12-2008

Rank-Data Distribution Method (R-D Method) for Daily Time-Series Bayesian Networks and Total Maximum Daily Load Estimation

Joon-Hee Lee
Utah State University

RANK-DATA DISTRIBUTION METHOD (R-D METHOD) FOR DAILY

TIME-SERIES BAYESIAN NETWORKS AND TOTAL

MAXIMUM DAILY LOAD ESTIMATION

by

Joon-Hee Lee

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

DOCTOR OF PHILOSOPHY

in

Civil and Environmental Engineering

Approved:

Dr. David K. Stevens Major Professor	Dr. R. Ryan Dupont Committee Member
Dr. Thomas B. Hardy Committee Member	Dr. Darwin L. Sorensen Committee Member
Dr. Wayne A. Wurtsbaugh Committee Member	Dr. Byron R. Burnham Dean of Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah

2008

ABSTRACT

Rank-Data Distribution Method (R-D method) for Daily Time-Series Bayesian Networks and Total Maximum Daily Load Estimation

by

Joon-Hee Lee, Doctor of Philosophy

Utah State University, 2008

Major Professor : Dr. David K. Stevens
Department : Civil and Environmental Engineering

Daily time series-based models are required to estimate the higher frequency fluctuations of nutrient loads and concentrations. Some mechanistic mathematical models can provide daily time series outputs of nutrient concentrations but it is difficult to incorporate non-numerical data, such as management scenarios, to mechanistic mathematical models. Bayesian networks (BNs) were designed to accept and process inputs of varied types of both numerical and non-numerical inputs.

A Rank-Data distribution method (R-D method) was developed to provide large time series of daily predicted flows and Total Phosphorus (TP) loads to BNs driving daily time series estimates of T-P concentrations into Hyrum and Cutler Reservoirs, Cache County, Utah. Time series of water resources data may consist of data distributions and time series of the ranks of the data at the measurement times. The R-D method estimates the data distribution by interpolating cumulative failure probability (CFPs) plots of observations. This method also estimates cumulative failure probability of predictions on dates with no data by interpolating CFP time series of observations. The R-D method estimates time series of mean daily flows with less residual between predicted flows and observed flows than interpolation of observed flows using data sets sampled randomly at varying frequencies.

Two Bayesian Networks, BN 1 (Bayesian Network above Hyrum Reservoir) and BN 2 (Bayesian Network below Hyrum Reservoir) were used to simulate the effect of the Little Bear River Conservation Project (LBRCP) and exogenous variables on water quality to explore the causes of an observed reduction in Total Phosphorus (TP) concentration since 1990 at the mouth of the Little Bear River. A BN provided the fine data distribution of flows and T-P loads under scenarios of conservation practices or exogenous variables using daily flows and TP loads estimated by R-D method. When these BN outputs were connected with the rank time series estimated by interpolation of the ranks of existing observations at measurement dates, time series estimation of TP concentrations into Cutler Reservoir under two different conservation practice options was obtained. This time series showed duration and starting time of water quality criterion violation. The TMDL processes were executed based on daily TP loads from R-D instead of mean or median values.

(263 pages)

ACKNOWLEDGMENTS

I would like to thank the LORD because I have worked using the knowledge and enthusiasm which the LORD has provided. I would like thank my advisor, Dr. David Stevens. He have advised, taught, and encouraged for a long time, more than 7 years for me. I would like to thank each of my committee members, Dr. Ryan Dupont, Dr. Thomas Hardy, Dr. Darwin Sorensen, and Dr. Wayne Wurtsbaugh. Their comments were very helpful to enhance my dissertation. In addition, taking many courses suggested by my committee members was very helpful to get a job.

My dissertation research was funded by the U.S. Department of Agriculture (USDA). I also got my post doctoral position at the Agricultural Research Services (ARS) in USDA. I would like to thank the USDA. I also thank Jeff Horsburgh, who gave me much professional assistance. I thank Dr. Bethany Neilson, who helped and encouraged me for my research. I would like thank to Dr. Guyoung Kang, who introduced me to USU and my advisor, Dr. Stevens.

Finally, I would like to thank my wife, Grace, for her support and patience. I would like to my patents for their prayers, patience, and support. I thank my daughter Joanne for refreshing me through her entertainment.

Joon-Hee Lee

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

OVERVIEW

Nutrient enrichment causes anoxia problems in lakes and reservoirs by stimulating the primary producers, primarily algae. An ecosystem becomes more productive by nutrient enrichment, a process known as eutrophication. In a eutrophic ecosystem, nutrient enrichment may cause toxic chemical releases from sediment and by cyanobacteria blooms as well as fish kills due to anoxia in the hypolimnion (Dodds, 2002). Therefore, nutrient load control is important in water quality management. We need to consider the magnitude of low oxygen problems and toxic chemical release, but also important are the frequency, duration and timing of these releases because some aquatic life is very sensitive to these fluctuations in water quality.

While we need to predict the fluctuations of nutrient concentration under different nutrient load management options, most water quality models calculate monthly, seasonal or yearly nutrient concentrations (Stronge *et al.*, 1997; Worrall and Burt, 1999; Hanrahan *et al.*, 2001). These models cannot provide good enough prediction of water quality and nutrient load to manage water quality because higher frequency fluctuations may be important within a specific model period. Daily time series-based models are required to provide more accurate estimates of frequency, duration, magnitude and timing of loadings. Some mechanistic mathematical models can provide daily time series outputs of nutrient concentration (Bowen and Hieronymus, 2000, Washington State, 2004). However, we cannot often calibrate those models accurately because of the large quantities of data required for daily time series simulation for input or calibration. We often have just monthly water quality data in many watersheds. Although most water quality data, model inputs and outputs and management information are quantitative when models are used by water quality managers, often we have non-numerical input information such as non-point source control options, stakeholder-based management scenarios or future land development options. It is desirable but difficult to incorporate non-numerical input data to mechanistic mathematical models.

Bayesian networks (BNs) were designed to accept and process inputs of varied types of information: observations, model results, expert judgment, scenario type and a variety of other non-numerical inputs. Those inputs are often expressed as the probability or chance of a particular scenario. Water quality model outputs must often be expressed as the probability or chance for frequency-based water quality criteria (USEPA, 2000). The primary requirement of a Bayesian network is that it can process those probabilistic inputs and provide probabilistic outputs. Therefore, many researchers have developed Bayesian network models to support decision making for water quality management strategies (Reckhow, 1999; Neilson *et al.*, 2002; Varis and Jussila, 2002; Borsuk *et al.*, 2003; Ha and Stenstrom, 2003; WRRI, 2004). Those models provide seasonal outputs or yearly outputs because the models' time increments are tied to the time increments of the data that are available to drive the models. Often those data are only available monthly or seasonally. To address higher frequency fluctuations introduced above, we require large time series of daily data to drive daily time series models. The goal of this research is to expand and develop our current view of Bayesian network models to include high frequency time series models of nutrient loading and surface water quality responses.

In this dissertation, a Rank-Data distribution method (R-D method) was developed to generate the high frequency (daily) flows, Total Phosphorus (TP) loads and concentrations using only monthly or weekly frequency observations and predictions. Using these daily values, the BNs resulted in the probability of 21 categories for TP concentration into Cutler Reservoir in the Little Bear River watershed (Utah) under different nutrient load management strategies. The time series of rank of TP was linked with these BN outputs to provide daily time series of TP concentration into Cutler Reservoir under different nutrient load management strategies. We were able to evaluate the frequency, duration, magnitude and even timing of TP concentration criteria violation in these surface waters using these daily time series. Our goal is to provide these daily outputs with high accuracy and reliability using a new calibration process. The R-D method was able to support the TMDL process concerning daily frequency TP loads.

RESEARCH GOAL AND OBJECTIVES

The goal of this research is to develop a time series[1]-based Bayesian network model of nutrient loading and surface water quality responses. The new time-series BN model framework consists of a database, BN software and Rank time series. The database stores and provides the water quality data and quantity data from the R-D method. Bayesian Network software predicts probabilistic results of flow and nutrient concentration into the reservoir using Bayesian theory under selected scenarios. The TP concentration data from BN probabilistic outputs are assigned to different simulation

[1] In this dissertation, the term 'time series' is used in a generic sense; as a descriptive term for the dynamic behavior of data. It is not to be confused with the more formalized statistical concept of time series analysis using autoregressive, integrated, moving average models, formalized by Box and Jenkins. (1976).

dates by Rank Time Series. The following five research objectives address the time series Bayesian Network's challenges.

1) Establish the graphical model structure that can a) expose causal relationships among the water quality, water quantity and ecological variables, and; b) connect nutrient load management options to water quality variables.

2) Establish a database of daily water quality values that can support daily based BN model simulation. The database should consist of both existing observations and derived value generated based on the statistical characteristics of existing water quality observations.

3) Develop BN data structures for study sites that can expose the conditional probability of the hypothesis variables, information variables, and mediating variables under nutrient load management options. The BN data structure will be the format of Conditional Probability Tables (CPTs) and probability graphs.

4) Demonstrate the BN software to predict the effect of nutrient load managements on water quality at study sites.

5) Construct TMDL processes for TP using daily predictions of flows, TP loads and concentrations from the R-D method.

We have software to calculate probabilities in Bayesian networks (e.g., NETICA ™, Norsys Software, 1997). We also collected many water quantity data related to nutrient management and water quality. The most important challenge is to establish the water quality database for the new time series BN model and then to produce the daily water quality outputs from the BN model.

SCOPE

The dissertation consists of five major components and others (an introduction, a literature review, a summary and conclusions, and recommendation for further research).

The first major component addressed characteristics of the study site, the Little Bear River watershed, Cache County, Utah. This component includes general information and water quality characteristics of Little Bear River Watershed. Statistical analysis presents statistical summaries, the correlation among each of the water quality parameters, seasonal patterns and trends of the water quality parameters at each sampling location and the correlations among sampling locations. The results of this statistical analysis support determination of information and mediating variables of TP concentration at the inlets of two reservoirs in the Bayesian Network model. Many TP concentration data of some streams are censored so that modified data analysis methods are required to estimate correlations, multi regressions, seasonal effects of water quality, and trend analysis. This component discusses Maximum Likelihood Estimation (MLE) and non parametric methods as alternatives to substitution of half detection limit as is often used.

The second major component focuses on missing data estimation to establish a database of daily water quality values that can support daily based BN model simulation. It is not easy to link water qualities among different locations due to different sampling times for each location. We need daily water quality data for daily based BN simulation to get quantitative results because BN is a statistical probability model. Therefore, this component applies a method for missing data problems, the R-D method, associated with non parametric data generation techniques.

The third major component addresses the use of the BN model to evaluate past nutrient management practices. In the Little Bear River watershed, federally funded best management practices (BMPs) began in 1990. It is not easy to evaluate the effect of BMPs on water quality of streams because many exogenous influences affect the water quality in addition to the BMPs. Many BNs in water quality management predict the effects of pollutant management on water qualities. This component discusses application of BN to evaluate the effect of BMP on the inlet TP concentration of Hyrum Reservoir and Cutler Reservoir. This BN model simulates the TP load and concentration under the same exogenous factor levels for each different watershed nutrient management scenario to help remove the effect of exogenous factors. These results may be an appropriate indicator to evaluate past BMPs.

The fourth major component predicts daily TP concentration of the Cutler Reservoir's inlet under different nutrient load management strategies by the daily based BN model using the derived daily water quality data set from the R-D method. This BN model links watershed nutrient management, exogenous factors and water quality. This component discusses validation and calibration of the R-D method using low frequency observations and predictions as well as daily time series of TP concentration by linking BN model output to rank time series.

The fifth major component calculates TMDLs for TP at the inlet of Cutler Reservoir under two different watershed nutrient management scenarios. The daily frequency flows, TP loads and concentration from R-D method are used to calculate the total annual load for each scenario instead of any statistic representative such as mean or median of observations with low frequency. In this component, two type of TMDL approaches including the violation frequency based approach and the total amount load based approach are tested and compared.

LITERATURE CITED

Box, G.E. and G.M. Jenkins, 1976. Time Series Analysis, Forecasting and Control, Revised Edition. Holden-Day, Okland, California, 575 pp.

Bowen, J.D. and J. Hieronymus, 2000. Neuse River Estuary Modeling and Monitoring Project Stage 1: Predictions and Uncertainty Analysis of Response to Nutrient Loading Using A Mechanistic Eutrophication Model. Department of Civil Engineering, University of North Carolina at Charlotte, Charlotte, North Carolina, 79 pp.

Borsuk, M.E., C.A. Stow and K.H. Reckhow, 2003. Integrated Approach to Total Maximum Daily Load Development for Neuse River Estuary Using Bayesian Probability Network Model (Neu-BERN). Journal of Water Resources Planning and Management 129:271-282.

Dodds, W.K, 2002. Freshwater Ecology: Concepts and Environmental Applications. Academic Press, San Diego, California. 569 pp.

Ha, H. and M.K. Stenstrom, 2003. Identification of land use with water quality data in stormwater using a neural network, Water Research 37 (17):4222-4230.

Hanrahan, G., M. Gledhill, W.A. House, and P.J. Worsfold, 2001. Phosphorus Loading in the Frome Catchment, UK: Seasonal Refinement of the Coefficient Modeling Approach. Journal of Environmental Quality 30(5):1738-1746.

Neilson,B.T., D.P.Ames and D.K. Stevens, 2002. Application of Bayesian Decision Networks to Total Maximum Daily Load Analysis. ASAE Conference Proceedings. Fort Worth, Texas, March 11-13, 2000.

Norsys Software Corp. 1997. Netica Application User’s Guide version 1.05, Norsy Software Corp., Vancouver, British Columbia, Canada, 91 pp.

Reckhow, K.H. 1999. Water Quality Prediction and Probability Network Models, Canadian Journal of Fishery and Aquatic Science 56:1150-1158.

Stronge, K.M., S.D. Lennox, and R.V. Smith, 1997. Predicting Nitrate Concentrations in Northern Ireland Rivers Using Time Series Analysis. Journal of Environmental Quality 26:1599-1604.

USEPA, 2000. Ambient Water Quality Criteria Recommendations: Information Supporting the Development of State and Tribal Nutrient Criteria. Rivers and Streams in Nutrient Ecoregion 14. EPA-822-B-00-022. Office of Water.

Varis, O. and S.F. Jussila, 2002. Water Resources Development in the Lower Senegal River Basin: Conflicting Interests, Environmental Concerns and Policy Options. International Journal of Water Resources Development 18(2): 245-260.

Washington State department of ecology, 2004. Model for TMDLs. http://www.ecy.wa.gov/programs/eap/models, *accessed* May 2004.

Worrall, F. and T.P. Burt, 1999. A Univariate Model of River Water Nitrate Time Series. Journal of Hydrology 214:74-90.

WRRI (Water Resource Research Institute at University of North Carolina), 2004. Neuse Modeling and Monitoring Project: Modeling for Nitrogen TMDL Development. http://www.ncsu.edu/ncsu/CIL/WRRI/neuseltm.html, *accessed* April, 2004.

CHAPTER 2

REVIEW OF LITERATURE

TOTAL MAXIMUM DAILY LOAD (TMDL)

A TMDL is a written plan and analysis established to ensure that an impaired water body attains and maintains water quality standards in the event of reasonably foreseeable increases in pollutant loads. Section 303 (d) of the Clean Water Act (CWA) requires States and Territories and authorized Tribes to identify and establish a priority ranking for water bodies for which technology-based effluent limitation required by section 301 of the CWA are not stringent enough to achieve the water quality standard and establish TMDLs for the pollutant causing impairment in those water bodies. States and Territories and authorized Tribes must establish TMDLs at the levels necessary to implement applicable water quality standards with seasonal variations and a margin of safety that aims to take into account any lack of knowledge concerning the relationship between effluent limitations and water quality (NARA, 2000).

The TMDL process consists of five activities. The first step is to select the pollutant to consider. The second step is to estimate the water body's assimilative capacity. The third step is to estimate the pollutant load from all sources to the water body. The fourth step is to predict pollution in the water body and to determine total allowable pollutant load. The fifth step is allocation of the allowable pollutant load among the different pollution sources (USEPA, 1991).

In the allocation step, the TMDL is calculated by this equation (USEPA, 1997)

$$TMDL = \sum LA + \sum WLA + MOS \tag{2.1}$$

in which TMDL= Allowable total maximum daily load found using the Assimilative capacity for a particular water body and contaminant, LA = pollutant load allocation for non-point sources, WLA= pollutant load allocation for point sources discharges, and MOS= Margin of safety. TMDLs must contain an expression of the pollutant load or load reduction necessary to assure that the water body will attain and maintain water quality standards (NARA, 2000).

Both flow and pollutant concentration from point sources, nonpoint sources and background levels into a water body should be predicted to estimate a reasonable load to meet the water quality standard in the TMDL process. We need predictive models for this purpose (USEPA, 1991). We also need appropriate pollutant load calculation methods. There are three categories for calculation of pollutant load in TMDLs.

The first category is a monitoring based method. Typically there is a single TMDL value for a specific location using this method. Since there is no national consensus for representing existing conditions, we select a type of statistical representation (mean, median, or geometric mean) or a type of critical condition (low or high flow event) to generate pollutant loads (Brannam *et al.*, 2005). A load is then calculated by multiplying a pollutant concentration by a flow. Then the TMDL is calculated by multiplying a pollutant water quality standard by a flow from a selected statistical representation or a critical condition. As an example, the TMDL project for the Little Bear River Cache County, UT used median of historic flows in 1998 and 1999 as the representative statistic. The TMDL of total phosphorus into a specific reservoir was then calculated by multiplying the median flow by the state water quality standard, 0.05 mg/l (Utah DEQ, 2000).

A monitoring based alternative with more flexibility is the load duration method. Here each actual pollutant load is found by multiplying average daily flow by the pollutant concentration on the same day. Each target pollutant load (TMDL) is then found by multiplying each average daily flow by the water quality standard concentration to make a cumulative frequency curve that shows each target pollutant load at the frequency with which a specific flow value is equaled or exceeded. Historical loads also are plotted on this graph by the same way (Figure 2.1). If the historical load falls on or below the TMDL curve, this means compliance with water quality criteria (Neilson *et al.*, 2005). The load reduction is determined based on allowable percentage of loads above

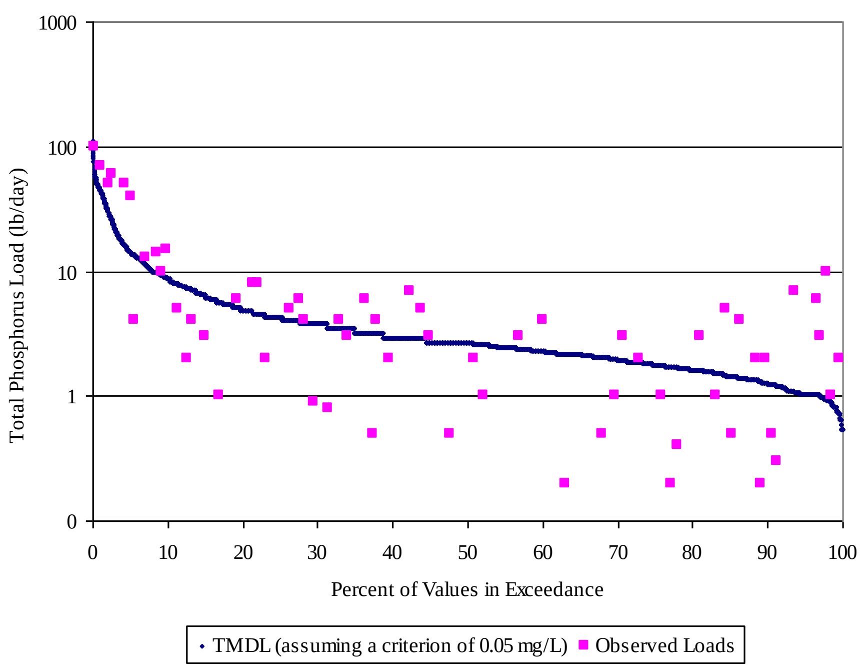


FIGURE 2.1. Load Duration Curve for Total Phosphorus (Neilson *et al.*, 2005).

the TMDL curve and the margin of safety on the load duration curve.

. The second category is to use empirical models. In many cases, we have only monthly data for water quality. This gap between measurements may cause large variance when we estimate pollutant loads (Preston *et al.*, 1989; Hodgkins, 2001). Sampling dates for flow are often not the same as the sampling dates for pollutant concentration so that we can not directly use these flow and pollutant concentrations to calculate loads. We need daily mean flow and daily pollutant concentration to avoid these problems. One approach is using regression models. In some cases, pollutant concentrations or loads have strong correlation with flow. North Carolina Department of Environment and Natural Resources used linear regression models between TP load and measured flow for summer and winter in Roberson Creek, North Carolina TMDL (NCDENR, 2004).

Summer:

$$TP\ (kg/day) = 1.79 + 7.45 * Flow\ (m^3/s) \tag{2.2}$$

$R^2 = 0.96$

Winter:

$$TP\ (kg/day) = 0.28 + 20.83 * Flow\ (m^3/s) \tag{2.3}$$

$R^2 = 0.90$

The third category is to use a loading model. Loading models require data of watershed characteristics including land use, soil information and land management practices to calculate pollutant load instead of in-stream pollutant or hydraulic routing (Neilson *et al.*, 2005). An example of this method is the Pollutant Loading Tool (PLOAD) (USEPA, 2001). In PLOAD, we calculate pollutant load using land use GIS data and export coefficients.

EPA proposed that weekly, monthly, seasonal or annual estimates may be used in the expression of TMDLs (NARA, 2000). However, the simple monthly or seasonal simulation output of flow and pollutant concentration cannot represent the characteristics of a pollutant load within the entire month or season because the flow and such pollutant concentration may have a large variation within a particular month or season. In addition, USEPA recommends identifying appropriate periods of duration and frequency of occurrence in addition to magnitude of pollutant concentration. USEPA does not encourage States and Tribes to identify nutrient concentrations that must be met at all times, rather a seasonal or annual averaging period is considered appropriate (USEPA, 2000). Therefore, the daily outputs of flow and pollutant concentration are required to predict pollutant loads more accurately in the TMDL process.

BAYESIAN NETWORK

A Bayesian Network (BN) is a probabilistic network model based on a graphical representation of the relationships among variables (Castillo *et al.*, 1997). In a BN, the relationships between parent variables and child variables are expressed by a link and node structure where the state of the parent node predicts the state of the child node (Jensen, 1996). BNs can satisfy our modeling needs, particularly by providing a useful communication medium that clearly displays major influences on water quality criteria; combines categorical and continuous variables; connects expert judgment to empirical data (Heckerman *et al.*, 1994); and expresses predicted outcomes as likelihoods as a basis for risk analysis and risk management (Marcot, 1998). In addition, a Bayesian Network can provide estimation of cost and benefit for risk management when we combine water quality probability and cost and benefit utility tables (Ames, 2002). Since a BN is a probability network model that expresses predictions probabilistically, it can support consideration of frequency-based standards such as violation on 10% of the days in a year (Smith *et al.*, 2001).

Conditional Probabilities

The basic concept in Bayesian propagation of certainties in causal networks is conditional probability (Jensen, 1996). The probability, P(A), of an event A is a number between 0 and 1. For example, 'the probability of a die turning up 5 is 1/6'. However, this statement is contingent on the unstated assumption that the die is fair. Therefore, the statement should be 'Given that the die is fair, the probability of the die turning up 5 is 1/6'. In the same way, a conditional probability statement is of the following kind:

Given the event B, the probability of the event A is x.

This probability is denoted by $p(A|B)=x$.

There are fundamental rules for probability calculation (Jensen, 1996).

$$P(A|B)P(B)=P(A,B) \tag{2.4}$$

$P(A|B)$ = The probability of the event A, given the event B

$P(A,B)$ = The probability of the joint event A∩B.

If the probabilities are conditioned by a context C, the formula should read

$$P(A|B,C)P(B|C) = P(A,B|C) \tag{2.5}$$

From Equation (2.4)

$$P(A|B)P(B)=P(B|A)P(A)=P(A,B) \tag{2.6}$$

Equation (2.6) yields Bayes' rule:

$$p(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)} \tag{2.7}$$

If the probabilities are conditioned by C, Bayes' rule reads

$$P(B / A,C) = \frac{P(A \mid B,C)P(B \mid C)}{P(A \mid C)} \tag{2.8}$$

Probability Calculus For Variables

The rules for calculation of probabilities related to each event can be applied to the probability of variables (Castillo *et al.*, 1997). We define $\{X_1,.....X_n\}$ as a set of discrete random variables and $\{x_1,.....x_n\}$ as a set of their possible realizations or instantiations. Also, we define $p(x_1,.....,x_n)$ as the joint probability distribution of variables in a variable set X. That is

$$p\ (x_1,\ldots,x_n)=p(X_1=x_1,\ldots,X_n=x_n) \tag{2.9}$$

We define X and Y as disjoint subsets of variables such that $p(y)> 0$. Then, the conditional probability distribution of X given $Y=y$ is

$$p(X{=}x|Y{=}y){=}p(x|y){=}\frac{p(x,y)}{p(y)} \tag{2.10}$$

We define x_i as a state of X and there are m different events in state x_i. If the events, $(x_i,y_1)\ldots.(x_i,y_m)$ are mutually exclusive, the marginal probability distribution is

$$p(x_i)=\sum_{j=1}^{m} p(x_i,y_j) \tag{2.11}$$

When X is a single variable and Y is a subset of variables, we obtain a particular case of Equation (2.10).

$$p(x_i \mid x_1\ldots\ldots x_k)=\frac{p(x_i,x_1\ldots\ldots.,x_k)}{p(x_1,\ldots\ldots..x_k)} \tag{2.12}$$

So that from Equation (2.11),

$$p(x_i \mid x_1\ldots\ldots x_k)=\frac{p(x_i,x_1\ldots\ldots\ldots,x_k)}{\sum p(x_i,x_1,\ldots\ldots\ldots x_k)} \tag{2.13}$$

which is the conditional probability distribution of X_i, given a subset of variables $\{X_1,\ldots.X_k\}$.

The marginal probability distribution is then obtained as

$$\begin{aligned} p(x_i)=p(X_i=x_i)&=\sum_{x_1,x_2\ldots..x_{i-1},x_{i+1}\ldots x_n} p(x_i,x_1,\ldots\ldots.,x_n) \\ &=\sum_{x_1,x_2\ldots..x_{i-1},x_{i+1}\ldots x_n} p(x_i \mid x_1\ldots,x_k)p(x_1,\ldots\ldots,x_k) \end{aligned} \tag{2.14}$$

In a Bayesian network water quality model, we use Equations (2.12), (2.13) and (2.14) to estimate the probability distribution of an endpoint using information variables. For example, if the endpoint is the total phosphorus concentration into a lake and we have data of headwater flow, headwater total phosphorus concentration, and point source flow and point source total phosphorus under Best Management Practices for the non-point source control and biological treatment for the point source, we can estimate the total phosphorus distribution into the lake using these variables with probability Equation (2.14) (Ames, 2002).

Building BN Models

The process to structure a Bayesian Network Model consists of problem definition, model inference and model validation (Ames, 2002). In problem definition, we have the following steps.

1) Identify hypothesis events and hypothesis variables.

   When we organize a Bayesian model for a decision support problem, the purpose is to give estimates of certainties for some event we are interested in, known as the hypothesis event. The hypothesis event is organized into a set of variables and the variables are hypothesis variables (Jensen, 1996). For example, we may be interested in the violation of a total phosphorus concentration criterion at the inlet to a reservoir. If the criterion is 0.05 mg/L, the hypothesis event is whether the total phosphorus concentration is over 0.05 mg/L or not and total phosphorus concentration is the hypothesis variable.

2) Identify the load management option.

The concentration of contaminants from runoff is dependent on land use (Ha and Stenstrom, 2003). We use headwater management such as Best Management Practices (BMPs) and point load management such as a biological process at a wastewater treatment plant (WWTP) to reduce contaminant load. Therefore, head water management options and point source management options should be identified as information variables.

3) Identify information variables and mediating variables.

We use much information in modeling tasks. Some information may reveal something about the state of some hypothesis variable. This revealing is done by certain variables and these variables are information variables (Jensen, 1996). For example, if we have data of headwater flow of the stream into a reservoir, these headwater flows are information variables for the TP concentration of the inlet to the reservoir because headwater flow affects the TP concentration of the inlet to the reservoir.

Some variables in a BN model are neither hypothesis variables nor information variables but these variables affect the state of hypothesis variables. We call these mediating variables. For instance, if the flow dominantly affects TP in a tributary and we do not have any information about TP under each flow category, the TP in the tributary is a mediating variable for TP of the inlet into a reservoir.

4) Identify data sources and categories of variables.

We use existing data, results of the model simulation and expert judgment as data sources in a Bayesian Network design (Varis and Jussila, 2002; Neilson et al., 2002).

Data sources should be identified and states or categories for all variables must be defined (Ames, 2002). For instance, Ticehurst *et al*. (2007) categorized phosphorus load from a catchment to the Cudgen Lake as <600 kg/yr, 600 – 800 kg/yr, 800 – 1000 kg/yr, 1000 – 1200 kg/yr and >1200 kg/yr.

5) Identify evaluation criteria related to probabilistic results.

After calculating the probabilities of each state of the hypothesis variables, we have to evaluate if this probability is acceptable. For instance, if such a model predicts a 25% probability of a major summer fish kill in a study of low oxygen effects on fish kills, we need to decide if this risk is acceptably low or not (Borsuk and Reckhow, 2004). Therefore, we need to identify criteria of acceptable probabilities.

6) Make the graphical model.

A graphical model represents the variables and relationships among variables using nodes and links (Figure 2.2). This graphical model is developed through a joint process of stakeholder involvement and scientific characterization (Borsuk *et al.*, 2001).

In model inference, we need the following steps.

1) Data arrangement.

Researchers arrange data for important variables according to categories which are made during problem identification. This helps us make conditional probability tables.

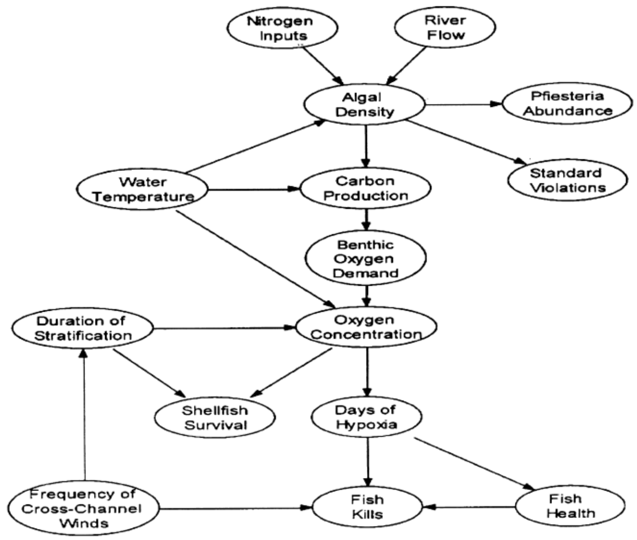


FIGURE 2.2. Graphical Model of Eutrophication in Neuse River Estuary (Borsuk *et al.*, 2003).

2) Make conditional probability table.

After collecting data, we need to compute conditional probability distributions. Usually, conditional probability tables (CPTs) represent the probability of empirical information, expert judgment and mechanistic model simulation results. CPTs are the tables that represent the probability or frequency with which a node takes on each discrete state, given the states of any parent nodes (Marcot *et al.*, 2001). Researchers make conditional probability tables for important variables based on data categories and the probabilities for each category combination (Table 2.1).

3) Calculate probability of a hypothesis variable.

BNs calculate the probability of hypothesis variables using CPTs and probability equations. The CPTs of variables that are given as conditions of hypothesis variable

TABLE 2.1. An Example of a Conditional Probability Table (Neilson *et al*., 2002). (Numbers are categories of each variable)

				In-stream Phosphorous Concentration (mg/L) (PH_ST)		
PH_HW	PH_TP	FL_HW	FL_TP	< 0.05	0.05 - 0.10	0.10 – 1.00
...	...	...	...	...	...	...
1	2	2	2	100%	0%	0%
1	3	1	1	91%	9%	0%
1	3	1	2	46%	35%	19%
1	3	2	1	95%	5%	0%
1	3	2	2	57%	31%	12%
1	4	1	1	30%	38%	33%
1	4	1	2	0%	0%	100%
...	...	...	...	...	...	...
3	2	1	1	82%	9%	9%
3	2	1	2	73%	27%	0%
3	2	2	1	8%	53%	39%
3	2	2	2	7%	40%	53%
3	3	1	1	45%	33%	22%
3	3	1	2	16%	43%	41%
...	...	...	...	...	...	...

are linked to CPTs of the hypothesis variable. For instance, Neilson et al. (2002) studied the effect of flow and Total Phosphorus (PH in Table 2.1) concentration of a point source and headwater of reservoir influent under nutrient load management options in East Canyon, Northern Utah. In this research, they constructed CPTs of head water flow (FL_HW) and wastewater treatment plant (WWTP) flow (FL_TP) for each season, total phosphorus concentration of the WWTP discharge (PH_TP) under each wastewater treatment technology option, and head water total phosphorus concentration (PH_HW) under non-point load management options. They made CPTs of reservoir influent TP concentration (PH_ST) related to FL_HW, PH_HW, FL_TP and PH_TP. They then estimated the probability to meet a 0.05 mg/L, total phosphorus water quality criterion in the inlet to East Canyon Reservoir (PH_ST) under each WWTP and headwater total phosphorus management option linking the upper CPTs.

4) Calculate probability of additional endpoint variables

BNs may use a hypothesis variable for a water quality parameter as a parent node for an ecological variable that stakeholders are interested in. For instance, the probability of a fish kill (ecological) can be predicted using the results of chlorophyll-*a* concentration and oxygen concentration (water quality) probabilities (Reckhow, 1999).

Validation is required for all models to declare the model to be reasonable for decision support (Stow *et al.*, 2003). Predictions of BN models consist of full probability distributions, rather than single values. For validation of probabilistic models, researchers have used median values as the point prediction (Scavia *et al.*, 1981; Stow *et al.*, 2003). They compared these predicted median values to observed values and evaluated the errors using the correlation coefficient, reliability index, and the average absolute error between the model predictions and observations (Stow *et al.*, 2003). Bowen and Hieronymus (2000) compared the frequency of predicted salinity to the frequency of observed salinity for model validation in their Neuse River Estuary modeling research. They found that the model underpredicted salinities by about 1.0 g/l for cumulative frequencies less than 0.9 and by 0.5g/l for cumulative frequencies in the upper 10% after calibration based on comparison between observations and predicted value using time series trends. They used a mechanistic model, CE-Qual-W2 (Cole and Buchak, 1995) in their research but this validation method may also be useful for probabilistic models.

Application to Water Quality Management

This section shows three specific examples of application of Bayesian Networks for water quality management.

East Canyon Creek Bayesian Network Case Study. East Canyon Creek is located in Northern Utah and flows north 26 km from Kimball Junction into East Canyon reservoir. High phosphorus loading caused eutrophication in the reservoir and the loading source are Snyderville Basin wastewater treatment plant (WWTP)'s discharge and non-point sources including agriculture, recreation and residential areas (UWRL, 2000).

Neilson *et al.* (2002) studied the effect of flow and total phosphorus (TP) concentration of the WWTP effluent and the headwater on reservoir influent TP under nutrient load management options in East Canyon. In this research, the hypothesis variable was reservoir influent total phosphorus concentration and the hypothesis event was whether the reservoir influent phosphorus was below 0.05 mg/L. WWTP management options were: 1) no biological treatment; 2) biological treatment; 3) advanced technology for effluent TP concentration of 0.1 mg/L and; 4) advanced technology for effluent TP concentration of 0.05mg/L. Headwater management options were: 1) no best management practice, and; 2) execution of best management practice. They assumed that WWTP TP concentration depended on WWTP management options, the headwater TP concentration depended on the headwater management option and flow of the headwater depended on season (Figure 2.3). They produced CPTs linking season to flow of the headwater, CPTs linking the WWTP management option to the WWTP TP concentration and CPTs linking the headwater management option to the headwater TP concentration. These loading CPTs were based on existing data but CPTs of reservoir influent TP concentration, under given headwater TP concentration and flow and WWTP TP concentration and flow were based on results of a water quality model (QUAL2E, Brown and Barnwell, 1987) simulation using these loads as inputs. They calculated probabilities of meeting 0.05mg/L TP in the inlet of East Canyon Reservoir using CPTs under WWTP and headwater TP management options during summer. In addition, combining these probabilities to a utility table for costs and number of visitors related to TP concentration, they calculated costs and benefits of sets of WWTP and headwater management options during summer (Neilson *et al.*, 2002).

Neuse River Estuary Bayesian Probability Network Model. The Neuse River Estuary, North Carolina has experienced severe consequences of eutrophication including algal blooms, fish kills and extensive hypoxia and anoxia. Nitrogen has been identified as the major factor limiting algal biomass in the Neuse River Estuary.

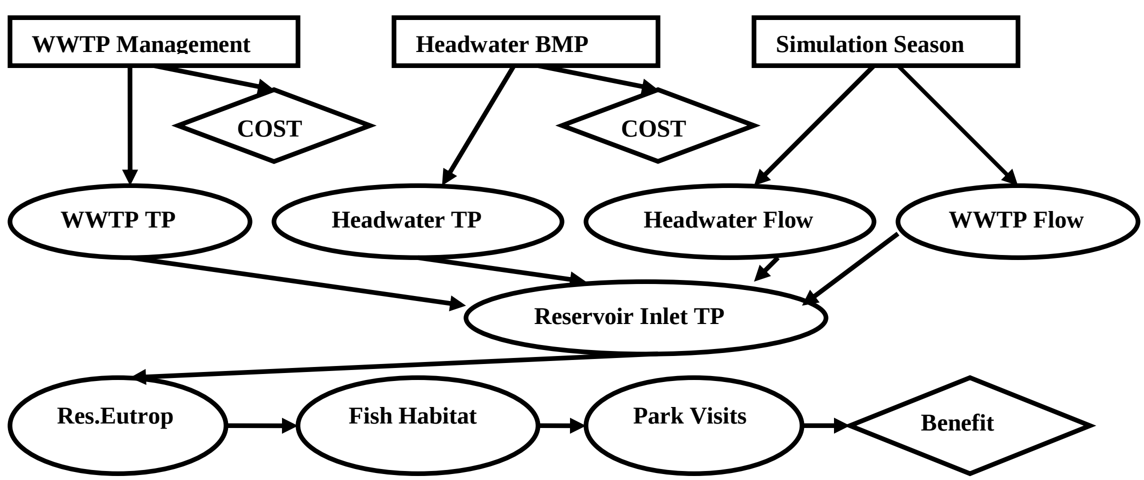


FIGURE 2.3. Complete BN for East Canyon Creek Watershed TP Management Issue (Neilson et al., 2002).

Therefore, a TMDL for total nitrogen is being established. Borsuk *et al.* (2003) established Bayesian network models for this TMDL. They established a graphical model representing the variables and relationships important to eutrophication. In this step, they used a joint process of stakeholder involvement and scientific characterization (Borsuk *et al.*, 2001). The hypothesis variable was fish kill and the hypothesis event was if fish kills are reduced under nitrogen input reduction. Each submodel simulated each variable and the probabilities characterized by various submodels were joined into one integrated Bayesian network. The information variables were nitrogen inputs, river flow, water temperature and duration of stratification. These variables were marginal variables without a parent node and were derived from historical data. When these variable distributions were specified, the predictive distributions of all the remaining variables were calculated from a network of regression submodels (Figure 2.4). They evaluated five possible scenarios corresponding to nitrogen input reduction of 0, 15, 30, 45 and 60% relative to 1991-1995 baseline inputs under the same conditions of all variables except nitrogen inputs. The integrated model prediction showed that the annual average chlorophyll-*a* concentration in the middle region of the estuary was expected to be around 20 μg/L but the state chlorophyll standard of 40 μg/L will most likely be violated on more than 10% of the days in the year under the base line conditions. As nitrogen inputs are reduced, frequency of chlorophyll standard exceedence was expected to decrease as well as average chlorophyll concentration.

Consequently, exceedence frequency met the state standard, the exceedence of chlorophyll concentration of 40 μg/L on 10% of day in a year at 45% nitrogen reduction with 50% confidence. However, the frequency of fish kill was not expected to change substantially with nitrogen reductions. While carbon production was predicted to decrease with nitrogen reduction, this effect was dampened-out through the down causal chain (Figure 2.4), so that the reduction in the number of days of resulting summer time hypoxia with nitrogen reduction was predicted to be insignificant (Borsuk *et al.*, 2003).

Water Resources Development In The Lower Senegal River Basin. The Senegal River conveys water over a distance of 1,800 km supplying water to 5 million people in a region of 500,000 km^2 in Western Africa. Lac de Guiers is the most important lake fed by the Senegal River and is used for city water supplies and

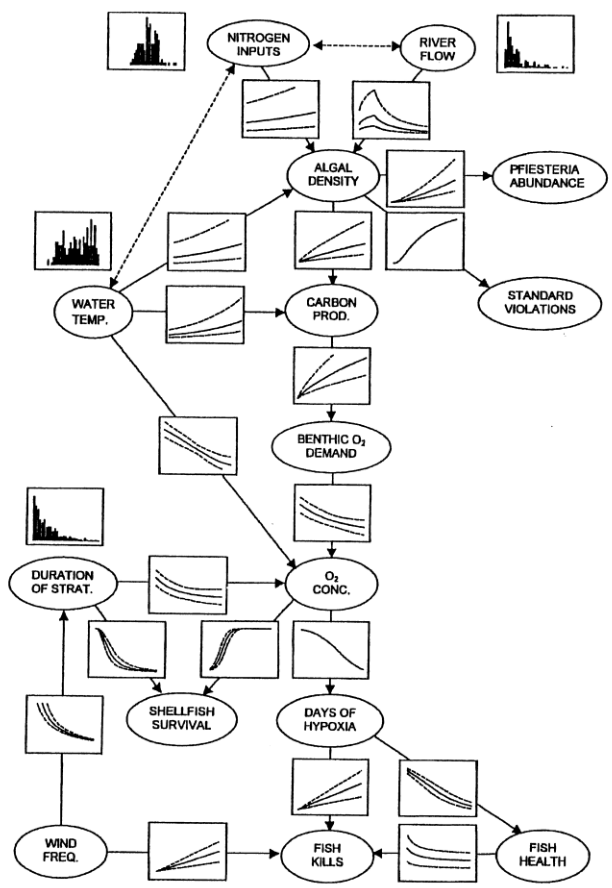


FIGURE 2.4. Fully Characterized BN for Neuse Estuary (Borsuk *et al.*, 2003).

agriculture. Until 1986, when Diama Dam was completed between the lake and river mouth, seawater entered the lake, causing an increase in salinity during the dry season while the lake was filled by low salinity water from the Senegal River during the rainy season. The expansion of sugar cane plantations caused eutrophication in the lake. Since the watershed around Lac de Guiers has high population growth, both development and conservation of water resources were required (Varis and Jussila, 2002).

Varis and Jussila (2002) used a Bayesian network model to study the conflicting interests among the various stakeholders, and the environmental and social concerns. They defined stakeholders concerns including the effect of city water supplies, agricultural water uses, cattle breeding and fishing as hypothesis variables, plus environmental factors such as plankton, macrovegetation, nitrogen, phosphorus and conductivity. They established five scenarios, combining water management policy options and tested these scenarios using a Bayesian network and documented data of the local ministries. The policy options were the methods of extension of water withdrawal for Dakar city, control method of the irrigated sugar cane industry, control of land use surrounding the lake, control of seasonal fluctuation pattern to the lake level and control of macrovegetation.

This Bayesian network model represents impacts of water management options such as conductivity increasing or fisheries being worse using probabilistic outputs. According to the results from this model, they recommended keeping the water level constant, implementing local aquatic vegetation management, locating the water intake in the southern part of the lake and construction of a diversion pipeline or canals to redirect the saline and nutrient rich water caused by the sugar cane industry from the lake to the Senegal River (Varis and Jusilla, 2002).

Limitation of Current Bayesian Network Models

Bayesian network models have been found to be useful for frequency based water quality evaluation, the effect of water management policy options and stakeholder's concern hypothesis in many applications. However, current BN models have limitations.

First, most Bayesian networks give seasonal- or yearly- based results (Neilson *et al.*, 2002; Borsuk *et al.*, 2003; WRRI, 2004). Seasonal BN models are not useful for duration- and timing-based evaluation of water quality in rivers and streams. We can realize how frequently the water quality parameter violates the water quality criteria in specific seasons using seasonal BN model but we cannot predict what month is most critical to water standard violation or the duration of continuous water quality criteria violation. We require daily-based simulation results and large time series of daily data to drive daily time series models.

Second, some BN models provide good information on the sensitivities and relative importance of different impacts of the water management options but not quantitative predictions of water management options due to a lack of data (Varis and Jussila, 2002). We need quantitative predictions to estimate costs and benefits as well as sensitivity analysis (Ames, 2002).

Third, in most reviewed BN models, validation was discussed but calibration was not discussed (Ames, 2002; Stow *et al.*, 2003). Most mechanistic mathematical models use one data set for calibration and another data set for validation. We need to calibrate

BN models as mechanistic mathematical models to increase the accuracy and reliability of their predictions.

These three problems are caused by a general lack of water quality data. New methods are required to reduce the uncertainties by a lack of data in BN models. The goal of this research is to develop a time series-based Bayesian network model of nutrient loading and surface water quality responses to overcome a lack of data. This new BN model will provide daily-based prediction of nutrient load and concentration from monthly water quality data using a data generation technique. The new BN model also will provide both sensitivity and quantitative predictions related to nutrient managements using daily probability output and have high accuracy and reliability by a calibration process. Therefore, the daily-based result will support duration- and timing-based water quality evaluation and TMDL processes.

CHARACTERISTICS OF WATER QUALITY DATA

Generally, water quality data have the following characteristics.

First, water quality data always have uncertainty. Since various errors occur during sampling, analysis and data collection, all water quality data are only estimates of some condition. Total error consists of sampling error and non-sampling error. Sampling error includes errors originating from inherent sample variability, errors originating from population variability, sampling design errors and field procedure errors. Non-sampling error includes laboratory errors (errors caused by improper sample storage, improper preparation or analysis procedure, and incorrect analytical data interpretation) and data management errors (Popek, 2003). Concerning the uncertainty of water quality data, a distribution of water quality data can represent the population more efficiently than one value.

Second, water quality data are usually dynamic in nature, but most water quality databases are populated with low-frequency data. It is easy and inexpensive to collect daily flow or temperature data but the frequency of the water chemical data is once or twice a month in historical data sets and sampling plans for many study sites. For instance, the sampling period of phosphorus and nitrogen was every 2 weeks in the sampling plan of Weber River Basin Project (UWRL, 1998) and every 3 weeks in the sampling plan of the Nooksack watershed TMDL project (Cusimano *et al.*, 2002). The WQ sampling sites at Weber River near Plain City have just 23 data points of Total phosphorus from October, 1999 to September, 2000 (Stevens, 2004) even though the USGS gage near the WQ sampling location has daily observations during the same period (USGS, 2004).

Third, many water quality data sets have non-normal distributions. While many statistical analyses assume data follow a normal distribution, water quality data often follow skewed data distributions (Helsel and Hirsch, 2002). Turbidity distributions of the Little Bear River at Mendon road from 1990 to 2004 are approximately log-normally (USEPA, 2005). Fecal coliform data of Nooksack River at Kamm Creek, Washington in the combination of 1994, 1995 and 1996 year data was very close to log-normal distribution (Mathews, 1994, 1995). Therefore, some data sets are required transformation before normal theory statistical analysis. Some researchers demonstrated nonparametric procedures, free from data distribution assumption, to environmental studies (Gilbert, 1987; Helsel and Hirsch, 2002).

Fourth, data are reported only above some threshold, while some observations are reported only as below that threshold (censored data) –that is we cannot know actual measurement values for a portion of the population. Censoring may occur when the pollutant concentration is very near or below the measurement limit of detection (Gilbert, 1987).

Fifth, water quality data often have seasonal patterns. The pollutant concentrations tend to be higher or lower in certain seasons of the year (Helsel and Hirsch, 2002). The seasonal cycles makes it difficult to detect long-term trends (Gilbert, 1987). For example, turbidity was low during winter but increased significantly in spring at Stoddard, Weber River, UT from October 1994 to September 1995 (UWRL, 1998). Nitrate concentration had seasonal cycle with high concentrations found in fall and winter for long term from 1960s and 1980s at North Cedarville of Nooksack River, Washington (USU, 2001).

PROBLEMS ASSOCIATED WITH LACK OF DATA

There are three types of lack of data problems that make it difficult to execute data analysis and modeling in water quality.

The first is unbalanced data sets. We have the problems of empty cells in designed experiments and the estimation of variance components with unequal cell frequencies (Hartley and Hocking, 1971). For example, we have data of TP, pH, Specific conductance and flow with same sample date in 1990 and 1991 but we don't have any turbidity data in those years at the location 4905740 of South Fork Little Bear River in

Cache County, UT (EMRG, 2005). In this case, it is not possible to get the correlation or regression among turbidity and other parameters.

The second is censored data. We have many specimens for which the concentration is reported as “not detected” or “below the analytical method detection limit”. These data are called censored data (Berthouex and Brown, 1994). Censored data make it difficult to summarize and compare data sets and can cause biased estimates of means, variances, trends and other population parameters. Some statistical analysis cannot work for data sets with censored values (Gilbert, 1987). Deleting censored data can obscure the information in graphs and numerical summaries (Helsel, 2005). We need better ways for statistical analysis of data sets with censored values. The most common method in environmental engineering for censored data is substitution of one-half the detection limits, multiplying limit values by 0.5 (Nehls and Akland, 1973) but some literature shows that this method is not good for interpreting censored data in comparison with other methods (Gleit, 1985; Helsel and Cohn, 1988). Helsel (2005) showed some application of Maximum Likelihood Estimation (MLE) and non parametric methods as alternatives for environmental data analysis with censored values.

The third problem is low and irregular frequency of measurements. The most common frequencies of water quality data are at weekly or monthly intervals. The low and, often, irregular frequencies of these measurements makes it difficult to fully characterize the dynamics of water quality in natural waters, and to calibrate and corroborate dynamic water quality models. Low frequency data are not sufficient to simulate high frequency statistical water quality models. High frequency sampling and measurement is one solution of this problem but this solution requires much money and time (Cusimano *et al.*, 2002). The better way is to estimate more unmeasured water quality values based on the distribution of water quality measurements.

The fourth is the gap between sampling dates. In water quality data sets, we often find long periods with rare water quality data between two specific times. For example, we have a lot of total phosphorus data from 1991 to 1993 and from 1997 to 1999 but we have no TP data from 1994 to 1997 at East Fork Little Bear River, Cache County, UT (USEPA, 2005). Trend analysis may be sensitive to this data gap in some cases. For example, if the study period is from 1970 to 1985 but we have records running from 1976 through 1984, it is probably prudent to use these records on this study because record for first 6 years and last one year were lost (Helsel and Hirsch, 2002).

In this dissertation, applications of several statistical methods to lack of data problems were studied. While some researchers have used these methods mainly for censored data problem, here these methods were applied to other lack of data problems as well as censored data. It will be shown how to use these methods in lack of data problems in Chapters 3 and 4.

Maximum Likelihood Estimation (MLE)

MLE is an optimization method to estimate parameters in distribution functions. MLE requires the specification of data distributions such as normal and lognormal. Optimization of the parameters of the data distribution function produces the specific distribution that best fits the observed data. MLE methods solve a likelihood function L, where for the data distribution with two parameters β_1 (mean) and β_2 (variance), L is a function of β_1 and β_2 and defines the likelihood of matching the observed distribution of data. The fit between the estimated distribution and the observed data is best when the function L is the maximum. In practice the natural logarithm ln(*L)* is used rather than L itself and maximizing ln(*L*) is calculated by setting the partial derivatives of ln*(L)* with respect to the two parameters equal to zero (Helsel, 2005).

$$\frac{d\ \ln[L(\beta_1)]}{d\beta_1} = 0 \ \text{ and } \ \frac{d\ \ln[L(\beta_2)]}{d\beta_2} = 0 \qquad (2.15)$$

Hartley (1958) estimated missing frequencies by MLE method assuming Poisson and binomial distributions. Some researchers applied MLE method to environmental studies with censored data under lognormal data distribution assumption (Shumway *et al.*, 1989: Charles and Stedinger, 1996). The MLE method for statistical studies with censored data uses three pieces of information: a) data above detection limits, b) the proportion of data below each detection limit, and c) the equation for an assumed distribution. In the general cases of environmental studies, we can consider *L* as the product of two pieces.

$$L = \Pi\ p[x_i]^{\delta i} \bullet F[x_i]^{1-\delta i} \qquad (2.16)$$

where *p[x]* is the probability density function (pdf) estimated from the observations equal to or above detection limits, $F[x]$ is the cumulative distribution function estimated from left-censored observations which are below detection limits and δ is the indicator of censored (0) or detected (1) data. $p[x_i]$ and $F[x_i]$ are determined by assumption of the distribution. The likelihood function L is used in the Equation (2.15) to estimate parameters, β_1 (mean) and β_2 (variance). β_1 (mean) and β_2 (variance) are the parameters that produce assumed distribution with the highest likelihood of producing the observed values for the detected observations and the observed portion of the data below each of the detection limits (Helsel, 2005).

MLE can be used for hypothesis test including comparing groups, correlation and regression as well as computing summary statistics concerning detection limits (Slymen *et al.*, 1994). However, MLE is strongly dependent on the data distribution type and does not give reasonable results when the detected sample size is small (Gleit, 1985; Shumway *et al.*, 2002).

Nonparametric Methods

A nonparametric method is a statistical method with certain desirable properties that hold under a relatively mild assumption regarding the underlying population from which the data are obtained. Nonparametric methods do not involve a specific data distribution. For this characteristic, nonparametric methods allow analysts to obtain more reliable p-values in hypothesis tests than parametric methods in many situations. Nonparametric methods are applicable in many situations where normal theory cannot be used by using ranks of observations rather than the actual magnitude of the observations. Nonparametric methods are relatively insensitive to outlying observations (Hollander and Wolfe, 1999). However, commercial computer software does not support many nonparametric methods even though development of the computer made time-consuming computations possible. For example, most software does not support Kendall's method or Akintia-Theil-Sen method to estimate trend slope in standard statistics packages (Helsel, 2005)

The Kaplan-Meier method (K-M method) produces survival function plots by nonparametric estimation. The Kaplan-Meier estimator of a survival function at time t is calculated using the following equation (Kaplan and Meier, 1958).

$$p(t_i) = \Pi \frac{1 - d_i}{n_i} \tag{2.17}$$

where n_i = number of patients who have not died or been censored before t_i, d_i = number of deaths (failure) at time t_i and $p(t_i)$= survival probability.

Even though the K-M is the fundamental method for survival analysis, it is often overlooked when a left or right censored data arises (Ware and Demets, 1976). Concentration below detection limits are left-censored while survival data are related to right censored. Many statistics software implement the K-M method for survival analysis recognizing right censored data but not left censored data. However, we can transfer left censored data to right censored data by flipping.

$$Flip_i=M_i-x_i \tag{2.18}$$

where, M_i= Flipping Constant, x_i=observation data, $Flip_i$=Flipped value.

Using the flipped values, we assign ranks to all detected observations from small to large, accounting for the number of censored data between each set of detected observations. In order to use Equation (2.17) for water quality data, redefinition of each term is required. The time t_i represents a flipped concentration value, n_i is the number of observations, both detected and censored, at and below each detected concentration, and d_i is the number of detected observations at that concentration (Helsel, 2005). We can make a survival function plot using probability from Equation (2.17). When we have several observations at the same value, the mean of n observations is follows.

$$\mu = \sum \frac{f_i}{n} x_i \qquad (2.19)$$

where f_i is the number of observation at each of the i unique values of x, n is total number of observation. This mean is same as the result of integrating under the K-M survival curve (Helsel, 2005).

Mann-Kendall's method is useful for regression, trend analysis and correlation analysis of water quality with censored data. First, the data are listed in the order by sampling time: $x_1, x_2, x_3, \ldots\ldots, x_n$, x_i is the datum at sampling time i. Then we calculate the sign of the differences between all possible pairs of data (Kendall, 1970).

$$\begin{aligned} \operatorname{sgn}(x_j - x_k) &= 1 \text{ if } x_j - x_k > 0 \\ &= 0 \text{ if } x_j - x_k = 0 \\ &= -1 \text{ if } x_j - x_k < 0 \end{aligned} \qquad (2.20)$$

The sum of all these $\operatorname{sgn}(x_j - x_i)$ is the Mann-Kendall statistics

$$S = \Sigma\, \Sigma \operatorname{sgn}(x_j - x_k) \qquad (2.21)$$

where, $j > k$

Variance of S is calculated by the following equation (Kendall, 1970)

$$VAR(S) = \frac{n(n-1)(2n+5) - \sum t_j(t_j-1)(2t_j+5)}{18} \qquad (2.22)$$

where n = number of data and t_j = the size of the jth tied group.

And then we calculate Z using S and VAR(S)

$$Z = \frac{S-1}{VAR(S)^{1/2}} \quad \text{if } S>0$$

$$= 0 \quad \text{if } S = 0$$

$$= \frac{S+1}{VAR(S)^{1/2}} \quad \text{if } S<0 \qquad (2.23)$$

If Z value is positive, there is an upward trend. If Z value is negative, there is a downward trend. We can reject null hypothesis, H_0 of no trend if the absolute value of Z is greater than $Z_{1-\alpha/2}$ (Gilbert, 1987). Since we may assign same the rank to censored data, censored data problem is reduced to the problem of tied data (Hirsch and Slack, 1984).

The seasonality is a common phenomenon with water quality data (Hirsch *et al.*, 1982). Since water quality parameters often have the pattern with one year period, comparing January data with May data does not give specific information about the existence of a trend (Smith *et al.*, 1982). The Seasonal Kendall test is the Mann-Kendall test generalized for the seasonality problem. In this method, we compute the Mann-Kendall statistics and it's variance, *VAR(S)*, separately for each season with data collected over years. Then we sum these seasonal statistics and calculate *Z* based on the summed *Z* and *VAR(S)* (Gilbert, 1987).

The slope of a trend for data set (x_i,Y_i) is estimated from individual sample slope for each pair of data (Theil, 1950).

$$S_{ij} = \frac{Y_j - Y_i}{x_j - x_i},\ 1 \leq i < j \leq n \tag{2.24}$$

$$\beta = \text{median } \{S_{ij},\ 1 \leq i < j \leq n\} \tag{2.25}$$

where S_{ij}=the individual slope between i th data and j th data, n=number of observations. If there is a significant seasonality in the data, after calculation of individual sample slope for each pair of data separately for each season with data collected over years, we select median slope as the overall slope from all individual sample slopes (Gilbert, 1987).

Mixed Method: Regression on Order Statistics

Regression on Order Statistics is a method to calculate summary statistics with a regression equation on a probability plot (Helsel, 2005). If the data or transformed data are normally distributed, the plotting position, p_i are converted to normal order scores.

$$R_i = F^{-1}(p_i) \tag{2.26}$$

where R_i is normal order score, F^{-1} is the inverse cumulative normal probability distribution and p_i is the plotting position of ith ranked observation. The linear regression model is constructed using normal order scores of observed values (Berthouex and Brown, 1994).

$$Y_i = \beta_1 + \beta_2 R_i + e_i \tag{2.27}$$

where Y_i is observed values, R_i is normal order score, e_i is the deviation for the fitted line and observed values and β_1, β_2 are coefficients. We use R_i of the non-censored portion of the data to estimate β_1, β_2. Helsel and Cohn (1988) showed the equation to calculate p_i, plotting position.

If the *i*th observation is uncensored,

$$p_{e,j} = p_{e,j+1} + \frac{A_j}{A_j + B_j}(1 - p_{e,j+1}) \tag{2.28}$$

$$p(i) = (1 - p_{e,j}) + \frac{p_{e,j} - p_{e,j+1}}{A_j + 1} r_i \tag{2.29}$$

If *i*th observation is censored,

$$pc(i) = \frac{1 - p_{e,j}}{C_j + 1} r'(i) \tag{2.30}$$

where

A_j = the number of uncensored observations between the jth threshold and j+1th threshold.

B_j = the number of observations, uncensored and censored, below the jth threshold.

$p_{e,j}$ = the probability of exceeding the jth threshold

r_i = the rank of the ith observation among the Aj observations above jth detection limit.

$r'(i)$ = the rank of the ith observation among the censored values known only to be less than jth threshold

$p(i)$ = plotting position of ith uncensored observation *pc*(*i*) = plotting position of ith censored observation among the censored values known only to be less than jth threshold.

After calculating each *p*(*i*), we can get a specific R_i value for each detected value (Y_i) and then estimate β_1, β_2 in Equation (2.26) and Equation (2.27). Censored data show interval values but not specific values. A specific value between intervals of each censored observation is estimated by inputting the plotting position of each censored observation into this regression and the summary statistics is calculated using the values estimated for each censored value as well as uncensored observations.

LITERATURE CITED

Ames, D.P., 2002. Bayesian Decision Networks for Watershed Management. Ph.D. Dissertation, Utah State University, Logan, Utah. 255 pp.

Berthouex, P. M. and L.C. Brown, 1994. Statistics for Environmental Engineers. Lewis Publisher, New York, New York, 335 pp.

Borsuk, M.E., R.T. Clemen, L.A. Maguire, and K.H. Reckhow, 2001. Stakeholder Values and Scientific Modeling in the Neuse River Watershed. Group Decision Negotiation 10:355-373.

Borsuk, M.E. and K.H. Reckhow, 2004. Summery Description of the Neuse Estuary Bayesina Ecological Response Network. http://www.ncsu.edu/ncsu/CIL/WRRI/neuseltm.html, *accessed* April, 2004.

Borsuk, M.E., C.A. Stow, and K.H. Rechow, 2003. Integrated Approach to Total Maximum Daily Load Development for Neuse River Estuary Using Bayesian Probability Network Model (Neu-BERN). Journal of Water Resources Planning and Management 129(4):271-282.

Brannam, K., S. Mostaghimi, G. Yagow, T. Dillaha, B. Benham, and R. Zeckoski, 2005. TMDL Case Studies. In Total Maximum Daily Load-Approaches and Challenges, T. Younos (Ed.), pp. 101-165. Penn Well Corporation, Tulsa, Okalahoma.

Brown, L.C. and T.O. Barnwell, 1987. The Enhanced Stream Water Quality Models QUAL2E and QUAL2E-UNCAS, EPA/600 /3-87/007, U.S. Environmental Protection Agency, Environmental Research Laboratory, Athens, Georgia.

Castillo, E., J.M. Gutiérrey, and A.S. Hadi, 1997. Expert Systems and Probabilistic Network Models. Springer-Verlag, New York, New York. 605 pp.

Charles, N. and J.R. Stedinger, 1996. Estimation of Moments and Quantiles Using Censored Data. Water Resource Research 32:1005-1012.

Cole, T.M. and E.M. Buchak. 1995. CE-QUAL-W2: A Two-Dimensional, Laterally Averaged Hydrodynamic and Water Quality Model, Version 2.0, Draft User Manual. Instruction Report EL-95-1. Corps of Engineers Waterways Experiment Station, Vicksburg, Mississippi.

Cusimano, R.F., S. Hood, and J. Lui, 2002. Quality Assurance Project Plan: Lake Whatcom TMDL Study. Washington State Department of Ecology, Olympia Washington, 47 pp.

EMRG, 2005. Little Bear River Watershed Data Viewer, Environmental Management Research Group, Utah Water Research Laboratory, Utah State University, Logan.

Gilbert, R.O., 1987. Statistical Methods for Environmental Pollution Monitoring. Van Nostrand Reinhold, New York, New York, 320 pp.

Gleit, A. 1985. Estimation for Small Normal Data Sets with Detection Limits, Environmental Science and Technology 19(12):1201-1206.

Ha, H. and M.K. Stenstrom, 2003. Identification of land use with water quality data in stormwater using a neural network. Water Research 37(17):4222-4230.

Hartley, H.O., 1958, Maximum Likelihood Estimation from Incomplete Data. Biometrics 14(2):174-194.

Hartley, H.O. and R.R. Hocking, 1971. The Analysis of Incomplete Data. Biometrics, 27(4):783-823.

Heckerman, D., D. Geiger, and D.M. Chikering, 1994. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence, July 29-31, Seattle, Washington. pp. 293-301.

Helsel, D.R., 2005, Nondetects and Data Analysis. John Wiley & Sons, Hoboken, New Jersey, 250 pp.

Helsel D.R. and T.A. Cohn, 1988. Estimation of Descriptive Statistics for Multiply Censored Water Quality Data. Water Resources Research 24(12):1997-2004.

Helsel, D.R. and R.M. Hirsch, 2002. Statistical Methods in Water Resources. U.S. Geological Survey Techniques of Water Resources Investigation, Book 4, 512 pp.

Hirsch, R.M. and J.R. Slack, 1984. Nonparametric Trend Test for Seasonal Data with Serial Dependence, Water Resources Research 20(6):727-732.

Hirsch R.M., J.R. Slack, and R.A. Smith, 1982. Techniques of Trend Analysis for Monthly Water Quality Data, Water Resources Research 18(11):107-121.

Hodgkins, G.A., 2001. Total Phosphorus Loads for Selected Tributaries to Sebago Lake, Maine. Water Resources Investigations Report 01-4003. U.S. Geological Survey, Augusta, Maine.

Hollander, M. and D.A. Wolfe, 1999. Nonparametric Statistical Methods (Second Edition), John Wiley & Sons, New York, New York, 787 pp.

Jensen, F.V., 1996. An Introduction to Bayesian Networks. Springer-Verlag, New York, New York. 178 pp.

Kaplan, E.L., P. Meier, 1958. Nonparametric Estimation from Incomplete Observation. Journal of the American Statistical Association 53(282):457-481.

Kendall, M.G., 1970. Rank Correlation Methods (Fourth Edition), Chales Griffin & Company Limited, London, U.K., 202 pp.

Marcot, B.G., 1998, Selecting Appropriate Statistical Procedures and Asking the Right Questions: A Synthesis. In Statistical Methods for Adaptive Management Studies, V. Sit, and B. Taylor (Ed.), pp. 129-142. British Columbia Ministry of Forests Research Branch, Victoria, British Columbia, Canada.

Marcot, B.G., R.S. Holthausen, M.G. Raphael, M.M. Rowland, and M.J. Wisdiom, 2001. Using Bayesian Belief Networks to Evaluate Fish and Wildlife Population Viability under Land Management Alternatives from an Environmental Impact Statement. Forest Ecology and Management 153:29-42.

Mathews, R.A., 1994. Kamm Creek Watershed Monitoring Project Second Annual Report: Jannuary- December 1995. Wastern Washington University, Institute for Watershed Studies, Bellingham, Washington.

Mathews, R.A., 1995. Kamm Creek Watershed Monitoring Project Third Annual Report: Jannuary- December 1995. Wastern Washington University, Institute for Watershed Studies, Bellingham, Washington.

NARA (National Archives and Records Administration), 2000. Federal Register Vol. 65. No. 135. Part 6 Environmental Protection Agency.

NCDENR (North Carolina Department of Environment and Natural Resources), 2004. Total Maximum Daily Load (TMDL) For Total Phosphorus Final Report- Reberson Creek TMDL, Raleigh, North Carolina. http://h20.enr.state.nc.us/tmdl/documents/reberson, accessed November, 2004.

Nehls G.J. and G.G. Akland, 1973, Procedures for Handling Aerometric Data. Journal of Airpollution Control Association 23:180-184.

Neilson, B.T., D.P. Ames and D.K. Stevens, 2002. Application of Bayesian Decision Networks to Total Maximum Daily Load Analysis. ASAE Conference Proceedings. March 11-13, 2002. Fort Worth, Texas.

Neilson, B.T., D.K. Stevens, and J.S. Horsburgh, 2005, TMDL Development Approaches, In Total Maximum Daily Load: Approaches & Challenges, T. Younos (Ed.), pp. 47-78. Penn Well Corporation, Tulsa, Oklahoma.

Popek, E.P., 2003. The Sample and the Error. In Sampling and Analysis of Environmental Chemical Pollutants, pp. 1-10. Academic Press, San Diego, California.

Preston, S.D., V.J. Bierman, and S.E. Silliman, 1989. An Evaluation of Methods for the Estimation of Tributary Mass Loads. Water Resources Research 25(6):1379-1389.

Reckhow, K.H. 1999. Water Quality Prediction and Probability Netwrok Models. Canadian Journal of Fishery and Aquatic Science 56:1550-1158.

Scavia, D., W.F. Powers, R.P. Canle, and J.L. Moody, 1981. Comparison of First-Order Error Analysis and Monte Carlo Simulation in Time-Dependent Lake Eutrophication Models. Water Resources Research 17:1051-1059.

Shumway R.H., A.S. Azari, and P. Johnson, 1989. Estimating Mean Concentrations Under Transformation for Environmental Data with Detection Limits, Technometrics 31(3):347-356.

Shumway, R.H., R.S., Azari, and M. Kayhanian, 2002. Statistical Approaches to Estimating mena water quality concentrations with detection limits, Environmental Science and Technology 36(15):3345-3353.

Smith, E.P., K. Ye, C. Hughes, and L. Shabman, 2001. Statistical Assessment of Violations of Water Quality Standards under Section 303(d) of the Clean Water Act, Environmental Science and Technology 35(3):606-612.

Smith, R.A., R.M. Hirsch, and J.R. Slack, 1982. A study of Trends in Total Phosphorus Measurements at NASQAN Station, United States Geological Survey Water-Supply Paper 2190. United States Government Printing Office, Alexandria, Virginia, 34 pp.

Stevens, D.K., 2004. Weber Basin Data Viewer, Utah Water Research Laboratory, Utah State University, Logan, Utah.

Stow, C.A., C. Roessler, M.E. Borsuk, J.D. Bowen, and K.H. Reckhow, 2003. Comparison of Estuarine Water Quality Models for Total Maximum Daily Load Development in Neuse River Estuary. Journal of Water Resources Planning and Management 129(4):307-314.

Slymen, D.J., A.D. Peyster, and R.R., Donohoe, 1994. Hypothesis Testing with values below Detection Limit in Environmental Studies, Environmental Science and Technology 28(5):898-902.

Theil, H., 1950. A Rank-Invariant Method of Linear and Polynomial Regression Analysis,3 Proceedings of Koninalijke Nederlandse Akademie Can Wethenschatpen Series A, 53:1397-1412.

Ticehurst J.L., R.A. Letcher, and A.J. Jakeman, 2007. A Bayesian Network Approach for Assessing the Sustainability of Coastal Lakes in New South Wales, Australia. Environemnatal Modeling and Software 22:1129-1139.

USEPA, 1991, Guidance for Water Quality-Based Decision: TMDL Process. EPA/440/4-91-001. Assessment and Watershed Protection Division.

USEPA, 1997. Technical Guidance Manual for Developing Total Maximum Daily Loads, EPA 823/B-97-002. U.S. Environmental Protection Agency, Office of Water, Washington D.C.

USEPA, 2000. Ambient Water Quality Criteria Recommendations: Information Supporting the Development of State and Tribal Nutrient Criteria. Rivers and Streams in Nutrient ecoregion 14. EPA-822-B-00-022. Office of Water. Washinton D.C.

USEPA, 2001. PLOAD version 3.0: An ArcView GIS Tool to Calculate Nonpoint Sources of Pollution in Watershed and Stormwater Projects- User’s Manual, Washington, D.C.

USEPA, 2005. STORET (Storage and Retrieval), http://www.epa.gov/stormoda/DW_stationcriteria. *accessed* November, 2005.

USGS, 2004. Water Resources of Utah, http://waterdata.usgs.gov/ut/nwis/sw. *accessed* May 2004.

USU (Utah State University), 2001. WRIA 1 Surface Water Quality Data Collection and Assessment: Phase 2 Summary Report (Final Draft).

Utah DEQ (Utah Department of Environmental Quality), 2000. Little Bear River Watershed TMDL.

UWRL (Utah Water Research Laboratory), 1998. Weber and Ogden River Basin Water Quality Management Study: Phase 1,2 Report-Volume 1.

UWRL, (Utah Water Research Laboratory), 2000. Technical support for Watershed Water Quality Evaluation for TMDL support for the East Canyon Creek/East Canyon Reservoir system, Summit Coumt,UT , 71 pp.

Varis, O. and S.F. Jussila, 2002. Water Resources Development in the Lower Senegal River Basin: Conflicting Interests, Environmental Concerns and Policy Options, International Journal of Water Resources Development 18:245-260.

Ware, J. H. and D.L. Demets, 1976. Reanalysis of Some Baboon Descent Data. Biometrics 32(2):459-463.

WRRI (Water Resource Research Institute at University of North Carolina), 2004. Neuse Modeling and Monitoring Project: Modeling for Nitrogen TMDL Development. http://www.ncsu.edu/ncsu/CIL/WRRI/neuseltm.html, *accessed* April ,2004.

CHAPTER 3

CHARACTERISTICS OF LITTLE BEAR RIVER WATERSHED

ABSTRACT

The Little Bear River, Cache County, UT, has two head waters, the South Fork and East Fork and three major reservoirs, Porcupine Reservoir, Hyrum Reservoir, and Cutler Reservoir. Data analysis focused on seven parameters, Total Phosphorus (TP), Dissolved Oxygen (DO), pH, turbidity, specific conductance, temperature, and flow, at the head water outlets and inlets of Hyrum and Cutler Reservoirs. The South Fork and East Fork subwatersheds above the study reaches are mainly forest and range land and have good water quality. The stream below the South Fork and East Fork confluence within the study reaches was impaired and agricultural land occupied 41 to 50% of these subwatersheds. In order to handle censored data (below detection limit) for phosphorus concentration, this chapter explains Maximum Likelihood Estimation (MLE), the Kaplan-Meier Method, the Kruska-Walis Method, the Kendall's tau, and the modified seasonal Kendall trend as alternative methods. The difference between medians and standard deviations were significant among different analysis methods in case of TP concentration in the East Fork subwatershed with up to a 71% censored rate. Most parameters were seasonality variable except TP concentration. Most parameters had significant correlations among parameters at the same analysis location and between downstream and upstream locations for the same parameter. Trend analysis showed significant downward trends of TP and dissolved phosphorus (DP) concentrations above

Cutler Reservoir since 1990, the starting point of the Little Bear River Conservation Project.

GENERAL INFORMATION

The Little Bear River (Figure 3.1) is in Cache County, Northern Utah. It flows from southeast to northwest, bounded by mountains and drains to Cutler Reservoir. This watershed has two major headwaters, East Fork and South Fork.

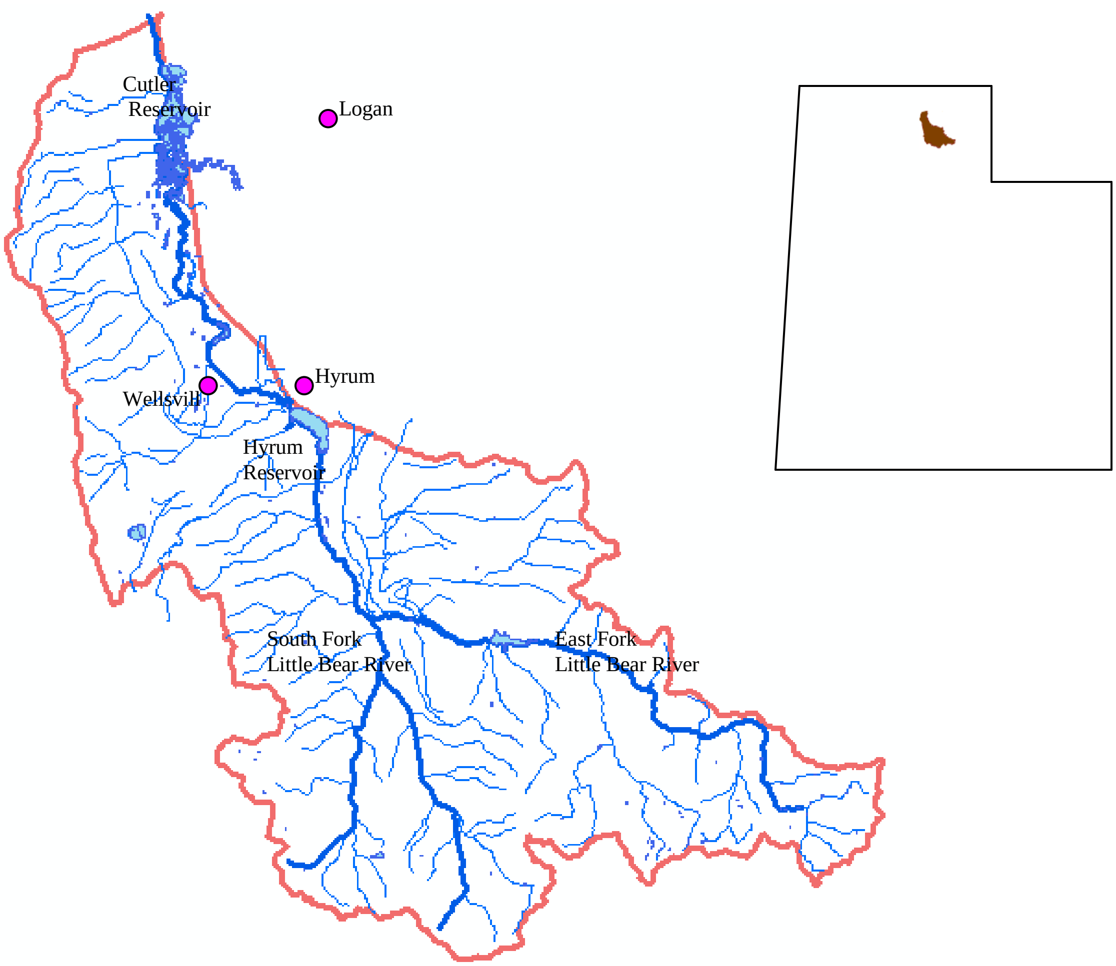


FIGURE 3.1. Little Bear River Watershed Located in Northern Utah.

The Little Bear Watershed has a drainage area of approximately 196,432 acres. Approximately 70% is range and forest, 26% is cropland (Chess, 2000). The range of Elevation is from 4401 ft to 9356 ft (DeLorme, 1993; EMRG, 2005).

The Little Bear River watershed is on a high-priority list of watersheds that are being adversely affected by nonpoint source pollution (Chess, 2000). The Little Bear River Steering Committee found cropland and pastures were significant sources of nutrients in the Little Bear River watershed (Chess, 2000). A Total Maximum Daily Load (TMDL) for the Little Bear River targets reduction of phosphorus. In 1990, USDA started the funding to establish a hydrologic unit area (HUA) planning effort to reduce non-point source pollution in the Little Bear River watershed (Chess, 2000). The Little Bear River Conservation Project included activities to reduce erosion and sediment loading such as stream channel and bank restoration as well as activities to reduce nutrient loading such as grazing land improvements and animal waste management systems (Chess, 2000).

SUBWATERSHEDS

The Little Bear River watershed has two main headwater drainages (East Fork and South Fork) and two main reservoirs (Hyrum Reservoir and Cutler Reservoir). The entire watershed was divided in four subwatersheds, East Fork, South Fork, Above Hyrum Reservoir, and Above Cutler Reservoir, by these main headwater drainages and main reservoirs to show the characteristics of flow and water quality. Water quality station data for each subwatershed outlet were used to identify flow and water quality characteristics for each subwatershed because water from all tributaries in the

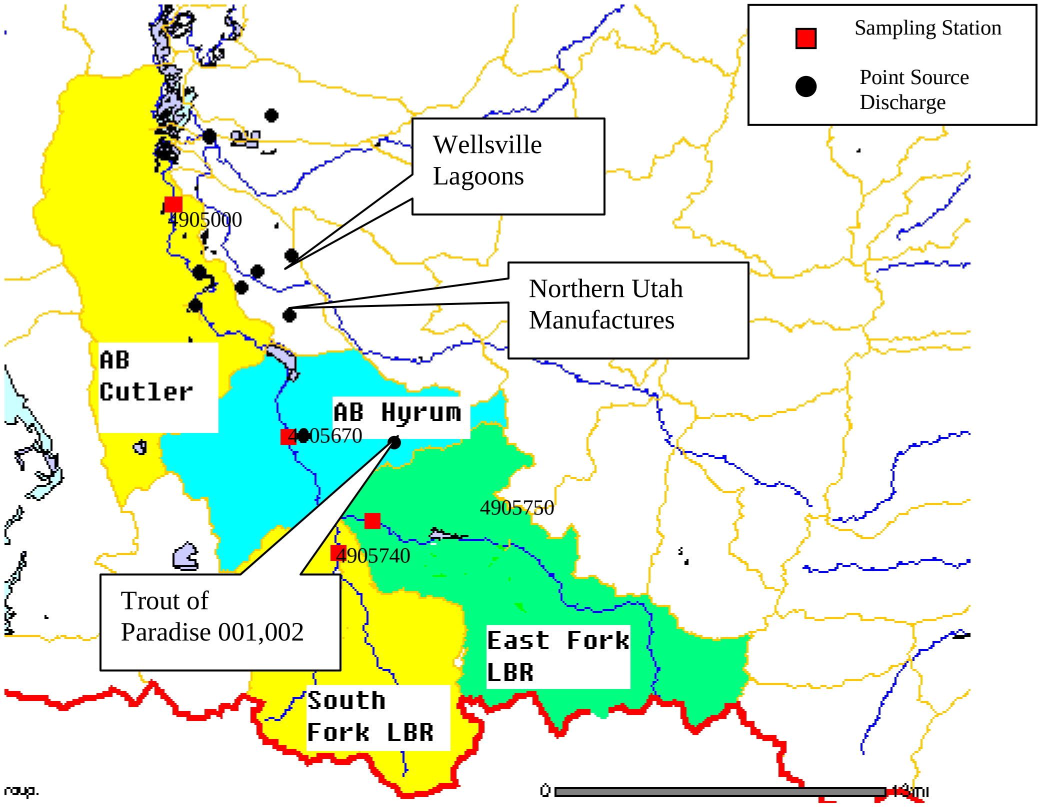


FIGURE 3.2. Sampling Location and Point Load Discharge at Subwatersheds in Little Bear River.

subwatershed comes together to the outlet point (Figure 3.2). Location 4905670 has more water quality data than any location closer to inlet to Hyrum Reservoir. For this reason, location 4905670 was main sampling station for the analysis. Other locations, location 4905000, 4905740 and 4905750 were the closest water quality stations to the outlet of the subwatersheds. Flow and water quality data used in this chapter were obtained by Utah Department Water Quality (Utah DEQ) from 1976 to 2004 and stored in USEPA STORET (USEPA, 2005).

There are no permitted point source discharges for compliance in East Fork or South Fork subwatersheds (Figure 3.2). The subwatershed above Hyrum had two permitted discharges, Trout of Paradise 001 and Trout of Paradise 002. The subwatershed above Cutler has two permitted discharges, Wellsville Lagoons and Northern Utah Manufacture, and there are three more point source discharges to Spring Creek, which meets the Little Bear River above the Cutler Reservoir.

While over 98% of East Fork and South Fork subwatersheds are forest and rangeland, subwatershed above the Cutler Reservoir has large amounts of agricultural and residential areas (Figure 3.3). The irrigated agricultural area is bigger than the non-

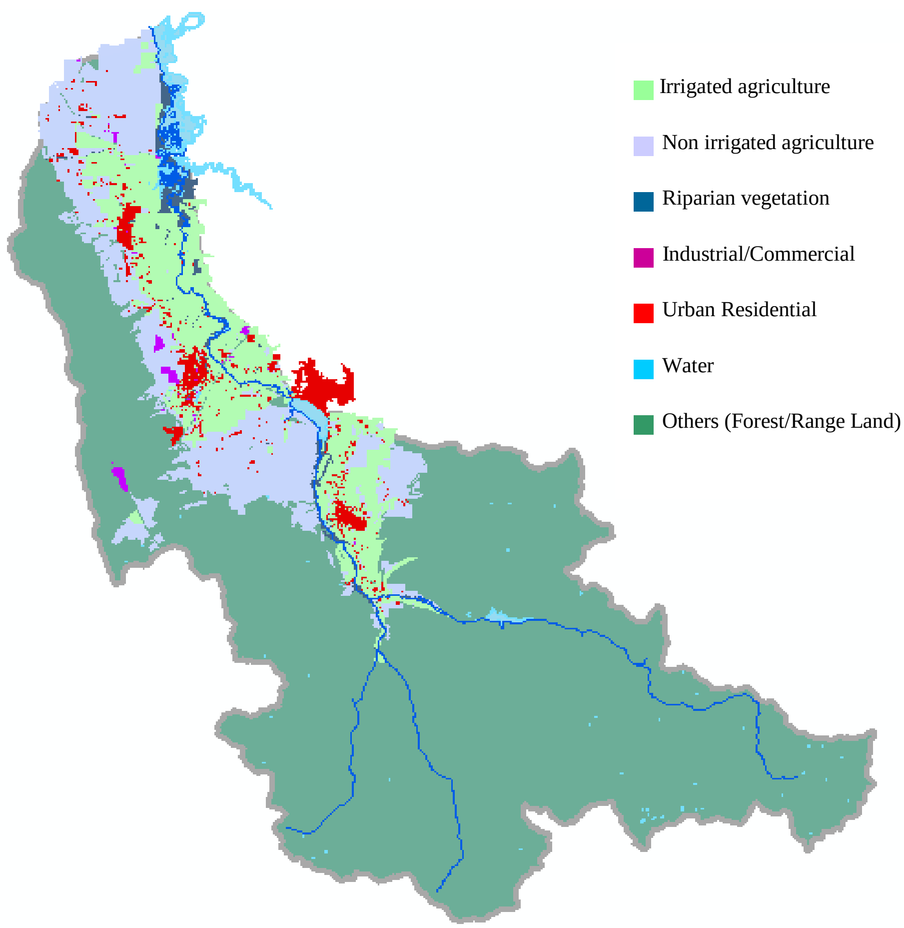


FIGURE. 3.3. Land Use in Little Bear River Watershed (Utah DNR, 2004).

irrigated agricultural area in above the Cutler Reservoir subwatershed while the irrigated agricultural area is smaller than the non irrigated agricultural area in the above Hyrum Reservoir subwatershed (Table 3.1). The above Cutler Reservoir subwatershed has especially large amounts of irrigated pasture (Table 3.1).

DATA ANALYSIS METHODS

Total phosphorus (TP) is the target pollutant identified in the TMDL process for the Little Bear River Watershed. The data analysis studied flow, turbidity, dissolved oxygen (DO), specific conductance (SC), temperature and pH as well as TP concentration because these parameters are associated to the phosphorus cycle in freshwater ecosystem.

TABLE 3.1. Land Use of Above Hyrum Reservoir and Above Cutler Reservoir Subwatersheds (Utah DNR, 2004).

	Above Hyrum Reservoir	**Above Cutler Reservoir**
Urban (Industrial, commercial and urban residential Area)	2.4 %	5.8 %
Others (Forest, Range Land)	56.3 %	43.8 %
Agriculture	41.3 %	50.4%
Irrigation	(41.8 %)	(61.2 %)
Grain	(10.7 %)	(8.9 %)
Alfalfa	(20.8 %)	(24.3 %)
Grass Hay	(2.9 %)	(5.7 %)
Pasture	(6.2 %)	(20.1 %)
Others	(1.2 %)	(2.2 %)
Non Irrigation	(58.2 %)	(38.8 %)
Dry Grain Beans/Sees	(31.2 %)	(21.6 %)
Dry Alfalfa	(11.5 %)	(7.7 %)
Dry Pasture	(9.1 %)	(3.9 %)
Dry Idle	(5.3 %)	(3.3 %)
Other	(1.3 %)	(2.3 %)

( ) is area percentage of agricultural land.

Primary production is limited by phosphorus in many aquatic habitats. Excessive phosphorus may cause algal growth (Dodds, 2002). The algal growth rate is associated with water temperature (Chapra, 1997). Algal growth may cause not only low oxygen and high turbidity but also high pH (Kann and Smith, 1999; Dodds, 2002). Phosphorus exists in natural water in three main forms, phosphate, dissolved organic phosphorus and particulate phosphorus (Dodds, 2002).

Data analysis included statistical summaries, the correlation among parameters, seasonal patterns and trends of each parameter at sampling locations of the four subwatersheds. Data analysis also included correlations among sampling locations. One challenge of data analysis is handling censored data for TP concentration in statistical procedures. When the concentration is reported as “not detected” or “below the analytical method detection limit,” these data are called censored data (Berthouex and Brown, 1994). The detection limit of the analytical method for TP concentration (USEPA method 365.1) is 0.01 mg/l (USEPA, 1993) but the practical quantitation limit is 0.02 mg/L. All of TP concentration below 0.02 mg/L are marked as ‘non-detect’ in USEPA STORET (USEPA, 2005).

The most common method for analysis of censored data is substitution of one-half the detection limits in environmental studies (Nehls and Akland, 1973) but some literature shows better methods (Gleit, 1985; Helsel and Cohn, 1988). This chapter shows how to apply modified data analysis methods to water quality data sets with censored data.

Statistical Summaries

Statistical summaries include mean, variance and median or other percentiles. We used one-half the detection limits and two alternative methods, Maximum Likelihood estimation (MLE), the Kaplan-Meier Method (K-M method) for censored data because MLE or K-M methods is recommended for summaries of over 50 observations with 50-80 % censoring (MLE) or less than 50% censoring (K-M method) (Helsel, 2005). The data analysis by ignoring all data below quantitation limit (BQL) are used to evaluate how much bias ignoring censored data can cause in statistical summaries.

Maximum Likelihood estimation (MLE) is a parametric method, which has some distribution type assumption such as normal or lognormal. We find the parameter values (mean and variance) at which the likelihood function is the maximum. The lognormal distribution has been shown to be an acceptable assumption for low river flows (Vogel and Kroll, 1989) and many types of water quality data (Gilliom and Helsel, 1986). Considering an ordered censored data set $X_1 \le X_2 \ldots\ldots \le X_C \le X_{C+1} \le X_n$, where the first C observations are censored, the likelihood function for the data lognormally and independently distributed (Cohen, 1991) is

$$L = \frac{n!}{c!(n-c)!}\left[\Phi\left(\frac{T-\mu_Y}{\sigma_Y}\right)\right]^c \prod_{i=c+1}^{n} \frac{1}{\sigma_Y}\phi\left(\frac{Y_i-\mu_Y}{\sigma Y}\right) \tag{3.1}$$

where Φ is distribution function (cumulative probability function) of standard normal variate, ϕ is density function of standard normal variate, Y_i is $\ln(X_i)$, T is the log of the measurement threshold, μ_Y is the mean of log-transformed data and σ_Y is the standard deviation of the log transformed data. One can get the best fit μ_Y and σ_Y by taking the logarithm of Equation (3.1) and setting the partial derivatives with respect to μ_Y and σ_Y to zero (Cohen, 1991). Following are traditional formulas recommended for re-conversion of μ_Y and σ_Y (Aitchison and Brown, 1957; Gilbert, 1987).

$$\mu_X = \exp\left(\mu_Y + \frac{\sigma_Y{}^2}{2}\right) \tag{3.2}$$

$$\sigma_X{}^2 = \mu_X \left[\exp(\sigma_Y{}^2) - 1\right]^{1/2} \tag{3.3}$$

The percentile of original scale is obtained by (Helsel, 2005)

$$p_k = \exp(\mu_Y + z_k \sigma_Y) \tag{3.4}$$

where p_k is the *k*th percentile value on the original scale, and z_k is the *k*th percentile of a standard normal distribution.

The K-M method is a nonparametric method, which produces a survival function. Because survival analysis is related to right censored data, flipping is required to transfer left censored data (concentration) to right censored data (Flipped data).

$$Flip_i = M_i - x_i \tag{3.5}$$

where, M_i= Flipping Constant, x_i=observation data (concentration), $Flip_i$=Flipped value. After flipping, the Equation (3.6) is used to estimate survival probability (Kaplan and Meier, 1958).

$$p(t_i) = \Pi\left(1 - \frac{d_i}{n_i}\right) \tag{3.6}$$

where t_i is flipped concentration, n_i is number of flipped observations which are not smaller than t_i, or are censored with flipped detection limit equal to or larger than t_i

(number at risk), d_i is the number of detects with t_i and $p(t_i)$ is survival function probability at t_i, which is the probability equal to or grater than *ti*. Because the survival function estimates the probability of each flipped value, percentiles are obtained directly on the survival curve which is survival probability vs. flipped value. The standard error of survival function is obtained by (Collett, 2003)

$$\text{s.e.}[p] = p \bullet \sqrt{\sum_j^k \frac{d_j}{n_j(n_j - d_j)}} \tag{3.7}$$

where s.e.[p] is the standard error of the survival function, n_j is the number at risk and d_j is number of detects at each of the k values for detected observations. The mean is obtained by

$$\mu = \sum \frac{f_i}{n} x_i \tag{3.8}$$

where is the mean, f_i is the number of observation at each of the i unique values of x, n is the total number of observations. This mean is found by integrating the area under the K-M survival curve (Helsel, 2005). The standard deviation is obtained by

$$\text{s.d.} = \text{s.e.[mean]} \times \sqrt{n} \tag{3.9}$$

where s.e.[mean] is standard error of mean and n is the sample size. For statistical summaries involved for TP concentration, Minitab® version 14 (Minitab Inc., 2006) is used to handle censored data. For other statistical summaries, Little Bear River Data Viewer (EMRG, 2005) was used. Data from 1990 to 2004 from STORET were used for the sampling locations 4905000 (Above Cutler), 4905670 (Above Hyrum), 4905740 (South Fork), and 4905750 (East Fork).

Comparisons for Seasonality

The comparisons among seasonal water parameter data show seasonality at each location. In this dissertation, January, February, and March were defined as winter, and April, May, and June were defined as spring, July, August, and September were defined as summer, and October, November, and December were defined as fall. Analysis of Variance (ANOVA) is the conventional method for comparisons but some cases violate the distributional assumption of ANOVA. In order to apply ANOVA to evaluation of the hypothesis of $\mu_1 = \mu_2 = \mu_3 = \mu_4$, the variance of each group should be the same ($\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4$) and each of the k samples should come from a normal population. Also, in order to apply a two-sample t-test to evaluation of the hypothesis of $\mu_1 = \mu_2$, both samples should come at random from normal populations with equal variances (Zar, 1999). These assumptions may be invalid in real water resources data in many cases. In this chapter, the Kruskal-Wallis test (Kruskal and Wallis, 1952), a nonparametric method, is used to compare groups (seasons) for each water parameter because this Kruskal-Wallis test is free from distribution assumptions of the same variance of all comparison groups. The Kruskal-Wallis test statistic, H is

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1) \tag{3.10}$$

where n_i is the number of observations in group i (season i), N is the total number of observations in all k groups (4 seasons), and R_i is the sum of the ranks of the n_i

observations in group *i*. If the calculated H value is greater than the H value for α= 0.05, the null hypothesis of no seasonality is rejected. *H* values for a specific α are in *H* distribution table for five or less groups but *H* may be considered approximated by χ^2 with *k*-1 degrees of freedom (Zar, 1999).

Using this method, all censored data are replaced by 0.01 to assign tie ranks to censored data for TP concentration. Minitab® 14 executed the Kruskal-Wallis test for comparisons among water parameter values of four seasons and comparisons between water parameter values of two seasons. Data from 1990 to 2004 from STORET were used for the sampling locations 4905000 (Above Cutler), 4905670 (Above Hyrum), 4905740 (South Fork), and 4905750 (East Fork).

Correlations

Because total phosphorus is a grab sampling parameter, it is not easy to measure TP concentration continuously with high frequency. Flow, turbidity, dissolved oxygen, specific conductance, temperature and pH may be measured continuously by on-line equipment. These parameters are associated with the phosphorus cycle in fresh water ecosystems (Dodds, 2002). Therefore, the purpose of correlation is to show the possibility of replacing TP sampling and analysis with continuous measurement of other parameters.

While the most traditional method of correlation between two parameters is Pearson's r, it is not appropriate for censored data (Helsel, 2005). The MLE (Allison, 1995), Kendall's tau-b (Kendall, 1970) and Spearman's rho (Spearman, 1904) methods were used for correlation between TP concentration and other parameters including flow, turbidity, dissolved oxygen (DO), specific conductance, temperature and pH for censored TP concentrations (Helsel, 2005). The *p*-values from those methods were compared to *p*-values from Pearson's methods. Because censored data are expressed by the range of values or an indicator (e.g. 0 means censored data 1 means detected data), those are not accessible to Pearson's correlation method. Replacing censored data with the half value of censored level is recommended by USEPA (USEPA, 1998). Censored TP concentrations were replaced with 0.01 mg/L (the half value of censored level) when the Pearson's r was calculated. There are two categories in correlation analysis, linear correlation and nonlinear monotonic correlation. When a dependent variable generally increases or decreases as the independent variable increases, the two variables are said to possess a monotonic correlation. This correlation is not always linear. Therefore, we used MLE as a linear correlation method and Kendall's tau and Spearmen's rho based on ranks, as the monotonic non linear correlation methods (Helsel and Hirsch, 2002).

The MLE method uses the likelihood r^2 as the criterion of correlation (Allison, 1995):

$$\text{Likelihood } r^2 = 1 - \exp(-G_0^2/n) \tag{3.11}$$

$$G_0 = -2\text{log-likelihood} = -2*(L_{null} - L_{full\ model}) \tag{3.12}$$

where n is number of paired observations (x,y), L is the log of the likelihood for each situation. MLE works for linear correlation assuming a normal residual distribution. We compare all residuals of data or transformed data above quantitation limits to normal distribution line on a probability plot for standardized residuals to check for distribution assumption violations. When some residuals were out of the 95% confidence interval

(Helsel, 2005), we determined that the data violated the distribution assumption and then we executed Kendall's tau and Spearmen's rho correlation analysis. In the MLE method, we evaluate the null hypothesis of the slope equals 0, just as in ordinary linear regression. When the p-value was equal to or greater than 0.05, we do not reject the null hypothesis. In this case, we cannot say " There is a significant linear correlation between two parameters."

Kandall's tau-b is a nonparametric and nonlinear correlation method described by (Kendall, 1970)

$$\tau_b = \frac{(N_c - N_d)}{\sqrt{\left[\frac{N(N-1)}{2} - T_X\right]\left[\frac{N(N-1)}{2} - T_Y\right]}} \tag{3.13}$$

where *Nc* is the number of concordant data pairs, *Nd* is the number of discordant data pairs, T_X is the number of ties in the *X* variable comparisons and T_Y is the number if ties in the *Y* variable comparisons. The significance of τ_b is tested by

$$S = N_c - N_d \tag{3.14}$$

$$VAR(S) = \frac{n(n-1)(2n+5) - \sum t_j(t_j-1)(2t_j+5)}{18} \tag{3.15}$$

where n = number of observations, m = the number of tied groups and t_j = the size of the jth tied group (Kendall, 1970).

We then calculate Z using S and $VAR(S)$

$$Z = \frac{S-1}{VAR(S)^{1/2}} \quad \text{if } S>0$$

$= 0$ if S = 0

$$= \frac{S+1}{VAR(S)^{1/2}} \quad \text{if } S < 0 \tag{3.16}$$

If the Z value is positive, there is an positive correlation. If Z value is negative, there is a negative correlation. We can reject the null hypothesis, H_0 of no correlation if $|Z| > Z_{1-a/2}$, where $Z_{1-a/2}$ is found in Gilbert (Gilbert, 1987).

Spearman's rho is simply the classical Pearson's r applied to the ranks instead of the actual observations (Hollander and Wolf, 1999). The Spearman's rho is calculated by Equation (3.17) (Spearman, 1904).

$$\rho = \frac{12\sum_{i=1}^{n}\left\{\left[R_i - \frac{n+1}{2}\right]\left[S_i - \frac{n+1}{2}\right]\right\}}{n(n^2-1)} \tag{3.17}$$

where R_i is the rank of ith independent variable (X_i), and S_i is the rank of ith dependent variable (Y_i).

In Kendall's tau and Spearmen rho, a tie rank is assigned to "all less than a specific threshold" and the tie rank is lower than any detected value at or above the threshold (Hirsch and Slack, 1984). Data from 1990 to 2004 at STORET were used for each location.

Trends Over Time

We evaluated the time trends for TP, Dissolved Phosphorus (DP), Flow and turbidity at sampling locations 4905000 (Above Cutler), 4905670 (Above Hyrum), 4905740 (South Fork) and 4905750 (East Fork). The nonparametric seasonal Kendall test is used as trend method because this method can handle trends in water resources data with seasonal cycles (Hirsch *et al*., 1982). In this method, the "less than a specific threshold" values are considered to be smaller than any numerical value equal to or greater than the specific threshold. A tied rank is assigned to all "less than" values. We compute the Mann-Kendall statistic and its variance, VAR(S), separately for each season with data collected over years (Equations 3.18 and 3.19) and then summed S and VAR(S) for each season to get S' and VAR(S') by Equations 3.19 and 3.20 (Gilbert, 1987).

$$S_i = \sum_{k=1}^{n_{i-1}} \sum_{j=k+1}^{n_i} \operatorname{sgn}(x_{ij} - x_{ik}) \tag{3.18}$$

$$VAR(S_i) = \frac{n_i(n_i-1)(2n_i+5) - \sum t_{ij}(t_{ij}-1)(2t_{ij}+5)}{18} \tag{3.19}$$

where $j>k$, n_i is number of data for the ith season and the t_{ij} is the size of the jth tied group for the ith season.

$$S' = \sum S_i \tag{3.20}$$

$$\mathrm{Var}[S'] = \sum \mathrm{Var}[S_i] \tag{3.21}$$

Z is calculated using S' in Equation (3.22). If $|Z| > Z_{1-a/2}$, we can reject the null hypothesis H_0 of no trend. $Z_{1-a/2}$ is found in Gilbert (Gilbert, 1987).

$$Z = \frac{S'-1}{VAR(S')^{1/2}} \quad \text{if } S' > 0$$

$$= 0 \quad \text{if } S' = 0$$

$$= \frac{S'+1}{VAR(S')^{1/2}} \quad \text{if } S' < 0 \tag{3.22}$$

One challenge in the trend analysis for water quality data is estimation of slopes with both censored data and seasonal cycles. The Theil-Sen slope is recommended as slope estimation with censored data (Theil, 1950; Helsel, 2005). The seasonal Kendall slope estimator is recommended for data with seasonal cycle (Gilbert, 1987). In this dissertation, two slope estimators are mixed to handle both censored data and seasonal cycle. We calculate slopes for all possible data pairs for each season separately.

$$S_{ijk} = \frac{Y_{ik} - Y_{ij}}{x_{ik} - x_{ij}}, \quad 1 \leq j < k \leq n \qquad (3.23)$$

where S_{ijk} = the individual slope between the the jth data and the kth data for the ith season, Y_{ik} is the k th parameter data for ith season, x_{ik} is the k the sampling period for the ith season and n is number of data for the ith season.

For censored data, the slopes may be interval values between 0 and a specific threshold (Helsel, 2005). After calculation of the slopes for each season, we made an interval table for all slopes from each season including the start and end points. When no censored data are involved, start point is equal to the end point. When any censored data are involved, start point may not be equal to the end point. We then estimate the median slope of all slopes by Turnbull's method, which can generate survival probabilities for interval data (Turnbull, 1976).

The Nooksack trend analysis program (Stevens, 2006) executed the seasonal Kendall Test based on a FORTRAN implementation by Gilbert (1987). Minitab® 14 was used to estimate the slope (Minitab Inc, 2007).

The Little Bear River Conservation Practice (LBRCP) began in 1990 and the trend was calculated for each parameter before LBRCP, from 1976 to 1989, and after starting LBRCP, from 1990 to 2004, separately (Helsel and Hirsch, 2002).

RESULTS

Summaries

The statistical summaries were calculated for each parameter comparing values for four sampling locations, East Fork, South Fork, Above Hyrum Reservoir, and Above Cutler Reservoir. Since sampling frequencies for each parameter among locations are different and there are often gaps in sampling times (Figure 3.4), the statistical summaries

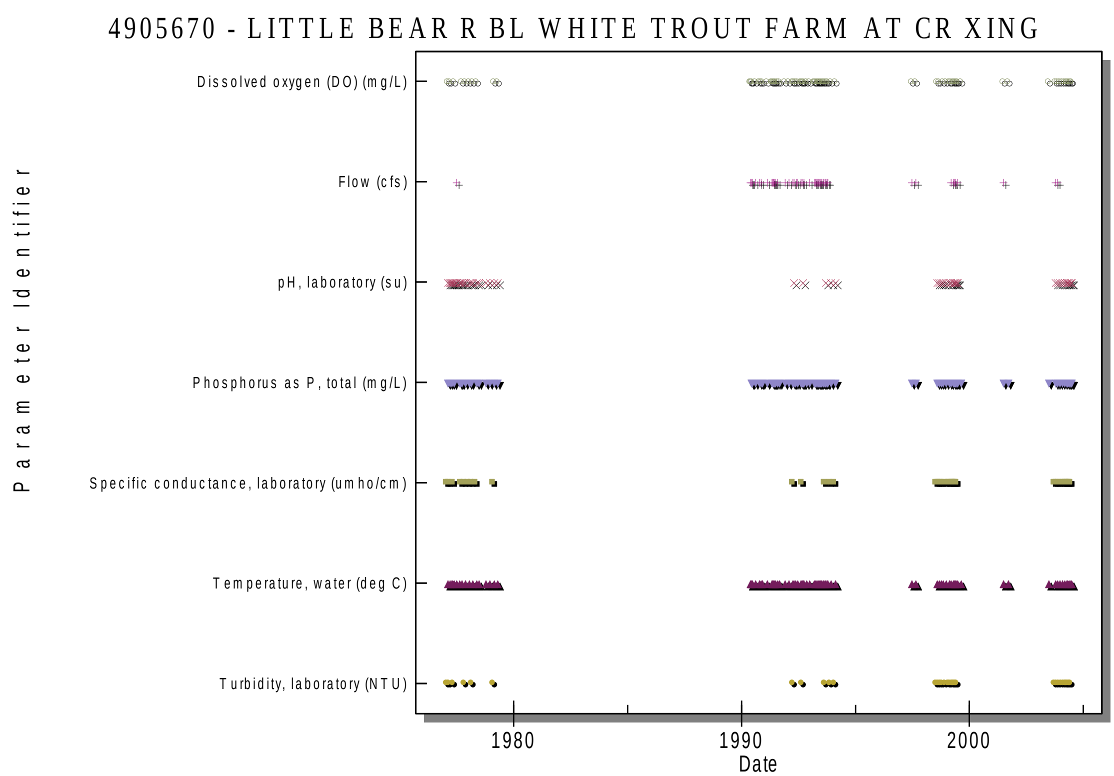


FIGURE 3.4. Sampling Time Distribution at Location 4905670 (EMRG, 2005).

may not be fully representative of water quality and quantity. Statistical summaries of TP concentration required special handling because of censored data. The results from the four approaches, 1) ignoring all data below the quantitation limit, 2) half of the quantitation limit, 3) MLE and 4) the KM methods are found in Appendix A.

The difference among the four analysis approaches for TP concentration is biggest at location 4905750 (Above Hyrum Reservoir), which has the most censored data. The range of each descriptive statistic from three approaches, half of the quantitation limit, MLE and the KM method are provide to help coarse evaluation of water characteristics (Table 3.2). According to the 90th percentiles of TP, the East Fork (4905750) and the South Fork (4905740) are relatively clean but the water below the confluence of the East and South forks is relatively contaminated. In location 4905750, the mean of the KM method was significantly larger than the MLE or half of below quantitation limit method, and the standard deviations of the MLE was significantly smaller than other methods (Appendix A). There was no large difference among 25th percentiles (Q1), 75th percentiles (Q3), means, medians and standard deviation of three methods at location 4905670 (Above Hyrum Reservoir) and location 4905000 (Above Cutler Reservoir) (Table 3.2). The censored rate was 71% at location 4905750, and the high censored rate may cause the large differences among means and deviations of half of below quantitation limit, MLE and the KM methods at location 4905750.

The arithmetic mean, median and geometric mean of specific conductance of location 4905000 (Above Cutler) were higher than those of the East Fork, the South Fork or Above Hyrum Reservoir (Table 3.3). This means the concentration of dissolved matter at location 4905000 was higher than at other locations. The median and geometric mean

TABLE 3.2. The Range of Summary Statistics for TP Concentration from Three Differences Methods. (Detail statistics are in Appendix A).

	# of Observa-tion	# of censored data	Range of Year	Mean	Median	Standard Deviation	Q1	Q3	90 %
4905740	68	26	1990-2004	0.0333-0.0371	0.0247-0.0250	0.0303-0.0327	BDL-0.0144	0.038-0.0423	0.068-0.069
4905750	75	53	1990-2004	0.0190-0.0296	BDL-0.0111	0.0262-0.0449	BDL-0.010	0.0223-0.0230	0.037-0.042
4905670	79	9	1990-2004	0.0582-0.0594	0.0450-0.0457	0.0465-0.0475	0.027-0.0284	0.0736-0.078	0.107-0.113
*4905000	145	3	1990-2004	0.1192-0.1194	0.1000-0.1000	0.1673-0.1676	0.0540-0.0560	0.1425-0.1410	0.176-0.182

* reject the result by MLE for distribution assumption violation.

TABLE 3.3 Specific Conductance Summaries. (Unit:umho/cm)

Location	# of Obser-vation	Range of Dates	Mean	Median	Standard Deviation	Geometric Mean
4905740	68	1990-2004	339.2	350.5	109.3	321.5
4905750	83	1990-2004	394.3	396	63.77	388.7
4905670	80	1990-2004	417.6	431.4	93.32	406.4
4905000	155	1990-2004	550	564	127.8	537.1

* Number of censored data is 0 at all four locations

TABLE 3.4. Turbidity Summaries. (unit: NTU)

Location	# of Obser-vation	Range of Dates	Mean	Median	Standard Deviation	Geometric Mean
4905740	28	1993-2004	10.67	3.645	19.12	4.0
4905750	26	1998-2004	21.8	1.85	63.96	3.0
4905670	29	1992-2004	15.04	4.5	31.29	4.9
4905000	132	1990-2004	13.4	11.85	9.23	10.5

* Number of censored data is 0 at all four locations

of turbidity of location 4905000 were higher than those of the East Fork, the South Fork or Above Hyrum Reservoir (Table 3.4). This means the concentration of particulate matter at location 4905000 was higher than at other locations.

Seasonality

The null hypothesis for the Kruskal-Walls test is that a parameter value is the same in all four seasons. The criterion to reject null hypothesis is $\alpha = 0.05$. Location 4905000 (Above Cutler) has significant seasonality for all seven parameters including TP, flow, DO, SC, temperature, pH, and turbidity.

At location 4905750 (East Fork), there were significant differences among the four seasons in DO, flow, SC, and temperature. At location 4905670 (Above Hyrum Reservoir) and 4905740 (South Fork), there were significant differences among four seasons in DO, flow, SC, temperature and turbidity (Table 3.5). The most interest thing is no significant seasonality of TP concentration at location 4905670, location 4905740 and location 4905750 (Figure 3.5 a; Table 3.5). For location 4905000, the TP in summer was significant higher than spring or winter according to the results of Kruskal-Walls test between summer and spring, and between summer and fall (Figure 3.5 b).

Correlations

The purpose of correlation is to show the possibility of replacing infrequent TP sampling and analysis with continuous measurement of other parameters. MLE was used to calculate correlation between TP and other parameters under the lognormal distributional assumption (except correlation between TP concentration and specific

TABLE 3.5. The results of Kruskal-Wallis Test for Seasonality (H-value, ( ) is p-value, /Number of Winter, Spring, Summer, Fall Observations).

Location / Parameter	4905740	4905750	4905670	4905000
TP	7.05 (0.070) /13,28,15,12	6.71 (0.082) /17,31,16,11	3.89 (0.273) /13,32,24,10	11.5 (0.009) /38,55,27,25
DO	22.15 (<0.001) /13,29,14,12	28.71 (<0.001) /18,35,19,12	17.29 (<0.001) /13,33,22,10	45.33 (<0.001) /38,57,31,28
Flow	12.66 (0.005) /12,28,13,11	11.38 (0.01) /16,35,23,12	20.98 (<0.001) /8,25,16,4	23.54 (<0.001) /26,37,28,21
pH	0.73 (0.866) /13,29,15,12	1.59 (0.661) /18,35,20,12	1.43(0.699) /13,33,23,10	11.47 (0.009) /39,57,32,28
Specific conductance	32.11 (<0.001) /13,29,14,12	21.60 (<0.001) /18,34,19,12	42.32(<0.001) /13,33,23,10	58.55 (<0.001) /39,57,31,28
Temperature	38.97(<0.001) /13,29,15,12	49.78 (<0.001) /18,35,20,12	45.62(<0.001) /13,33,23,10	108.86 (<0.001) /39,57,32,28
Turbidity	15.54(0.001) /6,10,6,6	*3.92 (0.271) /6,10,6,4	11.02(0.012) /7,11,6,5	20.61(<0.001) /33,45,30,24

*Minitab 14 ® (Minitab Inc., 2006) gave small sample note.

conductance at location 4905000) because lognormal distribution was suitable for MLE correlation of TP concentration according to the probability plots of residuals. For example, all standardized Residuals of TP concentration at location 4905740 were inside the 95% confidence interval of a normal line after transformation of TP concentration to a log scale (Figure 3.6). The normal distribution was suitable for correlation between TP concentration and specific conductance at location 4905000.

The criterion to reject the null hypothesis of no correlation was p = 0.05. Kendall’s Tau-b, Spearman’s rho and MLE under the lognormal assumption showed

a) b)

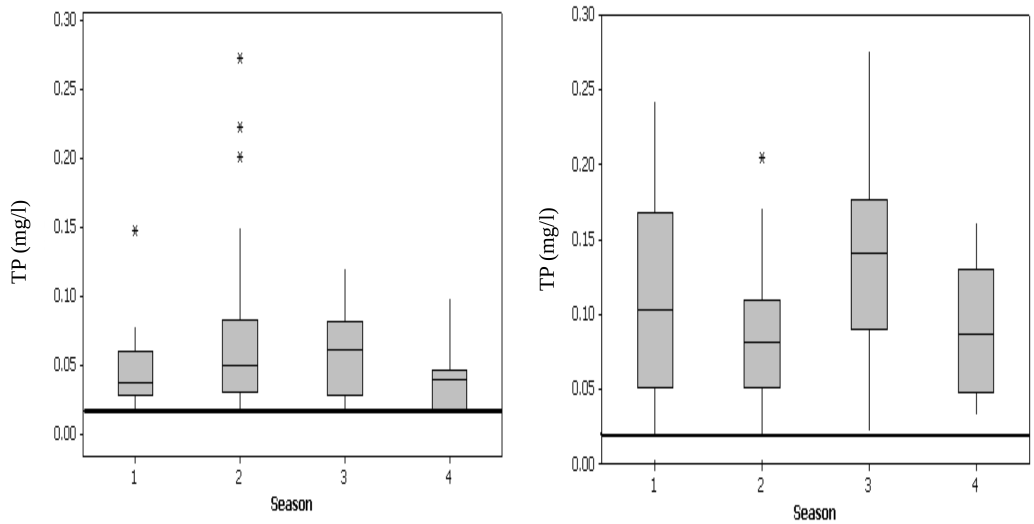


FIGURE 3.5. Seasonal Box Plots for TP Concentration a) at Location 4905670 and b) at Location 4905000. Season 1 is Winter, season 2 is Spring, Season 3 is Summer, and Season 4 is Fall. (▬: Censored Level :0.02 mg/l, * outlier beyond whiskers, box: range from 25 (Q1) to 75 (Q3) Percentiles, Center Line in Boxes: Median, Upper Limit of Whisker : Q3+1.5(Q3-Q1), Lower Limit of Whisker : Q1-1.5(Q3-Q1)).

a) b)

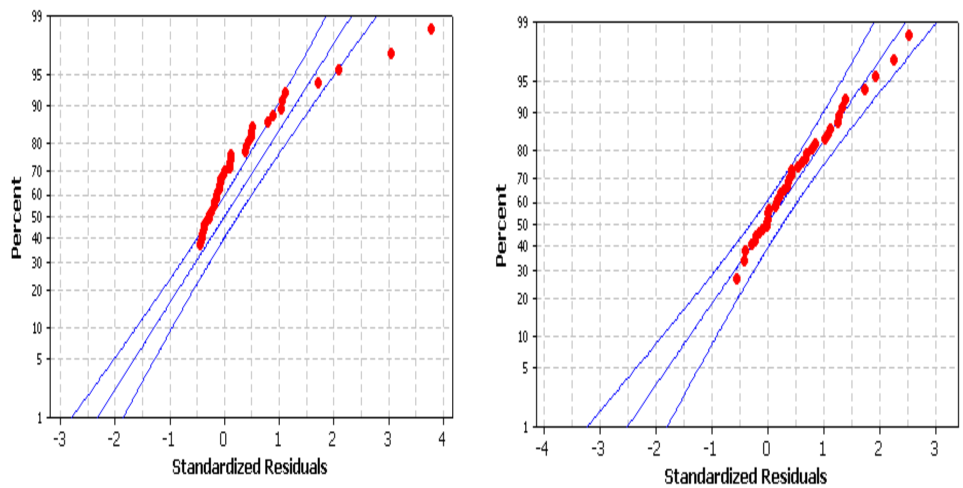


FIGURE 3.6. Probability Plot of Standardized Residuals for MLE Correlations of TP Concentration at Location 4905740, a) Assuming Normal Distribution b) Assuming Lognormal Distribution.

significant correlation between flows and TP concentrations at each sampling station except correlation between flow and TP at location 4905670 (Table 3.6). DO has a significant correlation with TP at locations 4905670 (Above Hyrum Reservoir) and 4905000 (Above Cutler Reservoir) but not at locations 4905740 (South Fork) or 4905750 (East Fork).

Flow had significant correlation to TP in all locations except 4905670 (Above Hyrum). Even though there were significant differences between nonparametric methods and the MLE in correlation between TP and flow in location 4905670, the scatter plot of TP versus Flow showed no significant correlation (Figure 3.7). The correlations between pH and TP concentration at locations 4905740 (South Fork), 4905670 (Above Hyrum) and 4905000 (Above Cutler) were significant, but not significant at location 4905750 (East Fork). Specific Conductance had a significant correlation with TP concentration at 4905000 but not a significant correlation at locations 4905750 and 4905670.

Even though there were differences between the nonparametric and MLE method in correlation between TP and SC at location 4905740 (South Fork), the scatter plot of TP versus specific conductance showed a non linear monotonic correlation (Figure 3.8). There were significant correlations between water temperature and TP concentration at location 4905000 (Above Cutler) and location 4905750 (East Fork) but not a significant correlation at location 4905670 (Above Hyrum) and location 4905740 (South Fork). There was strong correlation between Turbidity and TP concentration at all sampling locations. Linear correlation between TP concentration and other parameters differ from non linear correlation at location 4905000 (Above Cutler) because of two outliers for TP concentration of 0.79 mg/l and 1.88 mg/l. After removing outliers, the linear correlations

TABLE 3.6. Correlation Between TP Concentration and Other Parameters at Each Location.

(p-values)

Location	**Statistic**	**TP vs DO**	**TP vs Flow**	**TP vs pH**	**TP vs S.C.**	**TP vs Temperature**	**TP vs Turbidity**
4905740	No.of pair[1]	25/67	25/63	26/68	26/67	26/68	15/28
	Pearson's r^2	<0.001 (0.985)	0.248 (<0.001)	0.264 (<0.001)	0.038 (0.115)	<0.001 (0.530)	0.808 (<0.001)
	MLE r^2	0.006 (0.25<p< 0.5)	0.177 (<0.001)	0.236 (<0.001)	0.017 (0.25<p <0.5)	0.000 (0.9<p <0.95)	0.487 (<0.001)
	Kendall's τ_b	-0.058 (0.513)	0.187 (0.038)	-0.340 (<0.001)	-0.182 (0.038)	0.026 (0.772)	0.445 (0.002)
	Spearman's ρ	-0.083 (0.505)	0.246 (0.050)	-0.477 (<0.001)	-0.238 (0.052)	0.054 (0.662)	0.648 (<0.001)
4905750	No.of pair[1]	52/74	51/73	53/75	53/74	53/75	19/26
	Pearson's r^2	0.006 (0.514)	0.009 (0.412)	<0.001 (0.999)	0.010 (0.390)	0.014 (0.315)	0.982 (<0.001)
	MLE r^2	0.01 (0.5<p <0.75)	0.424 (0.025<p <0.01)	0.101 (0.05<p <0.1)	0.027 (0.1<p <0.25)	0.215 (0.025<p <0.05)	0.548 (<0.001)
	Kendall's τ_b	-0.004 (0.970)	0.314 (<0.001)	-0.164 (0.064)	-0.097 (0.2758)	-0.182 (0.038)	0.614 (<0.001)
	Spearman's ρ	0.013 (0.912)	0.421 (<0.001)	-0.204 (0.079)	-0.134 (0.256)	-0.244 (0.035)	0.718 (<0.001)
4905670	No.of pair[1]	9/76	5/52	9/77	9/77	9/77	7/29
	Pearson's r^2	0.078 (0.015)	0.205 (0.001)	0.04 (0.081)	0.026 (0.163)	0.013 (0.328)	0.876 (<0.001)
	MLE r^2	0.176 (<0.001)	0.136 (0.005<p <0.01)	0.238 (0.025<p <0.05)	0.012 (0.25<p< 0.5)	0.072 (0.1<p<0 .25)	0.571 (<0.001)
	Kendall's τ_b	-0.226 (0.005)	0.062 (0.522)	-0.157 [2] (0.143) [2]	-0.019 (0.809)	0.151 (0.055)	0.472 (<0.001)
	Spearman's ρ	-0.332 (0.003)	0.103 (0.465)	-0.228 (0.046)	-0.017 (0.880)	0.211 (0.066)	0.557 (0.002)
4905000 [3]	No.of pair[1]	2/139	0/99	3/141	3/141	3/141	1/123
	Pearson's r^2 [4]	0.096 (<0.001)	0.104 (0.001)	0.065 (0.002)	0.173 (<0.001)	0.059 (0.004)	0.121 (<0.001)
	MLE r^2	0.114 (<0.001)	0.095 (0.001<p <0.005)	0.066 (0.001<p <0.005)	0.138 (<0.001)	0.076 (<0.001)	0.149 (<0.001)
	Spearman's ρ	-0.296 (<0.001)	-0.360 (<0.001)	-0.3 (<0.001)	0.48 (<0.001)	0.249 (0.003)	0.421 (<0.001)

Lognormal distributional assumption for TP concentrations was used for MLE. Statistic values and p- values are shaded when p< 0.05. [1] Number of pairs (Censored TP/Total Pairs). [2] 1993-2004. [3] No Kendall's τ_b was obtained because Minitab Macro would not converge (over 12 hours). [4] Outlier for 4905000: TP 0.79 mg/l , 1.88 mg/l.

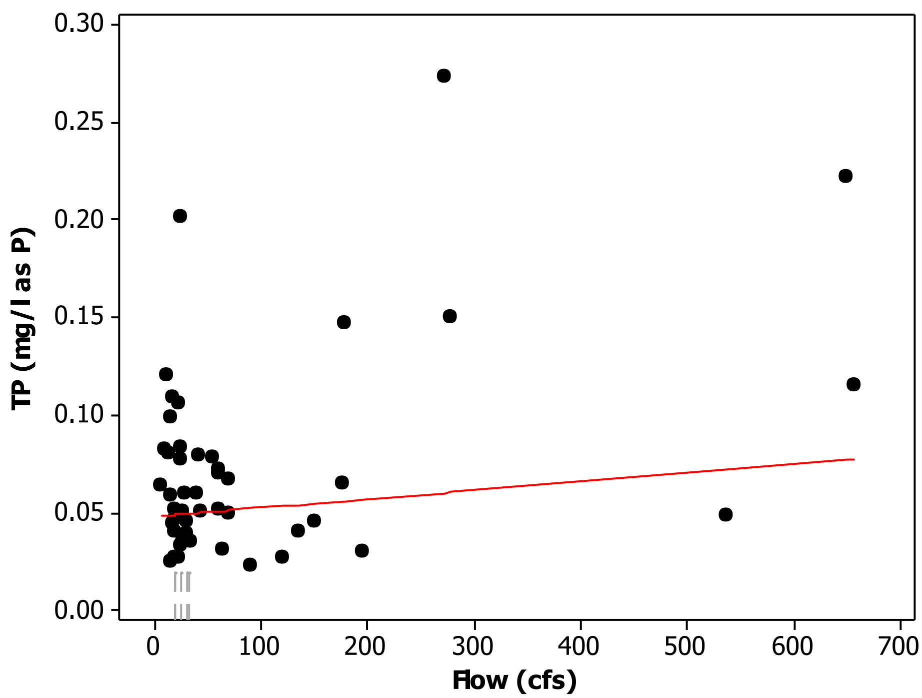


FIGURE 3.7. Scattered Plot between Flow and TP Concentration at Location 4905670. (● : pairs of detected (non censored) flow and TP concentration, --- pairs of detected flow and censored TP concentration).

are similar to non linear correlations at location 4905000 (Table 3.6).

The p-values from the Pearson's correlation coefficient differed from other correlation methods for location 4905750, but there was no significant difference between Pearson's correlation and other methods for locations 4905670 and 4905000. High censored percentage may affect the correlations at location 4905750 (Table 3.6).

Correlations between two locations for TP concentration have censored data not only for the response variable (Y) but also for the predicted variable (X). Kendall's tau-b test can handle these cases called doubled censored. Since Minitab 14 ® did not support

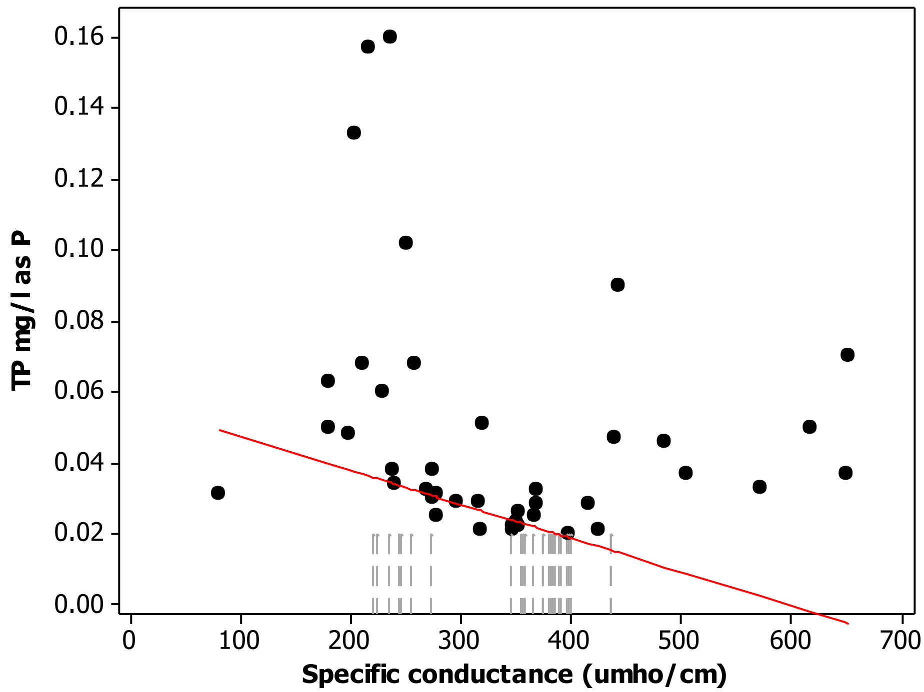


FIGURE 3.8. Scattered Plot between Specific Conductance and TP Concentration at Location 4905740 (● : pairs of detected (non censored) flow and TP concentration, --- pairs of detected flow and censored TP concentration).

MLE for doubled censored correlations, Pearson's r is used for the linear correlations. Half of quantitation limit is used for censored TP concentration using Pearson's r. All linear and nonlinear correlations between two locations' parameter were significant except turbidity between location 4905000 and location 4905670, and specific conductance between location 4905670 and 4905750 (Table 3.7; Appendix B).

Trends

Locations 4905740 (South Fork) and 4905750 (East Fork) have no data before 1990, and these locations have two sampling time gaps of more than 3 years from 1991 to 2004 for TP concentration. Location 4905670 (Above Hyrum) has data before 1990 but

TABLE 3.7. Correlation Between Two Locations for Each Parameter (p-values by Kendall's tau-b test).

Location	Method	TP	Flow	Turbidity	Specific Conductance
4905740 vs 4905670	Pearson's r	0.599 (<0.001)	0.807(<0.001)	0.859(<0.001)	0.85(<0.001)
	Kendall's τ_b	0.477(<0.001)	0.660(<0.001)	0.633(<0.001)	0.741(0.001)
	# of Data Pair	59	40	24	59
4905750 vs 4905670	Pearson's r	*0.548(<0.001)	0.904(<0.001)	*0.813(<0.001)	0.140(0.290)
	Kendall's τ_b	0.267(0.007)	0.444(<0.001)	0.464(0.002)	0.0469(0.605)
	# of Data Pair	59	46	23	59
4905670 vs 4905000	Pearson's r	0.584(0.001)	0.689(0.002)	0.117(0.666)	0.547(0.001)
	Kendall' τ_b	0.393(0.002)	0.607(<0.001)	0.167(0.392)	0.479(0.0002)
	# of Data Pair	31	17	16	31

* Outlier :TP 0.39 mg/l, Turb 325 mg/l at location 4905750.

the number of observations is small, only 23 observations for TP concentration from 1977 to 1991. Location 4905670 has one or two sampling time gaps of more than 3 years from 1991 to 2004 (Figure 3.9). On the other hand, location 4905000 has many observations without significant sampling gaps from 1977 to 2004 (Figure 3.10). Therefore, trends are emphasized for TP at location 4905000.

The result of Seasonal Kendall trends showed no trend in TP concentrations before starting the Little Bear River Conservation Project (1974-1989) but a downward trend in TP concentrations for 15 years after starting the Little Bear River Conservation Project (1990-2004). There was a downward trend in DP concentrations after starting the

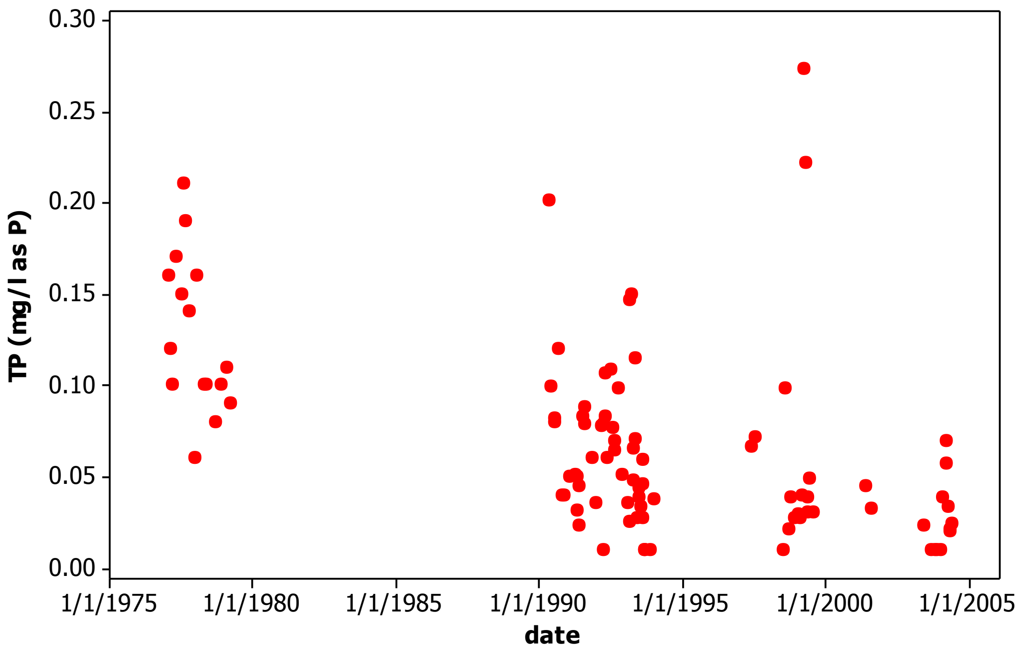


FIGURE 3.9. TP Concentration Time Series at Location 4905670.

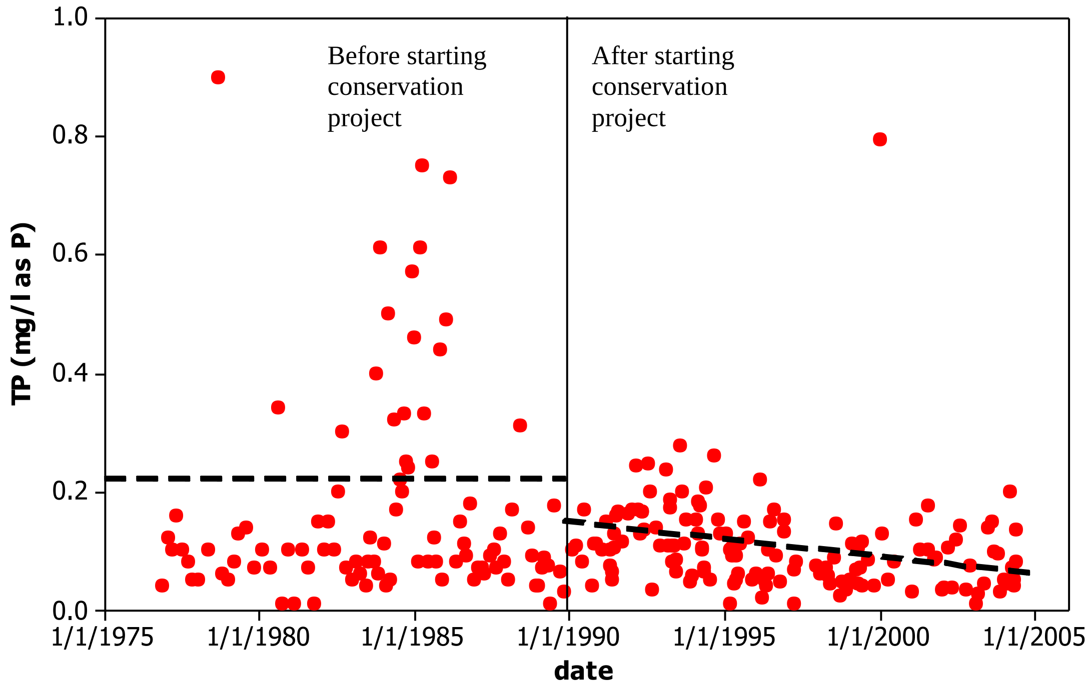


FIGURE 3.10. TP Concentration Time Series at Location 4905000.

Little Bear River Conservation Practice (LBRCP). There was no trend in turbidity before and after starting the LBRCP (Table 3.8). Flow had a downward trend during from 1983 to 1989 before starting the LBRCP.

The TP concentration decrease rate after 1990 was 0.00434 mg/l/yr. The DP concentration decrease rate was 0.00275 mg/l/yr. (Table 3.8). PP concentrations were calculated by subtracting DP concentration from TP concentration. The decrease rate of these calculated PP concentrations was 0.0015 mg/l/yr.

DISCUSSION

The summary statistics showed a big difference between water quality above and below confluence of the South fork and East fork of the Little Bear River. The different land use patterns may cause this difference. Forest and range land mainly occupy the South and East Fork subwatersheds while agricultural land occupied the Above Hyrum Reservoir and Above Cutler Reservoir subwatersheds. Especially, 41.1 % of the agricultural area is irrigated crops and 20.1 % of agricultural area is irrigated pasture in the Above Cutler Reservoir subwatershed.

TABLE 3.8 Trends for Above Cutler Reservoir (Location 4905000) Comparing Before and After Starting Little Bear River Project. (p-value : probability of slope=0)

	1977-1989 (Before Project)			**1990-2004 (After Project)**		
	# of obs.	p-value	Slope (/yr)	# of obs	p-value	Slope(/yr)
TP (mg/l)	95	0.554	-0.0009	144	<0.01	- 0.00434
DP (mg/l)	0	N/A	N/A	138	<0.01	- 0.00275
PP (mg/l)	0	N/A	N/A	125	0.029	-0.00150
Flow (cfs)	*49	*<0.01	*-15.09	111	0.012	3.75
Turbidity (NTU)	91	0.459	- 0.12	131	0.232	-0.174

*1983-1991 data, N/A: Not available.

The TP concentration's summary statistics by the half of detection limit method, MLE and the KM are close to one another at locations 4905000 (Above Cutler), 4905670 (Above Hyrum) and 4905740 (South Fork) (Appendix A). The mean of TP concentrations by the KM method was significantly larger than the mean by half of limit or MLE at location 4905750 (East Fork). The standard deviation of TP concentrations by the MLE method was significantly smaller than the mean by half of limit or KM at location 4905750. The percentage of censored data was 71% for TP concentration at location 4905750. Some literature shows that estimation errors increases dramatically between 60% and 80% censoring (Gilliom and Helsel, 1986; Kroll and Stedinger, 1996; Shumway *et al.*, 2002). The K-M method did not provide median TP concentration at location 4905750. Since over 50% of TP concentrations were censored at this location, the median was "below detection limit" and the nonparametric KM method could not find a specific median TP concentration.

TP concentration had strong linear correlation to turbidity at each location. This correlation shows that easily measured turbidity may be an indicator of TP concentration in the Little Bear River watershed. Turbidity is the parameter which can be measured continuously in the field while TP concentration is a grab sampling parameter. If turbidity is used as an indicator of TP concentration, the sampling time gap problems may be relieved.

All linear and nonlinear parameter correlations between the upstream and downstream sampling locations were significant except turbidity between locations 4905000 (Above Cutler) and 4905670 (Above Hyrum), and specific conductance between locations 4905670 and 4905750 (East Fork). This suggests the possibility that the correlation between downstream and upstream for each parameter may be used to fill in data gaps.

Trends at location 4905000 (Above Cutler Reservoir) showed a significant TP concentration downward trend after starting LBRCP and no significant TP concentration trend before starting LBRCP. If all of conditions are the same between before starting LBRCP and after starting LBRCP at location 4905000 except non point source management, it may be concluded that LBRCP have reduced TP concentration. Analysis for other factors to affect TP concentration is required to conclude that LBRCP have reduced TP concentration. The sampling time gap increased the uncertainty of trends at all locations except location 4905000.

Some literature recommends step trends, a comparison of two non-overlapping sets of data, to avoid the sampling time gap problem (Helsel and Hirsch, 2002). In this method, the U value from the Mann-Whitney test (Zar, 1999) is the criterion to reject the null hypothesis of no difference between data in the “early” and “late” period. There is no specific rule about how long the gap should be to make this method the preferred procedure (Helsel and Hirsch, 2002). When a known event has occurred at a specific time, we can determine the specific “early” and “late” periods. For example, the record may be divided into “before” and “after” period at the time of completion of wastewater treatment plant improvements (Helsel and Hirsch, 2002).

In the trend for location 4505000 (Above Cutler), the slope for TP concentration was -0.00450 mg/l/yr and the slope for DP concentration was -0.00322 mg/l/yr. The slope for PP was -0.00150 mg/l/yr (Table 3.8). The DP reduction was faster than PP reduction during the conservation practice. The turbidity had not decreased significantly after LBRCP even though turbidity had significant correlation to TP concentration (Table 3.8). This is because, even though the correlation was statistically significant, it is not strong enough ($r^2 = 0.121$) to reflect the downward trend in TP concentration (Table 3.6). Therefore, it can be conclude that DP reduction mainly decreased TP concentration rather than Particle reduction.

SUMMARY AND CONCLUSIONS

The two head waters, East Fork and South Fork of the Little Bear River, have good water quality according to summary statistics while the Little Bear River below the confluence of these waters was impaired with TP.

Location 4905000 (Above Cutler Reservoir) had significant seasonality in all parameters. Other locations did not show a significant seasonality in TP or pH while there was significant seasonality in other parameters.

TP concentration had strong linear correlation to turbidity at each location. This correlation shows that turbidity may be an indicator of TP concentration in the Little Bear River watershed. There were significant correlations between water quality of upstream and down but the turbidity did not have any significant correlation between above and below Hyrum Reservoir.

Trend analysis showed a significant downward trend of TP and DP concentration after starting LBRCP, but did not show any trends of TP concentration before LBRCP. DP reduction was faster than PP reduction at location 4905000 (Above Cutler Reservoir).

LITERATURE CITED

Aitchison, J. and I.A.C. Brown, 1957. The Lognormal Distribution. Cambridge University Press, Cambridge, U.K., 176 pp.

Allison, P.D., 1995. Survival Analysis Using the SAS@ System: A Practical Guide. SAS Institute inc., Cary, North Carolina, 292 pp.

Berthouex, P. M. and L.C. Brown, 1994. Statistics for Environmental Engineers. Lewis Publisher, New York, New York, 335 pp.

Chapra, S.C., 1997, Surface Water-Quality modeling. Mc Graw Hill. New York, New York, 844 pp.

Chess, D., 2000. Utah Watershed Review Vol. 8.

Cohen, A.C. 1991. Truncated and Censored Samples. Marcel Dekker Inc., New York, New York, 282 pp.

Collett, D., 2003. Modeling Survival Data in Medical Research (2nd ed.). Chapman and Hall/CRC, London, 391pp.

Dodds W.K., 2002. Freshwater Ecology: Concepts and Environmental Applications. Academic Press, San Diego, California, 569 pp.

DeLorme (DeLorme Mapping Company), 1993. Utah Atlas and Gazetteer. DeLorme Mapping, Freeport, Maine, 64 pp.

EMRG, 2005. Little Bear River Watershed Data Viewer, Environmental Management Research Group, Utah Water Research Laboratory, Utah State University, Logan, Utah.

Gilbert R.O., 1987. Statistical Methods for Environmental Pollution Monitoring, Van Nostrand Reinhold, New York, New York, 320 pp.

Gilliom, R.J. and D.R. Helsel, 1986. Estimation of Distribution Parameters for Censored Trace Level Water Quality Data: 1. Estimation Techniques. Water Resources Research 22(2): 135-146.

Gleit, A. 1985. Estimation for Small Normal Data Sets with Detection Limits. Environmental Science and Technology 19(12):1201-1206.

Helsel D.R. 2005. Nondetects and Data Analysis, John Wiley & Sons, Hoboken, New Jersey, 250 pp.

Helsel D.R. and T.A. Cohn, 1988. Estimation of Descriptive Statistics for Multiply Censored Water Quality Data. Water Resources Research 24(12):1997-2004.

Helsel, D.R. and R.M. Hirsch, 2002. Statistical Methods in Water Resources. U.S. Geological Survey Techniques of Water Resources Investigation, Book 4, 512 pp.

Hirsch, R.M. and J.R. Slack, 1984. Nonparametric Trend Test for Seasonal Data with Serial Dependence. Water Resources Research 20(6): 727-732.

Hirsch, R.M., J.R. Slack and R.A. Smith, 1982. Techniques of Trend Analysis for Monthly Water Quality Data, Water Resources Research 18(11):107-121.

Hollander, M. and D.A. Wolfe, 1999. Nonparametric Statistical Methods (second ed.), John Wiley and Sons, New York, New York., 787 pp.

Kann, J. and V.H. Smith, 1999. Estimatinf the Probability of Exceeding Elevated pH Values Critical to Fish Populations in a Hypereutrophic Lake. Canadian Journal of Fisheries and Aquatic Science 56:2262-2270.

Kaplan, E.L. and P. Meier, 1958. Nonparametric Estimation from Incomplete Observation. Journal of the American Statistical Association 53(282), 457-481.

Kendall, M.G., 1970. Rank Correlation Methods (Fourth Edition), Chales Griffin & Company Limited, Lodon,U.K. 202 pp.

Kroll, C.N. and J.R. Stedinger, 1996. Estimation of Moments and Quantiles Using Censored data. Water Resources Research 32:1005-1012.

Kruskal, W.H. and W.A. Wallis, 1952. Use of Ranks in One-criterion variance analysis. Journal of American Statistical Association 47(260): 583-621.

Minitab Inc., 2006. Minitab® 14, http://www.minitab.com/products/minitab/, *accessed* July, 2006.

Minitab Inc., 2007, Minitab Macros, http://www.minitab.com/support/macros/ default.aspx?action= display &cat=non, *accessed* January, 2007.

Nehls, G.J. and G.G. Akland, 1973, Procedures for Handling Aerometric Data. Journal of Airpollution Control Association 23:180-184.

Theil, H., 1950, A Rank-Invariant Method of Linear and Polynomial Regression Analysis, 3 Proceedings of Koninalijke Nederlandse Akademie Can Wethenschatpen Series A, 53:1397-1412.

Shumway, R.H., R.S., Azari and M. Kayhanian, 2002. Statistical Approaches to Estimating mena water quality concentrations with detection limits, Environmental Science and Technology 36(15):3345-3353.

Spearman, C., 1904. The Proof and Measurement of Association Between Two Things. American Journal of Psychology 15: 72-101.

Stevens, D.K., 2006. Trend Analysis Visual Basic Program for Nooksack Watershed, Utah Water Research Laboratory, Utah State University, Logan, Utah.

Turnbull, B.W., 1976. The Empirical Distribution Function with Arbitrarily Grouped, Censored and Truncated Data. Journal of the Royal Statistical Society. Series B (Methodological), 38(3):290-295.

USEPA, 1993. Method 365.1 Determination of Phosphorus by Semi-automated colorimetry. Environmental Monitoring System Laboratory, Office of Research and development, US. Environmental Protection Agency, Cincinnati, Ohio.

USEPA, 1998. Guidance for Data Quality Assessment. Practical methods for data analysis; EPA/600/R-96/084.

USEPA, 2005. STORET (Storage and Retrieval), http;//www.epa.gov/stormoda /DW_stationcriteria. *accessed* November, 2005.

Utah DNR (Utah Department of Natural Resurces), 2004. Land Use/Water Related Use, www.agr.state.ut.us/agrc/html/data/layers1 html, *accessed* November, 2004.

Vogel, R.M. and C.N. Kroll, 1989. Low-flow Frequency Analysis using Probability-Plot Correlation Coefficients, Journal of Water Resource Planning and Management 115 (3):338-357.

Zar, J.H., 1999. Biostatistical Analysis, (Fourth Edition), Prentice Hall, Upper Saddle River, New Jersey, 661 pp.

CHAPTER 4

FILLING DATA GAPS BY A RANK-DATA DISTRIBUTION METHOD (R-D METHOD)

ABSTRACT

High frequency data are required to populate high frequency statistical water quality models, but the most common intervals between water quality samples are weekly or monthly and are irregular. The Rank-Data distribution method (R-D method) was developed based on the concept that time series of water resources data consist of data distributions and time series of the ranks of the data at the measurement times, and that the distribution of a full high frequency data set, including both observations and unknown values, is identical to the distribution of the observations. Cumulative Failure Probabilities (CFPs) of unknown values for dates with no observations were estimated by interpolating time series of the CFPs of the observations to create a daily time series of CFPs. This estimated time series of CFPs was then linked to the data distribution to obtain the flow time series. In tests, time series of mean daily flows from the R-D method were better estimates of time series of measured flows from the original daily data set than from simple interpolation between observations. These tests demonstrated the promise of generating time series of water quality or water quantity by combining the probabilistic results from Bayesian Networks to the CFP time series from the R-D method.

INTRODUCTION

The most common frequencies of water quality samples are weekly or monthly. These low frequency data are not usually sufficient to populate high frequency statistical water quality models. Especially, Bayesian Networks (BNs) require large amounts of data because BNs use data under specific combinations of conditions to build contingency tables. For example, when a BN is used to estimate the total phosphorus distribution in a specific stream under high precipitation and high agricultural land use conditions, the BN uses only TP load and flow data under these conditions. High frequency sampling and measurement is one solution to this problem but this solution requires much money and time (Cusimano *et al.*, 2002). A better way may be to estimate unmeasured water quality values based on the distribution of a smaller number of water quality measurements.

This chapter concerns estimation of high frequency time series of water-related variables using low frequency observations at the same location, and consists of three sections: Ideas, Methods, and Validation using specific data sets. **Ideas** develops the new approach to estimating these time series. **Methods** shows the detail of the procedure, and **Validation** shows how well the new approach works.

Our goal is to estimate high frequency water quality constituent fluxes in streams using low frequency observations. Though high frequency stream water quality observations are required to validate this new approach for our goal, those observations are not available for constituents that require manual sampling and lab analysis. However, high frequency flow observations are widely available and they will be used here to develop and test this new approach. If the pattern of seasonal cycles and fluctuations of flow are similar to those of nutrient loads, the new approach to flow is applicable for nutrient load in a similar fashion.

IDEAS

Water resources data consist of four components: magnitude, duration, timing and frequency. In order to evaluate all of these four components simultaneously, time series are required in addition to distributions. The key to this new approach is to decompose the time series in two components, the data distribution and the time series of rank. As a test case, 3603 daily mean flow observations at USGS gage 10128500, Weber Basin, Utah from November 1991 to October 2000 were considered.

Water quality data may fall below the detection limit (censored data) (USEPA, 1998). Because censored data provide information, deleting censored data may cause errors (Hesel, 2005). Helsel (2005) recommended the non parametric Kaplan-Meier method (K-M method (Kaplan and Meier, 1958)) for summary statistics with censored data. In this chapter, the K-M method was used to estimate flows on dates with no data (data gaps). The K-M method was used for the data distribution. The cumulative failure probabilities (CFPs) of flows are calculated using the K-M method.

$$f(t_i) = 1 - \prod \frac{n_i - d_i}{n_i} \qquad (4.1)$$

where t_i is flow, n_i is the number of flow observations which are not smaller than t_i or are censored with detection limit equal to or larger than t_i, d_i is the number of detects with t_i, and $f(t_i)$ is Cumulative Failure Probability (CFP). The Cumulative Failure plot (CF plot) of these daily mean flows shows the probability associated with values not larger than a specific flow (Figure 4.1).

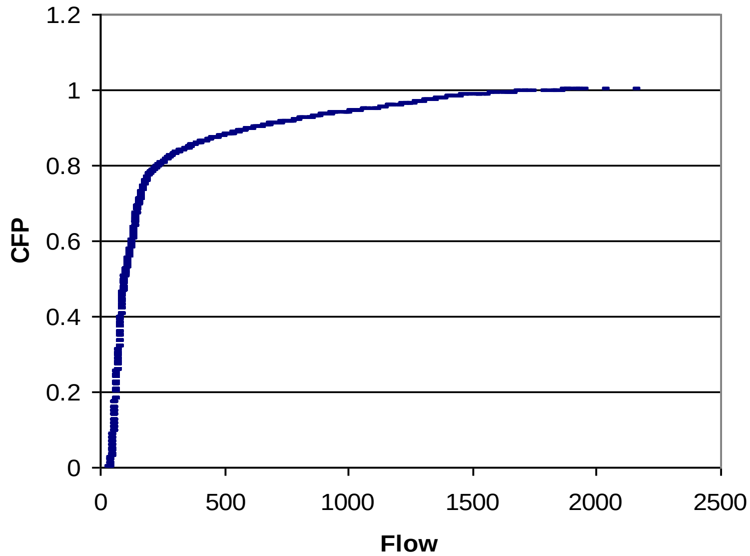


FIGURE 4.1. Cumulative Failure Plot (CF Plot) for the Entire Time Series of Daily Mean Flow (cfs) at USGS Gage 10128500.

Each flow has a specific rank among all flows. For example, the flow on June 14, 1991, 1640 cfs is the 33rd largest flow (3571st smallest flow) and the CFP is $0.99\,(=1-\frac{3603-3571}{3603})$. The time series of the rank shows the timing of maximum and minimum values (Figure 4.2). If we know the all ranks of the measurement times and the data distribution, the flow time series may be reconstructed. For example, if the CFP on July 2nd, 1995 is 0.98 on the time series of the rank and the flow for CFP 0.98 is 1370 cfs on the CF plot, the flow on July 2nd , 1995 is 1370 cfs. This process is shown in Figure 4.3.

Two assumptions are used to estimate the values for dates on which data are not available. The first assumption is that the distribution of all values including both

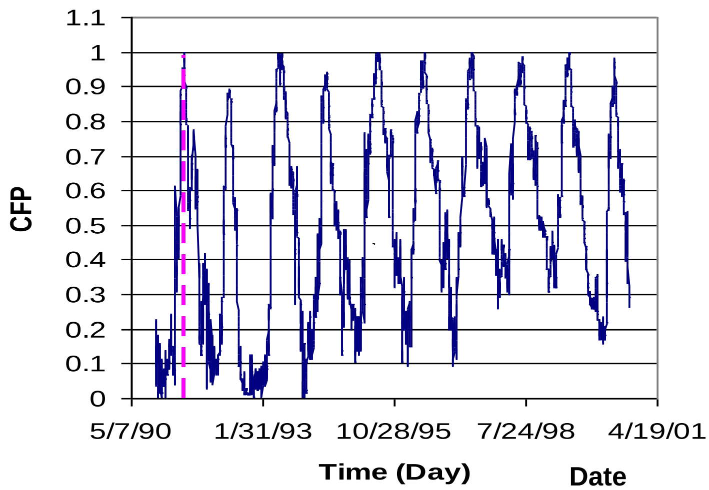
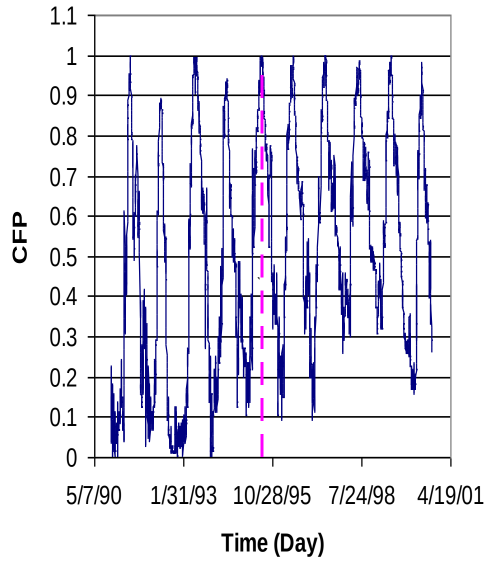


FIGURE 4.2. CFP Time Series of Flows at USGS Gage Station 10128500. (▬: CFP, ----: June 14, 1991).

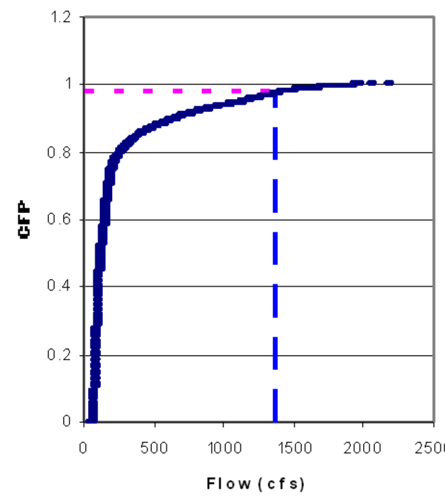
observations and unmeasured values is identical to the distribution of the observations alone. Under this assumption, we can generate an unlimited number of values using the CF plot of observations. The second assumption is that the interpolation or extrapolation between the CFPs of two observations on the CFP time series is identical to the true CFP for the prediction date (Equation 4.2).

$$p_i = \frac{p_2 - p_1}{t_2 - t_1}(t_i - t_1) + p_1 \qquad (4.2)$$

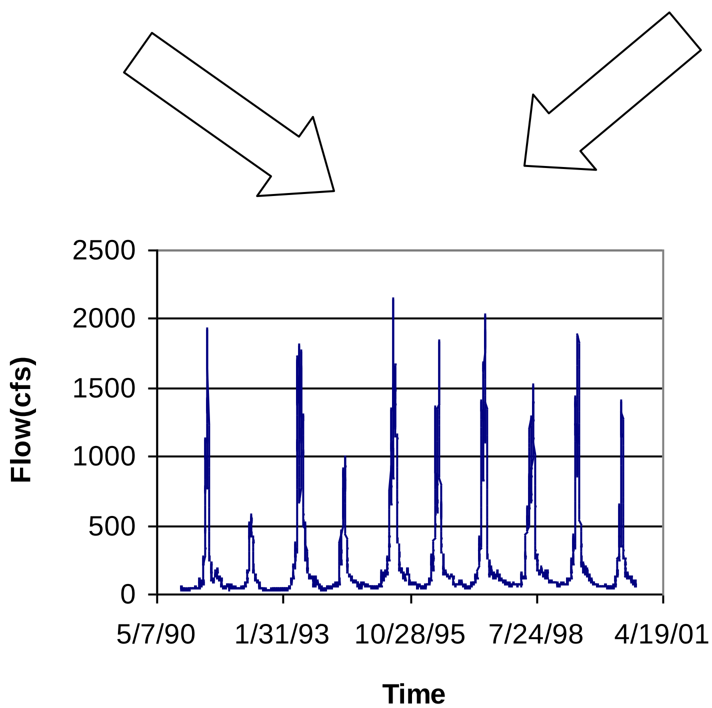
where t_1 and t_2 are two sampling dates of closest observations from prediction date, t_i is the date of prediction, p_1 is CFP of observation t_1, p_2 is CFP of observation at t_2 , p_i is the CFP of prediction at t_i. For example, if the flow CFP on November 5 (309th day), 1990 is 0.133 and the flow CFP of December 8 (342nd day) is 0.017, the flow CFP on November



a) Rank Time Series



b) Data distribution of flows as CFP.



c) Constructed Time Series of Values

FIGURE 4.3. The Work Flow to Make a Time Series by Combining Data Distribution with CFP Time Series.

25 (329th day) is 0.063 ($= \frac{0.017 - 0.133}{342 - 309}(329 - 309) + 0.133)$, the interpolation value between flow CFPs on November 5 and on December 8. In the same way, all the CFPs of predictions may be calculated. The section 4.4 will show how well these two assumptions work to construct the flow time series.

This Rank-Data distribution method (R-D method) is beneficial for BNs and TMDL calculation. BNs provide the data distribution under the specific conditions. If these BN outputs are connected with the rank time series estimated by interpolation of the existing rank under similar conditions, we can generate time series values under the specific conditions. This time series BN output may show the effect of changing condition on the timing and duration of water quality as well as frequency and magnitude. Chapter 7 will describe the use of the R-D method for Bayesian Networks.

Some monitoring-based Total Maximum Daily Loads (TMDLs) used statistical representations, mean and geometric mean to estimate loads (Brannan *et al.*, 2005). If we estimate daily flows and nutrient loads from monitoring data by the R-D method, we may calculate TMDLs based on daily observations and predictions but not statistical representations. Chapter 6 will describe how to estimate true TMDLs using the R-D method.

METHODS

We used two approaches of estimation for data distributions, parametric and non-parametric. In this dissertation, the non parametric KM method was used to avoid violating distribution assumptions required by parametric methods. Daily mean flow data at USGS gage 10128500, Weber River near Oakley, UT from October, 1991 to

September, 2000 were used to develop the R-D method. The MINITAB version 14 (Minitab Inc, 2006) statistical software was used to calculate cumulative failure probability (CFP) for flows and Microsoft Excel (Microsoft Inc, 2004) was used to generate flows using the CF plot. Sixty flows (Observations) were selected from 3653 flows (Original data). Each flow was randomly selected within each 2 month sampling block. The time blocks were October to November, December to January, February to March, April to May, June to July and August to September in each water year. The purpose of the procedure is to develop an approach for reconstructing the full time series of daily flows for 10 years (3603 flows) from these 60 selected observations.

Estimated Distribution
from Cumulative Failure Plot

The simulation period is from November 5, 1990, the earliest date, to September 15, 2000, the latest date, of selected sampling flows (Observations). The goal was to reconstruct the time series of 3603 flows from November 5, 1990 to September 15, 2000 using 60 observations. These 60 observations on which the reconstruction was based were regarded as “detected” in the context of K-M method, keeping their original value. The remaining flows were regarded as censored flows with censoring level at the highest observation flow, 1,870 cfs.

The K-M method was designed for right censored (above detection limit) and detected data only, but not left censored (below detection limit) data. Because 3,543 (= 3,603-60) unmeasured values are left censored, the flows were flipped to make a CF plot by the K-M method (Helsel, 2005).

$$Flip_i = M_i - x_i \quad (4.3)$$

where $Flip_i$ is the Flipped value, M_i is the Flipping Constant, x_i is the observation.

Minitab™ (Minitab Inc, 2006) was used to calculate the survival probability (= 1-CFP) of the flipped values using the K-M method. After the flipped values were returned to the unflipped scale, the CFPs were estimated (Figure 4.4). Both detected (observations) and censored value (unmeasured values) were estimated based on this cumulative failure plot. For example, if the simulation period is 3,603 days and we estimate mean daily flow for each day, the failure probability of the smallest flow equals 1-(3602/3603)= 0.00028. The CFP of the 2nd smallest flow equals 1-(3601/3602)*(1-0.00028) = 1-(3601/3603)= 0.00056, and so on. The flow with CFP 0.00028 on the cumulative failure plot is assigned to the smallest flow and the flow with CFP 0.00056 is

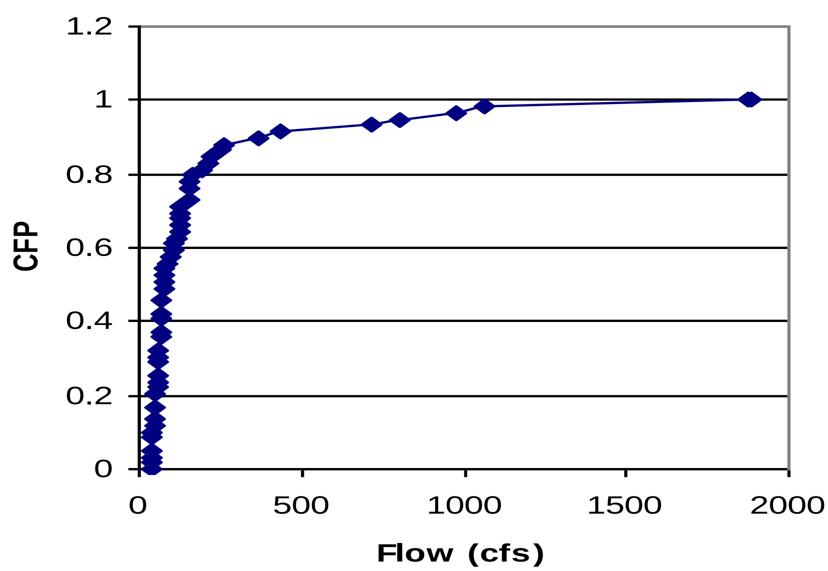


FIGURE 4.4. Cumulative Failure Plot of 60 Know Flows at USGS Gage 10128500.

assigned to the 2nd smallest flow. The 3603 CFPs and flows (estimated distribution) may be estimated from the smallest flow to largest flow in the same manner.

Estimation of CFPs and Conversion to Original Values

The CFP for each value was calculated within the observations using Equation (4.4)

$$\text{CFP} = 1 - \frac{r-1}{N} \tag{4.4}$$

where r is rank and N is total number of observations (= 60).

The interpolation of the CFPs of observations (observed CFPs) estimates the CFP of the unmeasured values (interpolated CFPs) between two measurement dates (Figure 4.5). All of the CFPs including observed CFPs and interpolated CFPs were converted to a single set of estimated ranks. The CFPs were then recalculated by these estimated ranks because the number of CFPs increase by interpolation (estimated CFPs). After finding the value with a rank within the estimated distribution, the value (prediction) was assigned to the date with the same rank within the estimated CFP time series. For example, the interpolated CFP for the mean daily flow on October 10, 1991 was 0.59 corresponding to rank 1644 within the 3603 estimated CFP time series. The 1644th flow (1644th prediction) within the estimated distribution, 81.84 cfs is then assigned to the flow on October 10, 1991 (Figure 4.6). The procedure is carried out for all data to produce a time series of daily mean flows.

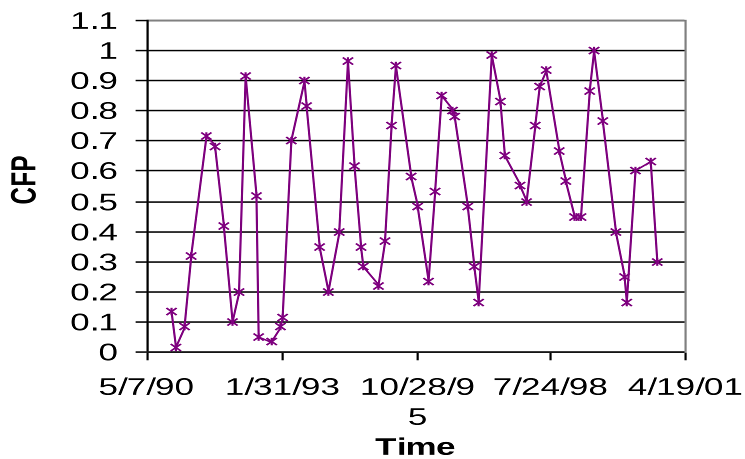


FIGURE 4.5. The CFP Time Series of Observations (x) and Unmeasured Values (Line).

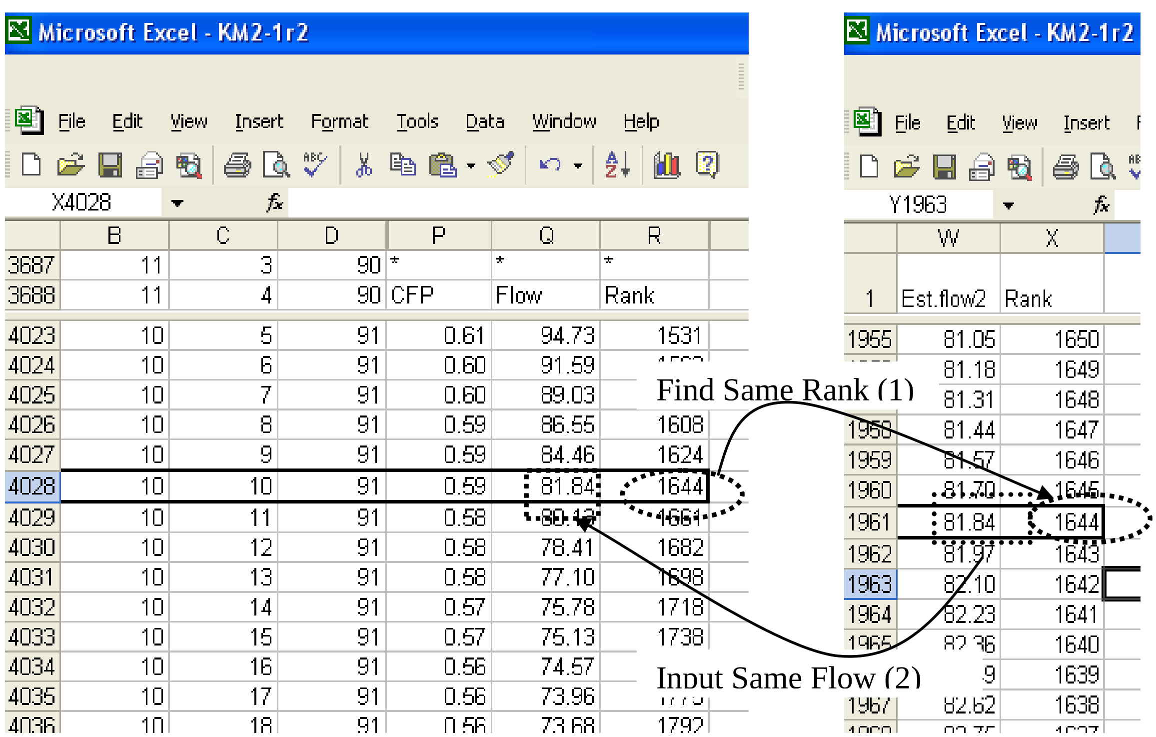


FIGURE 4.6. The Steps to Assign Estimated Values to Simulation Dates.

It is recognized that the data set with flow observations for a particular year may not necessarily include the peak flow during spring runoff for that year. Peak CFPs were approximately added to observed CFP time series before interpolation in order to improve the estimates of these peaks. The peak CFPs were estimated under the assumption that the dates of the peak CFP for each cycle are known. The steps to estimate those peaks and improve CFP time series

1) Calculate of the steeper slope of the two sides of the largest observed CFP in each yearly cycle of observed CFP time series (Sa_1, Sa_2, Sa_3,……Sa_n in Figure 4.7).
2) Estimate the peak CFPs. The largest observed CFP in each yearly cycle is extended to the peak CFP with the minimum slope of yearly steeper slopes calculated in 1)

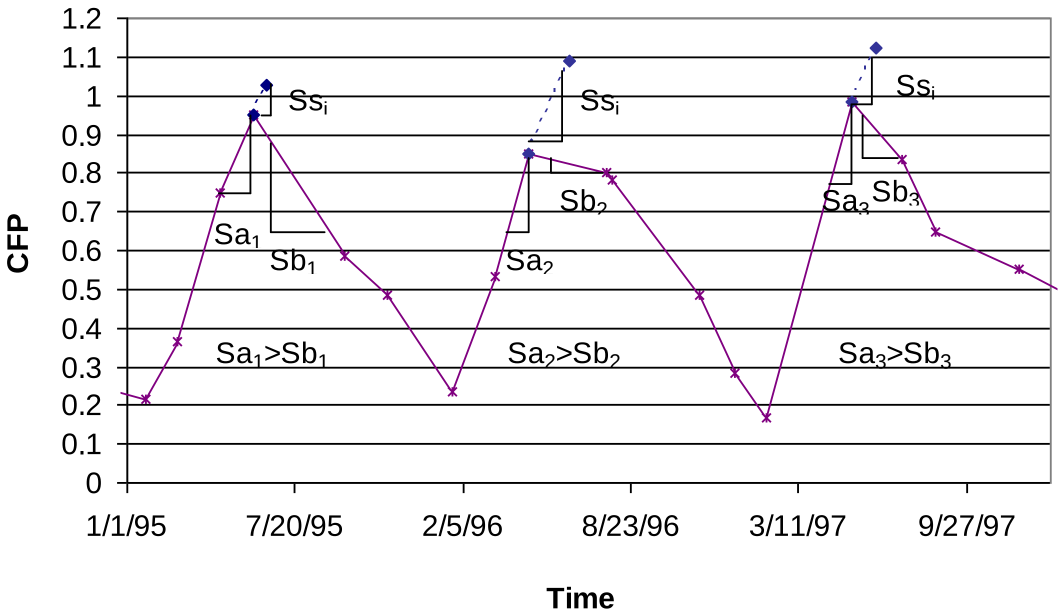


FIGURE 4.7. Estimation of Peak CFP of Each CFP Cycle Extending Largest CFP of Known Data Within Each Cycle. The Slopes of ---- Are Extended Slopes.

(Ss_i = minimum of Sa_1, Sa_2, Sa_3,……Sa_n in Figure 4.7). The line from the largest observed CFP to the estimated peak CFP is defined as the extended CFP line and the slope of these extended lines is defined as extended CFP slopes. Because we know the date of the peak CFP in each yearly cycle, the estimated peak CFP is where the extended line meets the peak date in each cycle. The same extended slope is used for all yearly cycles (Figure 4.7).

3) Rank adjustment. The estimated peak CFP may be greater than 1.0 because the CFP is calculated by extension of the largest observed CFP in each yearly cycle. Since the total number of CFPs increases after adding the approximate peak CFP, the CFPs are recalculated based on the rank of CFPs in this extended flow CFP set (extended CFPs).

4) Evaluate the initial extended CFP slope. After getting time series for all simulation dates by connecting the extended CFP time series with the estimated distribution, the sum of absolute differences between the largest observation and the prediction for that date in every yearly cycle (sum of largest observation residual) are calculated (Figure 4.8).

5) Optimize extended CFP slope. The extended CFP slope value is optimized by increasing Ss_i (Figure 4.7) to minimize sum of the largest observation residual (automatically using a program written in Visual Basic for Applications in MS Excel). The extended CFP slope is varied from the minimum to maximum values calculated in 1). Steps 2), 3) and 4) are repeated replacing extended CFP slopes.

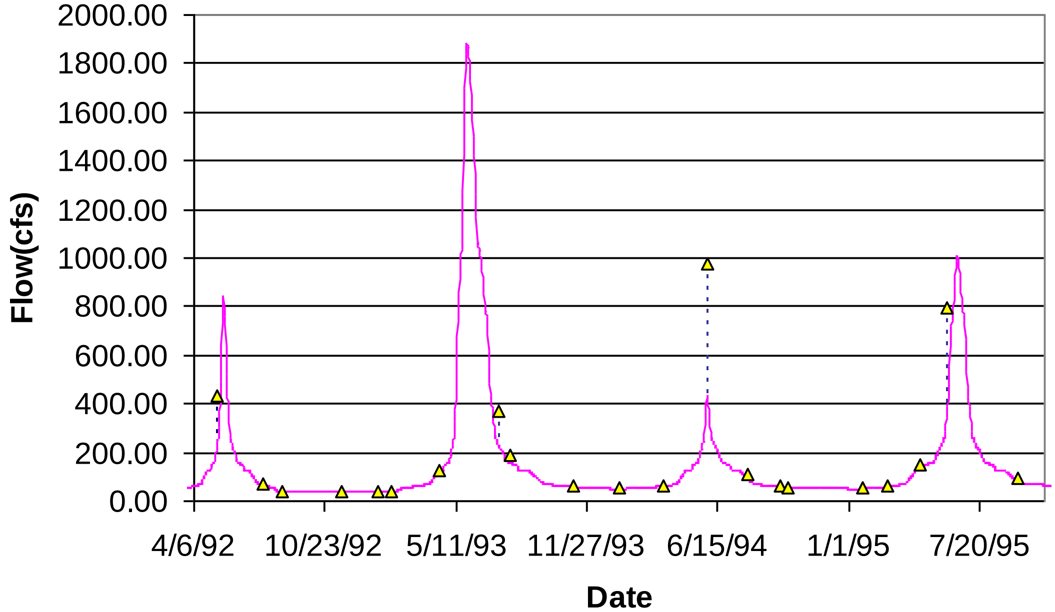


FIGURE 4.8. Example Daily Mean Flows for Each Date at USGS Gage 10128500 (-: Estimated Values, Δ: Obersvations, ----: Largest Observation residual between Observation and Prediction).

The time series of flow CFPs using the peak CFPs estimated by the extended CFP slope to minimize sum of the largest observation residual is defined as optimized CFP time series. This optimized CFP time series is used for predictions.

VALIDATION

USGS gage station 10105900, Little Bear River near Paradise, UT has 3,652 daily mean flows from October 1, 1992 to September 30, 2002 (original data). We sampled these flows (observations) randomly using three different sampling period blocks. The first is a two month block- data were selected randomly from every two month block of original data's time series (data set 1). For example, the daily mean flow on November

30, 1992 is selected randomly from the time block of October and November, 1992. The second set used a 1 month block (data set 2), and the third used a 2 month block for low flow season and a 1 month block for high flow season (March, April, May, and June) (data set 3). These three different data sets were handled independently (Table 4.1). We then reconstructed the 3554 original data's time series (original time series) from November 5, 1990, the earliest date to September 15, 2000, the last date of observations. All steps in section 4.3 were followed.

Minitab (Minitab Inc, 2006) was used to estimate the CFP for each observation by the K-M method after flipping the flows (Appendix C). A program written in Visual Basic for applications in Microsoft Excel estimated the 3554 estimated distributions from the CF plot using data sets 1, 2 and 3 by interpolation (Appendix D).

In order to evaluate the agreement of estimated distribution with distribution of original data, log scale Quantile-Quantile plots (USEPA, 1998) or Q-Q plots were used because log scale graph may be a better expression for extremely high flows. According to the Quantile-Quantile plots (USEPA, 1998) or Q-Q plots (Figure 4.9), the agreement

Table 4.1. Sampling Summary for Test Data Sets 1992-2002, USGS Gage 10105900.

	Data Set 1	**Data Set 2**	**Data Set3**
Sampling Period Block	2 month	1 month	1 month (Mar.-Jun.) 2 month (Jul.-Feb.)
Total number of estimated distributions	3554	3554	3554
Number sampled observations	60	118	78
Maximum Flow (cfs)	436	685	685

of the estimated distributions from data set 1 with the distribution of original data was appropriate when the range of prediction was 11 cfs (1 percentile value) to 425 cfs (96 percentile value) but estimation of extreme low or high flows were biased (Figure 4.10 a). While the agreement of the estimated distribution with the distribution of original data was appropriate over most the prediction's range, the estimated distribution from data set 3 deviated from the distribution of original data over most of the prediction's range (Figure 4.9 b, c). The inappropriate estimation may be caused by collecting more observations during high flow. Because the prediction flows are estimated by linking the flow distribution and CFP time series in the R-D method, an erroneous flow distribution may cause large error (difference between predictions and original data set). For this reason, the estimated distribution from data set 1 (instead of the estimated distribution from data set 3) was connected with CFP time series of data set 3 to estimate predictions in data set 3.

A program written in Visual Basic for Applications in Microsoft Excel was used to estimate CFP time series interpolating observation CFPs (Appendix E). The optimum annual peak CFPs at known peak times were estimated by selecting the optimum extended CFP slope (Figure 4.7) to minimize the sum of the largest observation residuals. These estimated annual peak CFPs are added to observed CFP time series to get optimized CFP time series (Table 4.2). The optimized CFP time series was then

Table 4.2. Estimating and Adding Peak CFPs.

	Data set1	**Data set2**	**Data set3**
Sampling Period Block	2 month	1 month	1 month+2 month
Slope Range (cfs/day)	0.003 – 0.025	0.001-0.026	0.002 – 0.025
Optimum Slope (cfs/day)	0.005	0.001	0.002
Added peak CFP (points)	14	14	14

converted to a rank time series. The optimized CFPs are recalculated by these ranks after the number of CFPs increased by interpolation to estimate unmeasured CFPs (Optimum CFP time series). After finding the value with a rank within an estimated distribution, the value was assigned to the date with that rank within optimum CFP time series.

The sum of the absolute values of the residuals was used as the criterion to evaluate the accuracy of the R-D method (Equations (4.5), (4.6)).

$$\text{Sum of residual} = \sum | O_i - P_i | \tag{4.5}$$

$$\text{Sum of largest observation residuals} = \sum | Q_y - P_y | \tag{4.6}$$

where Q_i is the observation at date i, P_i is the prediction from R-D method at date i, Q_y is the largest observation during year y and P_y is the prediction from the R-D method corresponding to Q_y.

Because the sum of the largest observation residuals (Figure 4.8) was positively correlated with the sum of the residuals (Figure 4.10), the extended CFP slope that minimized the largest observation residual may be used to estimate the optimum predictions.

Three different methods were used to construct flow time series. Method 1 was simple interpolation of observations, method 2 was an R-D method connecting data distribution to estimated CFP time series without peak CFPs and method 3 was an R-D method connecting the data distribution to optimum CFP time series with peak CFPs.

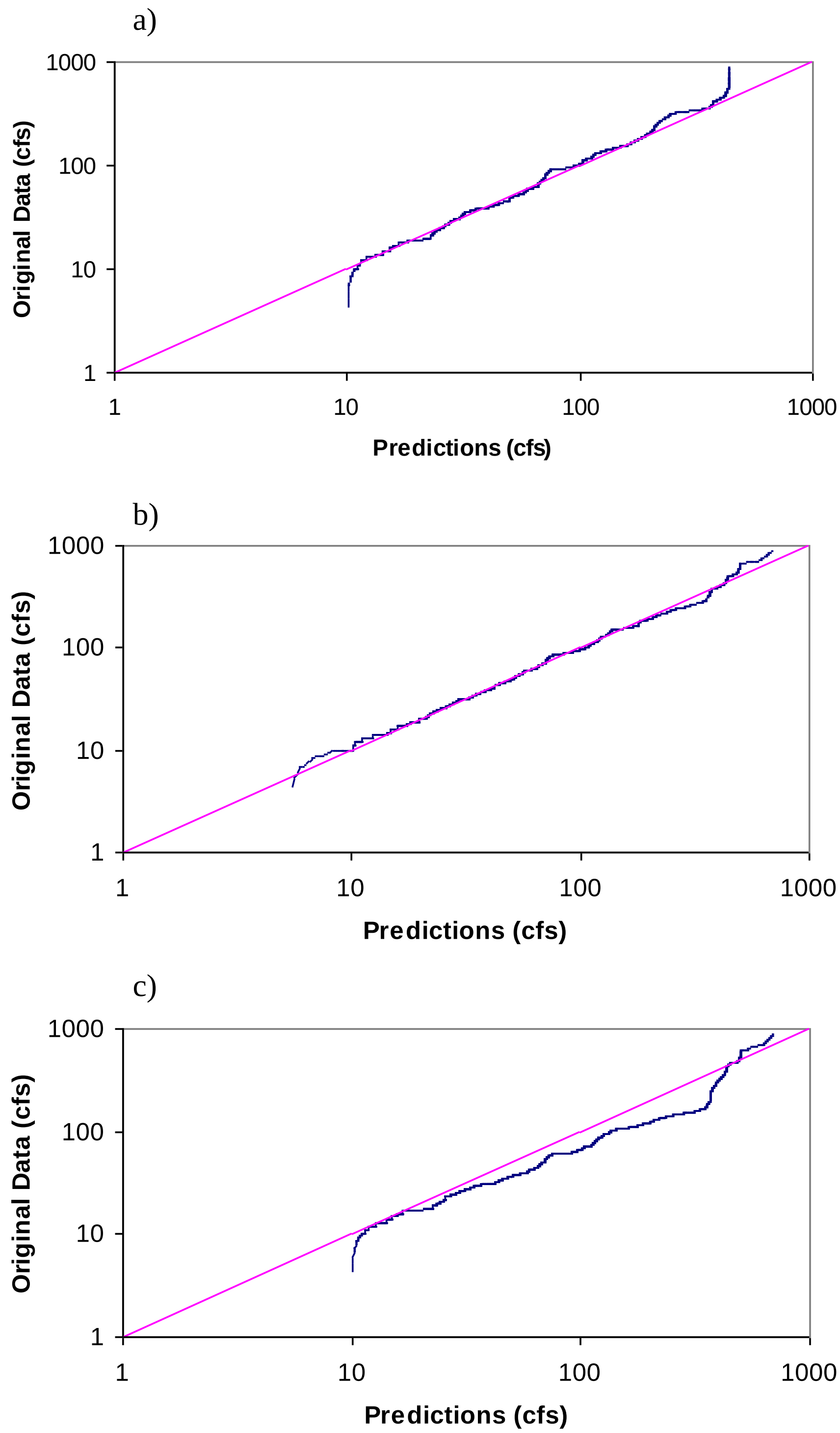
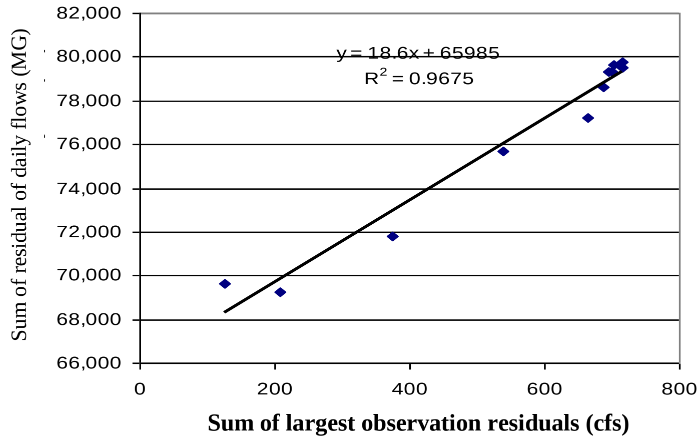


FIGURE 4.9. Log Scale Q-Q Plot Predictions from Data Sets versus Original Data. a): Data Set 1 (2 Month Sampling Block), b) Data Set 2 (1 Month Sampling Block), c) Data Set 3 (1 Month + 2 Month Sampling Block).

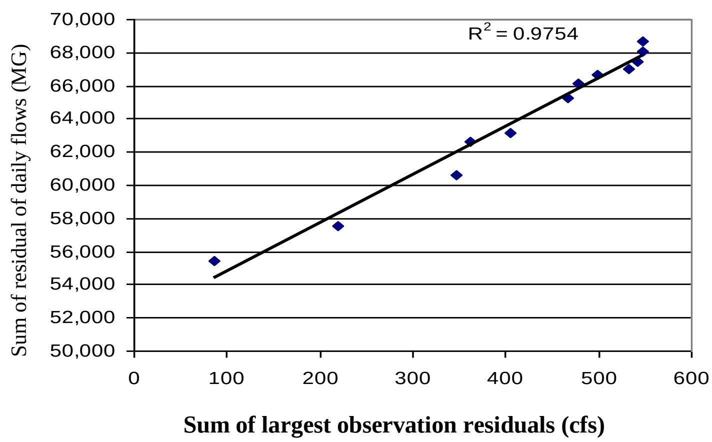
The agreement of the CFP time series from the R-D method with original CFP time series for 3554 original data was improved by method 3 (using peak CFPs), using data set 1 (Figure 4.11). The agreement of the time series of predictions by R-D method with the time series of original data was improved (Figure 4.12) and the sum of residuals decreased for data set 1 when the optimum CFP time series (method 3) was used (Table 4.3). Because the agreement of the CFP time series from R-D method with the original CFP time series was not significantly improved by method 3, using data sets 2 and 3 (Figure 4.11), the agreement of the time series of predictions with original time series was not improved (Figure 4.12) and the sum of residual did not decrease (Table 4.3). When comparing the sum of residuals from method 2 (R-D method without peak CFPs) to that from method 1 (simple interpolation), flow estimation might not be improved by method 2 (Figure 4.12, Table 4.3).

The total flow for 3554 days was 227,760 million gallons (MG). The flow time series reconstructed by the R-D method (methods 2 and 3) is closer than the flow time series by simple interpolation (method 1) of observations to original flow time series on the test using data set 1 and data set 3 but there was no significant change of flow time series on the test using data set 2 (Figure 4.12).

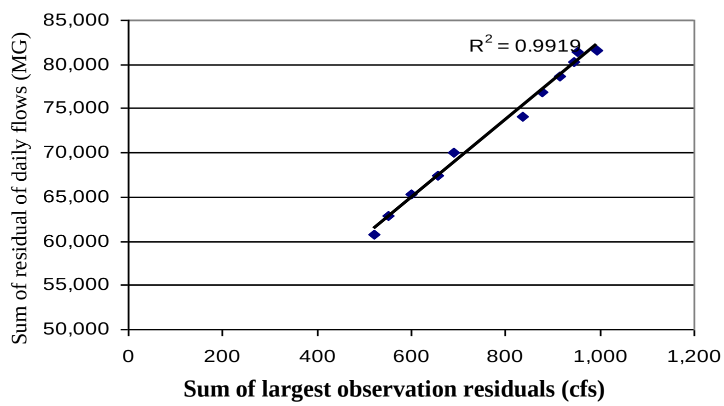
The differences between total flows from the original data set and the predictions (error) for 3554 simulation days from the R-D method (methods 2 and 3) was larger than those from simple interpolation using data sets 1 and 3 (Table 4.4), even though the flow time series from methods 2 and 3 was a better estimation than method 1. There was no significant difference in the total flow between the method 1 and the R-D methods



a) Data set1

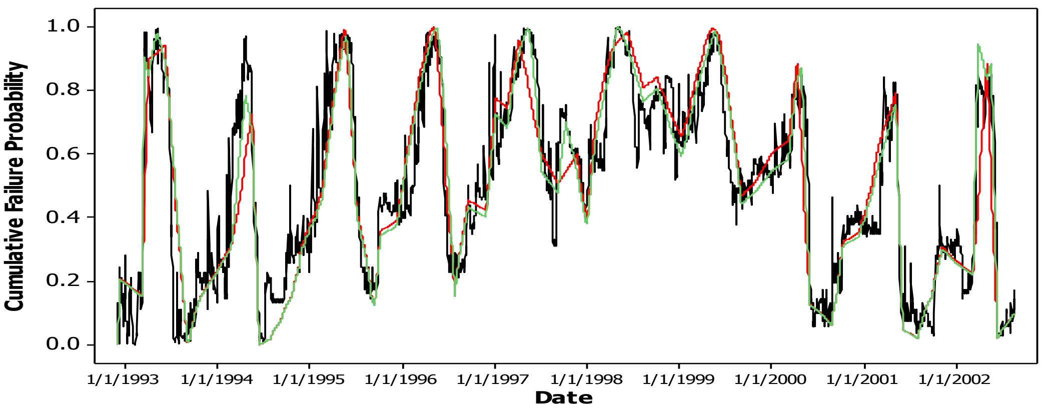


b) Data set 2

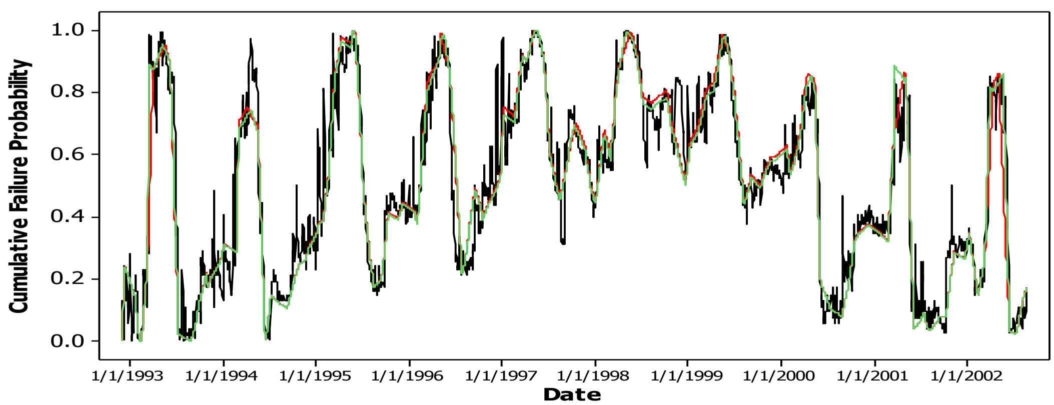


c) Data set 3

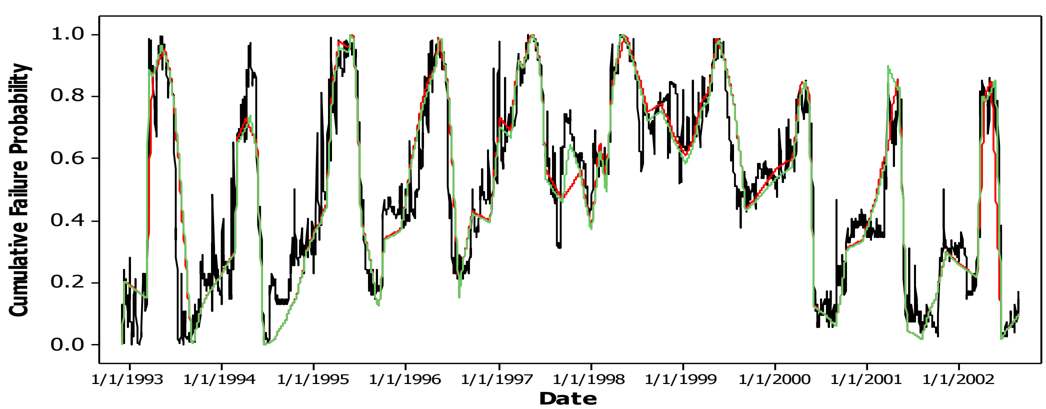
FIGURE 4.10. The Linear Relation Between Sum of Residual between Sum of Largest Observation Residuals and Sum of Residual Using Data Set 1, 2, and 3.



a) Data set 1 (2 month block)

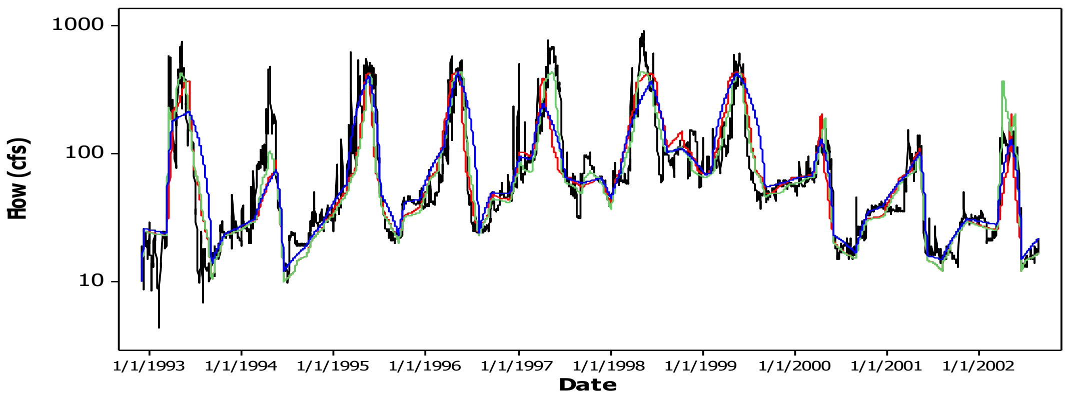


b) Data set 2 (1 month block)

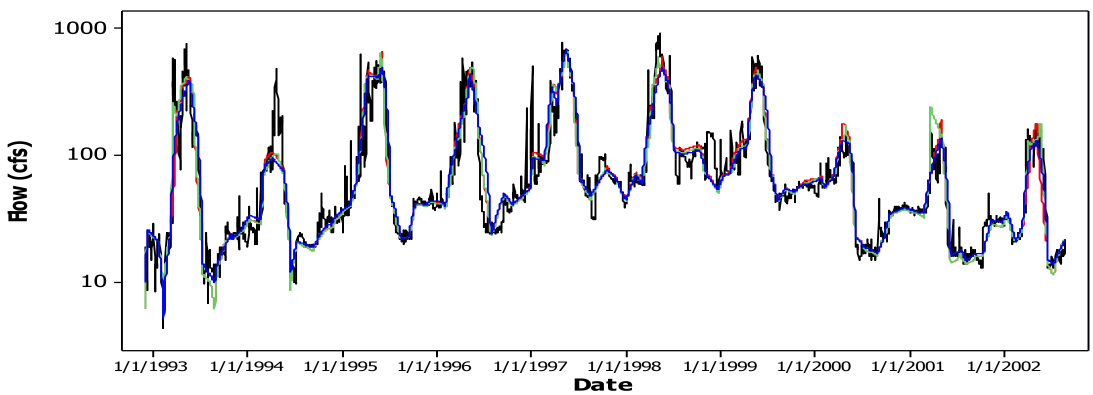


c) Data set 3 (1 month + 2 month block)

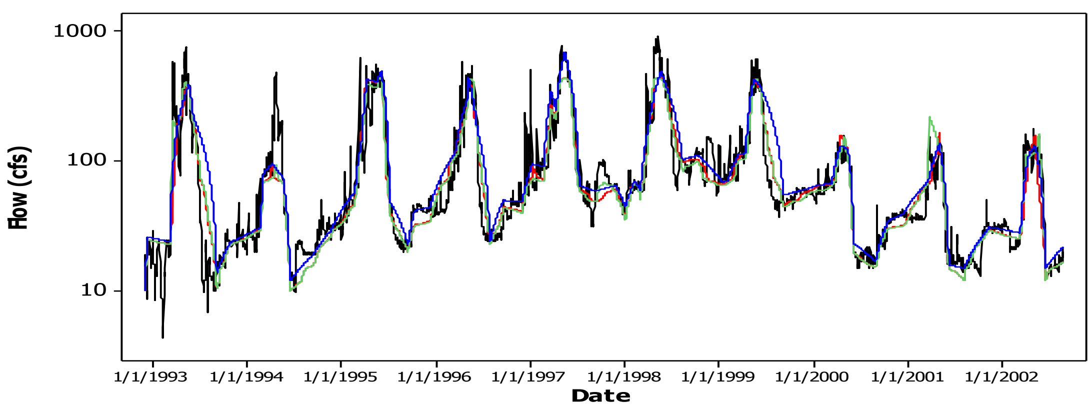
FIGURE 4.11. CFP Time Series for each Data Sets (Black: Original CFP Time Series, Red: Using Estimated CFP Time Series (Method 2), Green: Using Optimum CFP Time Series (Method 3).



a) Data set 1 (2 month block)



b) Data set 2 (1 month block)



c) Data set 3 (1month + 2 month block)

FIGURE 4.12. Time Series for each Data Sets (Black: Original Time Series, Blue: Estimated by Simple Interpolating Observations (Method 1), Pink: Estimated Using Estimated CFP Time Series (Method 2), Green: Estimated Using Optimum CFP Time Series (Method 3)).

Table 4.3. The Summation of Residuals of Daily Flows.

(Unit: MG)

	Data set 1	**Data set 2**	**Data set 3**
Method 1	92,439	55,053	68,808
Method 2	90,974	60,213	63,429
Method 3	69,643	55,413	60,738

Table 4.4. Total Flow for 3554 Days, ( ) is Error by Subtracting Observed Total Flows by Predicted Total Flows. (Unit:MG)

	Data set 1	**Data set 2**	**Data set 3**
Method 1	216,067 (-11,693)	221,502 (-6,258)	238,694 (10,935)
Method 2	204,585 (-23,175)	223,017 (-4,742)	204,597 (-23,163)
Method 3	204,589 (-23,171)	223,044 (-4,716)	204,613 (-23,147)

(method 2 and 3) using data set 2 (Table 4.4). The range of percent error for 10 year flow total was 2.1% to 10.2% for methods 2 and 3, and 2.7% to 5.1% for method 1 (Table 4.4). Because the positive and negative residuals canceled each other, the error of estimated total flow for 3554 days was reduced. It is possible for method 1 to estimate the total flow with less error even though method 1 caused a large sum of flow residuals using data sets1 and 3.

DISCUSSION

The validation of the R-D method shows the following. First, the estimated distribution from the CF plots of data sets did not deviate significantly from the distribution of the original data using data sets 1 and 2. The *p* values were very high for data sets 1 and 2 using the Kruskal-Wallis test (Kruskal and Wallis, 1952; Chapter 3) where the null hypothesis was that the flow was the same in the estimated distribution and original data (Table 4.5). This means the distribution of observations for a long time period such as for 10 or 20 years at low sampling frequency may be a good estimate of

Table 4.5 The Results of Kruska-Wallis Test for Predictions from Data Sets 1, 2, and 3.

Basis data set	**Data set1**	**Data set2**	**Data set3**
Sampling Period Block	2 month	1 month	1 month+2 month
Number of Predictions	3554	3554	3554
Median [1)]	52.58 (50.00)	50.53 (50.00)	68.04 (50.00)
Average Rank [2)]	3535.1 (3573.9)	3541.3 (3567.7)	3828.8 (3280.2)
H-value	0.63	0.29	127.05
p-value	0.426	0.588	<0.001

1) The median of predictions from each data set (median from original data).
2) The average rank of predictions from each data set (average rank from original data).

the distribution of the daily observations. It is very difficult to get daily frequency water quality data but in many cases low frequency water quality data for 10 years are available. The estimated distribution of a mixed block (data set 3), in which data were collected with a one month block for high flow and with a two month block for low flow, was a poor strategy for estimating the distribution of original data because the p value was < 0.001 in the Kruskal-Wallis test (Table 4.5). More frequent measurement during high flow shifted the cumulative failure plot for the 95 percentile (433 cfs) or smaller observations to right side in Figure 4.9 (c).

Second, the disagreement of the optimum CFP time series with the original CFP time series had a large negative effect on the sum of residuals while the disagreement of the estimated distribution with the original data distribution had a small negative effect. The estimated time series using the original data distribution and original CFP time series (category D0T0) is identical to the time series of original data and the sum of residual is 0 (Figure 4.13). While the sum of residual of the estimated time series using original data distribution and optimum CFP time series of data set 1 or 2 (category D0T1 or D0T2) was very high, the sum of the residual of the estimated time series using the estimated

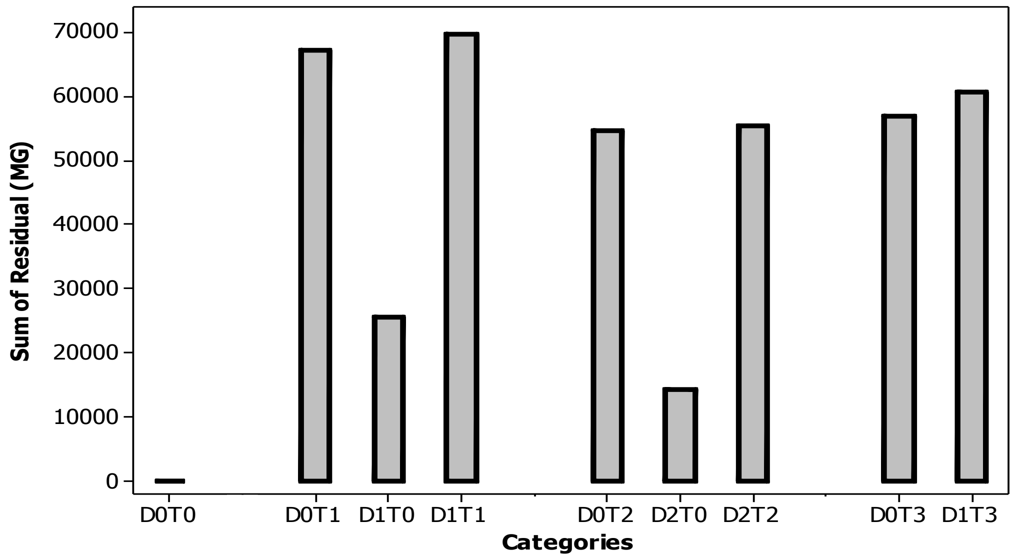


FIGURE 4.13. Residual Sum Between Predictions and Original Data for Combinations of Distribution Type and CFP Time Series Type. (D0,D1,D2: Using Data Distribution from Original Data Set, Data Set 1 and Data Set 2; T0,T1,T2,T3: Using CFP Time Series from Original Data (T0) and CFP Time Series by Method 3 from Data Sets 1 (T1), 2 (T2), and 3 (T3).)

data distribution of data set 1 or 2 and original CFP time series (category D1T0 or D2T0) was low. It may be concluded that the effect of the disagreement of the estimated distribution over the 93th percentile value with the original data distribution on sum of residual was small as long as the optimized CFP time series is accurate.

Third, optimizing the CFP time series by adding annual peak CFPs was effective to enhance the simulation accuracy of the R-D method. This means we may be able to predict good CFP time series of flow or water quality even though the CFP time series of observations suffers from low frequency or high randomness of the sampling period.

Fourth, the R-D method was beneficial when observations were more random and less frequent. The difference between the sum of the residual from the R-D method and interpolation is large for the data set with a two month sampling block. The difference is significant for the data set with combination of a one month sampling block during high flow and a two month sampling block during low flow but there is no significant difference in the data set with one month sampling block. The R-D method appears to be less sensitive to sampling conditions than interpolation.

Fifth, when the R-D method is used, we may reduce the sampling frequency and keep same flow sum of residuals. The measuring frequencies were 6 times, 12 times and 8 times per year for data sets 1, 2, and 3 (Table 4.1). Plots of sum of residuals versus measuring frequency show the frequency of interpolation and the R-D method corresponding to the same sum of residuals (Table 4.3, Figure 4.14). For example, in the validation data sets, the sampling frequency of the R-D method should be 8 times/year corresponding to 61,000 MG sum of residual for 10 years while the sampling frequency

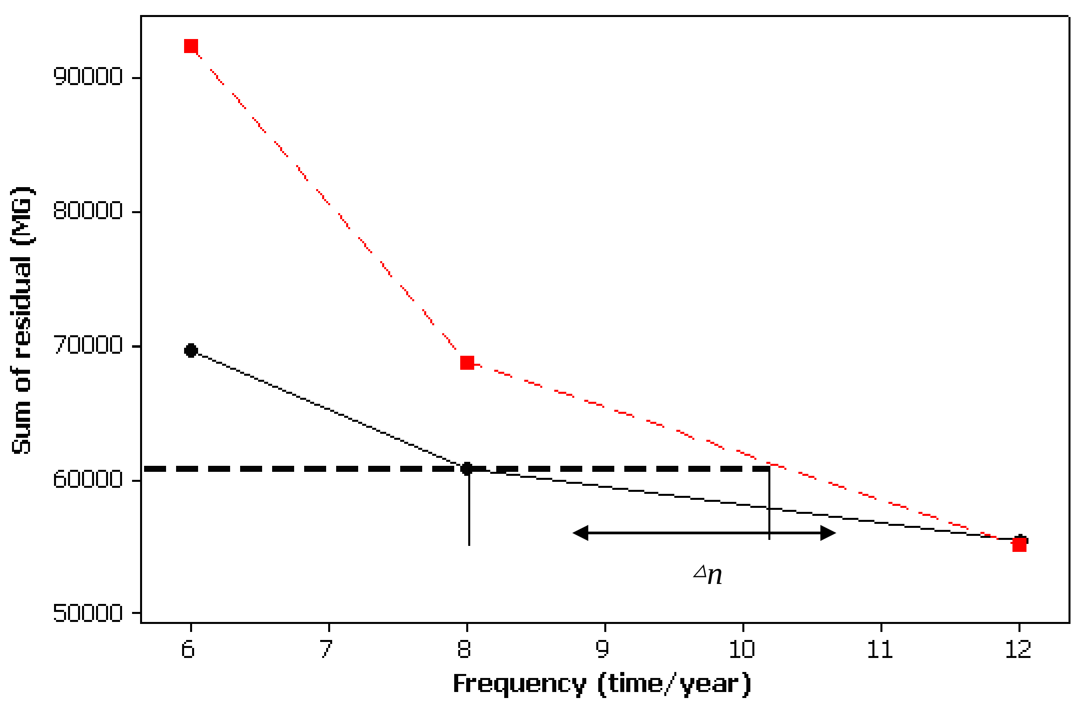


FIGURE 4.14. Sum of Flow Residual for Each Sampling Frequency (■: Interpolation, ●: R-D method). Each value is in Table 4.3.

of interpolation should 10.3 times/year. Therefore, sampling frequency difference (Δn) of flow between two methods equal to 10.2 - 8 =2.2 time/year for 61,000 MG sum of residuals. Less sampling reduces the cost of sampling and measurement or analysis (Figure 4.14).

$$B = C \times \Delta n \times p \qquad (4.7)$$

where B is benefit as cost reduction ($), C is cost per sample or measurement ($/sample or measurement), Δn is the sampling frequency difference (time/year) and p is period (year).

It is an interesting issue to determine the frequency of water quality sampling. Mesner *et al.* (2007) measured turbidity at every 30 minutes at the Little Bear River near Paradise, UT in 2006. The turbidity observations were converted to Total Suspended Solid (TSS) through regression between turbidity and TSS. TSS values were subsampled twice a day, daily, weekly and monthly from TSS estimations at a 30 miniute interval (continuous TSS). Because of the many combinations of data collection for each sampling interval, multiple data sets were collected for each case. TSS yearly load from all 30 minute values was 8.9 x 10^7 kg/yr and that from daily values was from 1.1 to 0.6 x 10^7 kg/yr (Figure 4.15). This means that the error (difference of yearly TSS load between using continuous values and daily values) was approximately ± 28 % of TSS yearly load from continuous values. This means if the yearly TSS estimated from the model is very close to the yearly TSS from daily observation, the estimated year TSS may be acceptable. The total flow from R-D method was very close to the total flow from daily mean flow observations during 10 years using monthly observations (data set

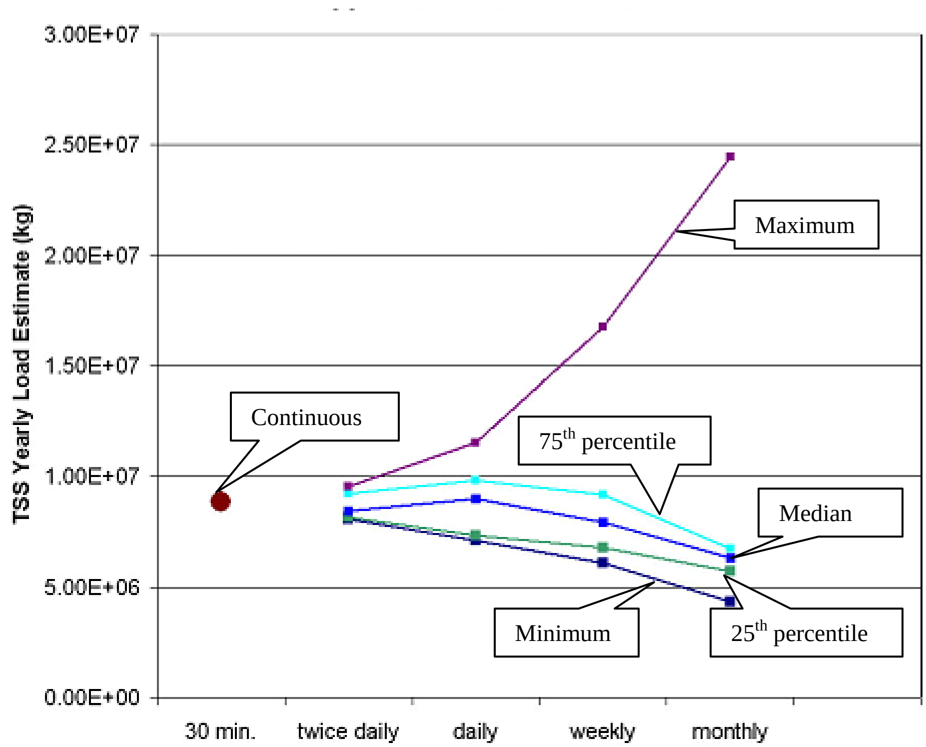


FIGURE 4.15. Total Suspended Solid Load Estimate for Each Sampling Interval (Mesner *et al.* 2007).

2 in Table 4.3). If the characteristics of yearly total flow are similar to those of yearly load of nutrient or particles, the yearly load from the R-D method may be close to the yearly load from daily observations and this estimated year load may be acceptable.

SUMMARY AND CONCLUSIONS

The R-D method consists of three steps: 1) creation of estimated distribution based on the distribution of observations, 2) estimation of time series ranking of predictions ,and 3) assignment of predictions to each date.

The first step creates CF plot from observations. A large number of predictions may be reconstructed based on this CF plot. The second step calculates the CFP time series of predictions based on the observed CFP time series by interpolation. The CFP time series may be improved by adding estimated annual peak CFPs before interpolation. The annual peak CFP is determined by optimizing the extended CFP slopes. The third step assigns predictions to simulation dates by matching the rank of prediction within an estimated distribution to the rank of optimized CFP time series.

The estimated distribution from CF plot of observations was similar to the distribution of original data. Optimizing the CFP time series by calibrating extended CFP slopes enhanced the agreement of time series of predictions with time series of original data.

The estimated time series by the R-D method were closer to the original time series than those estimated by simple interpolation, and the R-D method was more powerful for the data set collected with a longer sampling block.

The R-D method may be used to reduce the sampling frequency keeping the same error and reducing the measurement cost.

LITERATURE CITED

Brannan, K., S. Mostaghimi, G. Yagow, T. Dillaha, B. Benham, and R. Zeckoski, 2005. Chapter 4 TMDL Case Studies, In Total Maximum Daily Load Approache and Challenges, Tamim Younos (Editor), Penn Well Corporation, Tulsa, Oklahoma, p101-168.

Cusimano, R.F., S. Hood, and J. Liu, 2002. Quality Assurance Project Plan: Lake Whatcom TMDL Study, Washington State Department of Ecology. 47 pp.

Helsel, D.R., 2005. Nondetects and Data Analysis. John Wiley & Sons, Hoboken, New Jersey, 250 pp.

Kaplan, E.L. and P. Meier, 1958. Nonparametric Estimation from Incomplete Observation. Journal of the American Statistical Association 53(282):457-481.

Kruskal, W.H. and W.A. Wallis. 1952. Use of Ranks in One-criterion Variance Analysis. Journal of American Statistics Association 47(260):583-621.

Microsoft Inc, 2004. Using Office XP, http://www.microsoft.com/education/officexptutorial.mspx. *accessed* November, 2004.

Minitab Inc., 2006, Minitab® 14 http://www.minitab.com/support/macros/default.aspx?action=display&cat=non

Mesner, N.O., J.S. Horsburgh, D.K. Stevens, D.L. Sorensen, R.J. Ryel, and D. Jackson-Smith (2007). Evaluation of the Strengths and Weaknesses of Water Quality Monitoring Techniques, USDA_CSREES National Water Conference Presentation, 2007. Savannah, Georgia.

USEPA, 1998, Guidance for Data Quality Assessment. Practical Methods for Data Analysis; EPA/600/R-96/084.

CHAPTER 5

BAYESIAN NETWORK TO EVALUATE EFFECTS OF THE LITTLE BEAR RIVER WATER QUALITY CONSERVATION PROJECT

ABSTRACT

The Little Bear River watershed, Northern Utah is on a high-priority list of watersheds that are being adversely affected by nonpoint source pollution. Two Bayesian Networks, BN 1 (Bayesian Network above Hyrum Reservoir) and BN 2 (Bayesian Network below Hyrum Reservoir) were built to simulate the effect of the Little Bear River Conservation Project (LBRCP) and exogenous variables (point sources, agricultural land use and annual precipitation) on water quality to explore the causes of an observed reduction of Total Phosphorus (TP) concentration since 1990 at the mouth of the Little Bear River. Although BN1 and BN2 provided evidence that Little Bear River Conservation Practice (LBRCP) reduced the TP load in the river, this reduction explains only small decreases of the TP load and TP concentration. BN1 and BN2 showed evidence that agricultural land use, point sources and annual precipitation variables significantly increase TP load from the subwatershed to the stream. They also provided evidence that the point source variable and annual precipitation variable increased TP loads including both upstream load and subwatershed load into the Hyrum and Cutler Reservoir. The effect of the LBRCP appeared to be larger for wet precipitation conditions than for dry precipitation conditions.

INTRODUCTION

The Little Bear River in Cache County, Northern Utah flows from southeast to northwest, bounded by mountains, and drains to Cutler Reservoir, west of Logan, UT (Figure 5.1). This watershed has two major headwaters, the East Fork and the South Fork. The Little Bear Watershed has a drainage area of approximately 195,096 acres of which approximately 72 % is range or forest, 23 % is cropland or pasture, 2% is urban area, and remainder is water body and riparian (Utah DNR, 2004). The Little Bear River watershed is on a high-priority list of watersheds that are being adversely affected by nonpoint source pollution (Chess, 2000). The Little Bear River Steering Committee found cropland and pastures were significant sources of nutrients (Chess, 2000). In many fresh waters, an increase of phosphorus may cause algal blooms because phosphorus is often limiting (Mason, 2002). A Total Maximum Daily Load (TMDL) for the Little Bear River targets reduction of phosphorus (Utah DEQ, 2000).

Since the USDA started the Little Bear River Conservation Project (LBRCP) in 1990 (USEPA, 2006), phosphorus concentrations have reduced gradually at the river outlet. Our goal is to explore what caused the nutrient concentration reduction because exogenous forces (e.g. land use changes and climate changes) and on farm nutrient management can affect the water quality at management sites.

Bayesian Networks (BNs) were designed to accept and process inputs from varied sources: observations, model results, expert judgment, scenario types and a variety of other non-numerical inputs (Marcot *et al.*, 2001; Varis and Jussila, 2002; Borsuk *et al.*, 2003). BNs are designed to evaluate the effects of two or more variable combinations

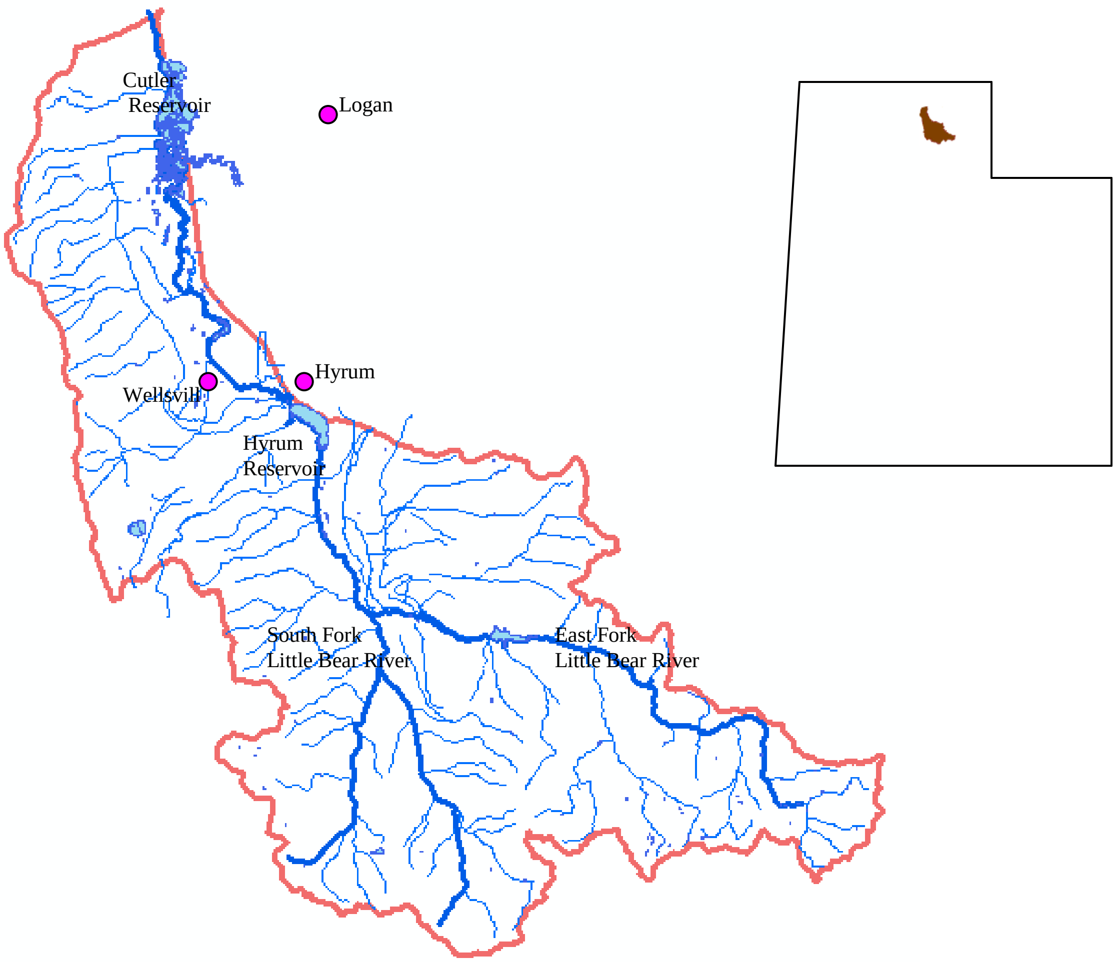


Figure 5.1 | Little Bear River Watershed located in Northern Utah (EMRG, 2004; USGS, 2004; USEPA, 2004).

(scenarios) on other variables (Marcot *et al.*, 2001; Ames, 2002). In this chapter, the effects of the combination of annual precipitation and conservation practices on the TP load and TP concentration in the stream were evaluated. BNs have been used to estimate the effect of future nutrient management activities using existing data (Marcot *et al.*, 2001; Borsuk *et al.*, 2003; Ames, 2005). In this chapter, BNs are discussed for evaluation of the effect of conservation practices and exogenous factors (precipitation, agricultural landuse area and point source loads) on phosphorus loads and concentrations into Hyrum and Cutler Reservoirs.

METHODS

Bayesian Network

A Bayesian Network (BN) is a probabilistic network model based on graphical relationships among variables (Castillo *et al.*, 1997). In a BN, the relationships between parent variables and child variables are logically expressed in a link and node structure where the state of the parent node predicts the state of the child node (Jensen, 1996). Conditional Probability Tables (CPTs) show the probability of each discrete state, given the states of any parent nodes (Marcot *et al.*, 2001). The marginal probabilities are calculated using CPTs.

BN have been used historically in water quality assessment. For example, Reckhow (1999) constructed a BN model for anoxia probability. In Reckhow's BN, arrows connected variables specifying the conditional dependences (Figure 5.2). For example, the percentage of nitrogen loading reduction depends on the percentage of forest buffer. Reckhow evaluated the effect of forest buffer on anoxia probability in the BN.

Conditional probabilities of child variables (x) given parent variables (y) are estimated by statistical model results based on observations and scientific judgment in Reckhow's BN. Marginal probabilities of child variables are obtained using joint probabilities of parent variables and conditional probability of the child variable given the parent variable (Equation 5.1b).

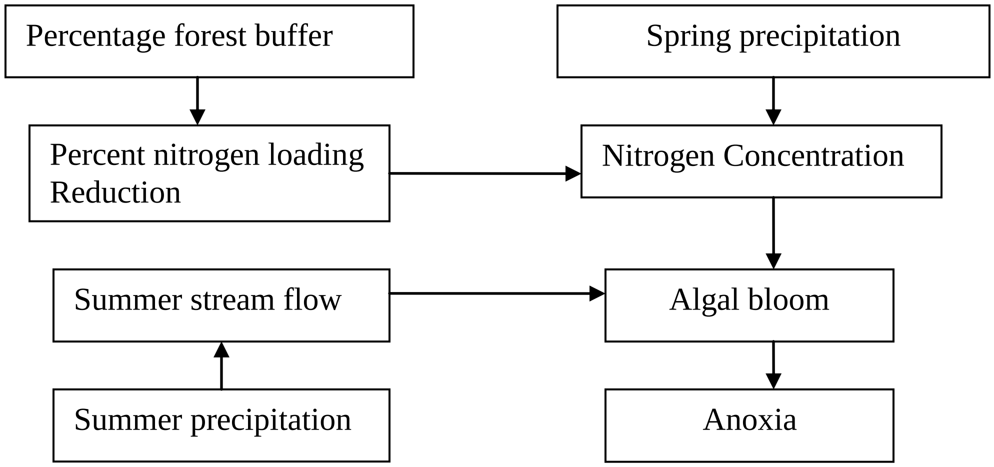


Figure 5.2 | Schematic of an anoxia model.

$$p(y_i) = \sum_{x_1, x_2 \ldots\ldots x_{i-1}, x_{i+1} \ldots x_n} p(y_i, x_1, \ldots\ldots\ldots, x_n) \tag{5.1a}$$

$$= \sum_{x_1, x_2 \ldots\ldots x_{i-1}, x_{i+1} \ldots x_n} p(y_i \mid x_1 \ldots, x_n) p(x_1, \ldots\ldots, x_n) \tag{5.1b}$$

where $p(x_1,\ldots..,x_n)$ is the joint probability (if number of parent variables >= 2) or marginal probability (if number of parent variable=1) of parent variables, $x_1,\ldots..,x_n$, and $p(y_i \mid x_1\ldots,x_n)$ is the conditional probabilities of child variables (y_i) given parent variables (Castillio *et al.*, 1997). For example, marginal probability of <5% nitrogen reduction is

P(<5%)= p(<5%| 70-80% forested buffer) × p(70-80% forested buffer)
+ p(<5%| 80-95% forested buffer) × p(80-95% forested buffer)
+ p(<5%| 95-100 % forested buffer) × p(95-100 % forested buffer).

Reckhow calculated the anoxia probability for 95-100 % forest buffer of the entire stream (p(95-100% forested buffer)=1) as 0.27, while the simple marginal probability at the proposed percentage of forest buffer was 0.3. In BN software, conditional probabilities are organized in CPTs specifying the relation between parent and child variables (Norsys Software Corp., 1997).

Water-related data

The databases were constructed to support BNs by calculating the CPTs. The existing and calculated inputs were TP concentration, flow, precipitation, agricultural land use area, and water quality conservation project. Because the water quality characteristics above and below a reservoir may significantly differ from each other, two separate networks were constructed: above Hyrum Reservoir (BN1) and below Hyrum Reservoir (BN2) (Figure 5.3). Each network was connected to separate data bases. In the data base for BN1, the sampling locations included East Fork (4905750), South Fork (4905740), confluence between East Fork and South Fork (4905700) and inlet into Hyrum Reservoir (4905670) and a point source from a fish farm. One permitted TP point source (one of two discharges from a fish farm) is ignored. There is no flow or TP concentration data for the point source before March 1991 in EPA STORET (USEPA, 2005). If that point source TP load is a parent variable of a child variable, it is difficult to calculate the reliable values of conditional probability of the child variable given the point source TP load and other parent variables before 1991 because of the lack of point source TP load data. The TP point load (median = 0.744 lb/d) was much smaller than main TP point load (median = 5.464 lb/d) so ignoring this point source TP load may not affect child variable significantly.

In the data base for BN2, the sampling locations included the outlet from Hyrum Reservoir (4905650), the inlet to Cutler Reservoir (4905000) and the discharge from the Wellsville Lagoons (Figure 5.3). In calculation of CPTs, all data sets which have data of child variables but not data of parent variables are ignored. Two permitted discharges' loads were removed from parent variables to avoid ignoring all data sets which have data of child variables but not data for these discharges. Many portions of Northern Utah Manufacturing discharges was near 0 (84 percentile = 0) with no TP concentration data at those measurement dates. The TP load of the other discharge from Northern Utah Manufacturing was small (median TP load= 0.72 lb/d). Therefore, removing these two discharges may be acceptable.

The TP loads were calculated by multiplying the flow by the TP concentrations and converting the units to lb/day. Stream flows of sampling location 4905670 (Above Hyrum) came from USGS flow data base for USGS gage 10105900 and 10106000 (USGS, 2006). Other flows and TP concentrations at the stream sampling locations and the point source outlets came from the EPA STORET data base (USEPA, 2005).

For some data, there was a mismatch in sampling or measurement time. If the time difference between observations was equal to or smaller than 1 day, they were used as if they were from the same date. For example, flow and TP data available on March 23, 1993 at Cutler, but flow and TP data for Hyrum release were available on March 22, 1993. Because time difference is ≤ 1 day, those data were used as if they came from the same day. Because flash flooding is rare in the Little Bear River Watershed, the flow and

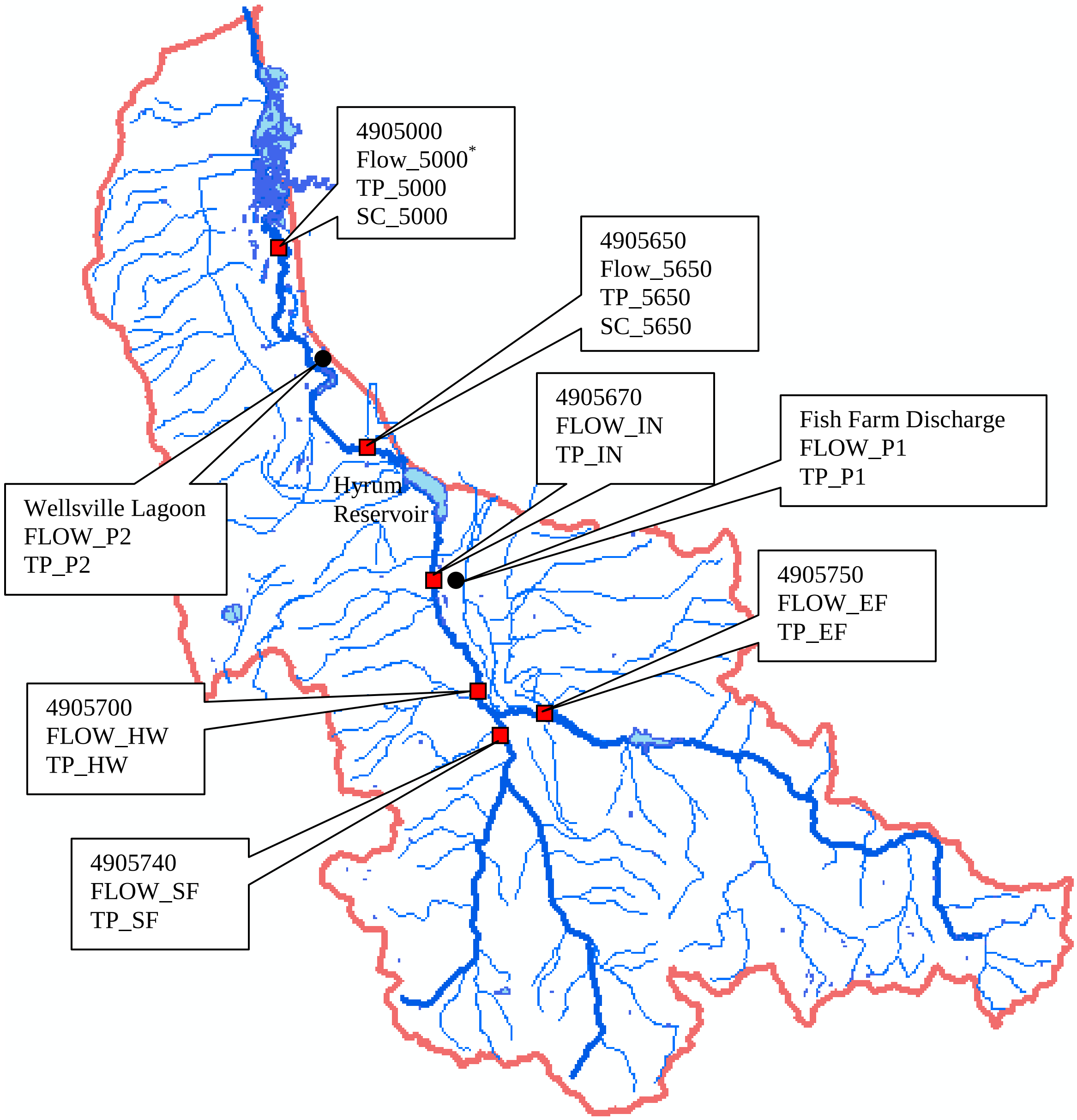


Figure 5.3 | Sampling locations of Little Bear River watershed. (EMRG, 2004; USGS, 2004; USEPA, 2004). (● : point source outlet sampling location, ■ :stream sampling location), * : 490 was removed from variable names for clarity. where Flow_xx = flow (cfs), TP_xx = TP concentration (mg/L), SC_xx = Specific Conductance (umho/cm), xx is last four digits of station ID.

water quality changes gradually. Therefore, merging data from date ≤ 1 day apart is acceptable.

It is possible to calculate the TP load when there are both flow and TP concentrations. Linear regression between upstream values and downstream values can be used to fill in the missing flows or TP concentrations (Table 5.1). The censoring level (below quantitation limit) of TP is 0.02 mg/L in EPA STORET (USEPA, 2005). Because both TP_HW and TP_SF values were below 0.02 mg/L on some sampling dates, the standard regression method (Pearson's) was not applicable to estimate slopes and

Table 5.1 | Regressions between variables of missing values (response variables) and predicted variables (Variables are defined on Figure 5.2)

Response variable (Y)	Predicted Variable (X)	Number of Pairs (censored/ total)	Regre-ssion Method	Slope	Inter-cept	Correlation Value, ( ): p-value
TP_SF (4905740)	TP_HW (4905700)	30/65	Theil-Sen	0.682	0.007	τ_b =0.569 (<0.001)
TP_EF (4905750)	TP_ HW (4905700)	48/71	Theil-Sen	0.382	0.001	τ_b = 0.481 (<0.001)
Flow_ IN1 (4905670)	Flow_HW (4905700)	0/155	Pearson	*0.996	*0.905	r^2= 0.998 (<0.001)
TP_ IN1 (4905670)	TP_ HW (4905700)	29/75	Theil-Sen	0.951	0.0196	τ_b = 0.371 (<0.001)
ln(TP_5650) (4905650)	TP_5000 (4905000)	4/84	MLE-lognormal	2.912	-3.231	r^2=0.166 (<0.001)
Flow_5650 (4905650)	Flow_5000 (4905000)	0/39	Pearson	*0.723	*-15.5	r^2=0.889 (<0.001)
Flow_5650 (4905650)	SC_ 5650 (4905650)	0/52	Pearson	*-0.203	*125	r^2=0.642 (<0.001)
Flow_ 5000 (4905000)	SC_ 5000 (4905000)	0/144	Pearson	*-0.317	*249	r^2=0.401 (<0.001)

*Regression equation after removing outlier with over 2.0 studentized residual twice.

r: Correlation coefficient , τ_b: Kendall tau-b correlation coefficient (Kendall, 1970) intercepts using these doubly censored pairs. In these cases, the Theil-Sen regression (Theil, 1950) was used to handle the censored pairs in estimation of slopes and intercept between TP_HW and TP_SF.

On four sampling dates, TP_5650 values were below 0.02 mg/L and TP_5000 values were uncensored. Maximum Likelihood Estimation (Slyman *et al.*, 1994) was used to estimate slope and intercept between the predicted variable (TP_5000) and the response variable (TP_5650) instead of Pearson's regression (Table 5.1; Figure 5.3).

The annual precipitation data (PRECIP) were obtained from the Western Regional Climate Center data base (WRCC, 2006). The precipitation measurement location was Logan Radio KVNU (Station No. 425182). Utah DNR (Utah Department of Natural Resources) provided water-related Landuse data files for 1986, 1996 and 2003 in GIS and tabular form for Landuse in the Little Bear River watershed (Utah DNR, 2004). Because the Little Bear River technical advisory committee consisting of local, state and federal resources agencies and representations from Utah State University defined cropland and pastures as significant sources of nutrients in the Little Bear River watershed (Chess, 2000), the agricultural land use area was emphasized. The agricultural area in the Above Hyrum subwatershed and Below Hyrum subwatershed were estimated for the year between 1986, 1996 and 2003 by interpolating agricultural area in 1986, 1996, 2003. The agricultural landuse areas before 1986 or after 2003 were estimated by extrapolating agricultural areas in 1986 or 2003 (Figure 5.4).

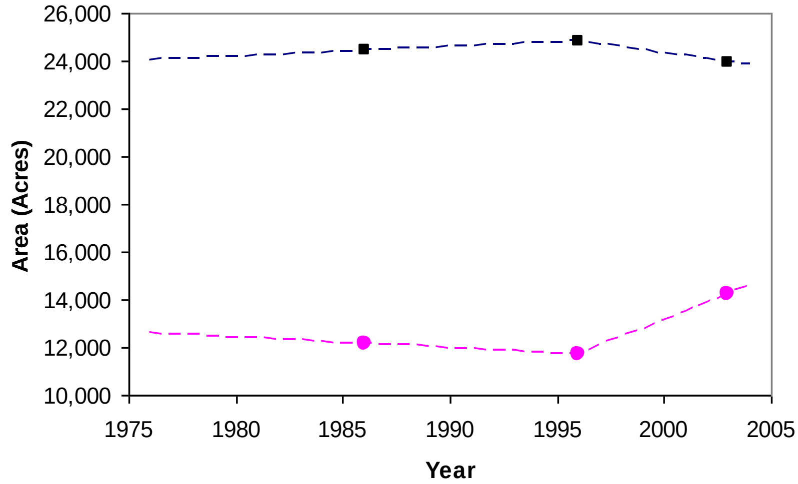


Figure 5.4 | Agricultural land use area in Little Bear River Watershed (● : Above Hyrum Reservoir, ■ :Below Hyrum Reservoir).

Bayesian Network (BN) construction

Netica version 3.17 (Norsys Software Corp., 1997) was used to build the two BNs. Contingency tables were automatically produced by Netica from each database. The first BN (BN1) estimated the effects of LBRCP and exogenous variables, agricultural landuse, annual precipitation and point source TP load on the TP load and TP concentrations at the inlet to Hyrum Reservoir (Figure 5.5, Table 5.2). The second BN (BN2) estimated the effects of LBRCP and the same exogenous variables on the TP load and TP concentrations at the inlet to the Cutler Reservoir (Figure 5.6, Table 5.3).

TP concentration, sum of dissolved organic phosphorus, particulate phosphorus and phosphate concentrations have been used as a water quality criterion even though only soluble reactive phosphorus (soluble inorganic phosphorus) is readily available to plants in aquatic ecosystems. Phosphate (PO_4^{3-}) is a dominant form of soluble reactive phosphorus but it is often difficult to determine the concentration of phosphate because the standard analysis method detects a variable group of phosphorus compounds as well as phosphate (Chapra, 1997; Dodds, 2002). Baker *et al.* (2008) suggested 0.04 mg/L as allowable TP concentration based on research about phosphorus uptake in East Canyon Creek. According to state water quality criterion of Utah (Utah DAR, 2000), the target TP concentration is 0.05 mg/L in the Little Bear River. This concentration is not far from Baker's suggestion. Because this TP concentration was also the target of the Little Bear River TMDL (Utah DEQ, 2000), the hypothesis variable of BN 1 is the TP

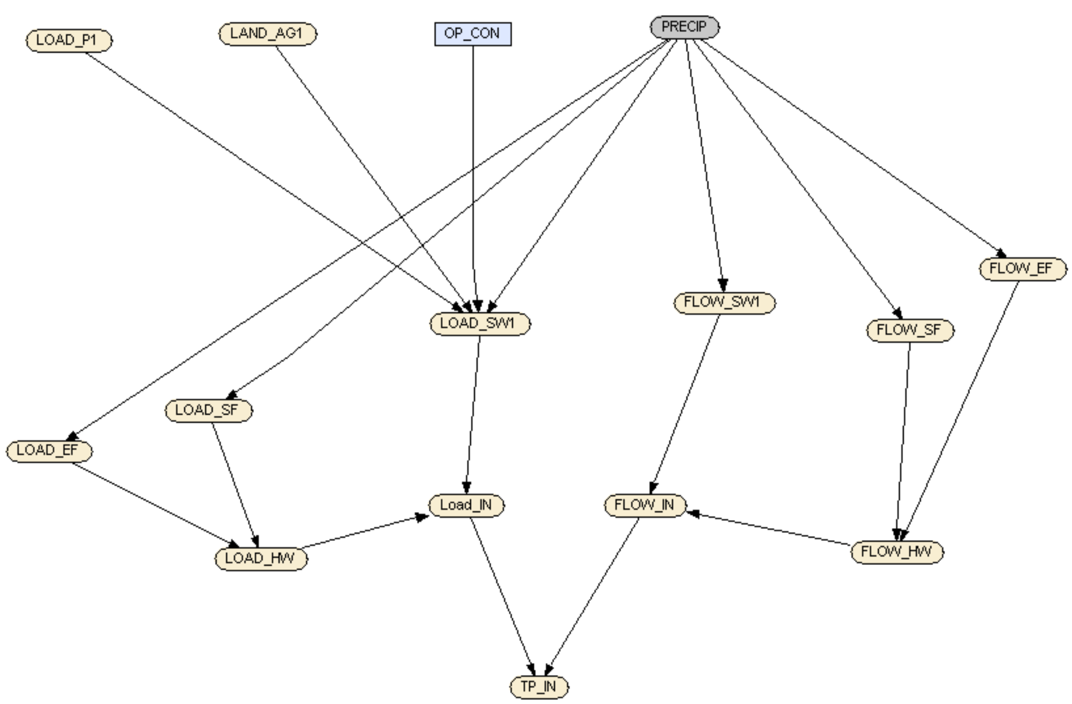


Figure 5.5 | Little Bear River BN 1 (above Hyrum Reservoir) for LBRCP and exogenous variable effect evaluation

Table 5.2 | Critical variables for evaluation of conservation project in the Little Bear River Watershed above Hyrum Reservoir

Variable Name	Description	Type	Sources
OP_CON	Conservation Project Option, Pre: Before starting the Project (1974-1989), Post: After starting the Project (1990-2004)	Decision	
LAND_AG1	Area of Agricultural Land Use	Exogenous	Utah DNR
PRECIP	Annual Precipitation for each water year	Exogenous	Western Region Climate center
LOAD_P1	TP load (lb/day) from point source	Exogenous	EPA STORET
LOAD_SF	TP load (lb/day) from the South Fork (Location 4905740)	State	EPA STORET
LOAD_EF	TP load (lb/day) from the East Fork (Location 4905750)	State	EPA STORET
LOAD_HW	TP load (lb/day) at the confluece of the South and East Fork (Location 4905700)	State	EPA STORET
LOAD_IN	TP load (lb/day) into the Hyrum Reservoir (Location 4905670)	State	EPA STORET
LOAD_SW1	TP point and nonpoint load from subwatershed above the stream reaches between location 4905700 (confluence of the South Fork and East Fork) and location 4905670 (inlet to the Hyrum Reservoir)	State	Estimated
FLOW_SF	Flow (cfs) from the South Fork (Location 4905740)	State	EPA STORET
FLOW_EF	TP load (lb/day) from the East Fork (Location 4905750)	State	EPA STORET
FLOW_HW	Flow (cfs) at the confluent of the South and East Fork (Location 4905700)	State	EPA STORET
FLOW_IN	Flow (cfs) into Hyrum Reservoir (Location 4905670)	State	USGS Flow data
FLOW_SW1	Flow from subwatershed above the stream reaches between location 4905700 (confluent of the South Fork and East Fork) and location 4905670 (inlet to Hyrum Reservoir)	State	Estimated
TP_IN	TP concentration (mg/L) into Hyrum Reservoir (Location 4905670)	State	EPA STORET

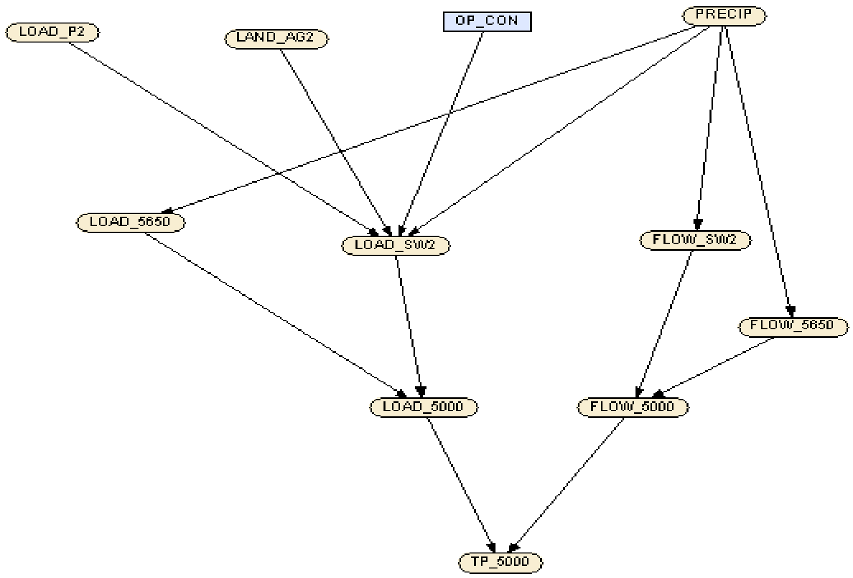


Figure 5.6 | Little Bear River BN 2 (below Hyrum Reservoir) for the LBRCP and exogenous variable effect evaluation.

concentration at the inlet to Hyrum Reservoir (TP_IN), and the hypothesis variable of BN2 is the TP concentration at the inlet to Cutler Reservoir (TP_5000).

There are two variable groups, the TP load group and the flow group in BN1 and BN2. Referring to Figure 5.3, the TP load group includes East Fork outlet load (LOAD_EF), South Fork outlet load (LOAD_SF), load at confluence between the East Fork and South Fork (LOAD_HW), and the inlet load to Hyrum Reservoir (LOAD_IN) in BN1. Upstream TP load variables were linked to down stream TP load variables. For example, LOAD_EF and LOAD_SF were linked to LOAD_HW (Figure 5.5). In the same way, the flow group included flows of all sampling locations in East Fork, South Fork and main stream Little Bear River, and upstream flow variables were linked to those downstream (Figure 5.5).

In BN1, the LBRCP option and exogenous variables were linked to subwatershed load (LOAD_SW1), the TP load from point sources and non point sources to the stream

Table 5.3 | Critical variables for evaluation of conservation project in the Little Bear River Watershed below Hyrum Reservoir

Variable Name	Description	Type	References
OP_CON	Conservation Project Option, Pre: Before starting the Project (1974-1989), Post: After starting the Project(1990-2004)	Decision	
LAND_AG2	Area of Agricultural Land Use	Exogenous	Utah DNR (2004)
PRECIP	Annual Precipitation for each water year	Exogenous	Western Region Climate Center (2006)
LOAD_P2	TP load (lb/day) from point source	Exogenous	EPA STORET (2005)
LOAD_5650	TP load (lb/day) at the effluent from Hyrum Reservoir (Location 4905650)	State	EPA STORET
LOAD_5000	TP load (lb/day) into Cutler Reservoir (Location 4905000)	State	EPA STORET
LOAD_SW2	TP point and nonpoint load from subwatershed above the stream reaches between location 4905650 and location 4905000 (inlet to Hyrum Reservoir)	State	Estimated
FLOW_5650	Flow (cfs) at the effluent from the Hyrum Reservoir (Location 4905650)	State	EPA STORET
FLOW_5000	Flow (cfs) into Cutler Reservoir (Location 4905000)	State	EPA STORET
FLOW_SW2	Flow from subwatershed above the stream reaches between location 4905650 and location 4905000	State	Estimated
TP_5000	TP concentration (mg/L) into Cutler Reservoir (Location 4905000)	State	EPA STORET

between location 4905700 and 4905670. LOAD_SW1 is calculated by subtracting the TP load at location 4905700 from TP load at location 4905670. LOAD_SW1 is a parent variable of LOAD_IN. In BN1, the annual precipitation (PRECIP) is linked to subwatershed flow (FLOW_SW1), which is the flow from the watershed to the stream between location 4905700 and 4905670 and is calculated by subtracting flow at location 4905670 by flow at location 4905700. Similarly, FLOW_SW1 is a parent variable of FLOW_IN. Finally, the TP loads (LOAD_IN) and flows (FLOW_IN) into Hyrum Reservoir were connected to TP concentration (TP_IN) into the reservoir (Figure 5.5). In BN2, variable are connected in the same way as BN1. Upstream TP load variables were linked to down stream TP load variables, and upstream flow variables were linked to those downstream (Figure 5.6). The upstream water in BN2 is the outlet from Hyrum Reservoir.

Categorizing variable state for BN 1

The conservation project option factor (OP_CON) is a decision variable. Pre conservation is defined as before starting the conservation practices and Post conservation is defined as after starting the conservation practices.

There are three exogenous variables, agricultural land use area (LAND_AG1), point source TP load (LOAD_P1) and precipitation (PRECIP). LAND_AG1 and LOAD_P1 were categorized as *H (High)* or *L (Low)*. PRECIP was categorized as *D* (Dry) and *W* (Wet) (Table 5.4)

The variables of TP load and flow groups have three categories, L (Low :smaller

Table 5.4 | The categories of exogenous variables in the Little Bear River Wastershed above Hyrum

Variable	Category		Unit
	L,D (<50 percentile)	*H,W*(>=50percentile)	
LAND_AG1	<12268	>=12268	Acres
LOAD_P1	<5.46	>=5.46	lb/day
PRECIP	<15.4	>=15.4	In

than 33 percentile value), M (Medium :from 33 percentile value to smaller than 67 percentile value) and H (High: equal to or larger than 67 percentile value). Conditional probabilities of L, M and H category of LOAD_SW1 were calculated for each category combination of OP_CON, LAND_AG1, LOAD_P1 and PRECIP in Netica (Appendix F (a)). In the same manner, conditional probabilities, *p(LOAD_EF|PRECIP)*, *p(LOAD_SF|PRECIP)*, *p(LOAD_HW|LOAD_SF, LOAD_EF)*, *p(LOAD_IN|LOAD_SW1, LOAD_HW)* were calculated in Netica (Appendix F (a)-(e)) .

The conditional probability of flow variables, *p(FLOW_SW1|PRECIP)*, *p(FLOW_EF|PRECIP)*, *p(FLOW_HW|FLOW_EF,FLOW_SF)*, *p(FLOW_IN|FLOW_SW1,FLOW_HW)* are calculated in the same manner of TP load variables (Apendix F (f)-(j)). The TP concentration at the inlet to the Hyrum Reservoir (TP_IN) was categorized as H,M and L using quantitation limit (0.02 mg/L) and target concentration (0.05 mg/L). (Appendix F (k)).

Categorizing variables of BN 2

The option of conservation project (OP_CON) is the only decision variable (Figure 5.6). Pre is before starting the conservation practices and Post is after starting the conservation practices. There are three exogenous variables, agricultural land use area (LAND_AG2), point source TP load (LOAD_P2), and annual precipitation (PRECIP). OP_CON and exogenous variables have no parent node (Figure 5.6).

Database of BN2 has no point source TP load data before 1990 (before starting LBRCP). Netica cannot calculate conditional probabilities of *p(LOAD_SW2|OP_CON, LAND_AG2, LOAD_P2, PRECIP)* for the case of OP_CON= Pre because all of

LOAD_P2 was empty in data combination of four variables, OP_CON, LAND_AG2, LOAD_P2, PRECIP when OP_CON is Pre. A challenge of BN2 is using LOAD_SW1 (Subwatershed TP load) contingency table of BN1 for conditional probability *p(LOAD_SW2|OP_CON, LAND_AG2, LOAD_P2, PRECIP)* of BN2.

BN1 has one point source, one major stream into the subwatershed below the confluence of the East and South Fork, and one stream into a reservoir. BN2 has one point source, one major stream into a subwatershed and one stream into a reservoir. In BN1, the major TP load sources within the subwatershed boundary are the point source (Fish Farm) and agricultural nonpoint source. In BN2, the major TP load sources within the subwatershed boundary are one point source (Discharge of Wellsville Lagoon) and agricultural nonpoint source. Because the pattern of flows and TP loads of BN2 is similar to those of BN1, it may be acceptable using CPT of *p(LOAD_SW1|OP_CON, LAND_AG1, LOAD_P1, PRECIP)* of BN1 for CPT of *p(LOAD_SW2|OP_CON,LAND_AG2,LOAD_P2,PRECIP)* of BN2.

The agricultural area below Hyrum (LAND_AG2) increased and decreased in the range of 23827 to 24803 acres from 1976 to 2004 acres while the agricultural area above Hyrum (LAND_AG1) decreased and increased in the range of 11686 to 14588 acres. Because all values of LAND_AG2 were out of range of any category of LAND_AG1, the category boundaries of LAND_AG1 needed to be adjusted before being used as category boundaries of LAND_AG2. The median of the agricultural area was 12,268 acres in BN1 and the median area was 24,370 acres in BN2. The ratio, 24,370/12,268 acres, is 1.99. The BN1 category boundaries of LAND_AG1, LOAD_P1 and LOAD_SW1 were multiplied by 1.99 and these scale up category boundaries was used in CPT for *p(LOAD_SW2|OP_CON, LAND_AG2, LOAD_P2, PRECIP)* of BN2 (Table 5.5). The conditional probabilities for each categorical combination, *p(LOAD_SW1|OP_CON, LAND_AG1, LOAD_P1, PRECIP)* of BN1, were used for the conditional probability for the same categorical combination, *p(LOAD_SW2|OP_CON, LAND_AG2, LOAD_P2, PRECIP)* of BN2 (Appendix F (a), G (a)). PRECIP was categorized as *D* and *W* using a median value (15.4 inch).

Each TP load or flow variable has three categories, L (Low: smaller than 33 percentile value), M (M: from 33 percentile value to smaller than 67 percentile value) and H (High: equal to or larger than 67 percentile value) (Appendix G). The TP concentration at the inlet to the Cutler Reservoir (TP_5000) was categorized as *H(high), M(medium)* and *L(low)* using target concentration (0.05 mg/L) and 75th percentile

Table 5.5 | Adjustment of category boundaries of variables for BN2 from those for BN1

Variables	Category	Above Hyrum (BN1)	Below Hyrum (BN2) after Scale adjustment	Below Hyrum (BN2) before Scale adjustment
Agricultural Land Use(LAND_AG) as Acres	L H	<12268 >=12268	<24370 >=24370	<24370 >=24370
Point Source TP load (Load_P) as lb/day	L H	<5.46 >=5.46	<10.86 >=10.86	<1.681 >=1.681
Total subwatershed TP load (LOAD_SW) as lb/day	L M H	< 3.53 3.53-9.51 >=9.51	<7.02 7.02-18.92 >=18.92	< 18.42 18.42-35.97 >=35.97

concentration (0.15 mg/L) (Appendix G (g)).

Bayesian Network simulation

The purpose of the Little Bear River BN is to evaluate the effects of the conservation practices and exogenous variables on the TP load and TP concentration into Hyrum and Cutler Reservoirs. LOAD_SW1 (the marginal probability distributions of categories of subwatershed TP load), LOAD_IN (the marginal probability of category of TP load into Hyrum Reservoir) and TP_IN (the marginal probability of category of TP concentration into Hyrum Reservoir) for Pre condition of OP_CON (conservation practice option) were compared to the marginal probability distribution of categories of LOAD_SW1, LOAD_IN and TP_IN (compared variables) for Post condition of OP_CON (evaluated variable) in BN1. When OP_CON was selected as Pre or Post in the conservation practice's effect tests in BN1, the probabilities of all variables directly linked (LOAD_SW1) and indirectly linked (LOAD_IN and TP_IN) to OP_CON in the network were calculated under the assumption of no conservation practice for Pre OP_CON or some conservation practice for Post OP_CON (Figure 5.7). The probability of each category for agricultural land area (LAND_AG1), Point TP load (LOAD_P1) and annual precipitation (PRECIP) came from all values of BN1 databases.

In the same manner, the effects of conservation practices on TP loads and concentrations were evaluated in BN2 (Figure 5.8). The effects of exogenous variables on TP load and concentrations were evaluated in BN1 and BN2 (Table 5.6). For example, in BN2, when annual precipitation conditions were selected as D (dry) or W (wet), the probabilities of all variables directly linked (LOAD_SW2, LOAD_5650, FLOW_SW2, FLOW_5650) and indirectly linked (LOAD_5000, FLOW_5000 and TP_5000) to

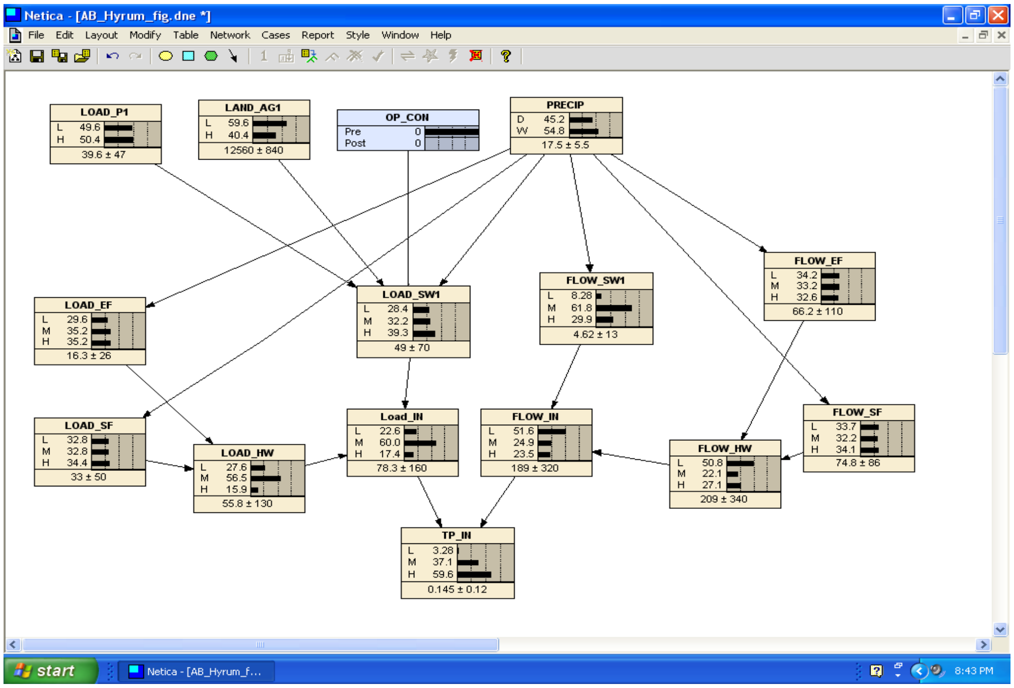


Figure 5.7 | The outputs from the Little Bear River BN above the Hyrum Reservoir (BN1) for pre conservation practices condition (OP_CON=Pre). A blue box is decision variable. Yellow boxes are exogenous and state variables. Each bar and number by each categories of each variable present the probability of that category.

PRECIP in the network were calculated under the assumption of all dry year (*p(D)*=1.0) for D PRECIP or all wet year (*p(W)*=1.0) for W PRECIP (Figure 5.8).

One interesting issue is under which annual precipitation condition (PRECIP) the conservation practices (OP_CON) had a larger effect on TP load and TP concentration. This is done by comparing how much the probability distribution of LOAD_SW1, LOAD_IN and TP_IN changed when selected OP_CON category was changed from Pre to Post under D and W condition of PRECIP in BN1 (Figure 5.9). For this task, OP_CON was selected as Pre or Post under D and W precipitation condition in BN1 or BN2. The probability of each category for agricultural land area (LAND_AG1 or

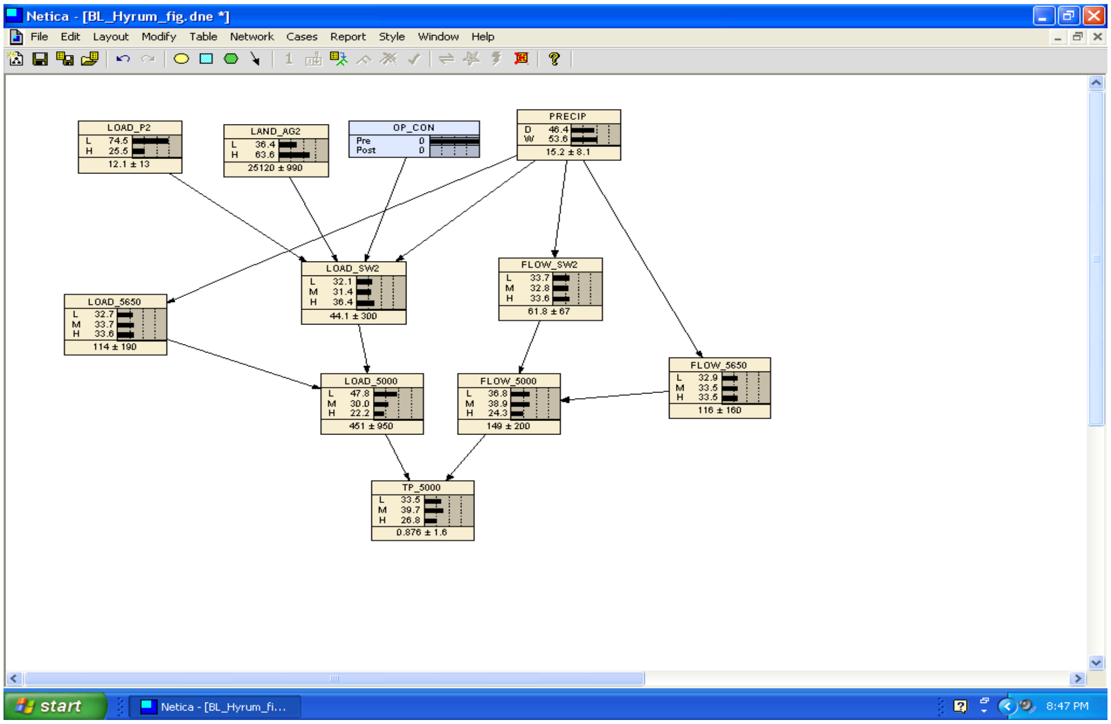


Figure 5.8 | The outputs from the Little Bear River BN below the Hyrum Reservoir (BN2) for pre conservation practices condition (OP_CON=Pre). A blue box is a decision variable. Yellow boxes are exogenous and state variables. Each bar and number by each categories of each variable present the probability of that category.

LAND_AG2) and Point TP load (LOAD_P1 or LOAD_P2) came from all values of BN1 or BN2 database because any category of these variables was not selected (Figure 5.9).

RESULTS

Effects of Conservation Project and exogenous factors on TP load and TP concentration on the stream above Hyrum Reservoir

The results of changing the BN1 variable for conservation practice (OP_CON) from Pre to Post are shown in Figure 5.10. Small differences are seen between the

Table 5.6 | The scenarios of variables for each simulation

	Evaluated Variable	Selected Category of variables				Compared variable
		OP_CON	LAND_AG1/ LAND_AG2	LOAD_P1/ LOAD_P2	PRECIP	
BN1	OP_CON	Pre/Post	None	None	None	LOAD_SW1, LOAD_IN, TP_IN
	LAND_AG1	None	L or H	None	None	LOAD_SW1, LOAD_IN, TP_IN
	LOAD_P1	None	None	L or H	None	LOAD_SW1, LOAD_IN, TP_IN
	PRECIP	None	None	None	L or H	LOAD_SW1, LOAD_IN, TP_IN, FLOW_SW1, FLOW_HW, FLOW_IN
BN2	OP_CON	Pre/Post	None	None	None	LOAD_SW2, LOAD_5000, TP_5000
	LAND_AG2	None	L or H	None	None	LOAD_SW2, LOAD_5000, TP_5000
	LOAD_P2	None	None	L or H	None	LOAD_SW2, LOAD_5000, TP_5000
	PRECIP	None	None	None	L or H	LOAD_SW2, LOAD_5000, TP_5000, FLOW_SW, FLOW_5650, FLOW_IN

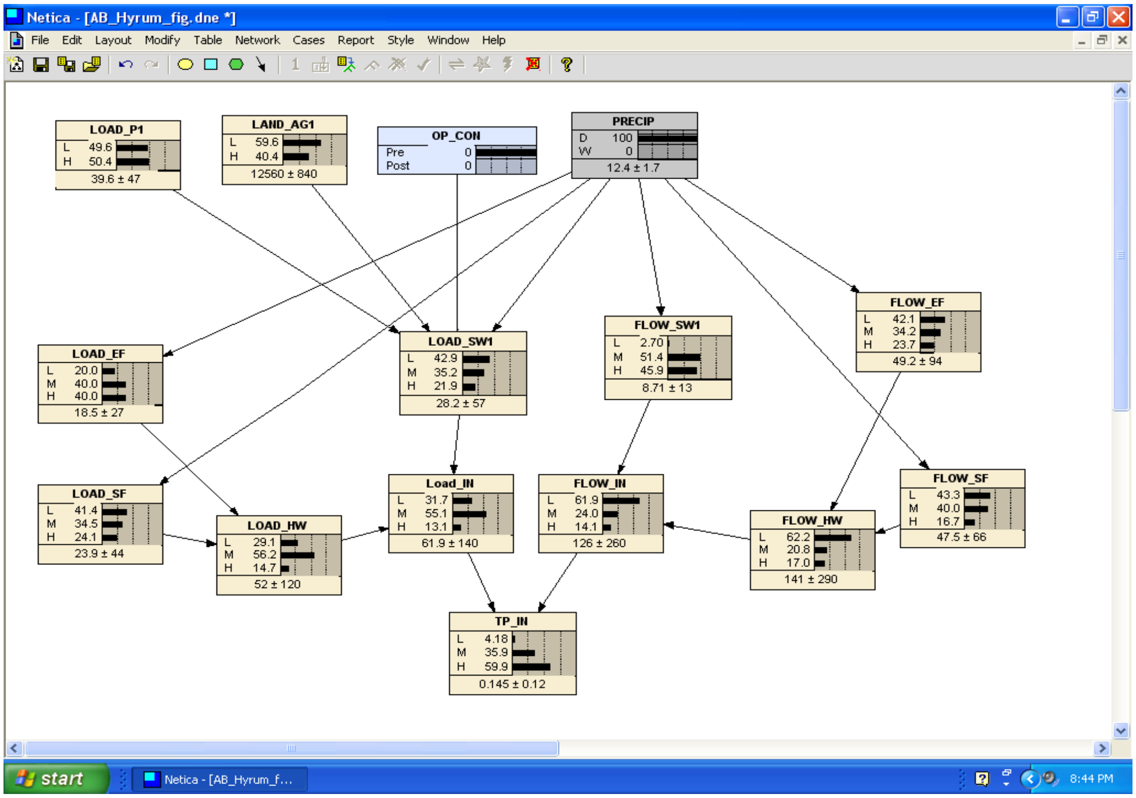


Figure 5.9 | The outputs from the Little Bear River BN above Hyrum Reservoir (BN1) for D annual precipitation and Pre LBRCP condition. A blue box is a decision variable. Yellow boxes are exogenous and state variables. Gray box means a specific category is selected. Each bar and number by each categories of each variable present the probability of that category.

predicted probabilities for each category of TP load (LOAD_IN) and TP concentration (TP_IN) into the Hyrum Reservoir even though the probability of the subwatershed TP load (LOAD_SW1) in low (L) category increased and the probability of medium (M) and high (H) for LOAD_SW1 decreased noticeably.

Figures 5.11, 5.12 and 5.13 show the effect of agricultural land area (LAND_AG1), point source load (LOAD_P1) and Precipitation (PRECIP) on the subwatershed load (LOAD_SW1), TP load (LOAD_IN) and TP concentration (TP_IN) at the reservoir's inlet. The agricultural land factor did not have a significant effect on

LOAD_IN or TP_IN, however, the *p(H)* for LOAD_SW1 increased and *p*(M) decreased, when the selection of category of LAND_AG1 was changed from *Low (L) to High (H)*. The LOAD_P1 had larger effect on LOAD_SW1, LOAD_IN and TP_IN than OP_CON or LAND_AG1, where *p(H)* for TP_IN (the probability of water quality criteria violation, TP_IN > 0.05 mg/L as TP) increased from 55.4% to 61.6% when LOAD_P1 was changed from L to H (Figure 5.12). More annual precipitation increased LOAD_SW1 and LOAD_IN because *p(H)* for LOAD_SW1 and LOAD_IN were higher and *p(L)* of those variables were lower for wet year than the dry year condition (Figure 5.13).

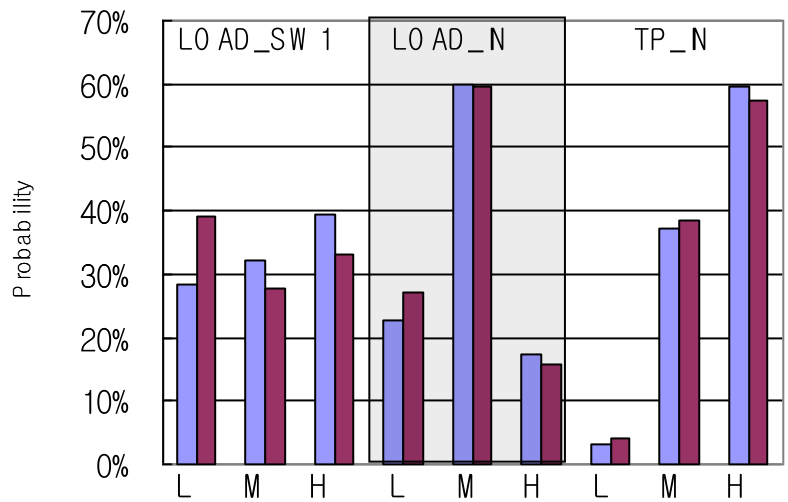


Figure 5.10 | The probabilities of TP load and TP concentration variables under selected conservation project options (OP_CON). Refer to Table 5.2 for variable descriptions. (■:OP_CON= Pre ■: OP_CON= Post, L,M, and H are categories for LOAD SW1, LOAD_IN and TP_IN).

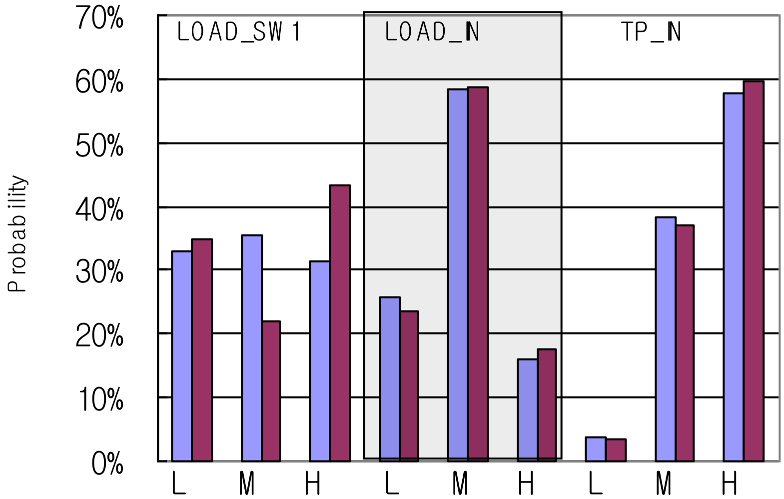


Figure 5.11 | The probabilities of TP loads and TP concentration variable under selected agricultural landuse category (LAND_AG1), (■:LAND_AG1= Low ■: LAND_AG1= High, L,M, and H are categories for LOAD SW1, LOAD_IN and TP_IN).

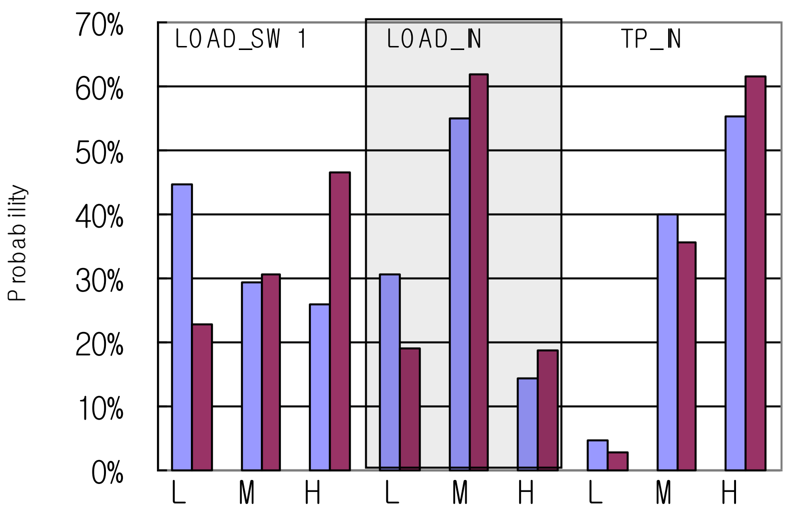


Figure 5.12 | The probabilities of TP loads and flow variable under selected point load category (LOAD_P1) (■:LOAD_P1= Low ■: LOAD_P1=High, L,M, and H are categories for LOAD SW1, LOAD_IN and TP_IN).

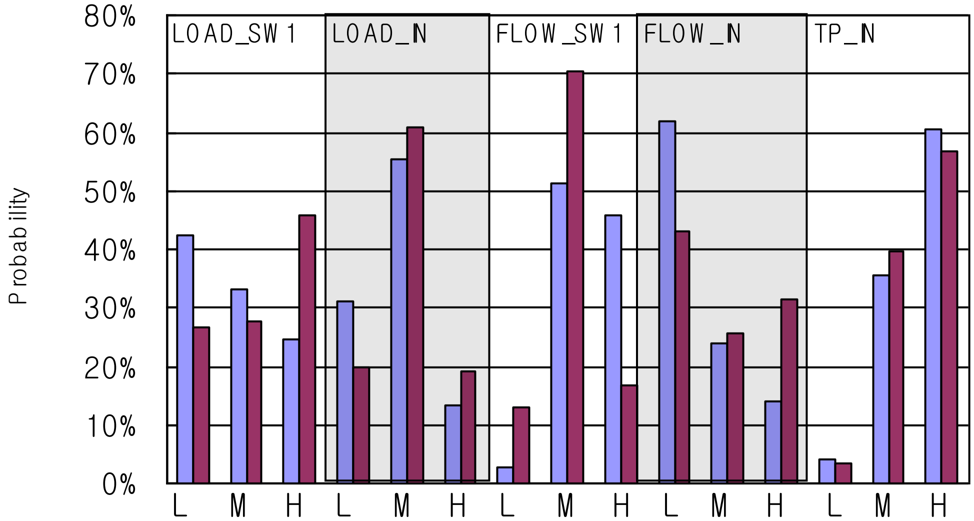


Figure 5.13 | The probabilities of TP load, TP concentration and flow variables under selected annual precipitation category (PRECIP) (■:PRECIP=Dry, ■: PRECIP=Wet, L,M, and H are categories for LOAD SW1, LOAD_IN, FLOW_SW1, FLOW_IN and TP_IN).

The probabilities *p(L)* and *p(M)* for subwatershed flow (FLOW_SW1) for dry years are noticeably lower than those for wet years (Figure 5.13). This seems to be unacceptable results. Because the 83.5 % of FLOW_SW1 values were between -1* and 1 cfs in the BN1 database and the category boundaries were very narrow (0 and 0.43 cfs), the categorizing of FLOW_SW1 was not effective to evaluate the effect of annual precipitation on FLOW_SW1. However, errors in FLOW_SW1 probability distribution may not cause significant error of the probabilities of the child variable, flow at the reservoir's inlet (FLOW_IN) categories. FLOW_IN is mainly controlled by confluence flow of the South Fork and the East Fork (FLOW_HW) because the FLOW_HW values are much higher than FLOW_SW1 values (Appendix F (f) and (i)). In a BN test, it is

* When the upstream flow is higher than the downstream flow, the value of FLOW_SW1 is negative (FLOW_SW1= FLOW_IN - FLOW_HW). In this case, we cannot estimate the flow from subwatershed exactly.

concluded that FLOW_IN is very sensitive to FLOW_HW while FLOW_IN is not sensitive to FLOW_SW1. For example, when L was selected as the category of all FLOW_HW values, the predicted probabilities of FLOW_IN were 99.2 %, 0.8 % and 0 % for L, M, and H, respectively. When M was selected as the category of all FLOW_HW values, the predicted probabilities of FLOW_IN were 5.3 %, 94.7 % and 0 % for L, M and H respectively. However, the change of predicted probability distribution of FLOW_IN was small, when changing selected category of FLOW_SW1 (Appendix H).

There is no noticeable difference of probabilities distributions between TP_IN for dry years and TP_IN for wet years (Figure 5.13). It may be concluded that the high annual precipitation cause more stream flow as well as more TP load and the annual precipitation does not noticeably affect the TP concentration in the stream.

When the selected OP_CON factor was changed from Pre to Post under wet conditions, the change of LOAD_SW1, LOAD_IN and TP_IN were larger than under dry conditions (Figure 5.14). It may be concluded that conservation practice had larger effects on the values of LOAD_SW1, LOAD_IN and TP_IN under wet conditions than under dry conditions.

Effects of Conservation Project and exogenous factors on TP load and TP concentration in the stream below Hyrum Reservoir

A small difference between the probabilities of TP load (LOAD_5000) and TP concentration (TP_5000) at the river outlet for Pre OP_CON (before starting conservation practices) vs. Post OP_CON (after starting conservation practices) were

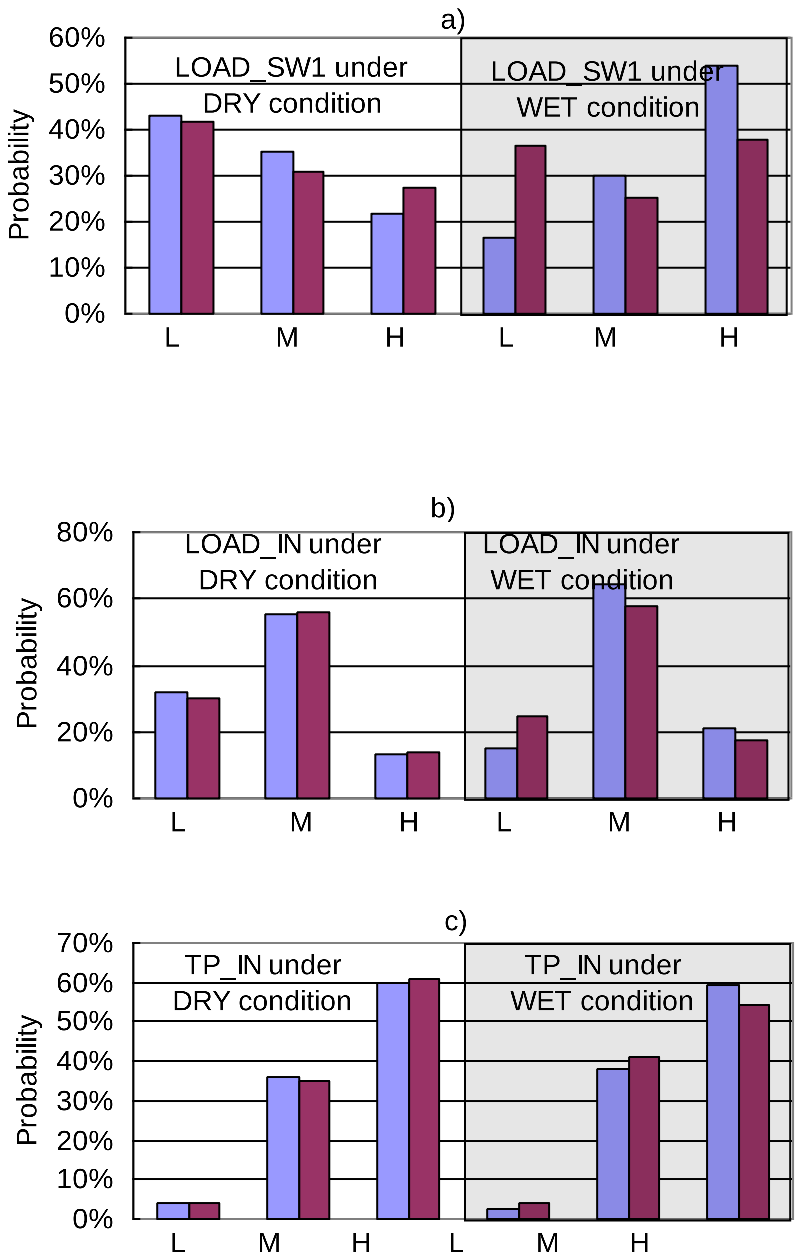


Figure 5.14 | The probabilities of subwatershed TP load (LOAD_SW1), TP load (LOAD_IN) and TP concentration (TP_IN) variables under selected annual precipitation category (PRECIP) and conservation project options (OP_CON). (■:OP_CON=Pre ■: OP_CON=Post, L,M, and H are categories for LOAD SW1 in a), LOAD_IN in b) and TP_IN in c)).

found while *p(L)* for subwatershed load (LOAD_SW2) increased dramatically and *p(M)* and *p(H)* for LOAD_SW2 decreased, changing OP_CON from Pre to Post (Figure 5.15). The probability of water quality criteria violation (*p(M)+p(H),* TP_5000 > 0.05 mg/L) decreased from 66.5% to 63.1%, when OP_CON was changed from Pre to Post (Figure 5.15).

Changing the selection of agricultural land area (LAND_AG2) category had a smaller effect on LOAD_SW2, LOAD_5000 and TP_5000 than conservation practices option (Figure 5.16). For example, the probability of water quality criteria violation increased from 63.4% to 65.6%, when LAND_AG2 was changed from low to high (Figure 5.16).

When the point load (LOAD_P2) category was changed from low to high, *p(H)* for LOAD_SW2 increased and *p(L)* and *p(M)* for LOAD_SW2 decreased dramatically, changing probability distributions of LOAD_5000 and TP_5000 (Figure 5.17). The

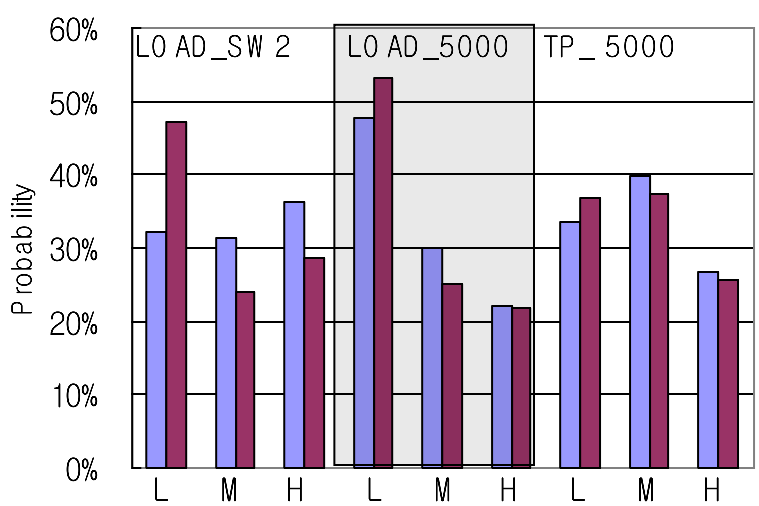


Figure 5.15 | The probabilities of TP load and TP concentration variables under selected conservation project options (OP_CON) (■:OP_CON=Pre, ■: OP_CON=Post, L,M, and H are categories for LOAD SW2, LOAD_5000 and TP_5000).

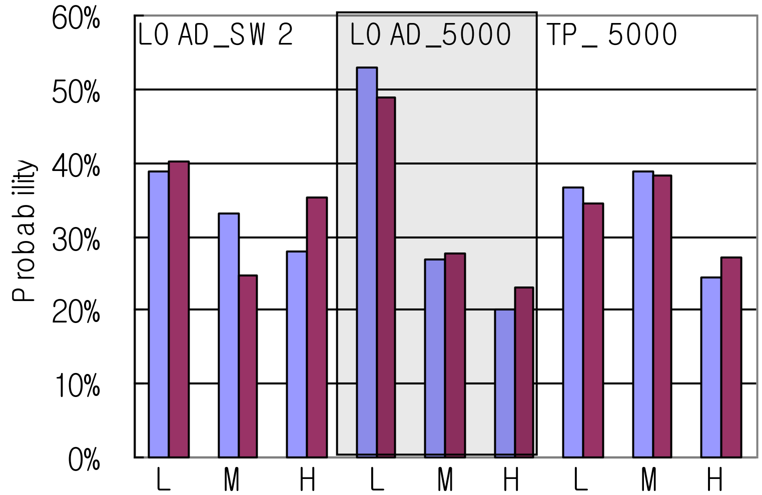


Figure 5.16 | The probabilities of TP loads and flow variable under selected agricultural landuse category (LAND_AG2) (■:LAND_AG2 = Low ■: LAND_AG2=High, L,M, and H are categories for LOAD SW2, LOAD_5000 and TP_5000).

probability of water quality criteria violation (*p(TP_5000 > 0.05 mg/L as TP)*) increased from 62.3% to 72.1% (Figure 5.17).

When the precipitation was changed from dry to wet, *p(L)* for LOAD_SW2, LOAD_5000, FLOW_SW and FLOW_5000 decreased and *p(M)* and *p(H)* for those variables increased. No significant change was observed in probability distributions for TP_5000 between dry and wet conditions (Figure 5.18). Apparently, the high annual precipitation causes more stream flow as well as more TP load into Cutler Reservoir so, the annual precipitation does not have significant effect on TP concentration.

When the selected conservation practices option (OP_CON) were changed from Pre to Post under wet conditions, the change of LOAD_SW1, LOAD_IN and TP_IN were larger than under dry conditions (Figure 5.19). It may be concluded that conservation practice has larger effects on the values of LOAD_SW2, LOAD_5000 and TP_5000 under wet conditions than under dry conditions.

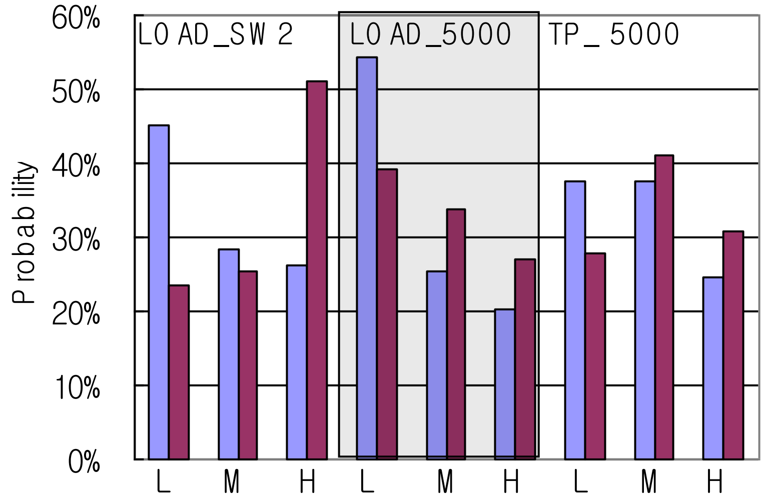


Figure 5.17 | The probabilities of TP loads and flow variable under selected point load category (LOAD_P2) (■:LOAD_P2= Low ■: LOAD_P2=High, L,M, and H are categories for LOAD SW2, LOAD_5000 and TP_5000).

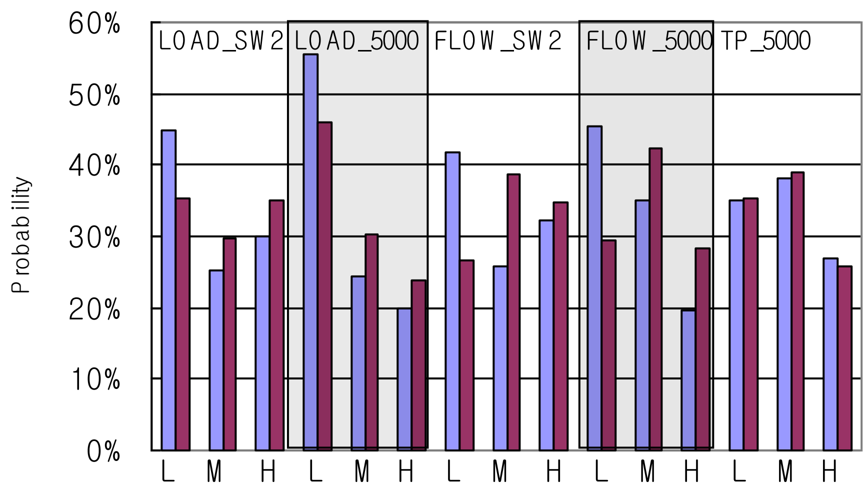


Figure 5.18 | The probabilities of TP loads and flow variable under selected annual precipitation category (PRECIP) (■:PRECIP=Dry ■: PRECIP=Wet, L,M, and H are categories for LOAD SW2, LOAD_5000,FLOW_SW2, FLOW_5000 and TP_5000)

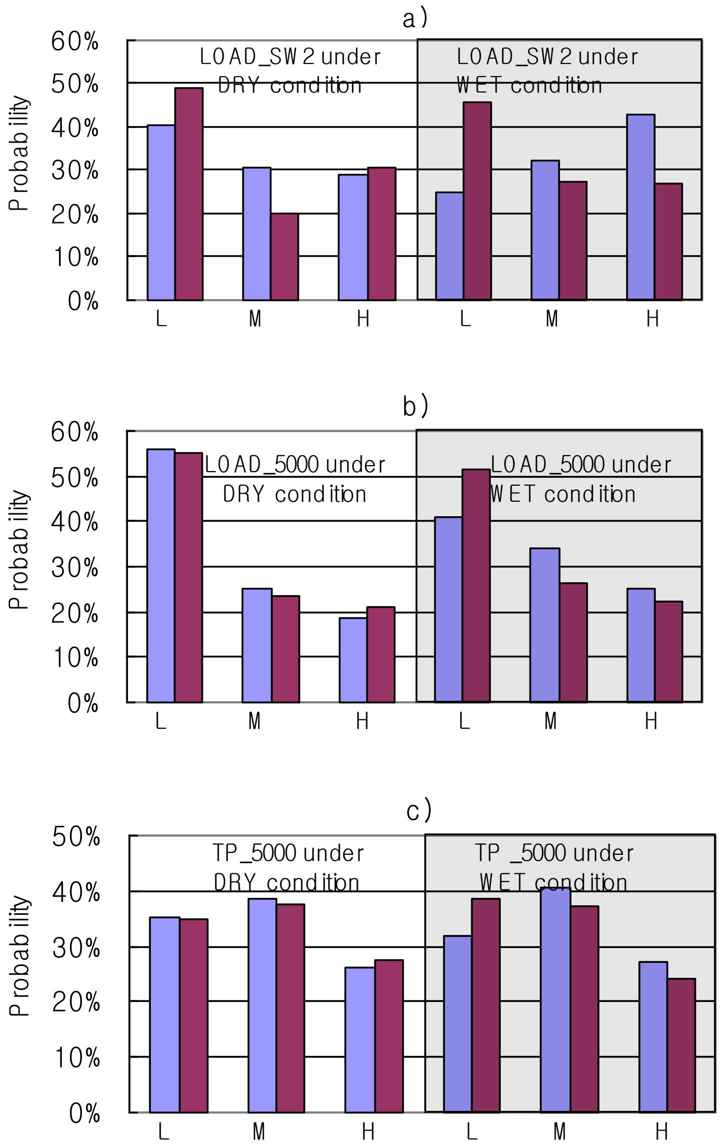


Figure 5.19 | The probabilities of subwatershed TP load (LOAD_SW2), TP load (LOAD_5000) and TP concentration (TP_5000) variables at the mouth of the Little Bear River under selected annual precipitation category (PRECIP) and conservation project options (OP_CON). (■:OP_CON=Pre ■: OP_CON=Post, L,M, and H are categories for LOAD SW2 in a), LOAD_5000 in b) and TP_5000 in c)).

DISCUSSION

Even though the subwatershed TP loads (LOAD_SW1,LOAD_SW2) were noticeably smaller under construction of conservation practices (OP_CON= Post) than before conservation practices were implemented, conservation practice options had only a small effect on TP load and TP concentration at the inlets of Hyrum and Cutler Reservoirs. The upstream water from East Fork and South Fork goes to Hyrum Reservoir taking subwatershed flow and TP load. The upstream water from the Hyrum Reservoir (effluent of the reservoir) goes to Cutler Reservoir taking subwatershed flow and TP load. Conservation practices below the confluence of East and South Fork have no effect on the upstream TP load. Conservation practices below Hyrum Reservoir have no effect on the release from Hyrum Reservoir. Because the upstream waters are mixed with waters from subwatersheds, the effect of conservation practice on TP loads at a reservoir's inlet may be reduced.

In BN1 (above Hyrum Reservoir), the 33rd and 67th percentile values of TP load at confluence of the South and East Fork were 4.75 lb/d and 16.84 lb/d, and the 33rd and 67th percentile values of subwatershed TP load were 3.53 lb/d and 9.51 lb/d. In BN2 (below Hyrum Reservoir), the 33rd and 67th percentile value of TP load at effluent of the Hyrum Reservoir were 2.96 lb/d and 12.9 lb/d, and the 33rd and 67th percentile value of subwatershed TP load were 7.02 lb/d and 18.82 lb/d. The upstream water TP loads were large enough to reduce the effect of subwatershed TP load reduction on downstream total TP load.

The Little Bear River below Hyrum Reservoir is controlled more by subwatershed TP load while the river above Hyrum Reservoir is controlled more by upstream TP load.

More water quality conservation projects were implemented below Hyrum Reservoir than above because the agricultural acreage below Hyrum Reservoir is two times larger than that above. The effect of conservation practices on the TP load and TP concentration into a receiving water was larger in BN2 than in BN1, comparing the probability distributions of categories for BN1 (Figure 5.10) to those for BN2 (Figure 5.15).

Comparing the probability distributions of subwatershed TP load for the low agricultural area condition to that for the high agricultural area condition showed that a larger amount of agricultural land caused more subwatershed TP load ,but agricultural area had no significant effect on TP load or TP concentration at the inlet of Hyrum or Cutler Reservoir. While there was no significant change from agricultural land area in 1986 to that in 1996 below the Hyrum Reservoir, the agricultural area above Hyrum Reservoir increased 22 % from 11,686 acres in 1996 to 14,225 acres in 2003 (Figure 5.4). This increase in agricultural land might be large enough to increase subwatershed TP load but this change did not increase significantly TP load and TP concentration into the reservoirs. TP concentrations of upstream water were relatively low (67^{th} percentile=0.05 mg/l), but the upstream water TP load was larger than subwatershed load because flows were relatively high (67^{th} percentile = 81 cfs). The effect of increased subwatershed load by larger amount of agricultural land above Hyrum Reservoir on TP load at the reservoir's inlet may be reduced because the upstream load is mixed with subwatershed load.

Comparing the probability distribution of subwatershed TP load for wet annual precipitation to that for dry showed more annual precipitation caused more subwatershed TP load. It may be because more annual precipitation caused larger non-point TP loads.

The increase of subwatershed TP load was large enough to increase the TP load significantly into both Hyrum Reservoir and Cutler Reservoir.

BNs suggested that the conservation practices had larger effect on subwatershed TP loads and TP loads and concentrations into receiving reservoirs under wet annual precipitation condition than under dry annual precipitation condition. Water quality conservation practices were executed to reduce the non point TP load. Non-point TP loads from subwatersheds to streams may be larger during wet years, and the conservation practices may reduce more TP loads from these non-point source loads.

The probability distribution of categories of TP load (LOAD_IN) and TP concentration (TP_IN) into Hyrum Reservoir showed a large change, changing the selected condition of subwatershed TP load (LOAD_SW1) as well as changing the selected condition of TP load at the confluence of the East Fork and South Fork (LOAD_HW) (Figure 5.20). For example, when L (low) category of LOAD_SW1 was selected (the probability that LOAD_SW1 value is in L category = 1.0), the probability of L category of LOAD_IN was 54 % and, when H(high) category of LOAD_SW1 was selected (the probability that LOAD_SW1 value is in H category = 1.0), the probability of L category of LOAD_IN was 0 % (Figure 5.20).

The probability distribution of categories of TP load (LOAD_5000) and TP concentration (TP_5000) into Cutler Reservoir also showed a large change, changing the selected condition of subwatershed TP load (LOAD_SW2) as well as changing the selected condition of TP load at the effluence of the Hyrum Reservoir (LOAD_5650) (Figure 5.21). TP loads into Hyrum (LOAD_IN) and Cutler (LOAD_5000) reservoirs are

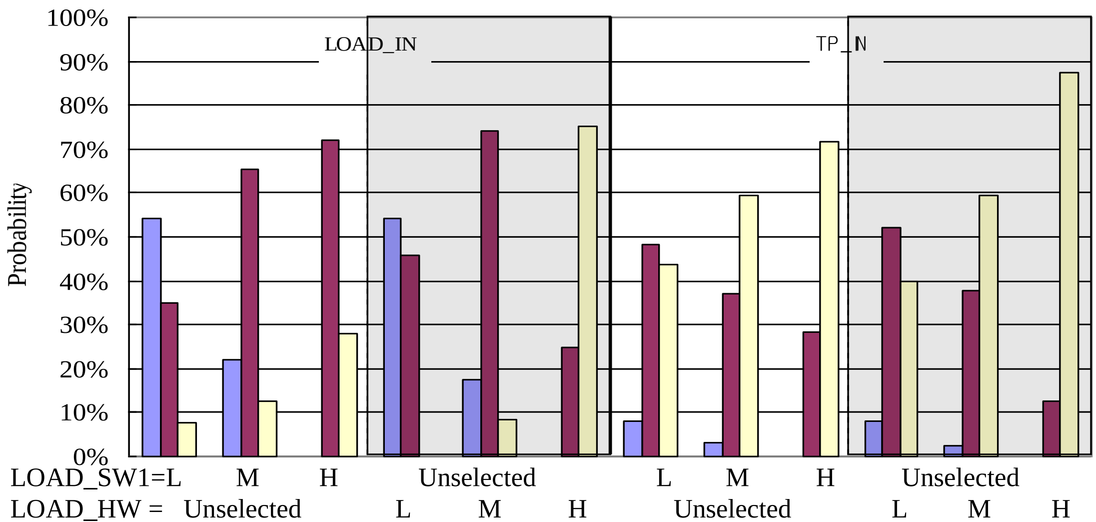


Figure 5.20 | The probabilities of TP load (LOAD_IN) and TP concentration (TP_IN) into the Hyrum Reservoir under selected subwatershed TP load (LOAD_SW1) and TP load at confluence of the East and South Fork (LOAD_HW) (■:probability of Low for LOAD_IN or TP_IN , ■: probability of Medium for LOAD_IN or TP_IN , □: probability of High for LOAD_IN or TP_IN).

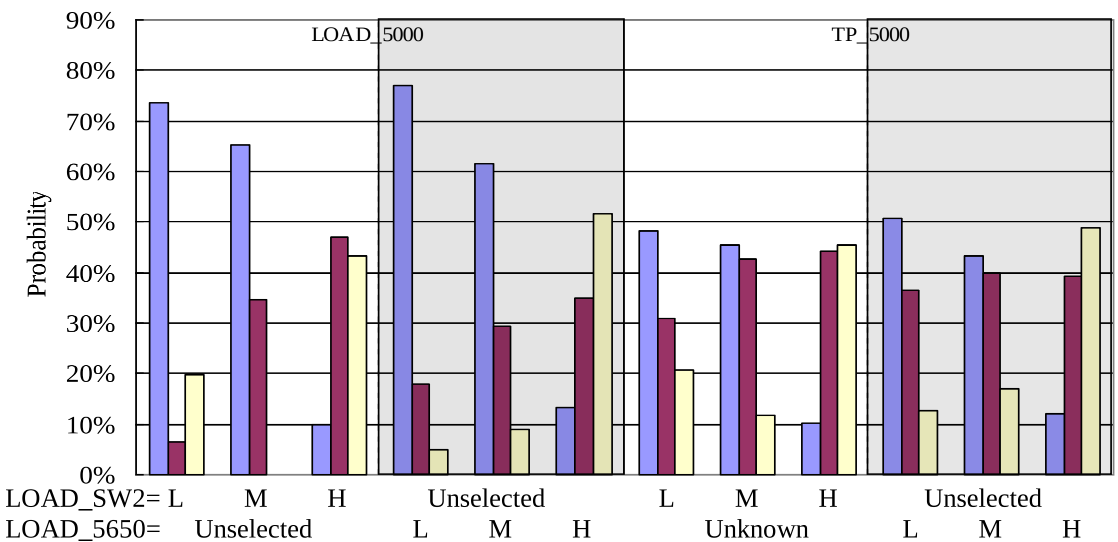


Figure 5.21 | The probabilities of TP load (LOAD_5000) and TP concentration (TP_5000) into the Cutler Reservoir under selected subwatershed TP load (LOAD_SW2) and TP load at effluence of the Hyrum Reservoir (LOAD_5650) (■:probability of Low for LOAD_5000 or TP_5000 , ■: probability of Medium for LOAD_5000 or TP_5000 , □: probability of High for LOAD_5000 or TP_5000).

sensitive to subwatershed TP load (LOAD_SW) as well as upstream TP loads (LOAD_HW, LOAD_5650). TP concentration into Hyrum Reservoir (TP_IN) and Cutler Reservoir (TP_5000) is sensitive to subwatershed TP load (LOAD_SW1, LOAD_SW2) as well as upstream TP load (LOAD_HW, LOAD_5650) (Figures 5.20, 5.21).

It may be concluded that if the TP load from subwatershed is reduced more than now, it is strongly possible to decrease TP load and TP concentration into Hyrum and Cutler Reservoirs. For example, the probability of the L category of TP concentration (=< 0.05 mg/L) into Cutler Reservoir is 48.4 % for the L condition of subwatershed TP loads (LOAD_SW2). The probability of L category of TP concentration (=< 0.05 mg/L) into the Cutler Reservoir is 36.9 % for Post LBRCP option (OP_CON). According to these results, it is concluded that if all subwatershed TP loads (LOAD_SW2) fall below the L category boundary (7.02 lb/d), the water quality standard violation frequency will be decreased from 63.1 % to 51.6 %. This suggests that more conservation practices or more point source controls to reduce the TP load from subwatershed to the stream will be helpful to reduce the TP concentration violation rate at the inlet of Hyrum Reservoir or Cutler Reservoir.

A Bayesian Network is a probabilistic model, in which each variable has two or three categories (Varis, 1998; Marcot *et al.*, 2001; Borsuk *et al.*, 2003) and probability distributions among these categories show the effect of a specific variable on other variables. Because of this characteristic, it is sometimes difficult to evaluate the effect of a specific variable on other variables. For example, in BN1, while the probability changes in the high (H) categories of LOAD_SW1, LOAD_IN and TP_IN were increases of

5.4 %, 0.9 % and 1.1 %, on changing OP_CON from Pre to Post under dry conditions, the same probability changes were decreases of 15.7 %, 2.5 % and 4.8 % under wet condition (Figure 5.14). It may be concluded that LBRCP reduced the TP non point source load and the TP load and concentration into Hyrum Reservoir for wet annual precipitation. However, implementing conservation practices (changing OP_CON from Pre to Post) under dry conditions increased *p(H)* for LOAD_SW1 by 5.4 % - it is unclear whether this increase is significant or ignorable. There is no specific rule to reject the null hypothesis of the differences in BN. It is one opinion that the 5.4 % probability increase of H category of LOAD_SW1 may be ignorable because this increase caused only a 0.9 % probability increase of H category of LOAD_IN, the child node (Figure 5.14), so the impact of the subwatershed load is attenuated by other variables. It may also be true that hidden factor may have caused the increase and are not included in the BN.

There has been no significant change of the agricultural area below Hyrum Reservoir since 1976 (Figure 5.4). BN2 showed that more annual precipitation might increase the TP load into the Cutler reservoir but not the TP concentration. We were not able to compare the TP point source loads before and after 1990 in the BN2 database because there were no TP point source load data prior to 1990. However, it is concluded that the TP concentration decrease since 1990 may have been caused by LBRCP because the LBRCP had significant effects on TP concentration into Cutler reservoir under wet annual precipitation conditions (Figure 5.19).

SUMMARY AND CONCLUSION

In order to evaluate the effect of conservation practices and exogenous variables on the TP load and TP concentration, BN1 (Above Hyrum) and BN2 (Below Hyrum) were constructed. Each BN used a different database. Some missing value of TP concentration and flow were filled with values estimated by regression between upstream and down stream data or by regression between two different variables.

BN simulations showed that conservation practices in the Little Bear River reduced subwatershed TP load above and below Hyrum Reservoir noticeably but the reductions were not large enough to reduce TP concentration into the receiving reservoirs noticeably, due to dilution of the effect by other factors. BNs suggested that the conservation practice have been working to reduce TP loads but more implementations of conservation practices are required.

There were three exogenous variables: agricultural landuse area, point source load and annual precipitation. Increased agricultural land area caused noticeably higher subwatershed TP load above and below Hyrum Reservoir significantly but not higher TP load and concentration into the receiving reservoirs, due to dilution of the effect by other factors. However, increased point source load caused significantly higher TP loads and concentrations into the Hyrum and Cutler reservoirs.

Increased annual precipitation caused a noticeably higher subwatershed TP load above and below Hyrum Reservoirs. These load increases were large enough to significantly increase TP loads into the Hyrum and Cutler reservoirs, but not TP concentration because more annual precipitation caused more flow and more TP load simultaneously.

The effects of conservation practices in the Little Bear River on the subwatershed TP load, TP load into Hyrum and Cutler Reservoirs and TP concentration into Hyrum and Cutler Reservoirs were larger for wet annual precipitation conditions than those for dry annual precipitation conditions.

It may be concluded that the TP concentration decreases since 1990 have been influenced by LBRCP (Little Bear River Conservation Project) only under wet annual precipitation conditions.

REFERENCES

Ames, D. P, Neilson, B. T., Stevens D. K. & Lall, U 2005 Using Baysian networks to model watershed management decisions: an East Canyon Creek case study, *J. Hydroinfomatrics* **7** (4), 267-282

Baker, M., Hochhanlter, S., & Lytle, E. 2008 Abstract, *Spring Runoff Conference*, March 13-April 1, 2008, Logan, UT.

Borsuk, M.E., Stow, C.A. & Reckhow K.H. 2003 Integrated Approach to Total Maximum Daily Load Development for Neuse River Estuary Using Bayesian Probability Network Model (Neu-BERN), *J. Water Resources Planning and Manage.* **129** (4), 271-282.

Castillo, E., Gutiérrey, J. M & Hadi, A. S. 1997 *Expert Szstems and Probabilistic Network Models.* Springer-Verlag, New York, NY. 605p.

Chapra, S.C. 1997 The Eutrophication Problem and Nutrients. *In Surface Water-Quality Modeling*, Lecture 28, Mc Graw-Hill, New York, NY, pp. 521-533.

Chess, D. 2000 Utah Watershed Review Vol 8, Utah Division of Water Quality.

Dodds,W.K. 2002 Nitrogen, Sulfur, Phosphorus and other nutrients. In *Freshwater Ecology* , Academic Press. San Diego, CA, pp 247-266.

EMRG (Environmental Management Research Group at Utah Water Research Laboratory) 2004 Little Bear River Watershed boundary shapfile.

Jensen, F.V. 1996 *An Introduction to Bayesian Networks,* Springer-Verlag. New York, NY.

Kendall, M.G. 1970 *Rank Correlation Methods*, 4th ed. Charles Griffin & Company Limited, London, U.K.

Marcot, B.G., Holthausen, R. S., Raphael, M. G., Rowland M. M, & Wisdiom M. R. 2001 Using Bayesian Belief networks to evaluate fish and wildlife population viability under land management alternatives from an environmental impact statement, *Forest Ecology and Manage.* **153**, 29-42.

Mason C. 2002 *Biology of Freshwater Pollutant,* 4th Ed. Pearson Education Limited, Essex, England.

Norsys Software Corp. 1997 *Netica Application User's Guide,* Vacouver, BC, Canada.

Reckhow, K.H. 1999 Water Quality Prediction and Probability network models, *Canadian Journal of Fishery and Aquatic Science*, **56**, 1150-1158.

Slymen, D.J., Peyster, A. D. & Donohoe, R. R. 1994 Hypothesis Testing with values below Detection Limit in Environmental Studies, *Environmental Science and Technology*, **28** (5), 898-902.

Theil, H. 1950 A Rank-Invariant Method of Linear and Polynomial Regression Analysis,3 Proceedings of Koninalijke Nederlandse Akademie Can Wethenschatpen Series A, **53**, 1397-1412

USEPA 2006 Little Bear River Project, http;//www.epa.gov/region8/water/nps/utlittlebear.pdf

USEPA 2005 STORET (Storage and Retrieval). http;//www.epa.gov/stormoda/DW_stationcriteria (Jan. 12, 2005)

USEPA 2004 Better Assessment Science Integrating Point and Non Point Sources (BASINS). http://www.epa.gov/waterscience/BASINS/ (Nov. 8, 2004).

USGS 2004 National Hydrograph Dataset. http://nhd.usgs.gov/data.html (Nov 10, 2004).

USGS 2006 National Water Information System:Web Interface, http://waterdata.usgs.gov/nwis/uv?10105900 (Apr. 10, 2006).

Utah DNR (Utah Department of Natural Resources) 2004 Land Use Program, http://www.agr.state.ut.us/agrc/html/data/layers1.html, (Nov. 2, 2004)

Utah DAR (Utah Division of Administrative Rules) 2000 R317-2-7, Water Quality Standards, http://www.rules.utah.gov/publicat/code/r317/r317-002.htm#T8

Utah DEQ (Utah Department of Environmental Quality) 2000 Little Bear River Watershed TMDL, http://www.waterqulaity.utah.gov/TMDL/index.htm#addinfo.

Varis. O. & Jussila, S. F. 2002 Water Resources Development in the Lower Senegal River Basin: Conflicting Interests, Environmental Concerns and Policy Options, *International J. Water Resources Development*, **18** (2), 245-260.

Varis, O. 1998 A belief network approach to optimization and parameter estimation: application to resource and environmental management. *Artificial Intelligence*, **101**, 135-163.

WRCC (Western Regional Climate Center) 2006 Logan Radio KVNU UT, Monthly Total Precipatiatation, http://www.wrcc.dri.edu/cgi-bin/cliMAIN.pl?utlkvn. (Feb 11, 2006).

CHAPTER 6

TOTAL MAXIMUM DAILY LOAD (TMDL) FOR TOTAL PHOSPHORUS AT THE MOUTH OF THE LITTLE BEAR RIVER

ABSTRACT

The Little Bear River watershed in Cache County, Northern Utah is on a high-priority list of watersheds affected by nonpoint source pollution. A Total Maximum Daily Load (TMDL) at the mouth of Little Bear River targeted reduction of phosphorus. In order to obtain daily frequency flows and TP loads from low frequency observations, a Rank-Data distribution connecting approach (R-D method) are used. Load duration curve are constructed based on daily flows and TP loads from the R-D method, and showed changes in TP loads associated with change of flow duration interval %, and the frequency of water quality standard violations. The TMDLs and historical TP loads were calculated from daily frequency flows and TP loads for low flow season (July to February) and high flow season (March to June) in wet year (97-98 water year) and dry year (02-03 water year). The allocations from a frequency targeted approach, in which the margin of safety (MOS) was adjusted for 10 % frequency exceeding water quality criterion (0.05 mg/L as TP), was higher than the allocation using total targeted mass, in which 0.2σ (0.2 times standard deviation of TP loads) was set as the MOS. The reduction percentage in a wet year was higher than that in a dry year. Appropriate reduced TP concentrations (0.044-0.047mg/L; TP concentration calculated based on

annual TP load allocation and flow) from total mass targeted approach may be an evidence to prove this more practical method.

Introduction

The Total Maximum Daily Load (TMDL) is a historical and watershed-based program to restore the surface water quality to a level that meets water quality standard (Younos, 2005). Section 303 (d) of Clean Water Act (CWA) requires States and Territories and authorized Tribes to identify and establish a priority ranking for water bodies for which technology-based effluent limitation required by Section 301 of the CWA are not stringent enough to achieve the water quality standard and establish TMDLs for the pollutant causing impairment in those water bodies. States and Territories and authorized Tribes must establish TMDL at the levels necessary to implement applicable water quality standards with seasonal variations and a margin of safety. The margin of safety aims to take into account any lack of knowledge concerning the relationship between effluent limitations and water quality (NARA, 2000).

Many traditional TMDL approaches have focused on targeting a single value depending on a water quality criterion and a design flow. This single number approach does not work well for impaired water caused by non-point source (NPS) pollutants (Stiles, 2001). Because stream flows cause different loading mechanisms to dominate under different flow regimes, variability in stream flows is an important concern regarding nonpoint sources (Cleland, 2002). The duration curve is a TMDL approach for characterizing water quality data under different flow regimes. The duration curve framework allows for easily presenting frequency and magnitude of water quality standard violations, allowable loadings, and size of load reductions (USEPA, 2007).

The duration curve is a monitoring based TMDL approach using observations. Research using duration curves has emphasized magnitude and frequency of water quality standard violations. A lack of observations may increase uncertainty in monitoring approaches. In order to calculate more reliable allocations or load reductions, daily load observations or predictions for all different simulation days are required, but daily water quality or flow values may have gaps of measurement dates (data gaps). The Rank-Data distribution method (R-D method, Chapter 4) is a statistical approach to fill in the data gaps of a variable during a simulation period.

The Little Bear River in Cache County, Northern Utah flows from southeast to northwest, bounded by mountains, and drains to Cutler Reservoir, west of Logan, UT (Figure 6.1). The Little Bear River watershed is on a high-priority list of watersheds that

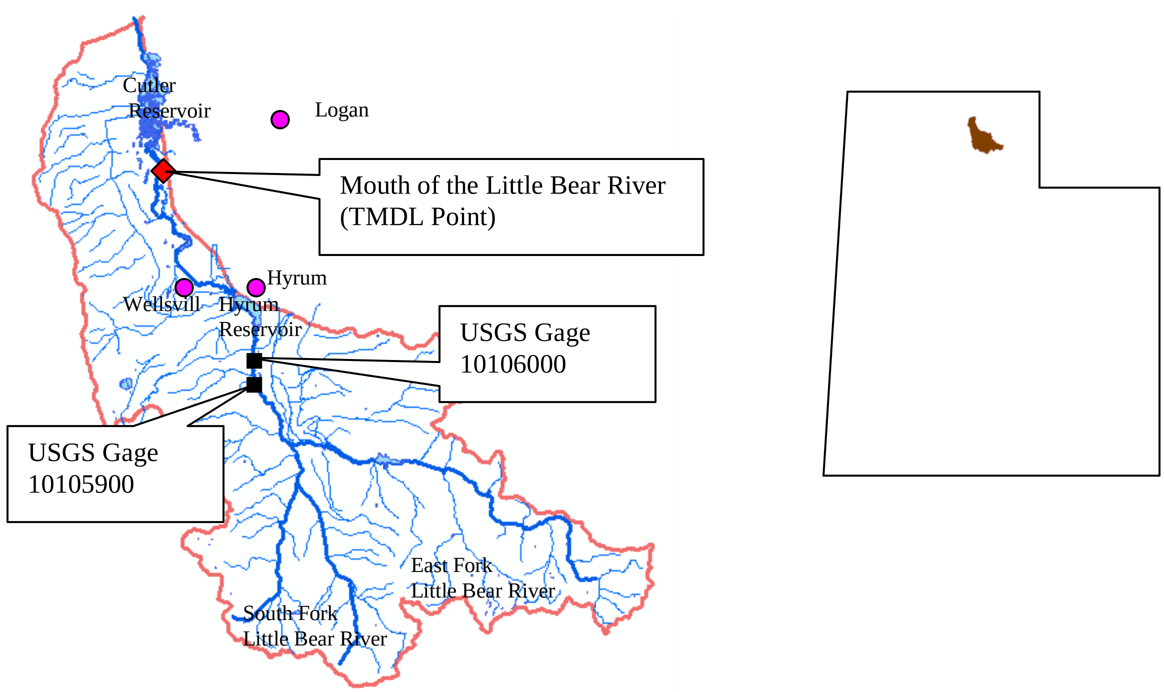


Figure 6.1— Little Bear River Watershed located in Northern Utah (EMRG, 2004; USGS, 2004; USEPA, 2004).

are being adversely affected by nonpoint source pollution (Chess, 2000). The Little Bear River Steering Committee found cropland and pastures may be significant sources of nutrients in the Little Bear River watershed (Chess, 2000). In many fresh waters, an increase of phosphorus may cause algal blooms because phosphorus is normally limited (Mason, 2002). A Total Maximum Daily Load (TMDL) for the Little Bear River targets reduction of phosphorus (Utah DEQ, 2000). In this chapter, the predictions from the R-D method were used to estimate the duration curve for Total Phosphorus (TP) TMDL at the mouth of the Little Bear River.

Using the load duration curve, we analyze what type of source (point or nonpoint source) may mainly contribute to exceedance of the water quality standard, and estimate TMDL, Margin of Safety (MOS), allocation, and load reduction percentage in this chapter.

Methods

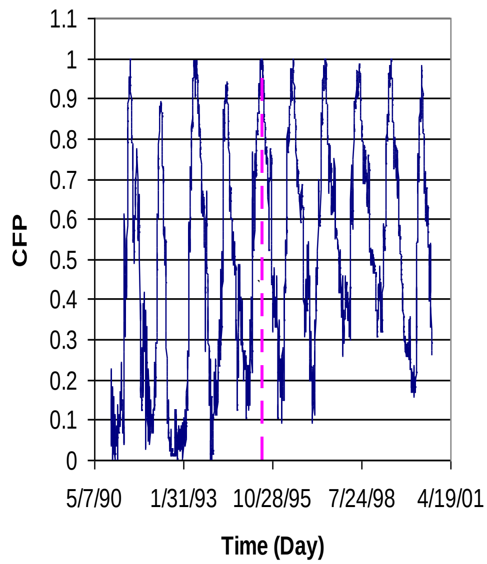
Data collection. The TP loads were calculated by multiplying the flow by the TP concentrations and converting the units to lb/day. The data collection period was from 1978 to 2004. Flows (164 observations) and TP concentrations (234 observations) at the mouth of the Little Bear River came from the EPA STORET data base (USEPA, 2005). Linear regression between flows and specific conductance and liner regression between the flows and upstream flows filled in the 85 flows (calculated values) at the mouth of the Little Bear River (Chapter 5, Table 5.1). Other flows and TP concentrations (missing values) on the non measuring dates were predicted by R-D method based on extended observations (observations plus calculated values). After filling in missing flows and TP concentrations, 9652 flows, 9652 TP loads and 9652 concentrations are obtained from 1978 to 2004.

Rank-Data distribution method (R-D method). The R-D method is an approach to fill in data gaps of a variable by linking the data distribution (cumulative probability plot) to the rank time series of the variable (Chapter 4). For example, if the Cumulative Failure Probability (CFP) (Kaplan and Meier, 1958) on July 2nd, 1995 is 0.98 on the time series of the rank and the flow for the CFP of 0.98 is 1370 cfs, the flow on July 2nd ,1995 is estimated to be 1370 cfs. The process is shown in Figure 6.2.

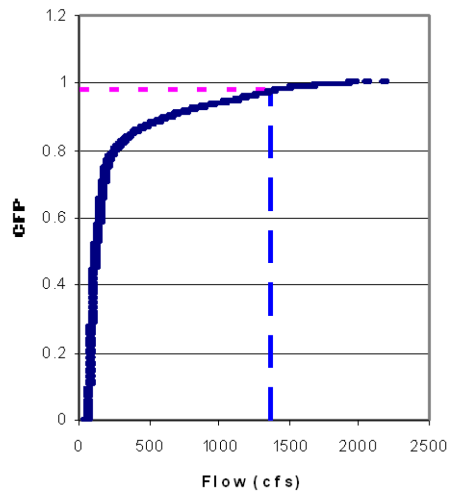
Two assumptions are used to estimate the values for dates on which data are not available. The first assumption is that the distribution of all values including both observations and missing values is identical to the distribution of the observations alone. Under this assumption, we can generate an unlimited number of values using the cumulative failure (CF) plot of observations. The second assumption is that the interpolation or extrapolation between the CFPs of two observations on the CFP time series is identical to the true CFP for the prediction date. If cumulative probability plot from observations and rank (cumulative probability) time series of all dates from interpolation of observation CFP time series are obtained, daily time series of values for the variable may be estimated under these two assumptions.

In this chapter, we tried to estimate daily time series of 9652 flows and 9652 TP load from January 5, 1978 to June 8, 2004 using extended observations. The Kaplan-Meier method (K-M method) was used to obtain the CF plot (Kaplan and Meier, 1958).

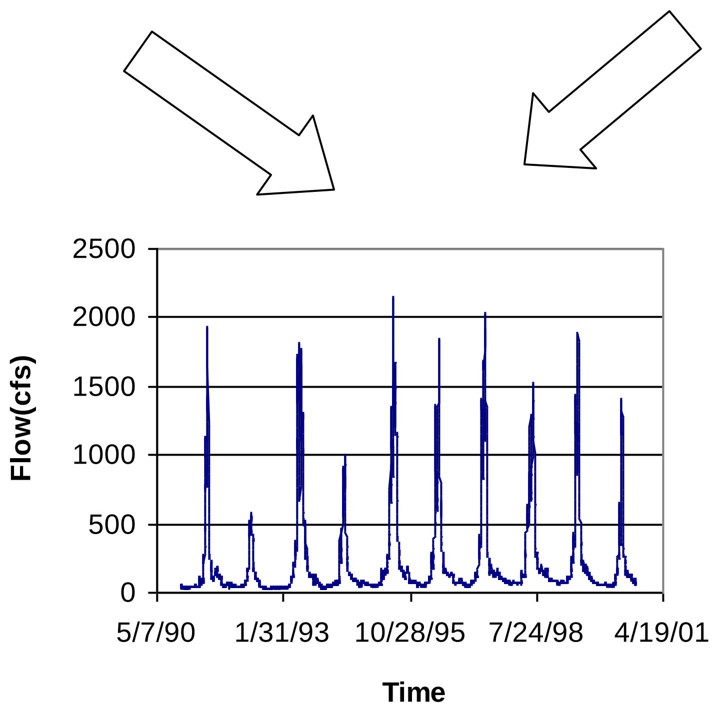
The K-M method was designed for right censored (Above Detection Limit) data and detected data only, rather than the usual left censored data (Below Detection Limit).



a) Rank Time Series



b) Data distribution of flows as CFP.



c) Constructed Time Series of Values

Figure 6.2—The idea to make time series combining data distribution with CFP time series (USGS gage 10128500, Weber River near Oakley, UT).

Because 9,428 missing TP load and 9,405 missing flows were set as left censored values (Below maximum of observations), all of flows and loads were flipped to make a CF plot by the K-M method (Helsel, 2005). Minitab statistical software (Minitab, 2006) was used to calculate the survival probability (= 1- CFP) of the flipped values using the K-M method. After the flipped values were returned to the unflipped scale, the CFPs were estimated. Flows and TP loads for 9652 simulation days were estimated based on this cumulative failure plot. For example, if the simulation period is 9652 days and we estimate daily TP load for each day, the failure probability of the smallest TP load equals 1-(9651/9652)= 0.000104. The CFP of 2nd smallest TP load equals 1-(9650/9651)*(1-0.000104)= 1-(9650/9652)= 0.000207, and so on. The TP load with CFP 0.000104 on the cumulative failure plot is assigned to the smallest TP load and the TP load with CFP 0.000207 is assigned to the 2nd smallest TP load. The 9652 CFPs and TP load (estimated distribution) may be estimated from the smallest TP load to largest TP load in the same manner.

The CFP for each value was calculated within the extended observations (observations and calculated values) using Equation 6.1.

$$1-\frac{r-1}{N} \tag{6.1}$$

where r is rank and N is total number of extended observations values.

The interpolating CFPs of the extended observations (observed CFPs) estimates the CFPs of the missing values (interpolated CFPs) between two measurement dates (Figure 6.3).

All of the CFPs including observed CFPs and interpolated CFPs for different dates (estimated CFP time series) were converted to a single set of estimated ranks. After

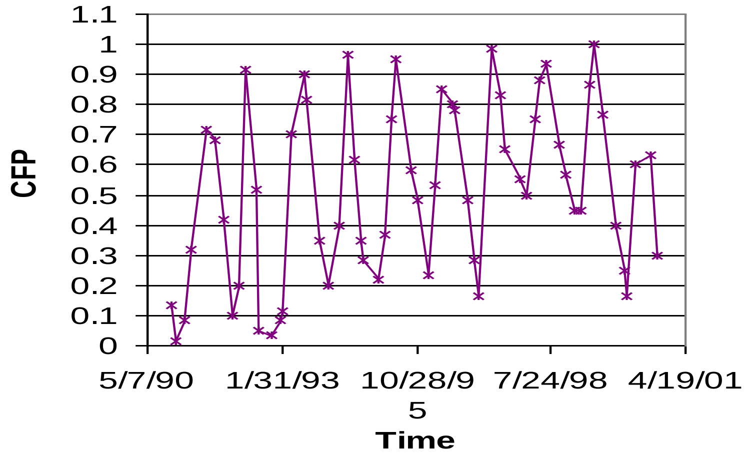


Figure 6.3—The CFP time series of observations (x) and unmeasured values (Line) (USGS gage 10128500, Weber River near Oakley, UT).

finding the value with a rank within the estimated distribution, the value (prediction) was assigned to the date with the same rank within the estimated CFP time series. For example, if the interpolated CFP for the mean daily flow on January 24, 1978 was 0.3165 corresponding to rank 6580 within the 9652 estimated CFP time series, the 6580th TP load (6580th prediction) within the estimated distribution, 24.94 lb/day is then assigned to the TP load on January, 1978. The procedure is carried out for all data to produce a daily time series of mean flows and TP loads.

It is recognized that a data set with extended observations of flows or TP loads per year may not include the peak flow or TP load during spring runoff for that year. Some approximate Peak CFPs of flows or TP loads are added between observed CFP time series before interpolating CFPs in order to get more accurate interpolated CFP.

The flow peak CFPs may be estimated under the assumption that dates of peak CFP of each cycle are known (Chapter 4). The daily mean flows during simulation period were not accessible in STORET. All of the peak flow dates at the TMDL point are the same or close to those at a USGS gage above Hyrum Reservoir (10106000) from 1942 to 1974. Because the flows at USGS gage 10106000 are accessible only until 1987, it was assumed that the flow peak dates at USGS gage 10105900 which is another USGS gage above Hyrum Reservoir close to gage 10106000 are the same as those at the TMDL point for 1978-2004 (Figure 6.1). In the same manner, the TP load peak CFPs may be estimated. It is assumed that dates of TP load CFP peak are identical to dates of known flow peak CFP of each cycle to estimate TP load peak CFP.

All predicted flows or TP loads were verified by comparing to extended observation flows or TP loads on the graph with predictions (X-axis) and extended observations (Y- axis) (Figure 6.4). The predictions were then calibrated by regression between extended observations and predictions (Table 6.1). For example, the predicted value, 370.2 lb/d for May 23, 1995 is far from the extended observation for same date, 247.89 lb/d. The regression curve is constructed using the predictions and extended observations and, the predicted value, 370.2 lb/d is inserted in the regression equation ($459.7 + 7.582 \times 370.2 - 0.02838 \times 370.2^2 + 0.000036 \times 370.2^3 = 284.19 lb/d$). The predicted value, 370.2 lb/d was calibrated as 284.19 lb/d by the regression. Because this calibrated value, 284.19 lb/d is closer to the extended observation, 247.89 lb/d, the predicted value, 370.2 lb/d was replaced by calibrated value, 284.19 lb/d. Other predicted values were calibrated by the same manner (Figure 6.5).

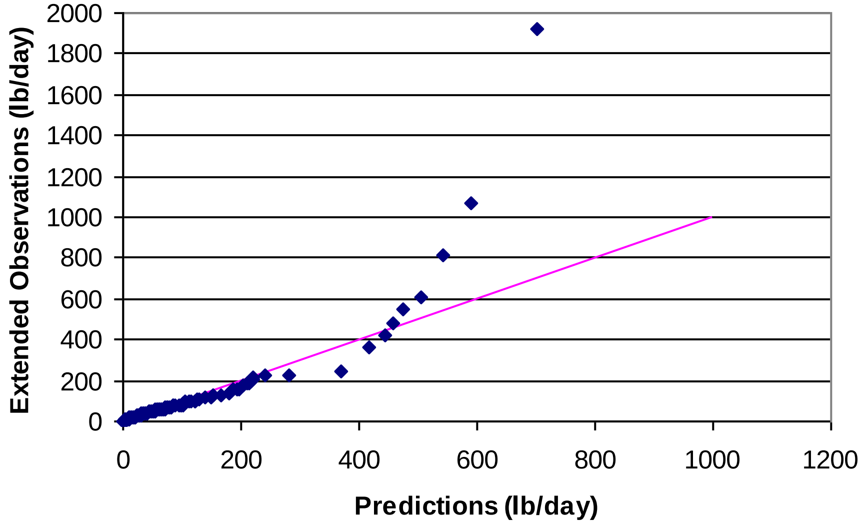


Figure 6.4—Symmetry graph between prediction and extended observations of TP load at the Little Bear River TMDL point.

Table 6.1— Regressions between extended observations and predictions.

Variable (Response variable)	Range of predictions before calibration	Number of Pairs	Equations (F: predictions)	Correlation Value, ( ): p-value
Flow	-5.45 cfs	11	$14.47 - 9.315F + 1.768F^2$	r^2=0.939 (<0.001)
Flow	8.4- 218.2 cfs	220	14.9 + 0.77F	r^2=0.991 (<0.001)
Flow	222.22-	14	503.1-3.299F + 0.008768F^2	r^2=0.946 (<0.001)
TP load	1.3-24.21 lb/d	76	0.3759 + 1.854F-0.03479F^2	r^2=0.980 (<0.001)
TP load	24.79-218.51	133	8.1 + 0.7897F	r^2=0.993 (<0.001)
TP load	221.68-457.78	7	-459.7 + 7.582F-0.02838F^2+0.000036F^3	r^2=0.985 (<0.001)
TP load	474-702.45	5	2209-10.03F+0.0137F^2	r^2=0.997 (0.001)
TP load	702.45-	2	1923 + 9.15F (Extrapolation)*	

*Simple extrapolating two largest points is applied to avoid too large load.

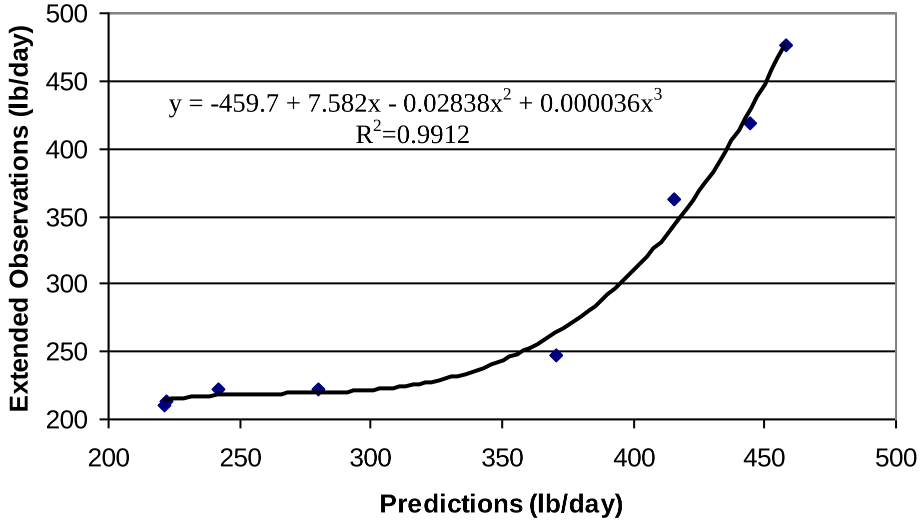


Figure 6.5—Regression between prediction and extended observations of TP load at the Little Bear River TMDL point.

TMDL using load duration curve and total sum of daily loads and flows.

A load duration method is a data-driven approach using long-term stream flow gauge station data (Neilson et al., 2005). Flow duration curves relate flow values (y- axis) to the percent of time (x-axis) that flow values are met or exceeded (Figure 6.6). The x-axis represents the percent of time or duration, as in a cumulative frequency distribution and the y-axis represents the flow value (typically daily average discharge rate) associated with that percent of time or duration (USEPA, 2007). Flow duration intervals are expressed as a percentage, with 0 corresponding to the highest flow value in the record, and 100 to the lowest. In Figure 6.6, a flow duration interval of 60 % associated with a flow value of 8.5 cubic feet per second (cfs) implies that 60% of all observed daily average stream discharge values equal or exceed 8.5 cfs (Neilson et al., 2005).

Water quality targets of TP concentration are translated into TMDLs through the

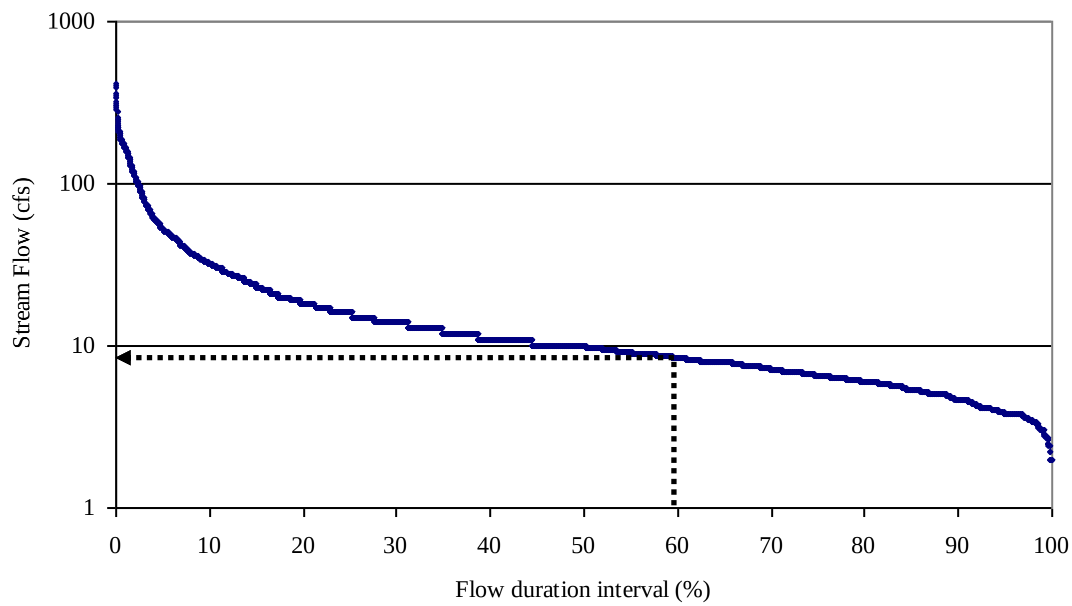


Figure 6.6—Flow duration curve at East Fork of Sevier River, UT, USGS 10183900 (Neilson et al., 2005).

loading capacity. USEPA (USEPA, 2007) currently defines loading capacity as "the greatest amount of loading that a water can receive without violating water quality standards." Therefore, a loading capacity duration curve (TMDL curve) is developed by multiplying flow duration curve with the numeric water quality target and a conversion factor for the target pollutant. For example, the TMDL corresponding to a flow duration interval of 60 % is calculated by 8.5 (cfs) x 0.05 (mg/L) x 5.389 (conversion factor for load in lb/day).

Each actual pollutant load is calculated by multiplying an average daily flow by the pollutant concentration on that day. These historical loads are also plotted on the graph similar to Figure 6.7 for every different flow duration interval. If the historical load falls on or below the TMDL curve, this means compliance with water quality criteria (Neilson et al., 2005).

The load reduction is determined based on allowable percentage of loads above TMDL curve and the margin of safety. USEPA classified the margin of safety (MOS) as two types, 'Explicit' and 'Implicit' (USEPA, 2007). In the explicit type, the safety factors are used. For example, the MOS was 10% of the criteria (200 cols/100ml) for fecal coliform in TMDL for the Upper Brindley Creek (Alabama DEM, 2005). In the implicit type, conservative assumptions are used. The 7 day consecutive low flow

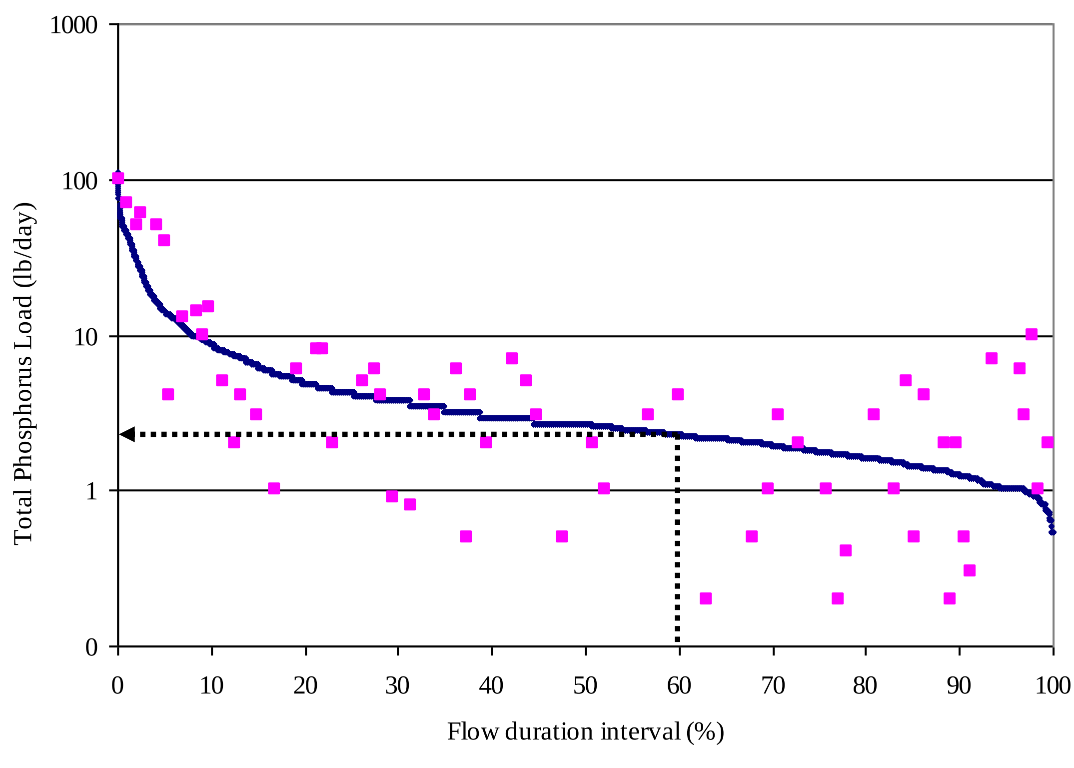


Figure 6.7—Typical loading capacity duration curve at East Fork Sevier River, USGS gage 10183900. — : TMDL, ■ Historical loads (Observations) (Neilson et al., 2005).

occurring in a 10 year period (7Q10) was used as a critical flow condition in the TMDL for the South Platte River, Colorado to result in the lowest dilution of pollutants (USEPA, 1993). In this chapter, two explicit types of MOS are used based on TP load distribution including 0.2S (0.2 times of standard deviation of the predicted loads), and flexible MOS corresponding to water quality criterion violation of 10 % of all frequency (Smith et al., 2001) to avoid excessive MOS and achieve flexibility. An allocation including both point and non point loads are calculated by Equation 6.2

$$\sum LA + \sum WLA = TMDL - MOS \qquad (6.2)$$

in which TMDL= Allowable total maximum daily load found using the Assimilative capacity for a particular water body and contaminant, LA = pollutant load allocation for non-point sources, WLA= pollutant load allocation for point sources discharges, and MOS= Margin of safety.

For this TMDL, the 1997-1998 water year (October, 1997 to September, 1998) was used as the simulation period for a wet year, and the 2002-2003 water year (October, 2002 to September, 2003) was used as the simulation period for a dry year. The hydrologic characteristics of the spring run off season are different from the low flow season. In the wet or dry year, the TMDL, MOS, allocation and load reduction were calculated for low flow (July to February) and high flow seasons (March to June) separately. The TMDLs were determined for four categories: 1) low flow (July to February, 243 days); 2) high flow season (March to June, 122 days) in a wet year; 3) low flow and 4) high flow season in dry year. Finally, annual historical TP loads, TMDLs and allocations were calculated by the weighted average for low flow and high flow seasons because the period of low flow season (243 days) differed from that of high flow (122 days).

Because it is difficult to access high frequency load data in the TMDL process, the historical loads are commonly expressed by statistical representative values such as median, geometric mean or highest value (Alabama DEM, 2005; USEPA, 2007; Utah DEQ, 2000). In this chapter, an estimated historical TP load (lb/day) during the simulation period was calculated using daily TP loads from the R-D method instead of using statistical representatives. In the same manner, the allocation was calculated using daily flows from the R-D method (Equation 6.3).

$$L_a = \sum \mathrm{LA} + \sum \mathrm{WLA} = \mathrm{TMDL} - \mathrm{MOS}$$

$$= \frac{1}{d}\sum Q \times 0.05mg / L \times f - x\sigma \qquad (6.3)$$

in which L_a = load allocation, TMDL= Allowable total maximum daily load found using the assimilative capacity for a particular water body and contaminant ($= \frac{1}{d}\sum Q \times 0.05mg / L \times f$ for TP), LA = pollutant load allocation for non-point sources, WLA= pollutant load allocation for point sources discharges, and MOS= Margin of safety, $\sum Q$ = sum of daily flows during simulation period, d = number of simulation days, f = conversion factor between concentration (mg/L), flow (cfs) and daily load (lb/day). σ = standard deviation of daily TP loads during the simulation period, and x= number of times for σ. Reduction percentage was calculated based on total amount of estimated historical TP loads and flows (Equation 6.4).

$$R(\%) = \frac{L_h - L_a}{L_h} \times 100 \tag{6.4}$$

in which R=load reduction percentage, L_h = average of historical TP daily loads (lb/d) and L_a=allocation (lb/d). If it is assumed that the daily mean flows are the same after load reduction, the concentration reduction is the same as the load reduction. The historical concentration and reduced concentration were calculated based on the total amount of TP loads and flows during the simulation period but not on statistics (average or mean of observations) in this chapter. (Equation 6.5).

$$C = \frac{\Sigma L}{\Sigma Q} \times f \tag{6.5}$$

in which $\sum L$= sum of daily TP loads (L=L_h for historical concentration and L=L_a for reduced concentration), $\sum Q$= sum of daily flows and f = conversion factor.

Two TP load reduction strategies are tested in this chapter. One is to reduce the TP load to allow 10 % of the predicted TP concentrations to violate the water quality criterion. The other is to reduce the estimated historical TP load to the allocation (TMDL - xσ) during a simulation period. The first (Frequency targeted approach) is based on statistics. MOS and the reduction percentage are determined by the only TP load corresponding to the 90th percentile of TP concentrations. Because the target is that the 90th percentile of TP concentration meets 0.05 mg/L, MOS are adjusted to satisfy this condition. The second approach (total mass targeted approach) is based on only allocation during simulation period. The MOS was 0.2 σ (standard deviation of loads for each season) and the total mass target (allocation) was TMDL – 0.2 σ.

Results

Most calibrated daily flows and daily TP loads from the R-D method and calibration are close to observed flows and TP loads from 1978 to 2004 (Figures 6.8, 6.9). It was not possible to verify all of the peak predicted loads and flows because the peaks of observations are not accessible in the EPA STORET or USGS data bases. According to these results, there were four peak loads from October, 1998 to September, 2000 (Figure 6.9). While the first and third peak represented annual peak loads during spring runoff, the second and forth peak loads were caused by unusually high TP concentrations. TP concentration observations at January 1, 2000 and August 10, 2000 were 0.79 mg/l and 1.88 mg/l (Figure 6.11). These estimated values, which were estimated by R-D method and calibration, are used for the TMDL as estimated historical values.

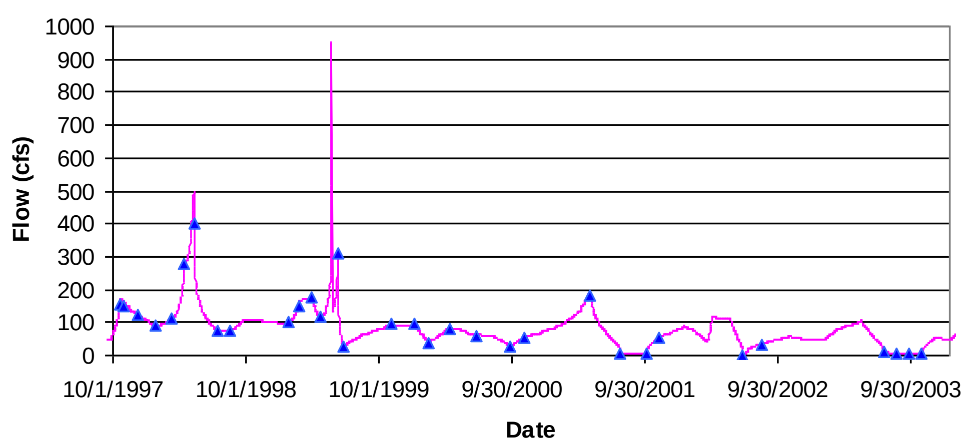


Figure 6.8—The flow time series at the TMDL location (▬ : estimated flows using R-D method, ▲ :observations).

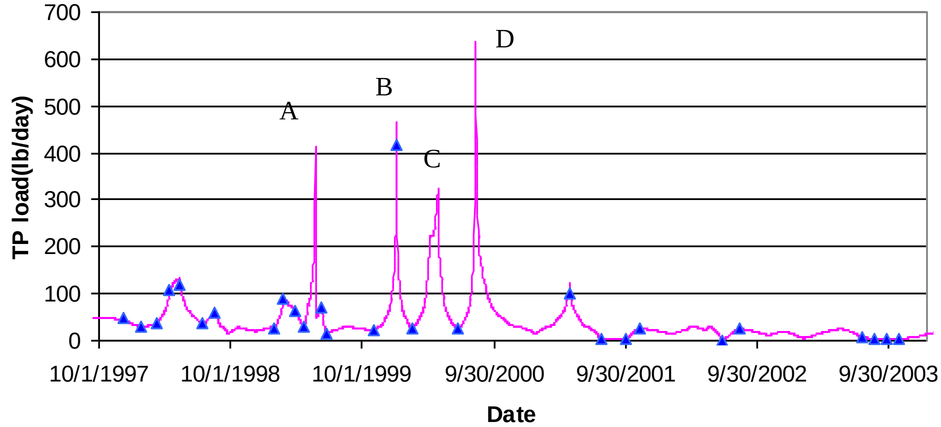


Figure 6.9—The TP load time series at the TMDL location (▬ : estimated loads using R-D method, ▲ :observations, A : 5/24/1999, B: 1/4/2000, C: 4/29/2000, D: 8/10/2000).

The TP load duration curves showed the trend of both flow and load change simultaneously along with the frequency of water quality standard violation. For example, the flow and TP load increased from March 1 to May 11, and then decreased from May 12 to June 30 simultaneously in 1998 (Figure 6.10 (b)). The load duration curve for March 1 to June 30 in 2003 showed same pattern of flows and TP load trend within high flow season as in 1998 (Figure 6.10 (d)). All plots showed the same direction, which means that the TP load and flow increased or decreased together, on the duration curve for high flow seasons. The plots on low flow season were irregular in 97-98 and 02-03 water years (Figures 6.10 (a) and 6.10 (c)). For example, the TP load decreased while flow increased from August 19 to September 30, 1998. Because the TMDL focuses on the violation frequency of water quality criterion (= 0.05 mg/L as TP), or allocation within a season using the sum of daily loads and flows in equation 6.5, these irregular trends of flow and TP load on low flow season may not significantly affect the

TMDL task. The vertical estimated load plots in the load duration curve means that the TP load increased without any change of flow (Figure 6.10 (a)). This unpractical prediction was caused by two observation points on July 15 and August 18, 1998. The flows are the same at two sampling dates while the TP load on August 18 (58.2 lb/d) is higher than July 15 (32.6 lb/d). In this case, the same rank is assigned to all dates between two dates for flows resulting in the same flow between two dates. Some possibilities for the irregular trend of flow and TP load will be discussed later.

There is no difference between the TMDL and the estimated historical load using the frequency targeted and total mass targeted approaches, but the different approaches provided the different MOS, allocation, and reduction percentages (Tables 6.2, 6.3). By the frequency targeted approach, the reduction percentage for low flow season was larger than high flow season in the 97-98 and 02-03 water years. The higher load reduction percentage was required to decrease the load by the TMDL in lower flow than the higher flow corresponding to the same difference between the historical load and TMDL in the frequency targeted approach (Figure 6.10 (a) and Figure 6.10 (c)). So, the high loads corresponding to the low flow duration interval caused high load reduction percentage in low flow seasons of the 97-98 and 02-03 water years when frequency targeted approach was used. According to the total mass targeted approach, the load reduction percentage for low flow season was smaller than for high flow season in the 97-98 water year (Table 6.3). Both approaches showed a significant difference between the load reduction percentages in wet and dry years.

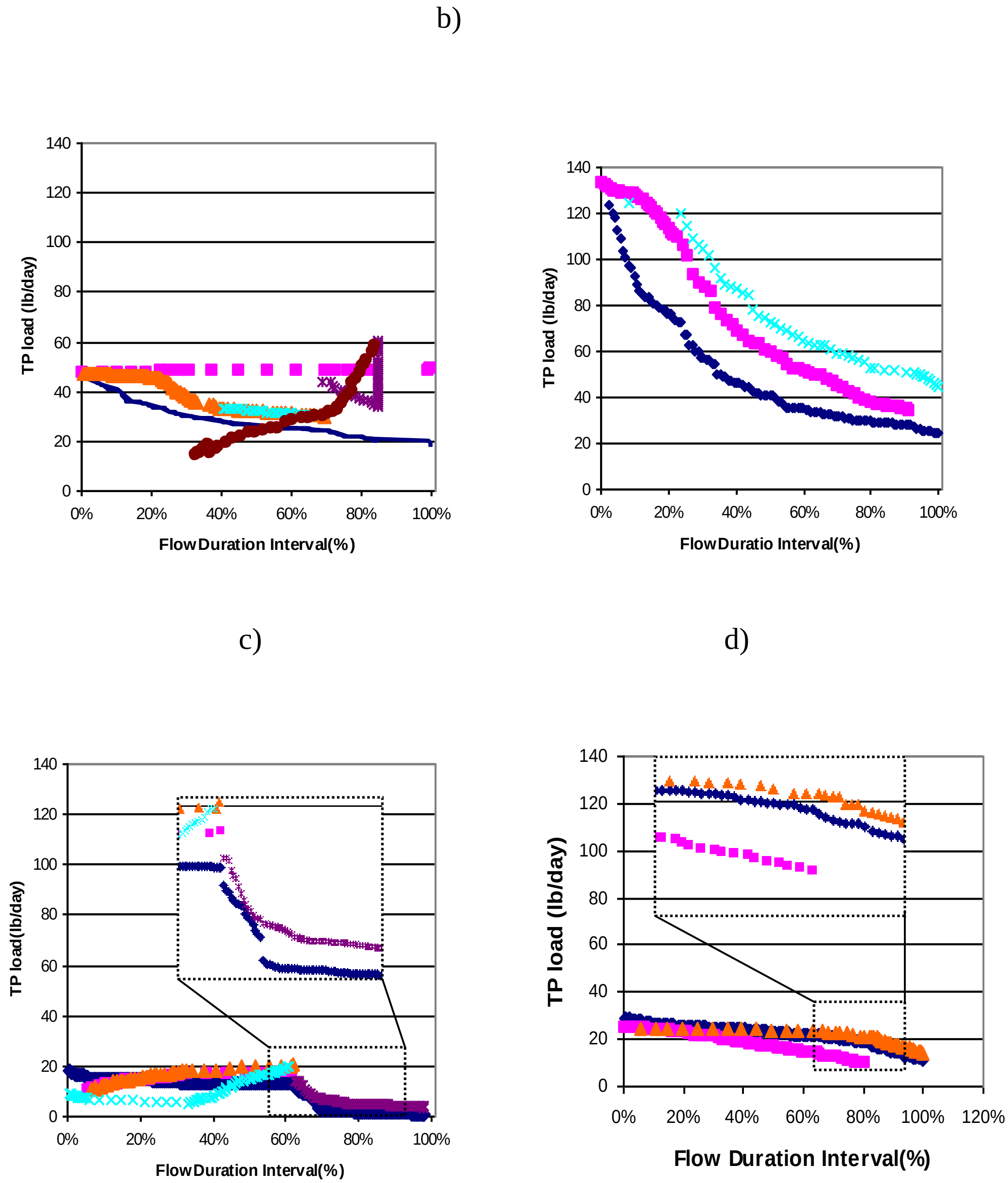


Figure 6.10—The TP load duration curve for TP: a) during low flow season in 1997-1998 water year (◆: TMDL, and ■▲xx◆ : calibrated TP loads for 10/1-10/21, 10/22-1/27,1/28-2/28. 7/1-8/19 and 8/20-9/30) , b) during high flow season in 1997-1998 water year (◆: TMDL, and ■x: calibrated TP loads for 3/1-5/11, 5/12-6/30), c) during low flow season in 2002-2003 water year (◆: TMDL, and ■, ▲,x,x: calibrated TP loads for 10/1-10/30, 10/31-12/13, 12/14-2/28 and 7/1-9/30), d) during high flow season in 2002-2003 water year (◆: TMDL, and ■, ▲: calibrated TP loads for 3/1-5/17, 5/18-6/30).

Table 6.2—TMDL, allocation, MOS and reduction percentage to meet the 10% frequency violation against numerical criterion (= 0.05 mg/L as P); Frequency Targeted mothod.

	97-98 water year			*02-03 water year*		
	Low flow season	High flow season	Total (Annual)	Low flow season	High flow season	Total (Annual)
Estimated historical TP load (lb/d)	39.16	76.61	51.67	10.51	19.19	13.41
TMDL (lb/d)	27.78	50.82	35.48	9.95	22.04	13.99
MOS (lb/d)	11.8 (=1.24 σ)	10.16 (=0.32 σ)	11.31	6.39 (=1.25 σ)	6.11 (=1.45 σ)	6.29
Allocation (lb/d)	15.98	40.66	24.17	3.56	15.93	7.70
Reduction percentage (%)	59.2	46.9	53.2	66.1	17.0	42.6
Historical TP Conc. (mg/L)	0.0705	0.0754	0.0728	0.0528	0.0435	0.0479
Reduced TP Conc. (mg/L)	0.0288	0.0400	0.0341	0.0179	0.0361	0.0275

* σ = standard deviation of TP loads.

Reduced TP concentration= (1-Reduction percentage) x Historical TP concentration

The historical TP concentration during a wet year is much higher than during a dry year. This means that the water was impaired in a wet year more than a dry year, and higher reduction percentage is required for wet years. The allocation (21.20 lb/day) is smaller than the estimated historical load (19.19 lb/day) for high flow season in the 02-03 water year using the total mass approach. This means that the stream has capacity to take a larger TP load of 2.01 lb/day and no TP load reduction was required. Most estimated TP concentrations were low in this season (Figure 6.11). The large reduction percentages were obtained in the frequency approach, and the reduction percentages seem to not be practical while the reduction percentages from the total mass targeted approach seem

Table 6.3—TMDL, allocation, MOS and reduction percentage using 0.2σ as MOS; total mass targeted method.

	97-98 water year			*02-03 water year*		
	Low flow season	High flow season	Total	Low flow season	High flow season	Total
Estimated historical TP load (lb/d)	39.16	76.61	51.67	10.51	19.19	13.41
TMDL (lb/d)	27.78	50.82	35.48	9.95	22.04	13.99
MOS (lb/d)	1.9 (=0.2σ)	6.35 (=0.2σ)	3.39	1.02 (=0.2σ)	0.84 (=0.2σ)	0.96
Allocation (lb/d)	25.9	44.47	32.09	8.93	21.20	12.36
Reduction percentage (%)	33.9	42.0	37.9	15.0	0	7.8
Reduced TP Conc. (mg/L)	0.0466	0.0438	0.0452	0.0449	0.0435	0.0442

* σ = standard deviation of TP loads.

Reduced TP concentration= (1-Reduction percentage) x Historical TP concentration.

practical. The reduced concentrations from the frequency targeted approach were very small (Table 6.2). This issue discussed below.

Discussion

Some estimated TP concentrations from the R-D method differed from observations (Figure 6.11). There were observed TP concentrations but not observed flows at April 11, 2000 (A in Figure 6.11) and March 12, 2001 (B in Figure 6.11) so that no observed TP load at those dates were involved in estimating loads by the R-D method. This caused overestimation of TP concentrations through overestimation of TP loads at those dates. This issue is recommended for future study.

The irregular trend of historical load duration plots may be caused by a point source load. A point source load affects the water quality significantly during the low

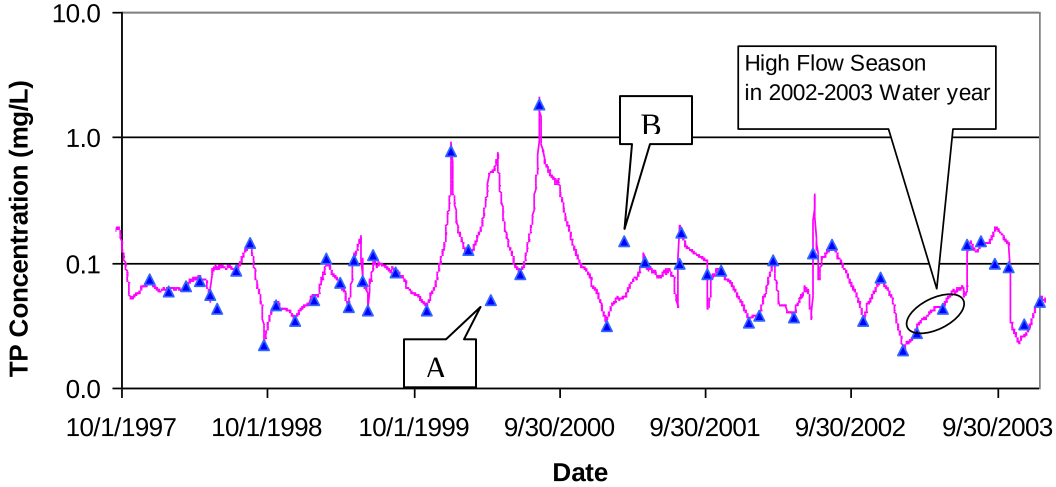


Figure 6.11—Total phosphorus concentration time series at TMDL point. —: estimated TP concentrations, ▲: observations.

flow season (Cleland, 2002). If a major point load, say the discharge from the Wellsville Lagoon decreased, and clean effluent from the Hyrum dam increased, the TP load may decrease without a decrease of flow.

The MOS is a critical issue of water quality management. If the MOS is too large, the allocation become too small and too much money must be paid to meet target load reductions. If the MOS is too small, allocation will be large and money will be saved, but the water will not have the appropriate quality. The frequency targeted TMDL approach depends on only one observation of all water resources data, the observation corresponding to the 90th percentile TP concentration for a "10% violation allowed strategy" or the observation corresponding to maximum TP concentration for a "no violation allowable strategy". For example, if the 90 percentile of all TP concentrations is 0.1 mg/L, the load reduction percentage should be 50% to meet the water quality criterion, 0.05 mg/L for 10% violation allowed strategy. Even though both the 90th

percentile and maximum concentration come from concentration data, the value may not be representative for some cases. Because the maximum or 90th percentile of TP concentration is a simple order statistic, the concentration distribution under or over the 90th percentile, the 90th percentile may be ignored. These characteristics of the frequency targeted TMDL approach may provide too large or too small MOS. In this study, the MOS seem to be too large in the frequency targeted approach because the reduced concentrations are very low (Table 6.2) and even the predicted load after reduction above TMDL curve are very close to TMDL curve (Figures 6.12, 6.13)

The total mass targeted TMDL approach may be more practical in the area with large variance of load or concentration. The purpose of this approach is to reduce the historical TP load (Total mass of TP) to the total allocation during the simulation period. In this approach, because the sum of loads on all dates during the simulation period is compared to the sum of allocations, the reduction percentage depends on all TP loads. In this approach, when the loads are more widely distributed with high σ, the possibility of a water quality standard violation is high under same historical load and TMDL. The larger MOS (larger 0.2σ) for more widely distributed load may produce the appropriate allocation with flexibility.

When $x\sigma$ is used as a MOS, the range of x may be another issue but 0.2 may be enough at the mouth of the Little Bear River based on two observations. The first is that the reduced TP concentrations seem to be small enough. The range of MOS was 6.8 to 12.5 % of TMDL, and the range of reduced concentration was 0.0438 to 0.0466 except during high flow season in the 2002-2003 water year. The MOS of the high flow season in the 2002-2003 water year is small but the load reduction is not required for this case.

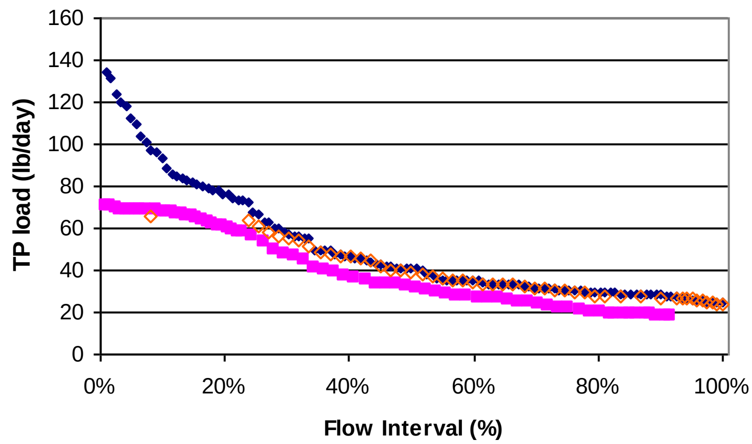


Figure 6.12—The TP load duration curve for TP during high flow season in 97-98 water year (◆: TMDL, and ■ ◇: calibrated TP loads for 3/1-5/11, 5/12-6/30 after reduction to meet 10% frequency violation against 0.05 mg/L).

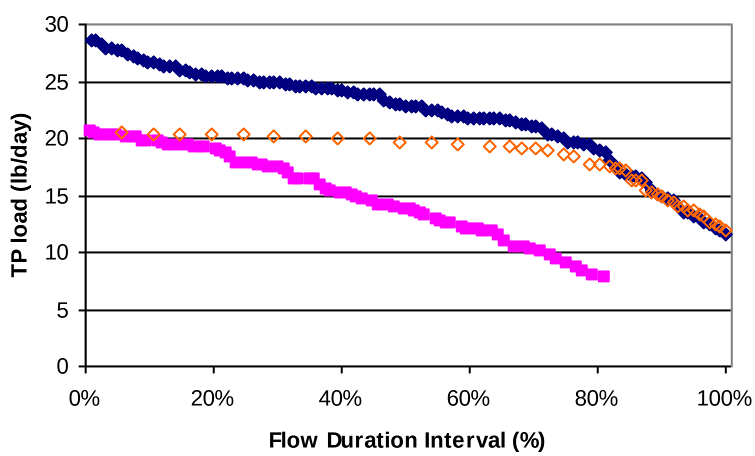


Figure 6.13—The TP load duration curve for TP during high flow season in 2002-2003 water year (◆: TMDL, and ■ ◇: calibrated TP loads for 3/1-5/17, 5/18-6/30 after reduction to meet 10% frequency violation against 0.05 mg/L).

The second is that the load reduction percentages were estimated separately for dry and wet years. If the reduction percentage for the wet year total is selected as a target, this may be a safe value.

The total mass targeted TMDL required daily mean flows and daily concentrations to estimate daily loads. Using low frequency data may cause inappropriate historical load, TMDL, or MOS. Even though many commercial hydrological and water quality models are available to predict daily frequency flow and TP concentration, data of some variables associated with prediction of flows or TP concentrations may not sometimes be accessible. In this case, the R-D method (Chapter 4) is an alternative because this method requires only low frequency flows, TP loads, and the times of peak flow and TP load to predict daily flows and loads.

Summary and Conclusions

The R-D method was used to fill in data gaps of flows and TP loads. The time series of cumulative failure probabilities (Rank time series) and cumulative failure plot (data distribution) of extended observations are required for this method. The estimation of flows and TP load was enhanced by the regression between extended observations and predictions from the R-D method.

Daily TP loads and daily mean flows from the R-D method were used for TMDL calculation at the mouth of the Little Bear River. The TMDLs and historical TP loads are calculated by these daily flows and TP loads, instead of the mean or median from low frequency data.

The reduction percentages were calculated for four different categories, low flow (July to February) and high flow (March to June) for a wet year (97-98 water year) and those for a dry year (02-03 water year) by two different approaches, the frequency targeted approach and the total mass targeted approach.

Higher reduction percentages are required for wet year than that for a dry year according to both approaches. The reduction percentage from the frequency targeted approach is higher than that from the total mass targeted approach because the large MOS was applied to reduce the frequency of water quality violation to 10 %.

When the 0.2σ of TP load are used as MOS, more practical reduction percentages and reduced concentration were obtained using the total mass targeted approach.

References

Alabama DEM (Alabama Department of Environmental Management) (2005) *Total Maximum daily Load for Upper Brindley Creek*. AL/03160109-030_01, Alabama Department of Environmental Management, Water Division, Montgomery, Alabama.

Chess, D. (2000) Utah Watershed Review Vol 8, Utah Division of Water Quality.

Cleland, B.R. (2002) TMDL development from the “bottom up” – Part II: Using Duration Curves to Connect the Pieces. *Proc. of the National TMDL Science and Policy 2002 Specialty Conf., November 2002*.

EMRG (Environmental Management Research Group at Utah Water Research Laboratory) (2004) Little Bear River Watershed boundary shapefile.

Helsel, D.R. (2005) Nondetects and Data Analysis. John Wiley & Sons, Inc.: Hoboken, New Jersey.

Kaplan, E.L.; Meier, P. (1958) Nonparametric Estimation from Incomplete Observation, *J. American Statistical Assoc.* **53** (238), 457-481.

Mason C. (2002) *Biology of Freshwater Pollutant*, 4th ed; Pearson Education Limited: Essex, England.

Minitab Inc. (2006) Minitab® 14 http://www.minitab.com/support/macros/default.aspx?action=display&cat=non *accessed* July.2006.

Neilson, T. B.; Stevens, D. K; Horsburge, J. S. (2005) TMDL Development Approach, in *Total Maximum Daily Load: Approaches & Challenges*, Younos, T. (Ed.), pp 47-78. PennWell Corporation: Tulsa, Oklahoma.

NARA (National Archives and Records Administration) (2000) Federal Register Vol 65. No. 135. Part 6 Environmental Protection Agency.

Smith, E.P.; Ye, K.; Hughes, C.; Shabman, L. (2001) Statistical Assessment of Violations of Water Quality Standards under Section 303(d) of the Clean Water Act, *Environmental Science and Technology* **35** (3), 606-612.

Stiles, T.C. (2001) A Simple Method to Define Bacteria TMDLs in Kansas. *National TMDL Science Issues Conference*, March 3-6, St. Louis, Missouri. pp.375-378.

Utah DEQ (Utah Department of Environmental Quality) (2000) Little Bear River Watershed TMDL. Salt Lake City, Utah.

USEPA (2004) Better Assessment Science Integrating Point and Non Point Sources (BASINS). http://www.epa.gov/waterscience/BASINS/ *accessed* November, 2004.

USEPA (2005) STORET (Storage and Retrieval), http;//www.epa.gov/stormoda/DW_stationcriteria. *accessed* January, 2005

USEPA (1993) TMDL case study: Denver Metro The South Platte River Segment, EPA841-F-93-001. http://www.epa.gov/owow/tmdl/cs1/cs1.htm.

USEPA (2007) An Approach for Using Load Duration Curves in the Development of TMDLs. EPA 841-B-07-006.

USGS (2004) National Hydrograph Dataset. http//nhd.usgs.gov/data.html. *accessed* November, 2004.

Utah DEQ (Utah Department of Environmental Quality) (2000) Little Bear River Watershed TMDL: Salt Lake City, Utah.

Younos, T. (2005) Preface, in *Total Maximum Daily Load: Approaches and Challenges*, PennWell Corporation: Tulsa, Oklahoma.

CHAPTER 7

HIGH RESOLUTION BAYESIAN NETWORK TO EVALUATE EFFECTS OF WATER QUALITY CONSERVATION PRACTICES

ABSTRACT

The Little Bear River watershed, Northern Utah is on a high-priority list of watersheds that are being adversely affected by nonpoint source pollution. Reduction of Total phosphorus (TP) concentration has been observed since 1990 at the mouth of the Little Bear River. A Bayesian Network was used to simulate the effect of the Little Bear River Conservation Project (LBRCP) and exogenous variables including point source load, agricultural landuse and annual precipitation on the reduction of TP concentration. In this chapter, the estimated 9652 flows, TP loads and concentrations from Rank-Data distribution method (R-D method) were used to support 21 categories of those variables in the Bayesian Network. The cumulative failure plots (distribution of predicted values) of TP concentration were obtained using the Bayesian Network under different scenarios. The LBRCP decreased the TP concentration significantly only in wet years. Increased agricultural land area and point source load caused higher TP concentration but the annual precipitation increases caused only small increases of TP concentration because TP load and flow increased simultaneously. The concentration's probabilities from the BN were linked to rank time series of TP concentration by the R-D method to obtain time series of TP concentration under a simulation scenario. According to this time series, the LBRCP caused a longer duration and earlier starting time of TP concentration equal to or below 0.05 mg/L under wet annual precipitation condition.

INTRODUCTION

The Little Bear River in Cache County, Northern Utah flows from southeast to northwest, bounded by mountains, and drains to Cutler Reservoir, west of Logan, UT (Figure 7.1). The Little Bear River watershed is on a high-priority list of watersheds that are being adversely affected by nonpoint source pollution (Chess, 2000). In the watershed below Hyrum Reservoir, approximately 50.4% is cropland and pasture, 43.8% is range, forest, water body and riparian, and 5.8% is urban area (Utah DNR, 2004).

Since the USDA started the Little Bear River Conservation Project (LBRCP) for best management practices (e.g., fencing, vegetation buffer and bank restoration) in 1990 (USEPA, 2006; EMRG, 2007), Total Phosphorus (TP) concentrations have decreased gradually at the mouth of Little Bear River. The water below Hyrum Reservoir is impaired more than that above Hyrum Reservoir (Utah DEQ, 2000). Our goal is to explore the causes of the nutrient concentration reduction at the mouth of the Little Bear River because exogenous forces including land use changes and climate changes as well as nutrient management (LBRCP) can affect the water quality at management sites.

Bayesian Networks (BNs) were designed to accept and process inputs of varied types of information: observations, model results, expert judgment, scenario types and a variety of other non-numerical inputs (Marcot *et al.*, 2001; Varis and Jussila, 2002; Borsuk *et al.*, 2003). BNs are designed to evaluate the effects of two or more variable combinations (scenarios) on other variables (Marcot *et al.*, 2001). However, most previous BNs provide the probabilities of a few categories for pollutant concentrations (Marcot *et al.*, 2001; Varis, 1998; Borsuk *et al.*, 2003). This characteristic has caused the difficulty in interpretation of probabilistic outputs.

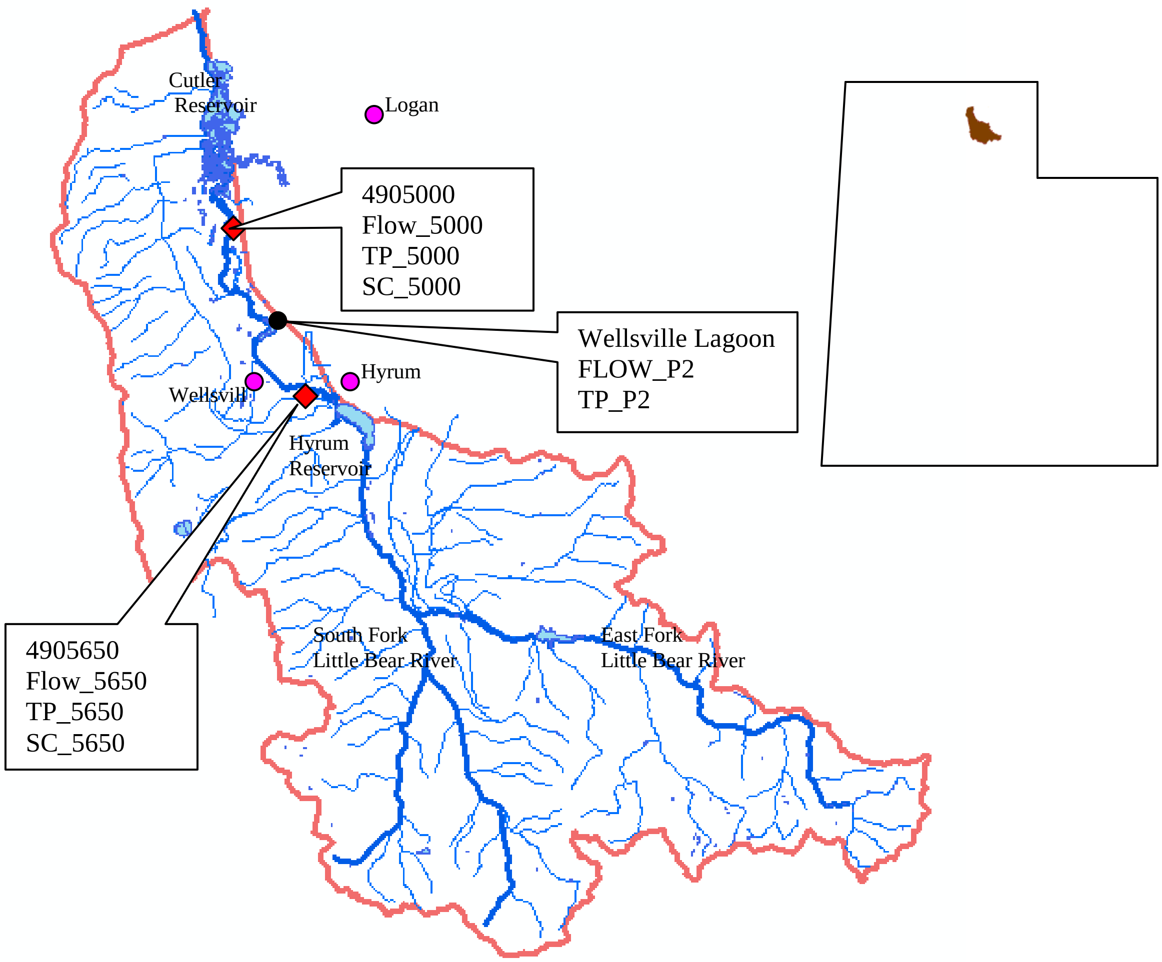


Figure 7.1 | Little Bear River Watershed located in Northern Utah (EMRG, 2004; USGS, 2004; USEPA, 2004). Flow_ xx = Flow (cfs), TP_xx = TP concentration (mg/l) and SC_xx = Specific conductance (umho/cm). xx is last four digits of station ID.

In this chapter, predicted flows, TP loads and concentrations from Rank- Data distribution method (R-D method) (Chapter 4) were collected and added to the BN database for the subwatershed below Hyrum Reservoir. Because the number of flows, TP loads and concentrations at the mouth of the Little Bear River in BN database was large, a large number of levels could be used for those variables (Appendix J). The probabilities from those categories may be converted to a cumulative failure plot for every different TP concentration and load.

The hypothesis variable was the TP concentration at the mouth of the Little Bear River. The effect of conservation practices on phosphorus concentrations at the mouth of the Little Bear River (at location 4905000 in Figure 7.1) were evaluated by comparing the cumulative failure plot of the TP concentrations during the conservation practice period to those before the conservation practices were implemented. The exogenous variables were annual precipitation, agricultural landuse area and point source loads. The effect of each exogenous variable on phosphorus concentrations was evaluated by comparing the cumulative failure plot of the TP concentrations for the one selected category of the exogenous variable to those for another selected category. The effects of the combination of annual precipitation category and conservation practices option on the TP concentration in the stream were evaluated in the same manner. The probabilistic outputs from the BN were connected with the rank time series from daily TP concentrations to obtain the concentration's time series during a simulation period by the R-D method. This time series supported evaluation of the duration and timing for the violation of water quality standards under each simulation scenario of conservation practice and exogenous variables.

METHODS

Bayesian Network

A Bayesian Network (BN) is a probabilistic network model based on graphical relationships among variables (Castillo *et al.*, 1997). In a BN, the relationships between parent variables and child variables are logically expressed in a link and node structure where the state of the parent node predicts the state of the child node (Jensen, 1996).

Conditional Probability Tables (CPTs) show the probability of each discrete state, given the states of any parent nodes (Marcot *et al*., 2001). The marginal probabilities are calculated using CPTs.

BN have been used previously in water quality assessment. For example, Reckhow (1999) constructed a BN model for evaluating the effect of forest buffer on anoxia probability in the Neuse River estuary, North Carolina (Figure 7.2).

The anoxia probabilities were calculated using conditional probabilities of combinations among percent forest buffer, nitrogen load reduction, precipitation conditions, flow and algal bloom variables (Figure 7.2). The probabilities of percent forested buffer translate to the fraction of the river reach that has a proposed percentage with a forest buffer. For example, if it is proposed that 20 % of a river reach has 70-80% of its length with a forested buffer, then *p(70-80 % forested buffer)* = 0.2 based on the

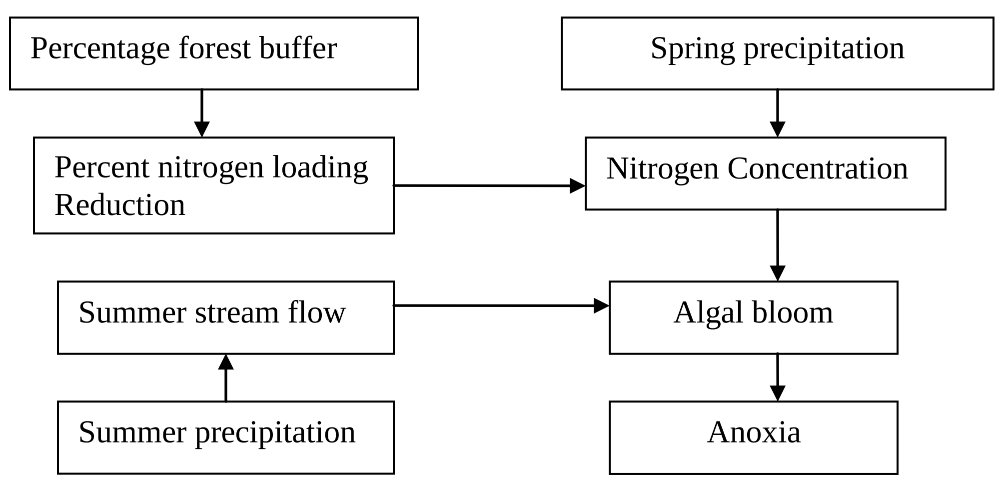


Figure 7.2 | Schematic of an anoxia model (Reckhow, 1999).

proposal of North Carolina Environmental Management Commission (NCDENR, 2008).

So, in Reckhow (1999), two cases were examined. In the first case, the entire river reach was predicted to be 95-100% buffered (*p(95–100% forested buffer)*=1). Under this percentage of forested buffer, *p*(anoxia) was 0.27. The second case was *p(70-80 % forested buffer)*=0.2, *p(80-95 % forested buffer)*=0.6 and *p(95-100 % forested buffer)*=0.2 (proposed forest buffer percentage). Under this proposed case, *p(anoxia)* was 0.3. Because the increase percentage of probability of anoxia was only 10% (from 0.27 to 0.3), changing the scenario of probabilities for percent forested buffer, proposed percent forested buffer might be more efficient.

Water-related data

A database was constructed to calculate CPTs supporting a BN. All existing and predicted inputs: TP concentrations, flows, precipitation, agricultural landuse area and water quality conservation option were organized by variable (Table 7.1), and the BN was connected to this data base. The sampling locations included the outlet from Hyrum Reservoir (4905650), inlet to Cutler Reservoir (4905000) and a discharge of Wellsville Lagoon (Figure 7.1). The numbers of flow and TP concentration data of two permitted discharges (discharges of Northern Utah Manufacturing) were small, but all of the TP loads from these point sources were small compared to the TP load from Wellsville Lagoon (Chapter 5). So, ignoring these point sources may be acceptable.

The TP loads were calculated by multiplying the flow by the TP concentrations and converting the units to lb/day at these two sampling locations. It is possible to calculate the TP load when there are both flow and TP concentrations. Linear regression

Table 7.1 | Critical variables for evaluation of conservation project in the Little Bear River Watershed below Hyrum Reservoir

Variables	Description	Type	References
OP_CON	Conservation Project Option, Pre: Before starting the Project (1974-1989), Post: After starting the Project(1990-2004)	Decision	
LAND_AG2	Area of Agricultural Land Use. Interpolating or extrapolating Agricultural land use area among 1986, 1996 and 2003.	Exogenous	Utah DNR (2004)
PRECIP	Annual Precipitation for each water year at Logan Radio KVNU (Station No. 425182).	Exogenous	Western Region Climate Center (2006)
LOAD_P2	TP load (lb/day) from point source (Wellsville Lagoon)	Exogenous	EPA STORET (2005)
LOAD_5650	TP load (lb/day) at the effluent from Hyrum Reservoir (Location 4905650)	State	EPA STORET
LOAD_5000	TP load (lb/day) into Cutler Reservoir (Location 4905000)	State	Prediction From R-D method
LOAD_SW2	TP point and nonpoint load from subwatershed above the stream reaches between location 4905650 and location 4905000 (inlet to Hyrum Reservoir)	State	Estimated from Flow and TP concentration in EPA STORET
FLOW_5650	Flow (cfs) at the effluent from the Hyrum Reservoir (Location 4905650)	State	EPA STORET
FLOW_5000	Flow (cfs) into Cutler Reservoir (Location 4905000)	State	Prediction From R-D method
FLOW_SW2	Flow from subwatershed above the stream reaches between location 4905650 and location 4905000	State	Estimated from flow in EPA STORET
TP_5000	TP concentration (mg/L) into Cutler Reservoir (Location 4905000)	State	Prediction From R-D method

between upstream and downstream values was use to fill in the missing flows or TP concentrations at locations 4905650 and 4905000 (Chapter 5). In order to support the 21 categories of probabilistic outputs at location 4905000, the low frequency observations or predictions of flow, TP load and concentrations (FLOW_5000, LOAD_5000 and TP_5000) were replaced by the 9652 predicted flows, TP loads and concentrations (high frequency database) from the R-D method for the period from 1978 to 2004 (Chapter 6).

Bayesian Network (BN) construction

Netica version 3.17 (Norsys Software Corp., 1997) was used to build the BN. Contingency tables are calculated by Netica in the BN based on the data in the high frequency database. The BN estimated the effects of conservation practice (LBRCP) and exogenous variables on the TP load and TP concentrations at the mouth of the Little Bear River. There are two variable groups, the TP load group and the flow group in the BN (Table 7.1). The TP load group includes effluent TP load from the Hyrum Reservoir (LOAD_5650), TP load at the mouth of the Little Bear River (LOAD_5000) and a point TP load from Wellsville Lagoon (LOAD_P2). Upstream TP load variables were linked to their downstream counterparts. In the same way, the flow group included flows of all sampling locations, and upstream flow variables were linked to those downstream (Figure 7.3).

The conservation practice option and exogenous variables were linked to subwatershed load (LOAD_SW2), the TP load from point sources and non point sources to the stream between location 4905650 and 4905000. LOAD_SW2 was calculated by subtracting TP load at location 4905650 (LOAD_5650) from TP load at location 4905000

(LOAD_5000). The annual precipitation (PRECIP) was linked to subwatershed flow (FLOW_SW), which is the flow from the watershed to the stream between locations 4905650 and 4905000 and was calculated by subtracting flow at location 4905650 from flow at location 4905000. Similarly, FLOW_SW is a parent variable of FLOW_5000. Finally, the TP loads (LOAD_5000) and flows (FLOW_5000) into Cutler Reservoir were connected to TP concentration (TP_5000) into the reservoir (Figure 7.3).

Categorizing variables of high resolution BN

The conservation project option (OP_CON) is the only decision variable (Figure 7.3). The level Pre is before starting conservation practices and Post is after starting conservation practices. There are three exogenous variables, agricultural land use area

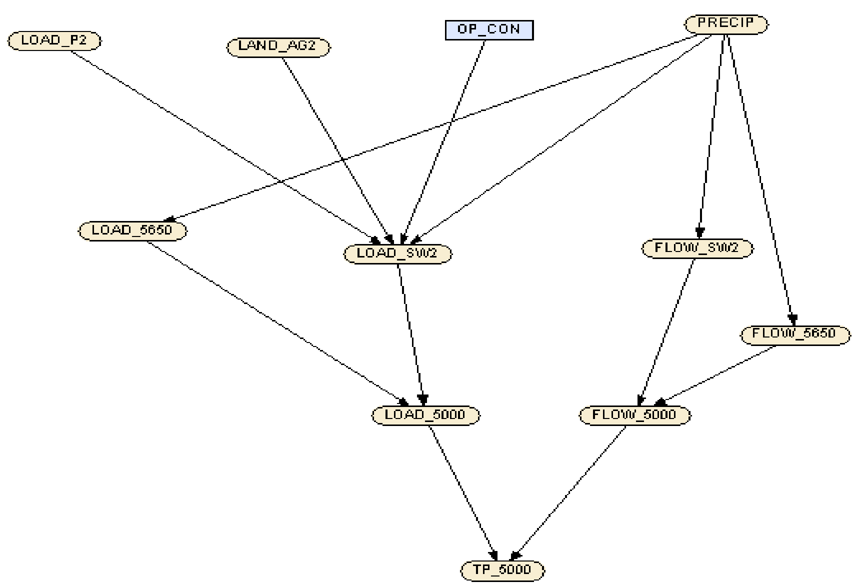


Figure 7.3 | Little Bear River BN for the LBRCP and exogenous variable effect evaluation (Definitions of variables are in Table 7.1).

(LAND_AG), point source TP load (LOAD_P), and annual precipitation (PRECIP). OP_CON and exogenous variables have no parent nodes (Figure 7.3). LAND_AG, LOAD_P and PRECIP have two categories each (Table 7.2).

The database for this BN has no point source TP load data before 1990 (before starting conservation practices). Netica cannot calculate reliable conditional probabilities, *p(LOAD_SW2|OP_CON, LAND_AG2, LOAD_P2, PRECIP)* for the case of OP_CON= Pre in this situation because all of LOAD_P2 was empty for data combinations of the four variables, OP_CON, LAND_AG2, LOAD_P2, PRECIP when OP_CON is Pre. A challenge of the BN is using the contingency table of subwatershed TP load in the upstream BN (BN for above Hyrum Reservoir) for conditional probability *p(LOAD_SW2|OP_CON, LAND_AG2, LOAD_P2, PRECIP)*. Because the pattern of flows and TP loads of two upstream and down stream BNs are similar each other (Chapter 5), it may be acceptable using the CPT of the upstream subwatershed load,

Table 7.2 | Category boundary of variables for the BN

Variables	Category	Range
Agricultural Land Use(LAND_AG) as Acres	L H	<24370 >=24370
Point Source TP load (Load_P) as lb/day	L H	<10.86 >=10.86
Annual Precipitation as inches	D W	<15.4 >=15.4

p(LOAD_SW1|OP_CON, LAND_AG1, LOAD_P1, PRECIP) for CPT of *p(LOAD_SW2|OP_CON, LAND_AG2, LOAD_P2, PRECIP)* of the down stream BN (Appendix I).

Each TP load or flow variable from the subwatershed and at location 4905650 has three categories, L (Low: smaller than 33 percentile value), M (M: from 33 percentile value to smaller than 67 percentile value) and H (High: equal to or larger than 67 percentile value) (Appendix I). The TP concentration, load and flow at the inlet to the Cutler Reservoir (TP_5000, LOAD_5000 and Flow_5000) were categorized as *A to S* (19 levels) using every 5th percentile values from the 5th percentile to the 95th percentile from the BN database. The category boundary between T and U level was the 99.5 percentile (Appendix J). Finally, this BN estimated high resolution probabilities of TP_5000, LOAD_5000 and Flow_5000 with 21 categories for each variable (Figure 7.4).

Bayesian Network simulation

The purpose of the Little Bear River BN is to evaluate the effects of the conservation practices (LBRCP) and exogenous variables on the TP load and TP concentration into Cutler Reservoir. In order to evaluate the effect of conservation practices, the conservation practice option (OP_CON, Evaluated variable) was selected as Pre or Post, and the probabilities of TP load and concentration variables (Compared variables) were calculated for these selections. The probability of each category for agricultural land area (LAND_AG2), Point TP load (LOAD_P2) and annual precipitation (PRECIP) came from the BN database because the categories of these variables were set at their marginal probabilities (Figure 7.4, Table 7.3). The probability of a category of an

.**Table 7.3** | The scenarios of variables for each simulation

Scena-rio No.	**Evaluated Variable**	**Selected Category of variables**				**Compared variable**
		OP_CON	LAND_AG	LOAD_P	PRECIP	
1	OP_CON	Pre or Post	None	None	None	TP_5000
2	LAND_AG2	None	L or H	None	None	TP_5000
3	LAOD_P2	None	None	L or H	None	TP_5000
4	PRECIP	None	None	None	D or W	LOAD_5000 FLOW_5000 TP_5000
5	OP_CON Under Dry Precipitation Year	Pre or Post	None	None	W	TP_5000
6	OP_CON Under Wet Precipitation Year	Pre or Post	None	None	D	TP_5000

exogenous variable was set as 1.0 to evaluate the effect of that variable. For example, the probability of High (H) or Low (L) category of annual precipitation (PRECIP) was set as 1.0 to evaluate the effect of PRECIP (Scenario No. 4 in Table 7.3).

One interesting issue is under which annual precipitation condition the LBRCP had a larger effect on TP concentration. This is done by comparing how much the probability distribution of TP_5000 changed when the OP_CON category was changed from Pre to Post under dry (D) or wet (W) conditions of annual precipitation (PRECIP) (Table 7.3).

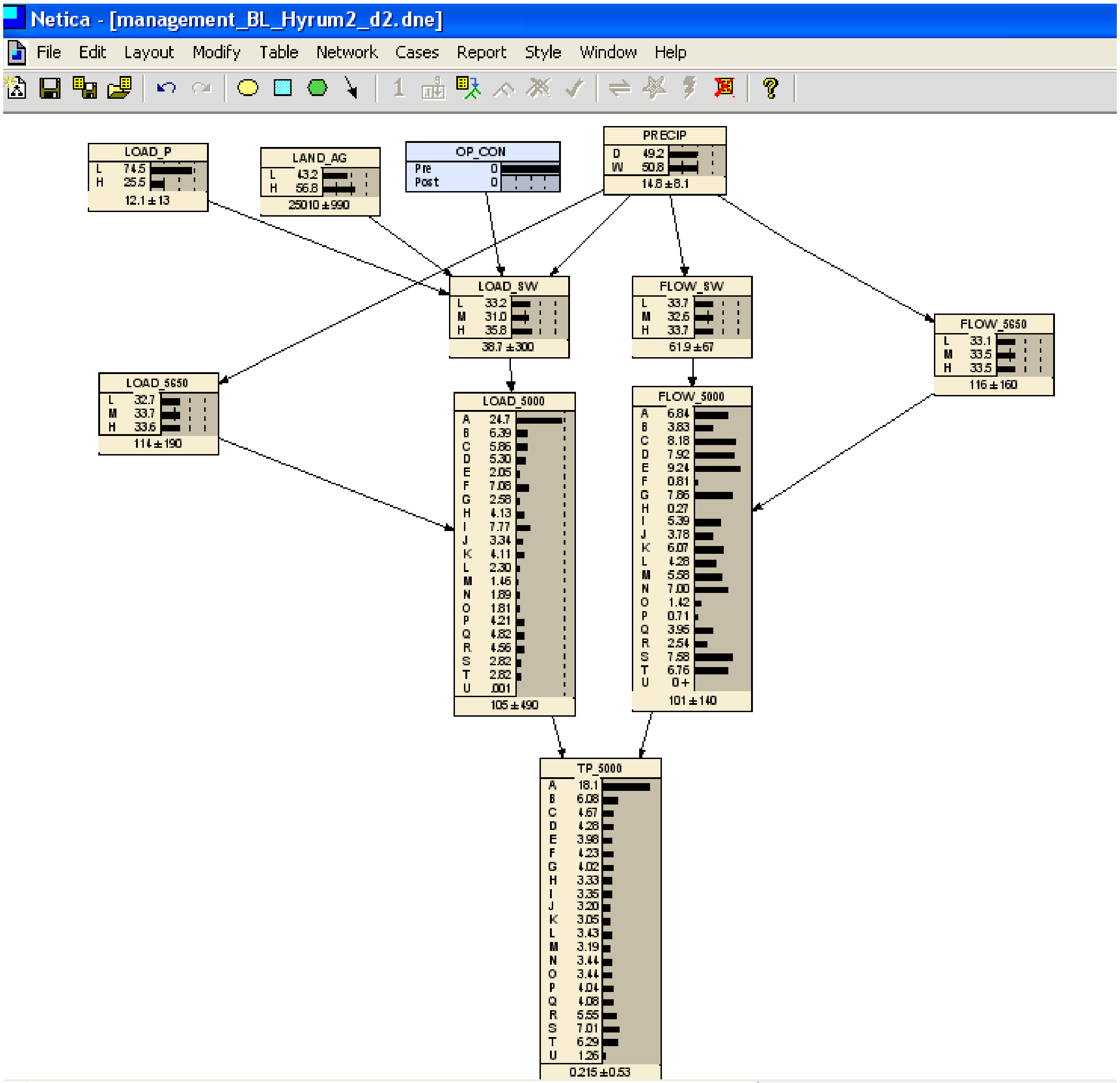


Figure 7.4 | The outputs of the high resolution Bayesian Network below Hyrum Reservoir for pre LBRCP condition (OP_CON=Pre). The blue box is the decision variable. Yellow boxes are exogenous and state variables.

The probabilities for every different categories of TP_5000 from BN simulations were converted to Cumulative Failure Probabilities (CFP)[2] (Kaplan & Meier, 1958) for the category boundaries by summing the probabilities. For example, the probability for

[2] Cumulative Failure Probability (CFP) is identical to cumulative probability (Sheskin, 2004) in this chapter, but CFP was used instead of cumulative probability for consistency among Chapters 4, 6 and 7.

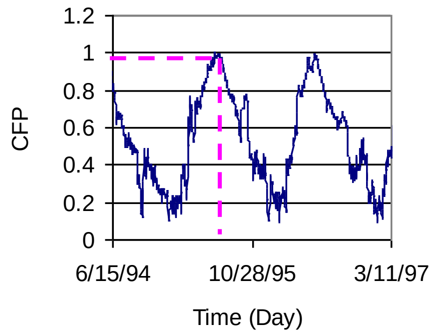
first category of TP _5000 from 0 to 0.0429 mg/L as TP was 18.1% and the probability for the second category from 0.0429 to 0.0527 mg/L was 6.08% in BN output under Pre OP_CON conditions, so the CFP for 0.0527 mg/L, the boundary of second category was 24.18 % (=18.1% + 6.08%). The plots of CFPs vs. the categorical boundary of TP concentration were constructed from these CFPs for each category boundary and then the 71 TP concentrations for every different percentile from the 25 percentile to the 90 percentile were found from the cumulative failure plots.

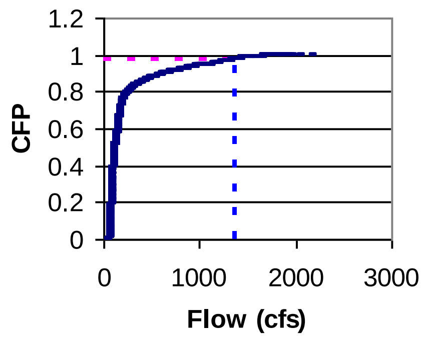
The Q-Q plots (quantile-quantile plots) between these pairs of 71 TP concentrations from the two different scenarios were used to evaluate the effect of conservation practice or each exogenous variable. A Q-Q plot is a statistical tool that plots the quantile of one data set against the same quantile of the other data set corresponding to the same cumulative probability (Gilchrist, 2000). For example, if the quantile of data set 1 at the cumulative probability of 50% is 0.05 mg/L and the same quantile of data set 2 is 0.045, a point is plotted at (0.05, 0.045). This type of plot is used to compare two distributions (Gilchrist, 2000). If a plot is on the line of perfect agreement, a TP concentration under a scenario is the same as the TP concentration under the other scenario corresponding to the same percentile. A plot far from the agreement line means that there is a big difference between the two TP concentrations corresponding to the same quantile.

Rank-Data distribution method (R-D method)

The R-D method is an approach to fill in data gaps of values of a variable by linking a data distribution (cumulative failure plot) to the rank time series of that variable

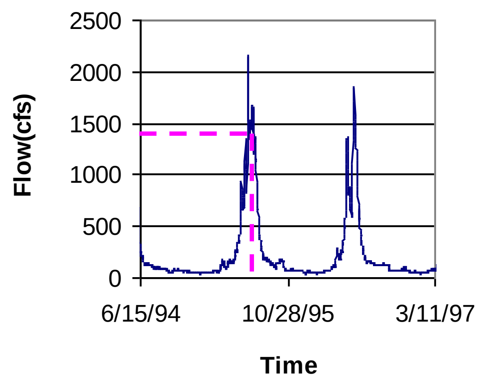
(Chapter 4). For example, if the Cumulative Failure Probability (Kaplan & Meier, 1958) on July 2nd, 1995 is 0.98 on the time series of the rank and the flow for CFP = 0.98 is 1370 cfs on the cumulative failure plot, the flow on July 2nd ,1995 is estimated to be 1370 cfs.





a) Rank Time Series

b) Data distribution of flows as CFP.



c) Constructed Time Series of Values

Figure 7.5 | The process of estimating time series by the R-D method combining data distribution with CFP time series. (USGS gage 10128500, Weber River near Oakley, UT).

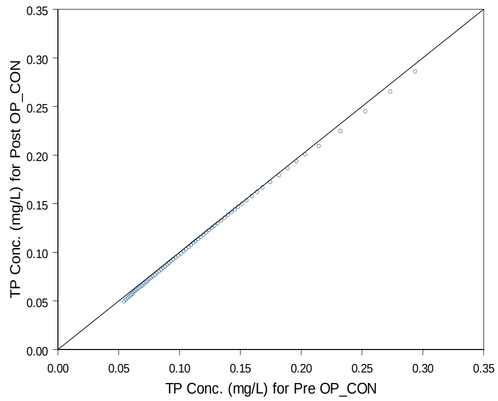
The process is shown in Figure 7.5. In this chapter, the R-D method was used to estimate time series of TP concentration under scenarios of conservation practices.

It is difficult to evaluate the effect of LBRCP on the timing or duration of the water quality standard violation by comparing the TP concentration time series in any year before starting LBRCP to those in any year after starting LBRCP. The exogenous variables may affect the TP concentration and the hydrological characteristics including timing and duration of high flow season (Spring runoff) in any year differs from other years. If it may be assumed that the rank time series of TP concentration depends on only hydrological characteristics, it may be acceptable as an evaluation method for timing or duration comparing one year TP concentration time series from the R-D method under one scenario to that under the other scenario. The cumulative failure plot for the probabilistic BN outputs of TP _5000 under Pre OP_CON (Scenario 1) or Post OP_CON condition (Scenario 2) was linked to rank time series of the TP concentration in the 97-98 water year to obtain 1 year time series of TP concentration under scenario 1 or scenario 2. The rank time series of TP_5000 came from the daily TP concentration from the BN database.

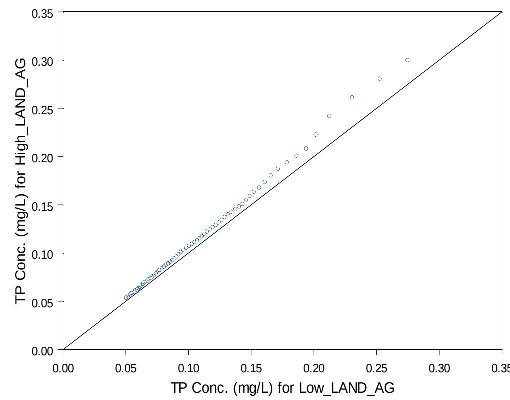
RESULTS

Effects of conservation practice (LBRCP) and exogenous variables

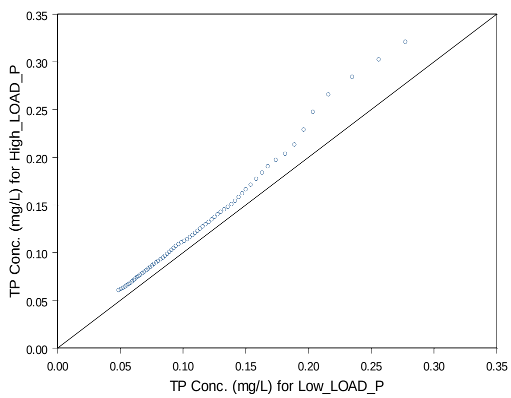
According to the Q-Q plot (Figure 7.6 a), the difference between TP concentration of TP _5000 under Post OP_CON (After starting conservation practices, scenario 1) and that under Pre OP_CON (Before starting conservation practices, scenario 1) corresponding to same CFP was small. The effect of PRECIP (Annual precipitation) a)



b)



c)



d)

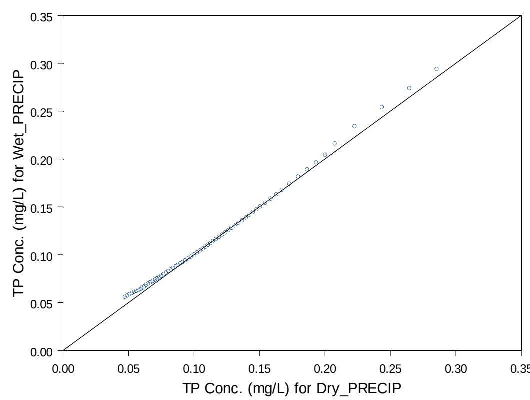


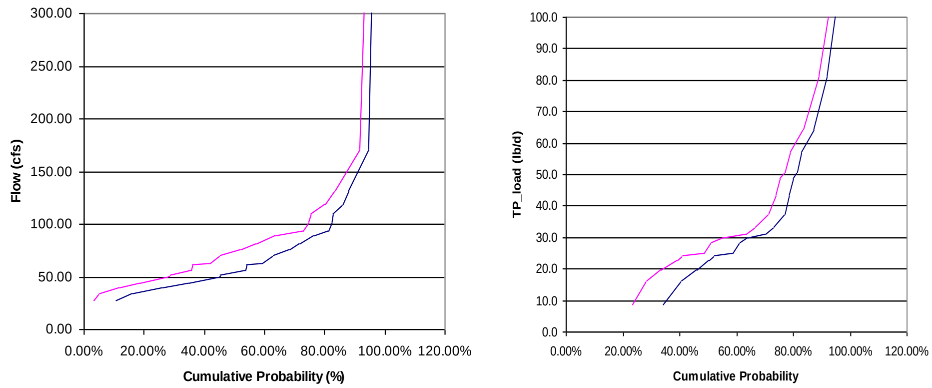
Figure 7.6 | Q-Q plot for Scenarios: a) Scanario1, comparing the same percentile's TP concentrations for Pre OP_CON and Post OP_CON, b) Scenario 2, comparing the same percentile's TP concentrations for L (low) LAND_AG and H (High) LAND_AG, c) Scenario 3, comparing the same percentile's TP concentrations for L (low) LOAD_P and H (High) LOAD_P. d) Scenario 4, comparing the same percentile's TP concentrations for D (Dry) PRECIP and W (Wet) PRECIP. (▬ : line of perfect agreement).

on the TP concentration was smaller than the effect of LAND_AG (Agricultural landuse area) or LOAD_P (point source load) when Q-Q plots among scenario 2, 3, and 4 were compared (Figure 7.6 b, 7.3.1 c, and 7.3.1 d). The PRECIP has two child variables, FLOW_SW2 and LOAD_SW2 while OP_CON and other exogenous variables have only child variable, LOAD_SW2. When the selected category of PRECIP was changed from D (Dry) to W (Wet) in scenario 4, each flow and TP load corresponding to the same CFP increased simultaneously (Figure 7.7 a and 7.7 b), and annual precipitation might not affect TP concentration.

When the selected category of OP_CON was changed from Pre to Post under wet annual precipitation conditions (scenario 5), each TP concentration corresponding to the same CFP decreased significantly (Figure 7.8 a). In the simulation of scenario 6, changing OP_CON from Pre to Post cause small decreases of TP concentration in the low percentiles and small increase of TP concentration in the high percentile (Figure 7.8 b). It may be concluded that LBRCP affected TP concentration on the stream during the wet year.

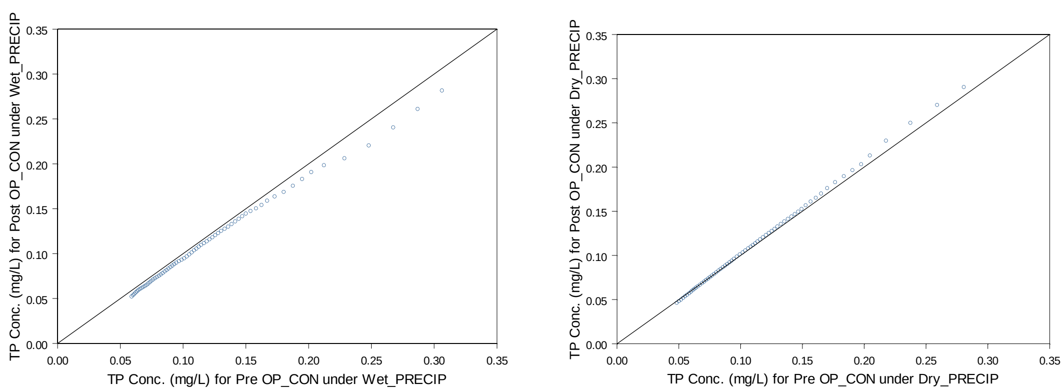
TP concentration time series by R-D method

Because the TP concentration violated the water quality criterion (0.05 mg/L) on many days, the timing and duration of TP concentrations equal to or below water quality criterion are evaluated. When the TP concentration from the BN outputs under Scenario 1 was connected with the rank time series of TP concentrations of the BN database in 97-98 water year, the magnitude of TP concentration for Post OP_CON was not similar to those for Pre OP_CON in the low concentration range, but the timing and duration below



a) b)

Figure 7.7 | Cumulative failure plots for Scenario 4 under D (Dry) PRECIP and W (Wet) PRECIP: a) Flow cumulative failure plot b) TP load cumulative failure plot (-- : W PRECIP, --:D PRECIP).



a) b)

Figure 7.8 | Q-Q plot for Scenarios 5 and 6 comparing the same percentile's TP concentrations for Pre and Post OP_CON under a) W (Wet) PRECIP b) under D (Dry) PRECIP.

or equal to 0.05 mg/L for Post OP_CON was very similar to those for Pre OP_CON (Figure 7.9). The timing and duration for Post OP_CON differed from those for Pre OP_CON as did the magnitude when the TP concentrations from BN outputs under Scenario 5 (under wet annual precipitation condition) were connected with rank time series (Figure 7.10). For example, the TP concentrations were equal to or below 0.05

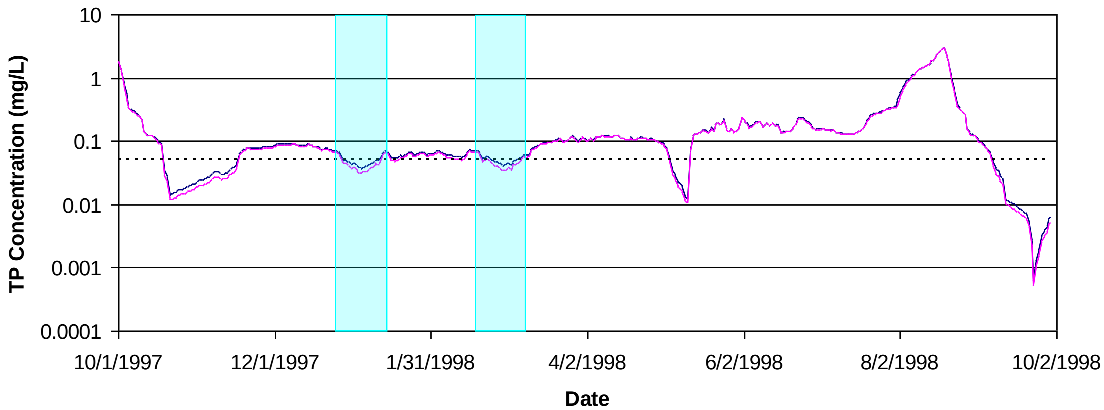


Figure 7.9 | TP concentration Time series from R-D method for scenario 1 (-- : Post OP_CON, --:Pre OP_CON)

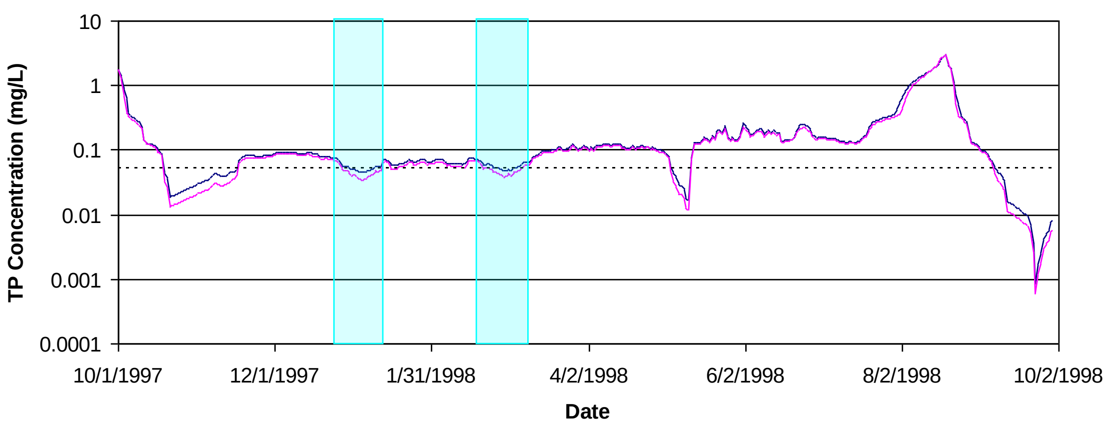


Figure 7.10 | TP concentration Time series from R-D method for scenario 5 (-- : Post OP_CON, -- : Pre OP_CON under wet annual precipitation)

mg/L for 16 days for Post OP_CON but for 9 days for Pre OP_CON from December, 1997 to January, 1998 (Figure 7.10).

The TP concentrations were equal to or below 0.05 mg/L for 12 days for Post OP_CON but for 3 days for Pre OP_CON from February, 1998 to March, 1998 (Figure 7.10). The date starting equal to or below 0.05 mg/L for Post OP_CON is 4 day earlier than the date for Pre OP_CON under wet precipitation condition. The time series of TP concentration for Post OP_CON were very similar to those for Pre OP_CON in all TP concentration range under the dry annual precipitation condition (Figure 7.11). This issue discussed below. It may be concluded that the conservation practice affected the timing and duration for the violation of the water quality standard under the wet year.

DISCUSSION

In the simulation of Scenario 6, changing OP_CON from Pre to Post caused small decreases of TP concentration in the low percentiles and small increases of TP

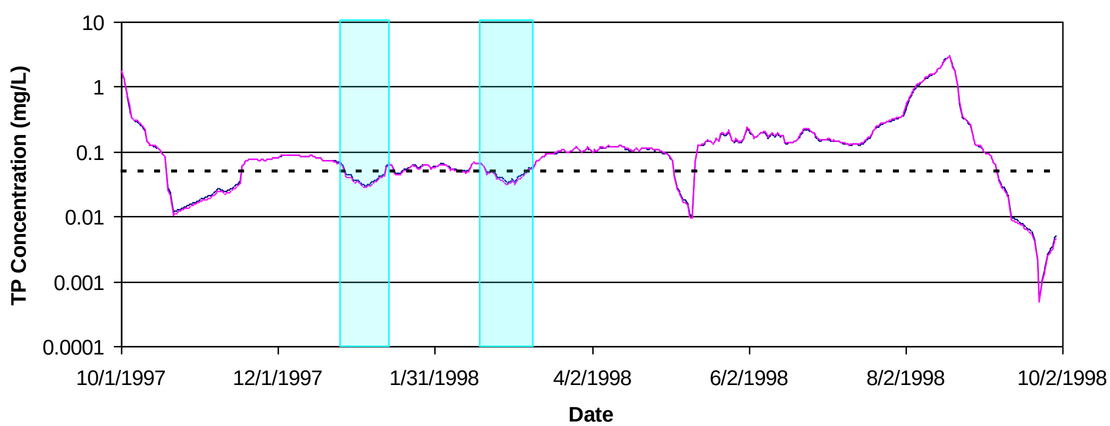


Figure 7.11 | TP concentration Time series from R-D method for scenario 6 (-- : Post OP_CON, -- : Pre OP_CON under dry annual precipitation).

concentration in the high percentile. It is difficult to interpret this result because two opposite direction of TP concentration change simultaneously occurred, but the gap between Q-Q plots and the line for dry annual precipitation condition in Figure 7.8 b seems ignorable comparing to Figure 7.8 a for wet annual precipitation condition. According to the estimated TP concentration time series for the dry year (Figure 7.11), the time series for Post OP_CON were very similar to those for Pre OP_CON. This may be evidence that the effect of conservation practices on TP concentrations was ignorable for the dry year. Because the time series came from same data distributions as those in Figure 7.8 b), it may be concluded that the gaps between Q-Q plots and the line for dry annual precipitation condition in Figure 7.8 b were ignorable.

The hypothesis test such as paired sample t-test (parametric) or Wilcoxon paired sample test (non-parametric) is more powerful than a graphical method such as Q-Q plot or cumulative failure plot to evaluate the effect of conservation practice on water quality. For example, in paired sample t-test between TP concentration for Pre OP_CON and for Post OP_CON corresponding to same percentile, we can say that the null hypothesis, 'mean of the differences between TP concentrations corresponding to same percentile under two different scenarios is 0' is rejected or not rejected at a given confidence level. There are some restrictions in this hypothesis test. For a paired sample t-test, the normal distribution assumption is used for the population of differences. For the Wilcoxon paired sample test, the distribution of differences must be symmetrical about the median (Zar, 1999). Because the distribution of the differences of TP concentrations from BN did not satisfy any distributional assumption, graphical methods were used. However, the high resolution BN output was more powerful than the conventional BN output because this allowed the evaluation of the conservation practices or exogenous variables using TP concentrations corresponding to the entire percentile range. In a conventional BN, we may say only that the probability of a category of a variable increased or decreased when changing the scenario. For example, we may say that the probability of water quality criterion violation (> 0.05 mg/L) decreased from 66.5% to 63.1%, changing OP_CON from Pre to Post in conventional Little Bear River BN (Chapter 5), but we may say all differences between TP concentrations for Pre and Post corresponding to the same percentile was very small from the BN using high frequency data (high resolution Bayesian Network). The high resolution BN may also support the R-D method to construct TP time series output.

SUMMARY AND CONCLUSIONS

The high resolution BN simulated the effects of the LBRCP and exogenous variables on the TP concentration at the mouth of the Little Bear River. High resolution BN provided 21 output categories for each of TP concentration, flow and TP load. This type of result was clearer than the result from conventional BN to evaluate the effect of the conservation practices or exogenous variables.

According to the Q-Q plot, the conservation practices (LBRCP) had only a small effect on the TP concentration when the all data in data base for both dry and wet years were used. However, the LBRCP decreased the TP concentration significantly in a wet year.

There were three exogenous variables, agricultural landuse areas, point source loads and annual precipitation. Increased agricultural land areas and point source loads caused higher TP concentrations but the annual precipitation increases caused only small increases of TP concentrations because TP loads and flows increased simultaneously.

The concentration's probabilities from high resolution BN were linked to rank time series of TP concentration by the R-D method. The LBRCP allowed longer duration and earlier starting of TP concentration below 0.05 mg/L under wet annual precipitation conditions while any noticeable effect of the LBRCP on TP concentration was not observed in time series for a dry year.

REFERENCES

Borsuk, M. E., Stow, C. A. & Reckhow, K.H. 2003 Integrated approach to total maximum daily load development for Neuse River Estuary using Bayesian Probability Network Model (Neu-BERN). *J. Planning and Manage.* **129** (4), 271-282.

Castillo, E., Gutiérrey, J. M. & Hadi, A. S. 1997 *Expert Szstems and Probabilistic Network Models.* Springer-Verlag. New York, NY.

Chess, D. 2000 Utah Watershed Review Vol 8, Utah Division of Water Quality.

EMRG (Environmental Management Research Group at Utah Water Research Laboratory) 2007 Database of water quality conservation activities in the Little Bear River Watershed.

EMRG (Environmental Management Research Group at Utah Water Research Laboratory) 2004 Little Bear River Watershed boundary shapfile.

Gilchrist, W.G. 2000 *Statistical Modeling with Quantile Functions*. Champ and Hall/CRC. Boca Raton, FL.

Jensen, F.V. 1996 *An Introduction to Bayesian Networks*, Springer-Verlag, New York, NY.

Kaplan, E. L., & Meier, P. 1958 Nonparametric estimation from incomplete observation. *J. American Statistical Assoc.* **53** (282), 457-481.

Marcot, B. G., Holthausen, R. S., Raphael, M.G. Rowland, M. M. & Wisdiom, M. J. 2001 Using Bayesian Belief networks to evaluate fish and wildlife population viability under land management alternatives from an environmental impact statement. *Forest Ecology and Manage.*, **153**, 29-42.

Norsys Software Corp. 1997 Netica Application User's Guide. Norsys Software Corporation, Vancouver, BC, Canada.

NCDENR (North Carolina Department of Environment and Natural Resources) 2008 Neuse River Basin - Nutrient Sensitive Water Management Strategy Agricultural Nitrogen Reduction Strategy, http://www.enr.state.nc.us/dswc/pages/intitiative.html, (Mar. 21, 2008)

Reckhow, K.H. 1999 Water quality prediction and probability network models, *Canadian J Fishery and Aquatic Sci.* **56**, 1150-1158.

Sheskin, D.J. 2004 The Single-Factor Between-Subject Analysis of Variance. In *Parametric and Nonparametric Statistical Procedures* (3rd Ed.), Chapman and Hall/CRC, Boca Raton, Florida, pp.667-756.

USEPA 2006 Little Bear River Project, http;//www.epa.gov/region8/water/nps/utlittlebear.pdf, (Aug.20, 2006)

USEPA 2004 Better Assessment Science Integrating Point and Non Point Sources (BASINS). http://www.epa.gov/waterscience/BASINS/, (Nov. 8, 2004).

USGS 2004 National Hydrograph Dataset. http//nhd.usgs.gov/data.html. (Nov.1, 2004)

Utah DNR (Utah Department of Natural Resources) 2004 Land Use Program, http://www.agr.state.ut.us/agrc/html/data/layers1.html, (Nov. 2, 2004)

Utah DEQ (Utah Department of Environmental Quality) 2000 Little Bear River Watershed TMDL, http://www.waterqulaity.utah.gov/TMDL/index.htm#addinfo.

Varis, O. and S.F. Jussila 2002 Water resources development in the Lower Senegal River Basin: Conflicting Interests, Environmental Concerns and policy options, *International J. Water Resources Development*, **18**, 245-260.

Varis, O. 1998 A belief network approach to optimization and parameter estimation: application to resource and environmental management. *Artificial Intelligence*, **101**, 135-163.

Western Regional Climate Center (WRCC) 2006 Logan Radio KVNU UT, Monthly Total Precipatiatation, http://www.wrcc.dri.edu/cgi-bin/cliMAIN.pl?utlkvn. (Feb 11, 2006).

Zar, J.H. 1999 *Biostatistical Analysis* (4th Ed.), Prentice Hall, Upper Saddle River, NJ.

CHAPTER 8

SUMMARY AND CONCLUSIONS

Alternative statistical methods for censored data are used to evaluate the water quality at the Little Bear River. The Rank-Data distribution method was developed to fill the data gaps and supported the Bayesian Networks and Total Maximum Daily Load (TMDL) processes.

Characteristics of Little Bear River Watershed

The two head waters, East Fork and South Fork of the Little Bear River, have good water quality according to summary statistics while the Little Bear River below the confluence of these waters was impaired with TP.

Location 4905000 (Above Cutler Reservoir) had significant seasonality in all parameters. Other locations did not show a significant seasonality in TP or pH while there was significant seasonality in other parameters.

TP concentration had strong linear correlation to turbidity at each location. This correlation shows that turbidity may be an indicator of TP concentration in the Little Bear River watershed. There were significant correlations between water quality of upstream and down but the turbidity did not have any significant correlation between above and below Hyrum Reservoir.

Trend analysis showed a significant downward trend of TP and DP concentration after starting LBRCP, but did not show any trends of TP concentration before LBRCP. DP reduction was faster than PP reduction at location 4905000 (Above Cutler Reservoir).

Filling data gaps by a Rank-Data distribution method (R-D METHOD)

The R-D method consists of three steps: 1) creation of estimated distribution based on the distribution of observations, 2) estimation of time series ranking of predictions and 3) assignment of predictions to each date.

The first step creates CF plot from observations. A large number of predictions may be reconstructed based on this CF plot. The second step calculates the CFP time series of predictions based on the observed CFP time series by interpolation. The CFP time series may be improved by adding estimated annual peak CFPs before interpolation. The annual peak CFP is determined by optimizing the extended CFP slopes. The third step assigns predictions to simulation dates by matching the rank of prediction within an estimated distribution to the rank of optimized CFP time series.

The estimated distribution from CF plot of observations was similar to the distribution of original data. Optimizing the CFP time series by calibrating extended CFP slopes enhanced the agreement of time series of predictions with time series of original data.

The estimated time series by the R-D method were closer to the original time series than those estimated by simple interpolation, and the R-D method was more powerful for the data set collected with a longer sampling block.

The R-D method may be used to reduce the sampling frequency keeping the same error and reducing the measurement cost.

Bayesian network to evaluate effects of the Little Bear River water quality conservation project

In order to evaluate the effect of conservation practices and exogenous variables on the TP load and TP concentration, BN1 (Above Hyrum) and BN2 (Below Hyrum) were constructed. Each BN used a different database. Some missing value of TP concentration and flow were filled with values estimated by regression between upstream and down stream data or by regression between two different variables.

BN simulations showed that conservation practices in the Little Bear River reduced subwatershed TP load above and below Hyrum Reservoir noticeably but the reductions were not large enough to reduce TP concentration into the receiving reservoirs noticeably, due to dilution of the effect by other factors. BNs suggested that the conservation practice have been working to reduce TP loads but more implementations of conservation practices are required.

There were three exogenous variables: agricultural landuse area, point source load and annual precipitation. Increased agricultural land area caused noticeably higher subwatershed TP load above and below Hyrum Reservoir significantly but not higher TP load and concentration into the receiving reservoirs, due to dilution of the effect by other factors. However, increased point source load caused significantly higher TP loads and concentrations into the Hyrum and Cutler reservoirs.

Increased annual precipitation caused a noticeably higher subwatershed TP load above and below Hyrum Reservoirs. These load increases were large enough to significantly increase TP loads into the Hyrum and Cutler reservoirs, but not TP concentration because more annual precipitation caused more flow and more TP load simultaneously.

The effects of conservation practices in the Little Bear River on the subwatershed TP load, TP load into Hyrum and Cutler Reservoirs and TP concentration into Hyrum and Cutler Reservoirs were larger for wet annual precipitation conditions than those for dry annual precipitation conditions.

It may be concluded that the TP concentration decreases since 1990 have been influenced by LBRCP (Little Bear River Conservation Project) only under wet annual precipitation conditions.

Total Maximum Daily Load (TMDL)
for Total Phosphorus
at the Mouth of the Little Bear River

The R-D method was used to fill in data gaps of flows and TP loads. The time series of cumulative failure probabilities (Rank time series) and cumulative failure plot (data distribution) of extended observations are required for this method. The estimation of flows and TP load was enhanced by the regression between extended observations and predictions from the R-D method.

Daily TP loads and daily mean flows from the R-D method were used for TMDL calculation at the mouth of the Little Bear River. The TMDLs and historical TP loads are calculated by these daily flows and TP loads, instead of the mean or median from low frequency data.

The reduction percentages were calculated for four different categories, low flow (July to February) and high flow (March to June) for a wet year (97-98 water year) and those for a dry year (02-03 water year) by two different approaches, the frequency targeted approach and the total mass targeted approach.

Higher reduction percentages are required for wet year than that for a dry year according to both approaches. The reduction percentage from the frequency targeted approach is higher than that from the total mass targeted approach because the large MOS was applied to reduce the frequency of water quality violation to 10 %.

When the 0.2σ of TP load are used as MOS, more practical reduction percentages and reduced concentration were obtained using the total mass targeted approach.

High Resolution Bayesian Network to evaluate effects of water quality conservation practices

The high resolution BN simulated the effects of the LBRCP and exogenous variables on the TP concentration at the mouth of the Little Bear River. High resolution BN provided 21 output categories for each of TP concentration, flow and TP load. This type of result was clearer than the result from conventional BN to evaluate the effect of the conservation practices or exogenous variables.

According to the Q-Q plot, the conservation practices (LBRCP) had only a small effect on the TP concentration when the all data in data base for both dry and wet years were used. However, the LBRCP decreased the TP concentration significantly in a wet year.

There were three exogenous variables, agricultural landuse areas, point source loads and annual precipitation. Increased agricultural land areas and point source loads caused higher TP concentrations but the annual precipitation increases caused only small increases of TP concentrations because TP loads and flows increased simultaneously.

The concentration's probabilities from high resolution BN were linked to rank time series of TP concentration by the R-D method. The LBRCP allowed longer duration and earlier starting of TP concentration below 0.05 mg/L under wet annual precipitation conditions while any noticeable effect of the LBRCP on TP concentration was not observed in time series for a dry year.

CHAPTER 9

ENGINEERING SIGNIFICANCE

In this research, methods to handle censored data are introduced. It is common to find left censored data (below detection limit) in the field of Environmental Engineering. Half of detection limit method has been used to substitute for censored values frequently. However, in this research, Maximum Likelihood Estimation (MLE), the Kalan-Meier Method, the Kruskal-Wallis method, Kendall's tau and modified seasonal Kendall trend were used as alternatives for censored data. This is important for evaluating historical or current water quality in a river where the number of censored data is not ignorable.

This research represents advances in the field of Environmental Engineering by providing unique approach to better estimating missing data in water quality monitoring with a limited data collection budget. The Rank-Data distribution method (R-D method) developed here estimates the daily frequency flow, TP load and concentration values in a river using low frequency observations and predictions of flow, TP load and concentration respectively at the same location. Because predictions from the R-D method are closer to the true values than the predictions by simple interpolation between observations, the R-D method may support a reduced sampling frequency and lower cost while keeping the same degree of uncertainty. The R-D method also provides support for Bayesian Network modeling efforts where high frequency probabilistic output with many categories is required, and Total Maximum Daily Load (TMDL) from daily predicted loads instead of annual statistical representatives.

It was demonstrated in this research that water quality conservation practices affect water quality. The construction of Bayesian Network represents an advance in the field of Environmental Engineering by providing a way to produce probabilistic results of water quality under different scenarios of pollutant management at a high resolution. In Bayesian Networks, a non-numeric variable (conservation practice option) is connected with numerical variables (flows, nutrient loads and concentrations). In this research, the predictions from the R-D method were added to the data base to obtain high frequency output with many categories. The data distribution from this high frequency output is an advance for interpreting the probabilistic result of water quality and to provide time series of water quality for evaluating the effects of conservation practices on the frequency, duration and timing of water quality standard violation.

An additional advance in Environmental Engineering resulting from the R-D method is TMDL process based on high frequency load predictions. This TMDL process provides more scientific margin of safety (MOS), allocation and reduction based on the high frequency predictions. This is important for planning pollutant load reduction to avoid wasting money.

CHAPTER 10

RECOMMENDATIONS FOR FUTURE RESEARCH

Future research efforts should:

1. Evaluate the effects of conservation practices on the water quality (Total Phosphorus concentration) inside of Hyrum Reservoir using the daily time series predictions of Total phosphorus (TP) concentration and flow from high frequency Bayesian Network at the reservoir's inlet.

   Because number of TP concentration observations is small for inside of Hyrum Reservoir, it is not easy to evaluate any enhancement of water quality inside the reservoir after starting conservation practices. The daily frequency TP load and concentration outputs from the R-D method may be helpful to simulate the water quality using a reservoir model.

2. Develop a computer module to simulate the water quality connecting estimation of missing values, Bayesian Network simulation and producing time series of TP concentration at the reservoir's inlet and in the reservoir.

   In this dissertation, developing the new approach to find missing values (R-D method) and applying the R-D method to BN and TMDL were the focus. Developing user friendly computer software to connect database, R-D method, a BN and a TMDL process is recommended.

3. Find or develop a paired sample hypothesis test releasing the assumption of a symmetric distribution or normal distribution.

In this dissertation, the graphical method was used to evaluate the TP concentration data distribution change from changing the TP load management scenario but a hypothesis test was not used because of the violation of data distribution assumption.

4. Enhance the accuracy of R-D method to reduce the error from true values.

   In this dissertation, predictions from regression between upstream and down stream values or between two different variables were added to the database (Chapter 5). Homogeneity, which means one slope is applied for all groups, was assumed for those regressions. If the homogeneity assumption is rejected, it is recommended to use different slopes for different groups for better regression (Sheskin, 2004). The pairs of two variables may be grouped by the trend pattern of those variables such as upward trend of both variables, upward trend of one variable with downward trend of the other variable and downward trend of both variables.

   In this dissertation, peak Cumulative Failure Probabilities (CFPs) were added to CFPs of observation to improve rank time series (CFP time series). The flows and TP loads estimated by R-D method were calibrated using regression between observations and predictions. If the methods for better estimation of peak CFP or better calibration of the values from the R-D method are developed, the better predictions may be expected.

5. Study how to handle unrealistically high predicted TP concentration.

   The high frequency TP loads are estimated from low frequency observations or predictions using the R-D method. When there are some unrealistically high TP concentrations in low frequency database, it is possible to obtain unrealistically high predictions of TP concentrations from R-D method. The load from maximum TP concentration observation (6 mg/L) and the estimated flow (119 cfs) on the same date was 3849 lb/day at the mouth of the Little Bear River. Because the upper 95 % confidence interval of TP load for 119 cfs was 100 lb/day on the regression line (flow vs. TP load), and the largest observed load was 817 lb/day corresponding to the flow 202 cfs, the load was obvious outlier. It is failed to correct this by replacing the unrealistic maximum TP concentration (6 mg/L) with half of that value (3 mg/L) for realistic prediction of maximum TP concentration from R-D method failed. After we obtained high frequency TP loads and flows by the R-D method from database using 3mg/L as maximum TP concentration and those flow and TP concentration predictions from the R-D method were calibrated (Chapter 5), unrealistic maximum TP concentration (6.05 mg/L) was still produced by the R-D method. When the cumulative failure plot (TP concentration distribution) was constructed from BN simulation outputs of TP _5000 to obtain 365 TP concentrations, the same procedure was needed, where 6.05 mg/L as TP was replaced by 3.0 mg/L as maximum TP concentration. These unrealistic values do not have a significant negative effect in the Bayesian Network or TMDL because of their short duration and small frequency, however, this issue merits further study.

6. Use more detail categories of hydrologic conditions

In this dissertation, four hydrologic categories (high flow season and low flow season in dry years and those in wet years) were used. Because various trends of temperature and precipitation such as early or late snow melting and small or large amount of spring precipitation may cause variable hydrologic conditions (Neitsch et al., 2003), it is recommended to use additional categories associated with hydrologic conditions.

Reference

Neitsch, S.L.; Arnold, J. G.; Kiniry, J. R.; Williams, J. R. (2003). SWAT *Technical Documentation Version 2003;* Grassland, Soil and Water Research Laboratory, Agricultural Research Service, USDA: Temple, Texas

Sheskin, D.J. (2004). The Single-Factor Between-Subject Analysis of Variance, in *Parametric and Nonparametric Statistical Procedures* (3rd Ed.), pp.667-756. Chapman and Hall/CRC: Boca Raton, Florida.

APPENDIX

Appendix A. The table of summary statistics for water quality

1) Dissolved Oxygen (mg/L)

	# of Observation	# of censored Data	Range of Dates	Mean	Median	Standard Deviation	Geometric Mean	Min.	Max.
4905740	68	0	1990-2004	9.882	10.15	1.412	9.778	6.4	13.45
4905750	84	0	1990-2004	10.13	9.96	1.262	10.05	7.2	14.19
4905670	79	0	1990-2004	10.01	9.9	1.577	9.9	7.5	16.58
4905000	154	0	1990-2004	8.97	9	1.838	8.767	3.8	13.65

2) Flow (cfs)

	# of Observation	# of censored Data	Range of Dates	Mean	Median	Standard Deviation	Geometric Mean	Min.	Max.
4905740	66	1	1990-2004	52.36	28.2	63.3	22.72	LD	304
4905750	86	3	1990-2004	41.61	11.1	81.66	12.2	LD	371
4905670	53	0	1990-2003	88.49	30	145	43.08	6	657
4905000	112	0	1990-2004	80.02	60	76.44	53.1	2	500

3) pH, field

	# of Observation	# of censored Data	Range of Dates	Mean	Median	Standard Deviation	Geometric Mean	Min.	Max.
4905740	69	0	1990-2004	8.149	8.3	0.4398	8.137	6.9	9
4905750	85	0	1990-2004	8.19	8.3	0.4415	8.177	7	9.54
4905670	80	0	1990-2004	8.062	8.15	0.3747	8.053	6.7	8.51
4905000	156	0	1990-2004	8.158	8.17	0.3056	8.153	7.4	10.5

5) Specific conductance (umho/cm)

	# of Observation	# of censored Data	Range of Dates	Mean	Median	Standard Deviation	Geometric Mean	Min.	Max.
4905740	68	0	1990-2004	339.2	350.5	109.3	321.5	80	652
4905750	83	0	1990-2004	394.3	396	63.77	388.7	163	619
4905670	80	0	1990-2004	417.6	431.4	93.32	406.4	203	590
4905000	155	0	1990-2004	550	564	127.8	537.1	336	1454

6) Temperature (deg C)

	# of Observation	# of censored Data	Range of Dates	Mean	Median	Standard Deviation	Geometric Mean	Min.	Max.
4905740	69	0	1990-2004	9.436	8.8	5.855	7.569	0.9	26
4905750	85	0	1990-2004	9.18	9.74	4.986	7.127	0.18	23.3
4905670	80	0	1990-2004	10.51	11.17	5.288	8.193	0.01	20.6
4905000	156	0	1990-2004	11.15	10.67	6.423	8.405	0.1	25.45

7) Turbidity, laboratory (NTU)

	# of Observation	# of censored Data	Range of Dates	Mean	Median	Standard Deviation	Geometric Mean	Min.	Max.
4905740	28	0	1993-2004	10.67	3.645	19.12	4.035	0.107	95.4
4905750	26	0	1998-2004	21.8	1.85	63.96	2.985	0.233	325
4905670	29	0	1992-2004	15.04	4.5	31.29	4.9	0.641	125
4905000	132	0	1990-2004	13.4	11.85	9.23	10.5	1.9	47.5

8) TP Concentration

4905000

	# of Observation	# of censored Data	Range of Dates	Mean	Median	Standard Deviation	90 percentile	Min.	Max.
Ignoring DL	142	0	1990 - 2004	0.1215	0.1	0.1686	0.176	0.02	1.88
Half of DL	145	3	1990 - 2004	0.1192	0.1	0.1676	0.176	0.01	1.88
*MLE-Lognormal	145	3	1990 - 2004	0.1135	0.0898	0.0878	0.216	BDL	1.88
KM	145	3	1990 - 2004	0.1194	0.1	0.1673 (0.0139)	0.182	BDL	1.88

* Distribution assumption violation. ( ) is standard error of mean

4905670

	# of Observation	# of censored Data	Range of Dates	Mean	Median	Standard Deviation	90 Percentile	Min.	Max.
Ignoring DL	70	0	1990 -2004	0.06443	0.0495	0.04704	0.1096	0.02	0.273
Half of DL	79	9	1990 -2004	0.05823	0.0450	0.04754	0.1066	0.01	0.273
MLE-Lognormal	79	9	1990 -2004	0.0587	0.04573	0.04717	0.113	BDL	0.273
KM	79	9	1990 -2004	0.0594	0.0450	0.0465 (0.00523)	0.109	BDL	0.273

4905740

	# of Observation	# of censored Data	Range of Dates	Mean	Median	Standard Deviation	90 Percentile	Min.	Max.
Ignoring DL	42	0	1990 -2004	0.04771	0.0335	0.03447	0.088	0.02	0.16
Half of DL	68	26	1990 -2004	0.03329	0.0250	0.03268	0.068	0.01	0.16
MLE-Lognormal	68	26	1990 -2004	0.03398	0.02474	0.03197	0.069	BDL	0.16
KM	68	26	1990 -2004	0.03711	0.0250	0.03028	0.069	BDL	0.16

4905750

	# of Observation	# of censored Data	Range of Dates	Mean	Median	Standard Deviation	90 Percentile	Min.	Max.
Ignoring DL	22	0	1990 -2004	0.0503	0.0320	0.0767	0.052	0.021	0.39
Half of DL	75	53	1990 -2004	0.02182	0.0100	0.04485	0.037	0.01	0.39
MLE-Lognormal	75	53	1990-2004	0.01898	0.01112	0.02622	0.042	BDL	0.39
KM	75	53	1990 -2004	0.0296	BDL	0.04374	0.041	BDL	0.39

Appendix B.

Correlation

Location	Method	TP	DP	Flow	Turbidity
4905740 vs 4905670	Pearson's r	<0.001	*0.084*	<0.001	<0.001
	Nonparametric	<0.001	*0.022*	<0.001	<0.001
4905750 vs 4905670	Pearson's r	*0.657/ <0.001*	0.001	<0.001	*0.668/ <0.001*
	Nonparametric	*0.007*	*0.006	<0.001	*0.002*
4905670 vs 4905000	Pearson's r	0.001	0.001	0.002	0.666
	Nonparametric	0.002	<0.001	<0.001	0.392

Outlinear :TP 0.39 mg/l, Turb 325 mg/l * Spearman's rho is used because of run time error for Kendall's tau-b test.

Location	Method	DO	pH	Temperature
4905740 vs 4905670	Pearson's r	0.705(<0.001)	0.711(<0.001)	0.925(<0.001)
	Nonparametric			
4905750 vs 4905670	Pearson's r	0.722(<0.001)	0.788(<0.001)	0.898(<0.001)
	Nonparametric			
4905670 vs 4905000	Pearson's r	0.590(0.001)	0.531(0.002)	0.918(<0.001)
	Nonparametric			

Appendix C. K-M estimation of flow distribution

a) Data set 1 (2 month sampling block)

Results for: VAL.MTW

Distribution Analysis: Flipped_Flow

Variable: Flipped_Flow

Censoring Information	Count
Uncensored value	60
Right censored value	3494

Censoring value: Flow_I = 0

Nonparametric Estimates

Characteristics of Variable

Mean(MTTF)	Standard Error	95.0% Normal CI Lower	95.0% Normal CI Upper
415.122	12.8630	389.910	440.333

Median = 450
IQR = 69 Q1 = 405 Q3 = 474

Kaplan-Meier Estimates

Time	Number at Risk	Number Failed	Survival Probability	Standard Error	95.0% Normal CI Lower	95.0% Normal CI Upper
64	3554	1	0.999719	0.0002813	0.999167	1.00000
68	59	1	0.982774	0.0168024	0.949842	1.00000
89	58	1	0.965830	0.0235549	0.919663	1.00000
130	57	1	0.948885	0.0285939	0.892842	1.00000
133	56	1	0.931941	0.0327209	0.867809	0.99607
252	55	1	0.914997	0.0362487	0.843951	0.98604
284	54	1	0.898052	0.0393389	0.820949	0.97516
298	53	1	0.881108	0.0420880	0.798617	0.96360
321	52	1	0.864164	0.0445591	0.776829	0.95150
368	51	2	0.830275	0.0488318	0.734566	0.92598
385	49	1	0.813330	0.0506899	0.713980	0.91268
389	48	1	0.796386	0.0523894	0.693705	0.89907
397	47	1	0.779442	0.0539453	0.673711	0.88517
398	46	1	0.762497	0.0553698	0.653974	0.87102
405	45	1	0.745553	0.0566728	0.634476	0.85663
409	44	1	0.728608	0.0578624	0.615200	0.84202
425	43	1	0.711664	0.0589456	0.596133	0.82720
429	42	2	0.677775	0.0608147	0.558581	0.79697
430	40	1	0.660831	0.0616097	0.540078	0.78158
431	39	1	0.643887	0.0623165	0.521748	0.76602
433	38	1	0.626942	0.0629381	0.503586	0.75030
434	37	1	0.609998	0.0634770	0.485585	0.73441
435	36	2	0.576109	0.0643148	0.450054	0.70216
440	34	1	0.559165	0.0646167	0.432518	0.68581

441	33	1	0.542220	0.0648422	0.415132	0.66931
442	32	1	0.525276	0.0649921	0.397894	0.65266
445	31	1	0.508332	0.0650669	0.380803	0.63586
450	30	1	0.491387	0.0650669	0.363858	0.61892
451	29	2	0.457498	0.0648422	0.330410	0.58459
455	27	1	0.440554	0.0646166	0.313908	0.56720
456	26	1	0.423610	0.0643147	0.297555	0.54966
458	25	1	0.406665	0.0639352	0.281354	0.53198
461	24	1	0.389721	0.0634769	0.265308	0.51413
465	23	1	0.372776	0.0629379	0.249420	0.49613
467	22	1	0.355832	0.0623163	0.233694	0.47797
469	21	2	0.321943	0.0608145	0.202749	0.44114
470	19	1	0.304999	0.0599278	0.187543	0.42246
471	18	1	0.288055	0.0589453	0.172524	0.40359
472	17	1	0.271110	0.0578621	0.157702	0.38452
474	16	2	0.237221	0.0553695	0.128699	0.34574
476	14	3	0.186388	0.0506894	0.087039	0.28574
477	11	2	0.152499	0.0467957	0.060782	0.24422
478	9	1	0.135555	0.0445584	0.048222	0.22289
481	8	1	0.118611	0.0420873	0.036121	0.20110
483	7	1	0.101666	0.0393381	0.024565	0.17877
484	6	1	0.084722	0.0362478	0.013678	0.15577
485	5	2	0.050833	0.0285927	0.000000	0.10687
486	3	1	0.033889	0.0235533	0.000000	0.08005
488	2	1	0.016944	0.0168002	0.000000	0.04987
490	1	1	0.000000	0.0000000	0.000000	0.00000

b) Data set 2 (1month sampling block)

Results for: VAL_MONTH.MTW

Distribution Analysis: Flipped_Flow

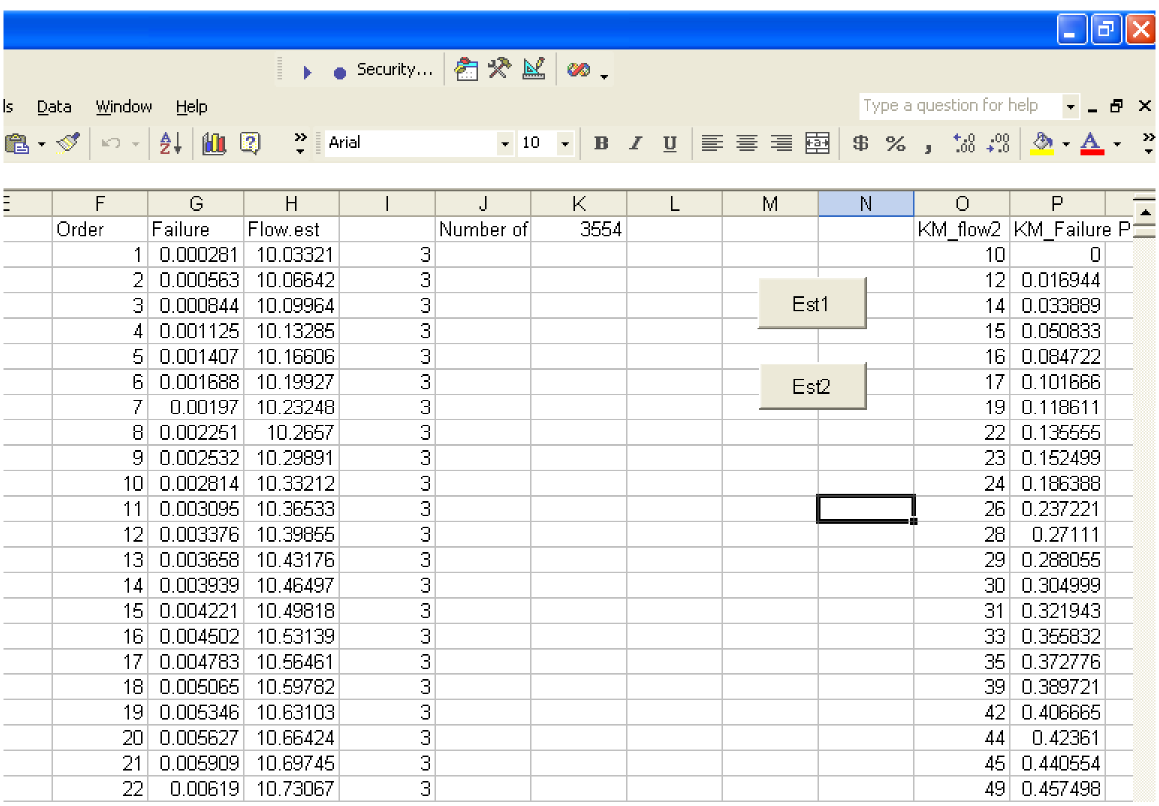
```
Variable: Flipped_Flow

Censoring Information  Count
Uncensored value         118
Right censored value    3436

Censoring value: Flow_I = 0


Nonparametric Estimates


Characteristics of Variable

            Standard   95.0% Normal CI
Mean(MTTF)     Error    Lower    Upper
   606.258   10.6370  585.410  627.107

Median = 650
IQR = 77  Q1 = 597  Q3 = 674


Kaplan-Meier Estimates

       Number
           at  Number     Survival   Standard   95.0% Normal CI
 Time    Risk  Failed  Probability      Error     Lower    Upper
```

15.0	3554	1	0.999719	0.0002813	0.999167	1.00000
202.0	117	1	0.991174	0.0085126	0.974490	1.00000
210.0	116	1	0.982629	0.0119834	0.959142	1.00000
264.0	115	1	0.974085	0.0146113	0.945447	1.00000
268.0	114	1	0.965540	0.0167967	0.932619	0.99846
275.0	113	1	0.956996	0.0186955	0.920353	0.99364
289.0	112	1	0.948451	0.0203879	0.908491	0.98841
311.0	111	1	0.939906	0.0219218	0.896940	0.98287
330.0	110	1	0.931362	0.0233284	0.885639	0.97708
333.0	109	1	0.922817	0.0246295	0.874544	0.97109
338.0	108	1	0.914273	0.0258412	0.863625	0.96492
354.0	107	1	0.905728	0.0269754	0.852857	0.95860
413.0	106	1	0.897183	0.0280415	0.842223	0.95214
452.0	105	1	0.888639	0.0290471	0.831707	0.94557
484.0	104	1	0.880094	0.0299983	0.821299	0.93889
498.0	103	1	0.871550	0.0309000	0.810987	0.93211
521.0	102	1	0.863005	0.0317565	0.800763	0.92525
523.0	101	1	0.854460	0.0325713	0.790622	0.91830
556.0	100	1	0.845916	0.0333475	0.780556	0.91128
564.0	99	1	0.837371	0.0340878	0.770560	0.90418
568.0	98	2	0.820282	0.0354694	0.750763	0.88980
576.0	96	1	0.811737	0.0361144	0.740954	0.88252
585.0	95	3	0.786104	0.0378853	0.711850	0.86036
587.0	92	1	0.777559	0.0384250	0.702247	0.85287
589.0	91	1	0.769014	0.0389412	0.692691	0.84534
592.0	90	1	0.760470	0.0394348	0.683179	0.83776
594.0	89	1	0.751925	0.0399067	0.673709	0.83014
597.0	88	1	0.743381	0.0403576	0.664281	0.82248
598.0	87	1	0.734836	0.0407882	0.654892	0.81478
605.0	86	2	0.717747	0.0415911	0.636230	0.79926
609.0	84	1	0.709202	0.0419645	0.626953	0.79145
618.0	83	1	0.700657	0.0423199	0.617712	0.78360
623.0	82	1	0.692113	0.0426577	0.608505	0.77572
625.0	81	1	0.683568	0.0429783	0.599332	0.76780
628.0	80	1	0.675024	0.0432821	0.590192	0.75986
629.0	79	2	0.657934	0.0438408	0.572008	0.74386
630.0	77	2	0.640845	0.0443361	0.553948	0.72774
631.0	75	1	0.632301	0.0445608	0.544963	0.71964
633.0	74	1	0.623756	0.0447703	0.536008	0.71150
634.0	73	1	0.615211	0.0449650	0.527082	0.70334
635.0	72	2	0.598122	0.0453106	0.509315	0.68693
640.0	70	3	0.572488	0.0457219	0.482875	0.66210
641.0	67	1	0.563944	0.0458309	0.474117	0.65377
642.0	66	1	0.555399	0.0459261	0.465386	0.64541
644.0	65	1	0.546855	0.0460076	0.456681	0.63703
645.0	64	2	0.529765	0.0461295	0.439353	0.62018
646.0	62	1	0.521221	0.0461700	0.430729	0.61171
648.0	61	1	0.512676	0.0461970	0.422132	0.60322
650.0	60	2	0.495587	0.0462105	0.405016	0.58616
651.0	58	2	0.478498	0.0461700	0.388006	0.56899
652.0	56	1	0.469953	0.0461294	0.379541	0.56037
655.0	55	1	0.461409	0.0460753	0.371103	0.55171
656.0	54	2	0.444319	0.0459260	0.354306	0.53433
657.0	52	1	0.435775	0.0458308	0.345948	0.52560
658.0	51	1	0.427230	0.0457218	0.337617	0.51684
659.0	50	3	0.401596	0.0453104	0.312790	0.49040
660.0	47	1	0.393052	0.0451449	0.304569	0.48153
661.0	46	1	0.384507	0.0449648	0.296378	0.47264
664.0	45	3	0.358873	0.0443359	0.271977	0.44577
665.0	42	2	0.341784	0.0438405	0.255858	0.42771
666.0	40	1	0.333240	0.0435692	0.247845	0.41863
667.0	39	1	0.324695	0.0432818	0.239864	0.40953

669.0	38	2	0.307606	0.0426573	0.223999	0.39121
670.0	36	1	0.299061	0.0423195	0.216116	0.38201
671.0	35	1	0.290517	0.0419641	0.208268	0.37276
672.0	34	2	0.273427	0.0411988	0.192679	0.35418
674.0	32	3	0.247794	0.0399062	0.169579	0.32601
676.0	29	3	0.222160	0.0384244	0.146849	0.29747
677.0	26	2	0.205070	0.0373204	0.131924	0.27822
678.0	24	4	0.170892	0.0347937	0.102698	0.23909
679.0	20	3	0.145258	0.0325705	0.081421	0.20910
680.0	17	1	0.136714	0.0317556	0.074474	0.19895
681.0	16	2	0.119624	0.0299973	0.060831	0.17842
682.0	14	1	0.111080	0.0290461	0.054151	0.16801
683.0	13	3	0.085446	0.0258399	0.034801	0.13609
684.0	10	1	0.076901	0.0246282	0.028631	0.12517
685.0	9	2	0.059812	0.0219202	0.016849	0.10278
686.0	7	3	0.034178	0.0167945	0.001262	0.06710
688.0	4	1	0.025634	0.0146087	0.000000	0.05427
690.0	3	2	0.008545	0.0085080	0.000000	0.02522
694.7	1	1	0.000000	0.0000000	0.000000	0.00000

Appendix D. MS Excel Macro to estimate flows from CFP curve

a) Work sheet



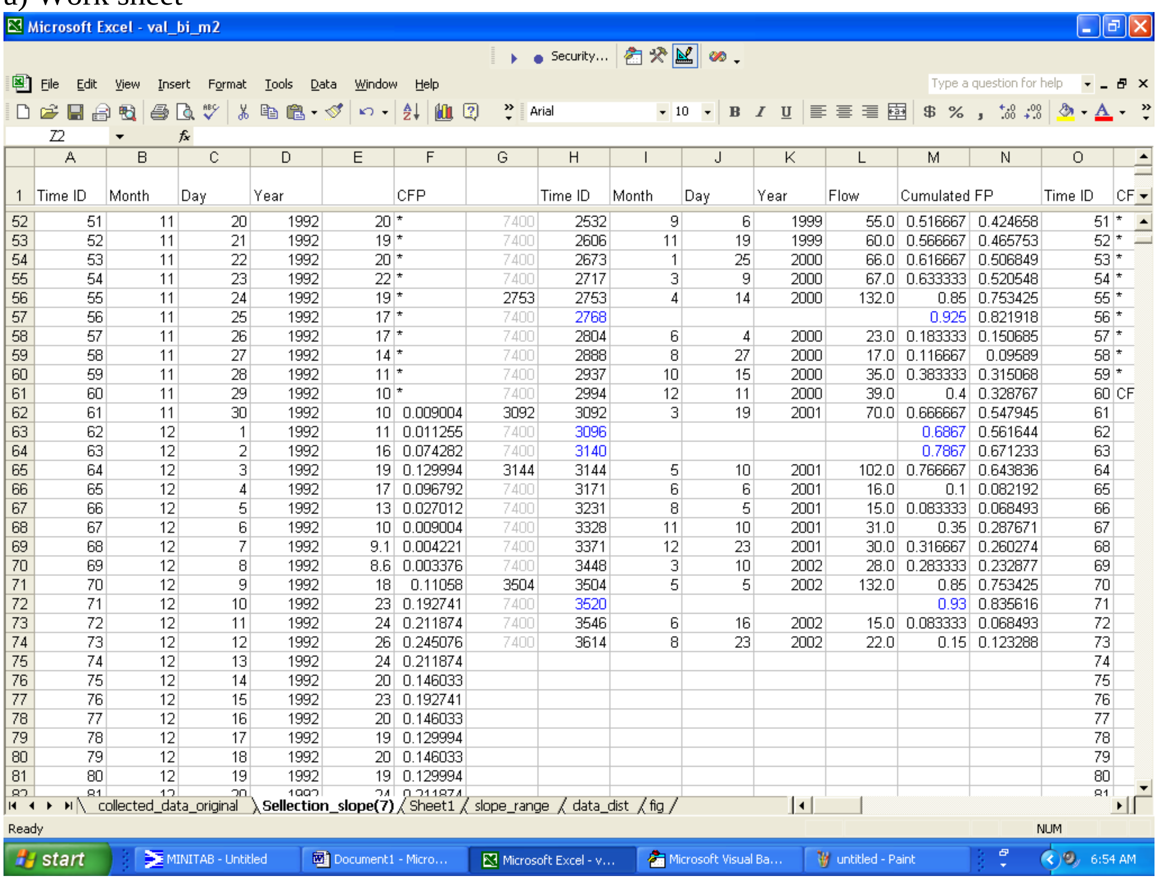
F	G	H	I	J	K	L	M	N	O	P
Order	Failure	Flow.est		Number of	3554				KM_flow2	KM_Failure
1	0.000281	10.03321	3						10	0
2	0.000563	10.06642	3						12	0.016944
3	0.000844	10.09964	3						14	0.033889
4	0.001125	10.13285	3						15	0.050833
5	0.001407	10.16606	3						16	0.084722
6	0.001688	10.19927	3						17	0.101666
7	0.00197	10.23248	3						19	0.118611
8	0.002251	10.2657	3						22	0.135555
9	0.002532	10.29891	3						23	0.152499
10	0.002814	10.33212	3						24	0.186388
11	0.003095	10.36533	3						26	0.237221
12	0.003376	10.39855	3						28	0.27111
13	0.003658	10.43176	3						29	0.288055
14	0.003939	10.46497	3						30	0.304999
15	0.004221	10.49818	3						31	0.321943
16	0.004502	10.53139	3						33	0.355832
17	0.004783	10.56461	3						35	0.372776
18	0.005065	10.59782	3						39	0.389721
19	0.005346	10.63103	3						42	0.406665
20	0.005627	10.66424	3						44	0.42361
21	0.005909	10.69745	3						45	0.440554
22	0.00619	10.73067	3						49	0.457498

b) Visual Basic Program (*Ittalic letters are explanations*)

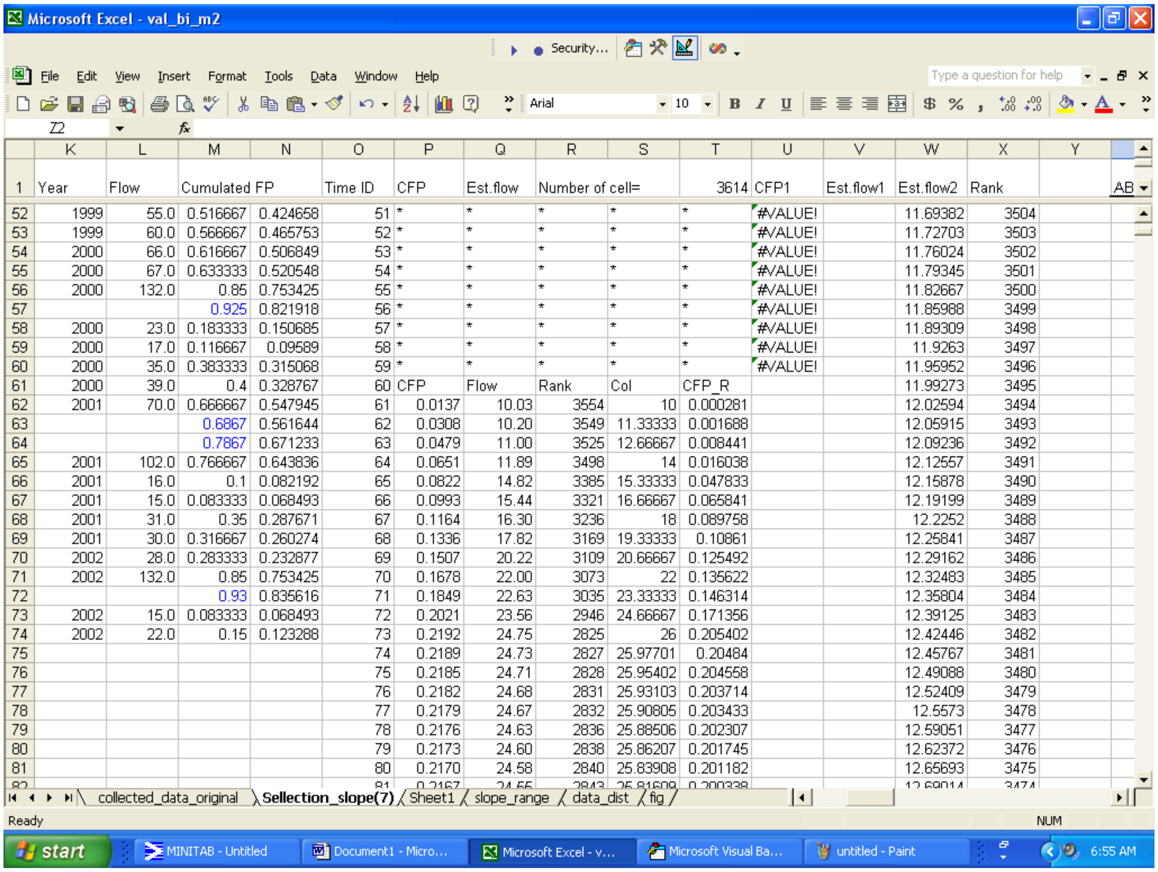
```
Private Sub CommandButton2_Click()

n = Range("k" + CStr(1)) 'number of predictions
r = 2
For f = 1 To n 'repeating for next routine 3554 times
x = Range("g" + CStr(f + 1))
If x > Range("p" + CStr(r)) Then ' comparing g to p. For example
r = r + 1  'g(2)>p(2). So r=2+1=3
End If
Range("h" + CStr(f + 1)) = (Range("o" + CStr(r)) - Range("o" + CStr(r - 1))) /
(Range("p" + CStr(r)) - Range("p" + CStr(r - 1))) * (x - Range("p" + CStr(r - 1))) +
Range("o" + CStr(r - 1)) ' estimating h(2), flow by interpolating (o(2),p(2)) and
(o(3),p(3)). Always, g(f+1) are between p(r-1) and p(r).
Range("i" + CStr(f + 1)) = r 'this means h(f+1) is estimated by interpolating (o(r), p(r))
and (o(r-1),p(r-1))
Next f

End Sub
```

Appendix E. MS Macro to estimate CFP time series and assign the estimated values

a) Work sheet


Microsoft Excel - val_bi_m2

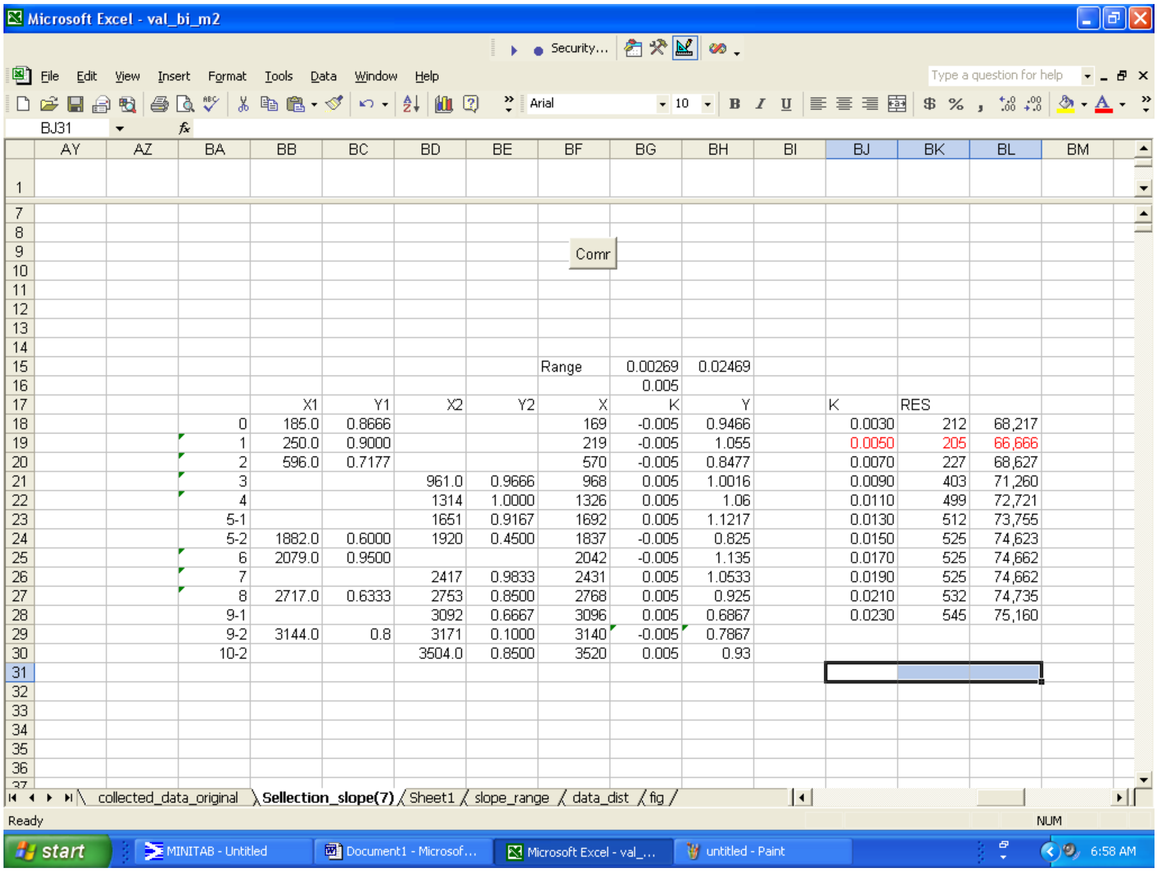
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	
1	Time ID	Month	Day	Year		CFP		Time ID	Month	Day	Year	Flow	Cumulated FP		Time ID	CF
52	51	11	20	1992	20	*	7400	2532	9	6	1999	55.0	0.516667	0.424658	51	*
53	52	11	21	1992	19	*	7400	2606	11	19	1999	60.0	0.566667	0.465753	52	*
54	53	11	22	1992	20	*	7400	2673	1	25	2000	66.0	0.616667	0.506849	53	*
55	54	11	23	1992	22	*	7400	2717	3	9	2000	67.0	0.633333	0.520548	54	*
56	55	11	24	1992	19	*	2753	2753	4	14	2000	132.0	0.85	0.753425	55	*
57	56	11	25	1992	17	*	7400	2768					0.925	0.821918	56	*
58	57	11	26	1992	17	*	7400	2804	6	4	2000	23.0	0.183333	0.150685	57	*
59	58	11	27	1992	14	*	7400	2888	8	27	2000	17.0	0.116667	0.09589	58	*
60	59	11	28	1992	11	*	7400	2937	10	15	2000	35.0	0.383333	0.315068	59	*
61	60	11	29	1992	10	*	7400	2994	12	11	2000	39.0	0.4	0.328767	60	CF
62	61	11	30	1992	10	0.009004	3092	3092	3	19	2001	70.0	0.666667	0.547945	61	
63	62	12	1	1992	11	0.011255	7400	3096					0.6867	0.561644	62	
64	63	12	2	1992	16	0.074282	7400	3140					0.7867	0.671233	63	
65	64	12	3	1992	19	0.129994	3144	3144	5	10	2001	102.0	0.766667	0.643836	64	
66	65	12	4	1992	17	0.096792	7400	3171	6	6	2001	16.0	0.1	0.082192	65	
67	66	12	5	1992	13	0.027012	7400	3231	8	5	2001	15.0	0.083333	0.068493	66	
68	67	12	6	1992	10	0.009004	7400	3328	11	10	2001	31.0	0.35	0.287671	67	
69	68	12	7	1992	9.1	0.004221	7400	3371	12	23	2001	30.0	0.316667	0.260274	68	
70	69	12	8	1992	8.6	0.003376	7400	3448	3	10	2002	28.0	0.283333	0.232877	69	
71	70	12	9	1992	18	0.11058	3504	3504	5	5	2002	132.0	0.85	0.753425	70	
72	71	12	10	1992	23	0.192741	7400	3520					0.93	0.835616	71	
73	72	12	11	1992	24	0.211874	7400	3546	6	16	2002	15.0	0.083333	0.068493	72	
74	73	12	12	1992	26	0.245076	7400	3614	8	23	2002	22.0	0.15	0.123288	73	
75	74	12	13	1992	24	0.211874									74	
76	75	12	14	1992	20	0.146033									75	
77	76	12	15	1992	23	0.192741									76	
78	77	12	16	1992	20	0.146033									77	
79	78	12	17	1992	19	0.129994									78	
80	79	12	18	1992	20	0.146033									79	
81	80	12	19	1992	19	0.129994									80	



Microsoft Excel - val_bi_m2

	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
1	Year	Flow	Cumulated FP		Time ID	CFP	Est.flow	Number of cell=		3614	CFP1	Est.flow1	Est.flow2	Rank	
52	1999	55.0	0.516667	0.424658	51	*	*	*	*	*	#VALUE!		11.69382	3504	
53	1999	60.0	0.566667	0.465753	52	*	*	*	*	*	#VALUE!		11.72703	3503	
54	2000	66.0	0.616667	0.506849	53	*	*	*	*	*	#VALUE!		11.76024	3502	
55	2000	67.0	0.633333	0.520548	54	*	*	*	*	*	#VALUE!		11.79345	3501	
56	2000	132.0	0.85	0.753425	55	*	*	*	*	*	#VALUE!		11.82667	3500	
57			0.925	0.821918	56	*	*	*	*	*	#VALUE!		11.85988	3499	
58	2000	23.0	0.183333	0.150685	57	*	*	*	*	*	#VALUE!		11.89309	3498	
59	2000	17.0	0.116667	0.09589	58	*	*	*	*	*	#VALUE!		11.9263	3497	
60	2000	35.0	0.383333	0.315068	59	*	*	*	*	*	#VALUE!		11.95952	3496	
61	2000	39.0	0.4	0.328767	60	CFP	Flow	Rank	Col	CFP_R			11.99273	3495	
62	2001	70.0	0.666667	0.547945	61	0.0137	10.03	3554	10	0.000281			12.02594	3494	
63			0.6867	0.561644	62	0.0308	10.20	3549	11.33333	0.001688			12.05915	3493	
64			0.7867	0.671233	63	0.0479	11.00	3525	12.66667	0.008441			12.09236	3492	
65	2001	102.0	0.766667	0.643836	64	0.0651	11.89	3498	14	0.016038			12.12557	3491	
66	2001	16.0	0.1	0.082192	65	0.0822	14.82	3385	15.33333	0.047833			12.15878	3490	
67	2001	15.0	0.083333	0.068493	66	0.0993	15.44	3321	16.66667	0.065841			12.19199	3489	
68	2001	31.0	0.35	0.287671	67	0.1164	16.30	3236	18	0.089758			12.2252	3488	
69	2001	30.0	0.316667	0.260274	68	0.1336	17.82	3169	19.33333	0.10861			12.25841	3487	
70	2002	28.0	0.283333	0.232877	69	0.1507	20.22	3109	20.66667	0.125492			12.29162	3486	
71	2002	132.0	0.85	0.753425	70	0.1678	22.00	3073	22	0.135622			12.32483	3485	
72			0.93	0.835616	71	0.1849	22.63	3035	23.33333	0.146314			12.35804	3484	
73	2002	15.0	0.083333	0.068493	72	0.2021	23.56	2946	24.66667	0.171356			12.39125	3483	
74	2002	22.0	0.15	0.123288	73	0.2192	24.75	2825	26	0.205402			12.42446	3482	
75					74	0.2189	24.73	2827	25.97701	0.20484			12.45767	3481	
76					75	0.2185	24.71	2828	25.95402	0.204558			12.49088	3480	
77					76	0.2182	24.68	2831	25.93103	0.203714			12.52409	3479	
78					77	0.2179	24.67	2832	25.90805	0.203433			12.5573	3478	
79					78	0.2176	24.63	2836	25.88506	0.202307			12.59051	3477	
80					79	0.2173	24.60	2838	25.86207	0.201745			12.62372	3476	
81					80	0.2170	24.58	2840	25.83908	0.201182			12.65693	3475	

collected_data_original | Sellection_slope(7) | Sheet1 | slope_range | data_dist | fig



	BA	BB	BC	BD	BE	BF	BG	BH	BI	BJ	BK	BL
15						Range	0.00269	0.02469				
16							0.005					
17		X1	Y1	X2	Y2	X	K	Y		K	RES	
18	0	185.0	0.8666			169	-0.005	0.9466		0.0030	212	68,217
19	1	250.0	0.9000			219	-0.005	1.055		0.0050	205	66,666
20	2	596.0	0.7177			570	-0.005	0.8477		0.0070	227	68,627
21	3			961.0	0.9666	968	0.005	1.0016		0.0090	403	71,260
22	4			1314	1.0000	1326	0.005	1.06		0.0110	499	72,721
23	5-1			1651	0.9167	1692	0.005	1.1217		0.0130	512	73,755
24	5-2	1882.0	0.6000	1920	0.4500	1837	-0.005	0.825		0.0150	525	74,623
25	6	2079.0	0.9500			2042	-0.005	1.135		0.0170	525	74,662
26	7			2417	0.9833	2431	0.005	1.0533		0.0190	525	74,662
27	8	2717.0	0.6333	2753	0.8500	2768	0.005	0.925		0.0210	532	74,735
28	9-1			3092	0.6667	3096	0.005	0.6867		0.0230	545	75,160
29	9-2	3144.0	0.8	3171	0.1000	3140	-0.005	0.7867				
30	10-2			3504.0	0.8500	3520	0.005	0.93				

b) Visual Basic Program *(Ittalic letters are explanations)*

Private Sub CommandButton5_Click()

a = 0
Dim d As Integer

For k = 0.003 To 0.025 Step 0.002 *'range of the extended slops*
a = a + 1
Range("BG16") = k *'setting the extended slope. After setting the slope, excel calculates the peak CFP using this extended slop*

n = Range("t" + CStr(1)) *'last data ID number*
r = 3
For f = 61 To *n ' repeating this routine from 61 ID number (November 30,1992) to 3614 ID number (August 23, 2002)*
x = Range("o" + CStr(f + 1))
If x > Range("h" + CStr(r)) Then
r = r + 1
End If

```
Range("p" + CStr(f + 1)) = (Range("n" + CStr(r)) - Range("n" + CStr(r - 1))) /
(Range("h" + CStr(r)) - Range("h" + CStr(r - 1))) * (x - Range("h" + CStr(r - 1))) +
Range("n" + CStr(r - 1))
Range("aa" + CStr(f + 1)) = 0
Next f 'This routine is estimating CFPs by interpolating CFPs of observations and
extended peaks

Stop

For i = 61 To n  ' This routine is finding flow in w matching r value (rank of CFPs in
CFP time series. That flow then come to q for the rank
  For j = 1 To 3554
If Range("r" + CStr(i + 1)) = Range("x" + CStr(j + 1)) Then
   Range("q" + CStr(i + 1)) = Range("w" + CStr(j + 1))
   GoTo 10
End If


   Next j
10 Next i

Stop

For x = 1 To 73 'this routine is calculating largest observation residuals

  d = Range("g" + CStr(x + 1))

 Range("AA" + CStr(d + 1)) = Abs(Range("e" + CStr(d + 1)) - Range("Q" + CStr(d + 1)))
Next x


 Range("BJ" + CStr(a + 17)) = k ' k is the extended slope
 Range("BK" + CStr(a + 17)) = Range("AA7294") 'AA7294 is largest observation
residuals sum
 Range("BL" + CStr(a + 17)) = Range("AG7294") 'AG7294 is residual sum

Stop
Next k



End Sub
```

Appendix. F. CPTs for BN1.

(a) CPT relating Conservation Program option (OP_CON), agricultural landuse area (LAND_AG1), point TP load (LOAD_P1) and annual precipitation (PRECIP) to subwatershed TP load (LOAD_SW1) above Hyrum.

OP_CON	**LAND_AG1**	**LOAD_P1**	**PRECIP**	**LOAD_SW1 (lb/d)**		
				L	**M**	**H**
				<3.53	**3.53 – 9.51**	**>= 9.51**
Pre	L	L	D	57.1%	14.3 %	28.6 %
Pre	L	L	W	33.3%	33.3 %	33.3 %
Pre	L	H	D	41.7 %	58.3 %	<0.1 %
Pre	L	H	W	<0.1 %	21.4 %	78.6 %
Pre	H	L	D	33.3 %	33.3 %	33.3 %
Pre	H	L	W`	33.3 %	33.3 %	33.3 %
Pre	H	H	D	33.3 %	33.3 %	33.3 %
Pre	H	H	W	<0.1 %	33.3 %	66.7 %
Post	L	L	D	55.6 %	33.3 %	11.1 %
Post	L	L	W	35.0 %	40.0 %	25.0 %
Post	L	H	D	25.0 %	62.5 %	12.5 %
Post	L	H	W	25.0 %	25.0 %	50.0 %
Post	H	L	D	55.6 %	11.1 %	33.3 %
Post	H	L	W	60.0 %	30.0 %	10.0 %
Post	H	H	D	33.3 %	0.01 %	66.7 %
Post	H	H	W	33.3 %	0.01 %	66.7 %

(b). CPT relating annual precipitation (PRECIP) to TP load from East Fork (LOAD_EF) .

PRECIP	● **LOAD_EF(lb/d)**		
	L	**M**	**H**
	< 0.32	**0.32-0.99**	**>=0.99**
D	20.0 %	40.0 %	40.0 %
W	37.5 %	31.3 %	31.3 %

(c). CPT relating annual precipitation (PRECIP) to TP load from the South Fork (LOAD_SF) .

PRECIP	**LOAD_SF (lb/d)**		
	L	**M**	**H**
	< 2.13	**2.13 – 7.84**	**>= 7.84**
D	41.4 %	34.5%	24.1 %
W	25.7 %	31.5 %	42.9 %

(d). CPT relating TP loads from the South Fork (LOAD_SF) and East Fork (LOAD_EF) to TP load at the confluence of the South and East Fork (LOAD_HW).

LOAD_SF	**LOAD_EF**	**LOAD_HW (lb/d)**		
		L	**M**	**H**
		<4.75	**4.75 – 16.84**	**>= 16.84**
L	L	>99.9 %	<0.1 %	<0.1 %
L	M	75.0 %	25.0 %	<0.1 %
L	H	< 0.1 %	> 99.9 %	<0.1 %
M	L	42.8 %	42.9 %	14.3 %
M	M	20.0 %	80.0 %	<0.1 %
M	H	<0.1 %	60.0 %	40.0 %
H	L	<0.1 %	> 99.9 %	<0.1 %
H	M	20.0 %	80.0 %	<0.1 %
H	H	7.7 %	7.7 %	87.6 %

(e). CPT relating TP load from subwatershed (LOAD_SW) and TP loads at the confluence of the South Fork and East Fork (LOAD_HW) to TP load at the inlet to the Hyrum Reservoir (LOAD_IN).

LOAD_SW1	**LOAD_HW**	**LOAD_IN(lb/d)**		
		L	**M**	**H**
		<9.80	**9.8 – 30.1**	**>= 30.1**
L	L	>99.9 %	< 0.1 %	< 0.1 %
L	M	46.7 %	53.3 %	< 0.1 %
L	H	< 0.1 %	50.0 %	50.0 %
M	L	66.7 %	33.3 %	< 0.1 %
M	M	6.06 %	94.0 %	< 0.1 %
M	H	< 0.1 %	20.0 %	80.0 %
H	L	< 0.1 %	>99.9 %	< 0.1 %
H	M	< 0.1 %	77.3 %	22.7 %
H	H	<0.1 %	6.0 %	94.0 %

(f). CPT relating annual precipitation (PRECIP) to subwatershed flow (FLOW_SW) above Hyrum.

PRECIP	**FLOW_SW1 (cfs)**		
	L	**M**	**H**
	< 0	**0 - 0.43**	**>=0.43**
D	2.7 %	51.4 %	46.0 %
W	12.9 %	70.5 %	16.7 %

(g). CPT relating annual precipitation (PRECIP) to flow from the East Fork (FLOW_EF).

PRECIP	**FLOW_EF (cfs)**		
	L	**M**	**H**
	< 3.3	**3.3 – 14.0**	**>=14**
D	42.1 %	34.2 %	23.7 %
W	27.7 %	32.3 %	40.0 %

(h). CPT relating annual precipitation (PRECIP) to flow from the South Fork

(FLOW_EF).

PRECIP	**FLOW_SF (cfs)**		
	L	**M**	**H**
	< 19.2	**19.2 – 49.8**	**>= 49.8**
D	43.3 %	40.0 %	16.7 %
W	25.7 %	25.7 %	48.6 %

(i). CPT relating flows from the East Fork (FLOW_EF) and South Fork (FLOW_EF) to flow at the confluence of the East Fork and South Fork.

FLOW_EF	**FLOW_SF**	**FLOW_HW (cfs)**		
		L	**M**	**H**
		<37	**37 – 81.3**	**>= 81.3**
L	L	87.5 %	12.5 %	< 0.1 %
L	M	75.0 %	25.0 %	<0.1 %
L	H	50.0 %	50.0 %	<0.1 %
M	L	81.8 %	18.2 %	<0.1 %
M	M	54.5 %	36.4 %	9.1 %
M	H	<0.1 %	50.0 %	50.0 %
H	L	>99.9 %	<0.1 %	<0.1 %
H	M	<0.01 %	<0.1 %	99.9 %
H	H	7.1 %	7.1 %	85.7 %

(j). CPT relating flow from subwatershed (FLOW_SW1) and flow at the confluence of the South Fork and East Fork (FLOW_HW) to flow at the inlet to the Hyrum Reservoir (FLOW_IN).

FLOW_SW1	FLOW_HW	FLOW_IN (cfs)		
		L < 36.8	M 36.8 – 88	H >= 88
L	L	> 99.9 %	< 0.1 %	< 0.1 %
L	M	62.5 %	37.5 %	< 0.1 %
L	H	< 0.1 %	< 0.1 %	> 99.9 %
M	L	> 99.9 %	< 0.1 %	< 0.1 %
M	M	< 0.1 %	> 99.9 %	< 0.1 %
M	H	< 0.1 %	10.8 %	89.2 %
H	L	97.7 %	2.3 %	< 0.1 %
H	M	< 0.1 %	> 99.9 %	< 0.1 %
H	H	< 0.1 %	25.0 %	75.0 %

(k). CPT relating TP load (LOAD_IN) and flow (FLOW_IN) to TP concentration (TP_IN) at the inlet of Hyrum Reservoir (Location 49805670).

LOAD_IN	FLOW_IN	TP_IN (mg/L)		
		L < 0.02	M 0.02 – 0.05	H > 0.05
L	L	11.1 %	65.3 %	23.6 %
L	M	6.7%	93.3%	<0.1%
L	H	33.3 %	33.3 %	33.3 %
M	L	<0.1 %	<0.1%	> 99.9%
M	M	<0.1 %	50.9%	49.1 %
M	H	<0.1 %	94.1 %	5.9 %
H	L	<0.1 %	<0.1 %	>99.9 %
H	M	<0.1 %	<0.1 %	>99.9 %
H	H	<0.1 %	22.9 %	77.1 %

Appendix G. CPTs for BN2.

(a). CPT relating Conservation Program option (OP_CON), agricultural landuse area (LAND_AG), point TP load (LOAD_P) and annual precipitation (PRECIP) to subwatershed TP load (LOAD_SW) below Hyrum.

OP_CON	**LAND_AG2**	**LOAD_P2**	**PRECIP**	**LOAD_SW2 (lb/d)**		
				L	**M**	**H**
				<7.02	**7.02 –18.82**	**>=18.92**
Pre	L	L	D	57.1%	14.3 %	28.6 %
Pre	L	L	W	33.3%	33.3 %	33.3 %
Pre	L	H	D	41.7 %	58.3 %	<0.1 %
Pre	L	H	W	<0.1 %	21.4 %	78.6 %
Pre	H	L	D	33.3 %	33.3 %	33.3 %
Pre	H	L	W`	33.3 %	33.3 %	33.3 %
Pre	H	H	D	33.3 %	33.3 %	33.3 %
Pre	H	H	W	<0.1 %	33.3 %	66.7 %
Pre	L	L	D	55.6 %	33.3 %	11.1 %
Pre	L	L	W	35.0 %	40.0 %	25.0 %
Pre	L	H	D	25.0 %	62.5 %	12.5 %
Pre	L	H	W	25.0 %	25.0 %	50.0 %
Pre	H	L	D	55.6 %	11.1 %	33.3 %
Pre	H	L	W	60.0 %	30.0 %	10.0 %
Pre	H	H	D	33.3 %	0.01 %	66.7 %
Pre	H	H	W	33.3 %	0.01 %	66.7 %

(b). CPT relating annual precipitation (PRECIP) to TP load from the Hyrum Reservoir (LOAD_5650).

PRECIP	**LOAD_5650 (lb/d)**		
	L	**M**	**H**
	< 2.96	**2.96 – 12.9**	**>= 12.9**
D	42.5 %	27.6 %	29.9 %
W	24.1 %	39.1 %	36.8 %

(c). CPT relating TP loads at the effluence of the Hyrum Reservoir (LOAD_5650) and TP load from subwatershed (LOAD_SW2) to TP load at the inlet to the Cutler reservoir (LOAD_5000).

LOAD_5650	**LOAD_SW2**	**LOAD_5000 (lb/d)**		
		L **<25.3**	**M** **25.3 – 52.6**	**H** **>= 52.6**
L	L	> 99.9 %	< 0.1 %	< 0.1 %
L	M	> 99.9 %	< 0.1 %	< 0.1 %
L	H	28.0 %	56.0 %	16.0 %
M	L	> 99.9 %	< 0.1 %	< 0.1 %
M	M	77.8 %	22.2 %	< 0.1 %
M	H	2.7 %	70.3 %	27.0 %
H	L	20.0 %	20.0 %	60.0 %
H	M	20.0 %	80.0 %	< 0.1 %
H	H	< 0.1 %	14.6 %	85.4 %

(d). CPT relating annual precipitation (PRECIP) to subwatershed flow (FLOW_SW) below Hyrum.

PRECIP	**FLOW_SW2 (cfs)**		
	L **< 28**	**M** **28 – 42.6**	**H** **>=42.6**
D	41.9 %	25.8 %	32.3 %
W	26.5 %	38.8 %	34.7 %

(e). CPT relating annual precipitation (PRECIP) to flow at the effluence from Hyrum Reservoir (FLOW_5650).

PRECIP	**FLOW_5650 (cfs)**		
	L **< 12**	**M** **12 – 48.41**	**H** **>=48.41**
D	44.0 %	28.0 %	28.0 %
W	23.2 %	38.4 %	38.4 %

(f). CPT relating flow at the effluence of the Hyrum Reservoir (FLOW_5650) and flow from subwatershed (FLOW_SW) below Hyrum to flow at the inlet to the Cutler Reservoir (FLOW_5000).

FLOW_5650	**FLOW_SW2**	**FLOW_5000 (cfs)**		
		L	**M**	**H**
		< 51.8	**51.8 – 97.2**	**>= 97.2**
L	L	>99.9 %	<0.1%	<0.1 %
L	M	>99.9%	<0.1%	<0.1 %
L	H	11.1 %	88.9 %	<0.1 %
M	L	92.9 %	7.1 %	<0.1 %
M	M	14.3 %	85.7 %	<0.1 %
M	H	<0.1 %	57.1 %	42.9 %
H	L	16.7 %	50.0 %	33.3 %
H	M	< 0.1 %	60.0 %	40.0 %
H	H	< 0.1 %	< 0.1 %	>99.9 %

(g). CPT relating TP load (LOAD_5000) and flow (FLOW_5000) to TP concentration (TP_5000) at the inlet of Hyrum Reservoir (Location 4905000).

LOAD_5000	**FLOW_5000**	**TP_5000 (mg/L)**		
		L	**M**	**H**
		=< 0.05	**0.05 – 0.15**	**>= 0.15**
L	L	26.1 %	60.9 %	13.0 %
L	M	81.8 %	18.2 %	< 0.1 %
L	H	> 99.9 %	< 0.1 %	< 0.1 %
M	L	< 0.1 %	36.4 %	63.6 %
M	M	3.0 %	90.9 %	6.1 %
M	H	25.0 %	75.0 %	< 0.1 %
H	L	< 0.1 %	< 0.1 %	> 99.9 %
H	M	< 0.1 %	21.1 %	78.9 %
H	H	4.0 %	58.0 %	38.0 %

Appendix H. Comparing the effect of selected FLOW_SW category change to the effect of selected FLOW_HW category change

(a). Marginal probabilities of categories of FLOW_IN for selected category of FLOW_SW.

		Selected FLOW_SW Category		
		L	**M**	**H**
FLOW_IN	L	58.6 %	49.2 %	54.5 %
	M	8.56 %	25.4 %	28.5 %
	H	32.8 %	25.5 %	17.0 %

(b). Marginal probabilities of categories of FLOW_IN for selected category of FLOW_HW.

		Selected FLOW_HW Category		
		L	**M**	**H**
FLOW_IN	L	99.2 %	5.3 %	0 %
	M	0.8 %	94.7 %	13.3 %
	H	0 %	0 %	86.7 %

Appendix I. CPTs for BN.

(a). CPT relating Conservation Program option (OP_CON), agricultural landuse area (LAND_AG), point TP load (LOAD_P) and annual precipitation (PRECIP) to subwatershed TP load (LOAD_SW) below Hyrum.

OP_CON	**LAND_AG**	**LOAD_P**	**PRECIP**	**LOAD_SW (lb/d)**		
				L	**M**	**H**
				<7.02	**7.02 –18.82**	**>=18.92**
Pre	L	L	D	57.1%	14.3 %	28.6 %
Pre	L	L	W	33.3%	33.3 %	33.3 %
Pre	L	H	D	41.7 %	58.3 %	<0.1 %
Pre	L	H	W	<0.1 %	21.4 %	78.6 %
Pre	H	L	D	33.3 %	33.3 %	33.3 %
Pre	H	L	W`	33.3 %	33.3 %	33.3 %
Pre	H	H	D	33.3 %	33.3 %	33.3 %
Pre	H	H	W	<0.1 %	33.3 %	66.7 %
Pre	L	L	D	55.6 %	33.3 %	11.1 %
Pre	L	L	W	35.0 %	40.0 %	25.0 %
Pre	L	H	D	25.0 %	62.5 %	12.5 %
Pre	L	H	W	25.0 %	25.0 %	50.0 %
Pre	H	L	D	55.6 %	11.1 %	33.3 %
Pre	H	L	W	60.0 %	30.0 %	10.0 %
Pre	H	H	D	33.3 %	0.01 %	66.7 %
Pre	H	H	W	33.3 %	0.01 %	66.7 %

(b). CPT relating annual precipitation (PRECIP) to TP load from the Hyrum Reservoir (LOAD_5650).

PRECIP	**LOAD_5650 (lb/d)**		
	L	**M**	**H**
	< 2.96	**2.96 – 12.9**	**>= 12.9**
D	44.1 %	27.9 %	27.9 %
W	22.5 %	38.8 %	38.8 %

(d). CPT relating annual precipitation (PRECIP) to subwatershed flow (FLOW_SW) below Hyrum.

PRECIP	**FLOW_SW (cfs)**		
	L	**M**	**H**
	< 28	**28 – 42.6**	**>=42.6**
D	41.9 %	25.8 %	32.3 %
W	25.8 %	39.2 %	35.0 %

(e). CPT relating annual precipitation (PRECIP) to flow at the effluence from Hyrum Reservoir (FLOW_5650).

PRECIP	**FLOW_5650 (cfs)**		
	L **< 12**	**M** **12 – 48.41**	**H** **>=48.41**
D	44.0 %	28.0 %	28.0 %
W	22.4 %	38.8 %	38.8 %

Appendix J. Category range of LOAD_5000, FLOW_5000 and TP_5000.

Category (percentile)	**Range**		
	LOAD_5000 (lb/day)	FLOW_5000 (cfs)	TP_5000 (mg/L)
A(0-5)	0 - 8.9	0 - 27.64	0 - 0.0429
B(5-10)	8.9 -16.3	27.64 - 34.44	0.0429-0.0527
C(10-15)	16.3 -19.7	34.44 - 39.75	0.0527-0.0596
D(15-20)	19.7 - 22.9	39.75 - 44.24	0.0596-0.0646
E(20-25)	22.9 - 24.1	44.24 - 49.09	0.0646-0.0707
F(25-30)	24.1 - 24.9	49.09 - 52.18	0.0707-0.0767
G(30-35)	24.9 - 28.4	52.19 - 56.15	0.0767-0.0835
H(35-40)	28.4 - 29.9	56.15 - 61.51	0.0835-0.0891
I (40-45)	29.9 - 31.2	61.51 - 63.07	0.0891-0.0947
J (45-50)	31.2 - 33.4	63.07 - 70.85	0.0947-0.1019
K(50-55)	33.4 - 37.4	70.85 - 75.43	0.1019-0.1083
L(55-60)	37.4 - 43.5	75.43 - 80.82	0.1083-0.1147
M(60-65)	43.5 - 48.8	80.82 - 88.68	0.1147-0.1222
N(65-70)	48.8 - 50.9	88.68 - 93.27	0.1222-0.1302
O(70-75)	50.9 - 57.2	93.27 - 102.54	0.1302-0.1401
P(75-80)	57.2 - 64.4	102.54-110.38	0.1401-0.1515
Q(80-85)	64.4 - 81.3	110.38-119.10	0.1515-0.1694
R (85-90)	81.3 - 110.3	119.10-131.38	0.1694-0.2098
S (90-95)	110.3 - 175.8	131.38-170.98	0.2098-0.3554
T (95-99.5)	175.8 - 5,055	170.98-934.96	0.3554-1.9411
U (99.5-)	5,055 -13,286	934.96-3,668	1.9411- 6.05 (3.0)

CURRICULUM VITAE

Joon- Hee Lee

PERSONAL DATA

Address : 124 USU Trailer Court Logan, UT 84341
Phone No.:435) 797- 6840 (Home), 435)797-3208 (Office)
E-mail : joon-hee.lee@aggiemail.usu.edu

EDUCATION

Ph.D. Civil & Environmental
Utah State University, Logan, Utah-June 2008, GPA 3.8/4.0
Adviser: David K. Stevens, Ph.D., P.E.
Dissertation : Rank-Data Distribution Method (R-D method) for Daily Time-Series Bayesian Networks and Total Maximum Daily Load Estimation

M.S. Environmental Science,
Hankuk University of Foreign Studies, Yongin, Korea-August 2000
Adviser : Guyoung Kang, Ph.D.
Thesis : A study on Electrochemical Oxidation Process for Treatment of Recalcitrant Organic Wastewater.

B.S. Chemical engineering, Kyung Won University, Sungnam, Korea
Senior Review Paper: A study on the Technology for String Hydrogen using hydrogenation of Metals

PROFESSIONAL TRAINING

- EPA BASINS modeling Training by Utah State University
- Physical Habitat Simulation System Training by Utah State University
- Groundwater Contamination modeling (MODFLOW, MODPATH, MT3D) by Koran Society of Groundwater Environment.

CERTIFICATE

- Environmental Engineer for air pollution control by Korean Government.

WORK EXPERIENCE

Dec. 2000 – Present Utah Water Research Laboratory, Logan, UT.
Research assistant

Duties : Visual Basic programming for nutrient load calculation.
Water quality data analysis (data organizing, statistical tests using computer program, interpreting and reporting the results of statistical tests.)
Evaluation of the effects of water quality conservation projects (building the

statistical model, building the database and simulating the statistical model and interpretation)

Jan. 2004- May 2004 Department of Civil and Environmental Engineering, Utah State University, Logan, UT
Teaching Assistant for Environmental Process Laboratory course

Duties : Preparing experiments (setting the experimental apparatuses, teaching the background and procedure of each experiment)

Mar. 1998 - Aug. 2000 Institute of Environmental Science, Hankuk Univ. of Foreign Studies, Korea
Research assistant.

Duties : Analysis of wastewater quality.
Operation of pilot-scale evaporator for night soil treatment
Operation, sampling, wastewater quality analysis and data analysis for Electrochemical Oxidation Treatment of industrial wastewater.

Apr. 1997 - Dec. 1997 DC Chemical Co., Gunsan plant, No.2 chemical production plant division, Korea
Chemical engineer.

Duties: Setting up wastewater treatment system.
Setting up water glass dissolving and purification systems

Dec. 1994 - Mar. 1997 DC Chemical Co., Gunsan plant, No.1 chemical production plant division, Korea
Chemical engineer.

Duties: Increasing yield of agricultural chemical product
Increasing the capacity of Chlor-Alkali Plant
- Electrolyzer test operation
- Caustic soda evaporator installation and test operation
- Sodium hypo chlorite reactor design and test operation

Air Pollution management
- Measuring Particulate Mater Concentration
- Management of organic chemical residue incinerator

TECHNICAL SKILLS

- Analysis of Quality for Water & Wastewater : Nitrogen, Phosphorus, COD, BOD
- Analysis of contamination metals in using AA.
- Simulation of Surface Water Quality using Basins v 3.0 including Qual2e, HSPF, PLOAD.
- Building and simulating Bayesian Network statistical model using NETICA and HUGIN.

- Programming Skill : Visual Basic, FORTRAN 77
- Skills of data analysis software: Sigma- plot, S-Plus, Minitab
- GIS Skill: Arc Hydro, Arc GIS. Arc View.

RESEARCH ACTIVITY

1. “Little Bear River Project” (Investigation of the change in the water quality in the Little Bear River Watershed in response to the implementation of Best Management Practices), funded by USDA
   - Period : Nov, 2004 – Present
   - Duties : Water Quality Data Analysis, Building Bayesian Networks to evaluate the effect of conservation practice and other factors on water quality.
2. "Nooksack Project" (Development of environmental management tool using modeling approach), funded by the Washington state department of ecology.
   - Period : Dec. 2000 – Aug. 2002
   - Duties : Programming to calculate the constituent concentration and load cause by non-point source and data analysis
3. "Development of wastewater treatment system using electrode contact method", funded by Korea Ministry of environment.
   - Period : Dec. 1998 - Aug. 2000
   - Duties : Operation of experimental apparatus, wastewater quality analysis and estimating equipment design factors.
4. "A study on the wastewater treatment system combining evaporation, nitrification and denitrification", funded by Korea Ministry of environment.
   - Period : Feb. 1998 - Nov. 1998
   - Duties : Operation of experimental apparatus and wastewater quality analysis.

PUBLICATIONS, PRESENTATIONS

Journal

- Lee,H, Y. Lee, H.Han, G. Kang and J. Lee, A Study on the Treatment of Nightsoil by Flash Evaporation, Journal of Korean Society of Environmental Engineers, Vol. 21, No. 11, pp 2071- 2078 (1999)

Conference Proceeding

- Lee, J., D.K. Stevens, Filling Data Gaps by a Rank-Data distribution method (R-D method), Spring Runoff Conference, Mar. 2008 , Utah State University, Logan, UT.
- Lee, J., G. Kang, J. Song, M. Jang and J. Lim, A Study on Electrochemical Oxidation Process for Treatment of Recalcitrant Organic Wastewater, Korean Society of Environmental Engineers conference, Nov.1999, Kwangju, Korea.
- Yoon.J., J.Lee, G. Kang, H. Lee, Y. Lee, Nitrification of night-soil, Using a Spiral Polymeric Membrane Reactor, Korean Society of Environmental Engineers conference, Nov. 1998, Kongju, Korea.
- Lee, J., G. Kang, Investigation into circulation of ground water by air sparging, Korea Soil Environment Society conference, Nov 1998, Seoul.

Presentation

- Lee,J., D.K. Stevens, Bayesian Networks to Evaluate the effect of Water Conservation Program on water quality at Little Bear River Watershed. Environmental Engineering Division Seminar, Jan. 2007, Utah State University, Logan, UT.
- Lee,J., G.Kang, A Study on Electrochemical Oxidation Process for Treatment of Recalcitrant Organic Wastewater, Environmental Engineering Division Seminar, Sep. 2001, Utah State University, Logan, UT

PROFESSIONAL ACTIVITIES

- Member of American Society of Civil Engineers (3/2005 - Present)
- Member of Korean Society for Environmental Engineers (6/1998 - 12/2000)
- Member of Korean Society for Chemical Engineers (5/1995-12/1998)

ASSISTANTSHIP, SCHOLARSHIP

- Scholarship from Paul Riley AWRA-Utah Student Conference and Scholarship Competition (May, 2008)
- Research and teaching assistantship from Utah State University (1/2001-present)
- Scholarship from Kyung Won Univ. (9/1991 - 2/1992)
- Scholarship from Hankook Univ. of Foreign Studies (9/1999 - 2/2000)