University of Ljubljana
Faculty of Civil and Geodetic Engineering
Institute of Sanitary Engineering

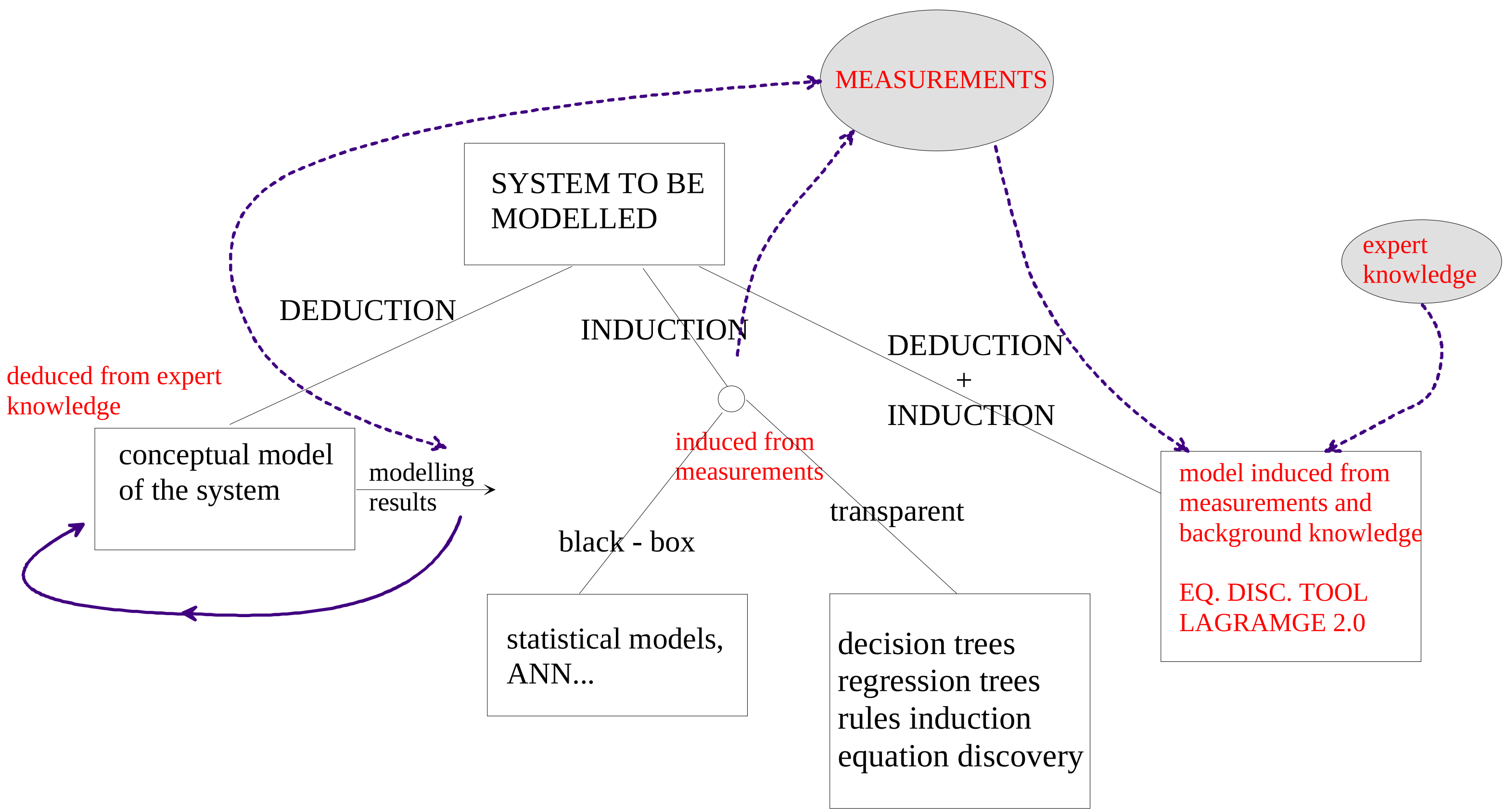
Domain library construction for knowledge-based equation discovery: An application in limnology

Nataša Atanasova, Ljupčo Todorovski, Boris Kompare, Sašo Džeroski

Overview

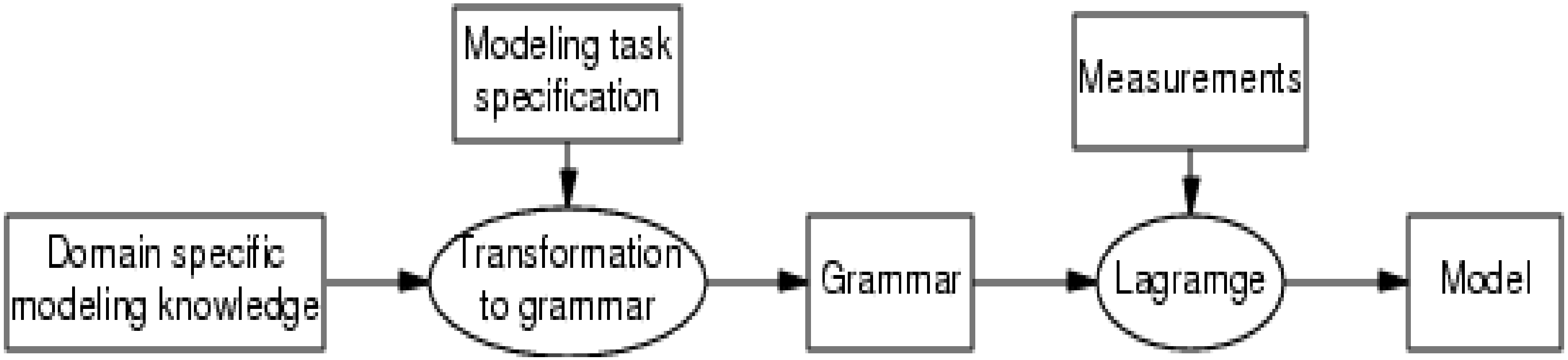
- Motivation for automated modelling
- Domain knowledge for modelling of population dynamics in lake ecosystems
- Formalization of the domain knowledge for Lagramge 2
- Application of Lagramge 2: Lake Bled

Motivation for automated modelling-
modelling procedures



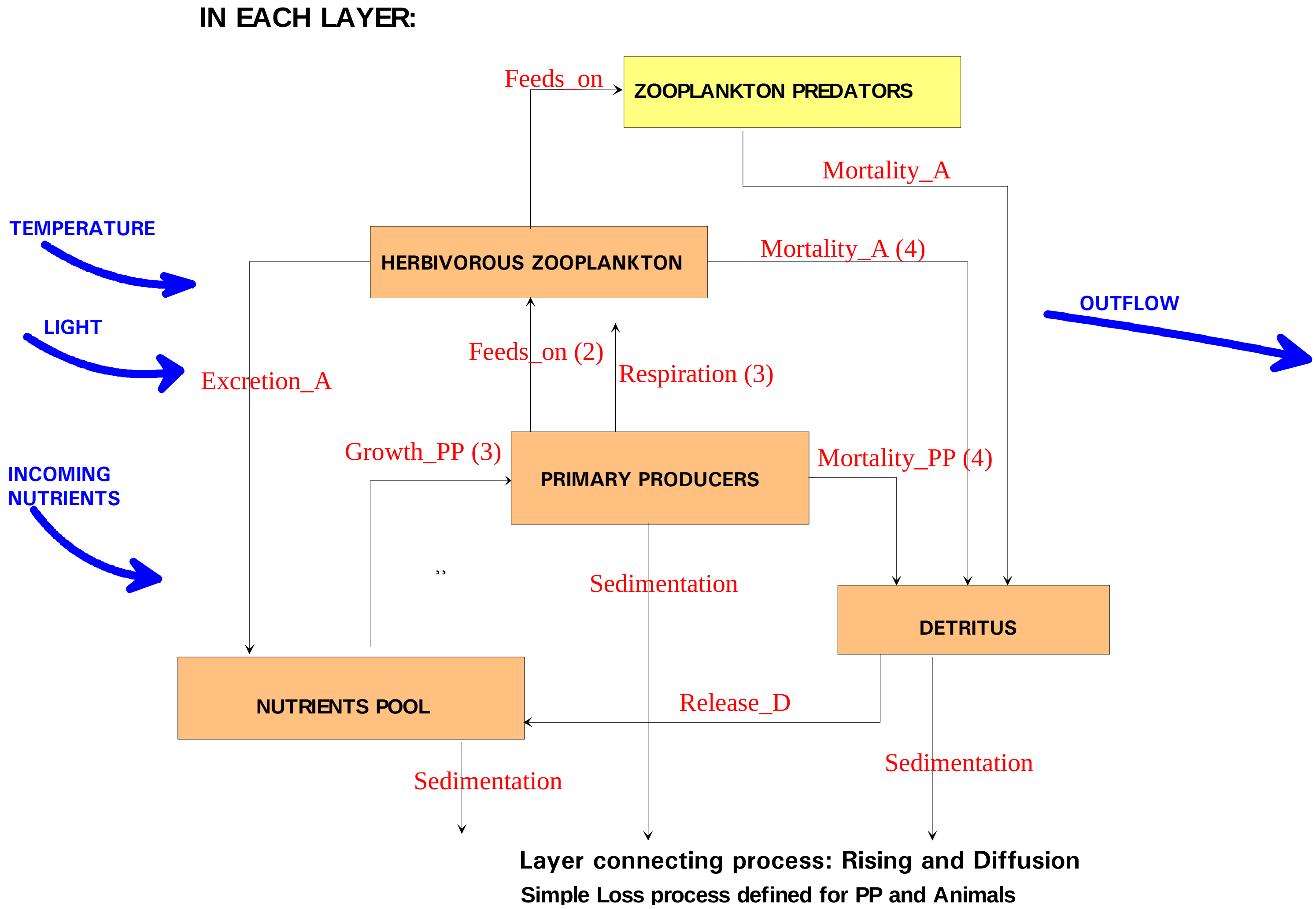
LAGRAMGE 2: how it works?

Todorovski (2003)



- from task specification and knowledge library to grammar
- using the grammar for equation discovery with lagramge

Generalized scheme of population dynamics processes and variables



Formalized population dynamics knowledge

an example for primary producer mass balance

Mass balance

dpp/dt

Processes

Growth(pp,{T},{L},{N})

type 1: pp*m_{max}* f(T)*f(L)*f(N)

type 2: pp*m_{max}* f(T, Topt)*f(L)*f(N)

type 3: pp*m_{max}* f(T, L)*f(N)

Mortality (pp, {T},{L},{N})

type 1:...

type 2:...

type 3:...

Sedimentation (pp, {T})

Respiration (pp, {T},{L},{N})

type 1:...

type 2:...

Feeds_on (Zoo,{pp},{T})

type 1: C_{fmax}* f(T) * f(food) *pp*zoo

type 2:C_{gmax}* f(T) * f(food) *zoo

Functions

$$f(T) = \frac{T - T_{\min}}{T_{ref} - T_{\min}}$$

$$f(T) = \Theta^{(T - Tref)}$$

$$f(T) = \exp\left(-2.3\left|\frac{T - T_{opt}}{T_{opt} - T_{\min}}\right|\right)$$

$$F(L) = \frac{I}{K_{sI} + I}$$

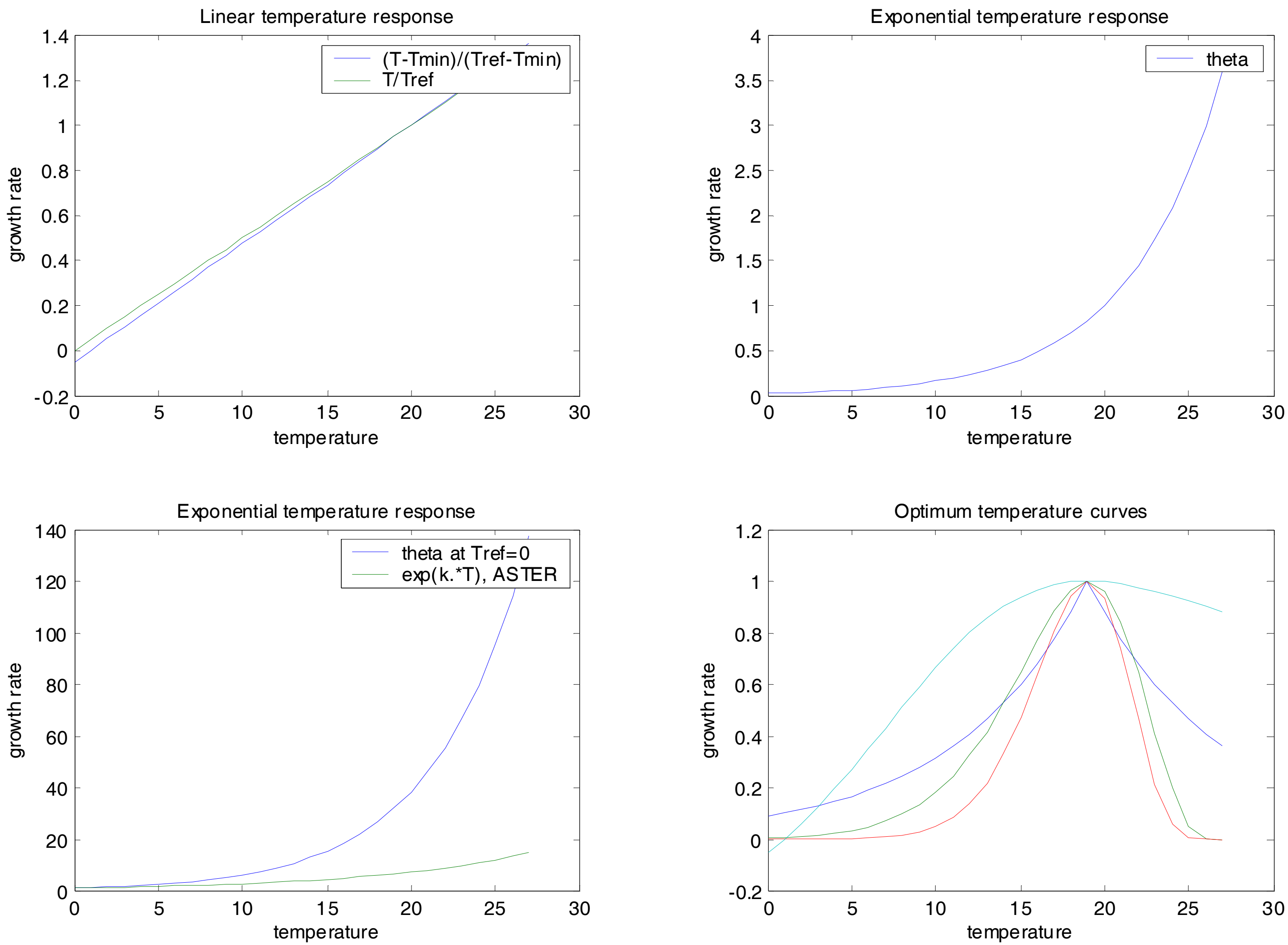
$$F(L) = \frac{I}{I_{opt}} e^{-\frac{I}{I_{opt}} + 1}$$

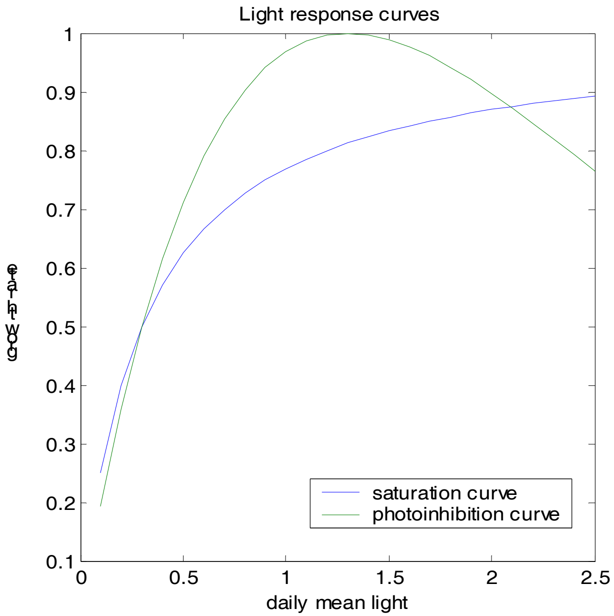
$$f(N) = \frac{N}{K_s + N}$$

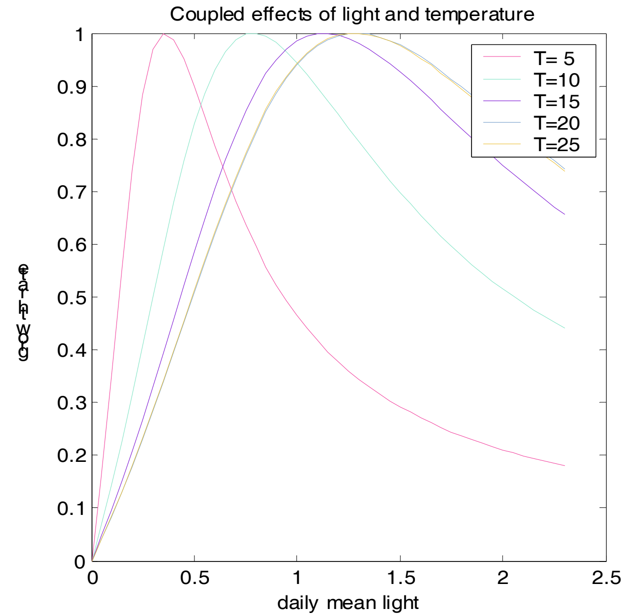
$$f(N) = \frac{N^2}{K_s + N^2}$$

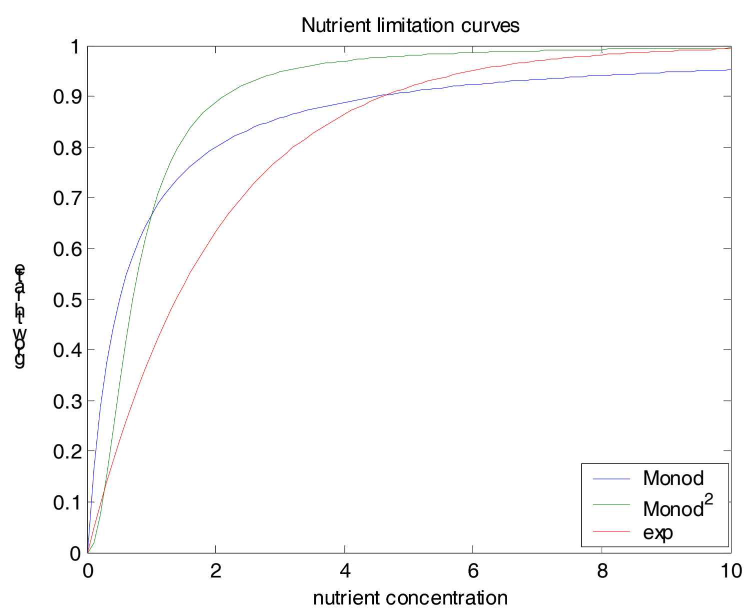
$$f(N) = 1 - \exp(-k * N)$$

Growth rate influences and limitations




Light response curves
daily mean light
saturation curve
photoinhibition curve


Coupled effects of light and temperature
daily mean light
T= 5
T=10
T=15
T=20
T=25


Nutrient limitation curves
nutrient concentration
Monod
Monod2
exp

Process formalisation in the library
the process Growth_PP

Growth of a primary producer (pp) is influenced by temperature and limited by light and nutrients

$$\frac{dpp}{dt} = \mu \cdot pp$$

most models $\mu = \mu_{\max}(T_{ref})\, f(T)\, f(L)\, f(P,N,C)$

Bendoricchio 1994, Topt is a variable $\mu = \mu_{\max}(T_{ref})\, f(T, Topt)\, f(L)\, f(P,N,C)$

Talbot et al., 1991: the Aster model, $\mu = \mu_{\max}(T_{ref})\, f(T, L)\, f(P,N,C)$

$$f(P,N,C) = f(P)\, f(N)\, f(C)$$

$$f(P,N,C) = \min\left[f(P), f(N), f(C)\right]$$

$$f(P,N,C) = \frac{f(P) + f(N) + f(C)}{n}$$

pp	primary producer concentration
μ	growth rate [1/time]
f(T)	temperature influence function
f(L)	light limitation function
f(N,P,C)	nutrients limitation function

Library

```
process class PP_growth (Primary_producer pp, Inorganics ns,
   Temperatures ts, Lights ls)

   process class PP_growth_type_1() is PP_growth
       expression pp*const(max_growth_rate, 0.01, 0.1, 4)*
       Food_limitations(ns)*Temp_influences(ts)*
       Light_limitations(ls)

   process class PP_growth_type_2() is PP_growth
       expression pp*const(max_growth_rate, 0.01, 0.1, 4)*
       Food_limitations(ns)*Light_limitations(ls)*
       product({t1,t2}, t1 in ts, Temp_influence_opt(t1,t2))

   process class PP_growth_type_3() is PP_growth
       expression pp*const(max_growth_rate, 0.01, 0.1, 4)*
       Food_limitations(ns)*
       product({t,l}, t in ts, l in ls, Light_temp(t, l))
```

Combining scheme (mass balances)

…

```
combining scheme Lake(Primary_producer pp)
    time_deriv(pp) =
        + sum({food, ts, ls}, true, PP_growth(pp, food, ts, ls))
        - sum({ts}, true, Loss(pp, ts))
        - sum({ts,ns}, true, Respiration_PP(pp, ts, ns))
        - sum({ts,ns}, true, Mortality_PP(pp, ts, ns))
        - sum({}, true, Outflow(pp))
        - sum({}, true, Sedimentation(pp))
        + sum({pp1}, true, Rising(pp,pp1))
        - sum({a, food, ts}, pp in food, Feeds_on(a, food, ts))*pp
```

The knowledge base

- Supports:
  - Food chain modelling in a lake
  - 0 dimensional models
  - N box models i.e., supports modelling of stratified lakes
  - Fixed internal nutrient levels in primary producers and animals
- Complexity of the models is defined by the expert, i.e. number of state variables and processes

Example of automated modeling

Case study: Lake Bled

Vol. = 25.7 * 10^6 m^3

Area = 1.47 * 10^6 m^2

Max. depth = 30 m
(western basin)

Avg. depth = 17.5 m

Ae = 0.98*10^6 m^2

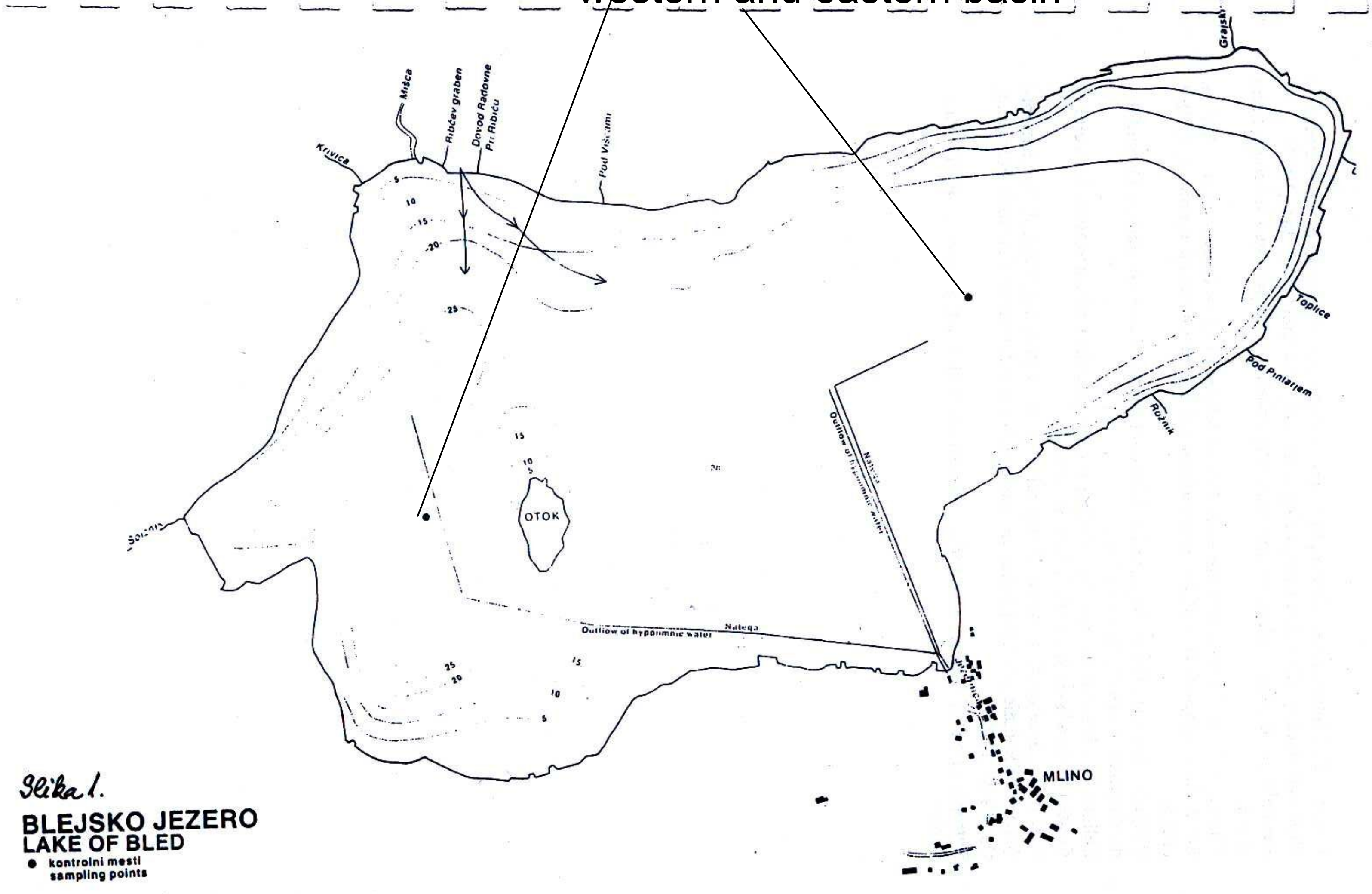
Aw= 0.49*10^6 m^2

Oscillatoria rubescens in Lake of Bled, November 1999. Photo by Mirko Kunšič

Inducing a food-chain model for Lake Bled

- Data:
  - Environmental Agency of the RS
  - monthly measurements of physical, chemical and biological data
  - Long term data set 1985 to 2002
  - Each variable is measured at two locations in the lake (western and eastern basin)
  - Samples are taken each two meters from the surface to the bottom
  - i.e, we have parameter measurements in two water columns

Sampling points in the western and eastern basin



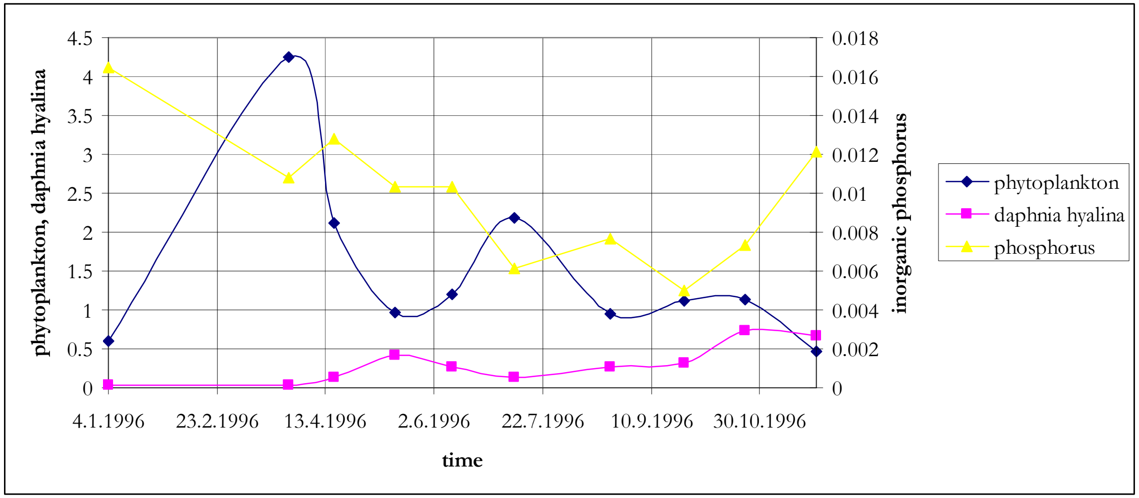
ARSO: Monitoring kakovosti ...

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

variable name	description	Units	Frequency
q_krivica	Inflow to the lake	m^3/day	daily
q_misca	Inflow to the lake	m^3/day	daily
q_radovna	Inflow to the lake	m^3/day	daily
q_jezernica	Outflow	m^3/day	daily
q_natega	Outflow	m^3/day	daily
ortp_krivica, ortp_misca, ortp_radovna	Nutrient (orthophosphate) concentration in the inflows	mg/l	monthly
temp	Water temperature of the streams and lake	°C	monthly
light	Light intensity	J/(m2*day)	monthly
ortp	Nutrient (orthophosphate) concentration in the lake	mg/l	monthly
phyto	Phytoplankton biomass concentration in the lake	mg/l	monthly
daph	Zooplankton (daphnia hyalina) biomass concentration in the lake	mg/l	monthly

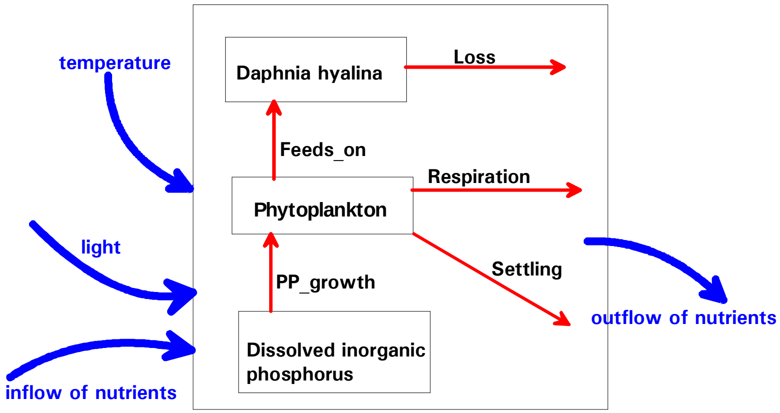
Inducing a model on real data

- Three d.e. model: ortp-phyto-daph
- Induced on eastern basin data, euphotic zone (upper 10 m), 1996



Expert knowledge

- General form of the equations



$$\frac{dortp}{dt} = \text{inflow - outflow - const} \cdot PP\_\text{growth}$$

$$\frac{dphyto}{dt} = PP\_\text{growth} - \text{respiration} - \text{settling} - \text{Feeds\_on}$$

$$\frac{ddaph}{dt} = Feeds\_on - loss$$

Task specification

variable Inorganic ortp_krivica
variable Inorganic ortp_misca
variable Inorganic ortp_radovna
variable Flow q_krivica
variable Flow q_misca
variable Flow q_radovna
variable Flow q_jezernica
variable Flow q_natega
variable Inorganic ortp
variable Primary_producer phyto
variable Animal daph
variable Temperature temp
variable Light light

process Inflow(ortp, ortp_krivica, q_krivica) inflow1
process Inflow(ortp, ortp_misca, q_misca) inflow2
process Inflow(ortp, ortp_radovna, q_radovna) inflow3
process Outflow(ortp, q_jezernica) outflow1
process Outflow(ortp, q_natega) outflow2
process PP_growth(phyto, {ortp}, {temp}, {light}) p1
process Feeds_on(daph, {phyto}, {temp}) p5
process Loss(phyto, {temp}) p6
process Respiration_PP(phyto, {temp},{}) p9
process Loss(daph, {temp}) p8

Step by step discovery of three d.e.

Phosphorus equation:

$$\frac{dortp}{dt} = (ORTP\_krivica \cdot \frac{q\_krivica}{7 \cdot 10^6} + ORTP\_misca \cdot \frac{q\_misca}{7 \cdot 10^6} + ORTP\_radovna \cdot \frac{q\_radovna}{7 \cdot 10^6})$$

$$-(ORTP \cdot \frac{q\_jezernica}{7 \cdot 10^6} + ORTP \cdot \frac{q\_natega}{7 \cdot 10^6}) - \quad \cdot \quad \cdot \frac{\qquad}{\quad + \quad} \cdot \quad {}^{-} \quad \cdot \frac{\qquad}{\quad + \quad}$$

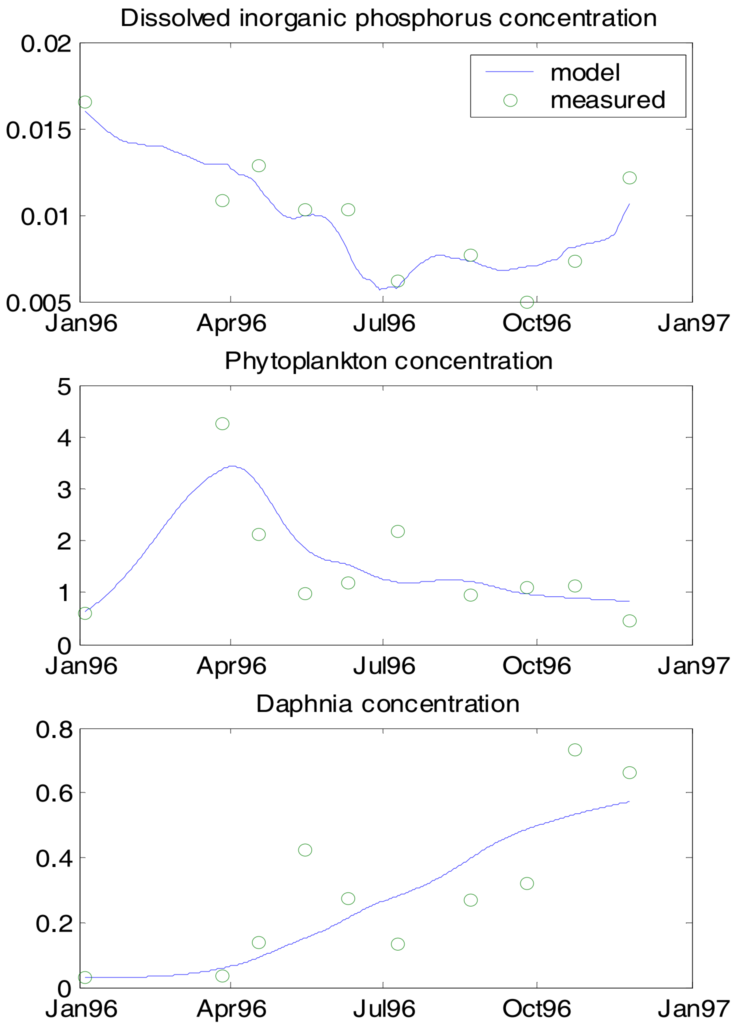
Using the growth term in the phosphorus eq., discover the rest of the processes in the phytoplankton eq.

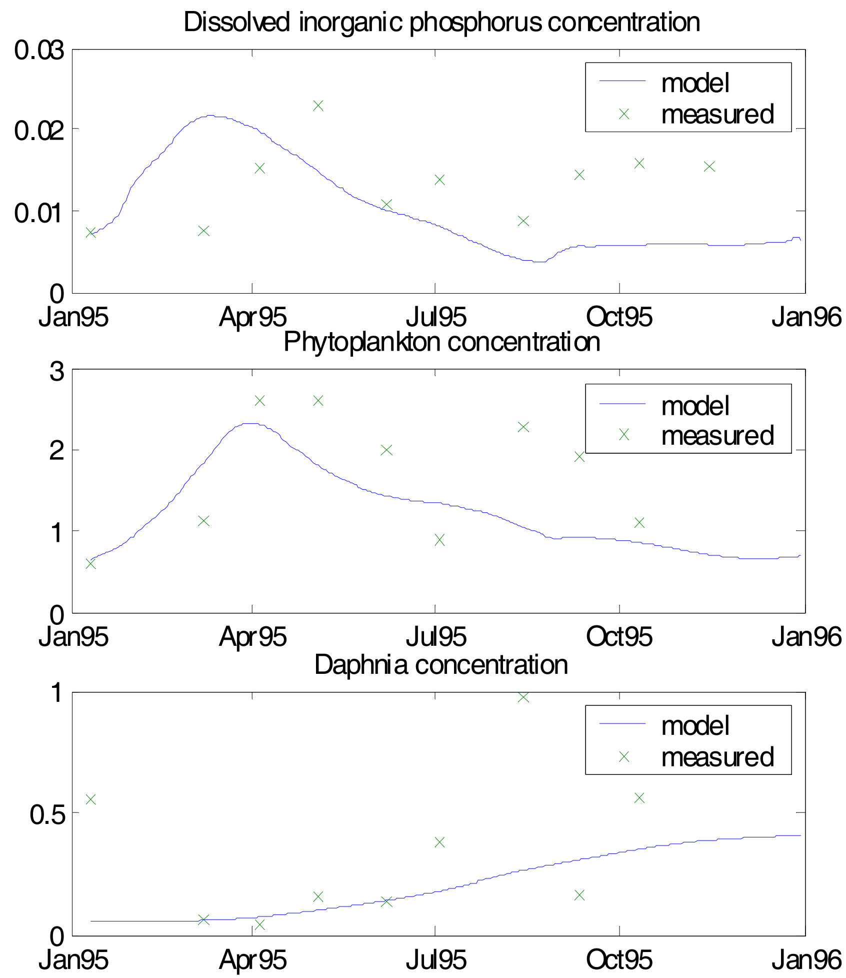
$$\frac{dphyto}{dt} = \quad \cdot \quad \cdot \frac{\qquad}{\quad + \quad} \cdot \quad {}^{-} \quad \cdot \frac{\qquad}{\quad + \quad} - phyto^2 \cdot 0.046 \cdot \frac{temp - 2.7}{18 - 4} -$$

$$phyto \cdot \frac{0.009}{10} - daph \cdot 1.01 \cdot \frac{temp}{20 - 3.5} \cdot \frac{phyto^2}{phyto^2 + 5} \cdot phyto$$

$$\frac{ddaph}{dt} = 0.02 \cdot daph \cdot 1.01 \cdot \frac{temp}{16.4} \cdot \frac{phyto^2}{phyto^2 + 5} \cdot phyto - daph^3 \cdot \frac{0.001}{5^2 + daph^2}$$

Model performance





Conclusions

- Domain knowledge can be successfuly introduced in the equation (model) discovery procedure
- Lake modelling library to support automated modelling of lakes with the Lagramge 2 machine learning tool was constructed.
- Library language formalism supports a precise formulation of the expert knowledge
- Lake Bled: induction of a simple model for complex dynamics
- Model improvement: increase of model complexity,

Acknowledgements

- Ministry of the Environment, Spatial Planning and Energy; Environmental Agency of the Republic of Slovenia
- Špela Rekar, Environmental Agency of the Republic of Slovenia