MODELING ENVIRONMENTAL SYSTEMS UNDER UNCERTAINTY:
TOWARDS A SYNTHESIS OF DATA-BASED AND THEORY-BASED MODELS

by

ZHULU LIN

(Under the direction of M. Bruce Beck)

ABSTRACT

A two-pronged modeling approach, in which both data-based modeling and theory-based modeling methods are jointly incorporated, has been developed. Its purpose is to gain a deeper understanding of complex, poorly-defined environmental systems. The Data-Based-Mechanistic methodology is employed in the data-based modeling procedure for structure identification and parameter estimation of a data-based model (DBM, in transfer function forms). In the theory-based modeling prong, the Recursive Prediction Error algorithm is modified and engaged in estimating time-varying parameters and detecting structural change for a theory-based model (TBM, in ordinary differential forms). Two concepts from linear processes in control system engineering, time constant and steady-state gain, are then examined in a synthesis of the two types of model in the parameter space spanned by these two lumped parameters.

Case studies on two environmental systems, an activated sludge system and an aquaculture pond, have been carried out to test the effectiveness of this proposed modeling approach. The results of both case studies have shown that: (1) more is gained through the joint application of the two separate modeling approaches to the same environmental system than the exclusive use of either model type; (2) to some extent, the structure identification and parameter estimation of one type of model can be readily improved by recourse to the modeling results of the other; and (3) reconciliation in the parameter space of the two types of models, based on data or theory, shows superiority over that in the state space.

INDEX WORDS: Activated sludge process, Aquaculture pond, Data-based modeling, Nonlinear system, Parameter estimation, Recursive estimation algorithm, System identification, Theory-based modeling, Time-varying parameter, Uncertainty analysis
MODELING ENVIRONMENTAL SYSTEMS UNDER UNCERTAINTY:
TOWARDS A SYNTHESIS OF DATA-BASED AND THEORY-BASED MODELS

by

ZHULU LIN

B.E., East China Institute of Chemical Technology, China, 1991
M.E., East China University of Science and Technology, China, 1996
M.S., The University of Georgia, 2003

A Dissertation Submitted to the Graduate Faculty
of The University of Georgia in Partial Fulfillment
of the
Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2003
Modeling Environmental Systems under Uncertainty:
Towards a Synthesis of Data-based and Theory-based Models

by

Zhulu Lin

Approved:

Major Professor: M. Bruce Beck

Committee: Wei-Jun Cai
C. Rhett Jackson
Robert B. Lund
Todd C. Rasmussen

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2003
DEDICATION

To Siew Hoon, James and my parents
The effort involved in writing this dissertation was not something I could undertake by myself alone. I am deeply indebted to my major professor, Dr. Beck, whose enthusiasm in understanding environmental systems and knowledge in modeling methodologies encouraged and guided me through the whole process and at each step. I also want to extend my thanks to those who helped and watched me grow in academia in the past years, but only a few out of many names are to be mentioned here. They are Ms. Jenny Yearwood, Mr. Stefano Ciavatta, Mr. Deek Cox, Mr. Chris Holliday, Mr. Feng Jiang, Dr. Hans Stigter, Dr. Wei Zeng, Dr. Rasmussen, Dr. Rong Liu, Dr. Femi Osidele, Dr. Xiao-Qing Zeng, and Dr. Amy Parker.

Growing in academia is not possible without growing with my dear family. I wish to give all credit to my aging parents, who brought me up in the face of adversity. Without their sacrifice, I myself would not have been able to jump to being in front of a computer screen from the back of an ox. Without the love from my wife, the joy from my son, and the support from my brothers and sisters, I would not have walked this far with light whistles instead of heavy groans. More importantly, without the light of Jesus, I would be still searching for nothing in the dark.

Last but not least, I want to acknowledge The University of Georgia Graduate School, Daniel B. Warnell School of Forest Resources and the Wheatley-Georgia Research Alliance foundation for financially supporting me throughout my graduate studies.
# Table of Contents

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acknowledgments</td>
<td>v</td>
</tr>
<tr>
<td>List of Figures</td>
<td>ix</td>
</tr>
<tr>
<td>List of Tables</td>
<td>xiv</td>
</tr>
<tr>
<td><strong>Chapter</strong></td>
<td></td>
</tr>
<tr>
<td>1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Models — reconciliation of theory with data</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Objectives of study</td>
<td>5</td>
</tr>
<tr>
<td>1.3 Approaches of the study</td>
<td>7</td>
</tr>
<tr>
<td>1.4 Structure of the dissertation</td>
<td>10</td>
</tr>
<tr>
<td>2 A Literature Review</td>
<td>11</td>
</tr>
<tr>
<td>2.1 Beyond toggling between theory- and data-based models</td>
<td>11</td>
</tr>
<tr>
<td>2.2 System identification</td>
<td>14</td>
</tr>
<tr>
<td>2.3 Methods for model structure identification</td>
<td>21</td>
</tr>
<tr>
<td>2.4 Methods for parameter estimation</td>
<td>25</td>
</tr>
<tr>
<td>3 Methodological Basis</td>
<td>32</td>
</tr>
<tr>
<td>3.1 Signal processing</td>
<td>32</td>
</tr>
<tr>
<td>3.2 Data-Based-Mechanistic modeling</td>
<td>37</td>
</tr>
<tr>
<td>3.3 Structure identification of mechanistic models</td>
<td>46</td>
</tr>
<tr>
<td>3.4 Time constant and steady-state gain</td>
<td>61</td>
</tr>
</tbody>
</table>
4 Case Study One: Ammonium-Nitrogen Removal in a Conventional Activated Sludge System

4.1 Introduction

4.2 Study site and data collection

4.3 Data-based modeling

4.4 Theory-based modeling

4.5 Towards the synthesis

5 Case Study Two: Algal Blooms in a Small Southeastern Piedmont Impoundment

5.1 A review of recent development in phytoplankton modeling

5.2 Site, experiment and data collection

5.3 Signal processing

5.4 Data-Based-Mechanistic modeling

5.5 A simple mechanistic model

5.6 Recursive parameter estimation and structural change

5.7 Towards the synthesis

6 Retrospective Remarks

6.1 The synthesis

6.2 Contributions, limitations and issues for future research

Bibliography

Appendix

A Algorithms of the Data-Based-Mechanistic Modeling

A.1 Simplified Refined Instrumental Variable Method

A.2 Fixed Interval Smoothing Algorithm
A.3 Model Order Identification (excerpted from [268]) . . . . 189

B The Recursive Prediction Error Algorithm . . . . . . . . . 192

C A Simulation Model for Aquaculture Pond Water Quality . . 195

D MATLAB Codes for the Modified RPE Algorithm . . . . . . . 203
List of Figures

1.1 Towards a synthesis of data-based and theory-based models through system identification. .................................................. 7

2.1 The definition of system and variables. .......................................................... 16

3.1 The raw data of DO concentrations taken from a fishery pond. ......................... 35

3.2 Determination of the AR order of the DO time-series. ....................................... 36

3.3 The spectrum of the DO time-series. ............................................................... 36

3.4 Signal extraction of the DO time-series: (a) trend, (b) seasonal component. .... 36

3.5 Interpolation and smoothing of the raw DO time-series. .................................. 36

3.6 Schematic diagram of the Data-Based-Mechanistic modeling procedure. ........ 38

3.7 SIMULINK flow diagram for the biomass-substrate model. ............................. 51

3.8 The simulated input and outputs of the biomass-substrate model. .................... 52

3.9 The reconstructed observations of states with time-invariant model parameters. 53

3.10 The reconstructed time-invariant model parameters (true value, dashed line; estimated value, continuous line; standard error, dotted lines). .................... 53

3.11 The deviations of the elements of the Kalman gain matrix with time-invariant model parameters. ................................................................. 54

3.12 The reconstructed observations of states with time-varying model parameter. .... 56

3.13 The reconstruction of (a) the time-varying parameter (true value, dashed line; estimated value, continuous line; standard error, dotted lines); and (b) the time-invariant parameter (true value, dashed line; estimated value, continuous line; standard error, dotted lines). ........................................ 56
3.14 The deviations of the elements of Kalman gain matrix with *time-varying* model parameter. ................................. 57
3.15 The comparison of model output and observations of the concentrations at downstream. ........................................... 59
3.16 The estimation of the parameters. ................................................. 60
3.17 The deviations of the elements in the $K$ matrix. ....................... 60

4.1 Schematic diagram of the activated sludge process in the AWTF No. 2. 65
4.2 Flow rates of (a) Crude sewage influent, and (b) Returned Activated Sludge (RAS). ......................................................... 67
4.3 Ammonia-nitrogen concentrations in (a) Crude sewage, (b) Outer channel, and (c) Inner channel. ......................... 67
4.4 Concentrations of (a) MLSS in Inner channel, and (b) DO in Middle channel. 68
4.5 Sub-system conceptualization for the activated sludge process of AWTF No. 2. 69
4.6 Linear TF model $[1, 1, 1]$ with constant parameters: (a) one-step-ahead prediction and observations, and (b) residuals plot. 72
4.7 Variations of TVP estimates of model $[1, 1, 1]$: (a) lag parameter ($\hat{a}_1(k|N)$), and (b) input parameter ($\hat{b}_0(k|N)$). ......................... 75
4.8 Comparison of variation patterns among: (a) *time constant*, (b) *steady-state gain*, and (c) trend of MLSS. ......................... 75
4.9 Linear TF model $[1, 1, 1]$ with TVP: (a) one-step-ahead prediction and observations, and (b) residuals plot. ......................... 76
4.10 Variations of TVP estimates of model $[1, 1, 1]$: (a) lag parameter ($\hat{a}_1(k|N)$), and (b) input parameter ($\hat{b}_0(k|N)$). ......................... 78
4.11 Comparison of variation patterns among: (a) *time constant*, (b) *steady-state gain*, and (c) Trend of MLSS. ......................... 78
4.12 Linear TF model [1, 1, 1] with TVP: (a) one-step-ahead prediction and observations, and (b) residuals plot. ................................................................. 79
4.13 Two types of conceptualization of nitrification process. ....................... 80
4.14 Re-conceptualized System One (Figure 4.5(a)) based on the results of the DBM step. ................................................................. 82
4.15 Recursive estimation results: (a) one-step-ahead prediction vs. observations of ammonium-N concentration, and (b) reconstructed nitrifying bacteria concentration. ................................................................. 86
4.16 (a) Removal efficiency of ammonium-nitrogen; (b) The biomass concentrations of the reconstructed nitrifying bacteria and the MLSS trend signal. ................................. 88
4.17 Recursive estimation results: estimates (solid line) and standard errors (dashed line, almost invisible) of the time-invariant parameters ($\mu_{max}$ and $k_d$) and the time-varying parameters ($r(t)$ and $u_2(t)$). ................................................................. 89
4.18 Recursive estimation results: estimates (solid line) and standard errors (dash line) of the elements of the $K$ matrix. ................................................................. 91
4.19 Comparisons of the two types of models (DBM(4.1): cyan dashed line; DBM(4.3): green dot-dashed line; DBM(4.5): blue solid line; and TBM (4.6): red dotted line) at: (a) time constants and (b) steady-state gains. ................................. 94
4.20 Comparisons of the two types of models with respect to model predictions. ................................. 94
4.21 (a) Components of the time constant of the TBM; (b) time constants of the TBM and DBM. ................................................................. 95

5.1 Whitehall pond with two EPCL trailers in the left background. ............... 108
5.2 Sampling locations at the Whitehall pond (reproduced from [273]). .......... 109
5.3 Examples of the raw data sets retrieved from the EPCL, fluorometers and Hydrolabs at location 2 (from May 23rd to October 16th, 2000): (a) from the EPCL, (b) from the fluorometers, (c) from the Hydrolabs. ................................. 110
5.4 Interpolated time-series with trends: (a) PAR below 0.5 surface, (b) Chlorophyll-
$\alpha$, (c) temperature, and (d) DO. .................................................. 111
5.5 Solar radiation serves as a driving force to the pond ecosystem (the vertical
black solid line indicates when algal biomass reaches its peak). ................. 112
5.6 Nonlinearity between input and output of Equation (5.1). ......................... 117
5.7 The variation of $\hat{b}_0(k|N)$ and its standard errors superimposed on the scaled
$y(k)$. ........................................................................................................ 119
5.8 Linear relationship between $\hat{b}_0(k|N)$ and $y(k)$. ................................. 119
5.9 Simulation and one-step ahead prediction of TF [1, 1, 2] with TVP (in Figure
5.7 against data). ....................................................................................... 119
5.10 The variation of SDP superimposed with scaled $y^*(k)$. ......................... 121
5.11 Linear relationship between SDP and $y^*(k)$. ........................................ 121
5.12 Forecasting and simulation of model (5.7) against data. ......................... 123
5.13 Cross-correlation function (CCF) between the deterministic residual $\nu(k)$ of
model (5.7) and temperature. ................................................................. 124
5.14 Forecasting and simulation of model (5.8) against data. ......................... 125
5.15 System block diagram of the photosynthesis process of algae as inferred from
the deterministic part of model (5.8). ...................................................... 126
5.16 Two stages of photosynthesis: Integration of the light reactions and Calvin
cycle. Redrawn from Figure 10.4 in [51]. ................................................. 127
5.17 Two segment relationship between $\hat{b}_0(k|N)$ and $y(k)$: (a) before $y(k)$ reaches
10.0 mg·L$^{-1}$ at the first time; (b) after $y(k)$ reaches 10.0 mg·L$^{-1}$ at the first
time. ........................................................................................................ 128
5.18 Forecasting and simulation of model (5.11) against data. ...................... 130
5.19 Temperature measurements at three levels of depth. ............................. 132
5.20 Concentrations of inorganic nitrogen and orthophosphate in Whitehall pond
(July 1st – August 5th, 2000). ................................................................. 132
5.21 Recursive estimation results: one-step-ahead prediction vs. observations of (a) algal biomass, and (b) dissolved oxygen. .................................................. 138

5.22 Recursive estimation results: estimates (solid line) and standard errors (dashed line, almost invisible) of the time-invariant parameters \((k_1, k_2, k_4)\) and the time-varying parameters \((k_3(t), k_5(t))\). .................................................. 139

5.23 Recursive estimation results: estimates (solid line) and standard errors (dash line) of the elements of \(K\) matrix. .................................................. 140

5.24 Model simulation results against field observations: (1) Biomass concentration of algae (upper panel) and duckweed (lower panel). ......................... 142

5.25 Model simulation results against field observations: (2) DO concentration (upper panel) and pH values (lower panel). ................................. 142

5.26 Model simulation results against field observations: (3) Concentration of organic P (upper panel) and orthophosphate (lower panel). ................. 143

5.27 Comparison of simulated algal biomass concentrations by different simulation models. ................................................................. 145

5.28 Comparison of simulated dissolved oxygen concentrations by different simulation models. ............................................................. 145

5.29 Comparison of simulated pH values by different simulation models. ........ 146

5.30 Comparison of the two models in parameter space: (a) time constant, (b) steady-state gain. ................................................................. 147

C.1 Conceptual model of an aquaculture pond. ................................. 196
<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Specifications of initial values and leading diagonal elements for the covariance matrices $P(t_0)$, $Q$ and $\Lambda$.</td>
<td>55</td>
</tr>
<tr>
<td>3.2</td>
<td>Specifications of initial values and leading diagonal elements for the covariance matrices $P(t_0)$, $Q$ and $\Lambda$.</td>
<td>58</td>
</tr>
<tr>
<td>4.1</td>
<td>The best 5 identified linear TF models with constant parameters (The sampling interval is 2 hours).</td>
<td>70</td>
</tr>
<tr>
<td>4.2</td>
<td>Model formulations, variables and parameters.</td>
<td>83</td>
</tr>
<tr>
<td>4.3</td>
<td>Specifications of initial values and the leading diagonal elements for the variance-covariance matrices $P(t_0)$, $Q$ and $\Lambda$.</td>
<td>85</td>
</tr>
<tr>
<td>5.1</td>
<td>Summary of recent theory-based models for eutrophication.</td>
<td>101</td>
</tr>
<tr>
<td>5.2</td>
<td>Summary of recent data-based models for algal growth.</td>
<td>106</td>
</tr>
<tr>
<td>5.3</td>
<td>Harmonics of all variables and their peak and trough times (24-hour clock) with dominant diurnal harmonics.</td>
<td>112</td>
</tr>
<tr>
<td>5.4</td>
<td>The best 5 identified linear TF models with constant parameters (The sampling interval is 2 hours).</td>
<td>116</td>
</tr>
<tr>
<td>5.5</td>
<td>The best 5 identified linear structures of DBM model (5.7) (The sampling interval is 2 hours).</td>
<td>122</td>
</tr>
<tr>
<td>5.6</td>
<td>The best 5 identified linear structures of DBM model (5.10) (The sampling interval is 2 hours).</td>
<td>130</td>
</tr>
<tr>
<td>5.7</td>
<td>Model formulations, variables and parameters.</td>
<td>134</td>
</tr>
</tbody>
</table>
5.8 Specifications of initial values and the leading diagonal elements for the variance-covariance matrices $P(t_0)$, $Q$ and $\Lambda$. ........................................ 137

C.1 Definition of state variables, environmental factors and parameters. ........ 200
C.1 (continued) ......................................................................................... 201
C.1 (continued) ......................................................................................... 202
Chapter 1

Introduction

Any model remains an intellectual plaything of limited impact ... ... unless it can be tested and verified by experiment, or by field observation, or both. A combined two-pronged approach — modeling coupled with direct study of the system or its components — holds the greatest promise of success.

— C. H. Mortimer ([152])

1.1 Models — reconciliation of theory with data

Water quality models of environmental systems have been extensively developed since the Streeter-Phelps equation describing the balance of dissolved oxygen in a stream first emerged in the 1920’s. These models are either research models, aiming at a better understanding of the system’s dynamics, or management models designed to assist the decision makers of the administration of water resources and pollution prevention. No matter what purpose a model is developed for, the success of the model is normally determined by how well it reconciles the theory governing the behaviors of the system with the data describing the dynamics of the system at an observable scale. Frequently, two polarized approaches to bridging the divide between theory and data are adopted by model builders. The first is a deductive reasoning approach: from an existing general theory we deduce the model relationships for the specific case study. The second is an inductive reasoning approach: assuming no a prior knowledge (theory) of process behavior, we attempt to develop the specific information acquired from the particular sample set of data into a general model ([19, 185]). These
two scientific research approaches employed among environmental scientists mainly result in two categories of environmental models — *theory-based* or *data-based* models\(^1\). It would be incorrect to suggest that the analyst has to make a choice between one or the other of these model types; it is better to view them as defining the two ends of the spectrum of models ([127]). A theory-based model (TBM) is closely associated with *a priori* information and with a deductive reasoning process, while the data-based model (DBM\(^2\)) is much more naturally oriented toward *a posteriori* information and inductive reasoning.

### 1.1.1 Dilemmas

It has been the dominant paradigm, however, to construct ‘large’, ‘complex’, theory-based models that are as comprehensive as possible. The modeling of environmental systems has been guided by the principle that were more details to have been included in a given structure, then this would presumably have brought about improved accuracy of prediction. There is a tendency to assume that rigorous science is not being done unless one is developing such a comprehensive model of the given systems ([22]). However, given the currently available field observations of the behavior of environmental systems, such comprehensive mechanistic models, which have become enormously complex assemblies of many hypotheses, cannot be effectively falsified, that is, they suffer from a lack of identifiability or over-parameterization. The model may contain descriptions either of a type of behavior not actually observed in the particular set of data, or of multiple types of behavior, the individual components of which cannot be disentangled from observations of their collective effect. The consequences

\(^1\)Many other pairs of names are used to label the two types of models for this kind of classification, e.g., internally descriptive vs. input-output, white box vs. black box, mechanistic vs. statistic, and physically-based vs. data-based, etc. By classifying them as ‘theory-based’ and ‘data-based’ models, we simply just created another classification scheme.

\(^2\)This acronym is also used by Young and his colleagues (e.g., [82, 259, 262], etc.) for the Data-Based-Mechanistic model(ing). In this research, the data-based model will be restricted to the model derived by the Data-Based-Mechanistic modeling procedure proposed by Young and his groups at Lancaster University, UK. In order to avoid unnecessary confusion, hereafter, the acronym DBM stands for the data-based model, while the Data-Based-Mechanistic modeling procedure is represented by its full name.
are usually apparent in the absence of a uniquely best combination of parameter values that fit the data (many combinations are equally good) and in parameter estimates with high error variances and covariances ([19]).

As an antithesis, a data-based modeling philosophy suggests that the dominant structure of an environmental system should be identified directly from the data using objective methods of statistical inference. In other words, wherever data availability allows, the model builder should progress from these measured data to a ‘small’, ‘simple’, model, which is identifiable, efficiently parameterized, and as objective as possible. This data-to-model approach contrasts with the alternative reductionist methodology and helps to avoid the possibility that prior scientific prejudice and an over-dependence on the current scientific paradigm will lead to an over-parameterized, statistically unvalidated model ([113, 218]). But if such a data-based approach to modeling is to be of maximum potential utility, it has to be carried out carefully. The simple input-output transfer function (TF) model, for instance, can provide a reasonable basis for time series forecasting and control but it lacks the kind of clear, mechanistic interpretation which is essential if the model is to be fully credible as a scientific theory of behavior.

Regarding the management (prediction) purpose for which environmental models are built, O’Neill ([166]) noted long ago that the error in the predictions from a model should decrease with a decreasing degree of model aggregation (in his case, aggregation of the behavior of several species of organisms into a single state variable of an ecological model). He also noted, however, that any increasing refinement of detail would tend to increase the prediction errors resulting from the necessarily increasing number of model parameters with uncertain values. The lack of predictive power, which we would rectify through an enlargement of the model’s state-parameter vector, may instead be substituted by an unacceptably high level of uncertainty ([25]). The following dilemma quoted directly from ([21]) appears just as intractable today as then:
With a large model it may well be possible to predict the “correct” future, but one would have little or no confidence in that prediction. In contrast, with a small [tightly identified] model it may be that a quite “incorrect” future is predicted. And, worse still, one might place considerable confidence in that prediction.

— M. B. Beck ([21])

1.1.2 Opportunities

None of this undesirable situation necessarily prevents us from making progress, however, especially in the respects of data collection and data-based modeling (see, e.g., [185]). The only valid reason for relying upon one of the two modeling approaches alone may be the inevitable difficulty of obtaining experimental or observational data. The inherent limitations in the available environmental time-series data seriously restrict the identifiability of the parameters in mechanistic models and the validity of the statistical assumptions of the data-based model. Fortunately enough, it is now relatively easy to be “data-rich”, if nevertheless “information-poor”, after an accumulating history of automated instruments being deployed in the water industry ([32]). In particular, data can now be sampled at a sufficiently high frequency and for long enough to illuminate and encompass the transient behaviors arising from the wide range of time-constants typical of environmental systems ([25]). The University of Georgia has acquired and commissioned a facility, the Environmental Process Control Laboratory (EPCL), designed to yield just such sets of data ([32]).

Rejecting the deterministic-reductionist philosophy, the DBM models derived through the Data-Based-Mechanistic modeling procedure retain the statistical advantages and parsimony of the data-based modeling approach but address the desire for a physical interpretation of the model. Over the evolution of the Data-Based-Mechanistic modeling philosophy and its methodologies since the 1970’s ([35]), this generic modeling procedure has been utilized over a wide range of disciplines over the last decade, including the fields of engineering, macro-economics and environmental sciences (see, e.g., [82, 266, 263, 259, 262], etc.).
Although the so derived data-based models are only considered credible if they can be interpreted in physically meaningful terms, in terms of seeking an understanding of the system’s observed behavior, the simpler statistical models are not a satisfactory end-point to the reconciliation of a priori theory with field data. It is instinctive to ask how and why certain types of behavior anomalies are observed, not merely to accept that they are. The interpretation of anomalies, and the revision of inadequate hypotheses are not matters that can be resolved without reference to a TBM ([19]). The theory-based model, however, characterizes how the inputs are connected to the states and how, in turn, the states are connected to each other and to the outputs of the system. In other words, they provide a more detailed description of the internal mechanisms of process behavior. Therefore, it is convenient to think of these models at the two extremes as intended for rather different, yet complementary, purposes: the larger, comprehensive, theory-based model as a formalized archive of hypotheses; the simpler, statistically-derived, data-based model as a primary vehicle for the analysis and interpretation of data sets ([19, 22]). Since the two approaches are complementary, more is to be gained from their joint application than from the exclusive use of either model type. As Reckhow and Chapra ([185]) put it:

Just as the theory must be representative of the system (lake) under study, so should the data. And just as the theory should have a basis in a set of rules for the conduct of scientific investigations (the scientific method), so should the analysis of the data.

— K. H. Reckhow and S. C. Chapra ([185])

1.2 Objectives of study

The objective of this dissertation research is to develop a general principle of joint application of the data-based and theory-based modeling approaches presented in Figure 1.1 for characterizing the behavior of complex environmental systems. Through the identification of
model structures and estimation of parameters, this proposed two-pronged approach moves
towards a synthesis of the data-based and theory-based models in parameter space. How
successfully this mission will be accomplished depends on how effectively we answer the
following three questions:

First, does the joint application of the two separate modeling approaches to the same
environmental system allow us to acquire more information about the system? In other words,
does the system analysis based on the data-based modeling provide us with some information
about the system that the analysis based on the theory-based modeling is otherwise not able
to deliver, and *vice versa*?

Second, even if more is to be gained from their joint application than from the exclusive
use of either model type, can we then go further to make use of the information obtained
from one type of model to solve problems of structure identification, parameter estimation
and/or interpretation in the other type of model. That is to say, can the results from the
theory-based modeling be utilized for building the data-based model in such a way that the
data-based model is more physically meaningful? Or conversely, can the identified data-based
model help to improve the identifiability of the corresponding theory-based model?

Third, even if the two types of model can benefit from one another as addressed in
preceding question, can we go far further to reconcile the data-based model and the theory-
based model in terms of model parameters? The parameters of the data-based model (in
a transfer function) are always lumped parameters while those of the theory-based models
(in ordinary differential equations) are not. Such a process of parameter reconciliation is,
therefore, not a one-to-one mapping function. To circumvent this difficulty, we will examine
the *time constant* and *steady-state gain* of each model instead. These two parameters are
familiar in characterizing a linear process in control systems engineering ([68]).

The answers to these questions can be discovered through working with some case studies
of applications of the two-pronged modeling approach to environmental systems. The first
case study is associated with an engineered environmental system, in which the proposed two-
The two-pronged modeling approach is used to analyze the ammonium-nitrogen removal in the aeration tanks of the municipal wastewater treatment plant in Athens-Clarke County, Georgia. The second case study is to apply the two-pronged modeling approach to understand the dynamics of algal blooms in early summer in an aquaculture pond in the southeastern Piedmont, which represents a more natural system.

Figure 1.1: Towards a synthesis of data-based and theory-based models through system identification.

1.3 Approaches of the study

On the construction of a model from, and by reference to, data, i.e., system identification, Ljung ([142]) has presented three basic entities of the system identification procedure: (1)
the data; (2) a set of candidate models; and (3) a rule by which candidates models can be assessed using the data. System identification is essentially the course of reconciling a priori theory and data. Therefore, in fulfillment of the objective of moving toward a synthesis of data-based and theory-based model set out above, the following steps are most crucial.

1. Data collection and experimental design. An appropriate record of the observed process dynamics is a prerequisite of system identification. Ideally, the input and output data are recorded during a specifically designed experiment, where the observer may determine which signals to measure, when to measure, how frequently to measure, how long the measurement campaign should last, and even which type of input signal to use ([19, 142]). The objective of experimental design is to make these choices so that the data obtained from the system become maximally informative, subject to constraints that a priori knowledge of the system may impose. However, in the cases of environmental systems, the experimenter may not have much freedom to affect the system. In other words, our experiments reduce simply to the observation of behavior without any human intervention other than enhancing the excitation signals of some forcing functions by artificial manipulation (e.g., fertilization).

2. Model structure identification. In more specific terms, this is a posteriori model structure identification, which attempts, by reference to the field data, to unambiguously determine how the measured system input disturbances are related to the state variables and how these latter are in turn related to both themselves and to the measured system output responses. In this research, there are two branches in the application of model structure identification. In one case, this is equivalent to testing the ‘success’ or ‘failure’ of any of the individual, constituent hypotheses in the theory-based model constructed from basic physical laws and other well-established relationships of the current paradigm, and to speculate on improved hypotheses if the former fails. In the other case, model structure identification determines an appropriate order of
the data-based model (i.e., the order of the relevant polynomials in the transfer function) and the nonlinear or non-stationary relationships between the input and output signals in the DBM models. In either branch, these are problems different from the straightforward matter of parameter estimation usually associated exclusively with the activity of model calibration. A posteriori model structure identification is no doubt the most important and, at the same time, the most difficult procedure of the system identification. The idea of recursive (state)-parameter estimation can be used to solve this problem ([23]).

3. Parameter estimation. Parameter estimation deals with the computation of values for the parameters that appear in the model (either data-based or theory-based) equations, once the structure of these relationships has been properly identified. A basic principle of parameter estimation is to obtain a set of estimates of the model parameters through the minimization of a certain type of objective function. However, the solution to this problem in this research involves not only the generation of accurate (point) estimates of the model parameters but also some quantification of the variance-covariance matrix of the uncertainty (or confidence) of the estimation. In addition to conventional batch parameter estimation scheme, the recursive scheme of parameter estimation will be also adopted in this research. Moreover, it should be also emphasized that: (1) model structure identification and parameter estimation are closely related and that the former problem can sometimes be solved by recourse to a parameter estimation routine; (2) an exercise in accurate parameter estimation is of dubious significance if the prior problem of model structure identification has not been satisfactorily resolved ([19]).

4. Synthesis in parameter space. It is intuitive for models to be cast and evaluated in the (observed) state space, i.e., the observations of the system’s behavior. However, our interest is in the synthesis of the two categories models — data-based and theory-based models, in the parameter space, through which we can examine the internal mechanism
behind the observed behavior. The parameters of the naturally expressed model of the system’s behavior will be brought out to the surface from being embedded within the state space ([33]). Therefore, through time-varying parameter estimation, the problem of apparent structural change can also be addressed. Once again, this problem is highly related with the previous two, and, in fact, is accomplished in the recursive scheme.

1.4 Structure of the dissertation

The remainder of this dissertation comprises five constituents. The following chapter (Chapter 2) briefly reviews the methods of structure identification and parameter estimation for both data-based and theory-based models. Chapter 3 illustrates the bases of the technicalities of signal processing, Data-Based-Mechanistic modeling and the Recursive Prediction Error (RPE) method. The modified RPE algorithm adapted to time-varying parameter estimation is considered to be a new contribution to this area. These first three chapters are followed by two case studies demonstrating the application of the proposed two-pronged modeling approach for environmental systems. One case study is concerned with nutrient reduction in point-source discharges, while the other addresses the problem resulting from the eutrophication of surface waters. In Chapters 4 and 5 the dynamics of ammonium-nitrogen removal in a municipal wastewater treatment plant and of algal blooms in a southeastern Piedmont impoundment are investigated, respectively. Chapter 6 engages in a retrospective assessment of the preceding two case studies and summarizes the contributions and limitations of this dissertation, on the subject of environmental system analysis, and opens up some further research opportunities for the future.
Chapter 2

A Literature Review

2.1 Beyond toggling between theory- and data-based models

Now that it is well recognized that both the (complex) theory-based and (simple) data-based models for environmental systems have their advantages and disadvantages for modeling poorly defined, complex environmental systems (or ecosystems, earth systems) (see Chapter 1), there has always been an interest in integrating these two basic modeling approaches to address environmental issues ([1, 55, 77, 100, 185, 186, 268]).

The first example of such an integration is a seminal event in modeling the relationship between lake trophic state and excessive nutrient loading. It was over 50 years ago that Sawyer ([197]) recommended inorganic nutrient thresholds above which excessive algal growths might be expected in northern temperate lakes. Since then, the two general types of models — empirical and theoretical — have been used by many modelers to describe the dependence of lake trophic state on external phosphorus loading. Significant developments occurred in both theoretical and empirical models of nutrient enrichment that provided the foundation and direction for much of the work that followed ([186]). The empirical models were derived from statistical treatments of data from large numbers of lakes, and theoretical models were based on more or less detailed mathematical descriptions of the kinetics of nutrient and population dynamics.

Empirical models of nutrient response in receiving waters owe much to Richard Vollenweider ([234, 235]), who found from his knowledge of north temperate lakes that nutrient concentration (trophic) state in a cross section of lakes is a simple function of annual nutrient
loading, lake mean depth ([234]), and water residence time ([235]). For example, after examining data on lake mean depth and lake nutrient loading from over 20 North American Lakes, Vollenweider ([234]) proposed nutrient loading criteria appropriate to the transition zones between trophic states. This result allows an analyst to determine a lake’s probable trophic state by simply referring to a table or graph if one knows its loading and mean depth. Interestingly, this loading table or plot has its theoretical basis ([229]).

There are many good review articles describing the burgeoning model development of both empirical and theoretical eutrophication models in the late 1960s and early 1970s. The empirical phosphorus loading models were treated by Reckhow ([184]), phosphorus-chlorophyll models by Nicholls and Dillon ([157]), several types of models by Kamp-Nielsen ([125]) and Straškraba ([222]); and dynamic theoretical eutrophication models by Scavia and Robertson ([200]), Jørgensen ([122]) and Straškraba and Gnauck ([223]). However, the most noteworthy work in the sense of integration of data-based and theory-based models for lake management was done by Reckhow and Chapra ([55, 185]). These two-volume textbooks not only introduced a variety of methods of empirical modeling (Volume 1, [185]) and principles of theoretical modeling (Volume 2, [55]), but also advocated the use of the combined two-pronged modeling approach (empirical and theoretical) for lake management and decision making.

In the past, while these two modeling approaches were employed almost accidentally to solve environmental problems by different modelers, recently, attempts to embed them within each other were carried out intentionally by the same researchers. In [233], van Tongeren proposed a concept of a “nested model”, after evaluating some of the vices and virtues of the regression, ordination and dynamic theoretical modeling techniques in limnology. The “nested model” is thought to be a possible solution for the problem of changing parameters and for the problem of parameters being different among lakes. This straightforward approach is to combine the virtues of regression analysis and those of dynamic modeling into one model. In such a model the local parameter estimates are obtained from a (descriptive) regression
of the (calibrated or measured) parameters of the same model applied to a number of lakes with different properties. A nested model thus consists of regression models embedded into a dynamic ecosystem model. The range of lakes used for these regression model defines the domain of the model. The full parameter space of the model can easily be derived from all independent sets of parameter estimates.

Reynolds and his colleagues developed the lake phytoplankton community model, PROTECH (Phytoplankton RespOnse To Environmental CHange), including nested regression equations to address the effects of surface/volume ratio, temperature and irradiance on growth rate ([79, 77, 78, 190]). The PROTECH model provides a detailed phytoplankton model with up to eight algal species from a library of eighteen possible species. Although the results for individual species do not correspond very well to observations, the results for the sum of all species lead to good agreement with data.

Young and Parkinson ([268]) developed an integrated statistical modeling procedure to combine the stochastically defined simulation models and data-based statistical models. They argued that this rational combination can provide valuable insight into the nature of poorly defined environmental systems. This general approach is also built on the assumption that the dynamic modeling of environmental systems should involve two basic model types: speculative and normally quite complex simulation models which represent the current, state-of-the-art, scientific understanding of the environmental system; and the DBM models obtained initially from the analysis of observational time-series, but only considered credible if they can be interpreted in physically meaningful terms. The objective statistical derivation of these much simpler DBM models contrasts with the rather subjective formulation of the complex simulation models. However, the two apparently quite different types of model are brought together in a novel approach where a general Data-Based-Mechanistic modeling methodology is used to simultaneously linearize and reduce the order of the complex simulation model, so exposing its ‘dominant modes’ of dynamic behavior. Five major steps of this general Data-Based-Mechanistic modeling strategy are illustrated in [267] or [268] in detail. In summary,
This approach involves three main methodological tools: uncertainty and sensitivity studies based on Monte Carlo simulation techniques; dominant mode analysis using a new method of combined linearization and model-order reduction; and Data-Based-Mechanistic modeling. This novel approach is demonstrated through two practical examples: modeling the global carbon cycle in relation to possible climate change; and modeling a horticultural glasshouse for the purpose of automatic climate control system design.

2.2 System identification

Most of the current comparisons in the literature between theoretical models and empirical models are made literally on the basis of the predictive (or simulative) performance of the models (e.g., [43, 198]). These are sufficient if our sole goal is to predict the behaviors of a system for managing purposes. However, we also want to go further to explore the mechanisms behind these behaviors — why the system gives these behaviors, not others. Therefore, the synthesis of the two types of modeling approaches in this research will be carried out in the framework of system identification, a subject comprising contributions from several disciplines, in particular, engineering, control theory, and mathematics. The rest of this chapter will give a brief review of the recent study of system identification in environmental systems.

System identification is interpreted as the complete process of deriving mathematical models from, and by reference to (experimental) data and prior knowledge or information ([17, 113, 142]). The data are usually time-series observations of system inputs and outputs. Prior knowledge available may include conservation laws, idealized physical equations, model parameter and noise values (as means, bounds or probability distributions), the nature of the system response and possible parameterizations. The whole process of system identification may include experimental design, model structure identification, parameter estimation, sensitivity and uncertainty analysis, model verification and validation. The field of system
identification has developed rapidly since it drew the attentions of systems analysts several decades ago. Bellman et al. ([39]), Koivo and Phillips ([133]), Parker ([170]), and Shastry et al. ([209])’s works are among the earliest contributions ([23]). Ever since, the literature on the problem of system identification has grown extensively. The following short paragraph provides a thumbnail introduction to the literature on the subject of system identification.

The book by Eykhoff ([80]) is to be recommended as providing a broad and comprehensive treatment covering several practical issues, while Söderström and Stoica ([211]) and Ljung ([142]) give more theoretically oriented presentations. Among the multitude of publications on methods of parameter estimation, the text by Young ([253]) provides an excellent introduction to recursive estimation techniques, or alternatively, these same techniques are given a rigorous treatment in Ljung and Söderström ([143]). The proceedings from the International Federation of Automatic Control (IFAC) series of Symposia on Identification and System Parameter Estimation, contain many articles on all aspects of the system identification problem since 1967. The book by Beck and van Straten ([34]) discusses the modeling and identification of environmental systems, while the recent product of the International Task Force on Forecasting Environmental Change led by Beck ([26]) provides a manifesto for how to generate longer-term environmental foresight and cope with structural change. Anderson and Bates ([7]) and Jakeman et al. ([112]) emphasize the role of model validation and model selection in their treatment of system identification. On the other hand, Box and Jenkins’ ([47] or [48]) detailed account of discrete-time, input-output, data-based modeling has had a most significant impact on the application of time series analysis in many diverse technical fields. Other statistical treatments of time series modeling, such as Anderson ([6]) and Brillinger ([50]), are also relevant for addressing the problem of system identification.

2.2.1 Environmental systems

A system is an object in which variables of different kinds interact and produce observable signals, oversimplified as Figure 2.1. The outputs $y$ are the observable signals that are of
interest to us. The inputs $u$ are the external signals that can be measured or even manipulated by the observers. The state variables $x$ determine the system’s status but are not necessarily observable. The parameters $\alpha$ represent how these variables interact with each other. The unmeasurable disturbances $\xi$ are the external stimuli whose influence on the system are observed only through the outputs. The output measurements are inevitably corrupted by the measurement noise $\zeta$, while the measurement errors in the inputs are encapsulated in the disturbance signal. All the variables are functions of time $t$. System identification thus can be broadly defined as the establishment of how the measured system inputs are related to the system’s state variables and how these latter are in turn related both to themselves and to the measured system outputs.

![Figure 2.1: The definition of system and variables.](image)

Since environmental systems are so poorly defined that the investigator is frequently ignorant of the true structure and relationships between the system variables, and the data obtained from the environmental systems are corrupted and not sufficient to identify system structure unambiguously, the task of system identification (or model structure identification, in particular) is thus of vital importance and fraught with enormous difficulties. Referring to the earlier discussion of the complementary roles of the data-based modeling (DBM) and theory-based modeling (TBM) for environmental systems in Chapter 1, one possible solution to reconciling the mechanisms governing the behavior of systems and the data
observed from them is to conduct system identification procedures within the frameworks of both DBM and TBM. Therefore, in the following paragraphs the problems of model structure identification and parameter estimation are reviewed from both perspectives. However, there are no clear-cut lines between the methodologies used for each modeling approach. Some parameter estimation methods (e.g., most recursive estimation methods) are applicable to both approaches; others were invented originally for one, but can be used for the other with modification (e.g., instrumental variables, [19]).

2.2.2 Data-based models

For explanatory purposes the discussions on model structure identification and parameter estimation for data-based models will be restricted to a general Single-Input-Single-Output (SISO) system; it can be extended to multivariate cases without difficulties. Multivariate models are discussed in greater detail in [271]. The behavior of environmental systems is intrinsically nonlinear, but can be approximated by the following time-variable parameter (TVP) transfer-function model ([257]):

\[
y(k) = \frac{B(k, z^{-1})}{A(k, z^{-1})} u(k - \delta) + \frac{E(k, z^{-1})}{A(k, z^{-1})} e(k) \tag{2.1}
\]

where \(y(k)\) and \(u(k)\) are the discretely sampled system output and input signals respectively at the \(k\)th time instant; \(\delta\) represents a pure time delay in the response between output and input; \(e(k)\) is a sequence of independent, Gaussian, random variables; \(A(k, z^{-1})\) and \(B(k, z^{-1})\) are the denominator and numerator polynomials respectively, defined as,

\[
A(k, z^{-1}) = 1 + a_1(k)z^{-1} + \ldots + a_n(k)z^{-n} \tag{2.2a}
\]

\[
B(k, z^{-1}) = b_0(k) + b_1(k)z^{-1} + \ldots + b_m(k)z^{-m} \tag{2.2b}
\]

where \(z^{-1}\) is the backward shift operator, that is \(z^{-i}y(k) = y(k - i)\); and the integers \(n\) and \(m + 1\) are the number of parameters in the respective polynomials. \(E(k, z^{-1})\) is a rational function acting as a ‘shaping filter’, passed through which a white noise \(e(k)\) is transformed.
to account for the combined effects of system noise $\xi$ and measurement error $\zeta$ that can be denoted by a lumped stochastic process $\nu(k)$ for convenience. Details of this stochastic process description will not concern us greatly in this research since we are trying to establish the nature of the deterministic relationship between $u$ and $y$.

Equation (2.1) states essentially that the current value of the output $y(k)$ is a (scalar) function of current and past measurements of the input $u(k)$, of past measurements of the output, and of current and past realizations of the stochastic process $\nu(k)$. Such a data-based model, being specific to the sample data set from which it is derived, is unlikely to be a universal description of a system’s dynamics. However, it provides a first attempt at elucidating any observed basic cause/effect relationships, such as which inputs affect which output, by how much, and how quickly. With assistance of the widely used Data-Based-Mechanistic modeling methods mentioned in the previous chapter, which will be expounded in the following chapters, the data-based (mechanistic) model in formulation (2.1) also delivers a fair reflection of our insight into the internal mechanisms of the system in physical terms.

Model structure identification (and/or parameter estimation) for a data-based model are therefore simplified to seeking input-output relationships, that is: (1) which input variables are in any way significantly related to the output variables; and (2) by how much and how quickly does the output variable respond to changes in a given input variable. In formal terms, we require a definition of the values $n$, $m$, and $\delta$ in Equations (2.1) and (2.2) ([17]); and rather more precisely, we need to determine the values of the time-invariant parameters and the variations of the time-variable parameters in the polynomials (2.2a) and (2.2b). Last it is also necessary to investigate the nonlinearity exhibited in the time-variable parameters.

### 2.2.3 Theory-based models

With *a priori* knowledge and observations, we assume a theory-based model of the environmental system (Figure 2.1) can be defined by the following representation of the state variable dynamics in a continuous-time ordinary differential equation format (denoted as $f$),
while the observation function (denoted by $h$) is represented by a discrete-time observation function as follows ([118, 217]):

$$\dot{x}(t) = f(x, u, \alpha; t) + \xi(t) \quad (2.3a)$$

$$y(t_k) = h(x, u, \alpha; t_k) + \nu(t_k) \quad (2.3b)$$

in which the dot denotes differentiation of the corresponding variable; $f$ and $h$ are vectors of nonlinear functions; $u$, $x$, and $y$ are the input, state, and output vectors, respectively; $\alpha$ is a vector of model parameters, or denoted as $\alpha(t)$ should it be a time-varying parameter (TVP); $\xi(t)$ and $\nu(t_k)$ are notional representations respectively of those attributes of system and output observation that are not to be included in the model in specific form, with both assumed to be white noise processes; especially, $\nu(t_k)$ accommodates the aggregated effects of the unmeasurable disturbances $\xi(t)$ and the measurement noises $\zeta(t)$ (see Figure 2.1); and $t$ and $t_k$ are continuous and discrete time, respectively. Should it be necessary, spatial variability of the system’s state can be assumed to be accounted for by, for example, the use of several state variables of the same attribute of interest at the several defined locations ([33]). The dimensions of $x(t)$, $u(t)$, $\alpha$, and $y(t)$ in Equation (2.3) are designated to $s$, $n$, $p$, and $m$ respectively, thus the dimensions of $\xi(t)$ and $\nu(t_k)$ are also defined.

Therefore, given an adequate number of field observations and fewer gross uncertainties about the observed nature of the system’s behavior, the task of model structure identification within the theory-based model framework is to reconcile the external description of the system, $[u(t_k), y(t_k)]$, with the internal description of the system, $[x(t), \alpha]$ ([24, 30]). The problem is thus to establish how the measured system input disturbances $u$ are related to the system state variables $x$, and how the state variables are in turn related both to each other and to the measured system outputs $y$. Solutions to this problem require a correct identification of the functional relationships $f$ and $h$ in equations (2.3) through iterative processes — alternately seeking to falsify confidently stated hypotheses and then speculating about relatively uncertain (but improved) hypotheses, on the basis of the recursive algorithms of
filtering theory ([17, 19, 23, 30]). In this sense, model structure identification is different from the straightforward parameter estimation which is usually associated with model calibration. Technically, the problem of model structure identification is much more difficult to solve than the latter, although the latter problem is frequently carried out implicitly in the solution of the former ([17]).

2.2.4 Parameter estimation

Parameter estimation deals with the computation of values of the parameters that appear in the models (2.1) and (2.3) once the structure of these relationships has been properly identified. A basic principle of parameter estimation is to obtain the estimates $\hat{\alpha}$ of the model parameters $\alpha$, either by minimizing some function of the errors between the output response observations and the model predictions of those output variables or by maximizing the probability of the observed system behavior. For example, the two most widely applied parameter estimation methods — (nonlinear) least squares and maximum likelihood — are equivalent when the measurement errors belong to white normal distributions with zero means and finite variances.

In terms of how the parameter estimation algorithms are implemented numerically, there are two classifications: recursive (on-line) schemes and batch (off-line) schemes ([19, 20, 23]). Essentially the former permits potential solutions to the prior problem of model structure identification while the latter is a matter of choosing a constrained optimization of parameter values in most cases. The following two sections give an account of these two parameter estimation schemes. The account is not comprehensive but focuses on recent developments and their applications in environmental modeling. Fuller details may be found in [23] and [80].
2.3 Methods for model structure identification

2.3.1 Cross correlation function

Engineering methods for estimating transfer functions, like Equation (2.1), are usually based on the choice of special inputs to the system, for example, step and sine wave inputs ([47]). These methods have been useful when the system is affected by small amounts of noise, but are less satisfactory otherwise, especially for environmental systems. One orthodox solution to the problem of identification of the data-based model (2.1) outlined in Section 2.2.2 is computing sample cross correlation functions $\rho_{uy}$ from the input/output observations $[u, y]$ of the system, as proposed originally by Box and Jenkins in the 1970’s ([47]). The benchmark of the cross correlation function method is that when the input signal $u(t_k)$ approximates a white noise sequence, it can be shown that $\rho_{uy}$ approximates the impulse response function between input and output (e.g., see [47]). In theory, it is then possible to determine, by inspection of the computed impulse response function, the pure time delay $\delta$ and appropriate orders $n$ and $m$ for the coefficients of the polynomials in Equation (2.1). In reality, the statistical properties of $u(t_k)$ do not generally approximate those of white noise, but then $u(t_k)$ may be prewhitened through a ‘shaping filter’ into a white-noise sequence. For the interested reader an exhaustive treatment of solving the identification problem in this manner is given in [47] or [48].

This method has been applied to an analysis of gas production dynamics in the anaerobic digestion of waste municipal/domestic sludge ([17]). However, most frequently, the sample cross-correlation function is employed as a prelude to working with the identification of theory-based models and is used to check if the stochastic residuals of a model are correlated with other variables (‘excluded’ input stimuli) of the system at the model verification step ([262], see [15, 16] for examples).
2.3.2 Instrumental variable series and Data-Based-Mechanistic modeling

In the original motivation to overcome the limitation of linear least squares estimator, a suite of instrumental variable (IV) algorithms has been developed (see e.g., [253]) that provides consistent unbiased parameter estimates which require no \textit{a priori} statistical information regarding the noise sequence. The Simplified Refined Instrumental Variable (SRIV) identification and estimation method is an extension of the original IV estimation procedure which was first introduced by Young ([251]) and a simplification of the Refined Instrumental Variable (RIV) approach (see e.g. [253]). It was shown that, under the assumption that the noise process is a serially uncorrelated series of white noise with Gaussian distribution, the complex RIV estimation algorithm could be reduced to the SRIV form ([254]).

The Data-Based-Mechanistic modeling procedure, proposed by Young and his colleagues (see e.g., [259]), is centered on the utilization of the data-based models (in transfer function form shown as Equation (2.1)) whose parameters are estimated by the combined usage of SRIV and Fixed Interval Smoothing (FIS) algorithms ([253, 257]). The associated methodologies will be illustrated in Chapter 3 and can be found elsewhere (e.g., [259]). Their application ranges from environmental science, through ecology and economics to engineering. For example, the Data-Based-Mechanistic modeling method has been applied successfully to the characterization of imperfect mixing in fluid flow processes, which has led to the development of the Aggregated Dead Zone (ADZ, or called Aggregated Mixing Zone, AMZ) model ([35, 37, 240, 266]). This model is almost always able to match the measured concentration profile much better than previous conventional models, such as the Advection Dispersion Equation (ADE), where the model structure is derived by physico-chemical reasoning. Much more interestingly, the Active Mixing Volume (AMV, see [266, 263]) concept, generalized from the ADZ models, has been applied widely to imperfectly mixing flow processes, such as solute transport in saturated soils, heat flow in soils of horticultural glasshouses, and ventilation in agricultural buildings ([178]).
Data-Based-Mechanistic modeling has been applied equally successfully to the identification of the nonlinear relationship between the effective rainfall and the river flow response over catchments ([256, 263]). The objectively identified transfer function model not only explains the rainfall-flow data in a parsimonious manner but also provides a reasonable physical interpretation ([262]). It suggests that the effective rainfall reaches the river and affects the flow via three pathways, and that the flow is also affected by the prevailing temperature variations. In particular, it indicates that the river flow is composed of five components: a very rapid instantaneous (i.e., within one day) effect; a ‘fast flow’ component with residence time 1.02 days, probably associated with surface processes; a ‘slow flow’ component with residence time 12.21 days, probably associated with subsurface and groundwater processes; and a very slow, ‘base flow’ component in the form of a ‘constant flow’ of 2.05 mm, probably due to deep aquifer effects. The fifth, temperature-dependent component accounts for long-term temperature-dependent effects, such as those arising from evapotranspiration processes.

These two methods described above are mainly designated for the model structure identification of data-based models, while the following two recursive estimation algorithms, the Extended Kalman Filter (EKF) and the Recursive Prediction Error (RPE) algorithms, refer to the problem of model structure identification of theory-based models.

2.3.3 **Extended Kalman Filter**

The approach to identifying environmental systems using the EKF has its origins in the early 1970’s (e.g., [36]), where the EKF was used to infer the structural nature of models for biochemical oxygen demand and dissolved oxygen in rivers. Since then, the use of the EKF in this role has been explored in detail by Beck in a number of papers and chapters in books (e.g., [20, 23, 58, 76, 130]).

The term *filter* can be qualitatively comprehended in the following two aspects ([20]): First, the filter can be seen as an algorithm for *translating* information about the observed
input-output behaviors of the real system into model-related estimates of the state variables and parameters, which is essentially the task of model structure identification. Second, the name filter also suggests the intuitive idea that the algorithm has the objective of filtering from the given field data the unwanted influences of measurement noise and uncertain disturbances. The Extended Kalman Filter ([118]) is an extended version of the linear Kalman Filter (LKF, [89, 124]) referring to the state estimation of linear models. For nonlinear systems this technique requires linearization of the model and simultaneously estimates dynamics of the states and parameters. As in all recursive estimation algorithms used for model structure identification through processing the discrete-time observations, the EKF consists of three components ([23]): (1) a prediction step across the sampling interval $t_{k-1} \rightarrow t_k$, (2) a correction step across the sampling instant $t_k$ as new observations $y(t_k)$ become available, and (3) a feedback gain matrix (or Kalman gain matrix), which provides a weighting procedure for taking account of the mismatch between the estimate and observed values of the system’s responses ($y$).

Despite its extensive applications in a variety of fields including environmental science (e.g., [64, 65, 58, 76, 130]), the EKF is not free of serious limitations ([23, 33, 217]). For example, it is well known that the on-line parameter estimates $\hat{\alpha}(t_k)$ generated by the EKF do not always converge to the true values $\alpha$ when the a priori noise characteristics of the system are not correctly specified ([80]). The failure of the EKF to deliver satisfactory results in identifying the structure of a model is well documented in [23, 24]. Since the convergence analysis of the EKF conducted by Ljung ([141]) came to light, a couple of remedies for the EKF have been proposed in the literature. Most of these remedies merely make small changes to the EKF, attempting to avoid severely biased estimates by either providing noise information of the system and measurements as accurately as possible ([172]) or simultaneously estimating the elements of the Kalman gain matrix ([218]), since the specification of the noises of the system and measurement eventually affects the value of the gain matrix. However, a significant improvement on the recursive estimator achieved by Ljung is the Recursive
Prediction Error algorithm, which would, in principle, eliminate some of the constraints of the EKF ([141, 143]).

2.3.4 Recursive Prediction Error algorithm

A full description of a continuous-discrete version of the RPE algorithm is presented in Appendix B. Originally, this new algorithm was not designed for model structure identification but for parameter estimation. The algorithm has been derived not from a background of filtering theory but as a recursive version of a more classical form of (non-recursive) optimization scheme ([33, 143]). However, when trained onto the problem of model structure identification the RPE algorithm performs really rather well ([217]), especially in illuminating structural inadequacies.

Being recursive, the RPE algorithm is not radically different from the equations of a Kalman filter, or extended Kalman filter ([33]). In summary, its principal advantages would appear to be: (1) performance that is considerably more robust than that of the EKF; (2) fewer arbitrary, prerequisite (prior) assumptions, which have themselves bedevilled performance of the EKF; (3) a better directed goal function; (4) an associated algorithmic structure focusing on the properties of the parameters, not the states; and (5) the potential to facilitate discrimination of structural inadequacy/change in the parameter space from that in the error space.

2.4 Methods for parameter estimation

2.4.1 Regionalized Sensitivity Analysis

It is not uncommon for many modeling exercises to be begun in the absence of an adequate set of field data. There are a few quantitative observations available (probably sampled irregularly and infrequently) together with less quantitative, more qualitative, empirical evidence of the system’s behavior. In such situations it is apparent that the problem of
parameter estimation is to discriminate key from redundant constituent mechanism; at most, perhaps a modest reduction in the spans of the candidate parameter distributions giving an acceptable match of the model with the observed behavior can be achieved. The approach to this problem, as described below, is due collectively to Hornberger, Spear, and Young ([23, 108, 109, 214, 252]) and has been labelled as Regionalized Sensitivity Analysis (RSA) or Generalized Sensitivity Analysis (GSA).

The observed behavior (knowledge) of the system of interest can be usefully exploited to distinguish if the output of the model would be accepted as being as observed ($B$), or as not being observed ($\neg B$). Each speculative constituent hypothesis of the model can be associated with one or more model parameters. A sample realization of the model parameter values can be classified as giving rise to either the behavior ($B$) or its complement, not-the-behavior ($\neg B$). Many such realizations of the parameters selected at random by Monte Carlo simulation will thus yield two marginal distributions $\{\alpha|B\}$ and $\{\alpha|\neg B\}$, the separation of which — as measured by the Kolmogorov-Smirnov statistic — permits a ranking of the relevance of the constituent model hypotheses (parameters) to the generation of observed behavior.

The RSA is stated most generally and succinctly in [109]. It is placed properly in the context of modeling ‘poorly defined’ systems by Young ([252]), and it is best illustrated by a case study of cultural eutrophication in Peel Inlet, western Australia ([108, 214]). Recently, it has also been applied (in a significantly improved form) to the assessment of the reachability of target futures through a case study of aquatic foodweb model in a Piedmont Impoundment (Lake Oglethorpe, Georgia, [29, 167, 168]).

2.4.2 Generalized Likelihood Uncertainty Estimation

The Generalized Likelihood Uncertainty Estimation (GLUE) methodology is a development of the RSA technique as described previously. The GLUE methodology has been developed from an acceptance of multiple behavioral models seen as a combination of a particular set of parameter values within some particular model structure. If there is no unique optimal
model then it will be possible to rank the model structure/parameter sets considered as simulators of the system on some relative likelihood scale. The likelihood measure $L(\alpha_i; y)$ for the $i$th model is associated with a particular set of parameters $\alpha_i$ conditioned on the observed data variables $y$. It is possible to evaluate the sensitivity of individual parameters, either by looking at the distributions of behavioral and non-behavioral models defined by the associated likelihood weights, as in the RSA procedures (e.g., [42]); or by evaluating the marginal distribution of likelihood for each parameter by integrating across the parameter space (e.g., [192]). However, it is worth stressing that it is the contribution of each parameter value within an individual set of parameters that is important in the acceptability of a model.

The GLUE procedure requires first that the sampling ranges be specified for each parameter to be considered. Normally the ranges can be set to be as wide as possible by physical argument or experience. However, it would be hoped that models at the edges of the parameter ranges will be rejected as non-behavioral. Secondly, the GLUE requires a method for sampling in the parameter space. Most of time, this is done by Monte Carlo simulation as in RSA, using uniform random sampling across the specified parameter range. Thirdly, the procedure requires a formal definition of the likelihood measure to be used and the criteria for acceptance or rejection of the models. The GLUE procedure was originally applied to catchment hydrological models ([42]), but has since found more extensive applications in flood studies ([192]), soil-vegetation-atmosphere water transfer (SVAT) studies ([84]), open-channel hydraulics ([96]) and hydro-geochemical modeling ([272]).

It is worth mentioning that the GLUE methodology also recognizes that as more data or different types of data are made available it may be necessary to update the likelihood weights associated with different models. This is achieved quite easily using Bayes’ equation which allows the $a$ priori distribution of model likelihoods to be modified by likelihood weights arising from the simulation of a new data set to produce an updated or $a$ posteriori likelihood distribution, which is essentially a recursive scheme of estimation ([41]).
2.4.3 Controlled Random Search

As multi-parameter sensors are now being widely used in the real-time monitoring of water quality, we can claim no more an impoverishment of data ([25]). However, the difficulties of reconciling large models with behavior as observed in high-volume data sets require an effective search algorithm of optimization. Published in 1979 as a generic mathematical global optimization procedure ([180]), Controlled Random Search (CRS) algorithm is suitable in this situation. The objective of the CRS is to search for the sub-region of the parameter domain that contains the global optimum parameter vector. The CRS starts with a Monte Carlo simulation, and calculates an objective function for each iteration of such a simulation. The CRS iteratively updates a stored set of randomly selected parameter vectors, by searching for and substituting a new trial parameter vector that better fits the model to the available data. The trial vector is obtained by reflecting the centroid of a subset of the current set of parameter vectors about another stored parameter vector. The iteration stops when a specified criterion is achieved, or until no better trial parameter vectors can be generated. The CRS was employed for calibration of a carbon-based ecological model of the Oosterschelde Estuary in The Netherlands ([131]). In [59], the CRS is also successfully applied to moving windows of the 80 days record of River Cam data for the purpose of detecting and diagnosing the possible paths of structural evolution in the observed behavior of the system.

2.4.4 Numerical Optimization Algorithms

In the ideal cases where ‘plenty of data’ are available, numerical optimization algorithms can be used to provide the parameter estimates that minimize the specified objective functions. A numerical optimization algorithm is a logical method that searches the response surface of the parameter hyperspace to locate the best set of parameters. There are many different
numerical optimization algorithms available (e.g., Nelder-Mead, Gauss-Newton, Newton-Raphson and genetic algorithm) and each uses a strategic mathematical method to obtain the estimates of the optimal parameters. Most methods work on the principle of creating an initial guess to the optimal solution and then make iterative attempts to enhance this guess following logical procedures defined by the strategy employed. These methods can be crudely categorized into two groups: *local* methods and *global* methods. The Nelder-Mead and Newton methods are examples of local methods while the genetic algorithm and the CRS described above are instances of the global methods.

Furthermore, the local methods can be broadly categorized in terms of the derivative information of the objective function that is, or is not, used. *Search* methods that use only function evaluation (e.g., the simplex search of Nelder-Mead [156, 147]) are most suitable for problems that are highly nonlinear or have a number of discontinuities. *Gradient* methods are generally more efficient when the function to be minimized is continuous in its first derivative. Gradient methods utilize information about the slope of the objective function to dictate the direction of search towards where the minimum is thought to lie. Most gradient search strategies are based on the following specific form as a so-called steepest-descent algorithm ([213]),

$$\alpha_{i+1} = \alpha_i - \rho \cdot \Gamma \cdot \left. \frac{\partial J}{\partial \alpha} \right|_{\alpha = \alpha_i}$$

where \(\alpha_{i+1}\) is the generated new point in the parameter hyperspace; \(\alpha_i\) is the current point; \(\left. \frac{\partial J}{\partial \alpha} \right|_{\alpha = \alpha_i}\) is the gradient matrix at the present point; \(\rho\) determines the step size; and \(\Gamma\) is a specifically chosen square matrix. It can be shown mathematically, that if \(\Gamma\) is a positive definite matrix, then the vector from \(\alpha_i\) to \(\alpha_{i+1}\) will give a better cost functional value ([213]). In the case of quasi-Newton methods, \(\Gamma\) is defined using an *approximation* of the Hessian matrix. If the newly defined point in the parameter space provides a lower functional value than the previous point, then it replaces the previous one. If the new functional value
is higher, the step size is reduced and a new point is evaluated. At the optimal point in the parameter hyperspace the gradient will be close to zero and the search is terminated. Local search methods are limited in cases where the parameter hyperspace has a complex and irregular response surface such that they often terminate in local rather than global minima. For this reason, the success of local search methods is often dependent upon the location of the search initiation.

Global methods avoid this sub-optimal, local optimization by continually looking away from the local optima that the algorithm is working towards. A generic algorithm (GA) ([103], see also [91, 241]) is a stochastic global search method which uses the biological concepts of ‘natural evolution’ and ‘survival of the fittest’ to obtain optimal parameter estimates. GA works with a collection of randomly selected populations of potential solutions rather than a single solution. Each population is comprised of a number of ‘chromosomes’ containing ‘genes’, which represent each parameter to be estimated. Each individual is assessed within a defined objective function and assigned a ‘fitness value’ which in turn, determines whether it is selected for breeding. The fittest individuals are selected and undergo a breeding process using ‘crossover’ and ‘mutation’ operators to generate a new population of ‘chromosomes’ which are subsequently evaluated. The process of mutation randomly alters the configuration of the chromosome maintaining its genetic diversity in order to prevent premature convergence as a consequence of a local optimum. The GA evolves a population of chromosomes over many cycles or generations until a specified termination criterion is met.

Another global method favored by many practicing hydrologists is the SCE-UA (abbreviation for the Shuffled Complex Evolution method developed at The University of Arizona) global optimization algorithm, developed by Duan et al. ([73, 74]) in the early 1990’s. The SCE-UA strategy combines the strengths of the simplex procedure of Nelder-Mead ([156]) with the concepts of controlled random search (CRS, [180]), competitive evolution in GA ([103]) and the complex shuffling concepts ([73, 74]). This method was designed specifically for the purpose of dealing with the peculiar problems encountered in conceptual watershed
model calibration. Extensive testing on a simplified research version of the NWSRFS-SMA\textsuperscript{1} model (i.e., the SIXPAR, six parameter, conceptual model) and the full-scale NWSRFS-SMA model revealed that the SCE-UA method was both effective and efficient, compared with other existing global methods, including the ARS (Adaptive Random Search) and the multi-start simplex method ([73, 212]).

The essence of the method is presented as follows. It begins with a population points sampled randomly from the feasible parameter space. The population is partitioned into several communities, each containing $2n + 1$ points where $n$ is the dimension of the problem. Each community is made to evolve based on a statistical “reproduction” process (concept from the GA algorithm) that uses the “simplex” geometric shape to direct the search in an improvement direction (procedures of the Nelder-Mead algorithm). At periodic stages in the evolution, the entire population is shuffled and points are reassigned to communities to ensure information sharing. As the search progresses, the entire population tends to converge toward the neighborhood of global optimum (concept from the CRS algorithm), provided the initial population size is sufficiently large. It is worth noting that the shuffled complex evolution method itself has been evolving to several new forms, such as the SCEM-UA (Shuffled Complex Evolution Metropolis, [237]), MOCOM-UA (Multi-Objective Complex Evolution, [250]) and MOSCEM-UA (Multi-Objective Shuffled Complex Evolution, [238]).

\textsuperscript{1}The abbreviation for the Soil Moisture Accounting portion of the National Weather Service River Forecast System.
Chapter 3

Methodological Basis

The objectives of synthesizing the data-based and the theory-based models laid out in Chapter 1 require systematic parameter estimation of both data-based and theory-based models. Among the variety of recursive and en bloc parameter estimation methods reviewed in Chapter 2, the Data-Based-Mechanistic modeling procedure using a derivative of the Instrumental Variable (IV) method as its core parameter estimation algorithm is employed to identify the time-variant nature of the parameters in the data-based model (DBM), while the Recursive Prediction Error algorithm, modified by the author to accommodate time-varying parameter estimation, is used in the theory-based model (TBM) structure identification (Sections 3.2 and 3.3). After the structures of both DBM and TBM are identified successfully, the time constant and steady-state gain are derived from their linear forms of structure (Section 3.4) for investigating the comparative performance of the two types of models. Before the raw data sets are used for parameter estimation procedures, however, they are pre-processed to interpolate missing data points, to remove outliers and to extract important signal components (Section 3.1).

3.1 Signal processing

Interpolating missing data, diagnosing outliers and extracting signal components (e.g., trend, seasonal variation) from environmental time-series are accomplished through application of the Dynamic Harmonic Regression (DHR) model to the time-series \( y(k) \). The basic DHR model contains the trend, cyclical, seasonal and white noise components expressed in (3.1)
\[ y(k) = T(k) + C(k) + S(k) + e(k); \quad k = 1, 2, \ldots, N \]  
(3.1)

where \( N \) is the total sample size; \( e(k) \) is an ‘irregular’ component, usually defined as a serially uncorrelated and normally distributed Gaussian sequence with zero mean and variance \( \sigma^2 \); \( T(k) \) is a trend or low frequency component; \( S(k) \) is a seasonal component with period different (normally shorter) from that of any cyclical or quasi-cyclical component, which is defined in (3.2); and \( C(k) \) is such a sustained cyclical or quasi-cyclical component. In this study, the cyclical component with longer period was not included in the DHR model (3.1) since the time-series under consideration is only a month long.

\[ S(k) = \sum_{i=1}^{R} \left\{ a_i(k)\cos(\omega_i t) + b_i(k)\sin(\omega_i t) \right\} \]  
(3.2)

where \( a_i(k) \) and \( b_i(k) \) are stochastic time-varying parameters and \( \omega_i, i = 1, 2, \ldots, R \) are the fundamental and harmonic frequencies associated with the seasonality in the time-series. The frequency values are chosen by reference to the empirical spectral properties of \( y(k) \) using standard methods of spectral analysis ([181]). The trend component \( T(k) \) can also be considered as a stochastic time-varying ‘intercept’ parameter in the DHR and therefore incorporated into the seasonal component as a zero-frequency term ([260]). This DHR model can be considered as a straightforward extension of the classical, constant-parameter, Harmonic Regression (or Fourier series) model, in which the gain and phase of the harmonic components can vary as a result of estimated temporal changes in the parameters \( a_i(k) \) and \( b_i(k) \).

In general, each of these TVP’s \( (a_i(k), b_i(k) \text{ and } T(k)) \) is modelled as a Generalized Random Walk (GRW) process (see discussions in Sections 3.2.2 and 3.3.2). The subsequent two-pass recursive TVP estimation procedures are the same as for the Dynamic Transfer Function models discussed in Section 3.2.2 and are embedded in the CAPTAIN Toolbox implanted in MATLAB® and provided by the Centre for Research on Environmental Systems (CRES) of Lancaster University (http://www.es.lancs.ac.uk/cres/). As mentioned above, the
identification of the number and values of the fundamental and harmonic frequencies associated with seasonality (3.2) is accomplished by reference to the empirical spectral properties of \( y(k) \) using standard methods of spectral analysis. The empirical Auto-Regressive (AR) spectrum of the data is established with the AR order identified from the Akaike Information Criterion (AIC: see [3, 181]) defined in Appendix A.3.

The results of an illustrative exercise are shown in Figures 3.1 through 3.5 for the signal processing of the DO concentration time-series measured in a fishery pond experiment (further described in Chapter 5). Figure 3.1 displays the 36 days, 432 observations of pond DO concentration time-series data. One of the advantages of the DHR approach is its recursive, stochastic, state-space formulation, which provides great freedom in the specification of the model and the smoothing characteristics ([260]). It can function very well, even if there are quite large segments of missing data, providing excellent interpolation over such gaps; and it handles other aberrations, such as outliers, sharp jumps in level, etc. In order to show the robustness and flexibility of the DHR analysis, a gap of two days observations has been artificially introduced into the series, which is drawn as the dashed line in the figure.

The first step of analysis is to determine the order of the AR spectrum of the data by the AIC. Shown by Figure 3.2, the AR(14) has the lowest AIC value. Therefore the AR(14) spectrum of the DO time-series is obtained in the spectral analysis (Figure 3.3). The pronounced power at a long period (> 2 days) arises from the presence of the trend; while the peaks at periods of 12 and 6 samples/cycle are associated with the diurnal and semi-diurnal oscillations in the DO time-series (the sampling intervals are 2 hours). Here 12 samples/cycle is the fundamental period of the oscillation, while the 6 samples/cycle is one of the harmonics of 12. Apparently, the peak at the one-day period is the most significant and dominates the spectrum. The trend and the seasonal components (sum of the diurnal and semi-diurnal variation) have been subsequently extracted from the original data (Figure 3.4). The trend signal shows that an increase in the magnitude of the DO concentration started around July 6th due to the algal bloom triggered by the fertilization applied on
Figure 3.1: The raw data of DO concentrations taken from a fishery pond.

July 5th. The variable magnitude of the seasonal component shows that the production and consumption of dissolved oxygen by phytoplankton are enhanced when algae were blooming after the fertilization. In addition, the power of the interpolation and smoothing of the DHR model is demonstrated in Figure 3.5.
Figure 3.2: Determination of the AR order of the DO time-series.

Figure 3.3: The spectrum of the DO time-series.

Figure 3.4: Signal extraction of the DO time-series: (a) trend, (b) seasonal component.

Figure 3.5: Interpolation and smoothing of the raw DO time-series.
3.2 Data-Based-Mechanistic modeling

Many environmental systems are non-stationary and nonlinear in nature. The Data-Based-Mechanistic methodology to modeling nonlinear systems, which has been successfully applied in the areas of economics, ecology, biology, engineering and environmental science ([257, 258, 263, 269]), is therefore presented. Referring to Figure 3.6, the first step of the Data-Based-Mechanistic methodology is to determine if the observed system is indeed a nonlinear system by either analyzing the residuals of the best SRIV identified constant-parameter linear TF model or allowing the parameters of a simple linear TF model to vary over time. If the system is ascertained to be nonlinear in the previous step then the (non-parametric) estimation of the time-varying parameters in the model can be carried out with the assistance of the powerful Fixed Interval Smoothing (FIS) algorithm (Appendix A.2) in the second step. Any parameter that is found to be significantly time-variant over the observational interval may reflect the non-stationary or nonlinear behavior of the system. Therefore, in the third step, the State Dependent Parameter (SDP) modeling procedure (illustrated in Section 3.2.3) proceeds to investigate whether the identified temporal parameter variations are state dependent and can be efficiently parameterized. Finally, having identified the structure of the nonlinear dynamic model, the fourth step is to re-estimate all the model parameters against the time-series via the SRIV method and nonlinear optimization. All the statistical methods used in the Data-Based-Mechanistic modeling procedure are also implemented in the captain Toolbox mentioned in Section 3.1.
Figure 3.6: Schematic diagram of the Data-Based-Mechanistic modeling procedure.
3.2.1 Parameter estimation of linear systems

Transfer function models

For a linear system, the general linear Single-Input-Single-Output (SISO) Transfer Function (TF) model takes the following discrete-time form:

\[ y(k) = \frac{B(z^{-1})}{A(z^{-1})} u(k - \delta) + \nu(k) \tag{3.3} \]

where \( y(k) \) and \( u(k) \) are the discretely sampled system output and input signals respectively at the \( k \)th time instant; \( \delta \) is the system pure time delay; \( \nu(k) \) represents uncertainty in the relationship arising from a combination of measurement noise, the effects of other unmeasured disturbances and modeling error; and \( A(z^{-1}) \) and \( B(z^{-1}) \) are the denominator and numerator polynomials respectively, defined as,

\[ A(z^{-1}) = 1 + a_1 z^{-1} + \ldots + a_n z^{-n} \tag{3.4a} \]
\[ B(z^{-1}) = b_0 + b_1 z^{-1} + \ldots + b_m z^{-m} \tag{3.4b} \]

where \( z^{-1} \) is the backward shift operator, that is \( z^{-i} y(k) = y(k - i) \); and the integers \( n \) and \( m + 1 \) are the number of parameters in the respective polynomials.

Parameter estimation using SRIV method

Providing sufficiently informative input and output time-series are retrieved from the system, the constant parameters of the TF model defined above must be estimated. There are a great number of different algorithms for parameter estimation, most of which are based upon the formulation of a objective function from the TF model equations which, when minimized, provides the optimal estimates of the model parameters (see Chapter 2, [23]). Perhaps the most well known method is the linear least squares algorithm, which provides good model parameter estimates, provided the measurements from the system are obtained without any stochastic noise. However, in practice, the observed data are usually corrupted by undesirable
structured ‘colored’ noise. The linear least squares algorithm acts to amplify the effects of the noise during the estimation process. As a result, the presence of any noise in the data will cause the parameter estimates to be asymptotically biased and statistically inconsistent\(^1\) regardless of the length of the time-series \((N)\) utilized. For that reason, several methods have been proposed that may lead to consistent estimates of the parameters. For example, the method of “instrumental variables (IV)” was presented in many regression textbooks (see [119]). Under this approach, an error-free variable that is highly correlated with the independent variable is used as a substitute independent variable for parameter estimation. An extension of the original IV procedure, the Simplified Refined Instrumental Variable (SRIV) estimation method (outlined in Appendix A.1), along with the Maximum Likelihood (ML) optimization procedure, has been adopted in the present research for the parameter estimation of the linear TF model (3.3).

**Model order identification**

Having estimated the parameters of a variety of different models, an optimal model structure must be selected. Model order identification, namely the process of identifying the most appropriate values of \(n, m\) and \(\delta\) in model (3.3), is chiefly undertaken, although not exclusively, with the assistance of carefully selected objective statistical criteria. In combination, these objective methods should provide both a measure of how well the model output explains the data and indicate the presence of model over-parameterization. In this study, model identification is based upon the combined reference to the coefficient of determination, \(R^2_T\), Young Identification Criterion, YIC, and Akaike Information Criterion, AIC, further illustrated in Appendix A.3. Additionally, the philosophy underpinning the Data-Based-Mechanistic modeling should not be disregarded during the identification stage. A model structure that has

\(^1\)A consistent estimate is one for which the bias and variance both approach zero as a \(n\) approaches infinity.
clear physical interpretation may be favored for selection over an alternative structure model with slightly superior identification statistics.

3.2.2 Parameter estimation of nonlinear systems

The behavior of a discrete-time, nonlinear dynamic systems can be generally represented by a stochastic, dynamic equation such as,

\[ y(k) = \mathcal{J}\{\chi(k), \mu(k)\} \tag{3.5} \]

where \( y(k) \) is the measured output of the system under consideration; and \( \mathcal{J}\{\cdot\} \) is a reasonably behaved and potentially nonlinear function of the variables in a non-minimum state space (NMSS) (see, e.g., [82], [258], and the references therein) defined by a NMSS state vector of the form,

\[ \chi(k) = \begin{bmatrix} y(k - 1), \ldots, y(k - n), u(k)^T, \ldots, u(k - m)^T, U(k)^T, \ldots, U(k - q)^T \end{bmatrix}^T \tag{3.6} \]

Here, \( \chi(k) \) is composed of the past values of \( y(k) \), as well as present and past values of a deterministic input (or exogenous) variable vector \( u(k) \), and the present and past values of a vector \( U(k) \) of other input variables. Finally, \( \mu(k) \) is an unobserved, zero-mean, stochastic process, which is considered as the source of all stochasticity in the system and is assumed to be independent of the input variables in \( u(k) \) and \( U(k) \). This model assumes that \( y(k) \) is causally related to the primary input variables in \( u(k) \), but the vector \( U(k) \) represents any other associated variables that may affect the system nonlinearly but whose relevance may not be clear prior to the analysis.

For explanatory purposes, we consider a nonlinear system with only one primary input variable. Following the arguments similar to those presented in [257], it is reasonable to assume that the nonlinear behavior of the system (3.5) can often be approximated by a
linear TF with *time-varying* parameters, which can be written as,

\[
y(k) = \frac{B(k, z^{-1})}{A(k, z^{-1})} u(k - \delta) + \nu(k) \\
= \frac{b_0(k) + b_1(k) z^{-1} + \cdots + b_m(k) z^{-m}}{1 + a_1(k) z^{-1} + \cdots + a_n z^{-n}} u(k - \delta) + \nu(k) \quad (3.7)
\]

The stochastic noise term \(\nu(k)\) in (3.7) arises from the stochastic disturbance \(\mu(k)\) in (3.5) and is assumed to be independent of the input variables in \(u(k)\) and \(U(k)\).

For the purpose of TVP estimation, (3.7) can be conveniently rewritten in the alternative vector format,

\[
y(k) = z(k)^T a(k) + \nu(k); \quad k = 1, 2, \ldots, N \quad (3.8)
\]

where \(\nu(k)\) is assumed initially to be white noise with variance \(\sigma^2\); and the state and parameter vectors \(z(k)\) and \(a(k)\) are defined respectively as,

\[
z(k) = [-y(k - 1), -y(k - 2), \ldots, -y(k - n), u(k), u(k - 1), \ldots, u(k - m)]^T \quad (3.9a)
\]
\[
a(k) = [a_1(k), a_2(k), \ldots, a_n(k), b_0(k), b_1(k), \ldots, b_m(k)]^T \quad (3.9b)
\]

The variations of TVP’s in \(a(k)\) are assumed to be characterized by the GRW processes defined in the following terms ([260]),

\[
a(k) = Fa(k - 1) + G\eta(k) \quad (3.10)
\]

where \(a(k)\) is a ‘surrogate’ state vector characterizing the complete temporal evolutions of the parameters; \(F\) and \(G\) are transition and input matrices respectively and can be time variable; and \(\eta(k)\) is a vector of serially uncorrelated, random noise with zero mean and covariance \(Q\). An overall State Space form of model can then be constructed for the *time-varying* parameter estimation by combining (3.8) and (3.10).

‘State’ Equations: \(a(k) = Fa(k - 1) + G\eta(k)\) \quad (3.11a)
‘Observation’ Equations: \(y(k) = H(k) a(k) + \nu(k)\) \quad (3.11b)
with the observation vector $H(k)$ set equal to $z(k)^T$.

It has been shown ([253]) that for estimation purposes, the ratio of $Q$ and $\sigma^2$ is more important than their own individual values. Therefore, the Noise Variance Ratio (NVR) matrix $Q_r$ is introduced into the estimation procedure and is defined as,

$$Q_r = Q/\sigma^2$$

(3.12)

The NVR matrix $Q_r$ is normally chosen to be diagonal and the optimal values of the $(n + m + 1)$ diagonal entities can be either estimated via Maximum Likelihood optimization or selected manually when ML occasionally fails due to the flat nature of the likelihood surface.

The time-varying parameters are then estimated using a two-pass operation (i.e., forward filtering and backward smoothing) on the system time-series, with the forward filtering routine being defined as (3.13) and (3.14) and the backward smoothing being the Fixed Interval Smoothing algorithm outlined in Appendix A.2.

1. Prediction step:

$$\hat{a}(k|k-1) = F\hat{a}(k-1)$$

(3.13a)

$$P(k|k-1) = FP(k-1)F^T + GQ_rG^T$$

(3.13b)

2. Correction step:

$$\hat{a}(k) = \hat{a}(k-1) + P(k|k-1)H(k)^T[I + H(k)P(k|k-1)H(k)^T]^{-1}$$

$$[y(k) - H(k)\hat{a}(k|k-1)]$$

(3.14a)

$$P(k) = P(k|k-1) - P(k|k-1)H(k)^T[I + H(k)P(k|k-1)H(k)^T]^{-1}$$

$$H(k)P(k|k-1)$$

(3.14b)

Subsequently, the FIS algorithm provides estimates $\hat{a}(k|N)$ of the model parameter vector $a$ conditional on the time-series data $\{y(k), u(k)\}$, over the whole observation interval. In addition, if the noise $\nu(t)$ is assumed to be a zero-mean sequence of serially uncorrelated
random variables (discrete white noise), it also yields a good estimate of the covariance matrix \( P^*(k|N) = E\{ \tilde{a}(k|N)\tilde{a}^T(k|N) \} \), where \( \tilde{a}(k|N) = a(k|N) - \hat{a}(k|N) \) is the estimation error. If \( \nu(t) \) is not a white noise, then the matrix \( P^*(k|N) \) will not provide an accurate estimate of the covariance properties of the FIS estimates \( \hat{a}(k|N) \). Nevertheless, it still provides information on the relative accuracy of the parameter estimates which can prove useful in evaluating the detailed nature of the estimated parameter variations ([262]).

### 3.2.3 State dependent parameter modeling

If all the estimated parameters appear relatively stationary over the observation interval, then it can be inferred that the system is predominantly linear. Otherwise, if any parameter exhibits significant temporal variation, further analysis should ensue with the aim of investigating the nature of the estimated parameter variation and attempting to identify whether the changes can be related to any of the variables in the NMSS vector \( \chi(k) \), defined by Formulation (3.6). One of the possible approaches to investigating such state dependency is to assume that \( a(k) \) is linearly related to possibly nonlinear functions of \( \chi(k) \) ([257, 258, 263]), that is,

\[
a(k) = M\{\chi(k)\}\theta
\]

where \( M(k) = M\{\chi(k)\} \) is a transformation matrix of possibly nonlinear functions that are dependent on the variables in \( \chi(k) \), and \( \theta \) is a time-invariant parameter vector.

The identification of appropriate nonlinearly transformed state variables from \( M(k) \) is fundamental to the success of the SDP modeling process. Statistical methods such as scatter plots and correlation analysis, in which the elements of \( M(k) \) are compared to the TVP vector \( \hat{a}(k|N) \), can be used in this selection process. However, it must be strongly emphasized that, in line with the Data-Based-Mechanistic modeling philosophy, the state variables and

---

2This is a general term; it could well be linear.

3However, some applications (e.g., adaptive control and forecasting) may actually require the parameter vector \( \theta \) to be time-variant ([82, 257, 263]). In that case, one of the simplest assumptions is to allow \( \theta(k) \) to vary as a vector random walk (RW).
nonlinear functions chosen must have some realistic physical interpretation in relation to
the nature of the system under consideration, but not necessarily comply with the current
scientific paradigms ([240, 263]).

Having determined the nature of the nonlinear elements in the transformation matrix
$M(k)$, we can estimate the parameters that characterize these nonlinear elements and the
time-invariant $\theta$ by assuming that,

$$\hat{a}(k|N) = M(k)\theta + \varepsilon(k) \quad (3.16)$$

where $\varepsilon(k)$ is a zero-mean, white noise vector with covariance matrix $P^*(k|N)$ as defined
earlier, which is introduced to allow for the uncertainty in the estimate $\hat{a}(k|N)$. Exploiting
the knowledge of the $P^*(k|N)$ matrix, $\theta$ can be estimated by minimizing the following
weighted least squares (WLS) objective function,

$$J = \sum_{k=1}^{k=N} \left[ \hat{a}(k|N) - M(k)\theta \right]^T W(k) \left[ \hat{a}(k|N) - M(k)\theta \right] \quad (3.17)$$

where $W(k) = P^*(k|N)^{-1}$ is normally based on the diagonal entities. The introduction of
$P^*(k|N)$ into (3.17) ensures that in estimating $\theta$, a changing estimate of the uncertainty
associated with the parameter vector $\hat{a}(k|N)$ is incorporated.

It is worth mentioning that the process of SDP modeling can sometimes be enhanced
by sorting the data set in a non-temporal order prior to the estimation of the time-varying
parameters. In this manner, the rapid natural variations between the input and output
time-series are effectively mitigated in the sorted data space such that the evolutions of the
estimated TVP’s are much smoother. The process of SDP modeling can then proceed to
investigate the relationship between the smoother estimated time-varying parameters with
appropriate variables taken from the NMSS vector (sorted in an identical manner) ([82, 262]).
3.3 Structure identification of mechanistic models

3.3.1 Model forms

The task of system identification in environmental systems analysis is to reconcile the observations of inputs and outputs, or the external description of the system, with the model structure, or internal description of the system, based on a priori theory. Unfortunately, this internal description is often an erroneous representation of the true system. It shall be represented in a continuous-time ordinary differential equation format (denoted as $f$), while the observation function (denoted by $h$) is represented by a discrete-time observation function as follows ([118, 217]):

$$\dot{x}(t) = f(x, u, \alpha; t) + \xi(t) \quad (3.18a)$$
$$y(t_k) = h(x, u, \alpha; t_k) + \nu(t_k) \quad (3.18b)$$

in which the notations and their definitions are the same as in Equations (2.3).

The innovations model

As the departure point of the derivation of the RPE algorithm, however, the innovations model of the behavior of the environmental system is defined as follows ([217]):

$$\dot{x}(t) = f(x, u, \alpha; t) + K\nu(t) \quad (3.19a)$$
$$y(t_k) = h(x, u, \alpha; t_k) + \nu(t_k) \quad (3.19b)$$

with all the notations and their dimensions defined as in the conventional definition (3.18) of the model, except that the dimension of the Kalman-like gain matrix $K$ that directly relates the observation noise sequence $\nu(t_k)$ to the system noise $K\nu(t)$ (counterpart of $\xi(t)$ in (3.18)) is $n \times m$. The innovations model is noted to be preferable for the purpose of identifiability ([217]). $\nu(t_k)$ is estimated using the innovations $\epsilon(t_k)$ — the difference between model output
\( \hat{y}(t_k) \) and the observations \( y(t_k) \), i.e.,

\[
\epsilon(t_k) = y(t_k) - h(\hat{x}, u, \hat{\alpha}; t_k) \tag{3.20}
\]

Therefore the variance-covariance matrix associated with \( \nu(t_k) \) includes both the measurement and system noise.

In the course of the derivation of the Recursive Prediction Error (RPE) algorithm the observation function (3.19b) is assumed linear in its argument for convenience, i.e.,

\[
y(t_k) = H(\alpha; t_k)x(t_k) + \nu(t_k) \tag{3.19b'}
\]

The RPE algorithm is outlined in Appendix B and the interested reader is encouraged to refer to [217] for the detailed derivation of the algorithm.

3.3.2 TVP estimation in RPE

As reviewed in Chapter 2, various methods of parameter estimation have been developed. Among the recursive schemes of parameter estimation, the RPE allows the estimates of parameters to converge rapidly to their true values ([217]). It also can be extended so that estimation of a time-varying parameter is possible. Solutions to this problem have been provided in [216]: one of them is to introduce an exponential weighting factor of the past data, better known as Exponential-Weighting-Into-The-Past (EWP) ([253]), so that the most recent data influence the estimate of the parameter vector most strongly. However, in practice, this kind of method can be proven to be a somewhat inflexible approach to the problem of estimating the time-varying parameters, since it only allows the analyst to specify a single forgetting factor for all TVP’s. If there happened to be a priori information available which suggested that certain of the parameters varied at different rates, or indeed, if it was known that only some of the parameters varied while others were constant, then it would be difficult for the analyst to make use of this information. From the author’s experience, this exponential weighting factor solution also causes computational instability of the algorithm.
These disadvantages can be circumvented by characterizing the variations of the TVP’s in some stochastic manner.

The evolution of each TVP is assumed to be described by a GRW process defined in the following State Space (SS) terms ([260]),

\[ \alpha_i(t_k) = F_i \alpha_i(t_{k-1}) + G_i \eta_i(t_k), \quad i = 1, 2, \cdots, (p + n \times m) \tag{3.21} \]

where,

\[ F_i = \begin{pmatrix} \omega & \beta \\ 0 & \gamma \end{pmatrix}, \quad G_i = \begin{pmatrix} \delta & 0 \\ 0 & \rho \end{pmatrix} \tag{3.21'} \]

\((p + (n \times m))\) is the number of the parameters in (3.19) including the elements in K matrix; \(\eta_i(t_k) = [\eta_{1i}(t_k), \eta_{2i}(t_k)]\), is a \(2 \times 1\), zero-mean, white-noise vector that allows for stochastic variability in the parameters and is assumed to be characterized by a (normally diagonal) covariance matrix \(Q_{\eta_i}\). This general model covers as special cases the Random Walk (RW: \(\beta = \gamma = \rho = 0; \omega = \delta = 1\)), the Integrated Random Walk (IRW: \(\omega = \beta = \gamma = \rho = 1; \delta = 0\)), and the intermediate case of a Smoothed Random Walk (SRW: \(0 \leq \omega \leq 1; \beta = \gamma = \rho = 1; \delta = 0\)), etc.

Having introduced the GRW models for the parameter variation, an observation equation linear in the parameters and formulated in discrete time \((t_k)\) would be required ([33]), i.e., in the form of\(^4\)

\[ y(t_k) = H(x, u; t_k) \alpha(t_k) + \nu(t_k) \tag{3.22} \]

Referring to the system defined in the form of (3.18), Beck et al. ([33]) proposes that \(H(x, u; t_k)\) can be obtained — in some manner (analytical or numerical) — from the equations of state (3.18a) and the structure of \(h\{\cdot\}\) (3.18b) for model representation (3.18).

Heuristically, the (time-varying) matrix \(H(x, u; t_k)\) in Equation (3.22) can also be obtained.

\(^4\)The careful reader should not confuse the observation matrix here with \(H(\alpha; t_k)\) in (3.19b’), which is defined in Appendix B, relating the observations \(y(t_k)\) to state variable \(x(t_k)\) instead of parameter \(\alpha\).
from the structure of \( h\{\cdot\} \) and the equations of state for the innovation model, i.e., (3.19a). An overall SS model can, therefore, be constructed straightforwardly by the aggregation of the subsystem matrices defined in (3.21), with the ‘observation’ equation defined by (3.22).

‘State’ Equations: \[
\alpha(t_k) = F\alpha(t_{k-1}) + G\eta(t_k) \quad (3.23a)
\]

‘Observation’ Equations: \[
y(t_k) = H(x, u; t_k)\alpha(t_k) + \nu(t_k) \quad (3.23b)
\]

If \( l = 2 \times (p + (n \times m)) \), then \( F \) is an \( l \times l \) block diagonal matrix with blocks defined by the \( F_i \) matrices in (3.21’); \( G \) is an \( l \times l \) matrix constructed by the concatenation of the corresponding subsystem matrices \( G_i \) in (3.21’) as well; \( H(x, u; t_k) \) is a *time-varying* matrix that relates the observations \( y(t_k) \) to the ‘state’ variables (model parameters, including \( K \) matrix entities) defined by (3.23a); \( \eta(t_k) \) is an \( l \) dimensional vector containing, in appropriate locations, the white noise input vectors \( \eta_i(t_k) \) (‘system disturbances’ in normal SS terminology) to each of the GRW models in (3.21). These white noise inputs, which provide the stochastic stimulus for parametric change in the model, are assumed to be independent of the observation noise \( \nu(t_k) \) (with variance-covariance matrix of \( \Lambda \)) and have a covariance matrix \( Q \) formed from the combination of the individual covariance matrices \( Q_{\eta_i} \).

Therefore, the forward pass recursive estimation algorithm for the *time-varying* parameters and the corresponding variance-covariance matrix has the following form:

1. Prediction step ([253, 260]):

\[
\hat{\alpha}(t_k|t_{k-1}) = F\hat{\alpha}(t_{k-1}) \quad (3.24a)
\]

\[
P(t_k|t_{k-1}) = FP(t_{k-1})F^T + GQG^T \quad (3.24b)
\]

2. Correction step ([217] and Appendix B):

\[
\hat{\alpha}(t_k) = \hat{\alpha}(t_k|t_{k-1}) + L(t_k)\epsilon(t_k) \quad (3.25a)
\]

\[
P(t_k) = \left[ I - L(t_k)\psi(t_k, \hat{\alpha}(t_k|t_{k-1})) \right] P(t_k|t_{k-1}) \left[ I - \psi(t_k, \hat{\alpha}(t_k|t_{k-1}))L^T(t_k) \right] + L(t_k)\Lambda(t_k)L^T(t_k) \quad (3.25b)
\]
All the the notations in (3.25) are defined in Appendix B for the RPE algorithm (Appendix B.2).

It is worth mentioning that, in the off-line situation, a backward pass smoothing procedure (e.g., Fixed Interval Smoothing algorithm (FIS) (referred to in Appendix A.2 and also [89, 260])) could be attached behind the two-step forward pass filtering procedure to conduct the task set by Beck et al. ([33]) to estimate the unobserved components (parameters $\alpha$) in the observation equations (3.19b'), so that the fulfillment of the overall goal of bringing out the inner parametric space of the model structure could be achieved. Furthermore, in this situation, $\psi^T(t_k)$ (see Appendix B) could serve as a surrogate of $H(x, u; t_k)$ since $\psi^T(t_k)$, a sensitivity coefficients matrix, relates the observations and parameters through the differentiation of the state with respect to parameters, which satisfies the requirement for the properties of $H(x, u; t_k)$. However the implementation of the backward pass smoothing of FIS is not entirely successful in practice, and therefore requires further research (beyond the scope of this dissertation).

### 3.3.3 A BIOMASS-SUBSTRATE HYPOTHETICAL SYSTEM

The foregoing modified RPE algorithm (3.24)–(3.25) for estimating the time-varying parameters has been tested on a hypothetical system reflecting a biomass-substrate biological relationship in a Continuously Stirred Tank Reactor (CSTR). The system consists of one CSTR and a nutrient or substrate $u_2(t)$ flowing into the tank with dilution rate $q_0$. The set of ordinary differential equations associated with this biological system is defined as follows ([104, 217]):

$$\frac{dx_1(t)}{dt} = -q_0x_1(t) + \mu x_1(t) \frac{x_2(t)}{K_S + x_2(t)} + q_0u_1(t)$$  \hspace{1cm} (3.26a)

$$\frac{dx_2(t)}{dt} = -q_0x_2(t) - \frac{1}{Y} \mu x_1(t) \frac{x_2(t)}{K_S + x_2(t)} + q_0u_2(t)$$  \hspace{1cm} (3.26b)

where, $u_1(t), x_1(t)$ are the concentrations of biomass in the influent and tank at time $t$, mg·L$^{-1}$; $u_2(t), x_2(t)$ are the concentrations of growth-limiting substrate in the influent and
tank at time $t$, mg·L$^{-1}$; $K_S$ is the saturation constant of the substrate, mg·L$^{-1}$; $\mu$ is the maximum growth rate, h$^{-1}$; $Y$ is the yield coefficient of biomass, (g biomass)·(g substrate)$^{-1}$; and $q_0$ is the dilution rate, h$^{-1}$.

The model (3.26) was simulated and the parameters in the relations were tuned using the SIMULINK simulation Toolbox in the MATLAB® software. The SIMULINK flow chart of the model is presented in Figure 3.7.

![Figure 3.7: SIMULINK flow diagram for the biomass-substrate model.](image)

As inputs to the system, the biomass concentration in the influent to the tank ($u_1(t)$) is set to zero, while the substrate concentration in the influent ($u_2(t)$) is defined as follows$^5$:

$$u_2(t) = \begin{cases} 30, & 0 \leq t \mod 40 < 20; \\ 10, & 20 \leq t \mod 40 < 40. \end{cases}$$

With the values for the parameters in the model (3.26) being taken as $\mu = 0.3$ h$^{-1}$, $K_s = 3.0$ mg·L$^{-1}$, $Y = 0.6$ (g biomass)·(g substrate)$^{-1}$, and $q_0 = 0.1$ h$^{-1}$, the input signal of substrate

\footnotetext[5]{The $t \mod 40$ means the remainder of $t$ divided by 40.}
and the simulated output of biomass and substrate are displayed in Figure 3.8. Note that the two output time-series were corrupted by Gaussian white noise with the variance-covariance matrix $\Lambda = \begin{pmatrix} 0.1 & 0 \\ 0 & 0.05 \end{pmatrix}$.

![Figure 3.8: The simulated input and outputs of the biomass-substrate model.](image)

**Reconstruction of time-invariant parameter using the RPE**

The application of the RPE leads to the reconstructed observations of states and time-invariant parameters presented in Figures 3.9 and 3.10. The unbiased estimates of the parameters, the fair goodness of fit and the small deviations of most elements of the Kalman-like gain matrix (shown in Figure 3.11) from zero indicate a high quality performance of the RPE for the purpose of parameter estimation (see also [217]).
Figure 3.9: The reconstructed observations of states with *time-invariant* model parameters.

Figure 3.10: The reconstructed *time-invariant* model parameters (true value, dashed line; estimated value, continuous line; standard error, dotted lines).
Estimation of time-varying parameters using RPE

For the purpose of demonstration we will now linearize the model (3.26) by substituting

$$\alpha(t) = -\frac{1}{Y}\mu \frac{x_1(t)}{K_S + x_2(t)}$$

(3.27)

into Equations (3.26a) and (3.26b). This leads to a linear system of the form:

$$\frac{dx_1(t)}{dt} = -q_0 x_1(t) - Y\alpha(t)x_2(t) + q_0 u_1(t)$$

(3.28a)

$$\frac{dx_2(t)}{dt} = -q_0 x_2(t) + \alpha(t)x_2(t) + q_0 u_2(t)$$

(3.28b)

in which, \(\alpha(t)\) is a time-varying parameter defined by (3.27).

From a priori information, \(Y\) is known to be time-invariant, while \(\alpha(t)\) is idealized as a random walk (RW) parameter. Each element in the \(K\) matrix, whose significant deviation from zero is indicative of an incomplete model for the system, is assumed to be time-invariant.
and is initialized to be zero. The specifications of the initial conditions of states and the covariance matrices $P(t_0)$, $Q$ and $\Lambda$ for this demonstration example are given in Table 3.1.

Table 3.1: Specifications of initial values and leading diagonal elements for the covariance matrices $P(t_0)$, $Q$ and $\Lambda$.

<table>
<thead>
<tr>
<th>State/Parameter</th>
<th>Initial Values</th>
<th>$P(t_0)$</th>
<th>$Q$</th>
<th>$\Lambda$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1(t)$ (mg·L$^{-1}$)</td>
<td>5.00</td>
<td>—</td>
<td>—</td>
<td>0.1</td>
</tr>
<tr>
<td>$x_2(t)$ (mg·L$^{-1}$)</td>
<td>2.53</td>
<td>—</td>
<td>—</td>
<td>0.05</td>
</tr>
<tr>
<td>$\alpha$ (h$^{-1}$)</td>
<td>0.001</td>
<td>2.0</td>
<td>0.001</td>
<td>—</td>
</tr>
<tr>
<td>$Y$ (g bio·(g sub)$^{-1}$)</td>
<td>0.6</td>
<td>0.0001</td>
<td>0</td>
<td>—</td>
</tr>
<tr>
<td>$K$ elements</td>
<td>0</td>
<td>0.001</td>
<td>0</td>
<td>—</td>
</tr>
</tbody>
</table>

Figures 3.12 through 3.14 show the estimation results for the application of the RPE to the linear model (3.28) with a single time-varying parameter. Among them, Figure 3.13 shows that the estimate of $\alpha(t)$ tracks its true variation with time with intermediate success although with a constant offset — the estimated value is always more positive than the true value. However, the goodness of fit (Figure 3.12) and the insignificant evolutions of $K$’s elements (Figure 3.14) imply that the estimation of the time-varying parameter through RPE is still fairly reliable.
Figure 3.12: The reconstructed observations of states with time-varying model parameter.

Figure 3.13: The reconstruction of (a) the time-varying parameter (true value, dashed line; estimated value, continuous line; standard error, dotted lines); and (b) the time-invariant parameter (true value, dashed line; estimated value, continuous line; standard error, dotted lines).
3.3.4 BOD-DO interactions in River Cam

The analysis of the interaction of biochemical oxygen demand (BOD) and dissolved oxygen (DO) through the data set collected from the River Cam, UK, has been extensively reported since the 1970’s ([20, 35, 217]). The time-varying parameter estimation capabilities of the modified RPE have therefore been tested in this well-examined real world case, before being applied to the two case studies in the following chapters.

The model (3.29) for describing the BOD-DO interaction in River Cam is directly adopted from [20] and [35] for the purpose of comparison. The specifications of the initial conditions of states and the corresponding covariance matrices $P(t_0)$, $Q$, and $\Lambda$ for this case are listed in Table 3.2. However, all the values of the diagonal elements in $\Lambda$ are determined by using
the method suggested in ([172, 247]), i.e., for a time-series \(y(t_k), k = 1, 2, \ldots, N\),

\[
\lambda = (0.8865 \frac{1}{N-1} \sum_{k=1}^{N} |y(t_k) - y(t_k - 1)|)^2.
\]

The model of BOD-DO interaction in River Cam is expressed as,

\[
\begin{align*}
\frac{dx_1(t)}{dt} &= -[\beta_1 + \frac{Q(t)}{V}]x_1(t) + \frac{Q(t)}{V}u_1(t) + L_A(t) \quad (3.29a) \\
\frac{dx_2(t)}{dt} &= -\beta_1 x_1(t) - [\beta_2 + \frac{Q(t)}{V}]x_2(t) + \frac{Q(t)}{V}u_2(t) + \beta_2 C_S(t) + D_B(t) \quad (3.29b)
\end{align*}
\]

where, \(x_1(t), x_2(t)\) are the concentrations of BOD and DO at the downstream end (output) of the reach, respectively, mg·L\(^{-1}\); \(u_1(t), u_2(t)\) are the concentrations of BOD and DO at the upstream end (input) of the reach, respectively, mg·L\(^{-1}\); \(Q(t)\) is the volumetric flow-rate in the reach, m\(^3\)·d\(^{-1}\); \(V\) is the constant volume of water in the reach, m\(^3\); \(C_S(t)\) is the saturation concentration of DO, mg·L\(^{-1}\); \(\beta_1\) is the BOD decay rate constant, d\(^{-1}\); \(\beta_2\) is the re-aeration rate constant of DO, d\(^{-1}\); \(L_A(t)\) is the rate of addition of BOD to the reach by, for example, local surface runoff mg·L\(^{-1}\)·d\(^{-1}\); \(D_B(t)\) is the net rate of addition of DO to the reach by the combined effects of photosynthesis, respiration, and decomposition of mud deposits, mg·L\(^{-1}\)·d\(^{-1}\).

Table 3.2: Specifications of initial values and leading diagonal elements for the covariance matrices \(P(t_0), Q\) and \(\Lambda\).

<table>
<thead>
<tr>
<th>State/Parameter</th>
<th>Initial Values</th>
<th>(P(t_0))</th>
<th>(Q)</th>
<th>(\Lambda)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(x_1(t)) (mg·L(^{-1}))</td>
<td>2.30</td>
<td>— (1.0(^a))</td>
<td>— (0.4)</td>
<td>0.2865 (0.4)</td>
</tr>
<tr>
<td>(x_2(t)) (mg·L(^{-1}))</td>
<td>8.00</td>
<td>— (1.0)</td>
<td>— (0.4)</td>
<td>0.1858 (0.2)</td>
</tr>
<tr>
<td>(\beta_1) (d(^{-1}))</td>
<td>0.32 (0.32)</td>
<td>0.001 (0.005)</td>
<td>0 (0)</td>
<td>—</td>
</tr>
<tr>
<td>(\beta_2) (d(^{-1}))</td>
<td>0.17 (0.17)</td>
<td>0.001 (0.005)</td>
<td>0 (0)</td>
<td>—</td>
</tr>
<tr>
<td>(L_A(t)) (mg·L(^{-1})·d(^{-1}))</td>
<td>0 (0)</td>
<td>2 (2)</td>
<td>0.05 (0.05)</td>
<td>—</td>
</tr>
<tr>
<td>(D_B(t)) (mg·L(^{-1})·d(^{-1}))</td>
<td>0 (0)</td>
<td>2 (2)</td>
<td>0.05 (0.05)</td>
<td>—</td>
</tr>
<tr>
<td>(K) elements</td>
<td>0</td>
<td>0.01</td>
<td>0</td>
<td>—</td>
</tr>
</tbody>
</table>

\(^a\) All the values in the parentheses were used in the EKF application ([20, 35]);
By assuming that $\beta_1$, $\beta_2$ and the entities in the $K$ matrix are time-invariant, but $L_A(t)$ and $D_B(t)$ are varying with time, the parameter estimation results shown in Figure 3.16 are similar to those of processing the same data set with the EKF ([20, 35]), but with two notable exceptions. First, the re-aeration rate constant $\beta_2$ for DO is continuously increasing during the observed period instead of decreasing around day 40. Second, before day 40 there is an implied addition of BOD to the reach, which may not be caused by dead algae. Compared with the EKF, one of the advantages of the RPE is the simultaneous recursive estimation of the weighting matrix $K$, which serves as a device to detect whether or not a model is a perfect description of reality. Its products with the innovations represent the acknowledged unknown of the system (see the lower panels of Figures 3.15(a) and 3.15(b)). An increase in $K\nu(t)$ and swift displacements of $K_{1,1}$, $K_{1,2}$, and $K_{2,2}$ (shown in Figure 3.17) around day 40 indicate that the model structure of form (3.29) has changed at day 40, since the contribution of algae to the system is no longer negligible ([20, 35, 217]).

Figure 3.15: The comparison of model output and observations of the concentrations at downstream.
Figure 3.16: The estimation of the parameters.

Figure 3.17: The deviations of the elements in the $K$ matrix.
3.4 Time constant and steady-state gain

In a linear, time-invariant system, two physically meaningful properties — the *time constant* and *steady-state gain* — can be derived from either the linear data-based model (e.g., Equation (3.3)) or the linear theory-based model (e.g., Equation (3.32)). The *time constant* $\tau$ is the time required for the system output to decay to $e^{-1} (\approx 36.8\%)$ of its maximum value in response to a unit impulse, while the *steady-state gain* $\kappa$ defines the relationship between the equilibrium output value when a step-change input is applied to the system. Since environmental systems are intrinsically nonlinear, in order to exploit these two lumped parameters to compare the performance of the data-based and theory-based models for environmental systems, we shall soon depart from the constraint of a linear, time-invariant, system.

3.4.1 Derivation from data-based model

For convenience, the deterministic part of the transfer function model of linear system (3.3) is restated in the following,

$$y(k) = \frac{B(z^{-1})}{A(z^{-1})}u(k - \delta)$$

(3.30)

with $y(k)$, $u(k)$, $\delta$, $A(z^{-1})$ and $B(z^{-1})$ defined as before. The *time constant* $\tau$, therefore, is defined as, in the case of a first-order TF,

$$\tau = \frac{-1}{\log_e(-a_1)}$$

(3.31a)

The *steady-state gain* $\kappa$ is calculated by setting the $z^{-1}$ operator in Equation (3.30) to unity such that,

$$\kappa = \frac{b_0 + b_1 + \ldots + b_m}{1 + a_1 + \ldots + a_n}$$

(3.31b)

As a result, the *steady-state gain* can be used to determine whether there have been any physical gains or losses across the system.
When the polynomials in Equation (3.30) have a higher-order (e.g., higher than second-order), the TF model can always be decomposed into lower order (e.g., first- or second-order) TF elements connected in serial, parallel or feedback form. The decomposition of the higher-order TF models can sometimes reveal useful information about the system’s internal behavior that would otherwise remain unknown.

In the case of a nonlinear system, the nonlinear behavior of the system can often be approximated by a linear TF with *time-varying* parameters ([257]), which are written in (3.7). Since the deterministic term of model (3.7) has the same formation as Equation (3.30), the corresponding *time constant* and *steady-state gain* of model (3.7) can be readily derived using Equation (3.31) yet with time variable polynomial parameters. Thus, the *time constant*, $\tau(t)$, and *steady-state gain*, $\kappa(t)$, vary with time as well.

### 3.4.2 Derivation from theory-based model

For explanatory purposes, we write a mechanistic model of a linear system, in which a reacting solute passes through a CSTR, in an ordinary differential equation (ODE) formation (3.32). In this linear system, a solute with concentration $u_0(t)$ in the influent passes through the tank (a CSTR) undergoing reaction with an assumed constant first-order rate of reaction $k$. The tank has a constant volume ($V$) and thus the assumed time-invariant influent and effluent flows ($F$) must be equal. Each of these parameters has an associated unit. The concentration of this single reactive solute in the tank is denoted as $x(t)$, hence the change rate of concentration, $dx(t)/dt$, can be described mechanistically as the following,

$$\frac{dx(t)}{dt} = -\frac{F + V k}{V} x(t) + \frac{F}{V} u_0(t) \tag{3.32}$$

---

*Time constant* and *steady-state gain* are concepts characterizing linear processes in control system engineering. When they are used to describe nonlinear systems, we shall give them different names to distinguish from their original usage. For example, we could add a prefix quasi- to them to generate two new terms: *quasi-time constant* and *quasi-steady-state gain*. However, in this dissertation, we are only dealing with nonlinear environmental systems, therefore, unless specified, we hereafter only refer to the *time constant* and *steady-state gain* derived from nonlinear systems.
Having formulated the behavior of this linear system in the form of (3.32), it is customary to define the system’s time constant, $\tau$, as

$$\tau = \frac{V}{F +Vk} \quad (3.33a)$$

and the steady-state gain, $\kappa$, as

$$\kappa = \frac{F}{F +Vk} \quad (3.33b)$$

So being defined, Equation (3.32) can also be written in terms of the lumped parameters – time constant and steady-state gain, i.e.,

$$\frac{dx(t)}{dt} = -\frac{1}{\tau}x(t) + \frac{\kappa}{\tau}u_0(t) \quad (3.34)$$

As in the nonlinear data-based model, the derivation of the quasi-time constant ($\tau(t)$) and quasi-steady-state gain ($\kappa(t)$) from the nonlinear theory-based model can be readily obtained by transforming the model’s nonlinear form into the linear form with time-varying parameters. Therefore, the two lumped parameters vary with time as well.
Chapter 4

Case Study One: Ammonium-Nitrogen Removal in a Conventional Activated Sludge System

4.1 Introduction

Having accounted for the details of the methodologies employed in the proposed two-pronged modeling approach, we now apply this generic modeling method to two complex environmental systems. Starting with a simpler case study — ammonium-nitrogen removal analysis in a conventional activated sludge system, we aim to illustrate: first, how the application of two separate modeling approaches to the same environmental system can help us to acquire more information about the system; second, how the results obtained from one modeling method can be utilized by the other in order to gain a deeper understanding about the system; third, how these two types of models derived from two opposite directions — data and theory, can be unified under parameter space. Although one of the critical procedures of the data-based modeling — state dependent parameter (SDP) modeling shown in Figure 3.6 — is omitted in this case study, its absence should not compromise the main purpose of providing a panoramic view of the two-pronged modeling method illustrated by this chapter.

4.2 Study site and data collection

The study site, the Athens Wastewater Treatment Facility (AWTF) No. 2 has two aeration tanks in the plant, each of which has three channels (Orbal facility). The preliminarily treated sewage flows in sequence from the outer channel through the middle into the inner channel. Effluent from the two tanks is combined and then split equally into three secondary clarifiers
for water-sludge separation and sludge settlement. Under normal operation the returned activated sludge is pumped from the bottom of the secondary clarifiers back to the outer channels of the aeration tanks. Figure 4.1 shows the scheme of the plant. The outer channel, middle channel and inner channel contain approximately 50 to 55%, 30 to 35%, and 15 to 20% of the total bioreactor volume, respectively ([140]). The annual average wastewater flow to the plant is about 5 MGD (788 m\(^3\)·h\(^{-1}\)), and the hydraulic retention time (HRT) in the aeration tank is about 24 hours. So the total volume of the aeration tank is about 19,000 m\(^3\).

![Figure 4.1: Schematic diagram of the activated sludge process in the AWTF No. 2.](image)

The University of Georgia’s Environmental Process Control Laboratory (EPCL) was deployed to collect data from the Athens Wastewater Treatment Facility No. 2 ([140]). The EPCL, commissioned in 1997 and manufactured by Capital Controls (Minworth Systems Limited, Birmingham, UK), is the core facility of the Environmental Informatics and Control Program at the University of Georgia. It can be used in many contexts, especially in the study of municipal and industrial wastewater treatment (see also [140]), surface water quality (see Chapter 5) and groundwater contamination. It comprises two mobile trailers, each of them equipped with on-line automatic monitors. One trailer houses a respirometer, ammonium monitor, TOC monitor, and turbidity monitor, coupled with a homogenizer and a debubbler.
The other trailer has a respirometer, $\text{NO}_x$ (nitrite and nitrate) monitor and orthophosphate monitor. Other than these on-board monitors, such out-board monitors as two DO probes, a mixed liquor suspended solid (MLSS) probe, and a sludge blanket level sensor can be mounted outside the trailers (see http://www.modeling.uga.edu/epcl/public_html/ and [140] for more information about the EPCL).

The sampling campaign started on February 1, 1998, and was completed on April 28, 1998. Up to 21 variables were monitored, including crude sewage flow rate, ammonia-nitrogen ($\text{NH}_4^+$-N) and total oxidized carbon (TOC) concentrations in the crude sewage, outer channel and inner channel, dissolved oxygen (DO) concentrations in the middle and inner channels, and mixed liquor suspended solids (MLSS) concentrations in the inner channel. All time-series collected by the EPCL were at 15-minute sampling intervals. However, for the same reasons that will be given in Section 5.3 in the next chapter, where the signal processing method introduced in Chapter 3 is applied for data analysis in detail, the data are decimated into 2-hour intervals when used for system analysis. Subsequently, the missing data were interpolated and the outliers were smoothed by the Dynamic Harmonic Regression (DHR) model in the CAPTAIN Toolbox. However, the returned activated sludge (RAS) flow rate was manually recorded by plant operators twice each shift (8 hours), hence at a 4-hour sampling interval. The smoothed data and the extracted trends of various time-series, such as the crude sewage and RAS flow rates, $\text{NH}_4^+$-N concentration in the crude sewage, outer and inner channels, MLSS concentration in the inner channel, and DO concentration in the middle channel, are shown in Figures 4.2 through 4.4, in which the time period covers from March 12th to March 30th, 1998.
Figure 4.2: Flow rates of (a) Crude sewage influent, and (b) Returned Activated Sludge (RAS).

Figure 4.3: Ammonia-nitrogen concentrations in (a) Crude sewage, (b) Outer channel, and (c) Inner channel.
The Orbal activated sludge process can be conceptually decomposed into two sub-systems in terms of sampling points (circles attaching to the dashed lines shown in Figure 4.1). System one, starting from the outer channel probe location to the inner channel probe location, is a Single Input Single Output (SISO) system, which covers the entire middle channel and a portion of the outer and inner channels of the aeration tank. It is shown in Figure 4.5(a). System two covers the remaining portions of the outer and inner channels and the clarifiers (Figure 4.5(b)). However, in this research we are only investigating the dynamics of the first sub-system, that is System one. For the ammonium-nitrogen removal analysis, the major input signal to the system one is the ammonia-nitrogen concentration monitored by the outer channel probe (plotted in Figure 4.3(b)), and the output signal is the ammonia-nitrogen concentration monitored by the inner channel probe (plotted in Figure 4.3(c)).

\footnote{Hereafter, in this chapter, we only refer to System one unless being specified otherwise.}
4.3 Data-based modeling

The Data-Based-Mechanistic modeling procedure, illustrated in Chapter 3, has been applied to explore the dynamics of the removal of ammonium-nitrogen from the bio-reactor system described in the previous section.

4.3.1 Linear TF model with constant parameters

Equation (4.1) describes the linear Transfer Function (TF) relationship between the \( \text{NH}_4^+ - \text{N} \) concentration entering the system (input \( u(k) \)) and that leaving the system (output \( y(k) \)).

\[
y(k) = \frac{B(z^{-1})}{A(z^{-1})} u(k - \delta) + \nu(k); \quad k = 1, 2, \ldots, N
\]  

(4.1)

where,

\[
A(z^{-1}) = 1 + a_1 z^{-1} + \ldots + a_n z^{-n}; \quad B(z^{-1}) = b_0 + b_1 z^{-1} + \ldots + b_m z^{-m}.
\]
This can be also written in the following alternative equation form,

\[ y(k) = -a_1 y(k - 1) - \ldots - a_n y(k - n) \]
\[ + b_0 u(k - \delta) + b_1 u(k - \delta - 1) + \ldots + b_m u(k - \delta - m) \]
\[ + \nu(k) + a_1 \nu(k - 1) + \ldots + a_n \nu(k - n) \]

where \( N \) is the total number of observations in the time-series; \( \delta \) represents the potential presence of any pure time delay in the system (in sampling intervals); \( \nu(k) \) represents uncertainty in the relationship arising from a combination of measurement noise, the effects of other unmeasured disturbances and modeling error; \( y(k) \) is the output signal of the system (i.e., the \( \text{NH}_4^+ \)-N concentration in the inner channel), mg·L\(^{-1}\); and \( u(k) \) is the input signal to system (i.e., the \( \text{NH}_4^+ \)-N concentration in the outer channel), mg·L\(^{-1}\).

The SRIV algorithm described in Appendix A.1 was employed to identify the order of the deterministic part of the linear TF model (i.e., Equation (4.1)). Table 4.1 lists the best 5 linear TF models with constant parameters for describing the relationship between the input \( \text{NH}_4^+ \)-N concentration and the output \( \text{NH}_4^+ \)-N concentration, from which the dynamics of the \( \text{NH}_4^+ \)-N removal efficiency in the system can be derived.

Table 4.1: The best 5 identified linear TF models with constant parameters (The sampling interval is 2 hours).

<table>
<thead>
<tr>
<th>No.</th>
<th>Denominator</th>
<th>Numerator</th>
<th>Delay</th>
<th>YIC</th>
<th>( R_T^2 )</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>-4.56</td>
<td>0.4357</td>
<td>1.34</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-4.05</td>
<td>0.7204</td>
<td>0.61</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>-3.45</td>
<td>0.7204</td>
<td>0.62</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>-2.51</td>
<td>0.7104</td>
<td>0.66</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>-2.10</td>
<td>0.7167</td>
<td>0.67</td>
</tr>
</tbody>
</table>

In terms of the three DBM selection criteria (YIC, \( R_T^2 \) and AIC) in Table 4.1, it is difficult to distinguish any one structure that is superior on all accounts. However, according to [140], the sewage is retained in the three channels of the aeration tank for about 24 hours.
(12 sampling intervals). A time delay is obvious for this system. Therefore, models [2, 3, 4] and [1, 1, 1] with pure time delays within their structures, rather than others without pure time delays, could be considered as candidates. Their parameter estimates are listed below accompanied by the corresponding standard errors in parentheses.

- For model [2, 3, 4]:
  \[
  \hat{a}_1 = -1.4438(0.0154), \quad \hat{a}_2 = 0.9262(0.0147), \\
  \hat{b}_0 = 1.4226(0.0969), \quad \hat{b}_1 = -2.1751(0.1828), \quad \hat{b}_2 = 1.1864(0.0959);
  \]

- For model [1, 1, 1]:
  \[
  \hat{a}_1 = -0.4483(0.0873), \quad \hat{b}_0 = 0.4986(0.0783).
  \]

Since model [2, 3, 4] is characterized by significantly complex poles (eigenvalues) and it has an inferior goodness of fit to the data, model [1, 1, 1] is chosen as the data-based model of this system. Figure 4.6 displays the one-step-ahead prediction of model [1, 1, 1] and the raw model residuals. The apparent pattern shown in the residuals plot indicates the inadequacy of the model structure [1, 1, 1] with constant parameters, so that further investigation is necessary.
Figure 4.6: Linear TF model \([1, 1, 1]\) with constant parameters: (a) one-step-ahead prediction and observations, and (b) residuals plot.

4.3.2 Linear TF model with TVPs

Therefore, a linear TF model with \textit{time-varying} parameters, rather than a linear TF model with \textit{constant} parameters, was used to describe the system’s behavior. The SISO model with TVP takes the form:

\[
y(k) = \frac{B(k, z^{-1})}{A(k, z^{-1})} u(k - \delta) + \nu(k); \quad k = 1, 2, \cdots, N
\]

(4.2)

where,

\[
A(k, z^{-1}) = 1 + a_1(k)z^{-1} + \cdots + a_n(k)z^{-n};
\]

\[
B(k, z^{-1}) = b_0(k) + b_1(k)z^{-1} + \cdots + b_n(k)z^{-n}.
\]
Thus, for the simplest linear structure $[1, 1, 1]$ identified previously, we have

$$y(k) = \frac{b_0(k)}{1 + a_1(k)} u(k - 1) + \nu(k)$$  \hspace{1cm} (4.3)

From the previous section it is evident that the standard errors of the estimates of lag parameter $a_1$ and the input (gain) parameter $b_0$ are similar. So it is reasonable to allow both of them to vary simultaneously over the observation interval. The Integrated Random Walk (IRW, see discussion in Section 3.3.2) model was chosen to model the variation of each of parameters $a_1(k)$ and $b_0(k)$ of model (4.3). The values of the corresponding NVR’s of $a_1(k)$ and $b_0(k)$ optimized by an ML algorithm are:

$$NVR(a_1(k)) = 0.2163 \times 10^{-5}, \quad NVR(b_0(k)) = 0.1570 \times 10^{-5}.$$

The two time-varying parameters estimated using the FIS algorithm and their standard errors are shown in Figure 4.7, while the resulting time variable time constant and steady-state gain, together with the trend of the MLSS in the aeration tank, are shown in Figure 4.8.

In this case, the time constant measures the rate of the ammonium-nitrogen removal from the system — a larger time constant indicates a slower consumption rate of the ammonium-nitrogen, and vice versa; while the steady-state gain gauges the removal efficiency of the ammonium-nitrogen of this system, that is, a smaller steady-state gain implies that a larger portion of the ammonium-nitrogen entering the system is removed and a smaller portion of the ammonium-nitrogen is left in the effluent of the system.

It is interesting to note that the parameter variations, especially that of the steady-state gain, show an oscillation pattern similar to that of the trend of the MLSS. The phase shift between the oscillations of the MLSS and the steady-state gain could be an evidence that the removal of ammonium-nitrogen from the system is achieved mainly through nitrification process. As we know, in a wastewater treatment system that accommodates nitrifying bacteria, the ammonium-nitrogen is removed through two major processes: (1) the nitrification involving the nitrifying bacteria (autotroph, *Nitrosomonas* and *Nitrobacter*); and (2)
the uptake for cell synthesis involving any living bacteria (autotroph and heterotroph). In addition, most of the MLSS is composed of heterotrophic bacteria which decompose the carbonaceous biodegradable matters in sewage for their growth, while only a very small portion of the MLSS is the nitrifying bacteria which utilize ammonium-nitrogen as their energy supplier through the nitrification process.

Now let us look at the characteristics of the variations of the MLSS trend and the *steady-state gain*\(^2\) shown in Figure 4.8. The peaks and troughs of the MLSS trend fluctuations are followed by those of the *steady-state gain*. That is to say that, when the MLSS is built up in the system, less ammonium-nitrogen is taken away from the system. This is because, in this system, the nitrification process, rather than the uptake for cell synthesis, was playing a dominant role for ammonium-nitrogen removal.

Figure 4.9 shows that the linear TF model with TVP displays a much better prediction capability than the linear model with constant parameters (see Figure 4.6). In addition, the residuals of model (4.3) no longer show any longer-term persistent patterns (other than a daily variation), although heteroscedasticity still exists.

\(^2\)Alternatively, the variation of the *time constant* can be used for the following observation; the result is the same.
Figure 4.7: Variations of TVP estimates of model [1, 1, 1]: (a) lag parameter \( \hat{a}_1(k|N) \), and (b) input parameter \( \hat{b}_0(k|N) \).

Figure 4.8: Comparison of variation patterns among: (a) time constant, (b) steady-state gain, and (c) trend of MLSS.
Figure 4.9: Linear TF model [1, 1, 1] with TVP: (a) one-step-ahead prediction and observations, and (b) residuals plot.

4.3.3 Impact of Hydraulic Retention Time

Both the time constant and steady-state gain tend to decline during the observed time period. Interestingly the declining trends of these two parameters mimic that of the influent flow rate shown in Figure 4.2. Less influent flow means longer hydraulic retention time (HRT) of sewage in the aeration tanks, which, in turn, results in a higher removal ratio of ammonium-nitrogen. Again, the higher removal ratio implies a smaller value of the steady-state gain. In order to account for the impact of the HRT, we weighted the original input signal ($\text{NH}_4^+\text{-N}$ concentration of the outer channel) to the system with the ratio of the instant total flow rate (i.e., sum of influent flow rate and RAS flow rate) to the mean of total flow rate. Therefore, the new resulting input signal is constructed as the following equation.

$$u(k) = \frac{Q_T(k)}{\bar{Q}_T(k)} C_{\text{NH}_3-N,\text{out}}(k)$$  \hspace{1cm} (4.4)
in which, \( Q_T(k) = Q_I(k) + Q_R(k) \), while \( Q_I(k) \) and \( Q_R(k) \) are the flow rates of influent and RAS, respectively; \( \bar{Q}_T(k) \) is the mean value of \( Q_T(k) \); and \( C_{NH_3-N,out}(k) \) is the \( NH_4^+\)-N concentration of the outer channel.

The same data-based modeling procedure described in the beginning of Section 4.3.2 was applied to the new pairs of input and output signals. However, the estimated NVR’s of \( a_1(k) \) (\( NVR = 0.9338 \times 10^{-6} \)) and \( b_0(k) \) (\( NVR = 0.1362 \times 10^{-12} \)) indicate that the input parameter \( b_0(k) \) is not changing too much with time, especially compared with the lag parameter \( a_1(k) \). Therefore, in the subsequent analysis, only \( a_1(k) \) is allowed to vary with time as an IRW process, while \( b_0(k) \) is assumed to be constant during the considered time period. The resulting linear TF model (4.3) takes the following form:

\[
y(k) = \frac{b_0}{1 + a_1(k)} u(k - 1) + \nu(k)
\]  

(4.5)

where \( u(k) \) is now defined by formulation (4.4).

The values of the corresponding NVR’s of \( a_1(k) \) and \( b_0(k) \) in model (4.5) are:

\[
NVR(a_1(k)) = 0.1023 \times 10^{-5}, \quad NVR(b_0(k)) = 0.
\]

The two parameters estimated using the FIS algorithm and their standard errors are shown in Figure 4.10, while the resulting time variable time constant and steady-state gain, together with the trend of MLSS in the aeration tank, are shown in Figure 4.11. When compared with the parameter estimation of the previous TF model with TVP (model (4.3)), the variation of either time constant or steady-state gain still remains, as it is presumably caused by the variation of the MLSS. However, the downward tendency shown in the time constant and steady-state gain of model (4.3) almost becomes negligible in the revised model (4.5), where the reducing influent and RAS flow rate during the time period of interest are taken into account. But Figure 4.12 shows that the improvements in the aspects of model prediction performance and residual plot are marginal.
Figure 4.10: Variations of TVP estimates of model [1, 1, 1]: (a) lag parameter ($\hat{a}_1 (k|N)$), and (b) input parameter ($\hat{b}_0 (k|N)$).

Figure 4.11: Comparison of variation patterns among: (a) time constant, (b) steady-state gain, and (c) Trend of MLSS.
4.4 Theory-based modeling

4.4.1 A model for nitrification dynamics

Qualitatively the basic biochemical model for nitrification shows that ammonium-nitrogen is oxidized in two stages to nitrate-nitrogen, each stage involves different species of nitrifying bacteria, figuratively shown by Figure 4.13. The first stage is the oxidation of ammonium to nitrite by *Nitrosomonas*, while nitrite is further oxidized to nitrate by *Nitrobacter* in the second stage. The conversion step from ammonium-nitrogen to nitrite-nitrogen takes place more slowly, thus it is generally believed to be the rate-limiting step for the overall process.

Several models for nitrification have been proposed and generally fall into two broad categories in terms of ways in which this process is viewed: as a two-step or a single-step conversion. As a representative of the former, the model for nitrification presented in [177]
employs Monod kinetics to depict the growth rate of the nitrifying bacteria and was later evaluated in [18] against the data collected at Norwich Sewage Works, England, in a recursive scheme. However, the Activated Sludge Models, developed by the Task Group of the International Association on Water Pollution Research and Control (IAWPRC), consider the autotrophic conversion of ammonium to nitrate simply as a single-step process. The mechanistic modeling of the nitrification process in wastewater treatment systems have been reviewed recently in [140].

In practice, the models for the nitrification of ammonium-nitrogen are believed to be easier to calibrate against experimental data than the corresponding models for carbonaceous biological oxygen demand (BOD) and suspended solid (SS) removal in an activated sludge process ([18]), but the ammonium-nitrogen removal process is still confounded by many other variables and processes besides the population of the nitrifying bacteria present in the MLSS. The nitrification process could be inhibited at a DO concentration as low as 0.2–0.5 mg·L\(^{-1}\) since the nitrifying bacteria have to compete with the abundant heterotrophic bacteria below a certain level of DO. For example, the expression for the nitrifying bacteria growth rate was modified in [52] to account for the limitation of growth under low dissolved
oxygen conditions (see also [101]). Other factors, such as heterotrophic bacterial conversion of dissolved organic nitrogen to ammonium-nitrogen, uptake and release of nitrogen compounds by phytoplankton and zooplankton ([97, 137]), and predation of nitrifiers by protozoa and rotifers ([138]), could also be included in the model to reconstruct the concentrations of ammonium-nitrogen and nitrifying bacteria in the mixed liquor of activated sludge system.

A simple theory-based model for the nitrification process in an activated sludge system is constructed for recursive parameter estimation purposes. This model simplifies the nitrification process as a single-step process. The state variables, parameters and the environmental functions pertaining to this model are listed in Table 4.2. The system, consisting of the partial inner and outer channels and the entire middle channel of the aeration tanks, was conceptualized as a Completely Stirred Tank Reactor (see Figure 4.5(a)) (CSTR). For the CSTR idealization, the following assumptions are necessities:

1. The concentration of the solute (e.g., state variable, \( x_1(t) \) or \( x_2(t) \)) in the tank is uniform — the same at each spatial location, i.e., does not vary with space;

2. The concentration of the solute in the effluent flow is the same as that in the system, and is a function of time.

However, the previous DBM modeling results suggest that a pure time delay of one sampling interval exists. This is not surprising because the channels of the aeration tanks in the AWTF are long corridors followed by each other, for which the sole CSTR idealization may not be sufficient. Therefore, a combination of a Plug Flow Reactor (PFR) and a CSTR in series was employed to idealize the real system. That is, the schematic representation of the system shown in Figure 4.5(a) is now modified into that of Figure 4.14. As a complement of the CSTR, the PFR idealization is convenient for the purpose of simulating solute movement of a purely advective, non-dispersive nature of a reactor. If the solute (e.g., ammonium-nitrogen) is subject to first-order chemical decay, with rate \( k(t) \), and enters the PFR with concentration \( u(t) \), then its concentration at the output of the PFR (or input to the followed
CSTR) will be \( u'(t) = u(t - \delta)e^{-k(t)\delta} \), assuming the flow rate does not change with time, in which, \( \delta \) is the pure time delay of the system.

\[
\dot{x}(t) = f(x, u, \alpha; t) + K\nu_1(t) \quad (4.6a)
\]
\[
y_1(t_k) = x_1(t_k) + \nu_1(t_k) \quad (4.6b)
\]

Figure 4.14: Re-conceptualized System One (Figure 4.5(a)) based on the results of the DBM step.

4.4.2 Recursive parameter estimation using RPE

In order to apply the RPE algorithm to the model for the dynamics of the nitrification process in the aeration tank of the wastewater treatment plant, the model with formulations (1)–(5) in Table 4.2 is rewritten in the innovations representation (4.6a) with a linear observation equation (4.6b) since the ammonium-nitrogen concentration in the aeration tank was monitored directly.
Table 4.2: Model formulations, variables and parameters.

**Formulations:**

\[
\frac{dx_1(t)}{dt} = -k(t)x_1(t) - \frac{Q(t)}{V}x_1(t) + \frac{Q(t)}{V}u_1(t - \delta)e^{-2\delta k(t)} + r(t)
\]  
(1)

\[
\frac{dx_2(t)}{dt} = Y_A k(t)x_1(t) - k_d x_2(t) - \frac{Q(t)}{V}x_2(t) + \frac{Q(t)}{V}u_2(t)
\]  
(2)

\[
k(t) = \frac{1}{Y_A} \mu_{max} f(T; t) f(DO; t) \frac{x_2(t)}{K_{NH} + x_1(t)}
\]  
(3)

\[
f(T; t) = \theta T(t) - 20
\]  
(4)

\[
f(DO; t) = \frac{DO(t)}{K_O + DO(t)}
\]  
(5)

**State variables and their initial values:**

- \(x_1(t)\) Ammonium-nitrogen concentration \(7.19\) mg·L\(^{-1}\)
- \(x_2(t)\) Nitrifying bacteria biomass concentration \(50.0\) mg·L\(^{-1}\)

**Environmental (forcing) variables:**

- \(Q(t)\) Wastewater flow rate, sum of the flow rates of the crude sewage and the RAS \(\text{m}^3\cdot\text{h}^{-1}\)
- \(T(t)\) Water temperature in the aeration tank \(^\circ\text{C}\)
- \(u_1(t)\) NH\(_4\)-N concentration of the inflow to the system \(\text{mg}\cdot\text{L}^{-1}\)
- \(DO(t)\) DO concentration in the middle channel \(\text{mg}\cdot\text{L}^{-1}\)

**Fixed parameters:**

- \(V\) Volume of aeration tanks \(6250\) m\(^3\)
- \(Y_A\) Yield coefficient of nitrifiers \(0.24\) g·g\(^{-1}\)
- \(K_{NH}\) Half saturation constant of ammonium-N \(1.0\) mg·L\(^{-1}\)
- \(K_O\) Half saturation constant of DO for nitrification \(0.4\) mg·L\(^{-1}\)
- \(\theta\) Coefficients for Arrhenius equation \(1.05\)
- \(\delta\) Pure time delay determined by DBM \(1.0\) 2hrs

**Time-invariant parameters and initial values:**

- \(\mu_{max}\) Specific growth rate of nitrifiers at 20\(^\circ\text{C}\) \(0.24\) d\(^{-1}\)
- \(k_d\) Decay rate of nitrifiers \(0.024\) d\(^{-1}\)

**Time-variant parameters and initial values:**

- \(r(t)\) Ammonium-N from other sources, e.g., mineralization of soluble organic nitrogen \(1.0\) mg·L\(^{-1}\)·h\(^{-1}\)
- \(u_2(t)\) Concentration of nitrifiers in the inflow to the system \(50.0\) mg·L\(^{-1}\)

\(a\) g nitrifying bacteria produced per g ammonium-N consumed.
where,

\[
\dot{x}(t) = \begin{bmatrix} \frac{dx_1(t)}{dt}, \frac{dx_2(t)}{dt} \end{bmatrix}^T;
\]

\[
x(t) = [x_1(t), x_2(t)]^T;
\]

\[
u(t) = [u_1(t), u_2(t)]^T;
\]

\[
\alpha(t) = [\mu_{\text{max}}, r(t), k_d, u_2(t)]^T;
\]

\[
K = \begin{pmatrix} K_{11} \\ K_{21} \end{pmatrix}.
\]

\( u(t) \) are the inputs to the system; \( y_1(t_k) \) is the observation of ammonium-nitrogen concentration in or leaving the system (mg·L\(^{-1}\)); \( \nu_1(t_k) \) is the measurement noise and disturbance reflected in the observations of ammonium-nitrogen concentration; \( K \) is the Kalman-like gain matrix, which distributes the impacts of mismatches among the constituent representations of the state variables; the two entities of vector \( f(x, u, \alpha; t) \) are Equations (1) and (2) respectively in Table 4.2; and all the other notations are the same as defined in Table 4.2.

As shown in Table 4.2, the maximum specific growth rate of the nitrifying bacteria at 20°C \( \mu_{\text{max}} \) and the decay rate of nitrifying bacteria \( k_d \) are assumed to be time-invariant over the period of study, while the parameters \( r(t) \) and the parameterized incoming concentration of nitrifying bacteria to the system \( u_2(t) \) are allowed to change with time. The variations of these \textit{time-varying} parameters are modeled by Random Walk (RW) stochastic processes since no \textit{a priori} knowledge is available with which to make a more sophisticated choice. Whether the evolutions of the recursive estimates of the \textit{time-invariant} parameters change with time will reveal the adequacy or inadequacy of the associated individual constituent hypothesis, which the parameter is representing. The trajectories of the \textit{time-varying} parameters also provide clues for constructing new hypotheses. The elements in \( K \) are also considered as \textit{time-invariant} parameters as well. Their values are set to be zeros initially, implying that
the mechanistic model describes the dynamics of nitrification process in the activated sludge system perfectly. Any excursions of the recursive estimation of \( K \)'s entities from zero suggest that the model fails to represent the real system correctly.

Table 4.3 specifies the corresponding initial values of states and parameters along with the leading diagonal elements for the variance-covariance matrices \( P(t_0) \), \( Q \) and \( \Lambda \), defined in Chapter 3. The fixed parameters, such as \( V, \theta \), take constant initial values specified in Table 4.2 and remain unchanged over the entire period.

Table 4.3: Specifications of initial values and the leading diagonal elements for the variance-covariance matrices \( P(t_0) \), \( Q \) and \( \Lambda \).

<table>
<thead>
<tr>
<th>State/Parameter</th>
<th>Initial Values</th>
<th>( P(t_0) )</th>
<th>( Q )</th>
<th>( \Lambda )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_1(t) ) (mg·L(^{-1}))</td>
<td>7.19</td>
<td>—</td>
<td>—</td>
<td>0.2129</td>
</tr>
<tr>
<td>( x_2(t) ) (mg·L(^{-1}))</td>
<td>50.0</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>( \mu_{max} ) (d(^{-1}))</td>
<td>0.24</td>
<td>0.00024</td>
<td>0</td>
<td>—</td>
</tr>
<tr>
<td>( k_d ) (d(^{-1}))</td>
<td>0.024</td>
<td>0.0000024</td>
<td>0</td>
<td>—</td>
</tr>
<tr>
<td>( r(t) ) (mg·L(^{-1})·h(^{-1}))</td>
<td>1.0</td>
<td>0.02</td>
<td>0.00001</td>
<td>—</td>
</tr>
<tr>
<td>( u_2(t) ) (mg·L(^{-1}))</td>
<td>50.0</td>
<td>5.0</td>
<td>0.0002</td>
<td>—</td>
</tr>
<tr>
<td>( K )'s</td>
<td>0.0</td>
<td>0.001</td>
<td>0.0</td>
<td>—</td>
</tr>
</tbody>
</table>

Having specified the initial conditions for all states and parameters, initial covariance matrices for all parameters, and the variance matrix for measurement errors, the modified RPE algorithm (illustrated in Section 3.3.2) has been applied to the simple mechanistic model of the dynamics of the nitrification process occurring in the aeration tank of activated sludge system. The reconstructed state variables, the estimates of parameters (including the parameterized input function of the nitrifying bacteria), and the estimated Kalman-like gain matrix entities are shown in Figures 4.15, 4.17, and 4.18, respectively. Before we move to analyze these recursive estimation results, however, it is worth mentioning that these reconstructed or estimated results for the state variables, the parameters and the elements of the Kalman-like gain matrix are not smoothed by the backward pass smoothing procedures (e.g., FIS algorithm; see the discussion in Chapter 3, [89] and [260] for detailed...
information). Caution must therefore be exercised when interpreting the transient recursive estimates during the initial stages of the recorded period. For instance, the sudden increase of the nitrifying bacteria biomass concentration on days 2 and 3, shown in Figure 4.15(b), is unlikely to have resulted from the natural growth of nitrifying bacteria, which is a rather slow process. In addition, the apparent phase lag between the one-step-ahead prediction and the observations of the ammonium-nitrogen concentration in the inner channel (output of our system) shown in Figure 4.15(a) is one of consequences of the recursive estimation without a coupled smoothing scheme.

![Recursive estimation results](image)

Figure 4.15: Recursive estimation results: (a) one-step-ahead prediction vs. observations of ammonium-N concentration, and (b) reconstructed nitrifying bacteria concentration.

First of all, let us scrutinize carefully the (reconstructed) dynamics of the nitrifying bacteria biomass concentration in the aeration tank, since it is crucial to the entire nitrification process. On March 18th (day 7), there was a small rainfall event (0.35 in) that caused, over the following three or four days, a slight increment in the crude sewage flow rate (see Figure 4.2(a)) and an excitation to the system (a step in the input signal, i.e., the increased ammonium-nitrogen in the outer channel, as shown in Figure 4.3(b)). Before this event, the
population of the nitrifiers remained at a level of about 50 mg·L\(^{-1}\) with healthy diurnal oscillations (omitting the sudden changes in the first three days). Shortly after the event hit the system, the biomass concentration of either the MLSS or the nitrifying bacteria experienced a small plunge, which may be caused by the flush-out of the MLSS, consisting of nitrifying bacteria, from the aeration tanks. During the period of the rainfall event and a few days thereafter, the activated sludge was taken out from the aeration tanks and stored in the following clarifiers, which is accomplished by manually decreasing the flow rate of the RAS (see Figure 4.2(b)) and prevents the activated sludge from being flushed out from the treatment plant due to an increased influent flow rate. This operation causes: (1) reduction of the activated sludge concentration in the aeration tank (see Figure 4.4(a)); and (2) a drop in the removal efficiency of the ammonium-nitrogen of the system (see Figure 4.16(a)).

After this event, the regular "dosage" of the returned activated sludge was restored, the "refuge" of the biomass in the clarifiers was then transferred back into the aeration tanks. Therefore, although the growing process of autotrophs is relatively slow, a matter of weeks, the population of the nitrifiers in the aeration tanks recovered very quickly, within a couple of days, after this disturbance. In turn, the nitrification process was restored as well, which is indicated by the recovery of the ammonium-nitrogen removal efficiency shown in Figure 4.16(a). Comparing the variation pattern of the ammonium-nitrogen removal efficiency with those of the MLSS and nitrifying bacteria (Figure 4.16) shows that the changing of the ammonium-nitrogen removal efficiency is more correlated to the dynamics of the nitrifying bacteria than the MLSS. This implies that, in this system, the nitrification process, rather than the uptake for organism cell synthesis, is more responsible for the ammonium removal, which reinforces the arguments set in Section 4.3.2 of the data-based modeling approach.

The recursive estimation results for the parameters, including the elements in \(K\) matrix, are presented in Figures 4.17 and 4.18. Note that the spikes appeared in the beginning of the estimates of the parameters probably are caused by setting the initial values for the leading diagonal elements in the variance-covariance matrices \(P(t_0)\) too large. They are not
Figure 4.16: (a) Removal efficiency of ammonium-nitrogen; (b) The biomass concentrations of the reconstructed nitrifying bacteria and the MLSS trend signal.

reflecting the true variation of the parameters. However, it is interesting to notice that, whether the parameter is assumed to be \textit{time-invariant}, like $\mu_{\text{max}}$ or $k_d$, or is assumed to be \textit{time-variant}, like $r(t)$ or $u_2(t)$, all of them experienced substantial changes (relative to their nominal values) shortly after the rainfall event disturbed the system. For $\mu_{\text{max}}$, even though environmental factors such as temperature and DO, have been taken into account in the model, its value increased by about 10% when there was sufficient substrate (ammonium-nitrogen) flowing to the system to feed the nitrifiers. The decay rate of the nitrifiers $k_d$ decreases for the same reason around the same time period, i.e., March 20th – 23rd, when the biomass concentration of nitrifiers was growing, as shown in Figure 4.15(b).

Parameter $r(t)$ measures the gain of ammonium-nitrogen in the system due to processes such as mineralization of soluble organic matters, which, in turn, is a function of the decay
process of bacteria and the concentration of organic matters in the system. Ideally, if the concentration of organic matters is not subject to considerable change, the variation of $r(t)$ reflects the changes of the decay rate of organisms in the activated sludge. Figure 4.17 shows that the variation pattern of $r(t)$ coincides with that of $k_d$, although the former is a function of the decay rate of both heterotrophs and autotrophs while the latter measures the decay rate of autotrophs only. Altogether with the above interpretation for $\mu_{\text{max}}$'s changing pattern, this means that, during March 20th – 23rd, the time of recovery, the system is subject to net gains of bacteria, which is confirmed by the plots of the simulated nitrifier’s biomass concentration and the measured MLSS concentration in Figure 4.16(b). The estimated dynamics of $u_2(t)$ is similar to the reconstructed dynamics of the nitrifying bacteria $x_2(t)$ in the system both in the variation pattern and magnitude.

Figure 4.17: Recursive estimation results: estimates (solid line) and standard errors (dashed line, almost invisible) of the time-invariant parameters ($\mu_{\text{max}}$ and $k_d$) and the time-varying parameters ($r(t)$ and $u_2(t)$).

In principle, any dramatic shift from zero of any element in $K$ is an indication of the inadequacy of the model’s representation of the real system. It can be caused for various
reasons. For example, the reasons can be: (1) wrong model structure, i.e., wrong constituent hypotheses about the processes governing the real system’s behaviors; (2) biased parameter estimates; (3) inappropriate initial conditions for state variables and/or starting values for parameters; and (4) unaccounted for substantial system changes during the course of the record; etc. The first jump displayed in $K_{1,1}$ (see Figure 4.18) is a result of the third reason of the above listed. If we change the starting value of $r(t)$ from 1.0 to 1.5, the sudden change of $K_{1,1}$ from 0 to 0.2 at day 2 would be replaced by a drifting of the $K_{1,1}$ from 0 to 0.2 spread over the first 6 days until the rainfall event occurred. The second jump of $K_{1,1}$ successfully indicates that the system was experiencing a remarkable disturbance that is beyond the model’s deterministic description of the real system. This disturbance was the influent flow rate increase starting on day 7. The fact that $K_{2,1}$ varies less than $K_{1,1}$ results from the fact that there are no observations of nitrifiers such that the RPE does not have to force $K_{2,1}$ to vary in order to minimize the mismatch between the prediction from the model and the observations. The relatively small values of $K_{2,1}$ also indicates that the reconstruction of the nitrifier’s dynamics relies more on the model itself than on data. In contrast, the reconstruction of ammonium-nitrogen relies more on the data than on model, which is implied by the large positive values of $K_{1,1}$. 
Figure 4.18: Recursive estimation results: estimates (solid line) and standard errors (dash line) of the elements of the $K$ matrix.

4.5 Towards the synthesis

In the beginning of this chapter (Section 4.1), we have asked ourselves three questions pertaining to the two-pronged modeling approach: (1) how the two separate modeling approaches can help us to acquire more information about the environmental system; (2) how the two separate modeling approaches can benefit from each other to gain a deeper understanding about the environmental system; and (3) how the two opposite modeling approaches can be reconciled under parameter space.
4.5.1 A SUMMARY OF THE TWO SEPARATE MODELING APPROACHES

The response\(^3\) to the first two questions has been made during the course of separately applying the data-based and theory-based modeling methods to this case study in the previous two sections. For example, firstly, in the data-based modeling procedures (Section 4.3) we found that the nitrification process is mainly responsible for the ammonium-nitrogen removal in the activated sludge system. Therefore, when building our theory-based model to approximate our real system in the theory-based modeling procedure (Section 4.4), we focused on the nitrification process and omitted the uptake of nitrogen for organism cell synthesis, thereby simplifying our theory-based model. Secondly, the analysis of the DBM approach identifies that a pure time delay of one sampling interval exists in the system, so that a combination of a PFR and CSTR, instead of a sole CSTR, was employed to idealize the system in the TBM approach. Thirdly, the DBM approach found that there exists some kind of relationship between the removal of the ammonium-nitrogen and the MLSS, a rough measure of bacteria biomass in the aeration tank. However the variations of these two signals are out of phase. Under the context of DBM, we have to resort to our speculation to squeeze a reasonable explanation for this phenomenon (see discussion in 4.3.2). Fortunately, in the TBM, we discovered the reasons behind this “out-of-phase” phenomenon by reconstructing the nitrifying bacteria biomass concentration, which constitutes the MLSS (see Figure 4.16). That is to say, making use of the information obtained by the DBM approach, the TBM approach rewards us with more insightful understanding of the activated sludge system.

4.5.2 SYNTHESIS UNDER PARAMETER SPACE

Figures 4.19 and 4.20 display comparisons of the two types of models that are derived, either from the DBM modeling approach in Section 4.3, or from the TBM modeling approach in Section 4.4. Three DBM models and one TBM model have been listed here. The DBM...
model (4.1) with [1, 1, 1] structure is a linear TF model with constant parameters; while the DBM model (4.3) is a linear TF model with \textit{time-varying} parameter; and the DBM model (4.5) is an extension of model (4.3) with its input ammonium-nitrogen weighted by the corresponding influent flow rate.

The one-step-ahead predictions of these four models are plotted in Figure 4.20, together with the ammonium-nitrogen observations of the output. On this comparison, with respect to the observed state variables, even a close inspection cannot tell which model is superior to the others except that the DBM model with constant parameters is marginally inferior. This is because the one-step-ahead prediction of the models are always forced by the algorithms to follow the trajectories of the observations. Comparison of the models with respect to their \textit{steady-state gains} in Figure 4.19(b) does not tell much of a story either, since the \textit{steady-state gain} essentially is a ratio of the output to input signals — functions (usually linear) of the observed state variables. However, the comparison of the models under the space of the parameter \textit{time constant} $\tau$ distinguishes the DBM model (4.5) from the other DBM models and makes it comparable to the TBM model. It is also interesting to note that Figure 4.19(a) displays the progress made by the three DBM models as they are developed in sequence. From model (4.1) to model (4.5), their \textit{time constants} become more and more close to that of the TBM.

Focusing on the two comparable models, DBM (4.5) and TBM (4.6), and their equivalent \textit{time constants} and \textit{steady-state gains}, we can make use of the knowledge of the \textit{time constant} to draw inferences about the other parameters in the TBM model, which contains some information about the system itself. Let us rewrite the first differential equation of the TBM model, which describes the dynamics of the ammonium-nitrogen concentration in the system:

\[
\frac{dx_1(t)}{dt} = -k(t)x_1(t) - \frac{Q(t)}{V}x_1(t) + \frac{Q(t)}{V}u_1(t - \delta)e^{-2sk(t)} + r(t)
\]

(4.7)

in which $Q(t)/V$ is the dilution ratio (h$^{-1}$), while $k(t)$ is the degradation rate of ammonium-nitrogen with unit h$^{-1}$, as defined by formulation (3) in Table 4.2. The \textit{time constant} of (4.7)
Figure 4.19: Comparisons of the two types of models (DBM(4.1): cyan dashed line; DBM(4.3): green dot-dashed line; DBM(4.5): blue solid line; and TBM (4.6): red dotted line) at: (a) time constants and (b) steady-state gains.

Figure 4.20: Comparisons of the two types of models with respect to model predictions.
Figure 4.21: (a) Components of the time constant of the TBM; (b) time constants of the TBM and DBM.

is calculated by the following equation:

\[
\tau(t) = \frac{1}{Q(t)/V + k(t)}
\]  

(4.8)

It is obvious that the value of the time constant of this system is determined by the two terms of the denominator in the RHS of equation (4.8), \(Q(t)/V\) and \(k(t)\). Both of them are functions of time and are plotted with each other in Figure 4.21(a). By comparing their patterns of variation with those of the time constants plotted in Figure 4.21(b), it is not hard to find out that the lower-frequency wavelike variations of the time constants are determined by those of \(k(t)\). Therefore, the magnitudes of \(Q(t)/V\) and \(k(t)\) should be at least at a comparative level (in other words, \(Q(t)/V\) should not dominate \(k(t)\) in value). That is to say,

\[
\frac{Q(t)}{V} = ck(t), \quad c \text{ is sufficiently close to } 1.
\]  

(4.9)
Therefore, of the three quantities $Q(t)$, $V$ and $k(t)$, given the information of two, we could identify roughly the range of the value for the third one. For example, in this case, having measured $Q(t)$ and predetermined the value of $k(t)$ through formation (3) in Table 4.2, then we can be assured that the effective volume of 6250 m$^3$ is a reasonable number for $V$. 
Chapter 5

Case Study Two: Algal Blooms in a Small Southeastern Piedmont Impoundment

5.1 A review of recent development in phytoplankton modeling

Phytoplankton play a major role in the dynamics of aquatic ecosystems. They are the “grass” of lakes and seas and most species have a worldwide distribution. As primary producers, they constitute the basis for total ecosystem production and exert direct effects on dissolved oxygen, pH, and nutrient levels in the water. However, excessive growth of the phytoplankton community causes many problems such as bad taste of fish, depletion of dissolved oxygen due to sudden massive phytoplankton die-offs and even ill-balanced ecosystems. The occurrence of phytoplankton blooms is a common problem in water bodies. Tremendous research attention has been drawn into trying to unravel the mysteries of the sudden formation and subsequent collapse of algal blooms and thus develop more successful water quality management methods. The broad topics and diverse approaches make any attempts to condense these fruitful efforts into a few pages certainly inappropriate. Comprehensive reviews of ecological modeling in the 1970’s and early 1980’s have been given in [1, 122, 223, 273]. Physical (or hydraulic, hydrological, hydrodynamic) processes driving stratification and mixing in a lake or transportation in a river strongly influence the variation of phytoplankton in water bodies. The complexity of the description of these physical processes varies considerably among different models. A short review of recent physical submodel development in lake
water quality has been given by [164]. The following two sections briefly review the development of the theory-based modeling and the data-based modeling of phytoplankton behavior in aquatic ecosystems, primarily over the past decade.

5.1.1 Theory-based modeling approaches

In general, the progress made recently on theory-based modeling comes from two aspects: modeling the complex community structure of the aquatic organisms and modeling the various processes regulating the dynamics of phytoplankton.

Complexity of community structure

In the 1970’s and early 1980’s, with few exceptions (e.g., [107, 199]), most of the ecological models simply described phytoplankton as one assemblage class, or at most a couple of functional classes (e.g., blue algae vs. non-blue algae, or algae vs. diatom), and zooplankton, if any, as one lumped group ([9, 15, 60, 83, 120, 121, 210, 228]; also see reviews by [122, 160, 223, 273]). Partly due to data restriction, these features of the “old” models are still inherited by some contemporary “big” physical-biological lake water quality models, such as CE-QUAL-W2 ([61, 63]), DYRESM Water Quality ([95, 203]), LIMNMOD ([126]), MINLAKE ([191]) and many others (e.g., [164, 165, 273, 274]). However, in an aquatic system, phytoplankton and zooplankton sometimes consist of several groups that, in turn, are composed of several species at the same time, which show different biological and ecological characteristics ([110, 116]). Therefore, to seek a more detailed description of the composition of organisms in lakes and streams is inevitable if data availability, thus identifiability of the model, permits ([123]).

In contrast to the simple description of phytoplankton and zooplankton in the models described above, in PROTECH ([77, 78, 79, 190]) a detailed phytoplankton model with up to eight algal species from a library of eighteen species can be used. This model is based on maximum specific growth rates of these species measured under ideal culture conditions. Although the results for individual species do not correspond very well to observations, the
results for the sum of all species lead to good agreement with data. This is interpreted as an indication of the need to model one level further down than required for interpretation ([78]). The one-dimensional model of a hyper-eutrophic, man-made lake (Lake Villerest, Loire, France [44]) takes into account all the five dominant phytoplankton species that were identified and counted in the lake ([4]), namely, two species of diatoms, two of chlorophyceae, and one cyanobacteria (*Microcystis aeruginosa*), but with zooplankton being introduced as a forcing variable instead of a separate state variable. Since cyanobacteria are able to regulate their migration in the water column ([111, 136, 189, 231]), this classification of phytoplankton allows a buoyancy regulation algorithm ([2, 162, 163]) to be introduced in the model to properly represent *M. aeruginosa* vertical motions. Hence, this model showed that buoyancy regulation is a major advantage for this species allowing its domination over the other species at least when the vertical stratification is well established.

Since biomanipulation is one of the most promising methods of preventing harmful algal blooms due to eutrophication ([208, 40, 174]), a lake ecological model can serve as a powerful tool to predict its effects. The model proposed by Suzuki *et al.* ([226]) is composed of nineteen state variables including two types of algae (non-blue-green algae and blue-algae) and three types of zooplankton (rotatoria, cladocera and copepoda). Its parameters were calibrated against mesocosm observations ([195]). The simulation study of this model indicated that the zooplankton composition has an important influence in suppressing algal growth in biomanipulation. However, the scope of the model is limited, since it only includes two types of algae, which is a small number relative to those in other structural dynamic models (e.g., nine species of algae are dealt with in [158, 159]). Another biomanipulation model of a shallow lake (Lake Bleiswijkszoom, the Netherlands, [116]) divided the fourteen types of phytoplankton into three groups depending on their behavior in response to external environmental factors — nutrient levels and solar radiation. The maximum net growth rates, and light attenuation coefficients, for these three groups were accordingly defined as nitrogen limited, phosphorus limited, or energy limited (solar radiation) species ([144, 201]).
A High Rate Algal Pond (HRAP) is an efficient treatment system for controlling wastewater pollution through the optimization of algal growth ([179]). Computer modeling provides a valuable vehicle to achieve a quantitative understanding of the relationship between key design/operation parameters and the growth of the algae ([81, 128, 145, 149]). When modeling the grazing process in a HRAP, Mesplé et al. ([149]) take into account both the plankton (i.e., phytoplankton and zooplankton) community structures and the interaction between these two compartments. There are ten phytoplankton taxa and five zooplankton taxa observed in the HRAP, three of which are considered in the model. Since it is impossible to calibrate a model including the ten phytoplankton taxa as state variables, a forcing variable — “Phyto Structure” — is used. This forcing variable represents the daily taxonomic composition of the phytoplankton population as a percentage of total biomass. Only the total mass of chlorophyll-a is computed, then distributed in each taxon in proportion to their biomass. However, the taxonomic determination is time consuming, tedious, and allows only a posteriori information of “Phyto Structure”. Meslé et al. ([149]) suggest that one may categorize the phytoplankton species in terms of cell size instead of taxonomy in order to succeed with fewer phytoplankton state variables. These regrouped phytoplankton classes in terms of cell size should have homogeneous physiologies (growth rate, death rate, etc.) and they must react in the same way to variations in temperature, radiation, nutrient concentrations, zooplankton grazing, etc. ([149, 223]). A summary of some of recent theory-based models for algal growth or eutrophication is given in Table 5.1.
Table 5.1: Summary of recent theory-based models for eutrophication.

<table>
<thead>
<tr>
<th>Models</th>
<th>DM&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Phy</th>
<th>Zoo</th>
<th>Mac</th>
<th>Sys</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>DYRESM Water Quality [95]</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>N&lt;sup&gt;b&lt;/sup&gt;</td>
<td>L&lt;sup&gt;c&lt;/sup&gt;</td>
<td>BW&lt;sup&gt;d&lt;/sup&gt;</td>
</tr>
<tr>
<td>Omlin et al. [164, 165]</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>N</td>
<td>L</td>
<td>M</td>
</tr>
<tr>
<td>NORECO [207]</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>N</td>
<td>L</td>
<td>M</td>
</tr>
<tr>
<td>MINLAKE [191]</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>N</td>
<td>L</td>
<td>—</td>
</tr>
<tr>
<td>PROTECH [77, 78, 79]</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>N</td>
<td>L</td>
<td>—</td>
</tr>
<tr>
<td>Rostherne [134, 135]</td>
<td>0</td>
<td>2–4</td>
<td>0</td>
<td>N</td>
<td>L</td>
<td>SEM&lt;sup&gt;e&lt;/sup&gt;</td>
</tr>
<tr>
<td>Bonnet &amp; Poulin [44]</td>
<td>1</td>
<td>5</td>
<td>FV&lt;sup&gt;f&lt;/sup&gt;</td>
<td>N</td>
<td>L</td>
<td>D</td>
</tr>
<tr>
<td>Asaeda &amp; Bon [10, 12]</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>Y</td>
<td>SL</td>
<td>M,Y</td>
</tr>
<tr>
<td>Frisk et al. [85]</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>N</td>
<td>L</td>
<td>W</td>
</tr>
<tr>
<td>Ginot &amp; Hervé [90]</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>N</td>
<td>SL</td>
<td>H</td>
</tr>
<tr>
<td>Zeng et al. [274]</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>N</td>
<td>SL</td>
<td>H</td>
</tr>
<tr>
<td>Osidele [167]</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>N</td>
<td>L</td>
<td>—</td>
</tr>
<tr>
<td>Jayaweera &amp; Asaeda [116]</td>
<td>0</td>
<td>14</td>
<td>1</td>
<td>Y</td>
<td>SL</td>
<td>M</td>
</tr>
<tr>
<td>Suzuki et al. [226, 195]</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>Y</td>
<td>SL</td>
<td>D</td>
</tr>
<tr>
<td>Mesplé et al. [149]</td>
<td>0</td>
<td>FV</td>
<td>3</td>
<td>N</td>
<td>HRAP</td>
<td>D,W</td>
</tr>
<tr>
<td>QSIM [205]</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>N</td>
<td>R</td>
<td>D</td>
</tr>
<tr>
<td>Griffin et al. [93]</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>N</td>
<td>E</td>
<td>H,W</td>
</tr>
</tbody>
</table>

<sup>a</sup> DM-Dimension; Phy-Phytoplankton <i>sp</i>; Zoo-Zooplankton <i>sp</i>; Mac-Macrophytes; Sys-System; Data-Sampling Frequency.

<sup>b</sup> N-Not Important; Y-Important.

<sup>c</sup> L-Lake; SL-Shallow Lake; R-River; E-Estuary; HRAP-High Rate Algal Pond.

<sup>d</sup> H-Hourly; D-Daily; W-Weekly; BW-Biweekly; M-Monthly; Y-Yearly.

<sup>e</sup> Scanning Electron Microscopy X-ray Microanalysis (SEM XRMA).

<sup>f</sup> Forcing Variable.
Variety of regulating processes

Phytoplankton in an aquatic ecosystem are regulated by processes such as growth, respiration, mortality, decomposition, sedimentation, and grazing; these in turn are governed by environmental parameters such as solar radiation, temperature, nutrient concentrations, and even iron (in the south Pacific Ocean [38, 155]). As demonstrated by [122], the mathematical expressions describing these ecological processes vary from model to model. It is generally possible in models or in the biological literature to find different mathematical descriptions of the same biological process. Although the estimation of the phytoplankton growth kinetics and relevant environmental factors remains a focus (e.g., [90, 128, 151, 221, 249]), an increasing number of studies suggest that “top-down” controls via the upper trophic levels (e.g., zooplankton grazing) may play an important role in regulating the phytoplankton community ([53, 67, 149, 202, 215, 205, 227]). There has been considerable debate about the role of zooplankton in controlling algal blooms through grazing ([45, 69, 175]). Various studies have shown the depletion of phytoplankton biomass to be dependent on the density and size of the zooplankton grazers ([232]), diel variation in zooplankton feeding rates ([176]), and phytoplankton abundance ([75]). Nonetheless, in modeling the eutrophication of inland waters and their coherent management, modelers have progressively included more and more sophisticated mechanistic representations of biological processes, such as phytoplankton development and food-chain interactions ([54]).

Biotic regulation and feedbacks of phytoplankton biomass in large rivers are more complicated, and much less well understood. Some investigations have shown a considerable to modest influence of grazing by zooplankton ([70, 71, 87, 92, 132]). And several studies suggest a significant impact from benthic filter feeders ([5, 14, 53, 87, 132, 204, 242]). Based on an extensive monitoring program of the River Rhine in 1990 including water quality parameters, plankton and the macrozoobenthos, the water quality model QSIM (version 8.3) ([205]) includes three ecological groups represented by green algae and diatoms for phy-
to plankton, by *Brachionus angularis* for rotifers, and by the zebra mussel for benthic filter feeders. Biomass growth of all three groups is simulated in the same way ([205]). The calculated losses of phytoplankton showed a seasonal dependency with a maximum grazing rate of about 0.04 d$^{-1}$ for rotifers and 0.2–0.3 d$^{-1}$ for zebra mussel. However, both grazing groups were unable to control algal development in the reach of River Rhine, where conditions are favorable for phytoplankton. The model CAEDYM (Computational Aquatic Ecosystem Dynamics Model) was used to simulate the dynamics of phytoplankton and zooplankton in the Swan River Estuary, Western Australia, where phytoplankton are differentiated as four groups and zooplankton as three size classes ([93]). The model clearly demonstrated the impact of zooplankton grazing on phytoplankton biomass. When examining the mechanisms behind loss of phytoplankton biomass, loss due to respiration was relatively constant and insufficient to cause the decline seen in the *in situ* data. But the net effect of zooplankton grazing increased over the 3-week simulation period and appeared to be the major factor contributing to the loss of phytoplankton biomass overall.

When modeling the HRAP’s described earlier ([149]), two HRAP’s with different retention times were compared (zooplankton species whose generation times are longer than the retention time are flushed out). The simulation results show that zooplankton grazing imposes vital effects on the growth and succession patterns of chlorophyll-a in the pond ecosystem. For low zooplankton abundance, a rough seasonal estimate of the variations in phytoplankton biomass is possible without consideration of zooplankton grazing in the model. In contrast, when zooplankton abundance is high, the error in the simulation is significant.

In addition, in shallow eutrophic lakes ([10, 12, 116, 246]) and rivers ([129, 169]), it is evident that the algal growth and DO changes are affected by macrophytes. Aquatic plants prevent the interception of nutrients by algae before nutrients infiltrate into the sediments ([13]), obtain nutrients directly from the sediments ([150, 182]), act as a refuge for zooplankton, and sometimes produce allelopathic substances that suppress algal growth ([105, 206]). Submerged plants can raise the sediment Redox potential, lower the pH and filterable Fe and
P proportions, enhance sediment P retention ([117]), and establish favorable conditions to reduce ammonium-N, and thus alter the release rate of P, NH$_4^+$ and NO$_3^-$ from the sediment into the overlying water ([57]). Heavy growth of macrophytes inhibit the production of phytoplankton by shading. When macrophytes are dominant, phytoplankton production decreases by as much as two orders of magnitude, and conversely, persistent, large phytoplankton growth may occur when macrophyte growth decreases ([150]). A number of models have been proposed that describe partly the process of the function of macrophytes in aquatic ecosystems ([62, 114, 115, 154, 173, 236, 239]). A dynamic model for eutrophication incorporating phytoplankton and nutrients in the overlying water, submerged macrophytes Potamogeton pectinatus L., and nutrient dynamics in sediments was developed to understand the effects of macrophytes on algal blooming in shallow lakes ([10, 12]). Its application to the Lake Veluwe (Netherlands) experiment is successful and the sensitivity analysis of important coefficients indicated that the maximum photosynthesis and respiration rate are the most sensitive parameters for macrophyte and algal development.

5.1.2 Data-based modeling approaches

Since the late 1970s, linear regression models, most of which are Single-Input-Single-Output (SISO) regression models, have been applied to phosphorus-limited lakes to predict the chlorophyll-a concentration ([161, 184, 185]), timing of algal bloom ([146]), and the relation between phytoplankton and zooplankton biomass ([148, 207]). Since many driving variables besides phosphorus (or nutrients) are known to control the behavior of algae, a multivariable time series analysis is therefore required. In the multi-variable regression model of predicting downstream chlorophyll-a concentration in the River Thames, developed by Young and Whitehead ([271]), the independent variables include upstream chlorophyll-a, flow and solar radiation. The model parameters were estimated using a recursive estimation algorithm ([244]). Using the time series analysis methodology proposed in [48], Stronge et al.
developed a multiple regression model for the mean April–June chlorophyll-a concentrations in Lough Neagh, Northern Ireland, for the time period 1974–92, by relating these concentrations to independent causal variables, i.e., the previous year’s spring chlorophyll-a concentration, soluble reactive P inputs for April–June and the particulate P concentration during the previous summer. The time series model explained 76% of the annual variation in spring chlorophyll-a concentration. Both Multiple-Input-Single-Output (MISO) linear regression and Stochastic Transfer Function (STF) models are used for reconstruction of daily algal concentration (chlorophyll-a) in the middle stretches of the Elbe River during the years 1985–2001 ([193]). The Stochastic Transfer Function model based on oxygen concentrations and temperature gives the best fit to the data, followed closely by the model additionally using pH values as another input (see Table 5.2).

In recent years the technique of neural networks has also been applied to water science to simulate algal concentrations in reservoirs ([187]) and rivers ([85, 245]). This approach is based on the concept of learning activities such that the procedure itself evolves an optimal model relating key variables. Like time series analysis, the neural networks approach takes into account no *a priori* process knowledge to determine the model structure or estimate parameters; it simply draws information out of the data by recognizing patterns of behavior. Thus in analyzing large data sets with no *a priori* knowledge of processes or causality, neural network methods are a more pragmatic approach. However, in terms of representing processes neural networks will be severely limited. Only generalized process statements can be made from the neural network models and for a process interpretation of algal growth behavior the dynamic mass balance and growth equation approach will always be required. Table 5.2 summarizes some of the data-based models for algal growth developed recently.

5.1.3 Discussion

No matter how much effort has been made to understand the mechanisms of phytoplankton behavior in aquatic systems, the formation and timing of algal blooms are still one of the
Table 5.2: Summary of recent data-based models for algal growth.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Method</th>
<th>Variablesa</th>
<th>System</th>
</tr>
</thead>
<tbody>
<tr>
<td>Young &amp; Whitehead [271]</td>
<td>Recursive estimation</td>
<td>Chla, PAR</td>
<td>River</td>
</tr>
<tr>
<td>Stronge et al. [224]</td>
<td>ARIMA</td>
<td>Chla, SP, PP</td>
<td>Lake</td>
</tr>
<tr>
<td>Romanowic &amp; Petersen [193]</td>
<td>Linear regression</td>
<td>DO, pH</td>
<td>River</td>
</tr>
<tr>
<td>Romanowic &amp; Petersen [193]</td>
<td>STF</td>
<td>DO, pH, Temp</td>
<td>River</td>
</tr>
<tr>
<td>Whitehead et al. [245]</td>
<td>Neural network</td>
<td>Chla, F, PAR, Temp</td>
<td>River</td>
</tr>
</tbody>
</table>

a Independent variables.
b Upstream chlorophyll-a concentration.
c Previous spring’s chlorophyll-a concentration.
d Soluble reactive P concentration.
e Previous summer’s particulate P concentration.

mysteries ecologists are trying to solve. Due to the complexity of the structure of aquatic ecosystems and lack of sufficient observations of them, it is inherently very difficult to make any progress.

When calibrating the long-term models, weekly, monthly or even yearly data sets are used on most occasions (see Table 5.1). However, Harris ([98]) reviewed the evidence for the existence and importance of both small-scale, high-frequency and large-scale, low-frequency variation in the planktonic environment. At the lowest levels processes are operating at high frequencies so at the lowest level subsystems appear to be at equilibrium over the time span of interest. At the largest scales (low frequency) the period may be so long that the properties of the system may appear to be constant over the period of study. But, in the middle range of level and frequency the observable dynamics are not assumed to be at steady-state and must therefore be studied and characterized. Obviously, most of the existing observations are not sufficiently frequent for the modelers to understand the dynamics of the processes such as photosynthesis, with a time constant of hours to days ([98, 99]). Sub-hourly measurements
taken by the Environmental Process Control Laboratory (EPCL, described in Section 4.2) and supporting instruments provide the solution to this problem.

In the following sections, the joint modeling approach sketched in Chapter 3 has been applied to the present case study of understanding phytoplankton blooms in a small south-eastern impoundment excited by deliberate fertilizations. Section 5.2 gives the details of the site and the experiment conducted in the summer of 2000. Sections 5.3 and 5.4 illustrate the implementation of the signal processing and the Data-Based-Mechanistic modeling techniques, respectively. Sections 5.5 and 5.6 exercise the modified RPE algorithm to estimate the time-varying parameters in a simple mechanistic model for algal growth and to detect the changes of structure in the system. Section 5.7 looks back and synthesizes the results and discussions of the DBM and TBM modeling approaches.

5.2 SITE, EXPERIMENT AND DATA COLLECTION

The study site is a small impoundment in the Southeastern Piedmont, located at the Whitehall estate of the Warnell School of Forest Resources, University of Georgia (83°24′W, 33°54′N, [66]). This aquaculture pond, adjacent to Oconee River, is fed by a first order stream draining 13 ha of forest, deer pens, barns, and gravel roads and parking lots within the Whitehall Experimental Forest. The volume of the pond is approximately 2000 m³, the surface area is less than 7000 m², and the maximum depth near the dam is about 2 m (see Figure 5.1 for a picture of the Whitehall pond). The outflow near the dam is regulated by a stand pipe ([171]). Low flow conditions obtained throughout the experiment due to summer drought.

The experiment was conducted from May 23rd through October 16th, 2000 (147 days). The pond was limed on July 2nd with 45 kg agriculture lime (Ca(OH)₂) broadcast from a Jon boat, which was followed by three fertilizations on July 5th, July 27th, and October 1st, respectively. The first two fertilizations involved distributing 8 liters of liquid ammonium
polyphosphate fertilizer of 11:37:0 (N:P:K) (approximately 0.9 kg NH$_4^+$-N and 3 kg PO$_4^{3-}$-P) to the pond and monitoring the system responses using real-time monitoring equipment described below. The third fertilization involved mixing 8 liters of liquid fertilizer aforementioned with 55 kg of Pacolet series Bt horizon soil (85 % fines, 15 % sand, see [171] for the detailed soil treatment), waiting for 24 hours before distribution of the resulting slurry to the pond, and monitoring the system responses.

Following the first fertilization, duckweed *Lemma* — a species of small free-floating macrophytes, commonly appearing in temperate and subtropical shallow, wind-protected ponds, irrigation ditches, and backwaters ([110]) — started to develop from the littoral zone at first and finally covered 80–90% of the pond. On September 27th, a few days prior to the third treatment of fertilization, the duckweed was removed mechanically by seining and raking from a kayak.
The Environmental Process Control Laboratory (EPCL, see Chapter 4), together with supporting monitoring facilities (Turner fluorometers, Hydrolab multi-parameter data-sondes), were deployed on-site (see Figure 5.2) in order to acquire comprehensive and detailed time-series observations with a sampling frequency of the order of 15 minutes or less. The variables measured by the EPCL consist of ammonia-nitrogen ($\text{NH}_4^+$-N), nitrite-nitrogen ($\text{NO}_2^-$-N), total oxidized nitrogen (TON$\approx\text{NO}_2^-$-N+$\text{NO}_3^-$-N), total oxidizable carbon (TOC), orthophosphate ($\text{PO}_4^{3-}$-P), suspended solids (SS), and temperature. All the measurements were taken at 15 minutes intervals. Two fluorometers were incorporated within the EPCL to measure chlorophyll-a concentrations at 15 minutes intervals. Water samples to the EPCL and fluorometers were pumped continuously from 0.5 m depth of the pond at each of the two sample locations (uptake 1 is the dock near the dam while uptake 2 is 

Figure 5.2: Sampling locations at the Whitehall pond (reproduced from [273]).
the platform near the inflow stream, see Figure 5.2). Equidistant between the two sampling intakes, a submerged drain baffle was constructed to accommodate the returning sample with the minimization of disturbance to system.

Other water quality variables, such as oxidation-reduction potential (ORP), pH, specific conductivity, photosynthetically active radiation (PAR), and dissolved oxygen (DO), were measured by two Hydrolabs at the same sampling locations as the EPCL intakes (see Figure 5.2). The sampling intervals are 15 minutes as well. Figure 5.3 shows three examples of the raw data sets retrieved by the EPCL, fluorometers and Hydrolabs, respectively, at the platform sampling location (uptake 2) from May 23rd to October 16th, 2000. However, only a small portion of these data — from July 1st, 2000 through August 5th, 2000 (36 days) recorded from the platform sampling location (uptake 2), covering the liming and first two fertilization — are examined in detail herein.

Figure 5.3: Examples of the raw data sets retrieved from the EPCL, fluorometers and Hydrolabs at location 2 (from May 23rd to October 16th, 2000): (a) from the EPCL, (b) from the fluorometers, (c) from the Hydrolabs.
5.3 Signal processing

Prior to frequency analysis, each time-series is decimated every eight samples so that the sampling intervals increase to 2 hours, i.e., 12 observations per day. So doing is because we know by plotting the data out in a spreadsheet that the major seasonal variation of each time-series from the Whitehall pond attaches to the diurnal frequency. Therefore, it requires at least two observations/day (samples/cycle) in order to describe diurnal oscillation ([56]). Moreover, the longest period (1/frequency) accounted for by the spectral analysis software used in this research (CAPTAIN) is 24 samples/cycle. That is, in order to get information about variation within one day, we need 2 to 24 observations/day. As described in Chapter 3, Dynamic Harmonic Regression (DHR) models are employed for the interpolation, smoothing and extraction of signal components (trend and harmonics) of the time-series of thirteen water quality variables, including five nutrient variables (TOC, NH$_4^+$-N, NO$_2^-$-N, TON, PO$_4^{3-}$-P), seven physico-chemical variables (temperature, PAR, SS, DO, conductivity, ORP, pH) and one biological variable (chlorophyll-$a$). Figure 5.4 displays four time-series (which will be used in the following sections) observed during the time period from July 1st to August 5th, 2000, with missing data points interpolated and outliers smoothed out.

Figure 5.4: Interpolated time-series with trends: (a) PAR below 0.5 surface, (b) Chlorophyll-$a$, (c) temperature, and (d) DO.
Table 5.3: Harmonics of all variables and their peak and trough times (24-hour clock) with dominant diurnal harmonics.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Harmonics</th>
<th>AR Order</th>
<th>Peak</th>
<th>Trough</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ammonia-Nitrogen</td>
<td>[20 10]</td>
<td>3</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Conductivity</td>
<td>[44 24 14]</td>
<td>22</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>DO</td>
<td>[24 12]</td>
<td>14</td>
<td>18</td>
<td>6</td>
</tr>
<tr>
<td>Nitrite-Nitrogen</td>
<td>[24 10]</td>
<td>16</td>
<td>22–0</td>
<td>14–18</td>
</tr>
<tr>
<td>ORP</td>
<td>[24 12]</td>
<td>30</td>
<td>6–8</td>
<td>18–20</td>
</tr>
<tr>
<td>Orthophosphorulate</td>
<td>[48 24 16 12 10]</td>
<td>26</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>pH</td>
<td>[24 12]</td>
<td>50</td>
<td>18–20</td>
<td>6–8</td>
</tr>
<tr>
<td>Solar Radiation</td>
<td>[24 12]</td>
<td>12</td>
<td>14</td>
<td>2</td>
</tr>
<tr>
<td>Suspended Solid</td>
<td>[24 12]</td>
<td>12</td>
<td>12</td>
<td>22–4</td>
</tr>
<tr>
<td>Temperature</td>
<td>[24 12]</td>
<td>26</td>
<td>16</td>
<td>8</td>
</tr>
<tr>
<td>TON</td>
<td>[48 24 13]</td>
<td>19</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>TOC</td>
<td>[16 7]</td>
<td>11</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

* Listed alphabetically.
* The periods of harmonics in hours.

Figure 5.5: Solar radiation serves as a driving force to the pond ecosystem (the vertical black solid line indicates when algal biomass reaches its peak).
The spectral analysis employing an Auto-Regressive (AR) spectrum of data is used to identify the number and values of the fundamental and harmonic frequencies associated with the trend and harmonics. The AR order of each variable is identified from the Akaike Information Criterion (AIC) ([3, 181]) and listed in the third column of Table 5.3. The periods of the identified harmonics contained in each variable’s behavior are listed in the second column of Table 5.3. Not surprisingly, it shows that most of these time-series contain diurnal and semi-diurnal harmonic components, except NH$_4^+$-N and TOC. Further inspection shows that chlorophyll-$a$, DO, ORP, pH, solar radiation, suspended solids and temperature merely contain dominant diurnal and non-dominant semi-diurnal harmonic components, in contrast to conductivity, PO$_4^{3-}$-P and TON, which also contain other harmonics and whose dominant harmonic is not diurnal. It is reasonable that the concentration of algal biomass in the water column (indicated by chlorophyll-$a$ concentration, if at all) and the water temperature are driven by the incoming irradiance. As the products or raw materials of the physiological processes of algae — DO concentration and pH, an indicator of CO$_2$ concentration — are highly correlated with algal biomass concentration, which, in turn, is linked to the oscillation of solar radiation. A nephelometric sensor is adopted in the EPCL to measure turbidity (suspended solids) ([140]), so that the turbidity is a measure of both suspended inorganic and organic matter in the water column ([8, 171]), with the latter mainly contributed by algal cells in the water column. Moreover, the diurnal component contained in the ORP indicates that DO is mainly responsible for the oxidation-reduction potential in the water column. After applying the DHR models to the time-series of PAR, temperature, DO, pH and chlorophyll-$a$, we can extract the signal components of the summation of two harmonics (i.e., diurnal and semi-diurnal) and superimpose them together (shown in Figure 5.5). Figure 5.5 shows that the solar radiation is indeed a driving force to the pond ecosystem. In each cluster of peaks the PAR reaches its peak first, followed in turn by temperature, DO, pH, and chlorophyll-$a$. Note also that the peaks of DO and pH appear at the same time of those of temperature, but several hours before those of chlorophyll-$a$. Apparently this does not mean
that DO and pH respond to the signal of incoming irradiance as quickly as temperature does since the former involves biological process with a time constant of hours to days while the latter engages physical processes with a time constant of minutes to hours. Therefore, it could be speculated that the sources of the concentration variations of DO and pH are not merely algal growth, but also other aquatic plant growth (e.g., macrophytes) and even re-aeration.

From the fact that the TOC and NH$_4^+$-N lack any daily oscillation components, it could be conjectured that during the observed time window the concentrations of TOC and NH$_4^+$-N in the water column were not affected by the physiological activities of primary production plants in the pond to a great extent. Nevertheless, the two harmonics in NH$_4^+$-N are close to being diurnal and semi-diurnal (shown in the second row of Table 5.3). The processes causing the presence of harmonics with more than one day (or less than half day) periods in conductivity, PO$_4^{3-}$-P and TON are not self-evident. Further investigation beyond pure signal analysis is suggested. The processes with longer time constants may relate to the lower frequency harmonics and ones with shorter time constants may be responsible for the higher frequency harmonics in these time-series.

5.4 DATA-BASED-MECHANISTIC MODELING

The Data-Based-Mechanistic modeling approach has been applied to explore the dynamics of the algal blooms occurring during the experiment conducted on the aquaculture pond.

5.4.1 LINEAR TF MODEL WITH CONSTANT PARAMETERS

Equation (5.1) is a linear Transfer Function (TF) relationship between the light limitation (input $u(k)$) and the algal biomass (output $y(k)$).

$$y(k) = \frac{B(z^{-1})}{A(z^{-1})}u(k - \delta) + \nu(k); \quad k = 1, 2, \cdots, N$$ (5.1)
where, 

\[ A(z^{-1}) = 1 + a_1 z^{-1} + \ldots + a_n z^{-n}; \quad B(z^{-1}) = b_0 + b_1 z^{-1} + \ldots + b_n z^{-m}. \]

where \( N \) is the total number of observations in the time-series; \( \delta \) represents the potential presence of any pure time delay in the system (in sampling intervals); \( \nu(k) \) represents uncertainty in the relationship arising from a combination of measurement noise, the effects of other unmeasured disturbances and modeling error; \( y(k) \) is the algal biomass concentration in water column, which is assumed to be 50 times chlorophyll-a measurements (see e.g., [273]), mg·L\(^{-1}\); and \( u(k) \) accounts for the light limitation of algal growth, to be discussed and defined below.

The relationship between algal growth (photosynthesis) and solar radiation (irradiance) (i.e., P-I curves) can be generally summarized as follows: photosynthesis increases linearly with light at low light intensities, becomes constant at higher intensities (photosaturation) and in some cases begins to decline at even higher intensities (photoinhibition). Numerous expressions relating photosynthesis to light are available, reviewed for example by [122], in which many of the equations include the extinction of light with increasing depth and use a Michaelis-Menten expression to describe the relationship between photosynthesis and irradiance. However, none of these relationships performs especially well, although linear models are found to be inadequate when applied to a macrophyte dominated stream in Cosby et al. ([65]). But in current research, the expression of Steele ([216]) is adopted. In addition, the small fishery experiment pond is considered as a CSTR (Completely Stirred Tank Reactor) and the PAR data are recorded along with other water quality data at 0.5 m below surface, so that the attenuation coefficient is then dropped from the original Steele expression. Therefore, the input time-series, \( u(k) \), is represented by \( f(I(k)) \) defined in Equation (5.2),

\[ f(I(k)) = \frac{I(k)}{I_S} e^{\left(1 - \frac{I(k)}{I_S}\right)^{1/3}} \quad (5.2) \]

in which, \( I(k) \) is the PAR measured at 0.5 m below surface, KJ·m\(^{-2}\)·h\(^{-1}\); \( I_S \) is the saturation solar radiation, KJ·m\(^{-2}\)·h\(^{-1}\). The solar radiation beyond this value will cause photoinhibition.
of algal growth. In this research, $I_S$ takes the optimal value of 70 KJ·m$^{-2}$·h$^{-1}$, by trial and error. From now on, $u(k)$ is called “Light Efficiency” because it symbolizes how efficiently the incoming irradiance can be utilized by aquatic plants (e.g., algae). Thus, the $y(k)$ and $u(k)$ time series are shown in Figure 5.6 (a) and (b), respectively.

The SRIV algorithm described in the Appendix A.1 was employed to identify the order of the deterministic part of the linear TF model (i.e., Equation (5.1)). Table 5.4 lists the best 5 linear TF models with constant parameters for describing the relationship between “Light Efficiency” and algal growth.

Table 5.4: The best 5 identified linear TF models with constant parameters (The sampling interval is 2 hours).

<table>
<thead>
<tr>
<th>No.</th>
<th>Denominator</th>
<th>Numerator</th>
<th>Delay</th>
<th>YIC</th>
<th>$R_I^2$</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>-5.18</td>
<td>0.0677</td>
<td>2.24</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>-3.68</td>
<td>0.118</td>
<td>2.18</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>2</td>
<td>17</td>
<td>-3.55</td>
<td>0.682</td>
<td>1.18</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>-0.856</td>
<td>0.091</td>
<td>2.22</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>-0.690</td>
<td>0.136</td>
<td>2.17</td>
</tr>
</tbody>
</table>

Not surprisingly, all of the linear TF models shown in Table 5.4 fail to explain the relationship of “Light Efficiency” and the growth of algae, since there exists a nonlinearity between these two variables, which is readily apparent from visual inspection of Figure 5.6 (a) and (b).

In order to proceed, we estimate the parameters of the model [1, 1, 2], which has the lowest value of the YIC. Recursive estimation results is the following,

$$\hat{a}_1 = -0.9989(0.000), \quad \hat{b}_0 = 0.2475(0.4074),$$

with standard errors included in parentheses. The standard errors associated with the parameter estimates imply little variation in the lag parameter $\hat{a}_1$ but considerable changes in the gain parameter $\hat{b}_0$. Therefore, in the following section the time-varying gain parameter within
the first-order TF model with linear structure \([1, 1, 2]\) is estimated by the FIS algorithm (see Appendix A.2 for illustration of the FIS algorithm).

5.4.2 Linear TF model with TVP

The SISO model with TVP takes the form:

\[
y(k) = \frac{B(k, z^{-1})}{A(k, z^{-1})} u(k - \delta) + \nu(k); \quad k = 1, 2, \ldots, N
\]  \hspace{1cm} (5.3)

where,

\[
A(k, z^{-1}) = 1 + a_1(k)z^{-1} + \ldots + a_n(k)z^{-n};
\]

\[
B(k, z^{-1}) = b_0(k) + b_1(k)z^{-1} + \ldots + b_n(k)z^{-n}.
\]
Thus, for the simplest linear structure \([1, 1, 2]\) identified previously, we have

\[
y(k) = \frac{b_0(k)}{1 + a_1(k)} u(k - 2) + \nu(k)
\]  

(5.4)

Bearing in mind the results of the previous section, it makes sense to constrain the lag parameter \(a_1(k)\) to be constant but allow the gain parameter \(b_0(k)\) to vary over the observation interval. An Integrated Random Walk (IRW, see discussion in Section 3.3.2) model was assigned to model the variation of parameter \(b_0(k)\). The values of the corresponding NVR’s of \(a_1\) and \(b_0(k)\) optimized by an ML algorithm are:

\[
NVR(a_1) = 0, \quad NVR(b_0(k)) = 0.7445 \times 10^{-5}.
\]

Thus the resulting linear structure of Equation (5.4) with *time variable* input parameter estimated by the FIS algorithm is shown by Equation (5.5).

\[
y(k) = \frac{\hat{b}_0(k|N)}{1 - 0.9650} u(k - 2) + \nu(k)
\]  

(5.5)

where the estimated *time-varying* parameter \(\hat{b}_0(k|N)\) and its standard error bounds are shown in Figure 5.7. Figure 5.9 shows that the linear TF model \([1, 1, 2]\) with *time variable* gain parameter performs exceptionally well. However, in accordance with the Data-Based-Mechanistic modeling philosophy, in addition to enhancing the model fit to data, it is essential that the identified states and parameters have a clear physical meaning in relation to the system under consideration. Hence, the following section will employ a State Dependent Parameter (SDP, see Section 3.2.3) modeling approach to provide the mechanistic interpretation of the nonlinearity exhibited in this system.
Figure 5.7: The variation of $\hat{b}_0(k|N)$ and its standard errors superimposed on the scaled $y(k)$.

Figure 5.8: Linear relationship between $\hat{b}_0(k|N)$ and $y(k)$.

Figure 5.9: Simulation and one-step ahead prediction of TF $[1, 1, 2]$ with TVP (in Figure 5.7 against data).
5.4.3 SDP Modeling

Inspection of Figure 5.7 shows that there exists a visible correlation between the estimated TVP’s variation ($\hat{b}_0(k|N)$) and algal biomass ($y(k)$) evolution. The scatter plot of $\hat{b}_0(k|N)$ against $y(k)$ (Figure 5.8) also suggests a linear relationship between these two quantities. Subsequently, the non-parametric State-Dependent Parameter (SDP) analysis yields the results shown in Figures 5.10 and 5.11. These two figures correspond respectively to Figures 5.7 and 5.8 (before the SDP analysis was applied). Note also that the rapid variation of estimated SDP $\hat{b}_0(k|N)$ is captured more easily and accurately when the time-series are sorted in ‘ascending order’ manner. Figure 5.11 is a plot of the estimated SDP against the algal biomass $y^*(k)$ (where the star superscript indicates that the observations of algal biomass have been sorted in ascending magnitude) and the FIS non-parametric estimate $\hat{b}_0(k|N)$ is shown as blue dots, with the standard errors shown dashed. The full red straight line is the Weighted Least Squares (WLS) estimate of the linear relationship. It might be better, in this case, to consider a two-segment linear relationship or a Michaelis-Menten relationship, rather than a single linear relationship, since Figure 5.11 suggests that the FIS estimate of the SDP approaches an asymptotic limit (approximately 0.7) after the algal biomass concentration is greater than 10 mg·L$^{-1}$. However, at first, we will select the single linear relationship due to its advantage of involving the least number of parameters. Nevertheless, it must be emphasized that further research is required and it is clear that the linear relationship may not always be optimal if the algal biomass concentration ranges from a very low value to a very high value (e.g., 10 mg·L$^{-1}$ here), which is not uncommon during algal blooms.

The DBM model accounting for the nonlinear relationship between $f(I(k))$ and $y(k)$ can now be given as Equation (5.6).

\begin{align}
  y(k) &= \frac{B(z^{-1})}{A(z^{-1})} u(k - \delta) + \nu(k) \\
  u(k) &= \mu_D y(k) f(I(k))
\end{align}  

(5.6a)  

(5.6b)
where $\mu_D$ is the slope of the red straight line in Figure 5.11; other notations are defined in the previous sections. The value of $\mu_D$ optimized by WLS is 0.1009. Application of the SRIV algorithm yields the best 5 linear structures for the model of Equation (5.6a) listed in Table 5.5. The first 3 model structures ([2, 1, 0], [2, 1, 1] and [2, 1, 2]) are comparable in terms of the three order identification criteria — YIC, $R_T^2$ and AIC. Note that all three models are characterized by the same second-order linear structures except with different pure time delays. This arises probably because the actual pure time delay in the system does not stay constant at one sampling interval (2 hours), but instead varies with time bounding between 0 and 2 sampling intervals, i.e., between 0 and 4 hours. Since [2, 1, 1] gives the best $R_T^2$ and AIC and second best YIC, we choose it as the best linear TF model structure.

The estimated parameters and associated standard errors of the favored TF [2, 1, 1] are as follows:

$$\hat{a}_1 = -1.7236(0.0204); \quad \hat{a}_2 = 0.7588(0.0190);$$

$$\hat{b}_0 = 1.3567(0.0967);$$
Table 5.5: The best 5 identified linear structures of DBM model (5.7) (The sampling interval is 2 hours).

<table>
<thead>
<tr>
<th>No.</th>
<th>Denominator</th>
<th>Numerator</th>
<th>Delay</th>
<th>YIC</th>
<th>$R^2_T$</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>-8.325</td>
<td>0.876</td>
<td>0.229</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>-7.570</td>
<td>0.912</td>
<td>-0.114</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>-7.271</td>
<td>0.889</td>
<td>0.116</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>-7.054</td>
<td>0.839</td>
<td>0.491</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>-6.406</td>
<td>0.846</td>
<td>0.455</td>
</tr>
</tbody>
</table>

so that as an explicit realization of Equation (5.6), we have

$$y(k) = \frac{1.3567}{1 - 1.7236z^{-1} + 0.7588z^{-2}}u(k - 1) + \nu(k)$$  \hspace{1cm} (5.7a)

$$u(k) = 0.1009y(k)f(I(k))$$  \hspace{1cm} (5.7b)

The nonlinear part (5.7b) of the DBM model (5.7) obtained by shifting the state variable (i.e., $y(k)$) from the SDP (i.e., $\hat{b}_0(k|N)$ in Equation (5.5)) into the input variable implies that the “Light Efficiency” — how efficiently the sunlight is utilized by algae — is a function of the algal biomass concentration (or cells) in the aquatic ecosystem. Let us therefore return to the discussion, at the beginning of this section, of the relationship between the SDP and the algal biomass concentration (see also Figure 5.11). When the algal biomass concentration in a water body is below a certain level (e.g., $10 \text{ mg·L}^{-1}$ in this case), the more active algal cells are there, the more efficiently the incoming irradiance is exploited by this aquatic ecosystem. Hence this relationship between the “Light Efficiency” and the algal biomass is approximately linear. However, when the algal biomass concentration in the water is beyond that threshold (i.e., $10 \text{ mg·L}^{-1}$), either the depletion of the nutrients or self-shading of algae will limit the rate of photosynthesis of the living algal cells. As implied in Figure 5.11, the efficiency of sunlight utilization by algae approaches a limit asymptotically when the algal biomass is higher than $10 \text{ mg·L}^{-1}$ (see Section 5.4.5).
Figure 5.12 presents the performance of the model (5.7). The model so defined functions very well indeed in a forecasting context, whereas the simulation persistently underestimates the magnitude of the algal biomass concentration in the pond, although it mimics the patterns of the evolution and oscillation of the algal biomass. This is not surprising since this simple model only takes account of the impact of the solar radiation on the pond ecosystem and ignores the other possible environmental factors (e.g., nutrients, temperature). For the purpose of simulation, the nonlinear part of model (5.7) can be defined as Equation (5.7b’), where only the solar radiation measurements are used to generate purely simulated algal biomass ($\hat{y}(k)$).

\[ u(k) = 0.1009\hat{y}(k)f(I(k)) \]  

(5.7b’)

Figure 5.12: Forecasting and simulation of model (5.7) against data.
Figure 5.13: Cross-correlation function (CCF) between the deterministic residual $\nu(k)$ of model (5.7) and temperature.

**Temperature impact**

Figure 5.13 presents the cross-correlation analysis between the deterministic residuals of model (5.7) and the water temperature. There is evidence of significant correlation at some lags between these two quantities by inspection of this figure. Therefore it is necessary to incorporate the temperature factor into the DBM model in order to pursue a further, more precise description of the dynamics of algal blooms through one of the following two ways: (1) considering temperature as another input variable, thence identifying the relevant transfer function (e.g., [262]); (2) assuming the ambient temperature affects the efficiency of algal utilization of sunlight through its impact on physiological processes. In line with the Data-Based-Mechanistic modeling viewpoint, the latter is implemented by including temperature into the nonlinear part of the model (see [263] for another example). The resulting modified
The DBM model is represented in Equation (5.8) and the slight improvement on forecasting performance is shown in Figure 5.14:

\[ y(k) = \frac{b_0}{1 + a_1 z^{-1} + a_2 z^{-2}} u(k - 1) + \nu(k) \]  
\[ (5.8a) \]

\[ u(k) = \mu_D \theta_D^{T(k)-20} y(k)f(I(k)) \]  
\[ (5.8b) \]

where, \( T(k) \) is the temperature, \(^\circ\)C; and
\( \hat{a}_1 = -1.7200(0.0204), \quad \hat{a}_2 = 0.7551(0.0190); \quad \hat{b}_0 = 1.5805(0.1131); \)
\( \hat{\mu}_D = 0.0555, \quad \hat{\theta}_D = 1.0166. \)

Figure 5.14: Forecasting and simulation of model (5.8) against data.

5.4.4 MECHANISTIC INTERPRETATION OF THE SECOND-ORDER DBM MODEL

Since the imaginary parts of the roots of the denominator of the TF part of the model (5.8), 0.86 ± 0.1245\( i \), are relatively small, the deterministic part of the model can be approximated by two real poles, both at 0.86. Therefore, if only considering the deterministic part of this model, the deterministic output \( \hat{y}(k) \) can be decomposed into two serial first-order TF’s, i.e.
(see [259], [266], etc. for discussion of the decomposition of DBM models):

\[
\hat{y}(k) = \frac{1.2572}{1 - 0.86z^{-1}} \times \frac{1.2572}{1 - 0.86z^{-1}} u(t - 1) \tag{5.9}
\]

Therefore it is sensible to speculate that the process of sunlight utilization by algae is composed of two serial stages, schematically expressed as the system block diagram of Figure 5.15.

![System block diagram of the photosynthesis process of algae as inferred from the deterministic part of model (5.8).](image)

Figure 5.15: System block diagram of the photosynthesis process of algae as inferred from the deterministic part of model (5.8).

One possible physical interpretation of the serial partition of the deterministic part of the DBM model (5.8) is the two-stage nature of photosynthesis, comprising a light reaction and a Calvin cycle (“dark reaction”) ([51]). The light reactions are the steps of photosynthesis that convert solar energy to chemical energy (NADPH and ATP). In this stage, water is split and oxygen is released as a by-product. Note that the light reactions produce no sugar; that happens in the second stage of photosynthesis. The Calvin cycle then reduces the fixed carbon to carbohydrate by the addition of electrons powered by the chemical energy (NADPH and ATP) generated from the first stage. The metabolic steps of the Calvin cycle are sometimes referred to as the “dark reactions” because no light is required directly. Figure 5.16, which is redrawn from Figure 10.4 in [51], illustrates the two stages of the photosynthesis schematically.

The two first-order transfer functions in Equation (5.9) have identical forms. Therefore they are characterized by the same time constants and steady-state gains, \( \tau = 13.26 \text{ hours} \) (6.6 sampling intervals) and \( \kappa = 8.90 \), respectively. However, doubts may reasonably be provoked on the discussion of the time constant \( \tau \) for the two separate stages of the photosynthesis.
It is almost impossible that the light reactions and Calvin cycle, which involve completely different processes (or reactions), take place with identical kinetic rates and thus the same time constants. This apparent misrepresentation may result from the failure of the algorithms of the Data-Based-Mechanistic modeling to further distinguish the two steps explicitly given the observed “macro-photosynthesis” on the pond. This “macro-photosynthesis” is composed of millions of micro-photosynthesis of cells with chloroplast, which do not necessarily occur at the same speed. Therefore, this second-order is not the ideal data-based representation of algal photosynthesis in terms of being synthesized quantitatively with theory-based model, although it has the appealing qualitative mechanistic interpretation summarized in Figure 5.16.

5.4.5 The first-order DBM model

As indicated by Figure 5.11, a two-segment linear regression should be superior to a single-segment linear regression in depicting the relationship between the estimated time-varying input parameter ($\hat{b}_0(k\mid N)$) and the sorted output observations ($y^*(k)$) (see the discussion in
Section 5.4.3). That is, the rate of change of the input parameter \( \hat{b}_0(k|N) \) with respect to the concentration of algal biomass in water column becomes slower after the algal biomass concentration reaches 10 mg·L\(^{-1}\) than before it does. Moreover, the input parameter essentially measures the “Light Efficiency” — how efficiently the sunlight is utilized by algae in an aquatic ecosystem. Therefore, it is reasonable to conjecture that this phenomenon is caused by: (1) the deficit of the resources (e.g., nutrients, light, etc.) for algal growth in the pond; (2) lower proportion of active algal cells among the algal pool as the algae aged. Therefore, we divided the time-series in terms of \textit{when} the algal biomass concentration reaches 10 mg·L\(^{-1}\) at the first time instead of \textit{whether} the algal biomass concentration reaches 10 mg·L\(^{-1}\). It is readily apparent that the algal biomass concentration reached 10 mg·L\(^{-1}\) at the 82nd sampling instant. Figure 5.17 shows the two-segment relationship between \( \hat{b}_0(k|N) \) and \( y(k) \).

![Figure 5.17: Two segment relationship between \( \hat{b}_0(k|N) \) and \( y(k) \): (a) before \( y(k) \) reaches 10.0 mg·L\(^{-1}\) at the first time; (b) after \( y(k) \) reaches 10.0 mg·L\(^{-1}\) at the first time.](image)

The corresponding nonlinear DBM model can now be given as Equation (5.10), in which the relationship between the SDP (\( \hat{b}_0(k|N) \)) and output observations (\( y(k) \)) is described as
a two-segment linear relationship:

\begin{align}
    y(k) &= \frac{b_0}{1 + a_1 z^{-1}} u(k - 2) + \nu(k) \quad (5.10a) \\
    u(k) &= \mu_{D1} T^{(k) - 20} y(k) f(I(k)), \quad \text{when } k < 82, \quad (5.10b) \\
    u(k) &= \mu_{D2} T^{(k) - 20} y(k) f(I(k)), \quad \text{when } k \geq 82. \quad (5.10c)
\end{align}

where $\mu_{D1}$ and $\mu_{D2}$ are the slopes of the blue straight lines in Figure 5.17 (a) and (b), respectively; other notations are defined in the previous sections. The values of $\mu_{D1}$ and $\mu_{D2}$ optimized by WLS are 0.1167 and 0.0806 respectively. Thus, the subsequent application of the identification procedure gives the best 5 linear model structures (Table 5.6). According to arguments similar to those in Section 5.4.3, we choose $[1, 1, 2]$ as the best linear structure for model (5.10). The estimated parameters and associated standard errors of the favored TF $[1, 1, 2]$ are as follows:

\begin{align*}
    \hat{a}_1 &= -0.9251(0.0044); \quad \hat{b}_0 = 1.9028(0.1081);
\end{align*}

so that as an explicit realization of Equation (5.10), we have

\begin{align}
    y(k) &= \frac{1.9028}{1 - 0.9252 z^{-1}} u(k - 2) + \nu(k) \quad (5.11a) \\
    u(k) &= 0.1167 \times 1.0115^{(k) - 20} y(k) f(I(k)), \quad \text{when } k < 82, \quad (5.11b) \\
    u(k) &= 0.0806 \times 1.0115^{(k) - 20} y(k) f(I(k)), \quad \text{when } k \geq 82. \quad (5.11c)
\end{align}

Figure 5.18 shows that the simulation performance of model (5.11) has been dramatically improved albeit the model’s one-step-prediction efficacy stays more or less unchanged.
Table 5.6: The best 5 identified linear structures of DBM model (5.10) (The sampling interval is 2 hours).

<table>
<thead>
<tr>
<th>No.</th>
<th>Denominator</th>
<th>Numerator</th>
<th>Delay</th>
<th>YIC</th>
<th>$R^2_T$</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>-8.057</td>
<td>0.902</td>
<td>-0.0174</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>-8.052</td>
<td>0.902</td>
<td>-0.0128</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-7.919</td>
<td>0.896</td>
<td>0.0467</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>-7.719</td>
<td>0.886</td>
<td>0.136</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>-7.701</td>
<td>0.892</td>
<td>0.0862</td>
</tr>
</tbody>
</table>

Figure 5.18: Forecasting and simulation of model (5.11) against data.
5.5 A simple mechanistic model

In order to achieve model identification and parameter estimation with a reasonable chance of success, the model must be simple, with a few parameters. Three main hypotheses are made. Firstly, it is assumed that pond water is completely mixed. This assumption was verified by the fact that the temperatures taken at 0.5m, 1m and 2m below water surface are almost the same from late June to the end of the record (see Figure 5.19). During the pond experiment the sampling pumps of the EPCL may act as a stirring force for circulating the pond water. Secondly, nutrients are supposed not to be rate limiting. During the considered time period (from July 1st to August 5th, 2000), two fertilizations have been applied to the pond ([171], or see description in Section 5.2). As shown in Figure 5.20, the summation of the concentrations of ammonia-nitrogen and nitrate-nitrogen is well above the half-saturation constant of 15 µg·L$^{-1}$ ([230]) and the concentrations of PO$_4^{3-}$-P also greatly exceed the half-saturation constant of 2.5 µg·L$^{-1}$ ([230]), except for a few data points near the end of this period. We therefore only consider light energy and temperature as growth-controlling factors. Thirdly, the pond volume is assumed to be constant throughout the period during which the model is applied. This assumption is moderately reasonable. During the considered time period, only a one-inch rainfall occurred in the afternoon of July 23rd, 2000 and a 1.5 inches rainfall in the afternoon of July 30th, 2000. The water level of the Whitehall pond almost never reached the outlet during the entire experiment and the loss of water by evaporation could be compensated through the small influent flow originating from springs along the sides of the channel.
Figure 5.19: Temperature measurements at three levels of depth.

Figure 5.20: Concentrations of inorganic nitrogen and orthophosphate in Whitehall pond (July 1st – August 5th, 2000).
The kinetic interactions of algae consist of growth and loss. Algal growth is accomplished by photosynthesis, and the losses occur via respiration, grazing, nonpredatory mortality, settling and flushing, etc. However, the hydrological flushing loss of algae can be ignored in the model since the water level of the pond almost never reached the outlet (see above discussion). While respiration and nonpredatory mortality of phytoplankton are often described as first-order reactions (e.g., [88, 273]), zooplankton grazing and settling of phytoplankton are much more complicated processes. They are functions of the physiological state of the algae, physical, chemical and biological properties of the surrounding water body and other state variables (e.g., zooplankton, protozoa, etc.). In the following simple mechanistic model (see Table 5.7), a first-order kinetic process is used to describe the collective effects of the respiration and mortality of algae and a time-varying parameter is employed to track the composite variation of the other sinks of algal biomass, including zooplankton grazing and settling to sediments.

However, the DO evolution, in a pond, depends on the balance between photosynthetic production, total consumption, and exchanges with atmosphere. The gain and loss due to respective algal photosynthesis and respiration are calculated based on the Redfield stoichiometric coefficients ([188]). The DO consumed by the degradation of other organic matter in the water column is simply represented by that participating in the decomposition of the biodegradable organic carbon (measured as a portion of Total Organic Carbon (TOC)). All the other sources of the dissolved oxygen (e.g., photosynthesis and respiration of macrophytes, etc.) are integrated into a time-varying parameter. Table 5.7 gives details of the state variables, parameters and the environmental functions associated with the mechanistic model elucidating the dynamics of algae and dissolved oxygen in the Whitehall pond.
Table 5.7: Model formulations, variables and parameters.

Formulations:
\[
\frac{dx_1(t)}{dt} = k_1 f(I; t) f_1(T; t) x_1(t) - k_2 f_2(T; t) x_1(t) - k_3(t) \quad (1)
\]
\[
\frac{dx_2(t)}{dt} = c k_1 f(I; t) f_1(T; t) x_1(t) - c k_2 f_2(T; t) x_1(t) + k_6 f_3(T; t) [C_S(T; t) - x_2(t)]
- k_4 f_4(T; t) C_{OC}(t) + k_5(t) \quad (2)
\]
\[
f(I; t) = \frac{I(t)}{I_S} \exp\left[1 - \frac{I(t)}{I_S}\right] \quad (3)
\]
\[
f_i(T; t) = \theta_T(t)^{20}; \quad i = 1, 2, 3, 4 \quad (4)
\]
\[
C_S(T; t) = 14.589 - 0.40 T(t) + 0.008 T(t)^2 - 0.000661 T(t)^3 \quad (5)
\]

State variables and their initial values:
\[
\begin{align*}
  x_1(t) & \quad \text{Algal biomass concentration} & 4.41 \text{ mg}\cdot\text{L}^{-1} \\
  x_2(t) & \quad \text{Dissolved oxygen concentration} & 7.04 \text{ mg}\cdot\text{L}^{-1}
\end{align*}
\]

Environmental (forcing) variables:
\[
\begin{align*}
  I(t) & \quad \text{Solar radiation (light intensity)} & \text{KJ}\cdot\text{m}^{-2}\cdot\text{h}^{-1} \\
  T(t) & \quad \text{Water temperature} & \circ\text{C} \\
  C_{OC}(t) & \quad \text{Portion of total organic carbon concentration} & \text{mg}\cdot\text{L}^{-1}
\end{align*}
\]

Fixed parameters:
\[
\begin{align*}
  c & \quad \text{Redfield stoichiometric coefficient} & 3.47 \\
  I_S & \quad \text{Saturation solar radiation} & 80.0 \text{ KJ}\cdot\text{m}^{-2}\cdot\text{h}^{-1} \\
  k_6 & \quad \text{Global transfer coefficient between air and water at } 20\circ\text{C} & 0.4 \text{ d}^{-1} \\
  \theta_{1,2,3,4} & \quad \text{Coefficients for Arrhenius equations} & 1.025, 1.02, 1.02, 1.02
\end{align*}
\]

Time-invariant parameters and initial values:
\[
\begin{align*}
  k_1 & \quad \text{Specific growth rate of algae at } 20\circ\text{C} & 2.4 \text{ d}^{-1} \\
  k_2 & \quad \text{Kinetic rate of loss of algae due to respiration and nonpredatory death} & 0.48 \text{ d}^{-1} \\
  k_4 & \quad \text{Total organic matter decay rate} & 0.6 \text{ d}^{-1}
\end{align*}
\]

Time-variant parameters and initial values:
\[
\begin{align*}
  k_3(t) & \quad \text{Loss of algae due to grazing and settling processes, et.} & 0 \text{ mg}\cdot\text{L}^{-1}\cdot\text{d}^{-1} \\
  k_5(t) & \quad \text{Sources of DO other than those listed explicitly} & 0 \text{ mg}\cdot\text{L}^{-1}\cdot\text{d}^{-1}
\end{align*}
\]
5.6 Recursive parameter estimation and structural change

5.6.1 Innovations format

In order to apply the RPE algorithm to the model for the dynamics of algae and DO in the Whitehall pond, the model with formulations (1)–(5) in Table 5.7 are rewritten in the innovations representation (5.12a) with a linear observation equation (5.12b) since both algal biomass (through chlorophyll-$a$) and DO concentrations are observed directly.

\[
\dot{x}(t) = f(x, u, \alpha; t) + K\nu(t) \tag{5.12a}
\]
\[
y(t_k) = H(\alpha; t_k)x(t_k) + \nu(t_k) \tag{5.12b}
\]

where,

\[
\dot{x}(t) = \begin{bmatrix} \frac{dx_1(t)}{dt} \\ \frac{dx_2(t)}{dt} \end{bmatrix}^T; \\
x(t) = [x_1(t), x_2(t)]^T; \\
y(t_k) = [y_1(t_k), y_2(t)]^T; \\
u(t) = [I(t)]; \\
\alpha(t) = [k_1, k_2, k_3(t), k_4, k_5(t)]^T; \\
K = \begin{pmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{pmatrix}; \\
H(\alpha; t_k) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}; \\
\nu(t_k) = [\nu_1(t_k), \nu_2(t_k)]^T.
\]

$u(t)$ are the inputs to the pond system; $y_1(t_k), y_2(t_k)$ are the observations of algal biomass and DO concentrations (mg·L$^{-1}$), respectively; $\nu_1(t_k), \nu_2(t_k)$ are the measurement noises and disturbances reflected in the observations of algal biomass and DO concentrations; $K$ is the
Kalman-like gain matrix, which distributes the impacts of mismatches among the constituent representations of the state variables; $H(\alpha; t_k)$ is the observations matrix; the two entities of vector $f(x, u, \alpha; t)$ are Equations (1) and (2) respectively in Table 5.7; and all the other notations are the same as defined in Table 5.7.

5.6.2 **Time-variant parameter estimation**

As shown in Table 5.7, parameters $k_1$, $k_2$, $k_4$ are assumed to be time-invariant over the period studied, while the values of $k_3(t)$ and $k_5(t)$ are allowed to change with time. The variations of these *time-varying* parameters are modeled by Random Walk (RW) stochastic processes since no *a priori* knowledge is available. Whether the evolutions of the recursive estimates of the *time-invariant* parameters change with time reveals the adequacy or inadequacy of the associated individual constituent hypotheses. The trajectories of the *time-varying* parameters also provide clues for constructing new hypotheses. The four elements in $K$ are also considered as *time-invariant* parameters as well. Their values are set to be zeros initially implying that the mechanistic model describes the dynamics of algae and DO perfectly. Any excursions of the recursive estimation of $K$’s elements from zero suggest the *apparent* structural change ([27]).

Table 5.8 specifies the corresponding initial values of states and parameters along with the leading diagonal elements for the variance-covariance matrices $P(t_0)$, $Q$ and $\Lambda$, defined in Chapter 3. For the *fixed* parameters such as $c$, $\theta$’s, they take constant initial values specified in Table 5.7 and remain unchanged over the entire period.

Having specified the initial conditions for all states and parameters, initial covariance matrices for all parameters, and the variance matrix for measurement errors, the modified RPE algorithm (illustrated in Section 3.3.2) has been applied to the simple mechanistic model of the dynamics of algal biomass and DO in the pond. However, it is worth mentioning that the recursive estimation results of the state variables, the parameters and the elements of the Kalman-like gain matrix are not smoothed by the backward pass smoothing
Table 5.8: Specifications of initial values and the leading diagonal elements for the variance-covariance matrices $P(t_0)$, $Q$ and $\Lambda$.

<table>
<thead>
<tr>
<th>State/Parameter</th>
<th>Initial Values</th>
<th>$P(t_0)$</th>
<th>$Q$</th>
<th>$\Lambda$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1(t)$ (mg·L$^{-1}$)</td>
<td>4.41</td>
<td>—</td>
<td>—</td>
<td>0.0456</td>
</tr>
<tr>
<td>$x_2(t)$ (mg·L$^{-1}$)</td>
<td>7.04</td>
<td>—</td>
<td>—</td>
<td>0.4287</td>
</tr>
<tr>
<td>$k_1$ (d$^{-1}$)</td>
<td>2.4</td>
<td>0.0012</td>
<td>0</td>
<td>—</td>
</tr>
<tr>
<td>$k_2$ (d$^{-1}$)</td>
<td>0.48</td>
<td>0.0012</td>
<td>0</td>
<td>—</td>
</tr>
<tr>
<td>$k_3(t)$ (mg·L$^{-1}$·d$^{-1}$)</td>
<td>0.0</td>
<td>0.012</td>
<td>0.00001</td>
<td>—</td>
</tr>
<tr>
<td>$k_4$ (d$^{-1}$)</td>
<td>0.60</td>
<td>0.0012</td>
<td>0</td>
<td>—</td>
</tr>
<tr>
<td>$k_5(t)$ (mg·L$^{-1}$·d$^{-1}$)</td>
<td>0.0</td>
<td>0.012</td>
<td>0.00001</td>
<td>—</td>
</tr>
<tr>
<td>$K$’s</td>
<td>0.0</td>
<td>0.012</td>
<td>0.0</td>
<td>—</td>
</tr>
</tbody>
</table>

procedures (e.g., FIS algorithm; see the discussion in Chapter 3, [89] and [260] for detailed information), therefore caution must be exercised when interpreting the transient recursive estimates during the initial stages.

Figure 5.21 displays the one-step-ahead predictions of the model against the observations of algal biomass and DO concentration. The model tracks the behavior of algae remarkably well (Figure 5.21a) — the peak and the diel characteristics of the algal growth are caught by the model, although the model results are those of the one-step-ahead predictions. However, the discrepancy between the model and the observations of DO is not insignificant (Figure 5.21b). The model captures the diurnal oscillation pattern quite precisely but misses the alternation of amplitude. A few matches or mismatches of DO dynamics seize our attention: (1) the model performs very well before the liming and fertilization operation occurred; (2) the model’s prediction overestimates the amplitude of the oscillation of the observations during the bloom, while persistently underestimates after the bloom; (3) after the bloom the model’s prediction shows a centric tendency and fails to follow the trend of the variation of
the time-series; (4) the highest DO concentration of the model’s prediction lags that of the observations by one cycle (24 hours).

Figure 5.21: Recursive estimation results: one-step-ahead prediction vs. observations of (a) algal biomass, and (b) dissolved oxygen.

As indicated in the previous sections, these mismatches between the behavior of the model and system could be rooted in the failure of the fixed model formation to track the structural change of the real system. We may search for the possible explanations for the above observations via inspecting the successful recursive estimation of the parameters (both time-invariant and time-variant, shown in Figure 5.22), through which the model structure is formed. Obviously, the presumed time-invariant parameters \( k_1 \) and \( k_2 \) stay stable for the rest of the period after the transient stages, while the decay rate of the total organic matter does not. This is not surprising because this first-order term is an extreme simplification for the consumption of the dissolved oxygen resulting from the degradation of all organic matter in the water column (it may also include the Sediment Oxygen Demand (SOD)). Therefore, a more detailed description of the decomposition of organic matter is indispensable, for example, by classifying the organic matters and the associated decay processes. The increase
in $k_5(t)$ — sources of DO other than re-aeration and algal photosynthesis — suggests that there exist other processes contributing (positively) to DO concentration in the water column. During the operation of the pond experiment duckweed *Lemma*, a species of free-floating macrophytes ([110]), were growing very fast after the first fertilization. Their presence started from rim of the pond and eventually almost covered the entire surface. That the significant production and consumption of DO by the photosynthesis and respiration of the duckweed are not explicitly included in the simple model may cause the model’s predictions always to underestimate the amplitude of the diurnal oscillation.

Figure 5.22: Recursive estimation results: estimates (solid line) and standard errors (dashed line, almost invisible) of the time-invariant parameters ($k_1, k_2, k_4$) and the time-varying parameters ($k_3(t), k_5(t)$).
Figure 5.23: Recursive estimation results: estimates (solid line) and standard errors (dash line) of the elements of $K$ matrix.

5.7 Towards the synthesis

Both data-based and theory-based modeling approaches have been applied to understand the dynamics of the algal blooms in the Whitehall experiment pond. Sections 5.3 and 5.4 are dedicated to the former and Sections 5.5 and 5.6 to the latter, respectively. There are no doubts that the employment of both modeling approaches for addressing an environmental problem acquires more information about the system than that of a single approach. At least, in this case study, two otherwise remaining unknown facts about the pond system were unravelled through the foregoing two-pronged modeling approach:

Fact one that the sunlight utilization by algae becomes less efficient as the algal biomass concentration reaches a threshold value, is discovered by the data-based modeling prong. That is, the SDP modeling technique characterizing the data-based modeling approaches
adopted in this research found that when the concentration of algal biomass in the water column is higher than a threshold (e.g., 10 mg·L$^{-1}$) the utilization efficiency of the solar radiation by the algae is bounded by an asymptotic limit (see discussion about Figure 5.11). The causes behind this fact could be various. For example, it could be caused by either or all of the following possible reasons: deficit of nutrients, self-shading or less active cells as algae age, etc.

*Fact two*, discovered through the theory-based modeling prong, is that a free-floating macrophyte, duckweed *Lemma*, plays a critical role in the pond dynamics (DO and pH dynamics, in particular) when algal growth in the pond becomes insignificant.

In a compartmental model aimed at simulating the dynamics of algal biomass and highly associated dissolved oxygen and pH in the pond ecosystem (described in Appendix C), the first fact is incorporated as a changing maximum specific growth rate of algae, while the second fact is taken into account by modeling the dry weight of duckweed biomass as a state variable. Figures 5.24 through 5.26 show the acceptable performance of this simulation model.
Figure 5.24: Model simulation results against field observations: (1) Biomass concentration of algae (upper panel) and duckweed (lower panel).

Figure 5.25: Model simulation results against field observations: (2) DO concentration (upper panel) and pH values (lower panel).
Figure 5.26: Model simulation results against field observations: (3) Concentration of organic P (upper panel) and orthophosphate (lower panel).

Now, let us pretend that we tried to understand the pond system only resorting to one of the two modeling approaches (DBM and TBM). Consequently, we would have been only aware of one or the other fact listed above about the Whitehall pond system, depending on which modeling approach was used. Then two other simulation models are constructed with each one only incorporating one of the two facts. That is, one model accounts for the changing maximum specific growth rate of algae but leaving out the duckweed biomass (denoted as “DBM model” for the sake of illustration convenience), while the other model does include duckweed as a state variable, but using a constant algal maximum specific growth rate (denoted as “TBM model”). Both models were calibrated manually to their best visual results. Some of the simulated water quality variables of the three simulation models (the third model is the model described in Appendix C accounting for both facts,
denoted as “Joint model”) are superimposed along with their field observations in Figures 5.27 through 5.29.

The superiority of the model resulting from the two-pronged modeling approach is obvious through the comparison with each of the other two models, which are the products of either data-based modeling or theory-based modeling approach alone. Although the “DBM model” catches reasonably the time and magnitude of algal blooming, it completely fails to mimic the dynamics of the dissolved oxygen and pH, especially after the crash of the algal bloom. The reason is that, upon the algal blooms collapsing, the duckweed dominated the primary production in the pond system that is responsible for the diurnal variation of the dissolved oxygen and pH values in the water column. Since the “DBM” model omitted the duckweed as a state variable, its failure to simulate DO and pH is inevitable. However, the “TBM” model has a reasonable performance of simulating the dynamics of DO and pH, but it failed to give an acceptable prediction of the algal bloom. Since the “TBM” model did not simulate the dynamics of algal bloom correctly, which is a critical process to the pond ecosystem, its simulation results for the dynamics of DO and pH are also questionable.
Figure 5.27: Comparison of simulated algal biomass concentrations by different simulation models.

Figure 5.28: Comparison of simulated dissolved oxygen concentrations by different simulation models.
Another important way to synthesize the data-based and theory-based models is to compare the quasi-\textit{time constant} and \textit{steady-state gain} of both models. The comparisons of the \textit{time constant} and \textit{steady-state gain} contrast the two models in the parameter space not only just in the observed state variable space (see discussion in Chapter 2, and [27, 33]). Figure 5.30 demonstrates that the data-based transfer function model and the theory-based model are equivalent in describing the dynamics of the algal blooms in the pond with respect to the two composite parameters. This equivalency of the \textit{time constants} and \textit{steady-state gains} of the two types of models is important. For a specific environmental system, it allows us to derive one type of model given the other one. We will return to this discussion in Chapter 6.

However, this application of the two-pronged modeling approaches to the addressed environmental issue is not flaw-proof. The data-based modeling did not purely start from the data as it should be. The inputs of the TF models identified by the SDP modeling method
are not the pure observations of the solar radiation but the transformed quantity of Steele’s equation ([216]), which is just one of the numerous expressions for the P-I curve. The numerical experiments by the author show that a different expression for the P-I curve, or even a different value that $I_S$ in Equation (5.2) takes, may affect the identified transfer function structure considerably.
Chapter 6

Retrospective Remarks

6.1 The synthesis

Environmental systems are extremely complex, poorly defined, and hard to manipulate with designed experiments. It is a very difficult task for us, modelers, to describe the system’s behavior and understand the mechanisms behind the behavior correctly, so that scientists can assist decision makers by providing the wisdom for natural resources management and pollution prevention. Two types of models are commonly used in these situations, but either of them is able to accomplish this mission with “guaranteed satisfaction”, due to their inherent weaknesses listed in Chapter 1. Therefore, this dissertation aimed to develop a general principle of joint application of the data-based and theory-based modeling approaches outlined in Chapter 1, which, it is to be hoped, will provide us with a better understanding about the system when the two-pronged modeling approach is applied to any one of the many environmental systems of interest. Two systems, activated sludge system and aquaculture pond, representing an artificial and a natural environmental system respectively, were selected to carry out the case studies of the application of the two-pronged modeling procedure. Whether or not these applications of the two-pronged modeling approach are successful is weighted in terms of answering the three benchmark questions asked at the beginning of this dissertation, which are restated here in abbreviated forms as follows:

1. Does the joint application of the two separate modeling approaches to the same environmental system allow us to gain more information about the system than the exclusive use of either model type?
2. Can the problems such as structure identification and parameter estimation of one model type be readily solved by recourse to the other?

3. Can the two types of model starting from opposite extremes (data and theory) be reconciled in the parameter space, which is spanned by time constant and steady-state gain?

The answers to these questions should lie in the two case studies contained in Chapter 4 and Chapter 5. The following paragraphs are dedicated to summarizing the application results in terms of how the abovelisted questions are addressed in those case studies.

*Answer to question one is yes.* Both case studies have demonstrated that more information about each system has been obtained through the joint application of the data-based and theory-based modeling approach, as we expected. Since the two different modeling methods examine the system from two completely different perspectives, it is no wonder that each modeling method can observe things that the other one cannot. In Case Study Two, especially a simulation model was built by accommodating the two findings, each of which is uniquely discovered by one or the other of the two modeling procedures, to demonstrate the importance of parallel usage of the two separate modeling methods to understand and/or describe the system correctly.

- In both case studies, the data-based modeling always has detected and identified the pure time delays existing in both systems, which are otherwise not obtained easily through the theory-based modeling, or at least to identify the values for these pure time delays. In Case Study One, the pure time delay is 1 sampling interval (2 hours); in Case Study Two, the pure time delay is 2 sampling intervals (4 hours).

- In Case Study One, in the data-based modeling prong, we found that the nitrification process is mainly responsible for the ammonium-nitrogen removal in the activated sludge system, which was also confirmed in the theory-based modeling.
• In Case Study Two, the signal processing step of the data-based modeling (which also can be conducted in Case Study One) separated the signals of all water quality time-series observed in the Whitehall pond into two components, trend and periodic oscillations. Subsequently, through comparing the periodic oscillation components, a general relationship among the corresponding water quality variables can be determined (see discussions in Section 5.3 and Figure 5.5). More interestingly, this signal processing procedure could allow us to have another modeling criterion, by which the periodic components of all the signals — their amplitudes, relative phasing, timing etc. — both those from the data and those from the TBM can be compared against each other to gain further insights and condition judgements about successful/unsuccesful model-building.

• In Case Study Two, the SDP modeling procedure of the data-based modeling has discovered that sunlight utilization efficiency of the phytoplankton approaches an asymptote after the phytoplankton biomass concentration is greater than a certain level in the water column (e.g., 10mg·L⁻¹), which has not been detected in the recursive estimation of the maximum specific growth rate of phytoplankton in the theory-based modeling procedure.

• The theory-based modeling approach is just as informative as the data-based modeling, as demonstrated in both case studies. In Case Study One, the theory-based model reconstructed the nitrifying bacteria biomass concentration in the aeration tank to reveal the secret of the “out-of-phase” phenomenon between the ammonium-nitrogen removal and the MLSS concentration, which can only be explained by resorting to speculation in the context of the data-based modeling.

• In Case Study Two, the recursive parameter estimation of the theory-based model found that the growth and respiration of duckweed in the Whitehall pond is mainly responsible for the dynamics of the dissolved oxygen concentration in the water column
after the algal blooms collapsed, which it would be impossible for the data-based model to determine quantitatively.

*Answer to question two is yes, but only to a small extent.* The second question is more valuable than the first one in a sense that it could offer an ultimate cure for the weakness of either the data-based model or the theory-based model addressed thoroughly in Chapter 1. To build a data-based model with physically meaningful structures or to build a theory-based model with identifiable parameters without compromising the model’s performance on describing the system’s behaviors is always an interest of most modelers. In the practice of environmental modeling, these are very difficult tasks to accomplish and, therefore, we only expected to make modest, incremental progress in this dissertation.

- In Case Study One, the finding of the data-based modeling that the nitrification process mainly contributed to the removal of ammonium-nitrogen in the activated sludge system helped to simplify the structure of the theory-based model, which, in turn, improved the identifiability of the latter by reasonably eliminating redundant processes, hence redundant parameters.

- Again in Case Study One, the determination of the pure time delay through the data-based modeling helps to build a theory-based model with a structure that maximally approximates the real system. That is, a combination of PFR and CSTR idealizations was used to conceptualize the Orbal facility.

- In Case Study Two, the ratio of the algal maximum specific growth rates before its biomass concentration reaches $10\text{mg}\cdot\text{L}^{-1}$ and after that in the water quality simulation model, which is also a theory-based model, is determined by the data-based modeling. Therefore, it is easy to see that, to some extent, the data-based modeling can help to improve the notorious problem: that the TBM lacks identifiability. On the other hand, the theory-based modeling for the algal growth helped to interpret the meanings of the
slopes of the linear regression lines between the SDP ($\hat{b}_0(k|N)$) and the algal biomass observations ($y(k)$) in the data-based modeling.

• In Case Study Two, the theory-based modeling discovered that the duckweed played an important role in influencing the dynamics of the aquaculture pond after most of the algae die off, which was not expected by the group of experimenters at all. Therefore, the discoveries from the theory-based modeling can help to design a better experiment or monitoring program to collect data, which is essentially the starting point of the data-based modeling.

\textit{Answer to question three is yes, yet more needs to be done.} Case Study One has demonstrated that the comparison of the two types of model in parameter space is more important than that in state space (referring to the discussion of Figures 4.19 and 4.20). The equivalent models, in terms of their performance in tracking the observations, become discernable when they are compared by their \textit{time constant}. Moreover, both case studies have shown that it is not difficult to find two models that are not only comparable in the state space but also equivalent in the parameter space. However, it becomes a question as to what can be achieved after finding the two equivalent models. Case Study One has shown that if the \textit{time constant} of the theory-based model is known, then the volume of the system can be approximately determined through equations (4.8) and (4.9).

One possible way to make use of the comparability of these two models in the parameter space spanned by the \textit{time constant} and \textit{steady-state gain} is to construct one type of model from the other. For example, if a linear data-based model like (3.30)\textsuperscript{1} has been identified from the observed data, from which the \textit{time constant} and \textit{steady-state gain} of the modeled process can be derived using equations (3.31). Therefore, on one hand, in the situation, where the mechanism of a process is not understood by the modeler, a linear sort of theory-based model with lumped parameters can be constructed through equation (3.34) as a substitute.

\textsuperscript{1}Nonlinear systems can be approximated by a linear model with \textit{time-varying} parameters.
On the other hand, through equation (3.34), one complicated TBM for a process can be simplified to a simple DBM-like model for the purposes of forecasting and real-time control, where simple models are preferred for computational efficiency.

6.2 Contributions, limitations and issues for future research

In addition to the conclusions drawn from the two case studies that are summarized in the above, the contributions of this dissertation to the course of environmental modeling are also manifested in the following perspectives:

- A general two-pronged modeling approach has been developed to gain a better understanding of complex environmental systems. The application case studies have shown a moderate success of this approach.

- The Recursive Prediction Error algorithm for recursive parameter estimation of the theory-based model has been modified to accommodate time-varying parameter estimation (Section 3.3.2). The RPE algorithm developed by Ljung ([141, 142]) was designed for parameter estimation in a recursive scheme, then modified and trained onto the problem of model structure identification by Stigter ([33, 217]) to illuminate model structural inadequacies. But the RPE algorithm has never been successfully applied for time-varying parameter estimation, where not only system structure changes and/or model structural inadequacies can be detected, but also new (improved) constituent hypotheses can be generated, through which, the structural mismatches between model and system could be eventually eliminated.

- To reconcile the data-based model and the theory-based model in a parameter space spanned by the time constant and steady-state gain is a novel idea. Intuitively, one evaluates a model by looking at the goodness of fitting against data, which is in the state space, since normally the model’s outputs are functions of state variables. However,
in a recursive estimation scheme, the model’s outputs are forced by the algorithm to follow the observations, therefore, it is possible that the predictions of all models to be compared are not discernible from the observations (manifested in Figure 4.20). Bringing the model’s parameter space out of the state space through \textit{time-varying} parameter estimation and then reconciling the DBM and TBM in terms of the two lumped parameters, \textit{time constant} and \textit{steady-state gain}, is an unprecedented exercise.

As new they are, the new things always evolve not without trembles. As readers have witnessed so far, the proposed two-pronged modeling approach did not answer the three key questions perfectly. However, it is unfair to load all the burdens of blame on the proposed joint modeling approach. It is the system being treated that is so intractable, such that no perfect answer exists, except the system itself. Besides the incompleteness of the general modeling approach, some of the methods and algorithms employed in this general modeling approach are not foolproof either. Being aware of their limitations, the author desires to work on these issues in the near future.

- As mentioned in Chapters 3, 4 and 5, the RPE estimates of the \textit{time-invariant} and \textit{time-variant} parameters are not currently smoothed by any smoothing algorithm. At the transient stage, the parameter estimates overshoot their final values before they eventually converge to them. This brings some difficulties and confusions when one tries to analyze the parameter estimation results since the overshoots cannot be easily differentiated from true variations of the parameter estimates. An FIS smoothing algorithm being attached to the preceding RPE algorithm should remedy this weakness. By so doing, parameter estimates of the theory-based models may be smoothed as nearly as those of the data-based models, since the FIS algorithm is also employed in the Data-Based-Mechanistic modeling procedure.

- In order to enable the RPE algorithm to estimate a \textit{time-varying} parameter, an RW time series model is used in the modified RPE, to model the temporal variation pattern
of that parameter, as in the EKF algorithm. But this also introduces uncertainty or even bias to the parameter estimates by arbitrarily specifying the diagonal entities in the covariance matrix $Q$ for the stochastic stimuli, which once was seen as one of the reasons that the results of the EKF are less reliable than those of the RPE. As in the Data-Based-Mechanistic modeling, this problem could be overcome by optimizing the Noise Variance Ratio (NVR) matrix $Q_r$ using the Maximum Likelihood method. Furthermore, the difficulty of arbitrarily specifying the initial values for the elements of the covariance matrix $P(t_0)$ of the parameters suffered by the Data-Based-Mechanistic procedure and the RPE can be mitigated by using the final values of $P(t)$ from a preliminary analysis, as suggested in [172].

- As mentioned in Chapter 5, the data-based model of the algal growth lacks identifiability due to the introduction of Steele’s expression for the P-I curve. The structure and the parameters of the data-based model are essentially not identified, so it is questionable when the estimated results are incorporated in the theory-based model to improve the identifiability of the latter. An exercise should be attempted for identifying the data-based model of algal growth referencing to the pure solar radiation and chlorophyll-a time-series.

- How to make good use of the synthesis results of the DBM and TBM is a problem that this dissertation has not addressed sufficiently. To carry on with this topic demands time, people and creativity. The author hopes more rather less research interests will be drawn onto it.
Bibliography


A.1 Simplified Refined Instrumental Variable Method

The Simplified Refined Instrumental Variable (SRIV) identification and estimation method for constant-parameter TF models of stochastic SISO systems is a simplification of the Refined Instrumental Variable (RIV) approach ([253]), which, in turn, is an extension of the original IV estimation procedure ([119, 251]). It was shown ([254]), that under the assumption that the noise process is a serially uncorrelated series of white noise with Gaussian distribution, the complex RIV estimation algorithm could be reduced to the SRIV form.

It is assumed that the discrete-time, linear TF model of a SISO system takes the following general form,

\[ y(k) = \frac{B(z^{-1})}{A(z^{-1})} u(k - \delta) + e(k) \]  \hspace{1cm} (A.1)

where \( y(k) \) and \( u(k) \) are the system output and input respectively at the \( k \)th time instant; \( \delta \) is the system pure time delay; \( e(k) \) is the response error representing the difference between the observed behavior and estimated model output; and \( A(z^{-1}) \) and \( B(z^{-1}) \) are the denominator and numerator polynomials respectively, defined as,

\[ A(z^{-1}) = 1 + a_1 z^{-1} + \ldots + a_n z^{-n} \]  \hspace{1cm} (A.2a)

\[ B(z^{-1}) = b_0 + b_1 z^{-1} + \ldots + b_m z^{-m} \]  \hspace{1cm} (A.2b)

where \( z^{-1} \) is the backward shift operator, that is, \( z^{-i} y(k) = y(k - i) \); and the integer \( n \) and \( m + 1 \) are the number of parameters in the respective polynomials.
For the purpose of estimation, Equation (A.1) can be conveniently rewritten in the alternative vector format,

\[ y(k) = z(k)^T a + \nu(k); \quad k = 1, 2, \ldots, N \]  

(A.3)

where \( \nu(k) = e(k) \); and the vector \( z(k) \) and parameter vector \( a \) are defined respectively as,

\[ z(k) = \begin{bmatrix} -y(k-1), -y(k-2), \cdots, -y(k-n), u(k), u(k-1), \cdots \end{bmatrix}^T \]

(A.4a)

\[ a = [a_1, a_2, \cdots, a_n, b_0, b_1, \cdots, b_m]^T \]

(A.4b)

with dimensions determined by the polynomials of (A.4).

For a sample size \( N \), the non-recursive (en-bloc) SRIV estimate \( \hat{a} \) of the parameter vector \( a \) is obtained by the solution of the following “IV normal equations”,

\[
\sum_{k=1}^{k=N} \hat{x}^*(k) z^*(k)^T \hat{a} = \sum_{k=1}^{k=N} \hat{x}^*(k) y^*(k)
\]

(A.5)

These are a modification of the associated and well known least squares normal equations for the same model. For each input-output pair, the data vectors in (A.5) are defined as follows,

\[
z^*(k) = \begin{bmatrix} -y^*(k-1), -y^*(k-2), \cdots, -y^*(k-n), u^*(k), \cdots \end{bmatrix}^T
\]

(A.6a)

\[
\hat{x}^*(k) = \begin{bmatrix} \hat{x}^*(k-1), \hat{x}^*(k-2), \cdots, \hat{x}^*(k-n), u^*(k), \cdots, u^*(k-m) \end{bmatrix}^T
\]

(A.6b)

Here, \( \hat{x}(k) \) is the instrumental variable, defined as an estimate of the ‘noise free’ system output and obtained from the previous iteration of the following adaptive “auxiliary model”,

\[
\hat{x}(k) = \frac{\hat{B}(z^{-1})}{\hat{A}(z^{-1})} u(k - \delta)
\]

(A.7)

where the polynomial \( \hat{A}(z^{-1}) \) and \( \hat{B}(z^{-1}) \) are adaptive estimates of the TF model polynomials; and the star superscripts indicate that the associated variables are adaptively pre-filtered, utilizing \( \hat{A}(z^{-1}) \) obtained in the previous iteration \( (j - 1) \), in the following manner,
The adaptive pre-filter $F$ acts to remove any undesirable high frequency components from the input signal, which would otherwise reduce the efficiency of the estimation results, while retaining those frequencies that are essential for system analysis.

In the first instance, an \textit{a priori} estimate of the IV vector $\hat{x}^*(k)$ can be obtained from the “auxiliary model” (A.7) using linear least squares parameter estimates. With this preliminary IV vector, the initial estimate of the parameter vector $a$ is therefore yielded from the first run of the algorithm. Subsequently, the statistical efficiency of $\hat{a}$ is further refined through subsequent iterations of this procedure, where at each iteration, the parameters of the adaptive pre-filter and “auxiliary model” are updated as well. It has been demonstrated that the optimal SRIV parameter estimates can be obtained within only three iterations ([253, 254]).

Having determined the optimal parameter vector $\hat{a}$ from the final iteration, the following two matrices are also generated:

(i) The inverse of the Instrumental Product Matrix (IPM),

\[
P(N) = \left[ \sum_{k=1}^{k=N} \hat{x}^*(k)z^*(k)^T \right]^{-1} \tag{A.9}
\]

(ii) The covariance Matrix,

\[
P^*(N) = \hat{\sigma}^2 \left[ \sum_{k=1}^{k=N} \hat{x}^*(k)\hat{x}^*(k)^T \right]^{-1} \tag{A.10}
\]

where $\hat{\sigma}^2$ is the variance of the model residuals $\hat{e}(k)$, i.e.,

\[
\hat{e}(k) = y(k) - \hat{x}(k); \quad \hat{\sigma}^2 = \frac{1}{N} \sum_{k=1}^{k=N} \hat{e}(k)^2 \tag{A.11}
\]
In the case of Gaussian white residuals, it can be shown that $P^*(N)$ is an estimate of the covariance matrix associated with the estimated parameter vector $\hat{a}(N)$ obtained at the final iteration. Consequently the square root of its diagonal elements provide estimates of the standard error on the elements of $\hat{a}(N)$. $P(N)$ is utilized to generate the YIC criterion, as discussed below in A.3.

A.2 Fixed Interval Smoothing Algorithm

In the off-line situation, where all the time-series data are available for analysis, the time-variable parameter (acting as surrogate ‘state’) recursive estimation in the SS formulations (3.11) and (3.23) can be accompanied by a smoothing operation. Here the estimates obtained from the forward pass filtering algorithm are updated sequentially whilst working through the data in reverse order using the Fixed Interval Smoothing (FIS) algorithm, outlined in Equation (A.12), where the ‘fixed interval’ is the interval covered by the total sample size ([260]).

$$\hat{x}(k|N) = F^{-1}[\hat{x}(k + 1|N) + GQrGT L_R(k)]$$  \hspace{1cm} (A.12a)

$$P(k|N) = P(k) + P(k)F^TP^{-1}(k + 1|N)[P(k + 1|N) - P(k + 1|k)]$$

$$P^{-1}(k + 1|k)FP(k)$$  \hspace{1cm} (A.12b)

$$L_R(k) = [I - P(k + 1)H^T(k + 1)H(k + 1)]^T \{F^TL_R(k + 1)$$

$$- H^T[y(k) - H(k + 1)F\hat{x}(k)] \}$$  \hspace{1cm} (A.12c)

with $L_R(N) = 0$ and $P^*(k|N) = \hat{\sigma}^2P(k|N)$ is the error covariance matrix associated with state estimates which, in the TVP context, define the estimated variations in the parameters. In theory, the vector $H(k)$ should contain only pure deterministic variables. However, in low noise situations, this vector is replaced ‘sub-optimally’ to contain stochastic variables contaminated by low measurement noises. In the case of the TVP estimation of the generic TF model (3.8), the vector $H(k)$ is represented by $z(k)^T$ and the ‘surrogate’ state vector
\( x \) is substituted by the polynomial parameter vector \( a \), while for the TVP estimation of mechanistic model (3.19), \( H(k) \) is proposed to be \( H(x, u; t_k) \) (see the discussion in Section 3.3.2) and \( x \) is substituted by the parameter vector \( \alpha \).

### A.3 Model Order Identification (excerpted from [268])

Model order identification is based around the \( R^2_T \), \( YIC \) and \( AIC \) criteria, which are defined as follows:

\[
R^2_T = 1 - \frac{\hat{\sigma}^2}{\sigma_y^2} \tag{A.13}
\]

\[
YIC = \log e \frac{\hat{\sigma}^2}{\sigma_y^2} + \log e \text{NEVN} \tag{A.14}
\]

\[
AIC(np) = N \log e \hat{\sigma}^2 + 2 \cdot np \tag{A.15}
\]

in which,

\[
\sigma_y^2 = \frac{1}{N} \sum_{k=1}^{N} [y(k) - \bar{y}]^2;
\]

\[
\bar{y} = \frac{1}{N} \sum_{k=1}^{N} y(k);
\]

\[
\text{NEVN} = \frac{1}{np} \sum_{i=1}^{np} \frac{\hat{\sigma}^2 \cdot \hat{p}_{ii}}{\hat{a}_i^2}
\]

and \( np = n + m + 1 \) is the number of estimated parameters in the \( \hat{a}(N) \) vector; \( p_{ii} \) is the \( i \)th diagonal element of the \( P(N) \) matrix (so that \( \hat{\sigma}^2 \cdot \hat{p}_{ii} \) can be considered as an approximate estimate of the variance of the estimated uncertainty on the \( i \)th parameter estimate); and \( \hat{a}_i^2 \) is the square of the \( i \)th parameter in the \( \hat{a}(N) \) vector.

We see that the coefficient of determination \( R^2_T \) is a statistical measure of how well the model explains the data. If the variance of the model residuals \( \hat{\sigma}^2 \) is low compared with the variance of the data \( \sigma_y^2 \), the \( R^2_T \) tends towards unity; while if \( \hat{\sigma}^2 \) is of similar magnitude to \( \sigma_y^2 \) then it tends towards zero. Note, however, that \( R^2_T \) is based on the variance of the model
errors $\hat{e}(k)$ and it is not the more conventional coefficient of determination $R^2$ based on the variance of the one-step ahead prediction errors. $R^2_T$ is a more discerning measure than $R^2$ for TF model identification: while it is often quite easy for a model to produce small one-step ahead prediction errors, since the model prediction is based on past measured values of the output variable $y(k)$, it is far more difficult for it to yield small model response errors, where the model output is based on only on the measured input variable $u(k)$ and does not refer to $y(k)$.

The Young Information Criterion (YIC) is a more complex, heuristic criterion. From the definition of $R^2_T$, we see that the first term is simply a relative measure of how well the model explains the data: the smaller the model residuals the more negative the term becomes. The second term, on the other hand, provides a measure of the conditioning of the IPM, which needs to be inverted when the IV normal equation (A.5) are solved. If the model is over-parameterized, then it can be shown (e.g., [243, 264]) that the IPM will tend to singularity. And then, because of its ill-conditioning, the elements of the inverse $P(N)$ will increase in value, often by several orders of magnitude. When this happens, the second term in the YIC tends to dominate the criterion function, indicating over-parameterization. An alternative justification of the YIC can be obtained from statistical considerations (see e.g., [255]). Although heuristic, the YIC has proven useful in practical identification terms over the past ten years: it should not be used as a sole arbiter of model order, however, and improvements in its statistical definition are being researched.

The Akaike Information Criterion (AIC) is a well known identification criterion for purely stochastic AR processes ([3]) and it is used to identify the order of AR models for the spectral analysis and the noise processes of the DBM models (although it can sometimes assist in identifying the TF model order if used carefully). Here, the first term is a measure of how well the model explains the data; while the second term is simply a penalty on the number of parameters in the model. Thus, as in the YIC, the AIC seeks a compromise between the
degree of model fit and the complexity of the model; i.e., it helps to ensure parsimony and simplicity in the overall TF model.
Appendix B

The Recursive Prediction Error Algorithm

The model for a system can be given by the following innovation format with the observation equation (B.1b) assumed linear in its arguments.

\[ \dot{x}(t) = f(x, u, \alpha; t) + K \nu(t) \]  \hspace{1cm} (B.1a)

\[ y(t_k) = H(\alpha; t_k)x(t_k) + \nu(t_k) \]  \hspace{1cm} (B.1b)

where, \( f \) is a vector of nonlinear functions; \( H(\alpha; t_k) \) is a vector relating observation vector \( y(t_k) \) to the state variables defined by (B.1a), normally with alternate unity and zero elements; \( u, x, y \) are the input, state, and output vectors, respectively; \( \alpha \) is a vector of model parameters; \( \nu(t_k) \) is the observation noise sequence and estimated using the innovation \( \epsilon(t_k) \), i.e., the difference between model output \( \hat{y}(t_k) \) and the observation \( y(t_k) \); \( K \) is the Kalman-like gain matrix amplifying the innovation process to arrive at the ‘system noise’ \( K \nu(t) \); and \( t, t_k \) are continuous and discrete time, respectively.

The states \( x \) and the parameters \( \alpha \) in Equation (B.1) and the associated covariance matrix \( P_{\alpha \alpha} \) of \( \alpha \) are recursively estimated by means of the Recursive Prediction Error (RPE) algorithm defined as follows:

\[ \hat{x}(t_k) = \int_{t_{k-1}}^{t_k} f(\hat{x}(t), u(t_{k-1}), \hat{\alpha}(t_{k-1}))dt + K(t_{k-1})\epsilon(t_{k-1}) \]  \hspace{1cm} (B.2a)

\[ \alpha(t_k) = \alpha(t_{k-1}) + L(t_k)\epsilon(t_k) \]  \hspace{1cm} (B.2b)

\[ P_{\alpha \alpha}(t_k) = [I - L(t_k)\psi^T(t_k, \hat{\alpha}(t_{k-1}))] P_{\alpha \alpha}(t_{k-1})[I - \psi(t_k, \hat{\alpha}(t_{k-1}))L^T(t_k)] \]
\[ + L(t_k)\Lambda(t_k)L^T(t_k) \]  \hspace{1cm} (B.2c)
where,

\[ e(t_k) = y(t_k) - \hat{y}(t_k) \]  

\[ \hat{y}(t_k) = H(\hat{\alpha}; t_{k-1})\hat{x}(t_k) \]  

\[ L(t_k) = P_{\alpha\alpha}(t_{k-1})\psi(t_k, \hat{\alpha}(t_{k-1}))S^{-1}(t_{k-1}) \]  

\[ S(t_k) = \psi^T(t_k, \hat{\alpha}(t_{k-1}))P_{\alpha\alpha}(t_{k-1})\psi(t_k, \hat{\alpha}(t_{k-1}) + \Lambda(t_k) \]  

\[ \psi^T(t_k, \hat{\alpha}(t_{k-1})) = H(\hat{\alpha}; t_{k-1})W(t_k) + D(t_k, \hat{\alpha}(t_{k-1})) \]  

\[ W(t) = \Phi_W(t_k, t_{k-1}; W(t_{k-1})) + \int_{t_{k-1}}^{t_k} \Phi_W(t, \tau; W(\tau)) \times (M(\tau, \hat{x}(\tau, \hat{\alpha}), u(\tau, \alpha)) \right) \]  

\[ -K(\hat{\alpha}(t_{k-2}))D(t_k, \hat{\alpha}(t_{k-2})))d\tau \]  

In turn, the relevant matrices in Equations (B.2) are illustrated and defined in the following paragraphs (refer to [217] for details).

\[ \psi(t_k, \alpha) \] denotes the gradient of the prediction \( \hat{y}(t_k; \alpha) \) with respect to \( \alpha \), i.e.,

\[ \psi(t_k, \alpha) \triangleq \frac{d(\hat{y}(t_k; \alpha))}{d\alpha} = -[\frac{d\epsilon(t_k; \alpha)}{d\alpha}]^T \]  

\[ W(t) \] is the sensitivity matrix that represents the sensitivity of the state with respect to each of the parameters, i.e.:  

\[ W(t) \triangleq \frac{d\hat{x}(t; \alpha)}{d\alpha} \]  

\[ \Lambda(t_k) \] is the measurement noise matrix, i.e.,

\[ \Lambda(t_k) = E\{\nu^T(t_k)\nu(t_k)\} \]  

\( \overline{M}(\hat{x}(t, \alpha), u, \alpha; t) \) is defined as:

\[ \overline{M}(\hat{x}(t, \alpha), u, \alpha; t) = \frac{\partial}{\partial\alpha} \{ f(\hat{x}(t; \alpha), u, \alpha; t) + K(\alpha)e(t, \alpha) \} \]  

\( \Phi_W(t_{k+1}, t_k; W(t_k)) \) is the state transition matrix that is associated with the homogeneous solution of (B.7) ([72])

\[ \frac{dW(t)}{dt} = (F_x(\hat{x}, \alpha; t) - K(\alpha)H(\alpha; t_k))W(t) \]
$D(\alpha; t_k)$ is defined as:

$$D(\alpha; t_k) = \frac{\partial}{\partial \alpha} \{ H(\alpha; t_k) x(t_k) \} \quad (B.8)$$
Appendix C

A Simulation Model for Aquaculture Pond Water Quality

This model, shown in Figure C.1, is composed of 7 state variables for the purpose of simulating the dynamics of algae, dissolved oxygen and pH in a small aquaculture pond. The environmental factors, state variables, and the parameter connecting these state variables are defined in Table C.1. The kinetic rates of all associated processes and the change rate of state variables are summarized in the following paragraphs.

Processes

The kinetic rate of each process appearing in Figure C.1 is defined as follows:

(1) Re-aeration of dissolved oxygen (DO):

\[ r_1 = K_L aO_2(C_{S,O_2} - x_3(t)) \quad (C.1) \]

(2) Carbon dioxide (CO₂) exchange with atmosphere:

\[ r_2 = K_L aCO_2(C_{S,CO_2} - CO_2(q)) \quad (C.2) \]

(3) Algal photosynthesis:

\[ r_3 = \mu_{max,AG}(t)\theta^{R(t)-20}_{\mu,AG} \frac{x_7(t)}{K_{P,AG} + x_7(t)} \frac{I_{0.5}(t)}{I_{s,AG}} \exp\left[1 - \frac{I_{0.5}(t)}{I_{s,AG}}\right] x_1(t) \quad (C.3) \]

where,

\[ \mu_{max,AG}(t) = \begin{cases} 
\mu_{max,AG,1}, & \text{before } x_1(t) \text{ reaches } 10.0 \text{ mg}\cdot\text{L}^{-1}; \\
\mu_{max,AG,2}, & \text{after } x_1(t) \text{ reaches } 10.0 \text{ mg}\cdot\text{L}^{-1}.
\end{cases} \]
State variables:

1. $x_1(t)$: Algal biomass concentration; $x_2(t)$: Duckweed biomass concentration;
2. $x_3(t)$: Total inorganic carbon ($C_T$) molar concentration, collective term of $[H_2CO_3^+]$, $[HCO_3^-]$, $[CO_3^{2-}]$; $x_4(t)$: Alkalinity (Alk) molar concentration, collective term of $[HCO_3^-]$, $2[CO_3^{2-}]$, $[OH^-]$ and $-[H^+]$; $x_5(t)$: Organic phosphorus (OP) concentration; $x_6(t)$: Orthophosphate ($PO_4^{-3}$) concentration.

Figure C.1: Conceptual model of an aquaculture pond.
(4) Algal respiration:

\[ r_4 = k_{r,\text{AG}} \theta_{r,\text{AG}}^{\theta(T)-20} x_1(t) \quad \text{(C.4)} \]

(5) Duckweed photosynthesis:

\[ r_5 = \mu_{\text{max,DW}} \theta_{\mu,\text{DW}}^{\theta(T)-20} \frac{x_7(t)}{K_p,\text{DW} + x_7(t)} \frac{I_o(t)}{I_{s,\text{DW}}} \exp\left[1 - \frac{I_o(t)}{I_{s,\text{DW}}}\right] x_2(t) \quad \text{(C.5)} \]

(6) Duckweed respiration:

\[ r_6 = k_{r,\text{DW}} \theta_{r,\text{DW}}^{\theta(T)-20} x_2(t) \quad \text{(C.6)} \]

(7) Carbonaceous matter aerobic degradation:

\[ r_7 = k_{\text{ox,C}} \theta_{\text{ox}}^{\theta(T)-20} TOC(t) \quad \text{(C.7)} \]

(8) Algal nonpredatory mortality and excretion:

\[ r_8 = k_{d,\text{AG}} x_1(t) \quad \text{(C.8)} \]

(9) Algal predatory mortality:

\[ r_9 = k_{p,\text{AG}} x_1(t) \quad \text{(C.9)} \]

(10) Algal settling to sediment:

\[ r_{10} = \frac{\omega_{\text{AG}}}{D_w} x_1(t) \quad \text{(C.10)} \]

(11) Duckweed mortality death:

\[ r_{11} = k_{d,\text{DW}} x_2(t) \quad \text{(C.11)} \]

(12) Duckweed photo-respiration (excretion):

\[ r_{12} = k_{e,\text{DW}} x_2(t) \quad \text{(C.12)} \]

(13) Mineralization of organic phosphorus:

\[ r_{13} = k_{m,\text{OP}} \theta_{m,\text{OP}}^{\theta(T)-20} \frac{x_1(t)}{K_{m,\text{OP,AG}} + x_1(t)} x_6(t) \quad \text{(C.13)} \]
(14) Organic phosphorus settling:

\[ r_{14} = \frac{\omega_{OP}}{D_w} x_6(t) \]  

(C.14)

(15) Ammonium-nitrogen nitrification:

\[ r_{15} = k_{nt} \theta_{nt}^{T(t)-20} \frac{x_3(t)}{K_{O_2,NT} + x_3(t)} NH_4(t) f_{pH} \]  

(C.15)

where,

\[ f_{pH} = \begin{cases} 
35.019 - 24.538 \times pH + 6.147 \times pH^2 \\
-0.647 \times pH^3 + 0.02435 \times pH^4, & \text{if } 4 \leq pH \leq 9; \\
0, & \text{otherwise.}
\end{cases} \]

in which,

\[ pH = -\log_{10} \left[ \frac{K_1 \Delta + \sqrt{(K_1 \Delta)^2 + 4K_1 K_2 x_5(t) [x_4(t) + \Delta]}}{2x_5(t)} \right] \]

with \( \Delta = x_4(t) - x_5(t) \).

(16) Nitrate-nitrogen denitrification:

\[ r_{16} = k_{dn} \theta_{dn}^{T(t)-20} \frac{K_{O_2,DN}}{K_{O_2,DN} + x_3(t)} NO_3(t) \]  

(C.16)

**Model**

1. **Algal biomass concentration** \( (x_1(t)) \), linearly related to the quantity of chlorophyll-\( a \) of phytoplankton, is mainly affected by photosynthesis (growth and respiration), zooplankton grazing, non-predatory death, settling processes. The change rate of algal biomass concentration can be described by Equation (C.17).

\[ \frac{dx_1(t)}{dt} = r_3 - r_4 - r_8 - r_9 - r_{10} - \text{discharge} \]  

(C.17)

2. **Duckweed biomass** \( (x_2(t)) \) dynamics is subject to the contributions from photosynthesis, dark respiration, photorespiration (excretion), and mortality processes (\( 62, 154, \ldots \)).
The change rate of its equivalent concentration in water column is depicted by Equation (C.18).

$$\frac{dx_2(t)}{dt} = r_5 - r_6 - r_{11} - r_{12} - \text{discharge}$$ (C.18)

3. **Dissolved oxygen** \((x_3(t))\) modeling involves processes such as re-aeration, photosynthesis and respiration of algae and aquatic plants, aerobic degradation of organic matters in water column and top layer of sediment, nitrification etc. The change rate of dissolved oxygen in water column can be modeled as Equation (C.19).

$$\frac{dx_3(t)}{dt} = r_1 + \frac{138}{106} \left[ (r_3 - r_4) + 1.2(r_5 - r_6) \right] - \frac{64}{14} r_{15} + \text{loading} - \text{discharge}$$ (C.19)

4. **Total inorganic carbon** \((x_4(t))\) dynamics in water column is mainly contributed by the change of carbon dioxide, which, in turn, is a function of processes such as photosynthesis and respiration, carbonaceous matters oxidation, denitrification, and exchange with atmosphere. Its change rate in water column can be written as Equation (C.20).

$$\frac{dx_4(t)}{dt} = r_2 + \frac{1}{12000} \left[ r_4 - r_3 + 1.2(r_6 - r_5) + \frac{12}{32} r_7 + \frac{5}{14} r_{16} \right] + \text{loading} - \text{discharge}$$ (C.20)

5. **Alkalinity** \((x_5(t))\) variation is also a function of the dynamics of carbon dioxide in the pond system, which involves photosynthesis and respiration processes of algae and duckweed. In addition, nitrification of ammonium-nitrogen reduces alkalinity in water column. Therefore, the change rate of alkalinity is model by Equation (C.21).

$$\frac{dx_5(t)}{dt} = \frac{1}{1000} \left[ \frac{18}{12} \left[ (r_4 - r_3) + 1.2(r_6 - r_5) \right] - \frac{2}{14} r_{15} \right] + \text{loading} - \text{discharge}$$ (C.21)

6. **Organic phosphorus** \((x_6(t))\) only refers to nonliving portion in the water column in this model. As algae and aquatic plants respire and die, biomass is decomposed to nonliving organic and inorganic matter (phosphorus in particular). The resulting inorganic phosphorus is available immediately to the primary producers in ecosystem, while the
organic phosphorus must undergo mineralization or bacterial decomposition into inorganic phosphorus before uptake. Therefore, the dynamics of organic P is modeled as Equation (C.22).

\[
\frac{dx_6(t)}{dt} = \frac{1}{106 12} f_{OP}(r_4 + r_6 + r_8 + r_{11} + r_{12}) - r_{13} - r_{14} + \text{loading} - \text{discharge} \quad (C.22)
\]

7. **Orthophosphate** \((x_7(t))\) dynamics is described by Equation (C.23).

\[
\frac{dx_7(t)}{dt} = \frac{1}{106 12} (1 - f_{OP})(r_4 + r_6 + r_8 + r_{11} + r_{12} - r_3 - r_5) + r_{13} + \text{loading} - \text{discharge} \quad (C.23)
\]

Table C.1: Definition of state variables, environmental factors and parameters.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition (Unit)</th>
<th>Value</th>
<th>Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>State Variables:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(x_1(t))</td>
<td>Algal biomass concentration (mgC·L(^{-1}))</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(x_2(t))</td>
<td>Duckweed biomass concentration (mgC·L(^{-1}))</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(x_3(t))</td>
<td>Dissolved oxygen concentration (mg·L(^{-1}))</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(x_4(t))</td>
<td>Total inorganic carbon molar concentration (M)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(x_5(t))</td>
<td>Alkalinity molar concentration (M)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(x_6(t))</td>
<td>Organic phosphorus concentration (mgP·L(^{-1}))</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(x_7(t))</td>
<td>Orthophosphate concentration (mgP·L(^{-1}))</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Environmental factors:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(I_{0.5}(t))</td>
<td>Photosynthetically active radiation at 0.5 m below water surface (KJ·m(^{-2})h(^{-1}))</td>
<td></td>
<td>Observed</td>
</tr>
<tr>
<td>(I_0(t))</td>
<td>Photosynthetically active radiation at water surface (KJ·m(^{-2})h(^{-1}))</td>
<td></td>
<td>Observed</td>
</tr>
<tr>
<td>(T(t))</td>
<td>Water temperature (°C)</td>
<td></td>
<td>Observed</td>
</tr>
<tr>
<td>(TOC(t))</td>
<td>Total oxidizable carbon concentration (mgC·L(^{-1}))</td>
<td></td>
<td>Observed</td>
</tr>
<tr>
<td>(NH_4(t))</td>
<td>Ammonium-nitrogen concentration (mgN·L(^{-1}))</td>
<td></td>
<td>Observed</td>
</tr>
<tr>
<td>(NO_3(t))</td>
<td>Nitrate-nitrogen concentration (mgN·L(^{-1}))</td>
<td></td>
<td>Observed</td>
</tr>
<tr>
<td>(p_{CO_2})</td>
<td>Partial pressure of CO(_2) in atmosphere (atm)</td>
<td>(10^{-3.5})</td>
<td>[225]</td>
</tr>
<tr>
<td>Parameters:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\cdot)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table C.1: (continued)

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition (Unit)</th>
<th>Value</th>
<th>Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_{\text{max,AG,1}}$</td>
<td>Algal maximum specific growth rate at 20 °C before biomass reaching 10 mg·L$^{-1}$ (d$^{-1}$)</td>
<td>1.57</td>
<td>Calibrated</td>
</tr>
<tr>
<td>$\mu_{\text{max,AG,2}}$</td>
<td>Algal maximum specific growth rate at 20 °C after biomass reaching 10 mg·L$^{-1}$ (d$^{-1}$)</td>
<td>1.08</td>
<td>DBM*</td>
</tr>
<tr>
<td>$\mu_{\text{max,DW}}$</td>
<td>Duckweed maximum specific growth rate at 20 °C (d$^{-1}$)</td>
<td>0.867</td>
<td>Calibrated</td>
</tr>
<tr>
<td>$k_r,\text{AG}$</td>
<td>Algal respiration rate at 20 °C (d$^{-1}$)</td>
<td>0.16</td>
<td>Calibrated</td>
</tr>
<tr>
<td>$k_r,\text{DW}$</td>
<td>Duckweed respiration rate at 20 °C (d$^{-1}$)</td>
<td>0.4</td>
<td>Calibrated</td>
</tr>
<tr>
<td>$k_d,\text{AG}$</td>
<td>Algal nonpredatory death rate (d$^{-1}$)</td>
<td>0.05</td>
<td>[273]</td>
</tr>
<tr>
<td>$k_d,\text{DW}$</td>
<td>Duckweed death rate (d$^{-1}$)</td>
<td>0.002</td>
<td>Calibrated</td>
</tr>
<tr>
<td>$k_p,\text{AG}$</td>
<td>Algal predatory death rate (d$^{-1}$)</td>
<td>0.002</td>
<td>Calibrated</td>
</tr>
<tr>
<td>$k_e,\text{DW}$</td>
<td>Duckweed photorespiration (excretion) rate (d$^{-1}$)</td>
<td>0.012</td>
<td>Calibrated</td>
</tr>
<tr>
<td>$I_{s,\text{AG}}$</td>
<td>Light intensity at 0.5 m below surface for algal maximum growth (KJ·m$^{-2}$·h$^{-1}$)</td>
<td>1500</td>
<td>Calibrated</td>
</tr>
<tr>
<td>$I_{s,\text{DW}}$</td>
<td>Light intensity at surface for duckweed maximum growth (KJ·m$^{-2}$·h$^{-1}$)</td>
<td>$5\times10^5$</td>
<td>Calibrated</td>
</tr>
<tr>
<td>$\omega_{\text{AG}}$</td>
<td>Algae settling rate (m·d$^{-1}$)</td>
<td>0.01</td>
<td>Calibrated</td>
</tr>
<tr>
<td>$\omega_{\text{OP}}$</td>
<td>Organic P settling rate (m·d$^{-1}$)</td>
<td>0.14</td>
<td>[273]</td>
</tr>
<tr>
<td>$K_{L,aO_2}$</td>
<td>Oxygen re-aeration coefficient (d$^{-1}$)</td>
<td>0.4</td>
<td>Calibrated</td>
</tr>
<tr>
<td>$K_{L,aCO_2}$</td>
<td>Carbon dioxide transfer coefficient (d$^{-1}$)</td>
<td>0.37†</td>
<td>[225]</td>
</tr>
<tr>
<td>$k_{ox,\text{C}}$</td>
<td>Carbonaceous matter aerobic degradation rate at 20 °C (d$^{-1}$)</td>
<td>0.03</td>
<td>Calibrated</td>
</tr>
<tr>
<td>$k_{nt}$</td>
<td>Nitrification rate at 20 °C (d$^{-1}$)</td>
<td>0.3</td>
<td>Calibrated</td>
</tr>
<tr>
<td>$k_{dn}$</td>
<td>Denitrification rate at 20 °C (d$^{-1}$)</td>
<td>0.1</td>
<td>[273]</td>
</tr>
<tr>
<td>$k_{m,\text{OP}}$</td>
<td>Organic P mineralization rate (d$^{-1}$) at 20 °C (d$^{-1}$)</td>
<td>0.1</td>
<td>[273]</td>
</tr>
<tr>
<td>$K_{\text{P,AG}}$</td>
<td>Half saturation constant of phosphorus for algal growth (mgP·L$^{-1}$)</td>
<td>0.075</td>
<td>[248]</td>
</tr>
<tr>
<td>$K_{\text{P,DW}}$</td>
<td>Half saturation constant of phosphorus for duckweed growth (mgP·L$^{-1}$)</td>
<td>0.005</td>
<td>Calibrated</td>
</tr>
<tr>
<td>$K_{O_2,\text{NT}}$</td>
<td>Half saturation constant of oxygen for nitrification (mg·L$^{-1}$)</td>
<td>2.0</td>
<td>[273]</td>
</tr>
<tr>
<td>$K_{O_2,\text{DN}}$</td>
<td>Half saturation constant of oxygen for denitrification (mg·L$^{-1}$)</td>
<td>0.1</td>
<td>[273]</td>
</tr>
<tr>
<td>$K_{\text{m,OP,AG}}$</td>
<td>Half saturation constant of algae for OP mineralization (mgC·L$^{-1}$)</td>
<td>1.0</td>
<td>[273]</td>
</tr>
<tr>
<td>$f_{\text{OP}}$</td>
<td>Fraction of algae debris decomposed to organic P</td>
<td>0.5</td>
<td>[273]</td>
</tr>
<tr>
<td>$\theta_{\mu,\text{AG}}$</td>
<td>Temperature coefficient for algal growth</td>
<td>1.047</td>
<td>[273]</td>
</tr>
<tr>
<td>$\theta_{\mu,\text{DW}}$</td>
<td>Temperature coefficient for duckweed growth</td>
<td>1.04</td>
<td>Calibrated</td>
</tr>
</tbody>
</table>

*The ratio of $\mu_{\text{max,AG,1}}$ and $\mu_{\text{max,AG,2}}$ is determined by the data-based modeling results.
†This figure is calculated by $(32/44)^{0.25}K_{L,aO_2}$. 
Table C.1: (continued)

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition (Unit)</th>
<th>Value</th>
<th>Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_{r,AG}$</td>
<td>Temperature coefficient for algal respiration</td>
<td>1.045</td>
<td>[273]</td>
</tr>
<tr>
<td>$\theta_{r,DW}$</td>
<td>Temperature coefficient for duckweed respiration</td>
<td>1.02</td>
<td>Calibrated</td>
</tr>
<tr>
<td>$\theta_{ox}$</td>
<td>Temperature coefficient for carbonaceous matter aerobic degradation</td>
<td>1.047</td>
<td>[273]</td>
</tr>
<tr>
<td>$\theta_{nt}$</td>
<td>Temperature coefficient for nitrification</td>
<td>1.08</td>
<td>[273]</td>
</tr>
<tr>
<td>$\theta_{dn}$</td>
<td>Temperature coefficient for denitrification</td>
<td>1.08</td>
<td>[273]</td>
</tr>
<tr>
<td>$\theta_{m,OP}$</td>
<td>Temperature coefficient for organic P mineralization</td>
<td>1.08</td>
<td>[273]</td>
</tr>
<tr>
<td>$K_H$</td>
<td>Henry’s law constant (M·atm$^{-1}$)</td>
<td>$10^{-1.47}$</td>
<td>[54]</td>
</tr>
<tr>
<td>$K_1$</td>
<td>The first dissociation constant of carbonic acid</td>
<td>$10^{-6.35}$</td>
<td>[54]</td>
</tr>
<tr>
<td>$K_2$</td>
<td>The second dissociation constant of carbonic acid</td>
<td>$10^{-10.33}$</td>
<td>[54]</td>
</tr>
<tr>
<td>$D_w$</td>
<td>Average depth of pond (m)</td>
<td>2.0</td>
<td>Calibrated</td>
</tr>
<tr>
<td>$C_{s,O_2}$</td>
<td>Saturation concentration of DO at standard condition (mg·L$^{-1}$)</td>
<td>‡</td>
<td>[273]</td>
</tr>
<tr>
<td>$C_{s,CO_2}$</td>
<td>Saturation concentration of CO$_2$ at standard condition (M)</td>
<td>§</td>
<td>[54]</td>
</tr>
</tbody>
</table>

‡Using the polynomial: $14.5412 - 0.3928T(t) + 0.007323T(t)^2 - 0.00006629T(t)^3$

§Using Henry’s law: $K_{HP}CO_2$
function [Dim, Date, State, Param, Var, Sen, SenNorm, Innos, Obs] ...  
  = Rpe4Tvp(file);  
%=================================================================
%RPE4TVP Recursive Prediction Error (RPE)
%algorithm for time-varying parameter (TVP) estimation.
%INPUT: 
%  - FILE: a .CSV file specifying system dimensions, file names 
%    for model and data, and initial values for estimation, etc. 
%OUTPUT: 
%  - DIM: system dimensions. 
%  - DATE: time or sampling instant. 
%  - STATE: state variables reconstructed by RPE. 
%  - PARAM: parameters including Kalman-like gain matrix, estimated 
%    by RPE. 
%  - VAR: diagonal elements of the variance-covariance matrix of the 
%    estimated parameters. 
%  - SEN: sensitivity coefficients. 
%  - SENNORM: normalized sensitivity coefficients. 
%  - INNOS: innovations. 
%  - OBS: observed time series. 
%==================================================================
[dimOfSys, nameOfFile, colOfData] = InitialFile(file);  
%Dimension of system
numOfState = dimOfSys(1);  
numOfParam = dimOfSys(2);  
numOfObserved = dimOfSys(3);  
numOfKalman = numOfState * numOfObserved;  
%  - Kalman-like gain matrix dimension
%File names of model and data
modelFile = nameOfFile{1};  
dataFile = nameOfFile{2};  
sheetName = nameOfFile{3};  
%Column specification of data file
timeCol = colOfData{1};
sampleCol = colOfData{2};
inputCol = colOfData{3};
outputCol = colOfData{4};
environCol = colOfData{5};
% Load time series from data file
[dataSet, headName] = eval(['xlsread' '(', 'dataFile', ')']);
time = eval(['dataSet', '(:,', timeCol, ')']);
sample = eval(['dataSet', '(:,', sampleCol, ')']);
U = eval(['dataSet', '(:,', inputCol, ')']);
Y = eval(['dataSet', '(:,', outputCol, ')']);
environ = eval(['dataSet', '(:,', environCol, ')']);
N = length(sample);
% Provide initial values for recursive estimation
[ivOfX, ivOfAlfa, ivOfVar, ivOfQ] = InitialValue(dimOfSys, file);
% Initialize matrices for storing results
stgState = zeros(numOfState, N);
stgParam = zeros(numOfParam+numOfKalman, N);
stgVar = zeros(numOfParam+numOfKalman, N);
stgSen = zeros(numOfState*(numOfParam+numOfKalman), N);
stgSenNorm = zeros(numOfState*(numOfParam+numOfKalman), N);
stgInno = zeros(numOfObserved, N);
% Initialization
xHat = ivOfX; % state variables
alfaHat = ivOfAlfa; % parameters
P = diag(ivOfVar); % variance-covariance matrix of parameter
W = zeros(numOfState, (numOfParam+numOfKalman)); % sensitivity coefficients
Epsn = [0.1; 0.05]; % innovations
% Some special vectors and matrices
rW = reshape(W, [1, numOfState*(numOfParam+numOfKalman)]);
% - row vector of W for ODE computation purpose
Lmbda = diag(Lambda(Y));
% - measurement noises
Qr = ivOfQ;
% - stimuli for TVP dynamics
TVP = [0 0];
% - Gauss-Markov models for TVP dynamics:
% 0 -- Random Walk (RW) model
% 1 -- Integrated Random Walk (IRW) model
% <1 -- Smoothed Random Walk (SRW) model
% RPE for TVP estimation loop
for i=1:N-1
    % Prediction step
    if i == 1
[alfaHatPre, PPre, alfaPre2, PPre2] = ... 
    Prediction(dimOfSys, TVP, alfaHat, diag(P), Qr);
else
    [alfaHatPre, PPre, alfaPre2, PPre2] = ... 
    Prediction(dimOfSys, TVP, alfaHat, diag(P), Qr, ... 
       alfaPre2, PPre2);
end;
K = reshape(alfaHatPre((numOfParam+1):(numOfParam+numOfKalman)), ... 
    [numOfState, numOfObserved]);
%Calculate state variables and sensitivity coefficients
[T, X] = ode23(modelFile, [time(i) time(i+1)], [xHat' rW], [], ... 
    dimOfSys, U(i,:), alfaHatPre, K, Epsn, envirn(i,:));
xHat = X(size(X,1), 1:numOfState)';
% - last values in this interval (nOfP*1)
rW = X(size(X,1), (numOfState+1):(numOfState ... 
    + numOfState * (numOfParam+numOfKalman)));
W = reshape(rW, [numOfState, (numOfParam+numOfKalman)]);
%Correction step
[alfaHat, P, Epsn] = Correction(dimOfSys, Y(i+1,:)’, ... 
    xHat, W, alfaHatPre, PPre, Lmbda);
%Calculate normalized SC
mW = W .* repmat(alfaHat’, [numOfState, 1]);
nW = reshape(mW, [numOfState*(numOfParam+numOfKalman), 1]);
%Store results
stgState(:, i+1) = xHat;
stgParam(:, i+1) = alfaHat;
stgVar(:, i+1) = diag(P);
stgSen(:, i+1) = rW’;
stgSenNorm(:, i+1) = nW;
stgInno(:, i+1) = Epsn;
end
%Returning results
Dim = dimOfSys;
Date = time; %+datenum(’30-Dec-1899’);
State = stgState;
Param = stgParam;
Var = stgVar;
Sen = stgSen;
SenNorm = stgSenNorm;
Innos = stgInno;
Obs = Y’;
%end-of-function
function [AlfaPre, PPre, AlfaGradnt, PGradnt] = Prediction(dim, tvp, alfa, var, qr, alfa2, var2)

% PREDICTION Prediction step of RPE algorithm for TVP estimation
numOfState = dim(1);
numOfParam = dim(2);
numOfObserved = dim(3);
umOfKalman = numOfState * numOfObserved;

% Construct F and G matrices for GRW models
[F, G] = FGMatrix(tvp);
for i = 1:numOfKalman;
    F = blkdiag(F, [1 0; 0 0]);
    G = blkdiag(G, [1 0; 0 0]);
end;

% Duplicate elements of vectors if IRW or SRW being used
qrDupt = Duplicate(qr);
if nargin <= 5
    alfaDupt = Duplicate(alfa);
    varDupt = Duplicate(var);
else
    alfaDupt = Duplicate(alfa, alfa2);
    varDupt = Duplicate(var, var2);
end;

% Construct parameter variance-covariance matrix and Q matrix
PDupt = diag(varDupt);
QDupt = diag(qrDupt);

% Predict the duplicated vectors
alfaDuptPre = F * alfaDupt;
PDuptPre = F * PDupt * F' + G * QDupt * G';
diagOfPDuptPre = diag(PDuptPre);

% Returning results
AlfaPre = alfaDuptPre(1:2:length(alfaDuptPre)-1);
PPre = diag(diagOfPDuptPre(1:2:length(diagOfPDuptPre)-1));
AlfaGradnt = alfaDuptPre(2:2:length(alfaDuptPre));
PGradnt = diagOfPDuptPre(2:2:length(diagOfPDuptPre));

end-of-function
numOfKalman = numOfState * numOfObserved;
%Initialize matrices in RPE algorithm
D = zeros(numOfObserved, (numOfParam+numOfKalman));
% - differentiation of linear observation equations
% w.r.t. parameter-to-be-estimated
psi = zeros((numOfParam+numOfKalman), numOfObserved);
% - gradient of innovations
S = zeros(numOfObserved);
L = zeros((numOfParam+numOfKalman), numOfObserved);
V = zeros(numOfParam+numOfKalman);
%RPE algorithm for time-invariant parameter estimation
H = eye(numOfObserved, numOfState);
yHat = H * xHat;
Epsn = y - yHat;
psi = (H * W + D)';
S = psi' * PPre * psi + lmbd;
L = PPre * psi * inv(S);
V = eye(numOfParam+numOfKalman) - L * psi';
%Returning results
AlfaHatCt = alfaHatPre + L * Epsn;
PCt = (V * PPre * V') + L * lmbd * L';
%end-of-function
function [FMat, GMat] = FGMatrix(tvp);
%FGMATRIX Construction of F and G matrices for GRW models
firstElmt = tvp(1);
if mod(firstElmt, 1) > 0;
    F = [firstElmt, ceil(firstElmt); 0, ceil(firstElmt)];
else
    F = [1, firstElmt; 0, firstElmt];
end
G = [1-ceil(firstElmt), 0; 0, ceil(firstElmt)];
for i = 2:length(tvp);
elmt = tvp(i);
if mod(elmt, 1) > 0;
    F = blkdiag(F, [elmt, ceil(elmt); 0, ceil(elmt)]);
else
    F = blkdiag(F, [1, elmt; 0, elmt]);
end
G = blkdiag(G, [1-ceil(elmt), 0; 0, ceil(elmt)]);
end
FMat = F;
GMat = G;
%end-of-function
function O = Duplicate(A, B);
%DUPLICATE Both A and B must be either ROW or COLUMN vectors
if nargin < 2;
    B = A;
end;
if (size(A, 1) ~= 1) & (size(A, 2) ~= 1)
    error('A and B must be ROW or COLUMN vectors.');
end;
if size(A) == size(B)
    if size(A, 1) > size(A, 2)
        C = [A, B];
        D = reshape(C', [numel(C), 1]);
    else
        C = [A; B];
        D = reshape(C, [numel(C), 1]);
    end
else
    error('Matrix dimensions must agree.');
end
O = D;
%end-of-function
function R = Lambda(x);
%LAMBDA Calculation of measurement noises
r1 = diff(x);
if r2 = sum(abs(r1));
    r3 = r2 ./ (length(x)-1);
    R = (0.8865 * r3) .^2;
%end-of-function
function [Num, Name, Col] = InitialFile(file);
%INITIALFILE Information about system, model and data.
[dimen, files, cols, ivSt, ivPar, ivVar, ivQ] = textread(file, ...
    'emptyvalue', 0, 'delimiter', ',');
%Number of state variables
numOfState = dimen(2);
%Number of the parameters to be recursively estimated
%(excluding fixed parameters)
numOfParameter = dimen(3);
%Number of observed state variables
numOfObserved = dimen(4);
%Name of model file
nameOfModelFile = files{2};
%Name of data file (Excel file)
nameOfDataFile = files{3};
%Name of the sheet of the data file
nameOfSheet = files{4};
%Specification of columns in the data file
colOfTime = cols{2};
colOfSample = cols{3};
colOfInput = cols{4};
colOfOutput = cols{5};
colOfEnvironFactor = cols{6};
%Returning results
Num = [numOfState; numOfParameter; numOfObserved];
Name = {nameOfModelFile; nameOfDataFile; nameOfSheet};
Col = {colOfTime; colOfSample; colOfInput; colOfOutput; ... 
        colOfEnvironFactor};
%end-of-function
function [IState, IAlfa, IAlfaVar, IQ] = InitialValue(dim, file);
%INITIALVALUE Initial values for state variables, parameter, parameter
%variance-covariance matrix, Kalman-like gain matrix, and Q matrix
[dimen, files, cols, ivSt, ivPar, ivVar, ivQ] = textread(file, ...
    '%*s%d%*s%s%*s%s%*s%n%*s%n%*s%n%*s%[^\n]' , ...
    'emptyvalue', 0, 'delimiter', ',');
%Extracting elements from dim
numOfState = dim(1);
numOfParameter = dim(2);
numOfObserved = dim(3);
numOfKalman = numOfState * numOfObserved;
%Initial values for parameters-to-be-estimated
IParam = ivPar(2:3);
IKalman = repmat(ivPar(4), numOfKalman, 1);
%Initial values for diagonal elements of covariance matrix
%of parameter-to-be-estimated
IParParamVar = ivVar(2:3);
IKalmanVar = repmat(ivVar(4), numOfKalman, 1);
%Returning results
IState = ivSt(2:3);
IAlfa = [IParam; IKalman];
IAlfaVar = [IParParamVar; IKalmanVar];
IQ = [ivQ(2); ivQ(3); repmat(ivQ(4), numOfKalman, 1)];
%end-of-function
%end-of-file