DEVELOPMENT OF A LAKE MODEL USING DATA AND EXPERT KNOWLEDGE - CASE STUDY: GREIFENSEE

Nataša Atanasova[1], Johanna Mieleitner[2], Sašo Džeroski[3], Ljupčo Todorovski[3], Boris Kompare[1]

[1]UL-FGG
Jamova 2, 1000 Ljubljana, Slovenia
Tel: +386 1 425 40 52; fax: +386 1 251 98 97; e-mail: natanaso@fgg.uni-lj.si

[2] Swiss Federal Institute of Aquatic Science and Technology (Eawag),
Überlandstrasse 133, 8600 Dübendorf, Switzerland

[3]Jozef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia

ABSTRACT

This paper deals with setting a simple lake model of lake Greifensee by using an automated modelling (AM) method. The method combines an empirical and a theoretical approach to modelling. Theoretical knowledge, encoded in a knowledge library, is used to guide the procedure of model induction from measured data. The Greifensee data set comprises long term measurements that are crucial for describing the lake behaviour and the lake's trophic state. Using the AM method and the measured data we discovered a lake model of Greifensee that consists of three ordinary differential equations. The model describes the temporal dynamics of phosphate, chlorophyll-a, and zooplankton, taking a one day time step. Validation of the model, performed on a four years time period, indicates fairly good fit to the measurements and long term model stability.

1 INTRODUCTION

In this paper we evaluate a modelling method that combines an empirical (data-driven) and a theoretical (knowledge-driven) approach to modelling (Todorovski, 2003). The background (theoretical) knowledge is introduced in the procedure of automated model induction from data (equation discovery) in the form of a knowledge library. As a result, the method discovers a set of models (equations) that follow the basic principles in the domain of interest.

In order to be used in the AM procedure the theoretical knowledge is coded in a knowledge library. For this research we use a knowledge library for lake modelling. It was estimated that the library comprises a great part of the existing modelling knowledge from this domain, i.e. ecosystems modelling with ordinary differential equations (Atanasova, 2005). Details about this kind of modelling can be found in e.g. (Jørgensen and Bendoricchio, 2001), (DeAngelis, 1992), (Chapra, 1997), and many others. Further, the models discovered from this library are structurally correct according to the expert modelling knowledge. The lake library was applied on real-world domains, i.e. discovering a phytoplankton model for lake Bled (Slovenia), lake Kasumigaura (Japan), lake Glumsø (Denmark) and Lagoon of Venice (Italy) (Atanasova, 2005). All of the discovered models consist of a single differential equation, except for the lake Bled where we discovered a model of three differential equations. Yet, the model was discovered step wise, i.e. one equation at a time, by strictly limiting and controlling the search space of the candidate models (Atanasova et al., 2005).

The main goal of this paper is to further evaluate the AM method and the existing knowledge library. Furthermore we make an attempt of simultaneously discovering a model of three differential equations. The data used in this study are measured on lake Greifensee (Switzerland).

Previous work was done on modelling this lake with a complex mechanistic model. This model is one-dimensional, resolving the depth of the lake and uses 13 state variables. The model was developed by Omlin et al. 2001and applied to Greifensee by Mieleitner and Reichert 2005. Current work on modelling Greifensee with this model approach is focused on developing a box model (i.e. the lake is described with four boxes) and on improving the plankton sub-model. In this research we reduce the lake description to 0-dimensional, i.e. one-box model, and three state variables. Thus, we are setting a model of three ordinary differential equations.

2 AUTOMATED MODELLING FRAMEWORK

The machine learning method, used in this paper, supports introduction of the background modelling knowledge in the procedure of model induction from data. The knowledge provides a recipe for building models in the domain of interest – it provides (1) taxonomy of basic process classes in the domain, (2) commonly used modelling alternatives for the processes in these classes, as well as (3) rules for combining the models of individual processes into the model of the whole observed system. The knowledge

library used here provides knowledge for modelling of food webs in lakes, following the mass conservation principle. The models are based on ordinary differential equations. For further details see (Atanasova, 2005).

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

3 LAKE GREIFENSEE DATA

3.1 LAKE DESCRIPTION

Greifensee is located in Switzerland with a watershed area of 163 km^2, and maximal and average depth of 32 m and 18 m respectively. The surface area of the lake measures 8.5 km^2, while the volume is 148 millions m^3. The lake has an average discharge of 4.08 m^3/s and an average retention time of 1.1 years. In the 1960s the lake was highly eutrophic with average phosphate concentrations of over 500 mg/m^3. The lake begun to recover around the 1970s after some measures have been taken to improve the water quality. Now, the average phosphate concentration in Greifensee has been reduced to 100 mgP/m^3. It is still relatively high and corresponds to eutrophic state (Bürgi 1994).

3.2 THE DATA: SOURCES AND DESCRIPTION

Input data to the lake. The input data obtained from the AWEL (Amt für Abfall, Wasser, Energie und Luft, Switzerland), include daily measurements of two river inflows, i.e. Aabach-Mönchaltdorf and Aabach-Niederuster. The measurements include the flow rates, temperature, pH, Oxygen, ammonia, nitrite, nitrate, total nitrogen, phosphate, and total phosphorous. Of the meteorological data we use the global solar radiation obtained from the Swiss Meteorological Institute (MeteoSchweiz). Values measured hourly were converted to daily averages.

*Chemical and Physical variables in the lake***:** Monthly measurements in the period of 1988 to 1999 were obtained from the Swiss Federal Institute of Aquatic Science and Technology (Eawag). We use averaged data of temperature, phosphate and chlorophyll (measured at the deepest location).

Biological variables in the lake: Monthly to weekly measurements of the years 1987-1999 were obtained from the Eawag. The data set comprised depth-integrated samples of phytoplankton and zooplankton (measured at the deepest location). Phytoplankton and zooplankton concentration data consist of counts of many different species. The total volume of all zooplankton species was used. The volumes were calculated by multiplying the counts of each species by the typical volume of one cell of this species. The volume was converted to wet weight (WW) using the density of water. For the conversion from wet to dry weight (DW), which is modelled, a factor of 10 was used for zooplankton (based on measurements of the Water Supply Authority of Zürich).

3.3 DATA PREPARATION FOR MODELLING

Previous experiences with Lagramge indicate that daily data are needed for discovering ordinary differential equations. Therefore we interpolated the monthly data by cubic spline interpolation to get a convenient data set of "*daily*" measurements of the variables measured in the lake for induction of differential equations with Lagramge. The variables used for model induction are depicted in Table 1.

Table 1: Description of the variables

Variable name	Description	Units
v	Volume of the lake	m^3
d	depth	m
q_uster	Inflow to the lake	m^3/day
q_moe	Inflow to the lake	m^3/day
po4_uster, po4_moe	phosphorus concentration in the inflows	mg/l
load_uster_po4, load_moe_po4	calculated load to the lake: po4*q for both inflows	m^3/day
temp	Water temperature	°C
light	Averaged daily light	W/m^2
po4	Inorganic phosphorus concentration in the lake	mg/l
Chl_a	Chlorophyll concentration	mg/l
zoo	Zooplankton biomass concentration	mg DW/l

4 EXPERIMENTAL SETUP

The experiments were aimed at discovering a simple lake model for prediction of the relevant state variables that describe the trophic state of the lake, i.e. phosphorus and chlorophyll_*a* concentrations. The basic concept of such a model can be represented as shown in Figure 1. The concept consists of three state variables: inorganic dissolved phosphorus (*po4*), phytoplankton, represented as chlorophyll_a (*Chl_a*), and zooplankton (*Zoo*). The state variables are influenced by the biological processes (arrows in Figure 1) that take place in the system. The phytoplankton concentration (*Chl_a*) is increasing due to consumption of the nutrient (*po4*) in the process

PP_growth and decreasing due to the processes of respiration, sedimentation and grazing by zooplankton, i.e. *Respiration_PP*, *Sedimentation*, and *Feeds_on* (Figure 1). The last contributes to the zooplankton concentration (*Zoo*). Zooplankton is lost due to the processes of *Respiration_A* and *Mortality_A*. In contrast, the nutrient concentration increases due to both respiration processes (*Respiration_PP* and *Respiration_A*), and decreases due to phytoplankton growth (*PP_growth*). The described modelling concept in Figure 1 represents the expert knowledge of this system that is introduced in the procedure of model induction from measured data.

The modelling knowledge library specifies how to combine the processes into a corresponding model of the whole system (Džeroski and Todorovski, 2003; Todorovski, 2003; Atanasova et al., 2005). According to the combining rules from the knowledge library, the processes defined in the system specification (Figure 1), will be composed in a model based on three differential equations (1), (2), and (3).

$$\frac{dpo4}{dt} = \text{(Loads) - (Outflows) } + \text{const} \cdot \text{Respiration\_A} + \text{const} \cdot \text{Respiration\_PP} + \text{Sediment\_release} - \text{const} \cdot \text{PP\_growth} \quad (1)$$

$$\frac{dChla}{dt} = \text{PP\_growth - Respiration\_PP - Sedimentation - Feeds\_on} \quad (2)$$

$$\frac{dzoo}{dt} = const \cdot \text{Feeds\_on - Respiration\_A - Mortality\_A} \quad (3)$$

Each of the processes is represented with several mathematical formulations in the knowledge library. This specifies the space of candidate models for this system, which are further fitted to the given measurements. The fitting (training) was performed on one year measured data from 1989.

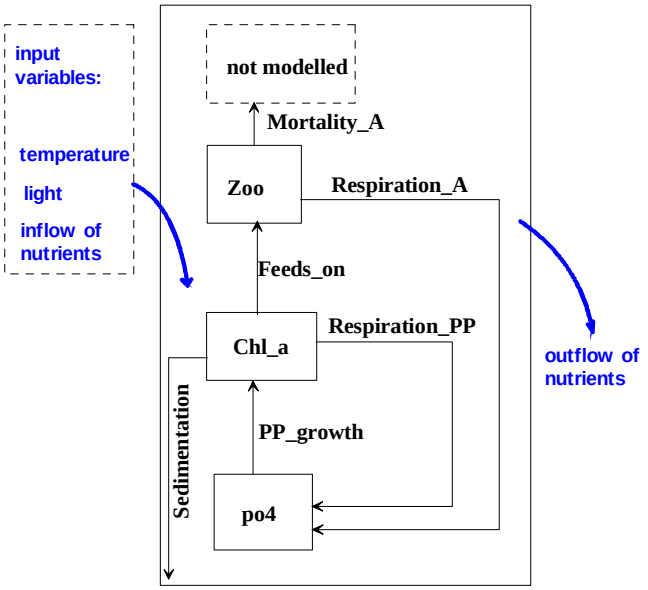

Figure 1: *The conceptual model of the Greifensee lake.*

5 RESULTS

Given the expert knowledge as described in section 4 (and using the lake modelling library) Lagramge discovered 28 224 candidate models, which were fitted and evaluated against the given measurement data from year 1989. The models were evaluated by two error measures. To rank the first ten models (best fitted to the measurements) we used the error measure included in Lagramge, i.e., MSE and MDL (see section 2). These best models were then evaluated according to the visual perception of the expert. Note that models with lowest MSE are not necessarily the best according to the domain experts. The best evaluated model using both error measures is presented in equations (4), (5), and (6).

The model equations follow the background knowledge as presented in equations (1), (2), and (3) in section 4. Each term in the equations (4), (5), and (6) represents the formulation of the processes in the equations (1), (2), and (3), correspondingly. For example, the process PP_growth is formulated as evident from the last term of the equation (4) and the first term of the equation (5).

$$\frac{dpo4}{dt} = \frac{load\_uster\_po4}{v} + \frac{load\_moe\_po4}{v} - po4 \cdot \frac{q\_uster}{v} - po4 \cdot \frac{q\_moe}{v} + 0.01 \cdot zoo \cdot 0.002 \cdot \frac{temp-4}{20-2.6} + 0.5 \cdot chla^2 \cdot 0.15 \cdot 1.13^{(temp-20)} + \frac{0.01+0.0021 \cdot temp}{d} - 0.088 \cdot chla \cdot 2.88 \cdot \frac{po4}{po4+10^{-7}} \cdot \frac{temp-2}{20-4} \cdot \frac{light}{light+30} \quad (4)$$

$$\frac{dchla}{dt} = chla \cdot 2.88 \cdot \frac{po4}{po4+10^{-7}} \cdot \frac{temp-2}{20-4} \cdot \frac{light}{light+30} - chla^2 \cdot 0.15 \cdot 1.13^{(temp-20)} - chla \cdot \frac{0.9}{d} - zoo \cdot 9.6 \cdot 1.11^{(temp-20)} \cdot \frac{chla}{chla+0.0014} \cdot chla \quad (5)$$

$$\frac{dzoo}{dt} = 0.99 \cdot zoo \cdot 9.6 \cdot 1.11^{(temp-20)} \cdot \frac{chla}{chla+0.0014} \cdot chla - zoo \cdot 0.0021 \cdot \frac{temp-4}{20-2.6} - zoo \cdot 0.047 \cdot \frac{temp-4}{20-3.6} \quad (6)$$

We simulated the model over a period of four years, i.e. 1988 to 1991. Recall that the model was calibrated on the one year (1989) measurements, while the rest of the simulation period is used for model validation. Figure 2 shows the simulated data together with the measurements. Phosphorus simulation performs with slight shift in time, which is acceptable from the expert point of view. The simulation of chlorophyll-a shows a distinctive local dynamics, unlike the other two state variables. The local dynamics of the state variable can be neither rejected nor confirmed from the measurements. Note, that the time scale of the measured data is one month, while the simulated data have daily time scale. Moreover, an expert would expect more dynamics then evident from the monthly measurements. It is especially encouraging that the model remains stable in spite of the distinctive local dynamics.

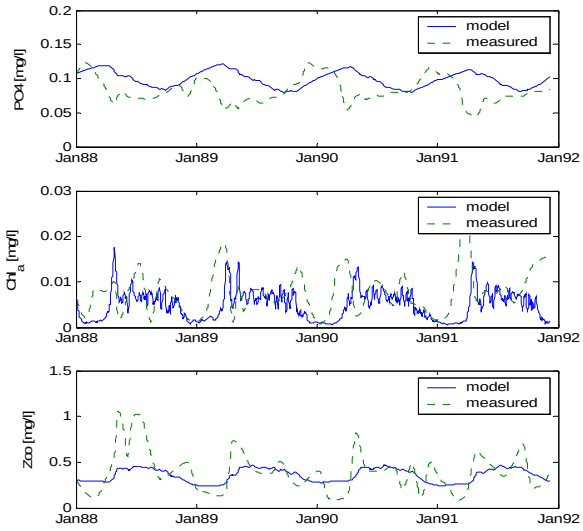

Figure 2: *Simulation of the model (equations 4, 5 and 6) over a period of four years. Simulated data (solid line) are plotted together with the measured data (dotted line)*

6 CONCLUSION

A modelling method that combines empirical and theoretical approach to modelling has been applied to real world data, i.e. the lake Greifensee. We discovered a simple model of three ordinary differential equations that simulates the behaviour of three state variables, i.e. phosphorus, chlorophyll-a, and zooplankton. The model was trained (identified) on one year's data and validated on a period of four years. The validation shows good behaviour of the model in terms of fitting to the data measurements and long term stability from the expert's point of view, regarding the complexity of the system.

ACKNOWLEDGEMENTS

We thank Hans Rudolf Bürgi form Eawag for providing data on chemistry and biology of Greifensee, the Swiss Meteorological Institute (MeteoSchweiz) for providing light intensities, and Pius Niederhauser from AWEL (Amt für Abfall, Wasser, Energie und Luft, Switzerland) for providing input data.

References

Atanasova, N. 2005. Preparation and use of the domain expert knowledge for automated modelling of aquatic ecosystems. PhD Thesis, University of Ljubljana, Ljubljana, Slovenia.

Atanasova, N., Todorovski, L., Džeroski, S., Rekar-Remec, Š., Recknagel, F., Kompare, B. 2005. Automated modelling of a food web in Lake Bled using measured data and a library of domain knowledge. *Ecological Modelling*. In press.

Bürgi, H. R., 1994. Seeplankton und Seesanierung in der Schweiz. Limnological Reports Danube.

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

DeAngelis, D. L. (1992): *Dynamics of Nutrient Cycling and Food Webs*. Chapmann & Hall. London 0 412 29830 9 (HB).

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

Jørgensen, S. E. and Bendoricchio, G. (2001): *Fundamentals of Ecological Modelling*. Elsevier 0-080-44028-2.

Mieleitner, J. and Reichert, P. (2005). Analysis of the transferability of a biogeochemical lake model to lakes of different trophic state. Ecological Modelling. In press.

Omlin, M., Reichert, P. and Forster, R. (2001). Biogeochemical model of Lake Zürich: Model equations and results. Ecological Modelling 141(1-3), 77-103.

Todorovski, L. 2003. Using Domain Knowledge for Automated Modeling of Dynamic Systems with Equation Discovery. Uni. of Ljubljana, Ljubljana, Slovenia.