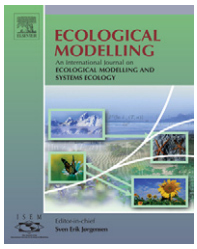
available at www.sciencedirect.com

ScienceDirect

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



Application of automated model discovery from data and expert knowledge to a real-world domain: Lake Glumsø

Nataša Atanasova[a,*], *Ljupčo Todorovski*[b], *Sašo Džeroski*[b], *Boris Kompare*[a]

[a] *Faculty of Civil and Geodetic Engineering, University of Ljubljana, Slovenia*
[b] *Jožef Stefan Institute, Slovenia*

ARTICLE INFO

Article history:
Published on line 26 November 2007

Keywords:
Automated modeling
Lagramge
Modeling knowledge base
Lake Glumsø
Phytoplankton model

ABSTRACT

In this paper, we apply automated modelling method Lagramge to the task of modelling phytoplankton dynamics in Lake Glumsø, Denmark. The approach is based on integrating expert knowledge in the process of automated model induction from measured data. It supports modelling of ecosystem dynamics with ordinary differential equations by following the mass conservation law. The data set used in this paper comprises 2 years daily measurements of data needed for phytoplankton modelling in lake. In order to have sufficient data set for training and testing the models, the entire data set was divided in two parts, each containing 1 year of daily measurements. The expert knowledge supplied to Lagramge consists of elementary models of the basic ecological processes related to the food web dynamics and rules for combining elementary into complex models of the whole system. By applying Lagramge on Lake Glumsø we discovered a set of phytoplankton models that showed good fit on the training data set. The models were evaluated by simulating them on testing data set, which revealed good performance of the models.

1. Introduction

Lake ecosystems are complex dynamic systems. Modelling of such ecosystems is a great challenge to the scientists, who are progressively improving and making more and more complex models. In general, we distinguish between two basic approaches to mathematical modelling. Following the deductive approach (knowledge driven), the model is derived from the basic background knowledge (e.g. basic physical, chemical and biological principles) from the domain of use. The second, inductive approach (data driven) is based on exploring some space of candidate models and asses their accuracy against measured data. The model that fits measured data best is the result of the induction.

In this paper we apply an approach to modelling, which combines advantages of both the domain expert knowledge and induction from measured data. The domain knowledge is gathered in a knowledge library, which is used to guide the process of induction. The library consists of a set of elementary models, mainly descriptions of basic generic processes as well as rules for building complex models as interactions of these processes. Knowledge encoded in the library typically follows the basic principles in the domain of interest (Todorovski and Dzeroski, 2001; Langley et al., 2002; Todorovski, 2003). In the early days of the development of these tools (Todorovski and Džeroski, 1997), the knowledge had to be provided as an explicit definition of the space of candidate models. Now, these tools allow the user to provide higher-level domain

* *Corresponding author*. Tel.: +386 1 425 4052x111; fax: +386 1 251 9897.
E-mail address: natanaso@fgg.uni-lj.si (N. Atanasova).
0304-3800/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2007.10.032

knowledge about building mathematical models of complex real-world systems.

In this paper we focus on the application of the newly developed knowledge library about modelling of lake ecosystems (Atanasova, 2005; Atanasova et al., 2006a) on a real-world domain, i.e., Lake Glumsø, Denmark. The library comprises many known concepts in the domain of lake modelling, which can be found in literature (e.g. Jørgensen and Bendoricchio, 2001; DeAngelis, 1992; Chapra, 1997; and so on). Together with the automated modelling method Lagramge 2.0 the library was already successfully applied to other lakes, i.e., Lake Bled (Atanasova et al., 2006b) and Lake Kasumigaura (Atanasova et al., 2006c).

Lake Glumsø has been tackled with machine learning methods previously (Todorovski et al., 1998; Todorovski, 2003) with the earlier version of Lagramge, i.e., the version V 1.0 that required a hand crafted grammar has been used to discover a phytoplankton model (Todorovski et al., 1998). The same model was (re)discovered with the latest version V 2.0 of Lagramge (Todorovski, 2003), by using a simple knowledge library. Slightly different model was discovered by implementing a complex knowledge library (Atanasova, 2005). All of these experiments were performed on a small data set that did not allow for successful model evaluation. Just recently we obtained additional data for Lake Glumsø (Jørgensen, 2004), i.e., 2 years data set of daily measurements. This data set allows for correct model induction and successful model evaluation as well, i.e., inducing models on 1 year data and testing on the other year (unseen) data.

The paper is organized as follows: in the next section we briefly explain the method and the procedure of introduction of the expert knowledge about a specific ecosystem to the model discovery tool, i.e., Lagramge (if not specified, version V 2.0 is meant). In Section 3 we present the data set and the experiments performed. Section 4 gives the results and discussion. Finally, the conclusions are summarized in Section 5.

2. The method: automated modelling framework

2.1. *Conceptual modelling of dynamic ecosystems*

In order to compose a mathematical model of a dynamic ecosystem we typically start with setting a conceptual model of the system model. In the conceptual model, ecological modeller determines (1) the relevant variables in the system and (2) the bio-geo-chemical processes that connect these variables. In the next step the modeller selects mathematical formulations for the processes included, and finally estimates the constant parameters' values in the mathematical model (by hand or numerically, by fitting the parameters to the measurements). In the later phase the modeller evaluates the obtained model by performing its simulations on test data sets. However there is an important issue to be raised here, i.e., has the modeller selected the correct mathematical structure (and parameters) for the selected concept? This issue can be addressed systematically using computational methods for modelling.

2.2. *Lagramge method*

Lagramge is an automated modelling tool for a given conceptual model, perform heuristic searches for optimal mathematical model structure and optimal model parameters, that is structure and parameters that fit measurement data best. The search is performed based on a knowledge

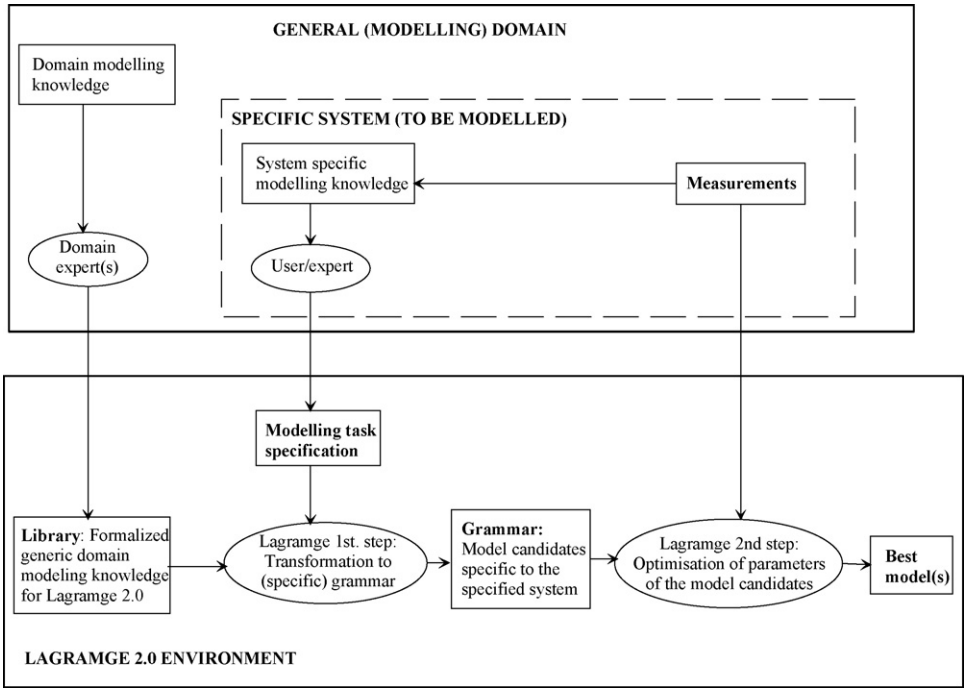


Fig. 1 – An automated modelling framework based on the integration of domain-specific modelling knowledge in the process of equation discovery.

library that contains general domain modelling knowledge, written in form of generic variables types, generic processes, list of alternative formulations for the generic processes, and combining schemes that represents a recipe for combining processes into a model of entire system, i.e., in ecological terms the combining schemes represent the mass balances of a state (system) variable. The procedure of automated modelling with Lagramge is shown in Fig. 1. In order to discover a suitable model for the observed system with Lagramge, the modeler must (1) submit the measured data of the system and (2) specify the modeling task. In the modelling task the modeller specifies observed system variables and processes that are expected to influence the behaviour of the observed system. Modelling task specification corresponds to conceptual model of the ecosystem.

Given a specification of modelling task at hand, Lagramge pre-processor can transform the high-level knowledge from the library into an operational form of a grammar in a following way. For each process in the task specification the algorithm first checks the consistencies of the types of the variables involved in the process. Then, the algorithm matches the process from task specification against the generic processes in the knowledge library. Once the match is found, we use the generic process definition to obtain the list of all alternative mathematical models for the particular process. Thus, the result of this match is the list of specific mathematical model structures that can be used to model the process specified in the task specification. This list of alternative models is encoded in the grammar as a list of alternative rules for generating arithmetical expressions. Combining scheme from the library is then used to compose the models of individual processes into a single model of the whole observed ecosystem. This grammar now completely specifies the space of candidate models of the observed system. This is illustrated in the left-hand side of Fig. 1.

Once we have the grammar, we can use equation discovery system Lagramge to heuristically search through the space of candidate models, 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 function is presented in Eq. (1), where $v_d(i)$ represents measured value of the variable v_d at point i, $\tilde{v}_d(i)$ represents the calculated value of the variable v_d at point i by the discovered equation of form $\dot{v}_d = E$, and m is the number of measured points.

$$\mathrm{MSE} = \frac{\sum_{i=1}^{m}(v_d(i) - \tilde{v}_d(i))^2}{m} \quad (1)$$

The other is minimum description length (MDL) function that takes into account model complexity and introduces preference towards simpler models. This function is shown in Eq. (2), where $l(E)$ represents the equation length (expressed in number of terms), l_{max} is the maximal equation length, which can be derived from the grammar and the simplest term in the grammar (E_0). Parameter 10 is selected based on experience. The second term in the function increases the MSE based on the equation length (longer equation will have higher error).

$$\mathrm{MDL}(v'_d = E) = \mathrm{MSE}(v'_d = E) + \frac{l(E)}{10 l_{max}}\mathrm{MSE}(v'_d = E_0) \quad (2)$$

Further details about the modelling framework from Fig. 1 can be found in (Todorovski, 2003).

2.3. *The knowledge library*

In this paper we are using a library that supports modelling food webs in lakes (Atanasova et al., 2006a). The library supports construction of zero-dimensional N-box models, by implementing the mass conservation law, i.e., modelling with ordinary differential equations (ODE modelling). It was estimated that the knowledge coded in the library covers great number of known lake models. Models of different complexity can be derived from the library, such as the simple Vollenweider's model (Vollenweider, 1968) or the fairly complex SALMO model (Bendorf, 1979; Recknagel, 1980). For more details see Atanasova et al. (2006a).

3. The data set and the experiments

3.1. *The data set*

Lake Glumsø (Jørgensen et al., 1986) is situated in a sub-glacial valley in Denmark. It is a shallow lake with average depth of about 2 m. Its surface area measures 266,000 m^2. For several years, it was receiving mechanically–biologically treated waste water from a community with 3000 inhabitants and a surrounding area which was mainly agricultural with almost no industry. The high-nitrogen and phosphorus concentration in the treated waste water has caused hypereutrophication. The lake contained no submerged vegetation, probably due to the low transparency of the water and oxygen deficit at the bottom of the lake.

The data set (provided by Jørgensen, 2004) includes 2 years of daily measurements from April 1973 to 1974 and from October 1974 to 1975. The data include daily measurements of through flow, daily sunlight intensity, water temperature, inorganic nutrient (dissolved phosphorus), phytoplankton concentration expressed as Chl-a, and zooplankton concentration (expressed in dry weight, DW). The data used for modelling are depicted in Table 1.

Table 1 – Measured variables in Lake Glumsø

Variable	Description	Unit
Temp	Water temperature	°C
Light	Light radiation	W/m^2
q_in	Inflow to the lake	m^3/day
q_out	Outflow	m^3/day
n_in	Nitrogen in the inflow	g/m^3
p_in	Phosphorus in the inflow	g/m^3
ps	Soluble phosphorus in the lake	g/m^3
phyto	Phytoplankton biomass concentration	g Chl-a/m^3
zoo	Zooplankton biomass concentration	g DW/m^3

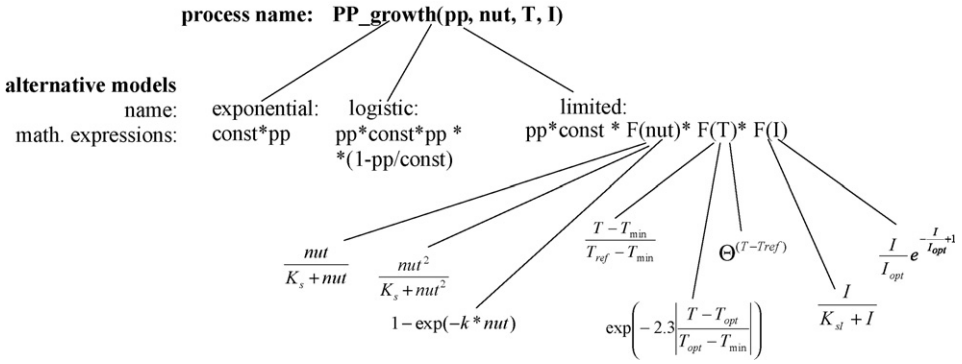


Fig. 2 – Schematic presentation of the modelling knowledge for the process of primary producer growth.

3.2. *Preparing Lagramge for model discovery*

3.2.1. *Knowledge preparation*

Our goal is to discover a phytoplankton model (expressed in Chl-a) for Lake Glumsø using the automated modelling method Lagramge. For this, we first need to prepare our specific knowledge about the system, i.e., modelling task specification that corresponds to conceptual model of the system. The processes that affect phytoplankton concentration were considered to be growth of phytoplankton, respiration, sedimentation, and grazing by zooplankton. Phytoplankton concentration increases due to the growth process and decreases due to respiration, sedimentation and grazing. This specific knowledge about the processes was introduced to Lagramge through task specification as shown in Table 2.

In the lines from 1 to 5 the variables' types are declared, i.e., *ps* (dissolved inorganic phosphorus) *phyto* (phytoplankton, expressed as Chl-a), *zoo* (zooplankton), *temp* (temperature), and *light* (light intensity). Note that phytoplankton is the system or state variable and the others are considered as independent variables. Therefore the word system stands in front of the variable declaration (Table 2, line 2). Processes are defined in the lines from 6 to 9. Phytoplankton growth is described in line 6. The process name is *PP_growth* and it has four arguments. The first is the name of the phytoplankton state variable. The arguments in the {} brackets, i.e., {ps}, {light} and {temp} define the influences and limitations of the process by nutrients, light, and temperature, respectively. Leaving one of them out would indicate no influence by the variable, which was left out. For instance, definition of a growth process that is influenced only by temperature and by two nutrients simultaneously (*ps* and *ns*) but is not light limited, would be:

process PP_growth(phyto, {ps, ns}, {temp}, {})

The process Feeds_on (line 7) stands for predatory loss of phytoplankton. Optional arguments of this process are the zooplankton food (*phyto*) and temperature (*temp*), which means that the growth of zoo can be or not influenced by the food concentration (none or many species) and temperature. Similarly, the rest of the processes in the system (Respiration_PP, and Sedimentation) are defined (see lines 8 and 9).

Note that each process in the task specification is declared in the knowledge library with several possible formulations among which Lagramge selects the most suitable one according to the given measurements. For example, the process of primary producer growth (*PP_growth*), schematically presented in Fig. 2, contains three different models (alternative formulations for this process), i.e., exponential model, logistic model, and growth model limited by nutrients (nut), temperature (*T*), and light (*I*). Furthermore nutrient, temperature, and light limitations are defined as functions that can also have several alternative formulations. Thus, according to Fig. 2 we have twenty possible formulations for the PP_growth process, i.e., one alternative for exponential model, one for logistic model and 18 for nutrient light and temperature limited model

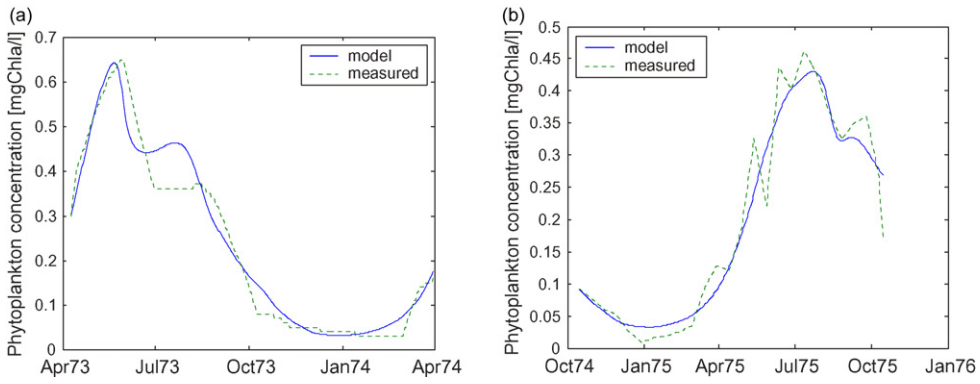


Fig. 3 – Best models performance on training data—(a) model induced on 1973/1974 data, Eq. (6) and (b) model induced on 1974/1975 data, Eq. (7).

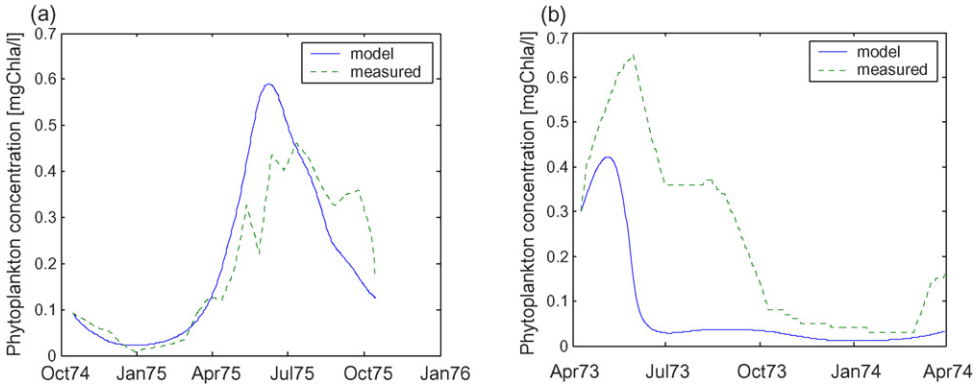


Fig. 4 – Evaluation of the phytoplankton models on test data set—(a) simulation of the model trained on 1973/1974 data (Eq. (6)) on the 1974/1975 data and (b) simulation of the model trained on 1974/1975 data (Eq. (7)) on the 1973/1974 data.

(3*3*2). We here present two variations of the growth-limited model (Eqs. (1) and (2)), where *pp* is concentration of primary producer, T_{min} stands for minimal water temperature for phytoplankton growth, T_{ref} is the reference temperature at which the constant parameters are given (typically 20 °C). Const, *k*1, and *k*2 are constant parameters that are fitted to measurements. Const represents the maximal growth rate at optimal conditions (nutrients, temperature and light), which is corrected by multiplication with nutrient, temperature, and light function, whereas *k*1 and *k*2 are half-saturation constant in the Monod function[1].

$$\text{PP\_growth} = \text{pp} \cdot \text{const} \frac{\text{nut}}{k1 + \text{nut}} \frac{T - T_{min}}{T_{ref} - T_{min}} \frac{I}{k2 + I} \quad (3)$$

$$\text{PP\_growth} = \text{pp} \cdot \text{const} \frac{\text{nut}^2}{k1 + \text{nut}^2} \frac{T - T_{min}}{T_{ref} - T_{min}} \frac{I}{k2 + I} \quad (4)$$

The first alternative (Eq. (1)) of the growth-limited model includes the Monod saturation function for nutrient and light limitation, and linear temperature function, whereas the second (Eq. (2)) includes a second order Monod for nutrient limitation. Similarly, we have a number of candidate formulations for the rest of the processes in this lake.

The library of modelling knowledge also specifies how to combine the processes into a corresponding model of the whole system (Dzeroski and Todorovski, 2003; Todorovski, 2003; Atanasova, 2005). The combining rules in the library support modelling with ordinary differential equations by following the mass conservation principle. According to the combining rules from the library, the processes defined in the task specification, will be composed in the following model based on one differential equation (5):

$$\frac{\text{d phyto}}{\text{dt}} = \text{PP\_growth} - \text{Respiration\_PP} - \text{Sedimentation} - \text{Feeds\_or}$$

[1] Monod function represents a limitation function. For example, Monod term $f(x) = x/(x + \text{constant})$ indicates limitation by *x*. The smaller the constant (half saturation coefficient) in the Monod term the smaller is the influence by *x*. A term with saturation coefficient of zero, i.e., $x/(x + 0)$ is equal to 1, which means no limitation (influence) by *x*.

3.2.2. *Data preparation*

Regarding lake data measurements, following needs to be stressed: (1) Some lakes have repeating (similar) patterns from year to year, but this is not necessary, so inducing general model from multiple seasons (years) typically does not lead to satisfactory results. Therefore, it is reasonable to fit a model on 1 year's data and validate it on others, and (2) measurements can be quite unreliable—thus some years may have more representative measurements. Models induced on such data are more general and can be easier validated on data from other periods.

Following this facts we prepared the 2 years data in two data set files, i.e., 1973/1974 and 1974/1975 data sets. In order to find the most general model (model that can also be validated on unseen data) we set Lagramge to train models on each year. Next, we take the best model from each year and simulate it on unseen data.

4. Results and discussion

4.1. *Model induction*

Lagramge discovered two sets of total phytoplankton models using (1) the expert knowledge specified in Table 2 and the data set from 1973/1974 and (2) the same expert knowledge and the data from 1974/1975. These models are ranked according to their goodness of fit to the training data set, i.e., MSE. For both years Lagramge found models that fit very well to the measurement data (see Fig. 3). The best model (model with lowest MSE on the training data) induced on 1973/1974 data is

Table 2 – Task specification for the Lake Glumsø

1	Variable inorganic ps
2	System variable Primary_producer phyto
3	Variable animal zoo
4	Variable temperature temp
5	Variable light light
6	Process PP_growth(phyto, {ps}, {temp}, {light}) p1
7	Process Feeds_on(zoo, {phyto}, {temp}) p3
8	Process Respiration_PP(phyto, {temp}, {ps}, {light}) resp0
9	Process Sedimentation(phyto, {temp}) sed0

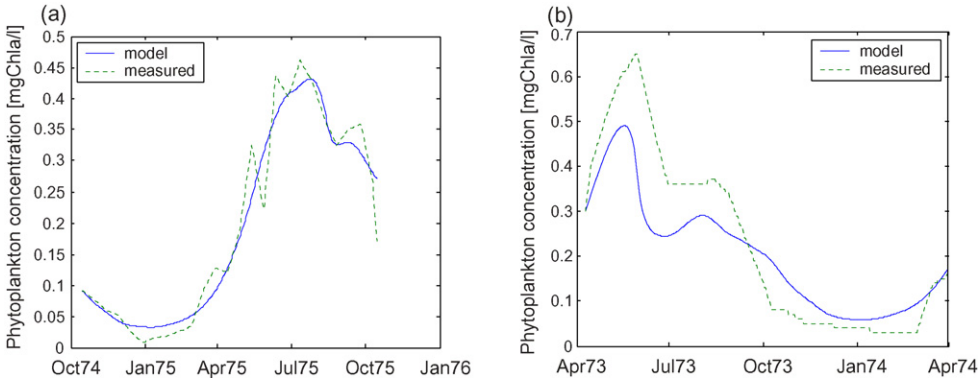


Fig. 5 – Phytoplankton model (Eq. (8)) performance on Lake Glumsø data. (a) Performance on the training set and (b) validation on unseen data.

presented in Eq. (6) and the best model induced on 1974/1975 data in Eq. (7).

$$\frac{\mathrm{d\,phyto}}{\mathrm{dt}} = \mathrm{phyto}\cdot 0.5\frac{\mathrm{ps}}{\mathrm{ps}+0.009}\frac{\mathrm{temp}}{20}\frac{\mathrm{light}}{225.2}e^{-((\mathrm{light}/133.4)+1)} - \mathrm{phyto}\cdot 0.14\frac{\mathrm{temp}-2.8}{20-4} - \mathrm{phyto}\frac{0.28}{d}\frac{\mathrm{temp}-4}{20-4} - \mathrm{zoo}\cdot 0.01\frac{\mathrm{temp}}{20}\frac{\mathrm{phyto}}{\mathrm{phyto}+4.45}\mathrm{phyto} \quad (6)$$

$$\frac{\mathrm{d\,phyto}}{\mathrm{dt}} = \mathrm{phyto}\cdot 0.25\frac{\mathrm{ps}^2}{\mathrm{ps}^2+0.006}\frac{\mathrm{temp}}{20}\frac{\mathrm{light}}{151.9}e^{-((\mathrm{light}/135)+1)} - \mathrm{phyto}\cdot 0.12\frac{\mathrm{temp}-4}{20-2} - \mathrm{phyto}\frac{0.18}{d}\frac{\mathrm{temp}-2}{20-2.3} - \mathrm{zoo}\cdot 0.01\cdot 1.13^{(\mathrm{temp}-20)}\frac{\mathrm{phyto}}{\mathrm{phyto}+5}\mathrm{phyto} \quad (7)$$

Regarding the transparency and clearness to domain experts both models follow the rules declared in the library. They contain four terms, each corresponding to the processes from the combining scheme (Eq. (5)), i.e., the first term represents phytoplankton growth, the second respiration, the third sedimentation, and the fourth grazing by zooplankton. Both years are identified with similar model structures. The only difference is in the nutrient limitation function, which in the first case (Eq. (6)) represents a simple Monod function, whereas in the second (Eq. (7)) Monod^2 function, and in the temperature function in the grazing term. Major difference between these two models is in the values of the models' parameters.

4.2. *Evaluation of the models*

In the previous section Lagramge found two models that fit very well to the training data set. In order to evaluate the models we simulated them on unseen data, i.e., the model trained on 1973/1974 data was simulated on 1974/1975 data and vice versa. The simulation of the models is shown in Fig. 4. It is evident that the model trained on 1973/1974 data shows better fit on the evaluation data set.

Models evaluation works in favour to the model discovered on 1973/1974 data, since the simulation on unseen data shows better fit to the measurements. This can lead to conclusion that the data set from 1973/1974 is more representative for the situation in the Lake Glumsø and therefore the models induced on this year data set can be successfully evaluated on other years.

However, these models are chosen according to their MSE on the training data set. This includes the possibility that models with slightly lower MSE may still perform good enough on the training set, but better on evaluation data set. Therefore, the evaluation analysis requires more detailed investigation on the other models in the model sets as well, i.e., models that have lower MSE on the training data. In this manner we simulated the first ten models in each model set on evaluation and training data set. This analysis reveals the following. The best model in the 1973/1974 model set performs best on the evaluation data set (1974/1975), whereas the best model in the 1974/1975 model set regarding evaluation is the one that was ranked on the third place according to MSE (Eq. (8)) on the training data.

$$\frac{\mathrm{d\,phyto}}{\mathrm{dt}} = \mathrm{phyto}\cdot 0.33\frac{\mathrm{ps}}{\mathrm{ps}+0.023}\frac{\mathrm{temp}}{20}\frac{\mathrm{light}}{198.5}e^{((-\mathrm{light}/137)+1)} - \mathrm{phyto}\cdot 0.12\frac{\mathrm{temp}-2.7}{20-4} - \mathrm{phyto}\cdot\frac{0.14}{d}\frac{\mathrm{temp}-4}{20-2.6} - \mathrm{zoo}\cdot 0.01\cdot 1.13^{(\mathrm{temp}-20)}\frac{\mathrm{phyto}}{\mathrm{phyto}+5}\mathrm{phyto} \quad (8)$$

Note that this model performs nearly as good as the best model (model with lowest MSE) on the training data and has very similar structure as the model induced on 1973/1974. They only differ in the parameters' values and the temperature function in the grazing term. Performance of this model on training (1974/1975) and evaluation (1973/1974) data set is shown in Fig. 5.

According to these results Lagramge can find good models in terms of goodness of fit on unseen data on both data sets. Note however, that the present information (data) about the system do not allow for final estimate about best model. In order to finally evaluate the models, we need additional data set to simulate the models, and investigate their behaviour on unseen data.

5. Conclusions

An approach to automated modelling (Lagramge), i.e., discovery of models in form of ODE's by using the expert knowledge and measured data, has been successfully applied on a real-world domain, i.e., Lake Glumsø. The data set comprising 2 years daily measurements was split in two parts in order to have sufficient data for training and evaluating models. Using the automated modelling tool a set of phytoplankton models was successfully discovered and evaluated. Evaluation of the models indicates that Lagramge can discover several good phytoplankton models in terms of fitting the training as well as the testing data. In order to make the final estimate about the best model, additional data set is needed to evaluate the models presented in this paper.

REFERENCES

Atanasova, N., 2005. Priprava in uporaba ekspertnega predznanja za avtomatizirano modeliranje vodnih ekosistemov (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. Ecol. Model. 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. Ecol. Model. 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, 2nd edition. Springer-Verlag, Berlin, New York, pp. 1–485.

Bendorf, J., 1979. A contribution to the phosphorus loading concept. Int. Revue. Ges. Hydrobiol. 64 (2), 177–188.

Chapra, S.C., 1997. Surface Water-Quality Modeling. McGraw-Hill.

DeAngelis, D.L., 1992. Dynamics of Nutrient Cycling and Food Webs. Chapmann & Hall, London.

Dzeroski, S., Todorovski, L., 2003. Learning population dynamics models from data and domain knowledge. Ecol. Model. 170, 129–140.

Jørgensen, S.E., 2004. Lake Glumsø data. Personal communication.

Jørgensen, S.E., Bendoricchio, G., 2001. Fundamentals of Ecological Modelling. Elsevier.

Jørgensen, S., Kamp-Nielsen, L., Christensen, T., Windolf-Nielsen, J., Westergaard, B., 1986. Validation of a prognosis based upon a eutrophication model. Ecol. Model. 32, 165–182.

Langley, P., Sanchez, J., Todorovski, L., Dzeroski, S., 2002. Inducing process models from continuous data. In: Paper Presented at the Nineteenth International Conference on Machine Learning, Sydney, Australia.

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

Todorovski, L., 2003. Using domain knowledge for automated modeling of dynamic systems with equation discovery. Ph.D. Thesis. University of Ljubljana, Ljubljana, Slovenia.

Todorovski, L., Dzeroski, S., 2001. Using domain knowledge on population dynamics modeling for equation discovery. In: Raedt, L.D., Flach, P. (Eds.), Proceedings of the Machine learning: ECML 2001. Springer, Berlin, str. 478–490 (Lecture notes in artificial intelligence, lecture notes in computer science, 2167).

Todorovski, L., Džeroski, S., 1997. Declarative bias in equation discovery. In: Paper presented at the 14th International Conference on Machine Learning, San Mateo, CA.

Todorovski, L., Dzeroski, S., Kompare, B., 1998. Modelling and prediction of phytoplankton growth with equation discovery. Ecol. Model. 113 (1/3), 71–81.

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.