8-2005

Application of Statistical Methods and Process Models for the Design and Analysis of Activated Sludge Wastewater Treatment Plants (WWTPs)

Jinsheng Huo
University of Tennessee - Knoxville

To the Graduate Council:

I am submitting herewith a dissertation written by Jinsheng Huo entitled "Application of Statistical Methods and Process Models for the Design and Analysis of Activated Sludge Wastewater Treatment Plants (WWTPs)." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Civil Engineering.

Chris D. Cox, Major Professor

We have read this dissertation and recommend its acceptance:

R. Bruce Robinson, William L. Seaver, Kung-Hui Chu

Accepted for the Council:

Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

To the Graduate Council:

I am submitting herewith a dissertation written by Jinsheng Huo entitled "Application of Statistical Methods and Process Models for the Design and Analysis of Activated Sludge Wastewater Treatment Plants (WWTPs)." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Civil Engineering.

Chris D. Cox

Major Professor

We have read this dissertation
and recommend its acceptance:

R. Bruce Robinson

William L. Seaver

Kung-Hui Chu

Accepted for the Council:

Anne Mayhew

Vice Chancellor and Dean of
Graduate Studies

(Original signatures are on file with official student records.)

Application of Statistical Methods and Process Models for the Design and Analysis of Activated Sludge Wastewater Treatment Plants (WWTPs)

A Dissertation

Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Jinsheng Huo

August 2005

DEDICATION

This dissertation is dedicated to my lovely wife Yan Jiang (Jenny) and our wonderful parents (Gangfu Huo, Xuexiu Zhao, Xianzhi Jiang and Aifen Liu) on both sides, the rest of our family and our friends, for always believing in me, inspiring me, and encouraging me to reach higher in order to achieve my goals**.**

ACKNOWLEDGMENTS

I wish to thank Dr. Chris D. Cox for his guidance and support through the development and writing of this dissertation. He spent countless time with me through this process and was very patient and understanding. It is him who inspires me and encourages me to get closer to my career goal step by step.

Many thanks are also extended to my committee members Dr. R. Bruce Robinson, Dr. William L. Seaver, and Dr. Kung-Hui Chu for their time spent with me on this dissertation.

Special thanks to Mr. Ken Glass from the Oak Ridge WWTP, and Mr. Tom Heikkinen from the Seneca WWTP, who helped me a lot collect plant data and other information. I also would like to thank Mark McKinney for his help on the WERF project.

Finally, thanks to my wonderful wife Yan Jiang (Jenny) for her understanding and strong support all through my study at the University of Tennessee, Knoxville.

ABSTRACT

The purpose of this study is to investigate statistical procedures to qualify uncertainty, and explicitly evaluate its impact on wastewater treatment plants (WWTPs). The goal is to develop a statistical-based procedure to design WWTPs that provide reliable protection of water quality, instead of making overly conservative assumptions and adopting empirical safety factors. An innovative Monte Carlo based procedure was developed to quantify the risk of violating effluent as a function of various design decisions. A simulation program called StatASPS was developed to conduct Monte Carlo simulations combined with the ASM1 model.

A random influent generator was developed to describe the statistical characteristics of the influent components of WWTPs. Prior to modeling, a two-directional exponential smoothing (TES) method was developed to replace those non-randomly missing data during weekends and holidays. The best models were selected based on various statistics and the ability to forecast future values. The time series models were then used to generate random influent variables with the same statistical characteristics as the original data.

The best Monte Carlo simulations were conducted using historical influent data and site-specific parameter distributions, according to the applications to both the Oak Ridge and Seneca WWTPs. This indicates that parameter uncertainty was more effective in predicting uncertainty in plant performance than influent variability. The ultimate

simulations were conducted using one-month's influent data, considering limitations of computing technologies. Application of the method to the two plants demonstrated that this method provided a reliable and reasonable estimate of the uncertainty of plant performance. The best predictions of plant uncertainty were obtained by determining the distribution for the most sensitive parameter and holding all other model parameters constant.

The StatASPS procedure proved to be a reliable and reasonable method to design cost-effective WWTPs. With further development, this procedure could provide engineers and regulators with a high degree of confidence that the plant will perform as required, without resorting to overly conservative assumptions or large safety factors.

TABLE OF CONTENTS

LIST OF TABLES

TABLE **PAGE**

LIST OF FIGURES

FIGURE **PAGE**

NOMENCLATURE

Symbols

Symbol	Definition, Unit	Symbol	Definition, Unit
X_I	Inert particulate organic matter, mg/L as COD	$\mu_{H,max}$	Specific growth rate of Heterotrophs, hr^{-1}
X_S	Slowly biodegradable substrate, mg/L as COD	K_S	Half-saturation coefficient, mg / L as COD
X_{BH}	Active heterotrophic biomass, mg/L as COD	Y_H	Yield of Heterotrophs, mg biomass formed/mg removed
X_{BA}	Active autotrophic biomass, mg/L as COD	b_{LH}	Decay coefficient of Heterotrophs, hr^{-1}
X_D	Debris from biomass death and lysis, mg/L	$\mu_{A,max}$	Specific growth rate of Autotrophs, hr^{-1}
S_I	Inert soluble organic matter, mg/L as COD	K_{NH}	Half-saturation coefficient for Nitrifiers, mg/L as N
S_S	Readily biodegradable substrate, mg/L as COD	Y_A	Yield of Autotrophs, mg biomass formed /mg N oxidized
S_O	Oxygen, mg/L as COD	b_{LA}	Decay coefficient of Autotrophs, hr^{-1}
S_{NO}	Nitrate Nitrogen, mg/L as N	k_h	Hydrolysis coefficient, mg COD/(mg biomass COD.hr)
S_{NH}	Ammonia nitrogen, mg/L as N	K_x	Switch coefficient for hydrolysis, mg / mg biomass
S_{NS}	Soluble biodegradable organic nitrogen, mg/L	η_H	Correct coefficient for anoxic hydrolysis, Dimensionless
X_{NS}	Particulate biodegradable nitrogen, mg/L as N	K_{OH}	Half-saturation coefficient for Heterotrophs, mg / L as O_2
S_{ALK}	Alkalinity, mole HCO^-_3/L	K_{OA}	Half-saturation coefficient for Autotrops, mg/L as O_2
TKN	Total Kjeldahl nitrogen, mg/L as N	K_{NO}	Half-saturation coefficient for Denitrifiers, mg/L as N
TCOD	Total COD, mg/L as COD	η_g	Correct coefficient for anoxic growth, Dimensionless
TSS	Total suspended solids, mg/L		

Abbreviations

ACF	Autocorrelation Function
ANO method	The missing values replacement method with the average of the nearest observation in the future and/or in the past.
ARIMA	AutoRegressive Integrated Moving Average
ASM1	Activated Sludge Model No.1
ASM2	Activated Sludge Model No.2
ASM3	Activated Sludge Model No.3
ASPS	Activated Sludge Process Simulator
Backward ES	Exponential Smoothing Method in a Backward Direction
BioCOD	Biodegradable Chemical Oxygen Demand
BOD_5	5-day Biological Oxygen Demand
BOD_u	Ultimate Biological Oxygen Demand
CAS	Conventional Activated Sludge
CDF	Cumulative Distribution Function
COD	Chemical Oxygen Demand
CSTR	Continuous-flow Stirred (completely mixed) Tank Reactor
CWA	Clean Water Act

DO	Dissolved Oxygen
DR	Dynamic Regression
EAAS	Extended Aeration Activated Sludge
EPA	Environmental Protection Agency (USA)
ES	Exponential Smooth
Forward ES	Exponential Smoothing Method in a Forward Direction
GoF	Goodness-of-Fit
GPD	US Gallons per Day
HRT	Hydraulic Retention Time
IWA	International Water Association (former IAWQ and IAWPRC)
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MC	Monte Carlo
MCAR	Missing Completely at Random
MGD	Million Gallon per Day
MLSS	Mixed Liquor Suspended Solids
MSE	Mean Square Error
NH_4-N	Ammonia Nitrogen
OUR	Oxygen Uptake Rate
PACF	Partial Autocorrelation Function
PDF	Probability Density Function
RAS	Return Activated Sludge
RMSE	Root Mean Square Error
SPD	Site-specific Parameter Distributions
SRT	Solids (Sludge) Retention Time
SS	Suspended Solids
StatASPS	Statistical Activated Sludge Process Simulator
Std Dev	Standard Deviation
SVI	Sludge Volume Index
TCOD	Total COD
TES	Two-directional Exponential Smoothing
TESWN	Two-directional Exponential Smoothing method with White Noise added
TKN	Total Kjeldahl Nitrogen (organic + ammonia nitrogen)
TN	Total Nitrogen
TON	Total Oxidized Nitrogen (NO^-_3-N + NO^-_2-N)
TSS	Total Suspended Solids
UPD	Universal Parameter Distributions
VSS	Volatile Suspended Solids
WEF	Water Environment Federation
WERF	Water Environment Research Foundation
WWTP	Wastewater Treatment Plant

CHAPTER I. INTRODUCTION

1.1 Purposes of the Study

The performance of wastewater treatment plants (WWTPs) is subject to uncertainties, including influent loads, design and operational parameters, and receiving water conditions, etc. Environmental engineers and regulators have to deal with this uncertainty at every stage of the design and permitting processes. Typically, when engineers and regulators lack enough information to deal with the uncertainty, they make overly-conservative assumptions and adopt large safety factors to ensure that water quality standards are met. This often leads to the WWTP being rated below its actual capacity. This conclusion is supported by the opinions of several consultants and practitioners, the frequent practice in recent years of re-rating activated sludge plants for greater capacity without significant facility improvement, and WERF (Water Environment Research Foundation) requests for project proposals in this area. Over-conservative designs and permitting requirements can be ill afforded in the wastewater treatment industry. Rather, cost effective plants that provide reliable protection of water quality are in high demand.

In this dissertation, the Activated Sludge Model No.1 (ASM1) and the Monte Carlo (MC) method are combined to assess and analyze the design and operation of WWTPs. The general goal of this dissertation is to develop design procedures that explicitly consider statistical uncertainty and variability, thereby minimizing the need for empirical safety

factors and providing quantitative statistical estimates of the plant performance. Our hypothesis is that statistical-based simulation methods can be used to describe the uncertainties in the plant performance of WWTPs. The developed methodology will offer the opportunity for the designer and regulators to meet and discuss the trade-off of risk and cost.

1.2 Objectives of the Study

Engineers continually face uncertainty in process model parameters and variability of plant influent during the design of wastewater treatment plants. Uncertainty means that the factors affecting the design and performance of the plant are not known precisely to the engineers. A common example of uncertainty is that usually only typical values or ranges of the various kinetic parameters used in the design models for activated sludge plants are known. Variability means that a factor influencing the design and performance varies over time. Examples of variability include the time-dependent nature of the influent load and receiving water conditions. Thus, environmental engineers and regulators must deal with this uncertainty at every stage of the design and permitting processes, most often by making conservative assumptions and adopting large safety factors to ensure adequate protection of water quality. Potentially, these design practices can result in a treatment plant that has a capacity far in excess of that required.

The development of more sophisticated means of dealing with uncertainty and variability is facilitated by the current availability of sophisticated models and large amounts of data that describe the performance of wastewater plants. In this dissertation, we explore the uncertainty in activated sludge plant performance that results from both uncertainty in the model parameters and variability in the plant influent that describe the process. Our goal is to develop a statistical-based design procedure to design more cost-effective WWTPs and more efficiently protect the water environment and quality.

Monte Carlo simulations combined with process models will be applied to quantify and analyze the uncertainty in WWTPs in this dissertation. The uncertainty of process model parameters will be described with two possible options: fixed calibrated parameter and parameter distributions (universal or site-specific). The variability of plant influent will be described with three options: historical data, predicted data from time series models, and randomly generated influent data. Base on the simulation results, we then can develop a statistical-based procedure to design more cost-effective WWTPs.

There are five specific objectives in this study:

- To determine the critical process model parameters that have the largest influence on the design and operation of wastewater treatment plants (WWTPs) using Monte Carlo simulation and sensitivity analysis.

- To generate the fixed calibrated parameters of WWTPs using a calibration procedure that leverages the previous sensitivity analysis results, and then generate the site-specific parameter distributions from the calibrated model parameters.
- To develop random influent generators for the influent of WWTPs based on the fitted time series models. The randomly generated influent data will have similar statistical characteristics as historical data but give different time histories. Those influent components include flow rate, five-day biological oxygen demand (BOD_5), suspended solids (SS) and ammonia nitrogen (NH_4-N). A missing data replacement method will also be developed to replace non-randomly missing values in time series data.
- To conduct Monte Carlo simulations combined with process models with uncertainty in process model parameters and variability in plant influent to determine the statistical characteristics (for example, mean, standard deviation, and percentiles) of the simulated effluents. A comparison of the simulated and measured effluent distributions will be used to validate the statistical-based design proposal.
- To determine design factors that result in high-reliability and low-cost plant performance. These main design factors, such as sludge residence time (SRT), bioreactor volume and oxygen supply, needed to maintain a certain percent (for example, 99.50 percent) of the effluent (or several days average effluent) of WWTPs below the effluent standards, will be determined. Those factors will be compared with those values commonly used in the traditional design

procedure. These determined design factors will be applied to assist environmental regulators in evaluating the ability of various plant designs to meet specific treatment requirements and the trade-offs between cost and risk.

CHAPTER II. REVIEW OF LITERATURE

2.1 Resources of the Literature Review

The literature review for this study has focused on four key areas:

- Applications of Activated Sludge Models (ASMs)
- Measurement and calibration of ASMs parameters
- Historical perspectives of Monte Carlo simulation of wastewater treatment processes
- Reviews of related statistical methods (Bayesian method, Time Series Models, Monte Carlo Method, Sensitivity Analysis, and Uncertainty Analysis)

2.2 Activated Sludge Process Models

2.2.1 Introduction

Wastewater treatment processes are commonly divided into two categories: suspended growth systems (for example, activated sludge systems) and attached growth systems (for example, biofilm systems) (Grady et al., 1999). The characteristic of the activated sludge process is that the biomass is suspended in the wastewater, consuming and degrading the organic pollutants. The activated sludge process has been developed over a period of

almost 90 years, both in theory and practice. Most large WWTPs are using activated sludge processes. It is also the predominant biological treatment featured in current research applications. Basically, the activated sludge process includes an aeration tank, clarifier, biomass return, and waste biomass disposal (Grady et al., 1999). In general, the activated sludge process is a continuous or semi-continuous aerobic and/or anoxic method for biological wastewater treatment, including carbonaceous oxidation, nitrification, denitrification and/or phosphate removal. Usually, the separation of the active biomass from the treated wastewater is performed by settling in clarifiers but may also be done by other methods, including flotation and membrane filtration.

Initially, the design and operation of WWTPs were mostly based on the empirical rules of thumb. Since the 1950's, many researchers and engineers have applied theories of reactor design and microorganism growth to wastewater treatment systems, making it possible to describe substrate degradation, microorganism growth, and plant performance in terms of mathematical models. In particular, the Eckenfelder (1955) and Lawrence-McCarty (1971) activated sludge models gained widespread practical application due to their ability to predict the plant performance. The Eckenfelder model, which was created by Professor Eckenfelder in 1955, is an experimental model that considers the relationship of WWTP loading and effluent quality based on research on the microorganism growth in a Sequencing Batch Reactor. A.W. Lawrence and P.L. McCarty (1971) developed the Lawrence-McCarty model, introducing Monod's formulation to the wastewater treatment field. The model shows that the effluent concentration is independent of the influent concentration. Furthermore, Solid Residence Time (SRT) is independent of Hydraulic

Residence Time (HRT) and is the most important parameter to the design and operation of WWTPs. SRT is the residence time of biomass that significantly affects the characteristics of the biomass in biological wastewater treatment systems.

Based on the models above, many other advanced models have been created. Among these models, the Activated Sludge Model No.1 (ASM1) is one of the most important models, created by the International Water Association (IWA, former IAWPRC and IAWQ) in 1986. It took many international experts in wastewater treatment more than three years to develop the matrix-format based Activated Sludge Model No.1 (ASM1) (Henze et al., 1986). Since it was developed, it has been well and widely applied by practicing engineers (Spanjers et al., 1995; Huo, 2001). The Activated Sludge Model No.2 (ASM2) was introduced to include the phosphorus and nitrogen removal (Henze et al., 1995). In 1999, the Activated Sludge Model No.3 (ASM3) revised some processes according to the new finds in the field of activated sludge processes (IWA, 2000). The ASM3 models 13 wastewater components and 12 biological processes. Compared with ASM1, ASM3 adds a new process: the storage of organic substrates and the lysis (decay) process is renamed as an endogenous respiration process. Furthermore, ASM3 can predict oxygen consumption, sludge production, nitrification and denitrification of activated sludge systems. However, considering the complexity of ASM2 and scarcity of applications of ASM3, only ASM1 is used in this project to evaluate the statistical characteristics of activated sludge WWTPs. Software that implements various Activated Sludge Models (ASMs) is shown in Table 2-1.

Table 2-1. Simulation software packages of activated sludge processes.

Package Name	**Features**	**Company or University**
GPS-X	ASM1, ASM2d, ASM3 including other unit process models.	Hydromantis, Inc., 1685 Main St. West, Suite 302, Hamilton, Ontario L8S 1G5 Canada
SSSP	ASM No.1 only.	C.P. Leslie Grady Jr., Environmental System Engineering, Rich Environmental Research Lab, Clemson University, Clemson, SC 29634-0919 USA
EFFOR	ASM No.1 and clarifier simulation.	Jan Peterson, I. Kruger AS, Gadsaxevej 363, DK-2860, Soborg, Denmark
BioWin	Include the ASM1, ASM2d, ASM3, and other unit operations.	EnviroSim Associates LTD. 7 Innovation Drive Suite 205 Flamborough, Ontario L9H 7H9, Canada

From Table 2-1, we can see there are many available implementations of the various ASMs. Among them, GPS-X and BioWin are most commonly used by practitioners. For example, the Hydromantis website (http://www.hydromantis.com/expertise02.html) discusses the application of GPS-X to the upgrade, design and assessment of four different treatment plants: 1) Queenston WPCP upgrade stage I (Regional Municipality of Niagara), 2) Niagara Falls WPCP upgrade – phase II detailed design, 3) Wheatly WPCP capacity assessment, and 4) Facility assessment of the Port Weller WPCP.

Daigger and Nolasco (1995) applied the steady-state ASM1 with and without Dold's excess biological phosphorus removal module to thirteen full-scale WWTPs. ASM1 proved to accurately predict full-scale plants performance and trends in performance even with the default model parameters. However, they also demonstrated that ASM1 cannot accurately describe the effects of the oxygen transfer system on the plant performance. In this paper, adjustments of half-saturation coefficients for oxygen and nitrate utilization were applied to obtain good simulation results. Kolisch and Londong (1998) also investigated the possibility of calculating the operational behavior of the WWTPs online by coupling the dynamic simulation with the online measurement of the process. The ASM1 based standard model SIMBA was developed in MATLAB® and SIMULINK®. This study showed that the dynamic simulation could be connected to the process system of WWTPs based on the continuously produced online data with average values in 15-minute cycles. According to the work performed, they found that the on-site practical problems were as important as the establishment and calibration of the model. Daigger and Barker (2000) studied methods to reduce the HRT and SRT, and corresponding

facility costs and space requirement. This study was based on ASMs simulations of biological wastewater treatment processes. Two important wastewater characteristics that significantly affect nitrification processes were identified: biological inert suspended solids and nitrifier maximum specific growth rate. The maximum specific growth rate, for a fixed temperature, varied by a factor of ±2. This variability was caused by several factors including the composition of the wastewater. Due to this variation of maximum specific growth rates, some commercial models (for example, BioWin) used the lower rates as default values, which results in overdesign of the processes. The ASMs simulation also showed that the use of an aerobic bioreactor with significant plug flow character could significantly reduce the required SRT. The plug flow character means either tanks with high length to width ratios, or tanks in series.

Cinar et al. (1998) evaluated the application of ASM2 using steady state data from four full-scale WWTPs. The model was calibrated with the plant data from the Mauldin Road WWTP. Only six model parameters were changed from the suggested default values. The determined model parameters of the Mauldin Road WWTP were also successfully confirmed with the data set from the Lower Reedy WWTP. However, the other two plants, the Durbin Creek and Gilder Creek WWTPs, could not be fully simulated by ASM2 because both are oxidation ditch processes. The researchers concluded that the inability of the model to describe the data was related to inaccuracies in describing the flow pattern in the oxidation ditch processes. Furthermore, this paper illustrated that precise ASM2 simulations required accurate DO concentrations because the simulations were very sensitive to the DO concentrations. Nolasco et al. (1998) combined process

modeling (ASM2) and pilot-plant techniques to identify and quantify the factors affecting the performance of the Step Bio-P process. Both on-line and off-line data were used to recalibrate the model ASM2, which was used as an experimental design and process optimization tool. The pilot plant performance data matched and verified the initial mathematical model results, which proved that ASM2 was an excellent tool for initially evaluating unique biological nutrient removal (BNR) processes and for designing more effective pilot testing programs.

Some important drawbacks observed after almost two decade's application experience of ASM1 (Gujer et al. 1999) are summarized as follows:

- ASM1 cannot deal with nitrogen and alkalinity limitations of heterotrophic biomass. This limitation can cause the computer code to predict negative concentrations (for example, ammonium) under some situations.
- ASM1 does not include the storage of poly-hydroxy-alkanoates and glycogen. This storage has been observed under aerobic and anoxic conditions in activated sludge processes in cases in which elevated readily biodegradable organic substrates are available.
- The hydrolysis process significantly affects the oxygen consumption and denitrification by heterotrophic biomass. Furthermore, the kinetic parameter for the hydrolysis process is difficult to measure.
- ASM1 distinguishes inert particulate organic matter based on their sources, i.e., influent or biomass decay. However, in reality, it is impossible to differentiate those two fractions via measurement.

To address these limitations, the IWA task group proposed activated sludge model No.3 (ASM3). ASM3, which is related to ASM1, can predict oxygen consumption, sludge production, nitrification and denitrification of activated sludge systems. However, ASM3 has not yet been widely tested against experimental or plant data. Thus, future improvement of the model structure might still be needed. The authors expected that ASM1 and ASM3 might well prove to be equivalent after improvements of both models (Gujer et al., 1999).

Mussati et al. (2002) compared ASM1 and ASM3 based on the simulations of the COST benchmark WWTP. The analysis of the process behavior to pulse and step disturbances showed that the ASM3 usually required a longer time to reach steady state than the ASM1 model. However, ASM3 simulation results are easier to interpret because of its more transparent model structure. Koch et al. (2000) illustrated that the ASM3 model better represents experimental data than ASM1 in certain situations. These situations include 1) the cases in which most of the incoming readily biodegradable substrate is being stored, 2) high COD loads in WWTPs caused by diurnal influent flow rate and COD variations, or 3) substantial non-aerated zones in WWTPs.

There are three important reasons why ASM1 is chosen instead of ASM3. Firstly, ASM1 has been widely applied in activated sludge systems for almost two decades. There are also paramount publications related to the application of ASM1, for example, developments of software packages (GPS-X and BioWin), experimental measurements of

model parameters, and calibration and application of ASM1 to the pilot-scale or full-scale WWTPs. Secondly, ASM1 and ASM3 are expected to be equivalent. In fact, the beginning experience of ASM3 might be highly based on the comparison to ASM1 and real plant data. Thirdly, because this dissertation mostly focuses on the uncertainty of activated sludge systems, it requires numerous published data to support our analysis of the simulated results. Thus, ASM1 was chosen as the primary model in this study.

2.2.2 Activated Sludge Model No.1 (ASM1)

With almost twenty years of practical application and further progress, ASM1 has become a very important reference model for further research work and is also the basis for many software packages (GPS-X, BioWin, etc.) applied in the design and operation of WWTPs (Gujer et al., 1995). In this section, the presentation of ASM1 and parameters in ASM1 are reviewed.

2.2.3 Presentation of ASM1

An important feature of ASM1 is introducing the matrix format to present the 8 processes and 13 components. As can be seen in Table 2-2 shown below, it is relatively easier to read and understand the processes and related components in the biological wastewater

Table 2-2. Activated sludge model No.1 (ASM1) in matrix format.
(After IAWPRC ASM1 report, 1986 and Grady et al., 1999)

Component, i → / Process, j ↓		1 X_I	2 X_S	3 $X_{B,H}$	4 $X_{B,A}$	5 X_D	6 S_I	7 S_S	8 S_O	9 S_{NO}	10 S_{NH}	11 S_{NS}	12 X_{NS}	13 S_{ALK}	Process rate, r_j
1	Aerobic growth of heterotrophs			1				$-1/Y_H$	$\frac{1-Y_H}{Y_H}$		$-i_{N/XB}$			$-i_{N/XB}/14$	$\hat{\mu}_H(\frac{S_S}{K_S+S_S})(\frac{S_O}{K_{O,H}+S_O})X_{B,H}$
2	Anoxic growth of heterotrophs			1				$-1/Y_H$		$-\frac{1-Y_H}{2.86Y_H}$	$-i_{N/XB}$			$\frac{1-Y_H}{14(2.86Y_H)}-i_{N/XB}/14$	$\hat{\mu}_H(\frac{S_S}{K_S+S_S})(\frac{K_{O,H}}{K_{O,H}+S_O})(\frac{S_{NO}}{K_{NO}+S_{NO}})\eta_g X_{B,H}$
3	Aerobic growth of autotrophs				1				$\frac{4.57-Y_H}{Y_H}$	$1/Y_A$	$-i_{N/XB}-1/Y_A$			$-i_{N/XB}-\frac{1}{7Y_A}$	$\hat{\mu}_A(\frac{S_{NH}}{K_{NH}+S_{NH}})(\frac{S_O}{K_{O,A}+S_O})X_{B,A}$
4	Death and lysis of heterotrophs		1-f'$_D$	-1		f'$_D$							$i_{N/XB}$-f'$_D$×$i_{N/XD}$		$b_{L,H}X_{B,H}$
5	Death and lysis of autotrophs		1-f'$_D$		-1	f'$_D$							$i_{N/XB}$-f'$_D$×$i_{N/XD}$		$b_{L,A}X_{B,A}$
6	Ammonification of soluble organic nitrogen										1	-1		1/14	$k_a S_{NS}X_{B,H}$
7	"Hydrolysis" of particulate organics		-1					1							$k_h\frac{X_S/X_{B,H}}{K_X+X_S/X_{B,H}}[(\frac{S_O}{K_{O,H}+S_O})+\eta_h(\frac{K_{OH}}{K_{O,H}+S_O})]X_{B,H}$
8	"Hydrolysis" of particulate organic nitrogen											1	-1		$r_7(X_{NS}/X_S)$
	Component rate, r_i								$r_i=\sum_{j=1}^{n}\psi_{ij}r_j$						

Note:

1. All organic components (1-7) and oxygen (8) are expressed as COD; all nitrogenous components (9-12) are expected as nitrogen.
2. Coefficients must be multiplied by –1 to express them as oxygen.

Example: the reaction rate of the second component X_S is expressed as:

$$r_2=\sum_{j=1}^{8}\psi_{2j}r_j=(1-f'_D)b_{L,H}X_{BH}+(1-f'_D)b_{L,A}X_{BA}-k_h\frac{X_S/X_{B,H}}{K_X+X_S/X_{B,H}}[(\frac{S_O}{K_{O,H}+S_O})+\eta_h(\frac{K_{OH}}{K_{O,H}+S_O})]X_{B,H}$$

treatment system. Furthermore, it is simple and direct to identify the rate of each component and calculate the overall reaction rate of each component.

It should be noted that there are three different units in Table 2-2, that is, components 1-8 are expressed in chemical oxygen demand (COD) units, whereas components 9-12 are given as nitrogen, and component 13, alkalinity, is in molar units. The 13 components can be classified into five categories: biomass, substrates, inert components, debris, and dissolved oxygen, which are listed across the top of Table 2-2 by symbol. In conformity with IAWPRC nomenclature (Grau et al., 1982), soluble components are characterized by S and particulate components by X. Subscripts are used to specify individual components: B for biomass, S for substrate, I for inert component, D for debris, and O for oxygen. Furthermore, the subscripts after the comma denote the different forms of biomass. For example, H represents heterotrophic biomass, while A represents autotrophic biomass. The index i is assigned to each component, and index j is assigned to each process. The particulate components are assumed to be associated with the activated sludge (flocculated onto the activated sludge or contained within the active biomass) within the activated sludge systems. The thirteen components listed in Table 2-2 are defined in detail in Table 2-3.

As shown in Table 2-2, the leftmost column lists the 8 fundamental processes, which can be basically classified as four categories: growth of biomass, decay of biomass, ammonification of organic nitrogen, and ‘hydrolysis’ of particulate organic, and the rightmost column lists their process reaction rate expressions.

Table 2-3. Definitions of component symbols in Table 2-2 of the ASM1 model.

Component No.	Component Symbol	Definition and Units
1	X_I	Inert particulate organic matter, mg/L as COD. This material cannot be degraded in ASM1; instead, it is always flocculated onto the activated sludge.
2	X_S	Slowly biodegradable substrate, mg/L as COD. Slowly biodegradable substrates are high molecular weight, soluble, colloidal and particulate organic substrates that must be transformed to readily degradable substrate by external hydrolysis (details in lysis:regrowth approach) before they are available to biomass. In the lysis:regrowth approach, the hydrolysis product of X_S is assumed to be readily biodegradable substrate (S_S) only.
3	X_{BH}	Active heterotrophic biomass, mg/L as COD. These organisms may be subjected to either aerobic or anoxic condition depending on what oxidatants are available. In the later case, readily biodegradable substrate serves as the terminal electron acceptor and the nitrate nitrogen serves as terminal electron acceptor.
4	X_{BA}	Active autotrophic biomass, mg/L as COD. These organisms are obligate aerobic and chemo-litho-autotrophic biomass, which are responsible for nitrification. Nitrite is treated as an intermediate compound of nitrification; therefore, it is assumed that ammonia nitrogen (S_{NH}) is directly nitrified to nitrate nitrogen (S_{NO}) by nitrifiers.
5	X_D	Debris from biomass death and lysis, mg/L as COD. This material has similar properties with X_I. The only difference is their origin: X_D is from biomass death and lysis, while X_I is directly from the influent.
6	S_I	Inert soluble organic matter, mg/L as COD. The main feature of S_I is that it cannot be further degraded in the WWTP. This material can readily be estimated from the residual soluble COD in the effluent of activated sludge systems under low organic load conditions (ASM1 report, 1987).
7	S_S	Readily biodegradable substrate, mg/L as COD. This fraction of substrate can be directly consumed by heterotrophic biomass. In practice, S_S is determined by respiration test (bioassay).

Table 2-3. Continued.

Component No.	Component Symbol	Definition and Units
8	S_O	Oxygen, mg/L as COD. S_O is the dissolved oxygen concentration, which can be directly measured in CSTRs. This value is taken as a constant to represent the aerobic or anoxic condition if we assume the oxygen input rate and oxygen consumption rate are equal. The rate of oxygen demand during biomass growth can be calculated from Table 2-2.
9	S_{NO}	Nitrate Nitrogen, mg/L as N. Since nitrite is not included in ASM1, S_{NO} represents both nitrate and nitrite nitrogen. However, S_{NO} is considered to be NO^-_3-N only for all stoichiometric computations (ThOD conservation), where S_{NO} is introduced as negative ThOD.
10	S_{NH}	Ammonia nitrogen, mg/L as N. S_{NH} includes ammonium plus ammonia nitrogen (NH^+_4-N + NH_3-N). However, S_{NH} is assumed to be all NH^+_4 for the balance of ionic charges. It should also be noted that S_{NH} does not have ThOD.
11	S_{NS}	Soluble biodegradable organic nitrogen, mg/L as N. This material is generated through the hydrolysis of particulate organic nitrogen and converted to ammonia nitrogen through ammonification.
12	X_{NS}	Particulate biodegradable organic nitrogen, mg/L as N. This material is formed from the death and lysis of both heterotrophic and autotrophic biomass and consumed by ammonification. It should be noted that X_{NS} should not be added to obtain the total concentration of the particulate components because it is a subset of these materials and has already been included in their concentrations.
13	S_{ALK}	Alkalinity, mM/L. In ASM1, the pH is constant and near neutrality. As we know, the pH influences many parameters in the model. However, little information is available to express the influences. Thus, in order to provide early detection of the possible changes in the pH that might greatly affect some biological processes, alkalinity is introduced to approximate the conservation of ionic charge in biological reactions. For all stoichiometric computation, S_{ALK} is assumed to be bicarbonate, HCO^-_3, only.

Process 1: Aerobic growth of heterotrophs: This process shows that the soluble substrate is degraded, resulting in the growth of heterotrophic biomass (X_{BH}). Furthermore, ammonia nitrogen will also be removed and incorporated into cell mass. In this case, reactant 1 is readily biodegradable substrate (S_S) and reactant 2 is dissolved oxygen (S_O). Note aerobic growth of heterotrophs reverts to the Monod equation if we removed the reactant 2. The primary purpose of the oxygen term is to work as a switch function that turns off aerobic growth at low DO concentration to allow anoxic growth to begin if nitrate is present, as shown in rate term of processes 1 and 2 in Table 2-2.

Process 2: Anoxic growth of heterotrophs: Compared with aerobic growth, where readily biodegradable substrate serves as the terminal electron acceptor, the nitrate nitrogen serves as terminal electron acceptor under anoxic conditions. The process rate item of process 2 only adds two new items; one is a switch function of S_{NO}, which presents the effect of nitrate nitrogen; the other is the empirical coefficient, η_g, which represents the difference between anoxic growth and aerobic growth. It should be noted that both heterotrophic rates should be zero under absolutely anaerobic conditions, that is, without both oxygen and nitrate.

Process 3: Aerobic growths of autotrophs: In this process, the growth of autotrophs (nitrifiers) occurs with ammonia nitrogen as the energy source and nitrate nitrogen as the end product. In addition, some ammonia is incorporated into the biomass. In the rate expression, there are two reactants: ammonia-N, and dissolved oxygen. There are two important things about the way the process is modeled that should be noted (Grady, C.P.,

Daigger, G.T., and Lim H.C., 1999). The first is that nitrification is considered a one-step reaction, that is, from ammonia-N to nitrate-N directly, without nitrite as an intermediate. This simplification is reasonable because the kinetic parameters for Nitrosomonas and Nitrobacter are similar and nitrite is consumed as fast as it is formed under balanced growth conditions, with the result that its concentration is usually very low and of little importance. However, it should be noted that the simplification is only good for bioreactors at steady state or those with relatively mild variations in dynamic load. The second important thing is that substrate and product inhibition is not considered because adequate kinetic relationships are not available. Thus, ASM No.1 is not appropriate to situations with excess high nitrogen concentration.

Figure 2-1 illustrates three important aspects of biomass decay: 1) there is no COD loss

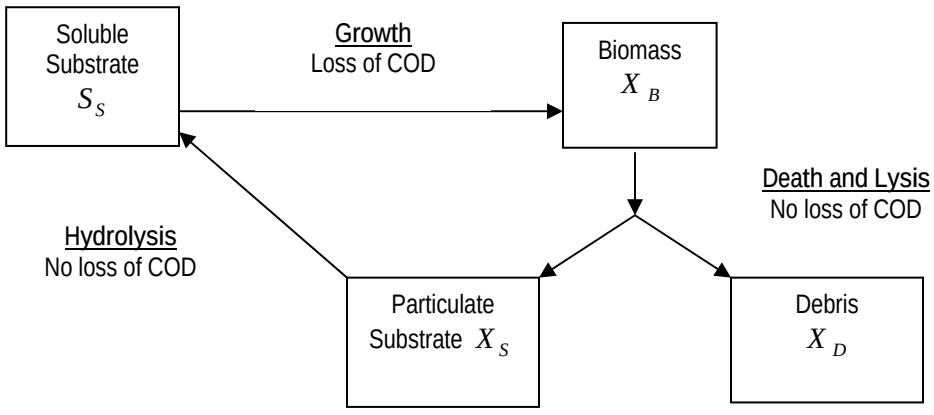


Figure 2-1. Schematic representation of the lysis:regrowth approach to modeling biomass decay. (After Grady et al., 1999)

and no electron acceptor utilization; 2) decay occurs at a constant rate regardless of the environmental conditions (i.e., b_H is not a function of the type of electron acceptor or its concentration); 3) the conversion from X_S to S_S can only occur under aerobic or anoxic condition; however under anaerobic conditions, it results in the accumulation of slowly biodegradable substrate (Grady, C.P.L., Jr.; Daigger, G.T.; Lim, H.C., 1999).

Process 4: Death and lysis of heterotrophs: This process characterizes the reaction of the loss of heterotrophs, which is also modeled by the lysis:regrowth approach in Figure 2-1. The loss rate of heterotrophs is assumed to be same regardless of the types of electron acceptor (Dold and Marais, 1986). However, the utilization rates of specific constituents will be influenced as shown in Table 2-2. The products of this process are biomass debris, slowly biodegradable substrate, and particulate biodegradable organic nitrogen.

Process 5: Death and lysis of autotrophs: The loss of autotrophs from the system is also modeled by the lysis:regrowth approach in a similar way as applied to heterotrophs. The only difference is that the values of the decay coefficient for autotrophic bacteria will be less than that for heterotrophs.

Process 6: Ammonification of soluble nitrogen: This process releases ammonia from nitrogen containing organic compounds. The expression is empirical but has proved to be useful for modeling the conversion (Dold and Marais, 1986).

Process 7: 'Hydrolysis' of particulate organics: This process converts slowly-biodegradable substrate to readily-biodegradable substrate. The correction factor η_h reflects decrease of the hydrolysis rate under anoxic conditions. Also, the rate of hydrolysis is assumed to be zero under absolutely anaerobic conditions, although hydrolysis is known to occur in anaerobic bioreactors (Haandel, 1981).

Process 8: 'Hydrolysis' of particulate organic nitrogen: The final process describes the conversion of particulate biodegradable organic nitrogen, X_{NS} into soluble, biodegradable organic nitrogen, S_{NS}. This rate is assumed proportional to the rate of hydrolysis of slowly biodegradable organic matter.

2.2.4 Parameters in ASM1

The symbols, units and typical values of parameters presented in the ASM1 are listed in detail in Table 2-4.

There are two ways to find the values of the model parameters. One way is to measure the parameters directly through experiments. There are many papers discussing the possible procedures to design the experiments (Ekama, G.A., Dold, P.L, and Maais, G.v.R., 1986; Kappeler, J. and Gujerr, W., 1992; Kabouris, J.C. and Geogakakos, A.P., 1996; Grady, C.P.L., Jr., Daigger, G.T., Lim, H.C., 1999). The other way is to calibrate the ASM1 parameters so that the model accurately mimics the performance of a

Table 2-4. Typical parameter values at 10°C and 20°C for domestic wastewater at neutral pH. (Henze et al., 1986; Grady et al., 1999)

Symbol	Definition	Units	Typical values in ASM1	
	Stoichiometric coefficients		***20 ℃***	***10 ℃***
Y_H	Yield for heterotrophs	mg biomass COD formed / mg COD removed	0.67	0.67
f'_D	Fraction of biomass leading to debris	mg debris COD / mg biomass COD	0.08	0.08
$i_{N/XB}$	Mass of nitrogen per mass of COD in biomass	mg N / mg COD in active biomass	0.086	0.086
$i_{N/XD}$	Mass of nitrogen per mass of COD in debris	mg N/mg COD in biomass debris	0.06	0.06
Y_A	Yield for autotrophs	mg biomass COD formed /mg N oxidized	0.24	0.24
	Kinetic parameters			
$\hat{\mu}_H$	Maximum specific growth rate for heterotrophs	hr^{-1}	0.25	0.125
K_S	Half-saturation coefficient for heterotrophs	mg / L as COD	20.0	20.0
K_{OH}	Oxygen half-saturation coefficient for heterotrophs	mg / L as O_2	0.20	0.20
K_{NO}	Nitrate half-saturation coefficient for denitrifiers	mg/L as N	0.50	0.50
$b_{L,H}$	Decay coefficient for heterotrophs	hr^{-1}	0.026	0.0083
η_g	Correction factor for μ_H under anoxic biomass growth	Dimensionless	0.8	0.8
η_h	Correction factor for anoxic hydrolysis	Dimensionless	0.4	0.4
k_a	Ammonification rate	L/(mg biomass COD·hr)	0.0033	0.0017
k_h	Maximum specific hydrolysis rate	mg COD/(mg biomass COD.hr)	0.125	0.042
K_X	Half-saturation coefficient for hydrolysis of slowing biodegradable substrate	mg COD/ mg biomass COD	0.03	0.01
$\hat{\mu}_A$	Maximum specific growth rate for autotrophs	hr^{-1}	0.033	0.013
K_{NH}	Ammonia-N half-Saturation coefficient for autotrophs	mg/L as N	1.0	1.0
K_{OA}	Oxygen half-saturation for autotrophs	mg/L as O_2	0.4	0.4
$b_{L,A}$	Decay coefficient for autotrophs	hr^{-1}	0.005	0.005

particular WWTP. Given the influent, effluent, design and operation parameters of a specific WWTP, the best available model parameters are identified based on the comparison of the simulated and measured plant performance (Wanner, O., Kappeler, J., and Gujer, W., 1992; Daigger, G.T. and Nolasco, D., 1995; Melcer, H., 1999).

The calibrated model parameters are from different calibration methods, for example genetic algorithms (Holland, et al., 1975). They are adaptive search algorithms based on the evolutionary ideas of natural selection and genetic. The concept behind this method is the representation of solutions a problem in an encoded format. The strength of genetic algorithms is that the certain patterns in the genes will be built after the operations are repeated and solutions become more and more fit (determined by a fitness function). Compared with other traditional methods, genetic algorithms are more robust to avoid getting stuck at local optimum. In practice, genetic algorithms work very well in both continuous and discrete problems and represent a powerful search tool in parameter optimization field (Goldberg, 1989). Gentry et al. (2001) applied genetic algorithms to determine both the spatial distribution and the flux represented by the accretion components of groundwater flow equation. They found the genetic algorithm could find an area or areas of accretion to a semiconfined aquifer given a relatively small set of observation data. This technique is less sensitive to the discontinuities in the accretion functions, compared with other traditional methods.

However, in this dissertation, the model parameters are treated as distributions, i.e., a range of numbers, instead of one specific group of values. In order to find the parameter

distributions, Cox (2004) applied Bayesian statistics to infer the universal distributions of 19 model parameters from published values reported for real WWTPs all over the world. The details of the model parameter distributions are shown in Table 2-5. Most of the parameters except η_g are assumed to be lognormal distributions in order to generate only positive values. The parameter distributions prove to be reasonable by the Kolmogorov-Smirnov Goodness-of-Fit test (Cox, 2004). Independence between these parameter distributions is also assumed due to the lack of published correlations. Significant correlations between model parameter distributions may exist in some situations.

The StatASPS (Statistical Activated Sludge Process Simulator) program is designed to introduce the uncertainty of the model parameters. That is, the 19 model parameters are represented by distributions instead of single values. In Cox's paper (2004), 15 parameter distributions are generated from the published parameter values from over the world through the Bayesian method. Most parameters are lognormal distributed except ηg, which is assumed to be a uniform distribution. Beyond the 15 parameters for which Cox (2004) specified distributions, there are four additional ASM1 parameters (f'_D, i_{NXB}, i_{NXD}, and k_a) which were not considered due to a lack of published values. In order to maintain consistency, these four parameters are also assumed to be represented as uniform distributions. The distributional model parameters are shown in Table 2-5, which is used in the Monte Carlo simulation of ASM1 in this dissertation. The parameter values in Table 2-5 are universal rather than site-specific. A goal of this research is to apply the Bayesian method developed by Cox (2004) to obtain site-specific parameter values of the parameters from plant operating data.

Table 2-5.The universal distributions of ASM1 parameters. (after Cox, 2004)

Parameter-Unit	Universal distribution	
	μ (mean ± SD)	σ (mean ± SD)
$\mu_{H,max}$– days^{-1}	1.14 ± 0.11	0.60 ± 0.066
K_s – mg COD/L	1.44 ± 0.16	0.76 ± 0.067
Y_H – mg COD biomass/mg COD substrate	-0.45 ± 0.035	0.12 ± 0.033
b_{LH} – day^{-1}	-1.06 ± 0.16	0.81 ± 0.15
$\mu_{A,max}$ – days^{-1}	-0.51 ± 0.096	0.44 ± 0.085
K_{NH} – mg N/ L	-0.675 ± 0.22	1.00± 0.11
Y_A – mg COD biomass/mg N substrate	-1.52 ± 0.18	0.55 ± 0.19
b_{LA} – days^{-1}	-1.97 ± 0.085	0.28 ± 0.081
k_h – day^{-1}	0.83 ± 0.12	0.36 ± 0.13
K_x –mg COD particulate substrate/mg COD biomass	-2.82 ± 0.46	1.34 ± 0.19
η_h – dimensionless	-0.86 ± 0.17	0.62 ± 0.17
K_{OH} – mg O_2/L	-1.46±0.323	0.83±0.21
K_{OA} – mg O_2/L	-0.82±0.25	0.96±0.11
K_{NO} – mg N/L	-1.55±0.34	1.01±0.20
η_g – dimensionless	U(0.10,0.90)	
f'_D – mg debris / mg biomass COD	U(0.04,0.12)*	
i_{NXB} – mg N / mg COD in active biomass	U(0.043,0.129)*	
i_{NXD} – mg N/mg COD in biomass debris	U(0.03,0.09)*	
k_a – L/(mg biomass COD·hr)	U(0.0034,0.0101)*	

Note: All distributions are assumed as lognormal distribution except η_g, f'_D, i_{NXB}, i_{NXD}, and k_a as uniform distribution. The symbol * denotes the added parameter distributions based on the experience and information available (Huo, 2004).

Rousseau et al. (2001) also gave their parameter distributions to describe the variability and uncertainty of the parameters. Most of the model parameters were assumed to be triangular distributions. These parameter distributions were determined by the mean/median values and uncertainty ranges. The remaining six parameters were considered temperature-dependent parameters. These parameters are the heterotrophic and autotrophic growth rates μ_H and μ_A, the heterotrophic and autotrophic decay constants b_H and b_A, the hydrolysis rate k_h and the half saturation coefficient for hydrolysis of slowly biodegradable substrate K_X. These six parameters were assumed to be truncated normal distributions in order to avoid negative values. However, this paper also recommended further research on how to choose parameter distributions (normal, triangular, etc.) and how this choice affects the effluent distributions.

2.3 Statistical Theories and Methods

2.3.1 Monte Carlo Method

The Monte Carlo (MC) method is a statistical simulation method that estimates possible outcomes from a group of random/pseudo-random variables by simulating a process a large number of times and analyzing the outcomes (http://csep1.phy.ornl.gov/mc/nodel1.html). In this dissertation, Monte Carlo analysis is used to evaluate the ASM1 model and characterize the uncertainty in WWTPs. Their uncertainty is obtained through performing multiple evaluations with randomly selected

model inputs, and then determining both the uncertainty in model predictions and the contribution of the input factors to this uncertainty. Compared to other methods of uncertainty analysis, such as first-order variance propagation, the accuracy of the Monte Carlo process is not affected by non-linear equations (Schuhmacher et al., 2001). Furthermore, it is also conceptually much simpler to understand and apply from a pragmatic viewpoint.

The Monte Carlo method is widely used to complete the uncertainty analysis of the system outputs and the sensitivity analysis of the system inputs, model parameters, and design/operation parameters. There are several examples in the literature of application of the Monte Carlo method to assess the performance of wastewater treatment systems. Von Sperling (1993) used simple uniform parameter distributions to run Monte-Carlo simulations of an activated sludge model with 11 parameters and 4 states (this was not the ASM series by IWA). According to Von Sperling, the Monte Carlo simulation was a simple, useful and robust method to estimate model parameters and complete the sensitivity analysis of the chosen model. He also mentioned certain subjectivities existing in the Monte Carlo simulation: selection of the range of parameter values and selection of the criteria in determining whether parameters were important or not. Vasquez et al. (1999) applied the Monte Carlo method to study the effects of uncertainty in thermodynamic data on chemical process design and simulation. The effects represented systematic errors as rectangular probability distributions and random errors as normal probability distributions. Potential applications of the Monte Carlo procedure include safety factor determination, process modeling optimization, and experimental design.

Abusam et al. (2002) applied the Monte Carlo method to a full-scale oxidation ditch WWTP. The uncertainties considered in their study include parameter values, influent loads, values of the initial states, simulation model, and seasonal changes in water temperature. The ranges of the model parameters were obtained from the calibrated values and reported values in the literature. This study showed how those uncertainty sources affected the plant performance. As discussed in the paper, uncertainties in influent loads and parameter values are the most important sources for the large deviation of the plant performance. The uncertainty in the model structure is negligible. Rousseau et al. (2001) also applied Monte Carlo simulation to the ASM1 model. The parameters were assumed to be triangular or truncated normal distributions. The results from the 300 Monte Carlo runs were shown in a histogram graph. However, authors still recommended further research on how to determine to appropriate number of Monte Carlo runs. Generally, 300 Monte Carlo runs were too few to be acceptable. Furthermore, Bixio et al. (2002) proposed that conventional design approaches employ larger-than-need safety factors to deal with the uncertainty in variations in plant loading. Thus, they proposed an alternative design approach coupled with the probabilistic Monte Carlo engine. Then, this approach was applied to the upgrade of a conventional WWTP. The results showed this approach could help the decision-making process that ensure the effluent standards are met without introducing above-normal capital investments. A risk-cost benefit concept was present because the acceptability of risk cannot be defined in isolation. For the WWTP Hove, the dimensions of the biological reactors were reduced by 21 percent. This large risk (5% or more) was accepted because about 43 percent of investment cost was

saved. Notice that all process parameters were assumed to be either triangular and truncated normal distributions. The truncation was set at 0.00001 to avoid negative meaningless values. The similar assumption of the process parameters was also applied in another Monte Carlo application of the ASM1 model within the GPS-X (Magbanua, 2004). The simulation results indicated that process uncertainty could be reduced with increasing of SRT and the value of HRT did not appear to be an important factor affecting the uncertainty of the plant performance.

The Monte Carlo method has also been applied in ASM3. Koch et al. (2001) calibrated the model with measured data from long-term full-scale and pilot-plant experiments for Swiss municipal wastewater. Furthermore, the uncertainty analysis and sensitivity analysis were completed through the Monte Carlo simulations. This study found that the confidence interval and the uncertainty of the predicted denitrification rate decrease significantly with increasing ration of COD to nitrogen.

In this dissertation, we will consider the variability in plant influent using historical influent data and/or randomly generated influent data, and the uncertainty in process parameters using the universal and/or parameter distributions. The randomly generated influent is an innovative idea to create an influent time series that has similar statistical characteristics but difference from the actual historical influent data. Another significant improvement is the determination of the uncertainty in process parameters. In this dissertation, the Bayesian method and published values all over the world were used to create universal parameter distributions (Cox, 2004). Most of the distributions are

lognormal distributions except for a few parameters with uniform distributions due to lack of information. The parameter distributions proved to be reasonable by the Kolmogorov-Smirnov Goodness-of-Fit test (Cox, 2004). The site-specific parameter distributions will be generated from the monthly-calibrated process parameters at a site-specific WWTP. These two parameter distributions are strongly backed by the statistical theories and tests, which provide the better confidence to be applied in Monte Carlo simulations.

2.3.2 Bayesian Method

The Bayesian method is used to generate a posterior distribution of the unknown parameters based on a known prior distribution and available measurements of those parameters (Meeker, W.Q. and Escobar, L.A., 1998). It is commonly expressed as follows:

Posterior Information = Prior Information + Sampling Information (2-1)

Using the Bayes' theorem, the following equation is obtained:

$$\pi(\theta \mid x) = \frac{L(\theta \mid x) \times \pi(\theta)}{\int [L(\theta \mid x) \times \pi(\theta)] d\theta} \quad (2\text{-}2)$$

where $\pi(\theta|x)$ is the posterior distribution of parameter θ given the sample observations x; $L(\theta|x)$ is the likelihood function; $\pi(\theta)$ is the prior probability density function; and $\int [L(\theta \mid x) \times \pi(\theta)] d\theta$ is the normalizing factor or constant.

One difficulty in the Bayesian method is to choose the appropriate prior distribution for the concerned parameter. Three prior distributions are commonly used at present (Bozdogan, 2002, Class notes).

- Natural conjugate priors are those whose prior and posterior distributions both belong to a standard family of distributions for any sample size and any values of the observations. They include convenience priors, reference priors, and informative priors.
- Vague priors are those for which little or no knowledge of the parameter is available. In this case, we also use the term vague, improper, or noninformative priors.
- The last option is data-based or data-adaptive priors. These are obtained from the available published or measured data.

The Bayesian method is commonly applied in many research fields to generate and regenerate more accurate parameter distribution from the prior distribution and available observations. Compared to point estimators (means, variances), the Bayesian method is used to find one better distribution from the several potential distributions instead of simply assuming a normal distribution. In Cox's paper (2004), he assumed the prior distributions of the parameters are lognormal or uniform distributions, and then used the available published data and Bayesian method to generate the posterior distributions. These generated distributions will be applied in the Monte Carlo simulation in this dissertation. For more details, please refer to Chapters III and V.

2.3.3 Time Series Models

A time series is a set of observations of a variable made at successive times. The goal of the time series analysis is to find a proper way (a "model") to present the time-structured relationship between several variables (Pankratz, 1991). In DeLurgio's book *Forecasting Principles and Applications* (1998), the author discusses the two main categories of time series models. The first is the univariate time series model, which assumes that the variable is only a function of time, receiving no effects from other variables. The second is the multivariate time series model, which introduces the correlation between variables (time series). The univariate time series models include Exponential Smooth (ES), AutoRegressive Integrated Moving Average (ARIMA), and others. The multivariate time series model includes the multiple regression of time series and Dynamic Regression (DR). The former is the simple regression of multiple variables. Its only difference from normal multivariate regression is that those multiple variables are time series. The dynamic regression model is usually chosen to explain the potential correlations between the related time series (Pankratz, 1991). If one or more independent variables (called input variables) are correlated to a dependent variable (called an output variable), the MARIMA (Multivariate ARIMA) model or DR model is preferred. Note that MARIMA is included in the DR model. It describes the attributes of both regression and ARIMA models, while DR model can describe the attributes of both regression and any other time series models. A dynamic regression (DR) model states how an output is linearly related to current and past values of one or more inputs. Basically, the DR model considers (1)

the possible time-lagged relationship between output and input, that is, the time structure of the input-output relationship; and (2) the possible time structure (autocorrelation pattern) of the time series (Pankratz, 1991).

The Exponential Smoothing (ES) method deals with the time series that is related to exponentially weighted moving averages. It includes simple ES, Holt's two-parameter ES, and Winters' three-parameter ES method (DeLurgio, 1998). The simple ES has the following equation.

$$F_t = \alpha A_{t-1} + (1-\alpha)F_{t-1} \tag{2-3}$$

where F_t is the exponentially smoothed forecast for period t, A_{t-1} is the actual in the prior period, F_{t-1} is the exponentially smoothed forecast of the prior period, and α is the smoothing constant.

The Winters' smoothing model is an extension of the Holt ES by adding a third smoothing operation to present seasonality. Thus it is also called the Holt-Winters method (DeLurgio, 1998). The equation of the Winters' model is listed below:

$$Y_{t+1} = (S_t + b_t)I_{t-L+1} + e_{t+1} \tag{2-4}$$

where Y_{t+1} is the forecast for period t+1, S_t is the smoothed nonseasonal level of the series at the end of t, b_t is the smoothed trend in period t, I_{t-L+1} is the smoothed seasonal index for period t+1, and e_{t+1} is the error of period t+1.

Ellis et al. (1993) applied the univariate and uncorrelated ARMA (autoregressive moving average) processes to the influent BOD_5 (5-day biological oxygen demand) mass loading, effluent BOD_5, and three water-quality additional variables. They found the ARMA models more accurately calculated design curves than the traditional design procedures that do not consider the time structure of the data.

Time series models have been used for forecasting wastewater treatment performance. Berthouex and Box (1996) describe a time series modeling procedure to calculate predictions of plant effluent quality with confidence intervals 1-5 days ahead. The plant effluent quality includes Biological Oxygen Demand (BOD) and Suspended Solids (SS). The time series model, which has the form of an exponentially weighted moving average (EWMA), is based on first differences of the dependent and independent log-transformed variables. This modeling procedure interprets that the change of the plant effluent (BOD or SS) is related to the current level of the moving average, and this change is a linear function of differences between the current levels and the moving averages of the independent variables.

2.3.4 Missing Data and Outliers Replacement

Missing values are almost inevitable in data recording necessitating procedures to account for them in the statistical modeling approach. For example, unexpected events and routine holidays may cause the measurements not to be taken. There are several

methods commonly used to deal with missing numeric values. However, most of those methods depend on the MCAR (missing completely at random) assumption, which means the probability that the observation is missing does not depend on the data set. Unfortunately, MCAR observations are less likely in time series data because observations are usually correlated with time series (for example, there are virtually no data available for Fridays, Saturdays, and other holidays for the Oak Ridge WWTP).

Various strategies are available for imputing missing values. 1) The simplest way is to replace all missing values for a given variable with the mean, median, or other location statistics (e.g., percentiles) of the non-missing values of the variable. 2) The second easiest way is to replace missing values with the average of the nearest observation in the future and/or in the past. 3) SAS® software deals with the situation in a more complicated but more valid way:

(1) Replace all missing values with a simple imputed value, like the series mean.

(2) With the updated estimates, forecast the missing values from the current time series model. Then update the original data set with the new forecasts of missing values.

(3) Repeat step 2 until the convergence occurs.

Just like missing data, outliers are very common in the data set. Outliers are defined as unusual data that are significantly different from the rest of the data set. The reasons leading to outliers might be recording errors or special events (also called “interventions”). The obvious examples of the potential interventions are variable

redefinition, regulation or law change, natural or artificial disasters, and unexpected system breakdown. It is very important to study the effects of the outliers, which sometimes can uncover the causes of changes (Pankratz, 1991). However, it might be difficult to determine whether the extreme low/high values are true outliers or legitimate random effects. Generally, there are three methods to deal with outliers, which are listed below:

1) Replace the outliers based on the original data. We can use the mean of two or more adjacent values. For example, $Y_{t,new} = (Y_{t-1} + Y_{t+1})/2$ or $Y_{t,new} = \frac{\sum_{\substack{i=-s \\ s\neq 0}}^{s} (w_{t+i} Y_{t+i})}{2s}$, where w_{t+i} is the weight of Y_{t+i} that affects the Y_t, s is half range of the adjacent values.

2) Replace the outliers based on the forecast data. This method works well if a reliable forecast method exists. The core of this method is to find the simple and effective time series models. Thus, the characteristics of the time structure of the time series might be reviewed and applied in the forecast models. After the proper models are found, the replacement is straightforward. First remove the outliers and treat them as missing data. Then substitute the forecasted values for the deleted outliers.

3) Preserve the outlier with the unusually high or low value by subtracting or adding some additional quantity to its original value. Normally, the adjusted value derived from this process would still preserve some of its extreme value. This can be valuable when it is believed that future values

will also be extreme but not quite as high or low as the outlier. For example, the maximum or minimum of the data set after excluding the outlier is commonly taken to replace the extremely high or low value, respectively.

Robinson et al. (2005) developed a multivariate outlier detection method to evaluate water quality data for outliers. Compared with commonly used univariate methods, this new method considers the correlation between variables. This method is believed to significantly decrease the number of possible outliers that have extreme ratios between correlated variables. For example, this paper indicates CBOD has some correlation with NH_4-N. If both CBOD and NH_4-N at the same time have extremely large values, the univariate methods may regard both of two points to be outliers. However, with this multivariate method, those two points might be OK considering those two variables are highly correlated.

In this dissertation, the outlier problem can be solved using a new univariate method: Two-directional Exponential Smoothing (TES) method, which will be discussed in details in Chapter IV.

CHAPTER III

MATHEMATICAL MODELING AND STATISTICAL APPLICATION

In this chapter, one program called StatASPS (written in both Visual Basic® and MATLAB®) is developed to apply the ASM1 model to wastewater treatment plants. Detailed numerical methods are applied to solve the steady and dynamic simulation of activated sludge processes. Time series models are also developed for use as random influent generators. Time series models for influent flow, temperature, BOD_5, SS, and ammonia-N are developed and used to simulate the influent to the plant. These models capture the statistical characteristics (trend, correlation, and seasonality, etc.) of the influent variables. The Monte Carlo procedure for this case is illustrated in Figure 3-1.

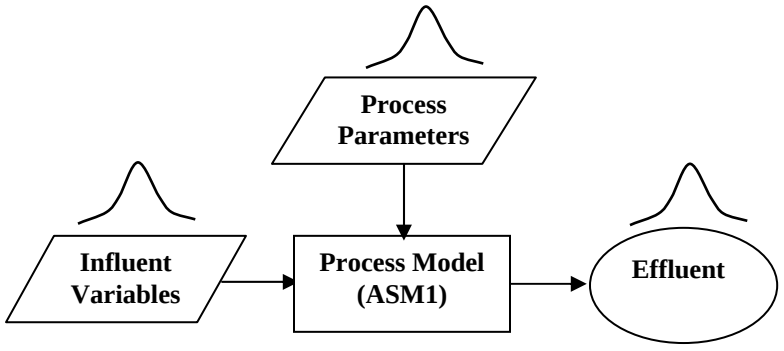


Figure 3-1. Monte Carlo simulation of ASM1 to evaluate uncertainty in activated sludge processes.

We have assumed that the values of the various process parameters have no correlations, because we were unable to find in the literature any reported relationships between the various ASM1 process parameters. As shown in Figure 3-1, there are two kinds of uncertainty of the ASM1 inputs. One is the variability of the influent variables, such as flow rate, temperature, BOD_5, SS, and Ammonia-N. This uncertainty is described by generated random influent data using suitable time series models, which capture the potential uncertainties in plant influent. The other is the uncertainty of the process parameters, which is mainly determined by the design and operation of WWTPs. Correspondingly, the uncertainty of the model inputs leads to the uncertainty of the model output, i.e., the plant effluent that defines the plant performance.

3.1 Application of Activated Sludge Model No.1 (ASM1)

The 8 processes, 13 components and 19 model parameters in the ASM1 are presented and explained in detail in Chapter II. In this chapter, we are going to focus on how to apply the ASM1 in a real WWTP. Specifically, the Monte Carlo method is introduced into ASM1 simulations to search the statistical characteristics in the plant performance. The universal parameter distributions and site-specific parameter distributions are applied in the following and future simulations.

3.1.1 Mass Balance Equations in ASM1

Integration of the ASM equations and mass balance equations is described in this section. The basic mass balance equation is given by equation (3-1):

Accumulation = input – output + generation (3-1)

Note that the unit of each term above is mass/time.

Initial development of the material balance equations is for a single CSTR; the case of multiple CSTRs in series will be developed subsequently. The following equation describes the mass balance of component i within a control volume V. For notation, please refer to Figure 3-2.

$$\frac{dC_i}{dt}V = (FC_{iO} + F_R C_{iR}) - [F_W C_{iW} + (F - F_W)C_{iE}] + r_i V \quad (3\text{-}2)$$

where, F, F_R and F_W are influent flow rate, recycle flow rate and wastage flow rate (Volume/Time), respectively; C_{iO}, C_i, C_{iR}, C_{iE}, and C_{iW} are the concentrations of

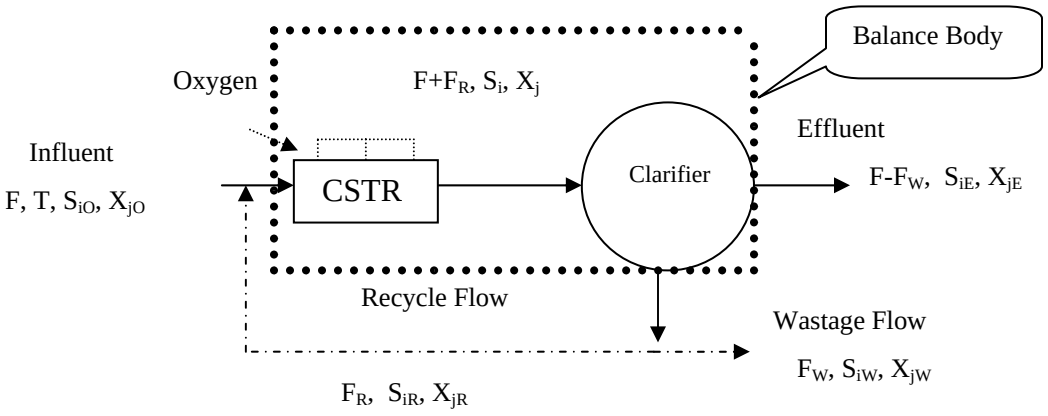


Figure 3-2. Basic flow diagram of activated sludge process in ASM1.

component i in the influent, CSTR, recycle flow, effluent, and wastage flow (Mass/Volume); V is the control volume of the system, which may be part of the total volume of the bioreactor if we consider multiple CSTRs; each of the r_{si} terms represents the reaction rate for that particular component as defined in ASM1 (Henze et al., 1986). From Figure 3-2, we can easily find that $C_{iW} = C_{iR}$. Separate material balances are needed for soluble and particulate components, denoted by S and X, respectively.

For each soluble component i, effluent, waste and recycle concentrations of each component are equal to the concentration in the aeration tank. Thus, the mass balance equation is:

$$V\frac{dS_i}{dt} = (FS_{iO} + F_R S_{iR}) - [(F - F_W)S_{iE} + F_W S_{iW} + F_R S_{iR}] + r_{Si}V \tag{3-3}$$

For each particulate component j, a similar material balance applies:

$$V\frac{dX_j}{dt} = (FX_{jO} + F_R X_{jR}) - [(F - F_W)X_{jE} + F_W X_{jW} + F_R X_{jR}] + r_{Xj}V \tag{3-4}$$

For particulate components the relationship between the effluent concentration and the component concentration in the aeration tank is given by $X_{j,E} = \beta X_j$, where $\beta = 1-\alpha$ and α is the efficiency of the clarifier. In this study we have assumed a constant clarifier efficiency of 99.75%, which represents relatively good clarifier performance.

The definition of SRT is

$$\Theta_C = \frac{VX_T}{F_W X_{TW} + (F - F_W)X_{TE}} \tag{3-5}$$

where $X_T = \Sigma X_j$ and the mass balance equation of the secondary clarifier is

$$(F + F_R)X_T - [(F - F_W)X_{TE} + F_W X_{TW} + F_R X_{TR}] = 0 \tag{3-6}$$

From the above three equations, we can derive the following mass balance equation for the particular components.

$$\frac{dX_j}{dt} = \frac{FX_{jO} - \beta(F - F_W)X_j - \gamma F_W X_j + r_{Xj}V}{V} \tag{3-7}$$

$$\gamma = \frac{X_{jR}}{X_j} = \frac{X_{jW}}{X_j} = \frac{X_{TR}}{X_T} = \frac{(F + F_R - \frac{V}{\Theta_C})}{F_R}, \tag{3-8}$$

and

$$F_w = \frac{\frac{V}{\Theta_C} - \beta F}{\gamma - \beta} \tag{3-9}$$

As mentioned above in simple ideal CSTR model, the components and transformation processes included in the ASM1 are described with indices i and j, respectively. Stoichiometric coefficients are presented as c_{ij} in the middle of the matrix, and the rate of process j is presented as r_{pj}, as listed in Table 2-2. Thus, the overall production rate of component i is calculated as:

$$r_i = \sum_{j=1}^{J} c_{ij} r_{pj} \tag{3-10}$$

where j is the total number of the processes.

For example, the mass balance equation of the readily degradable substrate (S_S) is shown as follows:

$$\frac{dC_7}{dt}V = (FC_{7O} + F_R C_{7R}) - [F_W C_{7W} + (F - F_W)C_{7E}] + r_7 V$$

$$= (FC_{7O} + F_R C_{7R}) - [F_W C_{7W} + (F - F_W)C_{7E}] + r_7 V \quad (3\text{-}11)$$

Figure 3-3 shows the basic structure of tanks-in-series in wastewater treatment. In fact, the influent here refers to the effluent from the primary clarifier. Basically, there are only two physical parts. One is the CSTRs in series (i.e., bioreactors), where the degradation of substrate and growth of biomass occur; the other is the secondary clarifier, where the solids and liquid are separated. There are also four important flows: influent, effluent, recycle flow, and wastage flow, which are shown as specific symbols in Figure 3-3. It should be noted that a complicated clarifier model is not considered in order to simplify the modeling process. Various models have their own advantages and disadvantages, and a consensus on the best model has not been achieved. Thus, the efficiency of the

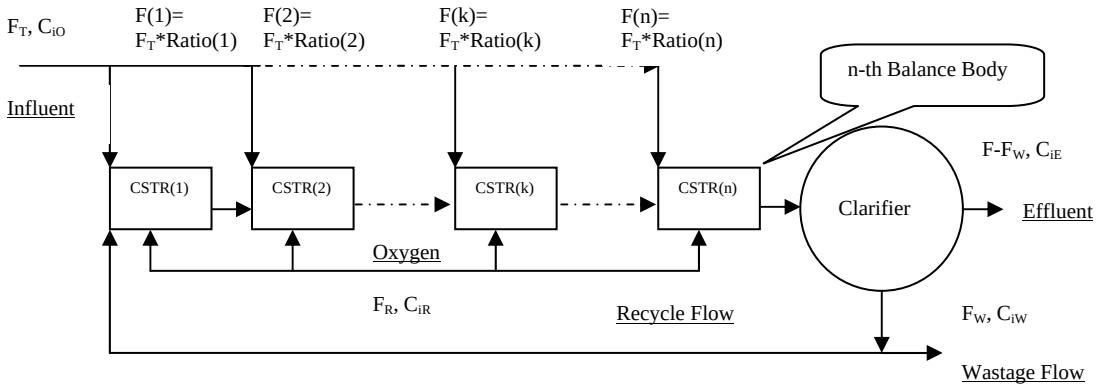


Figure 3-3. Basic flow diagram of activated sludge process tanks-in-series.

secondary clarifier is simply assumed to be a constant because the site-specific information is inadequate. Based on the plant data from Kuwahee WWTP, Ojai Valley WWTP, Oak Ridge WWTP, and Seneca WWTP, we assume that the efficiency of the secondary clarifier is 99.75 percent. Note that the value of the secondary clarifier efficiency might be changed if a specific WWTP is applied and related information is available.

3.1.2 Numerical Algorithms for the Simulation

3.1.2.1 Steady state solutions

For the steady state simulation, the differential terms are all equal to zero. The classic Newton-Raphson method for solving nonlinear equations is chosen, which is thought to be the most efficient multidimensional root finding method for reasonable close initial guesses (Sprott, 1998). The matrix format of the Newton-Raphson method in multiple dimensions is shown as follows:

$$\text{Given } \vec{F}(\vec{X}) = [f_1(x_1, x_2, ..., x_n), f_2(x_1, x_2, ..., x_n), ..., f_n(x_1, x_2, ..., x_n)] = \vec{0}, \quad (3\text{-}12)$$

Newton-Raphson method uses an iterative procedure to converge to the solutions. Each iteration solves the linear system of equations:

$$\vec{J}(\vec{X}) \cdot \Delta\vec{X} = -\vec{F}(\vec{X}) \quad (3\text{-}13)$$

where $\Delta\vec{X}$ =correction to $\vec{X}$, for example, $x_{i,j} = x_{i-1,j} + \Delta x_j$; i = ith iteration; j = jth variable; and $\vec{J}(\vec{X})$ = Jacobian matrix of $\vec{F}(\vec{X})$ =

$$\begin{bmatrix} \partial f_1/\partial x_1 & \partial f_1/\partial x_2 & \ldots & \partial f_1/\partial x_n \\ \partial f_2/\partial x_1 & \partial f_2/\partial x_2 & \ldots & \partial f_2/\partial x_n \\ \ldots & \ldots & \ldots & \ldots \\ \partial f_n/\partial x_1 & \partial f_n/\partial x_2 & \ldots & \partial f_n/\partial x_n \end{bmatrix} \quad (3\text{-}14)$$

3.1.2.2 Dynamic solutions

The main difference between dynamic and steady state simulations is that the first-order differential terms are no longer equal to zero. The fourth-order Runge-Kutta method is consequently used in the StatASPS package because it has proved to be an adequate method for most problems (Sprott, 1998). The Runge-Kutta method numerically integrates ordinary differential equations by using a trial step at the midpoint of an interval to cancel out lower-order error terms. The classical fourth-order Runge-Kutta formulas are shown below.

$$K_1 = h \times f(t_n, y_n)$$

$$K_2 = h \times f(t_n + h/2,\ y_n + K_1/2)$$

$$K_3 = h \times f(t_n + h/2,\ y_n + K_2/2)$$

$$K_4 = h \times f(t_n + h,\ y_n + K_3)$$

$$y_{n+1} = y_n + K_1/6 + K_2/3 + K_3/3 + K_4/6 + o(h^5) \quad (3\text{-}15)$$

As shown above, the fourth-order Runge-Kutta method requires four evaluations of the right-hand side per step h. However, it should be noticed that higher order does not always mean higher accuracy, although fourth-order Runge-Kutta is generally superior to second-order.

Figure 3-4 shows the flow diagram of the Monte Carlo simulation of the ASM1 in the StatASPS package. The comparison of simulated results between the StatASPS and the GPS-X® shows that the simulations from the StatASPS software both in Visual Basic® and MATLAB® are very close to those from the GPS-X® software. The advantages of Visual Basic StatASPS program are its flexibility and faster speed, which can be complied to an executable file and run under other environment besides Windows®. However, the advantages of MATLAB® StatASPS program are its powerful built-in algorithms and stronger graphic functions. Considering the advantages and disadvantages of both programs, we decided to use both according to different requirements. For example, for the single CSTR's Monte Carlo simulations, we use the StatASPS program in Visual Basic® due to its faster calculating speed. As for calibration purposes and multiple CSTRs' Monte Carlo simulations, we use the StatASPS program in MATLAB® because we can take advantage of the built-in statistical functions and powerful graphic functions.

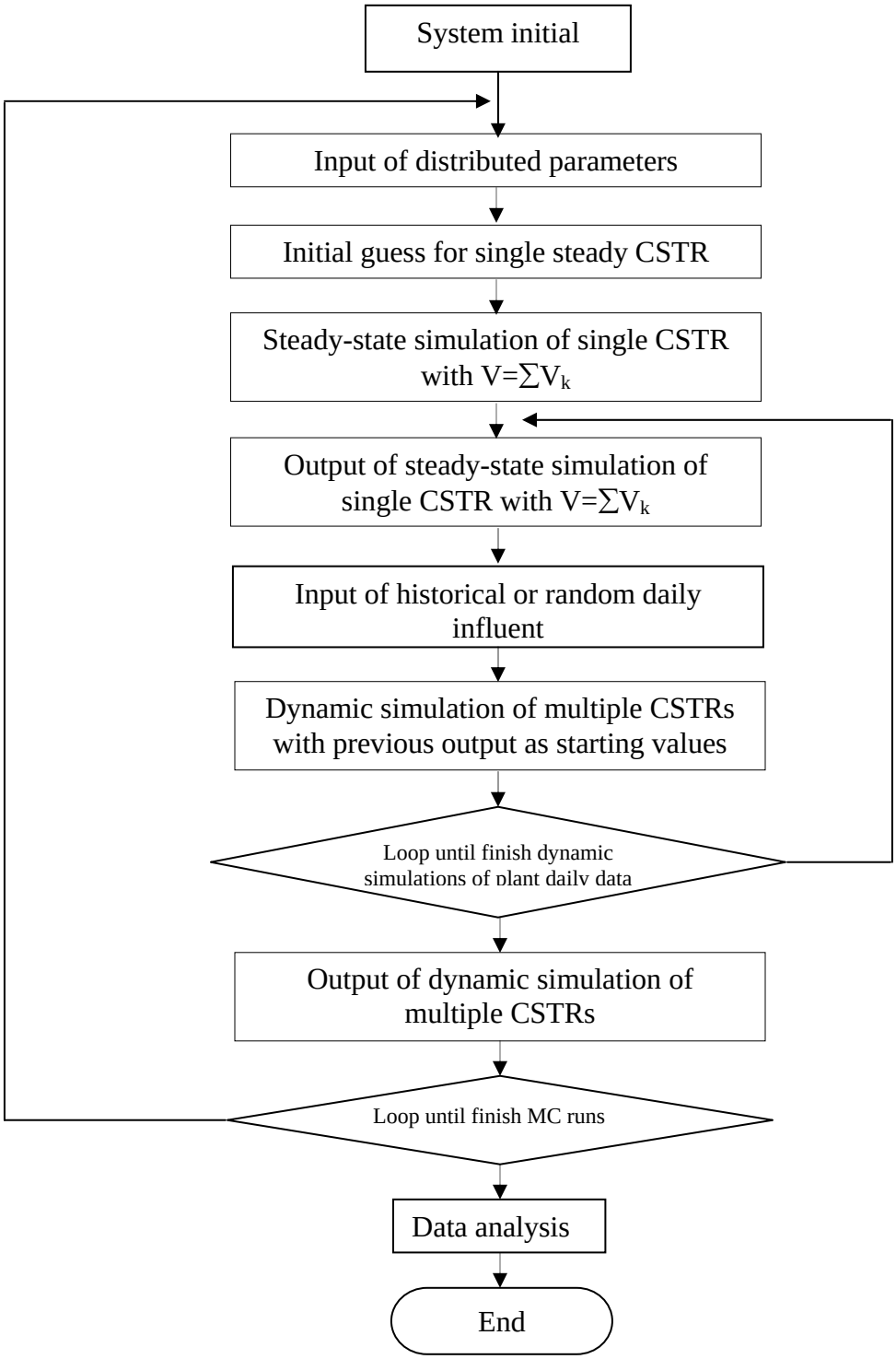


Figure 3-4. The overall flow diagram of Monte Carlo simulations of the ASM1 in the StatASPS package.

For dynamic simulations, the assumed initial conditions may have undue influence on the simulation predictions for early time points in the simulations. It is well known (Grady et al., 1999) that it takes some time (3-4 SRTs) for the plant to respond to a step change in input conditions. However, in this dissertation, we take the steady-state operation values as the starting values for the dynamic simulation. This significantly decreases the influence of the initial condition and allows realistic operating conditions to be reached after a much shorter time. A test simulation was conducted for 99 days with repeating diurnal influent pattern. Under conditions of SRT = 5 days and HRT = 5 hours, 5 days (i.e., 120 hours) were required for the system to reach quasi-steady state. Thus, the valid data for analysis only include those data after 5 days dynamic simulation. Figure 3-5 also shows how dynamic simulations of BOD_5, SS, and NH_3-N reach quasi-steady state rapidly after taking steady state simulated results as a starting point.

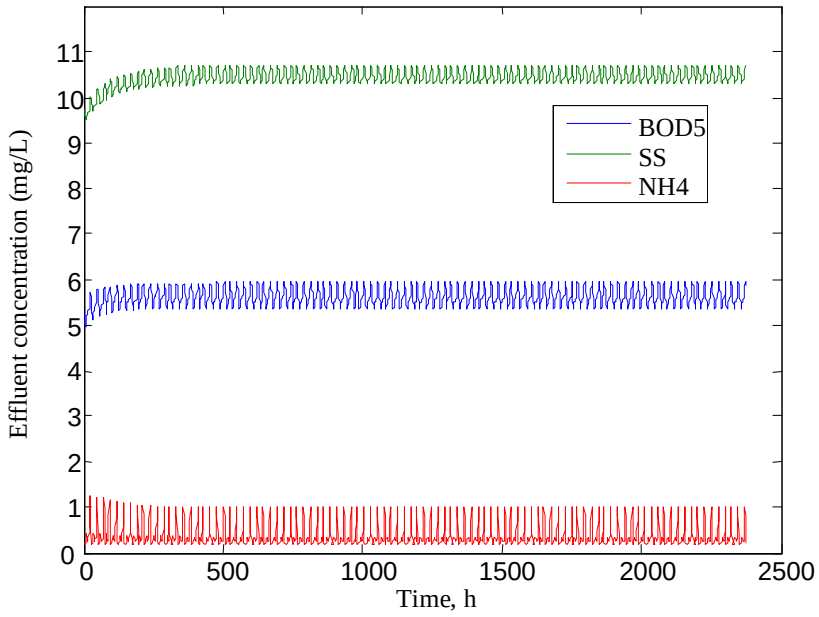


Figure 3-5. Approach to quasi-steady state from the initial conditions.

3.1.3. ASM1 Influent and Effluent Conversion Methods

3.1.3.1. Influent conversion method

The inputs of the ASM1 include the influent flow rate and 13 components. However, the model inputs of 13 components are not normally measured during routine operation of activated sludge plants. Thus, a method for conversion of the traditional measurement (BOD_5, COD, SS, TKN, Ammonia-N, etc.) to the 13 model input components is need. Grady et al. (1999) recommended one conversion method that we have adapted in slightly modified form here. The method allows computations of the 13 influent components from measured values of TSS, BOD_5, and Ammonia-N (S_{NH}). The Total COD (TCOD) consists of four components: (1) particulate biodegradable COD (X_{SO}), (2) soluble biodegradable COD (S_{SO}), (3) particulate inert COD (X_{IO}), and (4) soluble inert COD (S_{IO}). The subscript O indicates the influent. The conversion method is summarized as follows:

$TCOD=2.1*BOD_5$

$COD_{BO}=1.71*BOD_5$

$VSS=0.75*SS$

$TKN=1.74*S_{NHO}$

$COD_{IO}=TCOD-COD_{BO}$

$X_{IO}=0.56*VSS$

$S_{IO}=COD_{IO}-X_{IO}$

$S_{SO}=0.35*COD_{BO}$ (The value shown in Grady et al.'s book is 0.43)

$X_{SO}=COD_{BO}-S_{SO}$

$ON_{TO}=TKN-S_{NHO}$

$ON_{TO}=S_{NSO}+S_{NIO}+X_{NSO}+X_{NIO}$

$X_{NIO}\approx i_{N/XD}*X_{IO}= 0.06*X_{IO}$

S_{NIO} = 1.5 mg/L (Suggested)

$S_{NSO}=(S_{NSO}+X_{NSO})*[(S_{SO})/(S_{SO}+X_{SO})]$ (3-16)

where SS is suspended solids, VSS is volatile suspended solids, ON_{TO} is total organic nitrogen, and TKN is total Kjeldahl nitrogen.

This method has proved to be useful in some cases (Grady et al., 1999). However, when we applied this method to the Oak Ridge WWTP, some unexpected negative values are produced, which have no meaning. There are two methods to deal with the negative values. Firstly, the negative value indicates that the component is very small and is therefore negligible. Thus, we might replace it with a very small positive number, for example 0.001 or even 0. Secondly, the other method is to change the ratio of the components, which directly results in the negative values. We adopted the former case in this work.

3.1.3.2 Effluent conversion method

We are interested in evaluating plant performance in terms of several conventionally defined effluent variables such as BOD_5, TCOD, SS, TKN, and ammonia nitrogen. The relationships between the 13 state variables and the conventional effluent variables are given by the following formulas:

$$TKN = S_{NHE} + S_{NSE} + X_{NSE}$$

$$BOD_5 = f_{bod} \times (S_{SE} + X_{SE} + X_{BHE} + X_{BAE})$$

$$TCOD\ (Total\ COD) = S_{IE} + S_{SE} + X_{SE} + X_{BHE} + X_{BAE} + X_{IE} + X_{DE}$$

$$VSS = XCOD/icv = (X_{SE} + X_{BHE} + X_{BAE} + X_{IE} + X_{DE})/icv$$

$$SS = VSS/ivt \qquad (3\text{-}17)$$

where $f_{bod} = BOD_5/BOD_u$, icv = XCOD/VSS, and ivt = VSS/SS. Those values are defined based on experience according to Table 3-1 (Hydromantis, Inc). The subscript E indicates the component concentration in the effluent.

Table 3-1. Typical coefficients of the influent, CSTRs, and effluent. (Hydromantis, Inc.)

Items	icv[mg COD/ mg VSS]	ivt [mg VSS/mg TSS]	fbod [BOD_5/BOD_u]
Influent	2.2	0.6	0.66
CSTRs	1.48	0.75	0.66
Effluent	1.48	0.75	0.66

3.2 ASM1 Parameters

As shown in Table 2-4, the ASM1 model has 19 parameters including 5 stoichiometric coefficients and 14 kinetic parameters. The parameters are traditionally treated as fixed. Experiments can also be conducted to measure site-specific values of a few of the most critical process parameters. If resources for the measurement of these parameters are not available, default (fixed) values recommended in the literature will be assumed to be sufficiently accurate. For the calibration procedure, in order to better simulate the plant performance, Daigger (1997) and Huo (2001) calibrated the process parameters that were divided into 3-4 fixed seasonal groups. The research work also proved that the parameters were changing through a year. There were also a few studies in which values of process parameters were measured repeatedly over a period of a few months. The results clearly showed time-dependent variability in the site-specific parameters (Drtil et al., 1993; Fillos et al., 2000; Orhon et al., 1999; Sözen et al., 1996 and 1998). Therefore, there is uncertainty in process parameters that cannot be described by simply making a single measurement and/or calibration. The causes of parameter variability are the changing characteristics of the biomass, changing seasonal wastewater temperature, changing influent organic and hydraulic loadings, and changing DO concentrations, etc. This uncertainty in process parameters can be quantitatively described by a statistical distribution of the parameter values that occur in wastewater treatment plants all over the world (universal distribution). Seemingly, parameter uncertainty can be greatly reduced by making site-specific measurements of the parameter value. With the introduction of

uncertainty in process parameters, we expect that this method can capture the uncertainty in DO concentration, configuration, clarifier performance, and process upsets.

As is expected, in this dissertation, the model parameters are not considered to be constant but are sampled from distributions. That is, there are uncertainties in those parameters. Cox (2004) generated a group of parameter distributions using Bayesian method. Because the values used in the Bayesian method were from the measured and/or simulated data all over the world, we call these distributed parameters the universal parameter distributions (UPD) (or non-site-specific parameter distributions). Besides the fixed calibrated parameters and the universal parameter distributions, there is still another choice of the model parameters: the site-specific parameter distributions (SPD). The site-specific parameter distributions are generated using the same Bayesian method and same prior distributions. The only difference is that the measured data are from 12-month calibrated parameters. These three kinds of model parameters will be discussed in the following three sections.

3.2.1 Universal Parameter Distributions

The UPD parameters, as shown in Table 2-5, were generated by Cox (2004) using Bayesian method. The data used were from the published values all over the world. Thus, they are also called non-specific parameter distributions.

Most UPD parameters are generated from a lognormal distribution (μ,σ), $x_p = \exp[\mu + \Phi_{nor}^{-1}(p)\sigma]$, where Φ_{nor}^{-1} is the inverse CDF function for the standardized normal distribution and p is a uniform random number sampled over the range 0 to 1. A few parameters are generated using a uniform distribution [a,b], $x_p = a + (b-a)p$, where p is a uniform random number sampled over the range 0 to 1.

In the simulations in this chapter, we have used a temperature of 20°C, a value of Θ_C of 10 days, a recycle ratio of 0.5, and a dissolved oxygen level of 2.0 mg/L. The daily influent is represented as a constant 24-hour mean condition with a superimposed re-occurring diurnal pattern (GPS-X® and Grady et al., 1999), as shown in Figure 3-6. The

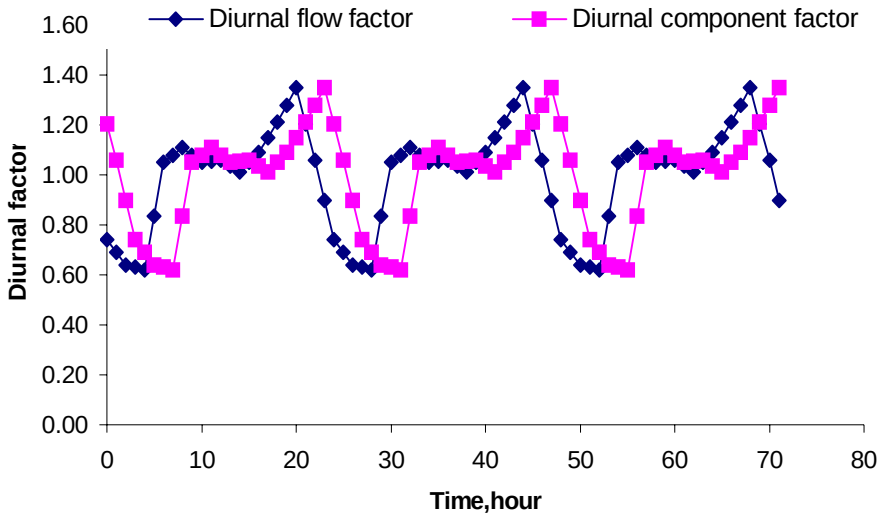


Figure 3-6. The diurnal factors for influent flow and components.

typical diurnal pattern of wastewater flow and concentration are similar and the concentration pattern is always about three hours delay than the flow rate pattern (Dold and Marais, 1996). In this section, we only focused on the study on how the uncertainty in process parameters affects the uncertainty of the plant performance. In Chapter V, we will conduct dynamic simulations with the effect from influent variability added. The influent concentrations in Table 3-2 were used. The results are based on 2000 Monte Carlo trials.

The statistical characteristics of TKN, TCOD, SS, BOD_5 and Ammonia nitrogen from the steady-state simulations are reported in Table 3-3 and Figure 3-7. The variability reported

Table 3-2. The typical characteristics of the domestic wastewater that has undergone primary sedimentation. (Henze et al.,1986; Grady et al., 1999)

Symbol	Unit	Influent Values
X_I – Inert particulate organic matter	mg/L as COD	40
X_S – Slowly biodegradable substrate	mg/L as COD	160
X_{BH} – Heterotrophic biomass	mg/L as COD	96
X_{BA} – Autotrophic biomass	mg/L as COD	10
X_D – Biomass debris	mg/L as COD	0
S_I – Inert soluble organic matter	mg/L as COD	40
S_S – Readily biodegradable substrate	mg/L as COD	64
S_{NO} – Nitrate nitrogen	mg/L as N	1
S_{NH} – Ammonia nitrogen	mg/L as N	12.5
S_{NS} – Soluble biodegradable organic nitrogen	mg/L as N	10.1
X_{NS} – Particulate biodegradable organic nitrogen	mg/L as N	18.28
S_{ALK} – Alkalinity	mM/L	6

Table 3-3. Statistical characteristics of effluent components of Monte Carlo simulation.

Items	Mean	Std Dev	Percentiles				
			Max	97.5%	50.0%	2.5%	Min
TKN	0.9406	1.6827	31.651	3.041	0.617	0.209	0.114
TCOD	52.2272	4.6339	160.82	59.93	51.52	48.00	46.47
SS	9.5270	2.0132	17.727	13.955	9.268	6.242	4.976
BOD_5	5.5817	3.0763	77.515	10.487	5.130	2.627	1.797
Ammonia-N	0.4643	1.6455	31.342	2.255	0.150	0.015	0.005

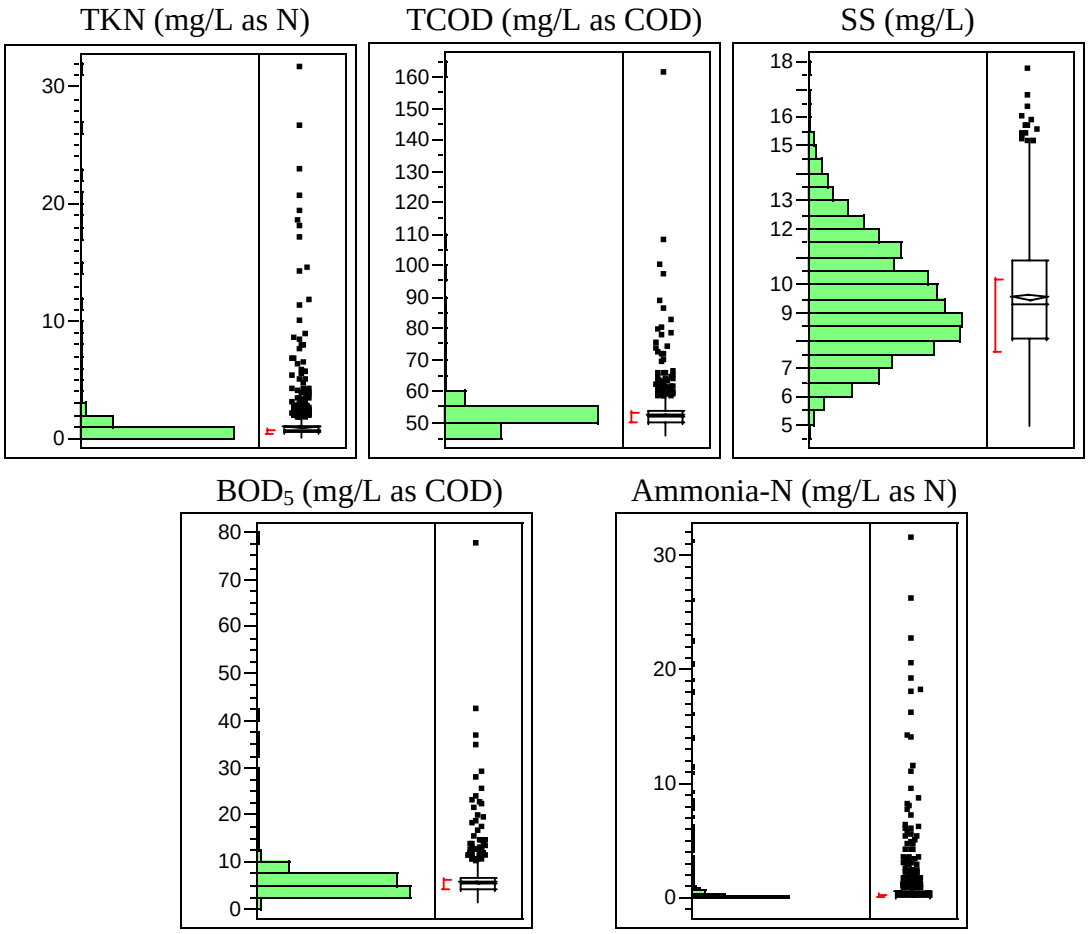


Figure 3-7. The histograms of TKN, TCOD, SS, BOD_5 and ammonia-N of Monte Carlo simulations (X-Axis: Frequency).

is due exclusively to uncertainty in the process parameters that characterize the plant performance. Variability in influent load would cause additional variability in the predicted effluent concentrations. The effluent variables are approximately lognormal in distribution. The mean value of each conventional pollutant is relatively low, indicating good removal of pollutants. However, the effluent distributions are also characterized by numerous extremely large values that may be of concern depending on the contaminant concentrations allowed in the permit. For example, if the discharge permit allowed an effluent ammonia concentration of 5 mg/L, there would be several combinations of process parameters that could result in an average plant performance that violates the permit, even though the mean predicted value is less than 5 mg/L. In order to be more certain of meeting the effluent permit, the engineer would need to either increase the SRT, use a reactor configuration that resulted in better treatment efficiency, or make site-specific measurements of the process parameters to better define their values.

The experimental determination of all 15 process parameters would be time consuming and expensive. We will now demonstrate how sensitivity analysis can be used to determine which variables contribute most to the uncertainty in the effluent. Only the relatively few variables contributing most to uncertainty would need to be determined experimentally.

We use the Spearman rank correlation method to determine the relationship between uncertainty in the process parameters and uncertainty in the effluent variables. The Spearman rank correlation has the advantage that it is not affected by non-linearities in

the process model. The Spearman rank correlation coefficient ranges from –1.0 to 1.0. The larger the absolute value of the correlation coefficient is, the stronger the relationship is. Positive coefficients indicate that an increase in the process variable is associated with an increase in the forecast, whereas negative coefficients imply the reverse situation. The sensitivity analysis considers both the uncertainty of the input parameter and how that uncertainty is propagated through the design model. The results of the sensitivity analysis, calculated in NCSS®, are shown in Table 3-4.

For the component ammonia-N, the significant parameters are, in order, K_{NH}, $\mu_{A,max}$ and Y_A. The analysis indicates that ammonia-N increases with increases in K_{NH} and Y_A and with decreases in $\mu_{A,max}$. The process rate for aerobic growth of autotrophs is $r_3 = \mu_{A,\max}(\frac{S_{NH}}{K_{NH}+S_{NH}})(\frac{S_O}{K_{O,A}+S_O})X_{BA}$, as shown in the Table 2-2. Thus, larger K_{NH} values lead to lower autotroph concentrations, which decrease the consumption of ammonia-N. The parameter $\mu_{A,max}$ has a similar effect except that larger $\mu_{A,max}$ values lead to higher autotroph concentrations, thereby increasing the consumption of ammonia-N. The parameter Y_A is the yield for autotrophs with the unit of mg biomass COD formed/ mg N oxidized. Thus, more yield of autotrophs will lead to more consumption of ammonia-N.

For the component TKN (Total Kjeldahl Nitrogen: organic + ammonia nitrogen), the most significant parameters are, in order, $b_{LH,}$ K_{NH}, $\mu_{A,max}$, and Y_A. These results suggest

Table 3-4. Sensitivity analysis between plant effluent and model parameters.

	S_{NH}	TKN	TCOD	SS	BOD_5
Y_H	-0.014	-0.089	**0.531**	**0.540**	**0.392**
Y_A	**0.271**	0.186	0.097	0.088	0.079
$\mu_{H,max}$	-0.015	0.008	**-0.330**	-0.005	**-0.332**
b_{LH}	-0.003	**0.651**	**-0.340**	**-0.686**	**-0.473**
η_g	0.000	-0.006	-0.009	0.008	-0.011
η_h	-0.006	0.003	-0.021	-0.020	-0.030
k_a	0.000	0.000	0.000	0.000	0.000
k_h	-0.032	-0.050	-0.075	-0.086	-0.084
$\mu_{A,max}$	**-0.439**	**-0.292**	-0.035	-0.040	-0.036
b_{LA}	0.124	0.074	0.005	-0.028	0.003
K_S	0.008	0.028	**0.319**	-0.018	**0.316**
K_{OH}	-0.073	-0.028	0.062	0.041	0.058
K_{NO}	-0.011	0.008	0.027	-0.018	0.028
K_X	0.010	0.065	0.096	0.093	0.102
K_{NH}	**0.760**	**0.447**	0.001	0.001	0.008
K_{OA}	0.200	0.145	-0.005	-0.017	-0.004

that TKN increases significantly with increases in b_{LH}, K_{NH} and Y_A and with decreases in $\mu_{A,max}$. Since ammonia-N is a component of TKN, it is not surprising that the parameters K_{NH}, $\mu_{A,max}$ and Y_A also significantly affect the effluent levels of TKN although they have different order of importance. An additional parameter b_{LH}, the decay coefficient for heterotrophs, is also important. In the lysis:regrowth approach, the decay of biomass produces particulate biodegradable organic nitrogen. Thus, the parameter b_{LH} becomes significant, because heterotrophic biomass is dominant in the bioreactor and its decay coefficient is much larger than autotrophic biomass.

For the component TCOD, the most significant parameters are, in order Y_H, $b_{LH,}$ $\mu_{H,max,}$ and K_S. These results suggest that the TCOD increases significantly with increases in Y_H and K_S and with decreases in b_{LH} and $\mu_{H,max}$. The process rate for aerobic growth of heterotrophs is $r_1 = \mu_{H,\max}(\frac{S_S}{K_S + S_S})(\frac{S_O}{K_{O,H} + S_O})X_{BH}$, as shown in the Table 2-2. Thus, larger K_S values lead to lower heterotrophic biomass concentrations, which decrease the consumption of TCOD. In contrast, increases in the value of the parameter $\mu_{H,max}$ result in increased consumption of TCOD. The parameter Y_H is the yield for heterotrophs with the unit of mg biomass COD formed/ mg COD removed. Thus, more yields of heterotrophs will lead to more consumption of TCOD. The parameter b_{LH} is the decay coefficient for heterotrophs. In the lysis:regrowth approach, the decay of biomass produces slowly biodegradable substrate, which is a main part of effluent TCOD.

For the component SS, the most significant parameters are, in order, b_{LH}, and Y_H. The analysis shows that SS tends to increase significantly with increases in Y_H and decreases in b_{LH}. The explanations for parameters b_{LH}, and Y_H are the same for effluent TCOD.

For the component BOD_5, the most significant parameters are, in order, b_{LH}, Y_H, $\mu_{H,max}$ and K_S. The analysis indicates that BOD_5 tends to increase significantly with the increases in Y_H and K_S, and decreases in b_{LH} and $\mu_{H,max}$. The explanations for parameters b_{LH}, Y_H, $\mu_{H,max}$ and K_S are the same for effluent TCOD.

Table 3-5 summarizes the effect of model parameters on plant effluent components, which can be used in calibration procedures. For example, suppose we want to calibrate effluent ammonia-N. The order of calibrated parameters is K_{NH} (ammonia substrate half-

Table 3-5. The revised sensitivity table for calibration purposes.

Parameter / Effluent	**No.1**	**No.2**	**No.3**	**No.4**
Ammonia-N	K_{NH}(+)	$\mu_{A,max}$(-)	Y_A(+)	N/A
BOD_5	b_{LH}(-)	Y_H(+)	$\mu_{H,max}$(-)	K_S(+)
SS	b_{LH}(-)	Y_H(+)	N/A	N/A
TKN	b_{LH}(+)	K_{NH}(+)	$\mu_{A,max}$(-)	N/A
TCOD	Y_H(+)	b_{LH}(-)	$\mu_{H,max}$(-)	K_S(+)

saturation constant, mg/L), $\mu_{A,max}$ (maximum specific growth rate of autotrophic biomass, hr^{-1}), and Y_A (autotrophic yield, mg COD biomass/mg N substrate). The plus sign means the positive correlation between effluent components and process parameters. That is, the effluent components would increase with increasing of process parameters. If our calibrated plant effluent data are less than the measured plant effluent data, the first parameter we need to concern is K_{NH}. In this case, we need increase the value of K_{NH}. The second parameter is $\mu_{A,max}$ and we need decrease the value of $\mu_{A,max}$. The last parameter is Y_A and we need increase the value of Y_A. This procedure can then determine the calibrated parameters for certain effluent components. The detailed application will be discussed in next section.

From the simulation and analysis above, we can make the following conclusions:

- The uncertainty in the process performance due to model parameter uncertainty is significant, and would only increase as variable plant loading is also considered.
- Model parameters contributing most to the uncertainty of the plant performance were identified using sensitivity analysis based on Spearman rank correlation. Several parameters contributing the most to the uncertainty in specific effluent variables include: for TKN, $b_{LH,}$ K_{NH}, $\mu_{A,max}$, and Y_A; for TCOD, Y_H, $b_{LH,}$ $\mu_{H,max,}$ and K_S; for SS, b_{LH}, and Y_H; for BOD_5, b_{LH}, Y_H, $\mu_{H,max}$ and K_S; and for ammonia-N, K_{NH}, $\mu_{A,max}$, and Y_A. These results reflect both the uncertainty of the input parameter and the way in which this uncertainty is propagated through the model.

- Model parameters identified by the sensitivity analysis as contributing significantly to uncertainty in plant performance can be determined experimentally, thereby decreasing the uncertainty in predicted performance and increasing the confidence of engineers and regulators that the plant will perform as required by the permit.

3.2.2 Calibration Procedure and Calibrated Parameters

As discussed, the previous Monte Carlo simulation used the universal parameter distributions, which are based on the published values all over the world. In order to better describe the plant performance for a specific WWTP, we need to find out the specific group of model parameters for this WWTP. The calibration procedure is needed to find the site-specific calibrated parameters. This procedure highly depends on the sensitivity analysis in Table 3-5. First, we need to figure out which effluent components will be the main concern in the calibration. For examples, if we consider the ammonia nitrogen, BOD_5, and SS to be the major effluent components, only 6 of total 19 parameters are needed to be included in this calibration procedure: K_{NH}, $\mu_{A,max}$, Y_A, b_{LH}, Y_H, and $\mu_{H,max}$. This will significantly decrease the workload of either calibration or measurement.

In this dissertation, only effluent ammonia nitrogen is considered due to the limitation of the secondary clarifier model. So far, there is no well-recognized clarifier model

available. Thus, instead of choosing one unreliable clarifier model, in this study we have assumed a constant clarifier efficiency of 99.75 percent, which represents relatively good clarifier performance. The clarifier efficiency is a ratio of the removal particulate concentration to the particulate concentration in the last CSTR. The historical plant data from both the Oak Ridge and Seneca WWTPs have proved that this clarifier efficiency is very reasonable. The effluent ammonia nitrogen approach has been used in the WERF research project (00-CTS-3). For details, please refer to the project report by Cox et al. (2003).

The ammonia nitrogen concentration in the Oak Ridge WWTP was measured by the Standard Method 4500-NH3-G: ammonia selective electrode method. According to the manufacturer of the Orion Ammonia Electrode, the detection limit is 0.01 mg/L ammonia nitrogen. As for the Seneca WWTP, the ammonia nitrogen concentration was measured by the automated phenate method, which is equivalent to the EPA method 350.1 or Standard Method 4500-NH3-G. This method is applicable over the range of 0.02 to 2.0 mg/L (American Public Health Association, American Water Works Association, and Water Environment Federation, 1998). However, the historical effluent ammonia nitrogen data of both Oak Ridge and Seneca WWTPs have the minimum value of 0.01 mg/L. Thus, the detection limit of the ammonia nitrogen method is considered to be 0.01 mg/L. Its effect on the calibration results will be discussed in Chapter V.

Because the effluent ammonia nitrogen is the only component used to our calibration procedure, Table 3-6 shows the simplified the calibration strategy. Only three model

Table 3-6. The revised sensitivity table for calibration purposes (NH_4-N only).

Parameter / Effluent	No.1	No.2	No.3
Ammonia-N	K_{NH}(+)	$\mu_{A,Max}$(-)	Y_A(+)

parameters will be calibrated: K_{NH}, $\mu_{A,max}$, and Y_A. The plant data used are from the one full year's historical data. The calibration is completed month by month. That is, finally we will obtain 12 groups of model parameters corresponding to each of the 12 months. The traditional way to consider the changing values of model parameters is based on the season differences. Huo (2001) divided the model parameters of Tianjin Ji-Zhuang-Zi WWTP into three categories: one is for Winter, another is for Summer, and the last one is for Spring and Fall. In our dissertation, we believe that the model parameters change more significantly than seasonal changes. Thus, we shorten the period to one month. The target function is the RMSE (root mean square error), which is calculated as the square of the difference between the predicted and historical effluent ammonia nitrogen data.

The enumerative search method is chosen in this calibration procedure. Compared with other searching methods, for example, genetic algorithm (GA), the enumerative search method finds the targeted values by searching a grid parameters with predefined intervals. In our case, the range for each parameter is [0.1 × Default Value, 1.9 × Default Value] with an interval of 15 percent of the default value. This search will require 13^3 =

2197 evaluations. Enumerative search methods are a global optimization method because they avoid getting stuck in local minima as can happen for gradient-based search methods. However, there are also limitations of this enumerative (exhaustive) search method: the searching range and searching intervals. There are two ways to solve this problem if needed. Firstly, we might use a larger range and smaller interval to ensure that reasonably accurate optima may be identified. Secondly, we can also use the regression methods and/or graphic information to find the best group of model parameters using the fitted function. Notice that the main purpose of the calibrated parameter is to generate the site-specific parameter distributions. In this case, we treat parameters as distributions instead of fixed values. That is, we are more interested in the right range of model parameters instead of the exact values.

3.2.3 Bayesian Method and Site-specific Parameter Distributions

Given a specific prior probability density function (PDF) and relative likelihood functions, it is reasonably easy to numerically calculate the posterior distribution by using numerical integration. However, it might be more difficult to those problems with two or more distributed parameters. In order to guarantee accurate and stable results by using numerical integration, it is generally necessary to understand the PDF and relative likelihood functions and then ensure that the integration is calculated under certain requirements (for example, integration intervals). This checking procedure of certain requirements is more complicated for two or more variables. For example, the most

commonly applied distribution in this dissertation is the lognormal distribution, whose PDF function is shown as follows:

$$f(t;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\cdot t\cdot\sigma}\exp\left(-\frac{1}{2}\left(\frac{\ln(t)-\mu}{\sigma}\right)^2\right), \text{ t>0} \qquad (3\text{-}18)$$

Where the median $t_{0.5} = \exp(\mu)$ is a scale parameter and $\sigma > 0$ is a shape parameter.

Simulation-based methods can be used to compute the posterior distributed parameters. To do this, we need to generate a sample from the posterior distributional parameter and then estimate the parameters. A large simulated sample theoretically provides a better estimation of the posterior parameter distributions. However, the number of samples is also highly dependent on current computing technologies and acceptable running time. Obviously, the first one is very objective, which is limited to current available technology development. This includes two aspects: computing hardware (faster computers, etc), and computing methods (better computing algorithms). The later one is really subjective although it is also related to the computing technologies. In this dissertation, we commonly take 24 hours (one full day) as a reasonable and acceptable running time constraint.

The general procedure is shown as follows, which requires only computable expressions of the relative likelihood and the inverse cumulative distribution function (CDF) of the independent marginal prior distributions.

- Step 1: Generate a random sample (θ_i), I=1,…, M, from the prior distribution f(θ).
- Step 2: Keep the ith sample point, θ_i with probability R(θ_i).
- Step 3: Repeat Step 2 until a random sample from the posterior PDF f(θ|DATA) is generated with a sample size M*.

In this case, the parameter θ is the concerned parameter. In our case it is either μ or σ for a lognormal distribution (K_{NH}, $\mu_{A,max}$, or Y_A). R(θ) is the relative likelihood function, as shown below.

$$L(\mu,\sigma)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\cdot t_i\cdot\sigma}\exp\left(-\frac{1}{2}\left(\frac{\ln(t_i)-\mu}{\sigma}\right)^2\right) \tag{3-19}$$

M is the number of original available data. Sample size M* can be calculated by $E(M^*)=M\int f(\theta)R(\theta)d\theta$. Generally, 2000-4000 sample sizes from the posterior distribution can give sufficient accuracy to get a smooth estimation of posterior distributed parameters (Meeker and Escobar, 1998; Cox 2004). The comparison and application of both the UPD and SPD parameters will be conducted in Chapter V.

3.3 ASM1 Influent Inputs and Time Series Models

3.3.1 Introduction

In this section, we will discuss variability (uncertainty) in the plant influent data. There are three options for the influent data. First, we can directly use the historical data. The advantage of this option is that the historical data are normally the easiest data to obtain. Even if the historical data are not available (especially for the new plant design), we can use similar historical data in this region as a reference. As a second option, we can use the predicted data from time series models, which are based on the historical data. The predicted data has advantage of better description of the trends of the historical influent data. However, the disadvantage of this method is that a long period of historical data (for example, three years) are needed to fit a valid time series model with correct trend, seasonality, and so on. Thirdly, we can use the randomly generated influent data. These random influent data are generated from the revised time series model. The advantage of this method is that the generated random influent data are different from either historical or predicted influent; while they have the same/similar statistical features with the historical influent data. This is the most important feature that we expect, especially for Monte Carlo simulations. Using this method, we can generate different random influent data for every Monte Carlo run. The discussion of these three methods will be conducted in details in this section.

As discussed above, this random influent variables generator is used to generate the random plant influent. The generating procedure is described as follows. First, fit the best-available time series model from the historical influent data. Then, modify the parameters of the fitted model based on preceding predicted values and error terms. Thus, the generated random plant influent component will be a random time series with the same statistical patterns with the original plant influent data.

The input data are from one existing plant, Oak Ridge WWTP. The Oak Ridge plant is designed for an average wastewater treatment flow of 735,000 L/h (17,604 m^3/day) and a peak flow of 1,577,255 L/h (37,854 m^3/day). We analyzed the primary clarifier effluent data, because our long-term goal is to study performance of the activated sludge treatment process under varying loads. Almost three-year's of daily data (from November 1, 1999 to July 31, 2002) for five influent variables (temperature, flow rate, BOD_5, suspended solids, and NH_4-N) were used to build the models. The data records for three of the variables (BOD_5, SS, and NH_4-N) are characterized by missing data points on Friday, Saturday, and other holidays, while the records for temperature and flow rate are virtually complete.

Several time series models, such as Exponential Smooth (ES), Autoregressive integrated moving average (ARIMA), and Dynamic Regression (DR) model, were applied to describe possible trends and seasonality (monthly and/or daily). An advantage of the DR model is its ability to consider potential correlations between the time series of the five variables.

3.3.2 Time Series Analysis

3.3.2.1 Missing values

Missing data points on Friday, Saturday, and holidays characterized the data records for BOD_5, SS, and NH_4-N. This situation is not uncommon at wastewater treatment plants, but must be accounted for in the statistical modeling approach, by replacing the missing values with appropriate estimates. There are several methods commonly used to estimate missing values. However, most methods depend on the MCAR (missing completely at random) assumption, which means the probability that a given observation is missing does not depend on the data set. Unfortunately, MCAR observations are less likely in time series data because observations are usually correlated with time series. The regular pattern of the missing data (Fridays, Saturdays, and holidays) at the Oak Ridge WWTP violates the MCAR assumption and renders many data replacement schemes invalid.

One potential solution is to replace all missing values for a given variable with the mean, median, or other location statistic (e.g., percentile) determined from the non-missing values of the variable (DeLurgio, 1998). However, this method can falsely introduce rapid temporal changes in variables in the data record. The second easiest method is to replace missing values with the average of the nearest observation in the future and/or the past. However, in this paper we have used a two-directional exponential smoothing (TES) record to replace the missing values. To apply the method, an exponential smoothing algorithm, provided in most commercial statistical software, is applied to the data in both

the forward and reverse directions. Estimates of missing values are determined by taking the mean of the forward- and reverse-direction exponential smoothing forecasts. The detailed instruction of the TES method will be elaborated in the section 3.4. Figure 3-8 shows a sample time series of the original data and the values after application of the TES method for estimating missing values. For other two variables (SS and NH_4-H), please refer to Chapter IV.

For influent flow rate and temperature, there are few missing data. No missing replacement method is needed. Figure 3-9 shows the daily data of the influent flow rate. From the plot of plant influent flow rate, it is hard to tell whether there is trend and/or

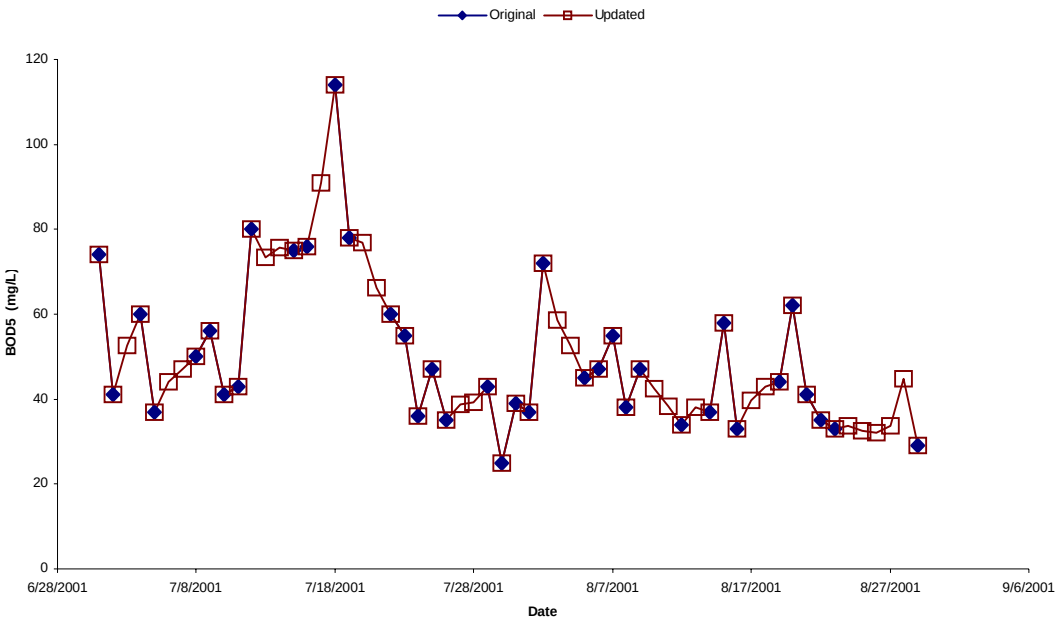


Figure 3-8. The scatter plots of the original and updated BOD_5 series for two months.

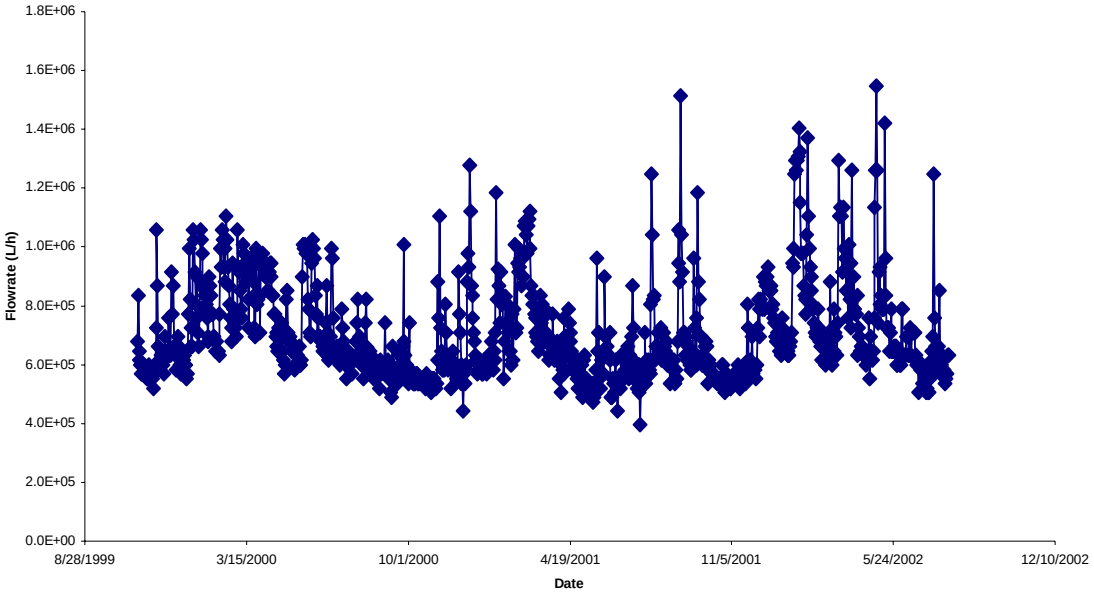


Figure 3-9. The scatter plot of the flow rate series.

seasonality. To find out the uncovered truth, ACF and PACF plots are introduced in the section 3.3.2.2.

3.3.2.2 Correlations

In order to investigate the possible trend and seasonality (daily or monthly), autocorrelation and partial autocorrelation functions (abbreviated as ACF and PACF) are introduced in this section. In the system identification procedure, ACF and PACF plots are used to suggest a tentative form of the model equations. Figures 3-10, 3-11, 3-12 and 3-13 are ACF and PACF plots of flow rate, BOD_5, SS, and NH_4-N. All four figures

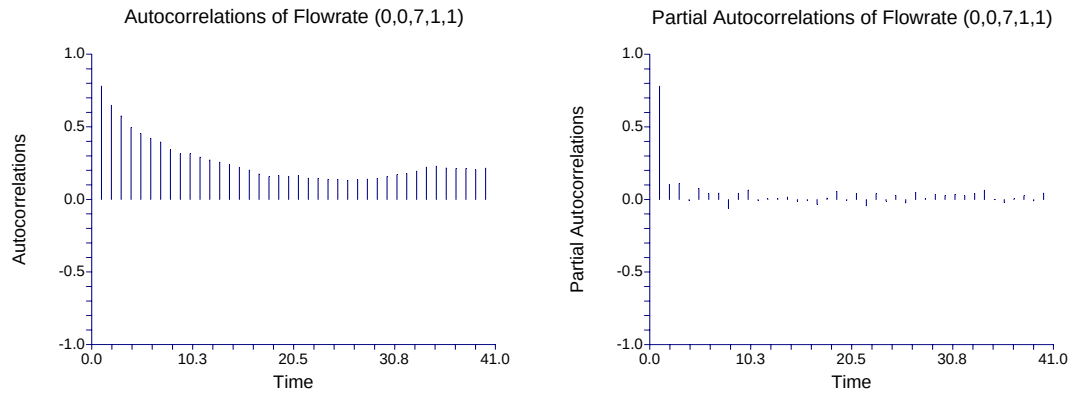


Figure 3-10. The autocorrelation plots of ACFs and PACFs of the daily Flow rate series.

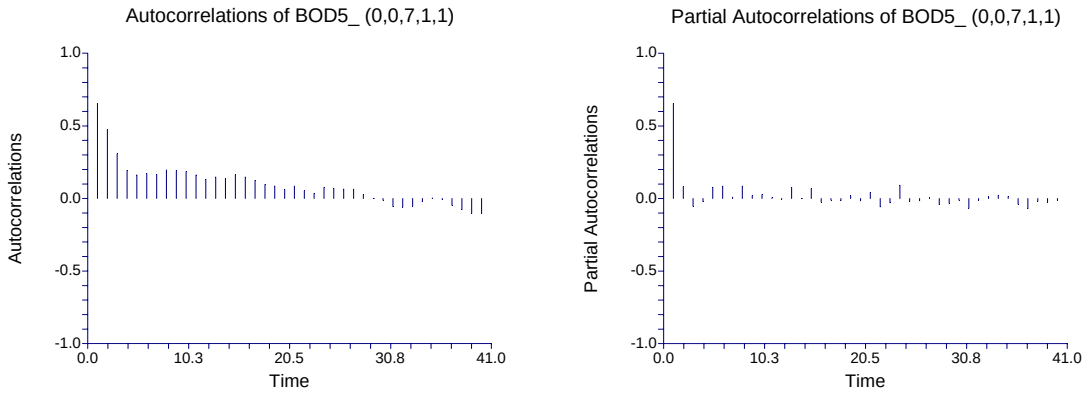


Figure 3-11. The autocorrelation plots of ACFs and PACFs of the daily BOD_5 series.

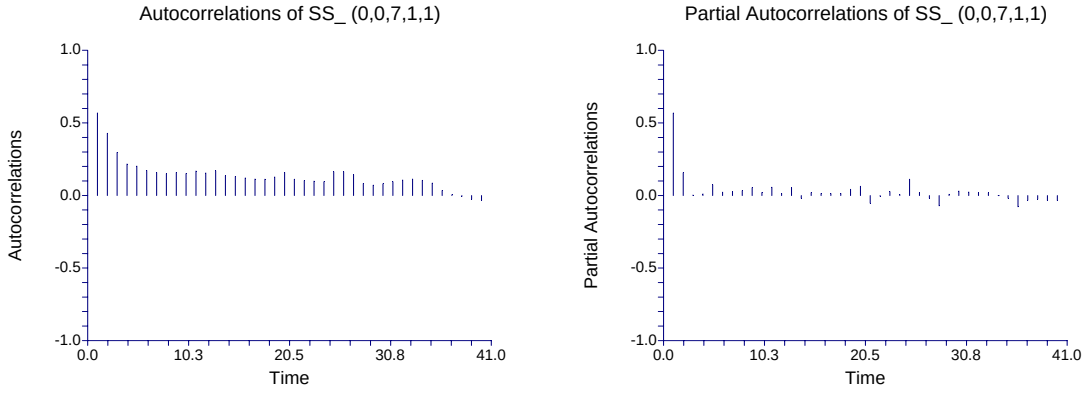


Figure 3-12. The autocorrelation plots of ACFs and PACFs of the daily SS series.

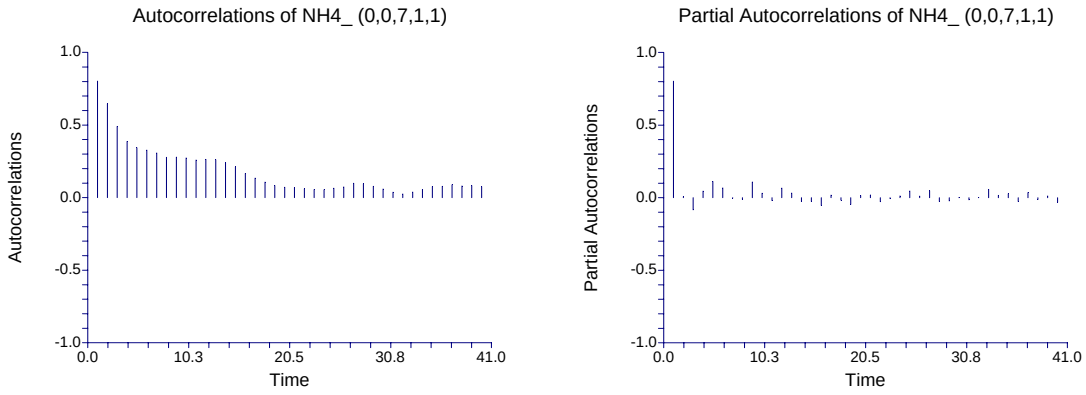


Figure 3-13. The autocorrelation plots of ACFs and PACFs of the daily NH_4-N series.

indicate that there is strong autocorrelation at lag 1 [AR (1)]. Neither daily nor monthly seasonality is significant. Those plots also show that there are no significant trends existing in those series. Notice that the ACF and PACF plots are very helpful to estimate which time series model might be good for the selected time series data. However, the final best-fitted model might be a little different from what we expected from ACF and PACF plots. For example, as shown in the section 3.3.3, the flow rate model is best fitted with a daily seasonality (i.e., the days of a week effect).

The models were designed to include cross correlation between variables, as well as autocorrelation within each variable. The Spearman correlation coefficients (Table 3-7) indicate that flow rate and temperature, BOD_5 and flow rate, NH_4-N and flow rate, and NH_4-N and BOD_5 are significantly correlated ($r>0.30$).

Table 3-7. Spearman correlations of the five time series.

	Temperature	Flow rate	BOD_5	SS	NH_4-N
Temperature	1.0000	-0.4557	0.0809	-0.0652	0.2836
Flow rate	-0.4557	1.0000	-0.3958	-0.0192	-0.5911
BOD_5	0.0809	-0.3958	1.0000	0.1527	0.3974
SS	-0.0652	-0.0192	0.1527	1.0000	0.1399
NH_4-N	0.2836	-0.5911	0.3974	0.1399	1.0000

The cross-correlation coefficients (Table 3-8) indicate that there is a lag 1 pattern existing in correlated groups: flow rate/temperature, BOD_5/flow rate and NH_4/flow rate. In reality, the typical diurnal pattern of wastewater flow and concentration are similar and the concentration pattern is about three hours delay than the flow rate pattern (Dold and Marais, 1986). However, the plant data are daily not hourly time series so that the hourly delay pattern is not expected in this research. Notice that flow rate, BOD_5, and NH_4 are correlated in a circled. Thus, only BOD_5/flow rate and NH_4/flow rate are considered in this research. Notice that the cross-correlation coefficients of Lag (0) (absolute values) are consistently smaller than Spearman correlation coefficients, as shown in Tables 3-6 and 3-7. This is due to the some outliers in the data series.

Table 3-8. Cross-correlation results of the possible correlated series.

Items	Lag(0)	Lag(1)	Lag(2)	Lag(3)
Flow rate/Temperature	-0.4117	-0.3943	-0.3943	-0.3957
BOD_5/Flow rate	-0.3575	-0.3944	-0.3770	-0.3474
NH_4-N/Flow rate	-0.5004	-0.5095	-0.4771	-0.4445
NH_4-N/BOD_5	0.3375	0.2824	0.2189	0.1475

Furthermore, temperature and flow rate are easily measured data, which can be recorded timely and sometimes automatically. However, as for SS, NH_4-N, and BOD_5 data, they are relatively more complicated and commonly their values are not available immediately. For example, the traditional method of BOD_5 takes five days to measure wastewater samples. As a result, temperature and flow rate are regarded as priory absolute inputs (independent variables). Notice that flow rate may be an output with temperature as an input, and be an input for both BOD_5 and NH_4-N, which depends how much the model is improved to decide whether we need this correlation.

3.3.3 Time Series Models

Once the forms of the model equations were known, the parameters were estimated using least-squares regression. Models were evaluated by employing statistics such as R^2, Root

Mean Square Error (RMSE), and Mean Absolute Percent Error (MAPE) to determine how well the model fit the data. In addition, the white noise and normality tests of the residuals were used to detect outliers and determine forecasting confidence intervals. Parameter estimation and model evaluation were conducted using two different procedures. In the first procedure, the full data record from November 1999 through July 2002 was used for both parameter estimation and performance diagnosis. In the second procedure, to verify the forecasting ability of the models, data were then split into a calibration data set spanning the period November 1999 through April 2002 and a holdout data set spanning the period May 2002 through July 2002. The calibration data set was used for parameter estimation and the holdout set was used to verify the ability of the models to forecast. The best models were selected on the basis of these evaluations. Additional details of the procedures used to build time series models can be found in standard textbooks (Prankratz, 1991; DeLurgio, 1998).

Table 3-9 shows the time series models of the five variables with November 1, 1999 to June 31, 2002 as both fit and evaluation range, which focus on the model fit. And Table 3-10 shows the time series models of the five variables with November 1, 1999 to April 30, 2002 as fit range and May 1, 2002 to July 31, 2002 as evaluation range, which focuses on the forecasting. Note that the variable temperature is well fitted without trend and seasonality, which is probably due to the slow-changing pattern in this series. The models of flow rate and NH_4 are also reasonable with $R^2>0.60$. BOD_5 and SS will be the key problems to be addressed.

Table 3-9. Time series models of the five time series with fit range November 1, 1999 to July 31, 2002 and evaluation range November 1, 1999 to July 31, 2002.

Series Name	Model Label	RMSE	MAPE	R^2
Temperature	*Damped Trend Exponential Smoothing	0.9511	3.6253	0.934
Flow rate	*Log ARIMA (2,0,0) $(1,0,0)_7$	102839.4	8.7492	0.620
	Log [Temperature +ARIMA(2,0,0) $(1,0,0)_7$]	101691.3	8.7174	0.628
	#Log [Temperature +ARIMA(1,0,0) $(1,0,0)_7$]	101873.2	8.7841	0.627
	Log {Temperature[N(1)/D(1)]+ARIMA(1,0,0)$(1,0,0)_7$}	101928.1	8.7978	0.627
	Log {Temperature[N(1)/D(1)]+ARIMA(2,0,0)$(1,0,0)_7$}	102261.6	8.8322	0.624
BOD_5	*Simple Exponential Smoothing	16.6540	18.3341	0.395
	# Flowrate[N(1)/D(1)]+AR(1)	15.3611	18.2972	0.486
	Flowrate[N(1)/D(1)]+AR(1) + Linear Trend	15.2139	17.6496	0.495
SS	Log Simple Exponential Smoothing	24.0172	19.0722	0.353
	*Simple Exponential Smoothing (M1)	24.2204	18.6755	0.342
	#Linear Trend + AR(1) +Point:31JAN2002 (M2)	20.5594	18.2940	0.526
	Combination: [0.12 M1 +0.88 M2]	20.5112	18.0015	0.529
NH_4	*Simple Exponential Smoothing (M1)	2.4940	15.4878	0.625
	# Flowrate[/D(1)]+AR(1) (M2)	2.3094	14.6447	0.678
	Flowrate[/D(1)] + AR(1) + Points: 02JUN & 03JUN2002[M3]	2.1848	14.7128	0.712
	Combination: [0.5 M1+0.5 M2]	2.3592	14.6556	0.665
	Combination: [0.5 M1+0.5 M3]	2.2565	14.5720	0.693

Table 3-10. Time series models of the five time series with fit range November 1, 1999 to April 30, 2002 and evaluation range May 1, 2002 to July 31, 2002.

Series Name	Model Label	RMSE	MAPE	R^2
Temperature	*Damped Trend Exponential Smoothing	1.2998	4.5511	0.772
Flow rate	*Log ARIMA (2,0,0) $(1,0,0)_7$	138175	9.6536	0.471
	Log [Temperature +ARIMA(2,0,0) $(1,0,0)_7$]	136309	9.6009	0.485
	#Log [Temperature +ARIMA(1,0,0) $(1,0,0)_7$]	135899	9.5680	0.488
	Log Temperature[N(1)/D(1)]+ARIMA(1,0,0)$(1,0,0)_7$	136536	9.7577	0.484
	Log Temperature[N(1)/D(1)]+ARIMA(2,0,0)$(1,0,0)_7$	137024	9.7757	0.480
BOD_5	*Simple Exponential Smoothing	12.2102	17.8990	0.529
	Flowrate[N(1)/D(1)]+AR(1)	12.7544	24.2151	0.486
	#Flowrate[N(1)/D(1)]+AR(1) + Linear Trend	12.4147	21.1794	0.513
SS	Log Simple Exponential Smoothing	29.6683	16.4552	0.420
	*Simple Exponential Smoothing (M1)	29.6079	16.0206	0.422
	# Linear Trend + AR(1) +Point:31JAN2002 (M2)	29.3667	13.8298	0.432
	Combination: [0.12 M1 +0.88 M2]	20.6017	18.0876	0.525
NH_4	*Simple Exponential Smoothing (M1)	3.8863	11.4872	0.682
	#Flowrate[/D(1)]+AR(1) (M2)	3.7713	10.5442	0.701
	Combination: [0.5 M1+0.5 M2]	2.3540	14.7499	0.666

After evaluating several potential models for each variable, recommended models were selected and are summarized in Table 3-11. Recommendations for simple univariate models and more complicated multivariate models are indicated in the table with by * and # symbols, respectively. In section 3.3.4, we will use the best model (complicated model with symbol #) to derive the equations needed for the random influent generator.

Predicted time series based on the simple models for flow rate, BOD_5, SS, and NH_4-N are compared to the actual data in Figures 3-14, 3-15, 3-16, and 3-17. The models are seen to track the data very well with great accuracy. However, there is a lag 1 problem between the predicted and real data. That is, it seems that the predicted data are one day delayed.

Table 3-11. Recommended time series models for the five variables. (Evaluation statistics are for the period November 1, 1999 to July 31, 2002)

Series Name	Model Label	RMSE	MAPE	R^2	W/N
Temperature	*#Damped Trend Exponential Smoothing	0.9511	3.63	0.934	Yes/No
Flow rate	*Log ARIMA (2,0,0) $(1,0,0)_7$	102839.4	8.75	0.620	Yes/No
	#Log[Temperature+ARIMA(1,0,0)$(1,0,0)_7$]	101873.2	8.78	0.627	Yes/No
BOD_5	*Simple Exponential Smoothing	16.6540	18.33	0.395	No/No
	#Flow rate[N(1)/D(1)]+AR(1)	15.3611	18.30	0.486	No/No
SS	*Simple Exponential Smoothing (M1)	24.2204	18.68	0.342	No/No
	#Linear Trend+AR(1)+Point:1/31/02 (M2)	20.5594	18.29	0.526	No/No
	Combination: [0.12 M1 +0.88 M2]	20.5112	18.00	0.529	No/No
NH_4	*Simple Exponential Smoothing	2.4940	15.49	0.625	No/No
	#Flow rate[/D(1)]+AR(1)	2.3094	14.64	0.678	No/No

Note:
1) * simple univariate model; # more accurate and more complicated multivariate model
2) W/N indicates the White Noise/ Normality tests of the residuals.

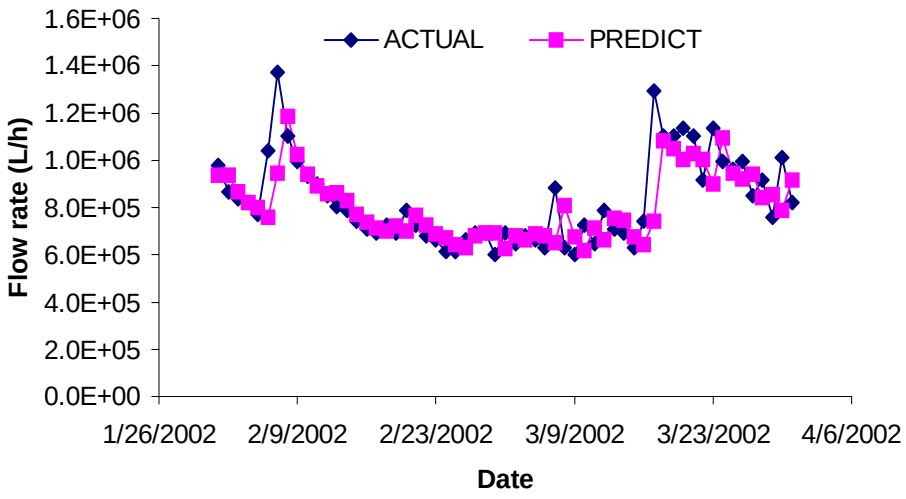


Figure 3-14. Flow rates forecasts (simple model) compared to the actual values.

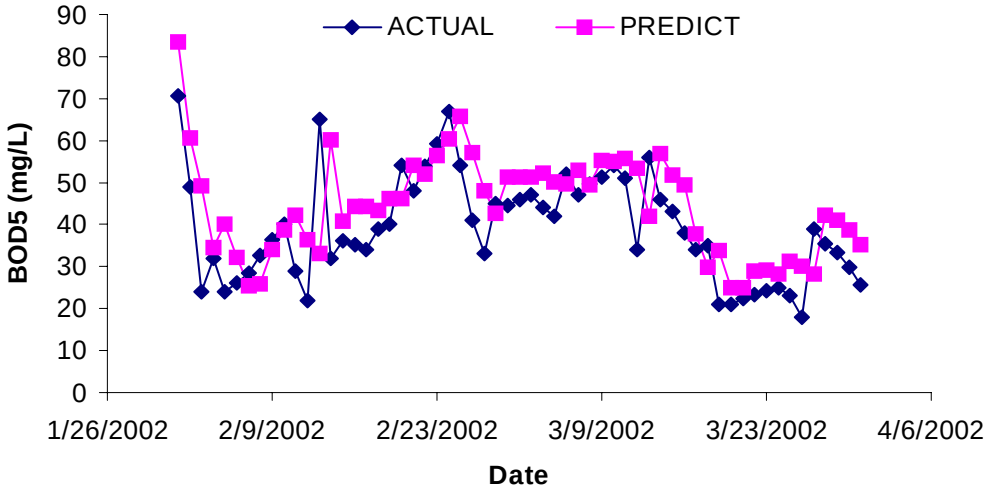


Figure 3-15. BOD_5 forecasts (complicated model) compared to the actual values.

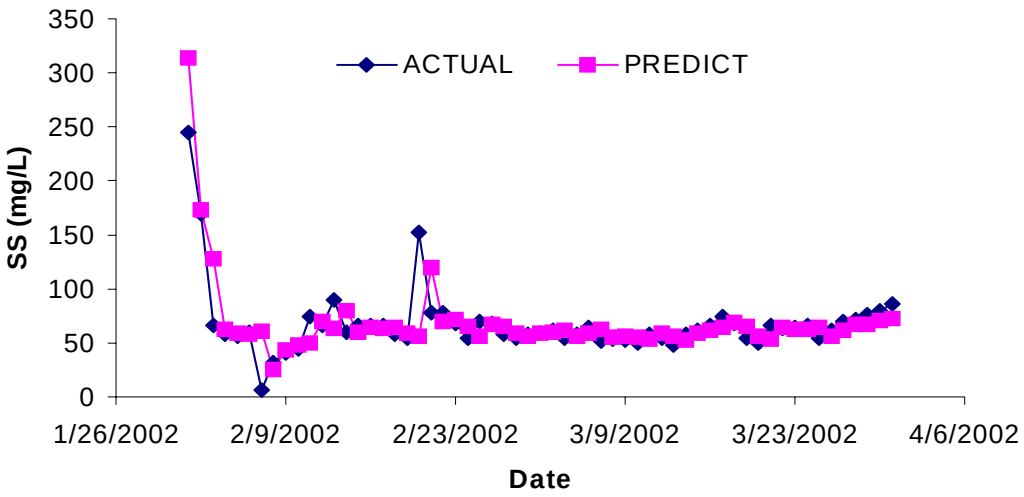


Figure 3-16. SS forecasts (complicated model) compared to the actual values.

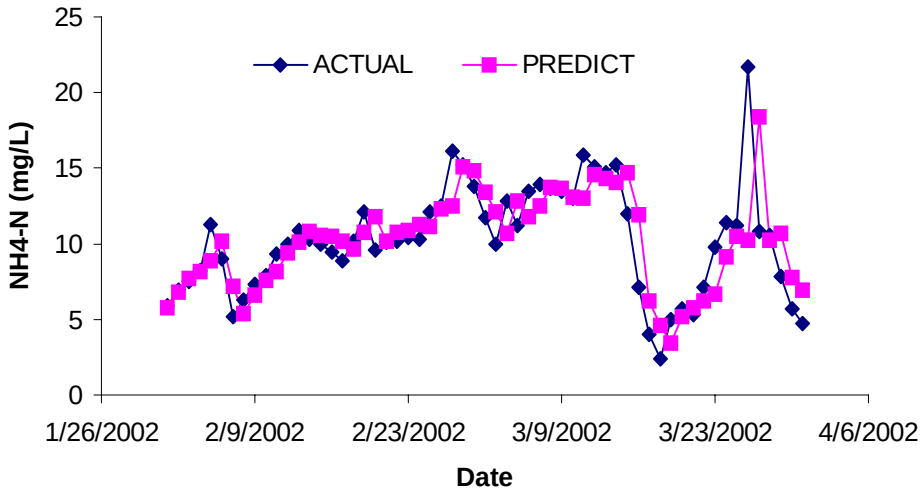


Figure 3-17. NH_4-N forecasts (complicated model) compared to the actual values.

This can be explained with the equations of the fitted time series models. All time series models are related with the values in the past, either its own past values, or past values of their correlated variables. This error-driven time series models determine that our predicted values forecast the values in the future based on the past values, thus making a strong lag 1 pattern as shown in Figures 3-14 to 3-17.

3.3.4 Random Generator of Wastewater Influent Conditions

We will now demonstrate how time series models, such as those described above can be used to generate random influent time series that have the same statistical characteristics as the historical influent data. We believe that these simulated time series can be in conjunction with dynamic wastewater simulation software to test plant performance under a wider variety of realistic operating conditions than possible using the historical data alone. In this example we will demonstrate the method for predicting realistic influent flow rate, BOD_5, SS, and NH_4-N concentrations. We are interested in capturing the correlations between flow and BOD_5 and between flow and NH_4-N, according to Table 3-7 and 3-8. Under these conditions, the univariate SS, and multivariate flow rate, BOD_5 and NH_4-N models are appropriate.

Table 3-12 shows the fitted parameters of appropriate time series models. The corresponding equations will be derived in the following sections.

Table 3-12. Parameter values for the fitted time series models.

Variable	Model	Parameter name	Parameter value
Flow rate	Log [Temperature + ARIMA(1,0,0)(1,0,0)$_7$]	Intercept (μ) Autoregressive, Lag 1 (ϕ_1) Seasonal Autoregressive, Lag 7 (ϕ_7) Temperature (β)	13.7573 0.7429 0.1352 -0.01646
BOD_5	Flow rate[N(1)/D(1)]+AR(1)	Intercept (μ) Autoregressive, Lag 1 (ϕ_1) Flow rate[N(1) / D(1)] (ω_0) Flow rate [N(1) / D(1)] Num1 (ω_1) Flow rate [N(1) / D(1)] Den1 (δ_1)	102.2267 0.6120 -1.2732E-05 1.4875E-05 0.5618
SS	Linear Trend +AR(1)+Point:1/31/02	Intercept (μ) Autoregressive, Lag 1 (ϕ_1) Point:31JAN2002 (γ) Linear Trend (β)	58.9363 0.6182 304.9423 0.02916
NH_4-N	Flow rate [/D(1)]+AR(1)	Intercept (μ) Autoregressive, Lag 1 (ϕ_1) Flow rate [/ D(1)] (ω_0) Flow rate [/ D(1)] Den1 (δ_1)	23.5297 0.7292 -3.8771E-06 0.7590

3.3.4.1 Random flow simulation

The Log [Temperature + ARIMA (1,0,0) $(1,0,0)_7$] notation (SAS, 1999) in Table 3-12 indicates that the natural-log of the flow rate is characterized by one significant auto-regression lags in flow rate, as well as a 7-day seasonality resulting from the daily pattern of sewage flow. The number 1 shown in the first bracket indicates the order of the autoregression is 1. And the number 1 in the second bracket means that the order of the seasonality autoregression is 1. Furthermore, this model is also a linear relationship with variable Temperature, which counts the seasonal pattern in the data set. The time series model of flow rate is:

$$\ln Q_t = \mu + \beta \times Temp + \frac{1}{(1-\phi_1 \mathrm{B}^1)(1-\phi_7 \mathrm{B}^7)}\alpha_t \qquad (3\text{-}20)$$

where Q_t is the flow, β is the coefficient for variable Temperature, α_t is the error, B is the backshift operator (for example, $B^nQ_t = Q_{t-n}$) and the subscript indicates the time t at which the variable is evaluated. Equation (3-20) can be expanded to yield:

$$\ln Q_t - \mu = (\phi_1 \mathrm{B}^1 + \phi_7 \mathrm{B}^7 - \phi_1\phi_7 \mathrm{B}^8)(\ln Q_t - \mu) + (1-\phi_1 \mathrm{B}^1 - \phi_7 \mathrm{B}^7 + \phi_1\phi_7 \mathrm{B}^8)\beta Temp + \alpha_t \qquad (3\text{-}21)$$

Parameter values for the model are given in Table 3-12. The predicted values of flow rate in Figure 3-14 were obtained by substituting the actual flow data for the lagged variables Q_{t-1}, Q_{t-2}, Q_{t-7}, Q_{t-8}, and Q_{t-9} in equation (3-21) and assuming that the error term is negligible.

We introduce two modifications of the application of equation (3-21) in order to generate a new random time series for Q_t that possesses the same statistical characteristics as the original time series: 1) we substitute *past predictions,* instead of *actual data* for the lagged values of Q_t and 2) we do not assume that the error α_t is negligible; rather we substitute a random value of the error sampled from an appropriate distribution. Without the latter modification, the errors will accumulate to the extent that the predicted values of Q_t will either increase unbounded or decrease to zero. An accurate estimate of the distribution of α_t is needed. In this work, we calculated α_t as the difference between the actual and predicted values in Figure 3-14. We then sampled randomly from this population of error estimates without replacement (i.e., each value of α_t is only used once). An example random time series generated by this procedure is compared to the actual time series in Figure 3-18. Recall that the goal is to produce a random time series

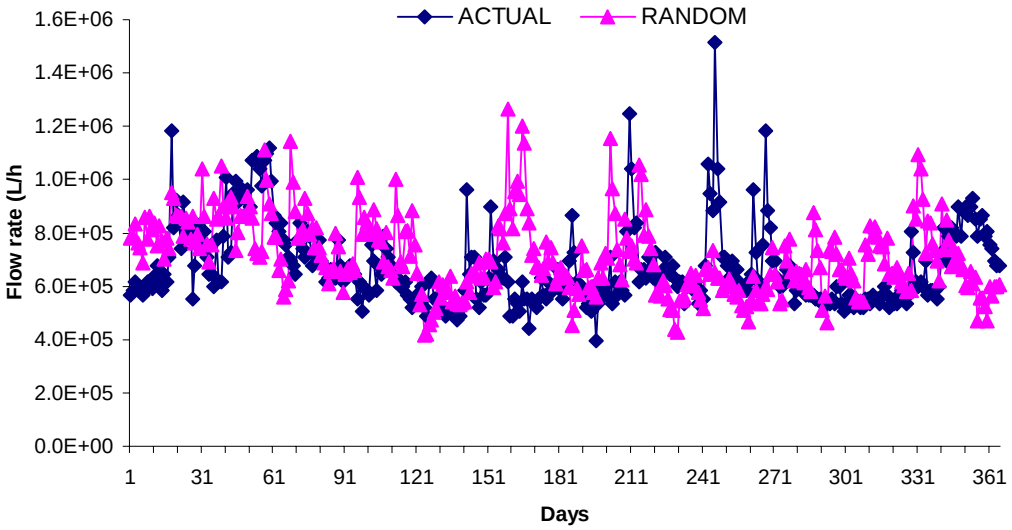


Figure 3-18. Comparison of randomly simulated and actual flow rates.

that is different from the original time series, but possesses the same statistical characteristics. We will evaluate the extent to which the goal has been achieved after developing a similar model for BOD_5, SS, and NH_4-N.

3.3.4.2 Random BOD_5 simulation

The multivariate model for BOD_5, summarized in Table 3-12 with the notation Flow rate [N(1)/D(1)]+AR(1), is a dynamic regression model with a flow rate transfer function characterized by single-lag numerator and denominator terms and a single-lag autoregressive term. The parameterized dynamic regression model is:

$$BOD_t = \mu + \frac{\omega_0 - \omega_1 \mathrm{B}^1}{1 - \delta_1 \mathrm{B}^1} Q_t + \frac{1}{\left(1 - \phi_1 \mathrm{B}^1\right)} \alpha_t \tag{3-22}$$

where BOD_t is the BOD_5 evaluated at time t. Equation (3-22) can be expanded to yield:

$$\begin{aligned} BOD_t - \mu &= \left[\left(\phi_1 + \delta_1\right)\mathrm{B}^1 - \phi_1 \delta_1 \mathrm{B}^2\right]\left(BOD_t - \mu\right) \\ &+ \left[\omega_0 - \left(\omega_1 + \phi_1 \omega_0\right)\mathrm{B}^1 + \phi_1 \omega_1 \mathrm{B}^2\right] Q_t + \left(1 - \delta_1 \mathrm{B}^1\right)\alpha_t \end{aligned} \tag{3-23}$$

Equation (3-23), with the parameter values in Table 3-12, is the basis for generating random time series of BOD_5 with similar statistical characteristics as the actual data. A procedure similar to that used for the flow model was used for BOD_5. Furthermore, values of Q_t obtained from the random flow model are used in equation (3-23). An example random time series for BOD_5 generated by this procedure is compared to the actual time series in Figure 3-19. The occasional negative and other exceeding low values

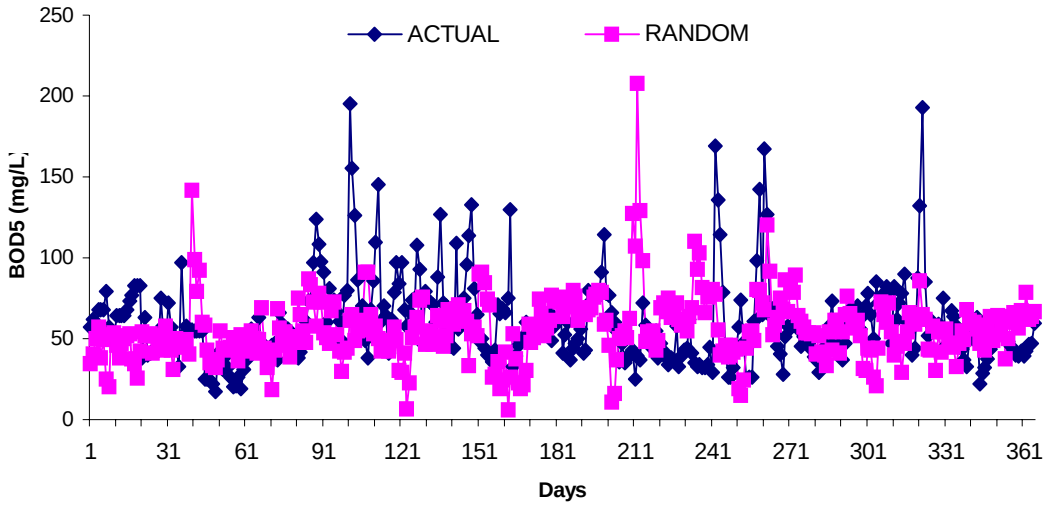


Figure 3-19. Comparison of randomly simulated and actual BOD_5 concentrations.

of BOD_5 in the randomly simulated time series could easily be replaced by some minimum value.

3.3.4.3 Random SS simulation

The univariate model for SS, summarized in Table 3-12 with notation: Linear Trend +AR(1) + Point: 1/31/2002. This model includes one linear trend term, one autoregressive term, and one intervention term. The intervention term is a point one, which means this intervention happened only at the date of January 31st, 2002. The rest pattern of the time series would not be affected. The parameterized regression model is:

$$SS_t = \mu + \beta SS_{t-1} + \frac{1}{1-\phi_1 B^1} a_t \qquad (3\text{-}24)$$

where SS_t is the SS evaluated at time t. Equation (3-24) can be expanded to yield:

$$SS_t - \mu = \phi_1 B^1 (SS_t - \mu) + (1-\phi_1 B^1)\beta SS_{t-1} + a_t + \gamma \times dummy \qquad (3\text{-}25)$$

where, $\gamma \times dummy = \begin{cases} 0, & Not \quad an \quad \text{int}\, ervention. \\ \gamma, & is \quad an \quad \text{int}\, ervention. \end{cases}$

Equation (3-25), with the parameter values in Table 3-12, is the basis for generating random time series of SS with similar statistical characteristics as the actual data. A procedure similar to that used for the flow model was used for SS. An example random time series for SS generated by this procedure is compared to the actual time series in Figure 3-20. The occasional negative and other exceeding low values of SS in the randomly simulated time series could easily be replaced by some minimum value.

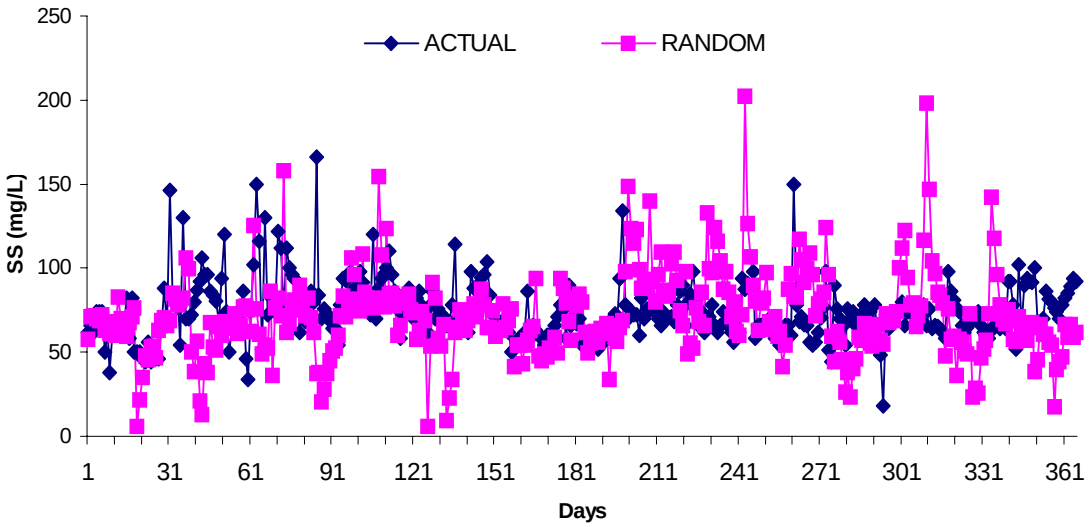


Figure 3-20. Comparison of randomly simulated and actual SS concentrations.

3.3.4.4 Random NH_4-N simulation

The multivariate model for NH_4-N, summarized in Table 3-12 with the notation Flow rate [/D(1)]+AR(1), is a dynamic regression model with a flow rate transfer function characterized by single-lag denominator term and a single-lag autoregressive term. The parameterized dynamic regression model is:

$$NH_t = \mu + \frac{\omega_0}{1-\delta_1 \mathrm{B}^1} Q_t + \frac{1}{\left(1-\phi_1 \mathrm{B}^1\right)} \alpha_t \qquad (3\text{-}26)$$

where NH_t is the NH_4-N evaluated at time t. Equation (3-26) can be expanded to yield:

$$NH_t - \mu = [(\phi_1 + \delta_1)\mathrm{B}^1 - \phi_1\delta_1 \mathrm{B}^2](NH_t - \mu) + (\omega_0 - \phi_1\omega_0 \mathrm{B}^1)Q_t + (1-\delta_1 \mathrm{B}^1)\alpha_t \qquad (3\text{-}27)$$

Equation (3-27), with the parameter values in Table 3-12, is the basis for generating random time series of NH_4-N with similar statistical characteristics as the actual data. A procedure similar to that used for the flow model was used for NH_4-N. Furthermore, values of Q_t obtained from the random flow model are used in equation (3-27). An example random time series for NH_4-N generated by this procedure is compared to the actual time series in Figure 3-21. The occasional negative and other exceeding low values of NH_4-N in the randomly simulated time series could easily be replaced by some minimum value.

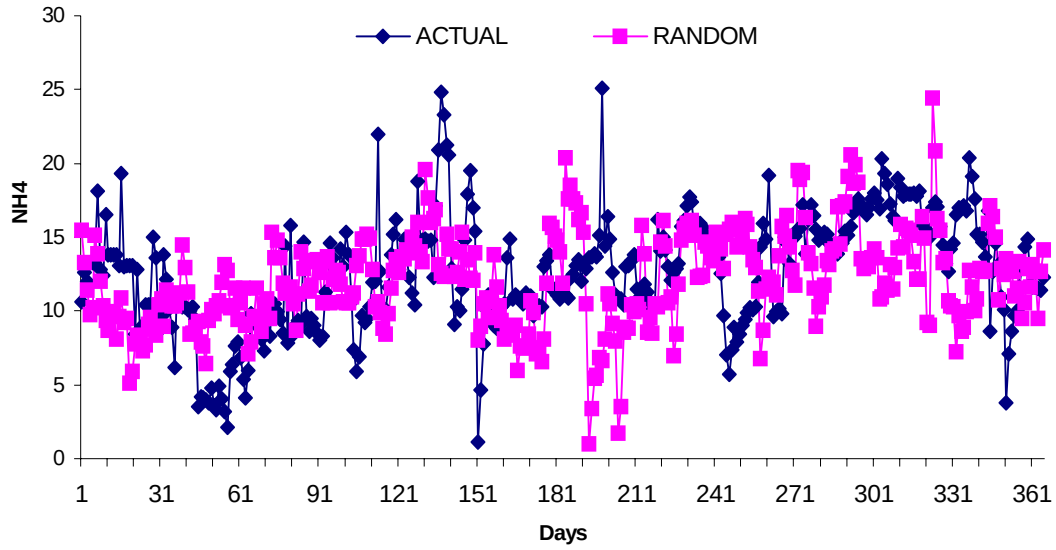


Figure 3-21. Comparison of randomly simulated and actual NH_4-N concentrations.

3.3.5 Model Evaluation and Discussion

Statistical characteristics of the randomly simulated time series for all four variables are compared in Table 3-13. The mean, standard deviation, and correlation coefficient between flow rate and BOD_5 and between flow rate and NH_4-N are in excellent agreement with the original data for all four variables. Auto-correlation functions for the randomly simulated and actual data are compared for all four variables (Figures 3-22, 3-23, 3-24 and 3-25). The agreement is quite good through the first several lags. Based on the similarity of the statistical characteristics, it can be concluded that the random influent generator models provide a realistic simulation of the influent to the Oak Ridge treatment plant.

Table 3-13. Comparison of descriptive statistics of actual, predicted, and randomly simulated data.

Variables	Statistics	ACTUAL	PREDICT	RANDOM
Flow rate (L/h)	Mean	707051.26	705966.36	702999.07
	Standard deviation	167584.84	130783.11	147636.78
BOD_5 (mg/L)	Mean	57.701	57.715	57.704
	Standard deviation	21.550	14.993	19.964
	Correlation (with flow rate)	-0.3553	-0.5113	-0.2837
SS (mg/L)	Mean	73.668	68.879	73.928
	Standard deviation	29.861	19.0126	30.2191
NH_4-N (mg/L)	Mean	12.148	12.153	12.148
	Standard deviation	4.086	3.368	3.762
	Correlation (with flow rate)	-0.4985	-0.6064	-0.3830

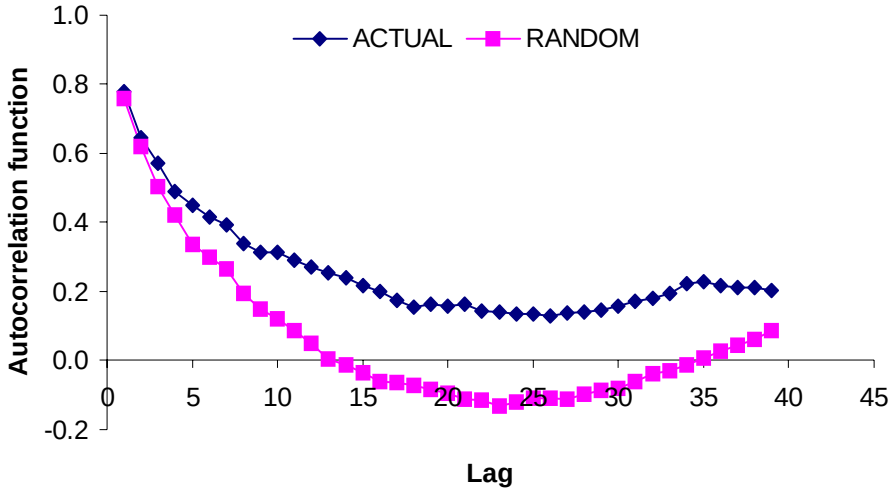


Figure 3-22. Comparison of autocorrelation function for randomly simulated and actual values of flow rate.

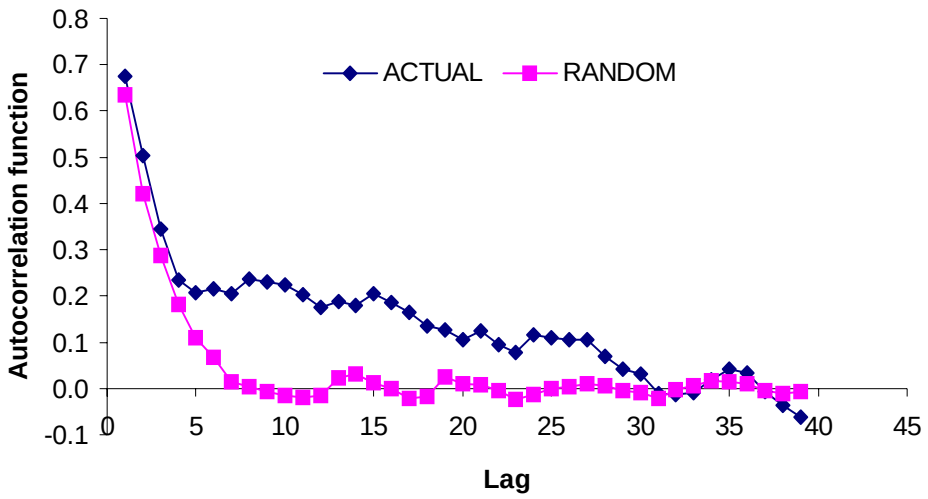


Figure 3-23. Comparison of autocorrelation function for randomly simulated and actual values of BOD_5.

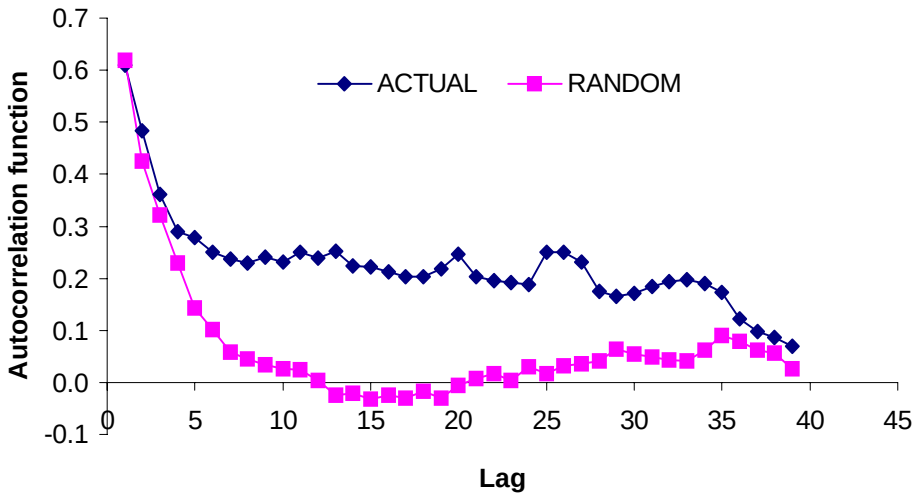


Figure 3-24. Comparison of autocorrelation function for randomly simulated and actual values of SS.

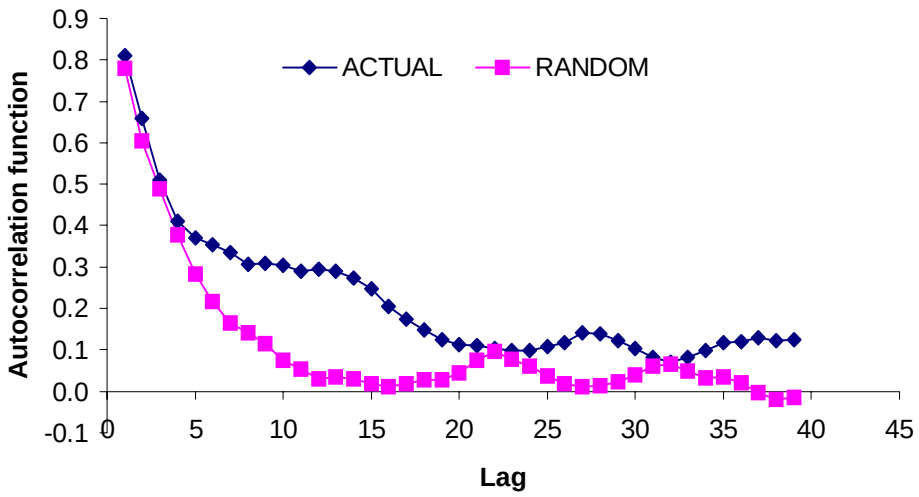


Figure 3-25. Comparison of autocorrelation function for randomly simulated and actual values of NH_4-N.

The random influent generator models would be used as simulators of plant influent in Chapter V. Models should also be developed for other treatment plants to validate their general applicability. Additional research is also needed to properly specify the error distribution. In this dissertation, a distribution consisting of the actual error values is used. This method is limited to fairly extensive data sets. A parametric distribution would have the advantage of allowing error values not present in the original data set. The errors from the time series models for the Oak Ridge WWTP were not normally distributed, but no other parametric distributions were investigated. Further research is also needed to define the minimum size data set needed to identify and calibrate the model. We also want to investigate how manipulations of the model parameters and the error term can be

used to simulate future influent conditions for which data may not yet exist (e.g., greater mean or greater variance).

As mentioned above, there are three methods to introduce the plant data to the simulation model. The first one is that we use the original data directly. The second one is that we use forecast values from the time series models. And the third one is that we use generated random influent based on the time series models. In this dissertation, only first and third methods are actually applied in the StatASPS package. This is because the second influent is virtually identical to the historical influent data. If we want our influent data as close to real data as possible, the real plant would be the first choice. In other words, there is no need taking this more complicated model without any significant benefits. In fact, randomly generated influent may be desired in some cases because it can generate influent data different from the historical data, yet still possessing similar statistical characteristics. The second method, time series model, provides the best estimate of real plant influent data, which can be used in one specific year's influent data. This estimate smoothes out possible outliers, and provides a relatively accurate base line for the third method: random influent generator. It is not uncommon that the original plant data have outliers (sudden extremely high or low values). The causes for outliers are probably from either recording mistakes or unexpected events (for example, storms, sudden operating failures, etc.). However, it is hard to tell the outlier sources unless more information is available. In order to bring back the uncertain information (for example, possible outliers) in the real plant influent data, the random influent generator is applied to generate one time series influent with reasonable white noise added. The goal of this

dissertation is to create a method to analyze the uncertainty of the plant performance. Thus, uncertainty (also called variability) in the plant influents should be quantified and included in the analysis. Time series model offers a method to well describe the time series characteristics of the plant influents.

CHAPTER IV

MISSING DATA AND OUTLIERS REPLACEMENT METHODS

4.1 Introduction

Missing values commonly exist in the operational records of wastewater treatment plants. For example, unexpected events cause the failure of measurements of plant data, or routinely holidays make the measurements unavailable. Missing values of wastewater treatment influent variables and effluent variables (BOD_5, SS, NH_4-N, etc.) are very common. These variables are also defined as time series variables because they are recorded at successive times (most of them are daily data). As discussed in previous chapters, time series data are used to describe the variability of plant influent data, which is one of two most important uncertainty sources in wastewater treatment systems. Furthermore, predicted influent data from time series models and randomly generated influent data are both based on the historical time series data. Thus, a reasonable and reliable missing data replacement method is highly needed to determine the correct variability of plant influent.

Three methods are commonly used to deal with missing values in time series. First, the simplest method is to replace all missing values for a given variable with the mean, median, or other location statistics (e.g., percentiles) determined from the non-missing

values of the variable (DeLurgio, 1998). The second easiest method is to replace missing values with the average of the nearest future and/or past observation, also called as the ANO (Average Nearest Observation) method in this dissertation. This method is employed as the default method for missing data replacement in the NCSS® software system. The third method, employed in the SAS® system, relies on fitting a satisfied time series model to the data to predict missing values. Applications of the later two methods are discussed below.

In the ANO method, missing values are replaced by the average of the nearest observation in the future and in the past. For example, for the time series [96, 88, missing1, missing2, 140, 148], the following steps in replacing the missing values would be made: [96, 88, missing1, missing2, 140, 148]$\Rightarrow$ [96, 88, missing1= (88+140)/2=114, missing2, 140, 148] $\Rightarrow$ [96, 88, 114, missing2=(114+140)/2=127, 140, 148] $\Rightarrow$ [96, 88, 114, 127, 140, 148]. Obviously, this method will give different replacements values if the time series occurs in the opposite order. For example, the new time series [148, 140, missing1, missing2, 88,96] becomes [148, 140, 114, 101, 88,96]. Notice that this missing value replacement method might be particularly poor for time series data with weak autocorrelation and/or strong daily seasonality. If this happens, manually entering a more reasonable estimate before using the algorithm is strongly recommended.

The application procedure of the third method is as follows: (1) replace all missing values with a simple imputed value, such as the series mean; (2) with the updated estimates added, forecast the missing values from the current time series model; (3) then update the

original data set with the new forecasts of missing values; (4) repeat step 2 until the convergence occurs.

However, the methods described so far depend on the MCAR (missing completely at random) assumption, which means that the probability that an observation is missing does not depend on the data set. Unfortunately, MCAR observations are less likely in time series data because observations are usually correlated with time series. For example, there are virtually no data available for Fridays, Saturdays, and other holidays of the influent data from the Oak Ridge wastewater treatment plant (Oak Ridge WWTP). This situation commonly exists in wastewater treatment systems, and is common in other data collection and recording systems as well. Thus, a new method to deal with those routinely missing data is badly needed.

Based on the information we have, these three methods are not suitable for use in those three series (BOD_5, SS and NH_4-N). The first two simple methods are too simple to be chosen for the procedure, and the last one is also not a good choice in this case because the missing values in these three variables routinely happen on Fridays, Saturdays, and other holidays. Thus, in order to find the suitable estimates of missing values, we use the simple exponential smoothing method, an automatic fitting model available in most of commonly used commercial software (SAS®, NCSS®, JMP®, etc.). Furthermore, in order to remove the effect of time order, a Two-directional Exponential Smoothing method (TES method) is used in which, forward ES and backward ES are applied independently and the mean values of those two forecasts are used to replace the missing

values. A TESWN (Two-directional Exponential Smoothing with White Noise added) method is also developed in which a random white noise term is added after the TES prediction is obtained. The goals upon which the various methods are evaluated are: 1) the original existing data should remain unchanged; and 2) the replaced values for the missing data should conform to the overall pattern of the time series. In order to evaluate the second goal, we will conduct the scatterplot of both original and updated (replaced) data to determine whether the updated data visually follows the pattern of original data. Furthermore, we will compare the statistical characteristics (for example, mean, standard deviation, and percentiles, etc.) of both original and updated data. Finally, we will also apply the updated influent data to the ASM model to determine how different the updated (replaced) data will make to ASM model predictions compared to the original influent data. The histogram plot, mean, standard deviation and percentiles will be the major concerns to evaluate missing data replacement methods.

Two plant influent data are used in this dissertation. The first plant data are from the Seneca WWTP, Germantown, MD. It is chosen because this plant has an almost complete record of daily data for a period of three years (1993-1995). We will create a test data set by intentionally removing data points for Fridays and Saturdays. This method can be tested by comparing the generated replacement points with the original data points. The influent component SS (suspended solid, mg/L) is discussed in details as a sample in this section. Other variables are performed following the same procedure.

The second plant is the Oak Ridge WWTP. Because the critical concern is the activated sludge system, the effluent from the secondary clarifier is considered as the influent to the activated sludge process. Almost three-year's daily data (from November 1, 1999 to July 31, 2002) were collected to analyze the statistical characteristics of five influent variables: Temperature (wastewater temperature in the bioreactor, °C), Flow rate (influent, L/h), BOD_5 (5-day biological oxygen demand, mg/L), SS (suspended solids, mg/L), and NH_4-N (ammonia nitrogen, mg/L). As mentioned above, this data set was the motivation for developing methods to replace missing data points. This data set has routinely missing values for Fridays, Saturdays, and other holidays. The TES data are expected to predict the influent characteristics on days with missing influent data in a manner consistent with the overall statistical characteristics of the data record.

In short, the Seneca influent data are a test data set applied in this section to prove how well our TES and TESWN methods perform. In practice, the Oak Ridge plant influent data with real missing values will be replaced using the TES and/or TESWN methods. This generated TES influent data will then be either directly used as a plant influent or used to generate random influent for the StatASPS package.

4.2 Methods

4.2.1 Flow Diagram of TES and TESWN Methods

In this dissertation, the Two-directional Exponential Smoothing (TES) and TES with White Noise (TESWN) methods are developed to replace routinely missing values from the data of wastewater treatment plants (WWTPs). Figure 4-1 describes the TES and TESWN procedures. Step 0-1 is applied to the Seneca data set in order to generate a data set with missing values for the purpose of evaluating the method. Normally, when the data set already has missing values, the procedure begins with step 0-2. The applications of TES and TESWN methods will be discussed in the following sections.

4.2.2 Introduction to Exponential Smoothing Methods

TES and TESWN methods both depend on exponential smoothing methods. In order to explain how these two methods can be applied to replace missing data, Holt's Linear Trend exponential smoothing algorithm is explained in detail. Holt's Linear Trend algorithm computes a local trend equation through the data using a weighting function that places the greatest emphasis on the most recent time periods. Both the trend and forecasting equations change from period to period. The formulas used in this algorithm are as follows:

$$F_{t+k} = a_t + b_t k \tag{4-1}$$

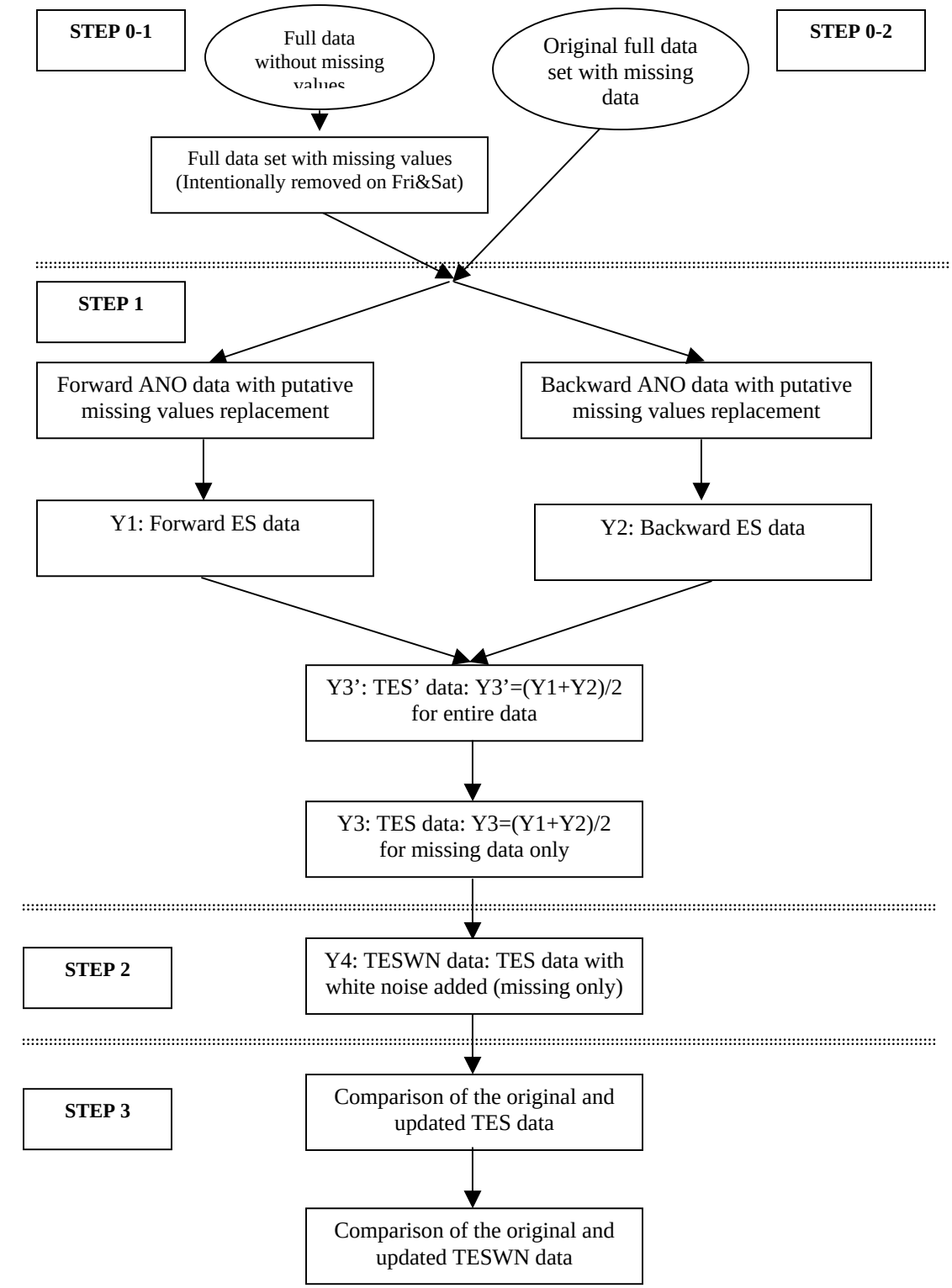


Figure 4-1. The flow diagram of the assessment and testing of the TES and TESWN missing data replacement methods.

where,

$$a_t = \delta Y_t + (1-\delta)(a_{t-1} + b_{t-1}) \quad (4\text{-}2)$$

$$b_t = \gamma(a_t - a_{t-1}) + (1-\gamma)b_{t-1} \quad . \quad (4\text{-}3)$$

and F_{t+k} is the forecast value at time t+k; Y_t is the actual value at time t; a_t and a_{t-1} are y-intercepts at time t and t-1, respectively; b_t and b_{t-1} are slopes at time t and t-1, respectively; and δ and γ are smoothing constants that are both between 0 and 1.

As shown in the equations above, initial values of a_{t-1} and b_{t-1} are needed since the forecast at time 1 requires the values at period 0. The backcasting method, which is currently known as one of the best methods, is adopted to generate initial values. This backcasting method simply reverses the direction of the series and forecasts it into the past. Thus, it gives the initial values for the intercept (a_{t-1}) and slope (b_{t-1}).

The smoothing constants δ and γ determine the weight given to the most recent past observations and therefore control the rate of smoothing or averaging. Values near 1 give virtually all the weight to the most recent data, while values near 0 distribute the weights to consider data from the more distant past data. The values of δ and γ can be determined either subjectively or objectively. If reliable experience with this or a similar time series is available, we can choose the values of the smoothing constants subjectively to tune the forecast to our own beliefs about the future of the series. There are several criteria to objectively determine the values of smoothing constants, such as the mean square error (MSE), the mean absolute error (MAE), and the mean absolute percent error

(MAPE). The criterion used in this dissertation is the mean square error (MSE) defined as:

$$MSE = \frac{1}{n}\sum e_t^2 = \frac{1}{n}\sum (Y_t - F_t) \qquad (4\text{-}4)$$

where e_t is the error at time t, Y_t is the actual value at time t, and F_t is the forecast value (TES') at time t. Equation 4-4 is applied only to forecast values for which the actual data values have been measured (e.g. Sunday – Thursday for the Oak Ridge data).

The goal is to find values of the smoothing constants that minimize MSE of the time series. A search for the proper values using an efficient grid-searching algorithm is accomplished in this dissertation.

4.2.3 TES Method

The TES method estimates of missing data points based on the autocorrelations of the time series to account for the fact that the missing values occur at non-random times. As the ANO method described above, exponential smoothing yields different values depending on the direction of the time series; while the TES method can minimize this directional bias. This TES method, which is designed to represent both forward and backward autocorrelation in the time series, can also decrease the difference above caused by different directions. To execute the method, firstly, generate a putative full data set using the ANO method. Secondly, forecast the missing values using the exponential

smoothing (ES) algorithm (in this case, Holt's Linear Trend exponential smoothing) in the forward direction (Forward ES). Next, the missing values are forecasted using Holt's Linear Trend exponential smoothing algorithm in the backward direction (Backward ES). The final replacement values for the missing data points are determined by averaging the forward and backward ES estimates. In fact, the TES method is a combination time series method.

The TES data are generated from the TES (Two-directional Exponential Smoothing) method. The equation is shown as follows:

$$TES_t = \begin{cases} TES'_t = \dfrac{(ES_{forward,t} + ES_{backward,t})}{2}, & \text{if data is missing,} \\ Original\ \ value_t, & \text{if data is available.} \end{cases} \quad (4\text{-}5)$$

Notice that the TES' data are the average of the forward ES data and the backward ES data for the whole data set. However, the TES data are the average of the forward ES data and the backward ES data for those missing values only.

4.2.4 TESWN Method

The TESWN method is known as the TES method with white noise added. It begins by applying the TES method and then adding to it a white noise term to account for random effects observed in the data but not captured by the autocorrelation function.

The TESWN data are generated from the TESWN method, that is, the TES data with white noise added for missing values only. The equation is shown as follows:

$$TESWN_t = \begin{cases} TES_t + WhiteNoise_t, & \text{if data is missing,} \\ Original\ \ value_t, & \text{if data is available.} \end{cases} \quad (4\text{-}6)$$

The white noise is randomly sampled from a normal distribution. The standard deviation of the normal distribution is calculated from the standard deviations of both the forward ES and backward ES forecast data. The characteristics of the white noise and the selection of the standard deviation will be discussed in Section 4.3.

4.3 Application of TES and TESWN Methods

TES and TESWN methods will be applied to the influent data from the Seneca WWTP. As mentioned above, the Seneca data are used to test these two methods following the step 0-1 in Figure 4-1. A test data set is generated by removing data points for Fridays and Saturdays of the Seneca data. The those two missing data replacement methods can be tested by comparing the generated replacement points with the original data points

4.3.1 TES Method

The TES method was assessed by comparing replacement values for the missing data to the actual data from the Seneca WWTP. A comparison of the original values and the replaced values using the TES method is shown in Figures 4-2 to 4-4. These three plots only show a short portion of the total Seneca WWTP time series to facilitate readability.

The replacement SS values for Fridays and Saturdays determined by the TES method are relatively close to the original values. It should be noted that the missing values determined by the algorithm are not necessarily bounded by the range between the known

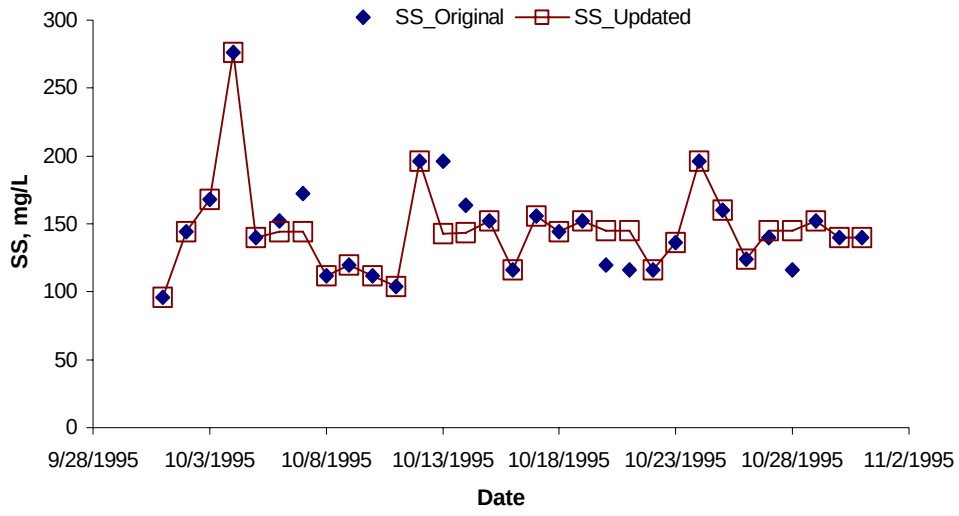


Figure 4-2. The partial scatter plot of the original and updated SS from the Seneca WWTP in October 1995.

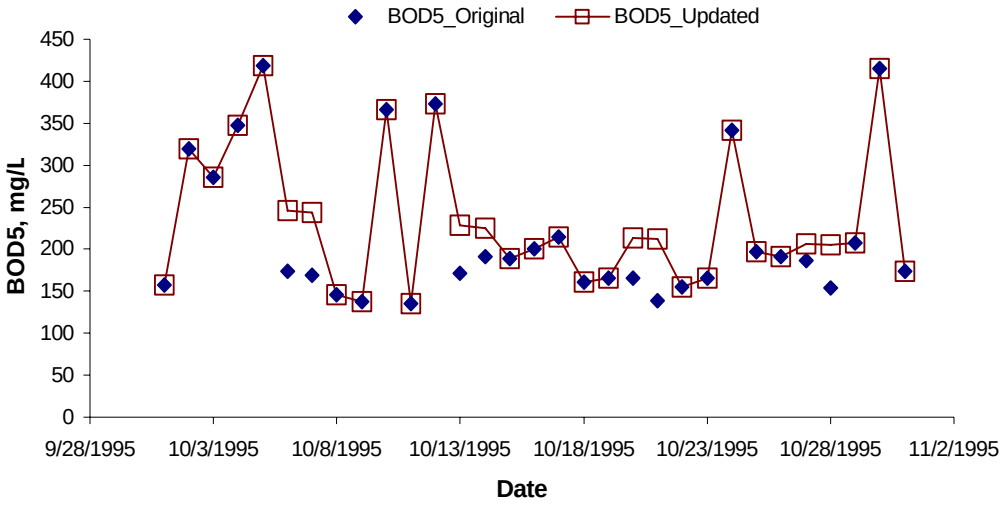


Figure 4-3. The partial scatter plot of the original and updated BOD_5 from the Seneca WWTP in October 1995.

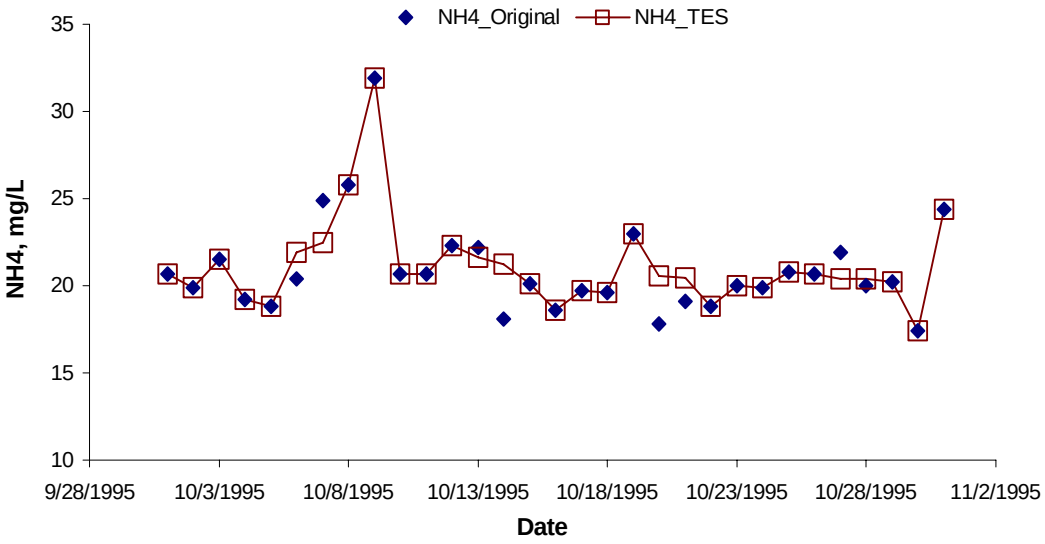


Figure 4-4. The partial scatter plot of the original and updated NH_4-N from the Seneca WWTP in October 1995.

data points preceding and following the section of missing data. For example, the updated values of October 6th and October 7th 2004 are 144.45 mg/L and 143.89 mg/L, respectively. These two values fall outside the range of the October 5th and 8th data points: 140 mg/L and 112 mg/L, respectively. Despite the relatively good estimates of the missing values, the TES method poorly represents the variability of the original data. For example, the missing data of November 10th, 1995 is 144.47 mg/L, instead of the original value of 100 mg/L. If such unexpected changes are common in the plant data, further corrections might be required (for example, seasonality). Overall, however, the comparison between the original values and the updated values appears adequate with no obvious evidences to deny our TES method. More rigorous evaluation of the TES method will be conducted in comparison with the TESWN method.

4.3.2 TESWN Method

As we discussed above, the TES method is only a combined forecast method to replace the missing values. In order to capture variability not represented in the overall trend, the TESWN method will be applied in the section.

The error between the TES' predictions and the actual data can be considered a source of white noise in the time series. It may be possible to characterize this white noise, then add it to the TES' predictions to obtain predictions of missing data that include the variability in the data that is not described by the local trend. We first determined the distribution of

the white noise for the Friday and Saturday missing values by calculating the error of the TES' predictions, which is shown in Figure 4-5 a). It can be seen that the white noise distribution for Friday and Saturday is different from the white noise distribution of the whole data set by comparison to Figure 4-5 f). Normally, the actual values will not be available for missing data points, so an alternative method of estimating the white noise distribution must be established. We will use the standard deviation of the distribution in Figure 4.5 a) as the target to evaluate various estimates of the white noise.

The best estimate of the white noise was obtained by estimating the error as the difference between the TES' predictions and the ANO predictions for the whole data set, shown in Figure 4-5 c). Both terms in the error estimate are available for each day of the week. The standard deviation of this error estimate (49.27) compares favorably with the target standard deviation of 46.68. The second best estimate of the white noise was obtained by estimating the error as the difference between the TES' predictions and the actual data points for the days on which they are available (Sunday through Thursday), shown in Figure 4-5 b). The standard deviation (53.17) of this error estimates was not quite as close to the target value, but this method has the advantage that actual data values, rather than a forecast, are used to predict the error. Other methods of predicting the error shown in Figure 4-5 d) and e) were less accurate.

An implicit assumption of characterizing white noise as the error from the days in which measurements are made is that the standard deviations of each days of the week are not significantly different. Figure 4-6 is a statistical test of equal variance. From these tables

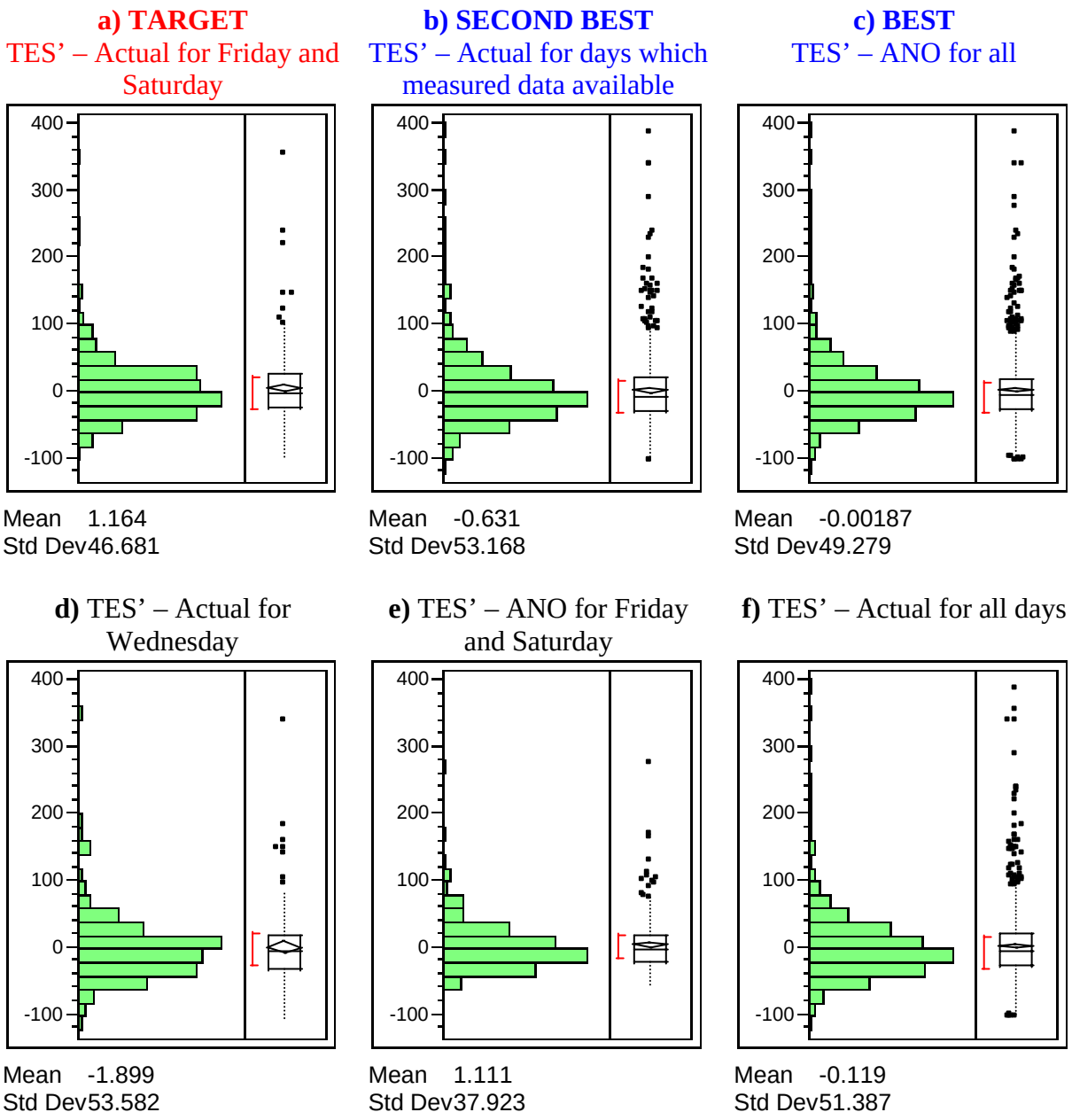
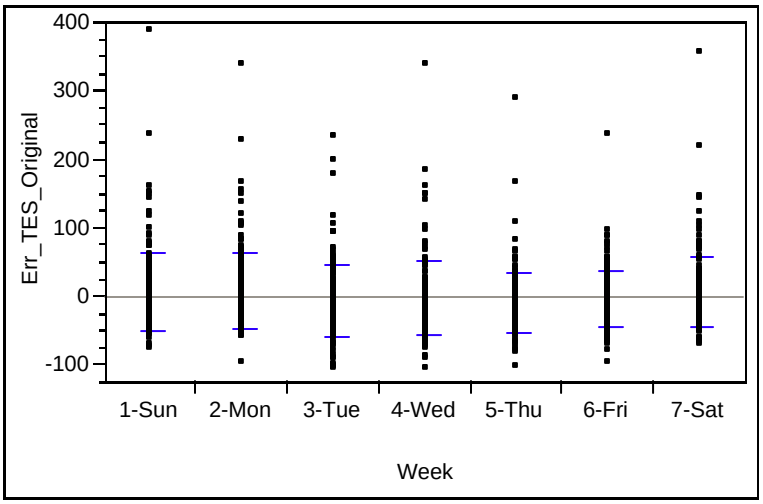
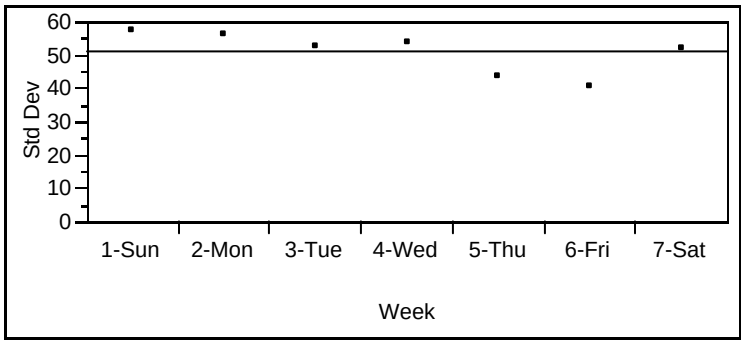


Figure 4-5. The histogram plots of error items with Fridays and Saturdays all missing.



Level	Count	Std Dev
1-Sun	155	57.10
2-Mon	155	56.36
3-Tue	156	52.38
4-Wed	156	53.58
5-Thu	155	43.61
6-Fri	155	40.29
7-Sat	155	51.92



Test	F Ratio	DFNum	DFDen	Prob > F
O'Brien[.5]	0.6861	6	1080	0.6609
Brown-Forsythe	1.4997	6	1080	0.1748
Levene	1.9732	6	1080	0.0667
Bartlett	5.0173	6	.	<.0001

Note: the p-values of “Prob>F” column show the probability of no significant difference.

Figure 4-6. The equal variance test for every 7 day in a week.

and plots, we can conclude that there is no significant difference among these 7-day standard deviations of error. This test gives the TESWN method strong support to use the standard deviation of error of the whole data set except missing values to estimate the standard deviation of error of missing values (in this case, all Fridays and Saturdays).

Based on the above analysis, we decided to estimate the error as the difference between the TES' predictions and the actual data for the days which measurements were taken. The white noise term for each prediction was then obtained by sampling from a normal distribution with mean zero and the standard deviation of this error estimate.

Statistical characteristics of TES and TESWN predictions for the Friday and Saturday suspended solids data from the Seneca plant are compared to the original data in Table 4-1 and Figure 4-7. For the TESWN, three different realizations are shown based on different random samplings of the white noise terms. In cases where the TESWN predictions were negative, this value was replaced by the minimum value in the data record. The TESWN method describes the statistical uncertainty (standard deviation) of original influent SS data better than the TES method. However, upper percentile values of the distribution of original data are much greater than those predicted by either the TES or TESWN methods. Apparently, the assumed normal distribution of the white noise term cannot fully capture the variability in the original data. The test of normal distribution is conducted with the error from the original data, as shown in Figure 4-8. From the plot, we can see the error term is not normally distributed. However, due to lack of information, it is the best available assumption we can make. Thus, in this paper, we

Table 4-1. The statistical characteristics of influent SS data: original, TES and three randomly generated TESWN data.

Items	**$SS_{Original}$**	**TES**	**$TESWN_1$**	**$TESWN_2$**	**$TESWN_3$**
Number	310	310	310	310	310
Mean	147.69	146.52	148.89	143.42	145.88
Std Dev	48.56	17.55	50.43	54.90	54.45
100.0%	516.00	194.65	287.04	294.47	285.96
99.5%	433.86	194.28	282.34	292.38	282.76
97.5%	271.60	190.71	250.40	263.94	247.36
90.0%	199.60	172.56	211.17	213.73	215.62
75.0%	169.00	156.77	183.39	176.22	185.03
50.0%	140.00	144.46	151.80	144.81	147.20
25.0%	116.00	133.42	117.16	108.63	109.61
10.0%	100.00	125.91	81.55	71.03	73.13
2.5%	84.00	116.47	39.76	36.67	31.96
0.5%	70.22	110.58	9.67	6.51	12.79
0.0%	68.00	110.36	9.66	5.97	8.81

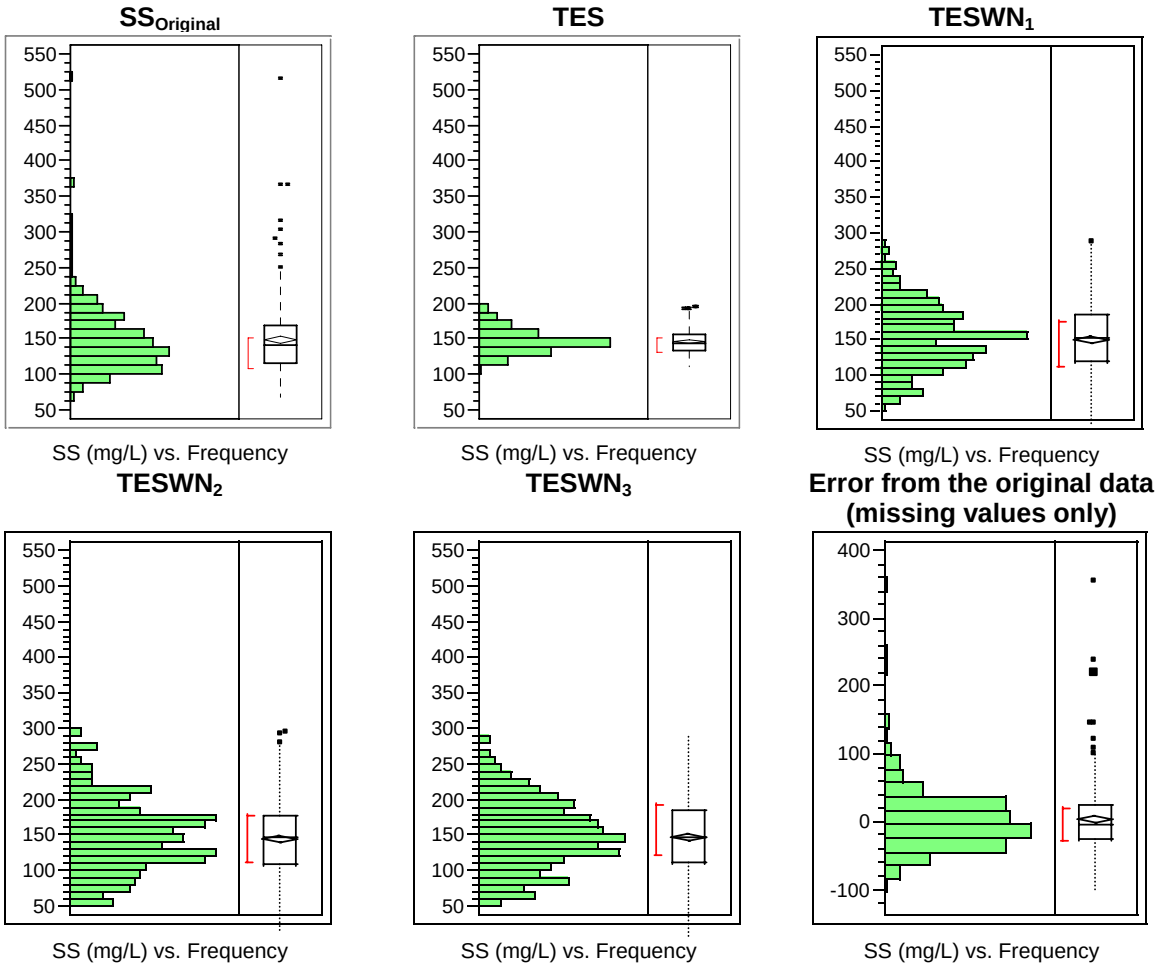


Figure 4-7. The histogram graphs of influent SS data: original data, TES data, three randomly generated TESWN data, and error from the original data (missing values only).

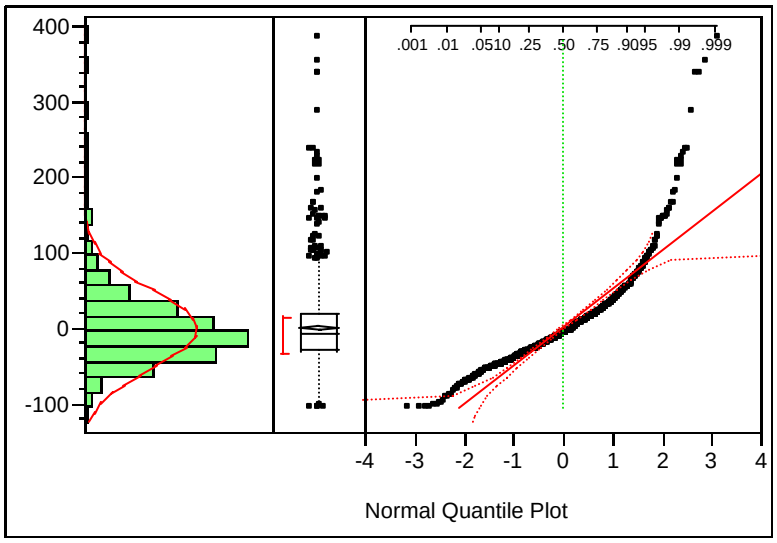


Figure 4-8. The histogram and normal quantile plot of the error between the TES' data and the original data.

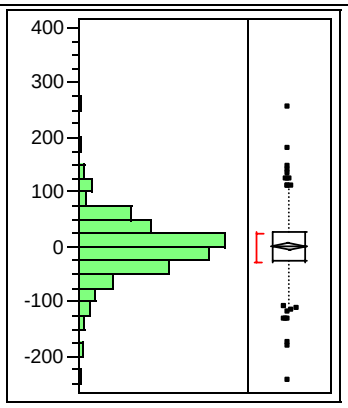
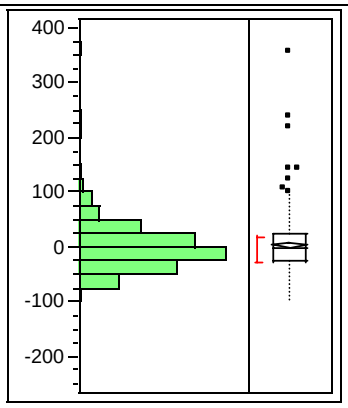
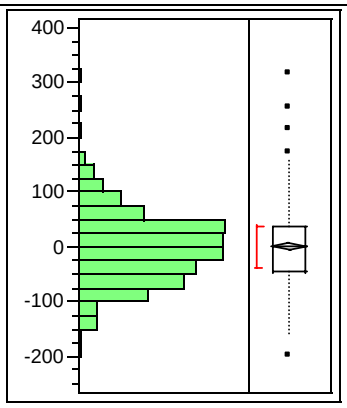
assume that the distribution of the error is a normal distribution, with the mean value (μ) equals zero and the standard deviation (σ) equals the standard deviation of error of the whole data expect for those missing values. It is shown in the following equation.

$$\text{Distribution of error} = \text{Normal Distribution}\ (0, \sigma^2_{\text{whole\_non-missing}}) \qquad (4\text{-}7)$$

4.4 Assessment of TES and TESWN Methods

4.4.1 Assessment of Error Terms

Figure 4-9 shows the comparison of the error term among three methods (ANO, TES, $TESWN_1$). The near zero mean and smallest standard deviation indicate the ANO might

Error of the ANO data	Error of the TES data	Error of the $TESWN_1$ data
		
SS (mg/L) vs. Frequency	SS (mg/L) vs. Frequency	SS (mg/L) vs. Frequency

Quantiles			Quantiles			Quantiles		
100.0%	maximum	255.0	100.0%	maximum	354.4	100.0%	maximum	317.0
99.5%		213.4	99.5%		289.0	99.5%		281.3
97.5%		121.2	97.5%		101.8	97.5%		130.2
90.0%		54.0	90.0%		48.1	90.0%		75.9
75.0%	quartile	26.3	75.0%	quartile	22.8	75.0%	quartile	35.6
50.0%	median	0.0	50.0%	median	-4.4	50.0%	median	-1.0
25.0%	quartile	-28.0	25.0%	quartile	-27.9	25.0%	quartile	-45.2
10.0%		-58.0	10.0%		-44.4	10.0%		-80.5
2.5%		-116.9	2.5%		-65.5	2.5%		-131.0
0.5%		-211.6	0.5%		-87.8	0.5%		-178.6
0.0%	minimum	-246.0	0.0%	minimum	-97.3	0.0%	minimum	-198.6

Moments		Moments		Moments	
Mean	-0.4355	Mean	1.1635	Mean	-1.2072
Std Dev	53.6834	Std Dev	46.6811	Std Dev	65.4503
N	310	N	310	N	310

Figure 4-9. The comparison of errors of the ANO, TES, and TESWN data.

be the best method to fill in the missing data if accurate estimates of each individual point is the goal. However, the TESWN method appears to be best if the goal is to capture the overall variability in the distribution, as described above. As for the TES and TESWN data, they have much higher absolute value of the maximum error than that of the minimum error. This might be caused by the possible outliers in the original SS data (the whole scatter plot shows there are some values higher than 400 mg/L, and even 500mg/L).

4.4.2 Assessment of Simulated Results

The comparison and testing above have demonstrated that the TES and TESWN methods are feasible alternatives to replace missing data. However, our goal is to introduce this plant influent data as the input data for the ASM1 model. In this case, we are most concerned with how the effluent predicted by the model is affected by the various estimates of missing data. Since ammonia-N is the main concern all through this dissertation, the TES and TESWN data of ammonia-N were generated using the similar procedure with SS data. The simulation conditions were set as: SRT = 5 days and HRT = 5 hours. Under these conditions, relatively frequent violations of effluent are expected. The influent components are from the Seneca WWTP. However, this time, the influent components are kept unchanged expect for ammonia-N. Five cases of influent data were simulated in this section: original, AVE, ANO, TES, and TESWN data. The AVE data was generated by replacing the missing values with the average values of the remained

data set. January-May 1994 and January-May 1995's data are chose as the input of the ASM1 model. Because we are most concerned about the effect of the missing data, the following analysis only consider those missing data for Fridays and Saturdays although the analyses for all available data are also completed.

The influent ammonia-N data are plotted in Figure 4-10. From the plot, we can see there were some variations in early 1994 and early 1995 as expected.

The statistical analysis and comparison of the simulated data from those five cases are shown in Figure 4-11 (Missing data only, hourly data). The mean effluent ammonia predicted by ASM1 when the original data are used as the influent is 1.755 mg/L. Obviously the AVE data provides the best estimates (1.736 mg/L) because the missing values are replaced by the mean value of the whole data set. The second and third best

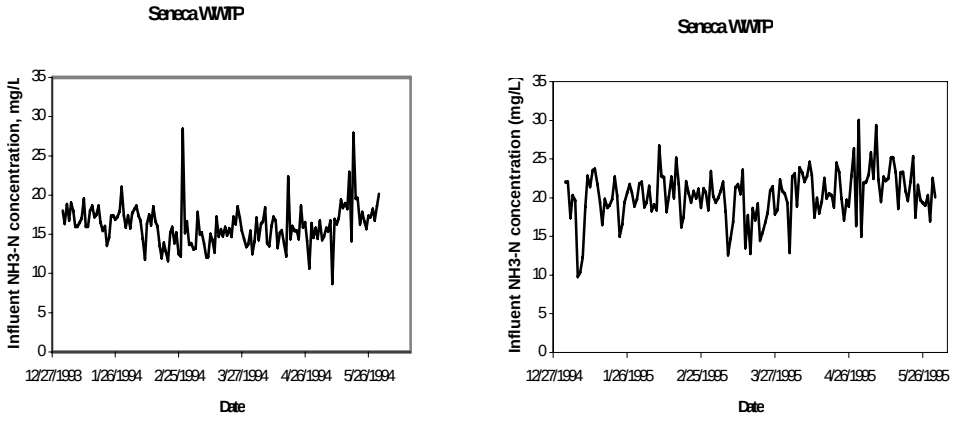
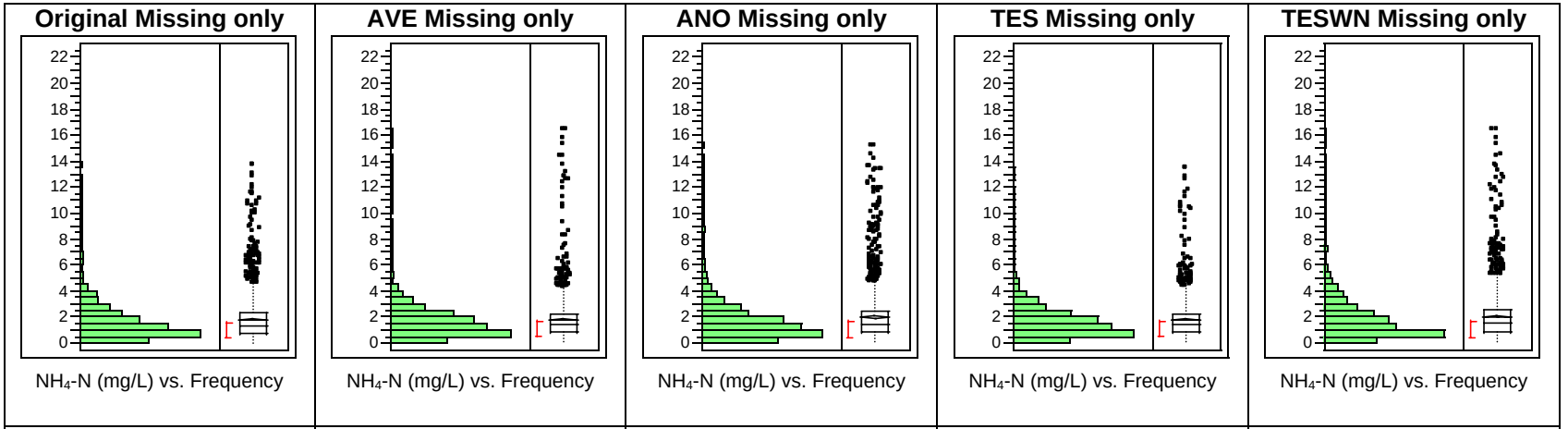


Figure 4-10. The scatter plot of the influent ammonia nitrogen data in the Seneca WWTP.



Quantiles	Original Missing only	AVE Missing only	ANO Missing only	TES Missing only	TESWN Missing only
100.0% maximum	13.725	16.436	15.166	13.494	16.435
99.5%	11.019	12.753	13.132	10.404	12.841
97.5%	6.633	5.135	8.561	4.940	7.063
90.0%	3.585	3.197	3.762	3.246	3.999
75.0% quartile	2.232	2.183	2.342	2.197	2.546
50.0% median	1.258	1.416	1.386	1.406	1.479
25.0% quartile	0.696	0.795	0.767	0.778	0.743
10.0%	0.425	0.480	0.422	0.480	0.485
2.5%	0.198	0.222	0.220	0.206	0.236
0.5%	0.000	0.033	0.065	0.045	0.042
0.0% minimum	0.000	0.000	0.000	0.000	0.000
Moments					
Mean	1.755	1.736	1.933	1.714	1.976
Std Dev	1.693	1.574	1.988	1.417	1.896
N	2040	2040	2040	2040	2040

Figure 4-11. The Statistical analysis and comparison of the ASPS simulation data. (1994 + 1995 hourly data, missing only)

would be TES and ANO data that have mean values of 1.714 and 1.933 mg/L, respectively. The standard deviation of original data are 1.693. The best estimate of 1.574 is still from the AVE data. The second and third best estimates are TESWN and TES data with values of 1.896 and 1.417. Till now, it seems that the best method would be the AVE method followed by the TES method. However, recall the motivation of introducing the missing data replacement method is to generate a plant influent data for the StatASPS model. The goal is to study on the uncertain characteristics of the plant performance. Thus, the greatest estimates of effluent ammonia nitrogen (e.g., 00, 99.5, 97.5, percentiles) should be considered in additional to the mean and standard deviation will be the most important factors to be considered besides the mean and standard deviation. For example, the 100, 99.5 and 97.5 percentile values of ammonia nitrogen predicted from the original data are 13.725, 11.019, and 6.633 mg/L. The closed estimates from the data replacement methods came from the TES data with values of 13.494, 10.404, and 4.940 mg/L, respectively. The second best is from AVE data, which has the values of 16.436, 12.753, and 5.135 mg/L, respectively.

The correlation table and scatter plot matrix are shown in Table 4-2. It shows the predicted effluents from the TES data had the largest correlation coefficient (0.8538) when compared with predicted effluents using the original data, followed by the AVE data (0.8288). In short, the simulated results indicate the best two methods are the TES and AVE methods. It is a surprise that the AVE data performs so well for the simulation results. Furthermore, the TES data replace the missing values very well, as shown visually in Figures 4-2 to 4-4 and provide a stronger theoretical support. The AVE

Table 4-2. The correlations between the five simulated effluent Ammonia-N.

	Original Data	**AVE Data**	**ANO Data**	**TES Data**	**TESWN Data**
Original Data	1.0000				
AVE Data	0.8288	1.0000			
ANO Data	0.6612	0.7422	1.0000		
TES Data	0.8538	0.9545	0.8484	1.0000	
TESWN Data	0.7481	0.8459	0.7221	0.8417	1.0000

method may be selected in cases where the influent is to be used for plant simulation and a simple-to-apply method is desired. Otherwise, we would strongly recommend the TES method as the best choice to replace the missing values.

4.5 Discussions and Conclusions

In this chapter, the TES and TESWN methods, compared with ANO and AVE methods, were conducted to replace the missing data (including those routinely missing values, like all Fridays and Saturdays and other holidays). The following conclusions are made based on application and evaluation of the methods using influent data from the Seneca WWTP:

- The TES method proves to be the best of the methods tested to replace missing values when the influent data are to be used for simulating plant performance. The generated TES data has better scatter plot, better correlation, and better mean and standard deviation. More important, it generates the best estimates of the three highest percentiles of the simulated effluent data using the StatASPS package.
- The AVE method is acceptable if simplicity of application is a high priority.
- The TESWN method describes the statistical characteristics of original influent SS data better than the TES method.

CHAPTER V

CASE STUDIES AND RESULT ANALYSIS

In this chapter, we develop and test methods to quantify uncertainty at two existing wastewater treatment plants: Oak Ridge WWTP, Oak Ridge, Tennessee, and Seneca WWTP, Germantown, Maryland. The purpose of this chapter is to how to best capture the plant performance using the ASM1 model and Monte Carlo method together with appropriate descriptions of the variability of the model parameters and plant influent.

As discussed in Chapter III, two sources of uncertainty in WWTPs are considered in this dissertation: process parameters and plant influent. For process parameters, we have three options: fixed calibrated parameters, universal parameter distributions, and site-specific parameter distributions. The calibrated parameters are obtained by fitting models to historical performance data. In the procedure, the RMSE (root mean square error) is minimized using the enumerative search method. The RMSE is calculated from the squared difference between the historical and simulated effluent data. Universal parameter distributions (UPD) are taken from Cox (2004). He generated the UPD from published parameter values representing WWTPs from all over the world using the Bayesian method. With calibrated parameters for one specific WWTP, we can also use the Bayesian method to obtain the site-specific parameter distributions (SPD). The difference between UPD and SPD is measured/calibrated data: UPD uses the values all

over the world; while SPD uses the site-specific calibrated values only. Generally, the SPD will have a narrower PDF (probability density function) plot than UPD. As for the plant influent, we have only two options: one is historical influent data, the other is randomly generated influent data, as discussed in Chapter III. The predicted influent data from time series models are not used because they are too close to the historical data.

The secondary clarifier model is simplified to a constant performance with a removal efficiency ratio of 99.75 percent for particulate components, which has proved by the plant data from Oak Ridge and Seneca WWTPs. Due to the limitation of the secondary clarifier model, only effluent ammonia nitrogen is considered as an example in this dissertation. If more reliable and effective clarifier model is available in the future, we can easily add it to our StatASPS program.

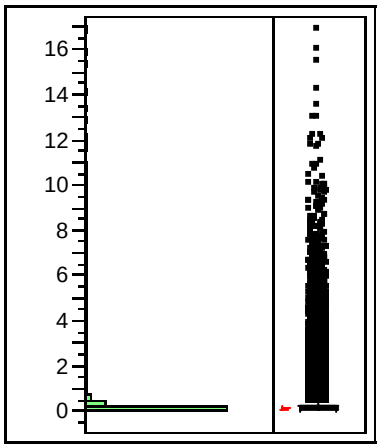
After the tests of the combination of model parameters and plant influent, we can find which combination of parameters and influent best describes the real plant performance with comparison to the historical effluent data. Then we can conduct the Monte Carlo simulations with different operational conditions (for example, different SRT).

A question that must be answered at the outset of the work is: how many Monte Carlo runs are needed to guarantee a valid simulation? According to the references (Meeker and Escober, 1998; Robert and Casella, 2004; Schuhmacher et al., 2001; Huo et al., 2004), the commonly used numbers of Monte Carlo runs are 1000-2000. If more accurate results are needed, we can conduct 5000, 10000 or even more Monte Carlo runs. In this

dissertation, we determined the number of Monte Carlo runs using the Monte Carlo simulations of the Oak Ridge WWTP with one-year influent data and parameter distributions (either universal or site-specific). The results of effluent ammonia nitrogen concentration are shown in Figure 5-1. Obviously, the simulation results of 1000 Monte Carlo runs have more uncertainty than 360 Monte Carlo runs. The maximum value increases from 16.838 to 38.343 mg/L. The standard deviation increases from 0.50 to 0.76. This is what we expected: the more Monte Carlo runs, the more uncertainty, and the more accuracy of the simulation results. Notice that the mean values and several highest percentiles (99.5, 97.5, and 90.0) are very close for both simulation results. These values are also the most important standards to determine whether the Monte Carlo simulation performs well. Thus, from the comparison results in Figure 5-1, we can conclude that 1000 Monte Carlo runs are sufficient to conduct a valid Monte Carlo simulation.

Another concern is reasonable program running times for the Monte Carlo simulations given current computing technologies. In this dissertation, we seek to limit the duration of Monte Carlo simulations to 48 hours (2 days) or better 24 hours (1 day). The Monte Carlo simulations in this dissertation generally follow this rule, although we expect much faster computing technologies in the near future to reduce the magnitude of this concern. Considering both accuracy and present computing speed, we decided to choose 1000 Monte Carlo runs due to the following three reasons. Firstly, it has a good histogram plot and much better performance than the one with 360 Monte Carlo runs. Secondly, it will save a lot of running time compared with the one with 2000 Monte Carlo runs.

360 Monte Carlo runs



NH_4-N (mg/L) vs. Frequency

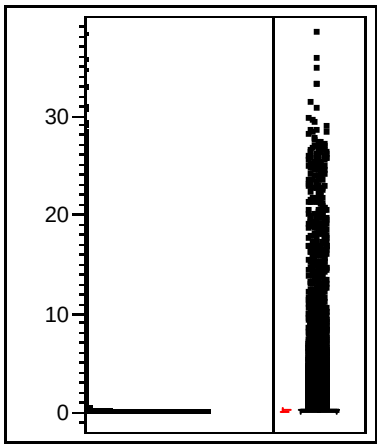
Quantiles

100.0%	maximum	16.838
99.5%		3.422
97.5%		1.242
90.0%		0.436
75.0%	quartile	0.210
50.0%	median	0.114
25.0%	quartile	0.070
10.0%		0.047
2.5%		0.023
0.5%		0.013
0.0%	minimum	0.0041

Moments

Mean	0.235
Std Dev	0.506
N	9360

1000 Monte Carlo runs



NH_4-N (mg/L) vs. Frequency

Quantiles

100.0%	maximum	38.343
99.5%		3.523
97.5%		1.139
90.0%		0.409
75.0%	quartile	0.205
50.0%	median	0.112
25.0%	quartile	0.069
10.0%		0.045
2.5%		0.024
0.5%		0.013
0.0%	minimum	0.0041

Moments

Mean	0.239
Std Dev	0.763
N	26000

Figure 5-1. The simulated effluent NH_4-N (mg/L) with one-year influent data and parameter distributions for the Oak Ridge WWTP.

Finally, our previous Monte Carlo simulation experience also indicates it is a reasonable number to choose.

Then, the question becomes how many days should we simulate for each Monte Carlo run? The dynamic simulation of the ASM No.1 model is chosen in this dissertation because it most accurately describes the real operational conditions in wastewater treatment systems. The main basis for this claim is the varied plant influent and plant operational conditions. It is obvious that a dynamic simulation takes much more simulation time than a steady state simulation. Table 5-1 shows simulation time as a function of the number of CSTR reactor and the length of the simulation. Based on Table 5-1, we selected dynamic simulations with 2 CSTRs and 36 days data. Notice that the simulation data of the first five days will be ignored because there is a time delay for the

Table 5-1. Monte Carlo performance analysis for dynamic simulations of a WWTP.

Case No.	Number of MC runs	Number of CSTR	Number of days	Running time (s)	Running time per MC run (s)
1	2	1	36	165	82.5
2	2	2	36	199	99.5
3	2	3	36	265	132.5
4	1	1	366	618	618
5	1	2	366	918	918

system to reach a realistic dynamic operation point. In other words, the valid data are only 31 days. The selection of a 5-day approach to quasi-steady state has been discussed in Chapter III. Initially, we planned for the duration of our Monte Carlo simulations to be one year. Unfortunately, this proved to be impractical given our desire to complete the simulations with 24 hours. In order to increase our confidence in the validity of simulations one-month in duration, we recommend using at least two months: each in the Winter and Summer. For comparison purpose, we will also run a limited number of Monte Carlo simulations with one-year data. The number of CSTRs actually could be determined using appropriate experiments or previous experiences. According to previous experience (Grady et al., 1999), 3-4 CSTRs are commonly used in practice. However, it is only an empirical number. For a specific plant, the number of CSTRs might change. In this dissertation, we only take two CSTRs due to lack of information. For further research, corresponding experiments might be needed to better describe the WWTPs.

So far, we have discussed the purposes of this chapter and planned steps to achieve the goals. In order to evaluate the plant performance, we also need to know the effluent requirements in environmental regulations. The comparison with the historical effluent data can be used to evaluate how well the Monte Carlo simulations do; while the comparison with required effluent standards will be used to evaluate how well the WWTPs perform. Table 5-2 presents the secondary treatment requirements for BOD_5 and SS. For effluent NH_4-N, effluent requirements are determined on a case-by-case basis depending on water quality requirement of the receiving stream.

Table 5-2. Secondary treatment requirements. (After Metcalf and Eddy, 2003)

Component	**Effluent Limitation***	**Memo**
BOD_5	30 (45) mg/L	Maximum 30-day average
	45 (65) mg/L	Maximum 7-day average
	85% (65) removal	Maximum 30-day average
Suspended Solids (SS)	30 (45) mg/L	Maximum 30-day average
	45 (65) mg/L	Maximum 7-day average
	85% (65) removal	Maximum 30-day average
pH	6.0-9.0	Range

*( ) denotes values applicable to treatment equivalent to secondary treatment. Adjustment available for effluents from trickling filter and waste stabilization pond facilities.

In some plants, for example, Oak Ridge WWTP, different requirements apply for "Summer' (May-October) and 'Winter' (November-April) conditions. The effluent requirements for the Oak Ridge WWTP are shown in Table 5-3. This table clearly indicates the effluent requirements in 'Winter' are significantly less stringent than the requirements in 'Summer'. For example, in 'Winter', the acceptable daily maximum ammonia-N concentration is 6.6 mg/L, compared to the value of 3.6 mg/L in 'Summer'. The values of monthly average and weekly average are 3.3 mg/L and 4.95 mg/L, respectively, in 'Winter'; while the corresponding values in 'Summer' are 1.8 mg/L and 2.7 mg/L, respectively. The required effluent data for BOD_5 and SS are more stringent than those shown in Table 5-2. The effluent ammonia nitrogen (mg/L) in Table 5-3 will be used in this dissertation. The daily maximum standard of Summer would be the key standard used in this dissertation because it is more stringent daily standard same as the simulation data. The analyses of 7- and 30-day average effluent will also be conducted.

Table 5-3. Effluent requirements of BOD_5, NH_4-N, and SS for the Oak Ridge WWTP.

Component	Monthly Average	Weekly Average	Daily Max
BOD_5 (May-October)	8	12	16
BOD_5 (November-April)	15	20	25
NH_4-N (May-October)	1.8	2.7	3.6
NH_4-N (November-April)	3.3	4.95	6.6
SS (suspended solids)	30	40	45

Note: Effluent limits under the NPDES Permit for the Biological Train (concentration reported in mg/L).

5.1 Oak Ridge Wastewater Treatment Plant

5.1.1 Description of the Oak Ridge WWTP

The Oak Ridge Wastewater Treatment Plant (Oak Ridge WWTP) is located in the city of Oak Ridge, Tennessee. Its primary purpose is to remove organics, solids and pathogenic organisms from the water or biodegrade them to stable minerals or organics that are compatible with the environment. The Oak Ridge WWTP operates 24 hours a day, 365 days a year and is designed for an average wastewater treatment flow of 735,000 L/h (17,604 m^3/day) and a peak flow of 1,577,255 L/h (37,854 m^3/day). The layout of the Oak Ridge WWTP is shown in Figure 5-2.

The process used in the Oak Ridge WWTP is a conventional activated sludge (CAS) process. This process generally applies a rectangular bioreactor with influent and return activated sludge (RAS) being added at one end and mixed liquor exiting at the opposite end. The flow pattern is quasi-plug-flow. The residence time distribution depends on the length-to-width ratio of the bioreactor, mixing provided by the oxygen transfer equipment, and the inlet and outlet configuration. This flow pattern can be modeled as series of continuous stirred tank reactors (CSTRs). Due to lack of information, the number of CSTRs is chosen as two. The calibration results in Section 5.1.2 prove that two CSTRs model can provide reliable and valid simulations. The HRT (hydraulic retention time) typically ranges from 4 to 8 hours, while the SRT ranges from 3 to 8 days,

Figure 5-2. The layout of the Oak Ridge WWTP.

but seldom exceeds 15 days. The MLSS concentration and composition vary little through the bioreactor because the SRT (solid retention time) is long relative to the HRT and the mixed liquor is recycled many times before it wasted. The benefits of CAS processes are that 1) its performance is well characterized and predictable, 2) the process and facility design well known, 3) operational parameters are well characterized, and 4) it is useful in a wide range of applications. However, the drawbacks are moderate capital and operating costs and moderate sludge settleability.

The operational conditions of the Oak Ridge WWTP are shown in Table 5-4. Those numbers are obtained from either historical plant data or operational information from

Table 5-4. The operational conditions of the Oak Ridge WWTP.

Items	Values
Q, influent flow rate	5.8 MGD
V, bioreactor volume	1.57MG
X, biomass concentration in bioreactors (average SS)	2,500 mg/L
X_W, biomass concentration in waste flow	26,000 mg/L
F_W, waste flow rate	15,000 gallon/day
HRT =V/Q	6 hours
SRT=$(V\times X)/(F_W\times X_W)$	10 days

practitioner (designer or operator) working at plants. The bioreactor volume is 1.57 MG, and the SRT is 10 days. The operational information will be used in the ASM1 model to conduct Monte Carlo simulations.

5.1.2 Calibration Procedure and Simulation Analysis

In order to better describe the plant performance for a specific WWTP, site-specific calibrated parameters for this WWTP are needed. The calibration procedure selected was elaborated in detail in Chapter III. As discussed, only effluent ammonia nitrogen is considered due to limitations of the chosen secondary clarifier model. According to the sensitivity analysis results, only 3 out of the 19 parameters are considered in the calibration parameters. These three parameters are: K_{NH} (half saturation coefficient for nitrifiers, mg/L), $\mu_{A,max}$ (specific growth rate of autotrophs, hr^{-1}), and Y_A (yield of autotrophs, mg/mg). The enumerative search method is applied to this calibration procedure to capture the group of calibrated parameters with the minimum RMSE (root mean square error) of the simulated and historical effluent data. The enumerative search range for each parameter is [0.1 × Default Value, 1.9 × Default Value] with an interval of 15 percent of the default value, yielding a total of $13^3 = 2197$ runs evaluations in the search space. According to our experience, it takes approximately 12 hours to complete one calibration of one month. Thus, in this dissertation, the calibration for each of 12 months will be conducted to determine a group of calibrated model parameters for one specific plant. Thus, this enumerative search method will not be limited to local optimum

values, which is not uncommon for traditional optimum methods. However, there are also limitations of this enumerative (exhaustive) search method: searching range and searching intervals. There are two ways to solve this problem if needed. Firstly, we might use a larger range and smaller interval to ensure that the optimum values will not be out of range. Secondly, we can also use the regression methods and/or graphic information to find the best group of model parameter using the fitted function. Notice that the main purpose of the calibrated parameter is to generate the site-specific parameter distributions. In that case, we treat parameters as distributions instead of fixed values. That is, we are more interested in the right range of model parameters instead of the exact values.

Figures 5-3 and 5-4 indicate that the calibrated parameters for the Oak Ridge plant data are well fitted. The missing measured data points are not plotted in these two figures. Figure 5-3 indicates the calibration for the 2001 full year data from the Oak Ridge WWTP. It clearly shows the simulated data from the calibrated parameters fit the measured plant effluent relatively well, capturing the overall trend at any particular time, even though some of the most extreme values of effluent ammonia are not well represented. In order to see the details, Figure 5-4 indicates the calibration for the August data in 2001 from the Oak Ridge WWTP. The calibrated data well captures the overall trend of real data. However, the outlier problem is obvious on the 221st day. The measured value is around 0.09 mg/L; while the simulated value is only around 0.06 mg/L. This might be caused by sudden increasing of plant influent loading (for example, the flow rate and/or ammonia-N concentration). That is, some of extremely variability of

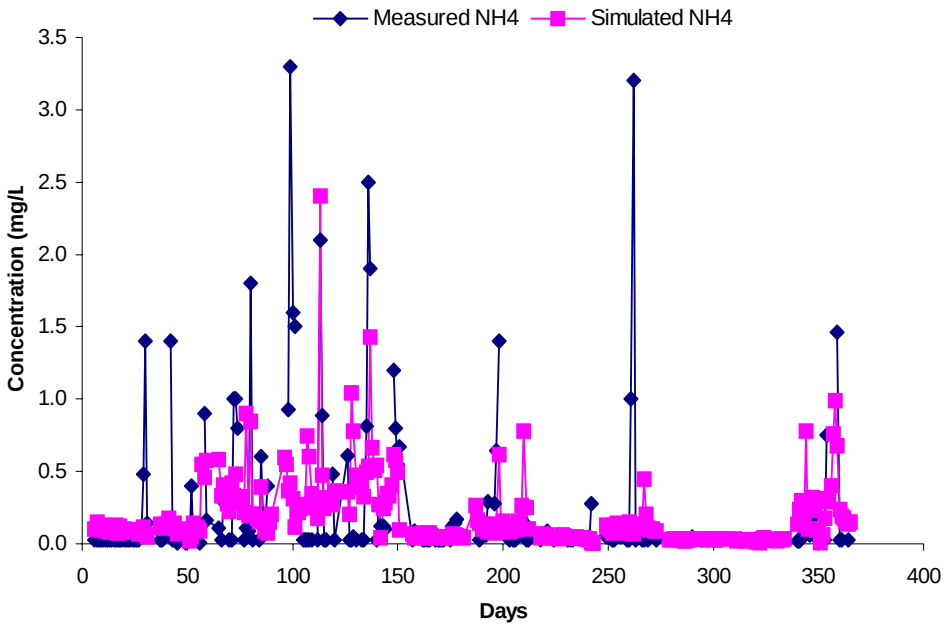


Figure 5-3. The comparison between calibrated and measured data in year 2001 from the Oak Ridge WWTP.

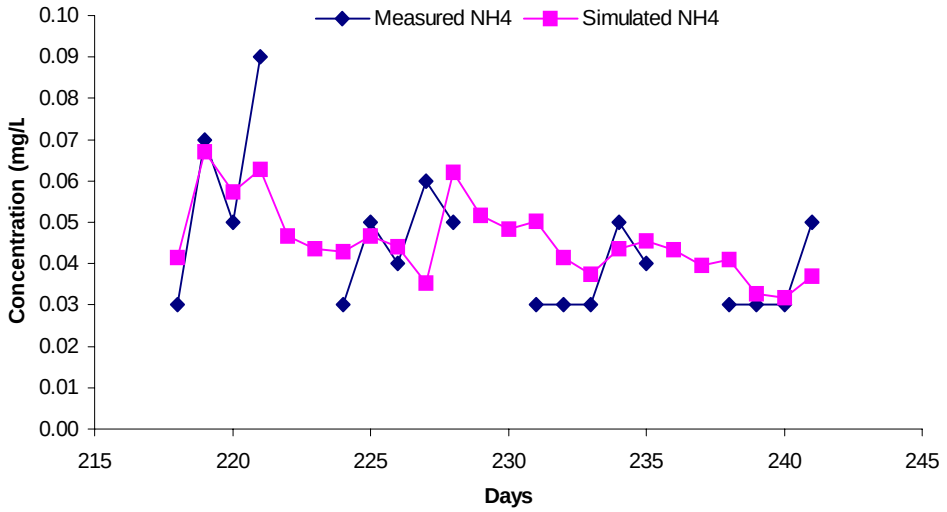


Figure 5-4. The comparison between calibrated and measured data in August from the Oak Ridge WWTP.

influent cannot be captured by the fixed calibrated parameters. In this case, outlier detection and replacement method might be preferred. As for the short-term variations in plant performance that cannot be captured by fixed-parameter models, they are not necessarily caused by easily observable deterministic causes.

Notice that the RMSE criteria might be affected by the detection limit of the ammonia nitrogen methods. As discussed in Chapter III, the ammonia nitrogen concentration is measured by the phenate method. This method is applicable over the range of 0.02 to 2.0 mg/L. Because the historical data have the minimum values of 0.01 mg/L, the calibration results above were obtained with the detection limit of 0.01 mg/L. In order to investigate how the detection limit affects the calibration results, we will consider all simulated effluent ammonia nitrogen concentrations that are lower than 0.01 mg/L to be 0.01 mg/L. That is, the minimum value of the simulated results will be equal to or larger than the detection limit of the selected method. The conducted calibration indicates that the calibrated values of the parameters K_{NH}, $\mu_{A,max}$, and Y_A, are still 0.5500 mg/L, 0.0464 $hour^{-1}$, and 0.4560 mg/mg, respectively, with the minimum RMSE of 0.0646. This result shows that the calibration standard of RMSE is not significantly affected by the low value of detection limit (0.01 mg/L). In fact, the more sensitive values would be the large effluent concentrations, which can also be explained in the equation of RMSE.

The calibrated parameters are shown in Table 5-5. The data shows that the third important parameter Y_A (yield of autotrophs) does not change as much on a month-to month as the other two parameters. Also, the minimum RMSEs for different months are

Table 5-5. The calibrated parameters for 12-month of the Oak Ridge WWTP.

Month/Year	K_{NH} (mg/L)	$\mu_{A,max}$ (hr^{-1})	Y_A (mg/mg)*	RMSE
1/2001	1.3000	0.0608	0.4560	0.2790
2/2001	1.7500	0.0608	0.4560	0.3710
3/2001	1.7500	0.0416	0.4560	0.4293
4/2001	1.9000	0.0416	0.4560	0.9286
5/2001	1.7500	0.0320	0.4560	0.6221
6/2001	0.7000	0.0560	0.0240	0.0446
7/2001	0.7000	0.0320	0.0240	0.2924
8/2001	0.5500	0.0464	0.4560	0.0646
9/2001	1.4500	0.0608	0.4560	0.8013
10/2001	0.5500	0.0560	0.3480	0.0049
11/2001	0.4000	0.0512	0.4560	0.0078
12/2001	0.5500	0.0224	0.0240	0.3108

Note: * indicates the nearly constant parameter, which would prove to be an insignificant parameter through correlation analysis and scatter plots.

significantly different. Large RMSEs are associated with the occurrences of some unusual high or low data points (or potential outliers). The following correlation analysis and scatter plots will explain why this happened.

As discussed above, Figures 5-3 and 5-4 demonstrate that our calibrated parameters are well fitted to the plant effluent data. However, the relationships between the RMSE and other three parameters should be examined to find/confirm the order of the importance of these three parameters. We also want to prove that calibrated parameters are sufficiently accurate to be used in the next Monte Carlo simulations. Before we go further, there is also another thing to consider. As shown in Table 5-5, some values of Y_A (June, July, and December) are 0.0240, which is much smaller than any published values to data (Cox (2004) reports that the minimum published values of Y_A is 0.24 mg/mg). As the results below will demonstrate, RMSE was very insensitive to Y_A. A peculiarity of the calibration program caused it to retain the lowest value of the parameter that resulted in the minimum RMSE. Therefore, the lowest values of Y_A should not be considered as representative of the actual values of this parameter.

Table 5-6 shows correlations between RMSE and the three calibrated parameters. Surprisingly, the most important parameter is $\mu_{A,max}$ instead of K_{NH}. The negative correlation between RMSE and $\mu_{A,max}$ is very strong with a coefficient of -0.7844. Parameter Y_A is the least important parameter. With an insignificant correlation coefficient of -0.0003. This also explains why the parameter Y_A has very small mistaken calibrated values as shown in Table 5-5.

Table 5-6. The correlations between RMSE and the three parameters. (April 2001, Oak Ridge WWTP)

Items	**K_{NH} (mg/L)**	**$\mu_{A,max}$ (hr^{-1})**	**Y_A(mg/mg)**
RMSE	0.0612	-0.7844	-0.0003

The following three plots (Figures 5-5, 5-6, and 5-7) also indicate the relationships between RMSE and these three parameters. Figure 5-5 confirms there are no relationships between RMSE and parameter Y_A at all. Thus, this parameter will be fixed at a default value of 0.24 mg/L (listed in Table 2-4) for all Monte Carlo simulations. The relationship between RMSE and K_{NH} is a little complicated. It has three different patterns: one is flat, another is linearly rapidly increasing, the last one is exponentially (or linearly) but slowly increasing. It indicates the parameter K_{NH} is not the only factor strongly affecting RMSE. In fact, Figure 5-6 clearly shows that K_{NH} is strongly affected by the interaction from $\mu_{A,max}$. For $\mu_{A,max}$ less than 0.01 hr^{-1}, which is extremely small value, there is almost no effect from K_{NH}. In the range 0.01 hr^{-1} < $\mu_{A,max}$ < 0.03 hr^{-1}, the optimum value of K_{NH} decreases with increasing value of $\mu_{A,max}$. For $\mu_{A,max}$ > 0.03 hr^{-1}, the value of K_{NH} has very little effect on RMSE. The whole plot also shows that the RMSE decreases with increase of $\mu_{A,max}$, although the tail part actually increases a little bit. The best calibrated parameters in this case is: $\mu_{A,max}$ = 0.0416 hr^{-1}, and K_{NH} = 1.9 mg/L. Notice that the tail part is almost flat, with little changes when $\mu_{A,max}$ > 0.03 hr^{-1}.

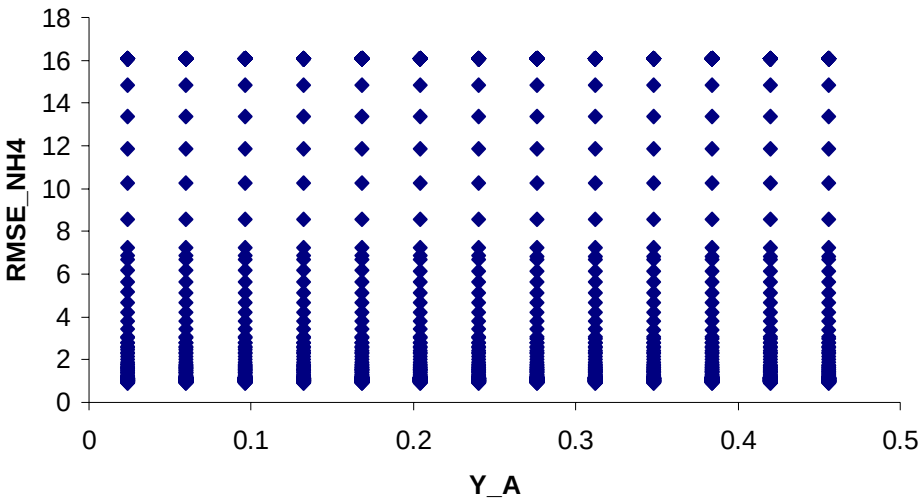


Figure 5-5. The scatter plot of RMSE against Y_A (Oak Ridge WWTP).

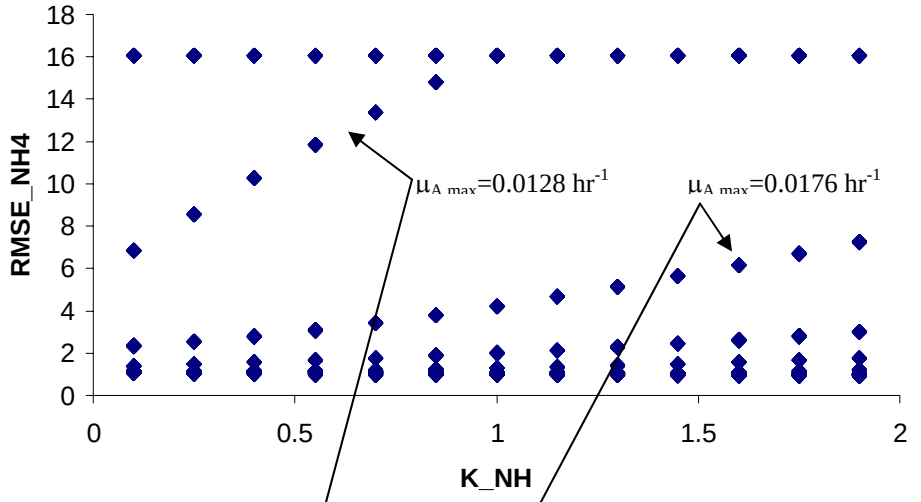


Figure 5-6. The scatter plot of RMSE against K_{NH} (Oak Ridge WWTP).

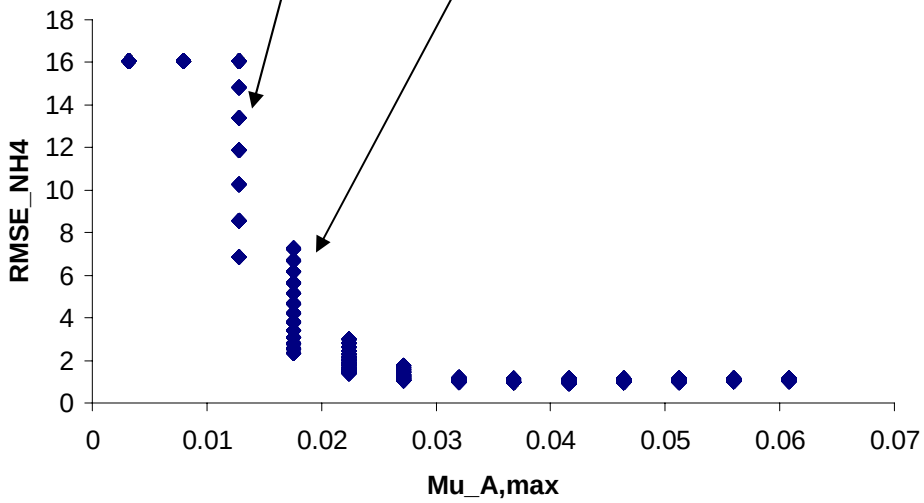


Figure 5-7. The scatter plot of RMSE against $\mu_{A,max}$ (Oak Ridge WWTP).

This demonstrates that the RMSE is relatively insensitive over a wide range of parameter values. This observation supports our approach in which parameters are represented as distributions instead of as fixed well-defined values.

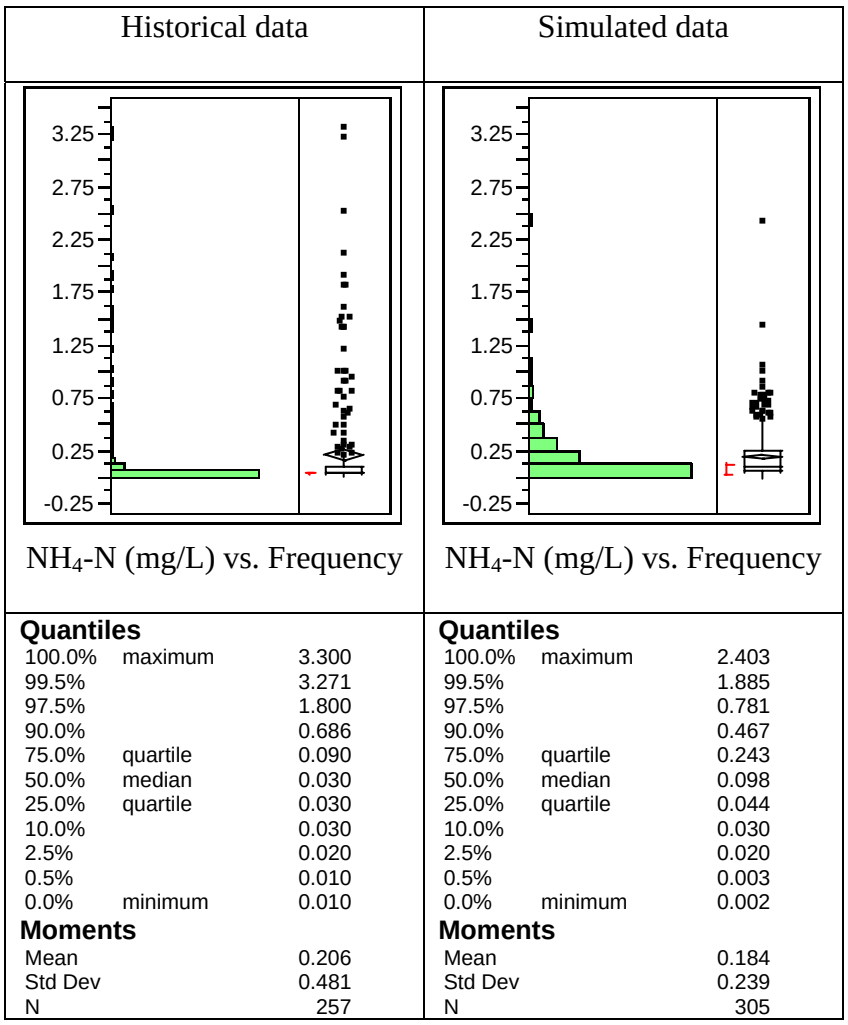
Notice that there is a relationship between $\mu_{A,max}$ and Y_A. When the substrate is being used at its maximum rate, the autotrophic biomass is also growing at their maximum rate. The maximum specific growth rate of the autotrophic biomass is thus related to the maximum specific substrate utilization rate as follows.

$$\mu_{A,max} = k \times Y_A \tag{5-1}$$

where, $\mu_{A,max}$ is the maximum specific growth rate of autotrophic biomass ($hour^{-1}$); k is the maximum specific substrate utilization rate [mg/(mg·hour)]; and Y_A is the yield of autotrophic biomass (mg/mg). Just like $\mu_{A,max}$, Y_A is affected both by the substrate and the microorganism in the bioreactors. However, Y_A represents the energy available in a substrate, while $\mu_{A,max}$ represents how fast a microorganism can process that energy and grow. Because they represent different features, there is no correlation between these two parameters (Grady, et al., 1999). For example, some slowly biodegradable substrates (low $\mu_{A,max}$) can still provide more energy to the degrading culture (i.e., higher Y_A) than rapidly biodegradable substrates do (Grady, et al., 1975) . This suggests that there is variability in parameter k that cannot be determined from data on $\mu_{A,max}$ alone, and vice versa. It is also important to determine the feature of the true growth yield: Y_A. Furthermore, in the ASM1 model, the process parameters do not include the parameter k: the maximum specific substrate utilization rate [mg/(mg·hour)], as shown in equation (5-

1). As shown in Table 2-4, the values of Y_A are different at two temperature: 20 and 10 °C, while the values of $\mu_{A,max}$ are same. This indicates that the wastewater system might have different k at different temperature. However, there is little information of the uncertainty in k available. Thus, even we have the determined equation (5-1), we still cannot conclude that these two parameters are certainly correlated. As a result of previous analyses, we only consider the parameters of K_{NH} and $\mu_{A,max}$.

The statistical characteristics of both historical and simulated results are shown in Figure 5-8. From the plot, the statistical characteristics of the simulated effluent ammonia nitrogen effluent data do not closely match those of the historical data. The mean value and standard deviation of the historical data are 0.206 mg/L and 0.481 mg/L compared to 0.184 mg/L and 0.239 mg/L for the simulated data. Furthermore, the highest percentiles of 100.0, 99.5, 97.5 and 90.0 for the historical data (3.300, 3.270, 1.800, and 0.667 mg/L, respectively) are much larger than those in the simulated data (2.403, 1.885, 0.781, and 0.467 mg/L, respectively). In general, we can conclude that the historical data are characterized by greater variability than the simulated data obtained with month-by-month calibrated parameters. Obviously, if we design a plant using the simulation from calibrated parameters, we fail to capture the true variability in plant performance. In other words, the traditional method using simulation results from calibrated parameter cannot provide a reliable design procedure without an extra safety factor added.



Quantiles

100.0%	maximum	3.300
99.5%		3.271
97.5%		1.800
90.0%		0.686
75.0%	quartile	0.090
50.0%	median	0.030
25.0%	quartile	0.030
10.0%		0.030
2.5%		0.020
0.5%		0.010
0.0%	minimum	0.010

Moments

Mean	0.206
Std Dev	0.481
N	257

Quantiles

100.0%	maximum	2.403
99.5%		1.885
97.5%		0.781
90.0%		0.467
75.0%	quartile	0.243
50.0%	median	0.098
25.0%	quartile	0.044
10.0%		0.030
2.5%		0.020
0.5%		0.003
0.0%	minimum	0.002

Moments

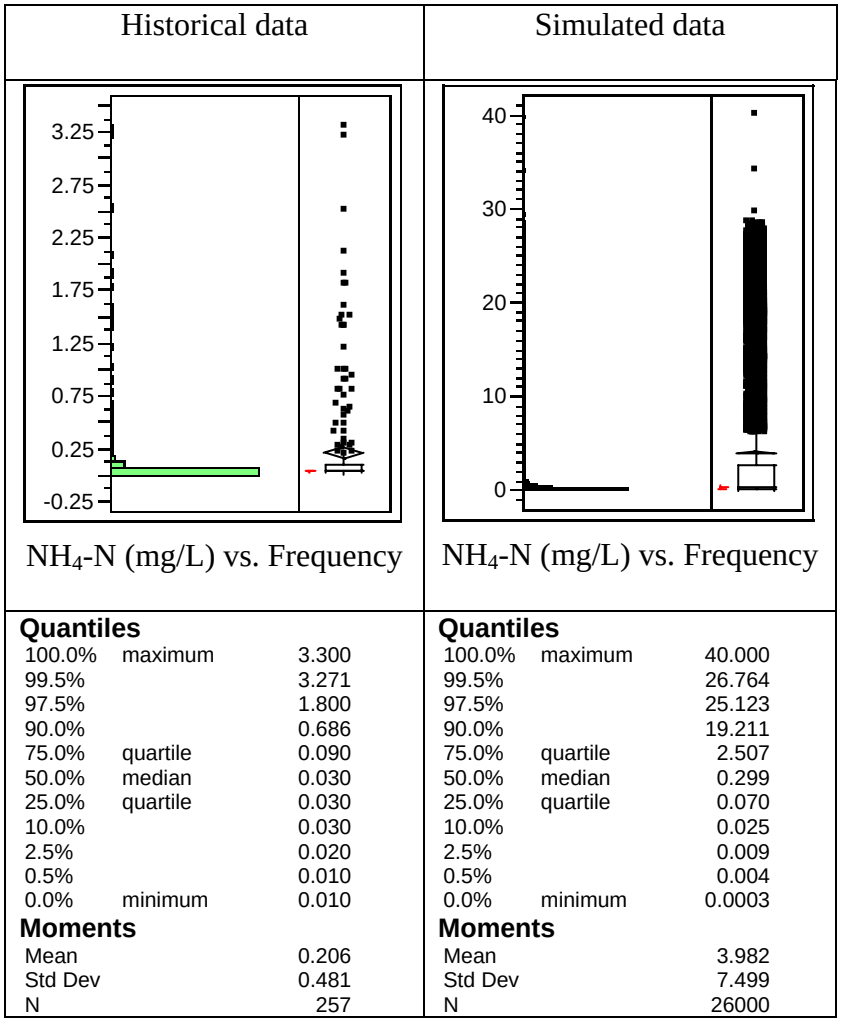
Mean	0.184
Std Dev	0.239
N	305

Figure 5-8. The comparison of one-year historical and simulated effluent NH_4-N (mg/L) of the Oak Ridge WWTP.

5.1.3 Simulations with Parameter Uncertainty

The previous section discussed the calibration procedure and calibrated parameters for the Oak Ridge WWTP. The comparison between historical and simulated effluent data indicates that the design methods that better represent the variability in the plant are needed. Thus, in this section, we will investigate the use of distributed parameters in conjunction with Monte Carlo simulations to assess the variability. As mentioned in Chapter III, there are two options for distributed parameters. One is the universal (or non-site-specific) distributed parameters generated by Cox (2004) based on published parameter values from all over the world; the other is the site-specific distributed parameter that will be generated using the Bayesian method and site-specific calibrated parameter.

Results of the Monte Carlo simulation with the universal parameter distributions is conducted are shown in Figure 5-9. The results are based on August 2001 historical influent data. Obviously, the Monte Carlo simulation results with the universal parameter distributions overestimate both the magnitude and the variability of the historical effluent. The mean value and standard deviation of the simulation results are 3.982 mg/L and 7.499 mg/L, respectively, compared to 0.206 mg/L and 0.481 mg/L for the historical data. As for the highest percentiles, they are unreasonably higher than the historical data with a maximum value of 40 mg/L and 90.0 percentile of 19.211 mg/L. In short, the universal parameter distributions introduce too much unrealistic uncertainty into wastewater treatment systems.



Quantiles		
100.0%	maximum	3.300
99.5%		3.271
97.5%		1.800
90.0%		0.686
75.0%	quartile	0.090
50.0%	median	0.030
25.0%	quartile	0.030
10.0%		0.030
2.5%		0.020
0.5%		0.010
0.0%	minimum	0.010

Moments	
Mean	0.206
Std Dev	0.481
N	257

Quantiles		
100.0%	maximum	40.000
99.5%		26.764
97.5%		25.123
90.0%		19.211
75.0%	quartile	2.507
50.0%	median	0.299
25.0%	quartile	0.070
10.0%		0.025
2.5%		0.009
0.5%		0.004
0.0%	minimum	0.0003

Moments	
Mean	3.982
Std Dev	7.499
N	26000

Figure 5-9. Monte Carlo simulations of effluent NH_4-N (mg/L) with the universal parameter distributions and August historical influent from the Oak Ridge WWTP.

The two previous simulations indicate that neither the fixed calibrated parameters nor universal parameter distributions accurately capture the variability observed in the actual plant. The fixed calibrated parameters can predict the average effluent well. However, the simulated results underestimate the variability observed in the historical effluent. The universal parameter distributions simply introduce too much unrealistic uncertainty, thus leading to inaccurate simulation results. Two key reasons are 1) too much uncertainty in the model parameters is considered; 2) and the likelihood of sampling extreme parameters from multiple distributions simultaneously. To reduce the uncertainty in the parameters, we will consider site-specific parameter distributions. These site-specific parameters can better describe the plant performance because they are generated from calibrated parameters while preserving the consideration of uncertainty.

The site-specific parameter distributions for K_{NH} and $\mu_{A,max}$ are generated from the monthly calibrated parameters using the Bayesian method detailed in Chapter III. The results are summarized in Table 5-7.

The detailed Bayesian procedure was discussed in Chapter III. The standard deviation of the prior distribution is calculated using the equation $\sigma = 0.25 \times \mu$ instead of $\sigma = 2 \times 0.25 \times \mu$ in order to narrow down the posterior distribution, especially for the site-specific parameter distributions. For details, please refer to Cox (2004).

Table 5-7. The universal and site-specific parameter distributions for the Oak Ridge WWTP.

Items	**K_{NH} (mg/L)**	**$\mu_{A,max}$ (hr^{-1})**
UPD: μ	-0.675	-3.688
UPD: σ	1	0.44
SPD: μ	-0.0844	-3.1791
SPD: σ	0.6531	0.3774
Percentiles (%)	**K_{NH} (mg/L)**	**$\mu_{A,max}$ (hr^{-1})**
1	0.2012	0.0173
5	0.3139	0.0224
10	0.3980	0.0257
25	0.5916	0.0323
50	0.9190	0.0416
75	1.4277	0.0537
90	2.1223	0.0675
95	2.6906	0.0774
99	4.1989	0.1001

Figure 5-10 shows the universal (UPD) and site-specific (SPD) parameter distributions of the parameter K_{NH}. It seems like the UPD is "narrower" than the SPD, however, it is actually not. Because all values of a lognormal distribution are larger than zero, the PDF curve of a lognormal distribution is not symmetric. We will discuss this together with the percentile and probability tables later. Figure 5-11 shows the comparison of the UPD and SPD for the parameter $\mu_{A,max}$. Both figures indicate that the curves of two SPD both move to right a little bit. In other words, the mean and median values of two SPD are both larger than those of the UPD.

Table 5-8 shows the percentile comparison between two kinds of SPD. Distributions labeled SPD calculate the prior standard deviation using formula: $\sigma = 0.25 \times \mu$, while parameters SPD' calculate the prior standard deviation using formula: $\sigma = 2 \times 0.25 \times \mu$ as originally recommended by Cox (2004). Obviously, SPD' will have a wider prior distribution and generate a wider posterior distribution with more uncertainty.

Until now, we obtained the two kinds of SPDs (SPD and SPD') for K_{NH} and $\mu_{A,max}$. However, we need to determine which one is better to conduct following Monte Carlo simulations. Table 5-9 compares the range of the monthly calibrated values to percentiles of SPD and SPD'. For example, for the parameter $\mu_{A,max}$, the minimum and maximum calibrated values are 0.0224 and 0.0608 day^{-1}, respectively. The minimum value falls at the lowest 5 percent of SPD' while the maximum value falls at the lowest 79 percent. Therefore, the probability of a random sample from SPD' to fall in the range of the

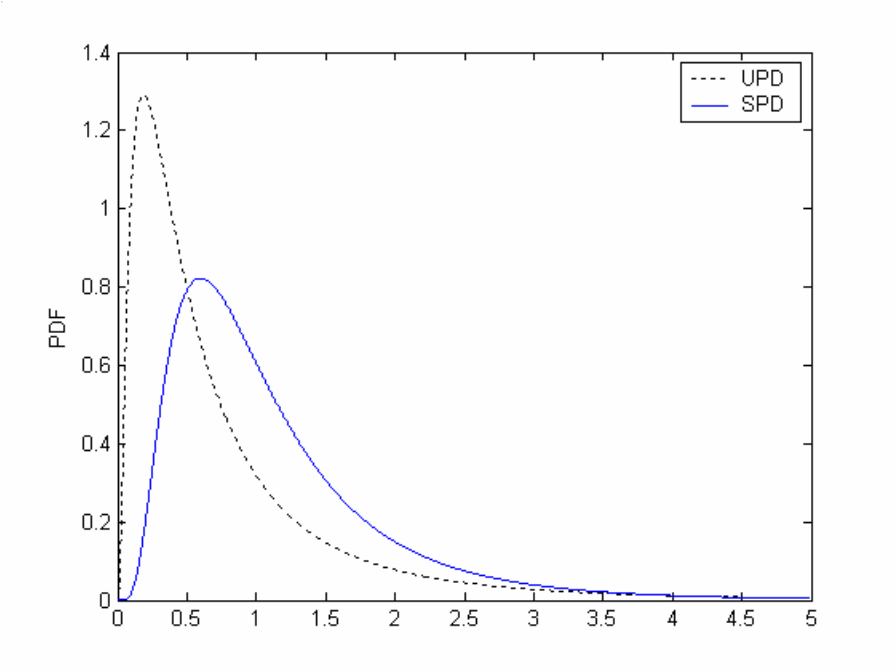


Figure 5-10. PDF plots of K_{NH} (mg/L) for the UPD and SPD for the Oak Ridge WWTP.

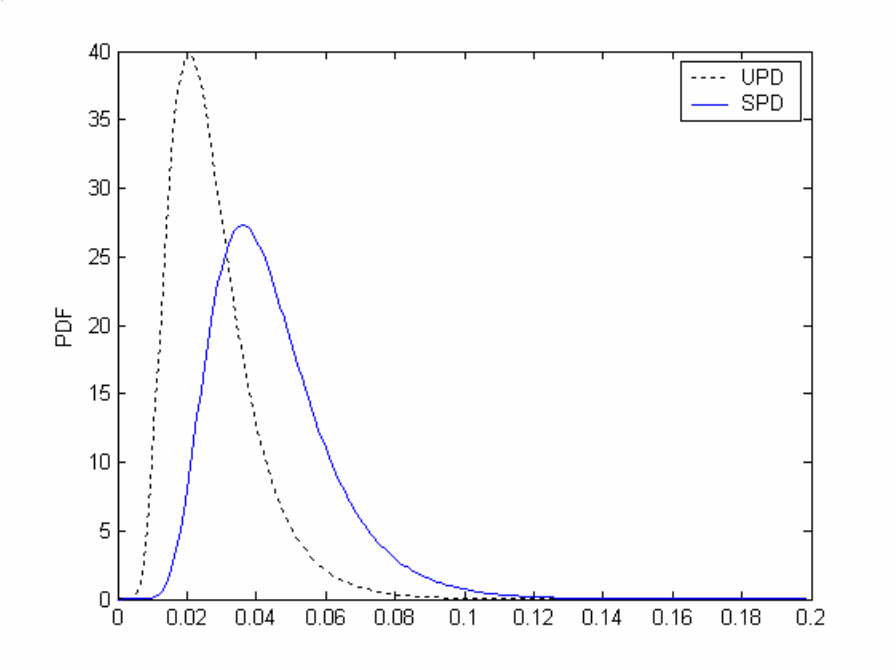


Figure 5-11. PDF plots of μ_{Amax} ($hour^{-1}$) for the UPD and SPD for the Oak Ridge WWTP.

Table 5-8. The percentile comparison between UPD and two SPDs.

Items	**K_{NH} (mg/L)**		**$\mu_{A,max}$ (hr^{-1})**	
	0.05	**0.95**	**0.05**	**0.95**
UPD	0.0983	2.6376	0.0121	0.0516
SPD’	0.3299	2.8541	0.0235	0.0841
SPD	0.3139	2.6908	0.0224	0.0774

Table 5-9. The probabilities of the data range between minimum and maximum calibrated parameters.

Items	**K_{NH} (mg/L)**		**$\mu_{A,max}$ (hr^{-1})**	
Range of calibrated values	**Minimum** **0.4**	**Maximum** **1.9**	**Minimum** **0.0224**	**Maximum** **0.0608**
PDF (SPD’)	0.0883	0.8472	0.0387	0.7906
Probability (SPD’)		**0.7589**		**0.7519**
PDF (SPD)	0.1014	0.8669	0.0503	0.8423
Probability (SPD)		**0.7656**		**0.7920**

calibrated value is approximately 75 percent. We believe that this probability is a little too low and adds too much uncertainty to the distributions; therefore, we used the SPD estimates to obtain slightly more narrow site-specific distributions.

As listed in Table 5-6, the correlation coefficient of $\mu_{A,max}$ and K_{NH} is 0.087847. That is, there are no strong correlations between those two parameters. Thus, in the following simulations, those two parameters are randomly and independently sampled. Notice that three kinds of parameters are considered: SPD#1-AP: all parameters are sampled from the UPD expect for $\mu_{A,max}$ and K_{NH}, which are sampled from their SPD; SPD#2-2P: all parameters are fixed except for $\mu_{A,max}$ and K_{NH}, which are sampled from their SPD; and SPD#3-1P: all parameters are fixed except for $\mu_{A,max}$, which is sampled from its SPD.

Figure 5-12 shows the comparison of Monte Carlo simulations to the historical data using the three different SPDs. Clearly, the simulation results with SPD#2-2P and SPD#3-1P are much better than the results with SPD#1-AP. Variability in fewer parameters yielded better simulation results. In fact, the SPD#3-1P best described the effluent ammonia-N when considering the mean, standard deviation and 99.5 percentile values. However, the simulation results with SPD#2-2P better represent the 97.5 and 90.0 percentiles. Nevertheless, we recommend SPD#3-1P, after considering the overall distribution of effluent of ammonia-N concentration. As for the maximum values, all three distributions predict much greater values than observed in the historical record; however, undue weight should not be given to these extreme points.

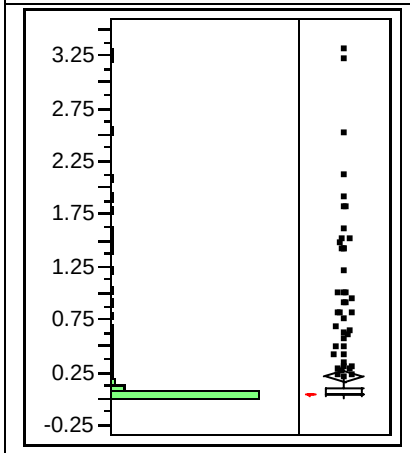
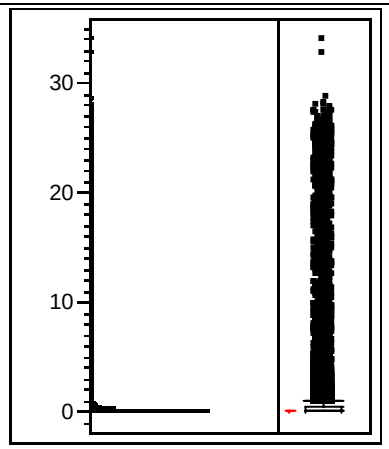
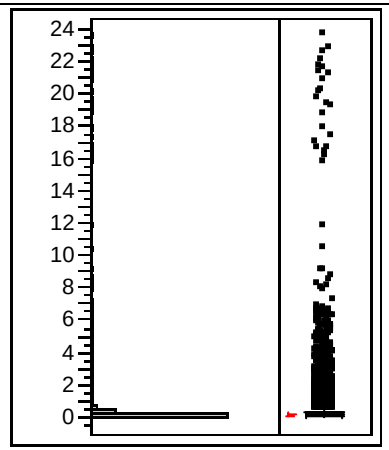
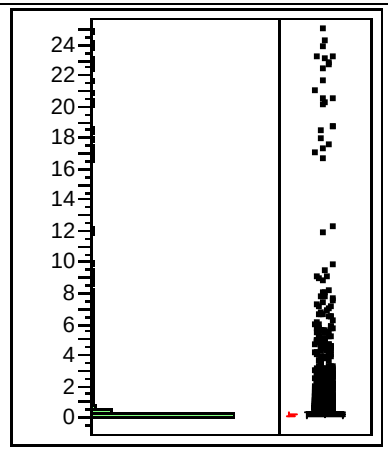
Historical data	SPD#1-AP	SPD#2-2P	SPD#3-1P
			
NH_4-N (mg/L) vs. Frequency	NH_4-N (mg/L) vs. Frequency	NH_4-N (mg/L) vs. Frequency	NH_4-N (mg/L) vs. Frequency
Quantiles	**Quantiles**	**Quantiles**	**Quantiles**
100.0% maximum 3.300	100.0% maximum 34.064	100.0% maximum 23.727	100.0% maximum 24.938
99.5% 3.271	99.5% 24.880	99.5% 4.055	99.5% 3.206
97.5% 1.800	97.5% 12.456	97.5% 1.438	97.5% 1.054
90.0% 0.686	90.0% 1.420	90.0% 0.519	90.0% 0.385
75.0% quartile 0.090	75.0% quartile 0.377	75.0% quartile 0.246	75.0% quartile 0.198
50.0% median 0.030	50.0% median 0.125	50.0% median 0.100	50.0% median 0.109
25.0% quartile 0.030	25.0% quartile 0.052	25.0% quartile 0.044	25.0% quartile 0.068
10.0% 0.030	10.0% 0.024	10.0% 0.022	10.0% 0.043
2.5% 0.020	2.5% 0.011	2.5% 0.010	2.5% 0.019
0.5% 0.010	0.5% 0.0053	0.5% 0.0051	0.5% 0.0089
0.0% minimum 0.010	0.0% minimum 0.0003	0.0% minimum 0.0019	0.0% minimum 0.0055
Moments	**Moments**	**Moments**	**Moments**
Mean 0.206	Mean 0.977	Mean 0.263	Mean 0.229
Std Dev 0.481	Std Dev 3.377	Std Dev 0.777	Std Dev 0.761
N 257	N 26000	N 26000	N 26000

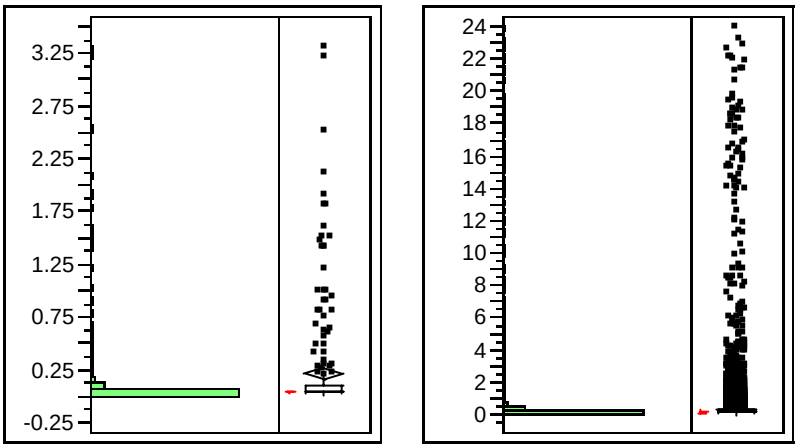
Figure 5-12. Monte Carlo simulations of effluent NH_4-N (mg/L) with SPDs and August historical influent for the Oak Ridge WWTP.

5.1.4 Simulations with Influent Variability

In last section, we discussed the effect of the parameter uncertainty on the Monte Carlo simulations. In this section, we will elaborate the effect of variability in the plant influent. Two options are available: (1) historical influent data, or (2) randomly generated influent data.

The previous simulation results in last section are the simulations from the August historical influent data with different parameter distributions. Since Winter and Summer may have different effects on plant performance, we would like to consider both in this dissertation to ensure a reasonable and reliable simulation. We take August and January typical months of Summer and Winter, respectively. Since August was already considered in the previous section, we will now conduct a similar simulation based on January historical influent data. The simulated results are shown in Figure 5-13. The mean value and standard deviation of the SPD#3-1P simulated data are smaller than the results from the August influent data. This leads to a slightly overestimation of the effluent ammonia nitrogen as compared to the historical data. However, the highest percentiles of 100, 99.5 and 97.5 are very similar. Thus, for the purpose of Monte Carlo simulations, we conclude that the SPD is adequate from both Summer and Winter conditions. Figure 5-13 also indicates that the performance in Winter is worse than the performance in Summer, which also explains why effluent requirement in Winter is less stringent than in Summer. We strongly recommend conducting simulations over at least

Historical data **SPD#3-1P**



NH_4-N (mg/L) vs. Frequency NH_4-N (mg/L) vs. Frequency

Quantiles

100.0%	maximum	3.300
99.5%		3.271
97.5%		1.800
90.0%		0.686
75.0%	quartile	0.090
50.0%	median	0.030
25.0%	quartile	0.030
10.0%		0.030
2.5%		0.020
0.5%		0.010
0.0%	minimum	0.010

Moments

Mean	0.206
Std Dev	0.481
N	257

Quantiles

100.0%	maximum	23.943
99.5%		3.122
97.5%		1.032
90.0%		0.403
75.0%	quartile	0.208
50.0%	median	0.116
25.0%	quartile	0.072
10.0%		0.049
2.5%		0.023
0.5%		0.013
0.0%	minimum	0.006

Moments

Mean	0.245
Std Dev	0.872
N	26000

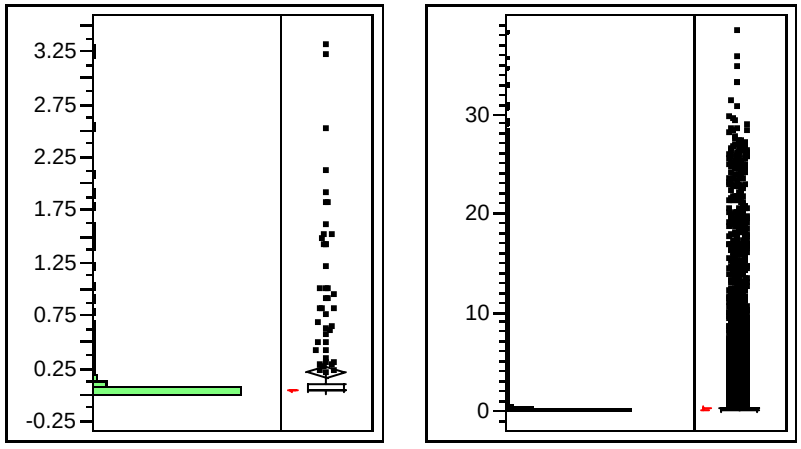
Figure 5-13. Monte Carlo simulations of effluent NH_4-N (mg/L) with the SPD#3-1P parameters, and January historical influent for the Oak Ridge WWTP.

Two months (for example, January and August) to indicate two extreme cases in the plant performance. However, in this dissertation, we would like to take August influent data as an example to study the further uncertainty in the plant performance in the following simulations.

In order to confirm whether the one months' influent data (either January or August) can provide reliable and reasonable simulations, we conduct 1000 Monte Carlo runs with one-year influent data. Figure 5-14 shows the Monte Carlo simulation with the SPD#3-1P parameters and one-year historical influent data from the Oak Ridge WWTP. The mean and standard deviation of 1000 Monte Carlo runs are 0.239 mg/L and 0.763. The mean and standard deviation of the historical plant effluent data are 0.206 mg/L and 0.481. More important, the percentiles of 99.50 and 97.50 of Monte Carlo simulation results are both very close to real plant effluent data. Overall Monte Carlo simulations using August and full-year data yield very similar results, thus justifying the use of the one month influent history.

Till now, Monte Carlo simulations have been conducted with distributed parameters and historical influent data. Next, we will discuss the simulations with randomly generated influent data. In Chapter III, we created a time series model of the Oak Ridge influent capable of generating random influent time series. We will now use this model to test the effects of influent variability on simulation performance. We consider three cases. First, we use randomly generated influent data for August with fixed calibrated parameters. Next, we use the randomly generated influent data for August with the SPD#3-1P

Historical data **SPD#3-1P**



NH_4-N (mg/L) vs. Frequency NH_4-N (mg/L) vs. Frequency

Quantiles		
100.0%	maximum	3.300
99.5%		3.271
97.5%		1.800
90.0%		0.686
75.0%	quartile	0.090
50.0%	median	0.030
25.0%	quartile	0.030
10.0%		0.030
2.5%		0.020
0.5%		0.010
0.0%	minimum	0.010
Moments		
Mean		0.206
Std Dev		0.481
N		257

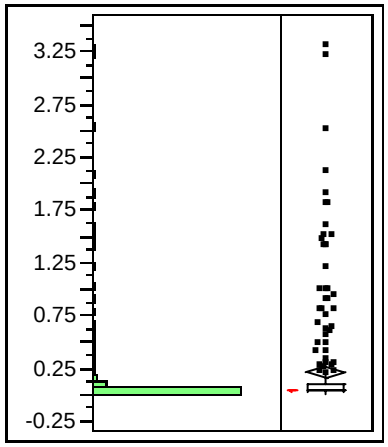
Quantiles		
100.0%	maximum	38.343
99.5%		3.523
97.5%		1.139
90.0%		0.409
75.0%	quartile	0.205
50.0%	median	0.112
25.0%	quartile	0.069
10.0%		0.045
2.5%		0.024
0.5%		0.013
0.0%	minimum	0.004
Moments		
Mean		0.239
Std Dev		0.763
N		360000

Figure 5-14. Monte Carlo simulation of effluent NH_4-N (mg/L) with the SPD#3-1P parameters and one-year historical influent for the Oak Ridge WWTP.

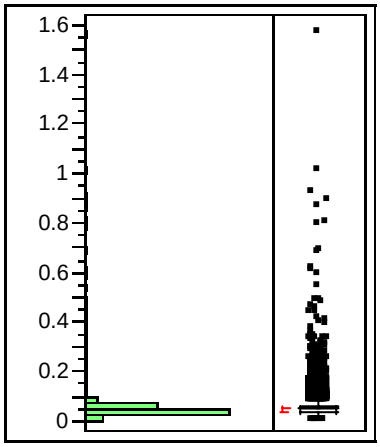
Parameters. Finally, we use randomly generated influent data for one-year with SPD#3-1P. Notice that for randomly generated influent, we will have different influent data for each Monte Carlo run. These random influent data are different from the historical data, while they have the same/similar statistical characteristics as the historical data. Figure 5-15 shows the statistical characteristics of simulated effluent Ammonia-N with random influent data for August and the fixed parameters. It clearly indicates that the simulated effluent data notably underestimate the variability in the real plant full year data. Thus, it is not acceptable to be applied in the future Monte Carlo runs. The simulation results with one-year random influent data and fixed calibrated parameters are shown in Figure 5-16. Obviously, this simulation is also bad because it underestimates the uncertainty of plant effluent, specifically in mean, standard deviation, and highest percentiles (99.5, 97.5, etc.). However, the simulation results are much better than the results with only one month influent, indicating that the one-year random influent data did introduce more variability into WWTPs.

Figure 5-17 shows the statistical characteristics of simulated effluent Ammonia-N with the SPD#3-1P parameters and random August influent data. The simulated results are not too bad. However, it predicts much higher mean value, standard deviation, and the highest percentiles (99.5, 97.5, etc.). That is, the simulation results overestimate the variability this time. Furthermore, compared with Figures 5-15 and 5-16, it suggests that the uncertainty in plant performance is better represented with uncertainty in model parameters rather than variability of the plant influent. Thus, it makes sense to use historical plant data instead of time-consuming randomly generated influent data.

Historical data **Simulated data**



NH_4-N (mg/L) vs. Frequency



NH_4-N (mg/L) vs. Frequency

Quantiles

100.0%	maximum	3.300
99.5%		3.271
97.5%		1.800
90.0%		0.686
75.0%	quartile	0.090
50.0%	median	0.030
25.0%	quartile	0.030
10.0%		0.030
2.5%		0.020
0.5%		0.010
0.0%	minimum	0.010

Moments

Mean	0.206
Std Dev	0.481
N	257

Quantiles

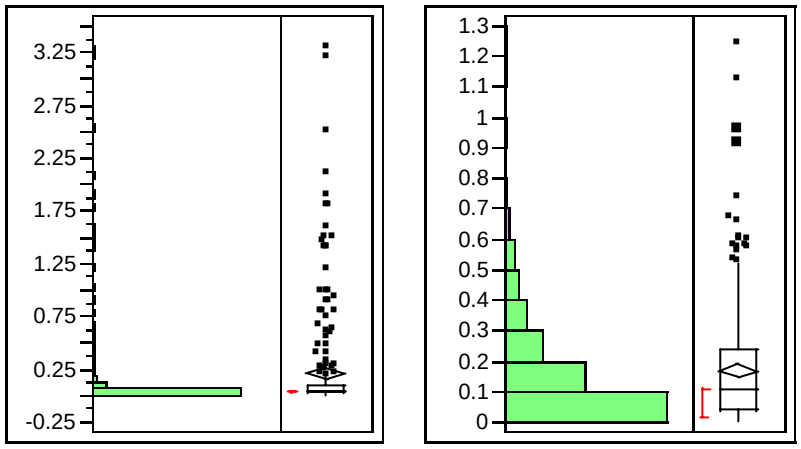
100.0%	maximum	1.5719
99.5%		0.1744
97.5%		0.1034
90.0%		0.0713
75.0%	quartile	0.0559
50.0%	median	0.0445
25.0%	quartile	0.0356
10.0%		0.0275
2.5%		0.0151
0.5%		0.0088
0.0%	minimum	0.0036

Moments

Mean	0.0488
Std Dev	0.02993
N	26000

Figure 5-15. Monte Carlo simulation of effluent NH_4-N (mg/L) with all fixed calibrated parameters and randomly generated influent for August for the Oak Ridge WWTP.

Historical data **Simulated data**



NH_4-N (mg/L) vs. Frequency NH_4-N (mg/L) vs. Frequency

Quantiles

100.0%	maximum	3.300
99.5%		3.271
97.5%		1.800
90.0%		0.686
75.0%	quartile	0.090
50.0%	median	0.030
25.0%	quartile	0.030
10.0%		0.030
2.5%		0.020
0.5%		0.010
0.0%	minimum	0.010

Moments

Mean	0.206
Std Dev	0.481
N	257

Quantiles

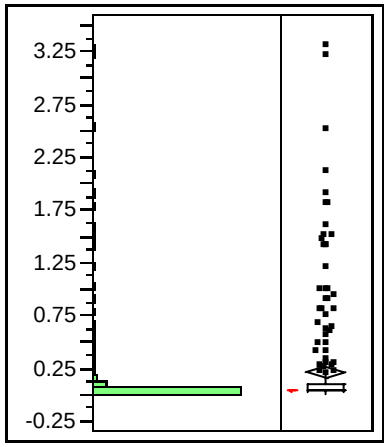
100.0%	maximum	1.2459
99.5%		1.1812
97.5%		0.6268
90.0%		0.4150
75.0%	quartile	0.2372
50.0%	median	0.1022
25.0%	quartile	0.0427
10.0%		0.0250
2.5%		0.0157
0.5%		0.0091
0.0%	minimum	0.0074

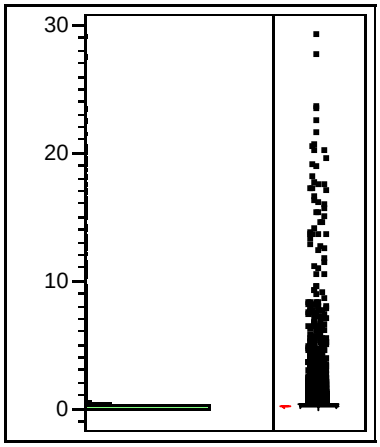
Moments

Mean	0.1669
Std Dev	0.1824
N	360000

Figure 5-16. Monte Carlo simulations of effluent NH_4-N (mg/L) with the fixed calibrated parameters and one-year randomly generated influent for the Oak Ridge WWTP.

Historical data **Simulated data**





NH_4-N (mg/L) vs. Frequency NH_4-N (mg/L) vs. Frequency

Quantiles

100.0%	maximum	3.300
99.5%		3.271
97.5%		1.800
90.0%		0.686
75.0%	quartile	0.090
50.0%	median	0.030
25.0%	quartile	0.030
10.0%		0.030
2.5%		0.020
0.5%		0.010
0.0%	minimum	0.010

Moments

Mean	0.206
Std Dev	0.481
N	257

Quantiles

100.0%	maximum	29.099
99.5%		4.446
97.5%		1.124
90.0%		0.409
75.0%	quartile	0.205
50.0%	median	0.112
25.0%	quartile	0.068
10.0%		0.045
2.5%		0.027
0.5%		0.015
0.0%	minimum	0.0055

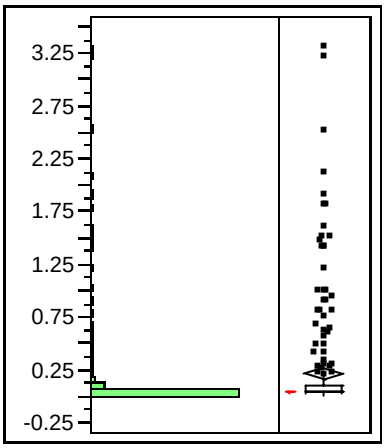
Moments

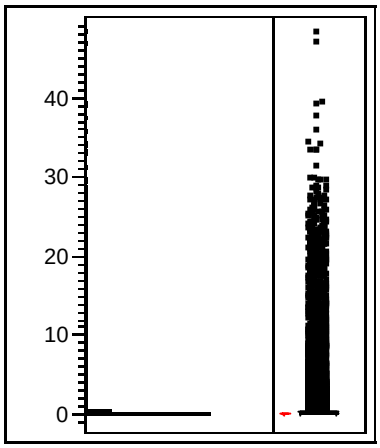
Mean	0.247
Std Dev	0.821
N	26000

Figure 5-17. Monte Carlo simulations of effluent NH_4-N (mg/L) with the SPD#3-1P parameters and randomly generated influent for August for the Oak Ridge WWTP.

In order to confirm whether the one months' random influent data (either January or August) can provide reliable and reasonable simulations, we also conduct 1000 Monte Carlo runs with the SPD#3-1P parameters and one year randomly generated influent data, as shown in Figure 5-18. As expected, the longer simulation introduces more uncertainty into the plant effluent data. However, for 99.5, 97.5, and 90.0 percentiles, the new simulation results (1000 Monte Carlo runs) have the values of 4.097, 1.173, and 0.421 mg/L, respectively. The corresponding values of the results with real plant influent data are 3.523, 1.139, and 0.409 mg/L, respectively. In fact, this might introduce too much uncertainty into our simulated results. Compared the results from the real plant effluent data, we can conclude that the random influent with the SPD#3-1P parameters introduce too much uncertainty that might leads to overly stringent design standards. This can also cause the overdesign problem of WWTPs. It is still difficult to make decision which simulation provides the best results. As we know, we expect the simulated results as close as possible to real plant effluent. However, on the safe side, we also want to make sure that our designed plant based on the simulated results can perform well with a certain probability (for example, 99.5 or 97.5 percent). Thus, it is normally acceptable if our simulated results are a little bit higher than the real plant effluent data. Considering these standards, we can conclude that the Monte Carlo simulation with randomly generated influent offers the safer plant effluent estimates and will be the one of the best choices to describe the uncertainty in the wastewater treatment system. However, it is clearly more time-consuming and expensive to conduct Monte Carlo simulations using the randomly generated influent data.

Historical data **Simulated data**





NH_4-N (mg/L) vs. Frequency NH_4-N (mg/L) vs. Frequency

Quantiles

100.0%	maximum	3.300
99.5%		3.271
97.5%		1.800
90.0%		0.686
75.0%	quartile	0.090
50.0%	median	0.030
25.0%	quartile	0.030
10.0%		0.030
2.5%		0.020
0.5%		0.010
0.0%	minimum	0.010

Moments

Mean	0.206
Std Dev	0.481
N	257

Quantiles

100.0%	maximum	48.202
99.5%		4.097
97.5%		1.173
90.0%		0.421
75.0%	quartile	0.208
50.0%	median	0.112
25.0%	quartile	0.067
10.0%		0.044
2.5%		0.025
0.5%		0.014
0.0%	minimum	0.0033

Moments

Mean	0.252
Std Dev	0.880
N	360000

Figure 5-18. Monte Carlo simulations of effluent NH_4-N (mg/L) with the SPD#3-1P parameters and the one-year randomly generated influent for the Oak Ridge WWTP.

5.1.5 Discussions

Table 5-10 summarizes the simulation completed in the previous sections. Obviously, the best simulation is conducted with one full year historical influent and the SPD#3-1P parameters. This best simulation has the best grade score: A+, considering the mean, standard deviation, the highest percentiles, and histogram plot. There are two simulations with the second best grade: A. One is the simulation with one month historical influent data and the SPD#3-1P parameters. The other is the simulation with one year randomly generated influent and the SPD#3-1P parameters.

Table 5-10. Summary table of Monte Carlo simulations of the Oak Ridge WWTP.

Influent	Month/year	Fixed	UPD	SPD#1-AP	SPD#2-2P	SPD#3-1P
Historical	8/2001		C	C+	A-	A (No.2)
	1/2001					A-
	2001	B+(*)				A+ (No.1)
Random	8/2001	C (*)				A-
	2001					A (No.3)

Note: * = Underestimate.

As expected, this simulation with the SPD#3-1P parameters and randomly generated influent would introduce more uncertainty into the Monte Carlo simulations of wastewater systems. Considering the simplicity, we recommend the simulation with the SPD#3-1P parameters and one-month historical data, which is the second best choice.

The best parameter distributions are the SPD#3-1P parameters, that is, all fixed calibrated parameters except the SPD $\mu_{A,max}$. It is an important finding that reliable Monte Carlo simulation can be achieved by considering only the most important parameter: $\mu_{A,max.}$ Future investigations may seek to determine if alternative methods of determining the distribution of the key parameter might result in better representations of plant variability.

Furthermore, Table 5-10 also indicates that the randomly generated influent data do not have a major impact on the uncertainty of the plant effluent. The random influent data combined with the fixed parameters usually remarkably underestimated the variability. The random influent data combined with the parameter distributions, however, introduce too much uncertainty, that is, at extreme conditions the effluent concentration was overestimated. This case is not favored either, because it might leads to the over-design of wastewater treatment plants. In short, from the Monte Carlo simulations of the Oak Ridge WWTP, we conclude that the best choice is the simulation with one year historical influent data and the SPD#3-1P; and the second best choice is the simulation with one month historical data and the SPD#3-1P.

In order to quantify the standard of evaluating simulated results, goodness-of-fit (GoF) tests are conducted. A goodness-of-fit test is a statistical test in which the validity of one hypothesis is tested without specification of an alternative hypothesis. The most common tests for goodness-of-fit tests are the chi-square test, Kolmogorov-Smirnov test, and Cramer-von Mises test. According to Conover (1999), the Cramer-von Mises test is one of the most powerful methods. This test can be applied in any case (either binned or unbinned data). The goal of the Cramer-von Mises test is to investigate the significance of the difference between two population distributions, based on two sample distributions.

The test statistic T_2 is calculated as follows:

$$T_2 = \frac{mn}{m+n}\left\{\sum_{i=1}^{n}[S_1(X_i) - S_2(X_i)]^2 + \sum_{j=1}^{m}[S_1(Y_i) - S_2(Y_i)]^2\right\} \quad (5\text{-}2)$$

where $S_1(x)$ and $S_2(x)$ are the empirical distribution functions of two samples; m and n are the sample sizes.

The hypotheses are:

H_0 : $F(x) = G(x)$, for all x;

H_1 : $F(x) \neq G(x)$, for at least one value of x. (5-3)

Reject H_0 at the approximate level α if T_2 exceeds the 1-α quantile $\omega_{1-\alpha}$ (for example, $\omega_{0.99}$=0.743 at α=0.05). These quantiles are obtained based on the asymptotic

distribution, valid for large m and n (Anderson and Darling, 1952). However, they are considered fairly accurate even if the sample sizes are small (Burr, 1964).

The quantified results of Table 5-10 are shown in Table 5-11, using the Cramer-von Mises goodness of fit test. Table 5-11 indicates that the best two options are still from the simulations with the SPD#3-1P parameters and one-year historical/random influent, with T_2 values of 0.6194 and 0.6035, respectively. Because both T_2 values are less than $\omega_{0.99}$=0.743, we conclude the null hypothesis of identical distribution functions is accepted at α=0.01. Surprisingly, the simulation with the SPD#3-1P parameters and August historical influent are not well presented as a result of the T_2 statistics. This method is ranked as the No.3 choice in Table 5-10. However, in Table 5-11, the

Table 5-11. GoF tests of Monte Carlo simulations of the Oak Ridge WWTP.

Influent	Month/year	Fixed	UPD	SPD#1-AP	SPD#2-2P	SPD#3-1P
Historical	8/2001		3.4185	1.3686	1.0620	1.7392
	1/2001					1.8216
	2001	2.2957				0.6194
Random	8/2001	1.5148				1.6770
	2001					0.6035

simulation with the SPD#2-2P parameters and August historical influent is much better and well deserves the No.3 choice, instead. Notice that the goodness-of-fit test is only based on a statistical standard. This standard might be inappropriate for the applications in practice. For example, the simulation results in Figure 5-12 indicate that the simulation with the SPD#2-2P parameters is only slightly different from the simulation with the SPD#3-1P parameters. However, the goodness-of-fit tests in Table 5-11 show that the simulation with the SPD#2-2P parameters is obviously much better than the simulation with the SPD#3-1P parameters. Thus, we believe that this Cramer-von Mises goodness-of-fit test cannot work for all the cases. In this dissertation, the evaluation of the simulation results mostly focuses on the highest percentiles (99.50 and 97.50 percent), mean and standard deviation, which are considered the most important factors affecting the plant performance. Further research is needed to find a better goodness-of-fit test emphasizing more on the highest percentiles, mean, and standard deviation.

Table 5-12 shows the approximate required time to conduct a simulation with 1000 Monte Carlo runs. The required time includes data preparation, program running, and

Table 5-12. The required time for 1000 Monte Carlo runs.

Influent	Time with SPD#3-1P
One-month data	1-3 days
One-year data	5-7 days

data analysis. As discussed, the second best choice with the one-month influent data can also provide a reasonable and reliable simulated results, compared with the real plant effluent and simulated results from one year influent data. From Table 5-12, one-month simulation only needs 1-3 days, which is much less than the simulation with one-year simulation (3-7 days). Thus, unless the best accuracy is needed, we can use the Monte Carlo simulation with one-month influent data (either Winter January or Summer August data) and the SPD#3-1P parameters. The following simulations will consider how changing SRT affects the plant effluent NH_4-N concentration. These simulations of effluent NH_4-N for different SRT = 6, 8, 10 and 12 days are conducted with the SPD#3-1P parameters and the August historical influent data, as shown in Figure 5-19. It obviously indicates that the mean values and the highest percentiles (99.50, 97.50, and 90.00) decrease with increasing SRT.

Figures 5-20 and 5-21 show the relationship between the percentiles of the effluent NH_4-N and SRT. The maximum effluent ammonia nitrogen is observed to increase with SRT over the range of 6 to 10 days; however, too much emphasis should not be placed on the most extreme values. The 99.50, 97.50 and 90.00 percentile ammonia nitrogen concentrations all decrease with increasing SRT. For example, for the case of SRT = 10 days, the 97.5 percent of data are under 1.054 mg/L, which is much less than the requirement (daily maximum concentration: 3.6 mg/L). Normally, regulatory limitations on plant effluent employ moving average methods to allow occasional short-term palnt upsets. Some extremely high effluent values will be smoothed out by 7-day or 30-day averages.

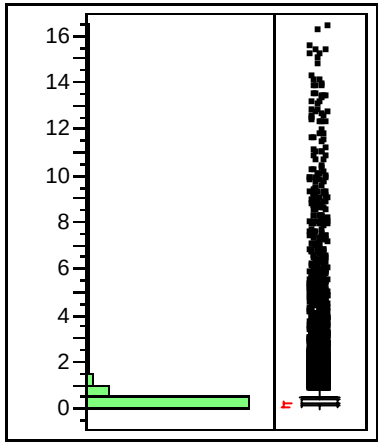
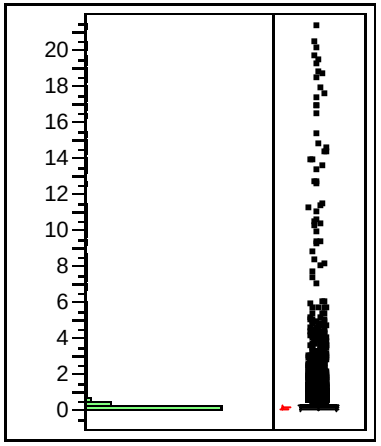
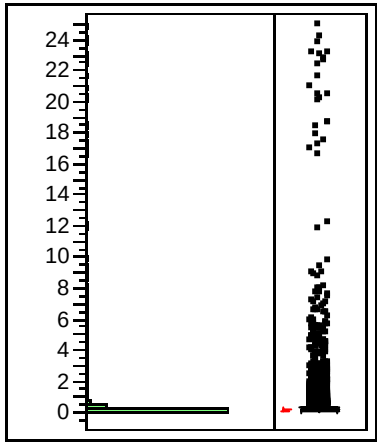
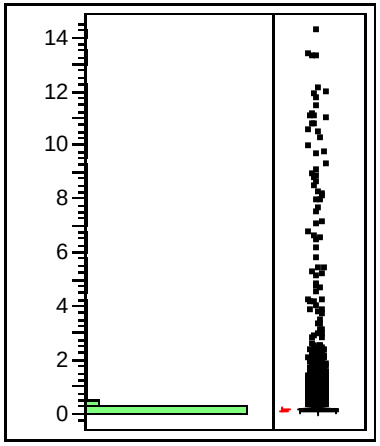
SRT = 6 days	SRT = 8 days	SRT = 10 days	SRT = 12 days
			
NH_4-N (mg/L) vs. Frequency	NH_4-N (mg/L) vs. Frequency	NH_4-N (mg/L) vs. Frequency	NH_4-N (mg/L) vs. Frequency
Quantiles	**Quantiles**	**Quantiles**	**Quantiles**
100.0% maximum 16.361	100.0% maximum 21.295	100.0% maximum 24.938	100.0% maximum 14.236
99.5% 7.900	99.5% 3.216	99.5% 3.206	99.5% 1.598
97.5% 2.741	97.5% 1.266	97.5% 1.054	97.5% 0.591
90.0% 0.908	90.0% 0.492	90.0% 0.385	90.0% 0.271
75.0% quartile 0.396	75.0% quartile 0.245	75.0% quartile 0.198	75.0% quartile 0.153
50.0% median 0.188	50.0% median 0.130	50.0% median 0.109	50.0% median 0.091
25.0% quartile 0.103	25.0% quartile 0.078	25.0% quartile 0.068	25.0% quartile 0.058
10.0% 0.064	10.0% 0.049	10.0% 0.043	10.0% 0.037
2.5% 0.026	2.5% 0.022	2.5% 0.019	2.5% 0.017
0.5% 0.019	0.5% 0.012	0.5% 0.0089	0.5% 0.0068
0.0% minimum 0.012	0.0% minimum 0.0084	0.0% minimum 0.0055	0.0% minimum 0.0052
Moments	**Moments**	**Moments**	**Moments**
Mean 0.460	Mean 0.265	Mean 0.229	Mean 0.157
Std Dev 1.027	Std Dev 0.705	Std Dev 0.761	Std Dev 0.445
N 26000	N 26000	N 26000	N 26000

Figure 5-19. Monte Carlo simulations of effluent NH_4-N (mg/L) with the SPD#3-1P parameters and the August historical influent for the Oak Ridge WWTP.

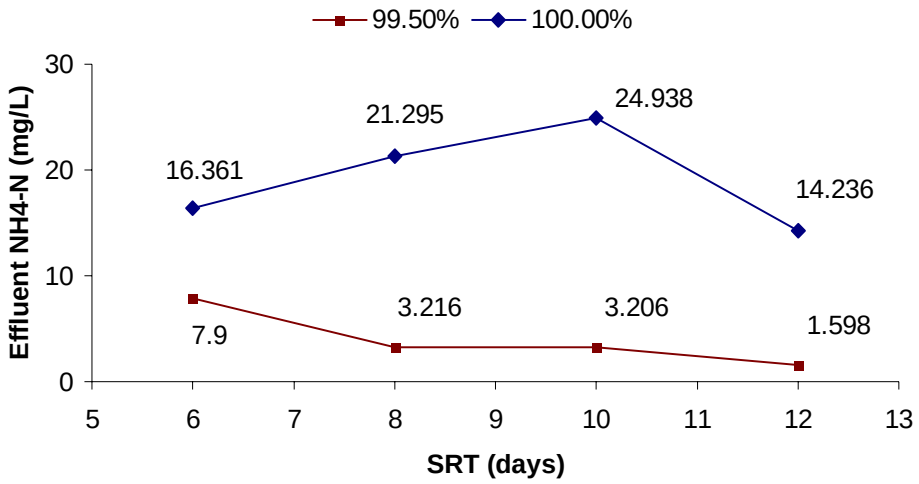


Figure 5-20. The percentile plot (100.00 and 99.50) against SRTs (Oak Ridge WWTP).

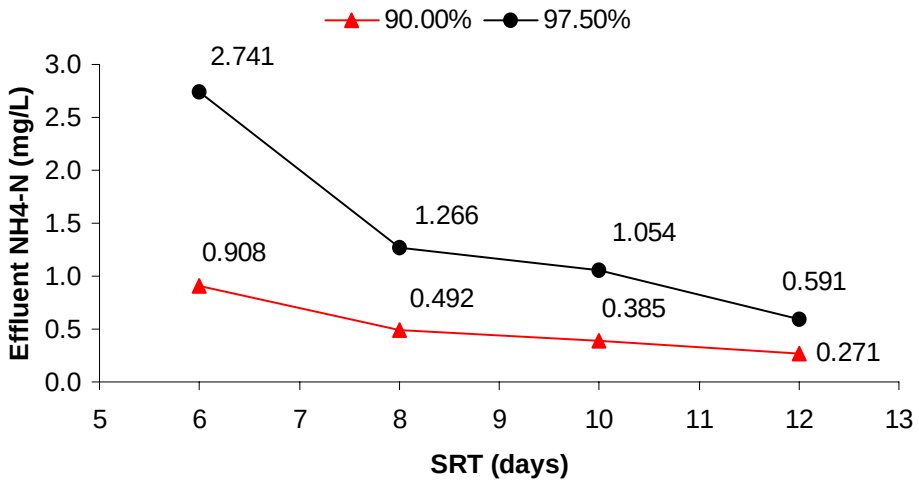
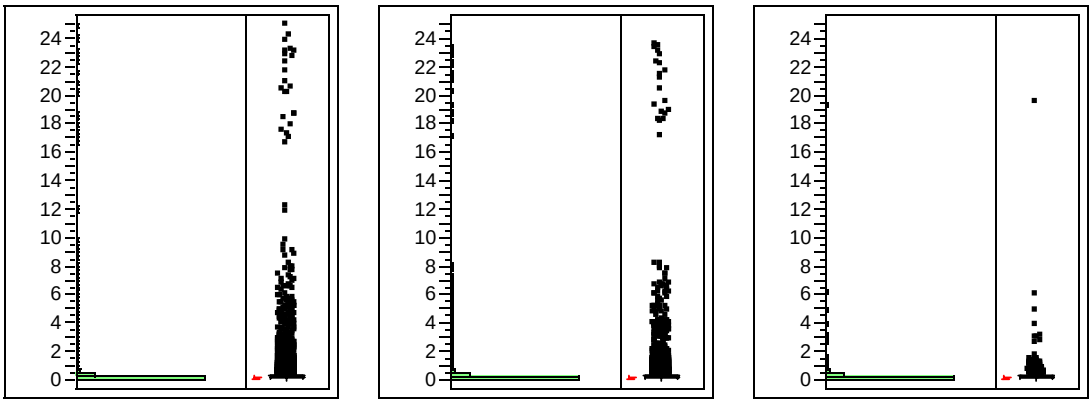


Figure 5-21. The percentile plot (97.50 and 90.00) against SRTs (Oak Ridge WWTP).

Effluent standards are based on daily max, 7-day average, and 30-day average are commonly employed. For example, these effluent ammonia-N standards of the Oak Ridge WWTP are 3.6, 2.7, and 1.8 mg/L, respectively. Figure 5-22 shows the comparison between the daily data, 7 and 30-day average data. As expected, the mean and standard deviation have no significant difference. Even for the percentiles, there is no significant difference either. However, there is something different on the histogram plots. It clearly indicates the average method smoothes out some largest values. This is also the reason why 30-day average has the most stringent standard, followed by the 7-day average. Because only one month data are considered, the 30-day average data for 1000 Monte Carlo runs have only 1000 data, which is too few to make reliable decision. Thus, we decide to take 7-day average data to make the decision chart. For comparison purposes, Figure 5-23 shows the similar plots of real plant effluent. It clearly indicates that the average method successfully smoothed out those large values, thus making the three highest percentiles decrease significantly from daily, to 7-day, to 30-day average. Compared with Figure 5-23, the big difference is that Figure 5-22 still has a few effluent data with extremely high values even for those 7-day and 30-day averages. Some of those values are likely caused by extremely low values of the parameter $\mu_{A,max}$. In this case, there would be continuous 30 days large values, which are impossible to be smoothed out by 7-day and 30-day average methods.

Figure 5-24 shows statistical characteristics of the same simulated results as obtained in Figure 5-19. The only difference is that Figure 5-19 uses daily effluent ammonia-N; while Figure 5-24 uses the 7-day average of effluent ammonia-N. Compared with Figure

Daily	7-day average	30-day average



NH_4-N (mg/L) vs. Frequency	NH_4-N (mg/L) vs. Frequency	NH_4-N (mg/L) vs. Frequency

Quantiles

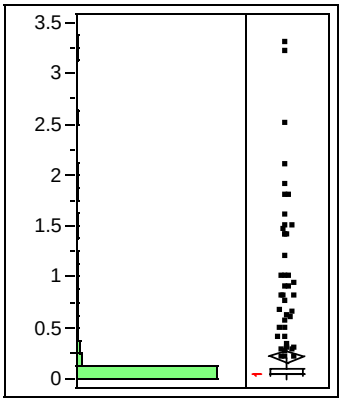
100.0%	maximum	24.938	100.0%	maximum	23.493	100.0%	maximum	19.464
99.5%		3.206	99.5%		3.598	99.5%		3.040
97.5%		1.054	97.5%		1.102	97.5%		1.072
90.0%		0.385	90.0%		0.385	90.0%		0.370
75.0%	quartile	0.198	75.0%	quartile	0.206	75.0%	quartile	0.204
50.0%	median	0.109	50.0%	median	0.114	50.0%	median	0.115
25.0%	quartile	0.068	25.0%	quartile	0.074	25.0%	quartile	0.074
10.0%		0.043	10.0%		0.053	10.0%		0.053
2.5%		0.019	2.5%		0.037	2.5%		0.037
0.5%		0.0089	0.5%		0.026	0.5%		0.026
0.0%	minimum	0.0055	0.0%	minimum	0.011	0.0%	minimum	0.014

Moments

Mean		0.229	Mean		0.235	Mean		0.229
Std Dev		0.761	Std Dev		0.772	Std Dev		0.721
N		26000	N		19000	N		1000

Figure 5-22. The comparison of daily, 7 and 30-day averages of simulated effluent NH_4-N of the Oak Ridge WWTP. (August, 2001)

Daily	7-day average	30-day average



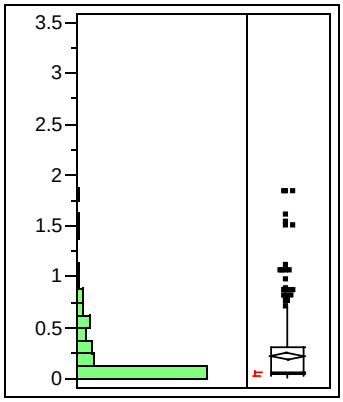
NH_4-N (mg/L) vs. Frequency

Quantiles

100.0%	maximum	3.300
99.5%		3.271
97.5%		1.800
90.0%		0.678
75.0%	quartile	0.090
50.0%	median	0.030
25.0%	quartile	0.030
10.0%		0.030
2.5%		0.020
0.5%		0.0029
0.0%	minimum	0.0000

Moments

Mean	0.206
Std Dev	0.480
N	258



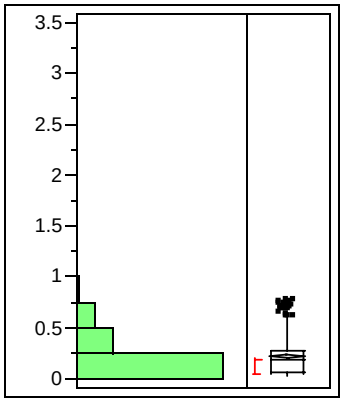
NH_4-N (mg/L) vs. Frequency

Quantiles

100.0%	maximum	1.833
99.5%		1.833
97.5%		1.055
90.0%		0.650
75.0%	quartile	0.298
50.0%	median	0.056
25.0%	quartile	0.030
10.0%		0.030
2.5%		0.026
0.5%		0.018
0.0%	minimum	0.018

Moments

Mean	0.219
Std Dev	0.320
N	358



NH_4-N (mg/L) vs. Frequency

Quantiles

100.0%	maximum	0.772
99.5%		0.765
97.5%		0.717
90.0%		0.499
75.0%	quartile	0.272
50.0%	median	0.178
25.0%	quartile	0.051
10.0%		0.031
2.5%		0.030
0.5%		0.029
0.0%	minimum	0.029

Moments

Mean	0.217
Std Dev	0.194
N	335

Figure 5-23. The comparison of daily, 7 and 30-day averages of historical effluent NH_4-N (mg/L) of the Oak Ridge WWTP.

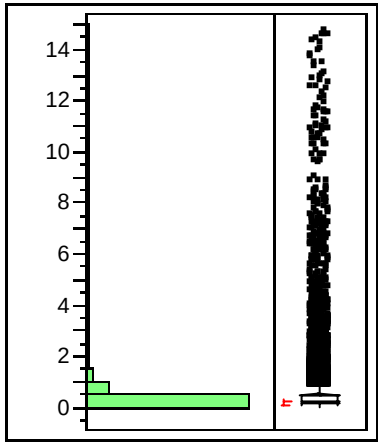
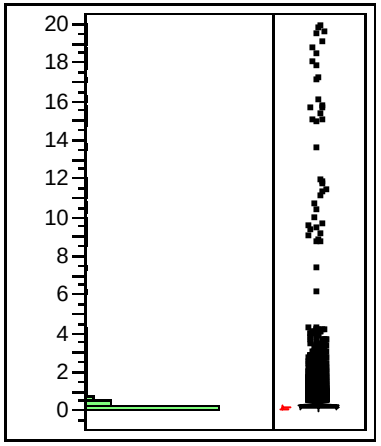
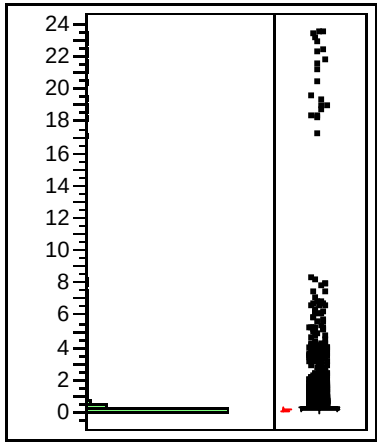
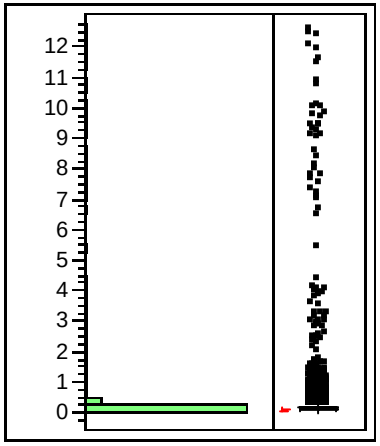
SRT = 6 days	SRT = 8 days	SRT = 10 days	SRT = 12 days
			
NH_4-N (mg/L) vs. Frequency	NH_4-N (mg/L) vs. Frequency	NH_4-N (mg/L) vs. Frequency	NH_4-N (mg/L) vs. Frequency
Quantiles	**Quantiles**	**Quantiles**	**Quantiles**
100.0% maximum 14.763	100.0% maximum 19.798	100.0% maximum 23.493	100.0% maximum 12.606
99.5% 7.456	99.5% 2.831	99.5% 3.598	99.5% 1.334
97.5% 2.766	97.5% 1.254	97.5% 1.102	97.5% 0.590
90.0% 0.935	90.0% 0.507	90.0% 0.385	90.0% 0.276
75.0% quartile 0.411	75.0% quartile 0.253	75.0% quartile 0.206	75.0% quartile 0.157
50.0% median 0.200	50.0% median 0.138	50.0% median 0.114	50.0% median 0.095
25.0% quartile 0.113	25.0% quartile 0.085	25.0% quartile 0.074	25.0% quartile 0.063
10.0% 0.077	10.0% 0.060	10.0% 0.053	10.0% 0.047
2.5% 0.051	2.5% 0.042	2.5% 0.037	2.5% 0.033
0.5% 0.037	0.5% 0.031	0.5% 0.026	0.5% 0.025
0.0% minimum 0.020	0.0% minimum 0.017	0.0% minimum 0.011	0.0% minimum 0.013
Moments	**Moments**	**Moments**	**Moments**
Mean 0.469	Mean 0.270	Mean 0.235	Mean 0.162
Std Dev 1.016	Std Dev 0.707	Std Dev 0.772	Std Dev 0.451
N 26000	N 26000	N 26000	N 26000

Figure 5-24. Monte Carlo simulations of 7-day average NH_4-N (mg/L) with the SPD#3-1P parameters and August historical plant influent from the Oak Ridge WWTP.

5-19, the mean and standard deviation are very similar, although the 7-day average data provide slightly smaller values for all four SRTs. The maximum values for all four SRTs significantly decrease thanks to the 7-day moving average. However, for 97.5 percentiles, there are no significant differences. The values of SRTs = 6, 8, 10, and 12 days are 2.766, 1.254, 1.102, and 0.590 mg/L, compared with values of 2.741, 1.266, 1.054, and 0.591 mg/L in Figure 5-19.

The similar plots of percentiles against SRTs are shown in Figures 5-25 and 5-26. Compared with Figures 5-20 and 5-21, there are no significant differences in the values and patterns of percentiles between the daily data and 7-day average data. The reason is that extreme model parameters result in a full month's worth of extreme values, as discussed above. In order to deal with this unrealistic uncertainty, more research on the distributional parameter is needed in the future. Due to equipment and time limitation, we will continue to use the daily effluent data with either 99.50 or 97.50 percentiles, which has proved to be a trustable and reliable method to study the uncertainty of WWTPs.

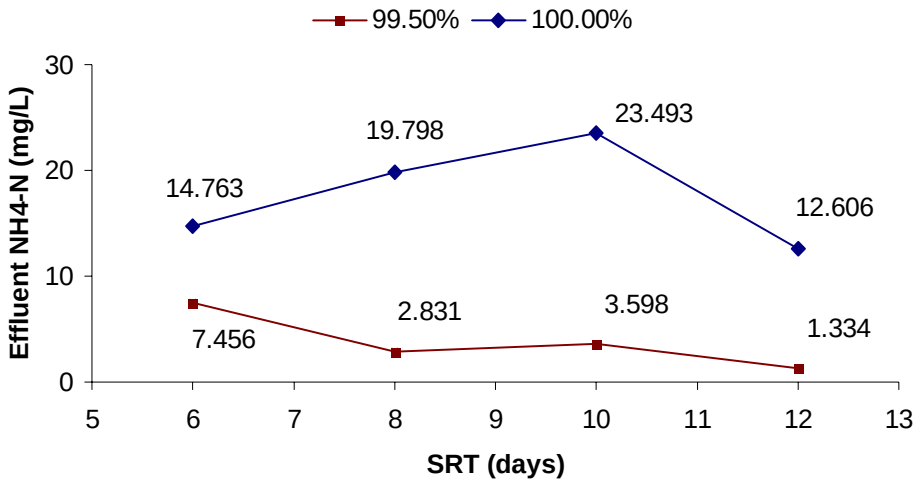


Figure 5-25. The percentile plots (100.00 and 99.50) against SRTs (Oak Ridge WWTP, 7-day average).

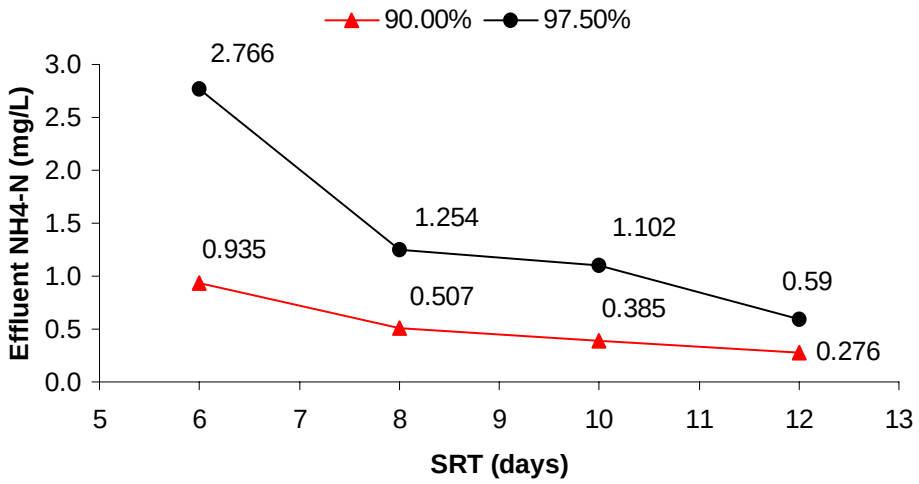


Figure 5-26. The percentile plots (97.50 and 90.00) against SRTs (Oak Ridge WWTP, 7-day average).

5.2 Seneca Wastewater Treatment Plant

5.2.1 Description of the Seneca WWTP

The Seneca Wastewater Treatment Plant is located in Germantown, Maryland. The design capacity is 5 MGD (1 million gallon per day = 3785.4 m^3/day). The influent flows directly into five parallel extended-aeration activated sludge tanks, each designed for 1 MGD. The plant layout is shown in Figure 5-27. There are also other processes, for example, sand filtration, following the activated sludge treatment process. However, because the key topic in this dissertation is the activated sludge system, we only consider the plant influent, the bioreactors, and the effluent from the secondary clarifiers. Furthermore, since we have five bioreactors, which are operated independently, we will take the average values of those five effluent components. It is reasonable because the effluent flow rates for those five bioreactors are all close to 1 MGD; therefore, no weighted average values are needed.

The Seneca WWTP applies an Extended aeration activated sludge (EAAS) process. EAAS processes utilize long SRTs to stabilize the biosolids resulting from the removal of biodegradable organic matter. SRTs of 20 to 30 days are typical, which means that HRTs around 24 hours are required to maintain reasonable MLSS concentrations. Long SRTs offer two benefits: reduced quantities of solids to be disposed of, and greater process stability. Furthermore, high quality, well-nitrified effluent can also be achieved. These benefits are obtained at the expense of the large bioreactors required to achieve the long

Figure 5-27. The layout of the Seneca WWTP.

SRTs, but for many small installations the benefits out-weight the drawbacks.

Table 5-13 shows the operational conditions of the Seneca WWTP. Those numbers are obtained from either historical plant data or operational information from practitioner (designer or operator) working at plants. From the table below, the SRT is chosen as 20 days, and total volume of the bioreactors is 5 MG (1.8927×10^7 Liters) for the Seneca WWTP. In order to consider the effect of SRT on the plant effluent, we also conducted Monte Carlo simulations for other three different SRTs: 5, 10, and 30 days, respectively.

Table 5-13. The operational conditions of the Seneca WWTP.

Items	**Values**
Q, influent flow rate	5×1MGD=5MGD
V, bioreactor volume	5×1MG=5MG
X, biomass concentration in bioreactors (average SS)	3,526 mg/L
X_W, biomass concentration in waste flow	2,7000 mg/L
F_W, waste flow rate	32,250 gallon/day
HRT =V/Q	24 hours
SRT=$(V\times X)/(F_W\times X_W)$	20 days

5.2.2 Calibration Procedure and Simulation Analysis

In order to better describe the plant performance for the Seneca WWTP, the site-specific calibrated model parameters are needed. The calibration procedure is the same as used for the Oak Ridge WWTP. For the Seneca WWTP, we also only consider the effluent NH_4-N concentration due to the limitation of the secondary clarifier model. The one-year plant historical data are used and the corresponding calibrated parameters for each of 12 months are determined, as shown in Table 5-14. The calibrated parameters for each of the 12 months are listed with minimum RMSE, which varies significantly from month to month. Large RMSEs are associated with the occurrences of some unusual high or low data points (or potential outliers). The following correlation analysis and scatter plots will explain why this happened.

Figures 5-28 and 5-29 demonstrate that the calibrated parameters for January and March plant data are well fitted to the plant effluent data. The missing measured data points are not plotted in the two figures. Figure 5-28 indicates the calibration for the January data from the Seneca WWTP. It clearly shows the simulated data from the calibrated parameters fit the measure plant effluent relatively well, capturing the overall trend at any particular time, even though some of the most extremely values of effluent ammonia are not well represented. The period between 24th and 28th day shows the simulated data are much higher than the measured data although they still following the same decreasing trend. There might be some unexpected sudden changes in plant influent or plant operational conditions. Figure 5-29 demonstrates the calibration for the March data from

Table 5-14. The calibrated parameters for 12-month of the Seneca WWTP.

Month/Year	K_{NH} (mg/L)	$\mu_{A,max}$ (hr^{-1})	Y_A (mg/mg)*	RMSE
1/1996	0.55	0.0176	0.3480	0.1052
2/1996	1.75	0.0416	0.4560	0.0870
3/1996	1.90	0.0512	0.4560	0.1176
4/1996	1.60	0.0608	0.4560	0.0468
5/1996	1.60	0.0608	0.4560	0.0774
6/1996	1.75	0.0368	0.4560	0.1078
7/1996	1.90	0.0464	0.4560	0.1130
8/1996	1.75	0.0608	0.4560	0.0385
9/1996	1.15	0.0560	0.4560	0.0863
10/1996	1.15	0.0608	0.4560	0.0884
11/1996	1.45	0.0608	0.4560	0.1160
12/1996	1.75	0.0480	0.4200	0.1756

Note: * indicates the nearly constant parameter, which would prove to be an insignificant parameter through correlation analysis and scatter plots.

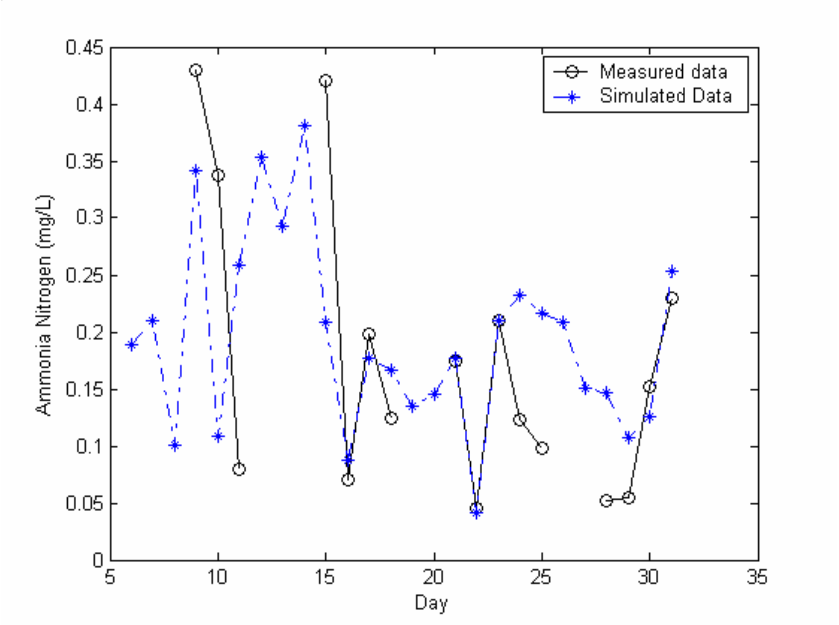


Figure 5-28. The comparison between calibrated and measured data in January from the Seneca WWTP.

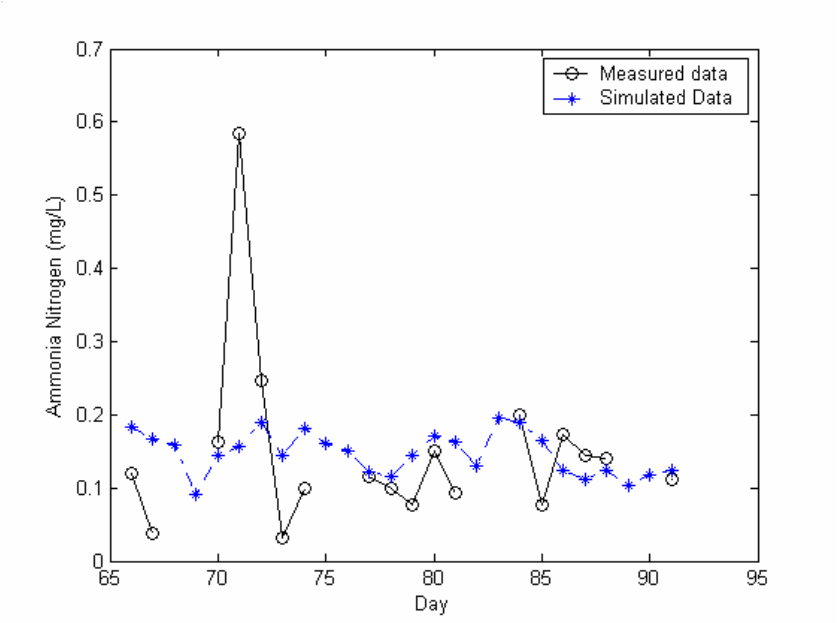


Figure 5-29. The comparison between calibrated and measured data in March from the Seneca WWTP.

the Seneca WWTP. Similar to the January calibrations, the calibrated model captures the trend of historical effluent data relatively well. However, the outlier problem is obvious on the 71th day. The measured data point is around 0.6 mg/L; while the simulated data point is only around 0.16 mg/L. This might be caused by sudden increasing of plant influent loading (for example, the flow rate and/or ammonia-N concentration). That is, some of extremely variability of influent cannot be captured by the fixed calibrated parameters. In this case, outlier detection and replacement method might be preferred. As for the short-term variations in plant performance that cannot be captured by fixed-parameter models, they are not necessarily caused by easily observable deterministic causes.

Table 5-15 shows the correlations between RMSE and the three calibrated parameters. Similar with the table of the Oak Ridge WWTP, the most important parameter is $\mu_{A,max}$ instead of K_{NH}. The negative correlation between RMSE and $\mu_{A,max}$ is very strong with a a value of –0.6446. Similar to the Oak Ridge plant, the correlation coefficient for Y_A is extremely small with a value of –0.00005. The plots of RMSEs against three parameters

Table 5-15. The correlations between RMSE and the three calibration parameters.

(Seneca WWTP)

Items	**K_{NH} (mg/L)**	**$\mu_{A,max}$ (hr^{-1})**	**Y_A(mg/mg)**
RMSE	0.0096	-0.6446	-0.00005

are also conducted to explain their relationships, as shown in Figures 5-30, 5-31, and 5-32.

Figure 5-30 confirms that there are no significant relationships between RMSE and the parameter Y_A. Thus, the parameter Y_A will fixed at the default value of 0.24 mg/mg (listed in Table 2-5) for all Monte Carlo simulations. The relationship between RMSE and K_{NH} is a little complicated, as shown in Figure 5-31. Compared to the same plot of the Oak Ridge WWTP, it only has two different patterns: one is flat, the other is linearly but slowly increasing. This plot indicates that the parameter K_{NH} is not the only factor strongly affecting RMSE. Instead, it is strongly affected by the interaction from $\mu_{A,max}$. This interaction can clearly be seen on Figure 5-32. For $\mu_{A,max}$ less than 0.01 hr^{-1}, which

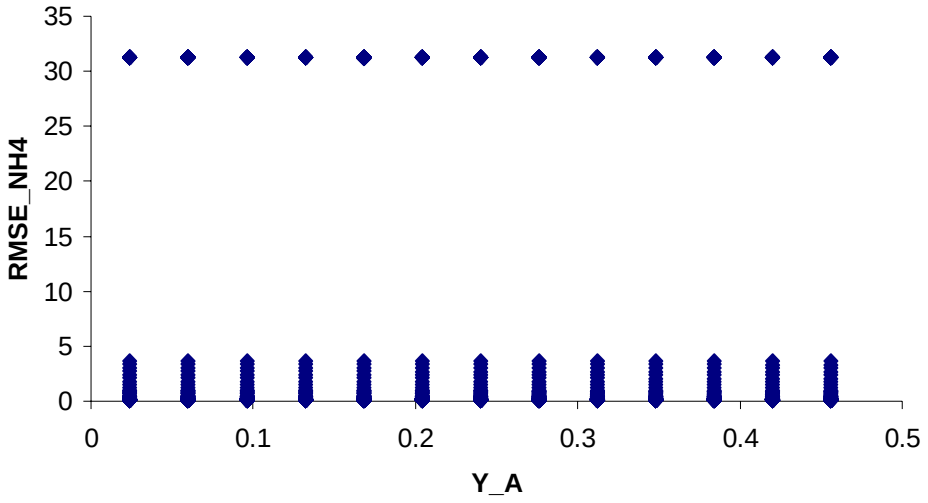


Figure 5-30. The scatterplot of RMSE against Y_A (Seneca WWTP).

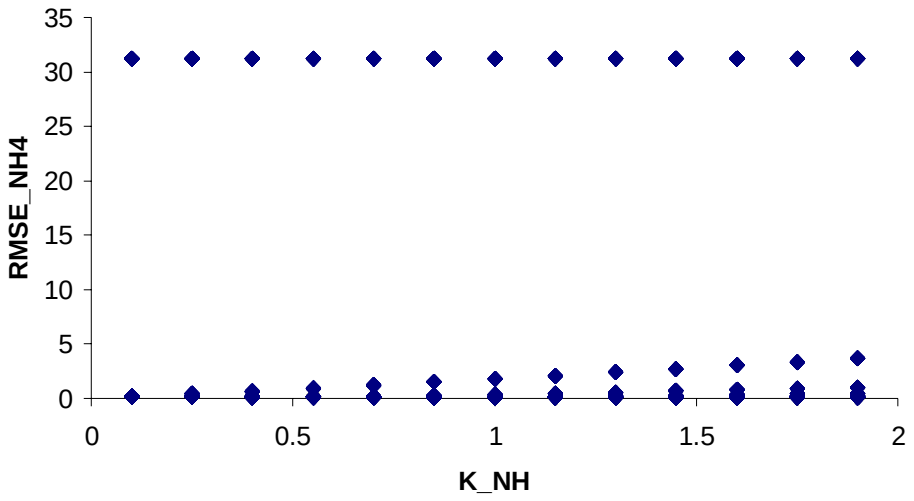


Figure 5-31. The scatterplot of RMSE against K_{NH} (Seneca WWTP).

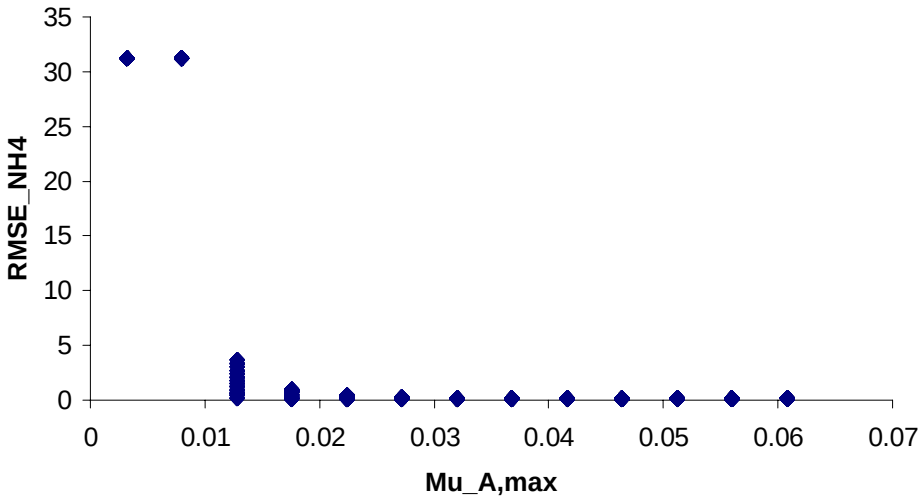
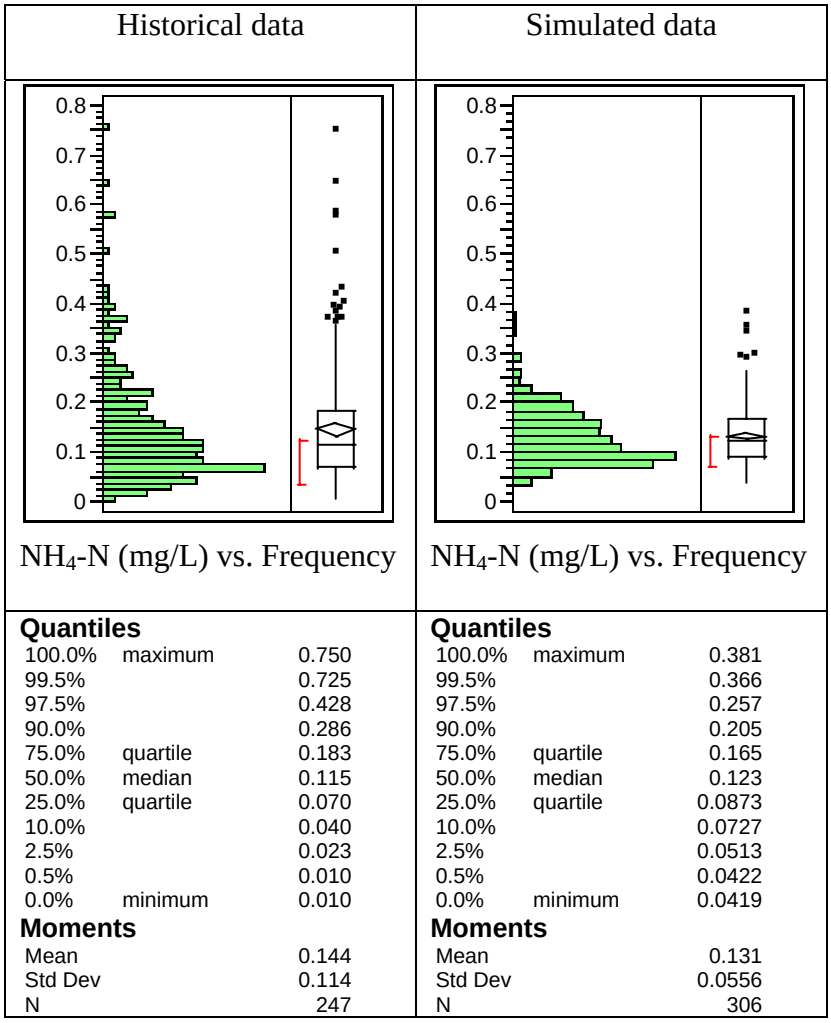


Figure 5-32. The scatterplot of RMSE against $\mu_{A,max}$ (Seneca WWTP).

is extremely small value, there is almost no effect from K_{NH}. For $\mu_{A,max} > 0.01$ hr^{-1}, the effect of K_{NH} keeps decreasing with increasing value of $\mu_{A,max}$ and tends to be no effect at all. The whole plot also shows that the RMSE decreases with increase of $\mu_{A,max}$, although the tailed part increases a little bit. The best calibrated parameters in this case is: $\mu_{A,max}$ = 0.0176 hr^{-1}, and K_{NH} = 0.55 mg/L. Notice that the tail part is almost flat, with little changes when $\mu_{A,max} > 0.02$ hr^{-1}. It is actually not very important to go further and search the best RMSE. This demonstrates that the RMSE is relatively insensitive over a wide range of parameter values. This observation supports our approach in which parameters are represented as distributions instead of as fixed well-defined values.

The statistical characteristics of both historical and simulation results are shown in Figure 5-33. From the plot, the statistical characteristics of the simulated effluent ammonia nitrogen effluent data do not closely match those of the historical data. The mean value and standard deviation of historical data are 0.144 mg/L and 0.114 mg/L compared to 0.131 mg/L and 0.056 mg/L for the simulated data. Furthermore, the highest percentiles of 100.0, 99.5, 97.5 and 90.0 for the historical data (0.750, 0.725, 0.428, and 0.286 mg/L, respectively) are much larger than those in the simulated data (0.381, 0.366, 0.257, and 0.205 mg/L, respectively). In general, we can conclude that the historical data are characterized by greater variability than the simulated data obtained with month-by-month calibrated parameters. Obviously, if we design a plant using the simulation from calibrated parameters, we fail to capture the true variability in plant performance. In other



Quantiles

100.0%	maximum	0.750
99.5%		0.725
97.5%		0.428
90.0%		0.286
75.0%	quartile	0.183
50.0%	median	0.115
25.0%	quartile	0.070
10.0%		0.040
2.5%		0.023
0.5%		0.010
0.0%	minimum	0.010

Moments

Mean	0.144
Std Dev	0.114
N	247

Quantiles

100.0%	maximum	0.381
99.5%		0.366
97.5%		0.257
90.0%		0.205
75.0%	quartile	0.165
50.0%	median	0.123
25.0%	quartile	0.0873
10.0%		0.0727
2.5%		0.0513
0.5%		0.0422
0.0%	minimum	0.0419

Moments

Mean	0.131
Std Dev	0.0556
N	306

Figure 5-33. The comparison of one-year historical and simulated effluent NH_4-N (mg/L) of the Seneca WWTP.

Words, the traditional method using simulation results from calibrated parameter cannot provide a reliable design procedure without an extra safety factor added.

5.2.3 Simulations with Parameter Uncertainty

The previous section discussed the calibration procedure and calibrated parameters for the Seneca WWTP. The comparison between historical and simulated effluent data indicates that the design methods that better represent the variability in the plant are needed. Thus, we will investigate the use of distributed parameters in conjunction with Monte Carlo simulations to assess the variability. Similar with the procedure used in the Oak Ridge WWTP, there are also two options for parameter distributions. One is the universal (or non-site-specific) distributed parameter generated by Cox (2004) based on the published values all over the world; the other is the site-specific distributed parameter that will be generated as follows using the Bayesian method and site-specific calibrated parameter.

Results of the Monte Carlo simulation with the universal parameter distributions is conducted are shown in Figure 5-34. One month historical influent data (January) are used for this simulation. Obviously, the Monte Carlo simulation results with the universal parameter distributions overestimate both the magnitude and the variability of the historical effluent. The mean value and standard deviation of the simulation results are

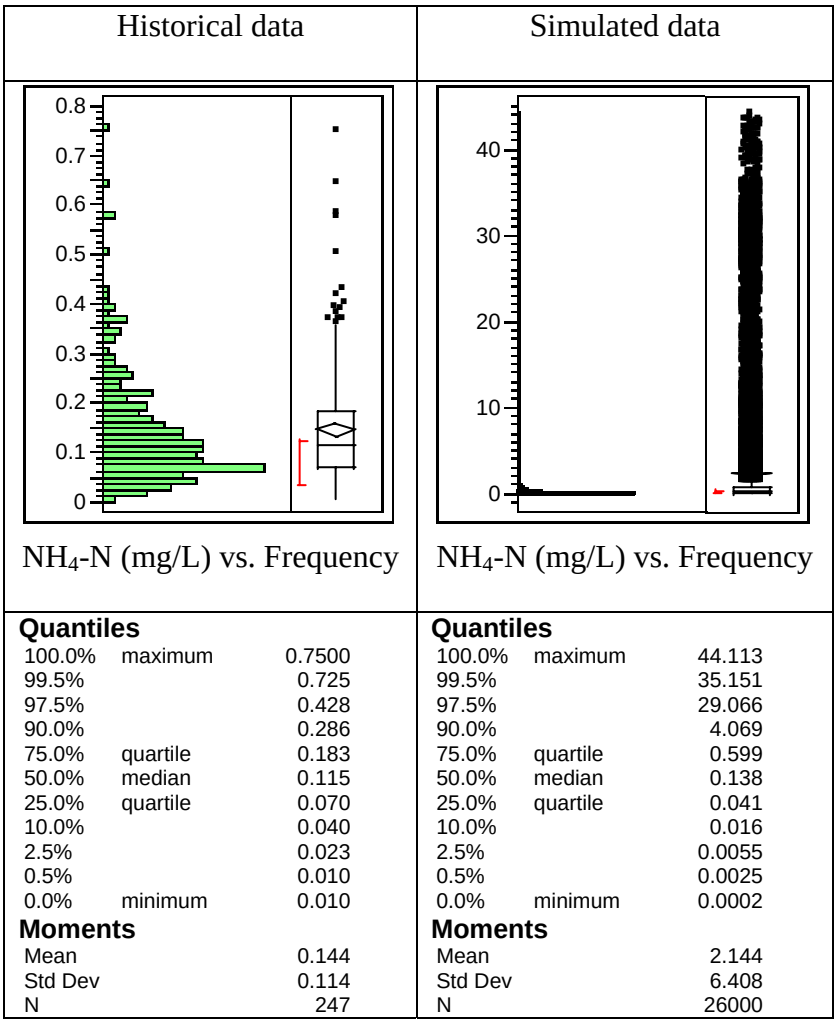


Figure 5-34. Monte Carlo simulations of effluent NH_4-N (mg/L) with the universal parameter distributions and January historical influent from the Oak Ridge WWTP.

2.144 mg/L and 6.408 mg/L, respectively, compared to 0.144 mg/L and 0.114 mg/L for the historical data. As for the highest percentiles, they are ridiculously higher than the historical data with a maximum value of 44.113 mg/L and 90.0 percentile of 4.069 mg/L. In short, the universal parameter distributions introduce too much unrealistic uncertainty into wastewater treatment systems.

The two previous simulations indicate that neither the fixed calibrated parameters nor universal parameter distributions accurately capture the variability observed in the actual plant. We have made the same conclusions for the Oak Ridge WWTP. The fixed calibrated parameters can predict the average effluent well. However, the simulated results notably underestimate the variability observed in historical effluent mostly because the ASM1 model with the calibrated parameters cannot capture all the sharp peaks due to sudden changes of the plant influent. The universal parameter distributions simply introduce too much unrealistic uncertainty, thus leading to inaccurate simulation results. Two key reasons are 1) too much uncertainty in the model parameters is considered; 2) and the likelihood of sampling extreme parameters from multiple distributions simultaneously. To reduce the uncertainty in the parameters, we will consider site-specific parameter distributions. These site-specific parameters can better describe the plant performance because they are generated from calibrated parameters while preserving the consideration of uncertainty.

The site-specific parameter distributions for K_{NH} and $\mu_{A,max}$ are generated from the monthly calibrated parameters using the Bayesian method. The results are summarized in Table 5-16. The detailed Bayesian procedure was discussed in Chapter III, which has been used in the Oak Ridge WWTP section. The standard deviation of the prior distribution is also calculated using the equation $\sigma = 0.25 \times \mu$ instead of $\sigma = 2 \times 0.25 \times \mu$ in order to narrow down the posterior distribution, especially for the site-specific parameter distributions. For details, please refer to Cox (2004) and previous section for the Oak Ridge WWTP.

Figure 5-35 shows the universal (UPD) and site-specific (SPD) parameter distributions of the parameter K_{NH}. It seems like the UPD is "narrower" than the SPD, however, it is actually not. Because all values of a lognormal distribution are larger than zero, the PDF curve of a lognormal distribution is not symmetric. We will discuss this together with the percentile and probability tables later. From there, we will see how site-specific parameters narrow down. Figure 5-36 shows the comparison of the UPD and SPD for the parameter $\mu_{A,max}$. Both figures indicate the curves of two SPD both move to right a little bit. In other words, the mean and median values of two SPD are both larger than those of UPD.

Table 5-16. The universal and site-specific parameter distributions for the Seneca WWTP.

Items	**K_{NH} (mg/L)**	**$\mu_{A,max}$ (hr^{-1})**
UPD: μ	-0.675	-3.688
UPD: σ	1	0.44
SPD: μ	0.2678	-3.1621
SPD: σ	0.4092	0.4203
Percentiles (%)	**K_{NH} (mg/L)**	**$\mu_{A,max}$ (hr^{-1})**
1	0.5045	0.0159
5	0.6668	0.0212
10	0.7736	0.0247
25	0.9918	0.0319
50	1.3071	0.0423
75	1.7225	0.0562
90	2.2082	0.0725
95	2.5622	0.0845
99	3.3863	0.1125

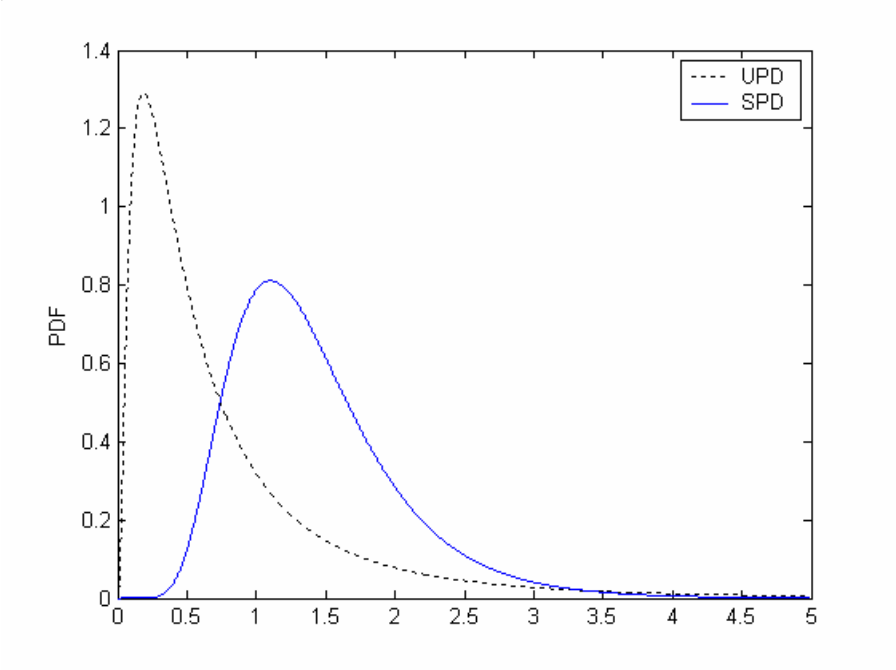


Figure 5-35. PDF plots of K_{NH} (mg/L) for the UPD and SPD for the Seneca WWTP.

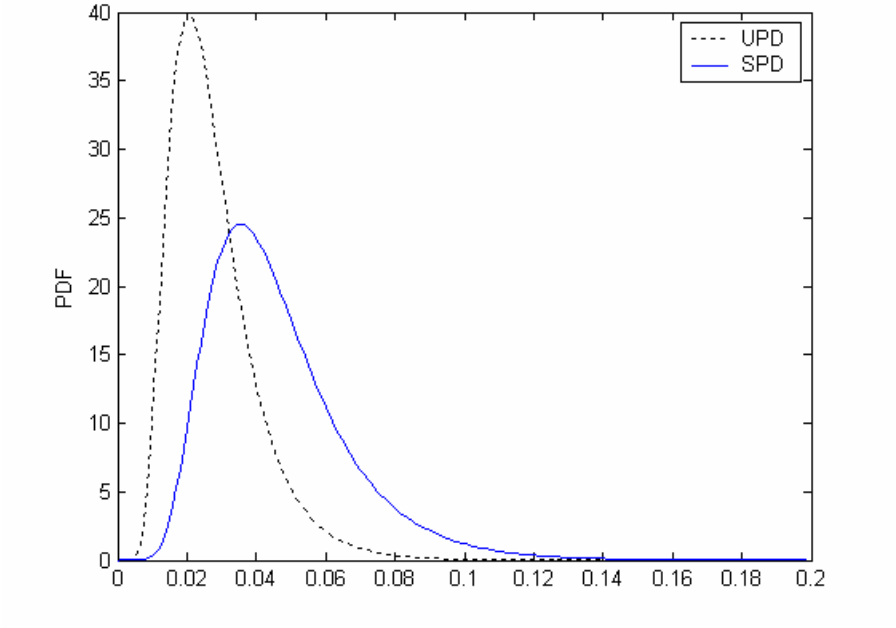


Figure 5-36. PDF plots of μ_{Amax} ($hour^{-1}$) for the UPD and SPD for the Seneca WWTP.

Table 5-17 shows the percentile comparison between two kinds of SDP. Distributions labeled SPD calculate the prior standard deviation using formula: $\sigma = 0.25 \times \mu$, while parameters SPD' calculate the prior standard deviation using formula: $\sigma = 2 \times 0.25 \times \mu$ as originally recommended by Cox 2005. Obviously, SPD' will have a wider prior distribution and generate a wider posterior distribution. That is, SPD' will generate a posterior distribution with more uncertainty.

For the parameter K_{NH}, the SPD parameter actually gets the narrowest range between 5 and 95 percentiles compared to either UPD or SPD' parameters. However, as for the parameter $\mu_{A,max}$, it seems that the SPD parameter is narrower than SPD' parameter although both of them are wider than the UPD parameter.

Table 5-17. The percentile comparison between UPD and two SPDs.

Items	**K_{NH} (mg/L)**		**$\mu_{A,max}$ (hr^{-1})**	
	0.05	**0.95**	**0.05**	**0.95**
UPD	0.0983	2.6376	0.0121	0.0516
SPD'	0.7227	2.8332	0.0231	0.0940
SPD	0.6668	2.5622	0.0212	0.0845

Until now, we obtained the two kinds of SPDs (SPD and SPD') for K_{NH} and $\mu_{A,max}$. However, we need to determine which one is better to conduct following Monte Carlo simulations Table 5-18 compares the range of the monthly calibrated values to percentiles of SPD and SPD'. For example, for the parameter $\mu_{A,max}$, the minimum and maximum calibrated values are 0.0176 and 0.0608 day^{-1}, respectively. The minimum value falls at the lowest 2 percent of SPD' while the maximum value falls at the lowest 74 percent. Therefore, the probability of a random sample from SPD' to fall in the range of the calibrated value is approximately 72 percent. We believe that this probability is a little too low and adds too much uncertainty to the distributions; therefore, we used the SPD estimates to obtain slightly more narrow site-specific distributions.

Table 5-18. The probabilities of the data range between minimum and maximum calibrated parameters.

Items	**K_{NH} (mg/L)**		**$\mu_{A,max}$ (hr^{-1})**	
Range of calibrated values	**Min** **0.5500**	**Max** **1.9000**	**Min** **0.0176**	**Max** **0.0608**
PDF (SPD')	0.0107	0.7526	0.0111	0.7329
Probability (SPD')		**0.7420**		**0.7218**
PDF (SPD)	0.0172	0.8197	0.0184	0.8054
Probability (SPD)		**0.8025**		**0.7870**

From the listed values in Table 5-14, the correlation coefficient of $\mu_{A,max}$ and K_{NH} is calculated to be 0.3858. That is, there are some but not significant correlations between those two parameters. Thus, in the following simulations, those two parameters are randomly and independently sampled. In fact, the following simulation results demonstrate that the Monte Carlo simulations are still very good even if no correlations are considered. The correlated random samples will continue in the future research if adequate information is available. Similar with the Oak Ridge WWTP, three kinds of parameters are considered: SPD#1-AP: all parameters are sampled from the UPD expect for $\mu_{A,max}$ and K_{NH}, which are sampled from their SPD; SPD#2-2P: all parameters are fixed except for $\mu_{A,max}$ and K_{NH}, which are sampled from their SPD; and SPD#3-1P: all parameters are fixed except for $\mu_{A,max}$, which is sampled from its SPD.

Figure 5-37 shows the comparison of Monte Carlo simulations to the historical data using three different SPDs for the Seneca WWTP. Clearly, the simulation results with SPD#2-2P and SPD#3-1P are much better than the results with SPD#1-AP Variability in fewer parameters yielded better simulation results. In fact, the SPD#3-1P best describes the effluent ammonia-N when considering the mean, standard deviation and 99.5 percentile values. However, the simulation results with SPD#2-2P better represent the 97.5 and 90.0 percentiles. Nevertheless, we recommend SPD#3-1P, after considering the overall distribution of effluent of ammonia-N concentration. As for the maximum values, all three distributions predict much greater values than observed in the historical record; however, undue weight should not be given to these extreme points.

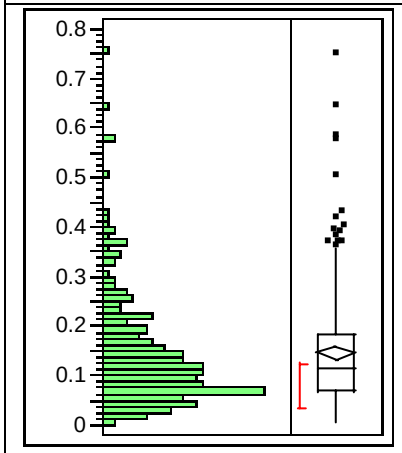
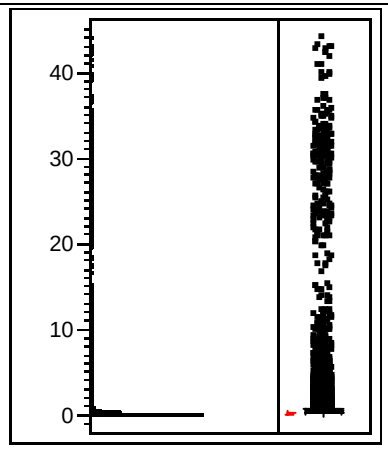
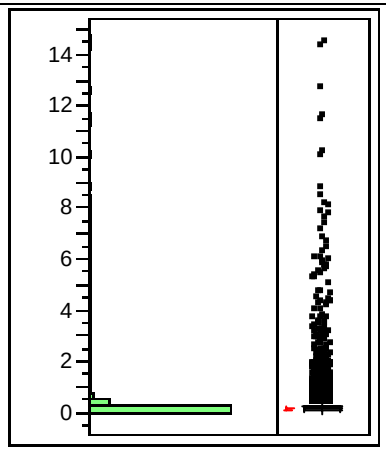
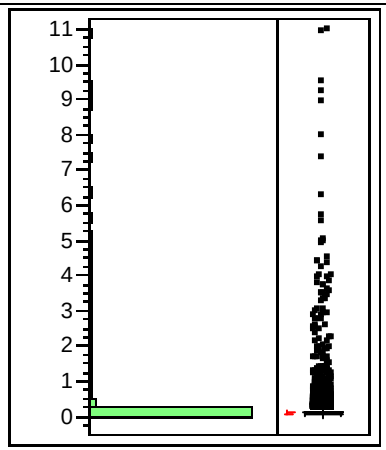
Historical data	SPD#1-AP	SPD#2-2P	SPD#3-1P
			
NH_4-N (mg/L) vs. Frequency	NH_4-N (mg/L) vs. Frequency	NH_4-N (mg/L) vs. Frequency	NH_4-N (mg/L) vs. Frequency
Quantiles	**Quantiles**	**Quantiles**	**Quantiles**
100.0% maximum 0.7500	100.0% maximum 43.976	100.0% maximum 14.503	100.0% maximum 10.952
99.5% 0.725	99.5% 26.107	99.5% 1.735	99.5% 0.770
97.5% 0.428	97.5% 2.350	97.5% 0.677	97.5% 0.338
90.0% 0.286	90.0% 0.674	90.0% 0.333	90.0% 0.184
75.0% quartile 0.183	75.0% quartile 0.296	75.0% quartile 0.193	75.0% quartile 0.118
50.0% median 0.115	50.0% median 0.133	50.0% median 0.110	50.0% median 0.078
25.0% quartile 0.070	25.0% quartile 0.066	25.0% quartile 0.066	25.0% quartile 0.053
10.0% 0.040	10.0% 0.037	10.0% 0.042	10.0% 0.038
2.5% 0.023	2.5% 0.018	2.5% 0.025	2.5% 0.025
0.5% 0.010	0.5% 0.0064	0.5% 0.015	0.5% 0.016
0.0% minimum 0.010	0.0% minimum 0.0011	0.0% minimum 0.0046	0.0% minimum 0.0056
Moments	**Moments**	**Moments**	**Moments**
Mean 0.144	Mean 0.523	Mean 0.179	Mean 0.110
Std Dev 0.114	Std Dev 2.570	Std Dev 0.352	Std Dev 0.223
N 247	N 26000	N 26000	N 26000

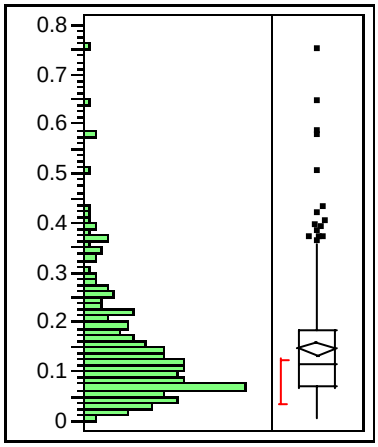
Figure 5-37. Monte Carlo simulations of effluent NH_4-N (mg/L) with SPD parameters and January historical influent from the Seneca WWTP.

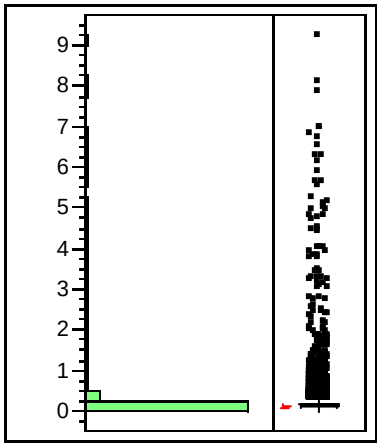
5.2.4 Simulations with Influent Variability

In last section, we discussed the effect of the parameter uncertainty on the Monte Carlo simulations. In this section, we will elaborate the effect of variability in the plant influent. In the Oak Ridge WWTP section, the randomly generated influent option has proved to be one of best choices to conduct the Monte Carlo simulations. However, compared to the simulations with historical influent, there are no significant differences. This also proved that the major uncertainty in the plant performance is from the uncertainty in model parameters but the variability in the plant influent. Therefore, we only considered the historical influent data in the Seneca WWTP section.

The previous simulation results in last section are the simulations from the January historical influent data with different parameter distributions. Since Winter and Summer may have different effects on plant performance, we would like to consider both in this dissertation to ensure a reasonable and reliable simulation. We take August and January typical months of Summer and Winter, respectively. Since January was already considered in the previous section, we will now conduct a similar simulation based on August historical influent data. The simulated results are shown in Figure 5-38. Surprisingly, the mean values and standard deviations of the SPD#3-1P simulated data are slightly larger than the results from the January influent data. The highest percentiles of 100, 99.5 and 97.5 are also slightly larger than the simulations with January data. The simulated data then are closer to the historical effluent data. Thus, for the purpose of Monte Carlo simulations, we conclude that the August data are the best choice if only one

Historical data **SPD#3-1P**





NH_4-N (mg/L) vs. Frequency NH_4-N (mg/L) vs. Frequency

Quantiles

100.0%	maximum	0.7500
99.5%		0.725
97.5%		0.428
90.0%		0.286
75.0%	quartile	0.183
50.0%	median	0.115
25.0%	quartile	0.070
10.0%		0.040
2.5%		0.023
0.5%		0.010
0.0%	minimum	0.010

Quantiles

100.0%	maximum	9.240
99.5%		1.061
97.5%		0.462
90.0%		0.250
75.0%	quartile	0.161
50.0%	median	0.106
25.0%	quartile	0.0725
10.0%		0.0519
2.5%		0.0338
0.5%		0.0219
0.0%	minimum	0.0081

Moments

Mean	0.144
Std Dev	0.114
N	247

Moments

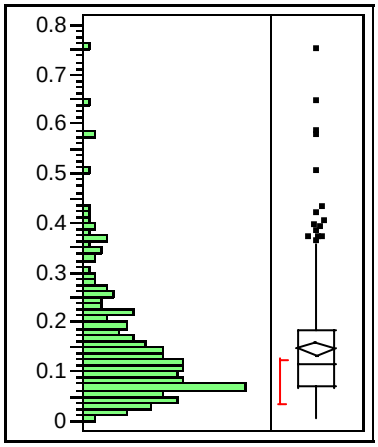
Mean	0.148
Std Dev	0.244
N	26000

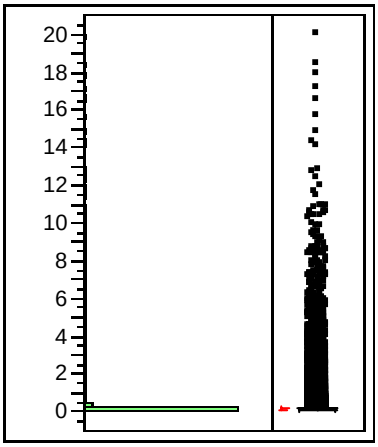
Figure 5-38. Monte Carlo simulations of effluent NH_4-N (mg/L) with the SPD#3-1P parameters and August historical influent for the Seneca WWTP.

Month influent data are required. We strongly recommend conducting simulations over at least two months (for examples, January and August) to indicate two extreme cases in the plant performance. However, in this dissertation, we would like to take August influent data as an example to study the further uncertainty in the plant performance in the following simulations.

In order to confirm whether the one months' influent data (either January or August) can provide reliable and reasonable simulations, we conduct 1000 Monte Carlo runs with one-year influent data. Figure 5-39 shows the Monte Carlo simulation with the SPD#3-1P parameters and one-year historical influent data from the Seneca WWTP. The mean and standard deviation of 1000 Monte Carlo runs are 0.124 mg/L and 0.248. The mean and standard deviation of the historical plant effluent data are 0.144 mg/L and 0.114. More important, the percentiles of 99.50 and 97.50 of Monte Carlo simulation results are both very close to real plant effluent data. Overall Monte Carlo simulations using August and full-year data yield very similar results, thus justifying the use of the one-month influent history.

Historical data **Simulated data**





NH_4-N (mg/L) vs. Frequency NH_4-N (mg/L) vs. Frequency

Quantiles

100.0%	maximum	0.7500
99.5%		0.725
97.5%		0.428
90.0%		0.286
75.0%	quartile	0.183
50.0%	median	0.115
25.0%	quartile	0.070
10.0%		0.040
2.5%		0.023
0.5%		0.010
0.0%	minimum	0.010

Moments

Mean	0.144
Std Dev	0.114
N	247

Quantiles

100.0%	maximum	19.996
99.5%		0.882
97.5%		0.383
90.0%		0.209
75.0%	quartile	0.134
50.0%	median	0.087
25.0%	quartile	0.059
10.0%		0.043
2.5%		0.030
0.5%		0.021
0.0%	minimum	0.0056

Moments

Mean	0.124
Std Dev	0.248
N	26000

Figure 5-39. Monte Carlo simulation of effluent NH_4-N (mg/L) with the SPD#3-1P parameters and one-year historical influent for the Seneca WWTP.

5.2.5 Discussions

Table 5-19 summarizes the simulations completed in the previous sections. Obviously, the best simulation is conducted with one full year historical influent and the SPD#3-1P parameters. This best simulation has the best grade score: A+, considering the mean, standard deviation, the highest percentiles, and histogram plot. The second best grade: A is given to the simulation with one-month historical influent data and the SPD#3-1P parameters. The best parameter distributions are the SPD#3-1P parameters, that is, all fixed calibrated parameters except the SPD parameter $\mu_{A,max}$. It is an important finding that reliable Monte Carlo simulation can be achieved by considering only the most important parameter: $\mu_{A,max.}$ Future investigations may seek to determine if alternative

Table 5-19. Summary table of Monte Carlo simulations of the Seneca WWTP.

Influent	Month/year	Fixed	UPD	SPD#1-AP	SPD#2-2P	SPD#3-1P
Historical	1/1996		C	C+	A-	A
	8/1996					A (No.2)
	1996	B(*)				A+ (No.1)

Note: * = Underestimate.

Methods of determining the distribution of the key parameter might result in better representations of plant variability.

Furthermore, Table 5-19 also indicates that the variability of the influent data are not the major effect on the uncertainty features of the plant effluent, similar with the results in the Oak Ridge WWTP section. In short, from the Monte Carlo simulations of historical influent data and the SPD#3-1P; and the second best choice is the simulation with one month historical data and the SPD#3-1P.

The Cramer-von Mises goodness-of-fit tests are also conducted to quantify the evaluation of the simulation results in Table 5-19. The T_2 statistic values are shown in Table 5-20. Surprisingly, Table 5-20 indicates that the best choice is the simulation with the SPD#2-2P parameters and January's historical influent, with the smallest T_2 value of 0.0292. The No.1 choice in Table 5-19, i.e., the simulation with SPD#3-1P parameters and one-year

Table 5-20. GoF tests of Monte Carlo simulations of the Seneca WWTP.

Influent	Month/year	Fixed	UPD	SPD#1-AP	SPD#2-2P	SPD#3-1P
Historical	1/1996		2.6072	0.4671	0.0292	0.7595
	8/1996					0.4142
	1996	0.3385				0.1699

historical influent, is ranked as No.2 choice here, instead. Because both T_2 values are less than $\omega_{0.99}$=0.743, we conclude the null hypothesis of identical distribution functions is accepted at α=0.01. Surprisingly, the simulation with the SPD#3-1P parameters and August historical influent are not well presented as a result of the T_2 statistics. This method is ranked as the No.3 choice in Table 5-19. However, in Table 5-20, the simulation with the SPD#2-2P parameters and January historical influent is much better and well deserves the No.1 choice, instead. The tests also indicate that the simulation with August historical influent (T_2=0.4142) is slightly better than the simulation with January historical influent (T_2=0.7595). As discussed in the Oak Ridge WWTP section, the goodness-of-fit test is only based on a statistical standard. This standard might be inappropriate for the applications in practice. For example, the simulation with the SPD#1-AP parameters is obviously much worse than the simulation with the SPD#3-1P parameters as shown in Figure 5-37. However, in Table 5-20, the goodness-of-fit tests indicate that this simulation (T_2=0.4671) is even better than the simulation with the SPD#3-1P parameters (T_2=0.7595). Thus, we believe that this Cramer-von Mises goodness-of-fit test cannot work for all the cases. In this dissertation, the evaluation of the simulation results mostly focuses on the highest percentiles (99.50 and 97.50 percent), mean and standard deviations, which are considered the most important factors affecting the plant performance. Further research is needed to find a better goodness-of-fit test emphasizing more on the highest percentiles, mean, and standard deviation.

Same with the decision for the Oak Ridge WWTP, the second best choice with the one-month influent data will be chosen as the best options for Monte Carlo simulations because it proved to provide a reasonable and reliable simulated results. From Table 5-12, one-month simulation only needs 1-3 days, which is much less than the simulation with one year simulation (3-7 days). Thus, unless the best accuracy is needed, we can use the Monte Carlo simulation with one-month influent data (either Winter January or Summer August data) and the SPD#3-1P parameters. The following simulations will consider how changing SRT affects the plant effluent NH_4-N concentration. These simulations of effluent NH_4-N for different SRT = 5, 10, 20 and 30 days are conducted with the SPD#3-1P parameters and the August historical influent data, as shown in Figure 5-40. It obviously indicates that the mean values and the highest percentiles (99.50, 97.50, and 90.00) decrease with increasing SRT.

Figures 5-41 and 5-42 show the relationship between the percentiles of the effluent NH_4-N and SRT. The maximum effluent ammonia nitrogen is observed to increase with SRT over the range of 10 to 30 days; however, too much emphasis should not be placed on the most extreme values. The 99.50, 97.50 and 90.00 percentiles provide a constant decrease with increasing SRT. For example, for the case of SRT = 10 days, the 97.5 percent of data are under 1.199 mg/L, which is much less than the requirement (Daily maximum: 3.6 mg/L). Normally, regulatory limitations on plant effluent employ moving average methods to allow occasional short-term plant upsets. Some extremely high effluent values will be smoothed out by 7-day or 30-day averages. Thus, a new statistical-based effluent

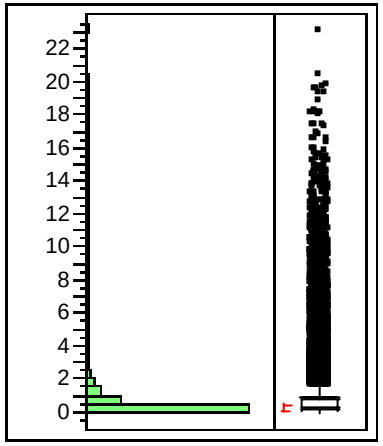
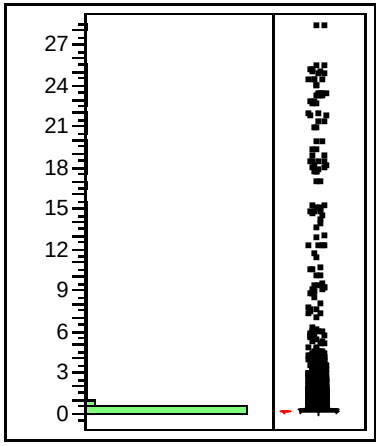
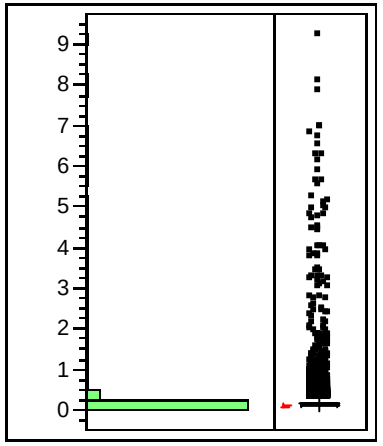
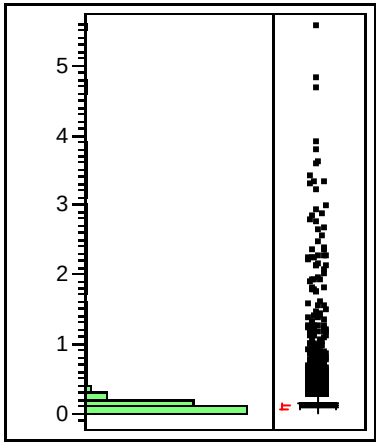
SRT = 5 days	SRT = 10 days	SRT = 20 days	SRT = 30 days
			
NH_4-N (mg/L) vs. Frequency	NH_4-N (mg/L) vs. Frequency	NH_4-N (mg/L) vs. Frequency	NH_4-N (mg/L) vs. Frequency
Quantiles	**Quantiles**	**Quantiles**	**Quantiles**
100.0% maximum 23.076	100.0% maximum 28.210	100.0% maximum 9.240	100.0% maximum 5.566
99.5% 11.744	99.5% 3.768	99.5% 1.061	99.5% 0.690
97.5% 5.710	97.5% 1.199	97.5% 0.462	97.5% 0.349
90.0% 1.993	90.0% 0.475	90.0% 0.250	90.0% 0.209
75.0% quartile 0.747	75.0% quartile 0.248	75.0% quartile 0.161	75.0% quartile 0.141
50.0% median 0.286	50.0% median 0.142	50.0% median 0.106	50.0% median 0.0955
25.0% quartile 0.139	25.0% quartile 0.090	25.0% quartile 0.0725	25.0% quartile 0.0666
10.0% 0.081	10.0% 0.061	10.0% 0.0519	10.0% 0.0486
2.5% 0.047	2.5% 0.036	2.5% 0.0338	2.5% 0.0328
0.5% 0.031	0.5% 0.022	0.5% 0.0219	0.5% 0.0221
0.0% minimum 0.012	0.0% minimum 0.0078	0.0% minimum 0.0081	0.0% minimum 0.0087
Moments	**Moments**	**Moments**	**Moments**
Mean 0.848	Mean 0.296	Mean 0.148	Mean 0.123
Std Dev 1.695	Std Dev 1.093	Std Dev 0.244	Std Dev 0.142
N 26000	N 26000	N 26000	N 26000

Figure 5-40. Monte Carlo simulations of effluent NH_4-N (mg/L) with the SPD#3-1P parameters and the August historical influent for the Seneca WWTP.

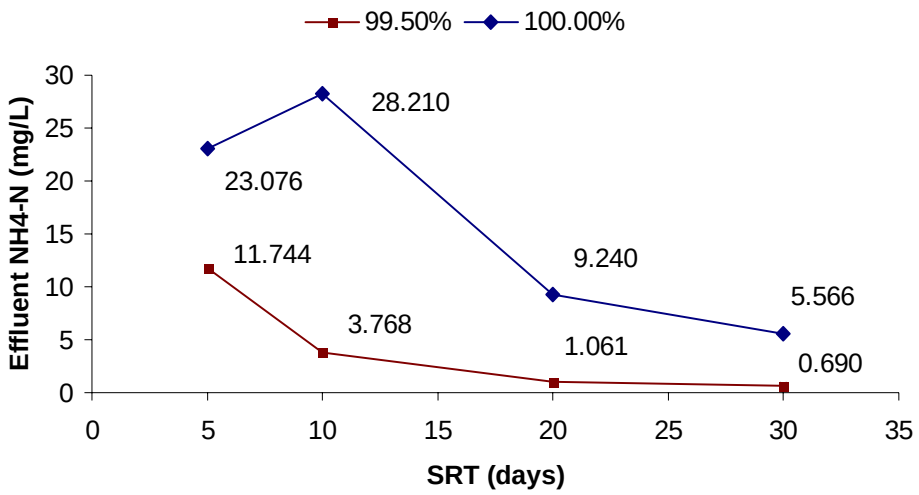


Figure 5-41. The percentile plot (100.00 and 99.50) against SRTs (Seneca WWTP).

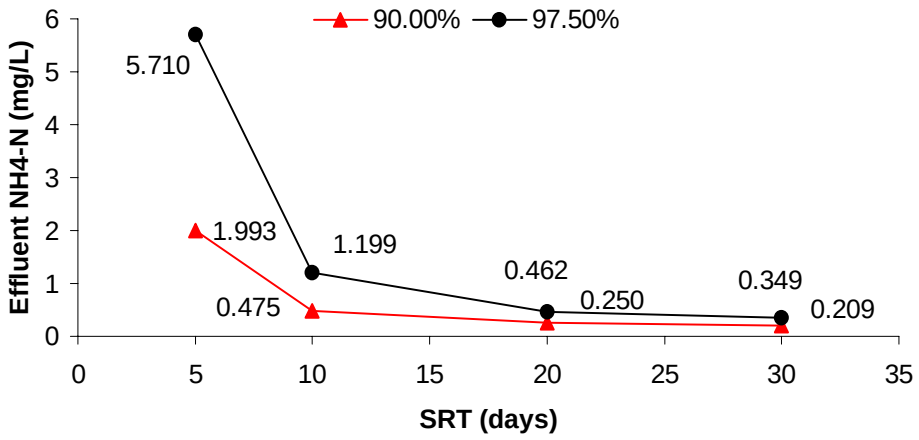


Figure 5-42. The percentile plot (90.00 and 97.50) against SRTs (Seneca WWTP).

regulation and corresponding statistical-based design procedure are needed to improve our design and operation of WWTPs in a more cost-effective way. In Chapter VI, we will elaborately discuss how to apply the StatASPS design procedure in practice based on the Monte Carlo simulations of WWTPs.

CHAPTER VI

CONCLUSIONS AND FUTURE WORK

In Chapter V, we conducted Monte Carlo simulations of two existing WWTPs: the Oak Ridge and Seneca WWTPs in the StatASPS program. In this chapter, we continue to discuss how we will apply the Monte Carlo simulated results to improve the design of WWTPs in practice.

Figure 6-1 shows the Monte Carlo based design procedure. As shown in the figure, we considered two kinds of uncertainty in the system: 1) influent data, and 2) model parameters. Furthermore, we considered two options for each kind of uncertainty. For influent data, we could choose either historical influent data or randomly generated influent data. Historical influent data (either one year or one month), marked as No.1 in Figure 6-1, has proved to be the best choice to be introduced into the Monte Carlo simulations based on the results of the previous chapter. If conditions permit, simulations based on one year's historical influent data are strongly recommended. However, due to limitations of current computing technologies, we may also choose one month's data instead. In this case, simulations based on both Winter and Summer influent conditions are strongly recommended, considering most of effluent standards are regulated based on these two cases. As for the randomly generated influent data, it has two obvious advantages: 1) it captures similar statistical features but is not limited to the historical

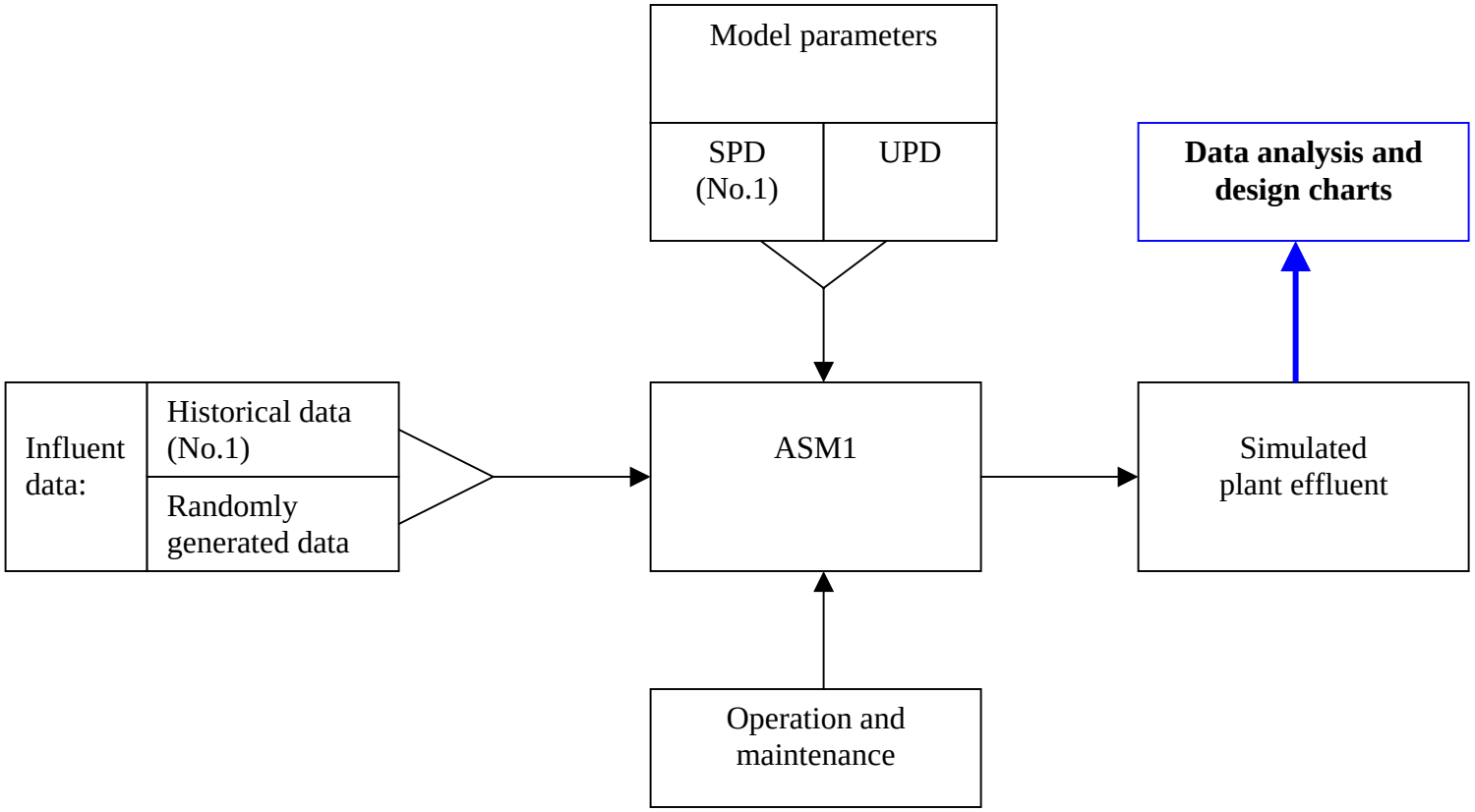


Figure 6-1. The Monte Carlo based design procedure.

data; and 2) it has slightly more uncertainty than the historical data, which provides an additional margin of safety.

For model parameters, we could choose either SPD (site-specific parameter distributions) or UPD (universal parameter distributions). SPD parameters, marked as the No.1 choice in Figure 6-1, are strongly recommended, which have proved to be the best choice, according to the simulations of both Oak Ridge and Seneca WWTPs. Even for SPD parameters, there are at least three types: SPD#1-AP: all universal parameter distributions except for the site-specific parameter distributions $\mu_{A,max}$ and K_{NH}; SPD#2-2P: all fixed parameters except for the site-specific parameter distributions $\mu_{A,max}$ and K_{NH}; and SPD#3-1P: all fixed parameters except for the site-specific parameter distribution $\mu_{A,max}$. The simulations of both Oak Ridge and Seneca WWTPs prove that the SPD#3-1P is the best choice choice of distributional model parameters to conduct reliable and relatively accurate Monte Carlo simulations. Notice that the SPD parameters are generated from the month-by-month calibrated parameters using the Bayesian method. It is very possible that calibrated or measured parameters are not available for the plant of interest. In this case, we might conduct experiments to determine them or perform the calibration procedure for some other WWTPs having similar processes and loading conditions. For the worst case, i.e., the similar WWTPs are also not available; we have no choice but choose the UPD parameters.

After conducting Monte Carlo simulations, we may analyze the simulated plant effluent data, which is in and of itself a distribution. If measured effluent data are available, we may compare the simulated results with real plant effluent. With the correct feedback, we can improve either the distributions of model parameters or the methods of generating/choosing influent data. This provides another chance to improve our design procedure. Traditional design approaches employ empirical safety factors that commonly lead to the over-design of WWTPs. If further developed, the Monte Carlo based procedure has potential to be a more robust design method.

Based on the simulated results from the StatASPS program, we can generate two kinds of plots for decision-making purposes. One is a plot of 99.50 and 97.50 percentiles ammonia nitrogen concentration against SRTs, also called percentile control chart. The other is a plot of plant effluent concentration against the violation percentage with different SRT values, also called decision-making chart. For the Oak Ridge WWTP, the percentile control chart is shown in Figure 6-2. The green line is the effluent standard, daily maximum concentration: 3.6 mg/L for effluent ammonia nitrogen. This standard is the most stringent value set for the Winter standard. Because we desire the system to be as reliable as possible, we would rather take this most stringent standard for decision-making purposes. From Figure 6-2, if our target is to control the 97.50 percent of the daily effluent data to be under the required standard (i.e., the green line), all four SRTs easily meet the

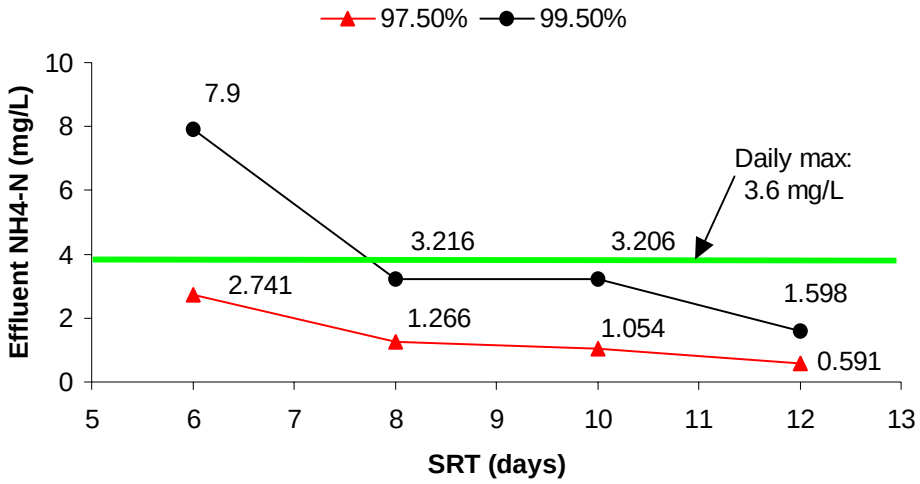


Figure 6-2. The percentile control chart of the Oak Ridge WWTP.

requirement. In other words, we may take SRT = 6 days as our design parameter if more cost-effective plant design is needed. However, if more reliable system is preferred, we may choose higher percent, for example 99.50 percent. With this design standard, we are 99.50 percent sure that the design plant will perform under the required effluent standard. From Figure 6-2, we can conclude that only simulations with SRT ≥ 8 days meet the requirement. The 99.50 percentile of the simulation with SRT = 6 days is 7.9 mg/L, which is above the green line (i.e., daily maximum concentration: 3.6 mg/L). In this case, we can only decrease the design SRT from current 10 days to 8 days if a more cost effective plant design is needed.

The decision-making chart of the Oak Ridge WWTP is shown in Figure 6-3. The x-axis is the percentage of effluent ammonia concentrations that exceed the effluent standard. The Y-axis is the effluent ammonia nitrogen concentration. To apply this decision-making chart, we need first to determine what the acceptable violation percentage. For example, suppose 0.5 violation percent (i.e., refer to the 99.50 percentile) is chosen, it clearly indicates that only simulations with SRT ≥ 8 days meet the requirement. Notice that the determination of acceptable violation percentage is very subjective. That is, it actually is a judgmental call. Different people may have different opinions. In this dissertation, we suggest that 99.50 or 97.50 percent compliance may be reasonable standards to design WWTPs. The corresponding violation percentages are 0.50 and 2.50 percent, respectively, which is believed to be safe enough for a well-performing system.

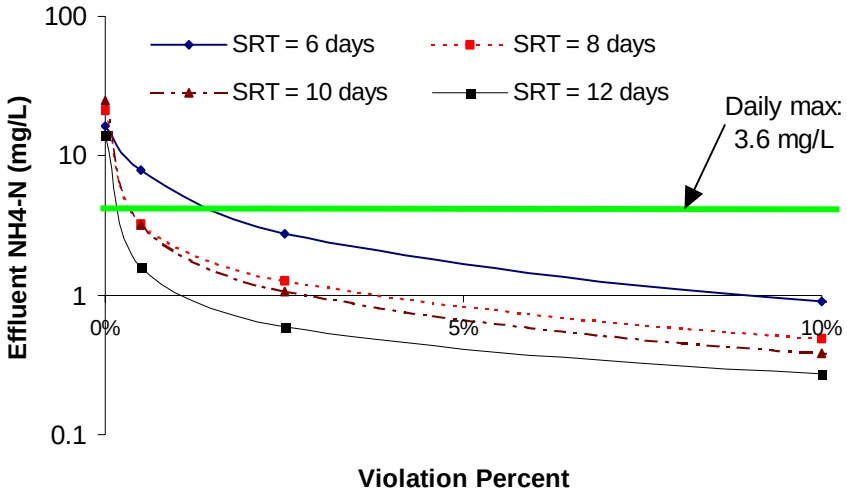


Figure 6-3. The decision-making chart of the Oak Ridge WWTP.

Two similar plots are constructed for the Seneca WWTP, as shown in Figures 6-4 and 6-5. Suppose the acceptable violation percentage is 2.50 percent, the simulations with SRT ≥ 10 days all meet the requirement. That is, we might decrease SRT from 20 days to 10 days for design purposes. In this case, we are 97.50 percent sure that this designed plant would perform under the required standard. However, if more stringent standard is required, we may choose the 99.50 percent as a design standard. From two previous figures, we can conclude that only simulations with SRT ≥ 20 days meet the requirement. In this case, we are 99.50 percent sure that this designed plant would perform under the required standard, that is, only 0.50 percent of effluent data would violate the required standard.

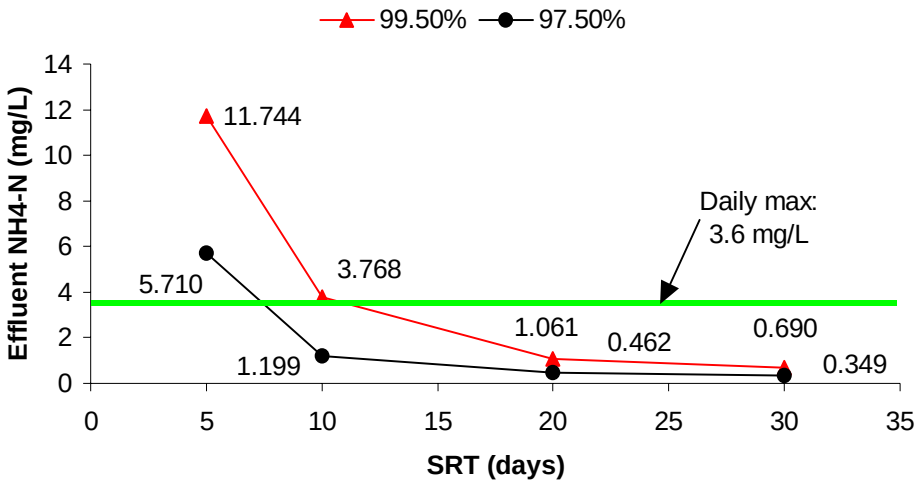


Figure 6-4. The percentile control chart of the Seneca WWTP.

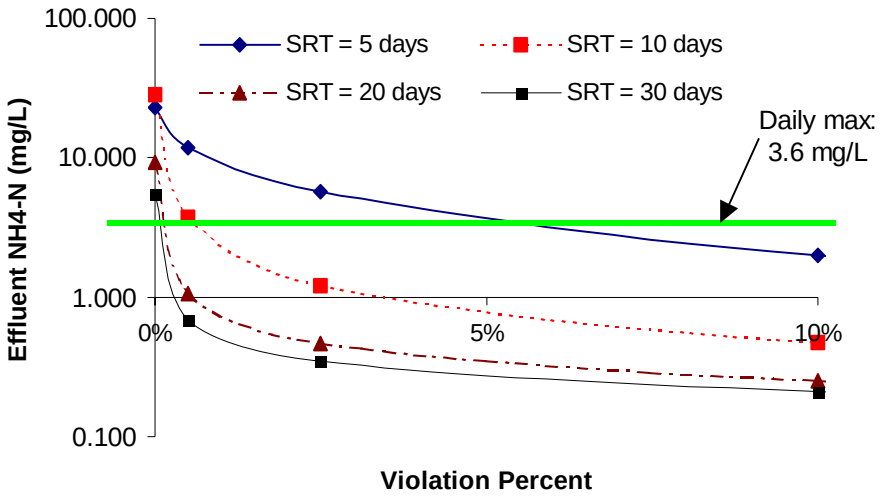


Figure 6-5. The decision-making chart of the Seneca WWTP.

As discussed above, we need get engineers, regulators, and communities together to have a roundtable discussion of the new regulations of plant performance. Based on the simulated results and corresponding percentile control and decision-making plots, we may create a more cost-effective design procedure to better protect our water environment.

The recommended the six steps of the StatASPS procedure can be summarized as follows:

Step 1: Replace missing influent data if needed;

Step 2: Search and replace outliers using the multivariate outlier detection method (Robinson et al., 2005);

Step 3: Calibrate the model parameters for the specific WWTP;

Step 4: Generate site-specific parameter distributions for the most sensitive parameters;

Step 5: Run Monte Carlo simulations with different operation conditions (for example, SRT);

Step 6: Make design charts (percentile control chart and decision-making chart).

To sum up, these following contributions were made in this dissertation:

- An innovative TES method was created to replace non-randomly missing values in time series data. The replaced TES data has similar statistical patterns of the real data.
- An innovative random influent generator was developed. Generated influent data have similar statistical characteristics as the historical influent data. The generator was based on a time series model fitted to the historical influent data.
- A Monte Carlo based simulation procedure called StatASPS was developed based on the ASM1 model. The procedure employs random influent generation, random model parameter generator, dynamic calibration procedure, and steady state and dynamic simulation of activated sludge processes.

Based on the Monte Carlo simulations of both Oak Ridge and Seneca WWTPs, we can make the following conclusions.

- The Monte Carlo based design procedure appears to more accurately capture the variability in activated sludge plant performance. With additional development, it may prove to be more robust design procedure than currently employed empirical safety factors.
- The best Monte Carlo simulations of both Oak Ridge and Seneca WWTPs use the one-year historical influent data and SPD#3-1P parameters. Parameter uncertainty was found to be more useful than influent variability in reproducing plant performance uncertainty. This also explains why we choose historical data instead of randomly generated influent data. In fact, the simulations conducted in Chapter V only used one-month's influent data, which has proved to be good enough. In this case, we strongly recommend conducting simulations representing both Summer and Winter conditions. As for the model parameters, the SPD#3-1P (i.e., all fixed parameters excepted for the site-specific parameter distribution $\mu_{A,max}$) proved to be the best parameter distributions, compared to other parameter distributions tested. The universal parameter distributions introduce too much uncertainty to wastewater treatment systems. Thus, if available, we only consider the most important parameters and keep other parameters constant with default values.
- For effluent ammonia nitrogen concentrations, sensitivity analysis determined the most important parameters to K_{NH} (half-saturation coefficient for nitrifiers), $\mu_{A,max}$

(specific growth rate of autotrophs), and Y_A (yield of autotrophs). However, the order of importance might change in practice. In the calibrations of both Oak Ridge and Seneca WWTP, we conclude that the most important parameter is $\mu_{A,max}$ and there is no significant relationship between the calibration standard RMSE and the parameter Y_A. K_{NH} has some interaction effects with the parameter $\mu_{A,max}$ on the calculated RMSE.

- The percentile control and decision-making charts demonstrate how changing SRT affects the plant performance. However, selection of acceptable risk is still a subjective decision. Thus, we strongly recommend that engineers, regulators and communities get together and discuss statistical- and risk- based effluent standards.

There are some recommendations for the future research work of the statistical-based design procedure using Monte Carlo simulations.

- Model parameters: the Monte Carlo simulations conducted in this dissertation demonstrated that uncertainty in model parameters is more important than variability in the plant influent, at least for effluent ammonia nitrogen concentration. Thus, additional work on defining model parameter distributions, especially correlations between parameters, may be considered the most important topic for future improvement of this design procedure.

- Design parameters: In this dissertation, we only considered the effect of changing SRT, which is the most important design and control parameter available to the engineer. In further research, other design parameters, for example, the oxygen concentration in CSTRs, bioreactor volume, number of CSTRs, etc. will be considered in StatASPS program.
- Process models: The models used in this dissertation are only for activated sludge processes. However, the Monte Carlo based design procedure can be applied to other water and wastewater treatment processes if appropriate process models are employed in the future. Furthermore, research on the clarifier models is also needed to better describe the performance of secondary clarifiers.

LIST OF REFERENCES

LIST OF REFERENCES

Abusam, A.; Keesman, K.J. and van Straten, G. (2001) Sensitivity Analysis in Oxidation Ditch Modeling: the Effect of Variations in Stoichiometric, Kinetic and Operating Parameters on the Performance Indices, Journal of Chemical Technology and Biotechnology, 76:430-438.

Abusam, A.; Keesman, K.J. and Spanjers, H. (2002) Estimation of Uncertainties in the Performance Indices of an Oxidation Ditch Benchmark, Journal of Chemical Technology and Biotechnology, 77 (9): 1058-1067.

American Public Health Association, American Water Works Association, and Water Environment Federation (1998) Standard Methods for the Examination of Water and Wastewater, 20th Edition, Washington, DC, American Public Health Association.

American Water Works Association (1999) Water Quality and Treatment: a Handbook of Community Water Supplies, 5th Edition, New York, McGraw-Hill.

Barker, P.S. and Dold, P.L. (1997) General Model for Biological Nutrient Removal Activated-Sludge Systems: Model Presentation, Water Environment Research, 69**,** 969-984.

Beccari M.; Pinto, A.C.; Ramadori, R. and Tomei, M.C. (1992) Effects of Dissolved Oxygen and Diffusion Resistances on Nitrification Kinetics, Water Research, 26(8), 1099-1104.

Benefield L.D. and Randall, C.W. (1980) Biological Process Design for Wastewater Treatment, Prentice Hall, Inc. Englewood Cliffs, NJ.

Berthouex, P.M. and Box, G.E. (1996) Time Series Models for Forecasting Wastewater Treatment Plant Performance, Water Research, 30 (8): 1865-1875.

Bisogni, J.J. and Lawrence, A.W. (1971) Relationships Between Biological Solids Retention Time And Settling Characteristics of Activated Sludge, Water Research, 5, 753-763.

Bixio D.; Parmentier G.; Rousseau D.; Verdonck F.; Meirlaen J.; Vanrolleghem P.A. and Thoeye C. (2002) A Quantitative Risk Analysis Tool for Design/Simulation of Wastewater Treatment plants, Water Science and Technology, 46(4-5), 301-307.

Brands E.; Liebeskind, M. and Dohmann, M. (1994) Parameters for Dynamic Simulation of Wastewater Treatment Plants with High-rate and Low-rate Activated Sludge Tanks, Water Science and Technology, 30(4), 211-214.

Bratley, P.; Fox, B.L. and Schrage, L.E. (1983) A Guide to Simulation. Springer-Verlag, New York, NY.

Cheremisinoff, N.D. (2002) Handbook of Water and Wastewater Treatment Technologies, Butterworth-Heinemann.

Cinar, O.; Daigger, G.T. and Graef, S.P. (1998) Evaluation of IAWQ Activated Sludge Model No.2 Using Steady-state Data from Four Full-scale Wastewater Treatment Plants, Water Environment Research, 70, 1216-1224.

Cokgör, E. U.; Sözen, S.; Orhon, D.; and Henze, M. (1998) Respirometric Analysis of Activated Sludge Behaviour- Assessment of the readily Biodegradable Substrate, Water Science and Technology, 32(2), 461-475.

Conover, W.J. (1999) Practical Nonparametric Statistics, 3rd edition, John Wiley and Sons, Inc., New York

Copp, J.D. and Murphy K.L. (1995) Estimation of the Active Nitrifying Biomass in Activated Sludge, Water Research, 29(8), 1855-1862.

Corbitt, R.A. (1999) Standard Handbook of Environmental Engineering, 2nd Edition, New York, McGraw-Hill.

Cox, C.D. (2004) Statistical Distributions of Uncertainty and Variability in Activated Sludge Model Parameters, Water Environment Research, 76 (7): 2672-2685.

Cox, C.D.; Robinson, R.B.; Huo, J.; McKinney, M.; Odom, K.; Chang, S. and Carrasco, I. (2004), WERF Report (Project 00-CTS-3): Tools for Rating the Capacity of Activated Sludge Plants, IWA publishing.

Daigger, G.T. and Nolasco, D. (1995) Evaluation and Design of Full-scale Wastewater Treatment Plants Using Biological Process Models, Water Science and Technology, 31(2), 245-255.

Daigger, G.T. and Parker, D.S. (2000) Enhancing Nitrification in North American Activated Sludge Plants, Water Science and Technology, 41(9), 97-105.

De la Sota, A.; Sherrard, J.H.; Novak J.T. and Randall, C.W. (1994) Performance and Model Calibration of R-D-N Processes in Pilot Plant, Water Science and Technology, 30(6), 355-364.

DeLurgio, S.A. (1998) Forecasting Principles and Applications, McGraw-Hill Book Co-Singapore.

Dincer, A.R. and Kargi, F. (2000) Kinetics of Sequential Nitrification and Denitrification Processes, Enzyme and Microbial Technology, 27, 37-42.

Dold, P.L. and Marais, G.v.R. (1986) Evaluation of the General Activated Sludge Model Proposed by the IAWPRC Task Group, Water Science and Technology, 18(6), 63-89.

Drtil, M.; Nemeth, P. and Bodfk, I. (1993) Kinetic Constants of Nitrification, Water Research, 27(1), 35-39.

Dupont, R. and Sinkjær, O. (1994) Optimization of Wastewater Treatment Plants by Means of Computer Model, Water Science and Technology, 30(4), 181-190.

Ekama, G.A.; Dold, P.L. and Marais, G.v.R. (1986) Procedures for Determining Influent COD Fractions and the Maximum Specific Growth Rate of Heterotrophs in Activated Sludge Systems, Water Science and Technology, 18, 91-114.

Ekama, G.A.; Barnard, J.L.; Gunthert, F.W.; Krebs, P.; McCorquodale, J.A.; Parker, D.S. and Wahlberg E.J. (1997) Secondary Settling Tanks: Theory, Modeling, Design, and Operation, IAWQ, London.

Ellis, G.W.; Grasso, D. and Ge, X. (1993) ARMA Processes and Reliability-Based Design of Wastewater-Treatment Facilities, Journal of Environmental Engineering, 119(3), 463-477.

Fillos, J.; Katehis. D.; Ramalingam, K.; Carrio, L.A. and Gopalakrishnan, K. (2000) Determination of Nitrifier Growth Rates in NYC Water Pollution Control Plants, WEFTEC 2000 Annual Conference Proceedings, CD-ROM, Water Environment Federation.

Garman, K.R.; Tetreault, M.J.; Dold, P.L.; Parker, D.S. and Finley, H.R. (1996) Evaluation of Critical Activated Sludge Parameters for Dynamic Process Models, Proceedings WEFTEC '96 69th Annual Conference & Exposition, Water Environment Federation, Vol 1. Part 1. 625-634.

Gee, S.S.; Pfeffer, J.T. and Suidan, M.T. (1990) Nitrosomonas and Nitrobacter Interactions in Biological Nitrification, Journal of Environmental Engineering, 116(1), 4-17.

Gentle, J.E. (2003) Random Number Generation and Monte Carlo Methods, Second edition, New York, Springer-Verlag New York Inc.

Gentry, R.W.; Camp, C.V.; and Anderson, J.L. (2001) Use of GA to Determine Areas of Accretion to Semiconfined Aquifer, Journal of Hydraulic Engineering, 127(9).

Goldberg, D.E. (1989) Genetic Algorithms in Search, Optimization, and Machine Learning, Addison-Wesley, Reading, Massachusetts.

Grady, C.P.L., Jr.; Daigger, G.T. and Lim, H.C. (1999) Biological Wastewater Treatment (2nd Edition), Marcel Dekker, Inc. New York, NY.

Grady, C.P.L., Jr.;Smets, B.F. and Barbeau, D.S. (1996) Variability in Kinetic Parameter Estimates: A Review of Possible Causes and a Proposed Terminology, Water Research, 30(3), 742-748.

Gujer, W.; Henze, M.; Mino, T. and Loosdrecht, M.v. (1999) Activated Sludge Model No.3, Water Science and Technology, 39(1), 183-193.

Hall E.R. and Murphy, K.L. (1979) Estimation of Nitrifying Biomass and Kinetics in Wastewater, Water Research, 14, 297-304.

Haugh, L.D. and Box, G.E.P. (1977) Identification of Dynamic Regression (Distributed Lag) Models Connecting Two Time Series, Journal of the American Statistical Association, 72(357), 121-130.

Henze, M. (1986) Nitrate Versus Oxygen Utilization Rates in Wastewater and Activated Sludge Systems, Water Science and Technology, 18, 115-122.

Henze, M. (1991) Capabilities of Biological Nitrogen Removal Processes from Wastewater, Water Science and Technology, 23, 669-679.

Henze, M.; Grady, C.P.L., Jr; Gujer, W.; Marais, G.v.R. and Matsuo, T. (1986) Activated Sludge Model No. 1. International Association on Water Pollution Research and Control, London, England.

Henze, M.; Gujer, W.; Mino, T.; Matsuo, T.; Wentzel, M.C. and Marais, G.v.R. (1995) Activated Sludge Model No. 2. International Association on Water Quality Science Technology Report No. 3, London, England.

Hipel, K.W. and McLeod A. (1994) Time Series Modelling of Water Resources and Environmental Systems, New York, Elsevier.

Holland, J.H. (1975) Adaptation in Natural and Artificial Systems, University of Michigan Press, Ann Arbor, Michigan.

Huang, J. and Hao, O.J. (1996) Alternating Aerobic-Anaerobic Process for Nitrogen Removal: Dynamic Modeling, Water Environment Research, 68, 94-104.

Huo, J. (2001) The Research and Application of Activated Sludge Process Model, Master Thesis, Tianjin University, Tianjin, China.

Huo, J.; Robinson, R.B. and Cox, C.D. (2004) Uncertainty and Sensitivity Analysis of Activated Sludge Model No.1 by Monte-Carlo Simulation for Single CSTR with Universal Distribution Parameters, EWRI '04 World Water & Environmental Resources Congress, Salt Lake City, Utah, June 27-July 1, 2004.

Huo, J.; Seaver, W.L.; Robinson, R.B. and Cox, C.D. (2005) Application of Time Series Models to Analyze and Forecast the Influent Components of Wastewater Treatment Plants (WWTPs), EWRI '05 World Water & Environmental Resources Congress, Anchorage, Alaska, May 15-19, 2005.

IWA (2000) Activated sludge models ASM1, ASM2, ASM2d and ASM3 (Scientific and Technical Reports, No. 9).

Johnson, D.E. (1998) Applied Multivariate Methods for Data Analysis, Durbury Press, Brooks/Cole Publishing Company.

Kabouris, J.C. and Georgakakos, A. P. (1996) Parameter and State Estimation of the Activated Sludge Progress-I. Model Development, Water Research, 30(12), 2853-2865.

Kanji, G.K. (1999) 101 Statistical Tests, SAGE Publications Ltd, Thousand Oaks, CA

Kappeler, J. and Gujer, W. (1992) Estimation of Kinetic Parameters of Heterotrophic Biomass Under Aerobic Conditions and Characterization of Wastewater for Activated Sludge Modelling, Water Science and Technology, 25(6), 125-139.

Kleijnen, J.P.C. and Helton, J.C. (1999) Statistical Analysis of Scatterplots to Identify Important Factors in Large-scale Simulations, 1: Review and Comparison of Techniques, Reliability Engineering and System Safety, 65, 147-185.

Koch, G.; Kuhni, M.; Gujer, W. and Siegrist, H. (2000) Calibration and Validation of Activated Sludge Model No.3 for Swiss Municipal Wastewater. Water Research, 34, 3580-3590.

Koch, G.; Kuhni, M. and Siegrist, H. (2001) Calibration and Validation of an ASM3-based Steady-state Model for Activated Sludge Systems - Part I: Prediction of Nitrogen Removal and Sludge Production, Water Research, 35(9), 2235-2245.

Kolisch, G. and Londong, J. (1998) An Online Simulation to Support the Operation of Municipal Sewage Plants, Water Science and Technology, 38(3), 71-78.

Kristensen, G.H.; Jørgensen, P.E. and Henze, M. (1992) Characterization of Functional Microorganism Groups and Substrate in Activated Sludge and Wastewater by AUR, NUR and OUR, Water Science and Technology, 25(6), 43-57.

Lawrence, A.W. and McCarty, P.L. (1970) Unified Basis for Biological Treatment Design and Operation. Sanitary Engineering Division, ASCE, 96, 757-778.

Larrea, L.; Garcia-Heras, J.L.; Ayesa, E. and Florez J. (1992) Designing Experiments to Determine the Coefficients of Activated Sludge Models by Identification Algorithms, Water Science and Technology, 25(6), 149-165.

Lesouef, A.; Payraudeau, M.; Rogalla, F. and Kleiber, B. (1992) Optimizing Nitrogen Removal reactor Configurations by Onsite Calibration of the IAWPRC Activated Sludge Model, Water Science and Technology, 25(6), 105-123.

Liebeskind, M.; Schäpers, C.; Bornemann, C.; Brands, E.; Freund, M. and Rolfs, T. (1996) Parameter Determination and Model Fitting- Two Approaches for Modelling Processes in Wastewater Treatment Plants, Water Science and Technology, 34(5-6), 27-33.

Magbanua, B.S. (2004) Probability Modeling of Activated Sludge Systems: Strategies for Reducing Process Uncertainty, EWRI '04 World Water & Environmental Resources Congress, Salt Lake City, Utah, June 27-July 1, 2004.

Makinia, J. and Wells, S. (2000) A General Model of the Activated Sludge Reactor with Dispersive Flow-I. Model Development and Parameter Estimation, Water Research, 34(16), 3987-3996.

Martinez, W.L. (2002) Computational statistics handbook with MATLAB, Chapman & Hall/CRC.

Massone, A.; Gernaey, K.; Rozzi, A. and Verstraete, W. (1998) Measurement of Ammonium Concentration and Nitrification Rate by a New Titrimetric Biosensor, Water Environment Research, 70, 343-350.

Mathieu, S. and Etienne, P. (2000) Estimation of Wastewater Biodegradable COD Fractions by Combining Respirometric Experiments in Various SO/XO Ratios, Water Research, 34(4), 1233-1246.

Meeker, W.Q. and Escobar, L.A. (1998), Statistical Methods for Reliability Data, New York John Wiley & Sons, Inc. (US).

Melcer, H. (1999) Full Scale Experience with Biological Process Models- Calibration Issues, Water Science Technology, 39(1), 245-252.

Metcalf and Eddy, Inc. (2003) Wastewater Engineering: Treatment, Disposal, and Reuse (4th edition), (Revised by Tchobanoglous, G.; and Burton F.L.), McGraw Hill, New York, NY.

Mussati, M.; Gernaey, K.; Gani, R. and Jorgensen, S.B. (2002) Computer Aided Model Analysis and Dynamic Simulation of a Wastewater Treatment Plant, Clean Technology Environment Policy, 4, 100-114.

Naidoo, V.; Urbain, V. and Buckley, C.A. (1998) Characterization of Wastewater and Activated Sludge from European Municipal Wastewater Treatment Plants Using the NUR Test, Water Science and Technology, 38(1), 303-310.

Nolasco, D.; Daigger, G.; Stafford, D.; Stephenson, J. and Kaupp, D. (1998) The Use of Mathematical Modeling and Pilot Plant testing to Develop a New Biological Phosphorus and Nitrogen Removal Process, Water Environment Research, 70(6), 1205-1215.

Nowak, O.; Schweighofer, P. and Scardal K. (1994) Nitrification Inhibition – a Method for the Estimation Actual Maximum Autotrophic Growth Rates in Activated Sludge Systems, Water Science and Technology, 30(6), 9-19.

Nowak, O.; Svardal, K.; Franz, A. and Kuhn, V. (1999) Degradation of Particulate Organic Matter- a Comparison of Different Model Concepts, Water Science and Technology, 39(1), 119-127.

Qasim, S.R. (1999) Wastewater Treatment Plants: Planning, Design, and Operation, Technomic Pub. Co., Lancaster, Pa.

Oh, J. and Silverstein, J. (1999) Oxygen Inhibition of Activated Sludge Denitrification, Water Research, 33(8), 1925-1937.

Oles, J. and Wilderer, P.A. (1991) Computer Aided Design of Sequencing Batch Reactors Based on the IAWPRC Activated Sludge Model, Water Science and Technology, 23, 1087-1095.

Olsson, G. (1999) Wastewater Treatment Systems: Modelling, Diagnosis and Control, IWA Publishing.

Orhon, D.; Cokgor, E.U. and Sozen S. (1999) Experimental Basis for the Hydrolysis of Slowly Biodegradable Substrate in Different Wastewaters, Water Science Technology, 39(1), 87-95.

Orhon, D.; Karahan, O. and Sozen, S. (1999) The Effect of Residual Microbial Products on the Experimental Assessment of the Particulate Inert COD in Wastewaters, Water Research, 33(14), 3191-3203.

Ossenburggen, P.J.; Spanjers, H. and Klapwik, A. (1996) Assessment of a Two-step Nitrification Model for Activated Sludge, Water Research, 30(4), 939-953.

Pankratz, A. (1991) Forecasting with Dynamic Regression Models, John Wiley & Sons Inc.

Paladino, O. and Tatto, M. (2000) Robust Stability and Sensitivity of Real Controlled CSTRs, Chemical Engineering Science, 55, 321-330.

Pigram, G.M. and Macdonald, T.R. (2001) Use of neural network models to predict industrial bioreactor effluent quality, Environmental Science Technology, 35,157-162.

Pollard, P.C.; Steffens, M.A.; Biggs, C.A. and Lant, P.A. (1998) Bacterial Growth Dynamics in Activated Sludge Batch Assays, Water Research, 32, 587-596.

Raltelli, A.; Chan, K. and Scott, E.M. (2000) Sensitivity Analysis, John Wiley & Sons Ltd., Wwat Sussex, England.

Rittmann, B.E.; Laspidou, C.S. and Flax, J. (1999) Molecular and Modeling analyses of the Structure and Function of Nitrifying Activated Sludge, Water Science and Technology, 39(1), 51-59.

Rittman, B.E. and McCarty P.L (2001) Environmental Biotechnology: Principles and Applications. McGraw-Hill, New York.

Robert, C.P. and Casella, G. (2004) Monte Carlo Statistical Methods, New York, Springer, U.S.A.

Robinson, R.B.; Cox, C.D. and Odom, K. (2005) Identifying Outliers in Correlated Water Quality Data, Journal of the Environmental Engineering Division, ASCE, Accepted for publication.

Rousseau, D.; Verdonck, F.; Moerman, O.; Carrette, R.; Thoeye, C.; Meirlaen, J. and Vanrolleghem, P.A. (2001) Development of a Risk Assessment Based Technique for Design/retrofitting of WWTPs, Water Science and Technology, 43(7), 287-294.

Rozich, A.F. and Castens, D.J. (1986) Inhibition Kinetics of Nitrification in Continuous-flow Reactors, Journal Water Pollution Control Federation, 58, 220-226.

Schuhmacher, M.; Meneses, M.; Xifro, A. and Domingo, J.L. (2001) The Use of Monte-Carlo Simulation Techniques for Risk Assessment: Study of a Municipal Waste Incinerator, Chemosphere, 43, 787-799.

Schutze, M.R.; Butler, D. and Beck, M.B. (2002) Modeling, Simulation and Control of Urban Wastewater Systems, Springer, 2002.

Siegrist, H.; Krebs, P.; Buhler, R.; Purtschert, I. and Rock, C. (1995) Denitrification in Secondary Clarifiers, Water Science and Technology, 31(2), 205-214.

Siegrist H. and Tschui M. (1992) Interpretation of Experimental Data with Regard to the Activated Sludge Model No.1 and Municipal Wastewater Treatment Plants, Water Science and Technology, 25(6), 167-183.

Siler, B.; Spotts, J. and Mckelvy, M. (1998) Using Visual Basic 6, Indianapolis, IN, USA.

Spanjers, H. and Vanrolleghem, P. (1995) Respirometry as a Tool for Rapid Characterization of Wastewater and Activated Sludge, Water Science and Technology, 31(2), 105-114.

Spéranddio, M.; Urbain, V.J.; Audic, J.M. and Paul E. (1999) Use of Carbon Dioxide Evolution Rate for Determining Heterotrophic Yield and Characterizing Denitrifying Biomass, Water Science and Technology, 39(1), 139-146.

Sprott, J.C. (1998) Numerical Recipes: Routines and Examples in BASIC, Cambridge University, New York.

Sollfrank, U. and Gujer W. (1991) Characterization of Domestic Wastewater for Mathematical Modelling of the Activated Sludge Process, Water Science and Technology, 23(4-6), 1057-1066.

Sorour M.T.; Olsson, G.; and Somlyody, L. (1993) Potential Use of Step Feed Control Using the Biomass in the Settler, Water Science and Technology, 28(11-12), 239-248.

Sözen, S.; Cokgör, E. U.; Orhon, D. and Henze, M. (1998) Respirometric Analysis of Activated Sludge Behaviour- Heterotrophic Growth Under Aerobic and Anoxic Conditions, Water Research, 32(2), 476-488.

Sözen, S. and Orhon, D. (1999) The Effect of Nitrite Correction on the Evaluation of the rate of Nitrate Utilization under Anoxic Conditions, Journal of Chemical Technology and Biotechnology, 74:790-800.

Sözen, S.; Orhon, D. and San, H.A. (1996) A New Approach for the Evaluation of the Maximum Specific Growth Rate in Nitrification, Water Research, 30(7), 1661-1669.

Sozen, S.; Ubay Cokgor, E.; Orhon, D. and Henze H. (1998) Respirometric Analysis of Activated Sludge Behaviour—II. Heterotrophic Growth Under Aerobic and Anoxic Conditions, Water Research, 32, 476-488.

Stokes, L.; Takacs, I.; Watson, B. and Watts, J.B. (1993) Dynamic Modelling of an ASP Sewage Works—a Case Study, Water Science and Technology, 28(11-12), 151-161.

Strotmann, U.J.; Geldern, A.; Kuhn, A.; Gendig, C. and Klein, S. (1999) Evaluation of a Respirometric Test Method to Determine the Heterotrophic yield Coefficient of Activated Sludge Bacteria, Chemosphere, 38(15), 3555-3570.

Tchobanoglous, G.; Burton, F.L and Stensel, B.H (2003) Wastewater Engineering: Treatment and Reuse/Metcalf & Eddy, Inc., 4th Edition, McGraw Hill.

Tillman, G.M. (1996) Wastewater Treatment: Troubleshooting and Problem Solving, Ann Arbor Press, Inc.

Urbain, V.; Mobarry, B.; Silva, V.D.; Stahl, D.A.; Rittmann, B.E. and Manem J. (1998) Integration of Performance Molecular Biology and Modeling to Describe the Activated Sludge Process, Water Science and Technology, 37(3), 223-229.

Vasquez, V.R. and Whiting,W.B. (1999) Effect of Systematic and Random Errors in Thermodynamic Models on Chemical Process Design and Simulation: A Monte Carlo Approach, Industrial and Engineering Chemistry Research, 38, 3036-3045.

Von Sperling, M. (1993) Parameter Estimation and Sensitivity Analysis of an Activated Sludge Model Using Monte Carlo Simulation and the Analysis's Involvement, Water Science and Technology, 28(11-12), 219-229.

Water Environment Federation and American Society of Civil Engineers (1998) Design of Municipal Wastewater Treatment Plants (4th edition), Water Environment Federation (MOP 8), Alexandria VA and American Society of Civil Engineers (MREP 76), Reston, VA.

West, D. and Dellana S. (2002) Transfer function modeling of processes with dynamic inputs, Journal of Quality Technology, 34 (3): 315-326.

Wanner, O.; Kappeler, J. and Gujer, W. (1992) Calibration of an Activated Sludge Model Based on Human Expertise and On a Mathematical Optimization Technique- a Comparison, Water Science and Technology, 25(6), 141-148.

Xu, S. and Hultman, B. (1996) Experience in Wastewater Characterization and Model Calibration for the Activated Sludge Process, Water Science and Technology, 33(12), 89-98.

Yantarasri, T.; Garcia, A. and Brune, D. (1992) Thermodynamic Model of Nitrification Kinetics, Journal of Environmental Engineering, 118(4), 568-584.

APPENDIX

APPENDIX-A

Equations of time series models

Flow rate model	*Log ARIMA (2,0,0) $(1,0,0)_7$

$$\ln Q_t = \mu + \frac{1}{(1-\phi_1 \mathrm{B}^1-\phi_2 \mathrm{B}^2)(1-\phi_7 \mathrm{B}^7)}\alpha_t$$

$$\Leftrightarrow \quad \ln Q_t - \mu = \frac{1}{1-(\phi_1 \mathrm{B}^1+\phi_2 \mathrm{B}^2+\phi_7 \mathrm{B}^7-\phi_1\phi_7 \mathrm{B}^8-\phi_2\phi_7 \mathrm{B}^9)}\alpha_t$$

$$\Leftrightarrow \quad (\ln Q_t - \mu)[1-(\phi_1 \mathrm{B}^1+\phi_2 \mathrm{B}^2+\phi_7 \mathrm{B}^7-\phi_1\phi_7 \mathrm{B}^8-\phi_2\phi_7 \mathrm{B}^9)] = \alpha_t$$

$$\Leftrightarrow \quad \ln Q_t - \mu = (\phi_1 \mathrm{B}^1+\phi_2 \mathrm{B}^2+\phi_7 \mathrm{B}^7-\phi_1\phi_7 \mathrm{B}^8-\phi_2\phi_7 \mathrm{B}^9)(\ln Q_t - \mu) + \alpha_t$$

Flow rate model	*Log [Temperature + ARIMA (2,0,0) $(1,0,0)_7$]

$$\ln Q_t = \mu + \beta \times Temp + \frac{1}{(1-\phi_1 \mathrm{B}^1)(1-\phi_7 \mathrm{B}^7)}\alpha_t$$

$$\Leftrightarrow \quad \ln Q_t - \mu - \beta \times Temp = \frac{1}{1-(\phi_1 \mathrm{B}^1+\phi_7 \mathrm{B}^7-\phi_1\phi_7 \mathrm{B}^8)}\alpha_t$$

$$\Leftrightarrow \quad (\ln Q_t - \mu - \beta \times Temp)[1-(\phi_1 \mathrm{B}^1+\phi_7 \mathrm{B}^7-\phi_1\phi_7 \mathrm{B}^8)] = \alpha_t$$

$$\Leftrightarrow \quad \ln Q_t - \mu = (\phi_1 \mathrm{B}^1+\phi_7 \mathrm{B}^7-\phi_1\phi_7 \mathrm{B}^8)(\ln Q_t - \mu) +$$

$$(1-\phi_1 \mathrm{B}^1-\phi_7 \mathrm{B}^7+\phi_1\phi_7 \mathrm{B}^8)\beta Temp + \alpha_t$$

SS	*Simple Exponential Smoothing (M1) **Linear Trend + AR(1) +Point:31JAN2002 (M2)** #Combination: [0.12 M1 +0.88 M2]

$$SS_t = \mu + \beta SS_{t-1} + \frac{1}{1-\phi_1 B^1}a_t$$

$$\Leftrightarrow \quad SS_t - \mu = \beta SS_{t-1} + \frac{1}{1-\phi_1 B^1}a_t$$

$$\Leftrightarrow \quad (SS_t - \mu)(1-\phi_1 B^1) = (1-\phi_1 B^1)\beta SS_{t-1} + a_t$$

$$\Leftrightarrow \quad SS_t - \mu = \phi_1 B^1 (SS_t - \mu) + (1 - \phi_1 B^1)\beta SS_{t-1} + a_t$$

$$\Leftrightarrow \quad SS_t - \mu = \phi_1 B^1 (SS_t - \mu) + (1 - \phi_1 B^1)\beta SS_{t-1} + a_t + \gamma \times dummy$$

where, $\gamma \times dummy = \begin{cases} 0, & Not \quad an \quad \text{int}ervention. \\ \gamma, & is \quad an \quad \text{int}ervention. \end{cases}$

BOD_5	#Flow rate[N(1)/D(1)]+AR(1)

$$BOD_t = \mu + \frac{\omega_0 - \omega_1 \mathrm{B}^1}{1 - \delta_1 \mathrm{B}^1} Q_t + \frac{1}{\left(1 - \phi_1 \mathrm{B}^1\right)} \alpha_t$$

$$\Leftrightarrow \quad (BOD_t - \mu)(1 - \delta_1 \mathrm{B}^1)(1 - \phi_1 \mathrm{B}^1) = (\omega_0 - \omega_1 \mathrm{B}^1)(1 - \phi_1 \mathrm{B}^1) Q_t + \left(1 - \delta_1 \mathrm{B}^1\right)\alpha_t$$

$$\Leftrightarrow \quad (BOD_t - \mu)[1 - (\phi_1 + \delta_1)\mathrm{B}^1 - \phi_1 \delta_1 \mathrm{B}^2] = [\omega_0 - (\omega_1 + \phi_1 \omega_0)\mathrm{B}^1 + \phi_1 \omega_1 \mathrm{B}^2] Q_t + \left(1 - \delta_1 \mathrm{B}^1\right)\alpha_t$$

$$\Leftrightarrow$$

$$BOD_t - \mu = [(\phi_1 + \delta_1)\mathrm{B}^1 + \phi_1 \delta_1 \mathrm{B}^2](BOD_t - \mu) + [\omega_0 - (\omega_1 + \phi_1 \omega_0)\mathrm{B}^1 + \phi_1 \omega_1 \mathrm{B}^2] Q_t + \left(1 - \delta_1 \mathrm{B}^1\right)\alpha_t$$

NH_4	*Simple Exponential Smoothing **#Flow rate[/D(1)]+AR(1)**

$$NH_t = \mu + \frac{\omega_0}{1 - \delta_1 \mathrm{B}^1} Q_t + \frac{1}{\left(1 - \phi_1 \mathrm{B}^1\right)} \alpha_t$$

$$\Leftrightarrow \quad (NH_t - \mu)(1 - \delta_1 \mathrm{B}^1)(1 - \phi_1 \mathrm{B}^1) = \omega_0 (1 - \phi_1 \mathrm{B}^1) Q_t + \left(1 - \delta_1 \mathrm{B}^1\right)\alpha_t$$

$$\Leftrightarrow \quad (NH_t - \mu)[1 - (\phi_1 + \delta_1)\mathrm{B}^1 + \phi_1 \delta_1 \mathrm{B}^2] = (\omega_0 - \omega_0 \phi_1 \mathrm{B}^1) Q_t + (1 - \delta_1 \mathrm{B}^1)\alpha_t$$

$$\Leftrightarrow \quad NH_t - \mu = [(\phi_1 + \delta_1)\mathrm{B}^1 - \phi_1 \delta_1 \mathrm{B}^2](NH_t - \mu) + (\omega_0 - \phi_1 \omega_0 \mathrm{B}^1) Q_t + (1 - \delta_1 \mathrm{B}^1)\alpha_t$$

APPENDIX-B

Statistical characteristics of parameter $\mu_{A,max}$: lognormal(-3.16214, 0.4203)

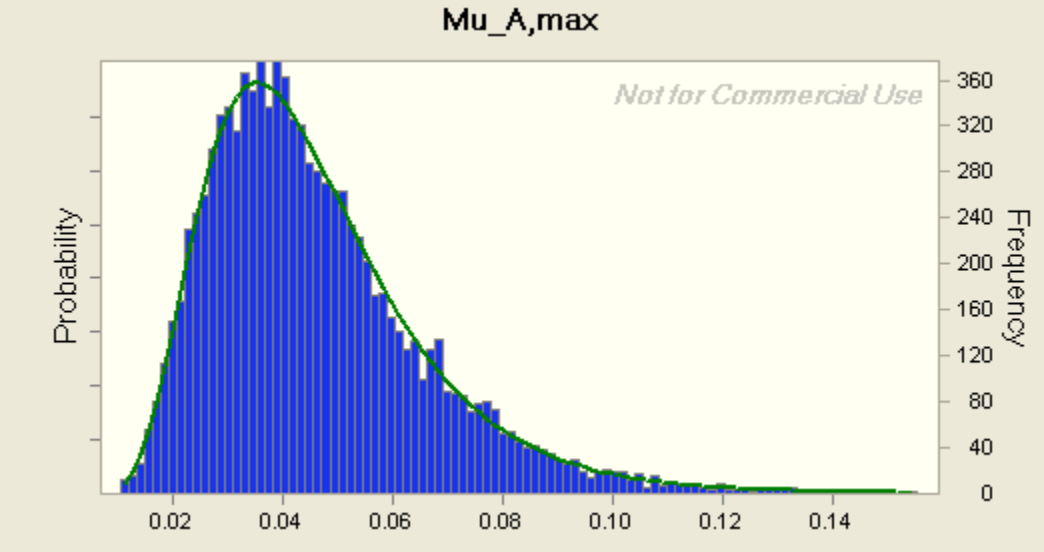


Figure A-1. Lognormal(-3.16214, 0.4203) with 10000 random numbers.

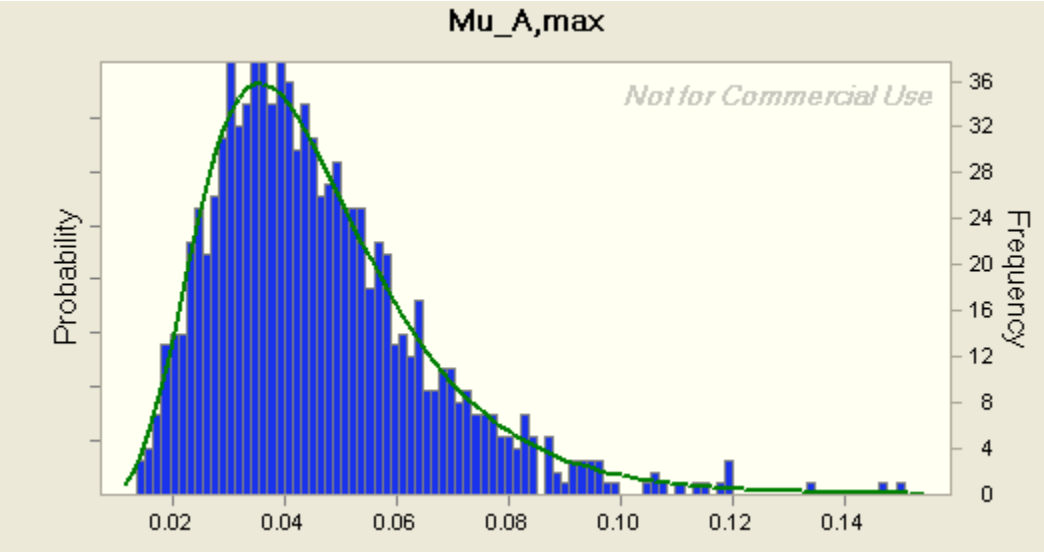


Figure A-2. Lognormal(-3.16214, 0.4203) with 1000 random numbers.

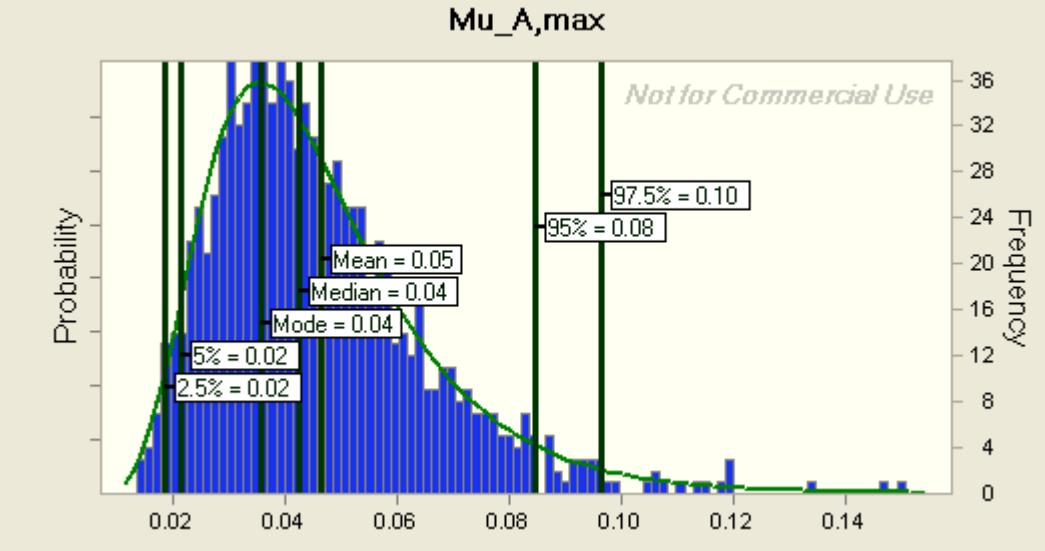


Figure A-3. The statistical characteristics of a Lognormal(-3.16214, 0.4203) with 1000 random numbers.

VITA

Jinsheng Huo was born in Xinxiang, Henan, P.R. China, on June 25, 1976 and is the son of Mr. Gangfu Huo and Mrs. Xuexiu Zhao. On December 19, 2002, he married his lovely wife Yan Jiang (Jenny).

In 1998, he received his Bachelor's degree in Environmental Engineering from Zhengzhou University of Technology (now known as Zhengzhou University). He won the Outstanding Undergraduate award honored by Henan Province Department of Education in 1998 and an Honorable Mention award in the American Mathematical Contest in Modeling in 1997.

Then he went to Tianjin University and received his Master degree in Environmental Engineering in March 2001. At Tianjin University, he designed, constructed, and operated one anaerobic EGSB (Expanded Granular Sludge Bed) reactor to treat municipal wastewater from Tianjin Ji-Zhuang-Zi Wastewater Treatment Plant. He also voluntarily worked on the simulation of hydraulic characteristics in a MBBR (Moving Bed Biofilm Reactor) reactor. His Master thesis is "Research and Application of Activated Sludge Process Model".

In May 2001, he enrolled in the doctoral program at the University of Tennessee, Knoxville under the direction of Dr. Chris D. Cox. During his study at the University of

Tennessee, he also received another Master degree in Statistics in December 2003 and became a licensed Professional Engineer in Environmental Engineering in December 2004. He also won the WMREI Fellowship for maximum three years in a row at the University of Tennessee. He has been a member of both American Society of Civil Engineers (ASCE) and Water Environmental Federation (WEF) since 2001.