton groups (Bruce et al., 2006). The model is fairly complex and, as most mechanistic ecological models, structurally fixed and cannot react to changes in the ecosystem structure.

Complex models are difficult to cope with, in particular when attempting to perform long-term simulations. On the other hand, no matter how complex the model is, it is still a rough simplification of the reality. This is incorporated in the model's parameters. Estimating the values of the model's parameters is another, even more difficult task since most of the time we are dealing with over-parameterised models. This means that we do not have a single (unique) set of parameters' values. There are therefore clear advantages to simplifying mathematical models as much as possible. In this context, Jakeman et al. (2006) suggested a 10-step procedure for the development of good models. Applications of this procedure can be found in Robson et al. (2008) and in Welsh (2008). Crout et al. (2009) suggest a methodology for model simplification (reduction of complexity), which is an extension to the previously suggested one

* Corresponding author. Tel.: +386 1 4217481; fax: +386 1 2519897.
E-mail addresses: natanaso@fgg.uni-lj.si (N. Atanasova), saso.dzeroski@ijs.si (S. Džeroski), ljupco.todorovski@fu.uni-lj.si (L. Todorovski), gal@ocean.org.il (G. Gal).

1364-8152/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.envsoft.2010.11.003

by Cox et al. (2006). Applying this methodology on real examples of existing process based models demonstrated that all of the models could be used in their reduced versions.

In this paper we are not dealing with simplification of existing models but rather with the direct discovery of the optimal (already simplified) model structure. To do so, we apply an automated knowledge discovery tool, called Lagramge (Džeroski and Todorovski, 2003) in order to determine the optimum model structure that can be used for conceptual modelling of the *Peridinium* population. Lagramge is a machine learning (ML, hereafter) algorithm based on equation discovery. But unlike the majority of the ML algorithms that are purely data-driven, i.e. they induce models from measured data only, Lagramge is capable of including background expert knowledge in the procedure of model induction from data. Domain knowledge is introduced in the form of generic processes. Todorovski (2003) developed formalism for encoding process based domain knowledge into a modelling library. Using the developed formalism, Atanasova et al. (2006a) elaborated a knowledge library for modelling of lake ecosystems. Using the library and the automated modelling method, Lagramge has constructed models of several real-world domains, e.g. Lake Bled, Slovenia (Atanasova et al., 2006b), Lake Kasumigaura, Japan (Atanasova et al., 2006c), Lake Glumsø, Denmark (Atanasova et al., 2008), and Lagoon of Venice, Italy (Atanasova, 2005).

The goal of this research is to discover an optimal model equation for *Peridinium* dynamics, which can (a) provide some indications as to which factors should be included in our models, and how to include them, i.e. suggest the necessary model complexity, and (b) suggest whether a single model structure (and set of parameters' values) is sufficient to model the *Peridinium* dynamics prior to, and after, the algal shift in the lake. In order to address these issues we model the *Peridinium* population prior to, and following, the changes that occurred in the lake in the early 1990s. Specifically, we compare the periods 1988–1992 and 1997–1999.

The remainder of this paper is organised as follows. In the next section we explain the automated modelling method Lagramge used for model discovery from measured data and the modelling background knowledge. Section 3 describes the Lake Kinneret dataset. Experimental setup for *Peridinium* model discovery is given in Section 4, followed by results and discussion in Sections 5 and 6. Finally, the conclusions are presented in Section 7.

2. Automated modelling with Lagramge

In this section, we present the method Lagramge, used for automated model discovery from both, measured data and background knowledge. We first explain the motivation for using Lagramge compared to the theoretical (or conceptual) approach to modelling, than we explain how Lagramge works, i.e., how it introduces the background knowledge (stored in a modelling knowledge library) into the procedure for model discovery from data, and finally we present a segment of the modelling knowledge library used in this particular case study.

2.1. Motivation for using the automated modelling method Lagramge

Lagramge is a ML tool based on equation discovery. Unlike the majority of the ML algorithms that discover models from measured data only, which in many cases results in black-box or semi-transparent models, Lagramge uses background domain knowledge to drive the procedure of model discovery from data. Thus, modelling with Lagramge is much closer to the theoretical or conceptual approach to modelling and is thus considered a hybrid approach to modelling.

In Fig. 1, we present a comparison between theoretical (conceptual) modelling and hybrid modelling with Lagramge. Theoretical modelling consists of many modelling steps (Fig. 1a), typically starting with conceptual modelling (elaborating a basic concept of the system by selecting state variables and processes that affect those variables), formulating suitable mathematical expressions for the conceptual model and testing the goodness of those selections (the conceptual and mathematical models) against measured data by simulating the mathematical model. These steps are performed iteratively. The modeller constantly adapts/changes the conceptual and/or mathematical model and/or mathematical model's parameters and performs the rest of the steps for each change. There are three important issues, which determine the number of repetitions of the modelling steps. (1) Has the modeller selected the correct conceptual model? (2) Has the modeller selected the correct mathematical structure for the selected concept? and, (3) the parameter estimation procedure.

These three issues can be addressed systematically using an automated modelling methodology, such as Lagramge, which is based on an equation discovery method. In contrast to the knowledge based (theoretical) approach, Lagramge performs all of the modelling steps automatically, except for the conceptual modelling. However, we can feed Lagramge with several conceptual models, as parallel runs of the software are possible, and perform the mathematical model development in a single iteration for each conceptual model (see Fig. 1b).

2.2. How Lagramge works

For a given conceptual model, Lagramge discovers a mathematical model, i.e., structure and its parameters' values based on (1) a knowledge library, where general modelling knowledge is encoded, (2) modelling task specification, which corresponds to a conceptual model of the system, where the user specifies important variables and processes that take place in the observed system, and, (3) time-series data of the measurements of the specified variables.

After reading the modelling task specification and the measurements, Lagramge performs a heuristic search through the set of candidate model structures composed following the knowledge encoded in the library. In particular, Lagramge composes a list of specific mathematical model structures that can be used to model the processes specified in the task specification, i.e., correspond to the given conceptual model. Lagramge can process this list of candidate models following two search strategies. When using exhaustive search strategy, each structure is being evaluated against time-series data. When dealing with complex conceptual models that lead to huge lists of candidate models, an alternative search strategy of beam search can be used, where only a portion of model structures is heuristically selected and evaluated.

For each candidate structure considered during the search, Lagramge uses non-linear optimization methods to fit the values of the model parameters against data. Parameter values are selected that minimize the discrepancy between model simulation and observed time-series data, using mean squared error (MSE) to measure the discrepancy. In addition, heuristic function for model selection can be based on the minimum description length (MDL) principle, which takes into account model complexity and introduces a preference towards simpler models. The model with the lowest value of the selected heuristics function (MSE or MDL) is considered as the best model for a given conceptual model (task specification) and dataset.

Let us illustrate the use of Lagramge on a simple example (Fig. 2). Suppose we have observations on phytoplankton (*phyto*) and zooplankton (*zoo*) in a lake. The task is to formulate a mathematical model for the observed variables. First, we specify

Please cite this article in press as: Atanasova, N., et al., Automated discovery of a model for dinoflagellate dynamics, Environmental Modelling & Software (2010), doi:10.1016/j.envsoft.2010.11.003

a conceptual model of the system. Since we do not have any other measurements we specify a very simple one, composed of two state variables (phyto and zoo) and three processes (growth, grazing, and loss), which influence the dynamics of the state variables (Fig. 2). Following the algorithm presented in Fig. 1b, Lagramge discovered a model that fits the observed data best. As evident from Fig. 2, Lagramge discovered the Lotka-Volterra's predator–prey model.

2.3. The modelling knowledge library

Currently the modelling knowledge library used by Lagramge supports lake modelling with ordinary differential equations (ODEs). The knowledge coded in the library includes a large number of bio-chemical and physical processes' formulations that can be used for deriving (inducing) known lake models. Models of different complexity can be derived from the library, such as the simple Vollenweider's (1968) model or the fairly complex model SALMO (Bendorf, 1979; Recknagel, 1980).

The generic conceptual model for modelling aquatic ecosystems, which is encoded in the knowledge library, is presented in Fig. 3. The boxes represent types of state variables, whereas the arrows indicate the bio-chemical and physical processes that influence the state variables. For example the dynamics of a primary producer can be affected by several processes, e.g., growth, respiration, mortality, excretion, sedimentation, and grazing. We refer the reader to Atanasova et al. (2006a) for more details. Part of the modelling knowledge library applied to a specific case (lake) can be expanded or reduced, based on the number of variables of each type we observe (model) in the ecosystem.

In the following sections, we present a segment of the modelling knowledge coded in the library that was used in our case study. Four basic processes, i.e. growth of primary producers, respiration, non-predatory mortality, and grazing were used to model the algal dynamics in our case.

2.3.1. Growth of primary producers

In general, the growth of a primary producer can be stated as:

$$\text{growth\_pp} = \mu \cdot \text{pp} \tag{1}$$

where pp is a primary producer concentration in [mass/volume] and μ is the primary producer growth rate in [1/time].

The library supports three general formulations of the growth process according to the growth rate assumptions: (1) the exponential model, which assumes a constant growth rate or unlimited growth, (2) the logistic model, which Verhulst (1845) suggests that the population growth is limited, i.e., it may depend on population density, and (3) a model which accounts for growth, limited by several factors, e.g., light, temperature and nutrients (such as the

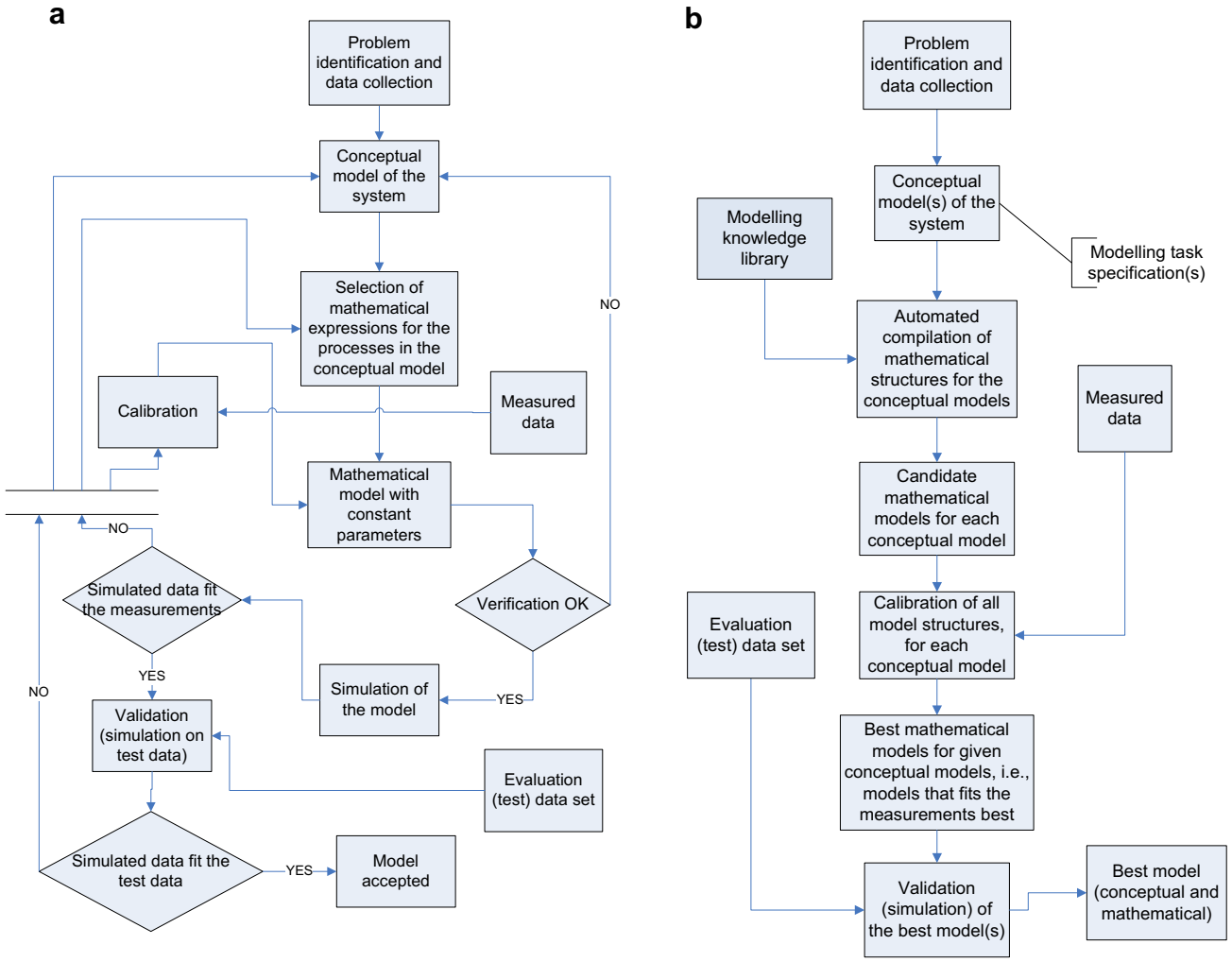

Fig. 1. Comparison of two approaches to process based modelling. (a) A knowledge based approach, and (b) a hybrid approach, combining the knowledge based and data-driven approaches.

Please cite this article in press as: Atanasova, N., et al., Automated discovery of a model for dinoflagellate dynamics, Environmental Modelling & Software (2010), doi:10.1016/j.envsoft.2010.11.003

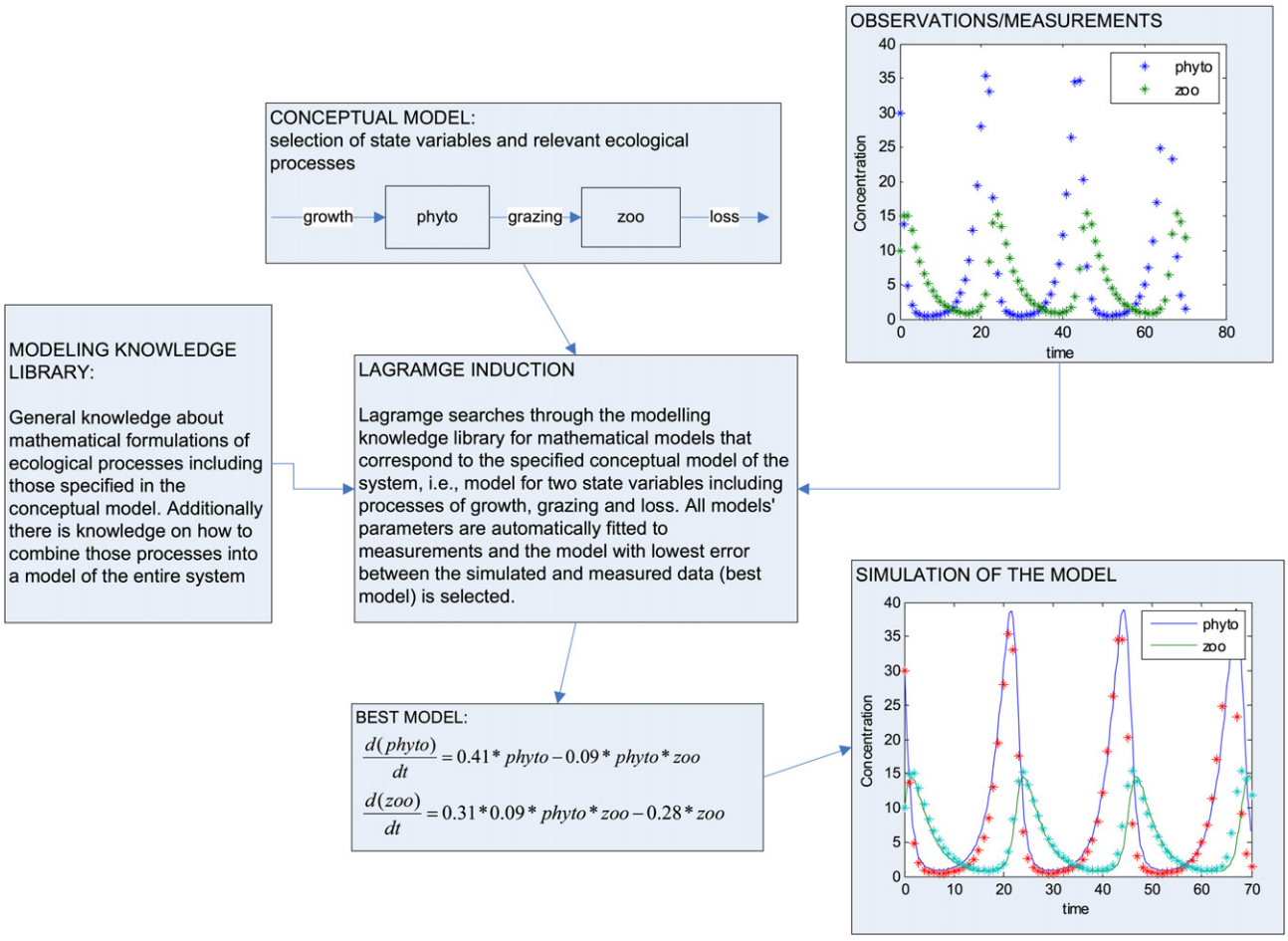

Fig. 2. Model induction with Lagramge: Simple example of predator–prey interaction. At input Lagramge takes the modelling knowledge library, the conceptual model of the observed ecosystem (i.e. user's specification of state variables and processes), and measurements of the state variables. At output Lagramge gives the best fitted model (identification of structure and parameters) to the measurements.

Droop model). The limiting factors or functions can be implemented differently in the expression for the growth rate. The most common formulation of growth rate is, however, a product of the limiting functions of temperature, light and nutrients. The growth rate therefore becomes:

$$\mu = \mu_{\max}\left(T_{\text{ref}}\right) \cdot f_1(T) \cdot f_2(L) \cdot f_3(N_1, N_2, N_3) \tag{2}$$

where $\mu_{\max}(T_{\text{ref}})$ is the maximum growth rate at the reference temperature (T_{ref}) and optimal conditions (in terms of food, temperature, and light), $f_1(T)$ is the temperature adjustment of the growth rate, $f_2(L)$ is the light limitation on the growth rate, and $f_3(N_1, N_2, N_3)$ models the nutrients limitation on growth. N_1, N_2, and N_3 are the limiting nutrients for the primary producer growth, such as phosphorus, carbon, and nitrogen.

The library contains six different expressions for the temperature function (two linear functions, an exponential function, and three optimal temperature functions), two functions for light limitation (Monod function and light inhibition curve) and three functions for nutrient limitation (Monod, Monod^2, and exponential function). For more details, see Atanasova et al. (2006a).

2.3.2. Respiration

Loss of mass due to respiration can be generally written as a first order (sometimes also second order) equation, where the rate coefficient is a function of temperature and/or the physiological conditions of the organism. Formulations included in the library for primary producers' respiration rates (r) are listed in Table 1, where r is included in the general form presented in Eq. (3) of the loss due to respiration.

$$\frac{\text{dpp}}{\text{d}t} = -r \cdot \text{pp} \tag{3}$$

2.3.3. Mortality

Non-predatory mortality of primary producers includes processes like senescence, bacterial decomposition of cells (parasitism), stress-induced mortality, severe nutrient deficiencies, extreme environmental conditions, or toxic substances. Commonly, this process is modelled when no other loss process is included. The mortality rate can be formulated as a constant or a temperature adjusted rate. Some models relate the mortality rate to the physiological conditions of the algal cells. The different formulations for the primary producers' mortality rates, included in the knowledge library, are shown in Table 2, where the general process equation is of the same form as in Eq. (3).

2.3.4. Grazing by secondary producers

Two formulations of the grazing process are included in the library. The first one uses an ingestion rate coefficient and the second uses a filtration rate, which states the amount of water filtered per unit of zooplankton per time, since most of the zooplankton are filter feeders. The two formulations are presented in Eqs. (4) and (5),

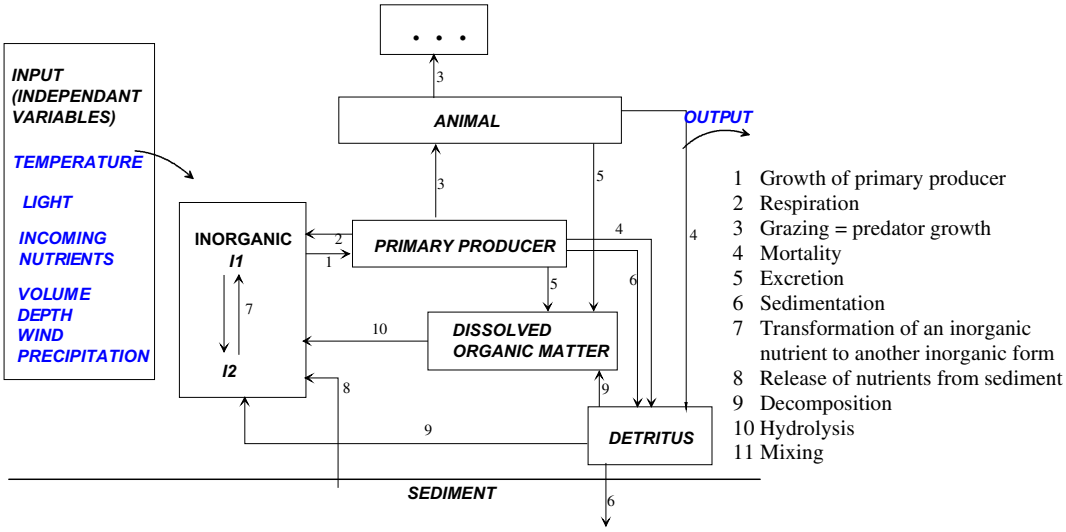

Fig. 3. Generalized scheme of state variables (boxes) and relations or processes (arrows) in aquatic ecosystem, as captured in the modelling knowledge library. As the scheme applies to a single compartment, process 11 is not presented here, since it represents mixing between compartments. It is, however, included in the modelling library.

$$\text{grazing} = c_{\text{gmax}} \cdot f_1(T) \cdot f_3(F_{\text{T}}) \cdot \text{sp} \tag{4}$$

$$\text{grazing} = \sum_{k=1}^{n} F_k \cdot c_{\text{fmax}} \cdot f_1(T) \cdot f_3(F_{\text{T}}) \cdot \text{sp} \tag{5}$$

where c_{gmax} is the maximal zooplankton ingestion rate coefficient in [mass pp/(mass sp × time)], sp is the secondary producer concentration [mass/volume], c_{fmax} is the maximal filtration rate coefficient [volume/(mass sp × time)], F_k is the individual prey concentration, F_{T} is the total available food concentration, $f_1(T)$ is a temperature adjustment function, and $f_3(F_{\text{T}})$ is a food limitation function.

3. Lake Kinneret dataset

The data used were collected by the Kinneret Limnological Laboratory (KLL) staff and extracted from the Laboratory's database (Kinneret Limnological Laboratory, 2001). The lake has been studied since 1969 and a wide range of physical, biological and chemical variables has been monitored routinely ever since (Berman et al., 1995). Data used included hydrological information such as inflow volumes and nutrient loading (ammonium – NH_4, nitrate – NO_3, dissolved organic nitrogen – DON, total dissolved phosphorus – TDP, total phosphorus – TP), physical (water temperature), chemical (Ca, conductivity, dissolved oxygen, CO_2, dissolved organic carbon – DOC, NO_3, NH_4, DON, TN, pH, PO_4, SRP, TDP, TP, total suspended solids, and turbidity) and biological (heterotrophic bacteria, copepods, cladocerans, and rotifers) factors. Chemical factors and temperature data were collected weekly while biological information was collected fortnightly.

All lake-based data were collected at a mid-lake station and at fixed depths. The depths used included 1, 3, 5, 7, 10 m for all biological and chemical factors and 1–10 m at 1 m resolution for the temperature data. Data for all biological factors were integrated over the 1–10 m depth range thus creating mass concentrations (m^{-3}) for the upper layer, while mean values for the upper 10 m of the water column were calculated for temperature and for the chemical factors. The fortnightly data were interpolated to create a dataset in which there was a fixed resolution between data points (weekly).

The periods examined in this study included 1 Jan. 1988–31 Dec. 1992 and 1 Jan. 1997–31 Dec. 1999. For more information on the sampling techniques the reader is referred to Serruya (1978). Table 3 lists the measured variables that we used for modelling the *Peridinium* dynamics.

4. Experimental setup

4.1. Experiments

In accordance with our goals, i.e., to discover a good model for *Peridinium* dynamics we formulated three experiments that correspond to three different modelling tasks for Lagramge. Our general approach was to discover a structure that is in line with existing theoretical modelling knowledge and expert knowledge about this particular case study. Thus, our general structure of the model to be discovered is presented in Eq. (6).

Table 1
Primary producer respiration rates (r), where r_{ref} is the respiration rate at reference temperature, r_{opt} is the respiration rate at optimal temperature, and k_r is maximum incremental increase in respiration under conditions of maximal growth.

Expression	Description and references
$r = \text{const}$	Exponential model
$r = r_{\text{ref}} \cdot f_1(T)$	Rate influenced by temperature, commonly used in models
$r = \text{pp} \cdot r_{\text{ref}} \cdot f_1(T)$	Second order kinetics
$r = r_{\text{ref}} \cdot f_1(T) + k_r \cdot f_1(T) \cdot f_2(N) \cdot f_3(L)$	Rate as a function of the physiological conditions of the organism (Scavia, 1980)

Table 2
Primary producer mortality rates (m), where m_{ref} is the mortality rate at the reference temperature.

Expression	Description and references
$m = \text{const}$	Exponential, non-limited model
$m = m_{\text{ref}} \cdot f_1(T)$	First order kinetics, temperature influenced model
$m = m_{\text{ref}} \cdot \text{PP} \cdot f_1(T)$	Second order kinetics, temperature influenced
$m = m_{\text{ref}} \cdot f_1(T) \cdot (1 - f_2(N) \cdot f_3(L))$	Scavia and Park (1976)
$m = m_{\text{ref}} \cdot f_1(T) \cdot \frac{\text{PP}}{k+\text{PP}}$	Nyholm (1978)

Table 3
Measured variables in lake Kinneret used for modelling of *Peridinium*.

Variable	Description	Unit
temp	Water temperature	°C
light	Light radiation	W/m^2
PO_4	Soluble phosphorus in the lake	g/m^3
NH_4	Ammonia concentration	g/m^3
NO_3	Nitrate concentration	g/m^3
ns	Sum of inorganic nitrogen (NH_4+NO_3)	g/m^3
dino	Dinoflagellate (*Peridinium*) biomass concentration	g WW[a]/m^3
zoo	Zooplankton biomass concentration	g WW/m^3

[a] WW stands for wet weight.

$$\frac{dpp}{dt} = \text{growth} - \text{respiration} - \text{mortality} \quad (6)$$

The task at hand is to find suitable mathematical expressions for each of the three processes in Eq. (6) by using the modelling knowledge library and the automated modelling tool Lagramge.

The alternative mathematical models in the library for the three selected processes are given in Section 2.2. Two experiments were performed by modifying the above concept (6), while one experiment added a zooplankton grazing process to this concept. In the first experiment (experiment 1), we formulate the modelling task specification for Lagramge to follow Eq. (6), where the growth process is modelled as nutrient, temperature and light limited growth Eq. (7). Three nutrients are included (in the nutrient limitation function): phosphorus (PO_4), ammonia (NH_4), and nitrate (NO_3).

$$\text{growth} = \text{algae}\cdot\mu_{\max}\cdot f(PO_4)\cdot f(NH_4)\cdot f(NO_3)\cdot f(\text{temp})\cdot f(\text{light}) \quad (7)$$

In experiment 2, we still follow Eq. (6), but we take nitrogen as total inorganic nitrogen, i.e., instead of NH_4 and NO_3 we take the sum (ns), since we have no information on the *Peridinium* preference for ammonia or nitrate:

$$\text{growth} = \text{algae}\cdot\mu_{\max}\cdot f(PO_4)\cdot f(\text{ns})\cdot f(\text{temp})\cdot f(\text{light}) \quad (8)$$

This is a better option for including nitrogen if we do not know *Peridinium* preferences. Using NH_4 and NO_3 separately in the growth process may cause unrealistic inhibition of growth if one of them is in very low concentration. In reality, in such a situation the algae would probably take the other nutrient instead of stopping the growth (Dortch, 1990).

Finally, in the last experiment (experiment 3) we formulate the modelling task specification to include grazing by zooplankton, as presented in Eq. (9). The growth process was formulated as in experiment 2.

$$\frac{dpp}{dt} = \text{growth} - \text{respiration} - \text{mortality} - \text{grazing} \quad (9)$$

4.2. Selection of the best model

The three experiments (represented as three modelling tasks), applied to the provided time-series data, produced a large number of mathematical models. Namely, each experiment (conceptual model) was applied to the lake-based data from both data periods (1988–1992 and 1997–1999). Recall from Section 2.1 that Lagramge induces a set of mathematical model structures for each conceptual model. These mathematical structures (more specifically the parameters in the mathematical structures) are then fitted to the provided data measurements and later ranked according to their goodness of fit.

In the task of model discovery from data it is of crucial importance to have an optimal training dataset and dataset to evaluate the discovered model. Optimal training data indicates that the dataset is representative for the ecosystem behaviour, contain as little noise as possible and consequently enable discovery of a model that can be successfully evaluated on unseen data. To select such dataset we considered subsets of data, so we could train models on a specific subset and validate it on others. Nine subsets of the entire dataset were used as training sets (see Table 4). These are the two periods 1988–1992 and 1997–1999, and subsets of these, i.e., 1988–1991, 1988–1990, 1988–1989, 1991–1992, 1990–1992, as well as 1997–1998 and 1998–1999. Thus, three modelling tasks applied to nine datasets result in 27 runs of Lagramge, each of which produced a series of models of which the best was selected based on its performance on the specified training dataset, i.e., by calculating the error (MSE) between the measured values and the values simulated by the model.

A very important issue here is the evaluation of model performance based only by MSE, as done by Lagramge. Sometimes models with low MSE or root MSE (RMSE) values indicating high model performance, do not fulfil the expert's expectations in terms of the quality of the output. This may be due to the model not correctly simulating peaks in the data, or the output has lower dynamicity then expected, and so on, whereas a model with higher MSE or RMSE on the same observations might perform better according to the expert's expectations. To overcome this issue, we introduce an additional expert's visual estimation (VE) in our procedure of model evaluation, thus the final selection of a best model is made by taking both (R)MSE and VE into account. The VE can have values from 1 to 5, where 1 stands for low estimation of model performance and 5 for best model performance according to the domain expert. The estimations are based on the model's capability to follow the dynamics of the data to which it has been fitted. A VE of 1 stands for no dynamicity at all, a VE of 2 indicates some dynamics, but too low to be considered as a good mode, a VE of 3 represents a model that is capable of identifying some (but not all) peaks and approximately follows their amplitudes, a VE of 4 stands for a model that successfully identifies all peaks and nulls in the measured dataset, but does not capture the amplitudes very accurately, and, finally, a VE of 5 is given to a model that captures all peaks and nulls and quite accurately the amplitudes. These criteria are schematically presented in Fig. 4.

The overall model evaluation (selection of best model) process is performed in three steps (Fig. 5). The first selection is performed by Lagramge (see also Section 2), i.e., models with lowest MSE from each run were taken into consideration for further evaluation. In the second step, these models are simulated on the training datasets, RMSE is calculated in addition to the MSE, and given a VE score by an expert. Thus, after this step the models contain two marks, i.e., RMSE and VE. The models with highest VE are selected for further evaluation. Note that multiple models can have similarly

Table 4
Evaluation of the best models from each experiment where a VE of 5 represents the best fit based on visual estimation by an expert. The italicized values indicate the models selected for further evaluation.

Model No.	Training data	Experiment 1		Experiment 2		Experiment 3	
		RMSE	VE (1–5)	RMSE	VE (1–5)	RMSE	VE (1–5)
1	1988–1989	7.6	3	8.04	2	10.66	2
2	1988–1990	9.06	2	8.4	1	15.98	2
3	1988–1991	6.7	2	7.2	1	7.33	1
4	1988–1992	6.99	1	6.4	2	11.3	2
5	1990–1992	*5.5*	*2*	5.16	3	5.2	3
6	1991–1992	6.4	3	*2.58*	*4*	51.23	1
7	1997–1998	*7.9*	*4*	7.99	4	16.3	2
8	1997–1999	13.3	3	*7.2*	*5*	89.47	1
9	1998–1999	5.8	2	5.8	3	18.32	3

Please cite this article in press as: Atanasova, N., et al., Automated discovery of a model for dinoflagellate dynamics, Environmental Modelling & Software (2010), doi:10.1016/j.envsoft.2010.11.003

high VE. In this case the expert selects either those having lower RMSE or all of them for further evaluation. Thus, the number of selected 'good' models depends on the expert's estimation. In the third step, the selected models are validated on an evaluation (unseen) dataset. Again, the models' performance is marked with the RMSE and the VE on the validation dataset. Finally the model with the highest VE is selected as best model. If multiple models have the highest VE, then the model with the highest VE and the lowest RMSE wins.

In order to determine whether a single model structure is applicable to the periods examined, both before and after the changes to the lake ecosystem, we prepared two evaluation datasets, 1988–1992 and 1997–1999. So, each model is simulated for both periods.

4.3. Analysis of the best model

The analysis is focused on evaluating the sensitivity of the model to specific processes in the model (or specific terms in the processes). Note that the structure of the models discovered by Lagramge is not done completely automatically as it is forced by the expert (by selecting a conceptual model) to some degree. For example, nitrogen and phosphorus are an obligatory input to the nutrients limitation, as provided in the modelling tasks for Lagramge (Eqs. (7)–(9)). If some of the nutrients are not important for the observed ecosystem, we expect Lagramge to select such parameters' values that the term has no influence on the entire simulation. The analysis is performed by simulating variations of the selected model by excluding processes and terms from the original model. Each variant is simulated on the training data, as well as on the evaluation dataset and compared to the simulation performed by the original model. If the difference is negligible, then the excluded term is considered as not important, i.e., the selected model is not sensitive to that particular term (process or specific term in the process). In this way we discover the model with minimal required complexity.

5. Results

In this section we present the optimal *Peridinium* model resulting from the induction and the evaluation procedure, described in Section 4. Given the nine data (sub)sets and the three modelling tasks Lagramge returned 27 sets of models, one for each pair of dataset and modelling task. More specifically, given the modelling

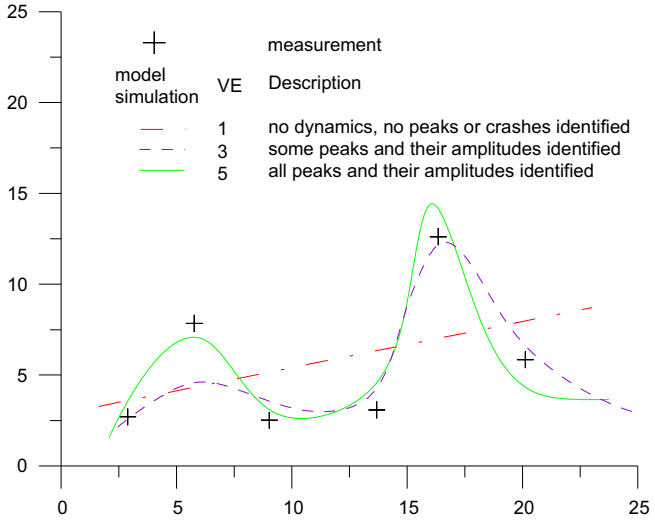

Fig. 4. Schematic representation of the criteria for expert's visual estimation (VE).

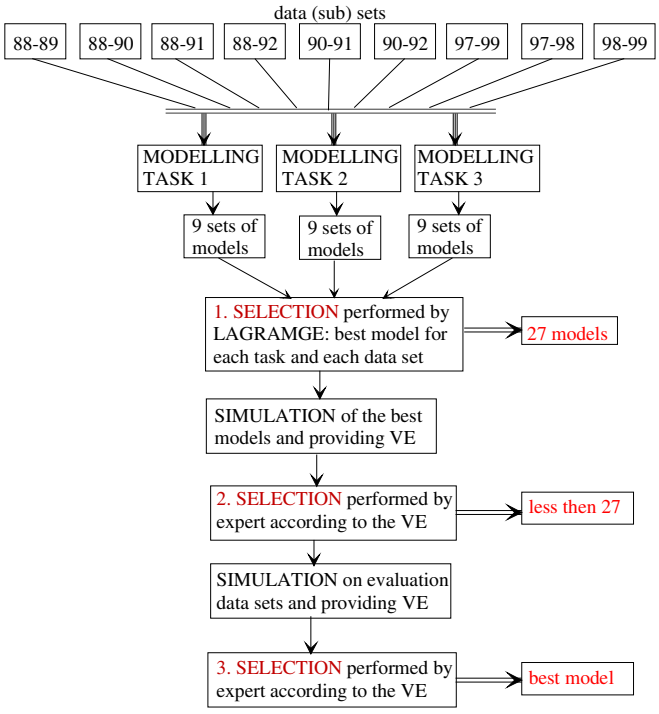

Fig. 5. Procedure of model induction with Lagramge and best model selection including the expert's visual estimation (VE).

knowledge library and the modelling task specifications, the space of candidate models includes 44,352 model structures for the modelling task 1, 12,960 structures for the modelling task 2, and 3.1×10^6 structures for the modelling task 3. To address the issue of large space of candidate models, we used the beam search strategy, which reduced the Lagramge search space to around 700 model structures for the modelling task 1, 650 for the modelling task 2, and 1200 for the modelling task 3. Lagramge ranked the models in each set of models according to their MSE between the simulated and the measured data. The RMSE of the best models (model with lowest MSE) from each model set is presented in Table 4. These models were taken for further evaluation according to the procedure in Fig. 5, and finally the best model was selected.

5.1. Selection of the best model for Peridinium dynamics

After taking the Lagramge's selection of *Peridinium* good models, each of the models was simulated and given a VE by an expert. Table 4 presents the two evaluation marks, RMSE and VE, of each model.

According to the results, it is yet again confirmed that a model with the lowest RMSE is not necessarily the best model according to the expert's expectations. In the second experiment for example, the model trained on the 1991–1992 dataset had the lowest RMSE (2.6), whereas the model trained on 1997–1999 data had higher RMSE (7.2). Simulation of the models indicates a good performance and thus high VE of both models (Fig. 6). However, the model with lowest RMSE was given a lower VE (4). Although it successfully identifies the peaks and crashes, it fails in simulating the amplitudes of the peaks. The model with the higher RMSE (Fig. 6, right) on the other hand successfully simulates the peaks and nulls, including the amplitudes of the peaks. The simulated curve is not a perfect fit to the measurements (higher RMSE), but according to the expert it simulates better the *Peridinium*, since it is also successful in simulating peaks with low and high amplitudes, and is

Please cite this article in press as: Atanasova, N., et al., Automated discovery of a model for dinoflagellate dynamics, Environmental Modelling & Software (2010), doi:10.1016/j.envsoft.2010.11.003

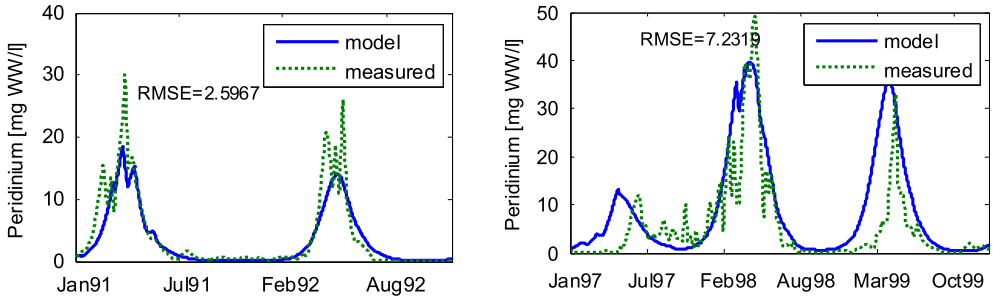

Fig. 6. Simulation of the model with lowest RMSE from experiment 2 (left) and the model with higher RMSE (right). The model on the left has a VE of 4, and the model on the right a VE of 5.

thus given a higher VE (5). Similarly, we selected two models with relatively low RMSE and high visual estimation from the first experiment (Table 4), i.e., the models trained on 1990–1992 and 1997–1998, respectively.

We did not select any models from the third experiment, since all of them had higher RMSE and lower visual estimation than the models from the first two experiments. Thus, including zooplankton grazing did not contribute to model improvement.

To summarize, according to the expert's VE we selected four models out of 27 that perform good on the training dataset. In order to select the best model of those four, the models were simulated on validation datasets. The four models include the following: Model No. 5 from experiment 1 (5/1) is given in Eq. (10) and model 7/1 is given in Eq. (11). The models 6/2 and 8/2 from the second experiment are given in Eqs. (12) and (13) respectively.

$$\frac{d(dino)}{dt} = dino\cdot 0.197\cdot\frac{PO_4}{PO_4+10^{-4}}\cdot\frac{NH_4}{NH_4+10^{-8}}\cdot\frac{NO_3^2}{NO_3^2+0.027}\cdot e^{-2.3\cdot\left(\frac{temp-16}{16}\right)^2}\cdot\frac{light}{182.2}\cdot e^{\left(-\frac{light}{127.4}+1\right)} - dino\cdot 0.0178\times 1.13^{(temp-20)} - dino^2\cdot 0.001\times 1.13^{(temp-20)} \quad (10)$$

$$\frac{d(dino)}{dt} = dino\cdot 0.11\cdot\frac{PO_4}{PO_4+10^{-8}}\cdot\frac{NH_4}{NH_4+10^{-8}}\cdot\frac{NO_3^2}{NO_3^2+0.0088}\cdot\frac{temp-4}{20-4}\cdot\frac{light}{light+2.5} - dino\cdot 0.009 - dino^2\cdot 0.001\cdot\frac{temp-4}{20-2} \quad (11)$$

$$\frac{d(dino)}{dt} = dino\cdot 0.08\cdot\frac{PO_4}{PO_4+10^{-8}}\cdot\frac{ns}{ns+10^{-8}}\cdot e^{-2.3\cdot\left(\frac{(temp-16)}{16}\right)^2}\cdot\frac{light}{light+10.05} - dino\cdot 0.038\cdot\frac{temp-4}{20-4} - dino^2\cdot 0.001\cdot\frac{temp-4}{20-2.8} \quad (12)$$

$$\frac{d(dino)}{dt} = dino\cdot 0.05\cdot\frac{PO_4}{PO_4+10^{-8}}\cdot\frac{ns}{ns+10^{-8}}\cdot e^{-2.3\cdot\left(\frac{(temp-18)}{16}\right)^2}\cdot\frac{light}{light+2.5} - dino\cdot 0.013\times 1.13^{(temp-20)} - dino^2\cdot 0.001\times 1.13^{(temp-20)} \quad (13)$$

As expected from the experimental setup, the model equations contain three terms for three processes; growth, respiration, and mortality. The models from the first experiment (Eqs. (10) and (11)) differ conceptually from the second experiment models (Eqs. (12) and (13)) in the growth term. The first experiment models contain three nutrients in the process formulation (PO_4, NH_4, NO_3), whereas the second experiment models contain two nutrients (PO_4, ns), where ns represents the sum of the nitrogen compounds. Further differences between the models comprise slightly different formulations of the processes and different values of the constant parameters.

The models were simulated and evaluated according to the third step of the evaluation process described in Section 4.2. According to the evaluation criteria, the best model for *Peridinium* in terms of performance on the training and evaluation datasets was a model trained in the second experiment on the 1997–1999 data (Eq. (13)). The model was given a VE of 4 when simulated on the validation dataset 1989–1992, which was the highest VE mark achieved on the validation datasets.

The growth term of the best model (Eq. (13)) contains nutrients (PO_4 and ns), temperature and light limitation functions. The nutrients and the light limitations are modelled with the Monod term, while for the temperature influence the optimal temperature function was selected. Respiration (second term in Eq. (13)) is modelled with temperature influenced first order kinetics and mortality with the temperature influenced second order kinetics. Note that this particular model was not validated on the data from the latter period, since the entire dataset from the second period was used for training. Yet, the model performs very well on the

Please cite this article in press as: Atanasova, N., et al., Automated discovery of a model for dinoflagellate dynamics, Environmental Modelling & Software (2010), doi:10.1016/j.envsoft.2010.11.003

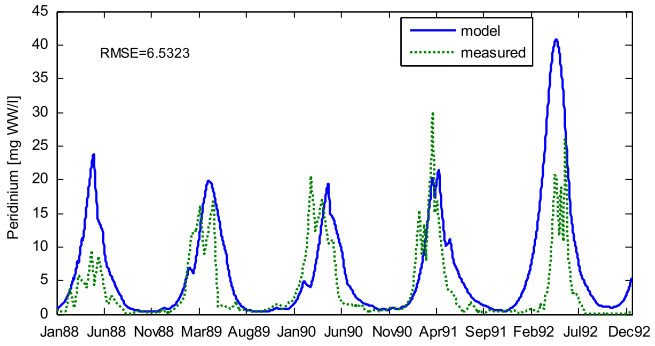

Fig. 7. Validation of the model trained on the data from 1997 to 99 (presented in Eq. (13)) on the data from period 1988 to 1992.

training data (Fig. 6, right) and was also very successful when applied to the 1988–1992 validation dataset (Fig. 7).

The best model successfully simulated the observed timing of the peaks and their amplitude in the validation dataset. It did, however, overestimate the peaks in 1988 and in 1992 (Fig. 7). While the reason for the unsuccessful simulation in 1988 is in the model itself, the 1992 peak *Peridinium* biomass can be successfully modelled with the model if simulated from Jan. 1992 (Fig. 8). Evidently the long-term simulation produces an incorrect initial value of the simulated variable (*Peridinium*) at the beginning of year 1992, which results in an overestimate of the *Peridinium* biomass in the spring.

5.2. Best model analysis

Analyses of the model's sensitivity to specific terms and processes show that the model is not sensitive to the nitrogen and light limitation function. The two Monod nutrient limitation functions contain a half-saturation constant. In both cases, the constant is very low (10^{-8}), indicating that this function basically works as a switch function, i.e., when the nutrient is present in the system the function value is close to 1, and has no influence on the value of the growth rate. When the nutrient concentration is zero the term is also equal to zero, causing the growth to stop. We can examine the sensitivity of the model outcome to the two nutrient functions by separately excluding them from the model (Eq. (13)). In the first case, we exclude the nitrogen limitation function (14), while in the second we exclude the phosphorus limitation function (15).

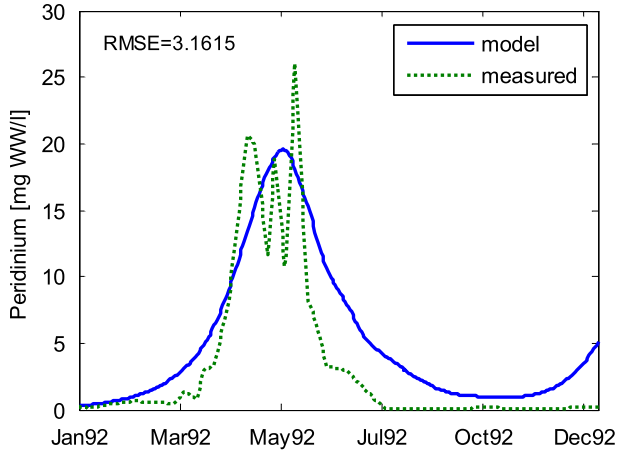

Fig. 8. Validation of the model trained on the data from 1997 to 1999 (Eq. (13)) on the data from year 1992, by starting the simulation with the measured value of *Peridinium* concentration.

$$\frac{d(dino)}{dt} = dino \cdot 0.05 \cdot \frac{PO_4}{PO_4 + 10^{-8}} \cdot e^{-2.3 \cdot \left(\frac{(temp-18)}{16}\right)^2} \cdot \frac{light}{light + 2.5} - dino \cdot 0.013 \times 1.13^{(temp-20)} - dino^2 \cdot 0.001 \times 1.13^{(temp-20)} \quad (14)$$

$$\frac{d(dino)}{dt} = dino \cdot 0.05 \cdot \frac{ns}{ns + 10^{-8}} \cdot e^{-2.3 \cdot \left(\frac{(temp-18)}{16}\right)^2} \cdot \frac{light}{light + 2.5} - dino \cdot 0.013 \times 1.13^{(temp-20)} - dino^2 \cdot 0.001 \times 1.13^{(temp-20)} \quad (15)$$

Simulating the model presented in Eq. (14) (Fig. 9), without the nitrogen limitation function provides results similar to the complete model (Fig. 6 right, Fig. 7). It is therefore possible to conclude that the nitrogen limitation function has little influence on the *Peridinium* growth. However, the results of the simulation of the model without the phosphorus limitation function (15) are quite different (Fig. 10). Evidently, the model is sensitive to the phosphorus limitation function, and thus, phosphorus should be included in the *Peridinium* model in spite of the very small half-saturation coefficient.

Similar analysis of the rest of the terms in the model shows that the model is not sensitive to the light limitation function. Thus, we can present the *Peridinium* model as in Eq. (16), showing that phosphorus and temperature are the minimal required information for modelling *Peridinium* dynamics in Lake Kinneret over the study periods.

$$\frac{d(dino)}{dt} = dino \cdot 0.05 \cdot \frac{PO_4}{PO_4 + 10^{-8}} \cdot e^{-2.3 \cdot \left(\frac{(temp-18)}{16}\right)^2} - dino \cdot 0.013 \times 1.13^{(temp-20)} - dino^2 \cdot 0.001 \times 1.13^{(temp-20)} \quad (16)$$

Note however, that though the impact of the removal of the nitrogen function from the model was small, some changes were observed (e.g. the RMSE in the second simulation period increased from 7.2, Fig. 6 right to 8.1, Fig. 9 right). Therefore, when simulating the *Peridinium* dynamics in periods, other then those presented in this paper, the original model (Eq. (13)) should be applied.

6. Discussion

In this research, we employ an automated modelling (AM) method that combines modelling knowledge and measured data to find an optimal model in a given modelling scenario. We further combine the output of the tool with a domain expert's evaluation of models to find a model of *Peridinium* dynamics in Lake Kinneret. In terms of integrated domain knowledge in the modelling process, the presented work is related to several other studies.

Whigham (1995) proposed an introduction of the domain knowledge in a form of context free grammars (CFGs) into a system based on genetic programming (GP). CFGs were also introduced in the earliest version of Lagramge (Todorovski and Džeroski, 1997). Although CFGs have been successfully used for solving various problems and introduced in various systems, they have one major draw back, i.e. CFGs are case specific, i.e., they can only be used for one modelling task at a time. Additionally, the knowledge representation using a CFG is completely different to the representation used by domain experts (typically in form of ODEs) and therefore not very popular among them. These draw backs are addressed in the present version of Lagramge by introducing formalism for higher-level knowledge representation, as described in Section 2.2, and automatically transforming it into grammars. The generated grammar is context-dependent, i.e. it allows the use of context-

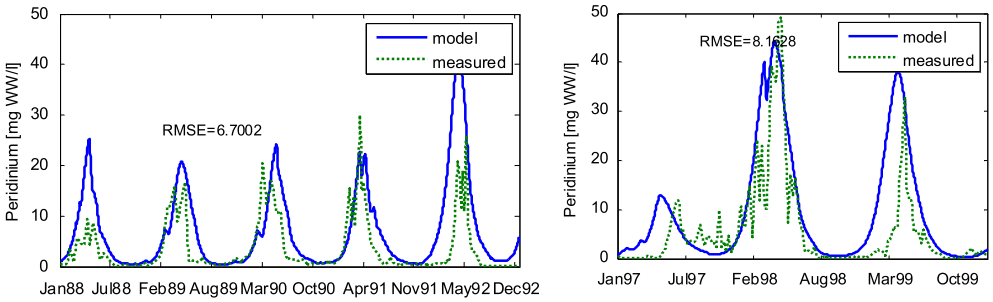

Fig. 9. Simulation of the *Peridinium* model without the nitrogen limitation function (Eq. (14)).

dependent constraints in the grammar specifying the space of possible equations.

Perhaps the most closely related system to Lagramge is the system called IPM (Bridewell et al., 2008). The domain knowledge in IPM is encoded in formalism similar to ours, which includes generic processes capturing the general domain modelling knowledge (Langley et al., 1997). IPM performs a heuristic search through the space of candidate models consisting of specific processes that can be derived from the generic processes. The search is performed directly, without generating grammars. Similarly as in Lagramge the selection of the best model is based on standard goodness-of-fit measures.

In our research, we included an expert's visual estimation of the model's performance. In fact, this research strongly suggests the use of additional model evaluation measures, since the standard goodness-of-fit measures (e.g. mean squared error or root thereof) many times fail in providing the best model according to the experts' criteria. They are, however, very beneficial in narrowing the space of possible solutions for given task, and thus of great value to domain experts.

Using the automated modelling system Lagramge combined with domain expert's criteria for model selection we propose a *Peridinium* model that captures the most crucial factors affecting *Peridinium* dynamics. However, the model discovery is based on the measured data in the system, which indicates that there might be other factors, not included in the database, affecting *Peridinium*.

Peridinium has been reported to be limited by an array of conditions including nutrients, micronutrients and physical conditions. Earlier studies conducted using both lake-based cells and cultured cells indicate that phosphorus is the main limiting nutrient especially during the period of the *Peridinium* bloom in the lake (Cavari, 1976; Criscuolo et al., 1981; Serruya and Berman, 1975). More recent studies have reported on the role of CO_2 limitation and its consequences in leading to a rapid decline in the *Peridinium* bloom and population (Berman-Frank et al., 1994; Vardi et al., 1999). There is also evidence, however to the role micronutrients play in governing the *Peridinium* bloom. The micronutrient selenium has been indicated as a key micronutrient vital to the development of *Peridinium* blooms (Lindstrom and Rodhe, 1978).

The selected model succeeded in emphasizing the factors, known as affecting *Peridinium* in the lake, out of the data on which it was based. Factors known to impact *Peridinium*, but not included in the database used for this study, such as selenium were therefore not selected by the model. Factors such as temperature and phosphorus availability, however, that were included in the database and also play a major role in the early stages of population development affecting the timing and magnitude of the blooms were identified by Lagramge as key factors. The only factor known to affect *Peridinium* that was also included in the dataset used for this study but not included in the model was $CO_{2(aq)}$. This may be a result of the timing of limitation by $CO_{2(aq)}$. This factor has been identified as the cause for the decline or crash in the bloom but not as a factor determining if and when the bloom would occur. In addition, its impact is mainly felt within the high density patches formed by the *Peridinium* in which $CO_{2(aq)}$ drops significantly but most likely undetected by measurements conducted at a fixed mid-lake station such as the one used for collecting the data applied to this study.

The results of the output sensitivity to the nutrient functions were somewhat surprising. In contrast to our expectation the low half-saturation value in phosphorus limitation function does not reflect a lack of influence of the nutrient on the *Peridinium* growth. The indications that nutrients do not influence the growth process according to the model are only true for the total inorganic nitrogen (*ns*). This is most likely an outcome of the higher nitrogen, than phosphorus, concentrations, at times by 1–2 orders of magnitude, while the half-saturation constant values in both functions are equal. In the case of PO_4, there are periods of very low phosphorus concentrations and in these periods the function is used to control the growth. In fact, these periods are regular in the 1988–1992, appearing also in the 1997 (Fig. 10).

Thus, the selected model is simple, robust and accurate enough for long-term simulation of *Peridinium* concentrations in spite of the changes that took place in the ecosystem. *Peridinium* dynamics

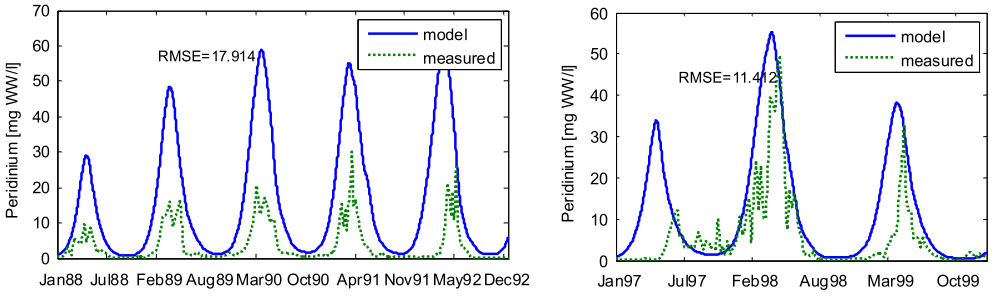

Fig. 10. Simulation of the *Peridinium* model without the phosphorus limitation function (Eq. (15)).

can be described with this simple model, and the change in algal succession has not influenced the *Peridinium* model structure. However, the model lacks descriptive power, as the dependencies between *Peridinium* and other nutrients (e.g. selenium and $CO_{2(aq)}$) and zooplankton species are not taken into account. As a consequence, more specific situations, like in the one 1988 cannot be successfully predicted with this model.

To better capture the dynamics of Dinoflagellate in Lake Kinneret, a more complex model, which can result from a more comprehensive database, as well as some further development of the modelling knowledge library, will be needed. One issue for example, would be a suitable introduction of nutrients' ratios (N:P) in the library for automated modelling. At present, they are not included due to difficulties with numerical instability that the ratios may cause, e.g., in the case when denominator (such as P values during the stratification period) is close to or zero. Also, other state variables would need to be included in the modelling tasks as well, where cyanobacteria *Microcystis* sp. and zooplankton are highest on the priority list. The induction of a more complex model will place higher computational demand on Lagremge, for which optimization and parallelisation will need to be considered.

7. Conclusions

In this paper we successfully discovered a long-term model for *Peridinium* dynamics in Lake Kinneret. This is of special importance, since the lake has undergone ecosystem changes, causing a shift in algal succession. The previously stable ecosystem with regular blooms of *Peridinium* has changed into an unstable ecosystem with reoccurring Cyanobacterial blooms. In contrast to the expectations that one model structure could not describe the *Peridinium* dynamics before and after the change, we managed to discover a model that simulates the *Peridinium* concentration over a longer period including both observed periods. The model was discovered by the automated modelling method Lagramge, capable of integrating the expert knowledge in the procedure of data-driven equation discovery. In its structure, the model includes very few independent variables, which makes it a simple model with good prediction power. The model reveals the most important factors for *Peridinium*, i.e., phosphorus and temperature. Such a model with good predictive power can be used for simulation of *Peridinium* dynamics as well as for integration into more complex lake models, such as CAEDYM (Bruce et al., 2006; Gal et al., 2009), which has already been applied to Lake Kinneret.

References

Atanasova, N., 2005. Preparation and use of the domain expert knowledge for automated modelling of aquatic ecosystems. Ph.D. thesis, University of Ljubljana, Faculty of Civil and Geodetic Engineering, Ljubljana, Slovenia.

Atanasova, N., Todorovski, L., Džeroski, S., Kompare, B., 2006a. Constructing a library of domain knowledge for automated modelling of aquatic ecosystems. Ecological Modelling 194 (1–3), 14–36.

Atanasova, N., Todorovski, L., Džeroski, S., Remec-Rekar, Š., Recknagel, F., Kompare, B., 2006b. Automated modelling of a food web in lake Bled using measured data and a library of domain knowledge. Ecological Modelling 194 (1–3), 37–48.

Atanasova, N., Todorovski, L., Džeroski, S., Recknagel, F., Kompare, B., 2006c. Computational assemblage of ordinary differential equations for chlorophyll-a using a lake process equation library and measured data of Lake Kasumigaura. In: Recknagel, F. (Ed.), Ecological Informatics, second ed. Springer-Verlag, Berlin, New York, pp. 1–485.

Atanasova, N., Todorovski, L., Džeroski, S., Kompare, B., 2008. Application of automated model discovery from data and expert knowledge to a real-world domain: Lake Glumsø. Ecological Modelling 212 (1–2), 92–98.

Bendorf, J., 1979. A contribution to the phosphorus loading concept. Internationale Revue der gesamten Hydrobiologie und Hydrographie 64 (2), 177–188.

Berman, T., Stone, L., Yacobi, Y.Z., Kaplan, B., Schlichter, M., Nishri, A., Pollingher, U., 1995. Primary production and phytoplankton in Lake Kinneret: a long-term record (1972–1993). Limnology and Oceanography 40, 1064–1076.

Berman-Frank, I., Zohary, T., Erez, J., Dubinsky, Z., 1994. CO_2 availability, carbonic anhydrase, and the annual dinoflagellate bloom. Limnology and Oceanography 39, 1822–1834.

Bridewell, W., Langley, P., Todorovski, L., Dzeroski, S., 2008. Inductive process modeling. Machine Learning 2008 (71), 32.

Bruce, L.C., Hamilton, D., Imberger, J., Gal, G., Gophen, M., Zohary, T., Hambright, K.D., 2006. A numerical simulation of the role of zooplankton in C, N and P cycling in Lake Kinneret, Israel. Ecological Modelling 193 (3–4), 412–436.

Cavari, B., 1976. ATP in Lake Kinneret: indicator of microbial biomass or of phosphorus deficiency. Limnology and Oceanography 21, 231–236.

Cox, G.M., Gibbons, J.M., Wood, A.T.A., Craigon, J., Ramsden, S.J., Crout, N.M.J., 2006. Towards the systematic simplification of mechanistic models. Ecological Modelling 198, 240–246.

Criscuolo, C.M., Dubinsky, Z., Aaronson, S., 1981. Skeleton shedding in *Peridinium cinctum* from Lake Kinneret – a unique phytoplankton response to nutrient imbalance. In: Developments in Arid Zone Ecology and Environmental Quality. Balban ISS, pp. 169–176.

Crout, N.M.J., Tarsitano, D., Wood, A.T., 2009. Is my model too complex? Evaluating model formulation using model reduction. Environmental Modelling and Software 24 (1), 1–7.

Dortch, Q., 1990. The interaction between ammonium and nitrate uptake in phytoplankton. Marine Ecology Progress Series 61, 183–201.

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

Gal, G., Hipsey, M.R., Parparov, A., Wagner, U., Makler, V., Zohary, T., 2009. Implementation of ecological modeling as an effective management and investigation tool: Lake Kinneret as a case study. Ecological Modelling 220, 1697–1718.

Jakeman, A.J., Letcher, R.A., Norton, J.P., 2006. Ten iterative steps in development and evaluation of environmental models. Environmental Modelling and Software 21, 602–614.

Kinneret Limnological Laboratory, 2001. Lake Kinneret Database: Limnological Data on Lake Kinneret Since 1968. Israel Oceanographic and Limnological Research, Yigal Allon Kinneret Limnological Laboratory, Tiberias.

Langley, P., Sanchez, J., Todorovski, L., Dzeroski, S., 1997. Inducing process models from continuous data. In: Sammut, C., Hofmann, A. (Eds.), Proceedings of the Nineteenth International Conference on Machine Learning. Morgan Kaufmann, San Mateo, CA, pp. 347–354.

Lindstrom, K., Rodhe, W., 1978. Selenium as a micronutrient for the dinoflagellate *Peridinium cinctum* fa. *westii*. Internationale Vereinigung für Theoretische und Angewandte Limnologie 21, 168–173.

Nyholm, N., 1978. A simulation model for phytoplankton growth and nutrient cycling in eutrophic, shallow lakes. Ecological Modelling 4 (2–3), 279–310.

Recknagel, F., 1980. Systemtechnische Prozedur zur Modellierung und Simulation von Eutrophierungs-prozessen in stehenden und gestauten Gewässern. Sektion Wasserwesen, TU Dresden, Dresden.

Robson, B.J., Hamilton, D.P., Webster, I.T., Chanc, T., 2008. Ten steps applied to development and evaluation of process-based biogeochemical models of estuaries. Environmental Modelling and Software 23, 369–384.

Roelke, D.L., Zohary, T., Hambright, K.D., Montoya, J.V., 2007. Alternative states in the phytoplankton of Lake Kinneret, Israel (Sea of Galilee). Freshwater Biology 52, 399–411.

Scavia, D., Park, R.A., 1976. Documentation of selected constructs and parameter values in the aquatic model cleaner. Ecological Modelling 2 (1), 33–58.

Scavia, D., 1980. An ecological model of Lake Ontario. Ecological Modelling 8, 49–78.

Serruya, C., 1978. Lake Kinneret. Dr. W. Junk, The Hague, 501 pp.

Serruya, C., Berman, T., 1975. Phosphorous, nitrogen and the growth of algae in Lake Kinneret. J. Phycol 11, 155–162.

Todorovski, L., 2003. Using Domain Knowledge for Automated Modeling of Dynamic Systems with Equation Discovery. Fakulteta zaračunalništvo in informatiko, University of Ljubljana, Ljubljana, Slovenia.

Todorovski, L., Džeroski, S., 1997. Declarative bias in equation discovery. In: Proceedings of the Fourteenth International Conference on Machine Learning. Morgan Kaufmann, Los Altos, CA, pp. 376–384.

Vardi, A., Berman-Frank, I., Rozenberg, T., Hadas, O., Kaplan, A., Levine, A., 1999. Programmed cell death of the dinoflagellate *Peridinium gatunense* is mediated by CO_2 limitation and oxidative stress. Current Biology 9, 1061–1064.

Verhulst, P.-F., 1845. Recherches mathématiques sur la loi d'accroissement de la population. Nouveaux Mémoires de l'Académie Royale des Sciences et des Lettres de Bruxelles 18, 1–41.

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 (OECD), Paris.

Welsh, W.D., 2008. Water balance modelling in Bowen, Queensland, and the ten iterative steps in model development and evaluation. Environmental Modelling and Software 23, 195–205.

Whigham, P.A., July 1995. Grammatically-based genetic programming. In: Rosca, J. (Ed.), Proceedings of the Workshop on Genetic Programming: From Theory to Real-World Applications. Morgan Kaufmann Publications, pp. 33–41.

Zohary, T., 2004. Changes to the phytoplankton assemblage of Lake Kinneret after decades of a predictable, repetitive pattern. Freshwater Biology 49, 1355–1371.

Please cite this article in press as: Atanasova, N., et al., Automated discovery of a model for dinoflagellate dynamics, Environmental Modelling & Software (2010), doi:10.1016/j.envsoft.2010.11.003