Uncertainty Analysis of Model Predictions for Environmental Systems: Concepts and Application to Lake Modelling

A dissertation submitted to the

SWISS FEDERAL INSTITUTE OF TECHNOLOGY ZÜRICH

for the degree of

Doctor of Natural Sciences

presented by
Martin Omlin
Dipl. Natw. ETH
born December 29, 1966
citizen of Sachseln, Obwalden and Dietlikon, Zürich

accepted on the recommendation of
PD Dr. Peter Reichert, examiner
Prof. Dr. Dieter Imboden, co-examiner
Prof. Dr. Roland Scholz, co-examiner

Zürich, 2000
Acknowledgements

This thesis was made possible by financial support from the Swiss Federal Institute of Environmental Science and Technology (EAWAG).

During the last 4 years I enjoyed the support and encouragement of many people whom I would like to thank.

A special thank goes to my supervisor Peter Reichert who always had time for me inspite of his tight time table. I would also like to thank Dieter Imboden for his support and his help in the administrative tasks.

Many people have contributed to the shape and substance of this work, both indirectly through discussions and directly through comments on earlier versions of the manuscript. At the end of each chapter the people who contributed to its content are mentioned.

I would like to thank my parents, Theres and Hans Omlin, who always encouraged me in my academic education and gave me financial support during this time. A special thank also to ASE Synneve Johnsen for her encouragement and moral support.
Contents

I  Concepts

1  On the Usefulness of Overparameterized Ecological Models  25
   1.1  Introduction ................................. 25
   1.2  Classical Model Identification - Principles and Problems  26
   1.3  Bayesian Inference - Advantages and Problems .................. 29
   1.4  Recommended Identification and Forecasting Procedure ............. 34
   1.5  Conclusions .................................. 37

2  A Comparison of Techniques for the Estimation of Model Prediction Uncertainty  39
   2.1  Introduction .................................... 39
   2.2  Identification and Prediction Procedures .......................... 41
      2.2.1  Classical Frequentist Analysis ...................... 41
      2.2.2  Bayesian Analysis .............................. 43
   2.3  A Simple Example of Ecological Relevance ....................... 45
      2.3.1  Model Equations ............................. 45
      2.3.2  Data Sets and Prior Distributions .................... 45
      2.3.3  Simulation Methods and Numerical Procedures ............... 46
      2.3.4  Identification and Prediction with Complete Data Sets ......... 47
      2.3.5  Identification and Prediction with Incomplete Data Sets ...... 48
   2.4  Summary and Conclusions ................................ 54

II  Application

3  Biogeochemical Model of Lake Zürich: Model Equations and Results  59
   3.1  Introduction ..................................... 60
   3.2  Lake Zürich .................................... 60
      3.2.1  Lake Description ............................ 60
      3.2.2  Data: Sources and Description .................. 60
   3.3  Model Description ................................ 62
      3.3.1  Overview ................................... 62
      3.3.2  Mass Balance Equations ...................... 63
      3.3.3  Physical Processes ........................... 64
      3.3.4  Biogeochemical Processes ...................... 66
      3.3.5  Model Implementation ........................... 75
   3.4  Results and Discussion ................................ 75
3.4.1 Concentration Profiles .................................................. 77
3.4.2 Sedimentation Fluxes .................................................... 85
3.4.3 Oxygen, Nitrate and Phosphate Conversion Rates ............... 85
3.4.4 Variable Stoichiometry and Phosphate Uptake ................. 86
3.5 Conclusions .................................................................. 88
3.6 Appendix .................................................................... 89

4 Biogeochemical Model of Lake Zürich: Sensitivity, Identifiability and Uncertainty Analysis 91
4.1 Introduction .................................................................. 92
4.2 Parameter Estimation ..................................................... 94
4.3 Sensitivity Analysis ........................................................ 95
4.4 Dependence Analysis ...................................................... 100
  4.4.1 Identifiability Analysis of Model Parameters .................. 101
  4.4.2 Influence of Fixed Parameter Values on Parameter Estimates 105
4.5 Prediction Uncertainty .................................................... 107
4.6 Summary and Conclusions .............................................. 110
Summary

A model is always a partial description of reality, but can be useful for the representation of the aspects for which it was designed. It is useful because developing models as a part of the scientific knowledge discovery process can give increased understanding, and given reasonable results, the outcome can be predicted. The value of these predictions, however, are dependent on the knowledge of their accuracy. It is the subject of this work to evaluate the uncertainty of predictions.

Causes for prediction uncertainty are best analysed by reviewing the processes of model development. This is shown in the first part of the work and starts with a presentation of the classical model identification approach. Problems in the context of prediction uncertainty are emphasized. An alternative Bayesian procedure is presented, which in many aspects is better but also has some drawbacks. The result of chapter 1 is a general framework of the modelling process which includes model identification and model prediction. In chapter 2 this framework is applied to a didactical example of microbial growth kinetics. In the example an important difference between the Bayesian approach and the classical method is discussed: the possibility of the Bayesian approach to handle non-identifiable model parameters. In Part II of the thesis the focus is changed from concepts to applied ecological modelling. A model which describes biochemical processes in a lake is presented and applied in the context of the proposed framework of chapter 1. The model and its development is shown in chapter 3. In chapter 4 the identifiability of model parameters and prediction uncertainty are analyzed.

PART I: Concepts

The classical approach to system identification consists of the following steps. From a set of proposed models inspired mainly by prior knowledge of the system, the most adequate model is selected based on a criterion which favours good agreement with data (quality of fit) and penalizes complicated model structures (parsimony). The parameters of the selected model are estimated with maximum likelihood point estimation and their uncertainty is estimated locally using a linear approximation to the model equations.

The parsimony principle provides a reasonable criterion for system identification and model selection by excluding overparameterized models. From the system identification point of view during the measurement periods it does not make sense to use a model structure for which parameters cannot be estimated with the available data. However, if the model is intended to make predictions that require an extrapolation outside of its identification domain, poorly identifiable parameters, relevant to processes that may be important in the extrapolation domain but are of less significance in the identification domain, are sometimes encountered. If prior knowledge of such processes is available, they should be included in the analysis because discarding such processes can lead to a
significant underestimation of prediction uncertainty.

Another problem with the classical identification procedure with respect to prediction and prediction uncertainty is the use of a single uniquely identified model as the "true" model. Ecological systems are extremely complex and all models of such systems are very drastic simplifications of reality. For this reason there is no obvious justification for the selection of a single model as the "true" description of an ecological system. The practical advantage of a unique identification is that it is easier to interpret a single model structure. On the other hand, this interpretation may be misleading if other model structures give similarly good descriptions of the data. In such a situation, selection of a single model neglects the uncertainty in model structure and results in the underestimation of the overall prediction uncertainty.

The Bayesian approach to identification and prediction can be used to eliminate the problems mentioned above. It's approach differs fundamentally from the classical procedure. In classical statistics the unknown parameters of a deterministic model are fixed values and only the estimators of their values that depend on the measurement are random variables. These estimators are generally used to calculate a point estimate of the true but unknown model parameter. In the Bayesian concept on the other hand, the model parameter itself is described with a probability distribution. These differences reflect a different meaning of probability in the two frameworks. Probability theory in classical statistics is only applied to random events that, at least in principle, can be sampled to determine empirical frequency distributions which approximate the true distribution asymptotically. In Bayesian statistics, the field of application of probability theory is extended. Here the notion of probability is also used to represent uncertainty or partial knowledge of a system that may be deterministic.

The main idea of the Bayesian approach is very simple: Bayes' theorem of conditional probability is used to update a prior probability distribution of model parameters to a new posterior distribution that includes the additional information from new measurements. These parameter distributions can then be used directly to calculate the distributions of model predictions. However, this step must be implemented with Monte Carlo simulations and is therefore computationally very demanding. Due to the calculation of probability distributions of parameters instead of point estimates, identifiability is not a necessary condition for the use of Bayesian techniques. Therefore, also overparameterized models can be used when there is strong evidence that these can give more reliable prediction uncertainty estimates. The Bayes model averaging technique expands this idea to the level of different model structures so that for each model a probability is calculated. A model prediction is not based on the selection of a unique model but consists of using a weighted average of the predictions of different models. The Bayesian approach solves some problems of the classical identification technique and therefore from a methodical point of view, it is recommended to use in situations where the inclusion of prior knowledge seems to be necessary and leads to overparameterized models. However, in consideration of the required computational effort this recommendation is too simple. For complex models with a lot of parameters it may be practically impossible to perform enough Monte Carlo simulations to obtain a reasonably accurate uncertainty estimate. To account for both the computational efficiency of the classical procedure, and superiority of the methodology of the Bayesian approach (for the range of applications mentioned above), the following framework for environmental system identification and forecasting is recommended:
A) The model parameters of all model structures are accurately identifiable.

→ Perform a classical parameter estimation for all the discretely different model structures.

According to the "selectivity" of the models, two cases can be distinguished:

a) One model structure has a much better value of the classical selection criterion.

→ Select this model and perform model prediction in the classical way (maximum likelihood point estimation of the parameters and linearized uncertainty estimation).

b) There is a subset of model structures with similarly good values of the selection criterion.

→ Do model prediction by Bayesian model averaging among this subset of models

B) There exist poorly identifiable model parameters.

→ Bayesian inference and model averaging should be applied.

According to the computational requirements, two cases can be distinguished:

a) There is enough computing power available to approximately calculate probability distributions by Monte Carlo simulation.

→ Do combined Bayesian parameter estimation and model averaging

b) There is insufficient computing power available to perform Monte Carlo simulations for the considered models.

→ Try to determine and isolate (by reparameterization) poorly identifiable model parameters. If this is possible, select different values of these parameters according to equal differences in the integral prior distribution function and treat models deviating in one of these poorly identifiable parameters as discretely different model structures in the other parameters. Then proceed according to A). If it is not possible to isolate non-identifiable parameters, apply the technique described above to subsets of the parameters in a cyclical way. The subsets of parameters with given values must be selected in such a way that the other parameters can be identified.

In chapter 2, an example is used as a didactical tool for illustrating the classical and the Bayesian techniques under different situations of identifiability. A simple model of microbial substrate conversion under different assumptions on prior knowledge and data availability is used. In order to focus on the issue of identifiability the model structure from which synthetic data is generated is assumed to be known. Three situations of data availability and prior knowledge are discussed. In the first situation a complete data set leads to identifiability of the model parameters. In cases where only a partial data set is available one model parameter remains non-identifiable. This results in two other situations where good or poor prior knowledge about the non-identifiable parameter is available.

Similar results were achieved regardless of the technique employed in the first situation where the data allowed precise identification of the model parameters. Considering
the much higher computational expense involved in employing the Bayesian approach in
such situations, the classical procedure is preferable from a practical point of view. In the
second situation, application of the classical technique to the identifiable parameter only,
while assuming fixed values for the non-identifiable parameter, yields reasonable results if
the value of the non-identifiable parameter is accurately known. However, the uncertainty
of the model predictions is underestimated if the uncertainty of the fixed non-identifiable
parameter is neglected. A more realistic prediction uncertainty estimate is calculated if
the prior knowledge about the uncertainty of the non-identifiable parameter in a Bayesian
sense is included in the calculation of the prediction uncertainty. If, in the third situation,
the available knowledge of the non-identifiable parameter is poor the traditional solution
is to simplify the model equations until all model parameters are identifiable (the parsi-
mony principle). This leads to reasonable data interpolation, but the uncertainty with
extrapolation is drastically underestimated. The Bayesian technique succeeds in situa-
tions involving non-identifiable parameters because parameter distributions are estimated
rather than just point values.

PART II: Application

A mathematical model for plankton, nutrient (phosphate, ammonia and nitrate) and oxy-
gen dynamics in lakes was developed which was to describe data sets available for lake
Zürich. These data sets of physical, chemical and biological parameters are monthly mea-
sured depth profiles over several years. The scope of modelling is the reproduction of the
seasonal nutrient dynamics of the lake. Sunlight data is averaged over daily periods so
that the simulation of the diurnal cycle is out of scope of the model. The model does not
describe detailed food-web interactions in the lake, but is an engineering-type model that
aggregates algal and zooplankton species as far as possible.

State Variables  The biological part of the model is represented with the three plank-
ton classes *Planktothrix rubescens* (most abundant blue-green algae in lake Zürich) other
algae and zooplankton. Biodegradable and inert organic matter summarize organic par-
ticles resulting from allochthonous sources, from death of algae and zooplankton, and
from zooplankton excretion as fecal pellets. Phosphate, ammonium and nitrate are the
most relevant nutrients and together with dissolved oxygen represent the dissolved state
variables of the model.

Physical Processes  Dissolved and suspended substances are transported in the water
column by vertical mixing. In addition, algae and suspended particles are transported by
sedimentation. Two sediment layers are used to approximately describe substances and
processes in the sediment where especially aerobic and anoxic mineralization of organic
compounds are of importance. A diffusive exchange with the sediment pore water affects
concentrations of dissolved substances.

Biogeochemical Processes  The biochemical processes considered in the model include
growth, respiration and death of zooplankton, algae and *Planktothrix rubescens*. Growth
of the phytoplankton depends on sunlight radiation and the concentration of nutrients
in the water whereas for the growth of the zooplankton algae are consumed. The respiration process describes the maintenance of the organisms at the expense of their own biomass. The three plankton classes are transformed by the death processes into inert and degradable organic material.

Aerobic and anoxic mineralization processes account for oxidation of the degradable organic matter and close the nutrient cycle of the lake by releasing phosphate and ammonium in dissolved form. Aerobic mineralization using dissolved oxygen leads to a significant reduction of oxygen concentrations in the deep hypolimnion. Anoxic mineralization using nitrate as an oxidant, also known as denitrification, contributes to the large elimination of nitrogen in the lake. Nitrification transforming ammonia to nitrate by consuming oxygen is also considered as a process in the model.

In order to bring the simulations into agreement with measured profiles, a variable stoichiometry of primary production with respect to phosphate was necessary, so that the phosphorus content of newly built algae can decrease when phosphate gets limited. An earlier model version with a constant stoichiometry of primary production produced less algal biomass which lead to a significantly smaller reduction of nitrate concentration in the epilimnion. In the current model with a variable stoichiometry, due to the smaller phosphorus content of newly built algal biomass, much more biomass can be produced under phosphorus limiting conditions, which leads to a higher nitrogen uptake in the epilimnion.

Moreover, consideration of a phosphate uptake process on sinking particles below the photic zone was necessary. Without this process, the model shows two important deficiencies: First, the thickness of the surface layer within which phosphate concentration becomes very low is significantly too small and second, the gradient of the phosphate profile in the hypolimnion is too small. The latter observation indicates a too small upward flux of phosphate released from the sediment. The introduction of the phosphate uptake process solves both problems. First, the uptake process decreases the concentrations below the epilimnion, and second, due to the increased downward flux of phosphate attached to sinking particles, a higher phosphate release from the sediment is possible.

After the calibration of model parameters which will be discussed in the next paragraph the model was able to reproduce the essential properties of the nutrient, oxygen and plankton dynamics in the lake over several years. Also the particulate fluxes of organic carbon and phosphorus were in agreement with the measurements. The predictions of the model to the years after the calibration period were also very close to the measurements. This is not as astonishing as it may seem, because the external driving conditions and therefore also the seasonal pattern of concentrations did not significantly change during this time.

Compared to the complexity of the real system, the biogeochemical model presented is relatively simple. However, from the point of view of identification and calibration of the parameters the model is rather complex. It is not possible to determine the values of all parameters with the monthly measured profiles of oxygen, nitrate, ammonia, phosphate, Planktonathrix rubescens, other algae and zooplankton. The computational requirements to perform a simulation with the lake model are too high for a Bayesian analysis. This situation therefore reflects the recommendation under B1b) in the framework presented in the conceptual Part I. As a start, literature values for all parameters were used. Then values
of those parameters that were assumed to be inaccurately known and had a significant influence on the model results were iteratively changed until the simulations led to a qualitative agreement with the measurements. Thereafter a formal sensitivity, identifiability and uncertainty analysis of the model based on the given layout of the measurement was performed.

The central question to be answered by this procedure was, which model parameters should be included into a formal parameter estimation process in order to minimize prediction uncertainty. The analysis consisted of the following steps:

- Based on prior knowledge and on the expected identifiability, model parameters were grouped into parameters potentially to be estimated from the data and parameters to be determined from other sources. Typically, kinetic parameters specific for the population in Lake Zürich were potentially to be estimated, stoichiometric, physical and input parameters were assumed to be known from other sources. This distinction had to be made because it is not possible to estimate input parameters, kinetic parameters and stoichiometric parameters at the same time.

- Based on prior estimates of parameter uncertainty and on linear error propagation of this uncertainty through the model, a sensitivity ranking of the model parameters according to their contribution to prediction uncertainty was made.

- If sensitivity functions for a given parameter set are approximately linearly dependent, changes in model results induced by a change in one parameter can be approximately compensated by appropriate changes in the other parameters of the parameter set. This means that a parameter set can be poorly identifiable although the results are sensitive to all of the parameters individually. Therefore parameter sets were selected for weighted least squares parameter fits based on this sensitivity ranking and on information on potential identifiability problems gained by an analysis of approximate linear dependence of sensitivity functions.

- The results were analysed for the quality of fit and for reasonableness of estimated parameter values and all the steps were iteratively repeated until convergence occurred. Iteration is necessary because both sensitivity ranking and approximate linear dependence analysis of sensitivity functions are local analyses that depend on the selected parameter values.

As a result of the sensitivity ranking, contributions to model prediction uncertainty of the model results (oxygen, nutrients and plankton concentrations in the water column of the lake) were identified 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 loading, discharge and mixing contributed most to the prediction uncertainty.

The parameters of algae and zooplankton growth, respiration and death, due to their high contribution to model prediction uncertainty, were strong candidates to be selected for the final parameter set. However, the analysis of approximate linear dependence of sensitivity functions revealed that due to the possibilities of compensating effects not all these parameters together could be estimated from the available data and the two parameters describing algal respiration and algal death had to be excluded from the parameter
set. Instead two parameters responsible for algal phosphate uptake and one parameter describing the sediment thickness were included to the set because of their relatively large error contribution on one hand and their insignificant increase of the degree of linear dependence of the sensitivity functions on the other hand.

With assumed values for respiration and most death parameters and a fit of growth parameters, a good correspondence of model results with data could be achieved. As compared with the literature, all model parameters seemed to be in a reasonable order of magnitude. But for small changes in the values of respiration and death parameters a similarly good fit could be achieved by adapting the growth parameters. This shows that the non-identifiability may not have severe consequences as long as net growth rates are similar. The model cannot easily be 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.

Finally, the results from the identifiability analysis above together with the prior uncertainty estimates of the parameters were used to estimate the uncertainty in model predictions. The computation time required for simulations with the lake model make it very expensive to do a regional uncertainty analysis with the aid of Monte Carlo simulations. In order to get a rough estimate of prediction uncertainty without such a high computational cost, linearized error propagation is used instead in the following way. First, the uncertainty of the results assuming uncorrelated parameters were calculated with the prior uncertainty of the parameters. Second, the prior uncertainties for the fitted parameters were replaced with the estimated standard deviations and correlations from the fit. While this approach is conceptually not a convincing technique their application in the didactical example of microbial growth kinetics in Part I showed similar results as a conceptually more convincing Bayesian analysis.

This uncertainty analysis led to very large estimates of the uncertainty of model predictions. The result is a contrast to the very good agreement of model simulations with measurements even 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 variables. For changes in driving conditions of the lake (discharge, loadings, light, etc.) a larger deviation of predictions from observations can be expected.

Key results:

1. A careful discussion of the problem of parsimonious models with respect to model predictions
2. A model that is able to reproduce the biochemical processes occurring in Lake Zurich
3. A careful identifiability and uncertainty analysis of the lake model
Zusammenfassung


Teil I: Konzepte


Das Sparsamkeitsprinzip ist ein sinnvolles Kriterium für die Selektion und die Identifikation von Modellen, bei dem überparameterisierte Modelle ausgeschlossen werden. Aus
der Sichtweise der Modellidentifikation macht es während der Messperiode keinen Sinn ein Modell zu bilden, dessen Parameter nicht mit den vorhandenen Messungen bestimmt werden können. Wird aber beabsichtigt das Modell für Prognosen außerhalb des Messbereichs zu verwenden, dann können schlecht identifizierbare Parameter auftreten, die relevant sind für Prozesse mit geringer Ausprägung im Identifikationsbereich, die im Extrapolationsbereich jedoch an Einfluss gewinnen. Falls Vorwissen über solche Prozesse verfügbar ist, sollten diese Prozesse in die Analyse einbezogen werden, weil deren Vernachlässigung unter Umständen zu einer signifikanten Unterschätzung der Prognoseunsicherheit führen kann.


Die Grundidee der Bayesschen Methode ist einfach: Das Bayessche Theorem über bedingte Wahrscheinlichkeiten wird verwendet um eine a priori Verteilung der Modellparameter unter Verwendung von neuen Messwerten zu einer a posteriori Verteilung zu aktualisieren. Die a posteriori Parameterverteilungen können dann verwendet werden, um die Verteilung der Modellprognosen zu berechnen. Dazu werden Monte Carlo Techniken verwendet, weshalb die Berechnungen sehr hohe Anforderung an Computerressourcen stellen. Weil Wahrscheinlichkeitsverteilungen an der Stelle von Punktschätzern verwendet werden, ist beim Bayesschen Ansatz die Identifizierbarkeit der Parameter keine notwendige Bedingung. Die Bayessche Modellmittelung erweitert die Idee der Bayesschen Identifikation auf die Ebene der Modellstrukturen, so dass für jedes Modell eine Wahrscheinlichkeit berechnet wird. Eine Modellprognose basiert nicht mehr auf der Wahl eines eindeutigen Modells,

A) Die Modellparameter aller Modellstrukturen sind genau identifizierbar.

→ führe eine klassische Parameterschätzung für alle verschiedenen Modellstrukturen durch.

Bezüglich der Selektion der Modelle, können zwei Fälle unterschieden werden:

a) Das Selektionskriterium einer einzelnen Modellstruktur ist gegenüber allen anderen Modellstrukturen deutlich besser.

→ Wähle dieses Modell und führe eine klassische Modelldarstellung durch (Maximum-Likelihood Punktsehätzer der Parameter und linearisierte Unsicherheitsschätzung).

b) Es gibt eine Teilmenge unter den Modellstrukturen mit ähnlich gutem Selektionskriterium.

→ führe eine Bayessche Modellmittellung mit den Modellen dieser Teilmenge durch.

B) Es existieren schlecht identifizierbare Parameter.

→ Bayessche Inferenz und Modellmittelung sollte durchgeführt werden.

Anhand der Rechenanforderungen können zwei Fälle unterschieden werden.

a) Es ist genug Rechenleistung verfügbar um die Wahrscheinlichkeitsverteilungen mit Monte Carlo Simulationen zu approximieren.

→ führe eine kombinierte Bayessche Parameterschätzung und Modellmittelung durch.

b) Es ist nicht genügend Rechenleistung verfügbar um Monte Carlo Simulationen von den in Betracht gezogenen Modellen durchzuführen.

→ Versuche schlecht identifizierbare Parameter zu bestimmen und zu isolieren (mittels Reparameterisierung). Falls dies möglich ist, wähle anhand gleicher Abstände in der integralen a priori Verteilungsfunktion verschiedene Werte für diese Parameter und behandle Modelle mit unterschiedlichen Werten für die schlecht identifizierbaren Parameter als separate Modellstrukturen. Führe weiter mit Punkt A). Falls nicht-identifizierbare Parameter nicht isoliert werden können, wende die oben beschriebene Technik in zyklischer Folge auf eine Teilmenge der Parameter an. Die Teilmengen der Parameter mit festgesetzten Werten müssen so gewählt werden, dass die anderen Parameter identifiziert werden können.
In Kapitel 2 wird ein didaktisches Beispiel verwendet, um die Eigenschaften des klassischen und des Bayesschen Ansatzes in Situationen unterschiedlicher Identifizierbarkeit zu illustrieren. Es wird ein einfaches Modell verwendet, das die Wachstumskinetik von Mikroorganismen in Abhängigkeit der Substratkonzentration beschreibt. Um sich auf das Problem der Identifizierbarkeit zu konzentrieren, wird die Modellstruktur, mit der die synthetischen Daten erzeugt wurden, als bekannt vorausgesetzt. Drei verschiedene Situationen mit unterschiedlichen Daten- und Vorwissensverfügbarkeiten werden diskutiert. Mit einem vollständigen Datensatz im ersten Fall sind alle Modellparameter identifizierbar. Bei nur teilweiser Verfügbarkeit der Daten ist ein Modellparameter nicht identifizierbar. Dies führt zu zwei weiteren Fällen, wobei unterschieden wird ob wenig oder viel Vorwissen über den nicht-identifizierbaren Parameter vorhanden ist.


Teil II: Anwendung


Zustand Variablen Der biologische Teil des Sees ist im Modell durch die drei Algenklassen *Planktothrix rubescens* (häufigste Blaualge im Zürichsee), andere Algen und Zooplankton vertreten. Organische Partikel allochthoner Herkunft, vom Tod der Algen und des


Als Ergebnis dieser Analyse sollte bestimmt werden, welche Modellparameter in einer formale Parameterschätzung einzubeziehen sind, damit die Prognoseunsicherheit möglichst gering ausfällt. Die Analyse bestand aus den folgenden Schritten:


- Aufgrund von a priori Schätzungen der Parameterunsicherheiten und linearer Fehlerfortpflanzung dieser Unsicherheiten auf die Modellrechnungen, wurde eine Sensitivitätsanalyse durchgeführt und die Parameter anhand ihres Beitrags zur Prognoseunsicherheit geordnet.

- Wenn die Sensitivitätsfunktionen für eine gegebene Gruppe von Parametern un- oder besser linearabhängig sind, dann kann die Änderung des Modellresultats aufgrund der Variation eines Parameters durch eine entsprechende Korrektur in den übrigen Parametern kompensiert werden. Dies bedeutet, dass eine Gruppe von Parametern schlecht identifizierbar sein kann, obwohl die Modellresultate auf die einzelnen Parameter sensitiv reagieren. Darum wurden die Parameter, die mit den Daten geschätzt werden sollten, aufgrund von zwei Kriterien ausgewählt: Der Sensitivität einerseits, kombiniert mit der Information über potentielle Identifikationsprobleme, die durch die Analyse der näherungsweisen linearen Abhängigkeit der Sensitivitätswerte erhalten wurden.
Die Resultate wurden auf die Übereinstimmung mit den Daten und die Parameterwerte auf ihre Plausibilität überprüft und die beschriebenen Schritte wurden iterativ wiederholt, bis die Resultate konvergierten. Ein iteratives Vorgehen war notwendig, weil sowohl die Sensitivitätsanalyse wie auch die näherungsweise lineare Abhängigkeit der Sensitivitätsfunktionen lokale Analysen sind, die von den Werten der gewählten Parametern abhängen.

Das Ergebnis der Sensitivitätsanalyse ergab, dass die Beiträge zur Prognoseunsicherheit der Modellresultate (Sauerstoff-, Nährstoff- und Planktonkonzentrationen in der Wassersäule des Sees) am grössten sind für die Halbsättigungskonstante des Algenwachstums bezüglich der Lichtintensität und für Parameter die proportional sind zum Wachstum, der Respiration und dem Tod von Algen und Zooplankton.

Koeffizienten der Mineralisierung und Nitrifikation, der Temperaturabhängigkeit biologischer Prozesse, so wie stöchiometrische Parameter lieferten einen deutlich kleineren Beitrag zur Unsicherheit der Modellprognosen. Unter den externen Parameter waren die Beiträge für die Stickstoff- und Phosphorzuflüsse, den Abfluss und die Mischung am grössten.

Die Parameter des Wachstums, der Respiration und des Todes von Algen und Zooplankton waren durch ihre hohen Beiträge zur Prognoseunsicherheit die Hauptkandidaten für die Wahl der zu schätzenden Parameter.


Schliesslich wurde das Ergebnis der Identifizierbarkeitsanalyse zusammen mit den a priori Unsicherheiten der Parameter verwendet um die Unsicherheit der Modellprognosen zu schätzen. Die benötigte Rechenzeit einer Simulation des Scenomells machte die Anwendung einer regionalen Unsicherheitsanalyse mittels Monte Carlo Simulationen praktisch undurchführbar. Um trotzdem eine grobe Schätzung zu bekommen, wurde stattdessen eine lineare Fehlerfortpflanzung durchgeführt. Dazu wurde zuerst die Unsicherheit der Modellresultate unter der Annahme unkorrelierter Parameter und der Verwendung der a priori Unsicherheiten der Parameter berechnet. Als nächstes wurden die a priori Unsicherheiten der geschätzten Parameter durch die geschätzten Standardabweichungen und Korrelationen ersetzt. Dieses Vorgehen stellt konzeptuell betrachtet keine sehr überzeugende Technik
Das didaktische Beispiel in Kapitel 2 ergab jedoch für dieses Vorgehen ähnliche gute Resultate wie bei der konzeptuell vielversprechenderen Bayesschen Analyse.


Schlüsselresultate:

1. Diskussion über die Wahl der Modellkomplexität im Hinblick auf Prognosen
2. Ein Modell, das die biochemischen Prozesse im Zürichsee beschreibt
3. Identifizierbarkeits und Unsicherheitsanalyse des entwickelten Seemodells
Part I

Concepts
Chapter 1

On the Usefulness of Overparameterized Ecological Models

Abstract: The parsimony principle is a useful criterion for reducing the non-uniqueness in classical system identification. However, if a uniquely selected model is used for prediction, the disregard of the uncertainty in model structure can lead to an underestimation of the uncertainty in model forecasts. This is particularly the case when processes become important during the prediction period that were insignificant during the identification period. If some knowledge of such processes is available, they should be included in the analysis. This requires an identification and forecasting technique that can use prior knowledge and can handle overparameterized, non-identifiable models. The Bayesian approach to statistical inference is such a technique. In this paper, the advantages and disadvantages of both the classical and the Bayesian methodology are discussed, and it is argued that from a methodical point of view, for poorly identifiable systems typical in ecological modelling, the Bayesian technique is the superior approach. Because of the huge computational requirements of the Bayesian technique a recommendation is given for an improved identification and forecasting procedure that, depending on the identifiability of the investigated system and on the power of the available computational facilities, uses the advantages of the appropriate method.

Keywords: Uncertainty; Prediction; Model; Identifiability; Bayesian statistics.

1.1 Introduction

The classical process of system identification consists of proposing a set of models mainly based on prior knowledge of the system to be investigated, selecting the most adequate of these models based on a criterion which favours good fitting models and penalizes complicated model structures, and estimating the parameter values and their uncertainties for
the selected model. The selected model is then used for forecasting and the uncertainty of model predictions is calculated by propagating the estimated uncertainty of the parameters to the predicted variables (Spriet 1985; Söderström and Stoica 1989, and many others). There is an increasing discussion in the statistical literature on the adequacy of this procedure, especially on using the same data for model selection and for the estimation of parameter uncertainty of the selected model and on neglecting the uncertainty in model structure (Chatfield 1995; Draper 1995). The conclusion is that in most cases this procedure leads to overoptimistic results by underestimating the uncertainty in model predictions. This has important consequences for ecological modelling, where the uncertainty in model structure is usually very large (O'Neill and Gardner 1979; Bartell, Cale, O'Neill, and Gardner 1982) and where the assumption of the existence of a single “true” model is not justified. An ecological system is too complicated to be modelled in all its detail. There may exist several models which approximate the investigated aspects of the behaviour of such a system equally well.

This paper begins with a brief review of the classical system identification methodology that ends with a discussion of the advantages and problems of this approach. Then, a brief review of Bayesian techniques is given. These techniques can deal with sets of models and overparameterized models, a feature that is important for estimating prediction uncertainty. These techniques can solve the problems of the classical identification procedure, but their application is computationally very demanding. For this reason, an identification and forecasting procedure is proposed that, whenever possible, is based on the classical technique, but switches to the discrete or to the continuous Bayesian approach as required. This procedure avoids the requirement of identifiable models in situations where there is strong evidence that their use would lead to an underestimation of the prediction uncertainty. It proposes a pragmatic discrete approximation to the continuous Bayesian case for situations where the unapproximated technique would be appropriate but is too computationally demanding.

1.2 Classical Model Identification - Principles and Problems

There exist many descriptions of the classical model identification methodology (Spriet 1985; Söderström and Stoica 1989, and many more). This section is based on the excellent review given by Spriet (Spriet 1985). Spriet gives the following Guiding Principles for models to be included into the set of models to be tested:

**GP1: Physicality:** The structure of a model should be related as directly as possible to the causal mechanisms acting in the system under investigation.

**GP2: Characterizability:** Model structures defining competing model classes should lead to a well-defined decision process for model structure selection.

**GP3: Identifiability:** It must be possible to estimate the unknown parameters of a proposed model structure to a reasonable accuracy using the available data.

Note that this first step of model selection depends not only on the system to be investigated but also on the prior knowledge and on the objectives of the modeller. The principles
GP2 and GP3 are based on the idea that it must be possible to uniquely determine the model and its parameter values. The principle GP3 has been expressed very distinctly by the following statement (Spriet 1985):

**S1:** “A model structure for which parameters cannot be estimated is useless”.

Once the set of competing models is determined, Spriet gives the following *Decision Criteria* for final model selection:

**DC1: Quality of fit:** The model structure should be able to represent the measured data to an adequate accuracy.

**DC2: Parsimony:** The model structure should be as simple as possible.

**DC3: Balanced accuracy:** The selection of the model structure must be based on an adequate compromise between available information and model accuracy.

The selection criterion is often based on a trade-off between quality of fit (DC1) and model complexity (DC2). The criterion DC3 is given more precisely by the following statement (Spriet 1985):

**S2:** “There is little sense in retaining an eventually more valid model structure, if the parameters in it can only be poorly estimated; an approximate structure with more precise parameter values may be more adequate”.

There are similar ideas behind the guiding principle GP3 and the decision criterion DC3.

There are some problems with the identification procedure described above. First, it is not really known how to formulate a criterion to make the desired trade-off between quality of fit (DC1) and model complexity (DC2). Several criteria have been proposed for application to special cases, but it is not yet known which procedure is most adequate for application to nonlinear regression models, as used in ecological modelling (Spriet 1985).

The second point is of a more fundamental nature: The techniques used to estimate model parameter uncertainties are based on the assumption that the model structure is known to be correct. This assumption is violated if the model is selected out of a set of possible models with the aid of the same data as is used for parameter uncertainty estimation. In other words: “it is well-known to be logically unsound and practically misleading to make inferences as if a model is known to be true when it has, in fact, been selected from the same data to be used for estimation purposes” or “the good fit of the best fitting model should not be surprising” (Zhang 1992; Chatfield 1995). The influence of this logical error on uncertainty estimates is very difficult to quantify and, according to Chatfield, there has been very little published work on statistical inference that explicitly considers the fact that the model was selected using the same data (Pötscher 1991; Zhang 1992, are some exceptions). The consequence of neglecting this fact is usually an underestimation of the model prediction uncertainty (Chatfield 1995).

The third problem with the identification procedure described above is the most fundamental one. Ecological systems are extremely complex. All models of such systems are very drastic simplifications of reality. For this reason, there is no obvious justification for the selection of a single model as the “true” description of an ecological system. It is much more typical to have several different model structures that describe the system in an adequate way. This idea has been expressed succinctly as “all models are wrong, but
some are useful" (Chatfield 1995). In this context “wrong” does not mean that there are no correct concepts included in the model, rather it means that each model has its limitations and features can be found in the real system that are not contained in any given model. Given this, from a conceptual point of view, it is not clear why one should insist on the dogma of a unique determination of model structure and associated parameter values ($GP_3$, $GP_3$, $DC_2$, $DC_3$). The practical advantage of a unique identification is that it is easier to interpret a single model structure. On the other hand, this interpretation may be misleading, if other model structures give similarly good descriptions of the data. In such a situation, selection of a single model neglects the uncertainty in model structure and results in the underestimation of the overall prediction uncertainty (Draper 1995).

The fourth problem of the methodology described above is in use of the identified model for prediction. The need for identifiable parameters in a selected model ($GP_3$) is obvious if the focus of the investigation is identification alone. However, if the model is intended for making predictions that require an extrapolation outside of its identification domain, poorly identifiable parameters, relevant to processes that may be important in the extrapolation domain but are of less significance in the identification domain, are sometimes encountered. Discarding such processes so that only models with identifiable parameters are considered can lead to a significant underestimation of the prediction uncertainty. In other words, the classical identification methodology is appropriate for the investigation of relevant processes in the investigated situation (this is what this methodology was developed for), but it may not be appropriate for predictions that require an extrapolation outside of the identification domain (this is how this methodology is often used).

Classical statistics is only suitable for describing the uncertainty of model predictions that are caused by the intrinsic randomness of a system. The uncertainty of forecasts for environmental systems has the additional (systematic) components of uncertainty due to the simplified description of the system (aggregation error) and due to lack of knowledge on the relevant processes. It is not evident that the first of these components is dominant.

The advantages and problems of the classical model selection procedure can be summarised as follows:

**Advantages:**

- Selection of a single model with unique parameter values is desirable from the point of view of ease of interpretation.

- Forecasts based on a point estimate of parameters of a single model have low computational requirements (as compared to the methods described in the next section).

**Problems:**

- Model selection criteria are, to a certain degree, arbitrary. This is particularly problematic for the general case of nonlinear regression, because most criteria with model complexity terms were derived for special situations such as order selection of autoregressive time series and it is not clear which of these is most adequate for the general case of nonlinear regression (Spriet 1985).

- Model selection and statistical inference using the same data set, and assuming the selected model structure to be true in the inference step, is logically incorrect and leads, in most cases, to an underestimation of uncertainty (Chatfield 1995).
Selection of a single model is often inappropriate because there is no single "true" model of an ecological system but there are several adequate simplifying descriptions. Furthermore, selection of a single model leads to a discard of the uncertainty in model structure and therefore also to an underestimation of total uncertainty (Draper 1995).

The requirement of identifiable model parameters can lead to the selection of overly simple models that underestimate prediction uncertainty in extrapolation domains.

The methods of classical statistics are suitable for describing the uncertainty caused by the intrinsic randomness of a system. However, without a new interpretation of probability it is not possible to describe the uncertainty due to lack of knowledge of a system with probability theory.

In order to solve these problems, it is necessary to replace the model selection process described above with a procedure that is able to combine the results of different models of similar adequacy, and to deal with overparameterized models which contain poorly identifiable parameters. Such a procedure cannot be based on a single point estimate of model parameters; it must be able to deal with probability distributions of parameters. Furthermore, it must be possible to include prior knowledge of processes and parameters into the analysis (in order to partially compensate non-identifiability). In the next section it is shown that Bayesian inference fulfills both of these requirements (at the expense of a decrease in the precision of the meaning of probability and an increase in computational load).

1.3 Bayesian Inference - Advantages and Problems

In classical statistics, probability theory is only applied to random events that, at least in principle, can be sampled to determine empirical frequency distributions, which approximate the true probability distributions asymptotically. In Bayesian statistics, the field of application of probability theory is extended. Here, the notion of probability is also used to represent uncertainty or partial knowledge of a system that may be deterministic. As an important example, in classical statistics the unknown parameters of a deterministic model are not random variables (only the estimators of their values that depend on the measurements are random variables) whereas in Bayesian statistics the uncertainty in parameter values is described with probability distributions. If this extension of the notion of probability is accepted, the main idea of the Bayesian approach to statistics is very simple: Bayes' theorem on conditional probabilities (which is a theorem of classical statistics) is used to update a prior probability distribution of model parameters to a new, posterior distribution that includes the additional information from new measurements (Bernardo and Smith 1994; Gelman, Carlin, Stern, and Rubin 1995, and literature cited therein).

From these probability distributions of the parameters, point estimators can be derived e.g. by taking the expectation value, the median or the maximum of the distribution. These parameter estimates, together with a measure of their uncertainty (e.g. their standard deviations) can be used analogously to the classical case to calculate model predictions. As an alternative to this procedure, the probability distributions of the parameters can be used directly to calculate probability distributions for model predictions. The latter possibility, although computationally much more demanding, is particularly attractive from a methodical point of view, because it is not based on simplifying assumptions such as
the linearization of the model, the assumption of normally distributed parameters or the
disregard of parameter correlations.

Bayesian estimation of model parameters is done with the aid of the following theorem
on conditional probabilities (continuous case of Bayes' theorem; cf. references given above):

\[
f_{\text{post}}(\mathbf{p}_M \mid \mathbf{y}_{\text{meas}}) = \frac{f_M(\mathbf{y}_{\text{meas}} \mid \mathbf{p}_M) f_{\text{pri}}(\mathbf{p}_M)}{f(\mathbf{y}_{\text{meas}})}
\]

(1.1)

where \( f_{\text{post}}(\mathbf{p}_M \mid \mathbf{y}_{\text{meas}}) \) is the conditional posterior probability density of the \( m \) parameters \( \mathbf{p}_M = (p_{M,1}, \ldots, p_{M,m}) \) of the model \( M \) under the condition that the \( n \) data points \( \mathbf{y}_{\text{meas}} = (y_{\text{meas},1}, \ldots, y_{\text{meas},n}) \) were measured, \( f_M(\mathbf{y}_{\text{meas}} \mid \mathbf{p}_M) \) is the conditional probability density for measuring \( \mathbf{y}_{\text{meas}} \) given the parameters \( \mathbf{p}_M \) (this is the likelihood function of the model), \( f_{\text{pri}}(\mathbf{p}_M) \) is the prior probability density of the parameters and \( f(\mathbf{y}_{\text{meas}}) \) is the probability density of the measurements. This theorem states (after multiplication with \( f(\mathbf{y}_{\text{meas}}) \)) that the product of the probability density of \( \mathbf{y}_{\text{meas}} \) and the probability density of \( \mathbf{p}_M \) conditional on the values of \( \mathbf{y}_{\text{meas}} \) is the same as the product of the probability density of \( \mathbf{p}_M \) and the probability density of \( \mathbf{y}_{\text{meas}} \) conditional on the values of \( \mathbf{p}_M \) (both products equal to the simultaneous probability density of \( \mathbf{y}_{\text{meas}} \) and \( \mathbf{p}_M \)). For deterministic models, the likelihood function \( f_M(\mathbf{y}_{\text{meas}} \mid \mathbf{p}_M) \) is usually constructed by assuming identically distributed independent measurements centered at the deterministic values \( \mathbf{y}_M(\mathbf{p}_M) \) of the model (in many cases normal distributions are assumed). A prior distribution of the parameters that represents the knowledge before the data \( \mathbf{y}_{\text{meas}} \) were measured must be provided (e.g. a uniform distribution of each parameter within a reasonable interval). Because the denominator \( f(\mathbf{y}_{\text{meas}}) \) of the above expression is independent of the parameters \( \mathbf{p}_M \), it can be determined by normalization of the probability distribution.

After the identification step described by equation (1.1) has been performed, probability densities for predicted observables can be calculated by integrating the product of the likelihood function of the model and the posterior distribution of the parameters over all possible values of the parameters: a

\[
f(\mathbf{y} \mid M, \mathbf{y}_{\text{meas}}) = \int f_M(\mathbf{y} \mid \mathbf{p}_M) f_{\text{post}}(\mathbf{p}_M \mid \mathbf{y}_{\text{meas}}) \, d\mathbf{p}_M
\]

(1.2a)

where \( \mathbf{y} = (y_1, \ldots, y_k) \) are the model predictions at \( k \) given locations in time and space. In case of a deterministic model, equation (1.2a) gives the probability distribution of measurements of predicted variables including the measurement error, the characteristics of which are contained in the likelihood function. In this case it may be of more interest to calculate the probability distribution of the "exact" predicted variables without measurement error and to consider the measurement error while interpreting the results. The probability density for \( \mathbf{y} \) is then given by propagating the posterior probability distribution of the parameters \( f_{\text{post}}(\mathbf{p}_M \mid \mathbf{y}_{\text{meas}}) \) through the deterministic model function \( \mathbf{y}_M(\mathbf{p}_M) \):

\[
f(\mathbf{y} \mid M, \mathbf{y}_{\text{meas}}) = \frac{\partial f}{\partial y_1 \ldots \partial y_k} \int_{\mathbf{y}_M(\mathbf{p}_M) \leq \mathbf{y}} f_{\text{post}}(\mathbf{p}_M \mid \mathbf{y}_{\text{meas}}) \, d\mathbf{p}_M
\]

(1.2b)

where \( \mathbf{y}_M(\mathbf{p}_M) \leq \mathbf{y} \) is an abbreviation for \( (y_{M,1} \leq y_1, \ldots, y_{M,k} \leq y_k) \). The equation (1.2b) is an analytical expression for the distribution of model predictions. In order to calculate this distribution numerically, Monte Carlo simulation with random sampling
of the parameters from \( f_{\text{post}}(\mathbf{p}_M \mid \mathbf{y}_{\text{meas}}) \) is done (Rubinstein 1981; Sobol 1994). This procedure does not make explicit use of equation (1.2b).

Note that Bayesian estimation of model parameters does not require the identifiability of the parameters. For an identifiable model the posterior distribution of the parameters \( f_{\text{post}}(\mathbf{p}_M \mid \mathbf{y}_{\text{meas}}) \) calculated according to equation (1.1) has a single sharp maximum. If the model is not identifiable, this is not the case. Instead, \( f_{\text{post}}(\mathbf{p}_M \mid \mathbf{y}_{\text{meas}}) \) may have a much more complex shape with various combinations of the parameters \( \mathbf{p}_M \) having a similarly high value of the probability density. This property of the probability density does not have a negative effect on model predictions using equation (1.2). For this reason, in the Bayesian analysis, the principle GP1 of physicality mentioned in the previous section may also be fulfilled in situations where it leads to a poorly identifiable model that violates the principle GP3. The usefulness of such a non-identifiable overparameterized model is that it can give more realistic uncertainty bounds for model extrapolations.

A discrete version of equation (1.1) can be used in a similar way to determine probabilities of different discrete model structures with known values of continuous parameters (semi-discrete case of Bayes’ theorem; cf. references given above):

\[
P_{\text{post}}(M_i \mid \mathbf{y}_{\text{meas}}) = \frac{f_{\mathbf{M}}(\mathbf{y}_{\text{meas}} \mid M_i) P_{\text{pri}}(M_i)}{f(\mathbf{y}_{\text{meas}})}
\]

(1.3)

where \( M = \{M_i\}_{i=1}^l \) is a set of model structures, \( P_{\text{post}}(M_i \mid \mathbf{y}_{\text{meas}}) \) is the posterior probability for the model structure \( M_i \), \( f_{\mathbf{M}}(\mathbf{y}_{\text{meas}} \mid M_i) \) is the likelihood function for all models conditional on the model structure \( M_i \), \( P_{\text{pri}}(M_i) \) is the prior probability for model structure \( M_i \) and \( f(\mathbf{y}_{\text{meas}}) \) is the probability density of the measurements. As above, \( f(\mathbf{y}_{\text{meas}}) \) can be determined by normalizing the probabilities.

The probability density for model predictions including the measurement error is calculated similarly to equation (1.2a) by the so-called “Bayesian model averaging” technique according to a

\[
f(\mathbf{y} \mid M, \mathbf{y}_{\text{meas}}) = \sum_{i=1}^l f_{\mathbf{M}}(\mathbf{y} \mid M_i) P_{\text{post}}(M_i \mid \mathbf{y}_{\text{meas}})
\]

(1.4a)

The (discrete) probabilities for model predictions of deterministic models without including the measurement error are given by

\[
P(\mathbf{y} \mid M, \mathbf{y}_{\text{meas}}) = \sum_{i: \mathbf{y}_M = \mathbf{y}} P_{\text{post}}(M_i \mid \mathbf{y}_{\text{meas}})
\]

(1.4b)

The discrete distribution of continuous results given by equation (1.4b) is atypical. This distribution occurs only in the absence of continuous parameters. If the models in the set \( M \) contain parameters that have been estimated classically before execution of the Bayesian model averaging process, equation (1.4a), with a likelihood function that only includes parameter uncertainty and not measurement uncertainty, should be applied instead of equation (1.4b). The original equation (1.4a) uses a likelihood function that includes both error contributions.

Note that the Bayesian model averaging procedure is not based on the selection of a unique model but it consists of using a weighted average of the predictions of different models. Models with a very small weight may be discarded, but the possibility of the
existence of similarly important models is given. In the discrete case, the prior distribution is usually assumed to be $P_{\text{pr}}(M_i) = 1/l$ for $i = 1, \ldots, l$.

A combination of the continuous case given by the equations (1.1) and (1.2) with the discrete case given by the equations (1.3) and (1.4) is the most universal technique. In this case the identification equation is

$$f_{\text{post}}(M_1, M_1 | y_{\text{meas}}) = \frac{f_M(y_{\text{meas}} | M_1, \mathbf{p}_{M_1}) f_{\text{pr}}(M_1, \mathbf{p}_{M_1})}{f(y_{\text{meas}})}$$ (1.5)

where it is natural to set the prior distribution to the product

$$f_{\text{pr}}(M_1, \mathbf{p}_{M_1}) = P_{\text{pr}}(M_i) f_{M_{i|\text{pr}}}(\mathbf{p}_{M_1})$$ (1.6)

Note that the posterior probability for the model $M_i$ is given by

$$P_{\text{post}}(M_i | y_{\text{meas}}) = \int f_{\text{post}}(M_1, \mathbf{p}_{M_1} | y_{\text{meas}}) d\mathbf{p}_{M_1}$$ (1.7)

If for some model structures this probability is very small, these models can be neglected for the calculation of predictions.

The probability density for model predictions including the measurement error is calculated according to a

$$f(y | M_1, y_{\text{meas}}) = \sum_{i=1}^{l} \int f_M(y | M_1, \mathbf{p}_{M_1}) f_{\text{post}}(M_1, \mathbf{p}_{M_1} | y_{\text{meas}}) d\mathbf{p}_{M_1}$$ (1.8a)

The probability density for model predictions of deterministic models without including the measurement error is obtained by

$$f(y | M_1, y_{\text{meas}}) = \sum_{i=1}^{l} \frac{\partial^k}{\partial y_1 \cdots \partial y_k} \int \frac{f_{\text{post}}(M_1, \mathbf{p}_{M_1} | y_{\text{meas}})}{y_{M_1}(\mathbf{p}_{M_1}) \leq y} d\mathbf{p}_{M_1}$$ (1.8b)

Again (cf. comment following equation 1.2b), for calculating numerical approximations to this distribution, Monte Carlo simulation is employed instead of direct use of equation (1.8b).

The combined discrete and continuous case of Bayesian model averaging, given by the equations (1.5) to (1.8), is recommended by Draper (Draper 1995) as a solution to the problems of the classical identification technique listed in the previous section.

Until now there have not been many applications of Bayesian identification and forecasting techniques to environmental systems. A relatively early publication on the usefulness of Bayesian parameter updating for ecological models is due to Tiwari and coworkers (Tiwari, Hobbie, and Peterson 1978). The approach of “regional sensitivity analysis”, developed by Hornberger and Spear (Hornberger and Spear 1981; Hornberger and Spear 1983) and later extended to forecasting and renamed “Monte Carlo filtering” (Rose, Smith, Gardner, Bunkert, and Bartell 1991), was not called Bayesian analysis by the authors. In fact, this procedure is a numerical implementation of a special case of Bayesian identification that uses only probability distributions of the measurements that are constant in an interval and zero elsewhere. This approach was also promoted by Beck (Beck 1987), who called it “HSY - algorithm” according to Hornberger, Spear and Young (Hornberger
and Spear 1981; Hornberger and Spear 1983; Young 1983). Recently, another aspect of the application of Bayesian techniques to environmental systems received attention (Dilks, Canale, and Meier 1992). These authors discussed the decrease in the estimates of prediction uncertainty by using the complete multidimensional probability distribution of the parameters instead of estimating the uncertainty with the uncorrelated marginal distributions. Although this is certainly true, it is not a major concern when Bayesian inference is used to include structural uncertainty of the model, because in this case the uncertainty estimates will usually become larger than conventional ones. Decision analysis for groundwater contamination sites is perhaps the field where Bayesian techniques have been most often applied to environmental systems. Here, a subjective prior interpretation of the hydrogeological environment is successively updated using new measurements that are planned with the aid of the current knowledge and an economic decision model (Marin, Medina, and Butcher 1989; Freeze, Massmann, Smith, Sperling, and James 1990; Gorelick, Freeze, Donohue, and Keely 1993). Recently, the use of Bayesian belief networks for environmental decision support has been proposed (Varis, Kuikka, and Taskinen 1994; Varis 1995).

The advantages and problems of the use of Bayesian inference techniques can be summarised as follows:

Advantages:

• Due to the extended interpretation of probability as a degree of belief in Bayesian statistics, probability theory can be applied to all sources of uncertainty, in particular also to lack of knowledge (however, this extension is combined with a decrease in precision of the meaning of probability).

• The Bayesian technique allows prior knowledge to be explicitly included in the analysis.

• The demand for a parsimonious model may be weakened if there is a strong prior evidence for behaviour that is not explicitly contained in the data used for model identification. The loss of parameter identifiability is not a problem in the Bayesian context.

• Bayesian model averaging does, in an appropriate way, consider the fact that there is no true model of an ecological system, but there exist several different adequate descriptions.

• Estimates of the uncertainty of Bayesian model predictions are more realistic (usually larger) than those based on the classical procedure.

Problems:

• The choice of the class of models to be included and of the prior probability distributions of the model parameters must be based on prior knowledge that may be difficult to formalize (it can be difficult to convert vague prior knowledge into the form of an exact probability distribution).

• The result of the identification process is not a single relatively simple model with a unique set of parameter values, but a weighted set of models with a probability distribution of the parameters. This is often a more realistic result, but it is more difficult to interpret.
The methods used for identification and prediction, especially in the case of continuous parameters, are computationally much more demanding than are the classical techniques (in most practical cases, the integrals in equations (1.1) to (1.8) must be numerically solved using Monte Carlo simulation).

Note that the use of prior knowledge is at the same time an advantage and a disadvantage of the Bayesian approach. On one side, overparameterized models cannot be included in the analysis without the possibility of using prior knowledge (because the non-identifiability of processes in such models requires another source of information). On the other side, if not enough information is available to formulate a prior distribution of parameters, the necessity of specifying such a distribution induces an arbitrariness into the procedure. If prior distributions are very uncertain, the sensitivity of the results to variations in the prior distributions can be used to assess this type of prediction uncertainty. This approach is called “robust Bayesian analysis” (Berger 1984). In case of very imprecise prior knowledge the Bayesian approach cannot directly be applied. In this case, statistical inference that is based on statistical reasoning with imprecise probabilities (Walley 1991), on second order probability (probability distributions of probability distributions) (Cheeseman 1986), or on the Dempster-Shafer theory of belief functions (Dempster 1968; Shafer 1976; Smets 1994) may be an alternative (Wasserman 1990). However, more research is necessary in order to prove the usefulness of the application of these theories for statistical inference.

1.4 Recommended Identification and Forecasting Procedure

As discussed by Draper (Draper 1995) and briefly reviewed in the previous section, the Bayesian approach solves most of the problems of the classical identification technique. Therefore, from a methodical point of view, it is recommended that the classical identification procedure be replaced by the combined Bayesian technique given by the equations (1.5) to (1.8). In consideration of the required computational effort, however, this recommendation is too simple. For complex models with a lot of parameters it may be impossible to perform enough Monte Carlo simulations to obtain a reasonably accurate uncertainty estimate. In this case, an alternative, computationally cheaper approach is required to consider model structure uncertainty and to deal with poorly identifiable model parameters. Classical identification of the identifiable parameters for selected values of the poorly identifiable parameters of a model, combined with discrete Bayesian model averaging according to the equations (1.3) and (1.4), is such a much less demanding procedure. To account for both the computational efficiency of the classical procedure, and the superiority of the methodology of the Bayesian approach in a pragmatic way, the following compromise is recommended for environmental system identification and forecasting:

Step 1: Pre-evaluation of Models

Propose a set of model structures according to the following criteria:

*Physicality:*  
The model structures should be related as directly as possible to the causal mechanisms acting in the system under investigation.

*Diversity:*  
For poorly known processes, models with different process formulations should be included (either by the possibility of switching continuously from one...
Possibility of extrapolation: Processes that are suggested to be important for planned extrapolations must be included independent of the identifiability of their parameters (in case of their nonidentifiability the prior distributions determine the behaviour in the extrapolation domain).

Step 2: Reparameterization

If necessary, the models should be reparameterized in such a way that different behavioral ranges of the models correspond to different parameters and that unnecessary nonlinearities in the parameters are avoided.

Step 3: Identification and Prediction

According to the identifiability of the parameters, two cases must be distinguished ("accurately identifiable" means that the nonlinearity of the model with respect to the parameters is small within their estimated uncertainty ranges):

A) The model parameters of all model structures are accurately identifiable.

→ Perform a classical parameter estimation for all the discretely different model structures.

According to the "selectivity" of the models, two cases can be distinguished:

a) One model structure has a much better value of the classical selection criterion.

→ Select this model and perform model prediction in the classical way (maximum likelihood point estimation of the parameters and linearized uncertainty estimation).

b) There is a subset of model structures with similarly good values of the selection criterion.

→ Do model prediction by Bayesian model averaging among this subset of models according to the equations (1.3) and (1.4). It may be necessary to smooth the transitions between regimes corresponding to different structures.

B) There exist poorly identifiable model parameters.

→ Bayesian inference and model averaging should be applied.

According to the computational requirements, two cases can be distinguished:

a) There is enough computing power available to approximately calculate probability distributions by Monte Carlo simulation with Latin hypercube sampling (McKay, Beckman, and Conover 1979; Sobol 1979; Stein 1987).

→ If different model structures are connected by continuous parameters then do Bayesian inference according to the equations (1.1) and (1.2), otherwise do combined Bayesian parameter estimation and model averaging according to the equations (1.5) to (1.8).
There is insufficient computing power available to perform Monte Carlo simulations for the considered models.

→ Try to determine and isolate (by reparameterization) poorly identifiable model parameters. If this is possible, select different values of these parameters according to equal differences in the integral prior distribution function and treat models deviating in one of these poorly identifiable parameters as discretely different model structures in the other parameters. Then proceed according to 3A. If it is not possible to isolate non-identifiable parameters, apply the technique described above to subsets of the parameters in a cyclical way. The subsets of parameters with given values must be selected in such a way that the other parameters can be identified.

Comments to step 1: The first criterion of physicality is the same as the criterion GP1 (refer to section 1.2 for a summary of the guiding principles GP1-GP3, the decision criteria DC1-DC3, and the statements S1 and S2 (Spriet 1985) cited in this section). The second criterion of model diversity accounts for lack of knowledge of the investigated system. This criterion is related to the criterion GP2 with the important difference that alternative hypotheses should also be included in situations in which it is not yet clear if they can be distinguished from each other using the available data. The third criterion is a very important extension of the classical procedure with regard to the uncertainty estimation of model predictions. This criterion is not necessary for pure identification of important processes in a system during the investigation period; however, it is essential for obtaining realistic uncertainty estimates in the extrapolation domain. In contradiction to the classical methodology (cf. GP5, S1 and S2), identifiability is no longer a model selection criterion. Note that by including all possible (non-identifiable) models in the selection set, the uncertainties of extrapolations increase without an upper bound (“the space of all models is too big” (Draper 1995)). Therefore, this step is critically dependent on a careful pre-selection of models according to the prior knowledge of the investigated system.

Comments to step 2: The problem of finding a good parameterization of a model was not addressed before. Although, due to the identifiability requirement, this is a more serious problem in classical identification, a good parameterization is also advantageous in the Bayesian approach. Reparameterizations that describe different behavioural ranges with different groups of parameters and that avoid unnecessary nonlinearities can lead to a much better convergence of the numerical algorithms and to an easier understanding of the origin of uncertainty (Ratkowsky 1986; Bates and Watts 1988).

Comments to step 3: As discussed in the previous section, from a methodical point of view it would always be possible to do combined Bayesian inference according to case 3Ba. However, for accurately identifiable model parameters, this is not necessary. In this situation, it is computationally much less demanding and leads to similar results, if one proceeds according to the classical identification scheme described in case 3Aa and does discrete Bayesian model averaging in case of bad “selectivity" of the model as described in case 3Ab. If there are models with poorly identifiable parameters and enough computational resources are available, Bayesian inference according to case 3Ba is the best technique. If this cannot be done due to computational limitations, the case 3Bb offers a pragmatic way of averaging between different (discrete) scenarios. Also this method gives more reasonable prediction uncertainty bounds than doing the analysis with a simplified
model that omits a non-identifiable but important process.

In order to clarify the application of the identification and forecasting procedure outlined above, it is used to analyze several synthetical data sets of a simple didactical example of ecological relevance (microbial kinetics) that covers all cases of the proposed procedure in a succeeding paper (Omlin and Reichert 1999).

1.5 Conclusions

Neglecting the uncertainty in model structure in classical system identification leads to an underestimation of the uncertainty in model predictions. The two most problematic steps of this procedure are:

- The use of a single uniquely identified model as the "true" model for predictions and for the estimation of prediction uncertainty ignores the uncertainty in model structure. This is a particular problem in ecological modelling where a single true model does not exist, because all models are crude simplifications of reality, and different descriptions of reality may be useful for different objectives.

- The exclusion of overparameterized models that include poorly identifiable processes by the parsimony principle is meaningful for system identification during the measurement period. However, if the identified model is used for prediction or extrapolation into domains where such processes may become important, their exclusion can lead to wrong results. If information on such processes is available, the parsimony principle leads to a lack of consideration of this prior knowledge.

These two problems of the classical identification procedure can lead to a significant underestimation of the uncertainty in model forecasts.

The Bayesian approach to identification and prediction can be used to eliminate the problems mentioned above:

- Instead of uniquely selecting one model, predictions can be based on averaging the forecasts of several competing models.

- Due to the calculation of probability distributions of parameters instead of point estimates, identifiability is not a necessary condition for the use of Bayesian techniques. Therefore, also overparameterized models can be used when there is strong evidence that these can give more reliable prediction uncertainty estimates.

- The use of prior knowledge in the form of prior parameter distributions is standard in Bayesian inference.

The disadvantages of Bayesian inference are the high computational requirements to numerically calculate the probability distributions, the necessity of formulating prior knowledge in the form of exact probability distributions, and the difficulty in the interpretation of the results that represent degrees of knowledge or beliefs rather than classical probabilities that can be approximated by frequency distributions. In spite of these problems in interpretation, the Bayesian technique is a mathematical formalism that can be successfully applied to obtain more reasonable (in most cases larger) uncertainty estimates compared to the classical technique. There exist other approaches (mentioned above) to
achieve the same goal. It is a common feature of all these approaches that they require either a similar extension of the notion of probability or the introduction of a new notion of a belief or something similar in order to be able to describe the contribution of lack of knowledge to total uncertainty.

The identification procedure recommended in this paper tries to profit from the strengths of the Bayesian approach while keeping the computational requirements to a minimum. It is based on the idea that in accurately identifiable situations the classical procedure is adequate, and that in many cases where Bayesian techniques should be applied, a rough, discrete approximation to probability distributions of non-identifiable model parameters leads to results close to a complete, continuous analysis.

It is very important to improve uncertainty analysis of ecological model forecasts as discussed in this paper. However, it should be noted that also such more realistic uncertainty estimates can never be rigorous bounds because a process not included in all competing model structures considered or even unknown at the time of modelling can become important in the future.

Acknowledgments

We would like to thank Rick Devlin, Manuel Gloor, Dieter Imboden, Roland Scholz, Werner Simon and Oskar Wanner for their helpful comments to a preliminary version of this paper.
Chapter 2

A Comparison of Techniques for the Estimation of Model Prediction Uncertainty

The basic concepts of frequentist and Bayesian techniques for the identification of model parameters and the estimation of model prediction uncertainty are briefly reviewed. A simple example with synthetically generated data sets of a model for microbial substrate conversion is used as a didactical tool for analyzing strengths and weaknesses of both techniques in the context of environmental system identification. The comparison results in the practical superiority of the frequentist technique in the case of identifiable model parameters (computational efficiency). However, in the case of poor parameter identifiability, the conceptual advantage of the estimation of parameter distributions and the use of prior knowledge make the Bayesian approach more recommendable. Because in environmental system identification prior knowledge often makes the use of overparameterized models necessary, Bayesian techniques are very important in this field and should more often be used.

Keywords: Bayes; frequentist; identifiability; overparameterized models; parsimonious models; uncertainty.

2.1 Introduction

There are two good reasons why modeling techniques should be used in environmental sciences: First, modeling is an important tool to gain understanding of how environmental systems work. Second, models can be used to predict the behaviour of natural systems.

Each of these two motivations puts emphasis on a different part of the modeling procedure. If the goal of modeling is to gain scientific knowledge, the task is to find the simplest model structure that is able to describe the measured data with reasonable accuracy and to find good estimates of the unknown model parameters (a more complicated model which does not significantly increase the quality of fit does not lead to greater knowledge (Spriet 1985)). On the other hand, if the motivation is prediction, what counts most is a good predictive capability and reasonable estimates of the uncertainty of the model predictions. The model structure and their parameters are in this case only an intermediate step and are not interesting in themselves (Hampel 1993). Extrapolations with a too simple model
structure, however, may lead to overoptimistic uncertainty estimates (Beck 1987; Reichert and Omlin 1997).

In order to identify the parameters of a given model, two concepts have to be distinguished: Point parameter estimation and regional parameter estimation. In the first case, a single value is chosen for each model parameter and the uncertainty of this value is estimated from the local properties of the deviations of the model results from the data at this point in parameter space. This can only be done if this point can be determined uniquely. Therefore, identifiability of model parameters is a necessary condition for this classical analysis (Spriet 1985). For this reason, if model parameters are not uniquely identifiable, the model structure is further simplified until identifiability is achieved. Regional estimation techniques, on the other hand, make estimates of parameter distributions instead of values and, therefore, do not require the selection of unique parameter values (Hornberger and Spear 1981; Hornberger and Spear 1983; Rose et al. 1991; Bernardo and Smith 1994; Gelman et al. 1995). In the case of good identifiable parameters, a narrow probability distribution is obtained, whereas the existence of one or more non-identifiable parameters leads to a wide distribution of these and correlated parameters if more precise prior knowledge is not available.

In ecological models, poor identifiability of parameters is typical. This is because natural systems are complex and prior knowledge of their mechanism requires complex model structures. The parameters of these systems are often not identifiable because the data used for parameter estimation is often sparse in relation to the model complexity. In these situations regional estimation techniques are more advantageous than classical point estimation techniques. Hornberger and Spear presented such a regional method and applied it to a eutrophication problem in a coastal estuary (Hornberger and Spear 1981; Hornberger and Spear 1983; Rose et al. 1991). However, Bayesian techniques (Bernardo and Smith 1994; Gelman et al. 1995) may be even better, because they are based on probability theory and they explicitly consider probability distributions for prior knowledge and for the measurement process (in the likelihood function) (Reichert and Omlin 1997).

In this paper, a simple, didactical example of microbial substrate conversion with Monod kinetics is used to compare the classical and Bayesian methods in both situations with identifiable and non-identifiable parameters. The influence of good and poor prior knowledge of the values of model parameters on the different strategies is also discussed. This example is very simple; for this reason, the results seem to be quite trivial. Note, however, that the simplicity of the example makes it possible to see the differences clearer and that the example nevertheless is of ecological relevance because this model is included as a submodel in many ecological models.
2.2 Identification and Prediction Procedures

In this section brief summaries of the classical and the Bayesian parameter estimation and model prediction techniques are given. Before the techniques are presented in the following subsections, a description of the conceptual differences between the classical and the Bayesian interpretation of probabilities and confidence or credibility intervals is given.

In classical (frequentist) statistics probability distributions are used to describe the stochastic behaviour of a system or at least of the measurement process. Probability distributions represent frequency distributions of measurements from reproducible random experiments in the limit of an infinite number of experiments. Model parameters are assumed to have true, fixed values. If these values are unknown, they can be estimated from data of model variables with the aid of statistical estimators. Although the model parameters are themselves not random variables, the estimators are, because they depend on measurements of the stochastic system (at least the measurement process is assumed to be stochastic). Similarly, the limits of confidence intervals are random variables. The interpretation of probabilities concerning the coverage of the true value by a confidence interval is the frequency distribution in the limit of many determinations of the confidence interval. No statement is possible on the probability of the true value being in a confidence interval determined from a single data set; this is because the parameter is not a random variable. This problem has been criticized as a lack of a behavioural interpretation of classical confidence intervals (Walley 1991).

In contrast to this frequentist interpretation, in Bayesian statistics probability distributions are used to describe the state of knowledge on the value of a model parameter (model parameters themselves are here random variables). Consequently, such probability distributions are not objective, but they change if additional data increases the state of knowledge. This broader definition of probability extends the applicability of probability concepts at the cost of a less precise definition of the term. A Bayesian credibility interval with a given probability content is interpreted as containing the model parameter with this probability to the present state of knowledge. This is a much more behaviourally usable interpretation than the meaning of a classical confidence interval. However, there is no frequentist test to assess the reliability of this probability statement.

Despite these large conceptual differences in interpretation, the practical use of classical confidence intervals and Bayesian credibility intervals as quantitative measures of parameter or model prediction uncertainty is very similar (in many contexts also the intervals are similar or even identical, see Bates and Watts (1988)). For this reason, a comparison of classical confidence intervals with Bayesian credibility intervals is of interest. In this paper, we focus on the practical aspects of such a comparison. For a discussion of the conceptual aspects in the context of ecology we refer to Ellison (1996) and Dennis (1996).

2.2.1 Classical Frequentist Analysis

There is a huge literature on classical system identification and parameter estimation. Some recent reviews on (mainly linear) system identification techniques are Ljung (1987) and Söderström and Stoica (1989). A description of parameter estimation techniques for nonlinear systems can be found in Bates and Watts (1988) and Seber and Wild (1989).

For the classical analysis, $n$ calculated values of the specified variables of a model $M$
at given locations in space and time can be represented by \( n \) random variables

\[
Y_M(p_M) = \{Y_{M,1}(p_M), \ldots, Y_{M,n}(p_M)\}
\]  

(2.1)

which depend on the values of \( m \) parameters \( p_M = (p_{M,1}, \ldots, p_{M,m}) \). These random variables can be described by the probability density function

\[
f_M(y, p_M)
\]

(2.2)

for the values \( y = (y_1, \ldots, y_n) \) of the variables. The probability density function \( f_M \) either describes a stochastic model or a deterministic model with a stochastic part due to the measurement process. In the latter case the probability density function describes the stochastic measurement centered around the values

\[
y = y_M(p_M)
\]

(2.3)

of the deterministic model described by the functions \( y_M \).

Parameter Estimation

If the values of the model parameters are unknown, they can be estimated with a statistical procedure from measurements of the model variables \( y_{\text{meas}} = (y_{\text{meas},1}, \ldots, y_{\text{meas},n}) \). Thereby the data \( y_{\text{meas}} \) are interpreted as realizations of the random variables \( Y_M \). The estimator \( \hat{p}_M \) is a function of the random variables \( Y_M \) and is therefore also a random variable in contrast to the model parameters \( p_M \) themselves. This procedure does not make use of prior knowledge on parameter values; the estimates depend only on the actual measurements.

**Maximum Likelihood Estimation.** The maximum likelihood principle is a general method to find such an estimator. The probability density function (2.2) gives the probability density for measuring the values \( y \) given the parameters \( p_M \). Viewed as a function of the parameters for given measurements \( y = y_{\text{meas}} \), it is called the likelihood function

\[
L_M(y_{\text{meas}}, p_M) = f_M(y_{\text{meas}}, p_M)
\]

(2.4)

The idea of the maximum likelihood principle is to select those parameter values \( \hat{p}_M \), which maximize the likelihood function (2.4) for the given measured data \( y_{\text{meas}} \).

**Least Squares Method.** The method of least squares is the most important special case of the maximum likelihood principle for parameter estimation of a given deterministic model \( y_M \). If the measurement process of the system variables at the given locations in space and time can be described by independent, normally distributed random variables with mean equal to the deterministic model results and known standard deviations \( \sigma_{\text{meas},i} \), maximizing the likelihood function is equivalent to minimizing the function

\[
\chi^2(y_{\text{meas}}, p_M) = \sum_{i=1}^{n} \frac{(y_{\text{meas},i} - y_{M,i}(p_M))^2}{\sigma_{\text{meas},i}^2}
\]

(2.5)

The value of this function is calculated as the weighted sum of the squares of the deviations between measurements and deterministic model results. The weight of each term in the
sum is given as the inverse variance of the corresponding measurement. The least squares estimates of the parameters are those parameter values \( \hat{p}_M \) for which this function takes its minimum.

The linear approximation of the estimates of the covariance matrix of the parameter estimators \( \text{Cov}(\hat{p}_M) \) is given by

\[
\text{Cov}(\hat{p}_M) \approx \left( \frac{\chi^2}{(n - m)} \left( \frac{\partial y_M}{\partial p_M} \right)^T \begin{pmatrix} \frac{1}{\sigma^{\text{meas},1}_M} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma^{\text{meas},2}_M} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma^{\text{meas},n}_M} \end{pmatrix} \frac{\partial y_M}{\partial p_M} \right)^{-1},
\]

(2.6)

where all functions are evaluated at the parameter estimates \( p_M = \hat{p}_M \).

**Uncertainty of Model Predictions**

The covariance matrix of the model parameter estimators can be used to calculate standard errors and confidence intervals of predicted values.

**Confidence Intervals.** As a first step, the covariance matrix of the expected model results \( \hat{Y}_M = (\hat{Y}_{M,1}, \ldots, \hat{Y}_{M,n}) \) is calculated by linear error propagation

\[
\text{Cov}(\hat{Y}_M) \approx \left( \frac{\partial y_M}{\partial p_M} \right) \cdot \text{Cov}(\hat{p}_M) \cdot \left( \frac{\partial y_M}{\partial p_M} \right)^T
\]

(2.7)

(again the derivatives \( \frac{\partial y_M}{\partial p_M} \) must be evaluated at \( p_M = \hat{p}_M \)). Then, the linear approximation to a \( 1 - \alpha \) confidence interval for the expected result for the \( i \)-th variable is given as

\[
\hat{y}_{M,i} \pm \sqrt{\text{Cov}(\hat{Y}_M)_{i,i}} \cdot t(n - m; \alpha/2)
\]

(2.8)

where \( t(n - m; \alpha/2) \) is the upper \( \alpha/2 \) quantile of the \( t \)-distribution with \( n - m \) degrees of freedom and \( \hat{y}_{M,i} = y_{M,i}(\hat{p}_M) \) is the model estimate for the variable \( i \) (Seber and Wild 1989; Bates and Watts 1988).

**2.2.2 Bayesian Analysis**

There are many excellent general books on Bayesian statistics. Some newer books are Bernardo and Smith (1994) and Gelman et al. (1995). There are also some reviews, discussions and applications of Bayesian statistics to environmental systems. Some of these are Tiwari et al. (1978), Marin et al. (1989), Freeze et al. (1990), Dilks et al. (1992), Gorelick et al. (1993), Varis et al. (1994), Aldenberg et al. (1995), Ellison (1996), Dennis (1996), Wolfsen et al. (1996) and Reichert and Omlin (1997).

The representation of models for Bayesian analysis is similar to that for classical analysis given by the equations (2.1) to (2.3). A minor difference is the interpretation of equation (2.2). Because model parameters are represented by random variables \( p_M = (p_{M,1}, \ldots, p_{M,m}) \) in Bayesian statistics, the probability density (2.2) for the \( n \) values \( y = (y_1, \ldots, y_n) \) for given values \( p_M = (p_{M,1}, \ldots, p_{M,m}) \) of the \( m \) parameters, changes to the conditional probability density function

\[
f_M(y|p_M) = f_M(y|p_M = p_M)
\]

(2.9)

for the values \( y \).
Parameter Identification

Bayesian parameter identification combines prior knowledge of model parameters with measured data to obtain an updated posterior distribution of the model parameters. This updating is done using Bayes’ rule

\[ f_{\text{post}}(P_M | y_{\text{meas}}) = \frac{f_M(y_{\text{meas}} | P_M) f_{\text{pri}}(P_M)}{f(y_{\text{meas}})} \]  

where \( f_{\text{pri}}(P_M) \) is the prior distribution of the model parameters, \( y_{\text{meas}} \) are the measured data, \( f_{\text{post}}(P_M | y_{\text{meas}}) \) is the posterior distribution of the model parameters conditioned with the measured data and \( f(y_{\text{meas}}) \) is the probability density of measured data. The conditional probability density \( f_M \) is evaluated at the measured values for the model response, this makes it similar to the likelihood function of the model (cf. section on classical analysis). Note, that Bayes’ rule only makes it possible to update parameter distributions. It does not allow one to derive distributions without prior knowledge. In many practical applications of Bayesian theory, in the absence of prior knowledge, a uniform prior distribution of the parameters is assumed. However, this procedure is not satisfying from a conceptual point of view because it depends on the parameterization of the model (Walley 1991; Bernardo and Smith 1991). From a pragmatic point of view, in the case of vague prior knowledge, it is recommended that one checks the sensitivity of the results to the selected prior distribution (Berger 1984).

The probability density distribution of measured data \( f(y_{\text{meas}}) \) in equation (2.10) is independent of the model parameters and can therefore be determined by normalization. Equation (2.10) then can be written in the form

\[ f_{\text{post}}(P_M | y_{\text{meas}}) = c_B f_M(y_{\text{meas}} | P_M) f_{\text{pri}}(P_M) \]  

where the factor \( c_B \) is calculated by normalizing the posterior distribution \( f_{\text{post}}(P_M | y_{\text{meas}}) \).

Model Prediction

The calculated posterior distribution of the model parameters makes it possible to determine the probability distribution of model predictions. The distribution of model results without measurement error for a deterministic model given by equation (2.3) is given by

\[ f(y|M, y_{\text{meas}}) = \frac{\partial^k}{\partial y_1 \cdots \partial y_k} \int_{y_M(P_M) \leq y} f_{\text{post}}(P_M | y_{\text{meas}}) dP_M \]  

This equation is not very useful for numerical calculations. Instead the distribution of model results is calculated with Monte Carlo Simulation techniques (Rubinstein 1981; Sobol 1994). With this method, parameter sets are drawn randomly from the posterior distribution and the corresponding model predictions are evaluated in a frequency distribution. In order to accelerate the convergence of this frequency distribution to the exact probability distribution (2.12), latin hypercube sampling techniques can be applied (McKay et al. 1979; Sobol 1979; Stein 1987).

For a single variable, the \( 1/2 \) and \( 1 - 1/2 \) quantiles of the posterior distribution are usually used as the limits of a \( 1 - \alpha \) credibility interval. In higher dimensions, the hypersurface of constant probability density with a probability content of \( 1 - \alpha \) is used to bound the credibility region. This credibility region is also called highest probability density region (Bates and Watts 1988).
2.3 A Simple Example of Ecological Relevance

In this example the results of classical (frequentist) and Bayesian identification and prediction techniques are compared for a simple model of microbial substrate conversion under different assumptions on prior knowledge and data availability. The use of synthetically generated data sets makes it possible to discuss the results of the techniques in the absence of model structure uncertainty and to check the dependence of inference results on data realization.

2.3.1 Model Equations

In many ecological and biotechnological systems substrate conversion rates by microbial organisms are linear for small substrate concentrations, but reach a state of saturation for large concentrations. One way to describe such a behaviour is given by the rate expression

\[ r = \frac{r_{\text{max}} \cdot C}{K + C}, \]  

(2.13)

where \( r \) is the actual substrate conversion rate, \( r_{\text{max}} \) is the maximum conversion rate, \( C \) is the substrate concentration, and \( K \) is the concentration at which the substrate conversion rate is half of its maximum value. This equation was originally derived by Michaelis and Menten (1913) to describe enzyme kinetics and it was later used by Monod (1942) to describe bacterial growth. It is now widely used as an empirical description of the transition of microbial growth and substrate conversion rates between the linear and the saturation range in ecological as well as biotechnological systems (Yang and Sykes 1998; Asaeda and Van Bon 1997; Orlob 1983; Henze et al. 1986; Henze et al. 1995; Güjer et al. 1995, are some examples).

For small concentrations the conversion rate given by equation (2.13) increases approximately linearly with the concentration

\[ r \approx \frac{r_{\text{max}}}{K} \cdot C \quad \text{for } C \ll K, \]  

(2.14)

whereas for large concentrations the conversion rate becomes constant

\[ r \approx r_{\text{max}} \quad \text{for } C \gg K. \]  

(2.15)

The convergence to this value is rather slow. If the concentration is 4 times \( K \) then the rate is 80 % of \( r_{\text{max}} \). An increase of the concentration to 19 \( K \) leads to a rate of 95 % of its maximum. Although the existence of a linear and a saturation range are compatible with many observations, the transition between these two ranges is often observed to be faster than in the Monod description. Other models exist, which differ mainly in the concentration dependence of the conversion rate in this transition range (Dabes et al. 1973; Bader 1982). Nevertheless, the Monod model is still most often used in practice (maybe for historical reasons) and therefore of practical relevance.

2.3.2 Data Sets and Prior Distributions

The synthetic data sets were produced by randomly sampling normal distributions centered at the exact values of the Monod kinetic model given by equation (2.13). The true model parameters are selected to be \( K = 1 \) and \( r_{\text{max}} = 1 \). The standard deviation of the
<table>
<thead>
<tr>
<th>Data Points</th>
<th>Concentration Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.0 - 0.5</td>
</tr>
<tr>
<td>10</td>
<td>0.5 - 3.0</td>
</tr>
<tr>
<td>10</td>
<td>3.0 - 10.0</td>
</tr>
</tbody>
</table>

Table 2.1: Number of synthetic measurements in different concentration ranges.

Normal sampling process is assumed to be 0.05 and simulates the measurement error of the data. Each data set consists of 30 data points and covers the concentration range as shown in Table 2.1.

For tests of the dependence on data sets, 10 sets were generated and parameter identification was performed for the complete data sets and for incomplete data sets covering the linear range only (first 10 data points). In order to demonstrate the dependence on prior information, all Bayesian analyses were performed for narrow and wide prior distributions for both parameters. These prior distributions were assumed to be lognormal distributions with a mean of 0.97 and a standard deviation of 0.05 for the narrow prior and a mean of 2.1 and standard deviation of 1.7 for the wide prior (see Fig. 2.1). The same prior distributions were applied independently to both parameters.

![Figure 2.1: Narrow and wide prior distributions used for Bayesian analysis.](image)

2.3.3 Simulation Methods and Numerical Procedures

Parameter identifications, predictions and calculations of confidence and credibility bands were all performed with the MATLAB numeric computation and visualization software (The MathWorks, Inc., Natick, Mass., USA).

For the classical (frequentist) analyses the nonlinear regression algorithm by Levenberg-Marquardt (Press et al. 1992) was used for the estimation of values and covariance matrix of the model parameters. The implementation of this algorithm is not part of the standard MATLAB library. An implementation of the algorithm by R. Schrager, A. Jutan and R. Muzic, which is freely available at ftp.mathworks.com, was used instead. Confidence bands were then calculated by applying the equations (2.7) and (2.8).
The algorithms for the application of the Bayesian analyses were implemented by the authors. The posterior distribution was approximated by evaluating equation (2.11) on a regular grid in parameter space. With the aid of contour plots of the posterior distribution, the parameter ranges were adapted to cover the region where the posterior distribution was significantly different from zero. Sets of model predictions were then generated for parameter values sampled from the posterior distribution. In order to accelerate convergence of the frequency distribution of the set of model predictions to the exact distribution, the Latin Hypercube sampling technique (Iman and Conover 1980) was applied. The median and the credibility interval of the predictions were then approximated by the median and the credibility interval of the frequency distribution of the set of predictions.

2.3.4 Identification and Prediction with Complete Data Sets

In this section the results of the classical (frequentist) and the Bayesian identification and prediction techniques are compared for complete data sets covering the linear as well as the saturation range of the rate function. This means that both model parameters should be identifiable from the data.

Classical Technique

Applied to ten different data sets each covering the whole concentration range, the classical least squares method shows that both parameters are accurately identifiable (relative standard errors of about 3 % for \( r_{\text{max}} \) and 12 % for \( K \)). The least squares estimates of the parameters for the different data sets were not significantly different from each other (not shown). This demonstrates the (approximate) independence of the results from the data realization. The parameter estimates for one data set are shown in Table 2.2, the prediction based on these estimates is shown in Fig. 2.2.

<table>
<thead>
<tr>
<th>parameter</th>
<th>estimate</th>
<th>se</th>
<th>corr</th>
</tr>
</thead>
<tbody>
<tr>
<td>( r_{\text{max}} )</td>
<td>0.98</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>( K )</td>
<td>0.97</td>
<td>0.12</td>
<td>0.82</td>
</tr>
</tbody>
</table>

Table 2.2: Least squares parameter estimates, standard errors and correlation coefficient for equation (2.13); complete data set.

Bayesian Technique

Figure 2.3 shows random samples together with 50 % and 95 % credibility regions of the posterior distributions for both prior distributions introduced in section 2.3.2. Although the parameter ranges are much stronger limited by the narrow priors, the difference in the posterior distributions are not very large. This shows that the parameter estimates depend primarily on the data and not on the prior. There is also no significant dependence on the data set (not shown). Figure 2.4 shows model predictions for the two posterior distributions shown in Fig. 2.3. Again, the credibility bands are only slightly narrower for the narrow prior.
Figure 2.2: Classical model prediction based on the parameter estimates given in Table 2.2; complete data set.

Figure 2.3: Posterior distribution of parameters (random sample with 50 % and 95 % credibility regions); complete data set; narrow prior (left) and wide prior (right).

Comparison
A comparison of Fig. 2.2 with Fig. 2.4 yields only insignificant differences between the results of the classical and the Bayesian prediction techniques and between the narrow and the wide prior distributions in this identifiable situation. This result can be expected to be valid also for other identifiable situations unless the prior distribution of the parameters are not extremely narrow. Given the much higher computational expense required for the Bayesian technique, from a practical point of view, the classical technique is preferable in such identifiable situations.

2.3.5 Identification and Prediction with Incomplete Data Sets
In this section the results of the classical and the Bayesian identification and prediction techniques are compared for incomplete data sets covering the linear range of the rate
Figure 2.4: Bayesian model prediction based on the posterior distributions shown in Fig. 2.3.4; complete data set; narrow prior (left) and wide prior (right).

function only. This means that the saturation rate \( r_{\text{max}} \) cannot be identified from the data. This uncertainty in \( r_{\text{max}} \) also makes the parameter \( K \) non-identifiable because the slope in the linear range is given by \( r_{\text{max}}/K \) (see equation 2.14). Whereas non-identifiability is not a conceptual problem for the Bayesian technique, the classical technique should not directly be applied to this situation. Instead, a simplified, identifiable model should be used. In order to demonstrate possible applications of the classical technique, the following alternatives are discussed: Holding \( r_{\text{max}} \) at a given, fixed value and estimating only \( K \) (a model with only one unknown parameter, \( K \), is an identifiable model), or estimating the parameter of a simplified, identifiable (linear) model. In addition, the problems of applying the complete, non-identifiable model are also discussed.

**Classical Technique**

**Good prior knowledge of \( r_{\text{max}} \).** If good prior knowledge on the value of the parameter \( r_{\text{max}} \) is available, this parameter can be held at a fixed value while the other parameter \( K \) is estimated. This means that the uncertainty in \( r_{\text{max}} \) is neglected for parameter identification. Note, that holding \( r_{\text{max}} \) at a fixed value solves the identifiability problem for \( K \) mentioned above. Holding \( r_{\text{max}} \) at the value of 0.97 (mean of the narrow prior distribution used for the Bayesian analysis) leads to the estimate shown in Table 2.3.

<table>
<thead>
<tr>
<th>parameter</th>
<th>estimate</th>
<th>se</th>
</tr>
</thead>
<tbody>
<tr>
<td>( K )</td>
<td>0.98</td>
<td>0.14</td>
</tr>
</tbody>
</table>

Table 2.3: Least squares parameter estimate and standard error for the parameter \( K \) of equation (2.13) based on a given value of \( r_{\text{max}} = 0.97 \); incomplete data set.

The relatively small value of the standard error confirms the identifiability of \( K \). Figure 2.5 (left) shows the corresponding model prediction.

The confidence band is rather narrow in the saturation range because in this procedure the uncertainty of the parameter \( r_{\text{max}} \) is neglected. Using the uncertainty of this parameter
Figure 2.5: Classical model prediction based on the posterior distributions shown in Fig. 2.3.4; partial data set only; $r_{max}$ fixed for identification and prediction (left) and $r_{max}$ fixed for identification but parameter uncertainty used for prediction (right).

for prediction (assuming the standard deviation of the narrow prior used in the Bayesian analysis) enlarges the confidence band (Figure 2.5, right). Note, however, that ignoring the uncertainty of $r_{max}$ for identification but using it for prediction is not a consistent technique.

Figure 2.6: Bayesian model prediction based on the posterior distribution shown in Fig. 2.9; partial data set only; narrow prior.

**Poor prior knowledge of $r_{max}$.** If only poor prior knowledge of the non-identifiable parameter is available, non-identifiability cannot be eliminated by holding $r_{max}$ at its known value. The classical solution to this problem is to use a simplified, identifiable model (Spriet 1985). It is evident that a linear model

$$r = k \cdot C$$

(2.16)

is adequate to describe the data. The least squares estimate for the parameter $k$ of this model is given in Table 2.4.
Table 2.1: Least squares parameter estimate for equation (2.16); incomplete data set.

<table>
<thead>
<tr>
<th>parameter</th>
<th>estimate</th>
<th>se</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k$</td>
<td>0.71</td>
<td>0.089</td>
</tr>
</tbody>
</table>

Figure 2.7 shows the prediction with this model. It is evident that the prediction within the data range leads to a reasonable data interpolation, whereas the extrapolation is wrong. This shows the danger of performing extrapolations with simple, identifiable models: The error propagation algorithm assumes the model structure to be correct, what cannot be tested in an extrapolation domain (and in our example is not true). Note that a classical statistician will not extrapolate a linear model as shown in Fig. 2.7, right, if he or she knows that the saturation regime exists. However, if model parameters are determined and published by one research group and the model is subsequently used by others, the information on the validity ranges is often lost, so that in practice, the danger for extrapolation errors as shown in Fig. 2.7, right, is large.

![Graph showing classical model prediction](image)

Figure 2.7: Classical model prediction with the simplified model (2.16) based on the parameter estimate given in Table 2.4; incomplete data set.

To demonstrate the consequence of using least squares estimation when model parameters are not identifiable, an analysis was performed for all 10 data sets under these conditions. Table 2.5 shows the parameter estimates and Fig. 2.8 shows the model prediction for the two data sets which led to the most extreme predictions. It becomes evident, that the estimates and model predictions are sensitive to the data set used. This is the case because in a non-identifiable situation the position of the poorly defined minimum of the sum of squares is very sensitive to statistical fluctuations of data points. Note, that such a dependence may be difficult to detect because in practice usually only a single data set is available (for the data set a, the problem is not detected by the algorithm; for the data set b, the large standard errors of the parameter estimates and the large value of the correlation coefficient show the identifiability problem clearly).
Table 2.5: Least squares parameter estimates, standard errors and correlation coefficients for equation (2.13); incomplete data sets; model not identifiable.

<table>
<thead>
<tr>
<th>data set</th>
<th>parameter</th>
<th>estimate</th>
<th>se</th>
<th>corr</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>$r_{\max}$</td>
<td>0.4</td>
<td>0.14</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td>$K$</td>
<td>0.21</td>
<td>0.18</td>
<td></td>
</tr>
<tr>
<td>b</td>
<td>$r_{\max}$</td>
<td>$2.14 \cdot 10^4$</td>
<td>$5.3 \cdot 10^8$</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>$K$</td>
<td>$3.14 \cdot 10^4$</td>
<td>$7.7 \cdot 10^8$</td>
<td></td>
</tr>
</tbody>
</table>

Figure 2.8: Classical model prediction based on the parameter estimates given in Table 2.5 for data set a (left) and data set b (right); incomplete data set; model not identifiable.

Bayesian Technique

**Good prior knowledge of $r_{\max}$.** Good prior knowledge of $r_{\max}$ results in a narrow prior distribution of this parameter. The random sample with credibility regions of the posterior parameter distribution shown in Fig. 2.9 shows a similar spread as the estimated distributions calculated under the condition of complete data set availability (Fig. 2.3). The lack of data in the saturation range is compensated by the information provided by the narrow prior. The uncertainty of the parameter $r_{\max}$ in the posterior distribution is determined by its prior distribution. The correlation between the parameters is very small because of the dominating influence of the independent prior distributions. The predictions and their uncertainties are shown in Figure 2.6.

**Poor prior knowledge of $r_{\max}$.** Figure 2.10 shows the posterior distribution for the wide prior distribution. The non-identifiability of the model parameters is reflected by a wide posterior distribution, which in this case indicates a large correlation between the two parameters. The model prediction (Figure 2.11) shows a wide credibility band in the extrapolation domain reflecting poor prior knowledge and non-identifiability of $r_{\max}$.

Comparison

If good prior knowledge of $r_{\max}$ is available, the classical analysis can be performed with a fixed value of this parameter. If the uncertainty of $r_{\max}$ is considered for model prediction,
Figure 2.9: Posterior distribution of parameters (random sample with 50 % and 95 % credibility regions); partial data set only; narrow prior.

Figure 2.10: Posterior distribution of parameters (random sample with 50 % and 95 % credibility regions); incomplete data set; wide prior (note the difference in scale in comparison to the Figs. 2.3 and 2.9).

nearly the same result is achieved as for the Bayesian analysis with the narrow prior (compare Figure 2.5, right with Figure 2.6, right). However, because the Bayesian analysis consequently uses the uncertainty in \( r_{\text{max}} \), whereas the classical analysis neglects it in the identification step, the Bayesian technique is much more convincing.

If only poor prior knowledge of \( r_{\text{max}} \) is available, no convincing classical technique is available. The use of the simplified, linear model is adequate for data interpolation, however, it underestimates prediction uncertainty in the extrapolation domain. The use of the complete, non-identifiable model leads to a sensitive dependence of estimates and predictions on the data set. In contrast, the Bayesian approach leads to similar results as the simplified model in the interpolation domain, but it also shows the uncertainty present in the extrapolation domain (compare Figs. 2.7 and 2.11).
Figure 2.11: Bayesian model prediction based on the posterior distribution shown in Fig. 2.10; incomplete data set; wide prior.

2.4 Summary and Conclusions

A comparison of classical (frequentist) with Bayesian identification and prediction techniques applied to a simple model of microbial substrate conversion showed similar results if the data allowed a precise identification of the model parameters. If the much higher computational expense required for the Bayesian approach is considered, from a practical point of view, the classical technique is preferable for such situations. However, if there are non-identifiable model parameters, it is difficult to apply the classical technique. An application that ignores the non-identifiability problem leads to model results that depend sensitively on random data fluctuations. Application of the classical technique only to the identifiable model parameters while assuming values of the non-identifiable parameters leads to reasonable results if these values are accurately known. However, this technique cannot be applied if only poor prior knowledge on the non-identifiable parameters is available. The traditional solution to this problem is to simplify the model equations until all parameters are identifiable. It is shown that this technique leads to a reasonable data interpolation, but to a drastic underestimation of the uncertainty in the extrapolation domain, (this is because in the example used for demonstration, the process neglected in the simplified model becomes relevant in the extrapolation domain). All these problems are more readily addressed with Bayesian techniques. Because parameter distributions and not point values are estimated, non-identifiability is not a problem for this technique. The explicit use of prior knowledge leads to narrow error bands in the extrapolation domain if good prior knowledge is available, and to wide error bands otherwise. The result is a more realistic uncertainty estimate in the extrapolation domain. In the interpolation domain, the results of the Bayesian technique are similar to those of the classical technique applied to a simplified model. These considerations make the Bayesian approach a better choice for model predictions when there are poorly identifiable model parameters.

The simple example discussed in this paper demonstrates the danger of underestimating the uncertainty of model predictions with parsimonious, identifiable models pointed out in a previous paper (Reichert and Omlin 1997). It shows that the Bayesian approach makes it possible to use an overparametrized model and thus to avoid this problem. These results are very obvious for the simple example discussed in this paper. However, it should
be noted that similar results can be expected in more difficult situations where more complex models and higher dimensions of parameter and response spaces make it much more difficult to detect the problem.

The problem of the underestimation of prediction errors with parsimonious, identifiable models is a general problem of forecasting. However, it is of special importance for modelling environmental systems for the following two reasons. (1) The complexity of environmental systems makes it necessary for the modeller to model in a very much simplified way a small subsystem of the investigated system. It is typical that in the prediction period aspects of the system become important that are not described by the model of the subsystem. This danger is especially large if extrapolations to changed external influences on the systems are made. (2) A huge amount of knowledge on environmental systems is available in the literature. For this reason, knowledge on processes that become important during the prediction period may be available, although it may not be contained in the data evaluated for model selection and parameter estimation. It is exactly this situation in which Bayesian techniques can be advantageously applied as discussed in this paper (if no knowledge is available, e.g. because processes not yet known become important, statistical techniques cannot solve the problem). For this reason, Bayesian techniques are very useful for the estimation of the uncertainty of model predictions for environmental systems.

Acknowledgments
We thank Dieter Imboden, Roland Brun, Peter Holzer and Tom Ravens for helpful comments.
Part II

Application
Chapter 3

Biogeochemical Model of Lake Zürich: Model Equations and Results

A mathematical model for plankton, nutrient (phosphate, ammonia and nitrate) and oxygen dynamics in lakes was developed. It is based on horizontally averaged changes in substance and organism concentrations due to vertical mixing, sedimentation, in- and outflows and biogeochemical conversion processes in the water column and in adjacent sediment layers. The biological part of the model was kept as simple as possible with the three plankton classes *Planktothrix (Oscillatoria) rubescens* (most abundant blue-green algae in the lake), other algae and zooplankton. Due to a strong phosphate limitation of algae growth in summer, the stoichiometry of primary production with respect to phosphorus had to be made variable in order to describe data sets available from Lake Zürich. In addition, a process of phosphate uptake of organic particles sinking through the hypolimnion had to be taken into account.

After adjusting some model parameters, the model was able to reproduce the key features of the nutrient and oxygen profiles and of algae-zooplankton interactions over several years. However, it was not able to predict occasionally occurring blooms of specific types of algae. This shows that the model is a good tool for studying the dynamics of the key processes governing plankton and nutrient dynamics in the lake, but that it is not detailed enough for the investigation of rarely occurring phenomena involving specific types of algae and for an accurate representation of zooplankton dynamics.

**Keywords:** Lake Zürich; water quality modelling; nutrients; plankton; variable stoichiometry; phosphate uptake.
3.1 Introduction

Water quality of lake Zürich is of importance for the local water supply authority (Wasser-
versorgung Zürich, WVZ) because the lake is an important source for drinking water. About 70 % of the drinking water in the region of Zürich originates from the lake. This water is used by almost 1'000'000 people.

In order to be able to early recognise problems with raw water quality from lake Zürich, WVZ conducts a lake water quality monitoring program.

In this paper, a model is described which is designed to quantitatively represent the processes leading to the dynamics of nutrients, oxygen and plankton in the lake. In the present stage, the main purpose of this model is understanding. However, it is planned to investigate its predictive capabilities in order to extend it to a water quality prediction tool for lake Zürich that supports the experts in their interpretation.

This paper is structured as follows. In section 3.2 a brief survey is given on lake Zürich and on the data used for simulation and model calibration. The model is described in detail in section 3.3 and the results are discussed in section 3.4. Finally, the key results with respect to model structure and results are summarized in section 3.5. Although the chosen engineering approach simplifies the system to a very high degree and does not include a detailed analysis of the food web, the model still contains a large number of uncertain model parameters. This makes a careful analysis of sensitivity, identifiability and prediction uncertainty mandatory. However, such an analysis goes beyond the scope of this paper and is delegated to a separate paper (Omlin et al. 2000).

3.2 Lake Zürich

3.2.1 Lake Description

Located at the border between the midland and the prealpine region of Switzerland Lake Zürich is enclosed by two ridges which protect the lake from direct exposure to the dominant west winds. The basin leads from the upper end at Rapperswil along a curved line in north west direction to Zürich 30 km downward. Its width varies between 2.5 km and 0.5 km (Figure 3.1). At Rapperswil Lake Zürich is separated by a dam from lake Obersee the outflow of which is the main inflow to Lake Zürich with an average discharge of approx. 76 m³/s. The only outflow is River Limmat which leaves the lake in Zürich with an average discharge of approx. 90 m³/s.

Lake Zürich has a surface of 65 km², a volume of 3.3 km³ and an average residence time of approx. 420 days. The deepest location with 137 m depth is close to Thalwil, where one of the 6 measurement sites of the water quality monitoring program is located (Figure 3.1). This paper is based on measurements at this site.

3.2.2 Data: Sources and Description

Physical, chemical and biological parameters are monthly measured in the water quality monitoring program by WVZ since 1972 at 6 locations in the lake. From this large collection of data only data used in this work is described below. A detailed description of the whole data set is available in the form of a technical report (Gammeter et al. 1997).
Chemical and physical parameters: Water temperature and concentrations of oxygen, ammonium, nitrate and phosphate are measured at the site close to Thalwil at depths 0, 1, 2.5, 5, 7.5, 10, 12.5, 15, 20, 30, 40, 60, 80, 90, 100, 110, 120, 130 and 136 m. Light attenuation depths to 10 %, 1 % and 0.1 % of the surface intensity are also measured by WVZ. An approximation to the light intensity at the lake surface was obtained by using solar radiation data from a measurement site of the Swiss Meteorological Institute (SMA) located in Zürich. Values measured hourly were converted to monthly averages for use in the model. Averaging sunlight over daily periods implies that a simulation of the diurnal cycle is not in the scope of the model and that light intensities and growth rates used in the model are daily averages that are smaller than instantaneous values during the day. This is an important point to be noted when model parameters are compared with published values.

Biological parameters: Between the water surface and 40 m depth phytoplankton concentrations are measured at the same depth levels as the chemical parameters mentioned above. Below 40 m phytoplankton concentrations are measured at 80, 120 and 136 m depth. From the detailed counting of algal species, in the present work the concentrations of the blue-green algae Planktothrix (Oscillatoria) rubescens and the sum of the concentrations of all other algae are used. Zooplankton concentrations are determined in spatially integrated samples between 0-20 m ('epilimnion') and 20-136 m ('hypolimnion'). From this data only Daphnia concentrations are used. All plankton concentrations were given in units of wet mass (WM). The modelled dry mass (DM) concentrations were multiplied by a factor of 5 for comparisons with measured wet mass concentrations.

Inflow: Organic particles and the nutrients phosphate, ammonia and nitrate are considered in the inflows from Lake Obersee at Rapperswil, from the sewage treatment plants around the lake and from several small rivers discharging into the lake. In addition, atmo-
spheric deposition of phosphate, ammonia and nitrate is considered. Data from WVZ and
reports from the local water protection authority (Amt für Gewässerschutz des Kanton
Zürich, AGW) (AGW 1985) and by the Swiss Federal Institute for Environmental Science
and Technology (EAWAG) (EAWAG 1992) were used to quantify the inflows from Lake
Obersee and from small rivers discharging into the lake; data to quantify inflows from
sewage treatment plants were obtained from AGW (1992).

3.3 Model Description

The model used in this paper is not a detailed food-web model of the lake, but an engineer-
ing type model that aggregates algal and zooplankton species as far as possible. It com-
bines and extends earlier approaches to one-dimensional water quality modelling of lakes
(Imboden and Gächter 1978; Karagounis et al. 1993; Ulrich et al. 1995). The main
extensions to these approaches are the use of several sediment layers within which aer-
ocic and anoxic mineralization of organic compounds is explicitly considered and which
allow a rough resolution of concentration profiles in the sediment (in some of the above-
mentioned models one sediment layer is considered) and the introduction of a phosphate
uptake process on organic particles sinking through the hypolimnion (Gächter and Mares
1985; Hupfer et al. 1995). In addition, a variable stoichiometry of primary production
with respect to phosphorus is used, as it is done in some of the above-mentioned models.
The biogeochemical processes are presented in a form that is common practice in the de-
scription of models for waste water treatment (Renz et al. 1986; Gujer et al. 1995). The
process formulations were highly influenced by recent developments in river water quality
modeling (Shanahan et al. 2000; Reichert et al. 2000; Vanrolleghem et al. 2000); the
values of parameters of the plankton submodel by Andersen (1997) and Reynolds (1997).

3.3.1 Overview

The model is based on horizontally averaged concentrations. This one-dimensional ap-
proach is justified if the typical time scale for transformation processes is much larger
than the time needed for horizontal mixing. Dissolved and suspended substances are
transported in the water column by vertical mixing. In addition, algae and suspended
particles are transported by sedimentation. Two sediment layers are used to approxi-
mately describe substances and processes in the sediment. Suspended particles interact
with the sediment layers by sedimentation and diffusive exchange with the sediment pore
water affecting concentrations of dissolved substances.

In Table 3.1 all state variables of the model are summarized. The biological part of the
model is represented by the three plankton classes *Planktothrix (Oscillatoria) rubescens*
(most abundant blue-green algae in Lake Zürich), other algae and zooplankton. The blue-
green alga *Planktothrix (Oscillatoria) rubescens* was selected as a separate plankton class
because in the average of many years it contributes up to about 20 % of the phytoplank-
ton biomass in Lake Zürich (Gammeter et al. 1997), its optimal growth conditions differ
significantly from that of other algae and it is of special interest for the water supply
authority. Biodegradable and inert organic matter summarize organic particles resulting
from allochthonous sources, from death of algae and zooplankton, and from zooplankton
excretion as fecal pellets. The phosphorus contents of algae, *Planktothrix (Oscillatoria)*
rubescens and organic matter are separate state variables because the variable stoichiom-
etry of primary production leads to a variable phosphorus content of these particles. *Zoo*
plankton, on the other hand, is described with a constant phosphorus content according to the Redfield stoichiometry (Redfield et al. 1966). In addition, the phosphorus content resulting from phosphate uptake by sinking particles is considered as a state variable. According to our favourite explanation of this process (Hupfer et al. 1995), this state variable is denoted inorganic phosphorus (this point is discussed in more detail in section 3.4). Phosphate, ammonium and nitrate are the most relevant nutrients and together with dissolved oxygen represent the dissolved state variables of the model.

### 3.3.2 Mass Balance Equations

In the water column (index: \(i\)), the following equation is solved for each dissolved state variable

\[
\frac{\partial S_i}{\partial t} = \frac{1}{A} \frac{\partial}{\partial z} \left( AK_z \frac{\partial S_i}{\partial z} \right) + \frac{q}{A} S_{in} - \frac{1}{A} \frac{\partial}{\partial z} \left( QS_i \right) + r_i - \frac{1}{A} \frac{D \theta}{Dz} \frac{\partial}{\partial z} (S_i - S_{a}) \quad (3.1)
\]

and the following equation for each particulate state variable

\[
\frac{\partial X_i}{\partial t} = \frac{1}{A} \frac{\partial}{\partial z} \left( AK_z \frac{\partial X_i}{\partial z} \right) + \frac{q}{A} X_{in} - \frac{1}{A} \frac{\partial}{\partial z} \left( QX_i \right) + r_i - \frac{\partial}{\partial z} (v_{sed} X_i) \quad . (3.2)
\]

In these equations \(S\) is the concentration of a dissolved compound (ML\(^{-3}\)), \(X\) is the concentration of a particulate compound (ML\(^{-3}\)), \(t\) is time (T), \(z\) is the vertical coordinate pointing downwards (\(z = 0\) at the lake surface) (L), \(A\) is the cross-sectional area of the lake (L\(^2\)), \(K_z\) is the vertical mixing coefficient (L\(^2\)T\(^{-1}\)), \(r\) are transformation rates (ML\(^{-3}\)T\(^{-1}\)), \(q\) is the volumetric lateral inflow (rivers and sewage treatment plants are assumed to

<table>
<thead>
<tr>
<th>State Variable</th>
<th>Unit</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Particulate</td>
<td></td>
</tr>
<tr>
<td>(X_{ALG})</td>
<td>gDM/m(^3)</td>
<td>algae without <em>Planktothrix rubescens</em> (dry mass without P)</td>
</tr>
<tr>
<td>(X_{PLR})</td>
<td>gDM/m(^3)</td>
<td><em>Planktothrix rubescens</em> (dry mass without P)</td>
</tr>
<tr>
<td>(X_{ZOO})</td>
<td>gDM/m(^3)</td>
<td>zooplankton (dry mass with Redfield composition)</td>
</tr>
<tr>
<td>(X_{S})</td>
<td>gDM/m(^3)</td>
<td>biodegradable organic material (dry mass without P)</td>
</tr>
<tr>
<td>(X_{I})</td>
<td>gDM/m(^3)</td>
<td>inert organic material (dry mass without P)</td>
</tr>
<tr>
<td>(X_{P,ALG})</td>
<td>gP/m(^3)</td>
<td>organic phosphorus in algae</td>
</tr>
<tr>
<td>(X_{P,PLR})</td>
<td>gP/m(^3)</td>
<td>organic phosphorus in <em>Planktothrix rubescens</em></td>
</tr>
<tr>
<td>(X_{P,S})</td>
<td>gP/m(^3)</td>
<td>organic phosphorus in biodegradable organic material</td>
</tr>
<tr>
<td>(X_{P,I})</td>
<td>gP/m(^3)</td>
<td>organic phosphorus in inert organic material</td>
</tr>
<tr>
<td>(X_{P,LS})</td>
<td>gP/m(^3)</td>
<td>phosphate attached to biodegradable organic material</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>State Variable</th>
<th>Unit</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dissolved</td>
<td></td>
</tr>
<tr>
<td>(S_{PHPO4})</td>
<td>gP/m(^3)</td>
<td>phosphate-phosphorus</td>
</tr>
<tr>
<td>(S_{NH4})</td>
<td>gN/m(^3)</td>
<td>ammonia-nitrogen</td>
</tr>
<tr>
<td>(S_{NO3})</td>
<td>gN/m(^3)</td>
<td>nitrate-nitrogen</td>
</tr>
<tr>
<td>(S_{O2})</td>
<td>gO/m(^3)</td>
<td>oxygen</td>
</tr>
</tbody>
</table>

Table 3.1: State variables of the lake model.
discharge into the epilimnion) \((L^2 T^{-1})\), \(S_{in}\) is the inflow concentration of a dissolved compound \((ML^{-3})\), \(X_{in}\) is the inflow concentration of a particulate compound \((ML^{-3})\), \(Q\) is the volumetric upward flow above lateral inflows \((L^3 T^{-1})\), \(D\) is the coefficient of molecular diffusion \((L^2 T^{-1})\), \(\theta\) is the porosity of the sediment \((-)\), \(h_{sed}\) is the thickness of the sediment layers \((L)\), and \(v_{sed}\) is the sedimentation velocity \((LT^{-1})\). In both equations, the first term on the right hand side describes mixing as an effective diffusion process, the second term describes lateral inflow, the third term advection in the water column resulting from lateral inflows (the outflow is at the lake surface) and the fourth term describes the net effect of all transformation processes. The last term of equation \((3.1)\) describes exchange of the dissolved substance with the pore water of the upper sediment layer (index: \(s_1\)), the last term of equation \((3.2)\) describes sedimentation of particles through the water column.

Equation \((3.1)\) is coupled to the following two equations for dissolved substances in the pore water of the two modelled sediment layers:

\[
\frac{\partial S_{s_1}}{\partial t} = \frac{1}{\theta h_{sed}} \left( \frac{D \theta}{h_{sed}} (S_{s_1} - S_{s_1}) + \frac{D \theta}{h_{sed}} (S_{s_2} - S_{s_1}) \right) + \frac{r_s}{\theta} \]

\[
\frac{\partial S_{s_2}}{\partial t} = \frac{1}{\theta h_{sed}} + \frac{D \theta}{h_{sed}} (S_{s_1} - S_{s_2}) + \frac{r_s}{\theta} \cdot \]

Here, the index \(s_2\) refers to the lower sediment layer. The equation for particulate substances \((3.2)\) is coupled with the two sediment equations:

\[
\frac{\partial X_{s_1}}{\partial t} = \frac{1}{h_{sed}} \left( X_{i, s_{sed}} - \frac{F_{vol,1}}{1 - \theta} X_{s_1} \right) + r_s \]

\[
\frac{\partial X_{s_2}}{\partial t} = \frac{1}{h_{sed}} \left( \frac{F_{vol,1}}{1 - \theta} X_{s_1} - \frac{F_{vol,2}}{1 - \theta} X_{s_2} \right) + r_s \cdot \]

where the volumetric particle fluxes between water column and the upper sediment layer, between the two sediment layers and between the bottom sediment layer and the permanent sediment (not modelled) are given as

\[
F_{vol,0} = \sum_X \frac{X_{i, s_{sed}}}{\rho} + F_{vol,ext} \]

and

\[
F_{vol,i} = F_{vol,i-1} + h_{sed} \sum_X \frac{r_s}{\rho}, \quad i = 1, 2 \]

where summation is over all particulate substances and \(F_{vol,ext}\) is an additional particle flux by substances that are not modelled.

### 3.3.3 Physical Processes

**Vertical mixing:** Information about diffusion coefficients was gained from the following sources: Strong mixing in the epilimnion was described with a large effective diffusion
coefficient in this zone. The heat budget method (Powell and Jassby 1974) was used to estimate diffusion coefficients in the metalimnion. Due to the low resolution of temperature data and only small variations between temperature profiles the uncertainty of calculated values was very large. In the hypolimnion for the same reasons the method could not be applied. Additional information about typical diffusion coefficients were taken from Lä (1973), where monthly temperature averages over the years from 1948 to 1957 were used for calculation. Phosphate profiles were used as follows, to get information about diffusion in the hypolimnion: During the stagnation period with small exchange between epi- and hypolimnion, the phosphate increase between two profiles was calculated. Assuming a constant phosphate flow and neglecting transformation processes, the heat budget technique was then applied to the phosphate concentration. All this information about diffusion coefficients was then summarized with constant values for summer and winter periods in the epi-, the meta- and the upper and lower hypolimnion. Then time dependent boundaries between these zones were taken from temperature data. A simulation of the temperature with these diffusion coefficients was then performed where the temperature above 10 m depth was given as a boundary condition and small corrections to the diffusion coefficients were made in order to get a good agreement between the simulation and the temperature data.

**Sedimentation:** Particles are assumed to have a constant sedimentation velocity, \( v_{\text{sed}} \) (LT\(^{-1}\)), which is different for different types of particles. The sedimentation velocity of *Planktothrix (Oscillatoria) rubescens*, which regulate their buoyancy with gas vacuoles and of zooplankton, which moves through the lake, was set to zero. The class of other algae was assumed to have a small sedimentation velocity, and the sedimentation velocity of dead organic particles was selected to be relatively large (cf. Table 3.5 in the appendix).

**Gas exchange at the lake surface:** Dissolved oxygen exchange is considered as a boundary condition at the lake surface with a flux proportional to the difference of the current oxygen concentration and saturation. The gas exchange velocity, \( v_{O_2,\text{atm}} \), is approximated to be constant.

**Light absorption:** Light intensity is assumed to decrease with water depth. The light extinction coefficient, \( \eta \) (L\(^{-1}\)), is assumed to depend linearly on the concentration of suspended particles \( X_{\text{part}} \) (ML\(^{-3}\)):

\[
\eta(z) = k_1 + k_2 \cdot X_{\text{part}}(z)
\]  

(3.9)

In this expression, \( k_1 \) (L\(^{-1}\)) and \( k_2 \) (M\(^{-1}\)L\(^2\)) are empirical parameters which were estimated with light and algae data from lake Zürich, and \( X_{\text{part}} \) is the sum of all particulate state variables given in Table 3.1. With these light extinction coefficients, the decrease of light intensity is

\[
I(z) = I_0 \exp \left( - \int_0^z \eta(z) \, dz \right)
\]  

(3.10)

where \( I_0 \) denotes the light intensity at the water surface.
3.3.4 Biogeochemical Processes

Table 3.2 gives an overview of the biogeochemical processes used in the model. Each row in this table represents one process with its stoichiometric coefficients for all state variables involved. Formulas for stoichiometric variables used in this Table are given in Table 3.3. With the process rates given in Table 3.4 the contribution of a process to the transformation rate of each component is calculated by multiplying the rate with the corresponding stoichiometric coefficient. Nutrient and oxygen dependence is formulated with Monod-type expressions for the transition from unlimited to limited rates. In some cases, multiplication of limiting factors has been shown to give poor results (De Groot 1983). For this reason, in the case of the dependence of a process rate on more than one nutrient concentration, in accordance with Liebig's law of the minimum, only the most limiting nutrient is considered for reducing the process rate.

The stoichiometric coefficients were calculated with the technique and the spreadsheet provided in Reichert et al. (2000), based on a Redfield composition of all organic particles. Some of the coefficients where then modified to account for the variable stoichiometry of primary production with respect to phosphorus. Because the phosphorus content is very small, the other mass fractions were not corrected in this step.

All conversion rates depend exponentially on temperature. All processes are active in the water column and in all sediment layers, however, some parameters have different values in the water column than in the sediment (e.g. the light intensity is zero in the sediment and mineralization rates are much smaller in the water column than in the sediment).

The values of the parameters of the biogeochemical processes used for the simulation are listed in the Tables 3.6 and 3.7 in the appendix.

Growth Processes: Algae, Planktothrix rubescens and Zooplankton

Algae: In order to account for the strong phosphorus limitation of algae growth during the summer, algae growth is modelled with a variable stoichiometry with respect to phosphorus. This allows algae to grow with a much smaller phosphorus content if phosphate is present in small concentrations than if it is readily available. In contrast to this consideration of seasonal changes in composition, changes in composition due to different dynamics of nutrient uptake and growth processes (Ketchum 1939), that could be modelled with the model proposed by Droop (1983), are neglected because they are not relevant on the time scale of operation of the model (months to years).

The stoichiometry of algae growth is described in row 1 of Table 3.2. It is based on the biomass composition according to Redfield et al. (1966) with the exception of the phosphorus content of the newly built biomass that is assumed to depend on the phosphate concentration in the water. The dependence of the phosphorus content of newly built algae, \( b_P \), on phosphate concentration in the water is formalized by the expression given in Table 3.3 and shown in Fig. 3.2. The variability in the stoichiometric coefficient for oxygen caused by the variability in the coefficient of phosphate is neglected. The dependence shown in Fig. 3.2 leads to algae growth according to the composition proposed by Redfield et al. (1966) if phosphate concentration in the water is high enough, whereas they grow with six times less phosphorus under strongly phosphorus limiting conditions. The necessity of using such a variability in the stoichiometry with respect to phosphorus is known from other lakes (Hupfer et al. 1995). Its necessity for the application of the model to lake Zürich
Figure 3.2: Dependence of the phosphorus content of newly built algae, $b_p$, on the phosphate concentration in the water column.

is discussed in section 3.4. Because of the low concentrations of ammonia in the epilimnion and the availability of nitrate, the main nitrogen source for algae growth is assumed to be nitrate. Under different circumstances it may be advantageous to introduce a switching function from ammonia to nitrate, depending on the concentrations of these compounds (Brown and Barnwell 1987).

The growth rate of algae is given in row 1 of Table 3.4. Light and nutrient limitations are described with Monod-type rate reduction factors. Only the reduction factor of the most limiting nutrient is considered.

**Planktothrix (Oscillatoria) rubescens:** The blue-green algae *Planktothrix (Oscillatoria) rubescens* is a low light specialist that keeps its vertical position with the aid of gas vacuoles at the lower border of the epilimnion (Lampeiff and Sommer 1997). In the most parsimonious lake model, *Planktothrix (Oscillatoria) rubescens* could be included in the class summarizing all other algae described above. *Planktothrix (Oscillatoria) rubescens* are considered as an own state variable because they are of special interest for the water supply authority and because their significantly different growth behaviour leads to different growth parameters.

The stoichiometry of growth of *Planktothrix (Oscillatoria) rubescens* given in row 2 of Table 3.2 is analogous to that of the other algae described above. Also for *Planktothrix (Oscillatoria) rubescens* nitrogen uptake is based on nitrate because of the availability of this nutrient in the epilimnion. An important difference in comparison to other algae in the model is that *Oscillatoria rubescens* are not exposed to grazing by zooplankton (see below).

The growth rate of *Planktothrix (Oscillatoria) rubescens* is given in row 2 of Table 3.4. It considers light inhibition using the light function proposed by Steele (1962) with a small value for the light intensity at maximum growth, $K_{I,PLR}$. The ability of holding its vertical position is accounted for by setting the sedimentation velocity to zero (cf. Table 3.5).
Zooplankton: The simplest possible zooplankton model is used which summarizes all organisms in one class that feeds on algae. In Lake Zürich Daphnia is the most abundant zooplankton species. Because this species has been shown to have a much greater impact on phytoplankton than the other species (Pace 1984; McQueen et al. 1986), modelled zooplankton concentrations are compared with measurements for Daphnia. Investigations of C, N and P content of natural zooplankton populations by Andersen and Hessen (1991) have shown little variations in composition and a constant composition was maintained even in starvation and food enrichment experiments. For this reason, in contrast to algae, zooplankton is modelled with constant elemental mass fractions.

The stoichiometry of growth of zooplankton in the model is given in row 3 of Table 3.2. The composition of zooplankton is assumed to remain constant at the Redfield values. In order to be able to maintain such a fixed composition under varying phosphorus content of the food \((a_{P, ALG})\) the yield \((Y_{ZOO})\) must decrease with decreasing phosphorus content of the food. This means that more algal biomass is consumed to build one unit of zooplankton biomass if the phosphorus content of the algae is low. Many studies support this idea of nutrient limited zooplankton growth (Elser and Hassett 1994; Sommer 1991; Urabe and Watanabe 1992; Hessen 1992). This phosphorus limitation is modelled with a yield that is proportional to the phosphorus content if the phosphorus content is below the Redfield value (Table 3.3). The variable \(f_c\) describes the fraction of consumed organic matter that is excreted as fecal pellets (the fraction \(Y_{ZOO}\) is converted to zooplankton biomass, the fraction \(1 - Y_{ZOO} - f_e\) is mineralized). Because the yield depends strongly on the phosphorus content of the food, \(f_c\) is parameterized with the parameter \(c_e\) which specifies the fraction of organic matter not used for zooplankton biomass that is excreted as fecal pellets (the fraction \(1 - c_e\) is mineralized). \(c_e\) rather than \(f_e\) is assumed to be independent of the phosphorus content of the food. It is assumed that all phosphorus that is not incorporated into zooplankton biomass is released in the form of phosphate so that the phosphorus content of fecal pellets is zero. This is a rather crude approximation to the low observed phosphorus content of fecal pellets (H.R. Bürgi, personal communication) which had probably to be changed for a less phosphorus limited lake.

The growth rate of zooplankton is given in row 3 of Table 3.4. It is proportional to the product of food and zooplankton concentrations. A limiting factor proportional to the phosphorus content of the food leads to a reduction of the growth rate if the phosphorus content is below the Redfield value. Note that for the grazing rate of algae this factor is compensated by the same factor in the yield (cf. Tables 3.2 and 3.3). This means that the grazing rate of algae is independent of their phosphorus content.

Algae and Zooplankton Respiration

The processes of algae and zooplankton respiration describe maintenance of these organisms at the expense of their own biomass. These processes are difficult to identify because (usually) only net growth is observed (identifiability becomes even worse with the introduction of death processes below). Nevertheless the process is introduced because spatial separation of growth and respiration can be important in lakes. In the epilimnion, respiration just leads to a reduction in observed growth rates. However, if algae sink and respire in the metalimnion, where light conditions do not allow them to grow, respiration can contribute to the observed metalimnonic oxygen minimum.

The stoichiometry of algae, Planktothrix (Oscillatoria) rubescens and zooplankton res-
piration are described in rows 4, 5 and 6 of Table 3.2. The three processes are similar with the minor difference, that phosphate release by respiring algae and Planktothrix (Oscillatoria) rubescens depends on their current phosphorus content, $a_{ALG}$ and $a_{OSC}$, respectively, whereas phosphate release by respiring zooplankton is constant because of the constant elemental composition of zooplankton.

The process rates of algae, Planktothrix (Oscillatoria) rubescens and zooplankton respiration are given in the rows 4, 5 and 6 of Table 3.4. The rates are proportional to the concentration of the respired species with limiting factors with respect to oxygen and with the same temperature dependence as the corresponding growth processes.
<table>
<thead>
<tr>
<th>no.</th>
<th>process</th>
<th>( S_{NH_4} )</th>
<th>( S_{NO_3} )</th>
<th>( S_{HPO_4} )</th>
<th>( S_{SO_2} )</th>
<th>( X_{ALG} )</th>
<th>( X_{P, ALG} )</th>
<th>( X_{PLR} )</th>
<th>( X_{P, PLR} )</th>
<th>( X_{ZOO} )</th>
<th>( X_S )</th>
<th>( X_{P, S} )</th>
<th>( X_{P, I} )</th>
<th>( X_I )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>growth alg</td>
<td>( -a_N )</td>
<td>( -a_N )</td>
<td>1.2</td>
<td>1.2</td>
<td>1.2</td>
<td>1</td>
<td>( b_P )</td>
<td>( b_P )</td>
<td>1</td>
<td>( f_p )</td>
<td>( f_p )</td>
<td>( f_p )</td>
<td>( f_p )</td>
</tr>
<tr>
<td>2</td>
<td>growth osc</td>
<td>( -a_N )</td>
<td>( -a_N )</td>
<td>1</td>
<td>( -1 ) Y_{ZOO}</td>
<td>( -a_P, ALG )</td>
<td>( -a_P, ALG )</td>
<td>1</td>
<td>( -a_P, OSC )</td>
<td>( -1 ) ( f_p ) Y_{ZOO}</td>
<td>0</td>
<td>( f_p )</td>
<td>( f_p )</td>
<td>( f_p )</td>
</tr>
<tr>
<td>3</td>
<td>growth zoo</td>
<td>( a_N )</td>
<td>( a_N )</td>
<td>( a_P, ALG )</td>
<td>( a_P, OSC )</td>
<td>( a_P, OSC )</td>
<td>( -1 )</td>
<td>( -a_P, ALG )</td>
<td>( -1 ) ( f_p )</td>
<td>( 1 ) ( f_p )</td>
<td>( f_p )</td>
<td>( f_p )</td>
<td>( f_p )</td>
<td>( f_p )</td>
</tr>
<tr>
<td>4</td>
<td>resp alg</td>
<td>( a_N )</td>
<td>( a_N )</td>
<td>( a_P, ALG )</td>
<td>( a_P, OSC )</td>
<td>( a_P, OSC )</td>
<td>( -1 )</td>
<td>( -a_P, ALG )</td>
<td>( -1 ) ( f_p )</td>
<td>( 1 ) ( f_p )</td>
<td>( f_p )</td>
<td>( f_p )</td>
<td>( f_p )</td>
<td>( f_p )</td>
</tr>
<tr>
<td>5</td>
<td>resp osc</td>
<td>( a_N )</td>
<td>( a_N )</td>
<td>( a_P, ALG )</td>
<td>( a_P, OSC )</td>
<td>( a_P, OSC )</td>
<td>( -1 )</td>
<td>( -a_P, ALG )</td>
<td>( -1 ) ( f_p )</td>
<td>( 1 ) ( f_p )</td>
<td>( f_p )</td>
<td>( f_p )</td>
<td>( f_p )</td>
<td>( f_p )</td>
</tr>
<tr>
<td>6</td>
<td>resp zoo</td>
<td>( a_N )</td>
<td>( a_N )</td>
<td>( a_P, ALG )</td>
<td>( a_P, OSC )</td>
<td>( a_P, OSC )</td>
<td>( -1 )</td>
<td>( -a_P, ALG )</td>
<td>( -1 ) ( f_p )</td>
<td>( 1 ) ( f_p )</td>
<td>( f_p )</td>
<td>( f_p )</td>
<td>( f_p )</td>
<td>( f_p )</td>
</tr>
<tr>
<td>7</td>
<td>death alg</td>
<td>( a_N )</td>
<td>( a_N )</td>
<td>( a_P, ALG )</td>
<td>( a_P, OSC )</td>
<td>( a_P, OSC )</td>
<td>( -1 )</td>
<td>( -a_P, ALG )</td>
<td>( -1 ) ( f_p )</td>
<td>( 1 ) ( f_p )</td>
<td>( f_p )</td>
<td>( f_p )</td>
<td>( f_p )</td>
<td>( f_p )</td>
</tr>
<tr>
<td>8</td>
<td>death osc</td>
<td>( a_N )</td>
<td>( a_N )</td>
<td>( a_P, ALG )</td>
<td>( a_P, OSC )</td>
<td>( a_P, OSC )</td>
<td>( -1 )</td>
<td>( -a_P, ALG )</td>
<td>( -1 ) ( f_p )</td>
<td>( 1 ) ( f_p )</td>
<td>( f_p )</td>
<td>( f_p )</td>
<td>( f_p )</td>
<td>( f_p )</td>
</tr>
<tr>
<td>9</td>
<td>death zoo</td>
<td>( a_N )</td>
<td>( a_N )</td>
<td>( a_P, ALG )</td>
<td>( a_P, OSC )</td>
<td>( a_P, OSC )</td>
<td>( -1 )</td>
<td>( -a_P, ALG )</td>
<td>( -1 ) ( f_p )</td>
<td>( 1 ) ( f_p )</td>
<td>( f_p )</td>
<td>( f_p )</td>
<td>( f_p )</td>
<td>( f_p )</td>
</tr>
<tr>
<td>10</td>
<td>aer miner</td>
<td>( a_N )</td>
<td>( a_N )</td>
<td>( a_P, S + a_P, I )</td>
<td>( -0.94 )</td>
<td>( -a_P, S + a_P, I )</td>
<td>( -1 )</td>
<td>( -a_P, ALG )</td>
<td>( -1 ) ( f_p )</td>
<td>( 1 ) ( f_p )</td>
<td>( f_p )</td>
<td>( f_p )</td>
<td>( f_p )</td>
<td>( f_p )</td>
</tr>
<tr>
<td>11</td>
<td>anox miner</td>
<td>( a_N )</td>
<td>( a_N )</td>
<td>( a_P, S + a_P, I )</td>
<td>( -0.94 )</td>
<td>( -a_P, S + a_P, I )</td>
<td>( -1 )</td>
<td>( -a_P, ALG )</td>
<td>( -1 ) ( f_p )</td>
<td>( 1 ) ( f_p )</td>
<td>( f_p )</td>
<td>( f_p )</td>
<td>( f_p )</td>
<td>( f_p )</td>
</tr>
<tr>
<td>12</td>
<td>nitrification</td>
<td>( a_N )</td>
<td>( a_N )</td>
<td>( a_P, S + a_P, I )</td>
<td>( -0.94 )</td>
<td>( -a_P, S + a_P, I )</td>
<td>( -1 )</td>
<td>( -a_P, ALG )</td>
<td>( -1 ) ( f_p )</td>
<td>( 1 ) ( f_p )</td>
<td>( f_p )</td>
<td>( f_p )</td>
<td>( f_p )</td>
<td>( f_p )</td>
</tr>
<tr>
<td>13</td>
<td>P-uptake</td>
<td>( -1 )</td>
<td>( 1 )</td>
<td>( -1 )</td>
<td>( -1 )</td>
<td>( -4.6 )</td>
<td>( -1 )</td>
<td>( -a_P, I )</td>
<td>( -1 ) ( f_p )</td>
<td>( 1 ) ( f_p )</td>
<td>( f_p )</td>
<td>( f_p )</td>
<td>( f_p )</td>
<td>( f_p )</td>
</tr>
</tbody>
</table>

Table 3.2: Biogeochemical conversion processes with stoichiometric coefficients for all state variables.
<table>
<thead>
<tr>
<th>Variable</th>
<th>Algebraic expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_{P,ALG}$</td>
<td>$\frac{X_{P,ALG}}{X_{ALG}}$</td>
</tr>
<tr>
<td>$a_{P,OSC}$</td>
<td>$\frac{X_{P,PLR}}{X_{PLR}}$</td>
</tr>
<tr>
<td>$a_{P,S}$</td>
<td>$\frac{X_{P,S}}{X_{S}}$</td>
</tr>
<tr>
<td>$a_{P,I,S}$</td>
<td>$\frac{X_{P,I,S}}{X_{S}}$</td>
</tr>
<tr>
<td>$a_{P,I}$</td>
<td>$\frac{X_{P,I}}{X_{I}}$</td>
</tr>
<tr>
<td>$b_{p}$</td>
<td>$\frac{b_{p_{min}} + b_{p_{max}}}{2} + \frac{b_{p_{max}} - b_{p_{min}}}{2} \cdot \tanh\left(\frac{S_{HPO_{4}} - S_{HPO_{4}^{crit}}}{\Delta S_{HPO_{4}}}\right)$</td>
</tr>
<tr>
<td>$Y_{ZOO}$</td>
<td>$Y_{ZOO_{max}} \cdot \min\left(1, \frac{a_{P,ALG}}{a_{P,red}}\right)$</td>
</tr>
<tr>
<td>$f_{e}$</td>
<td>$c_{e}(1 - Y_{ZOO})$</td>
</tr>
<tr>
<td>$v_{ZOO,N}$</td>
<td>$a_{N} \cdot \frac{1 - Y_{ZOO} - f_{e}}{Y_{ZOO}}$</td>
</tr>
<tr>
<td>$v_{ZOO,P}$</td>
<td>$a_{P,ALG} \cdot \frac{1 - a_{P,red}}{Y_{ZOO}}$</td>
</tr>
<tr>
<td>$v_{ZOO,O}$</td>
<td>$0.93 \cdot \frac{1 - Y_{ZOO} - f_{e}}{Y_{ZOO}}$</td>
</tr>
</tbody>
</table>

Table 3.3: Mathematical expressions for stoichiometric variables used in Table 3.2.
<table>
<thead>
<tr>
<th>no.</th>
<th>Process</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>growth alg</td>
<td>( k_{\text{gro,alg},T_0} \cdot \exp(\beta_{\text{alg}}(T - T_0)) \cdot \frac{I(z)}{K_{I,\text{alg}} + I(z)} \cdot \min \left( \frac{S_{\text{NO}<em>3}}{K</em>{\text{NO}<em>3,\text{alg}} + S</em>{\text{NO}<em>3}}, \frac{S</em>{\text{HPO}<em>4}}{K</em>{\text{HPO}<em>4,\text{alg}} + S</em>{\text{HPO}<em>4}} \right) \cdot X</em>{\text{alg}} )</td>
</tr>
<tr>
<td>2</td>
<td>growth osc</td>
<td>( k_{\text{gro,plr},T_0} \cdot \exp(\beta_{\text{plr}}(T - T_0)) \cdot \frac{I(z)}{K_{I,\text{plr}} \exp \left( \frac{1 - I(z)}{K_{I,\text{plr}}} \right) \cdot \min \left( \frac{S_{\text{NO}<em>3}}{K</em>{\text{NO}<em>3,\text{plr}} + S</em>{\text{NO}<em>3}}, \frac{S</em>{\text{HPO}<em>4}}{S</em>{\text{HPO}<em>4}} \right) \cdot X</em>{\text{plr}} )</td>
</tr>
<tr>
<td>3</td>
<td>growth zoo</td>
<td>( k_{\text{gro,zoo},T_0} \cdot \exp(\beta_{\text{zoo}}(T - T_0)) \cdot X_{\text{alg}} \cdot \min \left( \frac{a_{P,\text{alg}}}{a_{P,\text{red}}}, \frac{S_{\text{NO}<em>3}}{K</em>{\text{NO}<em>3,\text{alg}} + S</em>{\text{NO}<em>3}}, \frac{S</em>{\text{NH}<em>4}}{K</em>{\text{NH}<em>4,\text{zoo}} + S</em>{\text{NH}<em>4}} \right) \cdot X</em>{\text{zoo}} )</td>
</tr>
<tr>
<td>4</td>
<td>resp alg</td>
<td>( k_{\text{resp,alg},T_0} \cdot \exp(\beta_{\text{alg}}(T - T_0)) \cdot \frac{S_{\text{O}<em>2}}{K</em>{\text{O}<em>2,\text{resp}} + S</em>{\text{O}<em>2}} \cdot X</em>{\text{alg}} )</td>
</tr>
<tr>
<td>5</td>
<td>resp osc</td>
<td>( k_{\text{resp,plr},T_0} \cdot \exp(\beta_{\text{plr}}(T - T_0)) \cdot \frac{S_{\text{O}<em>2}}{K</em>{\text{O}<em>2,\text{resp}} + S</em>{\text{O}<em>2}} \cdot X</em>{\text{plr}} )</td>
</tr>
<tr>
<td>6</td>
<td>resp zoo</td>
<td>( k_{\text{resp,zoo},T_0} \cdot \exp(\beta_{\text{zoo}}(T - T_0)) \cdot \frac{S_{\text{O}<em>2}}{K</em>{\text{O}<em>2,\text{resp}} + S</em>{\text{O}<em>2}} \cdot X</em>{\text{zoo}} )</td>
</tr>
<tr>
<td>7</td>
<td>death alg</td>
<td>( k_{\text{death,alg},T_0} \cdot \exp(\beta_{\text{alg}}(T - T_0)) \cdot X_{\text{alg}} )</td>
</tr>
<tr>
<td>8</td>
<td>death osc</td>
<td>( k_{\text{death,plr},T_0} \cdot \exp(\beta_{\text{plr}}(T - T_0)) \cdot X_{\text{plr}} )</td>
</tr>
<tr>
<td>9</td>
<td>death zoo</td>
<td>( k_{\text{death,zoo},T_0} \cdot \exp(\beta_{\text{zoo}}(T - T_0)) \cdot X_{\text{zoo}} )</td>
</tr>
<tr>
<td>10</td>
<td>aer miner</td>
<td>( k_{\text{miner,aero},T_0} \cdot \exp(\beta_{\text{bac}}(T - T_0)) \cdot \frac{S_{\text{O}<em>2}}{K</em>{\text{O}<em>2,\text{miner}} + S</em>{\text{O}<em>2}} \cdot X</em>{\text{S}} )</td>
</tr>
<tr>
<td>11</td>
<td>anox miner</td>
<td>( k_{\text{miner,anox},T_0} \cdot \exp(\beta_{\text{bac}}(T - T_0)) \cdot \frac{S_{\text{NO}<em>3}}{K</em>{\text{NO}<em>3,\text{miner}} + S</em>{\text{NO}<em>3}} \cdot \left( 1 - \frac{S</em>{\text{O}<em>2}}{K</em>{\text{O}<em>2,\text{miner}} + S</em>{\text{O}<em>2}} \right) \cdot X</em>{\text{S}} )</td>
</tr>
<tr>
<td>12</td>
<td>nitrification</td>
<td>( k_{\text{nitr},T_0} \cdot \exp(\beta_{\text{bac}}(T - T_0)) \cdot \min \left( \frac{S_{\text{O}<em>2}}{K</em>{\text{O}<em>2,\text{nitr}} + S</em>{\text{O}<em>2}}, \frac{S</em>{\text{NH}<em>4}}{K</em>{\text{NH}<em>4,\text{nitr}} + S</em>{\text{NH}<em>4}} \right) \cdot X</em>{\text{S}} )</td>
</tr>
<tr>
<td>13</td>
<td>P-uptake</td>
<td>( k_{\text{upt}} \cdot \frac{1}{A} \left</td>
</tr>
</tbody>
</table>

Table 3.4: Process rates of biogeochemical conversion processes.
Death Processes

Death processes transform algae, *Planktothrix (Oscillatoria) rubescens* and zooplankton into degradable and inert organic material. Although it is not relevant if an algae sinking to the sediment is dead or alive (this means that it would again start primary production if exposed to light), the death processes simplify the model because they cause the sediment to consist (nearly) only of the model compounds of degradable and inert organic matter.

The stoichiometry of the death processes is given in rows 7 to 9 of Table 3.2. Algae, *Planktothrix (Oscillatoria) rubescens* and zooplankton are converted to a fraction \(1 - f_p\) of degradable organic material and a fraction \(f_p\) of inert organic material.

The process rates of the death processes is given in the rows 7 to 9 of Table 3.4. They are proportional to the current concentration and have the same temperature dependence as the growth processes.

Mineralization Processes

Mineralization processes account for bacterially mediated oxidation of degradable organic matter. Since bacteria are not modelled as state variables, a constant maximum mineralization rate is assumed and only limitation terms for degradable organic matter and oxidants are considered. Only aerobic mineralization using dissolved oxygen and anoxic mineralization using nitrate as the oxidant (denitrification) are used in the model. These two processes are known to be very important for the lake. The first process leads to a significant reduction of oxygen concentrations in the deep hypolimnion, the second to a large nitrogen elimination in the lake (Mengis et al. 1997). It is assumed that the turnover rates of all other oxidation processes (taking place in deeper sediment layers) are small compared to the turnover rates of these two oxidation processes. While this assumption seems reasonable for lake Zürich, in highly eutrophic lakes with very small oxygen concentration in the deep water anaerobic mineralization processes can become relevant.

The stoichiometries of aerobic and anoxic mineralization are given in the rows 10 and 11 of Table 3.2. The nitrogen balance is not closed for anoxic mineralization because molecular nitrogen gas (\(N_2\)) escaping from the lake is not considered as a state variable in the model. Because the desorption rate of phosphate adsorbed to organic particles is not significantly different from the mineralization rate in the sediment (Hupfer et al. 1995), desorption of adsorbed phosphate was included in the stoichiometry of the mineralization process (see below for a discussion of phosphate adsorption on organic particles).

The process rate of aerobic mineralization given in row 10 of Table 3.4 has a Monod-type limitation factor with respect to oxygen. The rate of anoxic mineralization given in row 11 of Table 3.4 has an inhibition factor with respect to oxygen and a limitation factor with respect to nitrate. Because oxygen concentrations in lake Zürich are never very low, significant anoxic mineralization (denitrification) only occurs in the sediment. The rate parameters \(k_{\text{miner,aero,T0}}\) and \(k_{\text{miner,anox,T0}}\) are much higher in the sediment than in the water column in order to account for the much higher bacterial density in the sediment.

Nitrification

The stoichiometry of nitrification is given in row 12 of Table 3.2. Ammonia is oxidized to nitrate with consumption of oxygen.
The process rate of nitrification is given in row 12 of Table 3.4. There are limitation factors for ammonia and oxygen. Similarly to mineralization discussed above, the maximum rate, \( k_{\text{min},T_0} \), is larger in the sediment than in the water column to account for the higher bacterial density.

**Uptake and Release of Phosphate**

In many lakes it can be observed that phosphate concentrations are very low during the summer not only within the photic zone, where it is consumed by growing algae, but also below. This can be explained with a phosphate uptake process on sinking particles. The existence of such a process has been confirmed by investigating organic particles from sediment traps at different water depths in Lake Sempach (Gächter and Mars 1985; Hupfer et al. 1995). It was shown that particles became enriched with inorganic phosphorus while settling through the hypolimnion, whereas the organic phosphorus content did not change significantly. One hypothesis for an explanation of this observation is that manganese and iron oxidising bacteria of the type *Metallogenium* form iron hydroxide surfaces on which inorganic phosphorus adsorbs abiotically (Gächter and Mars 1985). However, there may be other processes contributing to the observed increase in inorganic phosphorus content of the sinking particles. In order to account for this observation a process of phosphate uptake by organic particles was included in the model and the corresponding phosphorus was denoted 'inorganic phosphorus'. Note, however, that the chosen macroscopic model formulation does not depend on the microscopic mechanism of this uptake process which needs to be further investigated.

The stoichiometry of this phosphate uptake processes is very simple. It is given in row 13 of Table 3.2. As mentioned above, desorption is included in the mineralization process given in rows 10 and 11 of Table 3.2.

The transformation rate of the phosphorus uptake process is given in row 13 of Table 3.4. It constitutes an empirical parameterization of the poorly understood process, that is able to account for its macroscopic effects. The formulation of a more universal transformation rate needs more investigations on the microscopic mechanisms leading to phosphate uptake. Because iron diffusing out of the sediment could be the source for the build up of iron oxide surfaces to which phosphate adsorbs, the uptake rate is assumed to be proportional to the sediment area per unit of lake volume, \( 1/A \cdot dA/dz \). In fact, the iron profiles at the deepest point of the lake close to Thalwil show such an increase in iron concentrations in the deepest part of the lake due to an iron release from the sediment. In these profiles there is, however, no indication for such a release in the upper parts of the lake. Figure 3.3 shows a comparison of time series of iron concentrations in 24 m depth at Stäfa (close to the sediment) and in 30 m at Thalwil (lake depth at Thalwil 137 m). This comparison shows that iron concentrations in Stäfa, where the measurement is close to the sediment, are often significantly higher than in Thalwil. Due to horizontal mixing, this horizontal inhomogenity is probably smoothed out very rapidly. Nevertheless, it demonstrates that the process of phosphate adsorption to iron hydroxide surfaces could also be of importance in the upper hypolimnion. The uptake rate contains a limiting term with respect to oxygen to account for the oxidation of the iron diffusing out of the sediment to iron hydroxide surfaces at which phosphate adsorbs.
Figure 3.3: Time series of total iron concentration in 24 m depth at Stäfa (lake bottom at 25 m) and in 30 m depth at Thalwil (lake bottom at 137 m) show a horizontal inhomogeneity with increased iron concentrations in Stäfa. This is probably caused by iron diffusing out of the sediment.

3.3.5 Model Implementation

The model was implemented using the simulation and data analysis software for aquatic systems, AQUASIM (Reichert 1994; Reichert 1998), version 2.1/3 (see http://www.aquasim.eawag.ch for topical details on this program). Within a spatial configuration consisting of compartments and links of the available types, this program allows its users to define any state variables and transformation processes and to perform simulations, parameter estimations and sensitivity analyses. In order to solve the partial differential equations (3.1) to (3.8) numerically, in a first step this program discretizes the space derivatives. In a second step, the spatially discretized partial differential equations are integrated in time with the implementation DASSL (Petzold 1983) of the implicit (backward differencing) variable-step, variable-order GEAR integration technique (Gear 1971). For the minimization of the sum of weighted squares of the deviations between model results and measurements required for parameter estimation, two numerical techniques were used. The simplex method of Nelder and Mead (1965) was used to find a direction of decrease of the sum of squares far away from the minimum and for parameter estimations with a large number of parameters, whereas the secant method ofRalston and Jennrich (1978) was used for final convergence of identifiable parameter sets. For both techniques the constrained minimization algorithms implemented in AQUASIM were used.

3.4 Results and Discussion

Compared to the complexity of the real system, the biogeochemical model presented in section 3.3 (Tables 3.2 to 3.4) is relatively simple. However, from the point of view of parameter identification, the model is rather complex. Obviously, it is not possible to determine the values of all model parameters with the monthly measured profiles of oxygen, nitrate, ammonia, phosphate, *Planktothrix (Oscillatoria) rubescens*, other algae and zooplankton.
For this reason, we started with literature values for all parameters and iteratively changed the values of those parameters that were assumed to be inaccurately known and had a significant influence on the model results. When the simulations led to qualitative agreement with the measurements, we started a formal sensitivity, identifiability and uncertainty analysis of the model based on the given layout of the measurement. This analysis is only briefly summarized here; a detailed description is given elsewhere (Omlin et al. 2000). The analysis was based on work by Brun et al. (2000) and consisted of the following steps:

- Based on prior knowledge and on the expected identifiability, model parameters were grouped into parameters potentially to be estimated from the data and parameters to be determined from other sources. Typically, kinetic parameters specific for the populations in Lake Zürich were potentially to be estimated, stoichiometric, physical and input parameters were assumed to be known from other sources (this distinction had to be made, because it is not possible to estimate input parameters, kinetic and stoichiometric parameters at the same time with the available data). Because the sharp spatial location of the *Planktothrix* (*Oscillatoria*) rubescens peaks, which often consisted of only one or two data points, kinetic *Planktothrix* (*Oscillatoria*) *rubescens* parameters were also excluded from the formal parameter estimation procedure, but they were adapted empirically.

- Based on prior estimates of parameter uncertainty and on linear error propagation of this uncertainty through the model, a sensitivity ranking of the model parameters according to their contribution to prediction uncertainty was made.

- Based on this sensitivity ranking and on information on potential identifiability problems gained by an analysis of approximate linear dependence of sensitivity functions (Belsley et al. 1980; Belsley 1991; Brun et al. 2000), parameter sets were selected for weighted least squares parameter fits based on data for the two years 1988 and 1989.

- The results were analysed for the quality of fit and for reasonableness of estimated parameter values and all the steps were iteratively repeated until convergence occurred (both sensitivity ranking and approximate linear dependence analysis of sensitivity functions are local analyses that depend on the selected parameter values).

This procedure finally led to the simulations and parameter sets shown and discussed in this paper. This seems to be a reasonable parameter set (comparing parameter values with the literature) and it leads to a relatively good approximation of the results with measurements (cf. subsection 3.4.1 and 3.4.2). However, this parameter set is not unique. More experimental investigations and comparisons with data from other lakes are necessary in order to find a more universal and unique parameter set.

In the first subsection (3.4.1), calculations based on these parameter values are compared with measurements (Figs. 3.4 to 3.9) in order to demonstrate the degree of agreement and to discuss the most important phenomena visible in the data. In the second subsection (3.4.2), sedimentation fluxes of organic particles and of particulate phosphorus are compared with data from an earlier measurement period (because no data on particle fluxes was available for the simulation period). In the third subsection (3.4.3), calculated conversion rates of oxygen, nitrate and phosphate are discussed. Finally, in subsection 3.4.4, by a comparison with simplified models, the necessity of using a variable stoichiometry of
primary production and of including a phosphate uptake process on sinking particles in
the model is justified.

3.4.1 Concentration Profiles

In this section, an overview is given of the results of the simulations for the years 1989 and
1990. These simulations were performed with the parameter values given in Tables 3.5 to
3.7 in the appendix. As described above, 1989 is the second year of the parameter fit and
1990 is the first year of extrapolation. For the year 1990, this is an extrapolation with
respect to the biogeochemical part of the model only. Input, light intensity and vertical
mixing parameters of the lake were determined as described in the sections 3.2.2 and 3.3.3
with measured data from the corresponding years.

Oxygen and Nutrients in the Water Column

Measured and simulated oxygen and nutrient profiles in the lake are compared in Figs. 3.4
to 3.6.

The oxygen data points in Fig. 3.4 show a strong decrease in oxygen concentration in
the metalimnion and close to the lake bottom while in the epilimnion oxygen concentra-
tions remain high. These effects are reproduced at least qualitatively by the simulation.
Systematic deviations of the calculation from the measurements in the year 1989 and in
the first few months of the year 1990 seem to be mainly caused by inaccuracies in the de-
scription of mixing processes during the winter (also in 1989 the oxygen is mixed down into
the deep hypolimnion during the first few months of the year). The metalimnic oxygen
minimum in the simulation is caused by respiration of sinking algae and by mineralization
of organic particles in the sediment at this depth. As diffusive vertical mixing is very small
in the metalimnion, these processes lead to a significant decrease in oxygen concentration.
Due to the large sediment surface at this depth, mineralization in the sediment is the
dominant cause of this oxygen depletion in the simulation.

The measured phosphate profiles show clearly the phosphate limitation of algal growth
during the summer (concentrations in the epilimnion are almost zero) and the significant
upward flux of phosphate released by mineralization in the sediment (increasing concen-
trations in the depth of the lake, especially in the second half of the year). During the
summer, phosphate concentrations are not only very low in the epilimnion, where phos-
phte is consumed by primary production, but also below in the metalimnion. This is an
indication for the significance of a phosphate uptake process on sinking particles (cf.
subsection 3.4.4 below). Simulated phosphate profiles represent all these effects and agree
very well with measurements in both years (Fig. 3.5). Again, the deviations between calcu-
lation and measurements in the first few months of both years indicate an overestimation
of mixing in the deep hypolimnion during the winter.
Figure 3.4: Simulation and data of monthly oxygen profiles in 1989 (second year of the fit, top) and 1990 (first year of extrapolation, bottom). Markers represent measurements, lines simulated concentrations.
Figure 3.5: Simulation and data of monthly phosphate profiles in 1989 (second year of the fit, top) and 1990 (first year of extrapolation, bottom). Markers represent measurements, lines simulated concentrations.
Figure 3.6: Simulation and data of monthly nitrate profiles in 1989 (second year of the fit, top) and 1990 (first year of extrapolation, bottom). Markers represent measurements, lines simulated concentrations.
The significant decrease of nitrate concentrations in the epilimnion during the summer is caused by incorporation of nitrogen into algal biomass and subsequent export by sedimentation. This phenomenon is qualitatively reproduced by the model (cf. subsection 3.4.4 below). However, there is a systematic deviation of the shape of the calculated nitrate profile from the measurements. The significant amount of denitrification in the sediment (cf. subsection 3.4.3) causes the calculated nitrate concentrations to decrease with increasing distance from the water surface. This phenomenon does not occur in the data, with exception of the deepest hypolimnion. The measured constant nitrate concentrations between 20 m and 120 m depth could only be reproduced with a significantly higher effective diffusion coefficient in this range of depth of the lake. Such an enlarged diffusion coefficient, however, would destroy the oxygen and phosphate profiles shown in Figs. 3.4 and 3.5.

Algae and Zooplankton

Measured and simulated profiles of algae and of *Planktothrix (Oscillatoria) rubescens* are compared in the Figs. 3.7 and 3.8. Fig. 3.9 shows a comparison of measured and simulated time series of zooplankton averaged over the epilimnion.

In early spring the stratification of the water column enables net growth of algae (Fig. 3.7) in the photic layer of the lake. After reaching a maximum in May 1989 and April 1990 respectively, the algae concentration is reduced drastically by grazing of zooplankton and recovers then during the following months. The simulation agrees very well with measurements, what is especially remarkable for the extrapolation in 1990.

*Planktothrix (Oscillatoria) rubescens* reaches its maximum growth rate at light intensities typical for the metalimnion and is usually light inhibited in the epilimnion. This leads to the sharp peaks visible in the measurements as well as in the simulation (Fig. 3.8). However, the seasonal variation of *Planktothrix (Oscillatoria) rubescens* is poorly described by the model. Whereas the simulations are in agreement with the simulation during the second half of both years, the observed high concentrations of *Planktothrix (Oscillatoria) rubescens* in May and June are not reproduced by the model.

The dynamics of the zooplankton population strongly depends on the algae concentration. The algae maximum in spring is followed by a maximum of zooplankton concentration (Fig. 3.9) that leads to a significant decrease in algae concentration. The order of magnitude of the zooplankton concentrations as well as the point in time of the first maximum is very well reproduced by the model. However, after the first maximum, measurements show a stronger variation with a significant decrease followed by a second peak. This dynamics is not reproduced by the simple zooplankton submodel selected for our model. A more detailed zooplankton submodel that distinguishes different classes of species and takes into account grazing of zooplankton by other species of zooplankton is probably necessary to describe the observations in more detail. Such a model should probably also consider more classes of algae and would therefore increase model complexity considerably.
Figure 3.7: Simulation and data of monthly algae profiles in 1989 (second year of the fit, top) and 1990 (first year of extrapolation, bottom). Markers represent measurements, lines simulated concentrations.
Figure 3.8: Simulation and data of monthly *Planktothrix (Oscillatoria) rubescens* profiles in 1989 (second year of the fit, top) and 1990 (first year of extrapolation, bottom). Markers represent measurements, lines simulated concentrations.
Figure 3.9: Simulation and data of zooplankton concentration in the epilimnion over the simulation period 1989-1990. Markers represent measurements for *Daphnia*, the line represents simulated concentrations.
3.4.2 Sedimentation Fluxes

Figure 3.10 shows a comparison of sedimentation fluxes for particulate organic carbon and for particulate phosphorus calculated for 1990 and measured by (Sigg et al. 1987) in 1984 (there are no data available for other years). These data have not directly been used for model calibration, however, the ratio of the fluxes for organic carbon and for phosphorus has been used for the selection of the value of the parameter $a_{P,\text{max}}$. The comparison shows that the qualitative behaviour of these fluxes is represented correctly by the model. Especially, the interannual variation is represented very well. The fluxes seem to be somewhat larger in the simulation for 1990 than measured in 1984. However, the difference is not larger than intraannual differences can be expected to be.

3.4.3 Oxygen, Nitrate and Phosphate Conversion Rates

Figure 3.11 shows calculated conversion rates of oxygen, nitrate and phosphate in September 1989 in the water column and in the upper sediment layer (the conversion rates in the lower sediment layer were much smaller). For better comparability, the conversion rates in the sediment layer have been converted to water column-equivalent rates.

The conversion rates of oxygen in the water column show a maximum of production in the epilimion (significantly larger than the scale of the diagram) and a smaller max-
minimum in about 12 m depth in the metalimnion. This lower maximum is caused by the growth of *Planktothrix* (*Oscillatoria*) rubescens, whereas the upper maximum is caused by the growth of the other algae. Everywhere else, oxygen production is negative due to respiration, grazing, mineralization and nitrification. The sediment layer leads to a higher contribution to oxygen consumption than the water column (mainly due to mineralization).

The nitrate conversion rates show an intensive nitrate consumption due to primary production in the epilimnion and a smaller consumption due to *Planktothrix* (*Oscillatoria*) rubescens growth in a depth of about 12 m. In the rest of the water column, nitrate is produced by nitrification of ammonia diffusing out of the sediment, where it is produced by aerobic and anoxic mineralization. In a more detailed model with a more accurate resolution of the sediment, part of the nitrification shown here in the water column (because there is not enough oxygen in the upper sediment layer) could also occur in the highest mm of the sediment. In the sediment layer, nitrate is denitrified by anoxic mineralization.

Finally, phosphate is consumed in the water column by primary production in the epilimnion and by phosphate uptake in the hypolimnion. It is produced in the sediment layer by aerobic and anoxic mineralization.

### 3.4.4 Variable Stoichiometry and Phosphate Uptake

Figure 3.12 shows a comparison of measured concentration profiles of nitrate and phosphate during a typical summer situation (September 1989) with simulations for the model described in this paper, for a simplified model without phosphorus uptake on sinking particles, and for an even simpler model without phosphorus uptake and with a constant (Redfield) stoichiometry of algal growth. The equations of the simplified models were the same as those of the full model with exception of the changes mentioned above and the values of the maximum specific growth rates of algae and zooplankton and the sediment.
Figure 3.12: Comparison of the resulting phosphate and nitrate profiles in September 1989 for the full model described in this paper (black lines), for a model with omission of the phosphate uptake process on sinking particles (dark grey lines) and for a model with omission of the phosphate uptake process and with a constant (Redfield) stoichiometry of algal growth (light grey lines).

layer thickness for which a new fit has been done to allow for an optimal correspondence with the measurements.

Omission of the phosphate uptake process does only insignificantly change the nitrate profiles (the changes are mainly due to the changes in model parameters during the fit). However, the phosphate profiles calculated with this model show two significant deficits: First, the thickness of the surface layer within which phosphate concentrations become very low is significantly too small and second, the gradient of the phosphate profile in the hypolimnion is too small as well. The first observation is a result of the fact that phosphate is only incorporated into algal biomass within the photic zone of the lake and that there is no other significant process that consumes phosphate below if the uptake process is turned off. The second observation indicates a too small upward flux of phosphate released from the sediment. The introduction of the phosphate uptake process on sinking particles solves both problems: First, the uptake process decreases the concentrations below the epilimnion, and second, due to the increased downward flux of phosphate attached to sinking particles, a higher phosphate release from the sediment is possible (in the model the attached phosphate is released at the same specific rate as the organic material is mineralized).

The phosphate profiles of the model that, in addition to the omission of the phosphorus uptake process, keeps the phosphorus content of growing algae constant are only insignificantly different from the results of the model that only turns off the uptake process. However, due to the higher phosphorus content of growing algae, much less algal biomass can be produced during this phosphorus limited period. This leads to a significantly smaller reduction of nitrate concentration in the epilimnion. In the model with a variable stoichiometry that decreases the phosphorus content of newly built algae during phosphorus limited periods, much more biomass can be produced under phosphorus limiting conditions. This leads to the observed higher nitrate uptake in the epilimnion. A variability of the phosphorus content of algae as implemented in the model is confirmed by
more direct measurements of the composition of algae in other lakes (Hupfer et al. 1995).

3.5 Conclusions

A biogeochemical lake model was presented and applied to data from Lake Zürich, Switzerland. The one-dimensional model resolves the depth of the lake and calculates nutrient, oxygen and plankton dynamics averaged over daily fluctuations. The model considers growth, respiration and death of two classes of algae and one class of zooplankton, mineralization, nitrification, mixing, sedimentation and phosphate uptake on sinking particles. Due to the consideration of a roughly discretized sediment layer, it is able to describe the accumulation of organic matter in the sediment and the succeeding release of nutrients originating from mineralization.

After the calibration of mainly kinetic model parameters (cf. Omlin et al. (2000) for more details of the applied calibration, identifiability and uncertainty analysis procedures), the model was able to reproduce the essential properties of the nutrient, oxygen and plankton dynamics in the lake over several years (cf. section 3.4.1). Also the particulate fluxes of organic carbon and of phosphorus were in agreement with measurements (cf. section 3.4.2). In order to achieve these results, a variable stoichiometry of primary production with respect to phosphate was necessary, so that the phosphorus content of newly built algae can decrease when phosphate gets limited. Moreover, consideration of a phosphate uptake process on sinking particles below the photic zone was necessary (cf. section 3.4.4). The predictions of the model to the years after the calibration period were also very close to the measurements. This is not as astonishing as it may seem, because the external driving conditions and therefore also the seasonal pattern of concentrations did not significantly change during this time.

The model was used to calculate conversion rates of oxygen, nitrate and phosphate that are not directly measured (cf. section 3.4.3). These results are very useful for understanding the relative significance of different processes in the model and, with the aid of additional, more specific measurements, to improve our understanding of nutrient dynamics in the lake.

Although the model results look very promising, further applications of the model to different periods of Lake Zürich and to other lakes are necessary to improve the predictive capabilities of the model also for situations with significant changes in driving conditions.

Acknowledgments

We thank Sonja Gammeter and Ulrich Zimmerman from Wasserversorgung Zürich, Switzerland, Colin Reynolds from Windermere Laboratory, Ambleside, United Kingdom, and Hans Rudolf Bürgi, Maria Dittrich, René Gächter, Gerrit Goudsmit, Dieter Imboden, Frank Peeters, Laura Sigg, Bernhard Wehrli and Alfred Wüest from EAWAG, Switzerland, for many helpful discussions and comments.

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.
3.6 Appendix

This appendix contains the values of the model parameters used for the final simulation shown in this paper. For each parameter, value and unit is given, and, instead of a verbal description, a reference to the section, equation or table, where its meaning can be found. A summary of the origin of the parameter values is given in the text.

In addition to the parameters given in this appendix, initial values of state variables and time series for external variables (inflows, light at the water surface, water temperature, diffusivities) had to be provided in order to be able to perform a simulation. The origin of these data is discussed in section 3.2.2.

Table 3.5 contains the values of the physical model parameters. Due to the very rough discretization of the sediment, the thickness of the two sediment layers, \( h_{\text{sed}} \), is an empirical parameter, which was fitted. The values of the light extinction parameters, \( k_1 \) and \( k_2 \), were fitted independently of the model on raw data from WVZ. The value of the oxygen exchange velocity, \( v_{\text{O}_2, \text{atm}} \), was selected to be in the range used for other lake simulations in Switzerland (Imboden and Gächter 1978; Karagounis et al. 1993). Finally, information on sedimentation velocity was gained from Imboden and Gächter (1978), Karagounis et al. (1993), Andersen (1997) and Lampert and Sommer (1997).

Table 3.6 contains the values of the stoichiometric model parameters. Basic nitrogen and phosphorus fractions, \( a_N \) and \( a_{P, \text{red}} \), were based on the mass composition according to Redfield et al. (1966). Information on \( a_{P, \text{max}} \) came from Sigg et al. (1987) and Hupfer et al. (1995). The parameters for variable phosphorus stoichiometry, \( b_{P, \text{min}} \) and \( b_{P, \text{max}} \), were selected using Hupfer et al. (1995) and Andersen (1997). The parameter \( c_v \) was adjusted during preliminary simulations. The mass fraction of inert organic matter during death processes, \( f_p \), was selected to have a similar value as in activated sludge sewage treatment models (Henze et al. 1986). The critical value for switching from Redfield

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Ref.</th>
<th>Parameter</th>
<th>Value</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>( h_{\text{sed}} )</td>
<td>0.0036 m</td>
<td>eqs. (3.1,3.3-3.6)</td>
<td>( v_{\text{sed}, \text{ORG}} )</td>
<td>10 md(^{-1} )</td>
<td>eqs. (3.2,3.5,3.7)</td>
</tr>
<tr>
<td>( k_1 )</td>
<td>0.31 m(^{-1} )</td>
<td>eq. (3.9)</td>
<td>( v_{\text{sed}, \text{OSC}} )</td>
<td>0 md(^{-1} )</td>
<td>eqs. (3.2,3.5,3.7)</td>
</tr>
<tr>
<td>( k_2 )</td>
<td>0.026 gWM(^{-1} )m(^2 )</td>
<td>eq. (3.9)</td>
<td>( v_{\text{sed}, \text{ZOO}} )</td>
<td>0 md(^{-1} )</td>
<td>eqs. (3.2,3.5,3.7)</td>
</tr>
<tr>
<td>( v_{\text{O}_2, \text{atm}} )</td>
<td>1 m(^{-1} )</td>
<td>sect. 3.3.3</td>
<td>( \theta )</td>
<td>0.95</td>
<td>eqs. (3.3,3.4)</td>
</tr>
<tr>
<td>( v_{\text{sed}, \text{ALG}} )</td>
<td>0.2 m(^{-1} )</td>
<td>eqs. (3.2,3.5,3.7)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.5: Physical parameters used in the lake model.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Ref.</th>
<th>Parameter</th>
<th>Value</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>( a_N )</td>
<td>0.063</td>
<td>tabs. 3.2,3.3</td>
<td>( S_{\text{HPO}_4, \text{crit}} )</td>
<td>0.0042 gPm(^{-3} )</td>
<td>tab. 3.3</td>
</tr>
<tr>
<td>( a_{P, \text{max}} )</td>
<td>0.009</td>
<td>tab. 3.4</td>
<td>( w_{\text{ALG}} )</td>
<td>5</td>
<td>sect. 3.2.2</td>
</tr>
<tr>
<td>( a_{P, \text{red}} )</td>
<td>0.0087</td>
<td>tabs. 3.2,3.3</td>
<td>( w_{\text{ORG}} )</td>
<td>5</td>
<td>sect. 3.2.2</td>
</tr>
<tr>
<td>( b_{P, \text{min}} )</td>
<td>0.0014</td>
<td>tab. 3.3</td>
<td>( w_{\text{ZOO}} )</td>
<td>5</td>
<td>sect. 3.2.2</td>
</tr>
<tr>
<td>( b_{P, \text{max}} )</td>
<td>0.0087</td>
<td>tab. 3.3</td>
<td>( \Delta S_{\text{HPO}_4} )</td>
<td>0.0013 gPm(^{-3} )</td>
<td>tab. 3.3</td>
</tr>
<tr>
<td>( c_v )</td>
<td>0.7</td>
<td>tab. 3.3</td>
<td>( Y_{\text{ZOO,max}} )</td>
<td>0.5</td>
<td>tab. 3.3</td>
</tr>
<tr>
<td>( f_p )</td>
<td>0.1</td>
<td>tabs. 3.2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.6: Stoichiometric parameters used in the lake model.
stoichiometry to reduced phosphorus stoichiometry, $S_{HPO_4,\text{crit}}$, was fitted. The selection of the $\Delta S_{HPO_4}$ was influenced by the fitted value of $S_{HPO_4,\text{crit}}$ and by personal discussions with C. Reynolds. Finally, the values of the conversion factors between dry mass and wet mass, $w$, and the maximum yield of zooplankton, $Y_{\text{ZOO, max}}$, was selected using Andersen (1997) and personal information from C. Reynolds.

Table 3.7 contains the values of the kinetic model parameters. The parameters $k$, in the left column, are maximum conversion rates at $T_0 = 20^\circ\text{C}$, the parameters $K$, in the right column, are half-saturation concentrations, and the parameters $\beta$, in the right column, are temperature dependence coefficients. Information on the values of these parameters came from many different sources. The parameters $k_{\text{death},\text{ZOO},T_0}$, $k_{\text{pro},\text{ZOO},T_0}$, $K_{\text{HPO}_4,\text{ALG}}$ and $K_{\text{HPO}_4,\text{ALG}}$ were fitted (cf. Omilin et al. (2000) for details of fit parameter selection). The parameters for Oscillatora rubescens growth, $k_{\text{death},\text{PLR},T_0}$, $K_{\text{ALG},\text{PLR}}$, $k_{\text{resp},\text{PLR},T_0}$ and $K_{\text{PLR},\text{PLR}}$ were adjusted manually due to problems with the sharp peaks in the fit. Information on other kinetic parameters came from Imboden and Gächter (1978), Karagounis et al. (1993), Cooper (1984), Henze et al. (1986), Andersen (1997) and Lampert and Sommer (1997). Temperature dependence for algae and bacteria was taken from Brown and Barnwell (1987) and for zooplankton from personal communication with H.R. Bürgi.
Chapter 4

Biogeochemical Model of Lake Zürich:
Sensitivity, Identifiability and Uncertainty Analysis

A model for the description of nutrient, oxygen and plankton dynamics in Lake Zürich, Switzerland has recently been developed. Because with this model the attempt is made to mechanistically describe 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 its parameters to be identifiable from the monthly measured profiles. This raises the question 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 fit with measured data, to analyze the effect of other, fixed parameters on the estimates of the selected parameters, and to estimate the uncertainty of model predictions.

Keywords: Lake Zürich; water quality modelling; parameter estimation; sensitivity analysis; dependence analysis; identifiability analysis; uncertainty analysis.
4.1 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 changing environmental conditions (Beck 1987; Reichert and Omlin 1997; Beck 1999; 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 not only include 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 Bruns et al. (2000) are applied to the biogeochemical model of Lake Zürich described in Omlin et al. (2000). The 52 parameters listed in the appendix of Omlin et al. (2000) can certainly not 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 adequateness of model complexity for the goal of describing nutrient, oxygen and plankton dynamics in the lake should be addressed. Table 4.1 summarizes state variables and processes used in the lake model. A mechanistic model for the above-mentioned purpose obviously requires the state variables $S_{HFO}$, $S_{NH_4}$, $S_{NO_3}$, $S_{O_2}$, $X_{ALG}$, and $X_{ZOO}$ (we think that a description of the macroscopic mechanisms is mandatory for having a chance to predict changes in system variables induced by changes in external driving forces). $X_I$ is necessary for modeling permanent storage of organic material in the sediment, $X_S$ is necessary for the consideration of input of organic material, for modeling fecal pellets and it is useful for summarizing dead algae and zooplankton. It has been shown in Omlin et al. (2000) that a variable stoichiometry and an uptake process of phosphate on sinking particles is necessary to describe the nutrient dynamics correctly. This explains the necessity for all phosphorus-related state variables. $X_{PLR}$ and $X_{P,PLR}$ remain the only state variables that could be omitted in a simplified description of the biogeochemical processes in the lake. Indeed the blue-green alga Planktothrix (Oscillatoria) rubescens has not been included because of its importance on the nutrient conversion processes but because it is of special interest for the local water supply authority. The processes of growth, mineralization and nitrification are obviously required in order to describe nutrient conversions.
State Variables

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description (Concentration of ...)</th>
<th>Processes</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_{HPO_4}$</td>
<td>phosphate</td>
<td>Growth of algae</td>
</tr>
<tr>
<td>$S_{NH_4}$</td>
<td>ammonia</td>
<td>Growth of Planktothrix rubescens</td>
</tr>
<tr>
<td>$S_{NO_3}$</td>
<td>nitrate</td>
<td>Growth of zooplankton</td>
</tr>
<tr>
<td>$S_{O_2}$</td>
<td>oxygen</td>
<td>Respiration of Algae</td>
</tr>
<tr>
<td>$X_{ALG}$</td>
<td>algae without Planktothrix rubescens</td>
<td>Respiration of Planktothrix rubescens</td>
</tr>
<tr>
<td>$X_{PLR}$</td>
<td>Planktothrix rubescens</td>
<td>Respiration of zooplankton</td>
</tr>
<tr>
<td>$X_{ZOO}$</td>
<td>zooplankton</td>
<td>Death of algae</td>
</tr>
<tr>
<td>$X_S$</td>
<td>biodegradable organic material</td>
<td>Death of Planktothrix rubescens</td>
</tr>
<tr>
<td>$X_I$</td>
<td>inert organic material</td>
<td>Death of zooplankton</td>
</tr>
<tr>
<td>$X_{P,ALG}$</td>
<td>organic phosphorus in $X_{ALG}$</td>
<td>Aerobic mineralization</td>
</tr>
<tr>
<td>$X_{P,PLR}$</td>
<td>organic phosphorus in $X_{PLR}$</td>
<td>Anoxic mineralization</td>
</tr>
<tr>
<td>$X_{P,S}$</td>
<td>organic phosphorus in $X_S$</td>
<td>Nitrification</td>
</tr>
<tr>
<td>$X_{P,I}$</td>
<td>organic phosphorus in $X_I$</td>
<td>Phosphate-uptake of sinking particles</td>
</tr>
<tr>
<td>$X_{P,I,S}$</td>
<td>phosphate attached to $X_S$</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.1: State variables (S=dissolved substances, X=particulate substances) and Processes used in the lake model described in Omlin et al. 2000.

Phosphate-uptake by sinking particles has already been mentioned to be necessary to explain the observed phosphate profiles. The potential for model simplifications is therefore limited to a combination of the mineralization process with respiration and death. However, the proposed approach that considers these processes separately is conceptually clearer and avoids the necessity for depth-dependent parameters as used in previous studies (Imboden and Gächter 1978; Karagounis et al. 1993). 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 identifiability of subsets of its parameters is an important task to be addressed in this paper.

This paper presents the results of the application of the procedure described in Brun et al. (2000) to the lake model mentioned above. The sections 4.3 and 4.4 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 4.2 a brief overview of the techniques used for parameter estimation is given. This section also introduces some aspects of notation. In section 4.3 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 4.4, 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 4.5 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 4.6 the results
are summarized and conclusions are drawn.

4.2 Parameter Estimation

The selection of the parameter subset to be estimated from the lake data is discussed in section 4.4. 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

\[ WSS(\theta) = \sum_{k=1}^{n_k} \sum_{i=1}^{n_{yi}} \sum_{j=1}^{n_{yji}} \left( \frac{y_{meas, k, j, i} - y_k(z_{yji}, t_j, \theta)}{sc_{y_k}} \right)^2 \]  

(4.1)

in order to find the parameter estimates. In this equation, the index \( i \) is used to distinguish spatial locations, \( z_{yji} \) (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 \( y_k \), \( n_{yi} \). 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 lake profiles were measured, \( n_{t} \). The index \( k \) is used to distinguish different model variables, \( y_k \); it runs from 1 to the number of model variables to be compared with data, \( n_k \). \( y_k(z_{yji}, t_j, \theta) \) is the result of the lake model for the model variable \( y_k \) calculated with the parameter values \( \theta \) and evaluated at location \( z_{yji} \) and at time \( t_j \). \( y_{meas, k, j, i} \) is the measured value corresponding to the model variable \( y_k \) sampled at location \( z_{yji} \) at time \( t_j \). In the sum of weighted squares (4.1) the residuals, \( y_{meas, k, j, i} - y_k(z_{yji}, t_j, \theta) \), 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 (4.1) nondimensional. 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 \( y_k \), \( \sigma_{y_k} \), then the sum WSS is equal to \( \chi^2 \). 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 4.2 lists the values of \( sc_{y_k} \) used for the fit. Profiles for the five model variables oxygen, phosphate, nitrate, algae (other than Planktothrix (Oscillatoria) rubescens) and zooplankton were used for parameter estimation. Data for Planktothrix (Oscillatoria) rubescens, which was

<table>
<thead>
<tr>
<th>Scale</th>
<th>Value</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>( sc_{SO_2} )</td>
<td>1.0</td>
<td>gO/m^-3</td>
</tr>
<tr>
<td>( sc_{PO_4} )</td>
<td>0.025</td>
<td>gP/m^-3</td>
</tr>
<tr>
<td>( sc_{NO_3} )</td>
<td>0.1</td>
<td>gN/m^-3</td>
</tr>
<tr>
<td>( sc_{X ALG} )</td>
<td>0.5</td>
<td>gWM/m^-3</td>
</tr>
<tr>
<td>( sc_{X ZOO} )</td>
<td>0.2</td>
<td>gWM/m^-3</td>
</tr>
</tbody>
</table>

Table 4.2: Scales used as empirical weights for making the contributions of different model variables to the weighted sum of squares (4.1) nondimensional.
also available, was not used for the fit because of difficulties with the spatial resolution of the sharp *Planktothrix (Oscillatoria)* rubescens peaks in calculation and measurements. Model parameters used for the description of *Planktothrix (Oscillatoria)* rubescens were adjusted by hand, instead. The values shown in Table 4.2 reflect the sharp metalimnic minimum in oxygen concentrations which had to be compensated for by a large value for $s_{G_O2}$ and the smaller number of zooplankton measurements in comparison to the algae measurements that were compensated by a smaller value of $s_{G_ALG}$ in comparison to $s_{G_ALG}$ (zooplankton measurements were spatially integrated over epi- and hypolimnion, whereas algae were sampled with a higher spatial resolution).

### 4.3 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, this 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. (2000); it focuses on the local sensitivity of model results to model parameters. A similar ranking was done by Nordhaus (1995) based on the nonlinear 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. (2000)). Because of the uncertainty in the rate expressions and in the parameter values of the biogeochemical model, it seemed not to be meaningful to estimate inflow parameters from lake data (inputs to a lake are often not measured precisely, however, their estimation from lake data would only be meaningful if the biogeochemical processes in the lake would be 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 primarily estimate 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 significantly be 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 4.3,
4.4 and 4.5 show the selected uncertainty class of all model parameters for 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.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Unc. Class</th>
<th>Value</th>
<th>Unc. Range</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h_{sed}$</td>
<td>3</td>
<td>0.0036</td>
<td>0.0018</td>
<td>m</td>
</tr>
<tr>
<td>$S_{HPO_4, crit}$</td>
<td>2</td>
<td>0.0042</td>
<td>0.0008</td>
<td>gPm$^{-3}$</td>
</tr>
<tr>
<td>$\Delta S_{HPO_4}$</td>
<td>2</td>
<td>0.0013</td>
<td>0.0003</td>
<td>gPm$^{-3}$</td>
</tr>
<tr>
<td>$k_{death, ALG, T}$</td>
<td>3</td>
<td>0.030</td>
<td>0.015</td>
<td>d$^{-1}$</td>
</tr>
<tr>
<td>$k_{death, ZOO, T}$</td>
<td>3</td>
<td>0.029</td>
<td>0.014</td>
<td>d$^{-1}$</td>
</tr>
<tr>
<td>$k_{gro, ALG, T}$</td>
<td>2</td>
<td>1.13</td>
<td>0.23</td>
<td>d$^{-1}$</td>
</tr>
<tr>
<td>$k_{gro, ZOO, T}$</td>
<td>2</td>
<td>0.30</td>
<td>0.06</td>
<td>gDM$^{-1}$m$^3$d$^{-1}$</td>
</tr>
<tr>
<td>$k_{mineral, aero, sed, T}$</td>
<td>3</td>
<td>0.10</td>
<td>0.05</td>
<td>d$^{-1}$</td>
</tr>
<tr>
<td>$k_{mineral, aero, wat, T}$</td>
<td>3</td>
<td>0.010</td>
<td>0.005</td>
<td>d$^{-1}$</td>
</tr>
<tr>
<td>$k_{mineral, ammo, sed, T}$</td>
<td>3</td>
<td>0.10</td>
<td>0.05</td>
<td>d$^{-1}$</td>
</tr>
<tr>
<td>$k_{mineral, ammo, wat, T}$</td>
<td>3</td>
<td>0.010</td>
<td>0.005</td>
<td>d$^{-1}$</td>
</tr>
<tr>
<td>$k_{nitril, sed, T}$</td>
<td>3</td>
<td>0.50</td>
<td>0.25</td>
<td>gNm$^{-3}$d$^{-1}$</td>
</tr>
<tr>
<td>$k_{nitril, wat, T}$</td>
<td>3</td>
<td>0.10</td>
<td>0.05</td>
<td>gNm$^{-3}$d$^{-1}$</td>
</tr>
<tr>
<td>$k_{resp, ALG, T}$</td>
<td>3</td>
<td>0.050</td>
<td>0.025</td>
<td>d$^{-1}$</td>
</tr>
<tr>
<td>$k_{resp, ZOO, T}$</td>
<td>3</td>
<td>0.0030</td>
<td>0.0015</td>
<td>d$^{-1}$</td>
</tr>
<tr>
<td>$K_{opt}$</td>
<td>3</td>
<td>1200</td>
<td>600</td>
<td>gDM$^{-1}$m$^4$d$^{-1}$</td>
</tr>
<tr>
<td>$K_I, ALG$</td>
<td>3</td>
<td>34</td>
<td>17</td>
<td>Wm$^{-2}$</td>
</tr>
<tr>
<td>$K_{NH_4, nitril}$</td>
<td>3</td>
<td>0.50</td>
<td>0.25</td>
<td>gNm$^{-3}$</td>
</tr>
<tr>
<td>$K_{NO_3, ALG}$</td>
<td>3</td>
<td>0.040</td>
<td>0.020</td>
<td>gNm$^{-3}$</td>
</tr>
<tr>
<td>$K_{NO_3, mineral}$</td>
<td>3</td>
<td>0.10</td>
<td>0.05</td>
<td>gNm$^{-3}$</td>
</tr>
<tr>
<td>$K_{O_2, opt}$</td>
<td>3</td>
<td>0.50</td>
<td>0.25</td>
<td>gOm$^{-3}$</td>
</tr>
<tr>
<td>$K_{O_2, mineral}$</td>
<td>3</td>
<td>0.20</td>
<td>0.10</td>
<td>gOm$^{-3}$</td>
</tr>
<tr>
<td>$K_{O_2, nitril}$</td>
<td>3</td>
<td>0.40</td>
<td>0.20</td>
<td>gOm$^{-3}$</td>
</tr>
<tr>
<td>$K_{O_2, resp}$</td>
<td>3</td>
<td>0.50</td>
<td>0.25</td>
<td>gOm$^{-3}$</td>
</tr>
<tr>
<td>$K_{HPO_4, ALG}$</td>
<td>3</td>
<td>0.0019</td>
<td>0.0009</td>
<td>gPm$^{-3}$</td>
</tr>
</tbody>
</table>

Table 4.3: Potential fit parameters with uncertainty class, parameter estimate and prior estimate of uncertainty range (depending on uncertainty class $1=5\%, 2=20\%$ and $3=50\%$). The horizontal lines separate physical, stoichiometric and kinetic parameters.

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 4.3 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 critical concentration for the transition and $\Delta S_{HPO_4}$ 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 4.3 are kinetic parameters. In fact, these are all kinetic parameters of the model with exception of the kinetic parameters for *Planktothrix* (*Oscillatoria*) *rubescens* which were excluded because of the problems with the insufficient spatial resolution for fit mentioned in section 4.2, and with exception of the temperature dependence coefficients of the microbiological processes. All parameters $k$ are rate constants directly proportional to the process rate, all parameters $K$ are half-saturation concentrations of processes with respect to limiting light intensities, oxygen concentrations or nutrient concentrations, and the indices death, gro, miner, nitr, resp and upt, refer to death, growth, mineralization, nitrification, respiration and phosphate uptake processes, respectively (Omlin et al. 2000). 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 4.4 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
Table 4.5: Input parameters with uncertainty class, parameter estimate and prior estimate of uncertainty range (depending on uncertainty class 1=5%, 2=20% and 3=50%).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Unc. Class</th>
<th>Value $\theta_1$</th>
<th>Unc. Range $\Delta \theta_1$</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_{P,\text{inflow}}$</td>
<td>2</td>
<td>0.0087</td>
<td>0.0017</td>
<td></td>
</tr>
<tr>
<td>$f_{NH_4\text{,ana}}$</td>
<td>3</td>
<td>0.50</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>$f_{NH_4\text{,rain}}$</td>
<td>3</td>
<td>0.50</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>$f_{X_1\text{,dam}}$</td>
<td>3</td>
<td>0.20</td>
<td>0.10</td>
<td></td>
</tr>
<tr>
<td>$f_{X_1\text{,river}}$</td>
<td>3</td>
<td>0.20</td>
<td>0.10</td>
<td></td>
</tr>
<tr>
<td>$F_{HPO_4}$</td>
<td>2</td>
<td>1.0</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>$F_{X_1}$</td>
<td>2</td>
<td>1.0</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>$F_N$</td>
<td>2</td>
<td>1.0</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>$F_{K_S}$</td>
<td>2</td>
<td>1.0</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>$F_{O_2}$</td>
<td>2</td>
<td>1.0</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>$F_{X\text{FOR}}$</td>
<td>2</td>
<td>1.0</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>$F_Q$</td>
<td>2</td>
<td>1.0</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>$F_{X_S}$</td>
<td>2</td>
<td>1.0</td>
<td>0.2</td>
<td></td>
</tr>
</tbody>
</table>

across the lake surface was also classified into class 2. The sedimentation velocities of algae and dead organic particles, $v_{\text{sed,ALG}}$ and $v_{\text{sed,ORG}}$, are very inaccurate because, depending on the nature of the particles, a wide range of sedimentation velocities has been measured. To describe 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 $a_{P,\text{red}}$), maximum fraction of adsorbed phosphate on organic particles ($a_{P,\text{max}}$), maximum and minimum phosphorus content of newly built algae during primary production ($b_{P,\text{max}}$ and $b_{P,\text{min}}$), coefficients for excretion ($e_e$) and for build-up of inert organic material during death ($f_p$), coefficients for the conversion of dry mass to wet mass ($w_{\text{ALG}}$, $w_{\text{ORG}}$, $w_{\text{ZOO}}$), and maximum yield of zooplankton growth ($Y_{\text{ZOO,max}}$). The kinetic parameters not to be fitted are those related to growth, respiration and death of Planktothrix (Oscillatoria) rubescens ($k_{\text{death,PLR,T_0}}$, $k_{\text{pro,PLR,T_0}}$, $K_{I,PLR}$ and $k_{\text{resp,PLR,T_0}}$), and coefficients for temperature dependence of growth and respiration processes of algae, bacteria, Planktothrix (Oscillatoria) rubescens and zooplankton ($\beta_{\text{ALG}}$, $\beta_{\text{BAC}}$, $\beta_{\text{PLR}}$ and $\beta_{\text{ZOO}}$).

Table 4.5 lists the input parameters for lake simulations. Similarly to the parameters listed in Table 4.4, these parameters are not to be estimated from lake data. The parameter $a_{P,\text{inflow}}$ describes the phosphorus content of organic particles discharged into the lake. The parameters $f$ describe fractions of dissolved inorganic nitrogen in the 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 factors $F$ 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 $\Delta \theta_1$ of the parameter $\theta_1$ to the uncertainty of the result of the model variable $y_k$ calculated
at time $t$ at position $z$ is given as

$$\Delta y_{k,\theta}(z, t; \theta) = \Delta \theta_i \cdot \frac{\partial y_k}{\partial \theta_i}(z, t) \quad . \quad \text{(4.2)}$$

This uncertainty measure has the same dimension as the model variable $y_k$. It can be made nondimensional by division with the scale, $se_{y_k}$, introduced in section 4.2 (for numerical values see Table 4.2). 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:

$$\delta^{\text{msgqr}}(\theta) = \sqrt{\frac{1}{n_{\text{tot}}} \sum_{k=1}^{n_y} \sum_{j=1}^{n_z} \sum_{i=1}^{n_{y;i}} \left( \frac{\Delta \theta_i}{se_{y_k}} \cdot \frac{\partial y_k}{\partial \theta_i}(z_{y_k,i}, t_j; \theta) \right)^2} \quad . \quad \text{(4.3)}$$

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, $se_{y_k}$, for making the dimensional contributions nondimensional as are used for making the residuals in equation (4.1) nondimensional, this sensitivity measure is compatible with the parameter estimation procedure. With the notation

$$s_{\{k,i,j\},i}(\theta) = \frac{\Delta \theta_i \cdot \partial y_k}{\Delta y_k \cdot \partial \theta_i}(z_{y_k,i}, t_j; \theta) \quad \text{(4.4)}$$

and

$$s_i(\theta) = (s_{\{1,1,1\},1}(\theta), \ldots, s_{\{n_y, n_y, z, n_z\},1}(\theta))^T \quad \text{(4.5)}$$

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

$$\delta^{\text{msgqr}}(\theta) = \frac{1}{\sqrt{n}} \| s_i(\theta) \| \quad \text{(4.6)}$$

as introduced in Brun et al. (2000).

Table 4.6 shows a ranking of the sensitivities of the model results calculated according to equation (4.3) 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 4.6 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 ten potential fit parameters the sensitivity decreases by roughly a factor of ten). In addition, the results shown in this table do not reflect 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.
<table>
<thead>
<tr>
<th>Pot. Fit Parameter</th>
<th>$\delta_{n}^{s_{m}}$</th>
<th>Fixed Parameter</th>
<th>$\delta_{n}^{s_{m}}$</th>
<th>Input Parameter</th>
<th>$\delta_{n}^{s_{m}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K_{1 \text{, ALG}}$</td>
<td>1.207</td>
<td>$k_{1}$</td>
<td>0.559</td>
<td>$F_N$</td>
<td>0.428</td>
</tr>
<tr>
<td>$k_{\text{death, ZOO, } T_0}$</td>
<td>0.843</td>
<td>$v_{\text{sed, ALG}}$</td>
<td>0.351</td>
<td>$F_Q$</td>
<td>0.304</td>
</tr>
<tr>
<td>$k_{\text{growth, ZOO, } T_0}$</td>
<td>0.828</td>
<td>$k_{\text{death, PLRT, } T_0}$</td>
<td>0.348</td>
<td>$F_{\text{HPO}_4}$</td>
<td>0.222</td>
</tr>
<tr>
<td>$k_{\text{resp, ALG, } T_0}$</td>
<td>0.763</td>
<td>$\beta_{\text{ZOO}}$</td>
<td>0.312</td>
<td>$F_{K_n}$</td>
<td>0.188</td>
</tr>
<tr>
<td>$k_{\text{death, ALG, } T_0}$</td>
<td>0.711</td>
<td>$\theta$</td>
<td>0.304</td>
<td>$v_{\text{O}_2 \text{, atm }}$</td>
<td>0.181</td>
</tr>
<tr>
<td>$K_{\text{HPO}_4, \text{ ALG}}$</td>
<td>0.596</td>
<td>$\delta_{\text{ALG}}$</td>
<td>0.304</td>
<td>$F_{\text{XO}}$</td>
<td>0.091</td>
</tr>
<tr>
<td>$S_{\text{HPO}_4, \text{ crit}}$</td>
<td>0.502</td>
<td>$w_{\text{ALG}}$</td>
<td>0.302</td>
<td>$F_{O_2}$</td>
<td>0.077</td>
</tr>
<tr>
<td>$h_{\text{sed}}$</td>
<td>0.429</td>
<td>$k_{\text{growth, PLRT, } T_0}$</td>
<td>0.268</td>
<td>$F_{\text{XO}_2}$</td>
<td>0.064</td>
</tr>
<tr>
<td>$\Delta S_{\text{HPO}_4}$</td>
<td>0.300</td>
<td>$b_{\text{p, min}}$</td>
<td>0.236</td>
<td>$\alpha_{\text{P, inflow}}$</td>
<td>0.051</td>
</tr>
<tr>
<td>$\beta_{\text{opt}}$</td>
<td>0.169</td>
<td>$\alpha_{\text{P, red}}$</td>
<td>0.176</td>
<td>$f_{\text{XO, dam}}$</td>
<td>0.030</td>
</tr>
<tr>
<td>$k_{\text{miner, aero, sed, } T_0}$</td>
<td>0.199</td>
<td>$\alpha_{\text{P, max}}$</td>
<td>0.147</td>
<td>$f_{\text{XO, iners}}$</td>
<td>0.024</td>
</tr>
<tr>
<td>$k_{\text{resp, ZOO, } T_0}$</td>
<td>0.077</td>
<td>$b_{\text{p, red}}$</td>
<td>0.142</td>
<td>$f_{\text{NH}_4, \text{ air}}$</td>
<td>0.024</td>
</tr>
<tr>
<td>$k_{\text{miner, ana, sed, } T_0}$</td>
<td>0.076</td>
<td>$b_{\text{p, max}}$</td>
<td>0.138</td>
<td>$f_{\text{NH}_4, \text{ rain}}$</td>
<td>0.020</td>
</tr>
<tr>
<td>$K_{\text{O}_2, \text{ miner}}$</td>
<td>0.053</td>
<td>$b_{\text{p, max}}$</td>
<td>0.138</td>
<td>$f_{\text{XO, rain}}$</td>
<td>0.018</td>
</tr>
<tr>
<td>$k_{\text{miner, aero, wat, } T_0}$</td>
<td>0.043</td>
<td>$K_{\text{I, PLRT}}$</td>
<td>0.115</td>
<td>$f_{\text{XO, rain}}$</td>
<td>0.018</td>
</tr>
<tr>
<td>$K_{\text{NH}_4, \text{ nitri}}$</td>
<td>0.040</td>
<td>$b_{\text{p, max}}$</td>
<td>0.123</td>
<td>$f_{\text{XO, rain}}$</td>
<td>0.018</td>
</tr>
<tr>
<td>$K_{\text{O}_2, \text{ miner}}$</td>
<td>0.035</td>
<td>$v_{\text{ZOO}}$</td>
<td>0.115</td>
<td>$f_{\text{XO, rain}}$</td>
<td>0.018</td>
</tr>
<tr>
<td>$K_{\text{O}_2, \text{ resp}}$</td>
<td>0.028</td>
<td>$\beta_{\text{PLR}}$</td>
<td>0.125</td>
<td>$f_{\text{XO, rain}}$</td>
<td>0.018</td>
</tr>
<tr>
<td>$K_{\text{HO}_2, \text{ ALG}}$</td>
<td>0.023</td>
<td>$\beta_{\text{PLR}}$</td>
<td>0.084</td>
<td>$f_{\text{XO, rain}}$</td>
<td>0.018</td>
</tr>
<tr>
<td>$K_{\text{O}_2, \text{ nitri}}$</td>
<td>0.021</td>
<td>$\beta_{\text{WOR}}$</td>
<td>0.064</td>
<td>$f_{\text{XO, rain}}$</td>
<td>0.018</td>
</tr>
<tr>
<td>$k_{\text{miner, ana, wat, } T_0}$</td>
<td>0.021</td>
<td>$a_{\text{N}}$</td>
<td>0.050</td>
<td>$f_{\text{XO, rain}}$</td>
<td>0.018</td>
</tr>
<tr>
<td>$K_{\text{O}_2, \text{ opt}}$</td>
<td>0.018</td>
<td>$\beta_{\text{BAC}}$</td>
<td>0.043</td>
<td>$f_{\text{XO, rain}}$</td>
<td>0.018</td>
</tr>
</tbody>
</table>

Table 4.6: Average error contributions of the parameters with respect to the measured values, which are the concentrations of algae, zooplankton, phosphate, nitrate and oxygen. The parameters are grouped according to the groups made by the tables 4.3, 4.4 and 4.5.

4.4 Dependence Analysis

In this section, the information on the sensitivity of model results on parameters shown in Table 4.6 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. (2000), 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 subsets 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 $s_i$ (4.5) for a given parameter set are approximatively 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 (4.5) are normed

\[ S_i = \frac{s_i}{||S||} \]

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 \( \gamma \), is used as the measure for approximate linear dependence:

\[ \gamma(\theta) = \min_{||\beta||=1} \frac{1}{||S_1\beta_1 + \ldots + S_m\beta_m||} = \frac{1}{\min_{||\beta||=1} ||S\beta||} = \frac{1}{\sqrt{\min\{EV[S^T S]\}}} \]  

(EV[.] means the operator calculating the set of eigenvalues of the argument; see Brun et al. (2000) 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. (2000), we prefer the collinearity index \( \gamma \) because of its more intuitive interpretation given by the first expressions in equation (4.8): 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, \( \gamma \), between 10 and 15.

Note that both analyses, the sensitivity analysis done in section 4.3 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.

4.4.1 Identifiability Analysis of Model Parameters

Fig. 4.1 gives an overview of possible values of the collinearity indices, \( \gamma \), for all subsets of potential fit parameters listed in Table 4.3. 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 4.2), there are
already sets of three parameters, that lead to serious identifiability problems caused by the approximate linear dependence of sensitivity functions. On the other hand, 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. 4.1 is of limited value because it leads to the result that there may exist identifiable and non-identifiable parameters sets of nearly any size (this is not a special feature of the present investigation but can typically be expected). The key question is according to the collinearity indices of subsets of parameters identified in the top rows of Table 4.6 to which the model results are very sensitive. Before proceeding with this question, Table 4.7 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 5 combinations of three parameters listed in Table 4.7 consist all of subsets of the 8 most influential parameters out of the group of potential fit parameters listed in the left column of Table 4.6.

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 is shown in Table 4.8. The first two columns of this table list the 9 most influential parameters from the left column of Table 4.6 together with the sensitivity measure $\delta^{\text{sq}}_p$ according to equations (4.3) or (4.6) (the list stops there because there is a significant decrease by nearly a factor of 2 in $\delta^{\text{sq}}_p$ between rank 9 and rank 10).

A first analysis of possible approximate linear dependence problems consisted in cal-
<table>
<thead>
<tr>
<th>Set Size</th>
<th>Comb.</th>
<th>γ-range</th>
<th>Combinations with highest value of γ: Parameter Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>300</td>
<td>1.0-9.2</td>
<td>9.2: ( K_{i,ALG} ) ( k_{resp,ALG,T_0} )</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>6.9: ( K_{i,ALG} ) ( k_{gro,ALG,T_0} )</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5.9: ( k_{gro,ZO0,T_0} ) ( k_{death,ZO0,T_0} )</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5.4: ( k_{gro,ZO0,T_0} ) ( S_{HPO4,crit} )</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4.3: ( k_{resp,ALG,T_0} ) ( k_{gro,ALG,T_0} )</td>
</tr>
<tr>
<td>3</td>
<td>2300</td>
<td>1.0-17.0</td>
<td>17.0: ( K_{i,ALG} ) ( k_{gro,ALG,T_0} ) ( k_{resp,ALG,T_0} )</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>14.0: ( K_{i,ALG} ) ( k_{gro,ALG,T_0} ) ( k_{death,ALG,T_0} )</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>11.8: ( K_{i,ALG} ) ( k_{resp,ALG,T_0} ) ( K_{HPO4,ALG} )</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>11.6: ( K_{i,ALG} ) ( k_{resp,ALG,T_0} ) ( S_{HPO4,crit} )</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>11.5: ( K_{i,ALG} ) ( k_{resp,ALG,T_0} ) ( k_{death,ALG,T_0} )</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>11.5: ( K_{i,ALG} ) ( k_{resp,ALG,T_0} ) ( \Delta S_{HPO4} )</td>
</tr>
</tbody>
</table>

Table 4.7: Identification of the parameter sets belonging to the largest collinearity indices for subset sizes 2 and 3. 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>
<thead>
<tr>
<th>( \delta_{\text{msgr}}^{i} )</th>
<th>analysis 1</th>
<th>analysis 2</th>
<th>analysis 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \delta_{\text{set to par}}^{i} )</td>
<td>( \gamma ) set to par</td>
<td>( \gamma ) x = par</td>
<td>( \gamma ) x + par</td>
</tr>
<tr>
<td>offset ( \gamma )</td>
<td>1.0</td>
<td>12.8</td>
<td>12.8</td>
</tr>
<tr>
<td>( K_{i,ALG} )</td>
<td>1.21</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>( k_{death,ZO0,T_0} )</td>
<td>0.84</td>
<td>1.1</td>
<td>x</td>
</tr>
<tr>
<td>( k_{gro,ALG,T_0} )</td>
<td>0.83</td>
<td>8.7</td>
<td>x</td>
</tr>
<tr>
<td>( k_{gro,ZO0,T_0} )</td>
<td>0.76</td>
<td>12.6</td>
<td>x</td>
</tr>
<tr>
<td>( k_{resp,ALG,T_0} )</td>
<td>0.71</td>
<td>18.0</td>
<td>18.0</td>
</tr>
<tr>
<td>( k_{death,ALG,T_0} )</td>
<td>0.60</td>
<td>19.0</td>
<td>16.5</td>
</tr>
<tr>
<td>( K_{HPO4,ALG} )</td>
<td>0.50</td>
<td>19.9</td>
<td>12.6</td>
</tr>
<tr>
<td>( S_{HPO4,crit} )</td>
<td>0.43</td>
<td>21.2</td>
<td>12.6</td>
</tr>
<tr>
<td>( h_{sed} )</td>
<td>0.30</td>
<td>21.6</td>
<td>12.7</td>
</tr>
</tbody>
</table>

Table 4.8: Collinearity indices of parameter subsets of the most influential potential fit parameters according to Table 4.6. The first two columns repeat the information given in Table 4.6 (names of the parameters and values of the nondimensional average sensitivity, \( \delta_{\text{msgr}}^{i} \)). The column labelled ‘analysis 1’ contains collinearity indices, \( \gamma \), 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, \( \gamma \), 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’.

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’ of Table 4.8 and are shown by a sloping line in Fig. 4.1. 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 pa-
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’ of Table 4.8. These results show that the parameters $k_{\text{resp}, \text{ALG}, T_0}$ and $k_{\text{death}, \text{ALG}, T_0}$ increase the collinearity index significantly, whereas the addition of one of the parameters $K_{\text{HPO}_4, \text{ALG}}$, $S_{\text{HPO}_4, \text{crit}}$, or $h_{\text{sed}}$ did not significantly increase the value of $\gamma$ (compare with the offset values in the top row). This result motivated the third analysis, the results of which are shown in the columns labelled ‘analysis 3’ of Table 4.8. Here, all three parameters $K_{\text{HPO}_4, \text{ALG}}$, $S_{\text{HPO}_4, \text{crit}}$, and $h_{\text{sed}}$ 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 $k_{\text{resp}, \text{ALG}, T_0}$ or $k_{\text{death}, \text{ALG}, T_0}$ 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 7 parameters $K_{\text{I, ALG}}$, $k_{\text{gro, ALG}, T_0}$, $k_{\text{death}, \text{ZOO}, T_0}$, $k_{\text{gro, ZOO}, T_0}$, $K_{\text{HPO}_4, \text{ALG}}$, $S_{\text{HPO}_4, \text{crit}}$, and $h_{\text{sed}}$ 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. 4.1 is shown by a solid square.

Estimation of the 7 parameters $K_{\text{I, ALG}}$, $k_{\text{gro, ALG}, T_0}$, $k_{\text{death}, \text{ZOO}, T_0}$, $k_{\text{gro, ZOO}, T_0}$, $K_{\text{HPO}_4, \text{ALG}}$, $S_{\text{HPO}_4, \text{crit}}$, and $h_{\text{sed}}$ by minimizing the sum (4.1) leads to the estimates already given in Table 4.3. If we assume normally distributed, independent measurement errors with standard deviations proportional to the scales given in Table 4.2 and assume the other model parameters to be given, approximate standard errors and correlation coefficients of the estimates of the 7 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 4.9. 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, $k_{\text{gro, ALG}, T_0}$, and the half-saturation light intensity for algal growth, $K_{\text{I, ALG}}$. This problem was already observed in Tables 4.7 and 4.8. The large positive correlation coefficient of these two parameters is an indication of the absence of a significant light saturation effect. This problem could partially be overcome with a reparameterization of this process replacing.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>s.e.</th>
<th>$K_{\text{I, ALG}}$</th>
<th>$k_{\text{death}, \text{ZOO}, T_0}$</th>
<th>$k_{\text{gro, ALG}, T_0}$</th>
<th>$k_{\text{gro, ZOO}, T_0}$</th>
<th>$K_{\text{HPO}_4, \text{ALG}}$</th>
<th>$S_{\text{HPO}_4, \text{crit}}$</th>
<th>$h_{\text{sed}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K_{\text{I, ALG}}$</td>
<td>25 %</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$k_{\text{death}, \text{ZOO}, T_0}$</td>
<td>10 %</td>
<td>0.50</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$k_{\text{gro, ALG}, T_0}$</td>
<td>14 %</td>
<td>0.96</td>
<td>0.30</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$k_{\text{gro, ZOO}, T_0}$</td>
<td>4 %</td>
<td>0.48</td>
<td>0.45</td>
<td>0.37</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$K_{\text{HPO}_4, \text{ALG}}$</td>
<td>21 %</td>
<td>0.16</td>
<td>0.63</td>
<td>0.07</td>
<td>0.12</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$S_{\text{HPO}_4, \text{crit}}$</td>
<td>15 %</td>
<td>0.17</td>
<td>0.62</td>
<td>0.01</td>
<td>0.29</td>
<td>0.82</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>$h_{\text{sed}}$</td>
<td>2 %</td>
<td>-0.10</td>
<td>-0.18</td>
<td>-0.07</td>
<td>-0.30</td>
<td>0.14</td>
<td>-0.02</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 4.9: Relative standard errors (estimates of the parameter values are given in Table 4.3) and correlation coefficients for the seven parameters fitted to the data ignoring the uncertainty in the other parameters.
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 $k_{gro, ALG,T_0}$) (Ratkowsky 1986). Because the identifiability problems are not too severe and the parameterization used by Omlin et al. (2000) 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 the introduction, 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 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.

4.4.2 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 rises the question on the dependence of the parameter estimates on parameters that were held 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 4.10 shows the collinearity indices of all parameters sets (of size 8) consisting of the 7 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, $k_{resp, ALG,T_0}$, 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 $k_{resp, ALG,T_0}$ can be expected to have the largest influence on the estimates of the 7 fitted parameters. Table 4.11 shows the change in the fit results caused by an increase of the value of the parameter $k_{resp, ALG,T_0}$ by one uncertainty unit, $\Delta k_{resp, ALG,T_0}$. The resulting change in the parameters $K_{I, ALG}$ and $k_{gro, ALG,T_0}$ is largest. These are the most strongly correlated parameters (see Table 4.9) and the parameters for which the collinearity indices with $k_{resp, ALG,T_0}$ are largest (9.2 and 4.3, whereas it is in the range between 1.0 and 2.9 for $k_{resp, ALG,T_0}$ together with any of the
Table 4.10: Collinearity indices, $\gamma$, of all sets of parameters consisting of the seven parameters used in the fit (cf. Table 4.9) plus exactly one additional parameter. The sets are identified by the name of the additional parameter. To simplify the discussion, the average sensitivity, $\delta_{\theta}^{\text{masq}}$, of the results with respect to the additional parameter is also given (data from Table 4.6).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$\gamma$</th>
<th>$\delta_{\theta}^{\text{masq}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k_{\text{resp},ALG,T_0}$</td>
<td>19.13</td>
<td>0.711</td>
</tr>
<tr>
<td>$k_{\text{death},ALG,T_0}$</td>
<td>18.16</td>
<td>0.506</td>
</tr>
<tr>
<td>$k_{\text{death},PLR,T_0}$</td>
<td>17.54</td>
<td>0.348</td>
</tr>
<tr>
<td>$k_{\text{gro},PLR,T_0}$</td>
<td>17.31</td>
<td>0.268</td>
</tr>
<tr>
<td>$\delta_{\text{PLR}}$</td>
<td>17.04</td>
<td>0.125</td>
</tr>
<tr>
<td>$K_{\text{resp},PLR,T_0}$</td>
<td>17.03</td>
<td>0.122</td>
</tr>
<tr>
<td>$W_{\text{ALG}}$</td>
<td>16.46</td>
<td>0.300</td>
</tr>
<tr>
<td>$K_{1,PLR}$</td>
<td>15.84</td>
<td>0.115</td>
</tr>
<tr>
<td>$c_{e}$</td>
<td>15.82</td>
<td>0.142</td>
</tr>
<tr>
<td>$F_{\text{XPORG}}$</td>
<td>15.40</td>
<td>0.091</td>
</tr>
<tr>
<td>$Y_{\text{ZOO},\text{max}}$</td>
<td>15.03</td>
<td>0.133</td>
</tr>
<tr>
<td>$k_{1}$</td>
<td>14.82</td>
<td>0.559</td>
</tr>
<tr>
<td>$F_{\text{HPOA}}$</td>
<td>13.99</td>
<td>0.222</td>
</tr>
<tr>
<td>$\beta_{\text{BAC}}$</td>
<td>13.86</td>
<td>0.043</td>
</tr>
<tr>
<td>$b_{\text{P,max}}$</td>
<td>13.83</td>
<td>0.084</td>
</tr>
<tr>
<td>$\psi_{\text{ZOO}}$</td>
<td>13.81</td>
<td>0.115</td>
</tr>
<tr>
<td>$a_{\text{P,inflow}}$</td>
<td>13.75</td>
<td>0.051</td>
</tr>
<tr>
<td>$\bar{F}_{\text{X,domm}}$</td>
<td>13.67</td>
<td>0.030</td>
</tr>
<tr>
<td>$\psi_{\text{sed,ORG}}$</td>
<td>13.65</td>
<td>0.134</td>
</tr>
<tr>
<td>$F_{\bar{X}}$</td>
<td>13.63</td>
<td>0.188</td>
</tr>
<tr>
<td>$\theta$</td>
<td>13.60</td>
<td>0.304</td>
</tr>
<tr>
<td>$k_{2}$</td>
<td>13.60</td>
<td>0.138</td>
</tr>
<tr>
<td>$\delta_{\text{ALG}}$</td>
<td>13.56</td>
<td>0.302</td>
</tr>
<tr>
<td>$\nu_{\text{ORB}}$</td>
<td>13.55</td>
<td>0.064</td>
</tr>
<tr>
<td>$k_{\text{miner},\text{aer},\text{p,org}}$</td>
<td>13.49</td>
<td>0.040</td>
</tr>
<tr>
<td>$F_{\bar{X}_{s}}$</td>
<td>13.48</td>
<td>0.064</td>
</tr>
<tr>
<td>$a_{\text{P,max}}$</td>
<td>13.47</td>
<td>0.147</td>
</tr>
<tr>
<td>$k_{\text{miner},\text{aer},\text{p,alg}}$</td>
<td>13.34</td>
<td>0.021</td>
</tr>
<tr>
<td>$k_{\text{miner},\text{aer},\text{p,alg}}$</td>
<td>13.34</td>
<td>0.021</td>
</tr>
</tbody>
</table>

Table 4.11: Change in fit results caused by an increase of the value of $k_{\text{resp},ALG,T_0}$ by one uncertainty unit, $\Delta k_{\text{resp},ALG,T_0}$, as listed in Table 4.3 (from 0.050 d$^{-1}$ to 0.075 d$^{-1}$).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Original Value</th>
<th>New Value</th>
<th>Unit</th>
<th>Change [%]</th>
<th>Change/$\Delta \theta_l$ [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K_{1,ALG}$</td>
<td>34</td>
<td>20</td>
<td>Wm$^{-2}$</td>
<td>-43</td>
<td>-85</td>
</tr>
<tr>
<td>$k_{\text{gro},ALG,T_0}$</td>
<td>1.1</td>
<td>0.94</td>
<td>d$^{-1}$</td>
<td>-16</td>
<td>-82</td>
</tr>
<tr>
<td>$k_{\text{death},ZOO,T_0}$</td>
<td>0.029</td>
<td>0.028</td>
<td>d$^{-1}$</td>
<td>-2.5</td>
<td>-4.9</td>
</tr>
<tr>
<td>$k_{\text{gro},ZOO,T_0}$</td>
<td>0.30</td>
<td>0.31</td>
<td>gDM$^{-1}$m$^3$d$^{-1}$</td>
<td>1.6</td>
<td>8.2</td>
</tr>
<tr>
<td>$K_{\text{HPOA},ALG}$</td>
<td>0.0019</td>
<td>0.0015</td>
<td>gPm$^{-3}$</td>
<td>-18</td>
<td>-37</td>
</tr>
<tr>
<td>$S_{\text{HPOA,crit}}$</td>
<td>0.0042</td>
<td>0.0042</td>
<td>gPm$^{-3}$</td>
<td>-0.3</td>
<td>-1.4</td>
</tr>
<tr>
<td>$l_{\text{sed}}$</td>
<td>0.0036</td>
<td>0.0036</td>
<td>m</td>
<td>0.1</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Table 4.11: Change in fit results caused by an increase of the value of $k_{\text{resp},ALG,T_0}$ by one uncertainty unit, $\Delta k_{\text{resp},ALG,T_0}$, as listed in Table 4.3 (from 0.050 d$^{-1}$ to 0.075 d$^{-1}$).

other parameters listed in Table 4.11). Note that the decrease in the value of $k_{\text{gro},ALG,T_0}$ together with the even larger decrease in the value of $K_{1,ALG}$ leads to an increase in the
production rate for small light intensities. This compensates for the increase in respiration.

4.5 Prediction Uncertainty

The computation time required for simulations with the lake model make it very expensive to do nonlinear or regional uncertainty analyses with the aid of Monte Carlo simulations and/or Bayesian inference (Hornberger and Spear 1981; Hornberger and Spear 1983; Beck 1987; Gelman et al. 1995; Beck 1999; Rockhow and Chapra 1999). In order to get 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 cost, because the sensitivity functions \( \frac{\partial y_k}{\partial \theta_i} \) required for calculating the estimate of the uncertainty range of the model prediction for the variable \( y_k \) at location \( z_{yk,i} \) and at time \( t_j \)

\[
\Delta y_k(z_{yk,i}, t_j) = \sqrt{\sum_{j=1}^{m} \text{Corr} \{ \theta_j \} \frac{\partial y_k(z_{yk,i}, t_j)}{\partial \theta_1} \Delta \theta_1 \left( \frac{\partial y_k(z_{yk,i}, t_j)}{\partial \theta_2} \Delta \theta_2 \right)^2}
\]

have already been calculated for the sensitivity and dependence analyses performed in the sections 4.3 and 4.4 (\( m \) in equation 4.9 is equal to the number of parameters and \( \text{Corr} \{ \theta \} \) is the correlation matrix of the parameters). The meaning of \( \Delta y_k \) is the same as that for \( \Delta \theta_i \); as for the uncertainty ranges, we use measures comparable with a standard deviation of the corresponding distribution. The most critical approximation made by equation 4.9 is that it is only valid as long as the nonlinearities 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 nonlinearity 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 nonlinearity 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 (4.9) is the assumption to be made for the uncertainty ranges \( \Delta \theta_i \) and for the correlation matrix \( \text{Corr} \{ \theta \} \) of the model parameters. We perform the following three calculations: First, we calculate uncertainty of the results assuming uncorrelated parameters with the prior uncertainties listed in Tables 4.3 to 4.5. Second, we replace the uncertainties and correlations for the fitted parameters by those estimated in the fit (see Table 4.9). And third, we set the uncertainties of all potential fit parameters listed in Table 4.3 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 4.4.1, 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 has the purpose to see what would be the maximum possible reduction in prediction uncertainty by an estimation of all parameters which are meaningful to fit with the data used for this investigation (mainly monthly profiles of oxygen, phosphate, nitrate, algae and Planktothrix (Oscillatoria) rubescens and depth-integrated samples of zooplankton). The remaining uncertainty is caused by the parameters listed in
Tables 4.4 and 4.5. These remaining sources of uncertainties are mainly limited knowledge of sedimentation velocities, stoichiometric coefficients, kinetic parameters of *Planktothrix* (*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 4.9, 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, lead 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 Figure 4.2 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 *Planktothrix* (*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 seems 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 be exceeded.

Besides this relative importance of different causes of uncertainty also the magnitude of the uncertainty is 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. 4.2. 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 uncertainty, the good correspondence of the simulations with measured data is puzzling. On one side, 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). On the other hand, 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 a test of all model mechanisms.
Figure 4.2: Parameter error propagation on model predictions for oxygen, phosphate, nitrate and algae concentrations for three months in 1990. Markers represent measured data, black lines model predictions and grey lines predictions $\pm \epsilon_{ijk}$ according to equation (4.9). Uncertainty bounds have been calculated for the prior uncertainties listed in Tables 4.3 to 4.5 (widest uncertainty ranges), for the uncertainties of the fitted parameters replaced by the information given in Table 4.9 (intermediate uncertainty ranges), and for all uncertainties of the potential fit parameters according to Table 4.3 set to zero (narrowest uncertainty ranges).
4.6 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. (2000) for the biogeochemical lake model described by Omlin et al. (2000) lead 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-identifiability 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 easily be 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 biochemical processes and of the phosphorus content of algae. It may not be possible to reduce some of these uncertainties without increasing 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 (Bock 1999; Reckhow and Chapra 1999).

Acknowledgments

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. (2000).

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.
Curriculum vitae

29.12.66 Born in Sarnen

1970-1983 Primar and Sekundarschule Dietlikon

1983-1987 Kantonsschule Büclrain Winterthur

1987 Matura Typus E

1989-1995 Study at the Swiss Federal Institute of Technology (ETH) Zürich, Department of Environmental Sciences

1995 Diploma in natural science (Dipl. Natw. ETH)

1995-1999 Ph.D. thesis at the Swiss Federal Institute for Environmental Science and Technology (EAWAG)
References


117


