DISSERTATION

PROBABILISTIC FORECAST MODELS FOR HYDRO-ENVIRONMENTAL CHARACTERISTICS AND RISK-BASED ADAPTIVE RESERVOIR OPERATION

Submitted by

Han-Goo Lee

Department of Civil and Environmental Engineering

In partial fulfillment of the requirements

for the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Spring 2007

COLORADO STATE UNIVERSITY

10 November 2006

WE HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER OUR SUPERVISION BY HAN-GOO LEE ENTITLED "PROBABILISTIC FORECAST MODELS FOR HYDRO-ENVIRONMENTAL CHARACTERISTICS AND RISK-BASED ADAPTIVE RESERVOIR OPERATION" BE ACCEPTED AS FULFILLING IN PART REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY.

Committee on Graduate Work

Brian P. Bledsoe

Jennifer A. Hoeting

Neil S. Grigg

Advisor Darrell G. Fontane

Department Head Luis A. Garcia

ABSTRACT OF DISSERTATION

PROBABILISTIC FORECAST MODELS FOR HYDRO-ENVIRONMENTAL CHARACTERISTICS AND RISK-BASED ADAPTIVE RESERVOIR OPERATION

This study was motivated by the desire to improve risk-based decision making and adaptive management of large-scale water resources systems centering on multi-reservoir system operation. Forecasting the dynamic behavior of a water resources system is inherently uncertain. Case studies were performed using data from the Geum River basin in Korea. The overall objective of the research was to develop a methodology for managing a water resources system in an adaptive manner accounting for risks and uncertainties of the hydro-environmental characteristics. The characteristics considered in this research are the stage-discharge relationship, reservoir inflow, and water qualities in terms of biological oxygen demand (BOD) and total phosphorus (TP).

First, stage-discharge ratings were developed and assessed using both deterministic and probabilistic methods at two stage-discharge measurement stations chosen because they exhibited hysteresis. For deterministic approaches, nonlinear programming (NLP), fuzzy rule-based modeling, and a one-dimensional hydrodynamic model were used. For the probabilistic approach, a Bayesian Markov chain Monte Carlo (MCMC) technique was employed. Based upon a comparison of the methods, a hybrid methodology which combines NLP and Bayesian MCMC was proposed as the best alternative.

Second, stochastic monthly inflow forecast systems were developed using stochastic artificial neural networks and nonparametric modeling. To determine whether or not a *k*-nearest neighbor (*k*-NN) bootstrap resampling method might be used in practice for daily inflow forecasts aimed at short term reservoir system operation, a daily forecast model was developed. In the context of practical applicability, it was concluded that the *k*-NN method was preferred due to its ease of application. In addition, it was demonstrated that this method can be applied successfully for daily inflow forecasting.

Third, probabilistic BOD and TP models were developed using Bayesian networks. The relationships between reservoir release and risk of violating the water quality standards were derived. The case study clearly demonstrated that the probabilistic models

overcome the weaknesses of deterministic water quality models by offering information about risks of violation of standards or failures to meet targets. Compared to the other methods for uncertainty analysis such as sensitivity analysis, first order second moment (FOSM) analysis, and Monte Carlo methods, the advantages of a Bayesian MCMC technique were identified.

Fourth, instead of relying on the classical rule curves for reservoir system operation, an adaptive sampling implicit optimization (ASISO) model was developed that considered multiple objectives of energy production, water supply, and water quality management in terms of BOD and TP. A decision support system (DSS) especially designed for interactively integrating the probabilistic inflow, BOD and TP forecast models to the ASISO model was developed. The ASISO based DSS demonstrated an alternative for reservoir system operation by combining simulation and optimization algorithms and incorporating the risk of water quality standard violation and adaptive sampling of the inflow series. The case study also showed that the reservoir inflow forecast systems played a very important role in terms of differences between the models considered.

This research contributed to implementation of adaptive reservoir system operation with consideration of risk and uncertainty by joining probabilistic forecast models for hydro-environmental characteristics to reservoir system operation, which has been considered a very daunting task. Probabilistic forecast models were proposed by comparing several popular methodologies. This research showed the possibility of application of stochastic artificial neural networks (ANNs) in the field of water resources.

It is recommended for further study that the developed reservoir operation system be reduced to a weekly or daily time step. For sophisticated short term inflow forecast models in addition to the *k*-NN method, Markovian autoregressive models should be investigated by incorporating a variety of exogenous variables such as temperature and humidity.

Han-Goo Lee
Department of Civil and Environmental Engineering
Colorado State University
Fort Collins, CO. 80523
Spring 2007

ACKNOWLEDGEMENTS

To my advisor, Professor Darrel G. Fontane, the root of my knowledge.

To my committee members, Professor Neil S. Grigg, Professor Brian P. Bledsoe, and Professor Jennifer A. Hoeting, their great assistance and instruction.

To *K*water (President Mr. Kwak, Kyol-Ho, former president Mr. Ko, Seokgu, Mr. Yoon, Bohun), my constant supporter.

To Professor John W. Labadie, Professor Jose D. Salas, and Professor Jorge A. Ramirez, their great teaching and advice.

To my parents, my roots.

To my wife, Jaymi, my peace.

To Changhee and Yeron, my future.

To Dick and Philias, and John and Lyla, a mirror of my future.

To Mr. Won, Inhee and Mr. Kim, Seokhyon, encouraging and guiding me in right direction in my study.

To Mr. Kim, Bongjae, Mr. Cha, Giuk, Mr. Kim, Hyonsik, Mr. Kim, Gijung, Mr. Jo, Yongduck, Mr. Shin, Yongho, Mr. Jung, Yongbae, Mr. Jo, Yongdae, Mr. Jo, Seongho, Mr. Lee, Dongjin, Mr. Song, Taejeong, Mr. Lim, Gwangsup, Mr. Lee, Taesam, and Ms. Ji, Un, Mr. Park, Daeryong, the great trouble-solving brokers when I had difficulties in my study.

To Dr. Noh, Jaekyong, Dr. Lee, Sangho, Mr. Seong, Yongdu, Mr. Oh, Byongdong, Mr. Park, Jaeyong, Ms. Jeon, Seonmee, their great advice and support.

To Dr. Park, Seongje, Dr. Lee, Jaeyeung, Dr. Park, Sanggil, Dr. Jang, Seokhwan, Dr. Shin, Hyonseok, their great help and advice.

To Mr. Jang, Phil, Mr. Ji, Sunjin, Mr. Choi, Gapsu, Dr. Kim, Junkyom, the unforgettable memories in Fort Collins.

To Rocky mountain, my root of inspiration.

To Fort Collins and Collindale (Jim and Pat), fresh incitement to study and pleasure.

To Brookview (#17) and Aggi Village (27-L), my place to rest when I was exhausted.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION AND MOTIVATION

The principles for embodying a comprehensive framework for water industry management can be classified into several groups: "*comprehensive approaches; river basin focus; stakeholder involvement and coordination mechanisms; voluntary, cooperative, regional action; public involvement; resolution of conflicts and disputes; local responsibility and accountability; organizational management and role settings; conservation approaches and environmental ethics; training, education, and capacity building; market focus, pricing, and incentives; risk management; adaptive management; decision support; finance; and regulation.*" *[Grigg, 1996]*. The suggested framework may be divided into political and technical groups. Typical components in the technical group include risk and adaptive management which are the focus of this research.

1.1 Risk-based Decision Making and Adaptive Management

1.1.1 Risk-Based Decision Making

The question about whether the future can be deterministically predicted, or whether it is arbitrary and random, had been a topic of discussion for a long time especially in physics. The French scientist, Laplace, publicly expressed the idea of

scientific determinism first. It was not until 1926, that German physicist, Werner Heisenberg formulated a new theory called quantum mechanics addressing uncertainty in the position and speed of a particle. He believed the universe evolved in an arbitrary way. However, Albert Einstein suspected that these events all had specific causes, but that the causes were somehow hidden away in some hard-to-find form. That's what he meant when he said "*God doesn't play dice with the universe.*" *[Stephen Hawking, public lecture: http://www.hawking.org.uk /lectures]*.

No matter what the reasons are, the uncertainties of scientific prediction for the future are inevitable in the real world, and the resources for making decisions are limited. These facts make it clear that failing to analyze and quantify the uncertainties can lead to wrong decisions. Uncertainty analysis is concerned with quantifying the expected variability of a dependent variable calculated as a function of one or more independent variables. Risk analysis is a procedure for determining the risks of standard violation or failures to meet targets due to the various sources of uncertainty *[Chow et al., 1988]*.

To support better decision-making for environmentally sound and sustainable development (ESSD) and management of water resources, many researchers have tried to build predictive deterministic models based on mathematical representations of systems *[Stow et al., 2003]*. Regardless of the sophistication and complexity of the physical, biological and chemical processes in deterministic models, uncertainties always exist due to two main factors: system error including insufficient and improper mathematical representation of natural phenomena, uncertainties of model parameters, misspecification of boundary conditions and natural variation that is Heisenberg's view, and; measurement error. In general, modelers have applied optimization techniques, maximum likelihood

estimates (MLE) and some heuristic devices like genetic *[Solomatine, 1995]* and evolutionary algorithms *[Duan et al., 1992]* to simulation models in order to minimize errors in the calibration process. Many engineers who favor Heisenberg's view may think that more detailed mathematical representation of nature helps decrease uncertainties. Although these endeavors might improve fitting of observed data, they may produce over-parameterized models that are difficult to use.

From the viewpoints mentioned above, the biggest disadvantage of deterministic models may be that it is very difficult for these models to produce information about risks. The most convincing measure for overcoming the drawbacks of the deterministic approach is to develop predictive, stochastic models, which are able to address uncertainties via probability distributions. As a leading surrogate for stochastic modeling in recent years, the Bayesian data modeling technique has received great attention especially in the ecological domain with many parameters which are difficult to identify *[Reckhow, 1999]*. However, there is an ongoing debate as to the relative merits of simple models that statistically fit the data versus more complex process models that attempt to simulate chemical, physical and biological processes. Bayesian modeling generally employs more simplified functional relationships than process models.

As for other stochastic modeling approaches, the following methods have had great popularity among hydrologists and ecologists: nonparametric models and seasonal or stationary ARMA models with Bayesian learning; state-space modeling with Kalman filtering, and; stochastic artificial neural network and fuzzy rule-based modeling for controlling the facilities for water resources management such as reservoirs.

1.1.2 Adaptive Management in Decision Making

In the water industry, it is common for multiple stakeholders including regulators on behalf of government, interest groups, environmentalists, and support groups to be involved in the decision process. This often creates complicated problems with objectives such that they can not be solved in a strictly technical way. The final resolutions often lie in the legal, financial, and political arena *[Grigg, 1996]*. This fact indicates what it takes to reach a decision may vary unexpectedly. It also indicates that the planning and management process is so dynamic as to lead to periodic reevaluation of goals, needs, and actions, which is defined as adaptive management.

Adaptive management is a sort of "learning by doing" strategy *[Stow et al., 2003]* in that decisions are made successively by learning from newly gathered observations with *a priori* information about model parameters and explanatory variables, and forecasting the future using a Bayesian approach. Water resources systems continuously change on a basin-wide scale and new information is often gathered in the form of time series. Furthermore, unexpected political events arise frequently and the information for decision-making is generally random. When these factors are considered in a system, it makes the system analysis process complicated. Consequently, decision makers want to be kept up-to-date promptly on new situations, and newly gathered information should be continuously assimilated into the decision-making process. More generally speaking, the prior information that is assumed or regarded as the best information at the current time is upgraded with newly gathered information in the next time step, which is called posterior information. The posterior information serves as the prior information in the next time step. In a recent report for developing total maximum daily load focusing on dealing with

non-point pollution sources *[National Research Council, 2001]*, the National Research Council Committee To Assess the Scientific Basis of the Total Maximum Daily Load Approach to Water Pollution Reduction put an emphasis on adaptive implementation to improve TMDL estimates over time.

1.2 Problem Identification

When hydrologic engineers engaged in water and environmental planning and management are asked what variables have the most influence on their decisions, they are likely to point out the measured, predicted or projected data of precipitation, stream flows, reservoir inflows, water qualities, water demand, and so on. Furthermore, the measurement errors and parameter estimation of the models for predicting the above-mentioned variables make the decision processes complicated. As a reason why these variables are noted, it is because they contain significant uncertainties and, these uncertainties are propagated through the decision-making process.

Figure 1-1 illustrates the big picture with the main factors involved in river basin planning and management and their relationships. The determination of each factor can be regarded as a sub-decision making process to reach the final goal, that is, river basin planning and management. Among the variables in Figure 1-1, the water demand projection might be pointed out as the most uncertain variable because it is so susceptible to the political and social issues. As for the stage, it may not have a crucial influence in that the measurement errors are not considerably large as compared to the discharge and water qualities, and the time-series of stage is often obtained by direct measurement

rather than model simulation. Furthermore, the stage is dependent on the discharge, and often predicted using hydraulic routing models as a function of the predicted discharge. The prediction of precipitation is another challenging issue in the decision making procedure. Prediction of precipitation, not addressed in this study, is dependent on the physical information from rainfall radar stations and weather satellites rather than simple stochastic models, and; the long term prediction of flow series can be made separately from the prediction of precipitation.

This research focuses on integrated river basin management centering on reservoir system operation on the basis of uncertainty and risk-based management of water resources in an adaptive way. In this study, the uncertainty and risk-based management is associated with uncertainty of the inflow forecast and the risk of water quality standard violation. In this study, the following three problems are selected as the main concerns, with which water resources management agencies including *K*-water (Korea Water Resources Corporation) have faced in common: prediction of reservoir inflow and water quality; production of discharge series in streams by using rating curves and; multi-objective risk-based reservoir operation.

- Forecast system of the reservoir inflow and water quality

K-water is responsible for construction and operation of multipurpose reservoirs, water supply, hydropower generation, water treatment, and so on, all over Korea. Although forecasting the flows in reservoirs and streams is recognized as the most influential process in reservoir system operation, *K*-water has not developed a standardized long-term (e.g. monthly or yearly) and mid-term (e.g. weekly) forecast

system which employs advanced stochastic techniques. During normal seasons, *K*-water has mainly used simple statistics such as mean and percentile-duration flows from empirical flow duration curves. During flood season, selected return period flows based on frequency analysis and flows from rainfall-runoff event models with the precipitation forecasted by Korea Weather Administration have been used. However, one of the problems related to mid-term and long-term forecasts is that they have been mainly based on marginal probabilities of the variables instead of forecasting using conditional probabilities obtained in previous time steps in an adaptive and progressive manner. Likewise, as for predicting monthly water qualities in reservoirs, the technical problems that *K*-water has are almost the same as those associated with the inflow forecast. Compared to flow forecasting, the uncertainties in water quality prediction may be worse because it is hard to consistently identify the parameters of water quality models. Furthermore, these parameters are often highly variable in time and space.

- Development and uncertainty analysis of ratings for stream flows generation

In general, stream flows are calculated as a product of flow velocity and flow area. In order to obtain discharge continuously in time, two methods of direct and indirect measurement are generally considered. The direct measurement is made by observing stages over the crests of the structures such as flumes using uniquely predetermined ratings, or by installing flow velocity sensors such as the acoustic velocity meters in streams. The indirect acquisition of stage data is dependent on a rating that represents the relationship between stage (h) and discharge (Q), which is developed using stage-discharge data measured intermittently. The discharges are generally measured using the

propeller velocity meter or some advanced velocity meters such as a microwave water surface velocity meter and acoustic Doppler current profiler. Conventionally, stream flows are indirectly obtained using ratings because the direct measurement is very inefficient in terms of costs, time and efforts, and further is difficult technically.

K-water is measuring *h*-*Q* data at about 60 sites for the purpose of water resources planning and reservoir system operation. The *h*-*Q* measured data are used for developing the ratings generally expressed in the form of power function $Q=a(h-e)^b$:

where *a*, *b*, and *e* are the parameters, and *e* is specifically called zero flow stage or gauge height of zero flow, reflecting the hydraulic property of *h*-*Q* relationships. The ratings are traditionally developed and assessed every hydrologic year. In this procedure, some hydraulic features such as zero-flow stage and change of flow controls may be considered, and the reliabilities of developed ratings can be assessed using hydraulic models. However, when hydrologists and water resources system operators develop a rating, they tend to depend on only some statistical properties like the correlation coefficient and MSE (mean square error) to fit it to *h*-*Q* measurements.

In order to come up with solutions to these problems, *K*-water has developed a spreadsheet-based analysis system called "*HydroToolkit*" for developing and assessing ratings since the end of the 1990's *[K*-water, *2003]*. The systems mainly rely on simple statistics and comparison using graphs, and nonlinear optimization for parameter estimation. As another approach to getting highly reliable ratings, *K*-water tried to combine both Parshall flume and 1-dimensional hydraulic channel routing *[K*-water, *2004]*. The Parshall flume and hydraulic routing method were limited to analyses of low flow ratings and high to medium flow ratings respectively. Even though the hydraulic

routing models can simulate the *h-Q* relationships exactly, it is very difficult to represent the reliabilities of ratings probabilistically.

- Risk-based reservoir operation

K-water has tried to develop rule curves for operating its multi-purpose reservoirs using implicit stochastic optimization techniques. However, there has been a continuous gap between theoretical developments of optimization techniques and practical implementation. *Wurbs [1993]* and *Labadie [2004]* discuss possible reasons for this gap. As the key to success in implementation of reservoir optimization models, *Yeh [1985]* and *Labadie [2004]* recommended combination of simulation and optimization models with consideration of hydrologic uncertainties. However, the *K*-water still tends to be dependent on mainly simulation models. Moreover, these models are not supported by advanced reservoir inflow forecast systems.

As public awareness of the environment has been emphasized in Korea, the government and NGOs have been trying to find solutions to ensure enough water to keep streams environmentally sound. As one of the solutions, the increasing release from reservoirs has been required. However, many problems such as meeting cost payments and expected deficits of drinking and industrial water supply in the near future lie ahead. There is a good example where this type of complicated problem was addressed by mutual concessions and arbitration among the stakeholders. The city of Seoul, the capital in Korea, prepared a project for restoration of the 'Cheonggye-chen' in 2003, which is a stream flowing through the central area of Seoul. One of the biggest problems was to secure the water to maintain the water depth for attaining the goals of ecological and

esthetic aspects. For a solution to this problem, the city of Seoul planned to intake water from the Han River where *K*-water has a exclusive right to use water of this river. This issue was submitted to arbitration by the central river management commission of Korea to settle the dispute between the city of Seoul and *K*-water. In 2005, the commission decided that Seoul could use the water for free since the project was planned to promote the public interest. However, there is still a remaining problem that the impact of water supply to Seoul and Gyonggi province due to the loss of water in the main stream of the Han River has not been thoroughly assessed.

Thus, the intensive requirement for water quality and environmental improvement in streams and rivers is placing considerable pressure on reservoir operation. In addition, what makes the problem more difficult to solve is conflict with other beneficial water uses such as municipal water supply, irrigation, hydropower generation, and so on. From this viewpoint, it is essential to incorporate a sub-model for water quality and ecological analysis in reservoir system operation. *Ko [1997]* tried to develop an integrated system for establishing river basin operational planning considering water quantity and quality in the Han River aiming at more advanced reservoir operation of *K*-water. For stream water quality analysis in the downstream area of the reservoir system in his research, the *QUAL2E* water quality simulation model was used. The model was developed by EPA in the United States, and is widely used throughout the world. The approach in his research was to combine interactively a reservoir optimization model and *QUAL2E* for deriving the reservoir optimal operation policy. For easy connection between *QUAL2E* and the optimization model, the reservoir releases and water qualities produced by *QUAL2E* were analyzed by regression analysis. The developed regression functions were embedded in

the optimization model: hence the water qualities conditioned on the reservoir releases were treated deterministically as a soft constraint. A drawback of this approach is that it can not handle risk of water quality standard violation in the sense that although a simulated water quality does not exceed a predefined standard, there is still risk to exceed it due to uncertainties of the model parameters and natural variation. From this viewpoint, risk-based reservoir operation is very valuable for decision-making, and a tradeoff analysis between risk level and deficit of water supply can provide an appropriate guide for decision making.

1.3 Research Objectives

In a broad sense, the objectives of the research are to develop and apply probabilistic models for simulation and forecast of hydrologic and environmental characteristics using a data-driven approach, and to reflect the risks or uncertainties of the variables into an adaptive decision-making process for river basin management. The following techniques were evaluated in this research:

- Bayesian networks with a Bayesian Markov chain Monte Carlo (MCMC) learning algorithm;
- Nonparametric models using kernel density estimation (KDE) and *k*-nearest neighbor (*k*-NN) bootstrap methods;
- Stochastic artificial neural networks with a Bayesian MCMC method, and;
- Fuzzy rule-based modeling.

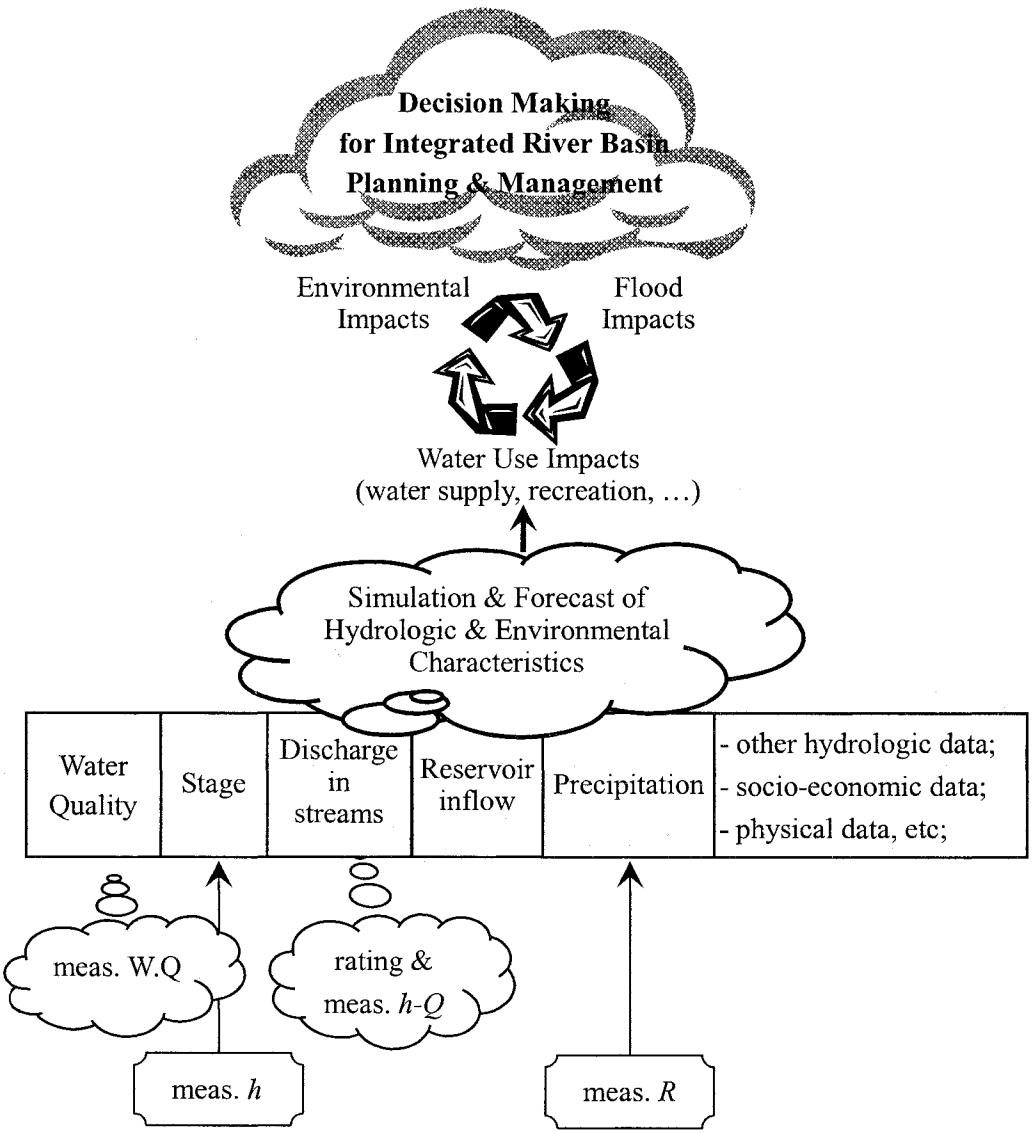


Figure 1-1 Uncertain factors that exist in decision making for integrated river basin planning and management, and each relationship: the variable and actions into cloud indicate they may have considerable uncertainties; "meas. W.Q", "meas. h-Q", "meas. h" and "meas. R" mean the measured water quality, stage-discharge, stage and precipitation data respectively.

As shown in Figure 1-1, the stage-discharge rating, water quality, and reservoir inflow (or stream discharge) are considered the most uncertain characteristics in integrated river basin planning and management. Therefore, this research focused on the uncertainty analysis of ratings and the development of probabilistic reservoir inflow and water quality forecast systems using these modeling techniques. The probabilistic water quality models were interactively joined with a reservoir optimal operation model for risk-based management in terms of biological oxygen demand (BOD) in streams and total phosphorus (TP) in reservoirs. The stochastic inflow forecast model was incorporated into a decision support system especially designed for the progressive operation of reservoir system in an adaptive way. A case study was performed in the Geum River basin in Korea. The components of the research are:

i) Development and uncertainty analysis of ratings:

- Deterministic approach:
  - 1-D unsteady flow analysis using the *HEC-RAS* software developed by the US Army Corps of Engineers;
  - NLP using *SOLVER* built into *MS-Excel*;
  - Fuzzy rule-based modeling using the *MATLAB-Fuzzy logic toolbox*;
- Probabilistic approach:
  - Bayesian data modeling using the *WinBUGS [Spiegelhalter et al., 1995]* software;
- Suggestion of the most relevant way by comparison of the methodologies in terms of ease of application, goodness of fit tests and facilitation of adaptive management;

ii) Development of a reservoir inflow forecast system:

- Stochastic artificial neural network modeling;
- Nonparametric modeling using KDE and *k*-NN bootstrap methods;
- Comparison of the methodologies and suggestion of the most appropriate one;

iii) Development of probabilistic BOD and TP models on the basis of:

- the classical Streeter-Phelp's BOD and Vollenweider's TP models with the help of a Bayesian network;

iv) Development of uncertainty and risk-based reservoir operation system:

- Development of an adaptive sampling implicit stochastic optimization model by coupling the probabilistic BOD and TP models and a reservoir inflow forecast model interactively with dynamic programming (DP);
- Development of a proto-type of decision support system including a data base, model base and user interface.

This research contributed to implementation of adaptive reservoir system operation with consideration of risk and uncertainty by joining probabilistic forecast models for hydro-environmental characteristics to reservoir system operation, which has been considered a very daunting task. Promising methods for probabilistic forecast models were proposed by comparing several popular methodologies. This research showed the possibility of application of stochastic artificial neural networks (ANNs) in the field of water resources, which has used almost exclusively deterministic ANN.

1.4 Organization of Dissertation

Chapter 2 describes the study area. Chapter 3 reviews the details of the stochastic modeling approaches including the literature review.

In Chapter 4, the principles of rating are reviewed first, then ratings are developed using deterministic and probabilistic methods. Finally, the methodologies are compared and the recommended method is suggested.

In Chapter 5, the monthly and daily reservoir inflow forecast systems are developed using stochastic artificial neural networks with a Bayesian MCMC learning method and nonparameteric time series modeling. Like the case of the rating analysis, the methodologies are compared and the most appropriate method is suggested.

In Chapter 6, the representative methodologies for uncertainty analysis are reviewed. The probabilistic water quality models of BOD at Gongju and TP in the Daecheong reservoir are developed, and the relationships between reservoir release and risk of violating the water quality standards are derived.

Chapter 7 presents an application of an adaptive sampling implicit stochastic optimization model by interactively coupling the water quality models. A generalized microcomputer dynamic programming (DP) package, *CSUDP [Labadie, 1999]* is used for optimal joint operation of the two reservoirs. A decision support system is developed, which helps to join the reservoir system operation model to the progressive optimization and inflow forecast models. Chapter 8 presents the summary and conclusions.

CHAPTER 2

AREA OF APPLICATION

2.1 Overview

Figure 2-1 shows the entire Geum River basin that is located in the southeast portion of the Korean peninsula. The study area starts from the upper area and ends at Gongju gage. The study area accounts for approximately 85% of the entire Geum River basin. There are two multi-purpose reservoirs, Daecheong (*DC*) and Yongdam (*YD*), in the Geum River basin, and they are managed by *K*-water.

The Daecheong multi-purpose reservoir was built in 1980 for the purpose of flood control, water supply to Daejon and Cheongju cities, and power generation. It is a concrete and rockfill combined type dam (72m in height, 495m in length and a volume of $1{,}234*10^6$ m^3).

The Yongdam multi-purpose reservoir is located upstream of Daecheong dam. It has a height 70m, a length of 498m and a storage of 815million m^3. Construction started in December of 1990 and ended in October of 2001. The *YD* dam was faced with serious regional conflicts between the lower basin and the areas of Jeonju, Iksan, Gunsan, Gimjea and the Gunjang industry complex that have benefits from the *YD* dam. The main issue was that the cities of the lower basin, which basically depend on *DC* reservoir for the life, did not want to share water because they believed the water quality of *DC* reservoir

would be deteriorated due to the decrease of inflow. These conflicts led to projects to analyze the effect of the diverted water on the water quality of *DC* reservoir, and a *DC-YD* reservoir optimal joint operation rule was prepared. Consequently, it turned out that the expected problem would not be so serious under the condition that the wastewater treatment plants be completed. Now, the two reservoirs are being operated normally.

2.2 Review of the sub-Areas for Rating Analysis

In this study, the rating analysis was performed at two stage-discharge measurement stations, the Hotan and Donghyang placed in the Daecheong dam and Yongdam dam basins respectively. The Hotan station is located on the main stream of the Geum River, while the Donghyang station on a comparatively small stream. They are very different in terms of hydraulic features: hence, these two stations were chosen as the study areas.

- Donghyang water level station

This site was especially selected for analysis for the purpose of determining the most relevant methodologies for discharge measurement by comparing a couple of options. The Donghyang station, equipped with a telemetry system, has been operated for managing the Yongdam reservoir since 1999, and the stage-discharge measurement has been made in the conventional way every year. For improving the accuracy of low flow measurements, a Parshall flume was built in 2002 at a 200m downstream location from the Donghyang station. Two supplementary stage gauges were installed at Daeya-kyo (1.5km upstream from the Donghyang station) and Somti-bo (1.6km downstream from the Donghyang station) in 2002 for computing medium to high discharge hydro-dynamically. Figure 2-2 illustrates the Yongdam dam basin with 930km^2 in area that is

composed of 8 sub-basins, and the Donghyang station with 164.4km^2 in area, the Parshall flume and two supplementary gauges located in the Guryang stream with 172.3km^2 of upstream area.

- Hotan water level station

This station was installed in 1979 for the purpose of managing the Daecheong reservoir at a place in between the Daecheong dam and the Yongdam dam. Like the Donghyang station, the stage-discharge measurement has been made in conventional way every year. Because the reach around this station is relatively wide and mild, the rating shows a hysteresis. Figure 2-3 shows a panoramic view of the station.

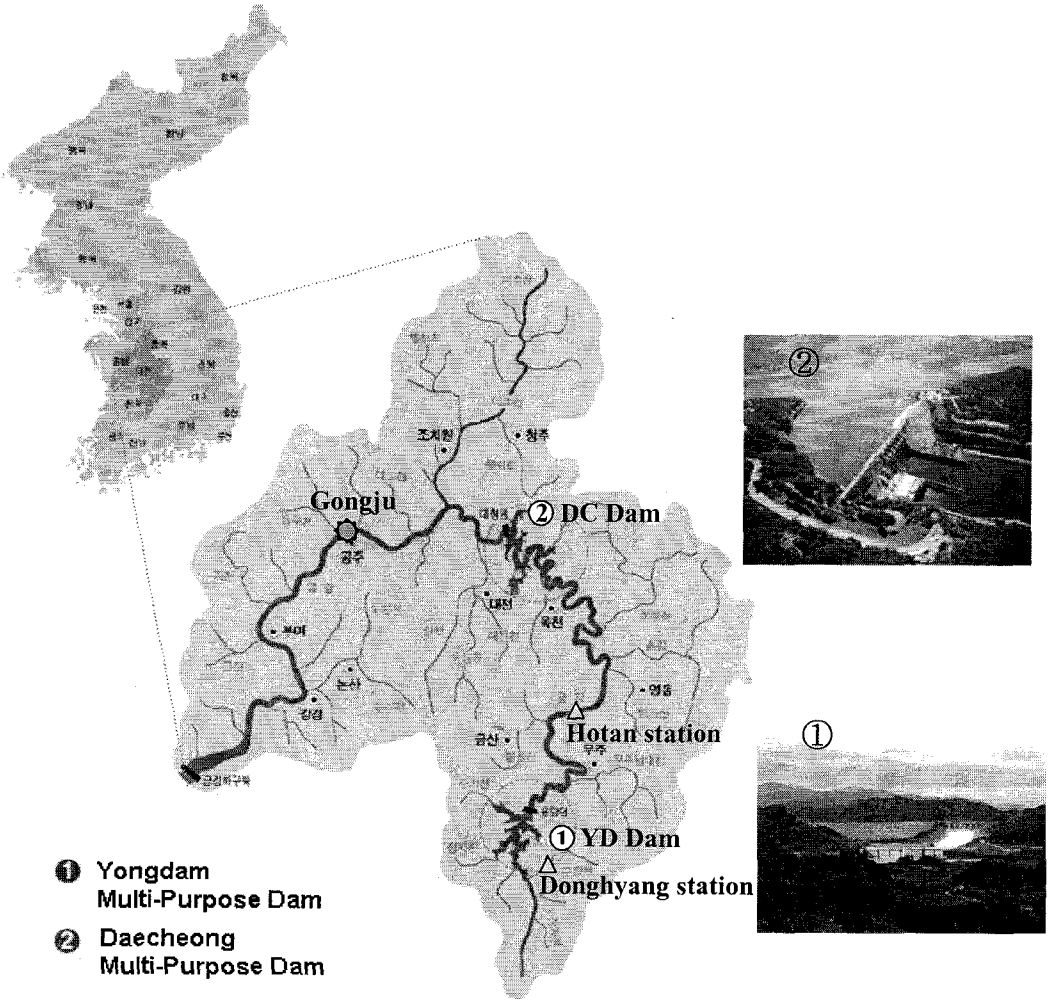


Figure 2-1 The study area composed of two multi-purpose dams, the Donghyang and Hotan stations for rating analysis, and Gongju city that is the control point for water quality management.

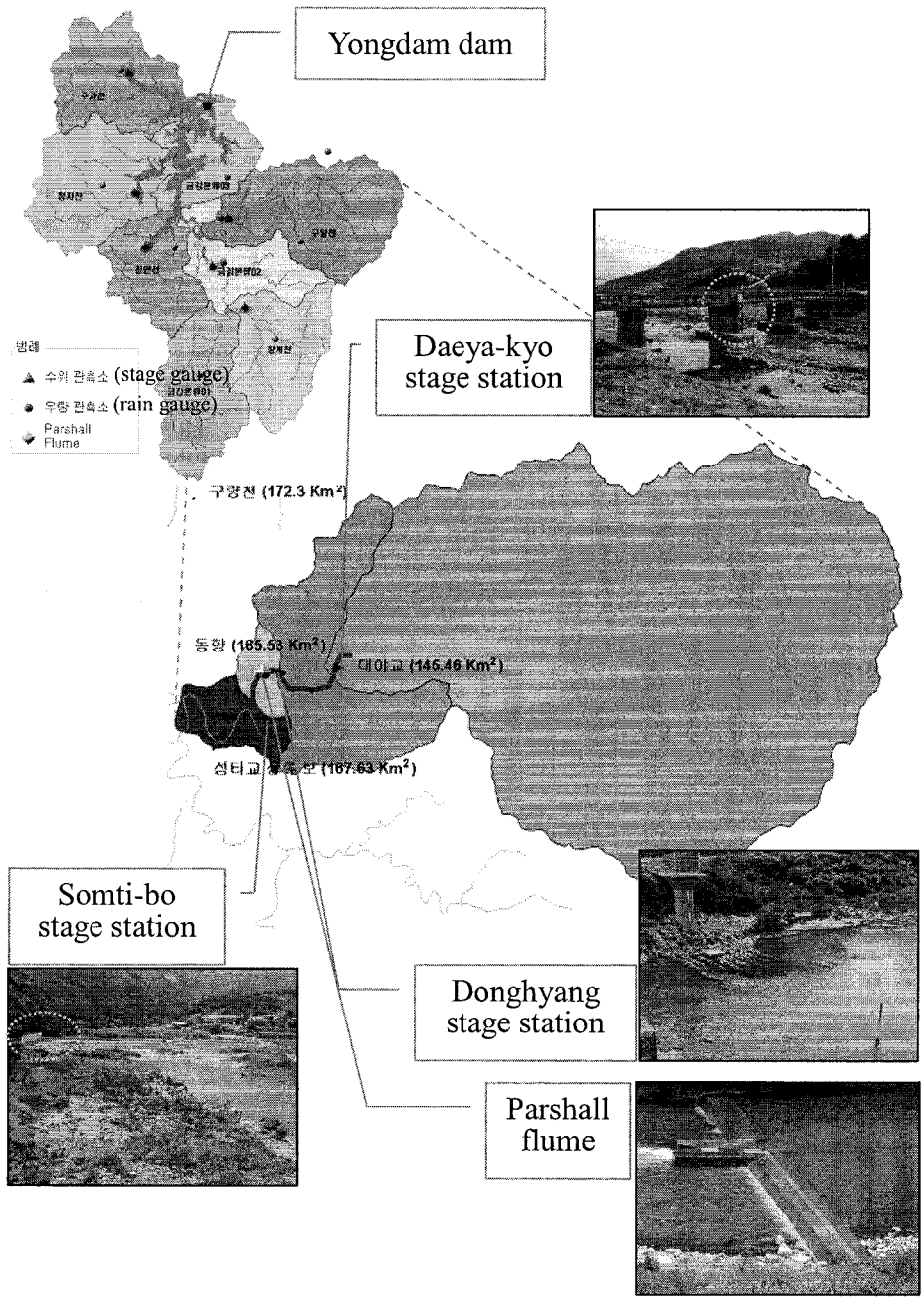


Figure 2-2 The study area centering on the Donghyang station for rating analysis.

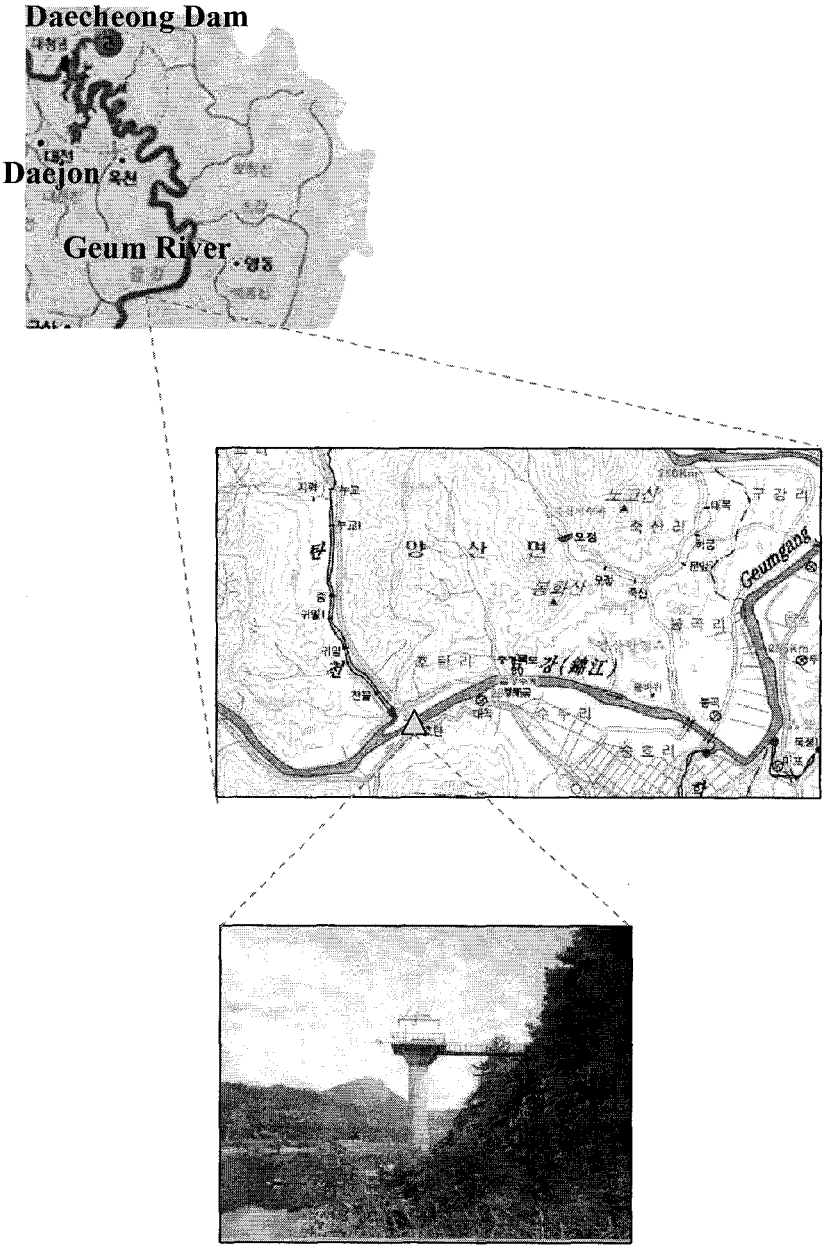


Figure 2-3 The study area centering on the Hotan station for rating analysis.

CHAPTER 3

DATA-DRIVEN STOCHASTIC MODELING FOR SIMULATION AND FORECAST OF HYDROLOGIC AND ENVIRONMENTAL CHARACTERISTICS

3.1 Data-driven and Behavior-driven Modeling

For simulation and forecast of various hydro-environmental characteristics such as precipitation, stream flows or water quality, a variety of models have been developed using three basic concepts: physically based behavior-driven, data-driven, and hybrid *[Solomatine, 2002]*. Traditionally, mathematical models have been widely used containing physical representations of the underlying natural processes such as infiltration, surface-subsurface-ground water flows, reaction kinetics of pollutants, and so on. For simple representation of the above-mentioned processes, lumped parameter conceptual models or empirical models have been developed. These models are often solved with finite difference or finite element methods in time and space. In this respect, the mathematical models are often called behavior-driven models.

In contrast, data-driven models basically recognize a certain pattern between inputs and outputs of a system that implicitly contain all information about the physical,

chemical and biological relationships, whether or not they are linear. In order to determine these relationships, the models employ pattern recognition engines with learning functions that have been devised in the areas of data mining, machine learning, and statistics rather than relying on mathematical descriptions of physical processes *[Solomatine, 2002]*. The learning process corresponds to the calibration process for model parameters in mathematical modeling. Data-driven models have been highlighted recently for solving water resources related problems and there has been a sudden rise of interest in data mining and pattern recognition using databases. The interest in data-driven models can be explained by the development of: cutting edge database management systems; communication technologies using telemetry systems and satellites, and; internet-based communication networks *[Velickov and Solomatine, 2000]*. These advanced technologies have helped develop data gathering and management systems, and the systematic databases have been coupled with techniques for extracting knowledge and patterns from the data about causes and impacts.

Hybrid models are somewhere in between behavior-driven and data-driven models. Bayesian network modeling employs relatively simple functional relationships and falls under this class of models. In Bayesian modeling underlying functions are often represented by relatively simple physical processes instead of directly describing the conditional distribution in presenting the dependent relationship between variables.

Taking a look at the features of the three modeling approaches, the behavior-driven models are very useful for analyzing "*what if*" type of queries: hence they might be more relevant to planning rather than managing water resources. For example, the hydrologic or hydraulic rainfall-runoff models can do simulation of the impacts

corresponding to the variation of hydrologic conditions (e.g. the impermeable area) due to urbanization and industrialization upstream. In this case, however, it is almost impossible to analyze this kind of assumed cause-impact relationships with data-driven models because the observed data reflecting theses relationships do not often exist. As opposed to behavior-driven models, data-driven models are often relevant for management and control of a system in that some of them are based on stochastic analysis, or can be easily extended to the stochastic modeling. This indicates it is possible to make risk-based decisions by performing uncertainty analysis of models and parameters *[Reckhow, 1999]*. However, some techniques such as nonparametric modeling are often limited to only interpolation in forecasting the system variables because they rely on observations absolutely. The hybrid models have the features of both modeling approaches, and can be used for both purposes.

The following sub-sections present some representative stochastic modeling approaches: Bayesian probability networks; nonparametric modeling; stochastic artificial neural networks (ANN); fuzzy rule-based modeling, and; stochastic state-space modeling based on Kalman filtering. Both Bayesian networks and stochastic ANN models are generally trained using a Bayesian MCMC algorithm. Table 3-1 shows the basic structure of each method. Given the underlying functional relationships, if the states of the independent system variables are known, the prime concern in the stochastic modeling is to obtain the conditional probability of the dependent variable no matter whether it is Gaussian or non-Gaussian. Another concern is to determine the posterior probabilities of the system parameters, which are useful for progressive adaptive management.

Table 3-1 Probabilistic representation of stochastic time series models.

Stochastic model	Representation
General probabilistic Expression for time series modeling	$p(x_t \mid x_{t-1}, x_{t-2}, \ldots) = \dfrac{p(x_t, x_{t-1}, x_{t-2}, \ldots)}{\int_{x_t} p(x_t, x_{t-1}, x_{t-2}, \ldots) dx_t}$
Bayesian probability networks	$p(x_t, x_{t-1}, x_{t-2}, \ldots) = p(x_t \mid x_{t-1}) \times p(x_{t-1} \mid x_{t-2}) \times p(x_{t-2} \mid x_{t-3}) \times \ldots$: Factorization of joint probability; $x_t = f(\theta, x_{t-1}, x_{t-2}, \ldots) + \varepsilon_t$: $\varepsilon_t \sim N(0, \sigma^2)$: functional relationship between system variables of interest; $p(\theta \mid x_t) = p(x_t \mid \theta, x_{t-1}) / p(x_t)$: Bayesian learning
Nonparametric modeling	$p(X_t \mid X_{t-1}) = \sum_{i=1}^{n} \dfrac{1}{(2\pi\lambda^2 \mathrm{S}')^{1/2}} w_i \exp\left(-\dfrac{1}{2} \dfrac{(X_{t-1} - b_i)^2}{\lambda^2 \mathrm{S}'} \right)$: Approximation of a conditional density function using Gaussian kernel function
Stochastic ANN	$\tilde{X}_t = f(\theta, X_{t-1}) = \theta_1 + \sum_{j=1}^{h} \theta_2 \tanh\left[\theta_3 + \sum_{i=1}^{n} \theta_{4,j}^{(i)} X_{t-1}^{(i)} \right]$ where, $\tanh(z) = \dfrac{e^z - e^{-z}}{e^z + e^{-z}}$; $p(\theta \mid x_t) = p(x_t \mid \theta, x_{t-1}) / p(x_t)$: Bayesian learning
Fuzzy rule-based modeling	*If* $(X_{t-1}\ \Theta X_{t-2}\ \Theta, \ldots)$ *then* X_t: fuzzy rule; $\mu_{Xt} = \sum \mu'_{i, Xt}$: combined fuzzy set; $M(X_t) = \dfrac{\sum v_i M(X_t^i)}{\sum v_i}$: defuzzification
State-space modeling	$X_t = A_t X_{t-1} + \Gamma w_{t-1}$, $w \sim N(0, Q)$: dynamic equation; $Z_t = H^* X_t + v_t$, $v \sim N(0, R)$: measurement equation; $K_t = P_{t\mid t-1} H^T (H^* P_{t\mid t-1} H^T + R)^{-1}$: Kalman filtering; where, $P_{t\mid t-1} = A_t P_{t-1\mid t-1} A_t^T + \Gamma Q \Gamma^T$: forecast error variance

3.2 Bayesian Probability Networks and Bayesian Learning

3.2.1 Fundamentals of Bayesian/Belief/Causal/Graphical Network

Bayesian (probability) networks are graphical, probabilistic models that allow the structured representation of a cognitive process. They have been accepted as tools for decision-making in complex situation within a variety of disciplines *[Pearl, 1988]*. A Bayesian network is designed using a graphical model representing the functional relationships among system variables with their probabilistic dependences. All linkages are quantified probabilistically using data and expert judgment *[Reckhow, 1999]*.

1) Structure of Bayesian Network

Bayesian networks begin with a graphical depiction with nodes linked by arrows. In the graph, nodes represent system variables and an arrow from a parent node to child node represents a dependency between two linked variables. There are two types of networks: feedback cycles and a directed acyclic graph (DAG). The scheme of feedback cycles provides both forward (cause-to-effect) and backward (effect-to-cause) information propagation so that information can arrive at any node, while DAGs consider only forward propagation. DAGs are mainly used to represent causal relationships, and known as Bayesian networks due to reliance on Bayes's theorem as the basis for updating information *[Pearl, 2000]*.

2) Application of Bayesian Network

In hydrologic and environmental modeling the Bayesian method has been used mainly for parameter estimation and uncertainty analysis *[Dilks et al., 1992; Borsuk et al., 2000; Thiemann et al., 2001; Qian et al., 2003; Kanso et al., 2003]*. When a system

is modeled with a nonlinear, non-Gaussian function of the form $Y=f(X, \theta)+\epsilon$ (where Y, X, θ and ϵ are a system variable vector, input vector, parameter vector and error respectively), the main concern is to determine the posterior probabilities of θ and uncertainties of Y due to uncertain parameter values. According to the dimension of system variable vector, the structure of a Bayesian model could be one layer or multi-layer on the basis of DAGs. A multi-layered model is referred to as Bayesian network.

For example, suppose a stream reach consists of a dam as the upper boundary, a water quality measurement station for managing water quality as the lower boundary, and two internal water quality measurement sites before and after the junction with a tributary. The water qualities at the three sites are affected by the release (Q) from the dam, the BOD load (L_1) and the discharge (Q_1) from the tributary. Figure 3-1 illustrates a Bayesian network for a BOD model with two inputs (L_0-Q_0, L_1-Q_1), three system variables (BOD_1, BOD_2, BOD_3), and two BOD reaction kinetic rates (K_{r1}, K_{r2}). This network shows directly linked cause-effect relationships among the system variables. From the network, it can be noticed that BOD_3 is directly affected by BOD_2 that results from L_1, Q_1 and BOD_1, and BOD_1 is influenced by Q_0 and L_0.

The network can be represented probabilistically with a joint probability that can be factored into several conditional and marginal probabilities (Equation 3.1). However, the joint probability is not practical because it is too complicated to get an analytical solution. On the contrary, the conditional probability offers very useful information because human reasoning typically is based on marginal or conditional situations. In this example, the most interesting variable is most likely to be BOD_3 at the control point conditioned on BOD_2, BOD_1, L_0, Q_0, Q_1, L_1 , K_{r1} and K_{r2}.

Probabilistically BOD_3 can be represented by a conditional probability $P(BOD_3|\ BOD_2, BOD_1, L_0, Q_0, Q_1, L_1, K_{r1}, K_{r2})$, which is very challengeable to solve. Taking a close look at the network, it can be noticed that the network implicates a spatial lag-1 Markovian property, that is once the value of BOD_2 is known, the values of BOD_1, L_0, L_1 and K_{r1} are not required to predict BOD_3. Given L_0, Q_0, L_1, Q_1 deterministically, this fact helps the full joint probability (Eq. 3.1) reduce to three independent conditional probabilities for the system variables and 2 marginal probabilities for the parameters (Eq. 3.2(a)). Finally, the joint posterior probability of the reaction kinetic rates is represented in Equation 3.2(b). This conditional independence greatly simplifies the BOD model, which ends up with three linked sub-models (layers): $P(BOD_1|L_0, Q_0, K_{r1})$; $P(BOD_2|L_1, BOD_1, Q_0, Q_1)$ and; $P(BOD_3|BOD_2, Q_0, Q_1, K_{r2})$. Each sub-model can be solved simply by using Bayesian learning.

$$\begin{aligned} P(BOD_3, BOD_2, BOD_1, L_0, Q_0, Q_1, L_1, K_{r1}, K_{r2}) = {} & P(L_0) \times P(Q_0) \times P(Q_1) \times P(L_1) \\ & \times P(K_{r1}) \times P(K_{r2}) \\ & \times P(BOD_1|\ L_0, Q_0, K_{r1}) \\ & \times P(BOD_2|\ BOD_1, L_0, Q_0, Q_1, L_1, K_{r1}) \\ & \times P(BOD_3|\ BOD_2, BOD_1, L_0, Q_0, Q_1, L_1, K_{r1}, K_{r2}) \end{aligned} \tag{3.1}$$

$$\begin{aligned} P(BOD_3, BOD_2, BOD_1, K_{r1}, K_{r2}) = {} & P(K_{r1}) \times P(K_{r2}) \\ & \times P(BOD_1|K_{r1}) \\ & \times P(BOD_2|\ BOD_1) \\ & \times P(BOD_3|\ BOD_2, K_{r2}) \end{aligned} \tag{3.2(a)}$$

$$\begin{aligned} P(K_{r1}, K_{r2}|\ BOD_3, BOD_2, BOD_1) \propto {} & P(K_{r1}) \times P(K_{r2}) \\ & \times P(BOD_1|K_{r1}) \\ & \times P(BOD_2|\ BOD_1) \\ & \times P(BOD_3|\ BOD_2, K_{r2}) \end{aligned} \tag{3.2(b)}$$

3) Quantification of conditional probability in Bayesian Networks

After construction of a Bayesian network, the next step is to quantify the causal relationships between the nodes with conditional probabilities, which are indicated by an arrow between a parent and child node in a network. If a child node has multiple parent nodes, the probability is conditional on every possible combination of the parent nodes. A node that has no incoming arrows is described probabilistically by a marginal probability. To perform a network analysis, the conditional probabilities should be elicited using expert judgment, statistical analysis of observed data and a functional relationship in the form of $X=f(P, \theta, \epsilon)$, where X, P, θ and ϵ are system variables, a vector of causes (or parents), a parameter vector and an error term respectively *[Reckhow, 1999; Borsuk, 2004]*. In the Bayesian network example for BOD a model, the classical *Streeter-Phelp*'s BOD-DO model might be considered for the functional relationship. Compared to directly assigning conditional probabilities, the functional characterization of the causal relationships leads to providing more direct and meaningful information for making a decision by preparing several alternatives using the 'if-then' type of analysis by changing the values of system variables *[Borsuk, 2004]*. For example, one can analyze how the probabilities of BOD_2 and BOD_3 would be affected by the intentional increase and decrease of BOD load (L_1) from the tributary.

3.2.2 Bayesian Learning/Inference/Analysis Using MCMC Techniques

1) Fundamentals of Bayesian Learning

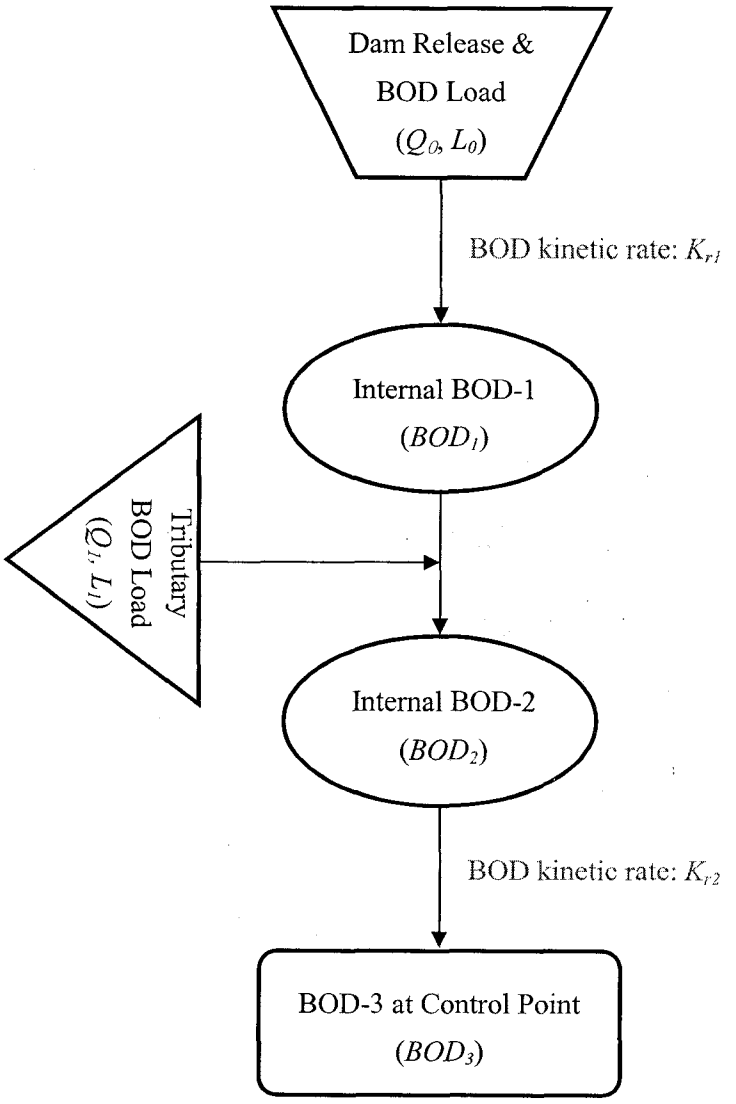


Figure 3-1 Graphical representation of a Bayesian network for a BOD model.

A deterministic learning approach provides a single set of posterior parameters, while Bayesian learning can do both the posterior distributions over the parameters and the predictive distributions for the system variables. In addition, Bayesian learning provides successive updates of prior information as new data become available. This is referred to as an adaptive method or data assimilation. The advantages of Bayesian inference can be summarized as *[Gelman, et al., 1995]*:

- to include important prior information;
- to handle uncertainty explicitly;
- to assimilate new information (observation) in adaptive way.

Bayes theorem lies at the heart of Bayesian learning, and can be employed in probabilistic modeling with underlying system equations in the form of $Y=f(X, \theta, \epsilon)$:

$$p(\theta \mid Y) = \frac{p(\theta, Y)}{p(Y)} = \frac{p(Y \mid \theta)p(\theta)}{\int p(Y \mid \theta)p(\theta)d\theta} \propto p(Y \mid \theta)p(\theta) \tag{3.3}$$

where:

θ	= uncertain model parameter vector;
Y	= observations;
$p(\theta \mid Y)$	= posterior distribution of parameter θ;
$p(\theta, Y)$	= joint distribution;
$p(Y \mid \theta)$	= likelihood function;
$p(\theta)$	= prior distribution of parameter θ;
$p(Y)$	= marginal distribution of data.

Bayesian analysis addresses the unknown model parameter vector θ randomly, which may have some prior beliefs described by probability functions $p(\theta)$. The priors can be generally determined by expert judgment, statistical analysis over historical data, conjugate priors for guaranteeing closed form of posteriors, or no information of θ. The first three priors are informative, and the latter, non-informative, which are often

expressed by Uniform distributions. The likelihood function $p(Y|\theta)$ reflects how well parameters allow a model to describe data: hence this is used to update the priors for parameters using newly gathered data. The term $p(Y)$ in Equation 3.3 is called a prior predictive distribution, which is not conditioned on previous observations and not dependent on θ. Thus this is considered a constant. Hence the posterior density of θ is generally described by the product of a likelihood function and prior distributions in an unnormalized form *[Gelman et al., 1995]*.

In association with determination of priors, one can criticize Bayesian analysis because posteriors can be affected intentionally by subjective priors. This can be true especially in some beginning steps in a stochastic process, if the priors are chosen subjectively. This problem can be lessened with asymptotic theory by using large samples and non-informative priors *[Dowd et al., 2003]*. Furthermore, the impact of subjective priors can be reduced through adaptively successive updates of priors.

As one of problems in association with Bayesian inference, most of applications in the literature have used simplified functional relationships instead of employing very sophisticated processes represented in mathematical models such as CE-QUAL-W2. This is probably due to difficulties in incorporating chemical, physical and biological processes into probabilistic models. Consequently, there is an ongoing debate as to the relative merits of simple models that statistically fit data versus more sophisticated models. However, the two approaches are different in their goals. Mathematical models can be used to present and assess scientific hypotheses, and therefore augment scientific judgment. Bayesian network models ideally provide the interface among engineers, stakeholders and decision makers in the context of decision making *[Reckhow, 1999]*.

2) Markov chain Monte Carlo (MCMC) Sampling Techniques

For years, Bayesian inference was not largely used in practice in favor of frequentist inference. The main reasons were computational difficulties and the formal use of subjective priors. In recent years, with the help of new computational approaches such as MCMC, the first problem has been greatly reduced *[Dowd et al., 2003; Gelman, et al. 1995]*. Bayesian inference using MCMC has been shown to be particularly useful for ecological models with poor parameter identification *[Qian et al., 2003]*. For probabilistic modeling using a Bayesian approach, the Bayesian Markov chain (BMC) demonstrated its potential for uncertainty analysis of ecological models *[Dilks et al., 1992]*. *Quin et al. [2003]* proved that a Bayesian MCMC method was superior to the BMC by showing that the BMC was very inefficient if it had many parameters. Figure 3-2 illustrates the modeling procedure of Bayesian inference. For Bayesian inference, two methods can be generally considered: direct simulation and the MCMC method.

- Direct simulation vs. MCMC method *[Gelman et al., 1995]*

In simple Bayesian models where conjugate priors are applied, it is often easy to sample from the joint posterior probability distribution directly. For more complicated model with no closed form for a joint posterior probability, the basic idea of direct simulation is to factor a known joint probability analytically into a conditional and marginal probability in closed form in sequence, and simulate in part in inverse order.

MCMC is often used when sampling θ from $p(\theta|Y)$ is not easy because $p(\theta|Y)$ is not represented in closed form, or because of computational complexity due to many parameters θ. Instead, MCMC generates a draw from a distribution that approximates some target distribution $p(\theta|Y)$: hence, the key to a MCMC simulation is to construct an

irreducible, aperiodic Markov process to ensure convergence to a stationary distribution. There are various strategies for constructing MCMC algorithms, but all of them are special cases of the Metropolis-Hastings (MH) algorithm, and the MH algorithm and Gibbs sampler have been mainly employed in practice.

- Metropolis-Hastings algorithm and Gibbs Sampler *[Gelman et al., 1995]*

The strategy of the MH algorithm is to construct a random walk by applying an acceptance-rejection rule so that the samples from a proposal distribution $g(\cdot)$ that surrounds target distribution $f(\cdot)$ may converge to a known target distribution $p(\theta|Y)$:

i) The method begins at t=0 with the selection of θ^0. Given θ^t, sample a candidate value θ^* from $g(\cdot)$;

ii) Compute the MH ratio $R(\theta^t, \theta^*)$;

$$R(\theta^t,\theta^*) = \frac{f(\theta^*)g(\theta^t \mid \theta^*)}{f(\theta^t)g(\theta^* \mid \theta^t)} \tag{3.4}$$

iii) Sample a value θ^{t+1} according to the following:

$$\theta^{t+1} = \begin{cases} \theta^* \text{ with probability } \min\{R(\theta^t,\theta^*),1\} \\ \theta^t \text{ otherwise;} \end{cases} \tag{3.5}$$

iv) Increment t and return to step *i)*.

The strategy of the Gibbs sampler is to sequentially sample from conditional posterior distributions that are often available in closed form. This method is specially adopted for multidimensional target distributions. The algorithm with p parameters is:

i) Select starting values θ^0 at t=0;

ii) Generate in turn,

$$\begin{aligned} \theta_1^{t+1} \mid \cdot \sim f(\theta_1 \mid \theta_2^t,\ldots,\theta_p^t) \\ \theta_2^{t+1} \mid \cdot \sim f(\theta_2 \mid \theta_1^t,\theta_3^t,\ldots,\theta_p^t) \end{aligned} \tag{3.6}$$

$$\vdots$$

$$\theta_p^{t+1} \mid \cdot \sim f(\theta_p \mid \theta_1^t, \theta_2^t, \ldots, \theta_{p-1}^t)$$

iii) Increment t and return to step *i)*.

3) Model and MCMC diagnostics/Sensitivity analysis *[Gelman et al., 1995]*

After determining posteriors, three basic assessments are generally required for checking goodness-of-fit of models and convergence to a stationary distribution of samples: model diagnostics, sensitivity analysis, and MCMC diagnostics.

Model diagnostics relates to assessment of models in terms of goodness-of-fit with a learning data set and accuracy of prediction with a validation data set. This process is referred to as a posterior predictive check conditioned on the historical data Y:

$$p(\tilde{Y} \mid Y) = \int p(\tilde{Y}, \theta \mid Y) d\theta = \int p(\tilde{Y} \mid \theta) p(\theta \mid Y) d\theta \qquad (3.7)$$

where, $\tilde{Y}$ is the simulated values obtained from the learning or the validation data set.

Sensitivity analysis is used for assessing how sensitive posteriors are when a current model is replaced with other models, and how sensitive changes of priors are to posteriors. MCMC diagnostics provides methods for assessing convergence to a stationary distribution of samples. Some useful measures for checking convergence are the degree of mixing within and between two chains, which is assessed by the graphs for history paths, autocorrelations of the chains, the *Gelman-Rubin* statistic R *[Gelman et al., 1995]* and correlation between the parameters. When a mixing rate is measured using the *Gelman-Rubin* statistic R, this value should be less than about 1.1 for convergence *[Gelman et al., 1995]*. If poor convergence is indicated from the MCMC diagnostics, a model should be improved via reparameterization process.

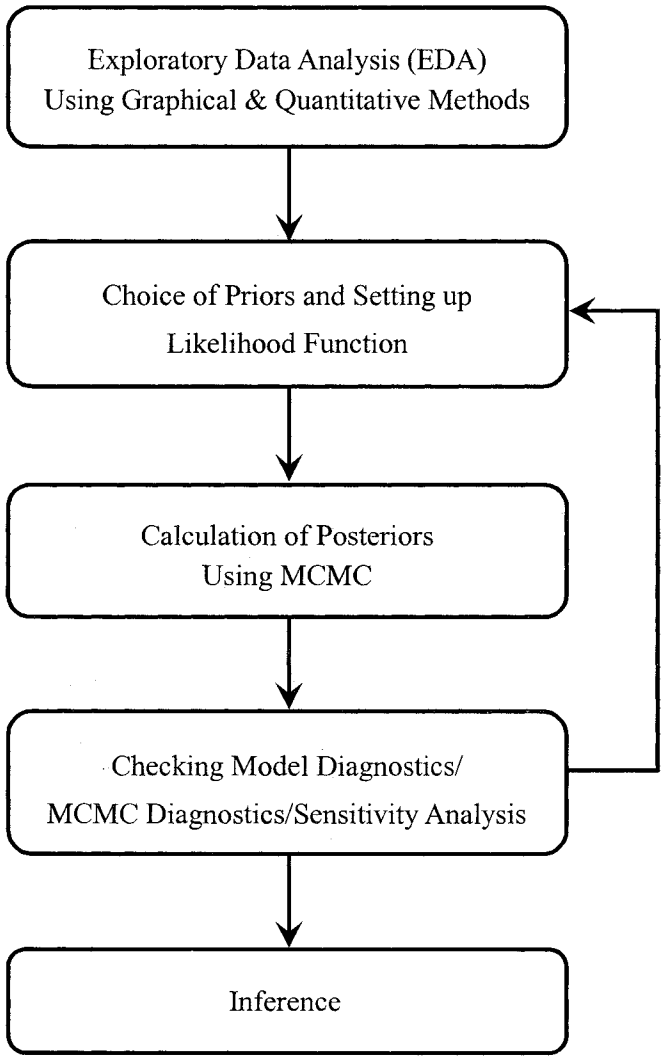


Figure 3-2 Summary of Bayesian Data Analysis.

3.3 Nonparametric Modeling

3.3.1 Comparison of Parametric and Nonparametric Approaches

A random process is a sequence of random variables indexed in time. The infinite set of all possible realizations is called the ensemble. A time series is completely described by a full joint probability distribution, and it can be factored into its conditional and marginal probabilities. The conditional probabilities play the most essential role in stochastic modeling, while the joint probability is not so meaningful in practice. For simplicity of the dependence structure of the conditional probability, a random process is often assumed to be Markovian, that is, dependent on only a finite set of prior values *[Bras and Rodriguez, 1985]*.

Considering a stochastic process $\{x_1, x_2, ..., x_t, ..\}$ for a random variable X, an order p Markov model is defined on the basis of conditional probability as:

$$P(x_t \mid x_{t-1},.....,x_{t-p}) = \frac{P(x_t,x_{t-1},.....,x_{t-p})}{P(x_{t-1},.....,x_{t-p})} = \frac{P(x_t,x_{t-1},.....,x_{t-p})}{\int P(x_t,x_{t-1},.....,x_{t-p})\,dx_t} \tag{3.8}$$

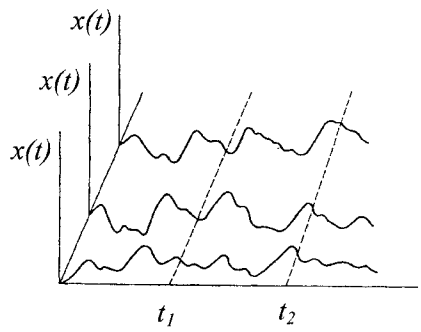


Figure 3-3 Ensemble of time series *[Bras and Rodriguez, 1985]*.

In order to obtain an analytical solution for Equation 3.8, the joint probabilities of both the numerator and denominator should be determined. However, this work is very difficult if the order p is high. Hence, hydrologists have contributed to developing two methodologies: 1) the classical time series analysis such as ARMA (auto regressive and moving average) based on a parametric approach and; 2) nonparametric approaches.

In parametric analysis, the Markovian dependence structure in conditional probabilities is simplified using functional relationships of the form: $X_t=f(X_{t-1},..,X_{t-p},\theta)+\epsilon$, where θ and ϵ are the parameter vector and residual respectively. Linear functions are widely used and residuals are often assumed to be normally distributed *[Salas, 1993]*. Unlike the Bayesian approach, the parameters of the models are treated deterministically. As an alternative for overcoming the drawback of linearity, neural network modeling has been studied *[Salas et al., 2000]*.

In contrast, nonparametric modeling, which is motivated by the theoretical background in time series modeling by *Yakowitz [1979]*, tries to obtain the conditional probability directly from the observed data. Specifically, the methods for density function estimation strive to approximate the underlying density locally using data from a small neighborhood around the point of estimation *[Lall, 1995]*. They impose only weak assumptions such as continuity of the target functions rather than *a priori* specification of the target function or assumption of a particular parametric standard distribution. The goal of nonparametric modeling is to completely eliminate *a priori* assumptions of specific probability distributions for random variables *[Adamowski and Feluch, 1991]*. However, it has been known that this method has drawbacks such as confidence bounds that are usually much wider than these from parametric analysis. Additionally predictions

outside the range of the observations are not possible *[Sharma et al., 1997]*. The representative methodologies for nonparametric modeling, which have won popularity in water resources planning and management, are *k*-NN techniques *[Lall and Sharma, 1996]* and kernel density estimation *[Sharma et al., 1997]*. Nonparametric methods can be employed in several areas such as time series analysis *[Tarboton et al., 1998; Lall and Sharma, 1996; Sharma et al., 1997]*, frequency analysis *[Gingras and Adamowski, 1994]*, spatial analysis *[Lall and Bosworth, 1993]* and general regression analysis with uncertainty considered probabilistically *[Adamowski and Feluch, 1991]*.

3.3.2 Nonparametric Modeling Techniques

Yakowitz [1973, 1979] applied a finite order Markov chain to hydrologic time series, however, he realized that the discretized transition probability matrix of the state space became unmanageable as the number of parameters became large. To overcome this weakness, a creative idea of a transition function was developed, which represents empirical conditional PDFs that are evaluated using a nearest neighbor or kernel density estimate method.

1) *k*-nearest neighbor (*k*-NN) resampling technique using bootstrap

The *k*-NN algorithm discussed in this section is based on the paper by *Lall and Sharma [1996]*. A generic *k*-NN density estimator is defined as *[Silverman, 1987]*:

$$f(x) = \frac{1}{r_k^d(x) n} \sum_{i=1}^{n} K\left(\frac{x - x_i}{r_k(x)}\right) \tag{3.9}$$

where:

K = kernel function;

d = the dimension of the state space, equals to the order of Markov chain;

n = the number of observed data points;

$r_k(x)$= Euclidean distance from a data point at time t to the k_{th} nearest data point.

The kernel function $K(\cdot)$ determines the behavior of the tail of the density function, and the number of neighbors k serves as smoothing factor.

In a historical time series $X=(x_0, x_1, \ldots, x_n)$, for a conditional probability defined in Equation 3.8, let the d dependent variables at time t, $(x_{t-1}, x_{t-2}, \ldots, x_{t-d})$ be called a feature vector and denote it D_t, and the simulated or forecasted value, x_t, called a successor and denote it S_t. A conditional probability can be represented as $p(S_t|D_t)$. Let an arbitrary time i be a current time, then the k nearest neighbors $(D_j : j=1..k)$ around D_i can be determined according to the Euclidean distance between D_i and D_j. The neighbors are composed of k pair of D_j and S_j as:

$$J_{i,k} = \left\{ \begin{pmatrix} S_1 \\ D_1 \end{pmatrix}, \begin{pmatrix} S_2 \\ D_2 \end{pmatrix}, \ldots, \begin{pmatrix} S_k \\ D_k \end{pmatrix} \right\}; \tag{3.10}$$

Once the k nearest neighbors are determined, they are sorted in ascending order in terms of distance. If there are some neighbors with the same distance, they can be treated randomly for resampling. The expected value of S_i conditioned on D_i can be estimated by averaging the successors S_j with weights. The weights are expressed as a discrete mass function depending on the distance between D_i and D_j. In practice, a conditional probability $p(S_i|D_i)$ is often evaluated by bootstrap resampling using a monotonically and rapidly decreasing empirical distribution function (EDF), which serves as the kernel function K. The probability of an EDF is in inverse proportion to the Euclidean distance. When a set of S_i is resampled using $K(\cdot)$, a S_i is sampled using a cumulative distribution

function of EDF by sampling a random value U from a Uniform distribution between 0 and 1. A discrete resampling kernel function K may be defined as:

$$K(i,j) = \frac{1/j}{\sum_{j=1}^{k} 1/j} \tag{3.11}$$

$$K(i,j) = \frac{\frac{1}{r_{ij}/r_{i1}}}{\sum_{j=1}^{k} \frac{1}{r_{ij}/r_{i1}}} = \frac{r_{i1}/r_{ij}}{\sum_{j=1}^{k} r_{i1}/r_{ij}} \tag{3.12}$$

Equation 3.12 considers a real distance r_{ij} between D_i and D_j and leads to a rough value of K, while Equation 3.11 leads to a smooth K (Figure 3-4).

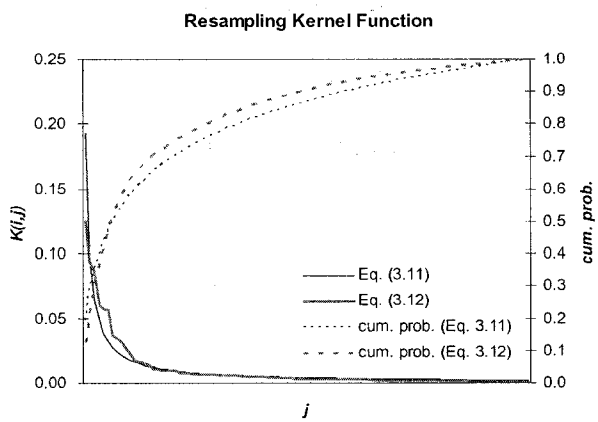


Figure 3-4 Two types of kernel functions for k-NN nonparametric modeling.

For k-NN modeling, the two parameters of k and model order d should be calibrated. The model order can be identified using classical methods such as correlogram, spectrum and Akaike information criteria (AIC) analysis *[Salas, 1993]*. According to *Silverman [1986]*, the bias-variance tradeoff plays an important role in choice of band width k, which indicates that the smaller the band width, the less the bias

and the higher the variance because this may result in a rough density estimate. For determining k, two methods may be considered: prescriptive choice of $k=n^{1/2}$ and optimization of k with a generalized cross validation (GCV) function:

$$\min_k f_{GCV}(k) = \frac{\sum_{i=1}^{n} e_i^2 / n}{\left(1 - 1/\sum_{j=1}^{k} 1/j\right)}; \quad e_i = x_t - \tilde{x}_t \tag{3.13}$$

Lall et al. [1996] mentioned that the prescriptive choice is a good alternative for $1 \leq d \leq 6$, and $n \geq 100$ because it has been shown that the sensitivity to the choice of k is small.

2) Kernel Probability Density Estimate (KDE)

The KDE algorithm described here is based on the paper by *Sharma et al. [1997]*. The basic idea of time series modeling using KDE is to derive a conditional probability density function from joint and marginal PDFs with kernel functions using a historical time series $X=(x_0, x_1, \ldots, x_n)$. A general univariate kernel probability density estimator is written *[Silverman, 1986]*:

$$f_{KDE}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - x_i}{h}\right) \tag{3.14}$$

where:

n = the number of observed data;
h = the band width (smoothing factor);
K = kernel function that must be integrate to 1.

From Equation 3.14, it is noticed that the KDE is formed by summing all sub-kernels centered at each observation x_i. For kernel functions, the Gaussian function is the most popular and practical choice *[Silverman, 1986]*. A multidimensional extension

of an univariate KDE is represented for a vector X_t in d dimensions (e.g. a bivariate joint KDE $p(x_t, x_{t-1})$ is 2 dimensional) with the Gaussian kernel function:

$$f_{KDE}(X) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{(2\pi)^{d/2}\det(\mathrm{H})^{1/2}} \exp\left(-\frac{1}{2}\left[(X - X_i)^T \mathrm{H}^{-1}(X - X_i)\right]\right) \tag{3.15}$$

where H serves as both the covariance matrix of the Gaussian kernel and the band width matrix, which must be a symmetric positive definite $d \times d$ matrix. Summing Gaussian sub-kernels with a covariance H centered at each observation forms a KDE. H is specified with the sample covariance matrix S and a scaling factor λ:

$$H = \lambda^2 S \tag{3.16}$$

In determining λ and h, the bias-variance tradeoff is a fundamental idea. In case the underlying distribution is known to be a Gaussian distribution with a Gaussian kernel function, *Silverman [1986]* gave an asymptotic optimal scaling factor of band width as:

$$\lambda = \left(\frac{4}{d+2}\right)^{1/(d+4)} n^{-1/(d+4)} \tag{3.17}$$

On the contrary, when the underlying distribution is not known, λ can be decided by a least square cross validation (LSCV) based on an integrated square error (ISE):

$$\min_{\lambda} f_{LSCV}(\mathrm{H}) = \frac{1 + (1/n)\sum_{i=1}^{n}\sum_{j\neq i}[\exp(-L_{ij}/4) - 2^{d/2+1}\exp(-L_{ij}/2)]}{(2\pi^{1/2})^d n \det(\mathrm{H})^{1/2}} \tag{3.18}$$

where, $L_{ij} = (X_i - X_j)^T \mathrm{H}^{-1}(X_i - X_j)$

In a seasonal time series $X=(x_0, x_1, \ldots, x_\tau, \ldots)$ for n years, for a lag-1 seasonal model corresponding to periodic autoregressive lag-1 model (PAR-1) of a parametric model, the KDE algorithm is implemented as:

i) Determine λ using Equation 3.17 or Equation 3.18 and H;

ii) Given an arbitrary data point $x_{\tau-1}$ at time $\tau-1$, determine the joint KDE for $X_{\tau,\tau-1}$ (Eq. 3.19) and the marginal KDE (Eq. 3.20) for $x_{\tau-1}$ from a n years seasonal series:

$$f(x_\tau, x_{\tau-1}) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{(2\pi)\lambda^2 \det(\mathrm{S})^{1/2}} \cdot \exp\left(-\frac{1}{2\lambda^2} \begin{bmatrix} (x_\tau - x_{i,\tau}) \\ (x_{\tau-1} - x_{i,\tau-1}) \end{bmatrix}^T \mathrm{S}^{-1} \begin{bmatrix} (x_\tau - x_{i,\tau}) \\ (x_{\tau-1} - x_{i,\tau-1}) \end{bmatrix} \right) \tag{3.19}$$

where, $\mathrm{S} = \begin{bmatrix} \mathrm{S}_{11} & \mathrm{S}_{12} \\ \mathrm{S}_{21} & \mathrm{S}_{22} \end{bmatrix}$

$$f(x_{\tau-1}) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{(2\pi\lambda^2 \mathrm{S}_{11})^{1/2}} \exp\left(-\frac{1}{2} \frac{(x_{\tau-1} - x_{i,\tau-1})^2}{\lambda^2 \mathrm{S}_{11}} \right) \tag{3.20}$$

Note that each observation contributes to the joint KDE with the distance from the observation $(x_{i,\tau}, x_{i,\tau-1})$ in year i to an arbitrary point $(x_\tau, x_{\tau-1})$, the scaling factor λ and the covariance S of $(x_\tau, x_{\tau-1})$;

iii) Given an arbitrary $x_{\tau-1}$, the conditional KDE is obtained in the form of a general univariate kernel probability density estimator of Equation 3.14 by substituting Equation 3.19 and Equation 3.20 into Equation 3.21(a):

$$f(x_\tau \mid x_{\tau-1}) = \frac{f(x_\tau, x_{\tau-1})}{f(x_{\tau-1})} \tag{3.21(a)}$$

$$f(x_\tau \mid x_{\tau-1}) = \sum_{i=1}^{n} \frac{1}{(2\pi\lambda^2 \mathrm{S}')^{1/2}} w_i \exp\left(-\frac{1}{2} \frac{(x_{\tau-1} - b_i)^2}{\lambda^2 \mathrm{S}'} \right) \tag{3.21(b)}$$

where:

$$w_i = \frac{\exp\left(-\frac{(x_{\tau-1} - x_{i,\tau-1})^2}{2\lambda^2 S_{22}}\right)}{\sum_{i=1}^{n} \exp\left(-\frac{(x_{\tau-1} - x_{i,\tau-1})^2}{2\lambda^2 S_{22}}\right)}$$

$$S' = \frac{\det(S)}{S_{22}} = S_{11} - \frac{S_{12}^2}{S_{22}}$$

$$b_i = x_{i,\tau} + (x_{\tau-1} - x_{i,\tau-1})\frac{S_{12}}{S_{22}}$$

Figure 3-5 illustrates a conditional KDE which is a slice of a bivariate joint KDE. Note that it is formed by summing the slices of n sub-joint KDEs. Each sub-conditional KDE has two parameters of b_i and $\lambda^2 S'$ representing the center and variance respectively. The weight w_i represents the area under each kernel slice which controls the contribution of $x_{i,\tau-1}$ to the conditional KDE, and is equal to the kernel function $K(\cdot)$ of the k-NN method;

iv) Sample a set of x_τ from $f(x_\tau|x_{\tau-1})$ given $x_{\tau-1}$ that is obtained from the marginal distribution in month τ-1 or from a historical record. In practice, one does not need to explicitly draw $f(x_\tau|x_{\tau-1})$ of Equation 3-21 since it is the sum of n Gaussian sub-conditional KDEs that contribute to $f(x_\tau|x_{\tau-1})$ as much as weight w_i. First, pick the i_{th} slice out of n sub-conditional KDEs with w_i corresponding to a random U from a Uniform distribution between 0 and 1; and then generate x_τ with mean b_i, variance $\lambda^2 S'$ and perturbation W in the m_{th} iteration using

$$x_\tau^{(m)} = b_i + \lambda (S')^{1/2} W_m \tag{3.22}$$

where $W_m \sim Normal(0,1)$. The statistics such as mean and variance can be inferred by iteratively sampling x_τ using Equation 3.22;

v) Simulation proceeds sequentially by season with updating of $x_{\tau-1}$. To do this, the conditional KDEs for each τ season should be built, and they are connected sequentially in time.

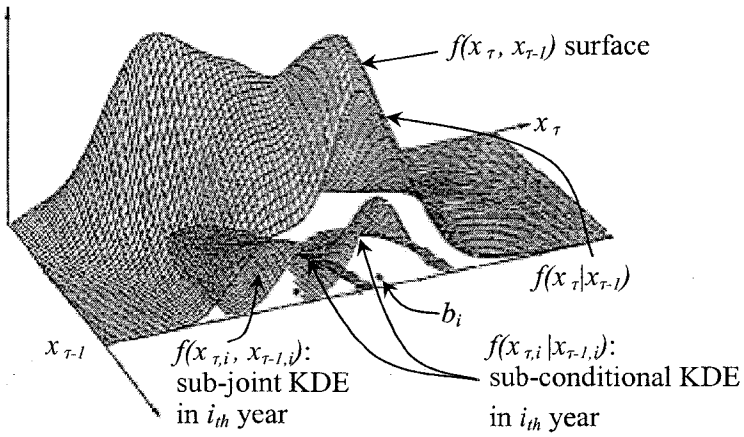


Figure 3-5. Illustration of conditional KDE *[Sharma, 1997]*.

3) Comparison of *k*-NN and KDE

The discrete kernel function K for k-NN and the weight w_i for KDE play the same role in sampling since picking the i_{th} sub-conditional KDE amounts to resampling S_i from k successors. However, the k-NN reproduces only historical data because it relies on a bootstrap resampling technique, while KDE adds the perturbation to the selected point. In the context of practical application of models, k-NN is preferred to KDE because: KDE is computationally difficult in the case of high order Markov chains such as daily forecast models, while; k-NN does not require any complicated computation process. Furthermore, the k-NN modeling can be applied to any high order Markov chain model.

3.4 Stochastic Artificial Neural Network Model

3.4.1 Fundamentals of Artificial Neural Networks

An artificial neural network is a nonlinear mathematical structure that is capable of representing arbitrarily complex nonlinear processes that relate the inputs and outputs of any system *[Hsu et al., 1995]*. Artificial neural networks (ANNs) have been used successfully to model complex nonlinear input-output relationships in a wide variety of fields. In recent years, ANN have become extremely popular for simulation and forecasting in water resources and environmental planning and management, and mainly treated in deterministic way *[Bhattacharya and Solomatine, 2000; Hsu et al., 1995; Maier and Dandy, 2000]*.

1) Architectures of Neural Networks

Many researchers have described the structures for ANNs using artificial neurons that are information processing units variously referred to as 'processing elements (PEs)', 'units', 'cells', or 'nodes' *[Haykin, 1993]*. In the 1950's, Rosenblatt's work resulted in a two-layer network, the perceptron, which was capable of learning certain classifications by adjusting connection weights. Recent work includes the three types of network architectures: the feedforward neural networks (FFN); the feedback (or recurrent) neural networks (FBN), and; the stochastic neural networks.

The FFN is a neural network where the connections between the units do not form a directed cycle, while the FBN makes bi-directional data flow possible. The multilayer perceptron (MLP) and radial basis function networks, and Hopfield networks are representative in the FNN and FBN respectively. The stochastic neural networks

(SNN) are built by introducing random variations into a network, either by giving the network's neurons stochastic transfer functions, or by giving them stochastic weights. This makes them useful tools for optimization problems since the random fluctuations help it escape from local minimums. Stochastic neural networks that are built by using stochastic transfer functions are often called Boltzmann machines (*http://www.answers.com/topic/neural-network*). From reviewing the real applications of ANN, it is indicated that MLP networks has been mostly used for prediction and forecasting of water resources characteristics *[Maier and Dandy, 2000]*.

2) Structure of Multilayer perceptron (MLP)

As shown in Figure 3-6, the PEs are arranged in three layers in MLPs: an input layer, one or more hidden layers, and an output layer. The input x from each PE in the former layer is multiplied by a connection weight w. These connection weights are adjustable and may be likened to the coefficients in statistical models. At each PE, the weighted input signals are summed and the threshold values of β and γ are added. This combined input I is then passed through a nonlinear transfer function $f(.)$ to produce the output of the PE y. The output of one PE provides the input to the PEs in the next layer. This process is represented as:

$I_j = \beta_j + \sum_{i=1}^{n} w_j^{(i)} x_t^{(i)}$: summation function;

$y_j = f(I_j)$: transfer function;

where:

I_j = activation level of node j;
$w_j^{(i)}$ = connection weight between node i and j;
$x_t^{(i)}$ = input from node I *at time* t, i=0,1,…,n;

β_j = bias or threshold for node j;

y_j = output of node j, and;

$f(.)$ = transfer (or activation) function.

Figure 3-6 illustrates an example of ANN with 2 input nodes (n), 2 hidden nodes (h), and 3 output nodes (m). An ANN with a hyperbolic tangent transfer function for the hidden layer and a linear transfer function for the output layer is expressed mathematically in Equation 3.23. For determining the best relationship between inputs and outputs, a set of the observed data are generally compared to the simulated values, and the error is computed. The inputs and outputs are mapped optimally in a network by adjusting the weights and biases until the global minimum in the error surface has been reached. This procedure is referred to as 'learning' or 'training' *[Mathworks, neural network toolbox user's guide]* and corresponds to 'parameter calibration' in deterministic conceptual models.

3.4.2 Training, Validation and Stochastic ANNs

1) Training (or Learning) :

Learning in ANNs falls under two categories: supervised and unsupervised. In the supervised learning, a network is modeled with a set of historical inputs and outputs. In the unsupervised learning, a network is presented only with inputs without outputs: hence, the network itself adjusts the parameters by clustering the inputs into classes of similar features *[Mathworks-Neural]*.

$$\tilde{y}_t^{(k)} = \gamma_k + \sum_{j=1}^{h} \alpha_{jk} \tanh\left[\beta_j + \sum_{i=1}^{n} w_j^{(i)} x_t^{(i)}\right], \quad k = 1,..,m \tag{3.23}$$

where, $\tanh(z) = \dfrac{e^z - e^{-z}}{e^z + e^{-z}}$

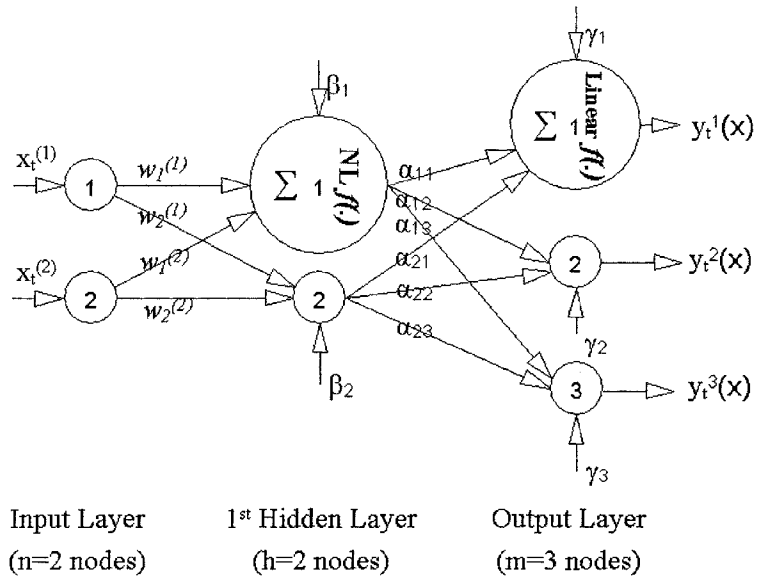


Figure 3-6 Typical structure of MLP. Where, NL*f(·)* is a nonlinear function.

The final purpose of training is to make predictions of the future with only inputs given to a network. The task of training is not always straightforward but may be complicated by the existence of local minima in the solution surface or the potential of overfitting the training data.

i) Generalization of ANNs : issue of overtraining

For training of a network, a set of training data that represents all the possible relationships between inputs and outputs should be prepared and loaded into a network. Once trained, the network should be able to predict using the input vectors that are not in the training data set, and be tolerant to noisy data to some extent. A network is said to be generalized satisfactorily when a trained network is reproduced properly within a predefined tolerance with the test data never used in training a network. On the contrary, it is overtrained or individualized only if it responds to the training data. This often takes

place when the training data are not well balanced so that they tend to contain the patterns concentrated on one part of the input domain range. To avoid overtraining, the training data set should be examined carefully and filtered through a preliminary analysis. For guaranteeing a robust generalization of a network, *Neal [1992]* suggested employing Bayesian stochastic learning. The methods for improving generalization are summarized:

- Tradeoff between the number of weights and training samples

If the number of parameters in the network is much smaller than the size of training set, then there is little or no chance of overtraining, and vice versa. *Amari et al. [1997]* show that overfitting does not occur if the ratio exceeds 30. When this condition is not met, there are clear benefits in using cross-validation.

- Cross-validation *[The Mathworks-Neural]*

In this method, the available data are divided into three subsets: training, test and validation data sets. The errors on the test set are monitored during the training process. The errors on both the training and test data sets will normally decrease during the initial phases. However, when a network begins to overfit the data, the errors on the test set begin to rise. When the test errors increase for a specified number of iterations, the training is stopped, and the parameters at the minimum of the test errors are returned.

- Regularization *[The Mathworks-Neural]*

One approach to this process is to use the Bayesian framework. One feature of this algorithm is that it provides a measure of how many network parameters are being effectively used. The effective number of parameters is approximately converged to for a sufficient number of iterations.

ii) **deterministic learning : issue of local minima**

The most common training algorithm is the backpropagation (BP) algorithm, which is a supervised learning algorithm for the FFN. It is basically a local optimization method on the basis of the steepest gradient method and conjugate gradient method. It is susceptible to be trapped in local minima in the error surface. Global optimization techniques such as genetic algorithms and the shuffled complex evolutionary algorithm (SCE) developed by *Duan et al. [1992]* are often used to escape from the local minima as they employ random search techniques. Although global optimization techniques are often more computationally intensive than local optimization techniques, they are gaining popularity with the help of advanced computer science and engineering.

2) Stochastic learning based on a Bayesian framework: issue of both overtraining and local minima

Conventional learning such as BP for the FFN can be interpreted in statistical terms such as the maximum likelihood estimation (MLE). The idea of the conventional learning is to find a single set of parameters for the network that maximize the fit to the training data. In a Bayesian framework, however, the parameters are assumed to be random variables with specified distributions. As benefits of the Bayesian approach, the avoidance of overtraining and the acquisition of the probability distributions for the predicted variables are pointed out *[Mackay, 1992; Neal, 1992]*.

3) Validation

Once the training is completed, the performance of a trained network has to be validated on an independent data set. If the validation errors are markedly significant, it is likely that either training and validation data set may not be representative out of the population, or the network is overfitted.

3.5 Fuzzy Rule-based Modeling

A common approach to describe uncertainty is to employ probability theory. Another method is the use of fuzzy logic. Fuzzy rule-based modeling has been successfully applied in fields such as automatic control, data classification, decision analysis *[Bijaya et al., 1996; Bhattacharya and Solomatine, 2000]* and expert systems. Fuzzy rule-based modeling has a number of different names such as fuzzy inference systems (FIS), fuzzy expert systems, fuzzy modeling, fuzzy logic controllers, and simply fuzzy systems *[Mathworks, Fuzzy Logic Toolbox User's Guide]*. Fuzzy rule-based methods are often compared and combined with neural networks since they can model the nonlinear system behaviors. However, in contrast to ANN that is a complete black box type, fuzzy rule-based modeling is very transparent as the rules are explicitly stated *[Bardossy and Duckstein, 1995]*.

3.5.1 Fundamentals of Fuzzy Rule-based Modeling

Fuzzy logic starts with the concept of a fuzzy set, which is associated with fuzzy numbers, membership, membership functions. A classical set is a container that wholly includes or excludes any given element. A fuzzy set has clear boundaries however it contains elements with only a partial degree of membership *[Bardossy and Duckstein, 1995]*. A fuzzy set can be represented in a mathematical terms as:

$$A = \{(x, \mu_A(x)); x \in X, \mu_A(x) \in [0,1]\}$$

where:

A = a fuzzy set of X;

$\mu_A(x)$ = membership function with a range of 0 to 1;

X = a set (universe), and;

x = membership.

The membership function, $\mu_A(x)$ represents the grade (or likeliness or possibility) of membership x in A. The closer $\mu_A(x)$ is to 1, the more x is considered to belong to A, and vice versa. Special cases of fuzzy sets are fuzzy numbers, which are fuzzy sets of real numbers *[Bardossy and Duckstein, 1995]*:

Fuzzy rules are represented in the form of '*if-then*' statement. The if-part of a rule is called the *premise* or *antecedent* and then-part is called the *conclusion* or *consequence*. A premise or consequence in a rule may have several clauses which relate to the input and output variables respectively:

$$\text{IF } X_1 \text{ is } a_1 \in A_{i,1} \ \Theta \ldots \Theta \ X_K \text{ is } a_K \in A_{i,K}$$
$$\text{THEN } Y_1 \text{ is } b_1 \in B_{i,1} \ \Theta \ldots \Theta \ Y_M \text{ is } b_M \in B_{i,M} \ : i^{\text{th}} \text{ rule}$$

Where:

X_k, a_k = input variable, value for the k^{th} clause in the premise of a rule, $k=1,..,K$;

Y_m, b_m = output variable, value for the consequence of a rule, $m=1,..,M$;

$A_{i,k}$ = a fuzzy set of X_k in the i^{th} rule;

$B_{i,m}$ = a fuzzy set of Y_m in the i^{th} rule;

Θ = fuzzy logic operator.

A fuzzy rule system $\Re$ is defined as a set of rules. For example, in a reservoir operation system, the premises may be composed of reservoir elevation, inflows, losses like evaporation and seepage, and water demands. The consequence is likely the releases to meet the various targets *[Bijaya et al., 1996]*. For another example, in stage-discharge relationships in streams, the input variables may be composed of stage at time t and variation rate of stage at time t, and the output is discharge at time t. A fuzzy rule system

consists of a variety of rules where all input and output are combined on the basis of expert knowledge. This is illustrated in Chapter 4.

3.5.2 Procedure of Fuzzy Rule-Based Modeling

1) Modeling Procedure by a FIS

A FIS can be represented by the two types of structures: Mamdany-type and Sugeno-type. The main difference is that the Sugeno output membership functions are either linear or constant, while this is not the case in the Mamdany type *[Mathworks-Fuzzy]*. The procedures of fuzzy rule-based modeling are:

i) generating a FIS structure: determination of the input-output nodes, the number and types of input-output membership functions, the number and structure of rules, and; selection of fuzzy and defuzzification operators;

ii) training the generated FIS: modification of expert knowledge or extraction of rules from observations, and;

iii) evaluating the trained FIS with validation data.

2) Operation of FIS

Once a FIS has been generated, the FIS is often operated with five processes in order *[The Mathworks-Fuzzy]*: 1) fuzzification of the input variables, 2) computation of the degree of fulfillment (DOF), 3) implication of the DOF and fuzzy set in each rule, 4) combination of all rule consequences, and 5) defuzzification.

Step 1) Fuzzification of Input Variables: k^{th} clause in the i^{th} rule

An input variable that is one of memberships in a fuzzy set is always fed into the membership function in the form of a crisp numerical value. The output of each membership function is a fuzzy degree of membership, and always has a range of 0 to 1. This process amounts to either a table lookup or function evaluation:

- input: crisp numerical value (a);
- output: a value between 0 and 1 of the k^{th} clause in the i^{th} rule ($\mu_{Ak,i}(a_k)$).

Step 2) Computation of Degree of Fulfillment (DOF): all clauses in the i^{th} rule

Once the inputs have been fuzzified, if a system is multivariate, the extent to which each clause is attributed in a rule is integrated by the DOF of a rule. The inputs for this step are the fuzzified input crisp values from the 1st step, and the output is a single value v_i in the i^{th} rule. There are two common fuzzy operators in determining the DOF:

- input : fuzzified input values = $\{\mu_{Ai,1}(a_1), \ldots, \mu_{Ai,k}(a_k)\}$;
- output : a value between 0 and 1, $v_i = \mu_{Ai,1}(a_1) \Theta \ldots \Theta \mu_{Ai,k}(a_k)$;
- product fuzzy operator:
  AND $(a1,a2)$: $\mu_{Ai,1}(a_1)*\mu_{Ai,2}(a_2)$;
  OR $(a1,a2)$: $\mu_{Ai,1}(a_1)+\mu_{Ai,2}(a_2)- \mu_{Ai,1}(a_1)*\mu_{Ai,2}(a_2)$;
- min-max inference fuzzy operator:
  AND $(a1,a2)$: $\min\{ \mu_{Ai,1}(a_1), \mu_{Ai,2}(a_2)\}$;
  OR $(a1,a2)$: $\max\{ \mu_{Ai,1}(a_1), \mu_{Ai,2}(a_2)\}$.

Step 3) Implication of the DOF and fuzzy set in a rule: consequence in the i^{th} rule

A consequence is a fuzzy set represented by a membership function. The membership function ($\mu_{i,B}(x)$) of a consequence in the i^{th} rule is reshaped by the DOF obtained from the 2nd step using two common fuzzy operators:

- input : v_i as DOF;
- output : $\mu'_{i,B}(x)$, a reshaped consequence fuzzy set = $v_i \Theta \mu_{i,B}(x)$;

- AND : $\min(v_i, \mu_{i,B}(x))$;
- PROD : $v_i * \mu_{i,B}(x)$.

Step 4) Combination of all consequences: combined consequence of all rules

Since model results are based on testing all of the rules, the rules must be combined. In this step, the reshaped consequence fuzzy sets in all rules obtained from the 3rd step are combined together. A combined fuzzy set is generated as an output by using the following common fuzzy operators:

- input : $\mu'_{i,B}(x)$, the reshaped consequence fuzzy set;
- output : $\mu_B(x)$, a fuzzy set of combined consequences;
- sum : $\sum \mu'_{i,B}(x)$, $i=1..I$;
- weighted sum: $\mu_B(x) = \dfrac{\sum_{i=1}^{I} v_i \mu'_{i,B}(x)}{\max_u \sum_{i=1}^{I} v_i \mu'_{i,B}(u)}$.

Step 5) Defuzzification

Defuzzification is the process to determine a represented crisp value for the combined fuzzy set that is obtained in the 4th step. This process is equivalent to estimation of mean, median or mode in probability distributions if the combined fuzzy set can be considered a sort of probability distribution. However, the combined fuzzy set is often such an irregular shape that it may be very difficult to infer parameters like confidence intervals and percentiles. Therefore, it is likely that a probabilistic modeling by a FIS is either almost impossible or computationally very demanding. The defuzzification is performed using the fuzzy means of $\mu'_{i,B}(x)$ and the normed weighted sum method. All the processes are illustrated in Chapter 4.

- input : $M(B_i)$, fuzzy mean of $\mu'_{i,B}(x)$ in the i^{th} rule;

- output : *M(B)*, an crisp result as an output for the system;
- Normed weighted sum: $M(B) = \dfrac{\sum_{i=1}^{I} v_i M(B_i)}{\sum_{i=1}^{I} v_i}$.

3.5.3 Derivation of Fuzzy Rules from Data and Training (Learning)

From the review of the FIS modeling procedure, it can be noticed that a Fuzzy rule-based modeling is affected by how well the rule system reflects expert knowledge and information from the observations. The different methods to derive a rule system are well summarized in *Bardossy and Duckstein [1995]*:

i) The rules are derived by the experts directly;

ii) The rules can be assessed by experts directly, but should be updated using data;

iii) The rules are not known explicitly, but the variables can be specified by experts;

iv) Only a set of observed inputs and outputs is available, and a rule system has to be constructed using the data set.

In real application of a fuzzy rule-based system in water resources and environmental planning and management, the last method might be encountered most commonly because the rules are more objective than in the other methods. However, this method requires very special care in system learning in case the observations have uncertainties to a significant extent. Although the first method is applicable to a system, it is very difficult to elicit the technical knowledge from experts without observations. Furthermore, the knowledge from many experts may not be consistent and the manually derived rules may lack numerical precision *[Abebe et al., 2000]*. The last method can be

also used for replication of simulation or optimization models. In this case, the outputs may be commonly obtained by a variety of models from simple regression models to very complicated mathematical models. For example, the results by an optimization model for reservoir operation can be replicated by a FIS, and the classical rules can be replaced. Compared to the first three methods, one thing that should be noted is that the number of rules is equal to the number of combinations for the membership functions of all input variables, while they are determined arbitrarily by the experts in the first three methods. The popular algorithms for the last method are the counting, weighted counting and optimization techniques *[Bardossy and Duckstein, 1995]*.

The fuzzy rule-based modeling is very similar to ANN in terms of the basic idea of data-driven modeling. Compared with ANN, the biggest advantages of the fuzzy rule-based modeling are that the knowledge from the experts can be added along with the rules abstracted from the data, and the rules from the data can be also modified by users.

Once FIS has been built by one of the three methods, it is assessed mainly in terms of the performance and overfitting. However, an optimized FIS by the last algorithm does not have to be trained again because it has been already trained in the optimization process. As reviewed in ANN, the cross-validation method can be used here to avoid the risk of overtraining. After training, the performance of the trained network has to be evaluated on an independent data set.

CHAPTER 4

RELIABILITY ANALYSIS OF DISCHARGE RATINGS USING PROBABILISTIC AND DETERMINISTIC MODELING

Water resources engineers are often concerned with the conversion of discharges at a given location along a stream into corresponding stages or converting measured stages into discharge estimates. This is accomplished via a rating that is the relationship between stage and discharge. One purpose of stream flow measurement is to establish ratings, and generally to produce continuous discharge records over time using measured stages. The generated discharges are mainly used for sustainable development and management for flood control, water use and environmental conservation. However, the biggest problem with ratings is that considerable uncertainty can exist due to the complexity of stage-discharge relationships and measurement errors of discharge.

Research dealing with the stage-discharge relationships is fairly limited. As a plausible reason, water resources researchers and engineers might think that the research for this area is very narrow, or not challenging enough to investigate. In this chapter, the principles of rating are reviewed first according to the guide *[ISO, 1998]* prepared by the

International Organization for Standardization (ISO). Second, ratings are developed using deterministic and probabilistic methods through a case study. For deterministic approaches, non-linear programming, fuzzy rule-based modeling, and a one-dimensional hydrodynamic model are applied. For the probabilistic approach, a Bayesian Markov chain Monte Carlo sampling technique is used. Finally, the methodologies are evaluated and compared in terms of suitability for adaptive management of ratings, consideration of uncertainty in the rating, reflection of hydraulic characteristics of the rating including hysteresis, and ease of application. The goodness-of-fit of rating curve is not considered as a definitive measure because the method that gives the smallest estimation error is not always the best due to uncertainties in the measured stage-discharge data.

The study was carried out at two stage-discharge measurement sites located near the Daecheong and Yongdam dams in the Geum River basin in Korea. The Donghyang station located in a small tributary of the Geum River where the hysteresis of the rating is not considerable (Figure 2-2), and the Hotan station, located on the main stream of the Geum River, where a loop shaped rating clearly exists (Figure 2-3). These two stations are operated to provide information for reservoir operation. Figure 4-1 shows the procedure for this study, and the relationships between the deterministic and probabilistic approaches.

4.1 Concepts of the Stage-Discharge Relation

4.1.1 Control, Complexity, Shift and Uncertainty of Ratings

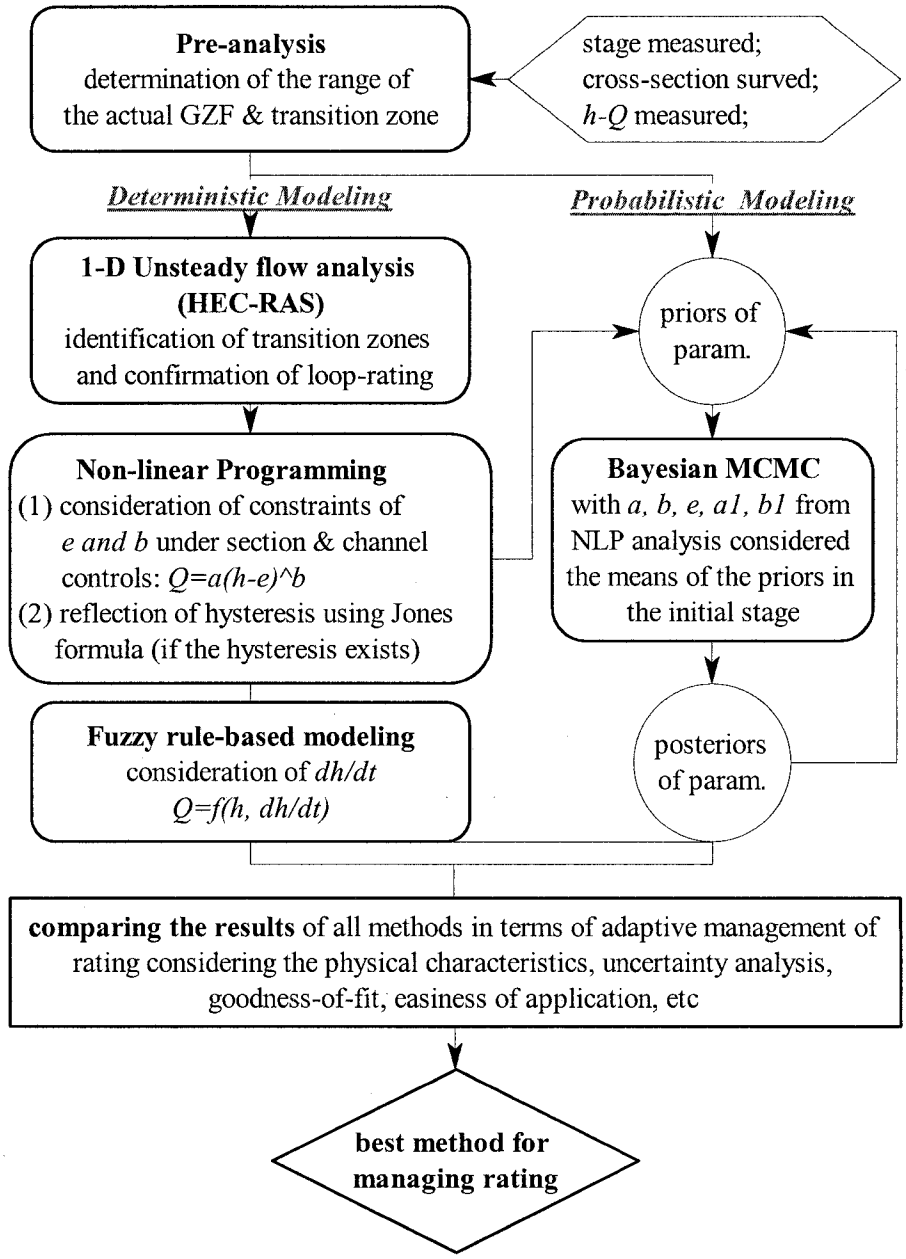


Figure 4-1 The procedure for rating analysis, and connection between the deterministic and probabilistic approaches.

1) Controls of flows in open channels

The ratings in open channels are governed by channel conditions in a reach located downstream from a stage-discharge gauging station. Depending on channel and flow conditions, the two types of control can be considered: section control and channel control. The ratings for low flow are generally governed by section control: hence, they are controlled by a particular downstream section such as naturally formed sand bars, accumulation of debris and rock ledge, and manmade structures like weirs. The section control is visually identified in the field by observing a riffle or a drop in water surface downward from a gauging station. As stage goes down, the flow eventually reduces to zero at a certain stage, which is referred to as the gauge height of zero flow (GZF). Figure 4-2 and 4-3 show a good example of section control. While the subcritical flow at reach 'A' changes to critical flow at 'B' and supercritical flow right after 'B' in turn, and returns to subcritical flow at 'C', the rating at reach 'A' is dominated by the section control formed by the rock ledge 'B'.

On the other hand, as the section control is submerged with increase of stage, a rating for high flow is usually controlled by a particular channel reach that is represented by channel curvature, roughness, slope and shape. A rating for medium flow may be controlled by either of the two controls, or a combination of both controls. Especially, the section where both controls exist is called a transition zone that represents the change from section control to channel control or vice versa *[ISO, 1998]*. In most stage-discharge stations, both section and channel controls are generally at work. If the two controls are identified clearly, it is evident that a transition zone should exist somewhere in between the two controls, whatever the bandwidth of the transition zone is. From this viewpoint, when a rating is developed, it should be separated according to the controls. If

the bandwidth of transition zone is large enough not to be negligible, a rating can be grouped into three parts corresponding to low, medium and high flows.

2) Complexities of stage-discharge relations

Complexity in a rating relationship is usually caused by backwater effect and hysteresis. Backwater arises due to obstructions of water flow by downstream manmade structures, tributaries and tides. This can disturb a rating at an upstream stage-discharge measurement site. Another complexity is hysteresis which results from the variation in energy slope due to flow accelerations. This is well known as a loop rating where each discharge matches two different stages rather than a single, unique stage-discharge relation *[ISO, 1998]*. It is most apparent that the loop rating arises in mild slope streams where dynamic flows are accelerated due to suddenly changing discharges in time, such as flood runoff and reservoir releases.

3) Shift of controls in stage-discharge relations

The shift of controls arises when channel conditions are unstable due to scour and deposition during flood events *[ISO, 1998]*. The shift can be temporary or permanent. If the shift is permanent, a group of consecutive stage-discharge measurements will deviate to the left from a current rating in the case of deposition, and to the right in the case of scour. When a shift is indicated, a new rating should be developed.

4) Uncertainties contained in the generated discharges

The final purpose of ratings is to compute a series of flow over time based on continuously measured stages. Taking a close look at the uncertainties of discharges, it is obvious that they are directly affected by the uncertainties of ratings and the measured stages, and the stage-discharge measurement data have a great influence on the uncertainties of the ratings. The errors in the stage-discharge measurement data are

propagated to the computed discharges. Furthermore, any statistical errors in establishing ratings will input the discharges. From this viewpoint, it may be worthwhile to produce discharges as probability distributions. When the information of uncertainties contained in discharges is offered to the water resources managers, this facilitates the management of water resources on the basis of risk. Figure 4-4 illustrates the process until the final product, discharges are obtained, and the propagation and accumulation of the errors into the discharges.

4.1.2 Hydraulic Characteristics of Ratings

Ratings can be represented as a function with some parameters reflecting the two types of controls. The ratings under section control can be expressed using a general equation for a weir or a flume, while the governing equations for the rating under channel control are much more complicated. For gradually varied uniform flow, the ratings can be simply expressed by the Manning equation, and for highly varied non-uniform and unsteady flow, equations such as the Saint-Venant equations would be appropriate. However, these are seldom used in the development of ratings. Instead, a type of power equation (Eq. 4.1 and log-transformed Eq. 4.2) is generally used because a rating curve segment for a given control tends to show a straight line when plotting logarithms of discharge versus effective depth. This fact gives analysts much valuable knowledge for shaping a rating curve and analytical purposes *[ISO, 1998]*.

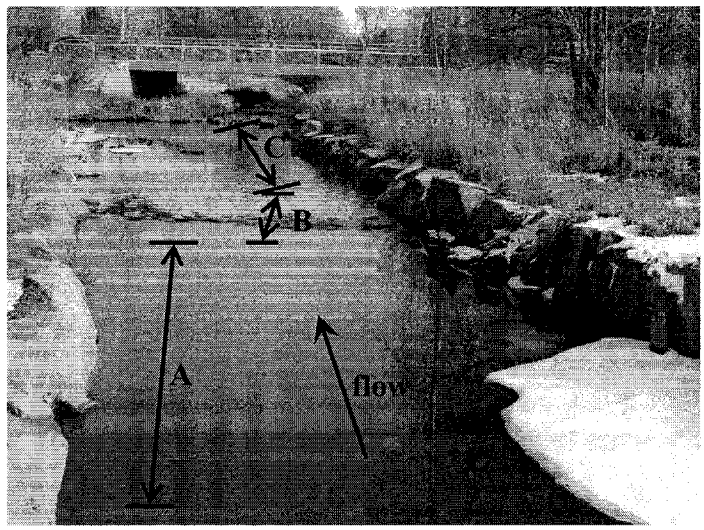


Figure 4-2 It shows 'A' reach under the section control formed by the rock ledge (B) in Spring creek in Fort Collins, Colorado, taken in December 2005.

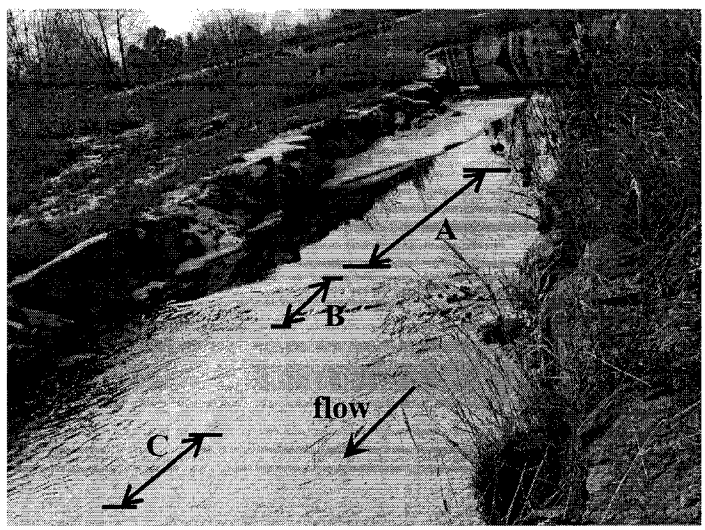


Figure 4-3 It shows the same reach taken in upstream direction as in Figure 4-2.

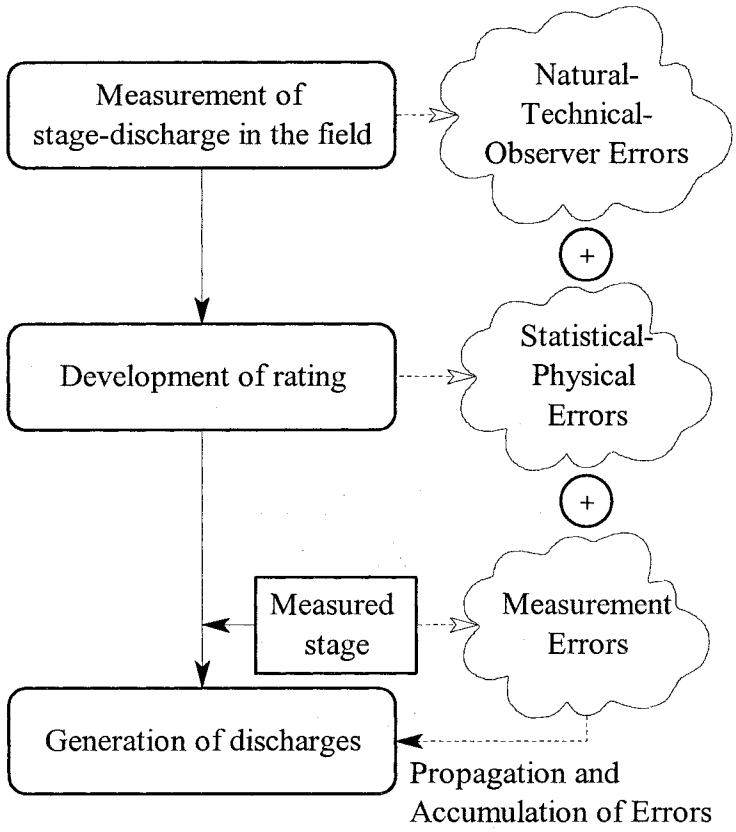


Figure 4-4 The propagation and accumulation of the errors into the generated discharges.

$$Q = a(h-e)^b \tag{4.1}$$

$$Log(Q) = b \cdot Log(h-e) + Log(a) \tag{4.2}$$

Where:

h is the stage;

e is the gauge height of zero flow;

$(h-e)$ is the effective water depth on a control section;

b is the slope of the rating curve;

a is a constant that is numerically equal to the discharge when $(h-e)$ is equal to 1.

To utilize this procedure, first stage should be transformed to effective depth by subtracting a predetermined GZF. The slope of a rating curve on a logarithmic scale is defined as the ratio ($\Delta Q/\Delta h$) of the horizontal distance to the vertical distance because the dependent variable, discharge, is always plotted as the abscissa of a rating curve *[ISO, 1998]*. In Equation 4.2, the above mentioned three parameters of a, b and e have their own hydraulic meaning, and they are characterized by the spatially unique hydraulic features at every stage-discharge measurement site. They are very susceptible to floods that cause channels to be unstable. Especially, b and e are very important for shaping and analyzing a rating curve, and can be obtained as follows:

1) Determination of the slope (*b*)

According to the ISO standard for developing rating curves *[ISO, 1998]*, the slope of a rating curve for section control is usually greater than or equal to 2.0, while the slope for channel control is between 1.5 and 2.0. Figure 4-5 shows that a rating is divided into the two segments of both section and channel controls centering on a transition zone.

2) Determination of GZF (*e*)

The GZF can be determined directly in the field by measuring the flow depth on the control section. However, this is not always the case because it may be hard to identify a natural control section if a man-made structure does not exist. In general, three methods used for determining the GZF are:

i) Determination of the GZF by field survey

The GZF can be obtained by subtracting the depth at the deepest point and the velocity head at the control section from the stage at a station in case the control section exists downstream from a stage-discharge station. The GZF by this method is referred to as the actual gauge height of zero flow *[ISO, 1998]*;

ii) Determination of the GZF in graphical way

With at least three data points measured accurately under section control plotted on logarithmic paper, the GZF is determined by adjusting the GZF until the line connecting three points becomes visually straight. The GZF by this method is referred to as the effective gauge height of zero flow. For irregular shaped channels, the effective GZF is greater than the actual GZF *[ISO, 1998]*;

iii) Determination of the GZF by optimization techniques

This method is almost identical to the graphical method except for using optimization techniques. The GZF can be optimized with an objective function subject to the slope range of each control, which is obtained in a field survey. This method is described in detail in the following section about determination of a rating using a non-linear programming technique.

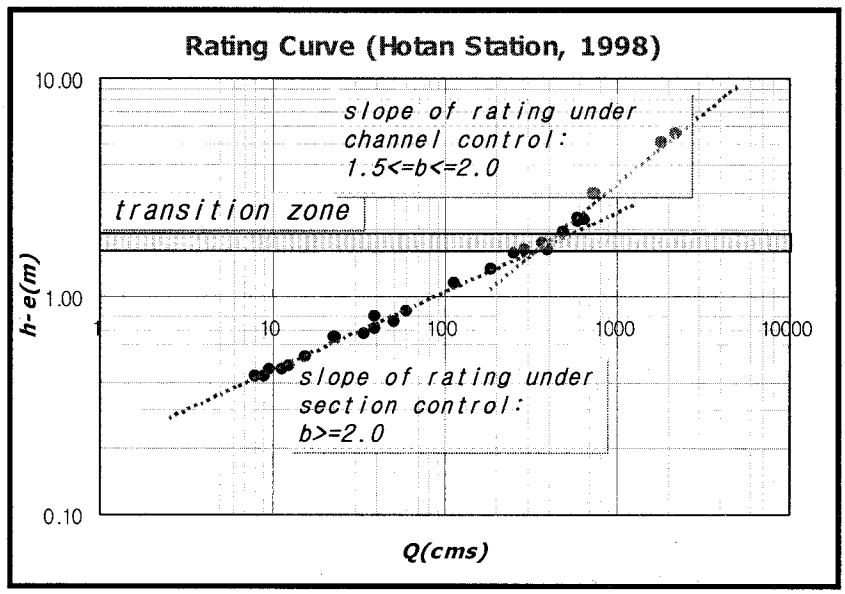


Figure 4-5 An example of slope change in a rating: When the effective depth *(h-e)* and discharge *(Q)* in natural open channels are plotted on the logarithmic paper, each rating curve segment by control is a straight line, and each slope has its own unique range.

4.2 Rating Analysis in Deterministic Approach Using a 1-D Unsteady Flow Analysis, NLP and Fuzzy Rule-based Modeling

4.2.1 Rating Analysis Using a 1-D Unsteady Flow Analysis

Once a set of physical data such as cross sections in reaches and hydrologic data are prepared in a study area, a rating can be developed indirectly by hydraulic modeling instead of using the directly measured *h-Q* data. Even though this method is considered to be very attractive in that ratings can be derived without stage-discharge measurements in the field, which is very hard work, this approach may be limited to the analysis mainly for channel control. This is because the results of numerical models such as the *HEC-RAS* (US Army Corps of Engineers) and *FLDWAV* (US National Weather Service) do not often converge to a unique solution or have numerical errors when modeling low flows. To obtain a complete rating using this method, a rating under section control may be analyzed in other ways. This fact could be pointed out as a drawback, however, this method provides very valuable information in shaping a rating curve considering hysteresis and measurement errors. Some strong points of this method can be summarized as:

- When the measured stage-discharge data are plotted on a graph, it is hard to separate hysteresis from measurement errors. The results from hydraulic models help screen the stage-discharge measurement data with very low reliability on the basis of the information about the hysteresis;
- When one wants to extend an existing rating with new values of observed stages beyond the existing range, this method provides a very reliable solution.

In this study a one-dimensional hydraulic model, *HEC-RAS*, was employed at the Donghyang and Hotan stations. The *HEC-RAS* system (ver. 3.1.3) now contains the two one-dimensional hydraulic analysis components for steady flow water surface profile computations and unsteady flow simulation. The hydraulic analyses of ratings were performed according to the successive processes shown in Figure 4-6. They consist of setting up *HEC-RAS* with a set of cross section data and boundary and initial conditions, calibrating the model by adjusting the Manning's n, verifying the calibrated model, and evaluating a developed rating with measured stage-discharge data.

1) Hydraulic analysis of the rating at the Donghyang station

The stream reach under study was modeled in *HEC-RAS* (Figure 4-7). The figure shows the Donghyang station, and two supplementary stage stations at Daeya-kyo and Somti-bo working as internal, upstream and downstream boundaries, respectively. The reach was schematized with 5 cross sections within a 1km long sub-reach upstream and 15 cross sections within a 3km long sub-reach downstream centering on the Donghyang station. These cross sections were surveyed at 200m interval in 2003.

For the upper (Daeya-kyo) and lower (Somti-bo) boundary conditions of the reach, the recorded stages were entered into *HEC-RAS*. For calibration, the computed stages at the Donghyang station were compared with the historical stages. The historical stages available at this station and the two supplementary stations were prepared during two flood events, which occurred on August 18, 2003 for 4 days and September 11, 2003 for 5 days (Figure 4-8 and 4-9). The former flood data were applied to calibrate the model and the latter to verify the model.

When the models were calibrated, the computed stages of the Donghyang station were compared with the historical stages. For the boundary conditions of the model, the historical stages observed at the two ends of the reach were used. The computed stages were consistent with the observed stages; however, the computed rating did not fit well to the measured stage-discharge data. Hence, the upper boundary condition was modified with the discharges computed by the rating, which was derived in a previous study *[K-water, 2004]*. For calibration in this case, a pair of stage-discharge measurement data were compared with the computed rating instead of comparison of the observed and computed stages. Figure 4-10 shows that the observed stage-discharge data were fitted well with Manning's n value of 0.031 except for the low to medium flows (approximately less than $20m^3/s$). It can be noticed that the rating needs to be separated into two segments around the height of the *Parshall* flume. The considerable deviation between the computed and the observed rating in the lower segment under section control again illustrates that section control is hard to analyze by unsteady flow models.

The calibrated model was verified with another flood event. Figure 4-11 shows that the computed rating from this step is consistent with the ratings obtained from the calibration process. In conclusion, two important facts can be drawn from the results for developing a rating at this station:

i) A loop rating does not exist, which means that a single relation of stage-discharge is good enough at this station;

ii) The governing control of the rating is shifted from section control to channel control within the range of 2.4m to 2.9m in stage (Figure 4-11). This can be explained by the fact that the computed rating fits well the observed stage-

discharge over the stage of 2.9m, and the rating below 2.4m is changed sharply. This result indicates that a transition zone is formed in this range: hence, the information for the transition zone was developed for the rating analysis by optimization and Bayesian MCMC techniques.

2) Hydraulic analysis of the rating at the Hotan station

Like the Donghyang station, the stream reach centering on the Hotan station was configured in the *HEC-RAS* with 5 cross sections upstream and 15 cross sections downstream from the Hotan station, which were surveyed in 2004 (Figure 4-12). For hydraulic analysis at this site, two flood events were prepared, which occurred on June 19, 2003 for 8 days and July 12, 2003 for 10 days (Figure 4-13 and 4-14). The first flood was used to calibrate the model and the second to verify the model.

Unlike the Donghyang station, the lower boundary condition was given in the form of a rating instead of as historical stages because the lower boundary condition did not exist firmly. First, a non-uniform flow analysis was performed to come up with a rating for the lower boundary condition using a series of arbitrary discharges as the upper boundary condition and the normal or critical depths as the lower boundary condition. One may think that the rating at the station could be influenced by the different conditions of the lower boundary; however it turned out that the results were not very sensitive to the lower boundary conditions. Next, an unsteady flow analysis was done using the upper boundary condition of the discharges generated by the rating developed in a previous study *[K-water, 2004]* and the lower boundary condition of the rating derived from the non-uniform flow analysis in the previous step.

When the model was calibrated, the computed rating was compared with the observed stage-discharge data similar to the Donghyang station. Figure 4-15 shows that the model was well calibrated with a Manning's n value of 0.046 except for flows less than approximately 200m^3/s. The figure shows clearly the hysteresis of the rating, which varies according to the sign of the variation rate of the stage. There are the two distinctive things that should be noted at this station: hysteresis in rating exists and; the rating is governed by three controls (Figure 4-16). The shift of controls can be explained by one of two reasons: the existence of section or channel controls downstream, and the sudden variation at the cross section of the gauging station. At the Hotan station, weirs for agricultural water withdrawal or naturally formed control sections may account for the two transition zones.

The verification of the calibrated model was performed with the verification data set (Figure 4-16), which shows the calibrated model works well. In conclusion, the two following facts should be taken into consideration in developing the rating at this station:

i) The loop rating does exist clearly above 300m^3/s in flow and 2m in stage: hence, the specially designed rating should be developed using a simple equation, sophisticated 1-D unsteady models or heuristic methods such as ANN and fuzzy rule-based modeling;

ii) The first transition zone takes place within a range of 0.6m to 0.7m in stage, and 65m^3/s in discharge; and the second one occurs within a range of 1.3m to 1.4m in stage and 150m^3/s in discharge (Figure 4-16).

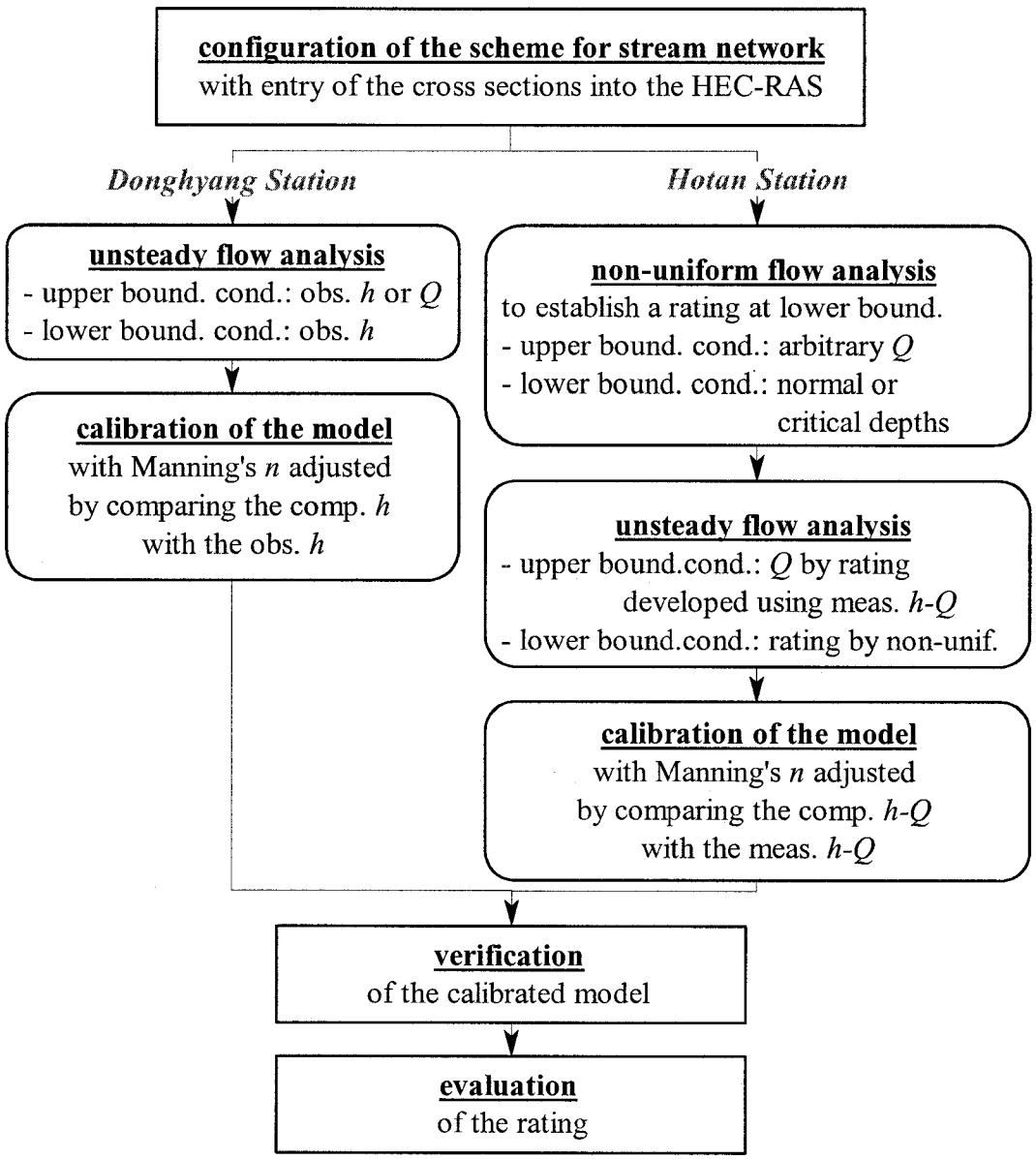


Figure 4-6 The procedure of the hydraulic analysis of the ratings at the Donghyang and Hotan stations, where "cond.", "comp.", "obs.", "meas.", "bound" and "non-unif" stand for condition, computed, observed, measured, boundary and non-uniform respectively.

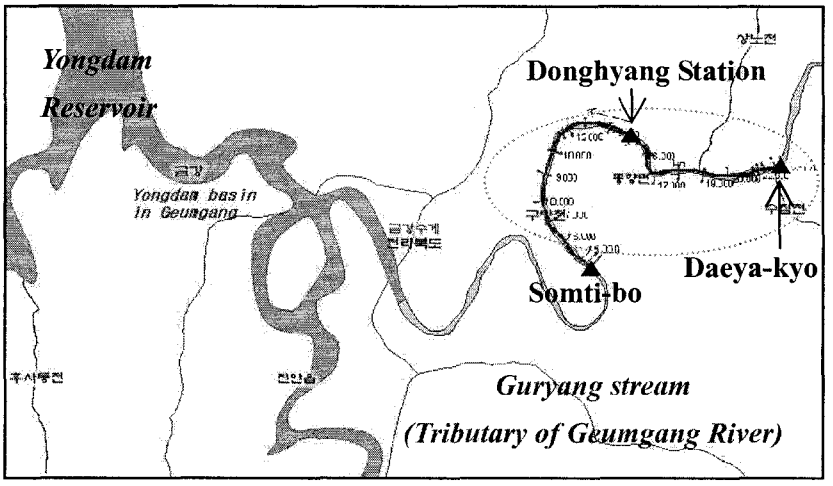


Figure 4-7 The modeled network in the *HEC-RAS* for hydraulic analysis of the rating at the Donghyang station in Guryang stream.

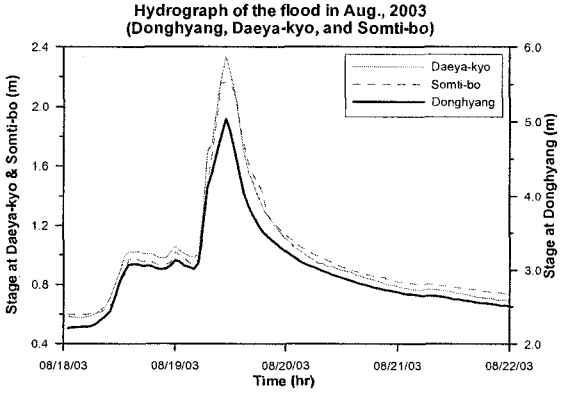


Figure 4-8 The hydrograph of the flood at the Donghyang station, Daeya-kyo, and Somti-bo station in Aug., 2003.

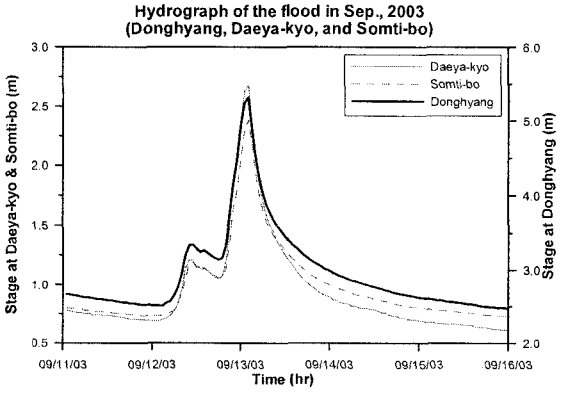


Figure 4-9 The hydrograph of the flood at the Donghyang station, Daeya-kyo, and Somti-bo station in Sep., 2003.

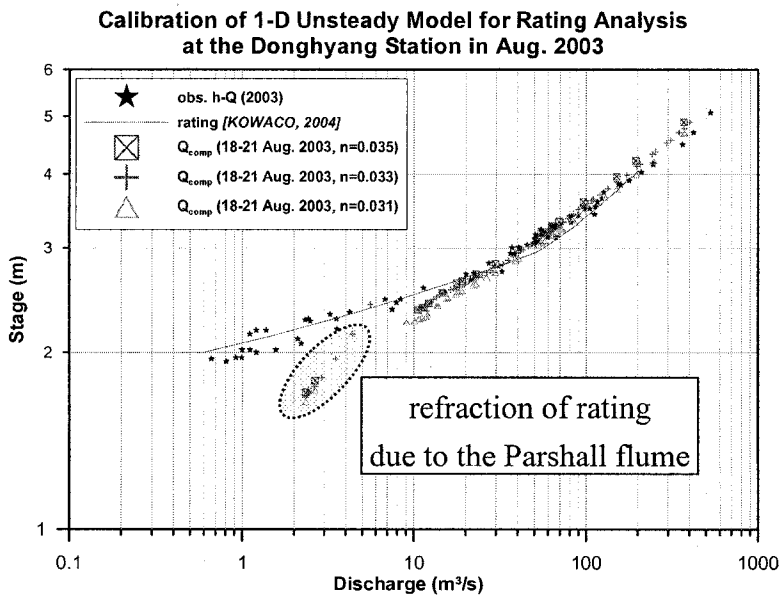


Figure 4-10 The calibration results for the August flood (August 18 to 21, 2003) at the Donghyang station. The rating *[K*-water, *2004]* on the graph means the rating developed in the previous study performed by *K*-water. in 2004.

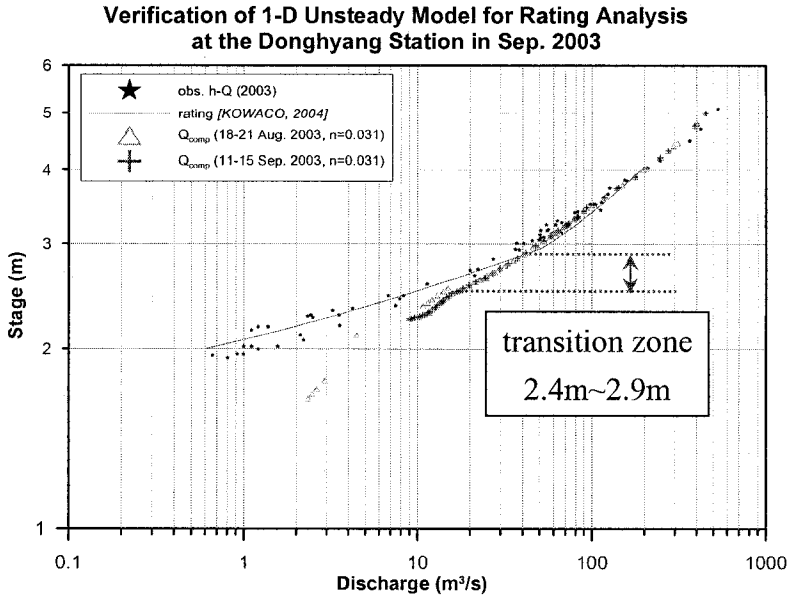


Figure 4-11 The verification of the calibrated model for the September flood (September 11 to 15, 2003) at the Donghyang station.

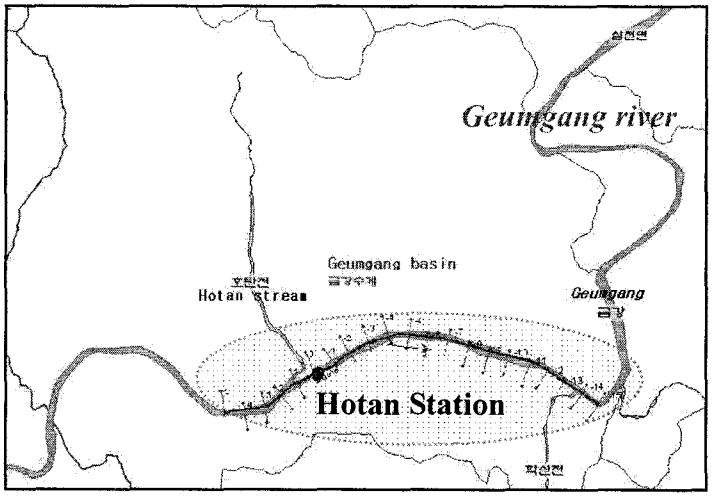


Figure 4-12 The modeled network in the *HEC-RAS* for hydraulic analysis of the rating at the Hotan station.

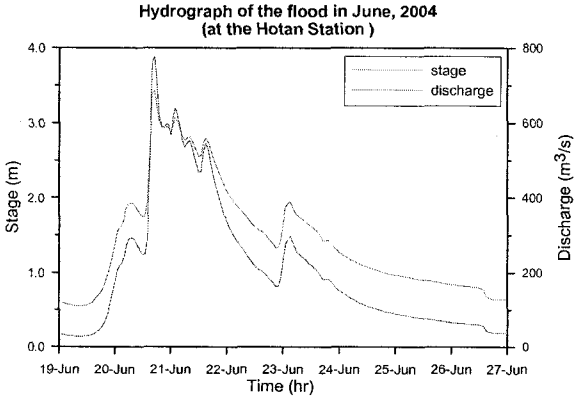


Figure 4-13 The hydrograph of the flood at the Hotan station in June, 2004. The discharge on the graph was computed using the rating developed in a conventional way.

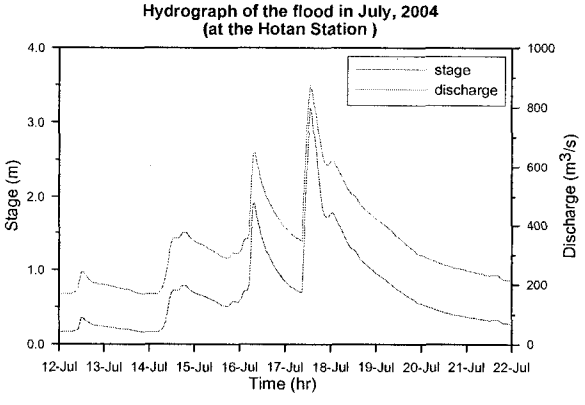


Figure 4-14 The hydrograph of the flood at the Hotan station in July, 2004.

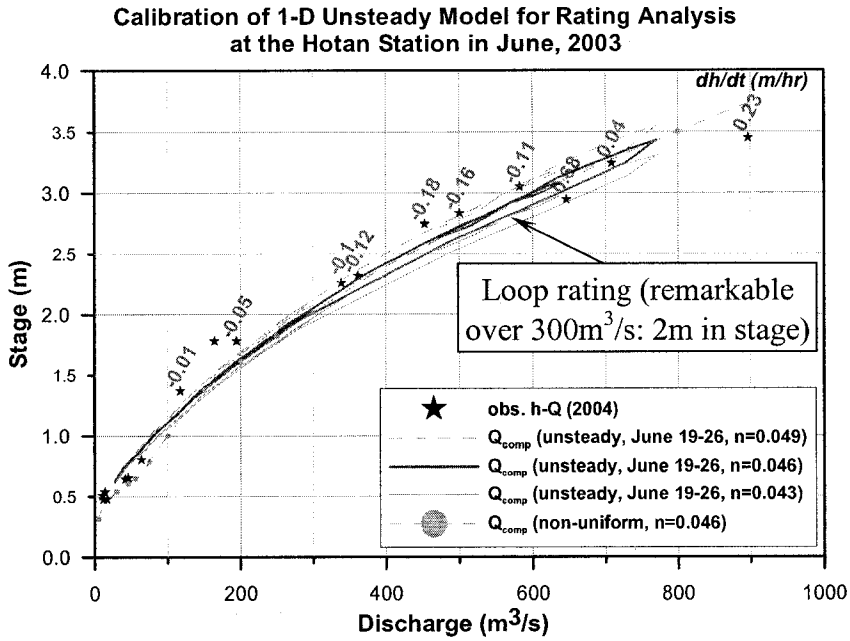


Figure 4-15 The calibration results for the June flood (June 19 to 26, 2004) at the Hotan station. The label on a point refers to the variation rate of the stage when stage-discharge was measured.

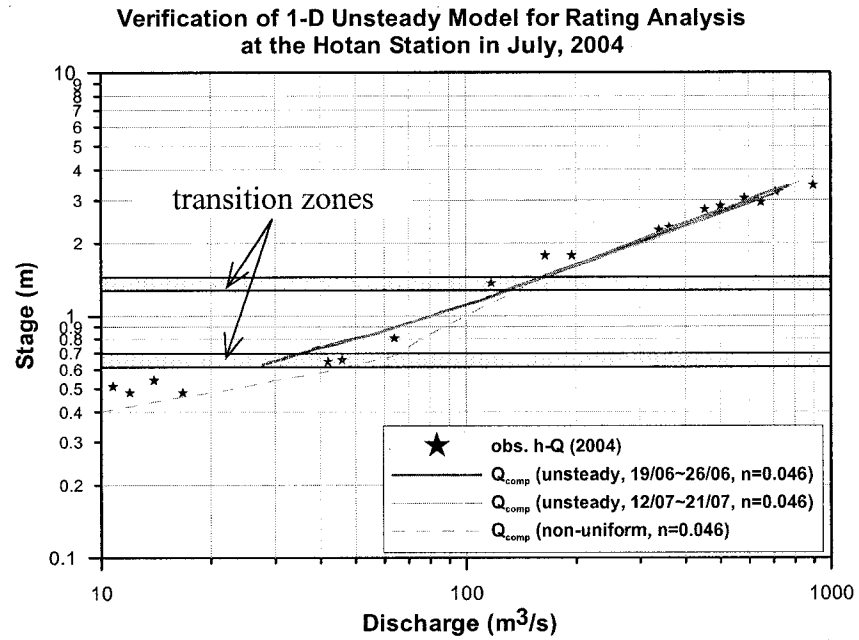


Figure 4-16 The verification results for the July flood data (July 12 to 21, 2004) at the Hotan station.

4.2.2 Rating Analysis Using Nonlinear Programming (NLP)

When a rating is represented with Equation 4.1 and 4.2, the concept of developing a rating using optimization techniques is that the linear relation between *Log(Q)* and *Log(h-e)* is dependent on the variation of the effective GZF, *e*: hence, the value of *e* that results in the best linear relation can be searched for with the help of optimization techniques.

In order to utilize this procedure, first the transition zones should be determined visually on a logarithmic graph. If the transition zones plainly exist, a rating should split into at least two segments. Next, the actual GZF should be obtained by field survey. This value may be represented with a range of a lower (*actual* GZF_{lower}) and upper (*actual* GZF_{upper}) limits rather than a single value. The actual GZF should be greater than the historical minimum stage and less than the transition zone. Once two values are determined, the effective GZF *e* is optimized over the parameters of *b* and *a* under section control, and then the remaining segments under channel or mixed controls are determined optimally without any constraints. Finally, each segment is combined within the transition zones to yield continuous rating. For a smooth connection, each segment needs to be refined by adjusting *e* slightly within its allowable range. Figure 4-17 shows the detailed procedure for this rating analysis.

In this study, if a rating needs to be divided into two segments, the objective function is set up to maximize the sum of the two linear correlation coefficients ($r_{section\ control}$ and $r_{channel\ control}$) between *Log(Q)* and *Log(h-e)* in order to guarantee the best

linearity (Equation 4.3). As the tool for optimization, the *SOLVER* that is embedded into *MS-Excel* was used with its nonlinear optimization option. This optimization problem is formulated as:

$$
\begin{aligned}
&\textit{objective function}: F = \max_{e}(r_{\substack{\text{section}\\\text{control}}} + r_{\substack{\text{channel}\\\text{control}}});\\
&\textit{subject to},\\
&\quad b_{\substack{\text{section}\\\text{control}}} \geq 2.0;\\
&\quad 1.5 \leq b_{\substack{\text{channel}\\\text{control}}} \leq 2.0;\\
&\quad \textit{actual } GZF_{lower} \leq e \leq \textit{actual } GZF_{upper};\\
&\quad \text{historical minimum } h \leq e \leq \text{transition zone.}
\end{aligned}
\tag{4.3}
$$

An optimized rating by Equation 4.3 is regarded as a single stage-discharge relationship in steady state. If a hysteresis exists, the steady state rating should be corrected to be loop-shaped. The best way to do this might be to apply the Saint-Venant equations. However, this is so complicated that a simplified equation such as the Jones formula is often used in practice *[Ogink, 1994]*:

$$
Q_u = \frac{1}{n} A_s R^{2/3} S_o^{1/2} \left(1 - \frac{1}{S_o}\frac{\partial h}{\partial x}\right)^{1/2} = Q_s \left(1 - \frac{1}{S_o}\frac{\partial h}{\partial x}\right)^{1/2};
$$

where, Q_u is unsteady state discharge;

Q_s is steady state discharge;

n is *Manning's roughness coefficient*; (4.4)

A_s is flow area;

R is hydraulic radius;

S_o is bed slope;

$\partial h/\partial x$ is stage variation along distance, and *approximated* as follows:

$\frac{\partial h}{\partial x} \approx -\frac{1}{c}\frac{\partial h}{\partial t}$: where, c is celerity;

hence, $Q_u = Q_s\left(1+\frac{1}{S_o c}\frac{\partial h}{\partial t}\right)^{1/2}$ (referred to as *Jones formula*).

In the Jones formula, it is very difficult to obtain $1/S_o c$ practically: hence, it is often approximated as a function of stage *[Delft Hydraulics, 1994]*:

$$\frac{1}{S_o c} \approx a1 + (b1 \times h) + (c1 \times h^2) \tag{4.5}$$

1) Rating analysis at the Donghyang station

The ratings were evaluated from 2003 through 2005 in order to trace the yearly shift of the physical parameters of *e* and *b*. Figure 4-18 to 4-19 show that the rating under channel control was significantly changed before and after the flood in August, 2003. Therefore, two ratings were developed, one before and one after the flood in 2003.

The transition zone for each period was identified visually on a logarithmic graph. Figures 4-18 through 4-21 show the transition zone continued to move up and down. This means the channel conditions were not stable due to the scour and deposition by the floods occurred during the three years. Table 4-1 shows how the transition zones were shifted year by year. One thing to be noted from this is that all transition zones are contained in the range of the transition zone (2.4m to 2.9m) determined in the unsteady flow analysis (Figure 4-11).

Because the Donghyang station has a Parshall flume (0.92m in height) downstream, which serves as a natural control section, the actual GZF was obtained by subtracting the

measured Parshall flume stages over 0.92m from the stages measured at the Donghyang station at the same time (Table 4-2). Finally, the effective GZFs were optimized in each period; however, the search of an optimal e to meet all constraints failed. This may be explained by the fact that the man-made Parshall flume might disturb the natural stage-discharge relationships. Hence, the optimization was made without the constraints for b. The results are summarized in Table 4-4, and Figure 4-22 shows the developed ratings with movement of e and the transition zone. Although the Parshall flume works as a firm section control, the effective GZF varied due to the measurement error of the stage-discharge. Therefore, the movements of the effective GZF and the transition zone may not occur in the same pattern.

2) Rating analysis at the Hotan station

The ratings were evaluated from 2004 through 2005. Because the measured stage-discharge data in 2004 were not sufficient to identify the hydraulic characteristics, the analysis in 2004 was made along with the data in 2003. The transition zones were identified on the basis of the results from the unsteady flow analysis, and confirmed visually on a logarithmic graph. However, it was almost impossible to recognize the transition zones without the help of the unsteady flow analysis due to the hysteresis of the rating and lack of the measured data around the transition zones. Figure 4-23 and 4-24 show the transition zones stayed in the same place throughout two years. This is summarized in Table 4-1.

The effective GZFs were optimized with the constraints for the slopes (b) of the lower two segments. The optimal GZFs to meet all the constraints were found in 2004 and 2005, and the results are summarized in Table 4-4. The results show that the effective GZF moved down from -0.505m to -0.539m due to the big flood with the peak flow of about 2,550m^3/s occurred on August 3, 2005. This shift can work as a constraint in the next period in the context of adaptive management.

Finally, the degree of the hysteresis was determined using the Jones formula. In this study, the parameters of $a1$, $b1$ and $c1$ in Equation 4.5 were obtained indirectly from the unsteady flow analysis results using the *SOLVER* optimization tool. The objective function was formulated to minimize the summation of squared errors between the unsteady flows by the *HEC-RAS* model and the computed discharges by the Jones formula during the flood event in June, 2004. The parameters are shown in Table 4-3. Figures 4-25 shows a 3-dimensional rating that is a function of h and dh/dt, and Figure 4-26 and 4-27 show the developed ratings in 2004 and 2005 with the unsteady state rating by *HEC-RAS*. Figure 4-27 shows that the *HEC-RAS* results of 2004 are not consistent with the results of 2005. This means the cross-section data surveyed in 2004 did not work well in 2005 because of big changes in channel shape: hence, care must be taken in applying the unsteady flow analysis in 2004 to another year.

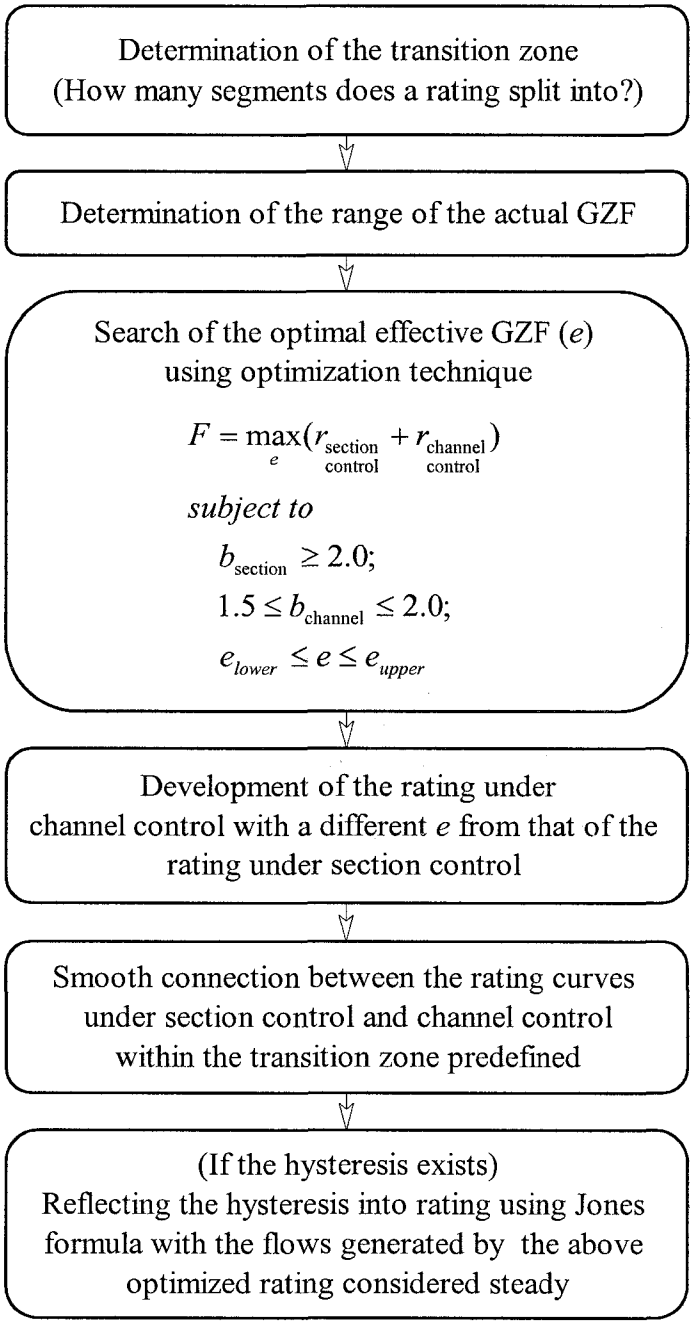


Figure 4-17 The procedure for rating analysis using an optimization technique.

Table 4-1 Identification of the transition zones using the measured stage-discharge data at the Donghyang and Hotan stations.

Station / Period		Transition zone	
		Donghyang	Hotan
2003	before the flood in Aug.	2.7~2.8	0.6~0.7 (lower zone) 1.3~1.4 (upper zone)
	after the flood in Aug.	2.65~2.75	
2004	-	2.4~2.6	
2005	-	2.55~2.65	

Table 4-2 Identification of the actual GZF using the measure stage-discharge data at the Donghyang and Hotan stations.

Station / Year	Actual GZF	
	Donghyang	Hotan
2003	1.5~1.75	No actual GZF because it was hard to identify a control section in the field.
2004	No actual GZF because the stages were not recorded in 2004 due to the trouble of the gauge.	
2005		

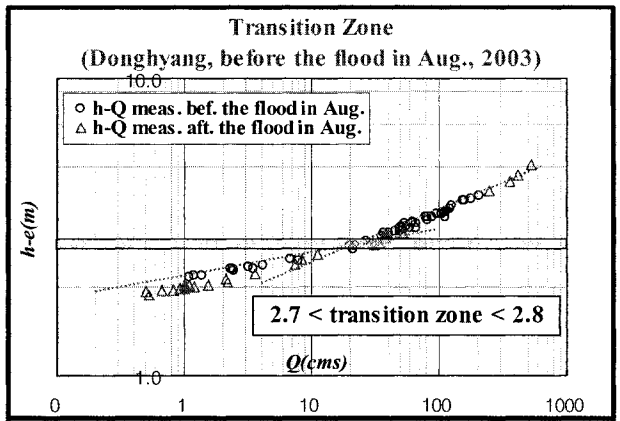


Figure 4-18 Identification of the transition zone (2.7~2.8m) at the Donghyang station before the flood in Aug., 2003.

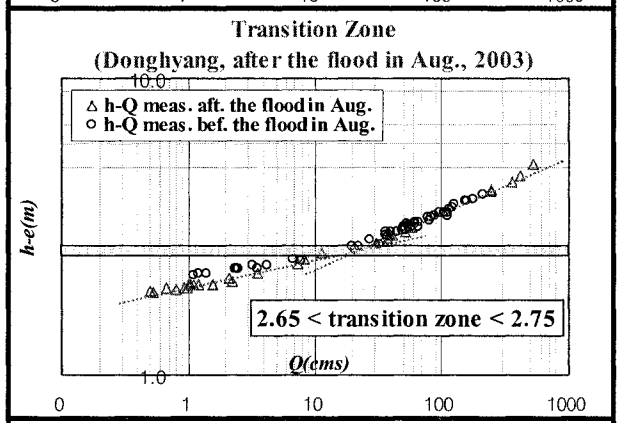


Figure 4-19 Identification of the transition zone (2.65~2.75m) at the Donghyang station after the flood in Aug., 2003: This shows the transition zone moved down compared to that before the flood in Aug., 2003.

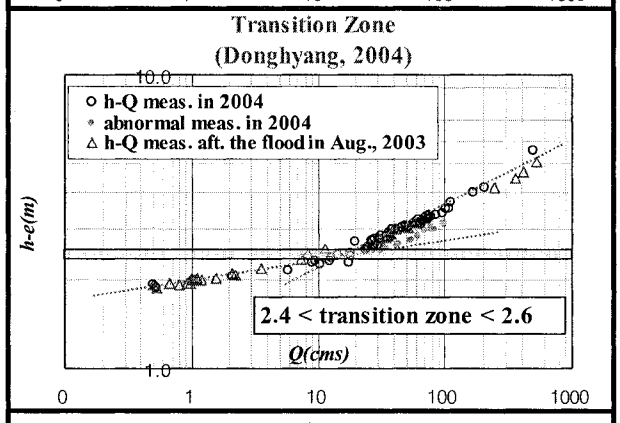


Figure 4-20 Identification of the transition zone (2.4~2.6m) at the Donghyang station in 2004: This also shows the transition zone moved down compared to that after the flood in Aug., 2003.

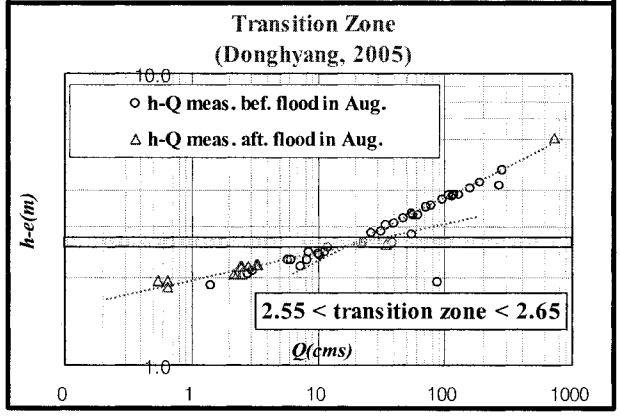


Figure 4-21 Identification of the transition zone (2.55~2.65m) at the Donghyang station in 2005: This shows the transition zone moved up compared to that in 2003.

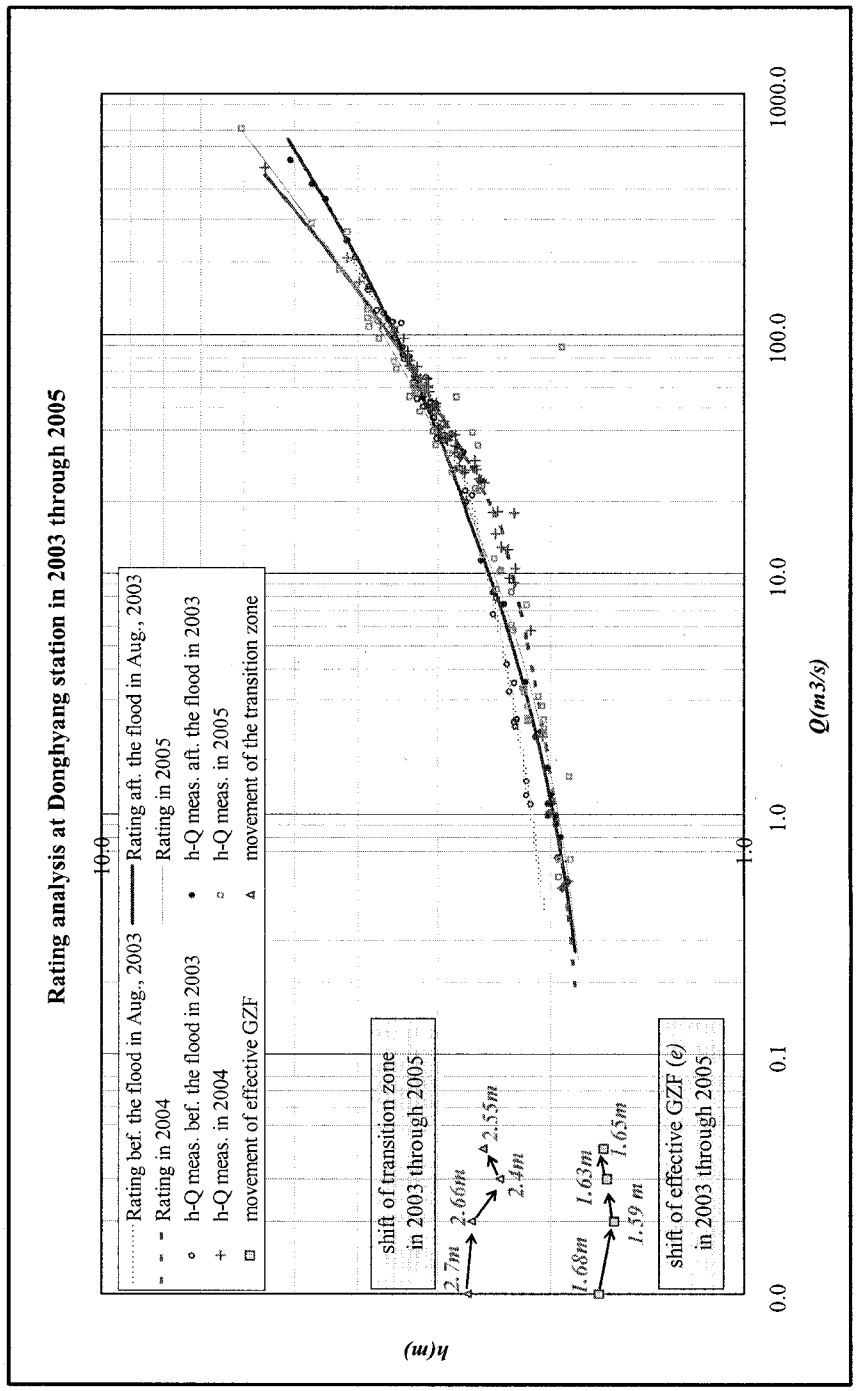


Figure 4-22 Evaluation of the rating with analyzing the shift of GZF and transition zone at the Donghyang station from 2003 through 2005.

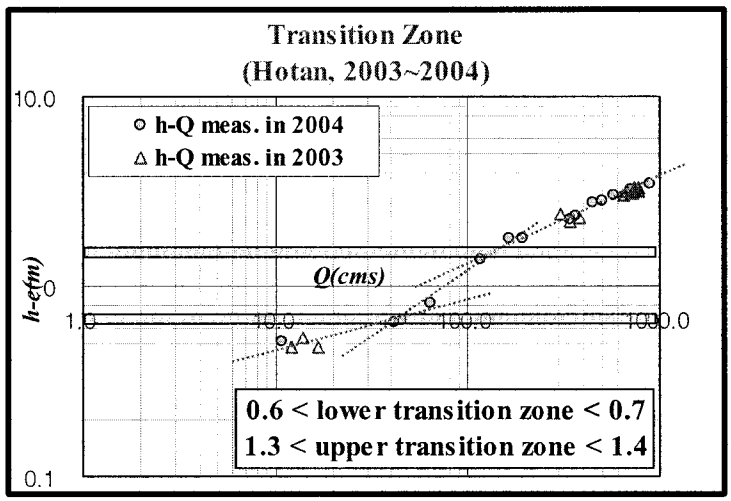


Figure 4-23 Analysis of the transition zone at the Hotan station in 2004.

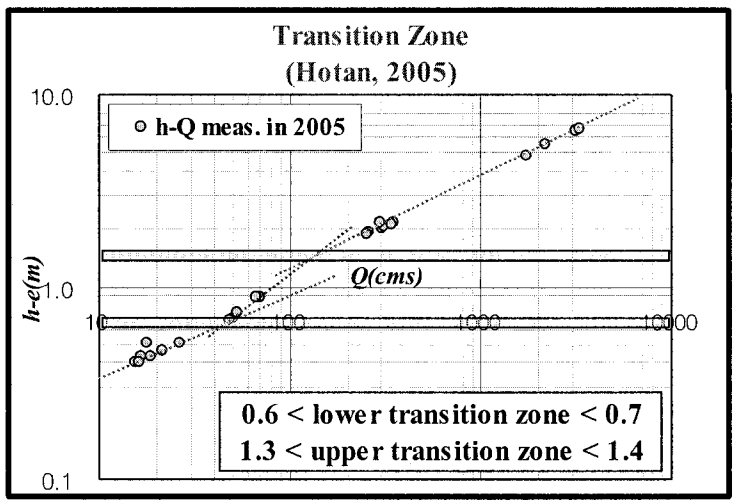


Figure 4-24 Analysis of the transition zone at the Hotan station in 2005.

Table 4-3 The parameters in Equation 4-5 for simplifying $1/S_o c$ of Jones formula.

a1	*b1*	*c1*
0.215	-0.0216	0.000

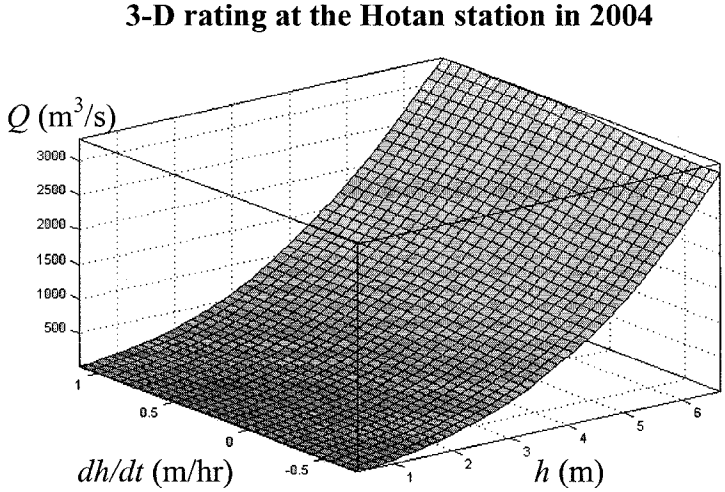


Figure 4-25 The 3-D rating curve at the Hotan station in 2004. Q is the function of h and dh/dt.

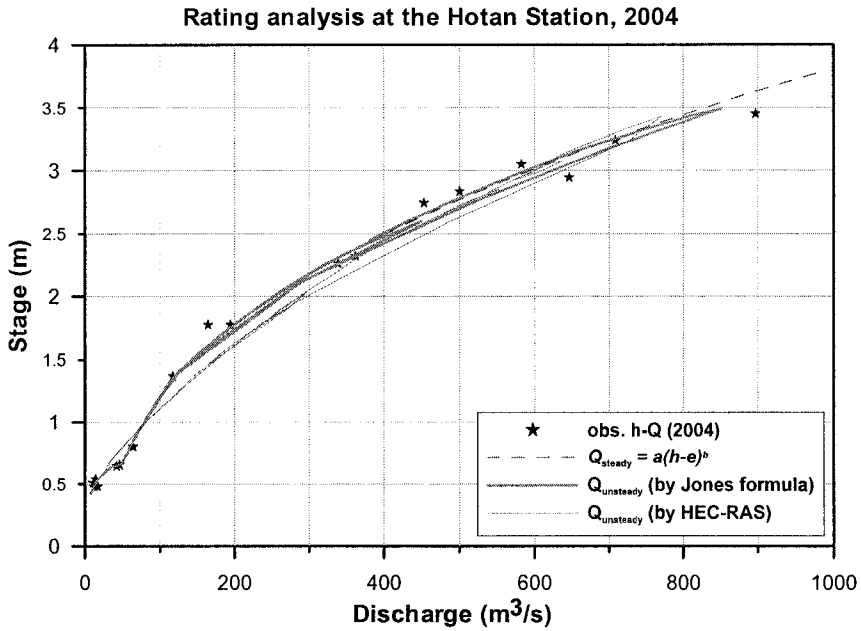


Figure 4-26 Rating evaluation the Hotan station in 2004, where, Q_{steady} refers to the flow computed using the rating $Q=a(h-e)^b$ developed through NLP, which is considered steady instead of the Manning equation.

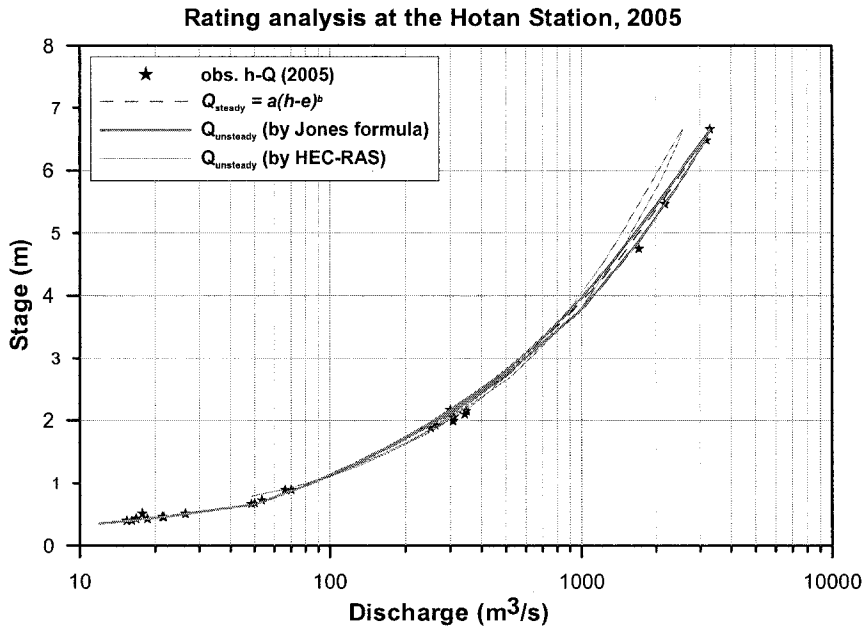


Figure 4-27 Rating evaluation at the Hotan station in 2005.

Table 4-4 The rating developed with NLP at the Donghyang and Hotan stations. Where, the slope for channel control ($b_{chn.ctrl}$) was determined when the effective GZF for section control was optimized: hence, it is different from b of the rating under channel control that is computed again after the optimal effective GZF was obtained.

year		Rating	r	his.min. h (m)	b (*sec. ctrl*)	b (*chn.ctrl*)
Rating by NLP at the Donghyang station						
2003	bef. flood	$Q=21.071(h-1.679)^{3.997}$: $h \leq 2.70$m	0.993	2.05	3.997	2.757
		$Q=18.683(h-1.62)^{2.646}$: $h \geq 2.70$m	0.978			
	aft. flood	$Q=12.948(h-1.59)^{2.717}$: $h \leq 2.66$m	0.986	2.08	2.717	2.741
		$Q=22.833(h-1.79)^{2.767}$: $h \geq 2.66$m	0.973			
2004		$Q=38.135(h-1.63)^{3.294}$: $h \leq 2.40$m	0.978	1.80	3.294	2.195
		$Q=17.580(h-1.44)^{2.300}$: $h \geq 2.40$m	0.977			
2005		$Q=16.276(h-1.65)^{2.281}$: $h \leq 2.55$m	0.927	1.65	2.281	2.463
		$Q=21.722(h-1.75)^{2.375}$: $h \geq 2.55$m	0.985			
Rating by NLP at the Hotan station						
2004		$Q=12.983(h+0.505)^{7.299}$: $h \leq 0.65$m	0.861	0.42	7.299	2.0
		$Q=36.196(h+0.505)^{1.890}$: $0.65 \leq h \leq 1.35$m	0.994			
		$Q=32.476(h+0.351)^{2.404}$: $h \geq 1.35$m	0.998			
2005		$Q=21.145(h+0.539)^{4.838}$: $h \leq 0.65$m	0.979	0.21	4.838	2.0
		$Q=21.365(h+0.739)^{2.479}$: $0.65 \leq h \leq 1.35$m	0.998			
		$Q=19.646(h+0.765)^{2.55}$: $h \geq 1.35$m	0.999			

4.2.3 Rating Analysis Using Fuzzy Rule-based Modeling

Unlike the classical methods that are based on the hydraulic features of the ratings, the fuzzy rule-based model is a typical data-driven method. Therefore, this approach does not depend on the hydraulic characteristics of the rating, instead, it tries to recognize the patterns between inputs and outputs. The only required data for a rating analysis by a fuzzy inference system are the stage-discharge measurements. Basically, fuzzy rule-based modeling can be considered a probabilistic approach because a finally combined fuzzy set has a similarity to a probability distribution. However, its irregular shape makes the parameters such as confidence interval and percentiles hard to determine. Therefore fuzzy rule-based modeling was used as a deterministic approach for this study.

The first step for fuzzy rule-based modeling is to generate a structure for the fuzzy inference system (FIS), which is composed of input-output nodes with selection of the number and types of input-output membership functions, the number and structure of rules, and fuzzy and defuzzification methods. Once the structure of FIS is designed, the next step is to extract some fuzzy rules from the observations and to train the FIS. The training process is associated with the trade off between performance and overfitting. After training, the trained FIS is tested using other independent data sets.

In this study, the *ANFIS* (Adaptive Neuro-Fuzzy Inference System) was used for the rating analysis. The ANFIS is built into the *MATLAB* package as a toolbox developed by *Mathworks*. When a FIS is trained using the counting algorithm *[Bardossy and Duckstein, 1995]*, the parameters of membership functions are tuned manually by the user (considering overfitting). Meanwhile, the *ANFIS* tunes the parameters using either a

backpropagation algorithm alone, or in combination with a least squared type of method *[Mathworks, Fuzzy Logic Toolbox User's Guide]*. These techniques provide a method for FIS to learn patterns or relationships from inputs and outputs. The *ANFIS* is designed for facilitating adaptive data modeling. Hence, a trained FIS structure can be saved in a user specified directory and becomes a prior structure for the next modeling with different data set. As constraints on the *ANFIS* (version 2.0), the software only supports a Sugeno-type system *[Mathworks, Fuzzy Logic Toolbox User's Guide]*:

The key issue in training FIS is overfitting. This is a phenomenon where the FIS is dominated by a certain data set. Some methods such as cross-validation and early stopping can be used to prevent overfitting. In this study, the cross-validation method was employed. For this method, a checking data set is required to control the potential for overfitting the data. The basic idea behind the cross-validation is that a model begins overfitting the training data set after certain iteration during training, and the model error for the checking data set tends to decrease until overfitting takes place, and then the error suddenly increases. The cross-validation makes training stop when the model error for checking data comes to a minimum *[Mathworks, Fuzzy Logic Toolbox User's Guide]*.

The inputs were composed of stage, and the variation rate of stages in case a hysteresis exists, and the output was always fixed to discharge. Figure 4-28 shows an example of the structure for fuzzy rule-based modeling with three membership functions for the inputs. Here, the rules are formed by combining the membership functions between input nodes. Therefore, the number of the rules and the membership functions of the output are nine, the same as in this example. Figure 4-29 shows how all the processes work into a FIS for rating analysis.

1) Rating analysis at the Donghyang station

Because the hysteresis was not considerable at this station, the input and output nodes were composed of only stage and the measured discharge respectively. The rating was analyzed using the stage-discharge data measured in 2003 when the unsteady analysis was done. When the *ANFIS*s for the rating analysis before and after the flood that occurred on August 18, 2003 were trained, the stage-discharge data measured in 2002 and 2004 were used respectively as the checking data. Finally, the ratings were simulated by the trained *ANFIS* using the evenly discretized stages ranging from the recorded minimum to the maximum, and compared with the data measured in 2003. The *ANFIS* was also compared with the counting and weighted counting algorithms programmed using Visual Basic for Applications (*VBA*) built into *MS-Excel*.

The training of the ANFIS was performed by changing the number and type of the input-output membership functions with the default fuzzy operators. The default fuzzy operators of *ANFIS* are: product operator for the 'AND' method and max operator for the 'OR' method in computing the degree of fulfillment; product operator for implication; max operator for aggregation, and; weighted average for defuzzification. The membership functions of stage were limited to the symmetric membership functions of the generalized bell and the Gaussian types. In contrast, for training with the counting and weighted counting algorithms, the length of rule support, the overlap size between rules, and the threshold (ϵ) were calibrated manually. The type of membership function was limited to the triangular fuzzy number for ease of computation.

Figure 4-31 and 4-33 show the training results for the stage-discharge data measured before and after the flood in August respectively. They indicate that the *ANFIS*

starts overfitting regardless of the type of the output membership function when the number of the input membership functions exceeds three. The extreme cases of overfitting are illustrated on the graphs. Contrary to general expectation, the checking error did not increase significantly during 100 training epochs. Figure 4-30 and 4-32 show the calibrated parameters of the input membership functions before and after training. From the comparison between the *ANFIS* and the counting-weighted counting algorithms (Figure 4-34), it can be noticed that the *ANFIS* is much more robust than the counting and weighted counting algorithms in that the two latter methods show very poor results for the low and high stage-discharges. Figure 4-34 also shows that the weighted counting algorithm is superior to the counting algorithm. One thing that should be noted is that the discharges simulated by the *ANFIS* were often negative when some low stages were fed into the *ANFIS*. Such values were not contained in the training data set. This fact indicates that fuzzy rule-based modeling has obvious limits in extending ratings beyond the training data set.

2) Rating analysis at the Hotan station

Unlike the Donghyang station, the variation rate of stage was added to the input nodes due to considerable hysteresis. The variation rate of stage at a time was obtained by averaging the variation rates for the past two hours. The rating was analyzed using the stage-discharge data measured in 2004. For analyzing the hysteresis of rating, the outputs from *HEC-RAS* were added into the training data set along with the measured data in 2003 and 2004 because the measured data containing the information about hysteresis were not sufficient. The stage-discharges over $100m^3/s$, where the hysteresis started, were

selected from the entire set of outputs computed by *HEC-RAS* using the flood data on June 19, 2004. Therefore, the hysteresis was analyzed indirectly through replicating the 1-D unsteady model. For checking the model, the data measured in 2005 and the simulated outputs for the flood on June 3, 2004 were fed into the system.

The training was performed by changing the number and type of membership functions. Figure 4-36 shows that the *ANFIS* was trained satisfactorily with: the *constant* output membership function, and; 4 and 2 *gbell* membership functions of the stage and variation rate of stage respectively. It is also recognized that the *ANFIS* starts overfitting regardless of the type of the output membership function when the number of membership functions for the stage variation exceeds three. Figure 4-36 shows the checking errors are less than the training errors. This is because the checking data set contains only the stage-discharge data below $100m^3/s$.

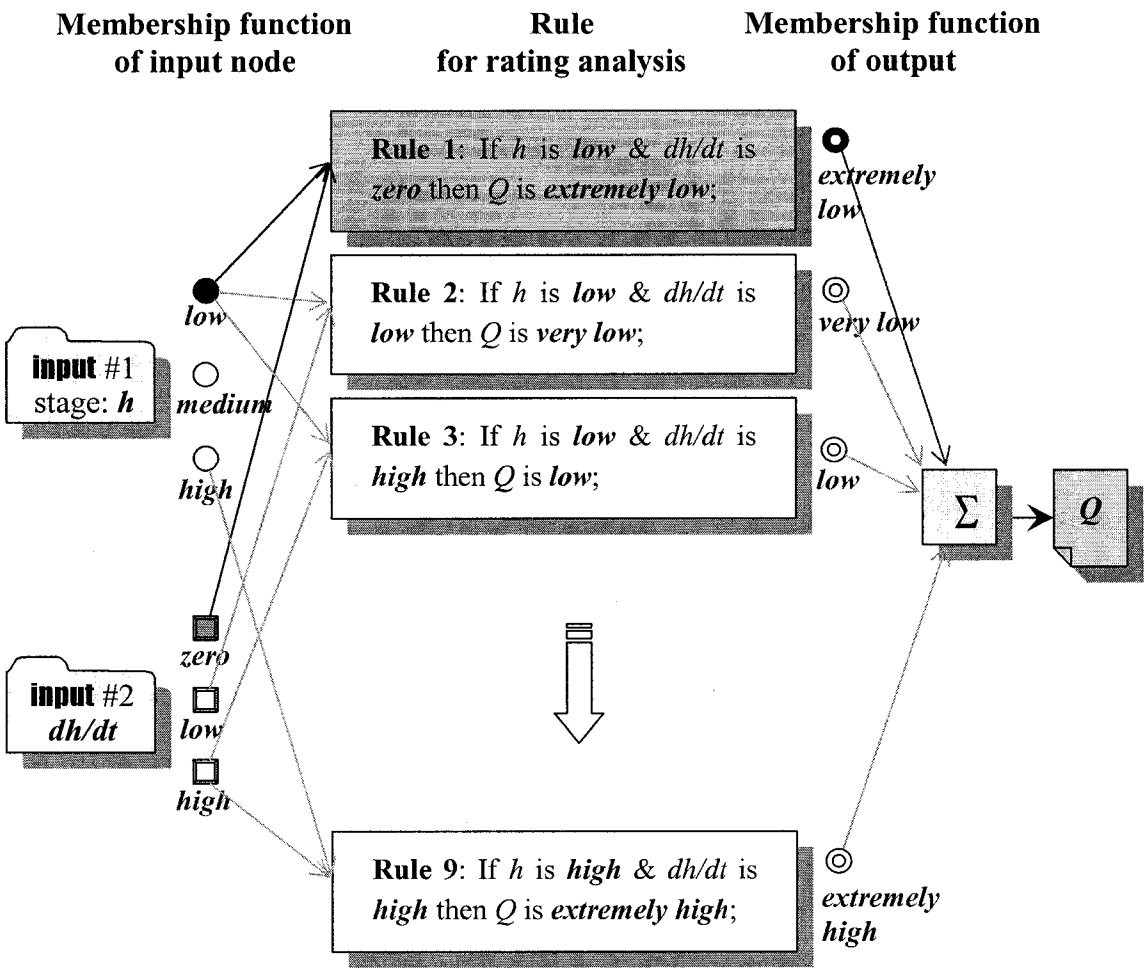


Figure 4-28 This shows an example of the structure of the fuzzy inference system for rating analysis in the case of the three membership functions per each input node. Where, h is stage (m), dh/dt is the variation rate of stage (m/hr), Q is discharge (m^3/s), and the symbol Σ refers to aggregation of rules.

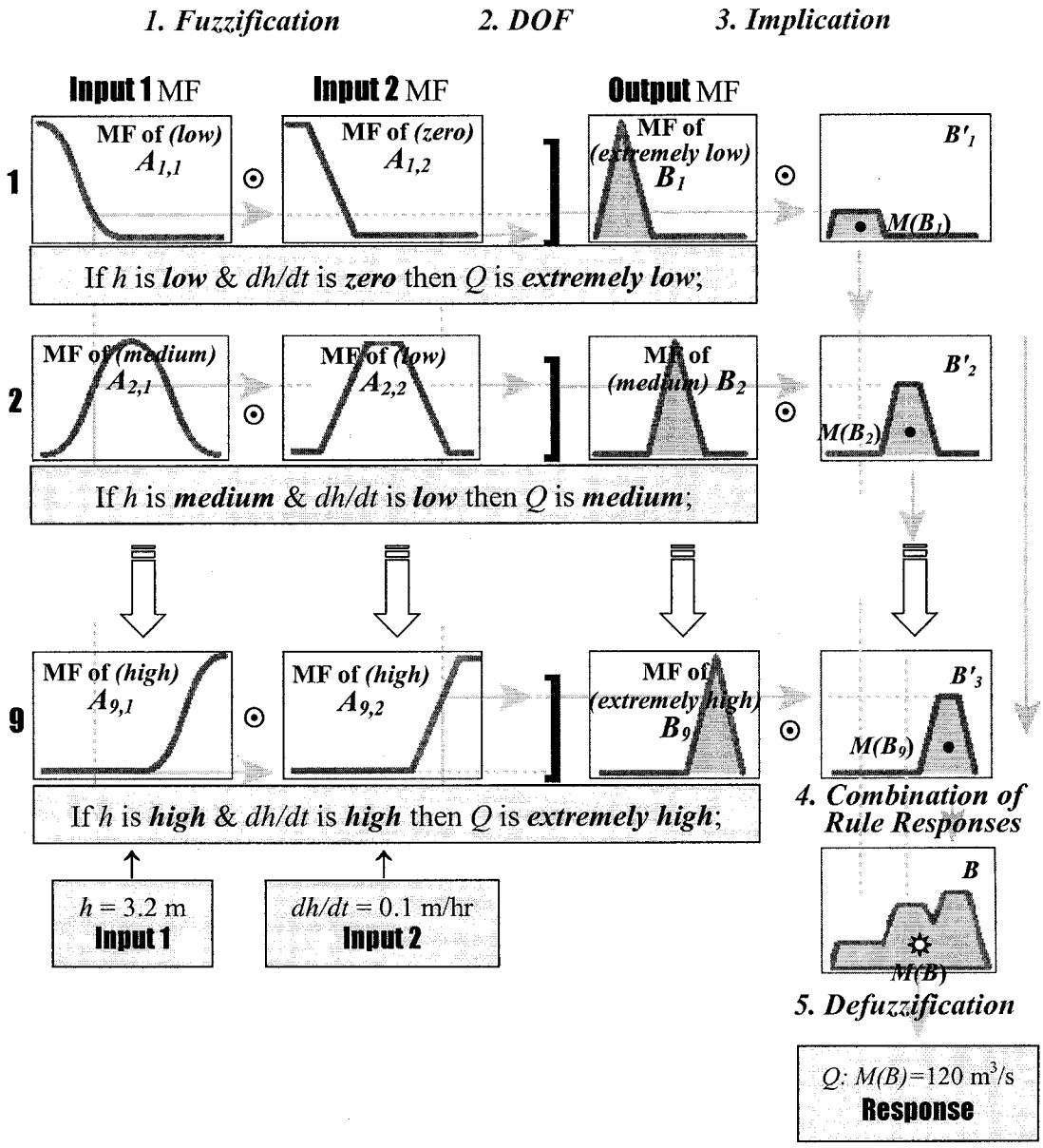


Figure 4-29 This illustrates how a FIS for rating analysis are operated in the case of the three membership functions per each input node. Where, MF stands for membership function, the symbol ⊙ refers to fuzzy operator, A_i and B_i are the membership functions of input-output, B'_i is the reshaped B_i by implication process, $M(B_i)$ is the mean of B'_i, a is the combined fuzzy set of all rules, and $M(B)$ is the mean of B that is the final response by the defuzzification process.

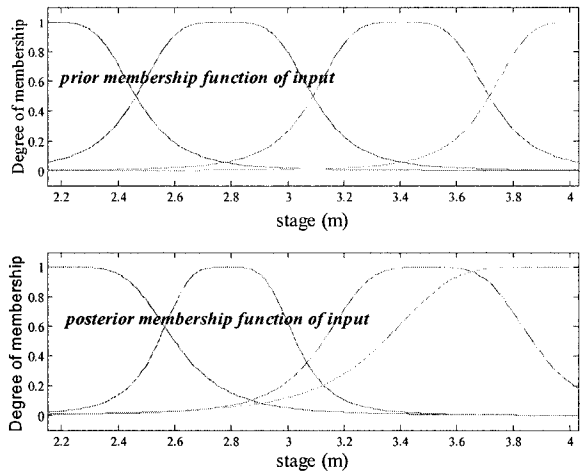


Figure 4-30 Variation of 4 input *gbell* membership functions before (*prior*) and after (*posterior*) training using the data measured before the August flood at the Donghyang station.

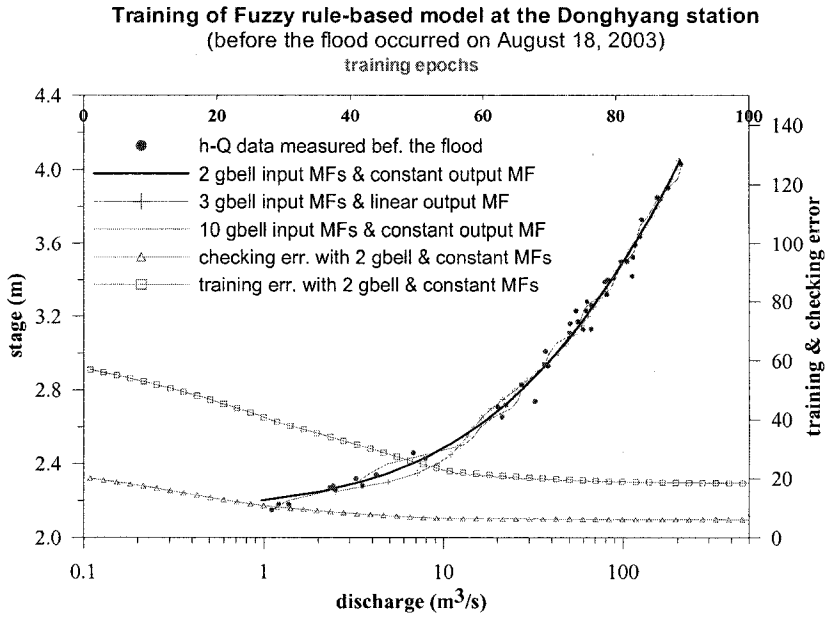


Figure 4-31 Training of the *ANFIS* for the data measured at the Donghyang before the flood occurred on August 18, 2003 by changing the membership functions. The *ANFIS* was trained satisfactorily with the *constant* output membership function and 2 *gbell* input membership function.

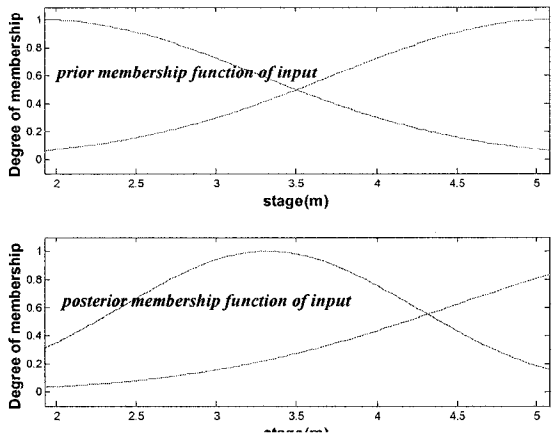


Figure 4-32 Variation of 2 input *gbell* membership functions before (*prior*) and after (*posterior*) training the data measured after the August flood at the Donghyang station.

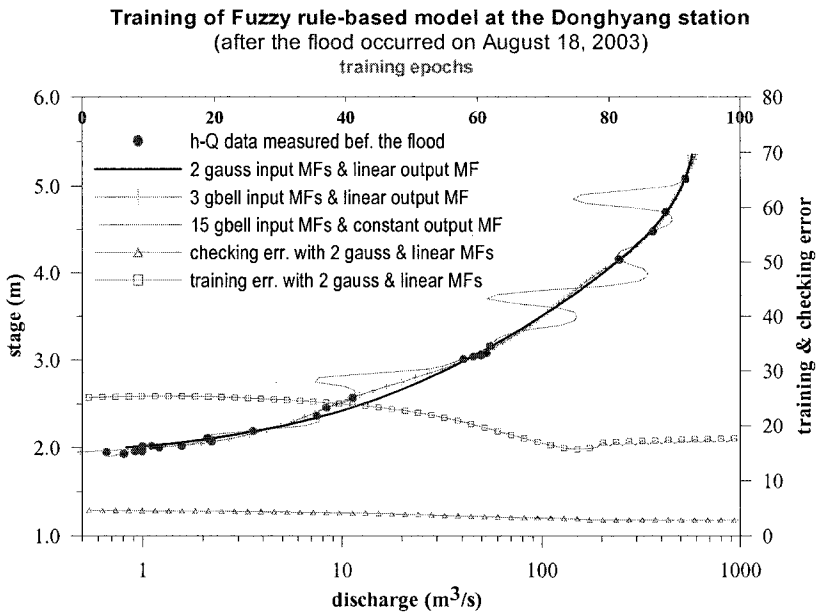


Figure 4-33 Training of the *ANFIS* for the data measured at the Donghyang after the flood occurred on August 18, 2003 by changing the membership functions. The *ANFIS* was trained satisfactorily with the *linear* output membership function and 2 *gauss* input membership function.

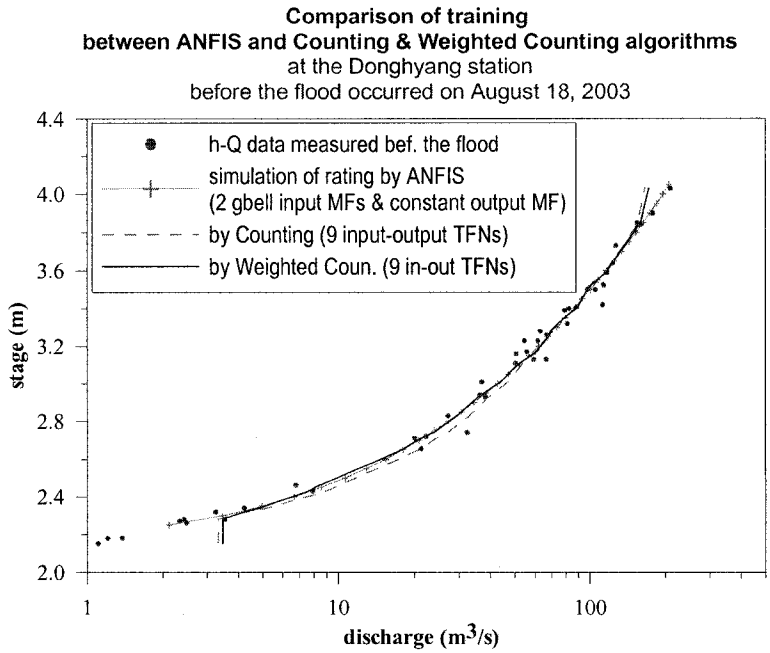


Figure 4-34 Comparison of training between the *ANFIS* and the counting/weighted counting algorithms at the Donghyang before the August flood in 2003.

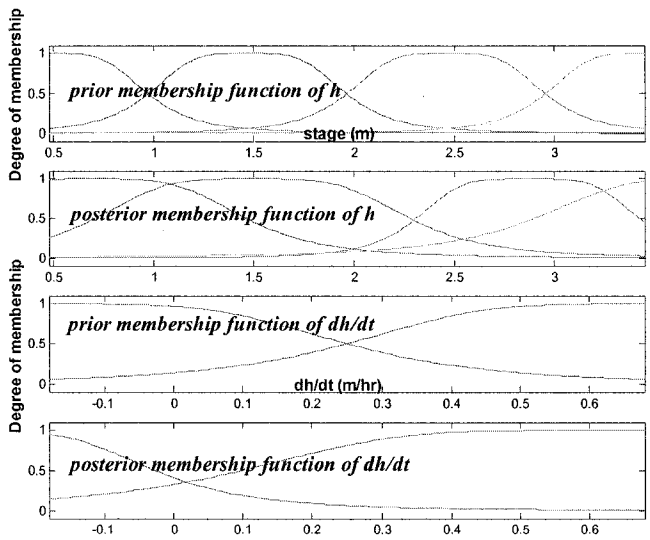


Figure 4-35 Variation of 4 and 2 *gbell* membership functions of stage and variation rate of stage respectively before (*prior*) and after (*posterior*) training at the Hotan station.

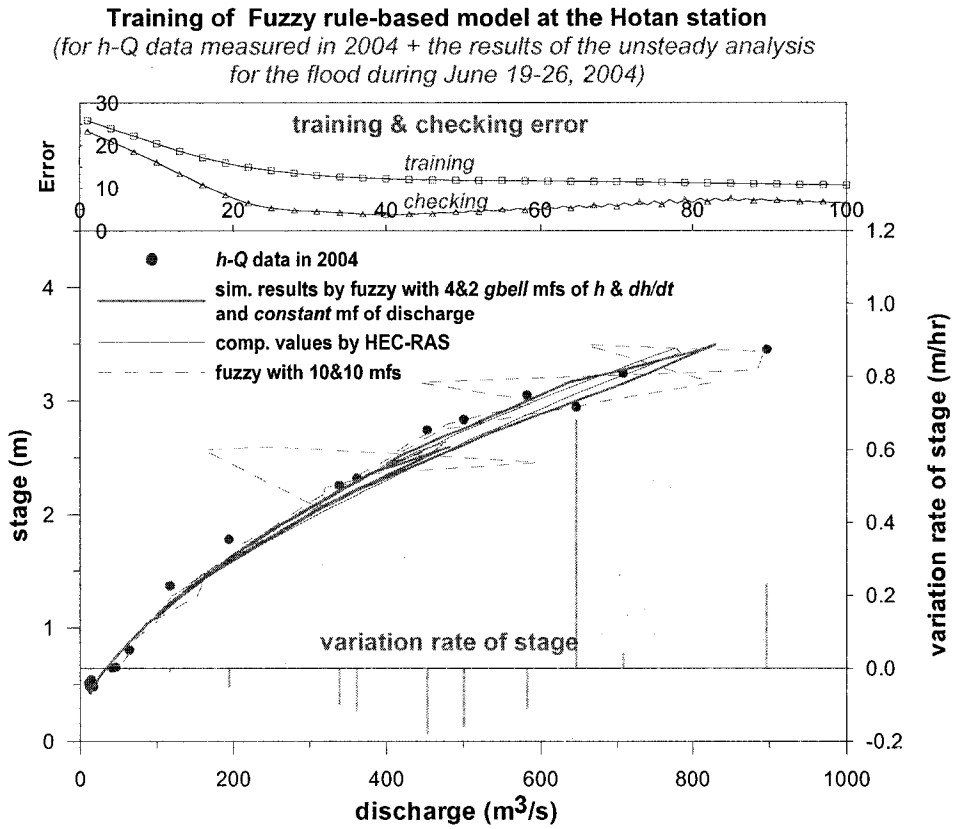


Figure 4-36 Training of the *ANFIS* on the data set containing both the data measured at the Hotan in 2004 and the results from *HEC-RAS* for the flood that occurred on June 19, 2004 by changing the membership functions.

4.3 Probabilistic Approach for Rating Analysis Using Bayesian MCMC

In this study, the rating analysis was performed at the Donghyang and Hotan stations with the help of the BMCMC sampling techniques currently experiencing an increase in popularity in water resources and environment engineering. As a tool for the BMCMC method, the *WinBUGS* package *[Spiegelhalter et al., 1995]*, developed by the Medical Research Council (MRC) and Imperial College of Science, Technology and Medicine U.K, was used. The purposes of the study are:

i) to analyze the uncertainties of ratings by deriving the probability distributions;

ii) to screen the erroneously measured discharges that lower the reliability of ratings on the basis of the derived probability distributions;

iii) to understand the behavior of the parameters as time goes by in the context of adaptive management.

Assuming the observations are independent and the errors ($\varepsilon = Q\text{-}Q^{rating}$) between the observed ($Q$) and computed discharge (Q^{rating}) are normally distributed with mean 0 and variance σ^2, ratings can be analyzed in a Bayesian approach using likelihood with the parameter vector $\theta_{steady}=\{a,e,b,\tau_{steady}\}$, $\theta_{unsteady}=\{a,e,b,a1,b1,\tau_{unsteady}\}$ and their prior probability distributions (Equation 4.6). If the data are actually dependent and assumed to be independent, then the variances are going to be underestimated. This can create a situation of over-confident inferences. One way to deal with this problem is to transform the original data. However, the observed stage-discharge data can be considered independent random variables because they are irregular time series and each observation is an independent and identically-distributed random variable.

$$\varepsilon = (Q - Q^{rating}) \sim Normal(0, \sigma^2);$$

(*likelihood* for rating analysis at the Donghyang station or for rating analysis for the low flow [steady state] at the Hotan station)

$$p(Q \mid \theta_{steady}) = \prod_{j=1}^{J} \frac{1}{\sqrt{2\pi\sigma^2_{steady}}} \exp\left(-\frac{(Q_j - Q_j^{steady})^2}{2\sigma^2} \right);$$

$$Q_j^{steady} = a(h_j - e)^b; \tag{4.6}$$

(*likelihood* for the hysteresis analysis [high flow] at the Hotan station)

$$p(Q \mid \theta_{unsteady}) = \prod_{i=1}^{I} \frac{1}{\sqrt{2\pi\sigma^2_{unsteady}}} \exp\left(-\frac{(Q_i - Q_i^{unsteady})^2}{2\sigma^2} \right);$$

$$Q_i^{unsteady} = Q_i^{steady} \times hysteresis\_correction(i);$$

$$Q_i^{steady} = a(h_i - e)^b;$$

$$hysteresis\_correction(i) = \left(1 + \frac{1}{S_o c_i} \frac{\partial h_i}{\partial t} \right)^{1/2}; \frac{1}{S_o c_i} = a1 + (b1 \times h_i) + (c1 \times h_i^2);$$

(priors in initial stage):

$$a \sim Gamma(a^{para1}, a^{para2});$$

$$b \sim Uniform(b^{para1}, b^{para2});$$

$$\tau_{steady} = \frac{1}{\sigma^2_{steady}} \sim Gamma(\tau^{para1}_{steady}, \tau^{para2}_{steady});$$

$$e \sim Uniform(e^{para1}, e^{para2});$$

censored with the following conditions:

- $e_{\substack{\text{section}\\ \text{control}}} < \min\{\text{historical minimum } h, \text{ lowest } h \text{ of the measured } h\text{-}Q \text{ data}\}$

  (in case of positive e in the rating segment under section control);

- $e_{\substack{\text{channel}\\ \text{control}}} <$ transition zone

  (in case of positive e in the rating segment under channel control);

$$a1 \sim Uniform(a1^{para1}, a1^{para2});$$

$$b1 \sim Uniform(b1^{para1}, b1^{para2});$$

$$c1 \sim Uniform(c1^{para1}, c1^{para2});$$

$$\tau_{unsteady} = \frac{1}{\sigma^2_{unsteady}} \sim Gamma(\tau^{para1}_{unsteady}, \tau^{para2}_{unsteady});$$

where, J and I are the numbers of observations prepared for steady and unsteady state rating analyses respectively; *hysteresis_correction* refers to the degree of unsteadiness based on the Jones formula; $p(Q|\theta)$ is the likelihood function, and; '*para*' stand for parameter. Figure 4-38 illustrates the graphical expression of Eq. 4.6 for the hysteresis analysis at the Hotan station. The source code of *WinBUGS* is described in *Appendix-A*.

In the initial stage of modeling, the non-informative prior distributions were given, based on the general information such as the range of b under channel and section controls. The Uniform distribution was used for the parameters of e and b; and the Gamma distribution was employed for the parameters of a and τ to ensure sampling positive values. The parameters of $a1$, $b1$ and $c1$ were used and assumed to be distributed uniformly in the case of a hysteresis analysis. The optimized parameters (Table 4-3) were included in the support of the Uniform prior distributions.

Once the posterior probabilities of the parameters were obtained in the initial stage, they were fed into the model as the prior distributions at the next stages. When the priors were determined, only the types and the variances of the probability distributions were borrowed from the posteriors. Then, the optimal parameters at a current stage were determined using NLP, and considered the means of prior distributions. In Equation 4.6, the effective GZF (e) was particularly censored because stages less than any positive e are not theoretically allowed in the rating of $Q=a(h-e)^b$.

The detailed processes for rating analysis using the BMCMC are shown in Figure 4-37. The ratings were separated based on the transition zones, and the abnormally measured data were excluded on the basis of the 95% confidence limit of the posterior distribution for the discharges obtained at the previous time step. For smooth connection

of the rating segments, the data included in the transition zones were added in each data set for analyzing each rating segment. In order to utilize this procedure, Equation 4.6 was coded using the programming language of the *WinBUGS* Software. When the model was run, two chains of the parameters with 30,000 samples and 10,000 burn-in iterations were generated with two different sets of starting points to check the convergence to a stationary distribution. The MCMC diagnostics were based on the graph of history path, autocorrelation of the chains, the *Gelman-Rubin* statistic R, and correlation between the parameters. A sensitivity analysis for the priors was not performed.

1) Rating analysis at the Donghyang station

The main purposes of BMCMC analysis at the Donghyang station are: to trace the variation of the parameters, especially the effective GZF (e); and to derive the probability distributions of ratings for the purpose of screening the erroneous measurements. Hence, the results were mainly compared with the results of NLP, and the analyses were performed for the same data as applied to the NLP. Figure 4-39 through 4-42 show that the results by the two methods were quite similar. Compared to NLP, the BMCMC can provide the various confidence limits of the rating. Figure 4-43 shows the variation of the effective GZF (e) over time from the uniform prior probability before the August flood in 2003 to the posterior probability in 2005. The noteworthy facts from the results are: the variance of the parameter e was decreasing in time even though the mean moved. However, it was actually dependent on the degree of the measurement errors contained in the data, and; the posterior probability was coming close to the Normal distribution.

The MCMC diagnostics were described in *Appendix-A*. Generally, the various MCMC diagnostics for the segment under channel control were not stable compared to

those of the segment under section control. This might be caused by the high measurement errors of the data under channel control. The parameters of b and e did not converge well, however, a and τ converged satisfactorily on the whole.

2) Rating analysis at the Hotan station

The ratings were assessed for both low (steady state rating) and high (unsteady state rating) flows identified centering on the stage of 1.35m with consideration of the transition zone in Table 4-1. Unlike the case of the Donghyang station, the main purpose of the study at the Hotan station was to confirm whether or not the BMCMC could analyze the hysteresis using the Jones formula with the parameters of $a1$ and $b1$ in Equation 4.5. Hence, the analysis results were mainly compared with the 1-D unsteady flow analysis in 2004. The confidence limits of the rating by BMCMC can be essential for identifying whether the measured data are acceptable or not especially at a place where a hysteresis is pronounced. The parameters of $a1$ and $b1$ shown in Table 4-3 were used as the means of the prior probabilities in the initial stage, and $c1$ was always fixed to zero. Figure 4-44 shows that the BMCMC are more consistent with the observed data than the 1-D unsteady analysis.

4.4 Comparison of the Methodologies and Conclusions

Rating analyses were performed in both deterministic and probabilistic ways at the Donghyang and Hotan stations. For deterministic approaches, NLP, fuzzy rule-based modeling and one-dimensional hydrodynamic models were employed, while the BMCMC method was used for probabilistic modeling. For reflecting the hysteresis of rating, the Jones formula was used.

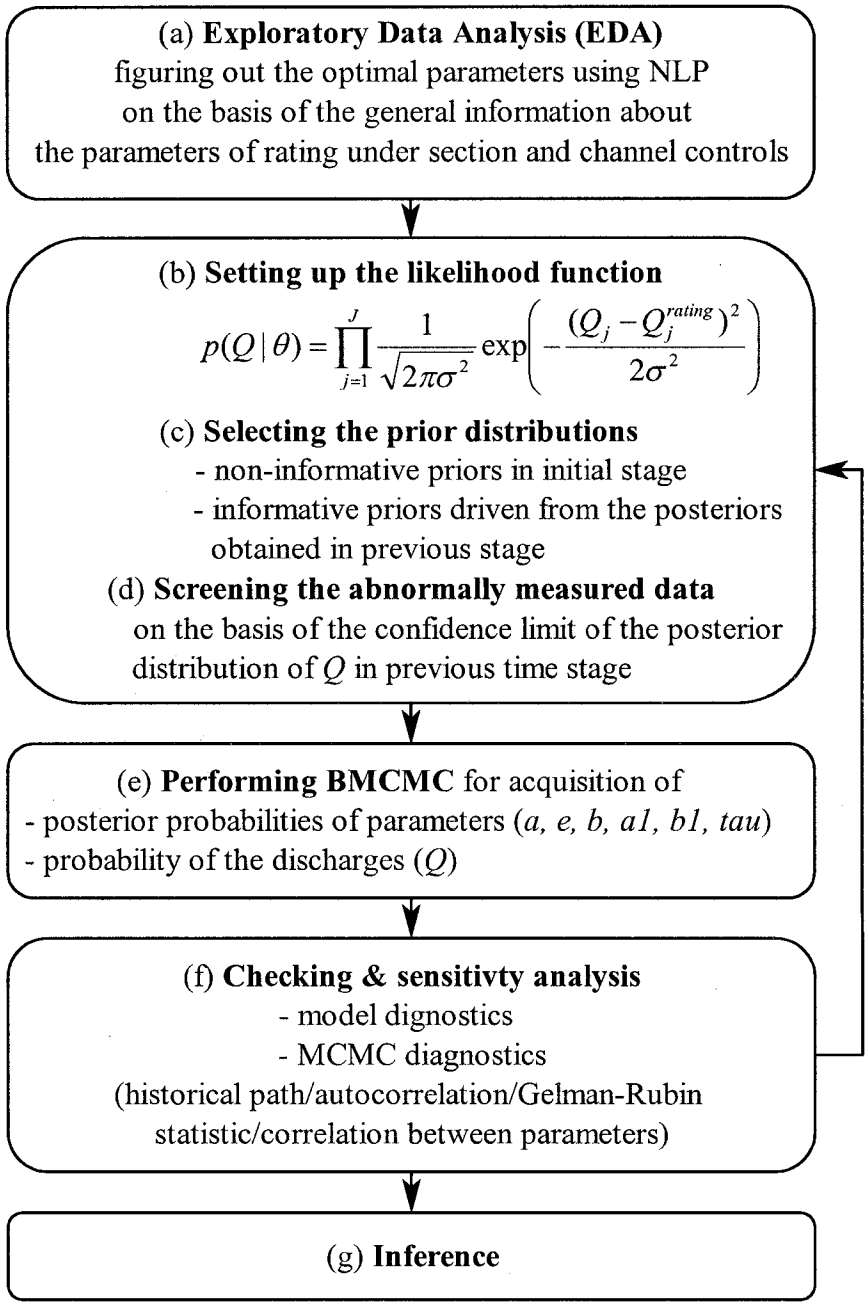


Figure 4-37 The procedure of Bayesian data analysis.

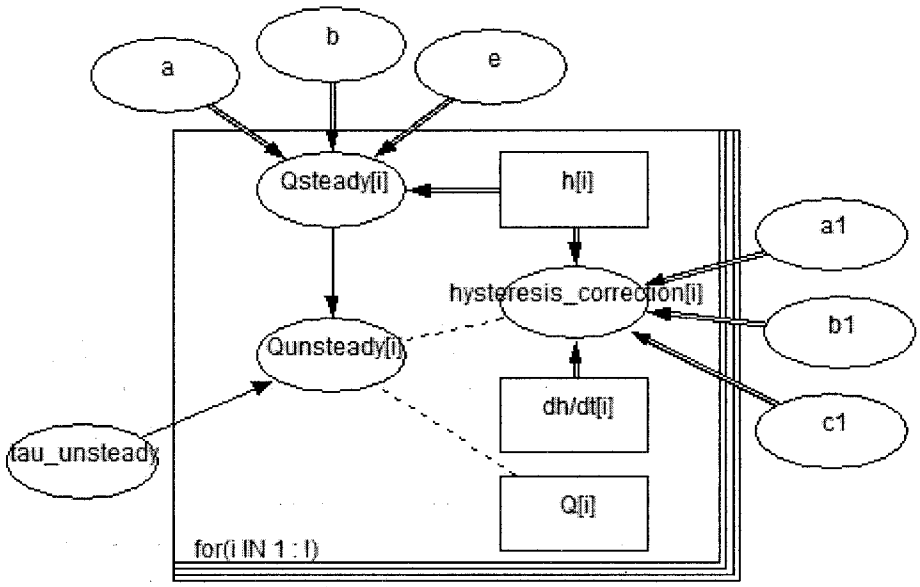


Figure 4-38 The graphical expression of the probabilistic modeling by BMCMC for analyzing the hysteresis at the Hotan station. *Qsteady* and *Qunsteady* stand for the steady state and unsteady state discharges respectively; *hysteresis_correction* is the degree of unsteadiness; *Q* is the measured discharges; *h* and *dh/dt* are the stages and stage variation rate in time, and; *tau_unsteady* is $\tau_{unsteady}$.

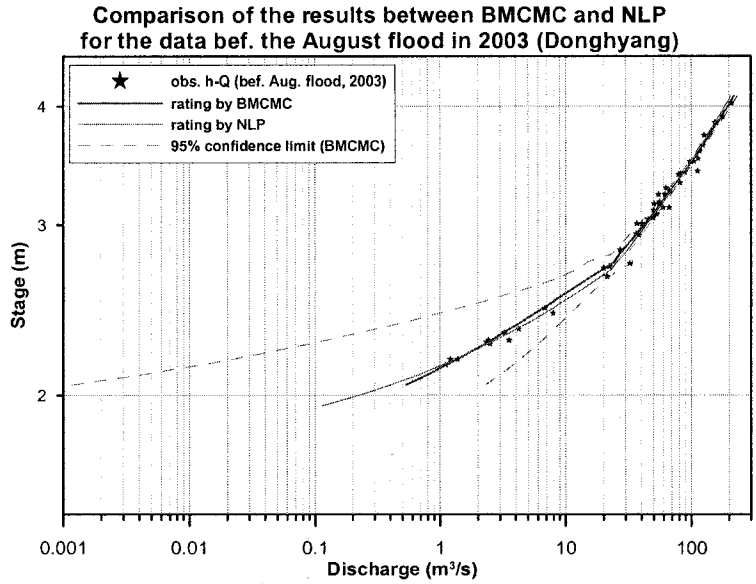


Figure 4-39 Comparison of the results between BMCMC and NLP at the Donghyang station for the data before the flood in August, 2003.

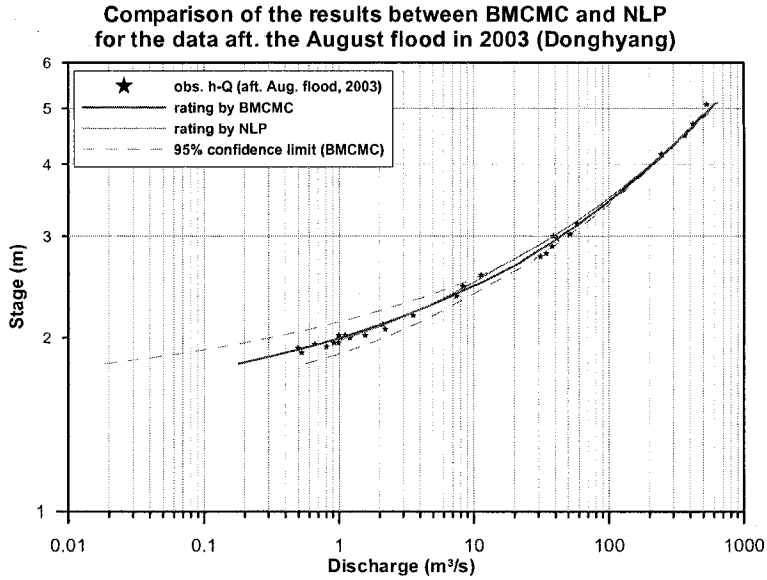


Figure 4-40 Comparison of the results between BMCMC and NLP at the Donghyang station for the data after the flood in August, 2003.

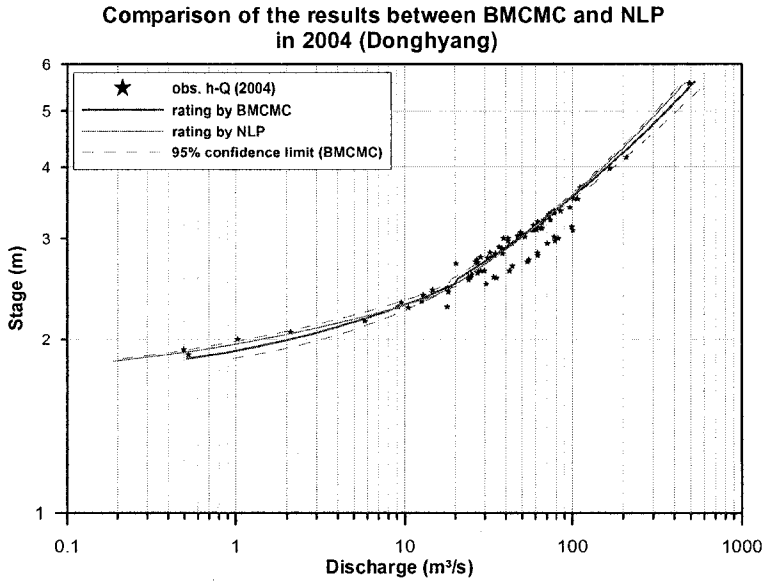


Figure 4-41 Comparison of the results between BMCMC and NLP at the Donghyang station in 2004.

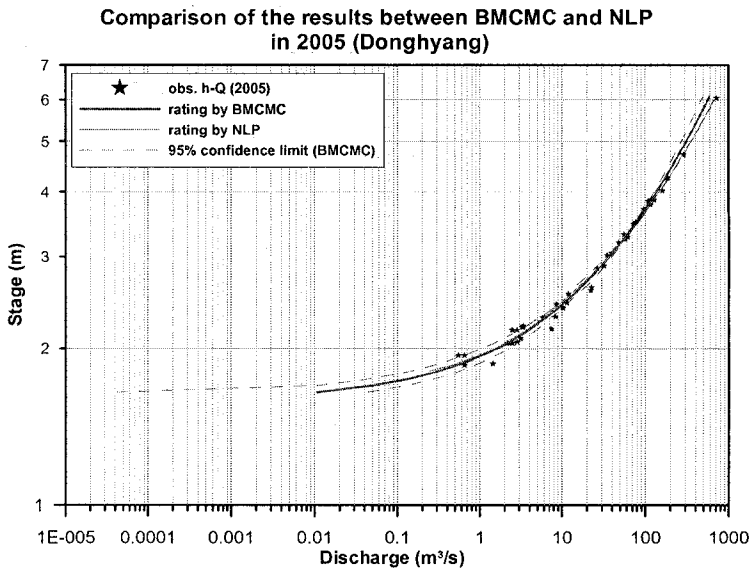


Figure 4-42 Comparison of the results between BMCMC and NLP at the Donghyang station in 2005.

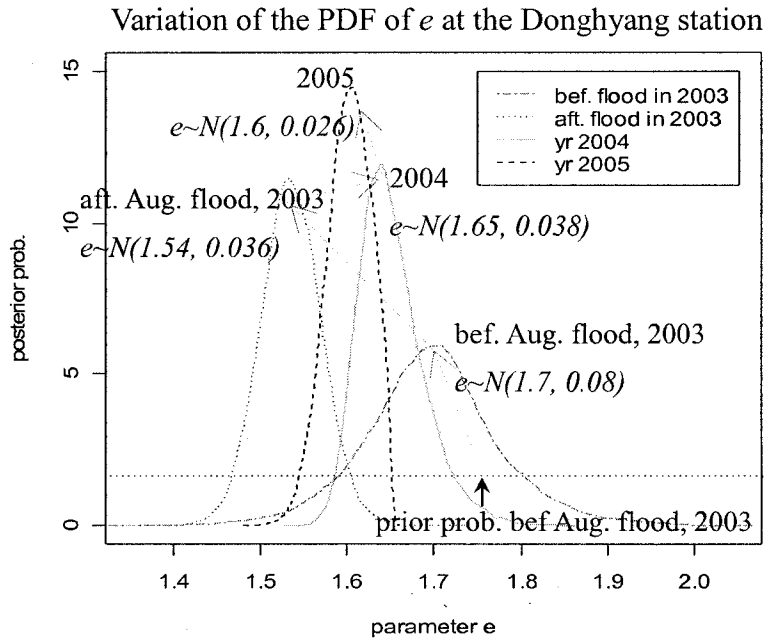


Figure 4-43 The variation of the parameter *e* at the Donghyang station.

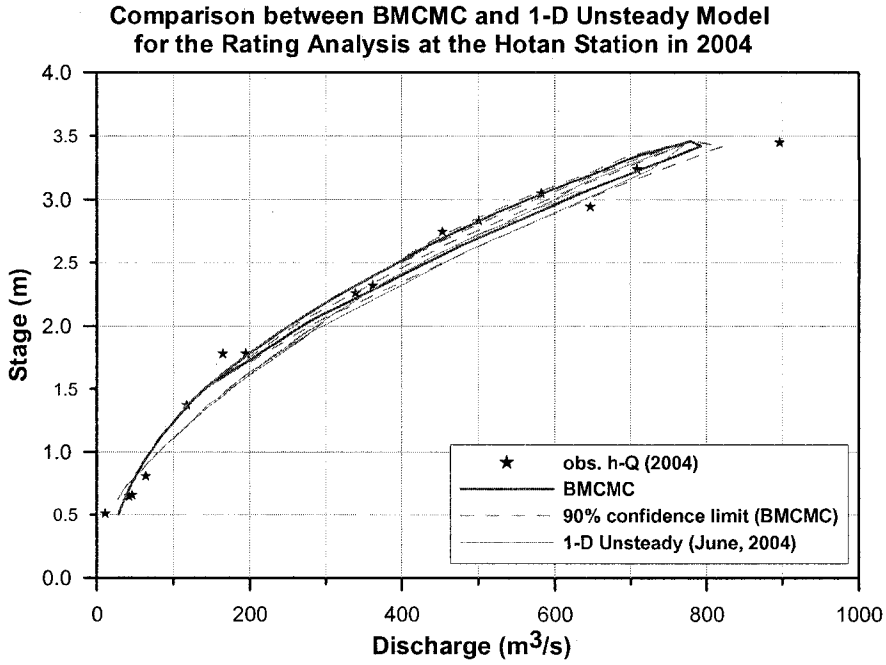


Figure 4-44 Comparison of the results between BMCMC and 1-D unsteady flow analysis at the Hotan station in 2004.

The comparison of the results was made at the Donghyang station in 2003 and at the Hotan station in 2004 where all four methods were used. For determining the most appropriate alternative in this study, the comparison of the models was based on both quantitative and qualitative aspects. However, although the quantitative comparison is very straightforward, this method has some problems in this case study: First, the goodness-of-fit depends on the degree of calibration. For example, the rating analysis by fuzzy rule-based modeling with many membership functions assures the least errors (see Figure 4-45); second, the measured discharge data are generally highly uncertain as reviewed in the previous section: hence, it is not so meaningful to assert that a method that is the best in terms of goodness-of-fit is the most appropriate because the hydraulic features of ratings might be lost.

Figure 4-45 and 4-46 show the comparison of the four methods at the Donghyang and Hotan stations respectively. In figure 4-45, the RMSE (root mean square error) for the 1-D unsteady flow analysis was not computed because the measured data did not match the values computed by the *HEC-RAS* exactly. In conclusion, NLP and BMCMC gave almost the same results, and the simulated values were consistent with the measured data on the whole. Meanwhile, it turned out that the fuzzy rule-based modeling and 1-D unsteady flow model showed very poor results around the low flow. If the Manning's roughness coefficients are calibrated more sophisticatedly by dividing a low flow domain into a couple of sections, this problem may be lessened. However, this may not be parsimonious in terms of the number of parameters. The features of each method can be summarized from a qualitative perspective as:

1. The remarkable advantages of the unsteady flow model are to be able to derive the loop rating almost exactly and to offer information for discerning the transition zones. If the cross-section data exist, the basic features of the rating should be examined by doing the unsteady flow analysis first. For extending existing ratings, this method provides a very reliable solution. As a drawback, this method may fail to represent the ratings for low flows under section control because of intermittent occurrence of numerical errors and divergence of the solution with low flows.

2. In the NLP method based on the classical rating analysis, it is very easy to reflect the hydraulic features such as the effective GZF, the slopes of rating segment and the changes of flow control in ratings. Furthermore, the process by this method is relatively simple. In order to take the hysteresis into account, supplementary methods such as the Jones formula should be employed. As the disadvantages of this method, it is difficult to obtain the uncertainties of both the rating and the generated flows.

3. The fuzzy rule-based modeling may be the most excellent in terms of the goodness-of-fit because it is capable of fitting to the data perfectly. This indicates that it is easy to get trapped into overtraining. As the typical disadvantages of this method: it is not easy to apply to real problems and embed into a decision support system for rating management because a finally trained FIS is a black box; it is difficult to get the information about uncertainties of the rating, and; the hysteresis may not be reproduced satisfactorily because the training of FIS requires lots of measured data containing information about the hysteresis, which is not generally available.

4. In contrast to the above three methods, the BMCMC works in probabilistic way. This method might be the most reasonable theoretically in that it can offer a rating with

its uncertainty and incorporate the prior information into a model. Furthermore, it can also reflect the hysteresis using the Jones formula. However, it is very difficult to apply practically. This might make engineers reluctant to apply it to real problems.

From the above reviewed features of the four methods, a hybrid method that combines the advantages of NLP and BMCMC could be recommended as a good alternative. After obtaining the posterior probabilities for the parameters using the BMCMC and screening the abnormal data, the rating can be refined using NLP with constraints based on the confidence limits of the parameters derived by the BMCMC analysis. Furthermore, this hybrid approach might be the best for evaluating ratings in case the measurement errors are high in that it can overcome its susceptibility to errors by introducing the prior information of the parameters.

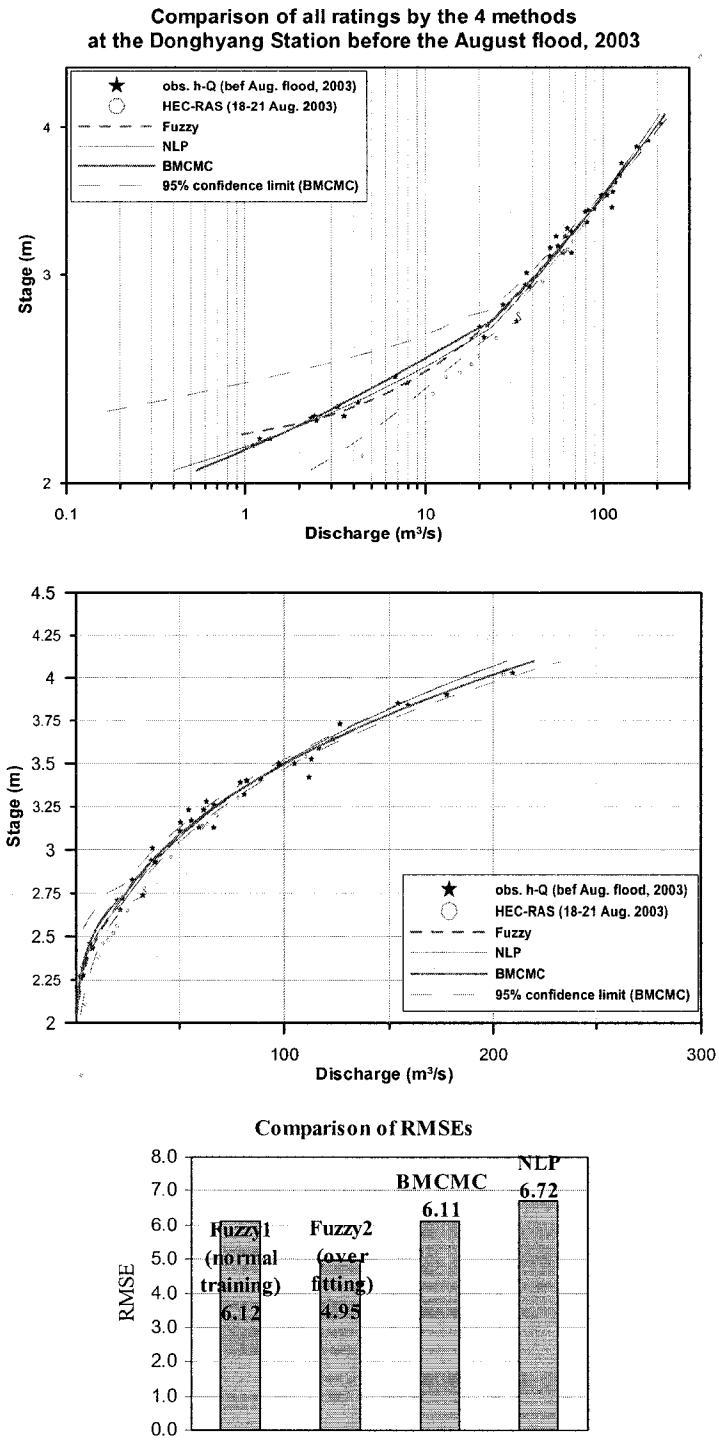


Figure 4-45 Comparison of all results by the 4 methods at the Donghyang station before the August flood in 2003 (log and normal scales), where Fuzzy1 and Fuzzy2 are the results by the *ANFIS* with normal training and overfitting (see Figure 4-31), and RMSE stands for root mean square error.

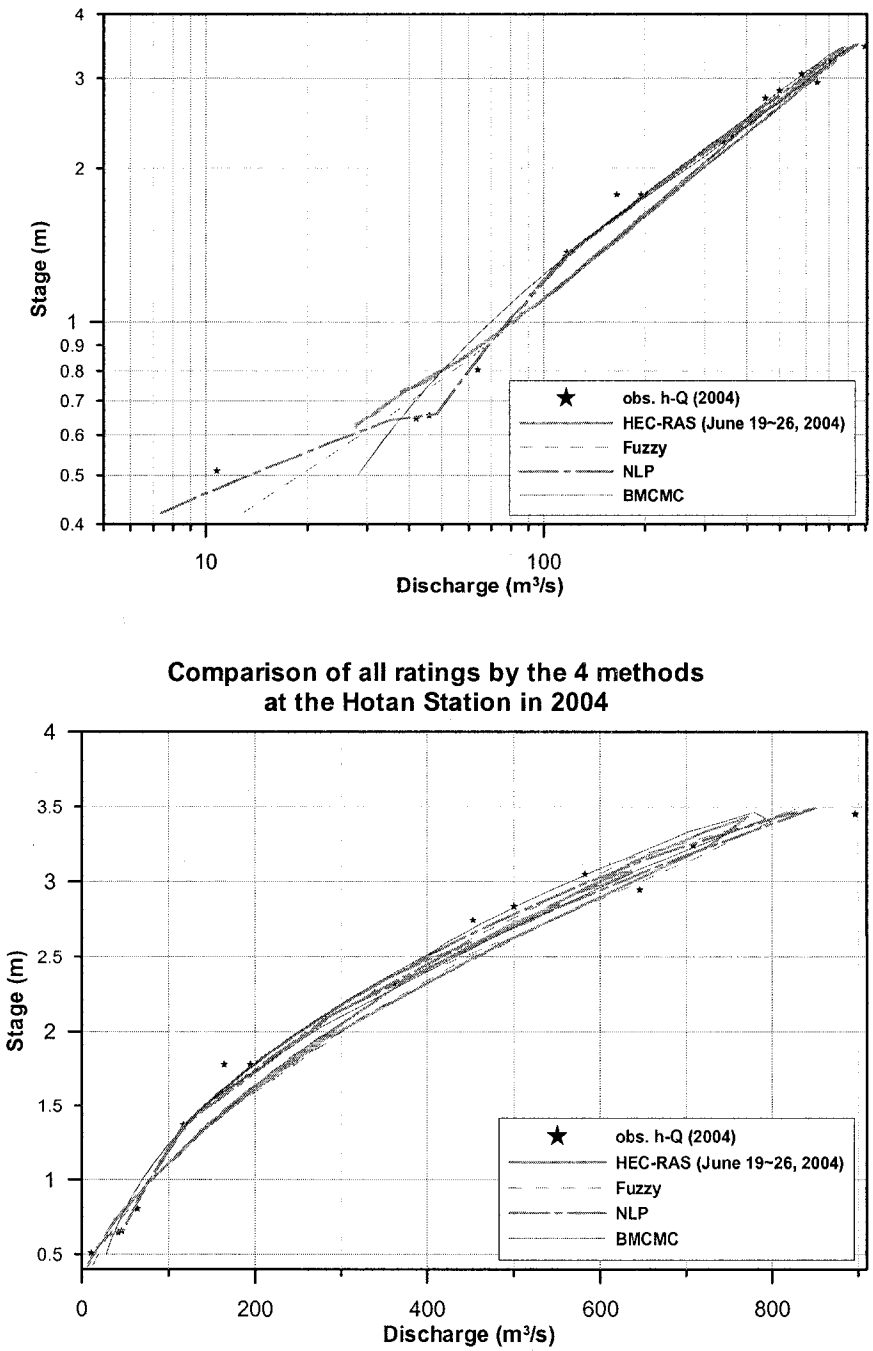


Figure 4-46 Comparison of all results by the 4 methods at the Hotan station in 2004 (log and normal scales).

CHAPTER 5

RESERVOIR INFLOW FORECAST USING STOCHASTIC ARTIFICIAL NEURAL NETWORKS WITH BAYESIAN MCMC AND NONPARAMETRIC METHODS

Hydrologic time series analysis is focused on both simulating long term synthetic flow sequences in streams or man-made reservoirs and forecasting the near future for controlling water resources systems. Elaborating on the differences between simulation and forecast of flow sequences, the simulation process tries to generate sequences that are statistically similar to the historical record *[Sharma et al., 1997]*, and goodness-of-fit of a model is assessed by the degree that marginal probabilities of synthetic sequences are close to those of observed sequences. On the contrary, forecast of future states is a process to come up with the expectations recursively from the conditional probabilities for a predefined lead time given the current situation. In addition, the forecast process can be updated successively in an adaptive way whenever new data set is obtained: hence, the success of forecast models is absolutely dependent on the degree that uncertainties in the forecast results are small during lead time. This is assessed after acquisition of new data sets. Figure 5-1 and 5-2 illustrate the above mentioned differences of the two approaches.

Because this research was motivated by risk-based and adaptive management of water resources, the study was performed with emphasis on developing probabilistic models. The objectives of the study addressed in this chapter are:

i) to develop monthly reservoir inflow forecast systems using: stochastic artificial neural networks with Bayesian Markov chains and Monte Carlo sampling techniques denoted as stochastic-ANN or ANN+BMCMC; and two nonparameteric time series modeling methods consisting of the kernel density estimate (KDE) and the *k*-nearest neighbor (*k*-NN);

ii) to determine a recommended model by comparing the models in terms of ease of application, goodness of fit, and ease of incorporation into a decision support system for water resources management;

iii) to determine whether or not *k*-NN modeling can be employed practically for daily inflow forecasts, and; whether or not it can provide an option compared to physical rainfall-runoff models for the short term operation of water resources systems.

The algorithms of these three modeling techniques are described in Chapter 3 in detail. The research was performed using a case study of the Chungju multipurpose dam and its reservoir located in the Han River basin. The Han River basin, which has the largest basin area in south Korea, is located at the center of Korean peninsular (36°N 30' ~ 38°N 55' and 126°E 24' ~ 129°E 02') and its area (26,219 km^2) accounts for about 26% of the whole area of South Korea (99,237 km^2). The length of the basin is 467.7 km. The Chungju multipurpose reservoir was constructed by *K*-water for the purposes of water supply for domestic use, industry, irrigation, power generation, and flood control.

The construction was initiated in June 1978 and finished in October 1986. The main features of the basin, dam and its reservoir are summarized in Table 5-1. The outline of the inflow forecast modeling is summarized in Table 5-2.

Table 5-1 Main features of the Chungju dam basin and reservoir.

	Description
Basin	(a) basin area : 6,648 km^2 (b) average inflow a year: 158.5 m^3/s (c) average rainfall a year : 1,111.4 mm (d) yearly amount of water supply : 3,380*10^6 m^3
Dam & Reservoir	(a) dam type : concrete gravity type (b) height (dam crest) : 97.5 m (EL.m 147.5) (c) length : 447 m (d) total storage capacity : 2,750*10^6 m^3 (e) flood control capacity : 616*10^6 m^3 (f) power generation capacity : 412.0*10^3 kw

Table 5-2 Summary of reservoir inflow forecast modeling.

	Monthly forecast	Daily forecast
Methods	Stochastic-ANN KDE *k*-NN	*k*-NN
Model architecture	Lag-1 Markov chain model	Lag-*p* to Markov chain model with exogenous variable
Lead time	1 month	1 day
Historical record period	1917~1940, 1956~2005 (74 years)	1986~2005 (20 years)
Programming language	Stochastic-ANN: *MATLAB*, *WinBUGS* KDE: "*R*" package *k*-NN: *VB-Excel*	*VB-Excel*

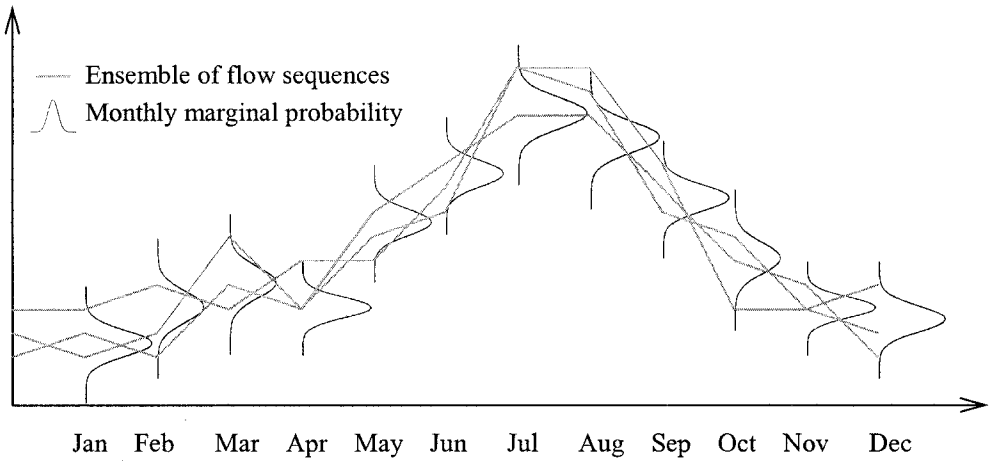


Figure 5-1 This shows an example of simulation of monthly flow sequences. With numerous generations of flow, the marginal probability of each month can be determined and the statistics of simulated and observed sequences are compared by month.

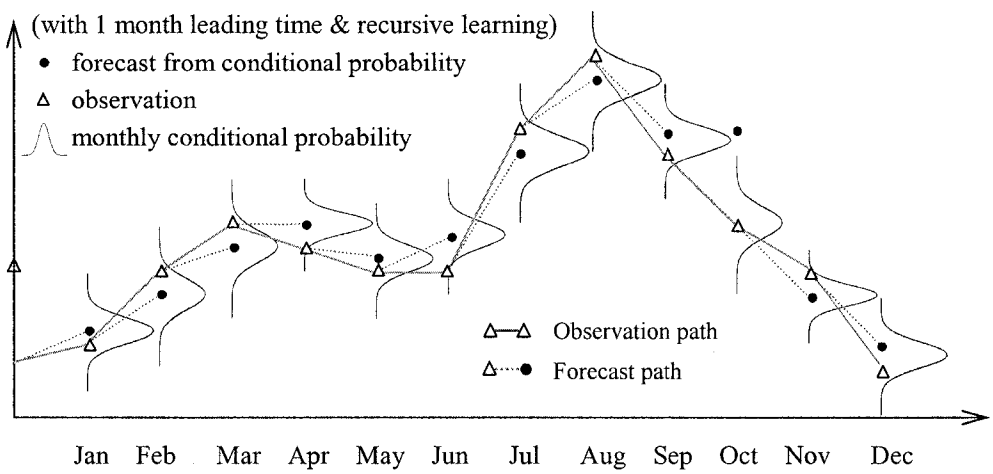


Figure 5-2 This shows an example of monthly flow forecast. With recursive learning of forecast models in every month, forecast is made recursively based on a conditional probability given the observations in previous time periods.

5.1 Monthly Reservoir Inflow Forecast for the Chungju Reservoir

5.1.1 Basic Statistics of the Monthly Inflow and Model Determination

As summarized in Table 5-2, the modeling for the monthly reservoir inflow forecasting was performed with a lag-1 autoregressive model using three methods. The forecast lead time is one month considering that observations are updated at the end of every month, that is, the forecast was updated every month. Figure 5-4 shows the autocorrelation function at this site, which indicates that the basin has very short memory in terms of temporal inertia of hydrologic features. From the autocorrelogram, it can be concluded that a time series model with lag-1 autoregressive term is sufficient to describe the temporal dependence relationships.

Figure 5-3 refers to the time series plot of the monthly inflow in the Chungju reservoir for 74 years from 1917 to 2005. Table 5-3 and Figure 5-5 show the basic statistics and box-whiskers plot of the historical monthly inflow in the Chungju reservoir. Figure 5-5 also shows temporally high variation of the hydrologic features with outliers beyond the 97.5 and 2.5 percentile inflows. Figure 5-6 shows the monthly marginal probabilities with kernel density estimators, which indicate this site has the probabilistic features of high variance and skewness, and an irregular distribution shape with multi-modality especially in July. Figure 5-7 shows the pairwise scatter plot in each month with smoothed lines by the locally weighted scatter plot smoother (LOWESS) that represents conditional expectations $E(x_t|x_{t-1})$. It indicates, along with the autocorrelation function in Figure 5-4, that the temporal dependence relationship of lag-1 is very weak except for the

pairs of January-February and November-December, and hydrologic events are highly random in time.

In order to evaluate the forecast ability of the three models, first of all, they were calibrated using 66 years of historical records from 1917 to 1997 with 3 years test data set from 1998 to 2000. The models were validated for 5 years of data from 2001 to 2005. For calibration, a trade-off analysis was employed between two contradictory factors such as bias-variance and the residuals between the calibration and test data sets. For statistical inference of the simulated inflows, 100 sets of 1,000 samples were drawn iteratively in every month in this study.

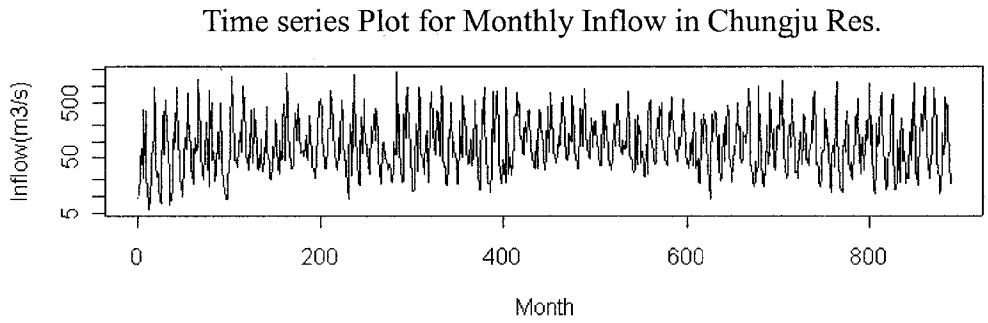


Figure 5-3 Time series plot of the monthly inflow in the Chungju reservoir for 74 years from 1917 to 2005.

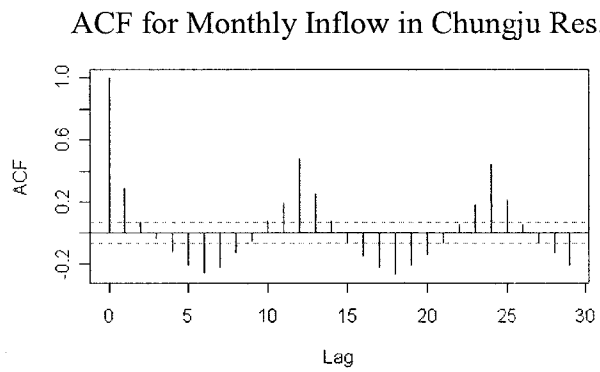


Figure 5-4 Autocorrelation function of the monthly inflow in the Chungju reservoir, used for identifying model architecture.

Table 5-3 Basic statistics of the monthly inflow in the Chungju reservoir.

statistics	Jan	Feb	Mar	Apr	May	Jun
mean	28.0	31.2	80.7	178.9	121.0	124.3
variance	217	704	4209	15928	6314	15980
skewness	0.88	2.85	2.18	2.00	1.06	2.31
statistics	Jul	Aug	Sep	Oct	Nov	Dec
mean	559.6	386.1	292.9	86.0	54.4	39.7
variance	135894	92181	62148	4088	1005	477
skewness	1.13	1.84	1.66	2.04	1.72	1.29

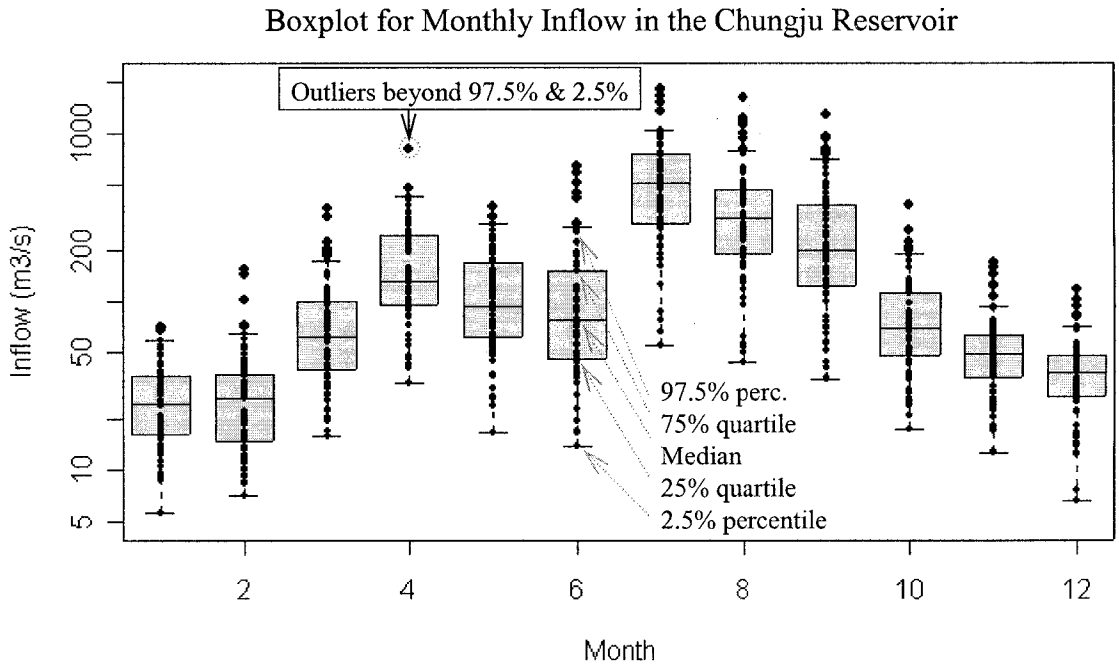


Figure 5-5 Box and whiskers plot of the monthly inflow in the Chungju reservoir for 74 years.

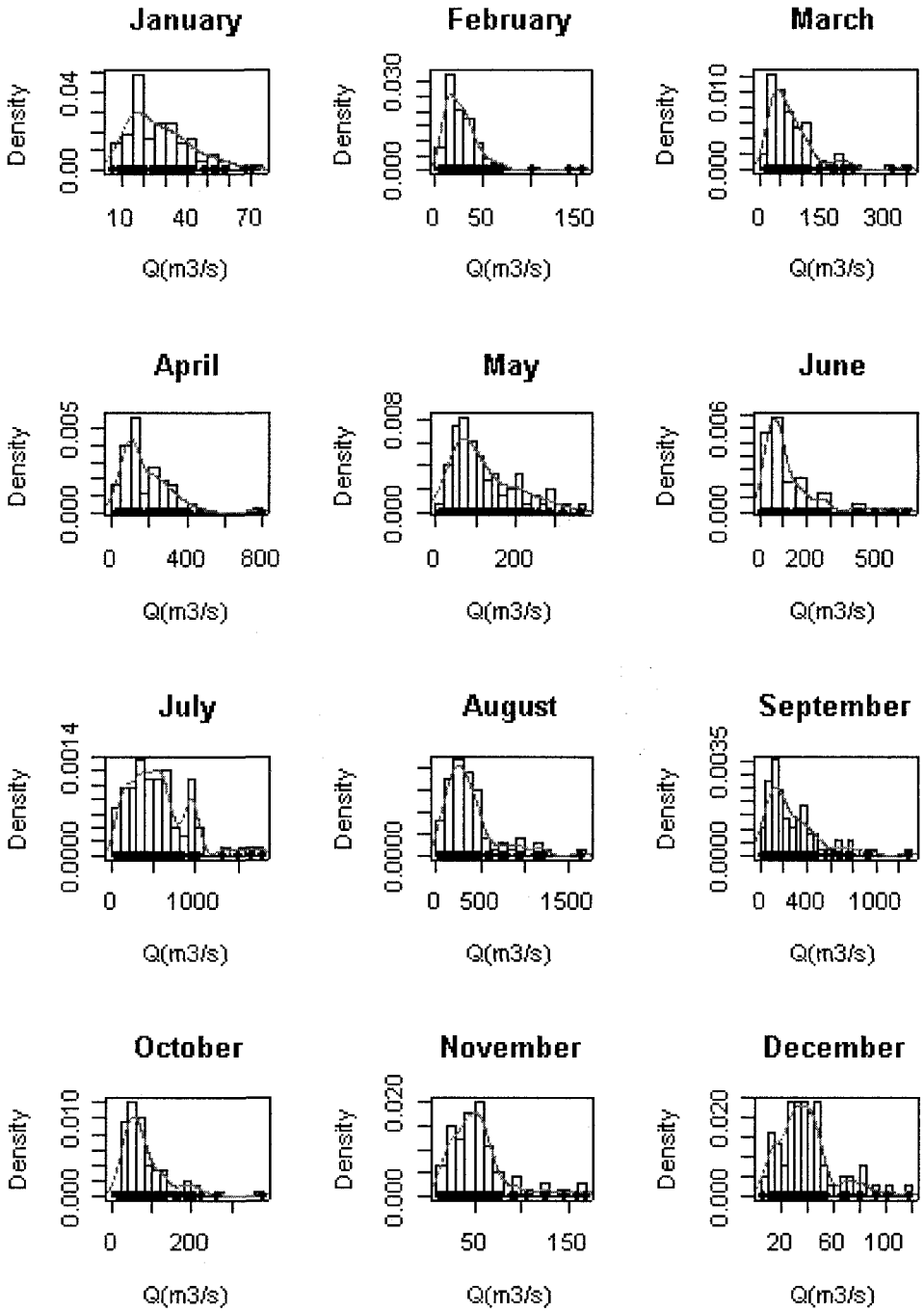


Figure 5-6 Histogram with Kernel density estimators of the monthly inflow in the Chungju reservoir for 74 years.

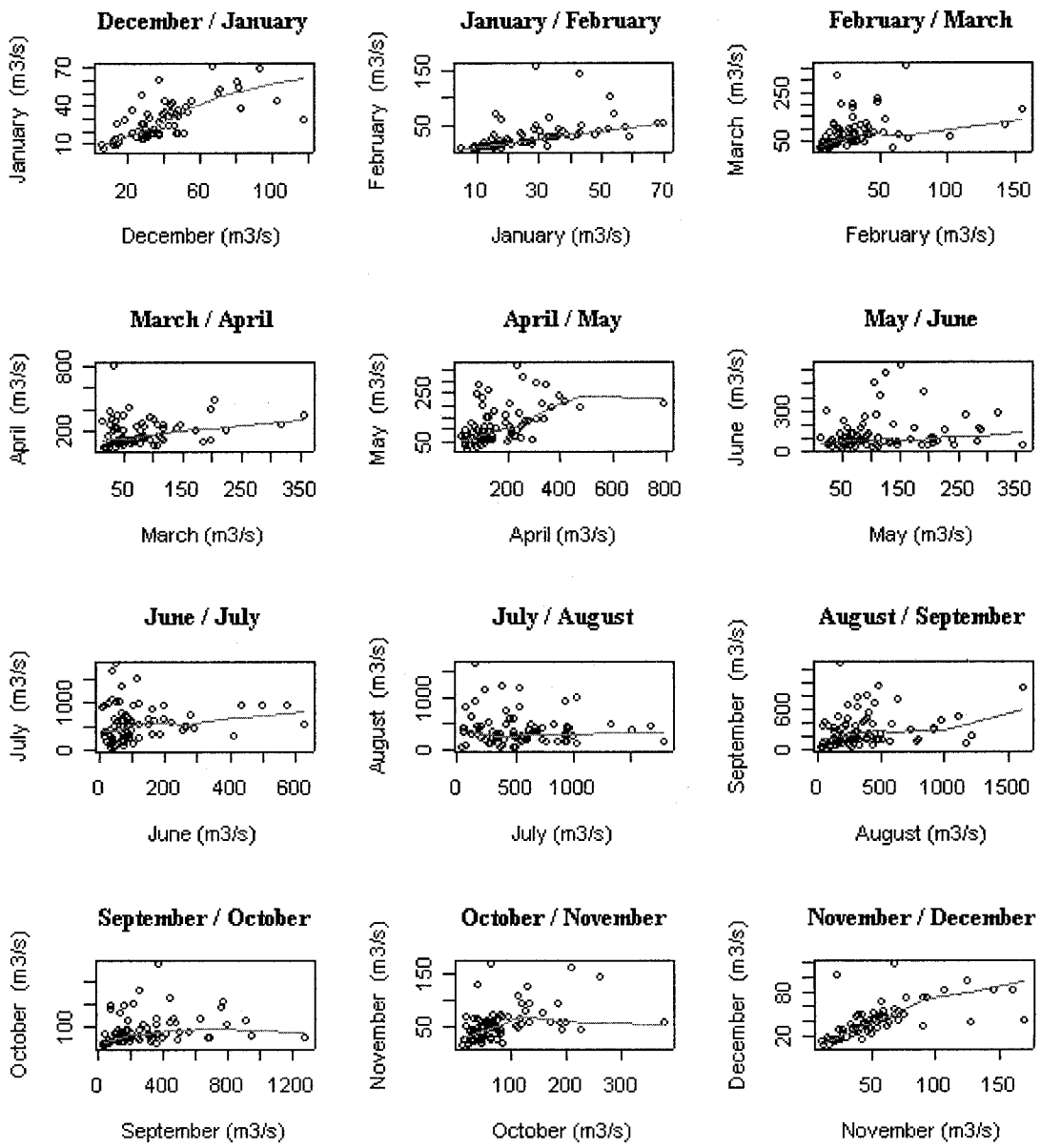


Figure 5-7 Pairwise scatter plot of the monthly inflow in the Chungju reservoir for 74 years with LOWESS (the locally weighted scatter plot smoother).

5.1.2 Monthly Reservoir Inflow Forecast Using a Kernel Density Estimate

As reviewed in Chapter 3, the basic idea of time series modeling using a KDE is to derive a conditional probability density function (Equation 3.8) from joint and marginal PDFs using kernel functions. In this study, a lag-1 forecast model using a kernel density estimate is developed for the monthly inflow in the Chungju reservoir. The KDE forecast model was developed using the "*R*" statistical package.

Figure 5-9 shows the flowchart for the forecast modeling using a KDE. As it is the same in every modeling approach, a significant amount of work may be required for calibration of model parameters. The parameters of the forecast model with KDE are the scaling factors (λ) of band width by month. However, the monthly values were into a single lumped scaling factor in this study. According to *Silverman (1986)*, the bias-variance trade-off plays an important role in choice of band width, which indicates that the smaller the band width is, the less the bias is and the higher the variance is because this can result in a rough density estimate. In this respect, the model can be calibrated over the band width matrix H by adjusting λ in Equation 3.18 to minimize the least square cross validation (LSCV) function in Equation 3.20. As another approach, the asymptotic optimal λ in Equation 3.19 that *Silverman (1986)* gave can be applied.

Because the study put an emphasis on the inflow forecast, the model was calibrated using a cross-validation technique with a test data set by adjusting λ around the asymptotic optimal λ ($\approx 0.494 = 69^{-1/6}$). First, the available historical data were divided into three subsets as shown in Table 5-4: calibration (or training) set, test set and validation set. The errors on the test set were computed along with those on the

calibration set. The bias-variance trade-off was reviewed with the errors between the expectation of simulated inflows and observed ones regarded as biases. Figure 5-8 shows the calibration result and bias-variance trade-off with the biases and variances of two data sets normalized to the same scale of 0 to 1. Contrary to general expectations, the relationship of tradeoff between bias and variance was not consistent over a λ value of 0.4. However, it can be noticed that the mean errors of the calibration set and test set maintain a distinct trade-off relationship over a λ range of 0.2 to 0.7, and the error of the test set increases steeply beyond λ of 0.7. In this respect, the parameter $\lambda = 0.4$ was chosen finally, on which the two lines of the calibration and test errors cross and the variances are smallest.

Once the parameter λ was determined, the next step is to review the marginal probability for each month and the bivariate joint probability for each pair of sequential months. While this step is not necessarily required in the modeling process, it helps understand the shapes of all conditional probability distributions in terms of multimodality and skewness. Figure 5-10 shows the bivariate joint probability in contour and surface plots with a scatter plot of the historical record in December and January, determined using Equation 3.21 and 3.22. In the graph, the smoothed line represents the conditional expectations $E(x_t|x_{t-1})$ of the historical record, which are estimated by LOWESS. The rest of the monthly marginal and joint probabilities are described in Appendix B.

After reviewing the marginal and joint probabilities, the remaining process is to determine the conditional probability distribution for every month, and to sample from it. For statistical inference of the simulated inflows, 100 sets of 1000 samples were drawn

iteratively in every month using Equation 3.23, and the statistics such as mean and variance were computed. During sampling, any negative simulated inflows were discarded: hence, the conditional probability may be censored. Figure 5-11 shows the conditional probability in January, 2001 with mean of 19.7 m^3/s, conditioned on the inflow of 14.44 m^3/s in December, 2000. This conditional probability is actually a slice, which is formed by cutting the joint probability distribution in Figure 5-10 along the conditional inflows in December. The same process was repeated sequentially to the end of the forecast period. The final results were compared with those by the other models.

Table 5-4 Data sets for calibration and validation.

Calibration data set	1917 ~ 1997 (67 years)
Test data set	1998 ~ 2000 (3 years)
Validation data set	2001 ~ 2005 (5 years)

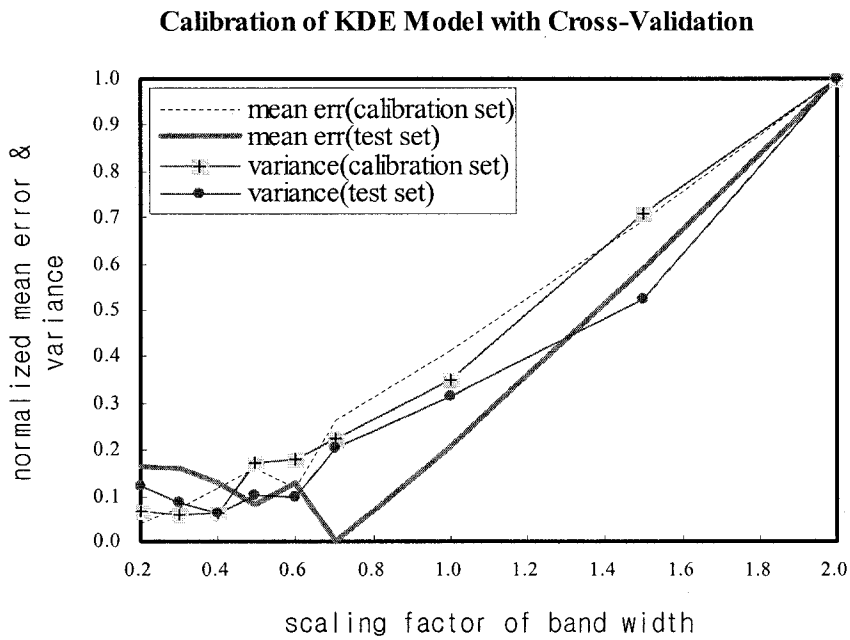


Figure 5-8 Calibration of the KDE forecast model over λ, where 'mean err' means the averaged error between the observed and computed inflows.

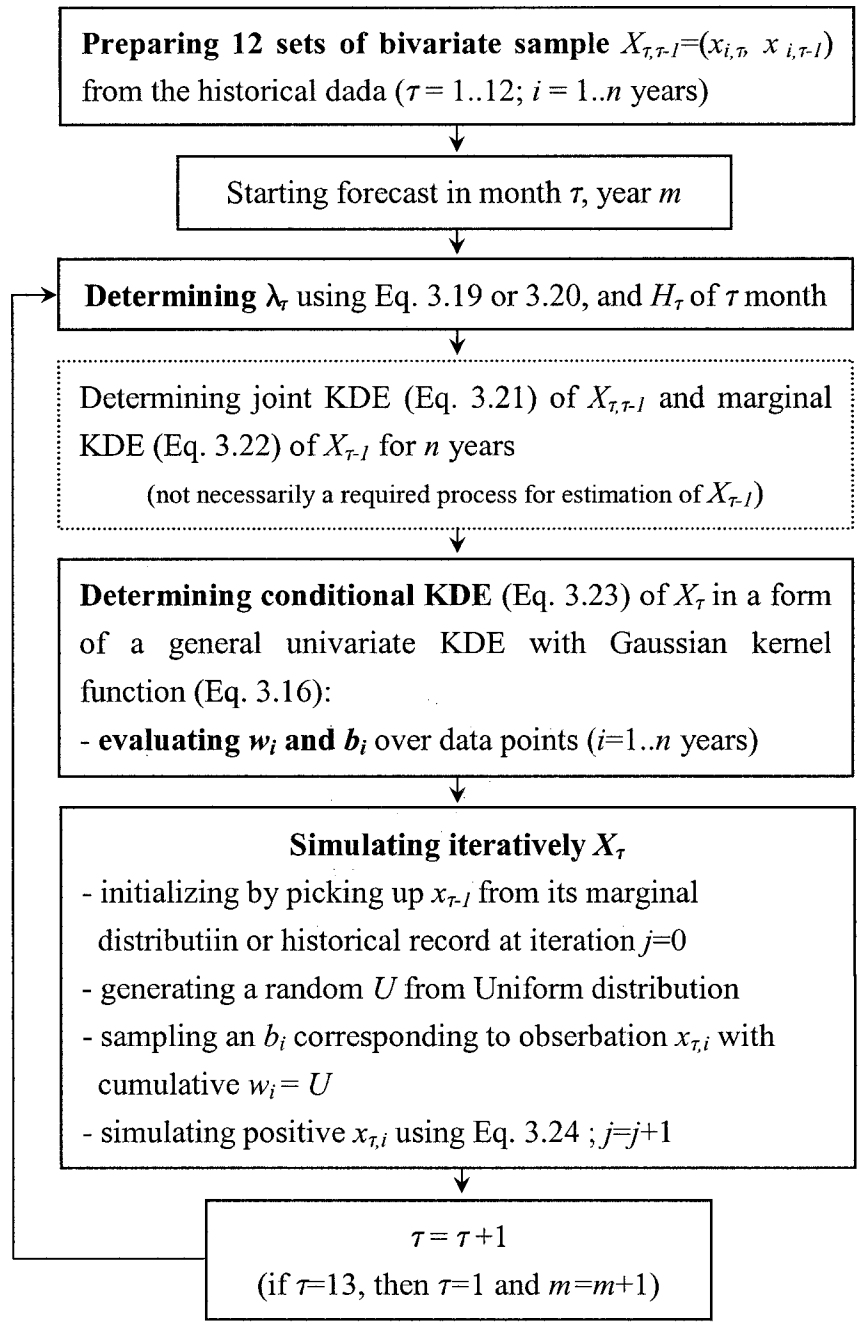


Figure 5-9. The flowchart for lag-1 monthly forecast model with KDE, where τ, i, m, and j mean index of month, year into the historical record except the forecast data set, year of the forecast data set, and sampling iteration number respectively.

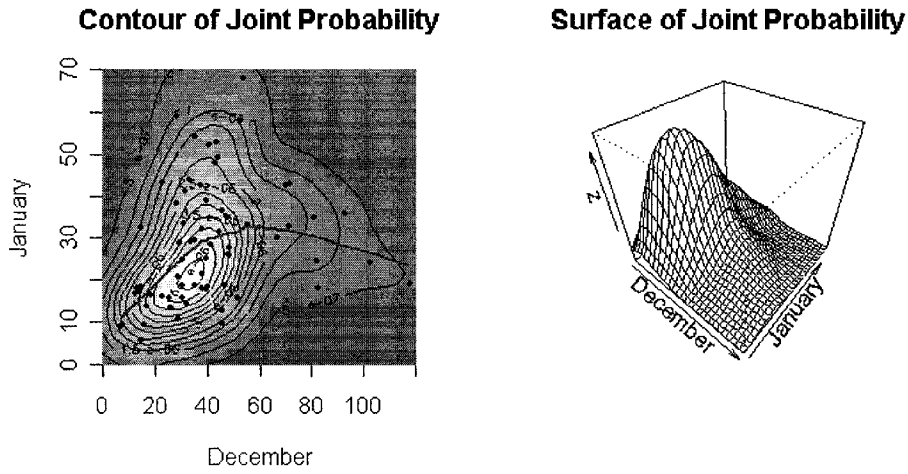


Figure 5-10 Joint probability of the inflow in contour and surface with a regressed line by the robust locally weighted regression and smoothing scatter plots (LOWESS) that represents the conditional expectations $E(x_t|x_{t-1})$ in December and January in the Chungju reservoir, determined using KDE.

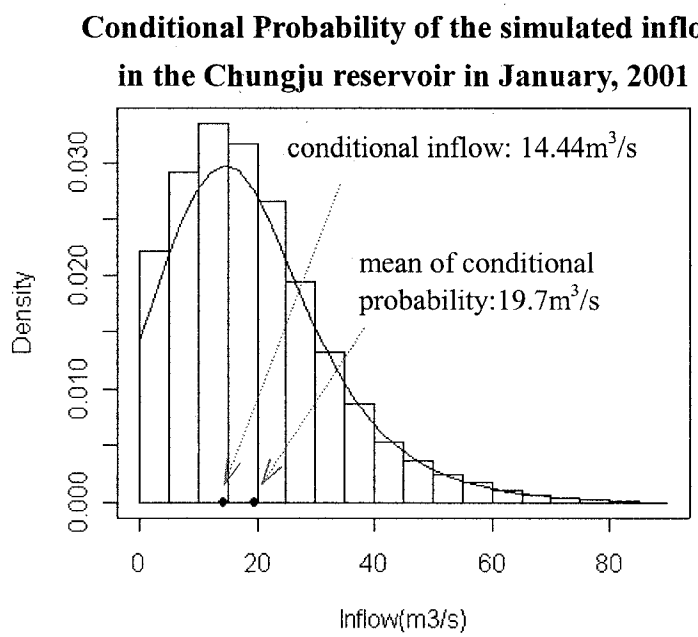


Figure 5-11 Conditional probability of the forecasted inflow determined using a KDE in January, 2001 in the Chungju reservoir, conditioned on the inflow of 14.44 m^3/s in December, 2000.

5.1.3 Monthly Reservoir Inflow Forecast Using a *k*-NN Density Estimate

As reviewed in Chapter 3, the basic idea of time series modeling using a *k*-NN method is to find the historical nearest neighbors of a current feature vector and resample with replacement from their successors using bootstrap and kernel functions representing the extent of similarity of the neighbor to the current feature vector. In this study, a lag-1 forecast model using a *k*-NN density estimator was applied and compared to the historical record. The same modeling processes and conditions such as grouping the historical record into three sub data sets for calibration, test and validation of the model as those for the KDE model were followed. The kernel function $K(\cdot)$ described in Equation 3-14 was used here. The *k*-NN forecast model was developed using Visual Basic for Applications (*VBA*) built into *MS-Excel*.

Figure 5-12 shows the flowchart for programming a forecast model using a *k*-NN density estimate. Like the KDE modeling, the calibration was performed over the model parameter first, the only parameter of the model is *k*. Likewise, a lumped *k* parameter was applied equally in every month in this study. The model was calibrated using cross-validation techniques by adjusting *k* from 5 to 60 around the *ad hoc* prescriptive choice of $k = n^{1/2}$ (≈ 8), where *n* means the total number of the data points of a monthly calibration data set (66 years from 1917 to 1997). Figure 5-13 refers to the calibration result and bias-variance trade-off with the biases and variances of two data sets normalized in the same scale of 0 to 1. Unlike the KDE model, the relationship of bias-variance trade-off was consistent, and the biases of two data sets show a good trade-off. From Figure 5-13,

it can be noticed that striking a balance between the bias and variance of the calibration set happens around k of 15: hence, the parameter k was determined to be 15 in this study.

The results by k-NN were compared with those of the KDE and the stochastic ANN (described in the next section). As one may guess, the remarkable advantage of k-NN modeling is that it is very easy to implement in computer programs. This fact indicates that it can be incorporated and assimilated directly into a decision support system for water resources management. This idea was embodied and applied to the reservoir optimal joint operation in the Daecheong and Yongdam reservoirs in the next chapter.

5.1.4 Monthly Reservoir Inflow Forecast Using Stochastic Artificial Neural Networks with BMCMC

Artificial neural networks have attracted great attention along with fuzzy rule-based modeling in simulation and forecast of hydrologic, hydraulic and environmental variables including rainfall-runoff, water qualities such as BOD and DO, and time series analysis of precipitation, water level and stream flow *[Maier and Dandy, 2000]*. In general, ANN has been applied in practice with the backpropagation training algorithm and multi layer feed forward network.

The challenges in applying ANN may be summarized as avoiding trapping in local minima, improvement of convergence speed, avoidance of overtraining, identification of the ANN structure, and incorporation of uncertainty into ANN models.

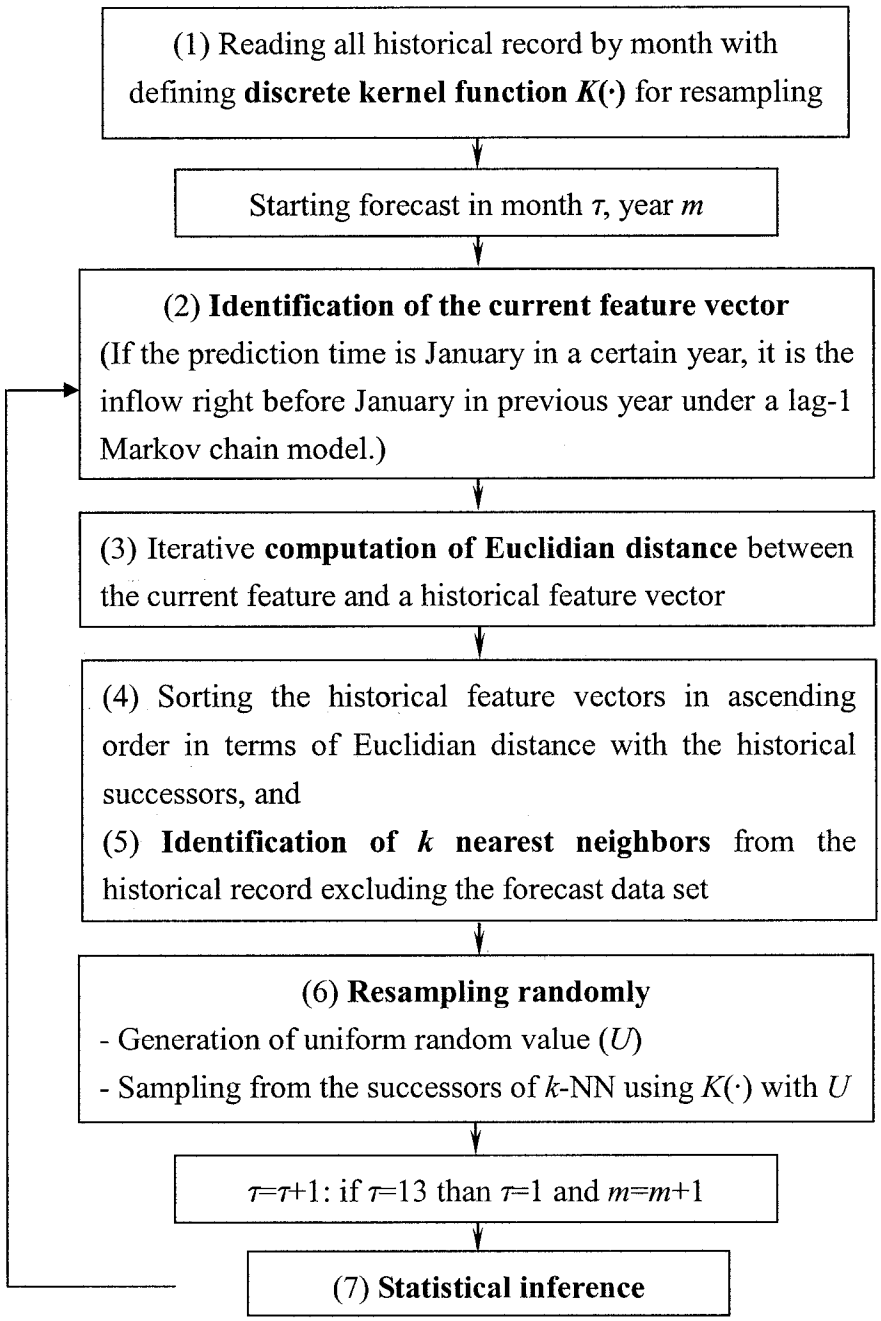


Figure 5-12 The flowchart for programming the forecast model using k-NN density estimate, where τ and m mean index of month and year of the forecast data set.

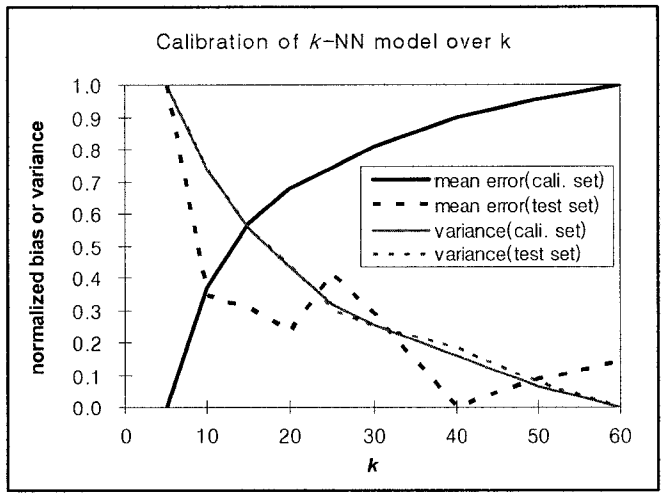


Figure 5-13 Calibration of k-NN forecast model over k with cross-validation technique.

Among the above difficult challenges, the biggest issues may be the first and last ones in terms of technical difficulty. In order to come up with solutions to the first problem, a lot of efforts of employing global optimization techniques instead of local search algorithms have been made, and the genetic algorithm has gained popularity as a tool for global optimization *[Hassoun, 1995]*. As for the last matter, ANNs in the field of water resources have been almost exclusively deterministic until now. The most promising method may be to employ a Bayesian Markov chain Monte Carlo (BMCMC) sampling technique in the training process to address network weights randomly.

In this study, the monthly forecast model was developed and evaluated using conventional ANNs and BMCMC as a training tool. This method is parametric,

nonlinear, and Gaussian if the residuals are assumed to be distributed normally. This is true since the nonlinear models are expressed by ANNs with the weights of nodes, and the likelihood function is composed of parameters from a certain type of probability density function. In this context, the stochastic ANN models are completely contrary to the nonparametric approaches. One may doubt that stochastic ANN models make a difference as compared with conventional time series models such as ARMA. However, conventional time series models do deal with their parameters deterministically while stochastic ANN models do.

Figure 5-14 represents a flowchart for establishing a stochastic ANN forecast model. Like nonparametric modeling, 12 sets of bivariate samples, $X_{\tau,\tau-1} = (x_{i,\tau}, x_{i,\tau-1})$ were prepared from the historical dada ($\tau = 1..12$; $i = 1..n$ years). Then, a deterministic ANN forecast model was developed using the *MATLAB* neural network toolbox for the purposes of determining the best architecture of the ANN and the optimal weights of the ANN by month. Once the network architecture and the weights were obtained, the weights were fed into a stochastic ANN as the means of the prior probability distributions for the network weights assuming a normal distribution. In this study, a stochastic ANN forecast model was developed by month using the *WinBUGS* described in Chapter 4. The sampling process for statistical inference was the same one used for the other two nonparametric models.

(1) Establishment of a deterministic ANN model

In this study, the deterministic ANN model was designed with a feedforward neural network composed of one hidden layer. The number of input and output nodes

should be one as long as a first order Markov chain is employed. The types of transfer functions in the hidden and output layers were fixed to the tangent sigmoid and linear functions respectively. For a training algorithm, the scaled conjugate gradient algorithm known to be very good at general purpose training was employed, which is one of many algorithms installed in the *MATLAB* neural network toolbox. The training epoch size was fixed to 5,000, and stopping criterion for generalization of the ANN was early stopping within the epoch size. The connection weights were initiated randomly. The model performance was measured using mean error (ME). The training data were not normalized considering complexity entailed in dealing with normalization in the stochastic ANN model.

The most strenuous processes of the deterministic ANN model may be determining the number of hidden nodes and training the connection weights between networks in every month. The number of the hidden nodes (regarded as the maximum number of nodes) was determined only for the ANN model in December (τ=12) using a cross-validation technique for simplicity. Finally, they were tuned by month. Figure 5-15 and 5-16 refer to the calibration results of the ANN model in December with five different numbers of hidden nodes. The MEs were normalized in the same scale of 0 to 1 for clear comparison between training and testing. Figure 5-16 shows that as the number of hidden nodes increases, the ME of the training data set decreases. However, the ME of the test data set increased suddenly over four hidden nodes. In this respect, four hidden nodes should be chosen normally to implement the first order Markov chain model for the Chungju reservoir. However, the number of hidden nodes was limited to one or two in this study to provide easy execution of the stochastic ANN

with the BMCMC approach and to create a parsimonious model. In the end the network architectures illustrated in Figure 5-17 ended up with one hidden node in January, April, May, June and September, and two in the other months. The details in association with the deterministic modeling are summarized in Table 5-5.

(2) Establishment of a stochastic ANN model

The same network architectures as the deterministic model were applied for the stochastic ANN model. Considering the lead time of one month, the model was developed independently in every month without forming a Bayesian network where all months are linked sequentially. However, it would be more convenient to establish a Bayesian network than monthly individual models in the case of lead time more than one month. Compared to the KDE and *k*-NN models, the main difference is that the stochastic ANN model employed temporally distributed parameters in that the parameters were treated individually by month, while the KDE and *k*-NN models were based on monthly lumped parameters.

The monthly stochastic ANN models with 2 hidden nodes were run for generating the inflow series with the help of Bayesian MCMC. Assuming the observations are independent, and the residuals between the observed (Q^{obs}) and simulated inflow (Q^{ANN}) are normally distributed with mean *0* and variance σ^2, the posterior probabilities of the parameter vector $\theta=\{w_1, w_2, w_3, w_4, b_1, b_2, b_3\}$ defined in Figure 5-17 can be described by combining the likelihood function and the prior probability distributions of θ using Bayes' theorem. Once the posterior probabilities of

the connection weights are obtained, the posterior predictive distribution of the inflows can be generated with the validation data set.

Equation 5.1 represents the detailed structure of the model, where n is the number of years in the historical record; $p(Q|\theta)$ is the likelihood function of the parameters; '*Normal*', '*Gamma*' and '*para*' mean the Normal, Gamma distributions and their parameters respectively. The parameters of w_i and b_i were assumed to be distributed normally, and the parameters trained in the deterministic ANN model were considered as means, which indicates the informative priors. The Gamma distribution was employed for the variance of the likelihood function with uninformative priors to ensure sampling positive values.

In order to embody this modeling, Equation 5.1 was coded using the programming language installed in the *WinBUGS*. When the model was run, two chains of the parameters with 20,000 samples and 10,000 burn-in iterations were generated with two different sets of starting points to check the convergence to a stationary distribution. Figure 5-18 refers to the pairwise scatter plot of the training results by the deterministic ANN model and the stochastic ANN model in each month. It shows that the means generated by the stochastic ANN model reproduced the results by the deterministic ANN model almost equally. The stochastic model shows a difference from the deterministic model in that it offers probability distributions. The MCMC diagnostics were based on the graph of history path, autocorrelation of the chains, *Gelman-Rubin* statistic R, and correlation between the parameters. The MCMC diagnostics were performed for the model in December when the structure of ANN

model was determined, and described in Appendix B. These show that the parameters of w_2 and b_2 are highly correlated: hence, this leads to very poor convergence. The detailed results were compared with those by nonparametric models in the next section.

$$\varepsilon = (Q^{obs} - Q^{ANN}) \sim Normal(0, \sigma^2);$$

A. *likelihood function* :

$$p(Q \mid \theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(- \frac{(Q_{i,\tau}^{obs} - Q_{i,\tau}^{ANN})^2}{2\sigma^2} \right) : \tau = 1...12; \qquad (5.1(a))$$

$$Q_{i,\tau}^{ANN} = w_3 \frac{1 - e^{-2\left(w_1 \times Q_{i,\tau-1}^{obs} + b_1\right)}}{1 + e^{-2\left(w_1 \times Q_{i,\tau-1}^{obs} + b_1\right)}} + w_4 \frac{1 - e^{-2\left(w_2 \times Q_{i,\tau-1}^{obs} + b_2\right)}}{1 + e^{-2\left(w_2 \times Q_{i,\tau-1}^{obs} + b_2\right)}} + b_3;$$

B. *prior probability* :

$$\begin{aligned} w &\sim Normal(w^{para1}, w^{para2}); \\ b &\sim Normal(b^{para1}, b^{para2}); \\ \sigma^2 &\sim Gamma(\alpha, \beta); \end{aligned} \qquad (5.1(b))$$

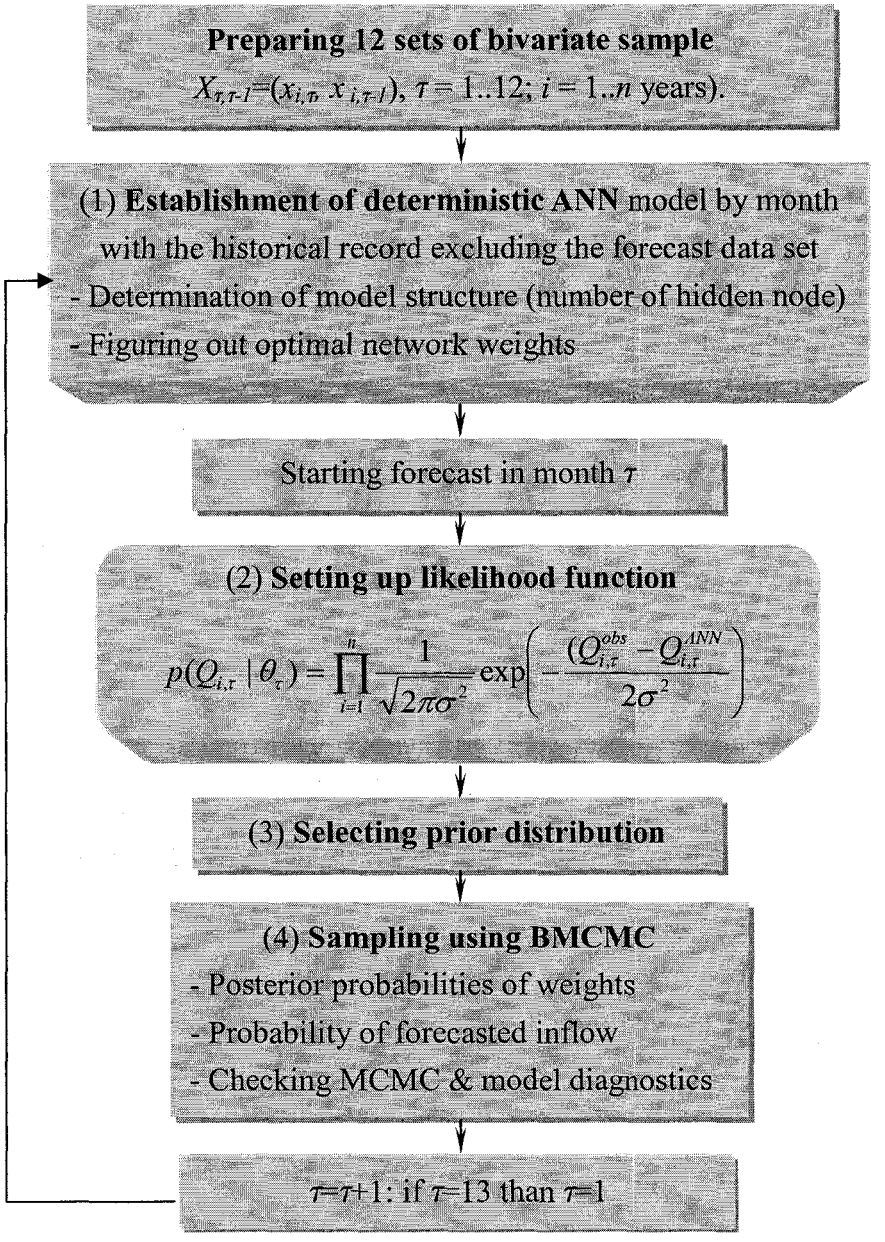


Figure 5-14 The flowchart for establishing a stochastic ANN forecast model, where τ, i, and m mean index of month, year of historical record except the forecast data set, and year of the forecast data set respectively.

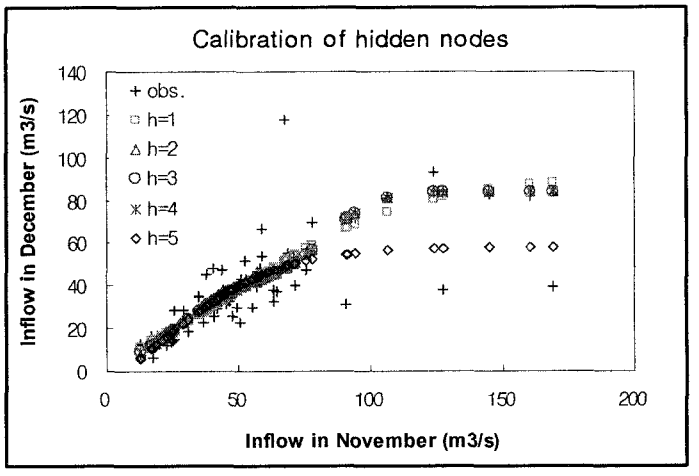


Figure 5-15 Calibration of the ANN deterministic model over the number of hidden nodes using cross-validation in December.

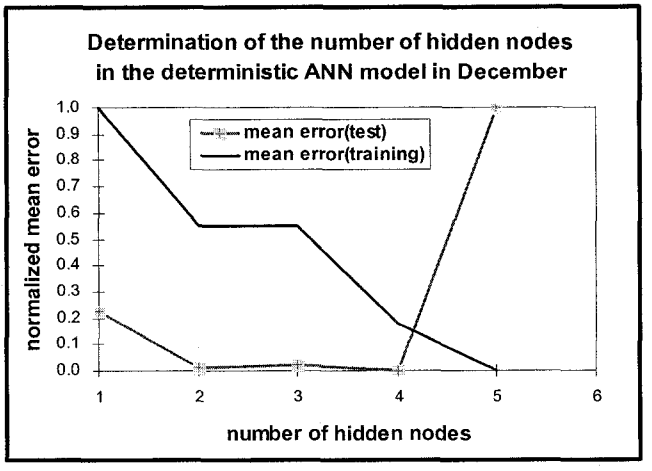


Figure 5-16 Determination of the deterministic ANN model structure over the number of hidden nodes in December.

Table 5-5 Configuration of the deterministic ANN forecast model.

Network architecture			Training	Transfer Function		Epoch size	Stopping criterion	Data
type	Number of hidden layer	Number of hidden nodes		Hidden layer	Output layer			
FF	1	1 (Apr, May, June, Sep) 2 (the other months)	BP	Tangent sigmoid	linear	5000	Early stopping	No norm

FF: feed forward network, BP: backpropagation training algorithm, norm: normalization

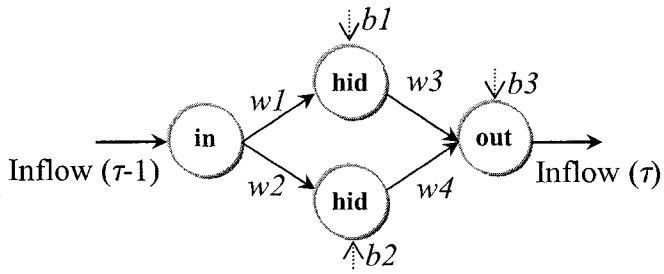


Figure 5-17 Network architecture in the case of 2 hidden nodes with weights, where "in", "hid" and "out" mean input, hidden and output nodes respectively.

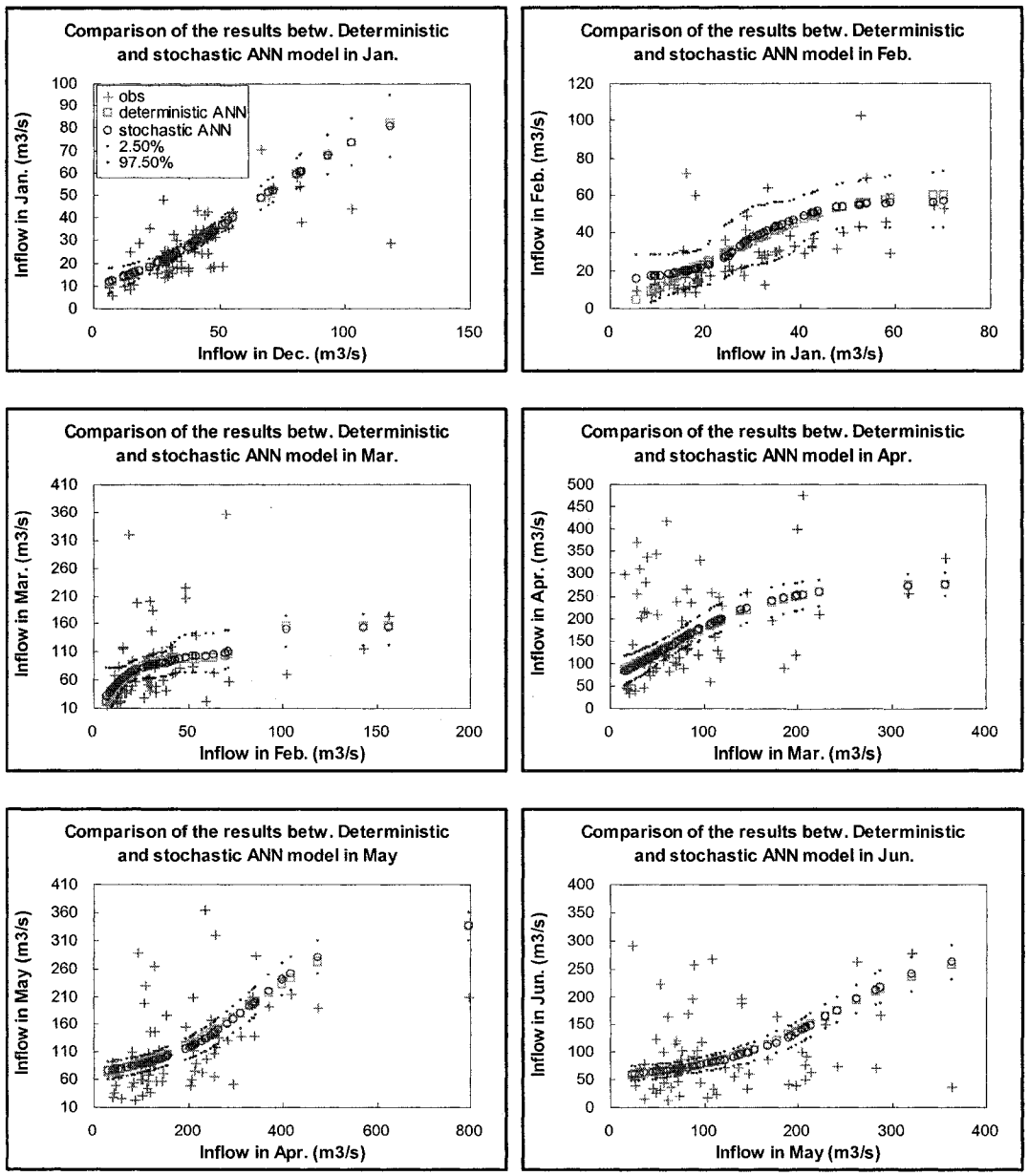


Figure 5-18 Monthly pairwise scatter plot for the training results by the deterministic ANN model by *MATLAB* toolbox and the stochastic ANN model by *WinBUGS* during training period (1917~2000).

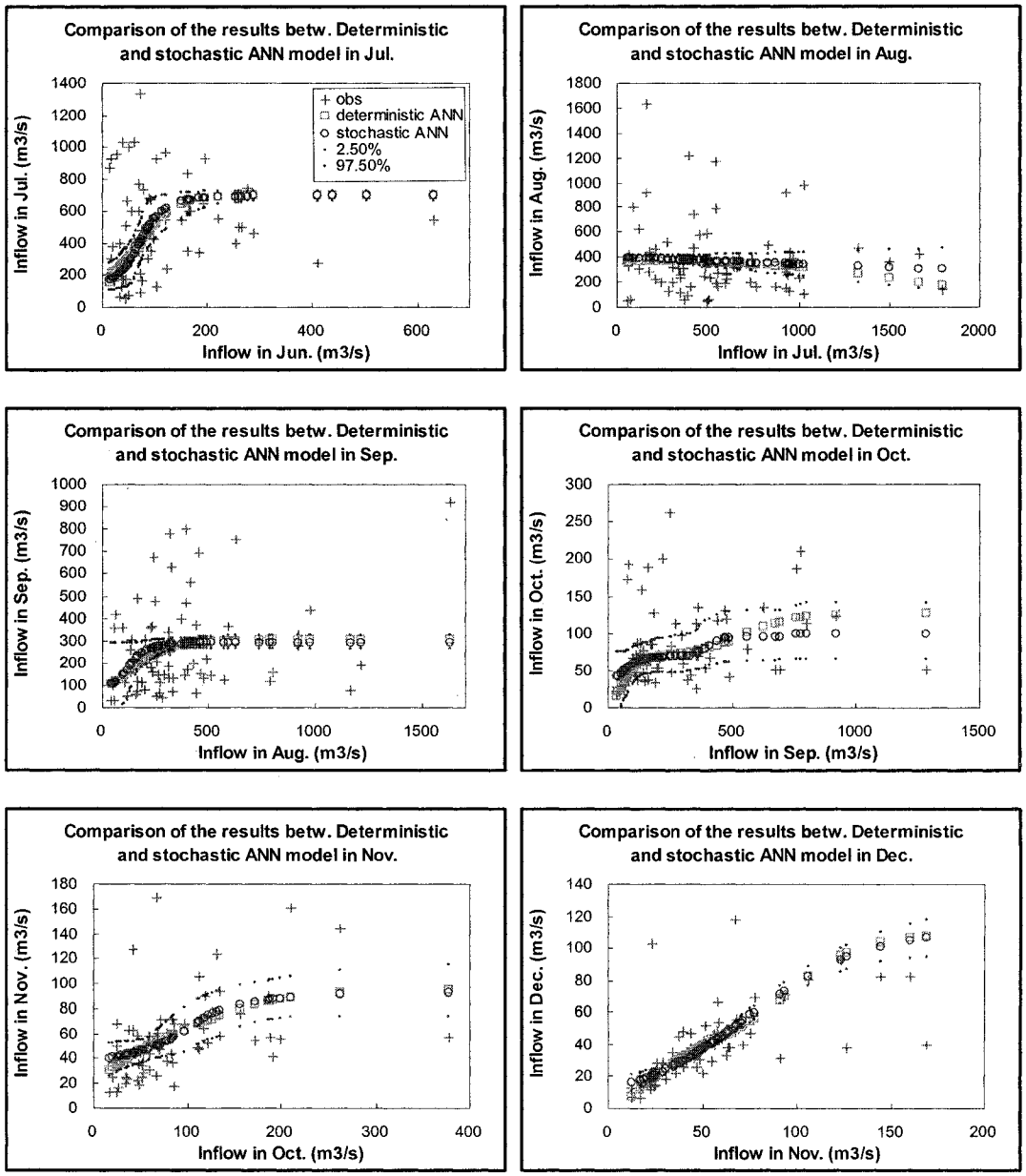


Figure 5-18 (continued.)

5.1.5 Comparison of the three monthly forecast models

Monthly reservoir inflow forecasts were developed using both nonparametric and stochastic ANN approaches for the Chungju reservoir where the historical monthly inflow data existed for 74 years from 1917 to 2005. For the methods of nonparametric modeling, the kernel density estimate and k-nearest neighbor density estimate were applied. For the stochastic ANN modeling, both conventional ANN and the Bayesian Markov chain Monte Carlo technique as a training tool were coupled in the *WinBUGS* package. The architecture of forecast model was limited to the first order Markov chain. As the first step of the modeling, calibrating or training the parameters of λ of KDE, k of k-NN and the weights (w_i, b_i) of ANN came first before using the training data set from 1917 to 1997 on the basis of cross validation techniques with the test data set from 1998 to 2000. Lastly, the model was validated with the validation data set from 2001 to 2005 in order to make sure of the forecast ability for future hydrologic events.

Figure 5-19 shows the calibration results of three models, where the mean calibration error by the k-NN model is the smallest, and the stochastic ANN and KDE models follow in order, while the mean test error of k-NN model is the highest. This indicates that the forecast errors by the k-NN model would be the highest. Figure 5-20 refers to the time series plot for the forecast results by three methods.

Figure 5-21 refers to the comparison of the forecast errors by the three models in terms of mean error. Figure 5-22 shows the scatter plot between the historical record and the generated inflow during the validation period with the slopes and the coefficients of determination that measure the linear relationship of two variables. Figure 5-23 shows the

conditional probabilities of the forecasts by the three models in 2001 with the marginal probability of the observations. From these results, the remarkable features and differences among the models can be identified as follows:

i) KDE and the stochastic ANN show smoothed results, while *k*-NN responds more roughly because there is no perturbation, which was pointed out by Sharma *et al.* (1997). The results of KDE and the stochastic ANN are so similar as to be difficult to discern any differences (Figure 5-20);

ii) In terms of model performance from Figure 5-21, it turned out that the stochastic ANN had the best forecast ability, and *k*-NN model was the worst as expected from the calibration results in Figure 5-19. Considering the slope and coefficients of determination of perfect linearity are equal to one, the stochastic ANN and KDE model show better results than *k*-NN model in Figure 5-22. However, it should not be asserted that *k*-NN model is always worse than two other models because the structures of models were different in that the parameters of the stochastic ANN model was monthly distributed, while those of the KDE and *k*-NN models were temporally lumped;

iii) In Figure 5-23, the nonparametric methods show irregular probability distributions with multi-modes, while the stochastic ANN model almost shows a normal distribution with single mode, and; the conditional probability by the KDE model is much closer to those of the marginal probabilities of the observations in general. However, this does not mean that the KDE model is the most superior in terms of reproducing the marginal probabilities because the results were not based on long term simulation but on only 5 years;

iv) As a common feature that was expected, Figure 5-7 shows very highly random hydrologic characteristics and that the three methods failed to forecast the abnormal hydrologic events such as severe floods and droughts satisfactorily (Figure 5-20).

On the basis of the above mentioned review, the main features of each model can be summarized in qualitative aspects as follows:

i) For the KDE model, the remarkable advantage is that it can smooth over the gaps between the data points in the density estimate by adding the perturbations to the picked observations with random generation from a standard normal distribution (Equation 3.24) as shown in Figure 5-20. In the context of simulation of inflow for a long time, it can also provide inflow realizations that are different but are similar to the historical record *[Sharma et al., 1997]*. As a drawback, it may be relatively difficult to develop required computer programming in the case of high order Markov chain models such as daily forecast models. Therefore, it may be difficult to incorporate into a data driven decision support system;

ii) Unlike KDE, *k*-NN is dependent on bootstrap without perturbation, which leads to rough density estimation. Consequently, it reproduces stream flow reservoir inflow by picking observed data with a conditional probability by discrete mass functions. As a great advantage of *k*-NN, it does not necessitate any sophisticated and complicated computation process, and it can represent any high order Markov chain model, which make coupling it into a decision support system easy;

iii) The stochastic ANN model with BMCMC differs in that: it is a parametric approach, and; it does not produce inflow directly from the observations using a conditional probability density function. Instead, it generates inflows probabilistically using a nonlinear function that expresses conditional relationships among variables while treating the model parameters such as node weights and residual variance of likelihood randomly. On the contrary, the nonparametric models resample directly from the historical record using a conditional probability estimated by a kernel function, and their parameters are treated deterministically. Compared to the nonparametric methods, a great advantage of this approach is although the historical records are not sufficient; this method can work well because it relies on a function that reflects the relationships between inputs and outputs. As for the drawbacks, the stochastic ANN model almost yields the normal distribution with single mode as the parametric models do generally. It is relatively difficult to embed it into an integrated decision support system because it uses special tools or software packages such as the "*R*" and *WinBUGS*. Such packages may not be easy for water resources engineers to learn. A computational problem may arise in case lead time is greater than one month. This can be solved effectively by forming a Bayesian network where all months are linked during the lead time with the weights of ANN. However, it may not easy to handle this problem in networks because of the need to consider many parameters. For instance, there are 84 parameters (7 parameters/month * 12 months) in the case of an architecture with 2 hidden nodes and lead time of one year.

In the context of practical application of models, one may say that water resources engineers might prefer *k*-NN considering the above comments about each method. Although *k*-NN showed the worst results among the three models in terms of goodness-of-fit in Figure 5-21 and 5-22, the ease of application may eclipse its drawbacks. It was practically applied to the reservoir optimal joint operation in the Daecheong and Yongdam dams, and described in the next chapter.

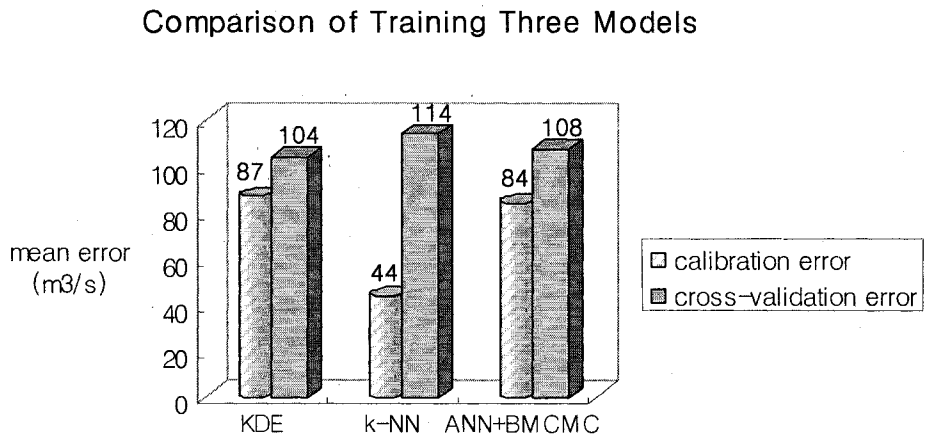


Figure 5-19 Comparison of the calibration results of the three models: KDE, *k*-NN and stochastic ANN with the validation data set from January, 2001 to December, 2005 in the Chungju reservoir.

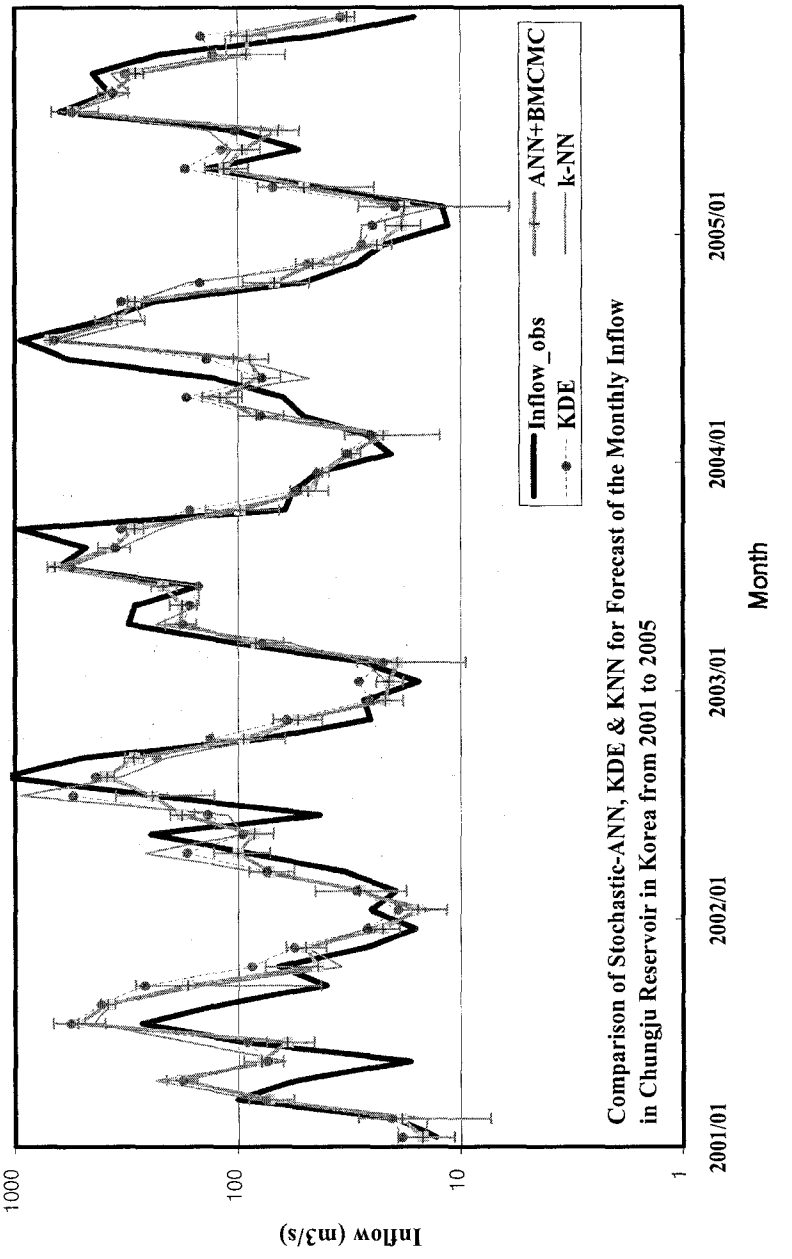


Figure 5-20 Comparison of the validation (forecast) results by KDE, *k*-NN and stochastic ANN with the validation data set from 2001 to 2005 in the Chungju reservoir, where the error bars belong to the stochastic ANN.

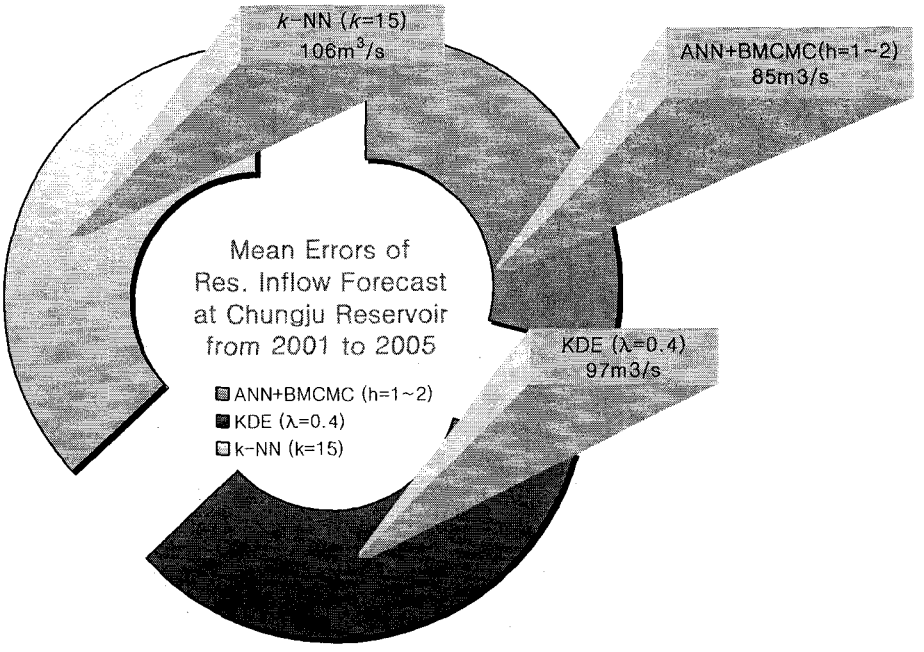
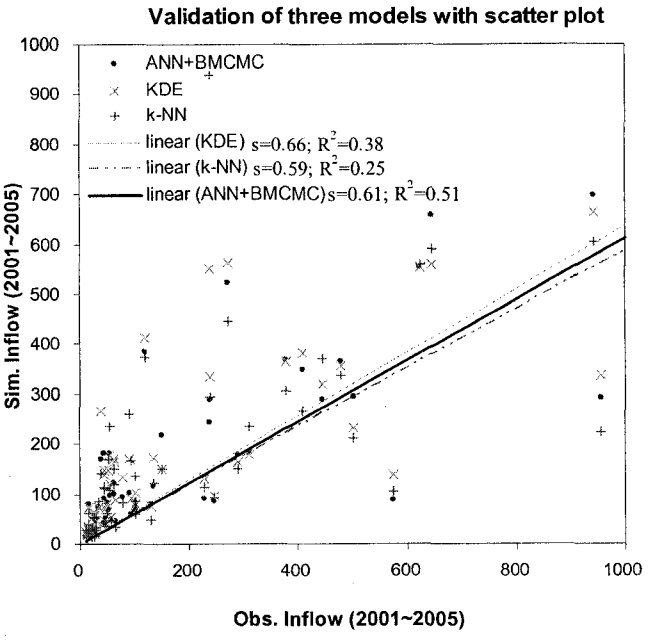


Figure 5-21 Comparison of the forecast errors in terms of RMSE by KDE, *k*-NN and stochastic ANN with the validation data set from January, 2001 to December, 2005 in the Chungju reservoir.



	ANN+BMCMC	KDE	*k*-NN
Mean forecast error	85	97	106
Slope	0.61	0.66	0.59
Coeff. of determination	0.51	0.38	0.25

Figure 5-22 Validation results of the three models at the Chungju dam in 2001 through 2005 with a scatter plot between the historical record and the mean of the simulated inflow, where 's' and 'R^2' mean the slope of linear relationship and the coefficient of determination respectively.

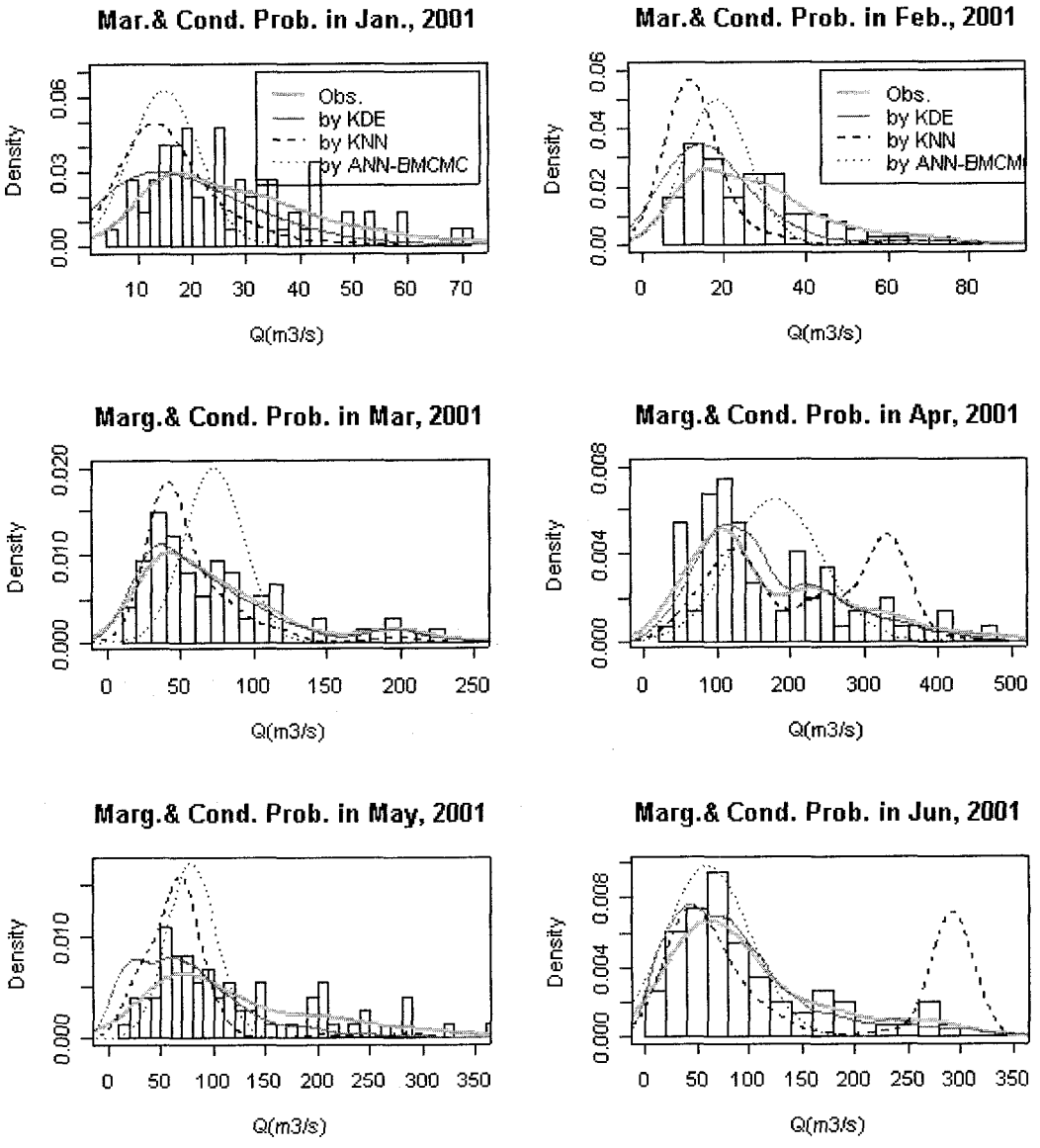


Figure 5-23 Comparison of the marginal probability of the observations and the conditional probabilities of the forecasts by the three models in 2001.

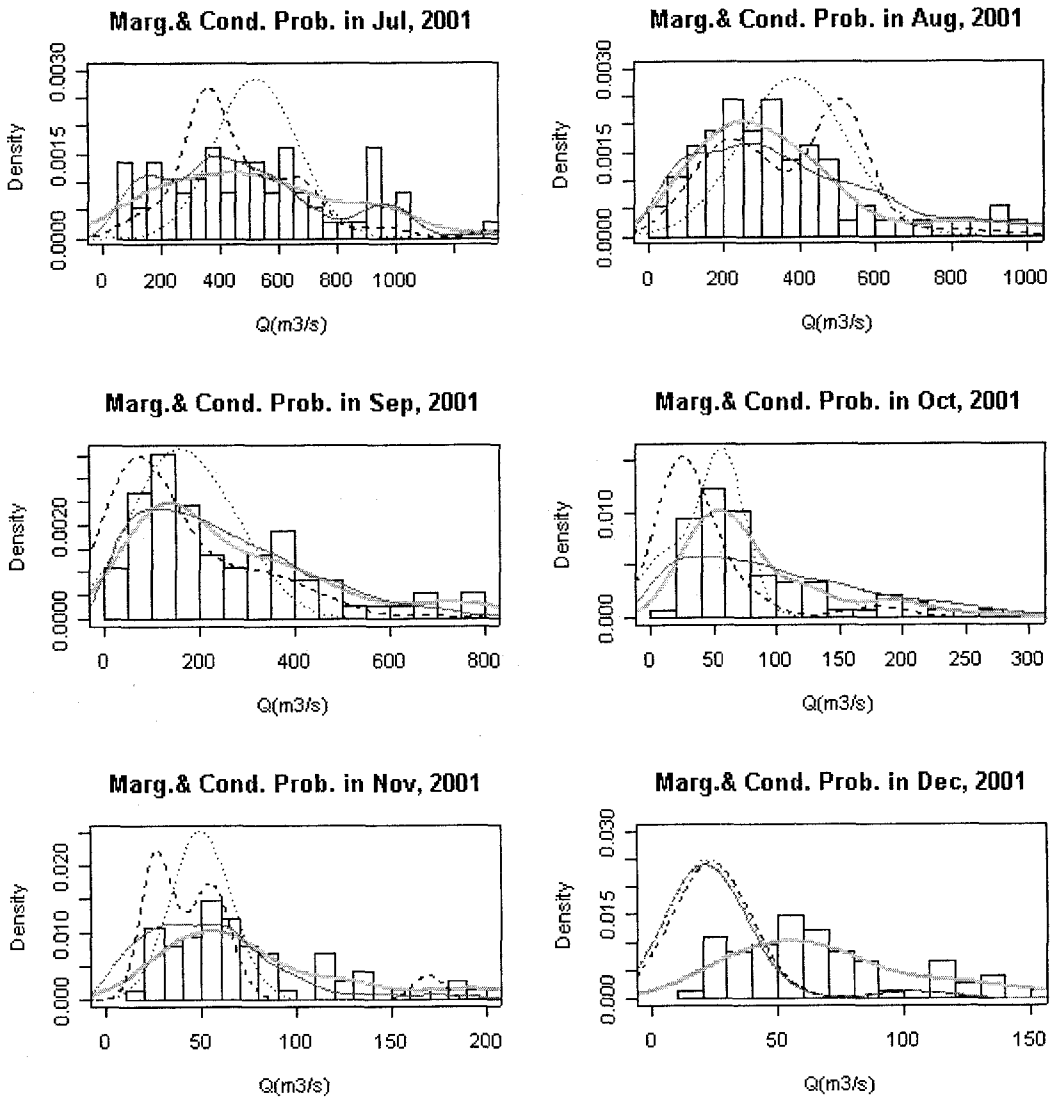


Figure 5-23 (continued.)

5.2 Daily Reservoir Inflow Forecast Using k-Nearest Neighbor Density Estimate for the Chungju Reservoir

As reviewed in the monthly inflow forecast models, water resources engineers may prefer the *k*-NN bootstrap method due to its ease of application. This study was motivated by such questions as whether or not *k*-NN bootstrap method can be employed in practice for daily inflow forecasting, and; whether or not it can provide an alternative to physical ly based rainfall-runoff models such as the Sacramento Soil Moisture Accounting (SAC-SMA) model, which are essential for short term operation of water resources systems.

Daily inflow time series models differ from monthly models in some aspects. Monthly models can be often represented sufficiently with a lag-1 autoregressive term for sites showing short memory in terms of temporal inertia. On the contrary, daily models often work with high order Markov chains because daily historical data show relatively long memory except during flood seasons. Furthermore, they can employ some exogenous input variables, such as rainfall, temperature, and so on, in the case the random variations of inflow can not be represented sufficiently with only an autoregressive term. This indicates that daily models can be formalized in many ways by combining the lags of autoregressive terms and exogenous variables, and physical rainfall-runoff models can be substituted by introducing precipitation as an exogenous variable. In the above mentioned respects, daily models are much more complicated than monthly models. Conventionally, autoregressive moving average models with exogenous variables, referred to as ARMAX, has been commonly used as daily time series models.

In this study, the daily reservoir inflow forecast model was developed using a *k*-NN bootstrap density estimate for the Chungju reservoir. Summarized in Table 5-2, the daily model was designed with a lag-*p* autoregressive term and a lag-*q* term of the average precipitation above the dam as an exogenous variable. This amounts to ARMAX (*p*, *q*) model. For example, Equation 5.2 shows a functional relationship for a model denoted as ARX (*p*, *q*)_*k*NN with *p* of 3and *q* of 5:

$$ARX(3,5)\_kNN: Q_{t+1} = f(Q_t, Q_{t-1}, Q_{t-2}; R_t, R_{t-1}, R_{t-2}, R_{t-3}, R_{t-4}) + \varepsilon \quad (5.2)$$

where, Q, R and ϵ mean inflow, precipitation and forecast error respectively.

Figure 5-24 shows the autocorrelation function of the daily inflow at this site. From the autocorrelogram, it can be said that the basin has about 10 day memory. In this study, autocorrelations over 0.2 were taken into account: hence, the maximum lag was fixed to 6. Finally, the parameters of *p* and *q* were calibrated in a trial and error way by changing two parameters using the calibration data set from 1986 to 2003. The number of nearest neighbors was obtained with the *ad hoc* prescriptive choice of $k = n^{1/2}$ (≈ 80), where n equals the total days of the calibration data set that is 6,574. As performance criterion, the statistic of root mean square error was used, and considered only for one year 1989 that represents normal hydrologic condition because it was too arduous to run the model for the whole period. Once the calibration had been finished, the model was validated using the validation data set from 2004 to 2005 for 2 years. The forecast lead time was one day, that is, the forecast was updated every day, and the 1,000 samples were drawn in each day for statistical inference of the generated inflows. Figure 5-25 and 5-26 refer to the time series plot for the calibration results over *p* and *q* parameters and the RMSEs of the models respectively. Figure 5-27 shows the scatter plot with the slope and coefficient of

determination between the historical record and the mean of the simulated inflow in 1989. They suggest some noteworthy information in determining the parameters of p and q as:

i) As the lag of the basin average precipitation decreases from 10 to 1 with the lag of inflow fixed to 0, the RMSE decreases: hence, the parameter q can be fixed to 1 (Figure 5-26);

ii) As the order of Markov chain increases from 1 to the maximum lag 6 with the lag of precipitation fixed to 1, the RMSE increases overall (Figure 5-26);

iii) The model ARX(1,1)_*k*NN shows better results than the model ARX(1,0)_*k*NN without the exogenous variable (Figure 5-26 and Figure 5-27). This indicates that addition of precipitation helps improve model performance, and;

iv) Without the autoregressive terms into a model, the daily inflow can not be forecasted sufficiently (Figure 5-26 and Figure 5-27).

Consequently, the model ARX(1,1)_*k*NN shows the best result in terms of RMSE in Figure 5-26 and linearity between the observations and the forecast values in Figure 5-27, and was selected for forecasting at this site. Figure 28 and 29 show the validation results for 2004 and 2005 respectively. Because the measurement errors for two years were considerably high below the low inflow of 10 m^3/s, the validation results look very poor, while the results of the medium and high inflows are favorable. In conclusion, this research shows that the daily inflow forecast models by *k*-NN nonparametric method with exogenous variables can be applied successfully in practice for short term management of water resources.

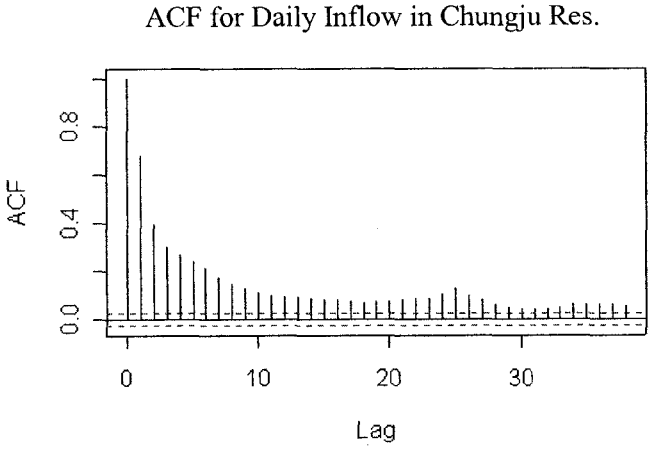


Figure 5-24 Autocorrelation function of the daily inflow in the Chungju reservoir for identifying model architecture.

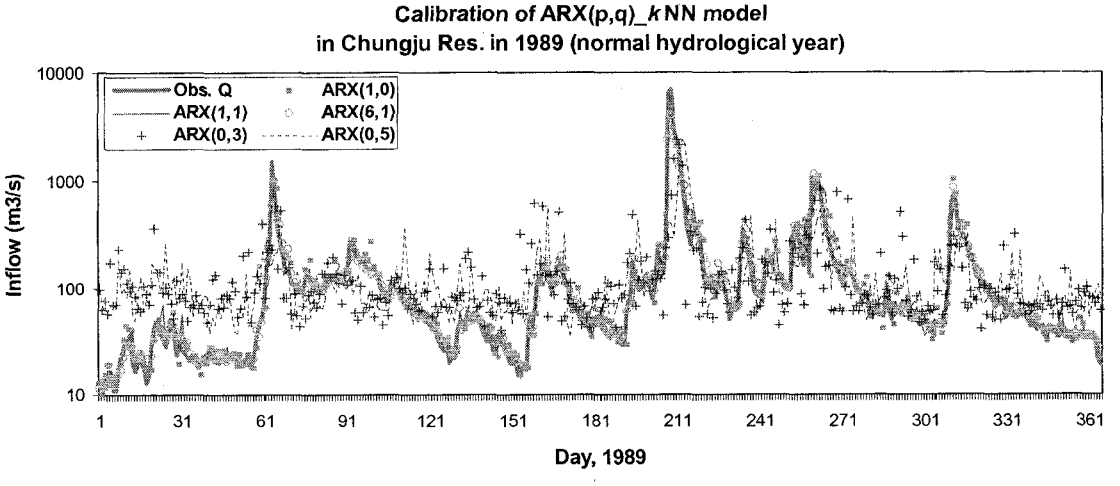


Figure 5-25 Calibration results of the ARX(p,q)_*k*NN models over p and q in the Chungju reservoir in 1989 (normal hydrologic year).

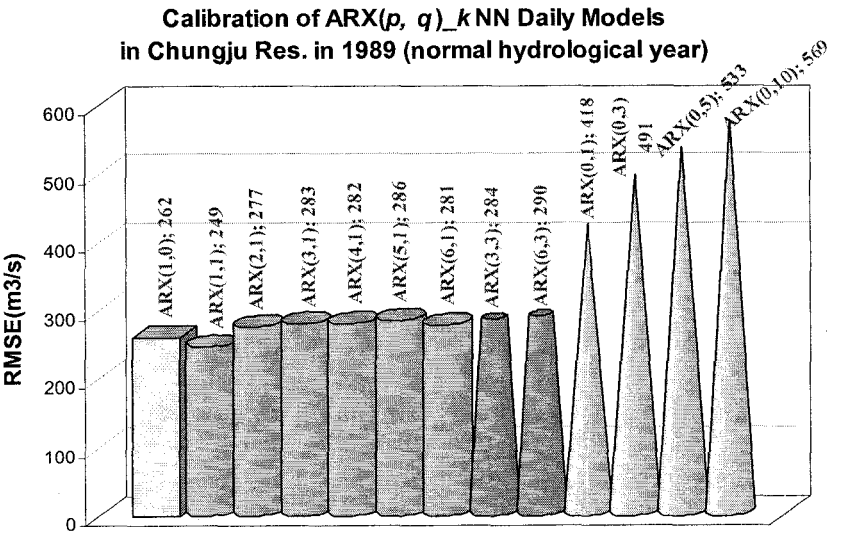


Figure 5-26 Comparison of RMSEs among the ARX(p,q)_*k*NN models for the purpose of model calibration over p and q in the Chungju reservoir in 1989 (normal hydrologic year).

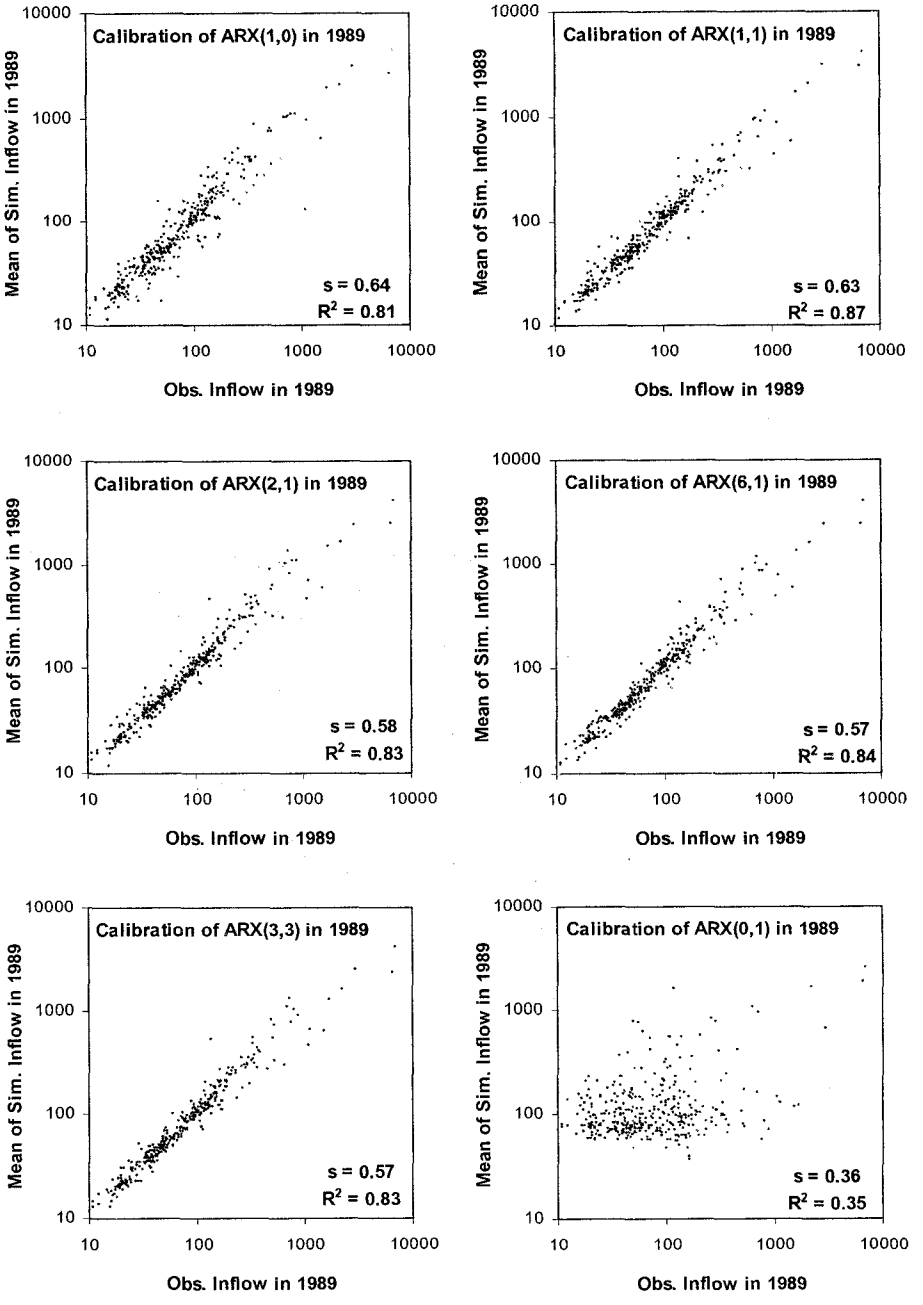


Figure 5-27 Scatter plots between the historical record and the mean of the simulated inflow in the Chungju reservoir in 1989 using ARX(p,q)_*k*NN model, used for model selection by comparing linear relationship, where 's' and 'R^2' mean the slope and the coefficient of determination respectively.

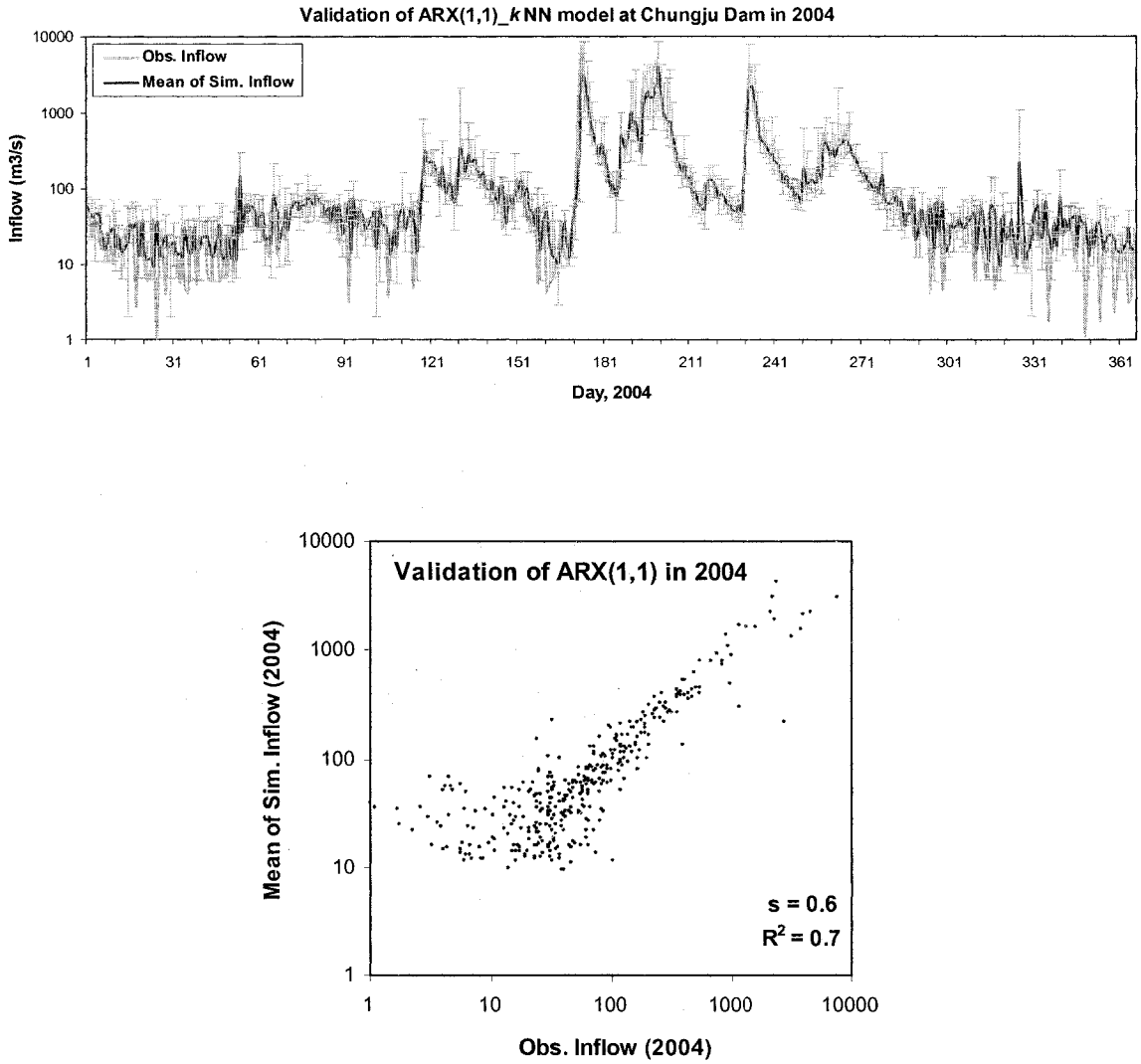
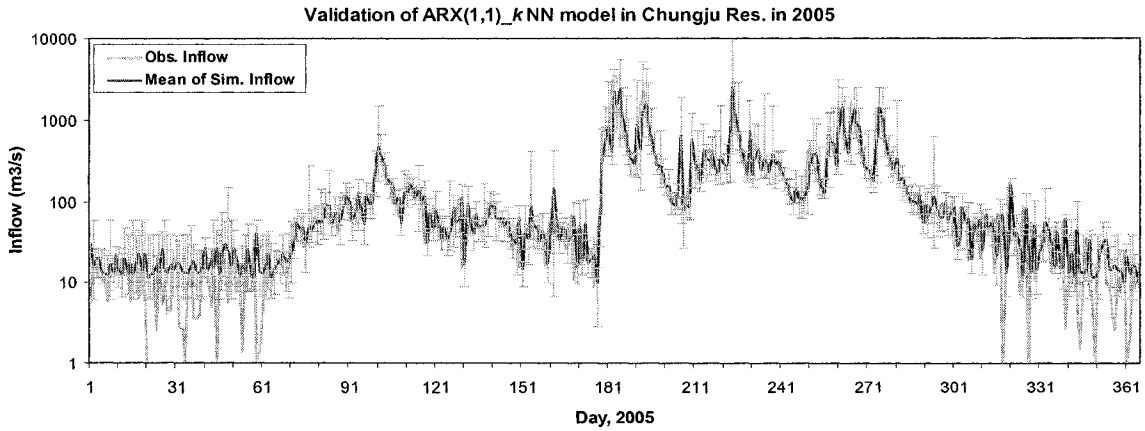


Figure 5-28 Validation of ARX(1,1)_*k*NN daily inflow forecast model in the Chungju reservoir in 2004 with time series plot and scatter plot between the historical record and the mean of the simulated inflow, where 's' and 'R^2' mean the slope and the coefficient of determination of linear relationship respectively.



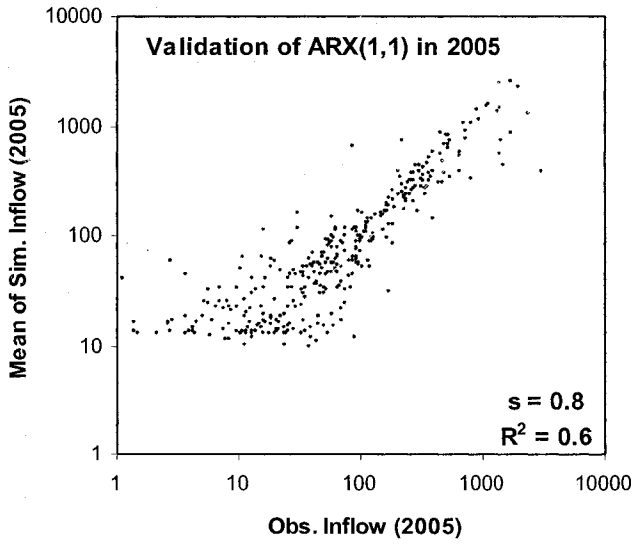


Figure 5-29 Validation of ARX(1,1)_*k*NN daily inflow forecast model in the Chungju reservoir in 2005 with time series plot and scatter plot between the historical record and the mean of the simulated inflow, where 's' and 'R^2' mean the slope and the coefficient of determination of linear relationship respectively.

CHAPTER 6

DEVELOPMENT OF PROBABILISTIC BOD and TP MODELS USING A BAYESIAN MCMC TECHNIQUE

As mentioned in chapter 1, decisions made in the processes of hydrologic design and water resources system management always entail risks resulting from the uncertainties due to errors in models, natural variations and measurement errors. In general, risk-based decisions are concerned with assessing the impact of hydrologic events on a water resources system and determining the design and key variables (generally called dependent variables) to meet given targets *[Chow et al., 1988]*. They are subject to the uncertainties of the explanatory variables and model parameters that are generally called independent variables. For example, in the case of designing the height of a bank, its variability is dependent on the uncertainties of the explanatory variables such as frequency of flood flows and parameters such as the roughness coefficient. In the case of modeling water quality in streams, the explanatory variables include the pollutant loads, flows from tributaries and parameters such as the biological-physical-chemical reaction rates. This indicates that the dependent variables are intrinsically random, and decisions should be made considering the risks of violating some established standards (criteria) for managing water resources. In this respect, the

uncertainty analysis between the dependent and independent variables is essential for risk-based decision making, and the probabilistic models are more valuable than deterministic ones in that they can provide various levels of risks based on uncertainty analysis.

The representative methodologies for uncertainty analysis, which have attracted great attention in water resources planning and management, are sensitivity analysis, derived distributions, first order second moment (FOSM) analysis, and Monte Carlo analysis. They have a common objective, that is, to quantify the probability distribution or the degree of variability for a dependent variable.

Sensitivity analysis, the simplest among these methodologies is concerned with evaluating the degree of variability of a dependent variable corresponding to the variability of independent variables. However, it has a drawback that it can not derive the probability distributions of dependent variables.

The derived distribution method is concerned with obtaining analytically the probability density function of a dependent variable when the probability density function of an independent variable or model parameter is given deterministically, or assumed to be known. However, it is limited only to analysis of an independent variable.

FOSM is a procedure for quantifying the expected variability of a dependent variable represented as a function of variability of one or more independent variables and model parameters. This method provides only the mean and variance instead of a probability density function for a dependent variable, and the variation of the dependent variable is absolutely affected by the statistical features of the independent variables. In general, the independent variables are assumed to be normally distributed: hence, the

dependent variable is represented with the normal distribution also, which is typically a Gaussian approach. Unlike the derived distribution method, it can be applied to problems with multiple independent variables.

Monte Carlo analysis is a process of iterative simulation of dependent variables by randomly sampling independent variables. Compared with the derived distribution and FOSM, it is computationally rigorous because it requires numerous random generations.

As a common drawback of the three methods, it is hard for the model parameters to be managed in an adaptive way on the basis of data assimilation: hence, they often work with one independent variable. In contrast to the above mentioned methodologies, a Bayesian Monte Carlo Markov chain tries to estimate not only time-varying states of the dependent variables but also the multiple model parameters along with their uncertainties in the context of the prior and posterior probability density functions. In addition, it can analyze the uncertainties between both dependent variables and explanatory variables.

In Korea, the government has been putting a national-wide plan for water quality conservation in streams, lakes and reservoirs in place since 2004, and is trying to prepare total maximum load standards for pollutants in the river basins (Han, Geum, Nakdong, Yongsan rivers). This concept is very similar to the TMDLs (total daily maximum loads) in the United States. TMDLs determine what level of pollutant load would be consistent with meeting water quality standards in river basins. In general, multi-purpose large storage projects are in operation in river basins in order to harness water and protect from floods. They tend to account for a large portion in the entire water management systems: hence, reservoir system operation plays an important role in managing water in an integrated manner.

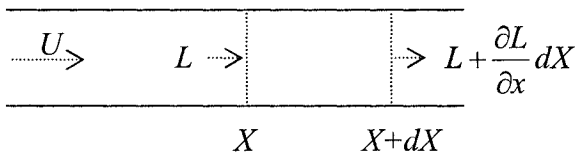
Figure 2-1 shows the study area composed of the two cascade dams and multiple tributaries. Regarding BOD and TP as the barometers for water quality management in this area, Figure 2-1 shows the spatial relationship between both BOD at the Gongju site and TP in the Daecheong reservoir and the operation of the two cascade reservoirs. Consequently, the two time-varying dependent variables of BOD and TP are directly influenced by the decision variables of the releases from the two reservoirs that are subject to the risks of quantifying the two water quality variables.

The objectives of the study in this chapter are to develop the probabilistic water quality models for BOD at Gongju and TP in the Daecheong reservoir, and come up with the relationships between reservoir release and risk of violating the water quality standards established for water quality management in the river basin. The study aims toward the risk-based optimal joint operation of the two reservoirs. In this study, the risk is defined as the cumulative probability to exceed a predefined BOD and TP criteria for water quality management. The criteria were set to 3mg/L in BOD and 0.03 mg/L in TP corresponding to the index of "2nd grade water quality" in streams and reservoirs respectively, in Korea. The two dependent variables were modeled by month using a Bayesian network with the help of the MCMC sampling method. Chapter 7 shows a procedure for risk-based reservoir operation by joining the probabilistic water quality models interactively.

6.1 Development of a Probabilistic BOD Model and Risk Analysis Using Bayesian Networks

6.1.1 Development of a Steady-State Monthly Probabilistic BOD Model

As shown in Figure 2-1, Gongju city is approximately 50km downstream from the Daecheong dam along the Geum River. The Gongju water quality measurement station is considered a control point for water quality management in this study. In general, BOD (biological oxygen demand) and DO (dissolved oxygen) have been considered the most important measures for assessment of water quality in streams, and traditionally modeled using *Streeter-Phelps* equations. As shown in Figure 6-1, DO is not problematic for the aquatic ecosystem in this area because it is almost always over 7.5 mg/L except in the Daecheong reservoir, which indicates a first grade water quality in streams in Korea: hence, DO was not considered in this study. BOD models can be developed using a mass balance with reaction kinetics where BOD is reduced by decomposition and settlement as it is carried downstream:



$$\frac{\partial LV}{\partial t} = LQ - \left(L + \frac{\partial L}{\partial x}dX\right)Q - k_r LV;$$

$$\frac{\partial L}{\partial t} = -\frac{\partial L}{\partial x}dX\frac{Q}{V} - k_r L; \qquad (6.1)$$

$$= -\frac{\partial L}{\partial x}U - k_r L;$$

where:

L = remaining BOD at time t (mg/L);

Q= discharge (m^3/d);

U= flow velocity (m/d);

X= distance (m);

V= volume of water body (m^3);

k_r= total removal rate (d^{-1}) including both decomposition and settling.

At steady state, Equation 6.1 becomes:

$$L = L_0 e^{-\frac{k_r}{U}X} \; : \; L_0 = \frac{Q_m L_m + Q_t L_t}{Q_m + Q_t}; \tag{6.2}$$

where, L_0 is the initial BOD or ultimate BOD, which can be calculated as the flow-weighted average of the BOD load (subscript t) from a tributary or any point source such as a waste water treatment plant and the BOD in a main stream (subscript m) right before a junction or outlet of pollutants *[Chapra, 1997]*.

Figure 6-2 refers to the schematic representation of the spatial relationships for water quality modeling along with the model parameters by reach. Centering on the water quality and stream flow measurement stations, the river section under study can be divided into three sub-reaches: Daecheong (subscript *DC*) dam to Bugang (subscript *Bg*); Bugang to Geumnam (subscript *Gn*); Geumnam to Gongju (subscript *Gj*). Each sub-reach has its own BOD model parameters of k_r and the coefficients of a and b for the rating between flow velocity (U) and discharge (Q), which is represented as:

$$U = aQ^b \tag{6.3}$$

The total removal rates k_r were determined probabilistically by month at the three reaches: hence, they were addressed in temporally distributed and spatially lumped ways. Unlike k_r, the coefficients of a and b are treated deterministically in this study. As shown in Equation 6.2, the coefficients of a and b for the rating should be determined for computing flow velocity U. They were determined using the historical stage-discharge measurement data at the three sites, and Figure 6-3 and Table 6-1 show the rating curve and the coefficients respectively.

Assuming the observations are independent, and the residuals between the observed BOD (L^{obs}) and simulated BOD (L^{sim}) are normally distributed with mean *0* and variance σ^2, the posterior probability of the parameter k_r can be described by combining a likelihood function and a prior probability distribution using Bayes' theorem. For the prior probabilities of k_r and the variance of the likelihood function, the normal and Gamma distributions were employed respectively. A Gamma distribution for the variance was chosen to ensure sampling positive values. In order to determine the means of the prior distributions for the monthly removal rates, a deterministic BOD model based on Equation 6.2 was developed in *MS-Excel*. The parameters (k_r) were determined optimally using the *SOLVER* optimization tool such that the predicted values matched observed values as closely as possible. The obtained k_r were regarded as the means of the prior probabilities, which indicates application of the informative priors. On the contrary, the prior probability of the variance was addressed uninformatively. Equation 6.4 refers to the structure of the steady state probabilistic BOD model at the Bugang station in month τ:

$$\varepsilon = (L_{i,\tau}^{obs} Bg - L_{i,\tau}^{sim}.Bg) \sim Normal(0, \sigma_\tau^2.Bg);$$

A. likelihood function :

$$p(L_\tau.Bg \mid k_{r\tau}.DC.Bg) = \prod_{i=1}^{n(years)} \frac{1}{\sqrt{2\pi\sigma_\tau^2.Bg}} \exp\left(- \frac{(L_{i,\tau}^{obs}.Bg - L_{i,\tau}^{sim}.Bg)^2}{2\sigma_\tau^2.Bg} \right) : \tau = 1..12 :$$

$$L_{i,\tau}^{sim}.Bg = L_{i,\tau}.Jn.Gp \times e^{-\frac{k_{r\tau}.DC.Bg}{U_{i,\tau}.Bg}(X.Gp.Bg)};$$

$$(a)\ L_{i,\tau}.Jn.Gp = \frac{(Q_{i,\tau}.DC \times L_{i,\tau}.bef.Gp) + (Q_{i,\tau}.Gp \times L_{i,\tau}.Gp)}{\omega(Q_{i,\tau}.DC + Q_{i,\tau}.Gp)};$$

$$(b)\ U_{i,\tau}.Bg = a.DC.Bg \times \omega(Q_{i,\tau}.DC + Q_{i,\tau}.Gp)^{b.DC.Bg}$$

$$(c)\ L_{i,\tau}.bef.Gp = L_{i,\tau}.DC \times e^{-\frac{k_{r\tau}.DC.Bg}{U_{i,\tau}.bef.Gp}(X.DC.Gp)};$$

$$(d)\ U_{i,\tau}.bef.Gp = a.DC.Bg \times Q_{i,\tau}.DC^{b.DC.Bg};$$

B. prior probability :

$$k_{r\tau}.DC.Bg \sim Normal(para1, para2);$$

$$\sigma_\tau^2.Bg \sim Gamma(\alpha, \beta); \qquad (6.4)$$

where: n is the number of years in the historical record; $p(L|k_r)$ is the likelihood function of the parameter k_r; the scripts of '*sim*', '*obs*', '*i*', '*Normal*', '*Gamma*' and '*para*' mean simulation, observation, a year, Normal, Gamma distributions and their parameters respectively; *L.bef.Gp* and *LJn.Gp* represent the initial BOD right before the junction with the Gap (subscript *Gp*) tributary and at the junction respectively; *U.bef.Gp* means the flow velocity right before the junction, and; ω is flow increment ratio reflecting lateral flow and return flow of irrigation, municipal and industrial water uses, which was determined using the historical flow record of *Gp* and *Bg*.

According to this modeling concept, the steady state probabilistic sub-models were also developed individually at the *Gn* and *Gj* stations, and a network model was formed by connecting the three sub models at the *Bg* through *Gj* in series. Once the network

model is established, the posterior probabilities of the model parameters are sampled first with the help of the MCMC technique, and then the posterior predictive distributions of the BODs at the three stations can be generated with the explanatory variables using the sampled posteriors. The monthly historical data available were divided into two subsets: a calibration (or training or learning) set from 2000 to 2001 and validation set in 2002 and 2004. In order to embody a probabilistic model based on Equation 6.4, the *WinBUGS [Spiegelhalter et al., 1995]* software was used, and the model was coded using its programming language. When the model was run, two chains of the parameters with 20,000 samples and 10,000 burn-in iterations were generated with two different sets of starting points to check convergence to a stationary distribution. The MCMC diagnostics were based on the graph of history path and autocorrelation of the chains, and are described in *Appendix-C*. The source code of *WinBUGS* for BOD model is described in *Appendix-C*.

Figure 6-4 displays the calibration and validation results showing that the steady state BOD model is well calibrated and consistent with the validation data set so that it can be applied to forecast of BOD in this area. The errors in the results mainly come from the uncertainties of the discharges from the Gap and Miho (subscript *Mh*) tributaries. Figure 6-4 also shows that the BOD generally starts to increase from April, the farming season when the application of fertilizers and the intakes for irrigation and discharges between rice paddies and streams are activated in Korea, and peaks in June (sometimes May) which is the flood season. Figures 6-5 shows the mean of the posterior probability of the total BOD removal rate with the prior mean by month, which peaks in August.

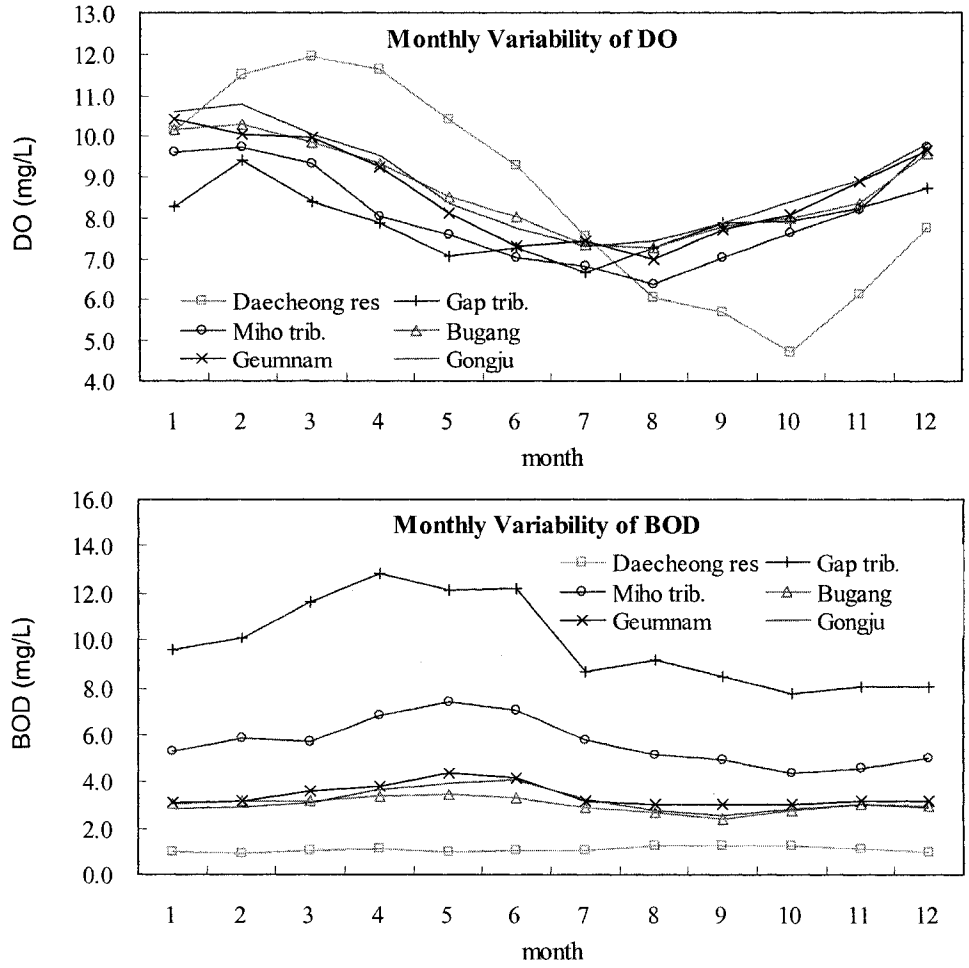


Figure 6-1 Monthly BOD and DO variability in the study area.

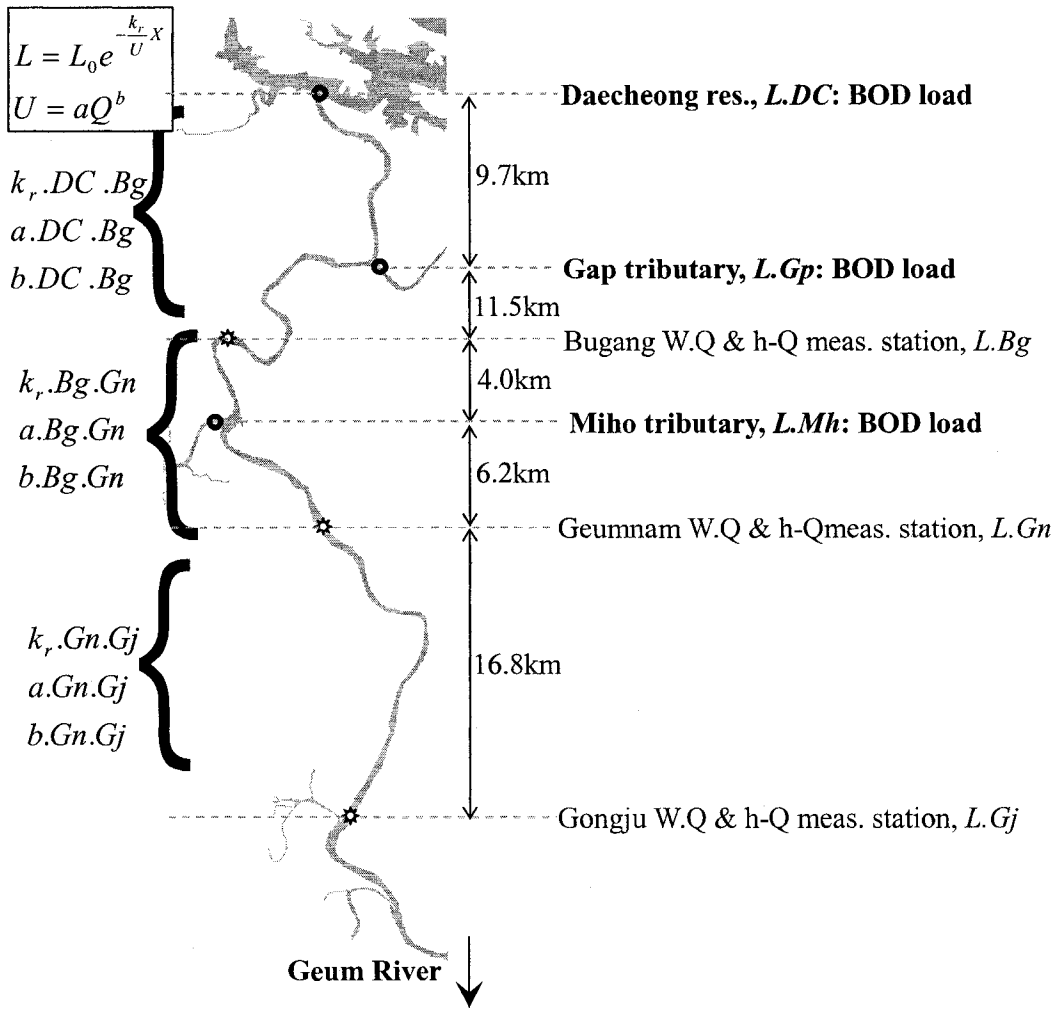


Figure 6-2 Schematic representation of the river basin for BOD modeling, which is composed of the Daecheong reservoir, 2 tributaries loading pollutants, and 3 water quality and discharge measurement stations, where W.Q and h-Q stand for water quality and stage-discharge respectively.

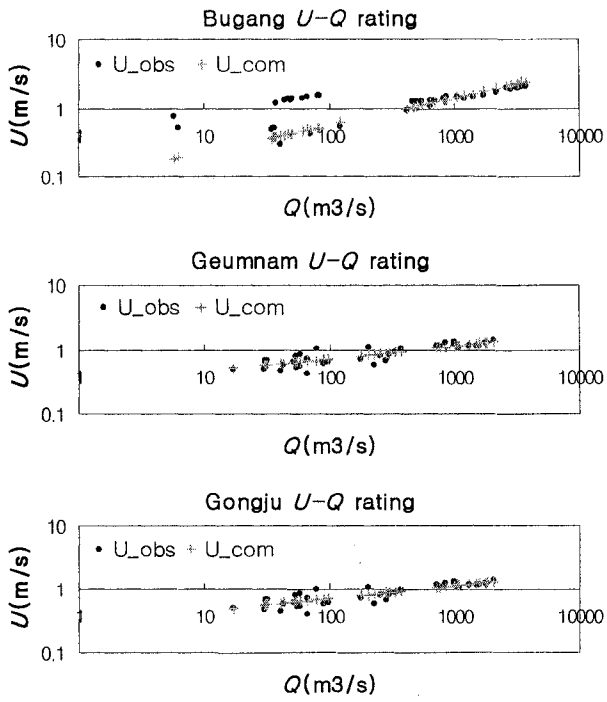


Figure 6-3 Rating curves between flow velocities and discharges at the Bugang, Geumnam and Gongju stations.

Table 6-1 Coefficients of the ratings for flow velocities and discharges at the Bugang, Geumnam and Gongju stations.

	Bugang	Geumnam	Gongju
a	0.089	0.118	0.285
b	0.400	0.400	0.200

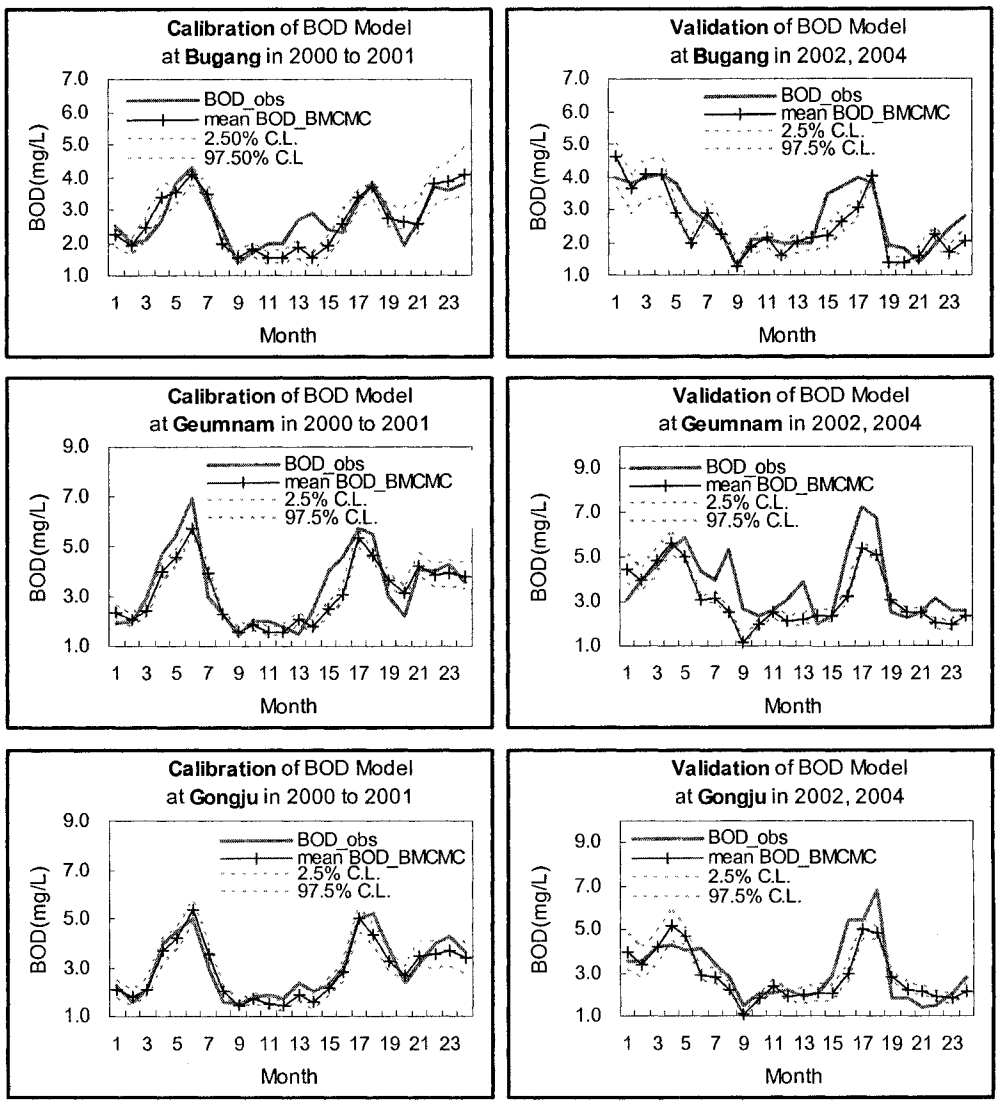


Figure 6-4 Calibration and validation results of the probabilistic BOD model, where BMCMC stands for Bayesian Markov chain Monte Carlo.

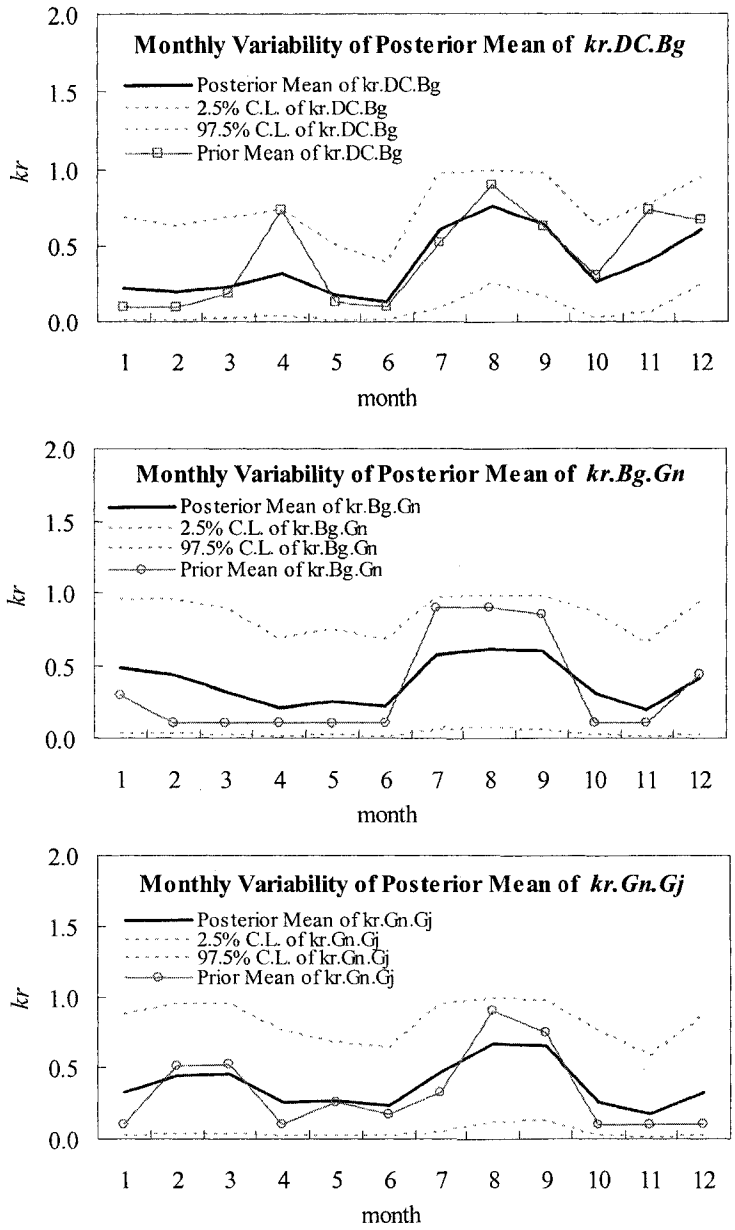


Figure 6-5 Monthly variability of the posterior mean of the total BOD removal rate with the prior mean by reach.

6.1.2 Risk Analysis of BOD Standard Violation at the Gongju Gage in association with the Operation of the Daecheong Reservoir

As reviewed in the introduction of this chapter, the prediction of BOD is intrinsically a part of a random process that is defined as a naturally occurring sequence of events in time or space. This implies that although a simulated BOD from a deterministic model does not exceed the BOD criterion, there may be still risk to exceed the standard. Risk analysis is a method of accounting for the risks resulting from the various sources of uncertainty *[Chow et al., 1988]*. The central concepts of this analysis are loading and capacity (or resistance) that correspond to the BOD loads and the water quality standards respectively in this study. When it comes to risk analysis, two methods can be considered: simple risk and; composite risk analysis. For the simple risk analysis, it is assumed that only the load is uncertain. On the contrary, for the composite risk analysis, both load and capacity are considered to be uncertain. Simple risk analysis was applied in this study because the stream BOD standard was firmly fixed to 3mg/L.

In this study, the relationships between the risk and the release of the Daecheong reservoir were established by month. First, the monthly releases were discretized evenly, and then the monthly averaged BOD loads from the Daecheong reservoir were computed with the monthly averaged BOD concentrations of the Daecheong reservoir. They were fed into the already calibrated and validated probabilistic model with the monthly averaged flows and BOD loads from the tributaries. Figure 6-6 refers to the posterior predictive probability distribution of the predicted BOD using the discretized release of the Daecheong reservoir with the concept of the risk in January. Figure 6-7 shows the three dimensional representation of the posterior predictive probability distribution in

January and July. The posterior predictive distributions and risks of the remaining months were obtained in the same way. Figure 6-8 shows the risk of the BOD standard violation and the mean of the posterior predictive probability distribution according to the releases of the Daecheong reservoir by month. Figure 6-9 and Table 6-2 refers to the monthly releases by risk at the Gongju, which were determined from Figure 6-8.

6.2 Development of a Probabilistic TP Model and Risk Analysis Using a Bayesian MCMC Technique

6.2.1 Development of a Steady-State Probabilistic TP Model

Eutrophication in reservoirs results from the input of organic and inorganic nutrients into a water body, which stimulates the growth of algae or rooted aquatic plants resulting in the interference with desirable water use of aesthetics, recreation, fish maintenance and water supply *[Thomann and Mueller, 1987]*. The main nutrients for aquatic plant growth are nitrogen, phosphorus and silica that are loaded into a water body by municipal, industrial and agricultural sources. The basic phenomena underlying the process of phytoplankton growth is that increasing solar radiation stimulates the photosynthesis reaction and increase of water temperature resulting in increase of phytoplankton biomass with uptake of the dissolved form of inorganic nutrients such as phosphate (PO_4^{3-}) and nitrate (NO_3^-). The detailed process for the cell synthesis and endogenous respiration can be represented as *[Chapra, 1997]*:

$$106CO_2 + 16NH_4^+ + HPO_4^{2-} + 108H_2O \rightleftarrows \underset{(algae)}{C_{106}H_{263}O_{110}N_{16}P_1} + 107O_2 + 14H^+ \qquad (6.5)$$

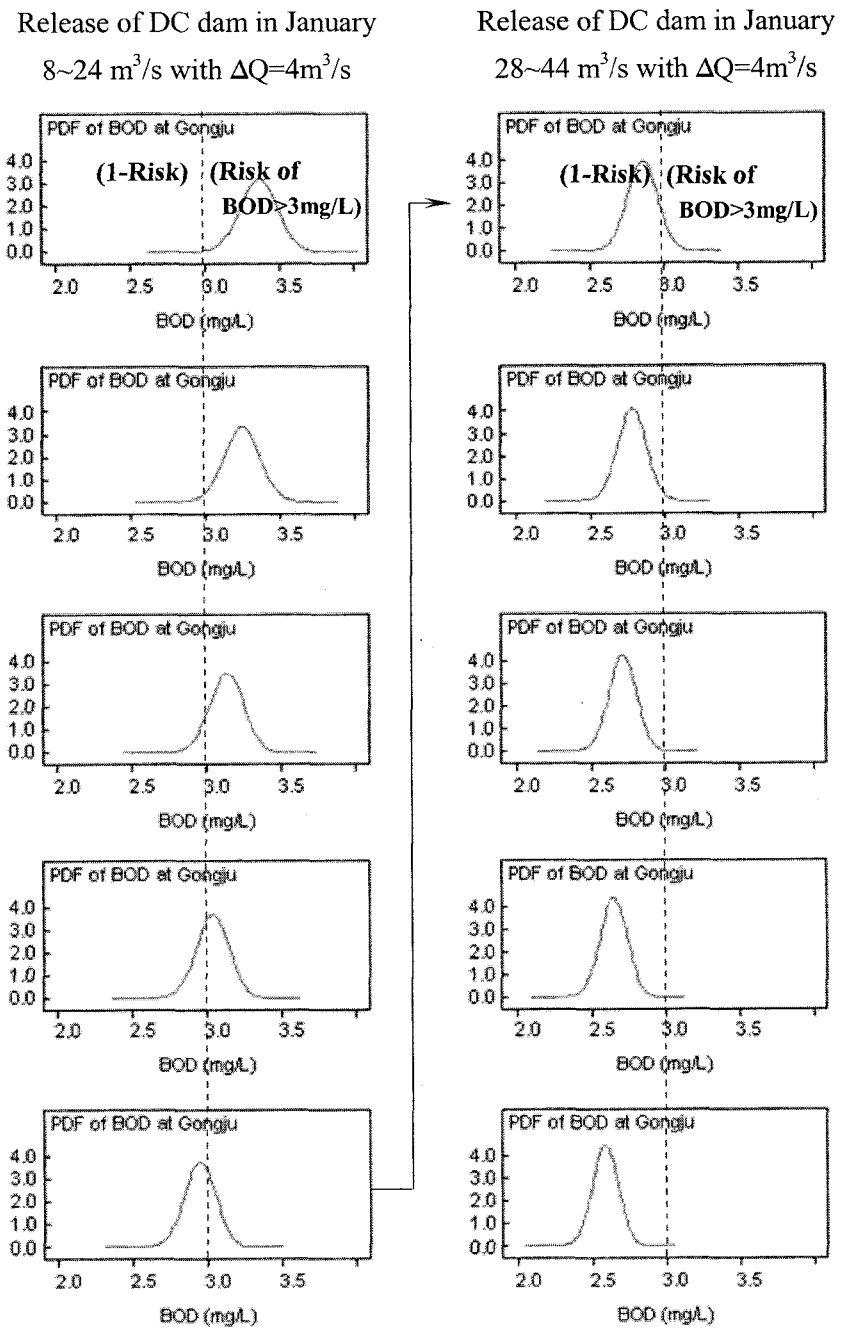


Figure 6-6 The posterior predictive probability distribution of the simulated BOD with the concept of the risk in January.

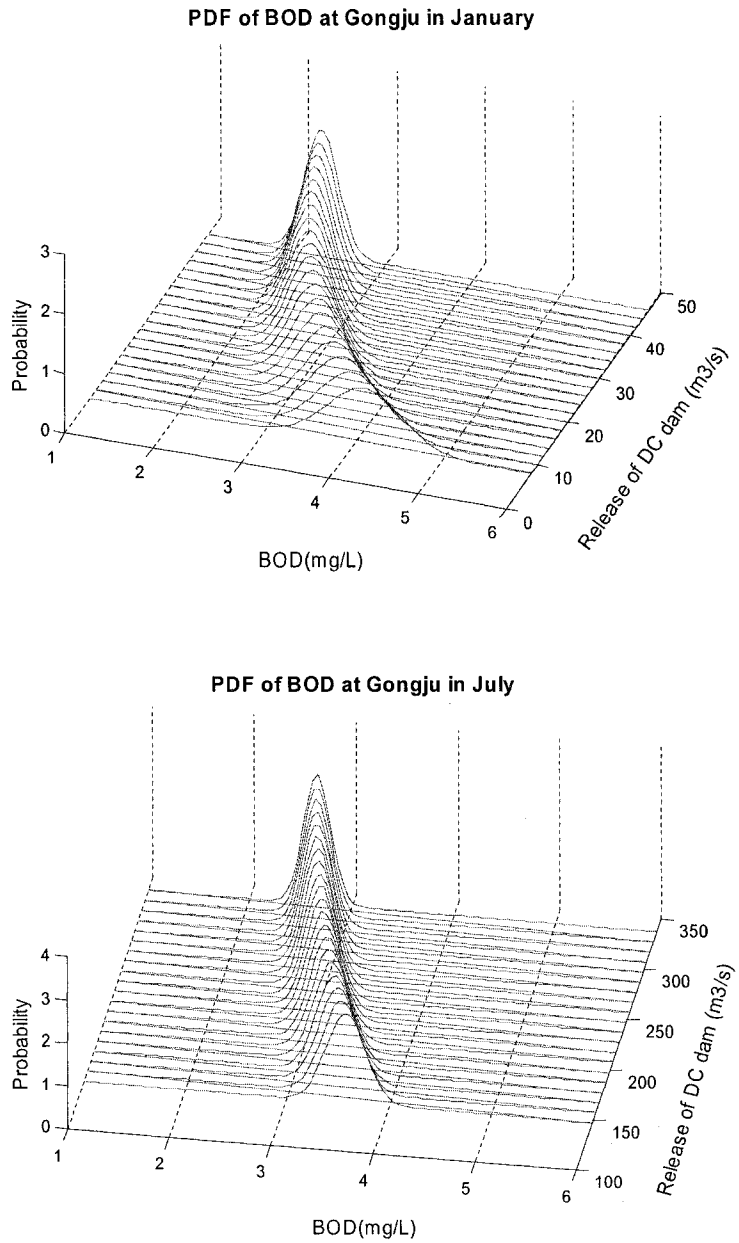


Figure 6-7 The three dimensional representation of the posterior predictive probability distribution of the simulated BOD in January and July.

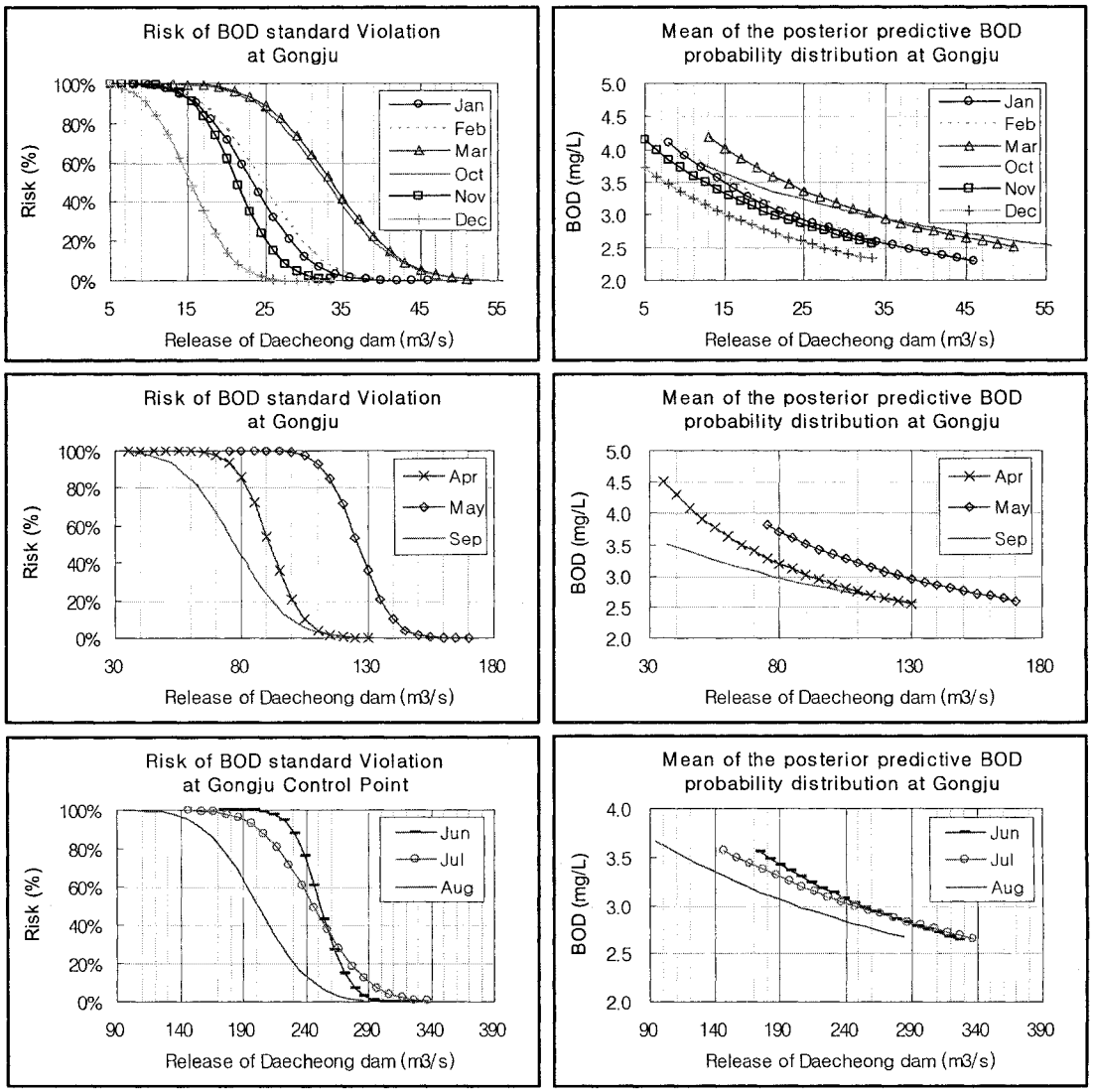


Figure 6-8 The risk of the BOD standard violation and the mean of the posterior predictive probability distribution of the simulated BOD versus the releases of the Daecheong reservoir by month.

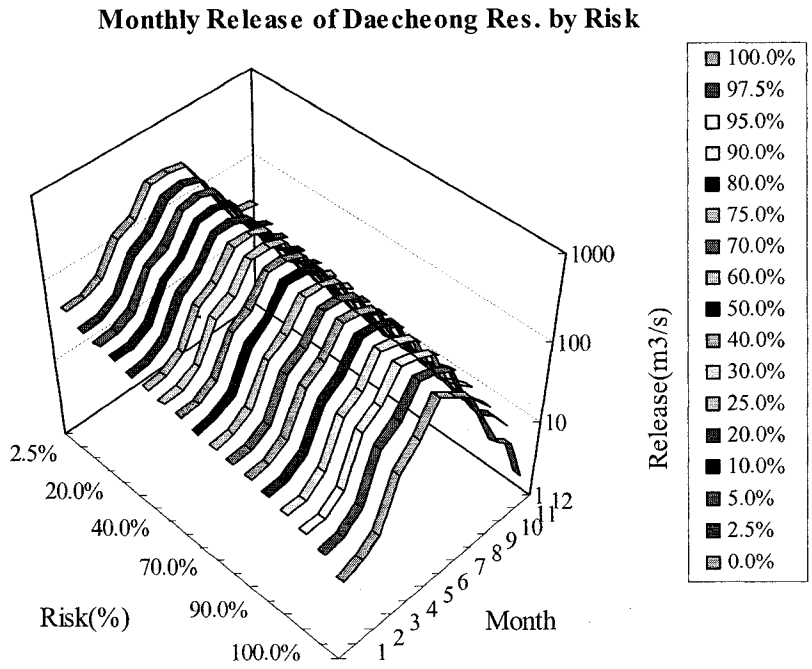


Figure 6-9 The monthly releases of the Daecheong reservoir by risk at the Gongju station.

Table 6-2 The monthly releases of the Daecheong reservoir by risk at the Gongju station.

	Release of the Daecheong Reservoir											
Risk (%)	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
100.0%	8.0	10.0	13.0	50.0	90.0	195.0	140.0	95.0	35.0	12.0	8.0	2.0
97.5%	12.0	12.5	20.0	70.0	105.0	216.0	179.0	136.0	44.0	19.0	12.5	6.5
95.0%	14.0	14.5	22.0	74.0	108.0	222.0	190.0	146.0	49.5	21.0	14.0	8.0
90.0%	16.0	16.5	24.5	77.0	112.0	229.0	202.0	158.0	55.5	24.0	15.5	10.0
80.0%	18.5	18.5	27.5	82.0	117.0	236.0	217.0	173.0	62.5	27.0	17.5	11.5
75.0%	19.5	20.5	29.0	84.0	119.0	240.0	223.0	179.0	65.5	28.0	18.5	12.5
70.0%	20.5	21.5	30.0	86.0	121.0	242.0	228.0	183.0	68.0	29.0	19.0	13.0
60.0%	22.0	23.0	32.0	88.0	123.0	247.0	238.0	193.0	73.0	31.0	20.0	14.5
50.0%	23.0	24.5	33.5	91.0	128.0	252.0	246.0	202.0	77.0	33.0	21.5	15.5
40.0%	25.0	26.0	35.0	94.0	129.0	256.0	255.0	210.0	81.5	34.5	22.5	16.5
30.0%	26.0	27.5	37.0	97.0	132.0	262.0	264.0	220.0	86.5	36.5	23.5	17.5
25.0%	27.0	28.5	38.5	99.0	134.0	264.0	269.0	225.0	89.0	38.0	24.5	18.5
20.0%	28.0	29.5	39.5	100.0	135.0	267.0	275.0	230.0	92.0	39.0	25.0	19.0
10.0%	31.0	32.5	42.5	105.0	140.0	276.0	291.0	245.0	100.0	42.0	27.0	21.0
5.0%	33.0	34.5	45.0	109.0	144.0	282.0	304.0	259.0	106.0	45.0	28.5	22.5
2.5%	34.5	36.5	47.5	113.0	148.0	287.0	314.0	270.0	111.5	47.5	30.5	24.0
0.0%	46.0	48.0	60.0	130.0	165.0	319.0	330.0	300.0	140.0	60.0	35.0	32.0

Equation 6.5 directly indicates that the algal growth is absolutely affected by controlling the loads of nutrients into reservoirs. When it comes to the limiting nutrient governing the eutrophication process, phosphorus tends to be identified because it is usually in short supply relative to nitrogen. Although total phosphorus is composed of several components, only the soluble inorganic phosphorus, also called orthophosphate ($H_2PO_4^-$, HPO_4^{2-}, and PO_4^{3-}), can be readily available to aquatic plants. However, total phosphorus measurement has been used widely to quantify eutrophication, and the correlations among eutrophication, TP and Chlorophyll has been studied *[Chapra, 1997]*. From this viewpoint, the TP models play a crucial role in managing reservoir water quality.

The monthly behavior of the phosphorus in reservoirs can be represented in Equation 6.6 in month τ by extending Vollenweider's yearly mass balance model for a well-mixed lake *[Vollenweider, 1976]*.

$$V_\tau \frac{d(TP_\tau)}{dt} = W_\tau - Q_\tau(TP_\tau) - k_{s\tau} V_\tau (TP_\tau); \tag{6.6}$$

where:

V = lake volume (m^3);

TP = total phosphorus concentration (mg/m^3);

t = time (month);

W = total phosphorus loading rate (mg/month);

Q = outflow (m^3/month);

k_s = a first-order settling loss rate in reservoirs ($month^{-1}$).

At steady state, Equation 6.6 becomes:

$$TP_\tau = \frac{W_\tau}{Q_\tau + k_{s\tau} V_\tau} \tag{6.7}$$

Figure 6-10 refers to the schematic representation of the study area for TP modeling. The section of the river has the three main sources of TP from: Yongdam (subscript *YD*) reservoir, the sub-basin (subscript *YD.Oc*) between Yongdam dam and Ockcheon (subscript *Oc*) station, and Bocheong (subscript *Bc*) tributary. As shown in Figure 6-10, one can notice that the phosphorus concentration in the Daecheong (subscript *DC*) reservoir is directly affected by the operations of the *DC* and *YD* reservoirs. The volume of the *DC* reservoir and TP loading rate into the DC reservoir in Equation 6.7 are represented in month τ as:

$$\begin{aligned} V_\tau &= V_{\tau-1} + (Q_\tau.YD + Q_\tau.YD.Oc + Q_\tau.Bc - Q_\tau.DC); \\ W_\tau &= W_\tau.Oc + W_\tau.Bc \\ &= (1-k_{r\tau}) \times W_\tau.YD + W_\tau.YD.Oc + W_\tau.Bc; \end{aligned} \tag{6.8}$$

where, the total loading rate into the *DC* reservoir was considered the sum of the loading rates of both the Bocheong tributary (*W.Bc*) and the Ockcheon basin (*W.Oc*) including the basin of the *YD* reservoir. In order to account for the impacts over TP in the *DC* reservoir in association with the releases from the *YD* reservoir, *W.Oc* should be divided into two components of: *W.YD* that is the loading rate from the *YD* reservoir, and; *W.YD.Oc* from the *YD.Oc* sub-basin, where *W.YD* is excluded from *W.Oc*. Considering decomposition and settling as the load is carried downstream, all of *W.YD* does not arrive at the Ockcheon: hence, the purely contributed loading rate out of *W.YD* can be represented by introducing k_r meaning the total phosphorus removal rate (%) as [(1-

k_r)*$W.YD$]. $W.YD.Oc$ can be determined by summing all the phosphorus loads from the main tributaries lying in the $YD.Oc$ sub-basin. The reservoir volume can be determined using a simple mass balance consisting of the YD release ($Q.YD$), the discharges from the Bocheong tributary ($Q.Bc$) and the $YD.Oc$ sub-basin ($Q.YD.Oc$). Figure 6-11 shows the monthly variability of TP in the study area. Unlike the BOD model, the first-order settling loss rate (k_s) in the DC reservoir and total removal rates (k_r) in the stream were addressed in a temporally lumped way because the monthly TP and flow data available were not consistent in the study area. Therefore, they were simply applied as constant by month, and the probabilistic TP model was calibrated using the monthly averaged TP loading rates and discharges from YD reservoir and $YD.Oc$ sub-basin without validation process.

Like the case of the BOD modeling, assuming the observations are independent, and the residuals between the observed TP (TP^{obs}) and simulated TP (TP^{sim}) are normally distributed with mean 0 and variance σ^2, the posterior probabilities of the parameters of k_s and k_r can be described by combining a likelihood function and their prior probability distributions using Bayes' theorem. In order to determine the means of the prior distributions for the parameters, the deterministic TP model based on Equation 6.7 and 6.8 was developed with *MS-Excel* and the parameters were determined optimally using the *SOLVER* optimization tool. The obtained rates were regarded as the means of the prior probabilities with a normal distribution. On the contrary, the variance was addressed with uninformative prior probability distribution. Equation 6.9 refers to the detailed structure of the steady state probabilistic TP model in month τ. Once the model is established, the posterior probabilities of the model parameters are sampled first, and

then the posterior predictive distribution of the dependent variable TP can be generated with the explanatory variables of *Q.DC* and *Q.YD*. All the modeling processes were applied exactly in the same way as was done in the BOD modeling. The MCMC diagnostics were based on the graph of history path, *Gelman-Rubin* statistic *R* and autocorrelation of the chains, and are described in *Appendix-C*. The source code of *WinBUGS* for TP model is described in *Appendix-C*.

$$\varepsilon = (TP_\tau^{obs}.DC - TP_\tau^{sim}.DC) \sim Normal(0, \sigma^2);$$

A. *likelihood function* :

$$p(TP_\tau.DC \mid k_s, k_r) = \prod_{\tau=1}^{12\ months} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(TP_\tau^{obs}.DC - TP_\tau^{sim}.DC)^2}{2\sigma^2} \right);$$

$$TP_\tau^{sim}.DC = \frac{W_\tau}{Q_\tau + k_s V_\tau};$$

$$(a)\ W_\tau = (1-k_r) \times W_\tau.YD + W_\tau.YD.Oc + W_\tau.Bc;$$

$$(b)\ V_\tau = V_{\tau-1} + (Q_\tau.YD + Q_\tau.YD.Oc + Q_\tau.Bc - Q_\tau.DC); \qquad (6.9)$$

B. *prior probability* :

$$k_s \sim Normal(k_s^{para1}, k_s^{para2});$$

$$k_r \sim Normal(k_r^{para1}, k_r^{para2});$$

$$\sigma^2 \sim Gamma(\alpha, \beta);$$

Figure 6-12 refers to the Bayesian learning results explaining that the steady state TP model is quite well calibrated in that the simulated values are consistent with the observations enough to be applied to forecasting TP in the *DC* reservoir. The errors of the results mainly come from the uncertainties of the discharges and phosphorus loads from the tributaries.

6.2.2 Risk Analysis of TP Standard Violation in the *DC* Reservoir in association with Operation of the *DC* and *YD* reservoirs

In this study, the relationships between the risk and the releases of the *DC* and *YD* reservoirs were established by month, which is a matter of two dimensional aspects consisting of two explanatory variables (*Q.YD* and *Q.DC*) with one dependent variable (*Tp.DC*). Like the BOD model, the monthly releases of the two reservoirs were discretized evenly first, and then the monthly averaged TP loads from the *YD* reservoir were computed with the monthly averaged TP concentrations of the *YD* reservoir. They were fed into the already trained probabilistic model with the monthly averaged flows and TP loads from the *Bc* tributary and the *YD.Oc* sub-basin. Figure 6-13 refers to the posterior predictive probability distribution of the predicted TP using the discretized releases of the *YD* and *DC* reservoirs in June. Figure 6-14 shows the three dimensional representation of the posterior predictive probability distribution in June. The rest of the posterior predictive distributions for all combined cases between the releases of the *DC* and *YD* reservoirs can be obtained in this way. Figure 6-15 shows the risks of the TP standard violation and the mean of the posterior predictive probability distribution in association with the releases of the two reservoirs in January, June and July. The results for the remaining months are described in *Appendix-C*. One can notice that the results show that the more the *YD* reservoir releases, the worse TP at the *DC* reservoir is. Finally, the risks (*R*) were replicated using an ANN with an architecture of 1 hidden layer and 5 hidden nodes for application. The model is represented in Equation 6.10 with a log-sigmoid transfer function where Q_1 and Q_2 refer to the releases of *YD* and *DC* reservoirs respectively, and Table 6-3 shows the parameters by month:

$$R = b_2 + \sum_{j=1}^{5} w_j \log sig\left[b_{1j} + \sum_{i=1}^{2} w_{ij} Q_i \right], \text{where } \log sig(z) = \frac{1}{1+e^{-z}}; \qquad (6.10)$$

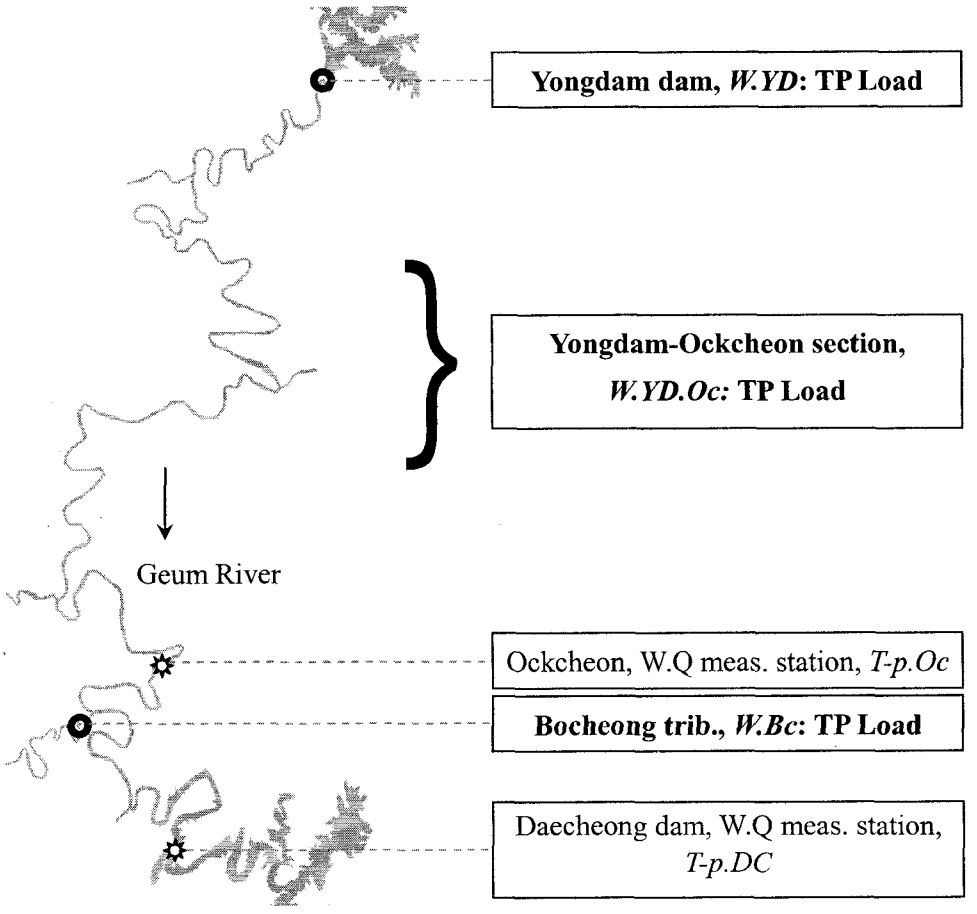


Figure 6-10 Schematic representation of the TP modeling system that is composed of the Yongdam and Daecheong multi-purpose dams and 2 water quality stations, where W.Q means water quality.

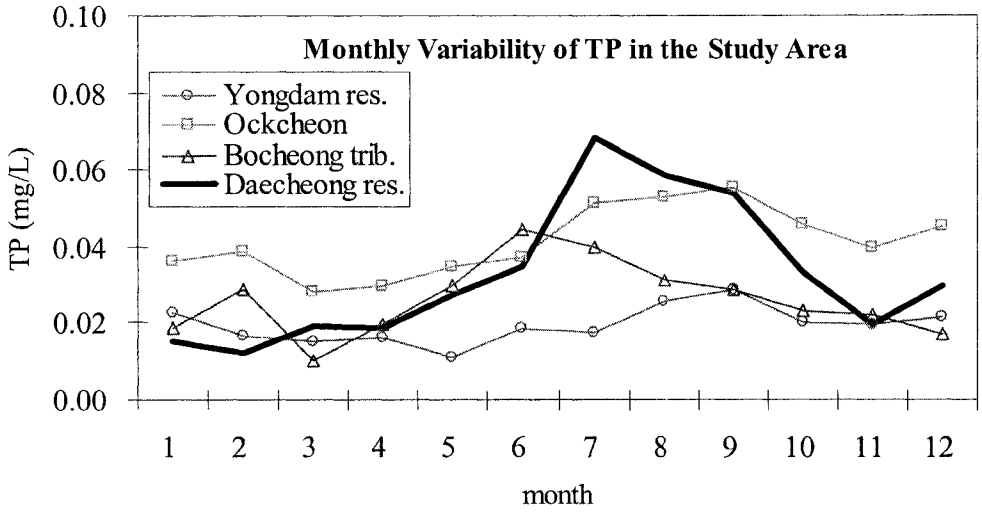


Figure 6-11 Monthly TP variability in the study area.

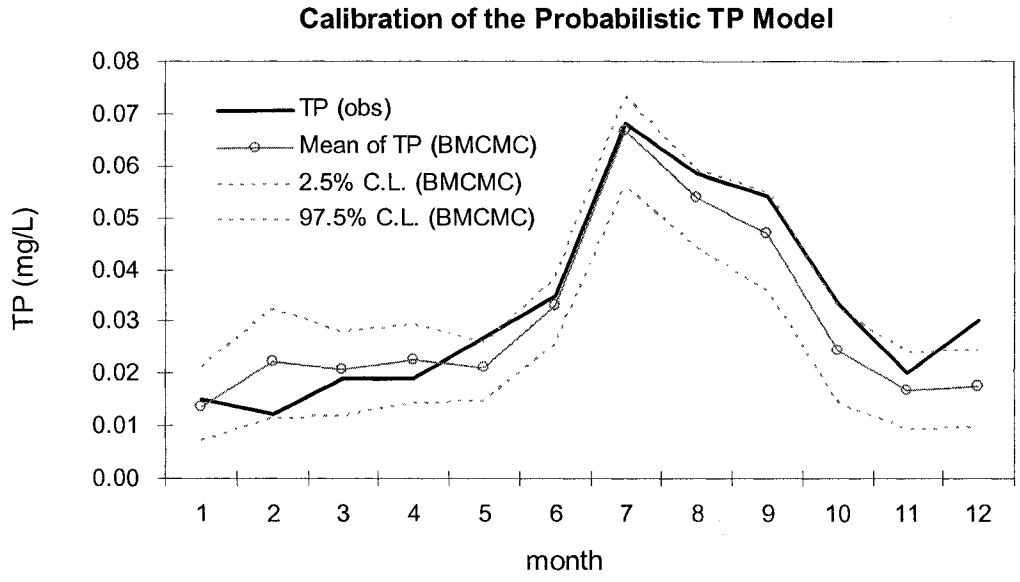


Figure 6-12 Calibration results of the probabilistic TP model.

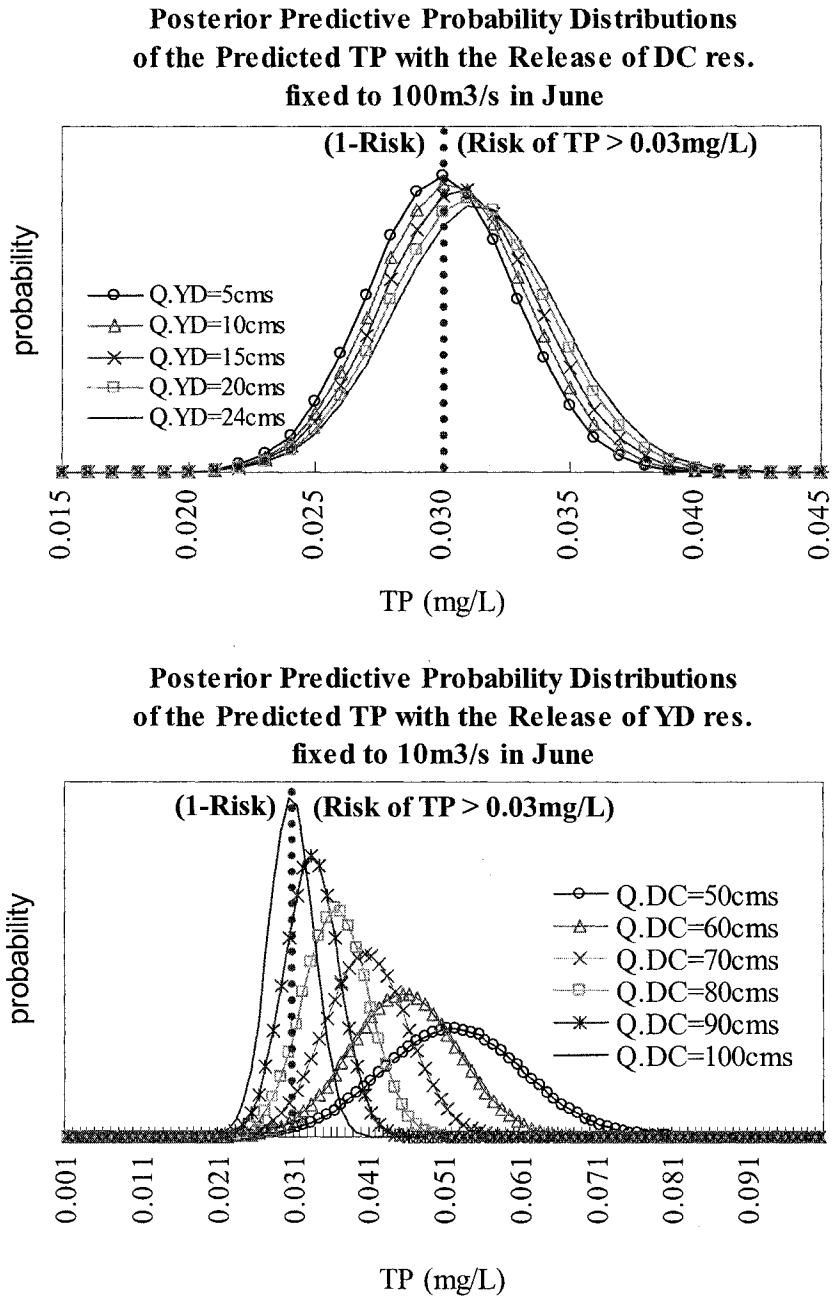


Figure 6-13 The posterior predictive probability distribution of the simulated TP with the release of Daecheong reservoir fixed to $100m^3/s$ (upper figure) and the release of Yongdam reservoir fixed to $10m^3/s$ (lower figure) with the concept of the risk in June.

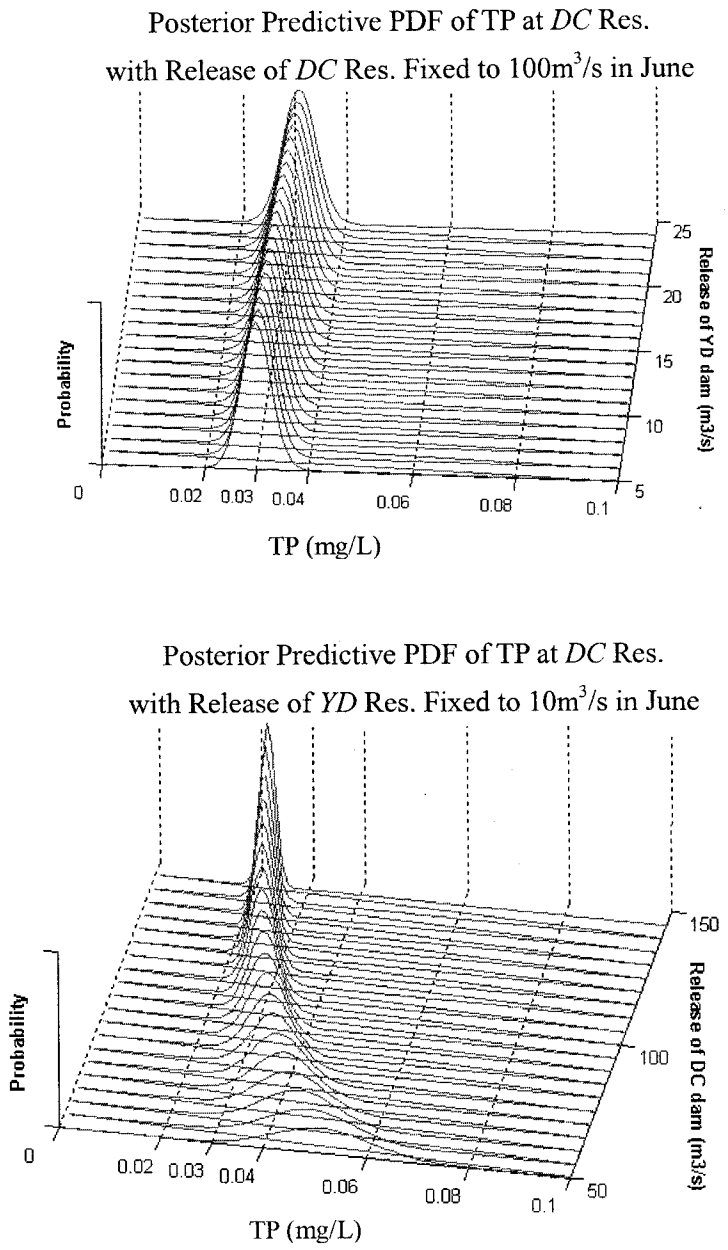


Figure 6-14 The three dimensional representation of the posterior predictive probability distributions of the simulated TP with the release of Daecheong reservoir fixed to $100m^3/s$ (upper figure) and the release of Yongdam reservoir fixed to $10m^3/s$ (lower figure) in June.

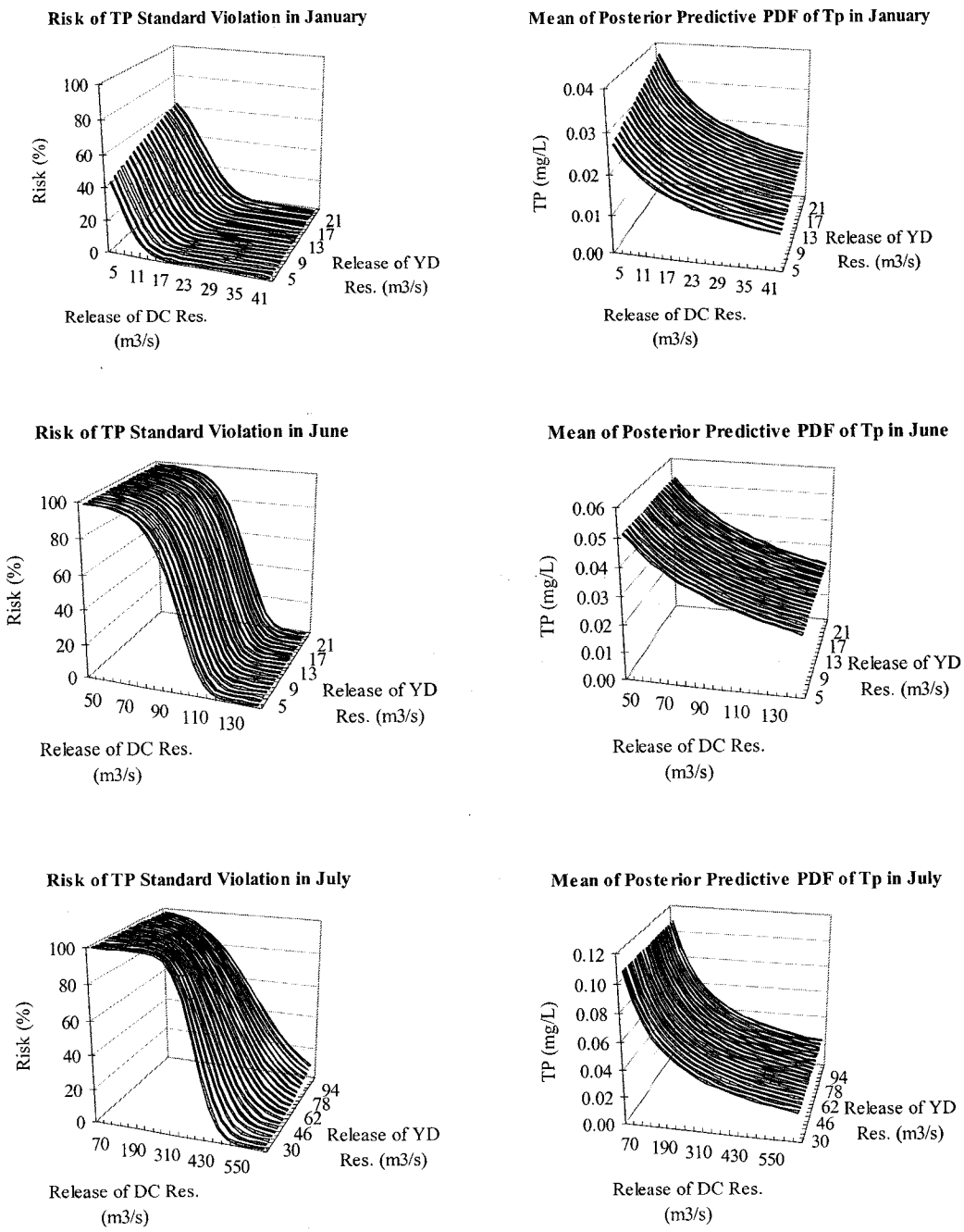


Figure 6-15 The risk of the TP standard violation and the mean of the posterior predictive probability distribution of the simulated TP versus the releases of the Daecheong and Yongdam reservoirs in Jan., Jun. and Dec..

Table 6-3 Parameters for the risks of TP standard violation represented in Equation 6.10 using ANN.

	w11	w21	w12	w22	w13	w23	w14	w24	w15	w25	
Jan	1.75	-0.63	0.42	-3.82	-0.86	3.49	0.76	-3.65	-1.20	1.09	
Feb	0.17	-2.04	0.06	-0.50	-0.02	0.72	0.01	-0.42	-0.33	2.40	
Mar	0.32	-2.67	-0.12	2.29	0.20	0.58	0.19	-0.62	0.53	-0.99	
Apr	-0.06	-0.91	-0.09	2.53	-0.28	-0.02	-0.15	0.45	0.28	-3.28	
May	0.07	1.14	0.18	0.44	-0.38	3.91	-0.12	0.94	0.07	-2.76	
Jun	-0.12	0.12	0.14	-3.08	0.03	0.21	0.29	-4.62	0.05	-1.03	
Jul	0.92	3.51	0.73	-3.35	1.11	-3.18	0.45	2.12	0.05	3.88	
Aug	-1.92	6.21	-1.13	-4.65	-0.20	4.70	-0.06	-4.89	-0.45	4.72	
Sep	0.63	-1.91	1.21	2.50	1.41	2.73	-0.43	2.15	-0.88	1.87	
Oct	0.10	-3.48	-0.30	4.66	-0.08	-0.26	-0.22	0.82	0.01	1.45	
Nov	0.08	-2.73	-0.03	-0.27	-0.13	-0.26	0.33	1.43	0.29	-2.92	
Dec	-0.53	2.47	0.16	0.54	0.21	-2.05	-0.05	-0.35	0.33	-0.57	
	b11	b12	b13	b14	b15	w1	w2	w3	w4	w5	b2
Jan	-2.89	-5.26	3.17	-1.87	-2.54	0.05	2.51	-2.27	-0.76	0.02	1.67
Feb	-1.04	0.45	-0.04	0.02	-0.26	1.64	-0.25	0.93	-0.33	-1.91	0.16
Mar	0.83	1.05	0.20	-0.16	-0.51	1.87	-1.55	0.52	-1.38	0.36	0.08
Apr	-0.83	1.15	0.02	0.09	0.85	-1.24	-1.44	-0.16	0.80	2.08	-0.21
May	1.20	0.18	-0.51	0.08	-1.41	1.13	-0.08	-2.22	0.56	1.14	-0.44
Jun	0.09	-0.56	0.09	1.43	0.85	-0.18	0.88	-0.19	1.76	-0.48	-0.92
Jul	0.96	2.50	1.02	-0.86	-1.01	-0.79	1.62	-0.66	0.84	-1.68	0.08
Aug	-7.43	1.10	-3.33	3.84	-3.01	-0.49	0.54	-16.91	-7.47	7.76	7.74
Sep	1.26	0.89	0.86	-1.82	-0.84	17.31	-5.56	4.46	7.99	7.12	-15.84
Oct	-2.61	1.09	0.23	0.54	1.76	1.32	-1.90	0.11	0.00	1.33	-0.39
Nov	-0.41	0.02	0.11	0.27	-2.11	-1.13	-0.20	-0.29	-0.40	3.90	-0.17
Dec	0.04	0.30	-1.17	0.08	0.46	-1.56	0.66	1.91	-0.32	-0.53	0.23

CHAPTER 7

RISK-BASED RESERVOIR SYSTEM OPERATION USING AN ADAPTIVE SAMPLING IMPLICIT STOCHASTIC OPTIMIZATION MODEL

This chapter presents an application of a progressive optimization method as an alternative approach to developing classical rule curves using the implicit and explicit stochastic optimization methods that have attracted lots of attention for reservoir operation. All the measures of system performance including energy production, water supply, and water quality management are important for the operation of the Yongdam and Daecheong reservoirs. To create an adaptive implicit stochastic reservoir optimization model, an adaptive sampling implicit stochastic optimization (ASISO) model was developed by coupling a reservoir model with the *k*-NN inflow forecast models. The ASISO model was evaluated by comparison with a sampling implicit stochastic optimization model using multiple inflow scenarios and with a deterministic dynamic programming optimization model using a single inflow scenario. A generalized microcomputer dynamic programming package, *CSUDP [Labadie, 1999]* was used to find the optimal joint operation of the two reservoirs. A decision support system (DSS)

framework was specially designed using *MS-Excel* to interactively connect the reservoir optimization model to BOD, TP and inflow forecast models.

7.1 Literature Review on Reservoir System Analysis

As far as management of reservoir systems is concerned, the following two main objectives considered in general are:

i) *Maximization of multiple benefits through optimal allocation of the limited water resources for various purposes;*

ii) *Minimization of multiple costs caused by flood damage, water shortages, etc.*

When many water resources system analysts try to develop models to provide decision support for managing reservoir systems, the hindrances that they are usually faced with result from the *1) high dimensional, 2) dynamic, 3) nonlinear, 4) multi-objective, and 5) stochastic* characteristics of the system. In order to determine the methodologies to cope with these challenging facets, a number of simulation and optimization models have been developed for reservoir system analysis. Network-flow models may be somewhere in between simulation and optimization models in the sense that they use optimization to perform the simulation. Network-flow models solve an optimization problem in each individual time period. Authors such as *Yeh [1985]*, *Wurbs [1993]*, and *Labadie [2004]* have described the theory and application of models to reservoir system operation, and presented state-of–the-art reviews related to reservoir system operation with a strong emphasis on optimization techniques.

Yeh [1985] reviewed state-of-the-art theories and applications of mathematical models applied for reservoir system analysis by classifying them into deterministic and stochastic approaches, and subdividing each of the two categories into optimization and simulation models. He provided critical reviews and in-depth analyses of the various models. According to *Wurbs [1993]*, although optimization and simulation are conceptually different, the distinction is somewhat obscured by the fact that most models, to various degree, contain elements of both approaches. He also states that most of the decision support systems within the water resources management agencies have focused on simulation models, while the academic community and research literature have put special emphasis on optimization techniques. *Labadie [2004]* pointed out possible reasons for the continuing gap between theoretical developments of optimization techniques and the real-world implementation: they are more mathematically complicated than simulation approaches; simulation models are ideal for "what-if" analysis that system operators are well versed in; simulation models can be easily combined with Monte Carlo techniques for uncertainty analysis, while many optimization models are not conductive to incorporating risk and uncertainty, and; simulation models can address the above mentioned five characteristics of reservoir system more flexibly than optimization models. However, one of the greatest features of prescriptive optimization models is that these models have an expanded capability to select optimal solutions systematically under agreed upon objectives and constraints *[Labadie, 2004]*.

As the key to success in implementation of reservoir optimization models: *Yeh [1985]* recommended intensive focusing on development of stochastic models with consideration of risk and uncertainty, combination of simulation and optimization models

with considering multi-dimensional and multi-objectives, and; *Labadie [2004]* pointed out improved linkage with simulation models and involvement of decision makers in system development. The most essential elements for reservoir system operation can be summarized as *1)* combination of optimization and simulation algorithms, *2)* incorporation of risks and uncertainties of water resources variables, and *3)* adoption of integrated river basin planning and management considering multi-objectives. The literature related to optimal reservoir operation is extensive, however most applications to reservoir system analysis involve linear programming (LP), non-linear programming (NLP), and dynamic programming (DP). DP is a popular approach because it is well suited for solving problems with a sequential decision structure, it handles nonlinearity, and it establishes feedback policies. Therefore, centering on the key elements to success in real world application of optimization models, this review focuses on reservoir optimization methods that use DP:

i) The handling of the stochastic nature of reservoir systems can be a relatively daunting task *[Marino et al., 1985]*. When it comes to stochastic characteristics of reservoir systems, both water demand and reservoir inflow are the main issues. However, most literature addressing stochastic optimization has focused on reservoir inflow because forecasting water demand is so susceptible to a variety of economic, social, and political issues, which are often hard to predict. For incorporating the stochastic nature of inflow into reservoir optimization models, two different techniques, implicit and explicit stochastic optimization are commonly used in conjunction with first-order Markov chain models. Implicit optimization employs Monte Carlo techniques deterministically with synthetically generated unregulated reservoir inflow sequences. Explicit optimization

deals with transition probabilities that present the discrete joint probabilities among sequential months' inflow series *[Labadie, 2004]*. Explicit stochastic optimization (ESO) tries to maximize the average system performances by assigning probabilities to inflow series. Computational difficulties in association with explicit stochastic optimization have led reservoir managers to rely on implicit stochastic optimization (ISO) techniques *[Labadie, 1997]*. Although the ISO technique is computationally simpler, the biggest drawback of ISO is that the resulting optimal policies are unique to the hydrologic time series used for the optimization. Repeated optimization for different sequences of historical or synthetic flows may produce significantly different optimal solutions unless the period-of analysis is extremely long. A variation of stochastic optimization, known as sampling stochastic dynamic programming (SSDP), was applied by *Kelman et al. [1990]*. The strategy of SSDP is to mix the deterministic and stochastic approaches with the aim to overcome the computational difficulties of ESO by representing the stochasticity of stream flow explicitly. SSDP determines optimal solutions by considering all possible stream flow scenarios simultaneously. For a single stream flow scenario, SSDP reduces to a deterministic model solved with perfect foresight;

ii) As for the combination of optimization and simulation algorithms, *Labadie [1993]* showed an outstanding application by developing the optimal monthly end-of-month storage guide curves for operation of Valdesia reservoir in the Dominican Republic using ESO, and incorporated them into a network simulation model for weekly real-time operations. In general, reservoir operations rely on release and storage rule curves developed by interpreting the results from ISO (or sometimes ESO) using multivariate regression or pattern recognition techniques such as artificial neural

networks and fuzzy inference systems. However, the developed rules are not adaptive, to various extents, to unexpected variations such as changes in the original water demands and climate. Therefore, a reservoir system with existing rule curves needs to be reevaluated periodically: when new unplanned water demands such as instream flows for environmental concern arise, and; when considerable changes of climate and severe droughts or floods occur. Furthermore, it may be difficult to develop multi-variate rule curves considering other variables such as water quality, and to reflect various users' options such as the ending storage into rule curves.

An alternative to the rule curves is to employ the progressive optimization based operations called naive feedback control. This approach uses successive running of an optimization model instead of fixed rule curves. At the beginning of each period, forecast information is updated, which is conditional on the current and historical information, and the optimization model is run. The entire process is repeated for each period until the end of the operational time horizon. Hence, this method makes use of optimization models in adaptive manner similar to a simulation process. Reservoir system managers should find this approach more flexible than using rule curves in that they can learn the behavior of reservoir system through simulation with successively updated information and users' options over operational guidelines. Furthermore, this does not require a reevaluation process of reservoir systems since reservoir operational policies are updated successively. The concept of this approach is very similar to the algorithm of Kalman filtering. Figure 7-1 illustrates the concept of the progressive optimization based operation.

iii) The hindrances in handling multi-objective optimization are: some objectives for reservoir operation are continuous and some are discrete, moreover; they are often non-commensurate and conflicting. One of the solutions to this problem is to combine multi-objective optimization (MOP) and multi-criteria decision analysis (MCDA). MOP is used for the continuous objectives for developing nondominated solutions defining a Pareto optimal surface of tradeoffs between the objectives. Techniques for generating nondominated solutions include the weighting method, ϵ-constraint method and goal programming. A nondominated solution is defined as a multi-objective solution that can not possibly be improved in one objective without adversely affecting one or more other objectives. Once a finite number of discrete alternatives are selected from the Pareto tradeoff region, they can be evaluated and ranked using the techniques for MCDA such as *ELECTRE*, *AHP*, and *PROMETHEE*. For more information on this approach and its application, see *Ko et al. [1992]*.

iv) In the context of integrated river basin management, many researchers have tried to incorporate water quality models into reservoir simulation or optimization models *[Labadie, et al., 1994, and Ko, 1997]*. A common feature of this research has been to establish reservoir operation policies to meet water quality standards by simulating water quality in deterministic way. However, the simulated downstream and reservoir water qualities affected by reservoir release is actually uncertain as discussed in Chapter 6: hence, as stated in Chapter 1, a risk-based water quality management might be more persuasive in the decision making process, and the probabilistic water quality models developed in Chapter 6 are worth consideration for replacing deterministic models.

This study was motivated by the importance of: reflection of hydrologic uncertainties into a reservoir optimization model; adaptive operation of a reservoir system; combination of simulation and optimization algorithms, and; integrated river basin management considering both water quantity and quality management. This research proposes an alternative for reservoir system operation for realizing this motivation. In order to treat adaptiveness and stochasticity, an adaptive sampling implicit stochastic optimization (ASISO) system was developed, which is coupled with a stochastic inflow forecast system based on a *k*-NN nonparametric technique. For integrated river basin management, the water qualities of BOD and TP are evaluated in the stream and reservoir along with other reservoir system performance measures. The strategy of the methodology is to mix deterministic and stochastic algorithms by running the optimization model progressively in a "simulation process" with the ensemble of inflow series forecasted adaptively.

The term of "adaptive sampling implicit" refers to successive execution of a deterministic optimization model using the ensemble derived by successive data-assimilation instead of explicit consideration of transition probabilities of the inflow series. ASISO may be preferred to ESO in that ASISO can come up with a lot of alternatives with multiple inflow scenarios, while ESO maximizes the average system performances by assigning probabilities to inflow series.

As a feature of the ASISO methodology, the ensemble is created by sampling a number of percentile inflows from the conditional probability distributions of the forecasted inflow series during the period-of analysis of 12. Like SSDP, for a single stream flow scenario, ASISO reduces to a deterministic dynamic programming (DDP)

model solved with perfect foresight. Because ASISO model is run in an adaptive mode, only the optimal release policy for the past or current time period is implemented.

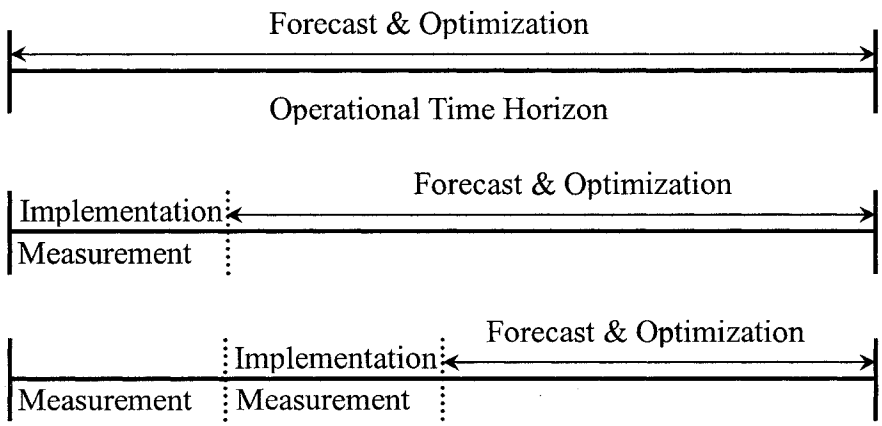


Figure 7-1 Illustration of progressive reservoir optimization

7.2 Case Study: Geum River Basin in Korea

For implementing the above mentioned ASISO based operation of reservoir system in practice, a case study was performed based on the Geum River reservoir system, Korea. Figure 7-2 shows the study area composed of the two cascade reservoirs of Daecheong (*DC*) and Yongdam (*YD*). The case study includes the following sequential sub-tasks:

- development of a probabilistic inflow forecast system using *k*-NN nonparametric modeling, which is essential for creating the ensemble;
- development of a monthly multi-objective ASISO model using *CSUDP*;

- development of a decision support system, which is composed of a data base, user interface, and model base integrating ASISO, inflow forecast and water quality models.

For water quality management in this area, the two time-varying variables of BOD at the Gongju gage and TP in the Daecheong reservoir were addressed, which interact with the operational policies of the two reservoirs. With the water quality standards of 3mg/L for BOD and 0.03 mg/L for TP, the risks of water quality standard violation were derived by month as described in Chapter 6. The risk of BOD standard violation at the Gongju gage described in Table 6-2 is a function of the release of the *DC* reservoir, while the risk of TP standard violation in the *DC* reservoir described in Table 6-3 is a function of the releases of both *DC* and *YD* reservoirs. The risks were interactively used in the reservoir system optimization model: hence it can be said that the probabilistic water quality models were incorporated implicitly.

7.2.1 Development of the Monthly Inflow Forecast Systems using a *k*-Nearest Neighbor Estimate in the *DC* and *YD* Reservoirs

Forecasting the reservoir inflows in an adaptive mode is a core process in the progressive optimization based operation. For implementing ASISO model, the lag-1 autoregressive forecast models were developed for the *YD* reservoir and the *YD-DC* sub-basin where the *YD* reservoir basin is excluded from the entire *DC* reservoir basin because the inflow of the *DC* reservoir is the sum of the release of the *YD* reservoir and

the flow from the *YD-DC* sub basin. The *k*-NN forecast models were developed and interactively joined to the reservoir system operation model in a decision support system. The stochastic inflow forecast models aim at producing the conditional probabilities in every month as final products. From the derived probabilities, a variety of percentile inflows can be sampled, or one or more selected by system managers to develop several scenarios. The forecast information is updated in every month on the basis of the observed inflow, and the update is repeated until the end of the operational time horizon. In the context of progressive operation, the forecast lead time was fixed to 12 months regardless of the current time for keeping pace with running the ASISO model.

First, the historical inflow record of the *DC* reservoir for 25 years from 1981 to 2005 was separated into the flow series of the *YD* reservoir and the *YD-DC* sub-basin on the basis of the historical inflow series of the *YD* reservoir for 4 years from 2002 to 2005. Figure 7-3 shows box-whiskers plot of the historical monthly inflow with outliers beyond the 97.5 and 2.5 percentiles. The entire processes described in Figure 5-12 were implemented for developing the forecast models in this study area. Like the case of the Chungju reservoir in Chapter 5, k was applied equally in every month in the sense of a lumped parameter, and k of 5, that amounts to $k = n^{1/2}$, was chosen, which is an *ad hoc* prescriptive choice. In this case study, n is the total number of the data points that is 25. Figure 7-4 refers to the calibration results, which shows that the simulation results match the calibration data set quite well. Unlike the case of the Chungju reservoir, validation of the models was not made because of an insufficient historical record. The *k*-NN forecast models were developed using *VBA* built into *MS-Excel*.

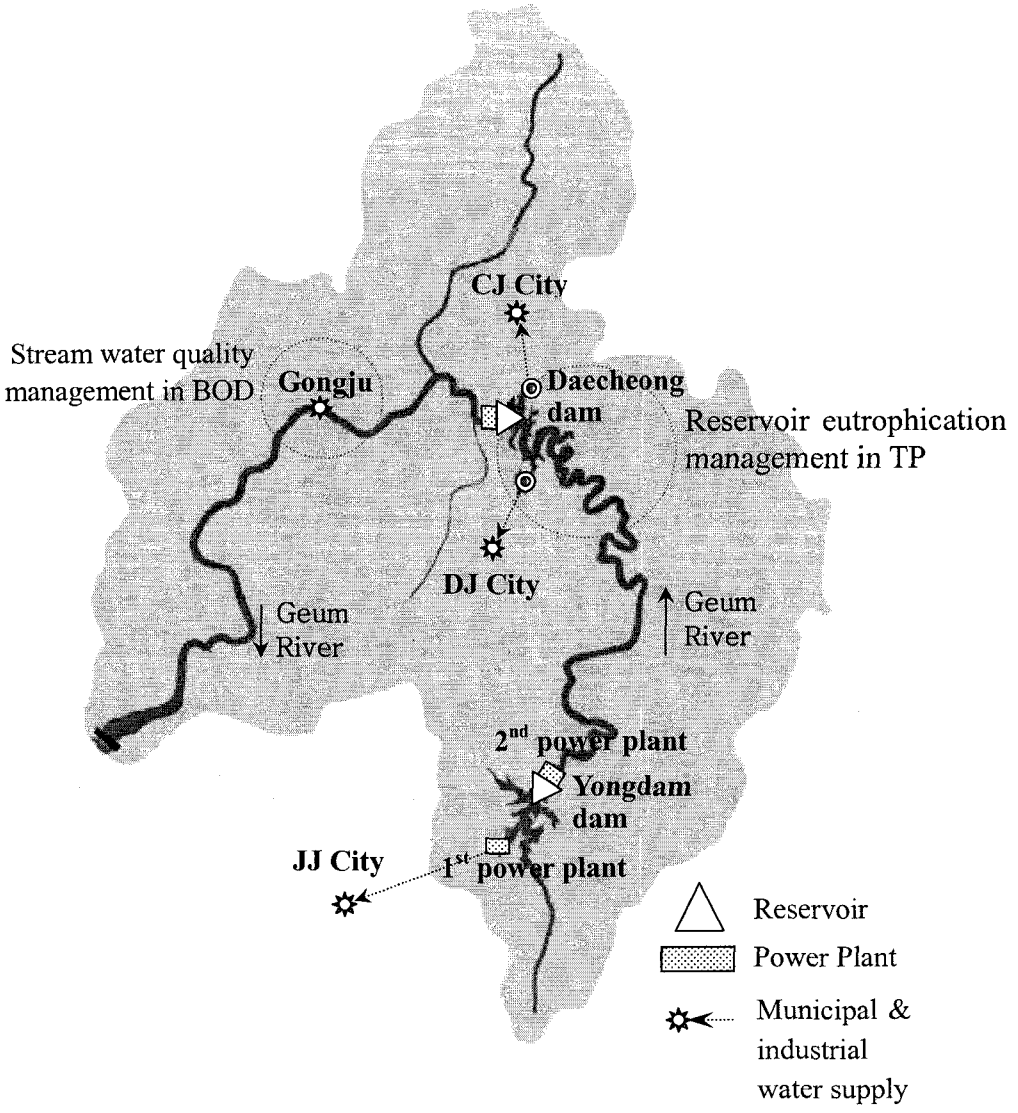


Figure 7-2 Configuration of reservoir system composed of the two cascade dams, the Yongdam and Daecheong.

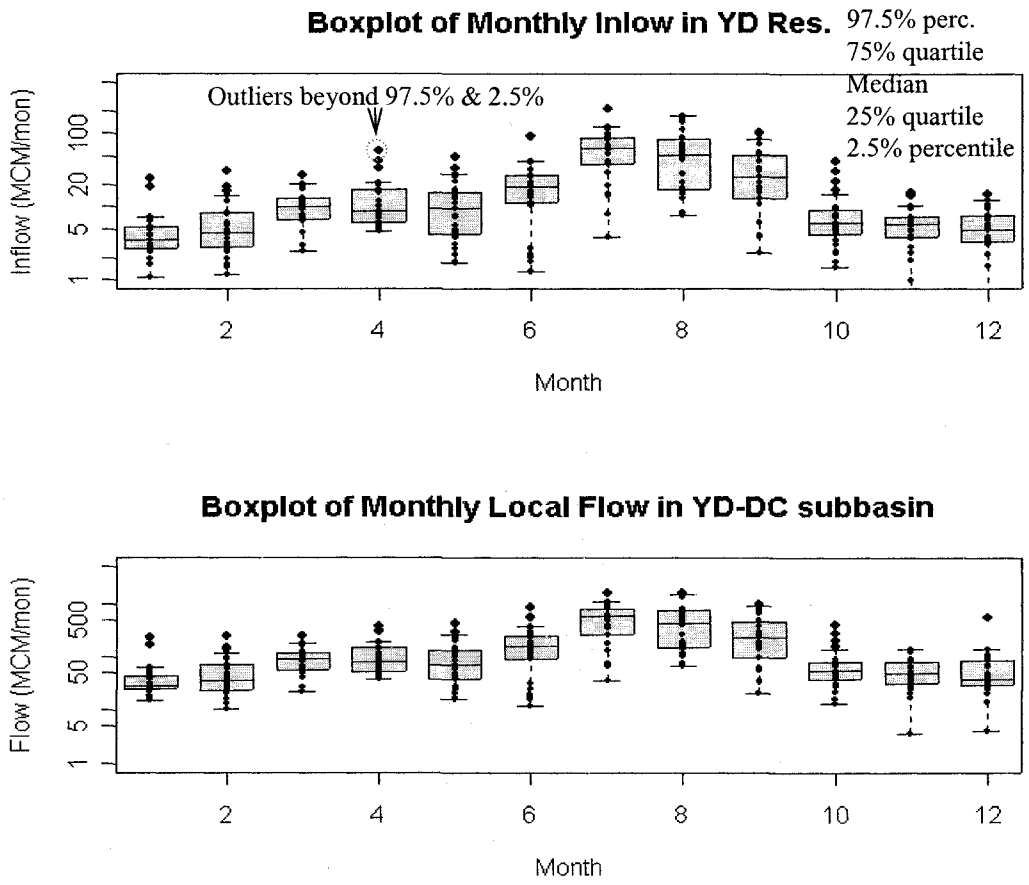


Figure 7-3 Box and whiskers plots of the monthly inflow in the Yongdam reservoir and the local flow in the *YD-DC* sub basin for 25 years.

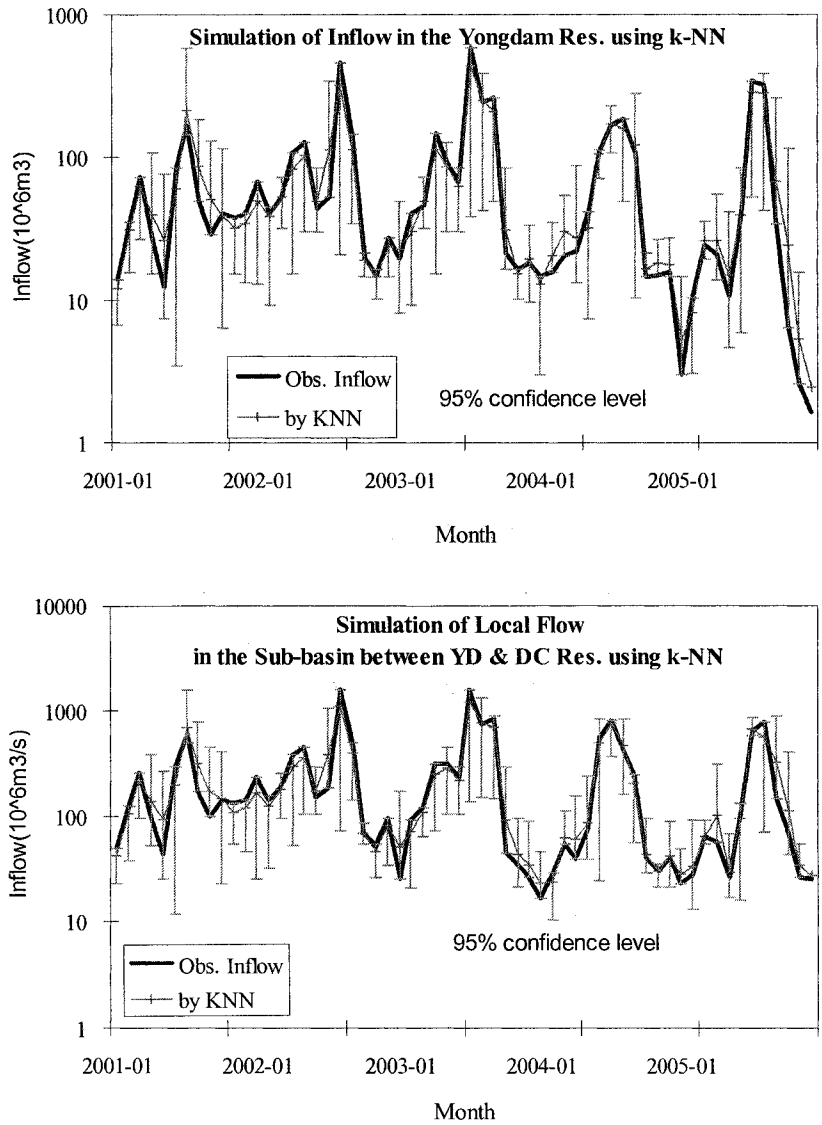


Figure 7-4 Calibration results of the *k*-NN models with the historical records from 1981 to 2005 in the Yongdam reservoir and the sub-basin between the *YD* and *DC* dams.

7.2.2 Development of the ASISO Model for Reservoir Optimal Operation

7.2.2.1 Formulation of the Reservoir System

Defining quantitative measures of reservoir system performance is a key element in formulating an optimization model. In this case study, these performance measures consist of: the total annual energy production; the firm water supply; the risk of BOD and TP standard violation at the Gongju gage in association with instream flow requirement and at the *DC* reservoir in association with eutrophication, respectively. Firm water supply is defined as the estimated maximum release that can be maintained continuously during an operational period *[Wurbs, 1996]*. The performance measures are functions of the releases of the two reservoirs. Multiple objectives addressed in this study are composed of maximization of the total annual energy, the firm water supply and the extent of satisfaction of BOD and TP target risks, subject to satisfying the strict or soft operational guidelines, summarized below. As for the BOD performance measure, once a user-selected target risk is given, a release (called the risk equivalence release in this study) of the *DC* reservoir can be obtained from the pre-analyzed relational table between BOD risk and the risk equivalence release (Table 6-2). The performance is assessed by comparing with the actual release of the *DC* reservoir. On the other hand, the case of TP is different in that an unique set of the risk equivalence releases of the two reservoirs does not exist, given a target risk. Therefore, TP risk in the *DC* reservoir should be first computed with the releases of the *YD* and *DC* reservoirs using the ANN function described in Table 6-3, and the TP performance measure is evaluated by comparing a user-selected TP risk with a computed risk. The operational guidelines are:

(1) For the purpose of securing the flood control pool, the water level of the *YD* reservoir should not exceed 261.5 El.m which amounts to 672.84 MCM (million cubic meters: $10^6 m^3$) during the flood season extending from the end of June to September. However, the *DC* reservoir does not necessitate a strict flood control pool in that it controls floods in conjunction with the *YD* reservoir. Hence, system managers can select the flood control pool around the normal full storage level of 76.5 El.m which amounts to 1,242.7 MCM. The properties of two reservoirs are summarized in Table 7-1.

(2) The water supply is divided into two parts: one is for the water demands in compliance with the traditional water rights for irrigation and minimum instream flow requirements in the downstream areas. The water supply is delivered through turbines for hydropower generation. The other part is for the municipal and industrial water demands in Daejon (*DJ*) and Cheongju (*CJ*) cities, and Jeonju (*JJ*) city, which are diverted through the outlet works intake structure of *DC* reservoir and the 1st hydropower plant of the *YD* reservoir, respectively. Figure 7-2 illustrates a configuration of the reservoir, and Tables 7-3 and 7-4 describe the water demands. The highest priority water right from the *YD* reservoir is the minimum instream flow of 5 m^3/s established for water quality management in the *DC* reservoir. The *YD* reservoir can release 16.1 to 25.8 m^3/s depending on the water level, and the release greater than 25.8 m^3/s should be spilled. The municipal and industrial water demands are programmed in the state function of *CSUDP* such as to be satisfied 100% of the time. Therefore, the release for hydropower generation and instream flow requirements become decision variables.

(3) As for the constraints on hydropower generation, the *YD* and *DC* reservoirs generate for 24 hours of base load and 5 hours of daily peak generation during the non-

flood season with the power capacity described in Table 7-1. The generation hours of the *DC* reservoir vary in accordance with flood conditions during the flood season.

(4) For water supply after the flood season, the conservation pools of the reservoirs should be recovered at the end of flood season, which is often given as target storage. However, determination of this target storage is very difficult due to the high uncertainty of flood events: hence, this storage was handled in the DSS such that system managers can select it. Similarly, the ending storage for December was provided as an option for maximization of the profit from hydropower generation.

In compliance with the above mentioned guidelines, the reservoir system was formulated with a following single scalar objective function that combines multiple objectives using the weighting method (Equation 7.1). The operational lead time was fixed to 12 months regardless of the current time in a progressive optimization process. For example, if the current period is the end of March, the end of the operational year is the next March. To quantify the objectives in commensurate units, the original objective values were transformed into the relative values with respect to the maximum capacity of hydropower generation, and the targets for water supply and water quality protection in terms of BOD and TP. The relative values reflect a surrogate of reliability, and were adjusted so as to have the range of 0 to 100. The reservoir system was programmed into *CSUDP* using the inverted form, and solved using the algorithm of incremental dynamic programming (IDP) that is a solver for multi-dimensional problems. The initial trajectory was obtained by averaging the historical operation record, and the initial and final discretization intervals of reservoir storage were set to 5 MCM and 0.1 MCM respectively for both the *YD* and *DC* reservoirs.

$$F = Max \sum_{j=1}^{4} \omega_j F_j, \text{ where:}$$

F_1 = maximization of the total annual energy

$$= Max\left[\sum_{t=1}^{T}\sum_{i=1}^{I} \frac{Energy_{i,t}(X_{i,t}, X_{i,t+1}, u_{i,t})}{pow.capa(i) \times \max.gen.hrs(i)} \times 100\right] : \begin{matrix} T = 12 \text{ months} \\ I = 3 \text{ power plants} \end{matrix}$$

$i = 1(YD\ 2nd \text{ power plant}), 2(DC \text{ power plant}), 3(YD\ 1st \text{ power plant})$

F_2 = maximization of the firm water supply

$$= Max\left[-\sum_{t=1}^{T}\sum_{d=1}^{D} \frac{S_{d,t}}{T.ws_{d,t}} \times 100\right] : \begin{matrix} D = 2 \text{ dimensions of } YD \text{ \& } DC \\ d = 1(YD \text{ reservoir}),\ 2(DC \text{ reservoir}) \end{matrix}$$

for $S_{d,t} = (T.ws_{d,t} - u_{d,t}) \geq 0;$

F_3 = satisfaction of the target risk of BOD standard violation

$$= Max\left[\sum_{t=1}^{T} \frac{u_{2,t}}{(T.u.BOD)_t} \times 100\right],$$

for $u_{2,t} \leq (T.u.BOD)_t;$

F_4 = satisfaction of the target risk of TP standard violation

$$= Max\left[\sum_{t=1}^{T} \frac{(T.risk.TP)_t}{Comp.risk.TP_t(X_{1,t}, X_{1,t+1}, u_{1,t}, u_{2,t})} \times 100\right],$$

for $(T.risk.TP)_t \leq Comp.risk.TP_t;$

subject to, (7.1)

$u_{1,t} = X_{1,t} - X_{1,t+1} + I_{1,t} - E(X_{1,t}, X_{1,t+1}) - WS_{JJ,t};$

$u_{2,t} = X_{2,t} - X_{2,t+1} + I_{2,t} + u_{1,t} - E(X_{2,t}, X_{2,t+1}) - WS_{DJ\_CJ,t};$

$\min.X_{d,t} \leq X_{d,t} \leq \max.X_{d,t};$

$\min.u_{d,t} \leq u_{d,t} \leq \max.u_{d,t};$

$Energy_{i,t}(MWh) = 9.81 \cdot \eta \cdot Q_{i,t} \cdot \bar{H}_{d,t} \cdot 720(hours/month);$

if $X_{d,t=flood\ periods} > X_d.T.flood\ control$, then high penalty

if $X_{d,t=11} \neq$ within 10% deviation range of $X_d.T.October$, then high penalty

if $X_{d,t=13} \neq$ within 10% deviation range of $X_d.T.December$, then high penalty

Where:

F = overall objective function;

F_j = j^{th} sub-objective function;

ω_j = weighting factor of j^{th} sub-objective;

$u_{d,t}$ (MCM) = release of d^{th} reservoir at period t;

$X_{d,t}$ (MCM) = beginning storage of d^{th} reservoir at period t;

$X_{d,t+1}$=ending storage of d^{th} reservoir at period t, or beginning storage at period t+1;

$I_{d,t}$ (MCM) = inflow of d^{th} reservoir at period t;

E_t (MCM) = evaporation at period t;

$WS_{JJ,t}$ (MCM) = water supply to Jeonju city at period t;

$WS_{DJ\_CJ,t}$ (MCM) = water supply to Daejon and Cheongju cities at period t;

pow.capa(i) (MW) = power capacity in MW of i^{th} power plant;

η = turbine efficiency;

$Q_{i,t}$ (m^3/s) = turbine discharge of i^{th} power plant at period t;

$H_{d,t}$ (m) = net head (m) of d^{th} reservoir at period t, computed using Table 7-2;

$S_{d,t}$ (MCM) = shortage of water supply of d^{th} reservoir at period t;

$T.ws_{d,t}$ (MCM) = water supply target (or, maximum shortage) of d^{th} reservoir;

$(T.u.BOD)_t$ (m^3/s) = release target to meet a user-selected BOD risk at period t;

$(T.risk.TP)_t$ (%) = user-selected target of TP risk at period t;

$(Comp.risk.TP)_t$ (%) = computed TP risk at period t;

$X_d.T.flood\ control$ = target storage for flood periods;

$X_d.T.October$ = target ending storage in October;

$X_d.T.December$ = target ending storage in December.

7.2.2.2 Development of Data-driven DSS for Progressive Optimization

A DSS was designed for embodying the ASISO based monthly reservoir operation in a data-driven way by interactively connecting the ASISO to the inflow, BOD and TP forecast models. The DSS was also designed so that users can select one of two choices of ASISO with multiple inflow scenarios and DDP with a single inflow scenario. The focal point of the DSS is to help system managers run the ASISO model sequentially like a simulation model with selection and adjustment of the following user options:

- the ending storage in December;
- the limited storage for flood control;
- the ending storage in October after the flood season;
- the reliability of the forecasted inflow in the case of running the DDP;
- the target risks of BOD and TP standard violation;
- the weighting factors for the multi-objectives.

The DSS was developed using *MS-Excel* and Visual Basic Application (*VBA*). The DSS consists of a model base, a flat file type of data base, and a user interface. Figure 7-5 illustrates the architecture of the DSS, and Figure 7-6 shows its screen capture. The procedures for running the DSS are:

(1) A year and month when the operation of the system starts are selected as an initial stage. This connects to the data base that contains the historical operation record and historical inflow series;

(2) The reservoir inflow forecast models are run, and the conditional probabilities of the inflows are obtained to the end of an operational year;

(3) The above mentioned options that affect the behavior of the system are selected by system managers;

(4) Running the ASISO model comes up with a lot of alternatives for operational policy, and users may select an averaged optimal policy. Instead of the ASISO, users can run DDP with a single inflow scenario corresponding to the reliabilities that users select;

(5) Post analysis of the performance measures is made in graphs and tables;

(6) If creating multiple alternatives is required in the case of running the DDP, the procedures from "*iii)*" to "*v)*" are run successively with a number of different inflow scenarios;

(7) Once a stage is finished, the next stage's operation is initiated by implementing step "*i)*" again.

7.2.2.3 Evaluation of the progressive optimization model

The developed ASISO and DDP model were evaluated by comparing the model results with the historical record of the reservoir system operation in terms of storage, release, inflow, BOD and TP for the years 2004 and 2005. The ASISO model was run with 100 iterations. The historical operations of the two reservoirs were based on the monthly rule curves developed using ISO with the performance measures of water supply and hydropower production. In order to compare the same conditions, the models were run with the ending storages of December fixed to those of the historical operation record. The values averaged from the historical storages of October were considered the

ending target storages of October for the two reservoirs. The *DC* and *YD* reservoirs were controlled not to exceed the normal pool storage and the flood control pool storage during the flood season respectively. The weighting factors of 1.0, 10.0, 1.0, and 1.0 were assigned for the objectives of hydropower production, water supply, and BOD and TP risks respectively aiming at meeting the downstream demands 100% of the time. The monthly forecasted inflows with the reliability of 50% were chosen in the case of running the DDP, and the following target risks of BOD and TP standard violation were selected arbitrarily.

	Target risks of BOD and TP standard violation (%)											
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
BOD	90	90	90	90	90	100	100	100	100	90	90	90
TP	30	30	30	30	50	90	100	100	100	70	30	30

Figure 7-7 shows the comparison of the results of the ASISO and DDP models with the historical operation record. From Figure 7-7 (a) and 7-7 (b), it can be noticed that as expected, shortages do not take place and almost all of the constraints for BOD and TP are met for the simulation years. The storages created from the ASISO shows a very small range in October to December because of the assigned target storages for these periods. The model results are different from the historical record because the forecasted inflow series was not the same as the observed inflow, and the performance measures were different. The remarkable difference between the two results is that the historical storages of two reservoirs were lower than the results from ASISO for the analysis period, especially the flood season. This difference can be understood by a fact that

reservoir system managers operated the reservoir system very safely with low water levels during the flood season due to highly uncertain flood characteristics in this area. This operation led to a failure to recover the normal pool after the flood season, and had a harmful influence on the normal operation during the non-flood season afterwards. Figure 7-7 also shows that the storages from the DDP with the inflows with the reliability of 50% and the mean storages from the ASISO with an ensemble scenario for the inflow are slightly different. However, the storages from DDP were almost within the minimum and maximum storages from ASISO of the time, which were created using an ensemble scenario. This indicates although the ASISO is better than DDP from a conceptual viewpoint, the adaptive DDP can be applied effectively in practice.

In conclusion, compared to the reservoir operation using classical rule curves, ASISO and adaptive DDP models make a difference in that: they can reflect the risks of water quality violation, and; system managers can create a lot of alternatives of operation policy by running them like simulation models with a variety of options.

Table 7-1 Properties of the two dams and reservoirs in the Geum river basin.

Item	Daecheong	Yongdam
F.W.L (El.m)	-	265.5 EL.m (672.84 MCM)
N.H.W.L (MCM)	1241.7	742.6
L.W.L (MCM)	451.7	68.7
Eff. Storage (MCM)	790	672
Pow. Cap.(MW)	90	22.1 (1^{st} power plant) 2.3 (2^{nd} power plant)
T.W.L (El.m)	29.0	76.7 EL.m (1^{st} power plant) 205.0 EL.m (2^{nd} power plant)
Efficiency of Pow. (%)	86	86
Max.Q.Turbine (m^3/s)	264.0	17.5 (1^{st} power plant) 6.2 (2^{nd} power plant)
Avg. Hrs of Pow. Gen.	5	24
Basin Area(km^2)	4134	930

Table 7-2 Relationship between water level (H) and reservoir storage (V).

Yongdam	H=1.561E-07*V^3-2.919E-04*V^2+1.989E-01*V+2.119E+02
Daecheong	H=6.670E-09*V^3-2.835E-05*V^2+5.362E-02*V+4.077E+01

Table 7-3 Municipal and industrial water demands. unit: MCM/mon

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
DJ&CJ	34.3	31.0	34.3	33.2	34.3	33.2	34.3	34.3	33.2	34.3	33.2	34.3
JJ	11.2	10.2	11.2	10.9	11.2	10.9	11.2	11.2	10.9	11.2	10.9	11.2

Table 7-4 Water demands for Irrigation and minimum instream flow in downstream area. unit: MCM/mon

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
DC	54.4	49.1	54.4	55.0	64.8	151.4	119.5	111.2	99.5	54.4	52.6	54.4
YD	13.4	12.1	13.4	13.0	13.4	13.0	13.4	13.4	13.0	13.4	13.0	13.4

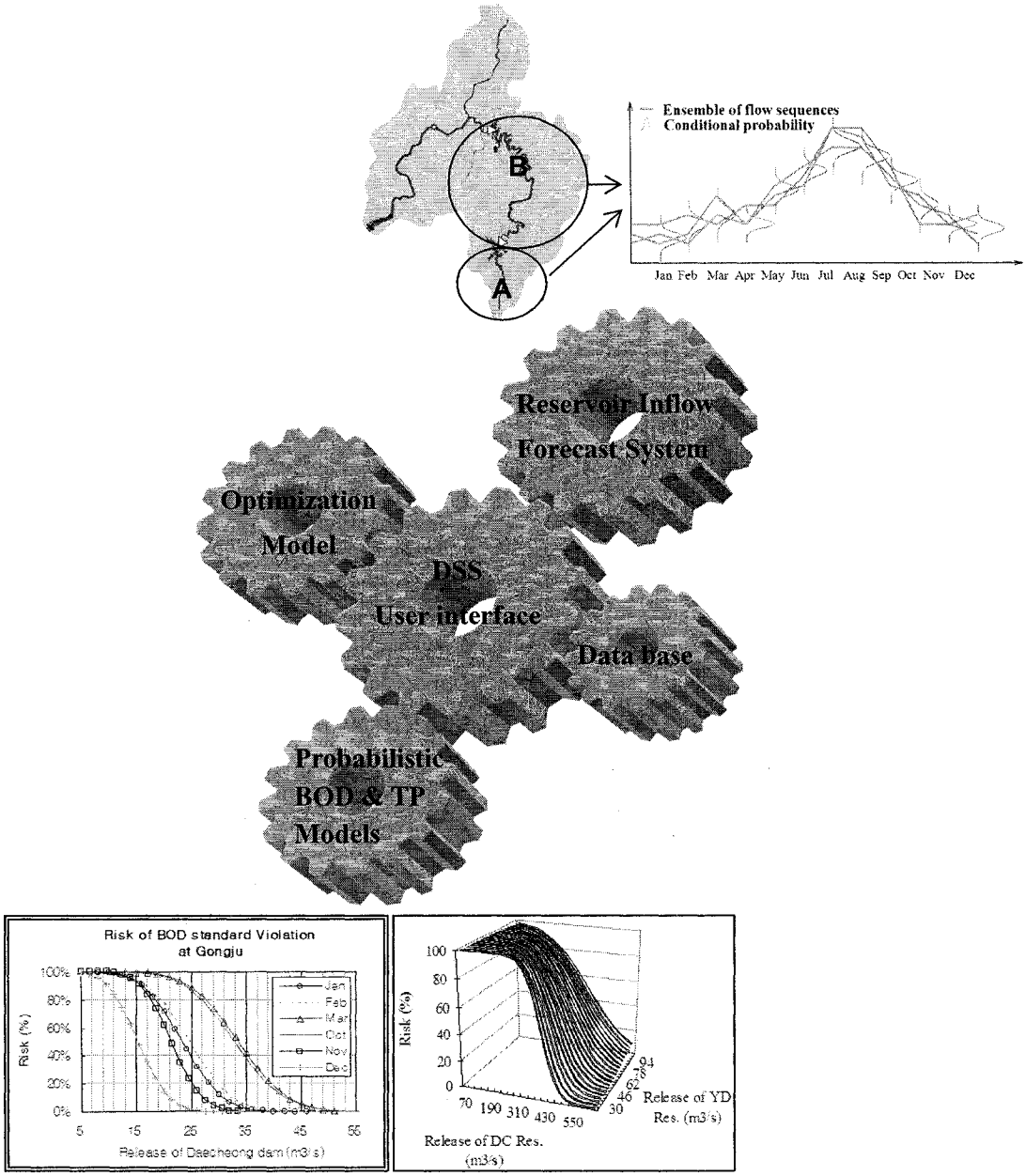


Figure 7-5 Configuration of DSS system that consists of the three sub-systems of model base, data base, and user interface.

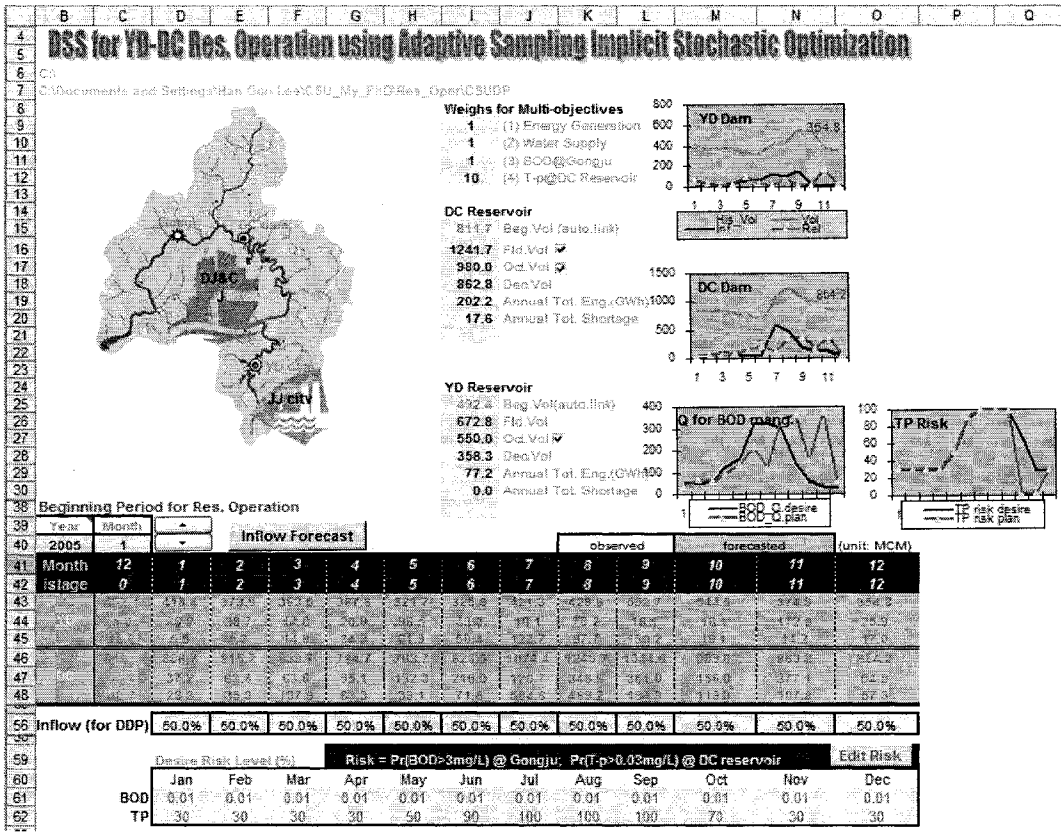


Figure 7-6 Screen capture of the DSS developed using *MS-Excel* and *VBA*.

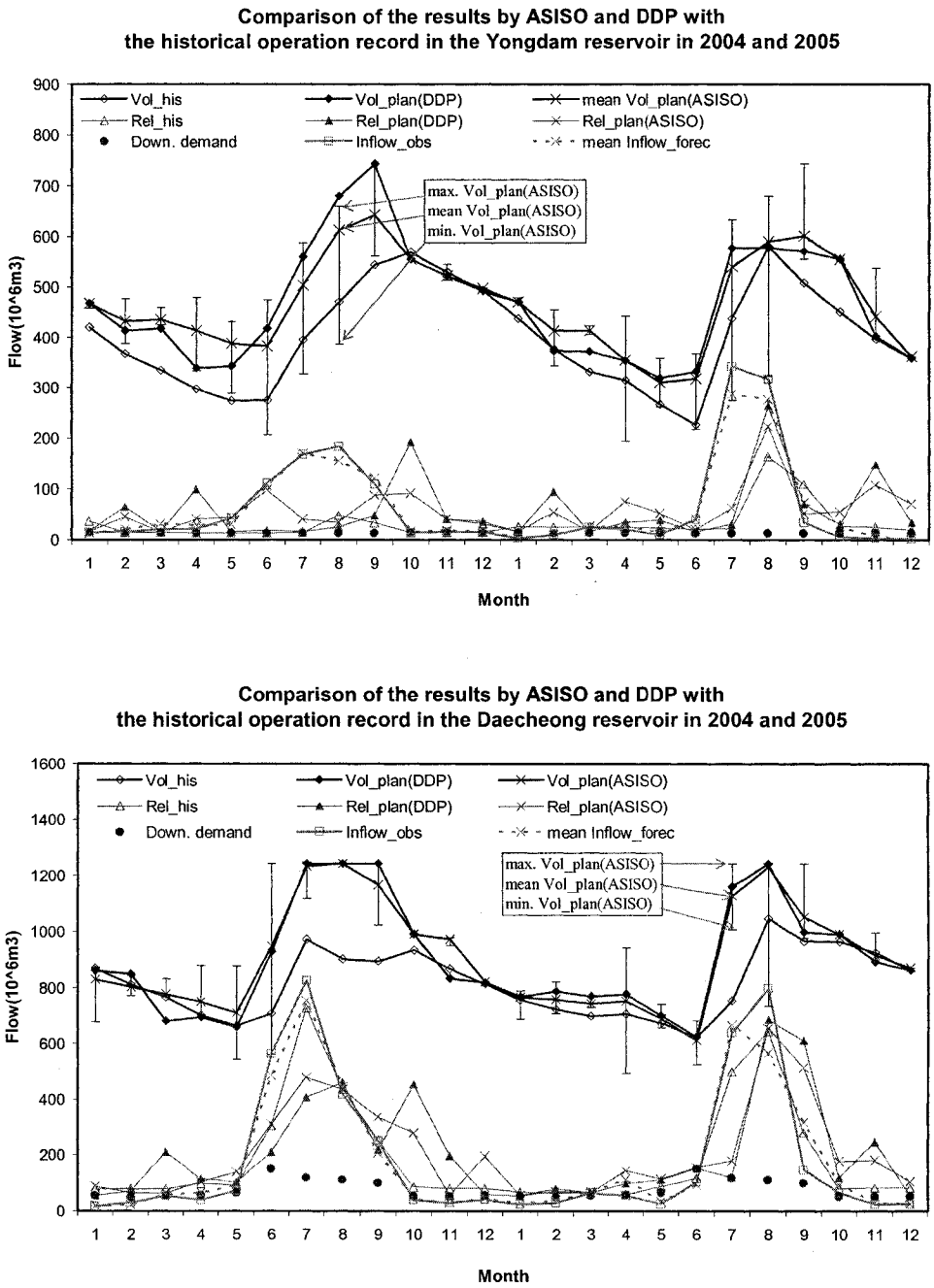


Figure 7-7 (a) Comparison of the results by ASISO and DDP with the historical record in the Yongdam and Daecheong reservoirs in 2004 and 2005.

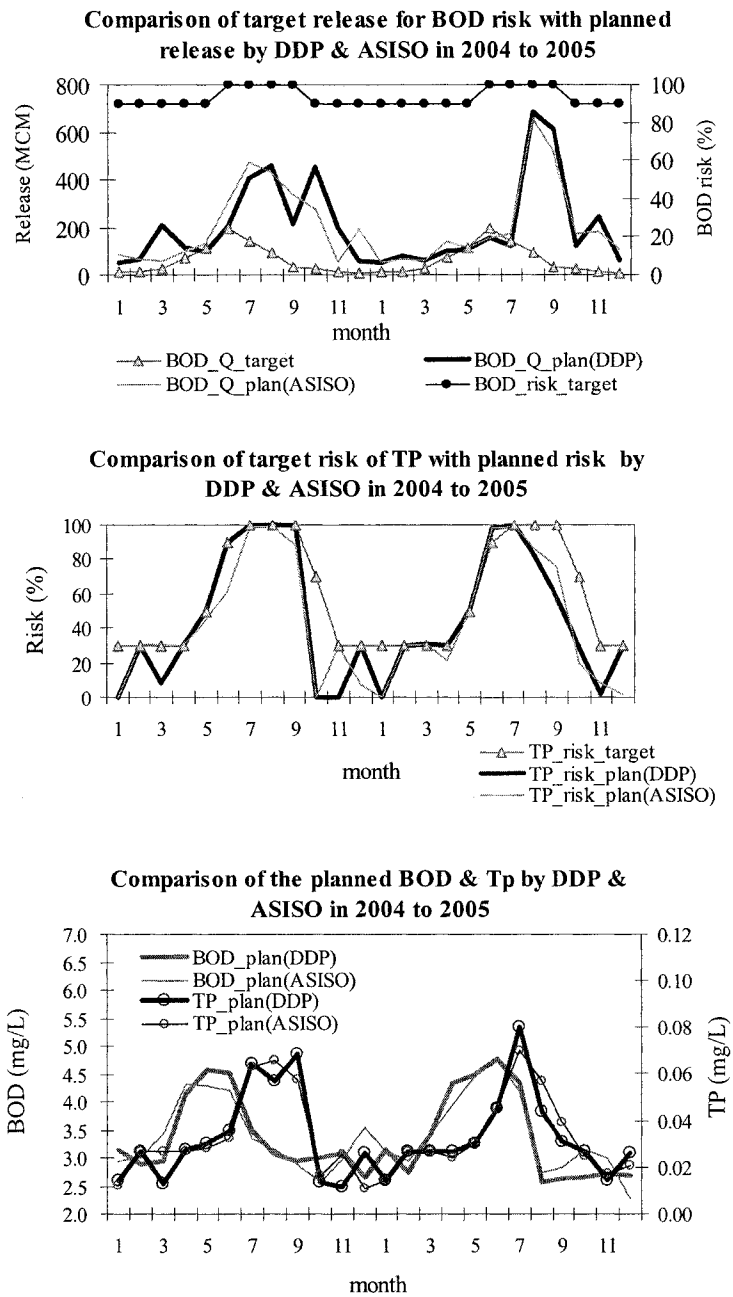


Figure 7-7 (b) Comparison of the results by ASISO and DDP in terms of BOD and TP in 2004 and 2005, where 'Q' is the release of the DC reservoir.

CHAPTER 8

SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

8.1 Summary

This study was motivated by risk-based decision making and adaptive management of large-scale water resources system centering on multi-reservoir operation. Water resources system are characterized by high dimensional, dynamic, nonlinear, multi-objective, and stochastic features that can perplex system managers. Among these, the handling of the dynamic and stochastic nature of the problem might be the most daunting task. The introduction of risk-based decision making and adaptive management to this study arose from the realization that forecasting the dynamic behavior of water resources system is inherently uncertain.

No matter what the reasons are, the uncertainties of scientific prediction for the future are inevitable in the real world and decisions are made on the basis of forecasting the future. In the end, risks are always incurred, to various degrees, in the process of decision making that aims at meeting the target levels of performance measures or scientifically established standards for environmental management.

Adaptive management refers to a successive "learning by doing" strategy that means forecasting the future is based on *a priori* information learned from the facts in the past using sequential data-assimilation, which is referred to as a Bayesian approach. Water resources systems continuously change dynamically on a basin-wide scale and new information is gathered in time series. Furthermore, unexpected political events arise frequently and the information for decision-making is generally random. Consequently, decision makers want to be kept up-to-date promptly on new situations, and newly gathered information should be continuously assimilated in a decision-making process.

An approach for realizing the two motivations is to manage the water resources system with reflection of uncertainties of the future, and update the uncertainties by sequential running of probabilistic or stochastic models with newly acquired information and the historical data base. As the leading surrogates for probabilistic modeling, nonparametric modeling, and Bayesian networks and stochastic artificial neural networks based on Bayesian MCMC learning algorithm have had great popularity among hydrologists and ecologists. This research studied the potential applicability to these emerging technologies to improving the ability to consider randomness. In order to demonstrate integrated river basin management based on consideration of uncertainties, a case study was performed in the Geum River basin in Korea.

There were two main objectives of the research:

i) to develop the probabilistic models for forecasting hydrologic and environmental characteristics using the data-driven approaches of Bayesian networks, nonparametric models, stochastic artificial neural networks, and fuzzy rule-based modeling. The variables considered in this research are the

stage-discharge relationship and reservoir inflow and water qualities of BOD and TP that are essential factors for reservoir system operation (Figure 1-1).

ii) to reflect risks and uncertainties of the variables into an adaptive decision-making process for reservoir system operation by joining the forecast models developed in step '*i)*' interactively to an adaptive sampling implicit stochastic optimization (ASISO) model (Figure 7-5).

First, ratings were developed and assessed using the deterministic and probabilistic methods in the two stage-discharge measurement stations of: the Donghyang station where the hysteresis of rating is not considerable and; the Hotan station located in the main stream of the Geum River where a loop shaped rating clearly exists. For the deterministic approach, non-linear programming, fuzzy rule-based modeling, and one-dimensional hydrodynamic models were used. For the probabilistic approach, a Bayesian MCMC technique was employed with the help of the *WinBUGS* software package. Finally, the methodologies were evaluated and compared in terms of possibility and superiority for adaptive management, uncertainty analysis and reflection of hydraulic characteristics including hysteresis, ease of application, and so on.

Second, stochastic monthly inflow forecast systems for the Chungju reservoir were developed using stochastic artificial neural networks and nonparametric modeling. To make sure whether or not the *k*-NN method can be employed in practice for daily inflow forecasts aiming at short term reservoir operation, this method was applied to a daily forecast model.

Third, a probabilistic BOD model was developed using Bayesian networks at the Gongju site for analyzing the impact on downstream water quality in BOD by the Daecheong reservoir release policy. The probabilistic TP model was also developed in

the Daecheong reservoir, where reservoir eutrophication is affected by the operation policies of the Daecheong and Yongdam reservoirs. The relationships between reservoir release and risk of violating the water quality standards were derived.

Fourth, instead of relying on the classical rule curves derived from the implicit and explicit stochastic optimization methods for reservoir operation, an adaptive sampling implicit stochastic optimization model was proposed with evaluation of the performance measures of energy production, water supply, and water quality management in terms of BOD and TP.

Lastly, a decision support system especially designed for embedding the adaptive sampling implicit stochastic optimization methodology was developed (see Figure 7-5 and 7-6). The DSS consists of the three sub-bases of 1) model base including the progressive optimization model, the BOD and TP models, and the inflow forecast model, 2) data base, and 3) user interface.

8.2 Conclusions and Recommendations for Further Study

The development and uncertainty analysis of ratings are very daunting tasks because of the considerably high measurement errors and complicated hydraulic features of ratings. From comparing the features of four methods in Chapter 4, it was hard to determine the best methodology. Instead, a hybrid methodology was suggested as a good alternative, which combines deterministic NLP and Bayesian MCMC. Bayesian MCMC is very useful in screening for abnormally measured stage-discharge data, and obtaining the information of the highly varying parameters of the ratings. NLP can be used to refine ratings with the constraints based on the confidence limits of the parameters derived by

the Bayesian MCMC analysis. It was demonstrated that both NLP and Bayesian MCMC can analyze the hysteresis of the rating using the Jones formula for the case study. For *K*-water's sake, the existed DSS for rating analysis named "*HydroToolkit*" should be upgraded to incorporate the hysteresis analysis functions and Bayesian MCMC.

As for practical application of reservoir inflow and stream flow forecast models, the *k*-NN method may be preferred to stochastic ANN and KDE modeling techniques due to its ease of application. Although the *k*-NN showed the worst results in terms of goodness-of-fit in the case study, the fit was still quite good and the applicability of the *k*-NN method was demonstrated in a real application for reservoir system operation in the Geum River basin. In addition, the results of the daily inflow forecast model also showed that the *k*-NN method could be applied for short term operation of a reservoir system.

Although there is ongoing debate as to the relative merits of using simple models, the probabilistic BOD and TP models demonstrated in the case study that they overcome the weaknesses of deterministic water quality models by offering the information about risks of standard violation or failures of meeting targets. Compared to the other methods for uncertainty analysis such as sensitivity analysis, the derived distribution method, FOSM, and the Monte Carlo method, the superiority of the Bayesian MCMC technique was also successfully demonstrated in the application of the adaptive sampling implicit stochastic optimization model. The results of the case study of the TP model showed that more release from the *YD* reservoir made TP at the *DC* reservoir worse because the phosphorus load from *YD* reservoir is greater than the increase of DC reservoir storage.

As for the methodologies for reservoir optimization, the importance of combining simulation and optimization algorithms, considering multi-objectives and operating the reservoir system adaptively were demonstrated by the case study. The case study also

showed that the reservoir inflow forecast systems played the most important role out of all functions included into the DSS in that the ensemble created by the forecast models made the greatest difference between the ASISO and ISO methods. From the comparison between the ASISO and adaptive DDP, the results indicate although the ASISO is much better theoretically, the adaptive DDP can be applied successfully in practice.

The typical data-driven models of nonparametric methods and stochastic ANN were clearly recognized as an excellent alternative for probabilistic modeling along with the hybrid model of Bayesian networks. The applications to stochastic inflow forecast and probabilistic water quality modeling showed their potential to be extended to various water resources projects. A pure data-driven modeling technique of stochastic ANN can be recognized as the best one in terms of goodness-of-fit as demonstrated in the case study for stochastic inflow forecast. Although fuzzy rule-based modeling may be included in the class of the data-driven probabilistic modeling techniques, it has limitations for practical application for probabilistic modeling.

However, the group of data-driven modeling has obvious limitations in representing physical characteristics of a system of interest because they are absolutely dependent on pattern recognition from data instead of relying on the system functions that describe a variety of physical-biological-chemical reactions. If data are not updated adaptively, this situation makes their limitations more serious. From this viewpoint, the hybrid Bayesian networks may be recognized as the most reasonable method in that it can both reflect the physical features of a system and analyze the system probabilistically in an adaptive way. Furthermore, this method can work even in the case where sufficient data are not available. However, the relatively complicated modeling processes may be an obstacle in

practical applications, and this fact was confirmed in the case study of the inflow forecast modeling by comparing with the *k*-NN nonparametric model.

In summary, for the projects that require interaction between an analysis model and a DSS, the pure data-driven *k*-NN method is recommended when sufficient data are available. For decision making based on 'what if' type of analysis, a Bayesian modeling is strongly suggested especially when data set is not large.

A monthly time step is the commonly accepted standard for the purpose of both designing a new storage project and operating the project. However, monthly average operational policies tend to overestimate the ability of the system to capture flood peaks and supply water. Furthermore, system constraints and operations criteria may be based on finer time scales: hence, it is recommended for further study that the developed reservoir operation system should be reduced to a weekly or daily real-time operation. For the expected weekly or daily systems, they should be based on adaptive management like the monthly ASISO model. Object oriented network-flow models linked to monthly operation models are strongly recommended in that it is relatively easy to describe a complicated reservoir system in detail. However, it may be problematic to incorporate stochasticity of the hydrologic variables into the system.

Forecasting reservoir inflow and stream flow series is surely a strenuous task. These forecasts however, may be the most important element out of all the procedures for reservoir system operation. For recommendations for further study about short term forecast models, Markovian autoregressive models could be made more sophisticated by incorporating a variety of exogenous variables such as temperature and humidity. As another alternative, a disaggregation forecast model may be considered, which is connected to monthly or yearly models.

REFERENCES

Abebe, A. J., Solomatine, D. P., Venneker, R. G. W. (2000). "Application of adaptive fuzzy rule-based models for reconstruction of missing precipitation events." *Hydrological Sciences Journal,* 45(2), 425-436.

Adamowski, K., Feluch, W. (1991). "Application of nonparametric regression to groundwater level prediction." *Can. J. Civ. Eng.*, 18, 600–606.

Amari, S.-i., Murata, N., Müller, K.-R., Finke, M., Yang, H. H. (1997). "Asymptotic statistical theory of overtraining and cross-validation." *IEEE Transactions on Neural Networks* 8(5), 985-996.

Bardossy, A., Duckstein, L. (1995). *Fuzzy Rule-Based Modeling with Application to Geophysical, Biological and Engineering Systems*, CRC Press, Inc.

Bhattacharya, B., Solomatine, D. P. (2000). "Application of artificial neural network in stage-discharge relationship." *Proceedings, 4th International Conference on Hydroinformatics, Iowa City, USA.*

Bijaya P. S., Duckstein, L., Eugene Z., S. (1996). "Fuzzy Rule-Based Modeling of Reservoir Operation." *J. of Water Resources Planning and Management,* 122(4), 262-269.

Borsuk, M. E. (2004). "A Bayesian network of eutrophication models for synthesis, prediction, and uncertainty analysis." *Ecological Modeling*, 173, 219-239.

Borsuk, M. E., Craig A. Stow (2000). "Bayesian Parameter Estimation in a mixed-order model of BOD decay." *Water Res.*, 34(6), 1830-1836.

Bras, R. L., I. Rodriguez-Iturbe (1985). *Random Functions and Hydrology*, Addison-Wesley, Reading, Mass..

Chapra, Steven C. (1997). *Surface Water-Quality Modeling*, McGraw-Hill.

Chow, Ven Te, David R. Maidment, Larry W. Mays (1988). *Applied Hydrology*, McGraw-Hill.

Delft Hydraulics (1994). *HYMOS User's Guide, Delft, The Netherlands.*

Dilks, D.W., Canale, R.P., Meier, P.G. (1992). "Development of Bayesian Monte-Carlo techniques for water-quality model uncertainty." *Ecol. Modelling*, 62, 149-162.

Dowd, Michael, Renate Meyer (2003). "A Bayesian approach to the ecosystem inverse problem." *Ecological Modeling*, 168, 39-55.

Duan, Q., Sorooshian, S., Gupta, V. (1992). "Effective and Efficient Global Optimization for Conceptual Rainfall-Runoff Models." *Water Res. Res.*, 28(4).

Gelman, A., Carlin, J., Stern, H.S., Rubin, D. B. (1995). *Bayesian Data Analysis, 2nd Edition*, CRC Press.

Gingras, D., Adamowski, K. (1994) "Performance of L-moments and nonparametric flood frequency analysis." *Can. J. Civ. Eng.*, 18, 600–606.

Grigg, Neil S. (1996). *Water Resources Management: Principles, Regulations, and Cases*, McGraw-Hill, ISBN 0-07-024782-X.

Hassoun, M. H. (1995). *Fundamentals of artificial neural networks*, MIT Press, Cambridge.

Haykin, S. (1993). *Neural network, a comprehensive foundation*, Macmillan College Publishing Company, USA.

Hsu, Kuo-lin, Gupta, H. V., Sorooshian, S. (1995). "Artificial neural network modeling of the rainfall-runoff process." *Water Resources Research,* 31(10).

ISO. (1998). *ISO 1100-2 Measurement of liquid flow in open channels – part 2: Determination of the stage-discharge relation.*

Kanso, A, M. C. Gromaire, E. Gaume, B. Tassin, G. Chebbo (2003). "Bayesian approach for the calibration of models: Application to an urban stormwater pollution model." *Wat. Sci. & Tech.*, 47(4), 77-84.

Kelman, J. J., R. Stedinger, L. A. Cooper, E. Hsu, S-Q Yaun (1990). "Sampling stochastic dynamic programming applied to reservoir operation," *Water Resources Research*, 26(3), 447-454.

Ko, I. H. (1997). *Integrated river basin operational planning considering water quaint and water quality*, Ph. D. dissertation, Colorado State University, Fort Collins, Colorado.

Ko, S.-K., Fontane, D., Labadie, J. (1992). "Multobjective optimization of reservoir systems operation." *Water Resour. Bull.*, 2(1), 111-127.

K-water (2004). *Stream Discharge Computation Using Hydraulic Channel Routing for the Yongdam Experimental Watershed*, Research report.

K-water (2003). *User's manual for the regular assessment system of hydrological data and Rating development system*, Technical document.

Labadie, John W. (2004). "Optimal Operation of Multireservoir Systems: State-of-the-Art Review." *J. of Wat. Res. Planning and Management Div., ASCE*, 130(2), 93-111.

Labadie, J. W. (1999). *Generalized dynamic programming package: CSUDP, Documentation and user manual*, Dept. of Civil Engineering, Colorado State Univ., Ft. Collins, CO.

Labadie, J. W. (1997). "Reservoir system optimization models." *Water Resources Update*, 108, 83-112.

Labadie, J. W., Fontane, D.G., Dai, T. (1994). "Integration of Water Quantity and Quality in River Basin Network Flow Modeling." *Water Policy and Management, Proceedings of the 21s Annual Conference, ASCE*, edited by Fontane, D.G., and Tuvel, H.N., Denver, CO. 61-64.

Labadie, J. W. (1993). "Combining simulation and optimization in river basin management." *Stochastic hydrology and its uses in water resources systems simulation and optimization*, J. Marco et al., eds., Kluwer Academic, Dordrecht, The Netherlands, 345-371.

Lall, U., Sharma, Ashish (1996), "A nearest neighbor bootstrap for resampling hydrologic time series." *Water Res. Res.*, 32(3), 679-693.

Lall, U. (1995). "Recent advances in nonparametric function estimation: Hydraulic applications." *U.S. Natl. Rep. Int. Union Geod. Geophys,* 1991–1994, Rev. Geophys., 33, 1093, 1995.

Lall, U., Bosworth, K. (1993). *Multivariate kernel estimation of functions of space and time hydrologic data, in Stochastic and Statistical Methods in Hydrology and Environmental Engineering*, edited by K. Hipel, Kluwer, Waterloo.

Mackay D. J. C. (1992). "A Practical Bayesian Framework for Backpropagation Networks." *Neural Computation*, 4, 448-472.

Maier, Holger R., Dandy, Graeme C. (2000). "Neural networks for the prediction and forecasting of water resources variables: a review of modeling issues and application." *Environmental Modeling & Software*, 15(1), 101-104.

Marino, M. A., H. A. Loaiciga (1985). "Dynamic model for multireservoir operation." *Wat. Res. Res.*, 21(5), 619-630.

Mathworks. *Fuzzy Logic Toolbox User's Guide, Ver. 2.0.*

Mathworks. *Neural Network Toolbox User's Guide, Ver. 3.0.*

National Research Council (2001). *Assessing the TMDL approach to water quality management*, National Academy Press, Washing D.C.

Neal, Radford M. (1992). *Bayesian Training of Backpropagation Networks by the Hybrid Monte Carlo Method*, Technical Report CRG-TR-92-1, Dept., of Computer Science, Univ. of Toronto.

Ogink, H. J. M. (1994). *Hydrogy 2*, IHE, The Netherlands, Lecture note.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*, Cambridge University Press, Cambridge, UK.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, San Francisco, CA:Morgan Kaufmann.

Qian Song S., Craig A. Stow, Mark E. Borsuk (2003). "On Monte Carlo Methods for Bayesian Inference." *Ecological Modeling,* 159, 269-277.

Reckhow, K. H. (1999). "Water quality prediction and probability network models." *Can. J. Fish. Aquat. Sci.,* 56, 1150-1158.

Salas, J. D., M. Markus, A. S. Tokar (2000). *Artificial Neural Networks in Hydrology*, edited by R. S. Govindaraju and A. Ramachandra Rao, Kluwer Academic Publishers, the Netherlands, 23-51.

Salas, J. D. (1993) *Analysis and modeling of hydrologic time series, in Handbook of Hydrology*, edited by D. R. Maidment, McGraw-Hill, New York.

Sharma, A., Tarboton, D. G., Lall, U. (1997). "Stream flow simulation: A nonparametric approach." *Water Res. Res.*, 33(2), 291-308.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, New York.

Solomatine, D. P., 2002. *Application of data-driven and machine learning in control of water reasources,* M. Mohammadian, R. A. Sarker and X. Yao (eds). Idea Group Publishing, 197-217.

Spiegelhalter, D. J., Thomas, A., Best, N. G., Gilks, W. R., (1995). *BUGS: Bayesian Inference Using Gibbs Sampling, Version 0.50*, MRC Biostatistics Unit, Cambridge.

Stow, C. A., Chris R., Borsuk, M. E., James D. B., Kenneth H. R. (2003). "Comparison of estuarine water quality models for total maximum daily load development in Neuse river estuary." *J. of Water Res. Planning and Management, ASCE.*

Tarboton, D. G., Sharma, A., Lall, U. (1998), "Disaggregation procedure for stochastic hydrology based on nonparametric density estimation." *Water Res. Res.*, 34(1), 107-119.

Thiemann, M., Gupta, T. H., Sorooshian, S. (2001). "Bayesian recursive parameter estimation for hydrologic models." *Wat. Res. Res.*, 37(10), 2521-2535.

Thomann, R. V., Mueller, J. A. (1987). *Principles of surface water quality modeling and control*, HarperCollins*Publishers* Inc.

Velickov, S., Solomatine, D. P. (2000). "Predictive Data Mining: Practical Examples." *Artificial Intelligence in Civil Engineering, Proc. 2nd Joint Workshop*, Cottbus, Germany. ISBN 3-934934-00-5.

Vollenweider, R. A. (1976). "Advances in defining Critical Loading Levels for Phosphorus Loading Concept in Limnology." *Schweiz. Z. Hydrol.* 37, 53-84.

Wurbs, R. A. (1996). *Modeling and Analysis of Reservoir Systems Operations*, Prentice Hall PTR, Upper Saddle River, NJ 07458.

Wurbs, R. A. (1993). "Reservoir-Systems Simulation and Optimization Models." *J. of Water Resources Planning and Management Div., ASCE,* 119(4), 455-472.

Yakowitz, S. (1979) "A nonparametric Markov model for daily river flow." *Water Res. Res.*, 15(5), 1035–1043.

Yakowitz, S. (1973). "A stochastic model for daily river flows in an arid region.", *Water Res. Res.*, 9(5), 1271–1285.

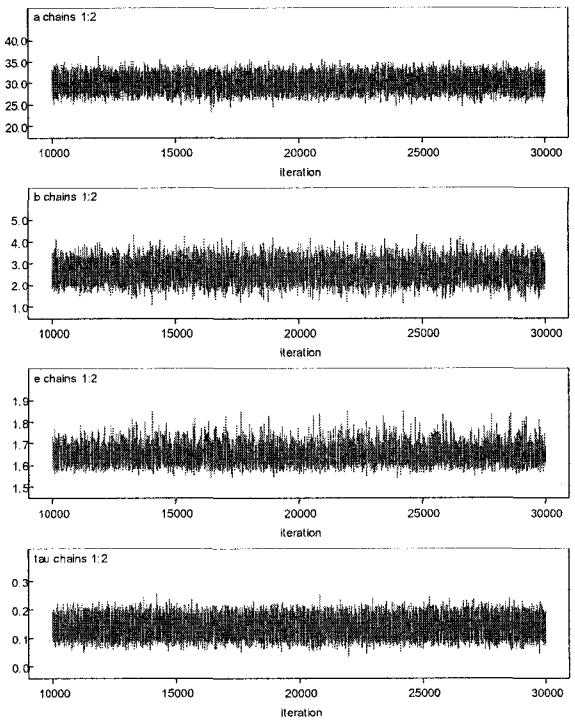
Yeh, W. (1985). "Reservoir management and optimization models: A state-of-the-art review." *Water Resources Research*, 21(12), 1797-1818.

APPENDIX-A. MCMC Diagnostics of Rating Analysis Using BMCMC

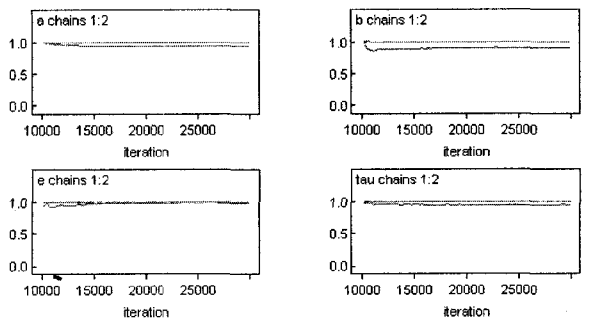
A-1) Donghyang station

(1) Analysis of the rating under section control in 2004

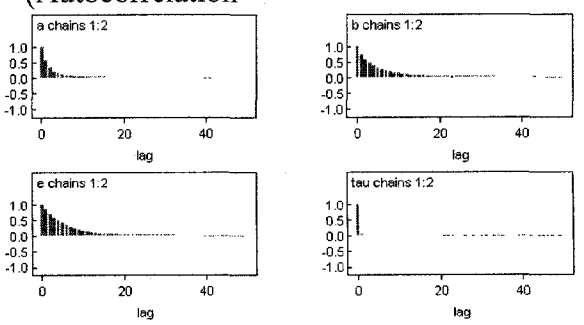
(History path)



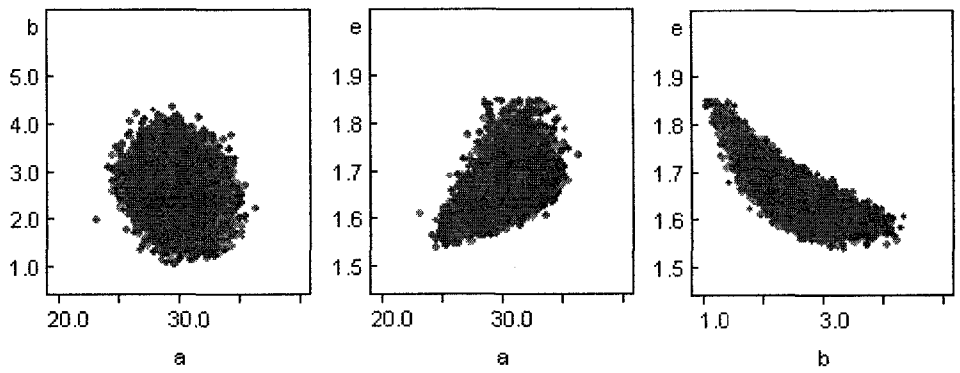
(*Gelman-Rubin R*: Within- and between-sequence)



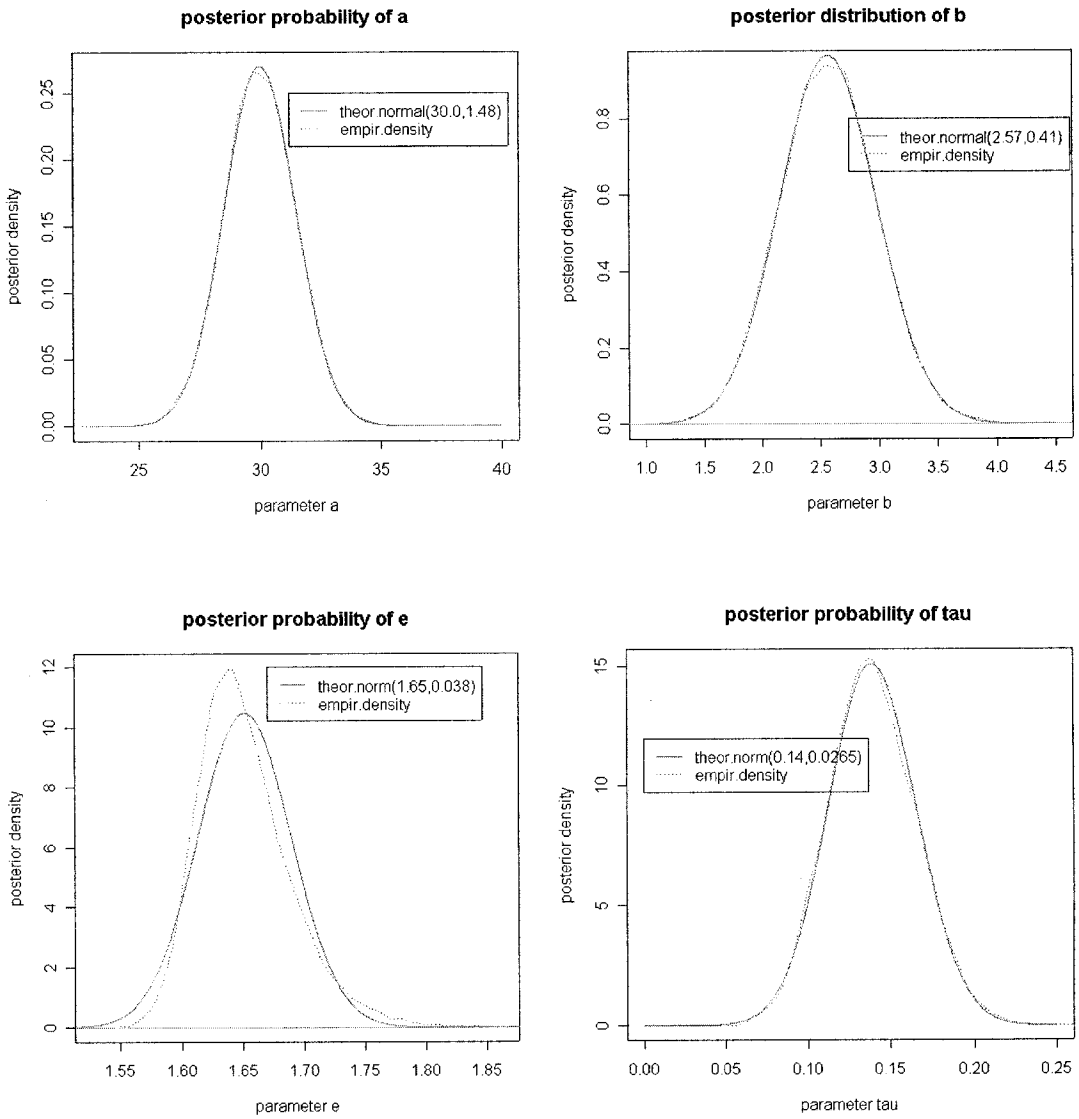
(Autocorrelation



(Correlation between two parameters)

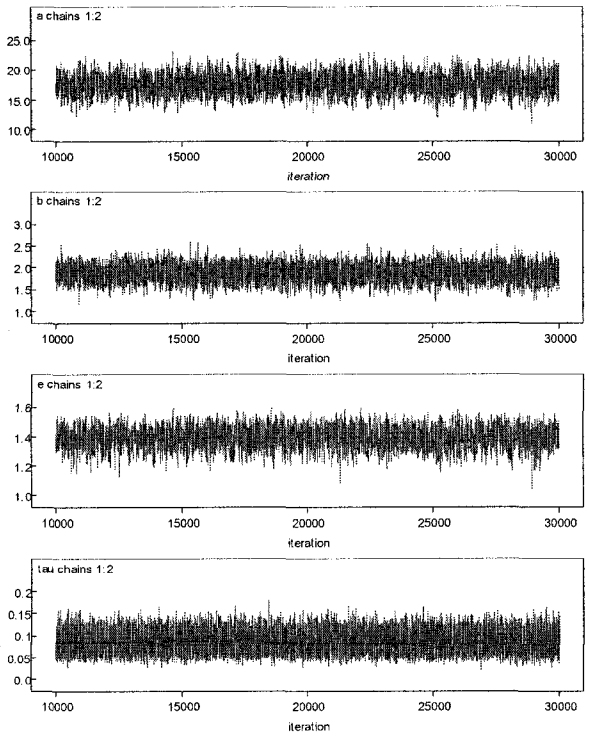


(posterior probabilities of 4 parameters (a, b, e, τ))

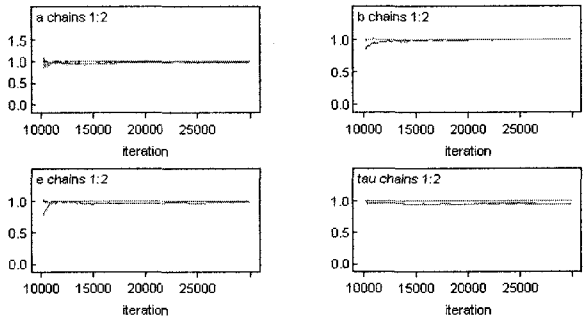


(2) Analysis of the rating under channel control in 2004

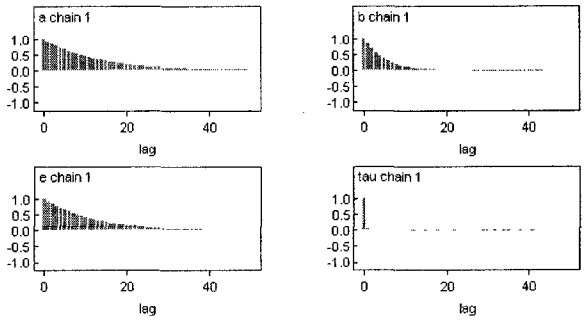
(History path)



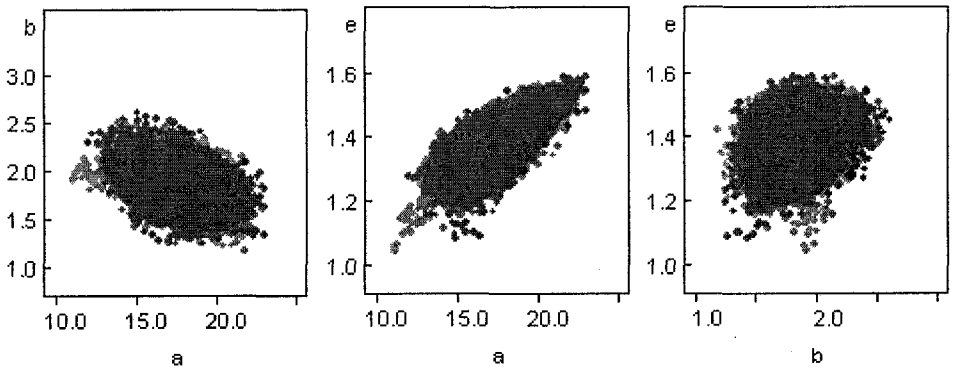
(*Gelman-Rubin R*: Within- and between-sequence)



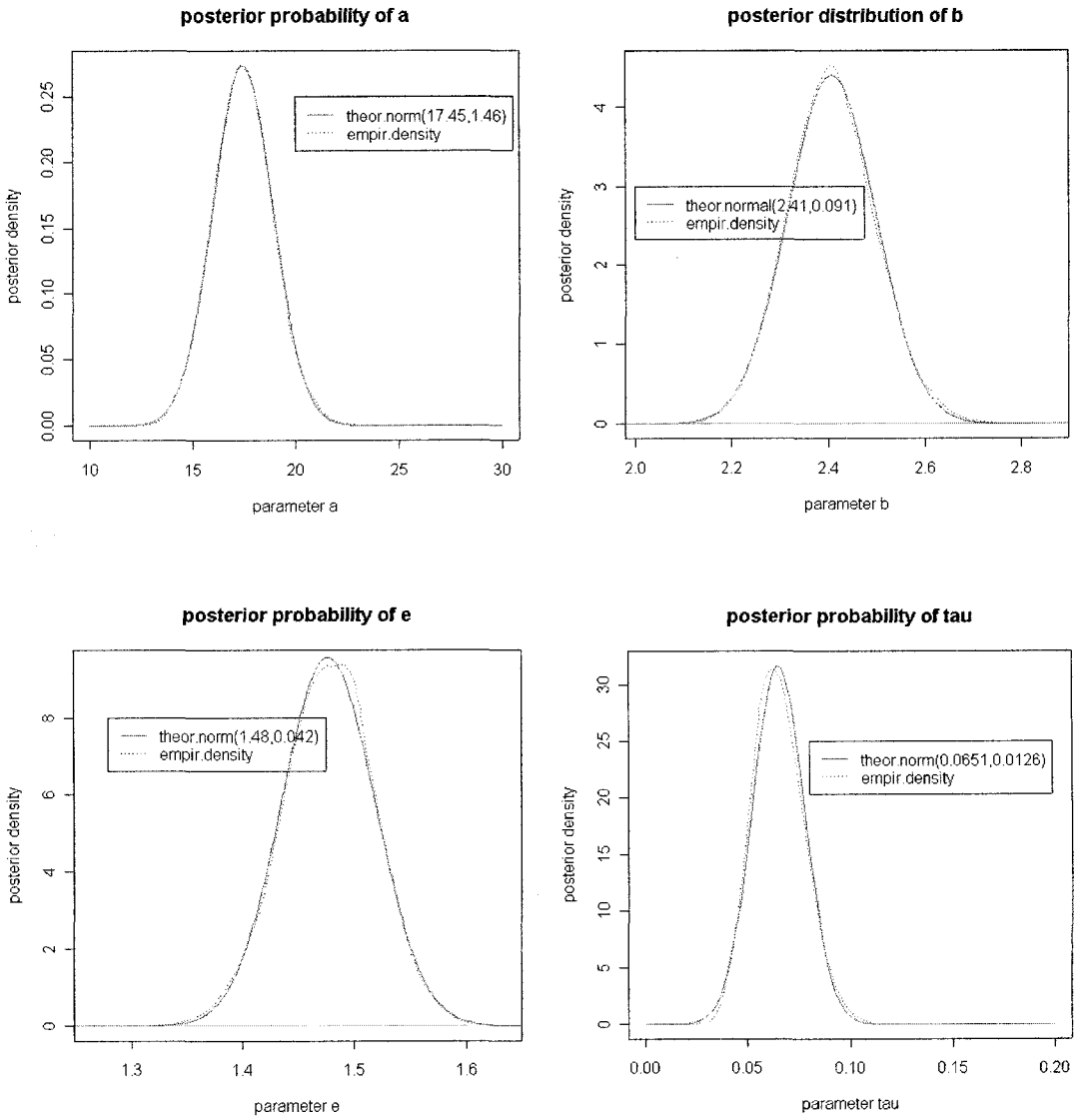
(Autocorrelation



(Correlation between two parameters)



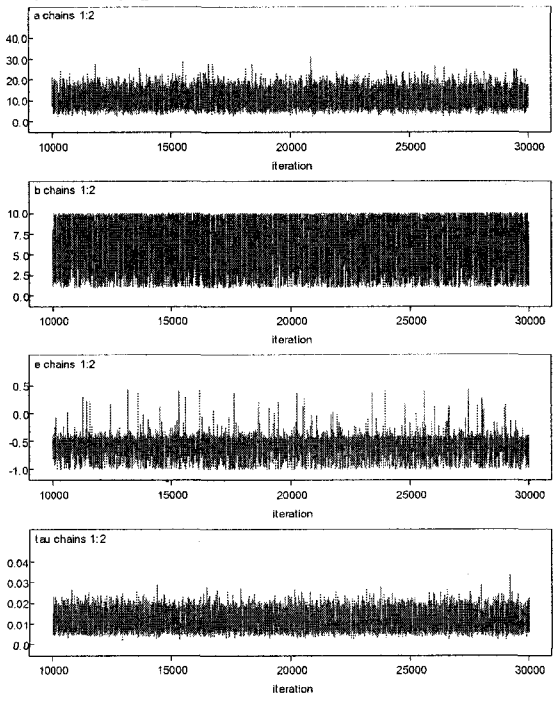
(posterior probabilities of 4 parameters (a, b, e, τ))



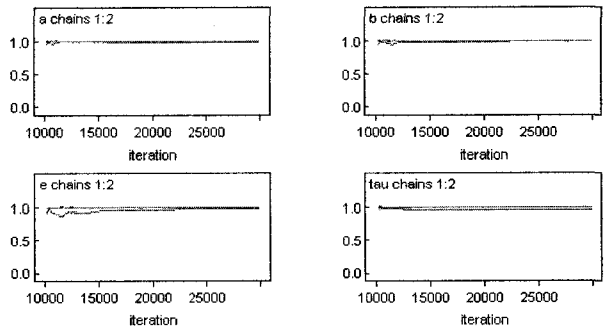
A-2) Hotan station

(1) Analysis of the rating under section control in 2004

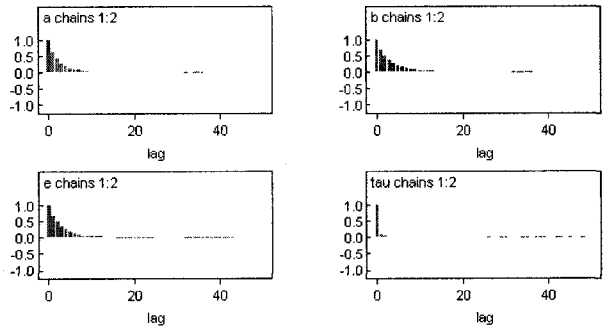
(History path)



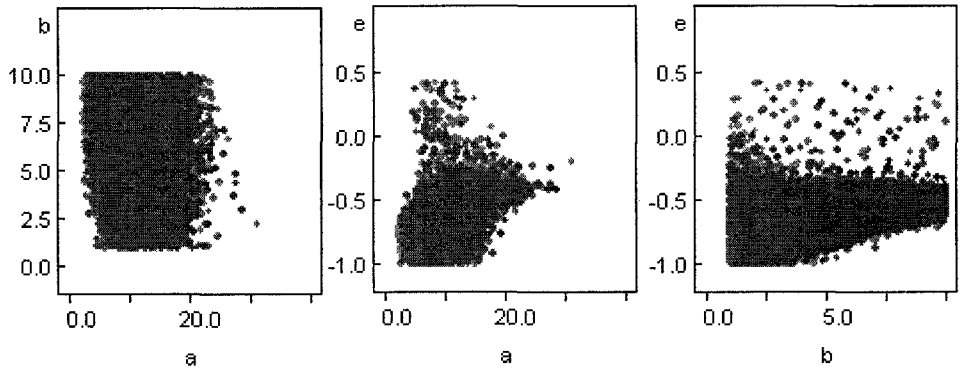
(*Gelman-Rubin R*: Within- and between-sequence)



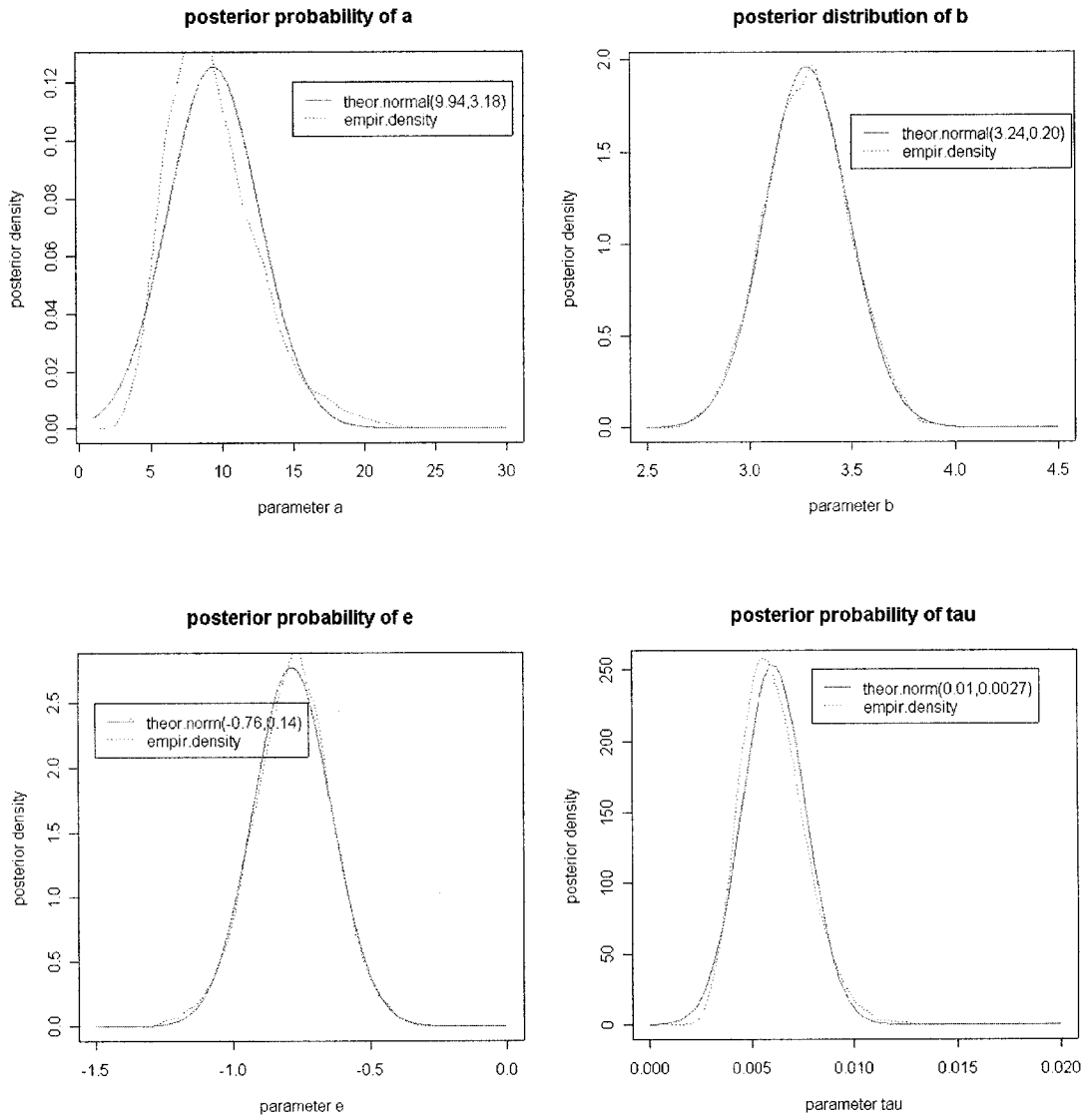
(Autocorrelation



(Correlation between two parameters)

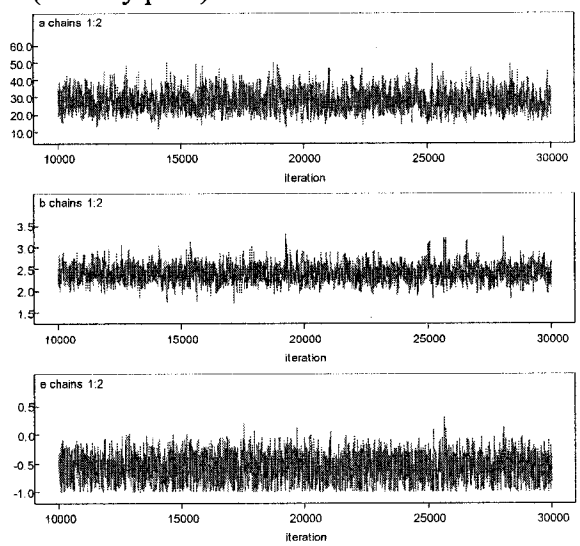


(posterior probabilities of 4 parameters (a, b, e, τ))

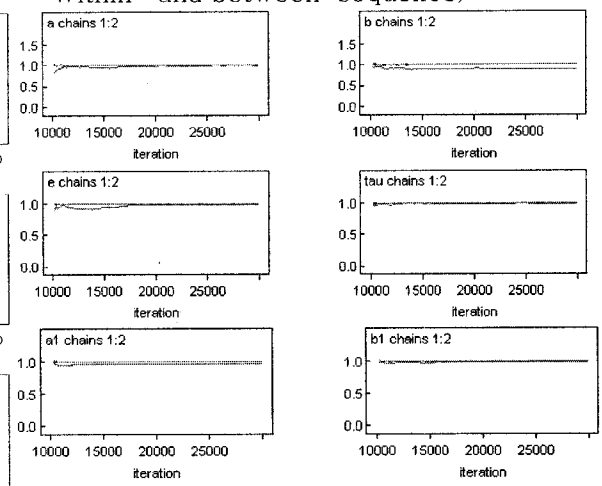


(2) Analysis of the rating under channel control in 2004

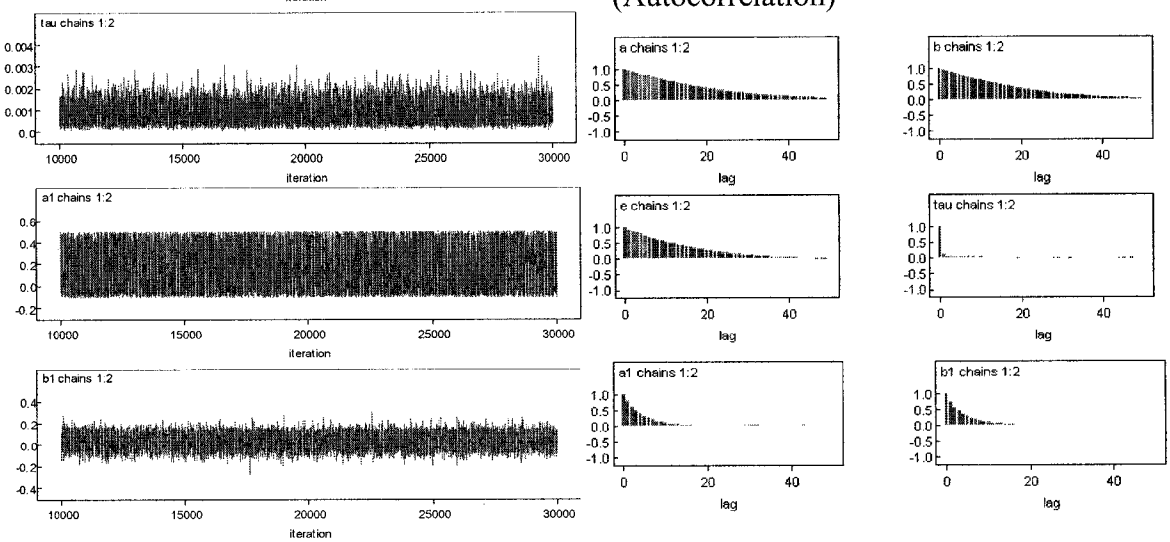
(History path)



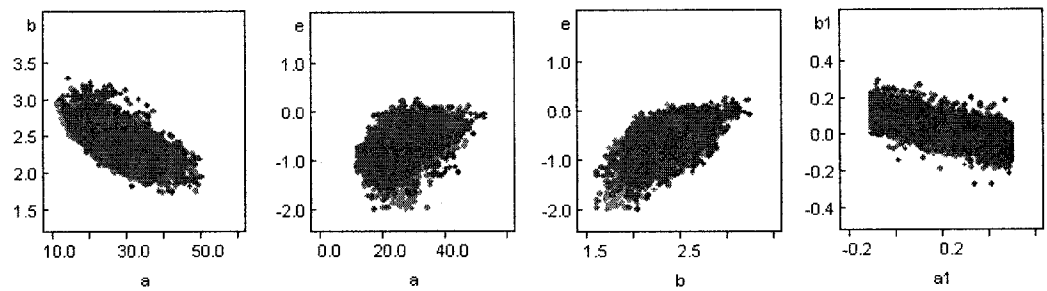
(*Gelman-Rubin R*: Within- and between-sequence)



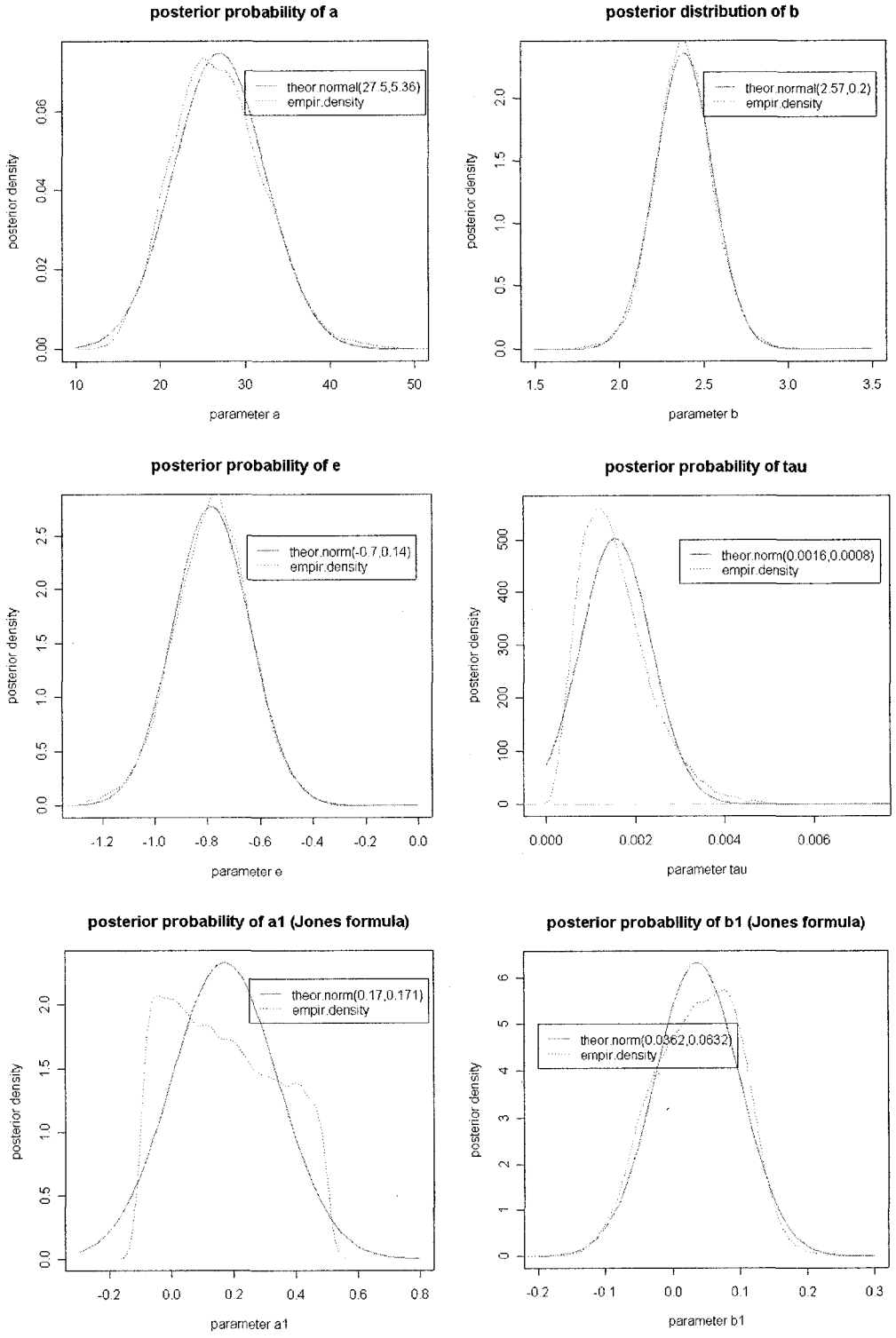
(Autocorrelation)



(Correlation between two parameters)



(posterior probabilities of 4 parameters (a, b, e, τ, $a1$, $b1$)

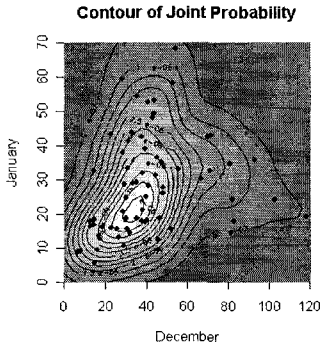


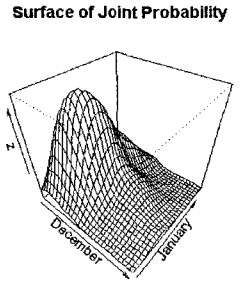
A-3) Source code of *WinBUGS* for the reliability analysis of ratings using MCMC

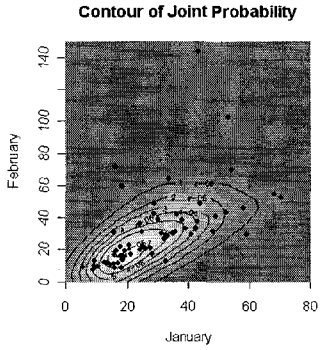
```
# Qu: the observed unsteady flow
# muu: the computed mean of unsteady flow
# Qs: the computed steady flow
# hu: the stage corresponding to the unsteady flow
# dhdt: the variation rate of stage in time
# hysteresis_correction: the degree of unsteadiness by the Jones formula
# U: the number of observations
# dnorm: a Normal distribution
# M: the number of data points for simulation of loop shaped raring curve

# Training the a,b, and e of Q=a(h-e)^b and a1 & b1 of the Jones formula
Model
{
   for (k in 1:U)
   {
         Qu[k]~dnorm(muu[k], tau)
         muu[k] <- hysteresis_correction [k]* Qs[k]
         Qs[k] <- a * pow((hu[k]-e), b)
         hysteresis_correction[k] <- pow(1+1/sv[k]*dhdt[k], 0.5)
         sv[k] <- 1/(a1+b1*hu[k]+c1*hu[k]*hu[k])
   }
   # Simulation of steady and loop shaped unsteady rating
   for (m in 1:M)
   {
         Qs_sim[m] <- a * pow((h_sim[m]-e), b)
         sv_sim[m] <- 1/(a1+b1*h_sim[m]+c1*h_sim[m]*h_sim[m])
         correct_sim[m] <- pow(1+1/sv_sim[m]*dhdt_sim[m], 0.5)
         Qu_sim[m] <- correct_sim[m]* Qs_sim[m]        #unsteady flow
   }
   # Prior Probabilities for the Parameters
         a~dgamma(a_1st, a_2nd)          #Gamma distribution
         b~dunif(b_1st, b_2nd)           #Uniform distribution
         e~dunif(e_1st, e_2nd)
         a1~dunif(a1_1st, a1_2nd)
         b1~dunif(b1_1st, b1_2nd)
         c1 <- 0
         tau~dgamma(tau_1st, tau_2nd)
         sigma <- sqrt(1 / tau)
         variance <- (1 / tau)
}
```

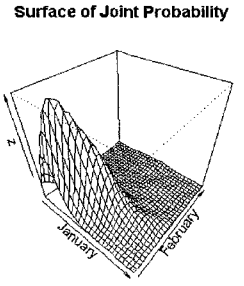
APPENDIX-B. Development of a Reservoir Inflow Forecast System

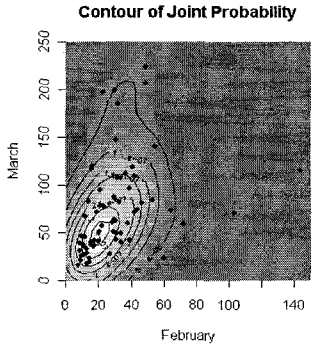
B-1) Joint probability of the monthly inflow in the Chungju reservoir using KDE

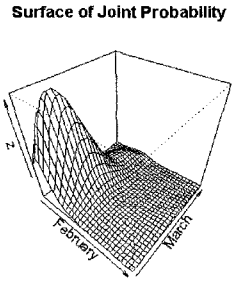


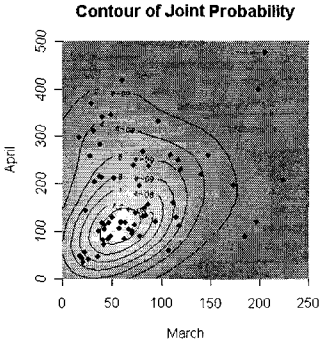


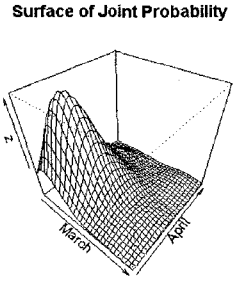


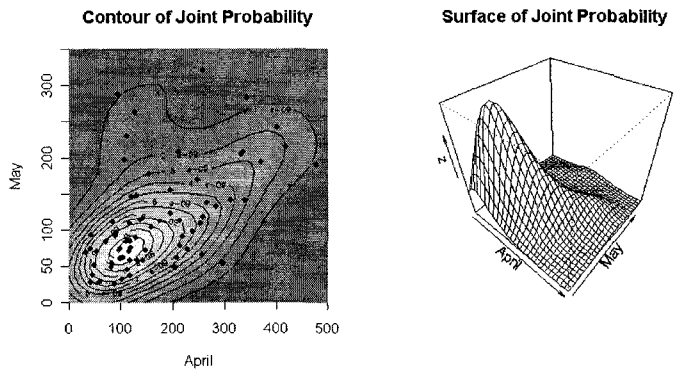


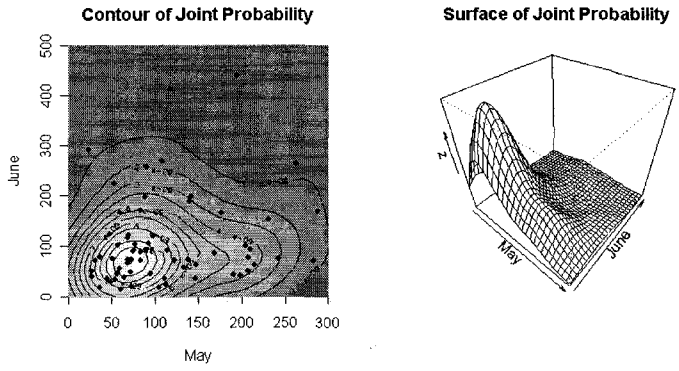


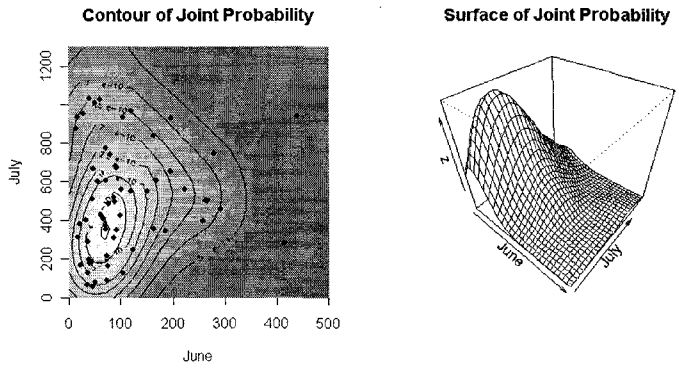


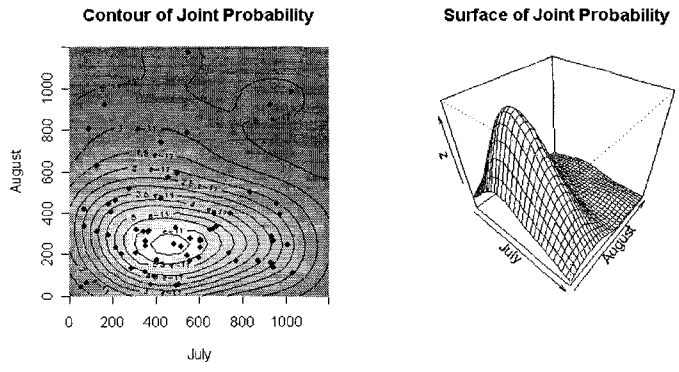


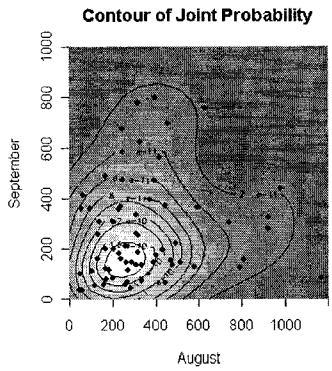


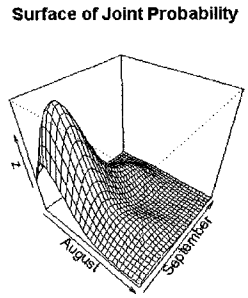


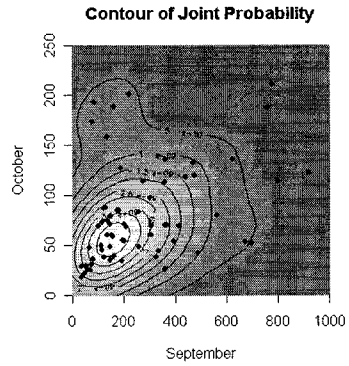


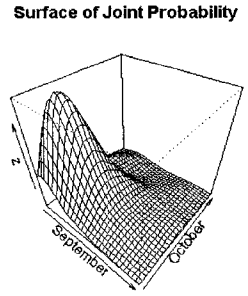


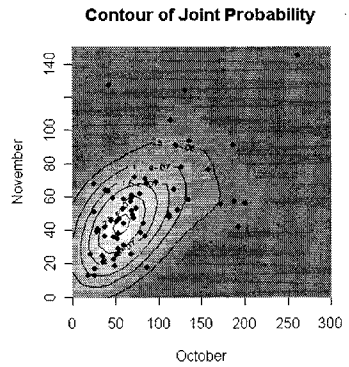


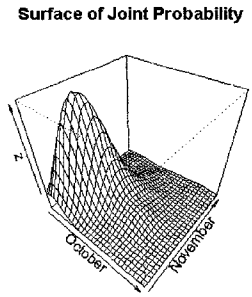


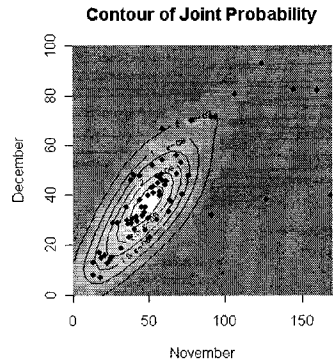


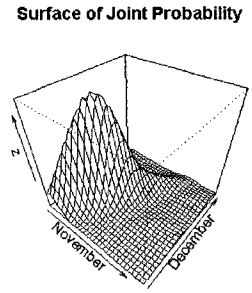






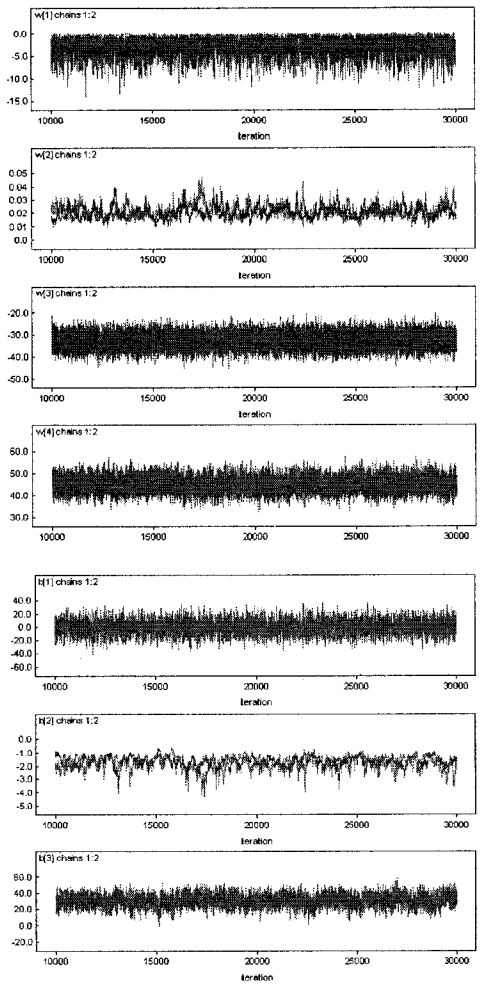




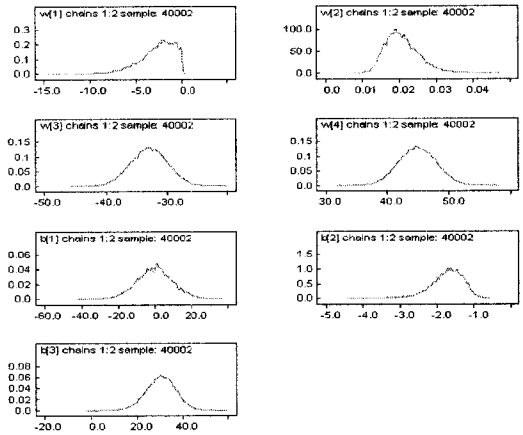


B-2) MCMC diagnostics of the stochastic ANN model in December

(History path)

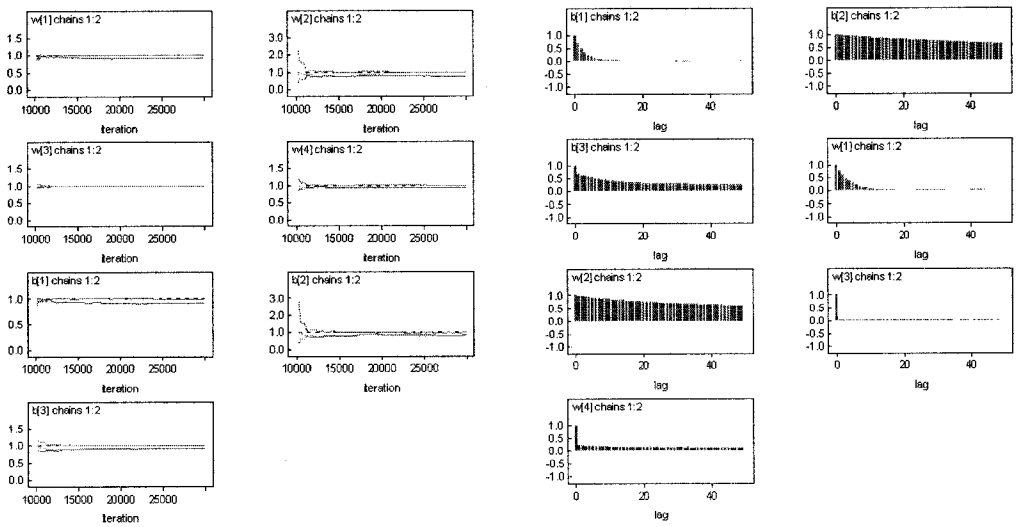


(Posterior probability of parameter)

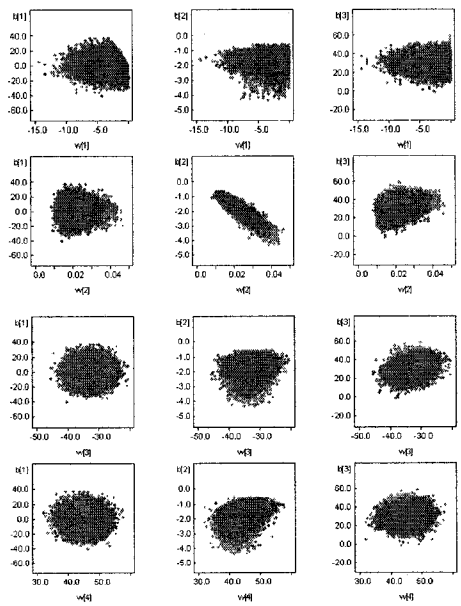


(Gelman-Rubin *R*: within- and between-sequence)

(Autocorrelation)



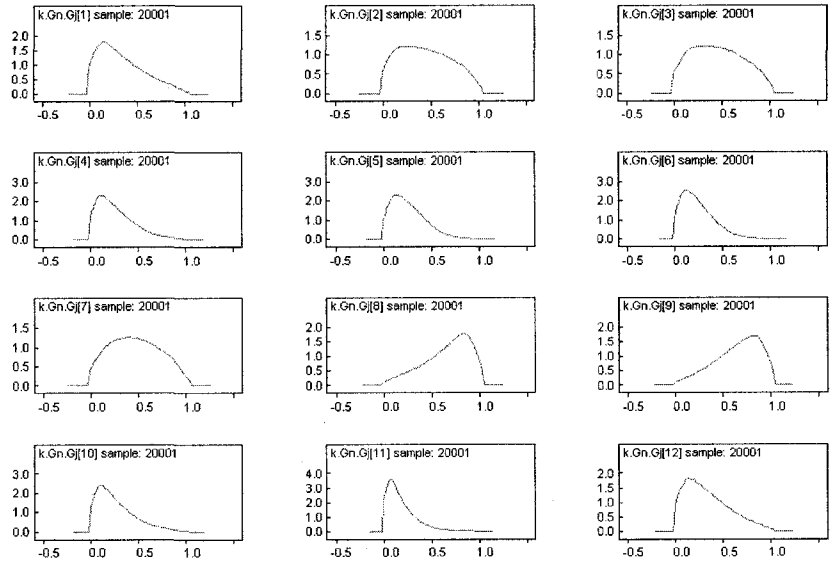
(Correlation between two parameters)



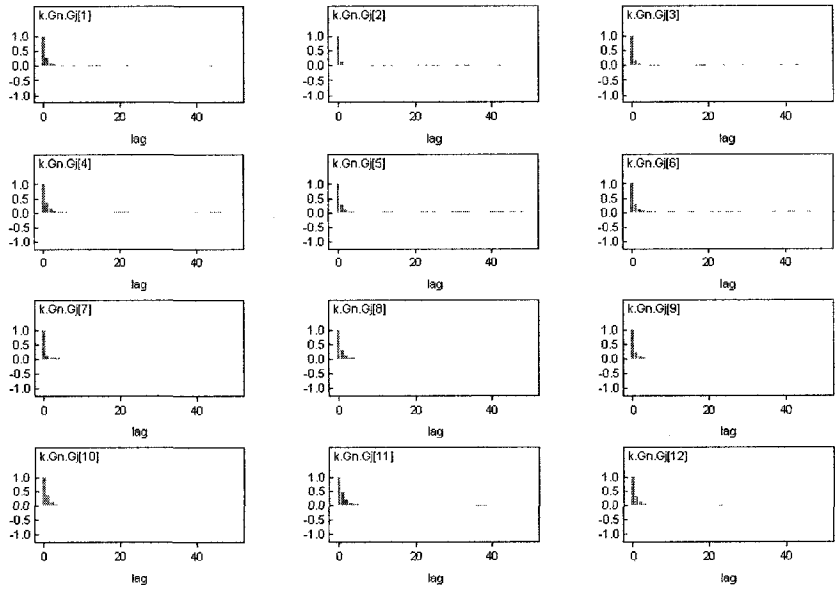
APPENDIX-C. Development of the Probabilistic BOD and TP Models

C-1) MCMC diagnostics for the BOD model at the reach of Geumnam to Gongju

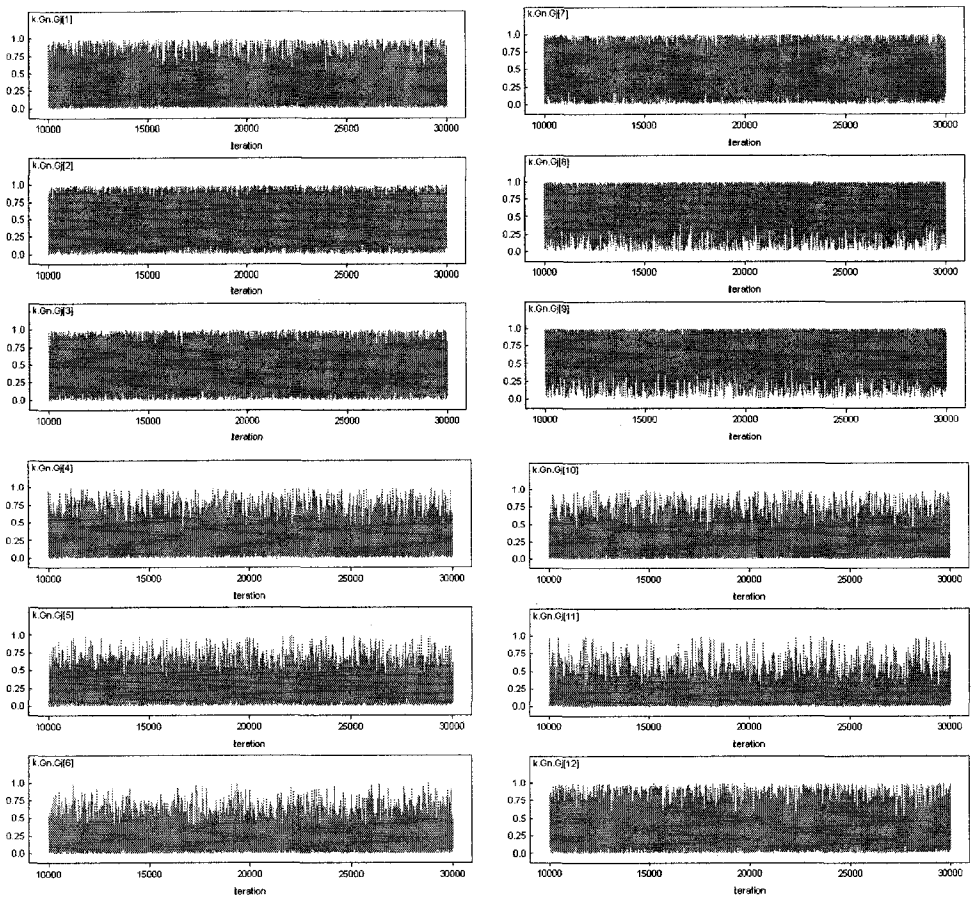
(Monthly posterior probability of $k_r.Bg.Gn$)



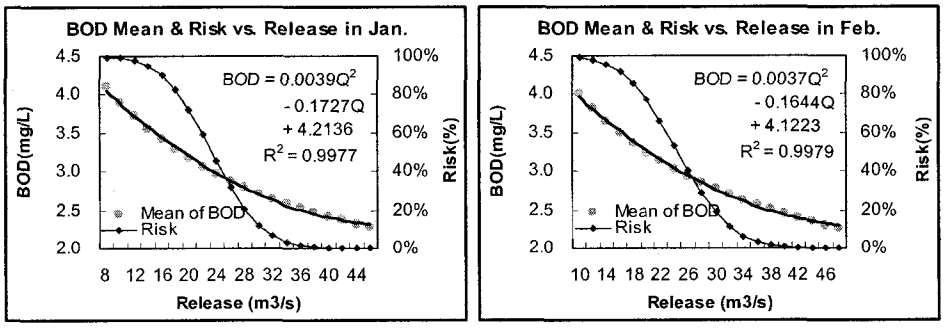
(Monthly autocorrelation of $k_r.Bg.Gn$)

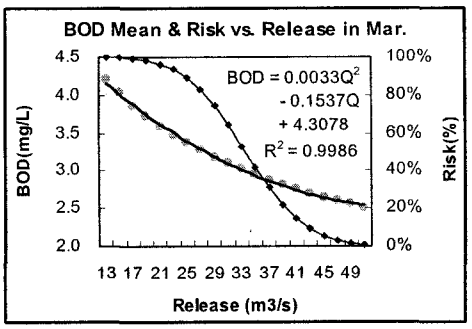


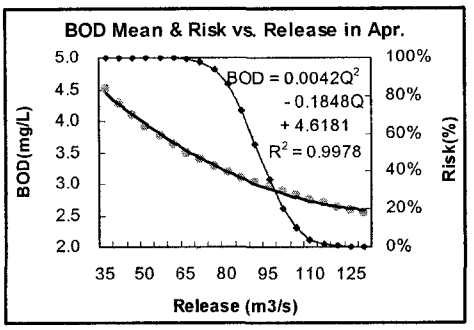
(Monthly history path)

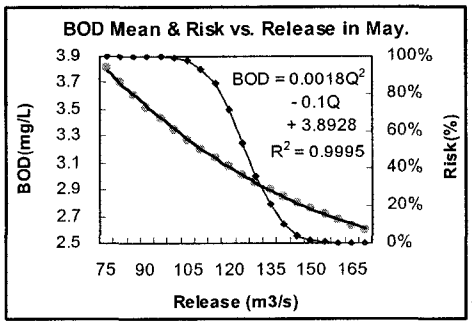


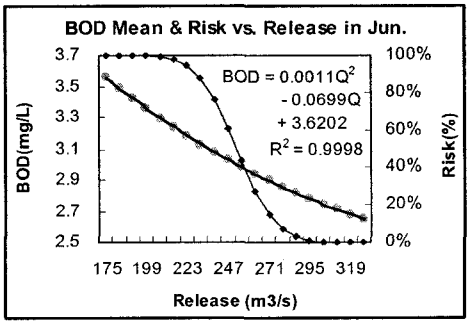
C-2) Risk of the BOD standard violation and the mean of the posterior predictive probability distribution of the simulated BOD at the Gongju by month

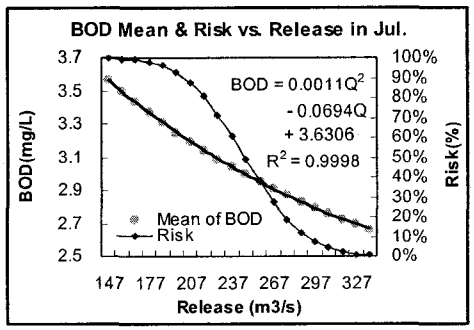


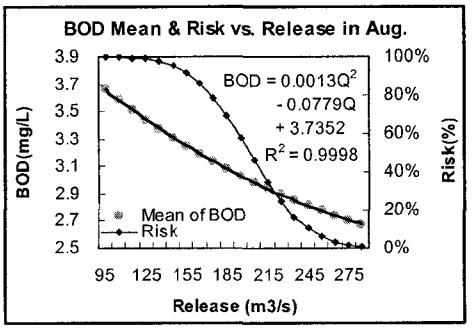


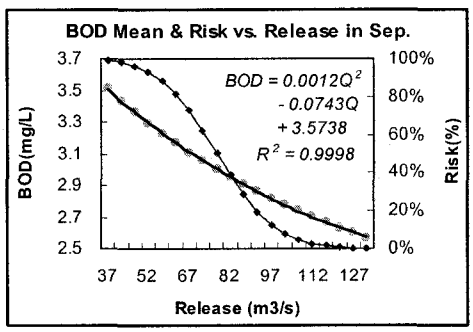


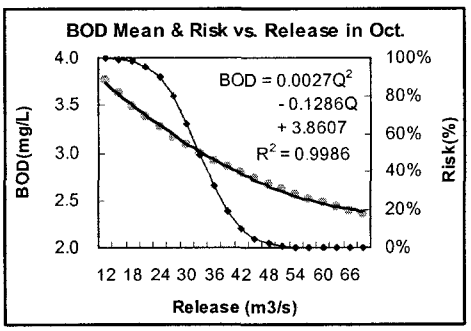


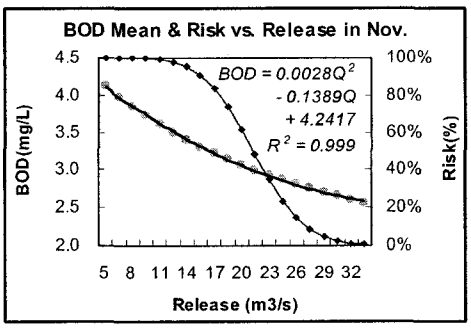


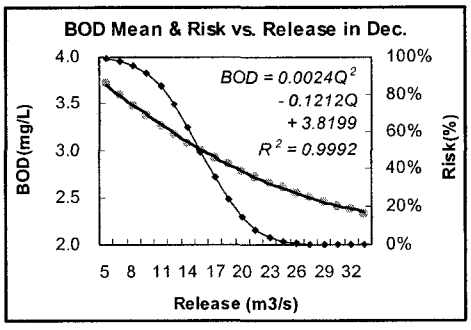












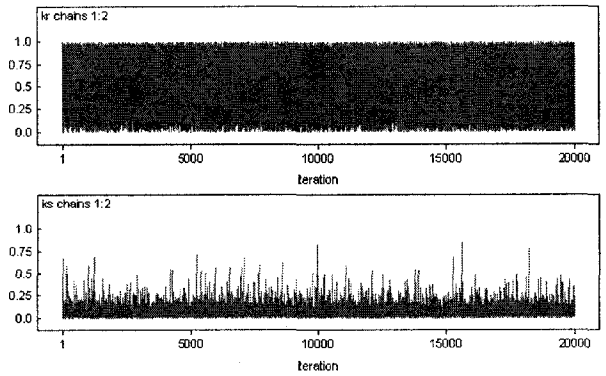
C-3) Source code of *WinBUGS* for the BOD model (DC dam ~ Gongju station)

```
# DC    : the Daecheong dam
# Gp    : the Gapcheon tributary
# Bg    : the Bugang water quality measurement site
# Mh    : the Mihocheon tributary
# Gn    : the Gumnam flow measurement site
# Gj    : the Gonhju control point (flow measurement site)
# Jn    : junction
# k     : decay + settling rate (/d)
# X     : the distance (m) between two points for water quality analysis
# U     : velocity (m/d) between two points for water quality analysis
# Q     : flow (m3/s)
# L     : BOD (mg/L), remaining oxidizable organic matter
# bef   : before
# J     : the number of observations

# Training or calibration of BOD model over reaction rate (kr)
Model {
  for (j in 1:J) {
  # right before junction between the Geum River and Gapcheon tributary
      L.bef.Gp[j] <- L.DC[j]*exp(-1*k.DC.Bg[j]*X.DC.Gp/U.bef.Gp[j])
      U.bef.Gp[j] <- 24*3600*(alp.DC.Bg*pow(Q.bef.Gp[j], bet.DC.Bg))
      Q.bef.Gp[j] <- Q.DC[j]
  # at the junction between the Geum River and Gapcheon tributary
      L.Jn.Gp[j] <- (L.Gp[j]*Q.Gp[j] + L.bef.Gp[j]*Q.bef.Gp[j])/Q.Jn.Gp[j]
      Q.Jn.Gp[j] <- Q.Gp[j]+Q.bef.Gp[j]
  # at Bugang station
      Obs.Bg[j]~dnorm(L.Bg[j], tau1)
      L.Bg[j] <- L.Jn.Gp[j]*exp(-1*k.DC.Bg[j]*X.Gp.Bg/U.Bg[j])
      U.Bg[j] <- 24*3600*(alp.DC.Bg*pow(Q.Bg[j], bet.DC.Bg))
      Q.Bg[j] <- Q.Jn.Gp[j]*inc_ratio.Bg
  # right before junction between the Geum River and Miho tributary
      L.bef.Mh[j] <- L.Bg[j]*exp(-1*k.Bg.Gn[j]*X.Bg.Mh/U.bef.Mh[j])
      U.bef.Mh[j] <- 24*3600*(alp.Bg.Gn*pow(Q.bef.Mh[j], bet.Bg.Gn))
      Q.bef.Mh[j] <- Q.Bg[j]
  # at the junction between the Geum River and Miho tributary
      L.Jn.Mh[j] <- (L.Mh[j]*Q.Mh[j] + L.bef.Mh[j]*Q.bef.Mh[j])/Q.Jn.Mh[j]
      Q.Jn.Mh[j] <- Q.Mh[j]+Q.bef.Mh[j]
  # at Geumnam station
      Obs.Gn[j]~dnorm(L.Gn[j], tau2)
      L.Gn[j] <- L.Jn.Mh[j]*exp(-1*k.Bg.Gn[j]*X.Mh.Gn/U.Gn[j])
      U.Gn[j] <- 24*3600*(alp.Bg.Gn*pow(Q.Gn[j], bet.Bg.Gn))
      Q.Gn[j] <- Q.Jn.Mh[j]*inc_ratio.Gn
  # at Gongju station
      Obs.Gj[j]~dnorm(L.Gj[j], tau3)
      L.Gj[j] <- L.Gn[j]*exp(-1*k.Gn.Gj[j]*X.Gn.Gj/U.Gj[j])
      U.Gj[j] <- 24*3600*(alp.Gn.Gj*pow(Q.Gj[j], bet.Gn.Gj))
      Q.Gj[j] <- Q.Gn[j]*inc_ratio.Gj }
```

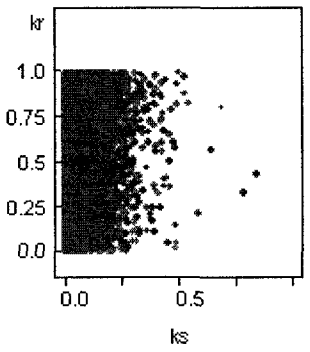
```
# Prior Probabilities for the Parameters
    k.DC.Bg[1]~dnorm(0.10, 0.01)I(0.01, 0.99)      #censoring
    k.DC.Bg[2]~dnorm(0.10, 0.01)I(0.01, 0.99)      #censoring
    k.DC.Bg[3]~dnorm(0.18, 0.01)I(0.01, 0.99)      #censoring
    k.DC.Bg[4]~dnorm(0.74, 0.01)I(0.01, 0.99)      #censoring
    k.DC.Bg[5]~dnorm(0.12, 0.01)I(0.01, 0.99)      #censoring
    k.DC.Bg[6]~dnorm(0.10, 0.01)I(0.01, 0.99)      #censoring
    k.DC.Bg[7]~dnorm(0.52, 0.01)I(0.01, 0.99)      #censoring
    k.DC.Bg[8]~dnorm(0.90, 0.01)I(0.01, 0.99)      #censoring
    k.DC.Bg[9]~dnorm(0.63, 0.01)I(0.01, 0.99)      #censoring
    k.DC.Bg[10]~dnorm(0.31, 0.01)I(0.01, 0.99)     #censoring
    k.DC.Bg[11]~dnorm(0.73, 0.01)I(0.01, 0.99)     #censoring
    k.DC.Bg[12]~dnorm(0.67, 0.01)I(0.01, 0.99)     #censoring

    k.Bg.Gn[1]~dnorm(0.90, 0.01)I(0.01, 0.99)      #censoring
    k.Bg.Gn[2]~dnorm(0.10, 0.01)I(0.01, 0.99)      #censoring
    k.Bg.Gn[3]~dnorm(0.10, 0.01)I(0.01, 0.99)      #censoring
    k.Bg.Gn[4]~dnorm(0.10, 0.01)I(0.01, 0.99)      #censoring
    k.Bg.Gn[5]~dnorm(0.10, 0.01)I(0.01, 0.99)      #censoring
    k.Bg.Gn[6]~dnorm(0.10, 0.01)I(0.01, 0.99)      #censoring
    k.Bg.Gn[7]~dnorm(0.90, 0.01)I(0.01, 0.99)      #censoring
    k.Bg.Gn[8]~dnorm(0.90, 0.01)I(0.01, 0.99)      #censoring
    k.Bg.Gn[9]~dnorm(0.85, 0.01)I(0.01, 0.99)      #censoring
    k.Bg.Gn[10]~dnorm(0.10, 0.01)I(0.01, 0.99)     #censoring
    k.Bg.Gn[11]~dnorm(0.10, 0.01)I(0.01, 0.99)     #censoring
    k.Bg.Gn[12]~dnorm(0.44, 0.01)I(0.01, 0.99)     #censoring

    k.Gn.Gj[1]~dnorm(0.10, 0.01)I(0.01, 0.99)      #censoring
    k.Gn.Gj[2]~dnorm(0.52, 0.01)I(0.01, 0.99)      #censoring
    k.Gn.Gj[3]~dnorm(0.52, 0.01)I(0.01, 0.99)      #censoring
    k.Gn.Gj[4]~dnorm(0.10, 0.01)I(0.01, 0.99)      #censoring
    k.Gn.Gj[5]~dnorm(0.26, 0.01)I(0.01, 0.99)      #censoring
    k.Gn.Gj[6]~dnorm(0.17, 0.01)I(0.01, 0.99)      #censoring
    k.Gn.Gj[7]~dnorm(0.33, 0.01)I(0.01, 0.99)      #censoring
    k.Gn.Gj[8]~dnorm(0.90, 0.01)I(0.01, 0.99)      #censoring
    k.Gn.Gj[9]~dnorm(0.75, 0.01)I(0.01, 0.99)      #censoring
    k.Gn.Gj[10]~dnorm(0.10, 0.01)I(0.01, 0.99)     #censoring
    k.Gn.Gj[11]~dnorm(0.10, 0.01)I(0.01, 0.99)     #censoring
    k.Gn.Gj[12]~dnorm(0.10, 0.01)I(0.01, 0.99)     #censoring

    tau1~dgamma(1.0E-0, 1.0E-3)
    tau2~dgamma(1.0E-0, 1.0E-3)
    tau3~dgamma(1.0E-0, 1.0E-3)
    sigma1 <- sqrt(1 /  tau1)
    sigma2 <- sqrt(1 /  tau2)
    sigma3 <- sqrt(1 /  tau3)
}
```

C-4) MCMC diagnostics for the TP model at the *DC* reservoir with monthly averaged TP loading rates and discharges from the *YD* reservoir and *YD.Oc* sub-basin

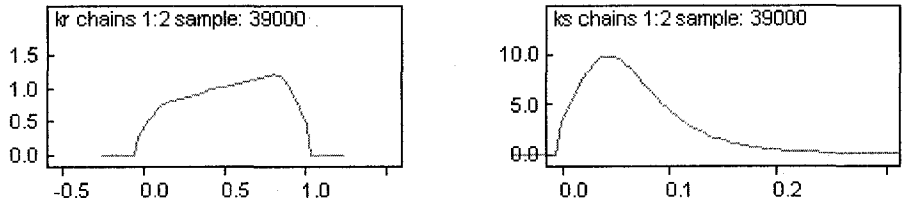
(history path)



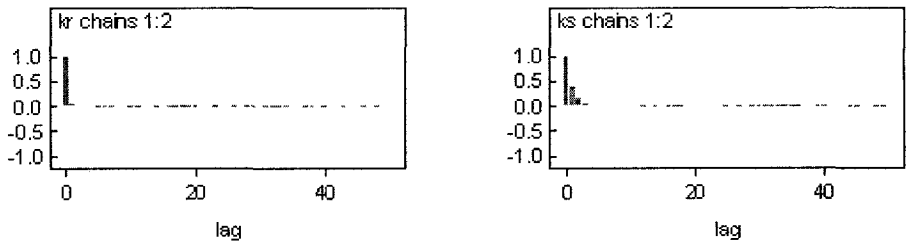
(correlation)



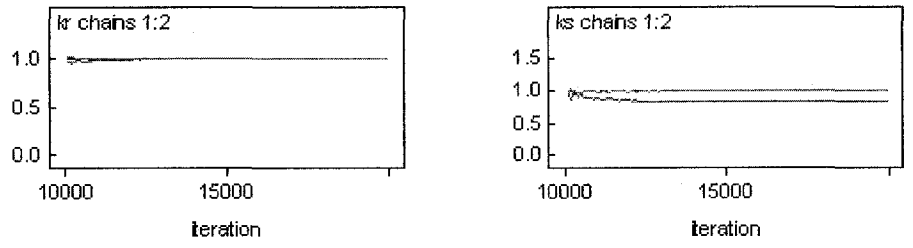
(posterior probability of k_s and k_r)



(autocorrelation of k_s and k_r)

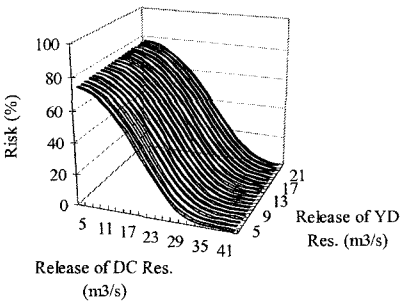


(Gelman-Rubin statistic R of k_s and k_r)

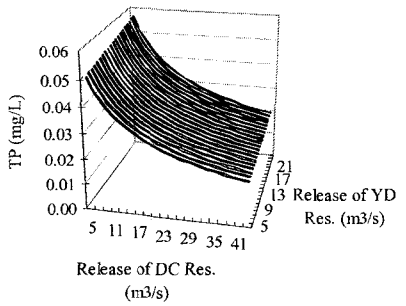


C-5) Risk of the TP standard violation and the mean of the posterior predictive probability distribution of the simulated TP in the *DC* reservoir by month

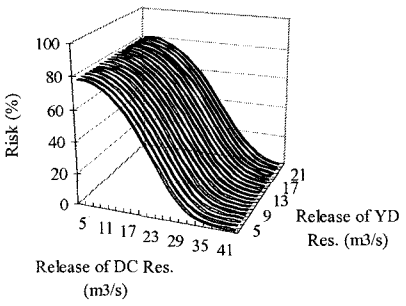
Risk of TP Standard Violation in February



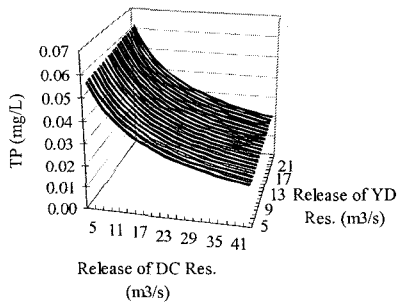
Mean of Posterior Predictive PDF of Tp in February



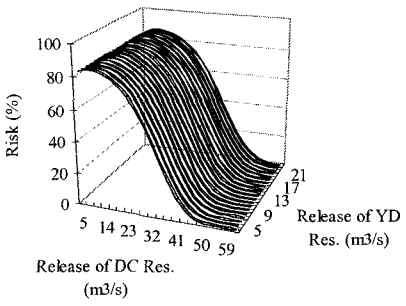
Risk of TP Standard Violation in March



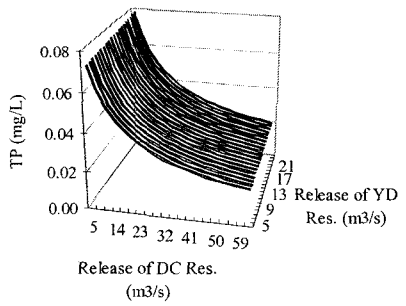
Mean of Posterior Predictive PDF of Tp in March



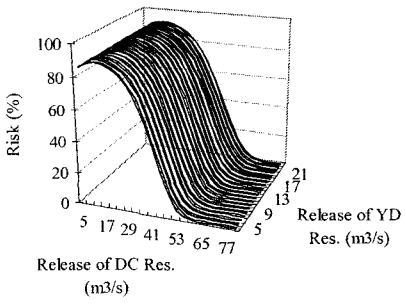
Risk of TP Standard Violation in April



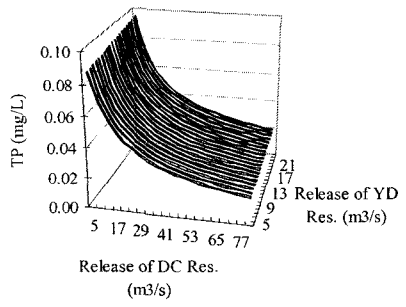
Mean of Posterior Predictive PDF of Tp in April



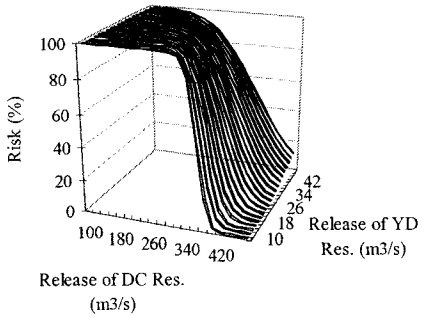
Risk of TP Standard Violation in May



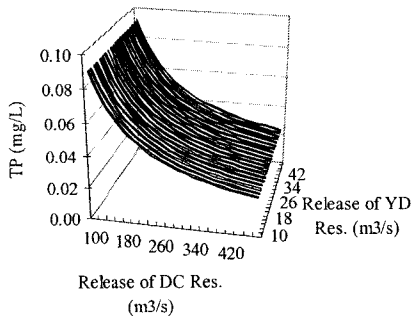
Mean of Posterior Predictive PDF of Tp in May



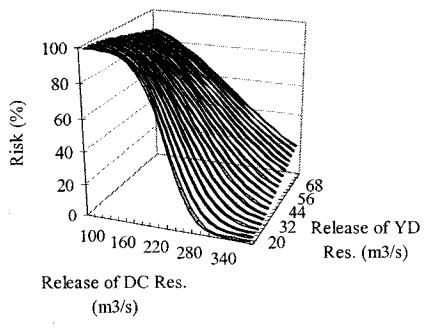
Risk of TP Standard Violation in August



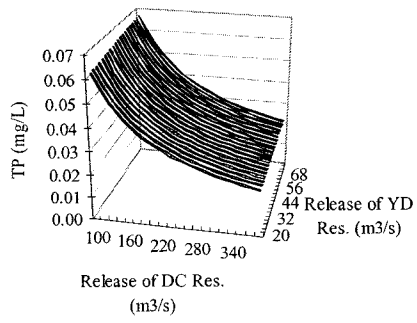
Mean of Posterior Predictive PDF of Tp in August



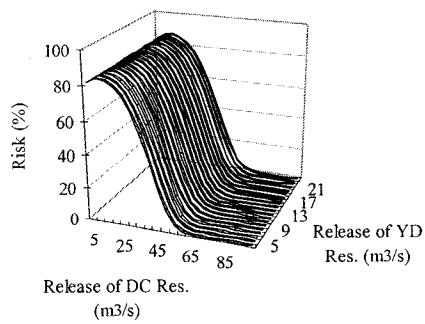
Risk of TP Standard Violation in September



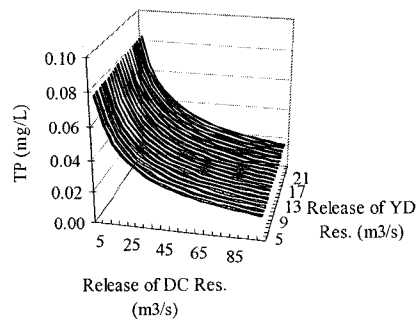
Mean of Posterior Predictive PDF of Tp in September



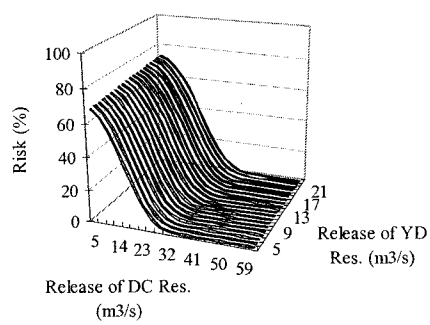
Risk of TP Standard Violation in October



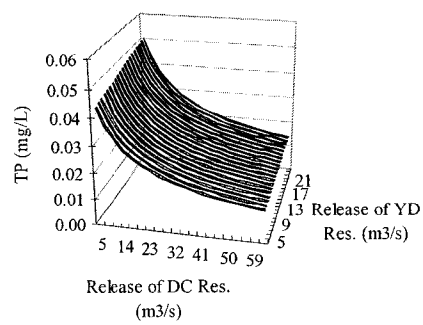
Mean of Posterior Predictive PDF of Tp in October

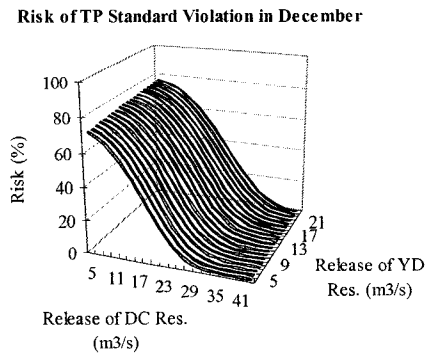


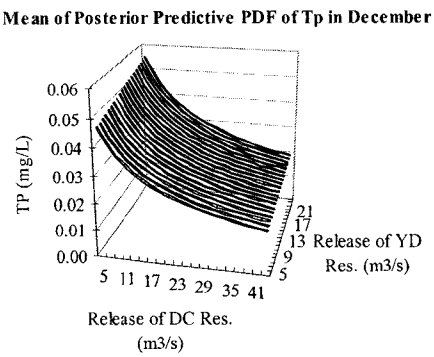
Risk of TP Standard Violation in November



Mean of Posterior Predictive PDF of Tp in November







C-6) Source code of *WinBUGS* for the TP model in the Daecheong reservoir

```
# DC          : the Daecheong dam
# YD          : the Yongdam dam
# OC          : the Ok-cheon station
# BC          : the Bo-cheong tributary
# TP          : the total phosphorous (mg/L)
# Q.YD, Q.DC  : the release of YD & DC dams (m3/s)
# Obs.Tp.DC   : the observed TP of DC dam (mg/L)
# W           : the load of TP (kg/month)
# ks          : a first order settling velocity (/month)
# kr          : the contributed rate of the TP load from YD dam at the Ockcheon
# J           : the number of observations

# Training or calibration of TP model over ks and kr
Model {
  for (j in 1:J) {
      Obs.TP.DC[j]~dnorm(TP.DC[j], tau)
      TP.DC[j] <- (W.DC[j]*1.0E6)/(Q.DC[j]*24*3600*days[j] +
                  ks*Vol.DC[j]*1.0E6)/1000

      W.DC[j] <- kr*W.YD[j] + W.YD.OC[j] + W.BC[j]
      W.YD[j] <- (TP.YD[j]*1000)*(Q.YD[j]*24*3600*days[j])/1.0E6 }
  # Prior probabilities for the parameters
      ks~dnorm(0.05, 0.001)I(0.001,0.9524)   : Normal distribution
      kr~dnorm(0.70, 0.001)I(0.001,0.99)
      tau~dgamma(1.0E-0, 1.0E-3)             : Gamma distribution
      sigma <- sqrt(1 / tau)
```