ELSEVIER

Ecological Modelling 141 (2001) 105–123

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecolmodel

Biogeochemical model of Lake Zürich: sensitivity, identifiability and uncertainty analysis

Martin Omlin, Roland Brun, Peter Reichert *

Swiss Federal Institute for Environmental Science and Technology (EAWAG), CH-8600 Dübendorf, Switzerland

Received 14 April 2000; received in revised form 13 December 2000; accepted 9 January 2001

Abstract

A model for the description of nutrient, oxygen and plankton dynamics in Lake Zürich, Switzerland has recently been developed. Because, with this model, an attempt is made to describe mechanistically the most important mass fluxes and conversion processes in the water column and sediment of the lake, it is already too complicated to allow all of its parameters to be identifiable from the monthly measured profiles. This raises the questions of how to select a subset of model parameters to be included in a formal parameter estimation process and how to estimate model prediction uncertainty. In this paper, a systematic approach to tackle this problem is applied to this model. The technique consists of the combination of an analysis of the sensitivity of model results to single parameters with an analysis of the approximate linear dependence of sensitivity functions of parameter subsets. It is demonstrated that the most severe parameter identifiability problems are caused by the parameterization of light dependence of algae growth, by competing effects of production, respiration and death of algae and zooplankton, and by the interactions between algae and zooplankton. The dynamics of dissolved variables is much easier to describe. The results of the analysis are used to select a parameter subset for a fit with measured data, to analyse the effect of other, fixed parameters on the estimates of the selected parameters, and to estimate the uncertainty of model predictions.

01 mingxi zhang

02 mingxi zhang

Keywords: Dependence analysis; Identifiability analysis; Lake Zürich; Parameter estimation; Sensitivity analysis; Uncertainty analysis; Water-quality modelling

1. Introduction

The parsimony principle of system identification states that a model should not be more complicated than necessary for the description of the data. This implies that its parameters are identifiable from the available data set (Spriet, 1985). While this principle is very important for gaining information on poorly known systems, there are also good reasons for using overparameterized, non-identifiable models for the description of environmental systems, especially in the context of prediction of expected changes under

* Corresponding author. Present address: Department of Systems Analysis, Integrated Assessment and Modelling, EAWAG, PO Box 611, 8600, Duebendorf, Switzerland. Tel.: +41-1-8235281; fax: +41-1-8235375.

E-mail address: reichert@eawag.ch (P. Reichert).

0304-3800/01/$ - see front matter

PII: S0304-3800(01)00257-5

changing environmental conditions (Beck, 1987, 1999; Reichert and Omlin, 1997; Reckhow and Chapra, 1999). The following goals of modelling often lead to detailed mechanistic descriptions of the system under consideration which, for typically available data sets, are overparameterized models:

- Summarizing the state of knowledge in a given field or even integrating knowledge from different fields.
- Creation of a model structure that is not specific to a given situation but transferable to similar systems.
- Design of a model with the goal to predict the behaviour of a system under different driving conditions than those observed.

In any case, careful thinking on the adequateness of a model structure used for an analysis is very important and should include not only the search for processes not yet included in the model that could become relevant in the future, but also the search for possibilities for model simplifications.

In this study, the parameter identifiability analysis techniques outlined in Brun et al. (2001) are applied to the biogeochemical model of Lake Zürich described in Omlin et al. (2001). The 52 parameters listed in the appendix of Omlin et al. (2001) certainly cannot all be determined from the monthly measured profiles of temperature, oxygen, nutrients and plankton in the lake and the information available on light and inflows. As mentioned above, before analysing the potential identifiability of parameter subsets, the question of the adequacy of model complexity for the goal of describing nutrient, oxygen and plankton dynamics in the lake should be addressed.

The model describes the dissolved compounds phosphate, ammonia, nitrate and oxygen and the particulate compounds algae without *Planktothrix rubescens*, *Planktothrix* (*Oscillatoria*) *rubescens*, zooplankton and biodegradable and inert dead organic particles. The model uses a variable stoichiometry of primary production with respect to phosphorus. Biogeochemical conversion processes considered are growth, respiration and death of algae, *P.* (*Oscillatoria*) *rubescens* and zooplankton, aerobic and anoxic mineralization, nitrification and phosphate uptake of sedimenting particles. It has been shown by Omlin et al. (2001) that a variable stoichiometry of phosphorus uptake by growing algae and a phosphate uptake process of sedimenting particles is required to describe the data adequately. The potential for model simplifications is thus limited. The mineralization processes could have been combined with respiration and death, and *P.* (*Oscillatoria*) *rubescens* could have been combined with the other algae to form a single algal class. However, the proposed approach of distinguishing respiration, death and mineralization is conceptually clearer and avoids the necessity for depth-dependent parameters as used in previous studies (Imboden and Gächter, 1978; Karagounis et al., 1993). We modelled *P.* (*Oscillatoria*) *rubescens* as a separate algal class not because this was a necessity for quantifying the dominant nutrient conversion processes but because this species was of special interest to the local water supply authority. This discussion shows that there is only a small potential for model simplifications that is not sufficient to make the model parameters identifiable from the available data. For this reason, the analysis of the identifiability of subsets of its parameters is an important consideration in this paper.

This paper presents the results of the application of the procedure described in Brun et al. (2001) to the lake model mentioned above. Sections 3 and 4 contain the main results of this procedure. Because the procedure is iterative, and only the results of the final pass through its steps are outlined, the results discussed in each section depend on the previous passes through all steps. The paper is structured as follows. In Section 2 a brief overview of the techniques used for parameter estimation is given. This section also introduces some aspects of notation. In Section 3, prior estimates of parameter uncertainty and linear error propagation techniques are used to produce a ranking of the model parameters according to their influence on mean model prediction uncertainty. In Section 4, subsets of parameters are analysed for potential identifiability problems caused by the possibility of partial compensation of an effect on the results caused by a change of one parameter in the set by an appropriate change

of the other parameters. The results of this analysis are used, together with the sensitivity ranking to select a subset of parameters to be estimated and to assess the effect of changes of the values of parameters that were not fitted on the estimates of fitted parameters. In Section 6, prior uncertainty estimates together with the results of the analysis performed in the previous sections are used to estimate the uncertainty of model predictions. Finally, in Section 7, the results are summarized, and conclusions are drawn.

2. Parameter estimation

The selection of the parameter subset to be estimated from the lake data is discussed in Section 4. In this section, the introduction of the sum of weighted squares of the residuals, which is minimized by the parameter estimation algorithm, is used to introduce some aspects of notation.

The weighted least-squares parameter estimation procedure applied in this study minimizes the sum

$$WSS(\boldsymbol{\theta}) = \sum_{k=1}^{n_y} \sum_{j=1}^{n_{t,fit}} \sum_{i=1}^{n_{y_k,z}} \left(\frac{y_{meas,k,j,i} - y_k(z_{y_k,i}, t_j, \boldsymbol{\theta})}{sc_{y_k}} \right)^2 \quad (1)$$

in order to find the parameter estimates. In this equation, the index i is used to distinguish spatial locations, $z_{y_k,i}$ (the locations are not the same for all relevant model variables, y_k, because dissolved substances are measured in a larger number of depths than plankton); it runs from 1 to the number of sample locations for model variable y_k, $n_{y_k,z}$. The index j is used to distinguish different points in time, t_j; it runs from 1 to the number of points in time at which measured lake profiles were used for the fit, $n_{t,fit}$. The index k is used to distingish different model variables, y_k; it runs from 1 to the number of model variables to be compared with data, n_y. $y_k(z_{y_k,i}, t_j, \boldsymbol{\theta})$ is the result of the lake model for the model variable y_k calculated with the parameter values $\boldsymbol{\theta}$ and evaluated at location $z_{y_k,i}$ and at time t_j. $y_{meas,k,j,i}$ is the measured value corresponding to the model variable y_k sampled at location $z_{y_k,i}$ at time t_j. In the sum of weighted squares (Eq. (1)), the residuals, $y_{meas,k,j,I} - y_k(z_{y_k,i}, t_j, \boldsymbol{\theta})$, are divided by the scale, sc_{y_k}, of the variable y_k. This scale has the same dimensions as the model variable y_k and makes the terms in the sum (Eq. (1)) non-dimensional. This is very important, because different model variables used for parameter estimation may have different dimensions. If the value of sc_{y_k} is selected to be equal to the standard deviation of the measurement of the variable y_k, σ_{y_k}, then the sum WSS is equal to χ^2. This makes it possible to use its value in order to assess the quality of the fit. In our application, however, the values of sc_{y_k} are used as empirical weights with the goal to make the contributions of different model variables to WSS similar in size and, therefore, give all measured variables a similar influence on the estimates of the parameters. Table 1 lists the values of sc_{y_k} used for the fit. Profiles for the five model variables oxygen, phosphate, nitrate, algae [other than *P. (Oscillatoria) rubescens*] and zooplankton were used for parameter estimation. Data for *P. (Oscillatoria) rubescens*, which was also available, were not used for the fit because of difficulties with the spatial resolution of the sharp *P. (Oscillatoria) rubescens* peaks in calculation and measurements. Model parameters used for the description of *P. (Oscillatoria) rubescens* were adjusted by hand, instead. The values shown in Table 1 reflect the sharp metalimnic minimum in oxygen concentrations that had to be compensated for by a large value for $sc_{S_{O_2}}$ and the smaller number of zooplankton measurements in comparison to the algae measurements that were compensated by a smaller value of $sc_{S_{ZOO}}$ in comparsion to $sc_{S_{ALG}}$ (zooplankton measurements were spatially integrated over the epi- and hy-

Table 1
Scales used as empirical weights for making the contributions of different model variables to the weighted sum of squares (Eq. (1)) non-dimensional

Scale	Value	Unit
$sc_{S_{O_2}}$	1.0	gO m^{-3}
$sc_{S_{HPO4}}$	0.025	gP m^{-3}
$sc_{S_{NO3}}$	0.1	gN m^{-3}
$sc_{X_{ALG}}$	0.5	gWM m^{-3}
$sc_{X_{ZOO}}$	0.2	gWM m^{-3}

polimnion, whereas algae were sampled with a higher spatial resolution).

3. Sensitivity analysis

In this section, the model parameters are grouped into parameters potentially to be estimated from lake data and parameters for which an estimation from the available data of Lake Zürich did not seem to be meaningful. Then, based on prior knowledge, for each parameter, a relative range of uncertainty is estimated subjectively. Finally, the uncertainty ranges of the parameters together with linear error propagation through the lake model are used to obtain a ranking of the sensitivity of model results with respect to all parameters. The analysis here is done as proposed by Brun et al. (2001), and it focuses on the local sensitivity of model results to model parameters. A similar ranking was done by Nordhaus (1995) based on the non-linear response of results to 'large' changes in model parameters and on relative changes of model results.

The parameters of the lake model can be classified into four types: physical parameters, stoichiometric parameters, kinetic parameters and parameters related to input fluxes into the lake (all parameters of the first three types are listed in the appendix of Omlin et al., 2001). Because of the uncertainty in the rate expressions and in the parameter values of the biogeochemical model, it did not seem to be meaningful to estimate inflow parameters from lake data (inputs to a lake are often not measured precisely, but their estimation from lake data would only be meaningful if the biogeochemical processes in the lake were known accurately). Stoichiometric and physical parameters are usually more accurately known or show a smaller variation from one system to another than parameters of process kinetics. This leads to the general guideline to estimate primarily kinetic model parameters.

It is very difficult to estimate the prior uncertainty of model parameters. In addition, the large number of parameters makes it extremely time-consuming to make a search on published values and uncertainties. In order to keep the expense within reasonable bounds while making the assessment not too subjective, only three classes of relative uncertainty were distinguished, and parameters were classified systematically into these classes. The three classes were accurately known parameters (class 1), very poorly known parameters (class 3) and an intermediate class of moderately inaccurate parameters (class 2). The uncertainty range was selected to be comparable with a standard deviation of the distribution, and the relative uncertainties were selected to be 5% for class 1, 20% for class 2 and 50% for class 3. This means that under normality assumptions, the value of class 3 variables would not be significantly different from zero at a 95% confidence level. A general criterion was to classify stoichiometric parameters and specific growth rates into class 2, and other kinetic parameters into class 3. Input-related parameters were typically classified into class 2, and physical parameters had to be treated individually. Tables 2–4 show the selected uncertainty class of potential fit parameters, model parameters not to be estimated from lake data and input parameters, respectively. The tables also contain the parameter value selected for the final model simulation. Note that the prior selection of a relative rather than an absolute uncertainty range makes the absolute uncertainty ranges dependent on the posterior estimate of the model parameter used for the final simulation.

Table 2 lists the parameters potentially to be estimated from lake data. The only physical parameter is the thickness of the sediment layers, h_{sed}. This is an important parameter because it limits the maximum possible flux of dissolved substances from the water column into the sediment (through a limitation of the maximum possible gradient given the concentration in the water column). This is an empirical parameter because it is only present due to the rough approximation of the sediment by only two layers. This parameter is assumed to be very inaccurately known (class 3). The stoichiometric parameters potentially to be estimated are related to the HPO_4^{2-} concentration in the water column that leads to algal growth with a smaller phosphorus stoichiometry than the Redfield phosphorus content. $S_{HPO_4,crit}$ is the criti-

Table 2
Potential fit parameters with uncertainty class, parameter estimate and prior estimate of uncertainty range (depending on uncertainty class 1 = 5%, 2 = 20% and 3 = 50%)[a]

Parameter	Uncertainty class	Value θ_l	Uncertainty range $\Delta\theta_l$	Unit
h_{sed}	3	0.0036	0.0018	m
$S_{HPO_4,crit}$	2	0.0042	0.0008	gP m^{-3}
ΔS_{HPO_4}	2	0.0013	0.0003	gP m^{-3}
k_{death,ALG,T_0}	3	0.030	0.015	d^{-1}
k_{death,ZOO,T_0}	3	0.029	0.014	d^{-1}
k_{gro,ALG,T_0}	2	1.13	0.23	d^{-1}
k_{gro,ZOO,T_0}	2	0.30	0.06	gDM^{-1} m^3 d^{-1}
$k_{miner,aero,sed,T_0}$	3	0.10	0.05	d^{-1}
$k_{miner,aero,wat,T_0}$	3	0.010	0.005	d^{-1}
$k_{miner,anox,sed,T_0}$	3	0.10	0.05	d^{-1}
$k_{miner,anox,wat,T_0}$	3	0.010	0.005	d^{-1}
k_{nitri,sed,T_0}	3	0.50	0.25	gN m^{-3} d^{-1}
k_{nitri,wat,T_0}	3	0.10	0.05	gN m^{-3} d^{-1}
k_{resp,ALG,T_0}	3	0.050	0.025	d^{-1}
k_{resp,ZOO,T_0}	3	0.0030	0.0015	d^{-1}
k_{upt}	3	1200	600	gDM^{-1} m^4 d^{-1}
$K_{I,ALG}$	3	34	17	W m^{-2}
$K_{NH_4,nitri}$	3	0.50	0.25	gN m^{-3}
$K_{NO_3,ALG}$	3	0.040	0.020	gN m^{-3}
$K_{NO_3,miner}$	3	0.10	0.05	gN m^{-3}
$K_{O_2,upt}$	3	0.50	0.25	gO m^{-3}
$K_{O_2,miner}$	3	0.20	0.10	gO m^{-3}
$K_{O_2,nitri}$	3	0.40	0.20	gO m^{-3}
$K_{O_2,resp}$	3	0.50	0.25	gO m^{-3}
$K_{HPO_4,ALG}$	3	0.0019	0.0009	gP m^{-3}

[a] The horizontal lines separate physical, stoichiometric and kinetic parameters.

cal concentration for the transition, and ΔS_{HPO_4} is the width of the transition range. Both parameters are assumed to belong to the uncertainty class 2 because there is knowledge on typical HPO_4^{2-} concentrations during the summer and during the other seasons from several lakes (the change in phosphorus stoichiometry must take place at concentrations between these values). All the other parameters listed in Table 2 are kinetic parameters. In fact, these are all kinetic parameters of the model with the exception of the kinetic parameters for *P.* (*Oscillatoria*) *rubescens*, which were excluded because of the problems with the insufficient spatial resolution for fit mentioned in Section 2, and with the exception of the temperature-dependence coefficients of the microbiological processes. All k parameters are rate constants directly proportional to the process rate, all K parameters are half-saturation concentrations of processes with respect to limiting light intensities, oxygen concentrations or nutrient concentrations, and the indices death, gro, miner, nitri, resp and upt, refer to death, growth, mineralization, nitrification, respiration and phosphate uptake processes, respectively (Omlin et al., 2001). In accordance with the strategy described above, all kinetic parameters with the exception of the specific growth rates of algae and zooplankton were classified into class 3.

Table 3 lists the model parameters that are not to be estimated from lake data. The two light-absorption parameters, k_1 and k_2, were determined by evaluations of light data from Lake Zürich and are therefore classified into class 2. The oxygen

exchange velocity across the lake surface was also classified into class 2. The sedimentation velocities of algae and dead organic particles, $v_{sed,ALG}$ and $v_{sed,ORG}$, are very inaccurate because, depending on the nature of the particles, a wide range of sedimentation velocities have been measured. A description of this spectrum of velocities by just two sedimentation velocities is a rather crude approximation made by the model. For this reason, the sedimentation velocities belong to class 3. The stoichiometric parameters include nitrogen and phosphorus fractions of biomass (a_N and $a_{P,red}$), maximum fraction of adsorbed phosphate on organic particles ($a_{P,max}$), maximum and minimum phosphorus content of newly built algae during primary production ($b_{P,max}$ and $b_{P,min}$), coefficients for excretion (c_e) and for build-up of inert organic material during death (f_p), coefficients for the conversion of dry mass to wet mass (w_{ALG}, w_{ORG}, w_{ZOO}), and maximum yield of zooplankton growth ($Y_{ZOO,max}$). The kinetic parameters not to be fitted are those related to growth, respiration and death of *P.* (*Oscillatoria*) *rubescens* (k_{death,PLR,T_0}, k_{gro,PLR,T_0}, $K_{I,PLR}$ and k_{resp,PLR,T_0}), and coefficients for temperature dependence of growth and respiration processes of algae, bacteria, *P.* (*Oscillatoria*) *rubescens* and zooplankton (β_{ALG}, β_{BAC}, β_{PLR} and β_{ZOO}).

Table 4 lists the input parameters for lake simulations. Similarly to the parameters listed in Table 3, these parameters are not to be estimated from lake data. The parameter $a_{P,inflow}$ describes the phosphorus content of organic particles discharged into the lake. The f parameters describe fractions of dissolved inorganic nitrogen in the

Table 3
Parameters not to be estimated from lake data with uncertainty class, parameter estimate and prior estimate of uncertainty range (depending on uncertainty class 1 = 5%, 2 = 20% and 3 = 50%)[a]

Parameter	Uncertainty class	Value θ_l	Uncertainty range $\Delta\theta_l$	Unit
k_1	2	0.31	0.06	m^{-1}
k_2	2	0.026	0.005	gWM^{-1} m^2
$v_{O_2,atm}$	2	1.0	0.5	m d^{-1}
$v_{sed,\ ALG}$	3	0.20	0.10	m d^{-1}
$v_{sed,\ ORG}$	3	10.0	5.0	m d^{-1}
θ	1	0.95	0.05	
a_N	1	0.063	0.003	
$a_{P,max}$	2	0.009	0.002	
$a_{P,red}$	1	0.0087	0.0004	
$b_{P,max}$	2	0.0087	0.0017	
$b_{P,min}$	2	0.0014	0.0003	
c_e	2	0.70	0.14	
f_p	3	0.10	0.05	
w_{ALG}	2	5.0	1.0	
w_{ORG}	2	5.0	1.0	
w_{ZOO}	2	5.0	1.0	
$y_{ZOO,max}$	2	0.50	0.10	
k_{death,PLR,T_0}	3	0.030	0.015	d^{-1}
k_{gro,PLR,T_0}	2	0.13	0.03	d^{-1}
k_{resp,PLR,T_0}	3	0.010	0.005	d^{-1}
$K_{I,PLR}$	3	2.0	1.0	W m^{-2}
β_{ALG}	2	0.046	0.009	$°C^{-1}$
β_{BAC}	2	0.046	0.009	$°C^{-1}$
β_{PLR}	2	0.046	0.009	$°C^{-1}$
β_{ZOO}	2	0.080	0.016	$°C^{-1}$

[a] The horizontal lines separate physical, stoichiometric and kinetic parameters.

Table 4
Input parameters with uncertainty class, parameter estimate and prior estimate of uncertainty range (depending on uncertainty class 1 = 5%, 2 = 20% and 3 = 50%)

Parameter	Uncertainty class	Value θ_l	Uncertainty range $\Delta\theta_l$	Unit
$a_{\mathrm{P,inflow}}$	2	0.0087	0.0017	
$f_{\mathrm{NH_4,ara}}$	3	0.50	0.25	
$f_{\mathrm{NH_4,rain}}$	3	0.50	0.25	
$f_{\mathrm{X_I,damm}}$	3	0.20	0.10	
$f_{\mathrm{X_I,rivers}}$	3	0.20	0.10	
$F_{\mathrm{HPO_4}}$	2	1.0	0.2	
$F_{\mathrm{X_I}}$	2	1.0	0.2	
F_{N}	2	1.0	0.2	
$F_{\mathrm{K_z}}$	2	1.0	0.3	
$F_{\mathrm{O_2}}$	2	1.0	0.2	
F_{XPORG}	2	1.0	0.2	
F_{Q}	2	1.0	0.2	
$F_{\mathrm{X_S}}$	2	1.0	0.2	

inflows to be assumed to be ammonia (the rest is nitrate) and fractions of organic particles in the inflow assumed to be inert (the rest is degradable). The F factors are artificial model parameters. They are multiplied with the input loadings of the substances indicated in their index and are used to account for the consequences of uncertainty in input loadings.

In linear approximation to the model equations, the contribution of the uncertainty $\Delta\theta_l$ of the parameter θ_l to the uncertainty of the result of the model variable y_k calculated at time t at position z is given as

$$\Delta y_{k,\theta_l}(z,t,\boldsymbol{\theta}) = \Delta\theta_l \frac{\partial y_k}{\partial \theta_l}(z,t). \tag{2}$$

This uncertainty measure has the same dimension as the model variable y_k. It can be made non-dimensional by division with the scale, sc_{y_k}, introduced in Section 2 (for numerical values see Table 1). This makes it possible to create a measure of global sensitivity of model results to a parameter by averaging the squares of the nondimensional error contributions for all state variables, all sampling locations and all points in time and taking the square root:

$$\delta^{\mathrm{msqr}}_{\theta_l}(\boldsymbol{\theta}) = \sqrt{\frac{1}{n_{\mathrm{tot}}} \sum_{k=1}^{n_y} \sum_{j=1}^{n_{t,\mathrm{sens}}} \sum_{i=1}^{n_{y_k,z}} \left(\frac{\Delta\theta_l}{sc_{y_k}} \cdot \frac{\partial y_k}{\partial \theta_l}(z_{y_k,i}, t_j, \boldsymbol{\theta}) \right)^2}. \tag{3}$$

Note that through the selection of model variables, sampling positions and sampling times, this measure reflects the sensitivity of model results conditional on the selected measurement layout. Due to the use of the same factors, sc_{y_k}, for making the dimensional contributions non-dimensional as are used for making the residuals in Eq. (1) non-dimensional, this sensitivity measure is compatible with the parameter estimation procedure. With the notation

$$s_{\{k,i,j\},l}(\theta) = \frac{\Delta\theta_l}{\Delta y_k} \frac{\partial y_k}{\partial \theta_l}(z_{y_k,i}, t_j, \theta) \tag{4}$$

and

$$\boldsymbol{s}_l(\boldsymbol{\theta}) = (s_{\{1,1,1\},l}(\boldsymbol{\theta}), \ldots, s_{\{n_y, n_{y_{n_y}z}, n_{t,\mathrm{sens}}\},l}(\boldsymbol{\theta}))^T \tag{5}$$

(all combinations of the three indices $\{k,i,j\}$ are converted to a single index for the components of the vector s_l), it becomes evident that

$$\delta^{\mathrm{msqr}}_{\theta_l}(\boldsymbol{\theta}) = \frac{1}{\sqrt{n}} \left\| \boldsymbol{s}_l(\boldsymbol{\theta}) \right\| \tag{6}$$

as introduced in Brun et al. (2001).

Table 5 shows a ranking of the sensitivities of the model results calculated according to Eq. (3) for all parameters. It becomes evident that the parameters of growth, respiration and death of algae and zooplankton have the strongest overall contributions to model prediction uncertainty. Mineralization and nitrification rate coefficients, coefficients

of temperature dependence of biological processes, and stoichiometry-related parameters contribute significantly less to the uncertainty of predictions. The most influential external parameters are nitrogen loadings, phosphorus loadings, discharge and mixing.

The values of the sensitivities given in Table 5 decrease quasi-continuously without especially large jumps. This makes it difficult to select the size of the set of most influential parameters to be fitted (within the first 10 potential fit parameters, the sensitivity decreases by roughly a factor of 10). In addition, the results shown in this table do not reflect the possibilities of the approximate compensation of a change in the results induced by a change in one parameter by appropriate changes in other parameters. Such possibilities, which would limit the identifiability of a set of parameters, also in the case that the model results are sensitive to each of the parameters, are addressed in the following section.

4. Dependence analysis

In this section, the information on the sensitivity of model results on parameters shown in Table 5 is combined with information on the degree of linear dependence of sensitivity functions in order to find parameter sets to be estimated from lake data and to assess the influence of the values of fixed parameters on the results for estimated parameters. The procedure follows the technique outlined in Brun et al. (2001), which combines a sensitivity analysis with an analysis of measures proposed for the degree of linear dependence of sensitivity functions. The technique is systematically applied to all possible combinations of sub-

Table 5
Average error contributions of the parameters with respect to the measured values, which are the concentrations of algae, zooplankton, phosphate, nitrate and oxygen[a]

Potential fit parameter	$\delta_{\theta_l}^{\mathrm{msqr}}$	Fixed parameter	$\delta_{\theta_l}^{\mathrm{msqr}}$	Input parameter	$\delta_{\theta_l}^{\mathrm{msqr}}$
$K_{\mathrm{I,ALG}}$	1.207	k_1	0.559	F_{N}	0.428
$k_{\mathrm{death,ZOO},T_0}$	0.843	$v_{\mathrm{sed,ALG}}$	0.351	F_{Q}	0.304
$k_{\mathrm{gro,ALG},T_0}$	0.828	$k_{\mathrm{death,PLR},T_0}$	0.348	$F_{\mathrm{HPO_4}}$	0.222
$k_{\mathrm{gro,ZOO},T_0}$	0.763	β_{ZOO}	0.312	$F_{\mathrm{K_z}}$	0.188
$k_{\mathrm{resp,ALG},T_0}$	0.711	θ	0.304	$V_{\mathrm{O2,atm}}$	0.181
$k_{\mathrm{death,ALG},T_0}$	0.596	β_{ALG}	0.304	F_{XPORG}	0.091
$K_{\mathrm{HPO_4,ALG}}$	0.502	ω_{ALG}	0.302	$F_{\mathrm{O_2}}$	0.077
$S_{\mathrm{HPO_4,crit}}$	0.429	$K_{\mathrm{gro,PLR},T_0}$	0.268	$F_{\mathrm{X_s}}$	0.064
h_{sed}	0.300	$b_{\mathrm{P,min}}$	0.236	$A_{\mathrm{p,inflow}}$	0.051
$\Delta S_{\mathrm{HPO_4}}$	0.169	$a_{\mathrm{P,red}}$	0.176	$f_{\mathrm{x_I,dam}}$	0.030
k_{upt}	0.109	$a_{\mathrm{P,max}}$	0.147	$f_{\mathrm{x_I,rivers}}$	0.024
$k_{\mathrm{miner,aero,,sed},T_0}$	0.077	c_{e}	0.142	$f_{\mathrm{NH_4,stp}}$	0.024
$k_{\mathrm{resp,ZOO},T_0}$	0.076	k_2	0.138	$f_{\mathrm{NH_4,rain}}$	0.020
$k_{\mathrm{miner,anox,sed},T_0}$	0.076	$v_{\mathrm{sed,ORG}}$	0.134	$F_{\mathrm{X_I}}$	0.018
$K_{\mathrm{O_2,miner}}$	0.053	$Y_{\mathrm{ZOO,max}}$	0.133		
$k_{\mathrm{nitri,wat},T_0}$	0.043	β_{PLR}	0.125		
$k_{\mathrm{miner,aero,wat},T_0}$	0.040	$k_{\mathrm{resp,PLR},T_0}$	0.123		
$K_{\mathrm{NH_4,nitri}}$	0.040	$K_{\mathrm{I,PLR}}$	0.115		
$K_{\mathrm{NO_3,miner}}$	0.035	ω_{ZOO}	0.115		
$K_{\mathrm{O_2,resp}}$	0.028	$b_{\mathrm{P,max}}$	0.084		
$K_{\mathrm{NO_3,ALG}}$	0.023	ω_{ORG}	0.064		
$K_{\mathrm{O_2,nitri}}$	0.021	a_{N}	0.050		
$k_{\mathrm{nitri,sed},T_0}$	0.021	β_{BAC}	0.043		
$k_{\mathrm{miner,anox,wat},T_0}$	0.018	f_{p}	0.024		
$K_{\mathrm{O_2,upt}}$	0.017				

[a] The parameters are grouped according to the groups made by Tables 2–4.

sets of the parameters of a given size. The analysis of the degree of linear dependence of sensitivity functions is similar to the technique recommended for collinearity analysis of influence factors in linear regression (Belsley et al., 1980; Belsley, 1991; Draper and Smith, 1998).

If the sensitivity functions $\boldsymbol{s}_l$ (5) for a given parameter set are approximately linearly dependent, changes in model results induced by a small change in one parameter can be approximatively compensated by appropriate changes in the other parameters of the parameter set (in linear approximation to the model equations). This means that a parameter set can be poorly identifiable, although the results are sensitive to all of the parameters individually. To gain a measure for this approximate linear dependence, the sensitivity functions (5) are normed

$$\tilde{\boldsymbol{s}}_l = \frac{\boldsymbol{s}_l}{\|\boldsymbol{s}_l\|},\ \tilde{\boldsymbol{S}} = (\tilde{\boldsymbol{s}}_1,\ldots,\tilde{\boldsymbol{s}}_m) \tag{7}$$

and their linear combination with minimum norm is calculated under the constraint that the squares of the coefficients sum up to unity. The inverse of the norm of this linear combination, the collinearity index γ, is used as the measure for approximate linear dependence:

$$\gamma(\boldsymbol{\theta}) = \frac{1}{\min\limits_{\|\beta\|=1} \|\tilde{\boldsymbol{s}}_1\beta_1 + \ldots + \tilde{\boldsymbol{s}}_m\beta_m\|} = \frac{1}{\min\limits_{\|\beta\|=1} \|\tilde{\boldsymbol{S}}\boldsymbol{\beta}\|} = \frac{1}{\sqrt{\min(\mathrm{EV}[\tilde{\boldsymbol{S}}^T\tilde{\boldsymbol{S}}])}} \tag{8}$$

(EV[.] means the operator calculating the set of eigenvalues of the argument; see Brun et al., 2001 for more details). Note that this measure is equal to unity if the sensitivity functions are orthogonal, and it tends to infinity in the limit of exact linear dependence of the sensitivity functions. This is also true for the condition number preferred by Belsley (1991) as a measure for approximate linear dependence. As outlined in Brun et al. (2001), we prefer the collinearity index γ because of its more intuitive interpretation given by the first expressions in Eq. (8): a change in the results caused by a change of one parameter can be compensated by the fraction $1 - 1/\gamma$ by an appropriate change of the other parameters. This is true in linear approximation and for the mean squared deviations of all model results. A collinearity index of 20 therefore means that a change in the results caused by a change of one parameter can be compensated by 95% by an appropriate change of the other model parameters (in the interpretation given more precisely above). According to our experience with the present application, we expect serious identifiability problems to start for a collinearity index, γ, between 10 and 15.

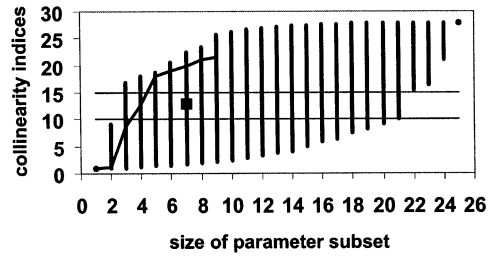

Fig. 1. Overview of the ranges of collinearity indices, γ, for all subsets of the potential fit parameters listed in Table 2. The ranges are plotted separately for all classes of subsets of given size (number of parameters). The sloping line connects the collinearity indices listed in the third column of Table 7, the marker shows the collinearity index of the parameter set selected for fit, and the horizontal lines bound the range of collinearity indices within which the transition from good to poor identifiability can be expected (for parameters to which the results are sensitive).

Note that both analyses, the sensitivity analysis done in Section 3 and the analysis of approximate linear dependence of sensitivity functions done in this section, are local analyses, the results of which depend on the values chosen for the model parameters. Because these analyses lead to the selection of parameters to be estimated, and therefore to new parameter estimates, the analyses have to be redone until convergence of the whole procedure is achieved. The results presented in the preceding section and in this section represent the results of the final step of this iterative procedure.

4.1. Identifiability analysis of model parameters

Fig. 1 gives an overview of possible values of the collinearity indices, γ, for all subsets of potential fit parameters listed in Table 2. Because of the extremely large number of parameter combinations evaluated for producing such a plot, typical

applications of the technique will limit the set sizes for the calculation of collinearity indices more strongly (this is possible without losing very important information). The plot shows that the maximum collinearity index increases very fast with increasing size of the parameter set. The largest collinearity index increases to 9.2 for parameter subsets of size 2, to 17.0 for parameter subsets of size 3, and to 18.1 for parameter subsets of size 4. This means that despite the large amount of data for five different model variables (see Table 1), there are already sets of three parameters that lead to serious identifiability problems caused by the approximate linear dependence of sensitivity functions. However, there are large subsets of parameters with collinearity indices below 5, where identifiability problems due to approximate linear dependence would not be expected.

As discussed above, the overview of collinearity indices shown in Fig. 1 is of limited value because it leads to the result that there may exist identifiable and non-identifiable parameters sets of almost any size (this is not a special feature of the present investigation but can be typically expected). The key question is according to the collinearity indices of subsets of parameters identified in the top rows of Table 5 to which the model results are very sensitive. Before proceeding with this question, Table 6 shows the parameter combinations leading to the highest values of the collinearity index for the subset sizes 2 and 3. These results show that for subset sizes of 2 and 3, there are only a few parameter combinations with collinearity indices close to the maximum. However, the composition of these parameter subsets is discouraging. The five combinations of two parameters and the top five combinations of three parameters listed in Table 6 all consist of subsets of the eight most influential parameters out of the group of potential fit parameters listed in the left column of Table 5.

Because this collinearity analysis leads to the identification of severe identifiability problems among the most influential model parameters, a systematic analysis of approximate linear dependence of the most influential parameters was performed. The results of this analysis are shown in Table 7. The first two columns of this table list the nine most influential parameters from the left column of Table 5 together with the sensitivity

Table 6
Identification of the parameter sets belonging to the largest collinearity indices for subset sizes 2 and 3[a]

Set size	Combinations	γ-range	Combinations with highest value of γ:			
			γ:	Parameter set		
2	300	1.0–9.2	9.2:	$K_{I,ALG}$	k_{resp,ALG,T_0}	
			6.9:	$K_{I,ALG}$	k_{gro,ALG,T_0}	
			5.9:	k_{gro,ZOO,T_0}	k_{death,ZOO,T_0}	
			5.4:	k_{gro,ZOO,T_0}	$S_{HPO_4,crit}$	
			4.3:	k_{resp,ALG,T_0}	k_{gro,ALG,T_0}	
3	2300	1.0–17.0	17.0:	$K_{I,ALG}$	k_{gro,ALG,T_0}	k_{resp,ALG,T_0}
			14.0:	$K_{I,ALG}$	k_{gro,ALG,T_0}	k_{death,ALG,T_0}
			11.8:	$K^{I,ALG}$	k_{resp,ALG,T_0}	$K_{HPO_4,ALG}$
			11.6:	$K_{I,ALG}$	k_{resp,ALG,T_0}	$S_{HPO4,crit}$
			11.5:	$K_{I,ALG}$	k_{resp,ALG,T_0}	K_{death,ALG,T_0}
			11.5:	$K_{I,ALG}$	k_{resp,ALG,T_0}	ΔS_{HPO_4}

[a] The first three columns show the size of the parameter subset (number of parameters), the number of possible combinations for such subsets out of the 25 potential fit parameters and the range of collinearity indices for these subsets. The other columns show the parameter sets leading to the largest value of the collinearity index.

Table 7
Collinearity indices of parameter subsets of the most influential potential fit parameters according to Table 5[a]

	$\delta_{\theta_l}^{msqr}$	Analysis 1 γ Set to par	Analysis 2 γ x+par	Analysis 3 γ x+par
Offset		1.0	12.6	12.8
$K_{I,ALG}$	1.21	1.0	x	x
k_{death,ZOO,T_0}	0.84	1.1	x	x
k_{gro,ALG,T_0}	0.83	8.7	x	x
k_{gro,ZOO,T_0}	0.76	12.6	x	x
k_{resp,ALG,T_0}	0.71	18.0	18.0	19.1
k_{death,ALG,T_0}	0.60	19.0	16.5	18.2
$K_{HPO_4,ALG}$	0.50	19.9	12.6	x
$S_{HPO_4,crit}$	0.43	21.2	12.6	x
h_{sed}	0.30	21.6	12.7	x

[a] The first two columns repeat the information given in Table 5 (names of the parameters and values of the non-dimensional average sensitivity, $\delta_{\theta_l}^{msqr}$). The column labelled 'Analysis 1' contains collinearity indices, γ, of the parameter sets consisting of the current parameter together with all parameters above. The columns labelled 'Analysis 2' and 'Analysis 3' contain collinearity indices, γ, of the parameter sets consisting of all parameters marked with an 'x' together with the current parameter. The row 'offset' contains the collinearity indices of the set of parameters marked with an 'x'.

measure $\delta_{\theta_l}^{msqr}$ according to Eq. (3) or Eq. (6) (the list stops there because there is a significant decrease by nearly a factor of 2 in $\delta_{\theta_l}^{msqr}$ between rank 9 and rank 10).

A first analysis of possible approximate linear dependence problems involved calculating the collinearity indices and condition numbers for the sets of most influential parameters of various size. The results of this analysis are listed in the columns labelled 'Analysis 1' in Table 7 and are shown by a sloping line in Fig. 1. These results show that already the four most influential parameters may be difficult to identify ($\gamma > 10$) and that serious identifiability problems can be expected for the five most influential parameters ($\gamma > 15$). In order to check if only the fifth parameter causes the problem also for the larger parameter sets, a second analysis was performed of sets of five parameters containing the top four and adding only one parameter from the ranks 5 to 9 at the same time. The results of this analysis are presented in the columns labelled 'Analysis 2' in Table 7. These results show that the parameters k_{resp,ALG,T_0} and k_{death,ALG,T_0} increase the collinearity index significantly, whereas the addition of one of the parameters $k_{HPO_4,ALG,SHPO_4,crit}$ or h_{sed} did not significantly increase the value of γ (compare with the offset values in the top row). This result motivated the third analysis, the results of which are shown in the column labelled 'analysis 3' of Table 7. Here, all three parameters, $k_{HPO_4,ALG,SHPO_4,crit}$ and h_{sed} were added to the four parameters already included in all parameter sets of analysis 2. This led to a slight increase in the collinearity index from 12.6 to 12.8 (see row 'offset'). The addition of one of the parameters k_{resp,ALG,T_0} or k_{death,ALG,T_0} increased this value to 19.1 or 18.2, respectively. To avoid the consequences of the identifiability problems caused by these two parameters, the parameter set consisting of the seven parameters, $K_{I,ALG}$, k_{gro,ALG,T_0}, k_{death,ZOO,T_0}, k_{gro,ZOO,T_0}, $K_{HPO_4,ALG}$, $S_{HPO_4,crit}$ and h_{sed} was selected for the formal parameter estimation procedure. The position of the collinearity index of this parameter set in the overview diagram presented in Fig. 1 is shown by a solid square.

Estimation of the seven parameters $K_{I,ALG}$, k_{gro,ALG,T_0}, k_{death,ZOO,T_0}, k_{gro,ZOO,T_0}, $K_{HPO_4,ALG}$, $S_{HPO_4,crit}$ and h_{sed} by minimizing the sum given by Eq. (1) leads to the estimates already given in Table 2. If we assume normally distributed, independent measurement errors with standard deviations proportional to the scales given in Table 1 and assume the other model parameters to be given, approximate standard errors and correlation coefficients of the estimates of the seven fitted parameters can be calculated (note that the above-mentioned restricting assumptions were not necessary for sensitivity and collinearity analysis). These results are given in Table 8. The results demonstrate that this parameter subset is already at the limit of identifiability (relative standard errors up to 25% and correlation coefficients up to 0.96) assuming the values of the other parameters being given. The results again show that the most severe identifiability problem within this set results from a correlation between the maximum specific growth rate of algae, k_{gro,ALG,T_0}, and the

half-saturation light intensity for algal growth, $K_{I,ALG}$. This problem was already observed in Tables 6 and 7. The large positive correlation coefficient of these two parameters is an indication of the absence of a significant light saturation effect. This problem could be partially overcome with a reparameterization of this process replacing the half-saturation light intensity, $K_{I,ALG}$, by a parameter equal to $k_{gro,ALG,T_0}/K_{I,ALG}$. Instead of two poorly identifiable parameters, this would lead to one accurately identifiable parameter (the new parameter $k_{gro,ALG,T_0}/K_{I,ALG}$, which measures the initial slope of the growth rate as a function of light intensity) and one poorly identifiable parameter (the maximum specific growth rate) (Ratkowsky, 1986). Because the identifiability problems are not too severe, and the parameterization used by Omlin et al. (2001) is much more commonly used, we did not change the parameterization.

What is the significance of the identifiability problems caused by the parameters k_{resp,ALG,T_0} and k_{death,ALG,T_0} if added to the set of selected model parameters? By omitting these parameters from the fit, only the identifiability problem for the fit algorithm is solved, and not the identifiability problem for the model as a whole. This latter problem can only be solved either by simplifying the model or by collecting additional data. The consequence of this identifiability problem is that for different values for these two parameters, a similarly good fit to the data could be achieved with adapted values of the fitted parameters. This result is not very astonishing, because respiration and death have similar consequences in reducing the effective growth rate of the algal population. As already mentioned in Section 1, the introduction of both processes respiration and death makes a simulation without explicitly depth-dependent parameters possible and allows to summarize all organic particles in the depth of the lake in the two classes of degradable and inert dead organic particles (into which algae and zooplankton are converted by the death process). For this reason, we did not want to further simplify the model. However, it is important to estimate the effect of changes of these parameters on the estimates of the fitted parameters as well as to check other parameters that were not fitted on the possibility of having a significant influence on the estimates of the fitted parameters. This is done in the next subsection.

4.2. Influence of fixed parameter values on parameter estimates

The model used in this paper contains more parameters than can be estimated from the available data. This raises the question of the dependence of the parameter estimates on parameters that were hold-fixed during the estimation process. If there is a high degree of approximate linear dependence of one of the fixed parameters with the set of fitted parameters, a significant influence of the value of the fixed parameter on the estimates for the fitted parameters can be expected. This is especially true for a parameter with a large error contribution to the simulation result.

Table 8
Relative standard errors (estimates of the parameter values are given in Table 2) and correlation coefficients for the seven parameters fitted to the data ignoring the uncertainty in the other parameters

Parameter	S.E.	$K_{I,ALG}$	k_{death,ZOO,T_0}	k_{gro,ALG,T_0}	k_{gro,ZOO,T_0}	$K_{HPO_4,ALG}$	$S_{HPO_4,crit}$	h_{sed}
$K_{I,ALG}$	25%	1						
k_{death,ZOO,T_0}	10%	0.50	1					
k_{gro,ALG,T_0}	14%	0.96	0.30	1				
$k_{gro,Z00,T_0}$	4%	0.48	0.45	0.37	1			
$K_{HPO_4,ALG}$	21%	−0.16	−0.63	0.07	−0.12	1		
$S_{HPO_4,crit}$	15%	−0.17	−0.62	−0.01	0.29	0.82	1	
h_{sed}	2%	−0.10	−0.18	−0.07	−0.30	0.14	−0.02	1

Table 9
Collinearity indices, γ, of all sets of parameters consisting of the seven parameters used in the fit (cf. Table 8) plus exactly one additional parameter[a]

Parameter	γ	$\delta_{\theta_l}^{\mathrm{msqr}}$	Parameter	γ	$\delta_{\theta_l}^{\mathrm{msqr}}$
$k_{\mathrm{resp,ALG},T_0}$	19.13	0.711	$f_{\mathrm{x_I,rivers}}$	13.21	0.024
$k_{\mathrm{death,ALG},T_0}$	18.16	0.596	$a_{\mathrm{P,red}}$	13.12	0.176
$k_{\mathrm{death,PLR},T_0}$	17.54	0.348	$K_{\mathrm{NH_4,nitri}}$	13.07	0.040
$k_{\mathrm{gro,PLR},T_0}$	17.31	0.268	β_{ZOO}	13.06	0.312
β_{PLR}	17.04	0.125	$K_{\mathrm{O_2,nitri}}$	13.02	0.021
$k_{\mathrm{resp,PLR},T_0}$	17.03	0.122	f_{p}	12.99	0.024
ω_{ALG}	16.46	0.300	$K_{\mathrm{O_2,upt}}$	12.99	0.017
$K_{\mathrm{I,PLR}}$	15.84	0.115	$k_{\mathrm{resp,ZOO},T_0}$	12.99	0.076
c_{e}	15.82	0.142	F_{Q}	12.95	0.304
F_{XPROG}	15.40	0.091	$K_{\mathrm{NO_3,ALG}}$	12.94	0.023
$Y_{\mathrm{ZOO,max}}$	15.03	0.133	$b_{\mathrm{p,min}}$	12.92	0.236
k_1	14.82	0.559	$F_{\mathrm{O_2}}$	12.89	0.077
$F_{\mathrm{HPO_4}}$	13.99	0.222	$k_{\mathrm{miner,aero,sed},T_0}$	12.88	0.077
β_{BAC}	13.86	0.043	F_{N}	12.88	0.428
$b_{\mathrm{P,max}}$	13.83	0.084	a_{N}	12.88	0.050
ω_{ZOO}	13.81	0.115	$k_{\mathrm{miner,anox,wat},T_0}$	12.86	0.018
$a_{\mathrm{P,inflow}}$	13.75	0.051	$K_{\mathrm{NO_3,miner}}$	12.86	0.035
$F_{\mathrm{x_I,damm}}$	13.67	0.030	$v_{\mathrm{O_2,atm}}$	12.86	0.181
$v_{\mathrm{sed,ORG}}$	13.65	0.134	$v_{\mathrm{sed,ALG}}$	12.85	0.351
$F_{\mathrm{K_z}}$	13.63	0.188	$k_{\mathrm{miner,anox,sed},T_0}$	12.85	0.076
θ	13.60	0.304	$F_{\mathrm{X_I}}$	12.85	0.018
k_2	13.60	0.138	$K_{\mathrm{O_2,miner}}$	12.85	0.053
β_{ALG}	13.56	0.302	$K_{\mathrm{O_2,resp}}$	12.85	0.028
ω_{ORG}	13.55	0.064	$\Delta S_{\mathrm{HPO_4}}^{*}$	12.85	0.169
$k_{\mathrm{miner,aero,wat},T_0}$	13.49	0.040	$k_{\mathrm{nitri,wat},T_0}$	12.85	0.043
$F_{\mathrm{X_s}}$	13.48	0.064	$f_{\mathrm{NH_4,rain}}$	12.85	0.020
$a_{\mathrm{P,max}}$	13.47	0.147	$f_{\mathrm{NH_4,stp}}$	12.85	0.024
$k_{\mathrm{nitri,sed},T_0}$	13.34	0.021	k_{upt}	12.85	0.109

[a] The sets are identified by the name of the additional parameter. To simplify the discussion, the average sensitivity, $\delta_{\theta_l}^{\mathrm{msqr}}$, of the results with respect to the additional parameter is also given (data from Table 5).

Table 9 shows the collinearity indices of all parameters sets (of size 8) consisting of the seven selected parameters plus one of the parameters that were held fixed during parameter estimation. The parameter sets are identified by the additional parameter and ordered according to a decreasing collinearity index. The results show that the specific respiration rate of the algae, $k_{\mathrm{resp,ALG},T_0}$, already identified in the previous subsection to cause identifiability problems, has the largest value of the collinearity index as well as of the sensitivity measure. For this reason, changes in the value of $k_{\mathrm{resp,ALG},T_0}$ can be expected to have the largest influence on the estimates of the seven fitted parameters. Table 10 shows the change in the fit results caused by an increase of the value of the parameter by one uncertainty unit, $\Delta k_{\mathrm{resp,ALG},T_0}$. The resulting change in the parameters $K_{\mathrm{I,ALG}}$ and $k_{\mathrm{gro,ALG},T_0}$ is largest. These are the most strongly correlated parameters (see Table 9) and the parameters for which the collinearity indices with $k_{\mathrm{resp,ALG},T_0}$ are largest (9.2 and 4.3, whereas it is in the range between 1.0 and 2.9 for $k_{\mathrm{resp,ALG},T_0}$ together with any of the other parameters listed in Table 10). Note that the decrease in the value of $k_{\mathrm{gro,ALG},T_0}$, together with the even larger decrease in the value of $K_{\mathrm{I,ALG}}$ leads to an increase in the production rate for small light intensities. This compensates for the increase in respiration.

5. Model validation

The word 'validation' can be misinterpreted by assuming that the model is checked for its truth (to be 'valid'). This is not possible. A model of an ecological system is always a rough simplification of reality that can only be 'invalidated' or 'falsified' for given environmental conditions by a comparison with data. In order to stress this point, the validation process is sometimes called a model confirmation or corroboration (Reckhow and Chapra, 1983). An increasing number of successful validation tests increase the degree of belief in the usefulness of a model for the types of model application covered by the tests. This makes the property of being validated not an objective property of a model but a property that depends on the purpose of modelling and on the power of tests that is judged to be adequate (Rykiel, 1996).

A useful technique of testing (validating) a model is using only a subset of all available data for calibration and testing model performance with the rest of the data. This procedure is called cross-validation (Snee, 1977). If data are available at different points in time and predicting future states of a system is a goal of modelling, then splitting the data at a given point in time and using data before this point for calibration and data after this point for validation is an especially powerful variant of cross-validation. This technique is called predictive validation (Power, 1993).

The purposes of modeling Lake Zürich were to gain a scientific understanding and to predict future states of the lake (Omlin et al., 2001). For this reason, a predictive validation was applied to test the model. We selected the Janus coefficient (Power, 1993), which is defined as the ratio of the mean variance of the deviation of model results from data in the prediction period to the mean variance of the same deviation in the fit period

$$J^2 = \frac{\frac{1}{n_{\mathrm{tot,pred}}} \sum_{k=1}^{n_y} \sum_{j=n_{\mathrm{t,fit}}+1}^{n_{\mathrm{t,fit}}+n_{\mathrm{t,pred}}} \sum_{i=1}^{n_{y_k,z}} \left(\frac{y_{\mathrm{meas},k,j,i} - y_k(z,i,t_j,\boldsymbol{\theta})}{sc_{y_k}} \right)^2}{\frac{1}{n_{\mathrm{tot,fit}}} \sum_{k=1}^{n_y} \sum_{j=1}^{n_{\mathrm{t,fit}}} \sum_{i=1}^{n_{y_k,z}} \left(\frac{y_{\mathrm{meas},k,j,i} - y_k(z,i,t_j,\boldsymbol{\theta})}{sc_{y_k}} \right)^2} \tag{9}$$

as an indicator for model validity. In this equation, $n_{\mathrm{tot,pred}}$ is the total number of data points in the prediction period, and $n_{\mathrm{tot,fit}}$ is the total number of data points in the fit period. All other variables have the same meaning as in Eq. (1). The Janus coefficient given by Eq. (9) can be expected to be significantly larger than unity if the model structure is too complicated and fits random fluctuations (overfitting).

In the present application, the model was fitted to the data from the years 1988 and 1989 and was validated using the data of the year 1990 and 1991. Using monthly spatial profiles of oxygen, nitrate, phosphate and algae and monthly averaged epilimnion and hypolimnion concentrations of zooplankton, as described in Omlin et al. (2001), the Janus coefficient was calculated based on 1713 fit data points and 1746 prediction data points. Its value of 0.8 gives us a high degree of confidence in the credibility of the model for this type of extrapolation. Note, however, that the following limitations of this validation procedure should not be forgotten:

Table 10
Change in fit results caused by an increase of the value of $k_{\mathrm{resp,\ ALG\ T}o}$ by one uncertainty unit, $\Delta k_{\mathrm{resp,\ ALG,\ T}o}$, as listed in Table 2 (from 0.050 to 0.075 d^{-1}).

Parameter	Original value	New value	Unit	Change (%)	Change/$\Delta\theta_l$ (%)
$K_{\mathrm{I,ALG}}$	34	20	Wm^{-2}	−43	−85
$k_{\mathrm{gro,ALG},T_0}$	1.1	0.94	d^{-1}	−16	−82
$k_{\mathrm{death,ZOO},T_0}$	0.029	0.028	d^{-1}	−2.5	−4.9
$k_{\mathrm{gro,ZOO},T_0}$	0.30	0.31	$\mathrm{gDM}^{-1}\ \mathrm{m}^3\ \mathrm{d}^{-1}$	1.6	8.2
$K_{\mathrm{HPO_4,ALG}}$	0.0019	0.0015	$\mathrm{gP\ m}^{-3}$	−18	−37
$S_{\mathrm{HPO_4,crit}}$	0.0042	0.0042	$\mathrm{gP\ m}^{-3}$	−0.3	−1.4
h_{sed}	0.0036	0.0036	m	0.1	0.2

- The external driving forces on the lake did not change significantly from the fit period to the prediction period. This means that this test mainly confirmed that the model runs in a stable mode and does not lead to significant trends that are not present in the data. This does not give an indication that the model is also able to predict the reaction of the lake to significant changes in driving forces (cf. discussion of prediction uncertainty in the subsequent section).
- Validation was only performed for Lake Zürich. This does not lead to conclusions with respect of the transferability of the model to different lakes.

Both of these properties, predictive capabilities for changes in external driving forces and for different lakes would be highly desirable properties of a lake model.

It should also be noted, that the measure given by Eq. (9) is not sensitive to systematic deviations between model results and data. In the present case, the major systematic deviations present in both the fitting and the prediction period are caused by an inaccurate description of mixing processes in the hypolimnion (Omlin et al., 2001). Mixing processes were quantified before the calibration procedure of the biogeochemical model was carried out. An iterative procedure of calibration of mixing and biogeochemical processes could decrease such systematic deviations. This was not done in the present study because such a procedure could lead to 'good' results by a compensation of different wrongly calibrated processes. Another problem of calibration by least squares and of validation using the Janus coefficient is the problem that minor time shifts (e.g. of algal peaks) in a qualitatively correct solution lead to a very large value of χ^2. This can lead to a least-squares solution that is different from a 'best' solution obtained by a manual optimization procedure based on a graphical assessment of the solutions (Elliott et al., 2000). For this reason, automatic fitting of parameters should always be accompanied by a visual inspection of the results.

6. Prediction uncertainty

The computation time required for simulations with the lake model make it very expensive to do non-linear or regional uncertainty analyses with the aid of Monte Carlo simulations and/or Bayesian inference (Homberger and Spear, 1981; Hornberger and Spear, 1983; Beck, 1987; Gelman et al., 1995; Beck, 1999; Reckhow and Chapra, 1999). In order to obtain a rough estimate of prediction uncertainty without such a high computational cost, linearized error propagation is used instead. This method is available at negligible computational costs, because the sensitivity functions, $\partial y_k/\partial\theta_l$, required for calculating the estimate of the uncertainty range of the model predition for the variable y_k at location $z_{y_k,i}$ and at time t_j

$$\Delta y_k(z_{y_k,i},t_j) = \sqrt{\sum_{l_1,l_2=1}^{m} \mathrm{Corr}[\boldsymbol{\theta}]_{l_1,l_2}\left(\frac{\partial y_k(z_{y_k,i},t_j)}{\partial\theta_{l_1}}\Delta\theta_{l_1}\right)\left(\frac{\partial y_k(z_{y_k,i,t_j})}{\partial\theta_{l_2}}\Delta\theta_{l_2}\right)} \quad (10)$$

have already been calculated for the sensitivity and dependence analyses performed in Sections 3 and 4 (m in Eq. (10) is equal to the number of parameters, and $\mathrm{Corr}[\boldsymbol{\theta}]$ is the correlation matrix of the parameters). The meaning of Δy_k is the same as that for $\Delta\theta_l$; as for the uncertainty ranges, we use measures comparable with a standard deviation of the corresponding distribution. The most critical approximation made by Eq. (10) is that it is only valid as long as the non-linearities in the system equations are not significant within the uncertainty ranges of the parameters. This assumption is certainly not fulfilled in our model because of the large uncertainty ranges of the parameters and the non-linearity of the model equations. Therefore, the uncertainty ranges calculated in this section are only rough approximations to a more accurate calculation that would consider the non-linearity of the model equations. However, because the estimates of the parameter uncertainty ranges are also only very rough, the computational expense for such an analysis would probably not be justified for the minor increase in the accuracy of the calculated uncertainty estimates.

The crucial point in the application of the error propagation formula (Eq. (10)) is the assumption to be made for the uncertainty ranges $\Delta\theta_l$ and for the correlation matrix Corr[$\boldsymbol{\theta}$] of the model parameters. We perform the following three calculations. First, we calculate the uncertainty of the results assuming uncorrelated parameters with the prior uncertainties listed in Tables 2–4. Second, we replace the uncertainties and correlations for the fitted parameters by those estimated in the fit (see Table 8). Third, we set the uncertainties of all potential fit parameters listed in Table 2 to zero. These calculations allow us to make a comparison between the prior uncertainty estimates, a crude approximation to the posterior uncertainty estimates after the estimation of the parameters selected in Section 4.1, and the limiting case of uncertainty estimates for exactly known values of those parameters we would ideally like to estimate from the lake data. This last calculation enables us to see what would be the maximum possible reduction in prediction uncertainty by an estimation of all parameters that are meaningful to fit with the data used for this investigation (mainly monthly profiles of oxygen, phosphate, nitrate, algae and *P.* (*Oscillatoria*) *rubescens* and depth-integrated samples of zooplankton). The remaining uncertainty is caused by the parameters listed in Tables 3 and 4. These remaining sources of uncertainties are mainly limited knowledge of sedimentation velocities, stoichiometric coefficients, kinetic parameters of *P.* (*Oscillatoria*) *rubescens*, temperature-dependence coefficients and input uncertainties (including mixing). These uncertainties can only be reduced by more specific investigations. The estimation of the posterior uncertainty bounds by using the fit results from Table 8, which were gained ignoring the uncertainty in the other parameters, is conceptually not convincing. However, it has been shown for a simple example in Omlin and Reichert (1999) that a combination of such fit results with the prior uncertainties for the parameters that were not fitted leads to similar results as a conceptually more convincing Bayesian analysis. Because computational requirements prevent us from doing a Bayesian analysis, we apply the simplified technique. The error introduced by this technique is at least bounded by the other two estimates so that the main conclusions of the paper are justified also if the simplified technique results in inaccurate results.

In Fig. 2 model predictions, data and all uncertainty estimates described above are shown for oxygen, phosphate, nitrate and algae profiles in different seasons during the prediction period 1990. It is obvious that the consideration of the fit results reduces the uncertainty to some degree. However, as the calculation with zero uncertainty for the potential fit parameters shows (to which the calculation that considers the fit results is astonishingly close), the maximum possible reduction in the uncertainty of the predictions that can be achieved by fitting the model parameters to the monthly profiles is limited. The uncertainty due to light extinction coefficients, kinetic parameters of *P.* (*Oscillatoria*) *rubescens*, stoichiometric assumptions, sedimentation velocities, input, mixing, etc., that can only be reduced by more specific measurements, has a considerable contribution to total uncertainty, especially if the fit results are considered. The most important systematic deviations seem to be due to an inaccurate description of the mixing processes in the lake (e.g. reflected by the March profiles for oxygen). The overall inaccuracy of 20% attributed to the vertical mixing coefficient, K_z, may be temporarily exceeded.

Besides this relative importance of different causes of uncertainty the magnitude of the uncertainty is also interesting. If we consider the fact that our uncertainty estimates are estimated as measures in the order of one standard deviation of the distributions characterizing the knowledge on the parameters, under the assumption of normally distributed uncertainties, 95% confidence intervals would be about twice as wide as the uncertainty intervals plotted in Fig. 2. This shows that the uncertainty bands are very large, especially for algae. This especially large uncertainty for algae is not astonishing, because the algal population changes significantly from year to year, and the maxima do not always occur in the same month of the year. The result that realistic uncertainty estimates for nutrient models are very large is not a new result of the present study (Reckhow and Chapra, 1999). Facing this large

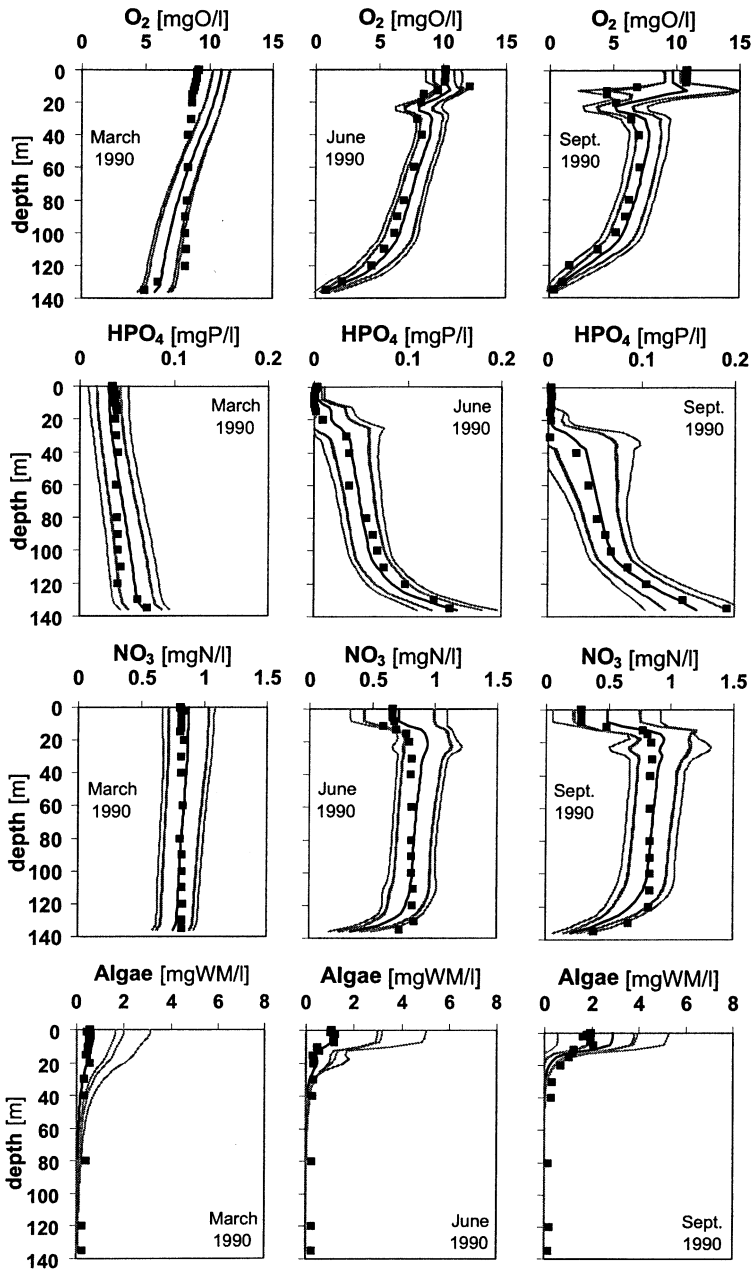

Fig. 2. Parameter error propagation on model predictions for oxygen, phosphate, nitrate and algae concentrations for 3 months in 1990. Markers represent measured data, black lines model predictions, and grey lines predictions $\pm\Delta y_k$ according to Eq. (10). Uncertainty bounds have been calculated for the prior uncertainties listed in Tables 2–4 (widest uncertainty ranges), for the uncertainties of the fitted parameters replaced by the information given in Table 8 (intermediate uncertainty ranges), and for all uncertainties of the potential fit parameters according to Table 2 set to zero (narrowest uncertainty ranges).

uncertainty, the good correspondence of the simulations with measured data is puzzling. On the one hand, this may be caused by a certain overestimation of the uncertainty due to conservative prior uncertainty estimates (however, the structural uncertainty of the model is neglected) and due to the omission of correlations among model parameters (Reckhow and Chapra, 1999). However, the predictions may be closer to the measurements than expected from their uncertainty because the external driving forces were very similar during the calibration and the extrapolation periods. This makes the validation exercise not to a test of all model mechanisms.

7. Summary and conclusions

A systematic investigation of the influence of the sensitivity of results to model parameters and of approximate linear dependence of sensitivity functions on model parameter identifiability, as proposed by Brun et al. (2001) for the biogeochemical lake model described by Omlin et al. (2001), led to the following results:

- Contributions to model prediction uncertainty of the model results (oxygen, nutrients and plankton concentrations in the water column of the lake) were largest for the half-saturation light intensity of algal growth and for parameters proportional to algae and zooplankton growth, respiration and death.
- Mineralization and nitrification rate coefficients, coefficients of temperature dependence of biological processes, and stoichiometry-related parameters contributed less to the uncertainty of model predictions.
- Among the external parameters nitrogen and phosphorus loadings, discharge and mixing contributed most to the prediction uncertainty.
- Due to the possibility of compensating effects, the parameters of algal and zooplankton growth, respiration and death could not all together be estimated from the available data (monthly profiles of nutrients, oxygen and plankton).
- With assumed values for most respiration and death parameters and a fit of growth parameters, a good correspondence of model results with data could be achieved. As compared to the literature, all model parameters seemed to be in a reasonable order of magnitude. For small changes in the values of respiration and death parameters, a similarly good fit could be achieved by adapting the growth parameters. This demonstrates again the non-identifiabilty of the parameters.

The last point shows that the non-identifiability may not have severe consequences as long as net growth rates are similar. The model cannot be easily reduced to net growth because respiration of sinking algae contributes to the metalimnic oxygen minimum, and death converts all living organic matter to dead degradable and inert organic matter building the major fraction of the lake sediment. The application of the model to longer time periods and to other lakes could lead to more information on parameters non-identifiable from the data set used for the present investigation.

A rough uncertainty analysis led to very large estimates of the uncertainty of model predictions. This result is in contrast to the very good agreement of model results with measurements also in the extrapolation domain of the model. There are probably three main reasons for this result: first, the prior uncertainty estimates were selected very conservatively, second, the omission of (unknown) correlations among the model parameters typically increases the calculated uncertainty, and third, the external conditions of the lake did not change significantly between the calibration and extrapolation periods. This leads to a continuation of the behaviour observed in the calibration period to the extrapolation period for most model variables. For changes in driving conditions of the lake (discharge, loadings, light, etc.), a larger deviation of predictions from observations can be expected. Although parameters that can be expected to be estimated from lake data lead to the largest single contributions to the estimated prediction uncertainty, the effect of all parameters that are not meaningful to be estimated from the data together was a significant contribution to the estimated prediction uncertainty. This limits the possibility of decreasing the predictive uncertainty

of the model without additional more specific measurements that lead to more accurate estimates of nitrogen and phosphorus loadings of the lake, of vertical mixing, of sedimentation velocities, of temperature dependence of the biogeochemical processes and of the phosphorus content of algae. It may not be possible to reduce some of these uncertainties without increasing the model complexity (e.g. the large uncertainty in the sedimentation velocity accounts for the existence of different particles that sediment at different speeds; this could be considered by introducing more classes of particles). This would increase the accuracy of subprocesses, but it is not clear whether it would also increase the accuracy of the predictions of the model as a whole (Beck, 1999; Reckhow and Chapra, 1999).

Acknowledgements

We thank Richard Forster, Sonja Gammeter and Ulrich Zimmerman from Wasserversorgung Zürich, Switzerland, for their collaboration in the development of the lake model described in more detail in Omlin et al. (2001). The local water supply authority (Wasserversorgung Zürich, WVZ), the local water protection authority (Amt für Gewässerschutz und Wasserbau, AGW), the Swiss Meteorological Institute (SMA) and the Swiss Federal Institute for Environmental Science and Technology (EAWAG) kindly provided data used for this study.

References

Beck, M.B., 1987. Water quality modeling: A review of the analysis of uncertainty. Water Resour. Res. 23 (8), 1393–1442.

Beck, M.B., 1999. Coping with ever larger problems, models, and data bases. Water Sci. Technol. 39 (4), 1–11.

Beisley, D.A., 1991. Conditioning Diagnostics — Collinearity and Weak Data in Regression. Wiley, New York.

Belsley, D.A., Kuh, E., Welsch, R.E., 1980. Regression Diagnostics — Identifying Influential Data and Sources of Collinearity. Wiley, New York.

Brun, R., Reichert, P., Künsch, H.-R., 2001. Practical identifiability analysis of large environmental simulation models. Water Resour. Res. 37 (4), 1015–1030.

Draper, N.R., Smith, H., 1998. Applied Regression Analysis, third ed. Wiley, New York.

Elliott, J.A., Reynolds, C.S., Irish, A.E., Tett, P., 2000. Modelling freshwater phytoplankton communities: an exercise in validation. Ecol. Modell. 128, 19–26.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 1995. Bayesian Data Analysis. Chapman & Hall, London.

Homberger, G.M., Spear, R.C., 1981. An approach to the preliminary analysis of environmental systems. J. Environ. Manage. 12, 7–18.

Hornberger, G.M., Spear, R.C., 1983. An approach to the analysis of behavior and sensitivity in environmental systems. In: Beck, M., van Straten, G. (Eds.), Uncertainty and Forecasting of Water Quality. Springer, Berlin, pp. 101–116.

Imboden, D.M., Gächter, R., 1978. A dynamic lake model for trophic state prediction. Ecol. Modell. 4, 77–98.

Karagounis, I., Trösch, J., Zamboni, F., 1993. A coupled physical–biochemical lake model for forecasting water quality. Aquat. Sci. 55 (2), 87–102.

Nordhaus, W.D., 1995. Managing the Global Commons — The Economics of Climate Change. MIT Press, Cambridge, MA.

Omlin, M., Reichert, P., 1999. A comparison of model extrapolation uncertainty estimation techniques. Ecol. Modell. 115 (1), 45–59.

Omlin, M., Reichert, P., Forster, R., 2001. Biogeochemical model of lake Zürich: Model equations and results. Ecol. Modell. 141, 77–103.

Power, M., 1993. The predictive validation of ecological and environmental models. Ecol. Modell. 68, 33–50.

Ratkowsky, D.A., 1986. A suitable parameterization of the Michaelis–Menten enzyme reaction. Biochem. J. 240, 357–360.

Reckhow, K.H., Chapra, S., 1999. Modeling excessive nutrient loading in the environment. Environ. Pollut. 100, 197–207.

Reckhow, K.H., Chapra, S.C., 1983. Confirmation of water quality models. Ecol. Modell. 20, 113–133.

Reichert, P., Omlin, M., 1997. On the usefulness of overparameterized ecological models. Ecol. Modell. 95, 289–299.

Rykiel, E.J., 1996. Testing ecological models: the meaning of validation. Ecol. Modell. 90, 229–244.

Snee, R.D., 1977. Validation of regression models: methods and examples. Technometrics 19 (4), 415–428.

Spriet, J.A., 1985. Structure characterization — an overview. In: Barker, H.A., Young, P.C. (Eds.), Identification and System Parameter Estimation 1985 — Proceedings of the seventh IFAC/IFORS Symposium. Pergamon Press, Oxford, pp. 749–756.

Biogeochemical model of Lake Zürich: Sensitivity, identifiability and uncertainty analysis

Omlin, Martin; Brun, Roland; Reichert, Peter

01 mingxi zhang Page 1
31/8/2023 3:13

02 mingxi zhang Page 1
31/8/2023 3:13

03 mingxi zhang Page 4
31/8/2023 5:12

04 mingxi zhang Page 4
31/8/2023 3:23

05 mingxi zhang Page 4
31/8/2023 3:54