MONTHLY AND SEASONAL STREAMFLOW FORECASTING IN THE RIO GRANDE BASIN

BY

ABUDU SHALAMU, B.S., M.S.

(Other name: Abdusalam Mamat)

A dissertation submitted to the Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy

Major Subject: Civil Engineering

Minor Subjects: Experimental Statistics, Geographic Information Systems

New Mexico State University
Las Cruces, New Mexico
August 2009

“Monthly and Seasonal Streamflow Forecasting in the Rio Grande Basin”, a dissertation prepared by Abudu Shalamu in partial fulfillment of the requirements for the degree, Doctor of Philosophy in Civil Engineering, has been approved and accepted by the following:

Linda Lacey
Dean of the Graduate School

James Phillip King
Chair of the Examining Committee

Date

Committee in charge:

Dr. James Phillip King

Dr. Salim A. Bawazir

Dr. Zohrab A. Samani

Dr. Robert Steiner

Dr. Michael N. DeMers

DEDICATION

This dissertation and all the work behind it are dedicated to…

To the memory of my father, Mamat Ismael, who inspired me all the time to become better man

To my mother, Kurwanbuwi, who is a hard-working and kind family woman, who raised me through countless difficulties of life

To my wife, Sophia Ibrahim, who made all of this possible, for her endless encouragement, patience and love

And

To my three precious daughters, Rena, Adina and Flora, the lights of my life

And

To the people who have supported me and were there for me throughout the entire doctorate program

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to one person, who made all this possible, through his professionalism, endless care for me and for my family, his countless hours of reflecting, reading, encouraging, and most of all, patience during past five years. He is as close to me as a friend; I respect him as my mentor; thank you Dr. James Phillip King! My thanks and appreciation to you for persevering with me as my advisor throughout the time it took me to complete this research and write the dissertation. Thanks for your guidance, advice, help, for believing in me even when I have doubts about myself; your support has been decisive in helping me to achieve my academic goals.

I express my special thanks to Dr. Salim Bawazir, both as my mentor and a friend, who inspires me all the time and has always been there whenever I have a difficult time, both academically and emotionally. He has consistently helped me to keep a perspective on what is important in life and has shown me how to deal with reality. His hard-working attitude and professionalism has influenced me during this time and will do the same for the rest of my career.

I wish to thank my dissertation committee members who were more than generous with their expertise and precious time. I am grateful to Dr. Zohrab Samani, for his encouragement and guidance for my research. Many thanks to Dr. Robert Steiner and Dr. Michael N. DeMers for agreeing to serve on my committee. Especially, I appreciate Dr. Steiner for his great help in solving statistical problems and guidance in writing SAS codes. I appreciate the useful feedback and guidance of Dr. Thomas Pagano, a hydrologic forecaster from Natural Resources Conservation Service, for his continuous and timely technical support throughout my dissertation research. He has been an excellent advisor away from the university.

I am grateful to the many people who offered help for me and my family. Especially, I need to express my gratitude and deep appreciation to Carroll and Jim Fox, who are like grandparents to my daughters, for their hard work in reviewing and editing every sentence of my dissertation and their consistent help and hospitality from the day we arrived in Las Cruces to the present with their kind hearts and endless love. Thanks to Rita King, a nice and kind person, who has helped me and my family through her encouragement and care for the last five years.

I must acknowledge as well the many friends, colleagues who assisted, advised, and supported my research and writing efforts over the years. Special thanks to Julie Moore for the editing work on this dissertation. Thanks to Osman Sawut, Peng Lixin, Cui Chunliang and Aliph Ablimit for their consistent support of my study for the past five years from China. I would also like to express my appreciation to my fellow graduate students, Jimmy Moreno, Cevik Yeliz and Kate Adams, for their valuable discussions during my study. Thanks to Elvia Cisneros and Lydia Madrid for their countless coordination works and for helping me every time I asked for it.

VITA

March 10, 1968 Born in Ili, Xinjiang Uyghur Autonomous Region, China

July, 1984 Graduated from No. 3 High School, Tokkuztara County, Xinjiang, China

1984-1986 Chinese Language, Northwest University of Nationalities, Lanzhou, China

1986-1990: Bachelor’s Degree in Irrigation & Drainage Engineering, Hohai Univ., Nanjing, China.

1990-1993: Master’s Degree in Irrigation & Drainage Engineering, Hohai Univ., Nanjing, China.

1993-2008: Senior Engineer and Vice-Director, Xinjiang Institute of Water Resources & Hydroelectric Science, China

2003-2004: Research Scholar at the Department of Biological and Agricultural Engineering, University of California, Davis

2005-Present: Doctoral Candidate/Graduate Assistant, Department of Civil Engineering, New Mexico State University, Las Cruces, New Mexico.

Professional and Honorary Societies

China Water Conservancy Society

Selected Awards

2008. Research project award “Investigation of Improved Operational Streamflow Forecasting in the Rio Grande Basin,” by Water Resource Research Institute, New Mexico.

2008. “Drainage System Design of Jiashi County,” Second place, as a principal designer, in the fourteenth regional contest of excellent project design, Xinjiang Uyghur Autonomous Region, China.

2008. "Investigation of Intelligent Cotton Drip Irrigation Key Techniques under Plastic Mulching," Second place, in the annual regional award in water resources science and technology research, Xinjiang, China.

2006. "Crop Evapotranspiration under Drip Irrigation Conditions," Second place, in Xinjiang regional young specialists contest, investigation report.

2003. "Heshuo County 1000ha Micro Irrigation Project," Second place, principal investigator, in the ninth regional contest of excellent project design, Xinjiang Uyghur Autonomous Region, China.

2001. "Outstanding worker and researcher" Bureau of Science and Technology of Xinjiang Uyghur Autonomous Region, China

Selected Publications

Shalamu, A., Bawazir, A. and King, J. P. "Infilling Missing Daily Evapotranspiration Data Using Neural Networks," *Journal of Irrigation and Drainage Engineering*, **under 2nd review.**

Shalamu, A., King, J. P., and Pagano, T. "Application of Partial Least Squares Regression in Seasonal Streamflow Forecasting," *Journal of Hydrologic Engineering*, **under 2st review.**

Shizhang, P., and Shalamu, A. (1998). Optimum Calculation of Leaching Water in Irrigation District. *Hohai University Academic Journal*, 26(2), 105-109.

Shalamu, A. (1997). Approaches to Several Problems in The Development of Drip Irrigation in Turpan Basin, *Journal of Xinjiang Water Resourses*, 94, 39-41.

Shizhang, P., and Shalamu, A. (1995). Brief Introduction Soil-Water-Crop Relationships under Salinity Conditions. *Journal of Irrigation and Drainage,*14(1), 19-23.

Shizhang, P., and Shalamu, A. (1995). An Optimal Irrigation & Drainage Model for Prevention of Soil Salinizations. *Journal of Advances in Water Sciences*, 6(3), 183-188.

ABSTRACT

MONTHLY AND SEASONAL STREAMFLOW FORECASTING IN THE RIO GRANDE BASIN

BY

ABUDU SHALAMU

Doctor of Philosophy, Civil Engineering

New Mexico State University

Las Cruces, New Mexico, 2009

Dr. James Phillip King, Chair

Improving the quality of streamflow forecasting has always been an important task for researchers and water resources managers. In this research, the seasonal and monthly streamflow forecasting using various data-driven statistical models was investigated for naturalized streamflow at Del Norte Gaging Station, Rio Grande, Colorado and observed Elephant Butte Reservoir net inflow, Rio Grande, New Mexico. The application of partial least squares regression (PLSR) and hybrid models in seasonal streamflow forecasting, the inclusion of snowpack and El Niño Southern Oscillation (ENSO) information in the monthly and seasonal streamflow forecasting were investigated. The modeling methods included autoregressive integrated moving average (ARIMA) models, transfer function-noise (TFN) models, artificial neural networks (ANN) models, principal components regression (PCR), and PLSR. Two hybrid modeling approaches, including TFN forecast modification using ANN (TFN+ANN) and a combination of principal components analysis (PCA) and ANN (PCA+ANN), were also applied in seasonal streamflow forecasting. The ARIMA models were used as a benchmark for the comparison of the performance of the models. Additionally, the forecasting results were compared to the Natural Resources Conservation Service (NRCS) official forecasts to evaluate the performance of the proposed models.

The results of seasonal flow modeling indicated that using a composite precipitation index is a relatively effective method in both improving forecast accuracy and developing parsimonious regression models with fewer and readily-available input variables. In comparison of PLSR and PCR, similar forecast accuracies were obtained for both methods in jackknife cross validation and test period (2003-2007) although PLSR has higher calibration coefficient of determination (R^2) and can reach its minimum prediction error with a smaller number of components than PCR. The comparison with NRCS official forecasts showed that the application of PLSR in seasonal streamflow forecasting is promising. The application of hybrid modeling approaches showed potential capability of hybrid models to improve forecast accuracy in seasonal streamflow modeling as compared to single models. For Elephant Butte net inflow modeling, the normalized root mean square errors (NRMSE) of forecasted and observed net inflow for April-July decreased from 0.36 to 0.19 from single TFN model to the TFN+ANN hybrid approach. The performance of PCA+ANN approach was also comparable to the TFN+ANN.

The results of monthly flow modeling suggested that the forecast modification using a combination of TFN and ANN methods (TFN+ANN) displayed better performance than the ANN models that were specifically calibrated for each month of the snowmelt season and was able to improve forecast accuracy significantly compared to other models. The normalized root mean square errors (NRMSE) for one-month-ahead forecasts for Del Norte Gaging Station were 0.46, 0.41, 0.24 and 0.21 for simple ARIMA, TFN model, ANN models and TFN+ANN approach respectively. These findings suggested that the TFN+ANN method is an advantageous approach in improving forecast accuracy and the ANN is a useful tool in forecasting monthly streamflow, whether it is used for direct modeling or used as a forecast modification technique.

The findings of this study may provide an impetus for streamflow forecasting by using hybrid modeling approach and PLSR method with various operationally available climatic variables. PLSR approach can be combined into NRCS's operational forecasting environment for possible forecast improvement.

TABLE OF CONTENTS

Chapter

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

LIST OF ABBREVIATIONS

A	Final Selected Number of Components
ACF	Autocorrelation Function
AHPS	Advanced Hydrologic Prediction System
AIC	Akaike Information Criterion
ANN	Artificial Neural Networks
ARIMA	Autoregressive Integrated Moving Average
BOR	Bureau of Reclamation
CCA	Canonical Correlation Analysis
CCF	Cross Correlation Function
E	Model Efficiency
ENSO	El Niño Southern Oscillation
ESP	Ensemble Prediction System
FFNN	Feed Forward Neural Networks
GIS	Geographic Information System
KAF	Thousand Acre-feet
M	Number of Component with p>0.1 in van der Voet's test
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MLP	Multilayer Perceptron

MLR	Multiple Linear Regression
MNN	Modular Neural Network
MRE	Mean Relative Error
MSE	Mean Squared Errors
N	Number of Component that has minimum PRESS
NRCS	Natural Resources Conservation Service
NRMPRESS	Normalized Root Mean Prediction Residual Sum of Squares
NRMSE	Normalized Root Mean Squared Error
NWCC	National Water and Climate Center
NWS	National Weather Service
P	Number of Component that passed sign test in sequence
PACF	Partial Autocorrelation Function
PC	Principal Components
PCA	Principal Components Analysis
PCA+ANN	Combination of Principal Components Analysis and Artificial Neural Networks
PCR	Principal Components Regression
PDO	Pacific Decadal Oscillation
PLSR	Partial Least Squares Regression
PRCP INDEX	The Composite Precipitation Index
PRCP	Precipitation
PRESS	Prediction Residual Sum of Squares

R^2	Coefficient of determination
RMPRESS	Root Mean Prediction Residual Sum of Squares
RMSE	Root Mean Squared Error
RS	Remote Sensing
SNOTEL	Snow Telemetry
SOI	El Niño Southern Oscillation Index
SRP	Salt River Project
SSE	Sum of Squared Error
SST	Sea Surface Temperatures
SWE	Snow Water Equivalent
TEMP	Temperature
TFN	Transfer Function-Noise
TFN+ANN	Combination of Transfer Function-Noise and Artificial Neural Networks
USGS	United States Geology Service
VIP	Variable Importance for the Projection
VIPER	Visual Interactive Prediction and Estimation Routines
WSOs	Water Supply Outlooks

LIST OF SYMBOLS

a	The number of factor used in the regression
a_t	White noise process
b	Delay parameter or dead time, is defined as $B^b x_t = x_{t-b}$,
B	The backward shift operator defined as $BX_t = X_{t-1}$
B	The matrix of coefficients, (m × p)
C	Constant parameter
C_i	The critical value, the difference between the PRESS statistics for *i* and i_{min} components
$D_{i,j}$	The critical value for van der Voet's test is based on the differences between squared predicted residuals
d	Order of consecutive differencing of the input variable
d'	Order of consecutive differencing of the dependent variable
E	Residual matrix of X
$E(n)$	The sum of squared errors of the network
E*	The residual matrix, (n × p)
F	The forecasted flow
$\overline{F}$	Mean of forecasted flow
F*	Residual matrix of Y
f_h	The activation function of the hidden neuron

f_0	The activation function for the output neuron
$F_{APR\text{-}SEP}$	Forecasted April-September runoff volume
F_{MJ}	Elephant Butte Reservoir March-July net inflow forecasts
F_{NRCS}	NRCS forecasts for March-July natural flow at San Marcial Gaging Station
i_{min}	The number of components for which PRESS is minimized.
m	The number of independent variables
n	The number of observations in calibration set
n	Number of samples
n'	Number of SNOTEL sites used in the forecast equation
m'	Number of months from October to forecast date of a water year
N_t	Stochastic disturbance (noise) term
$\mathbf{P'}$	The matrix of X loadings, (a × m)
p	The number of dependent variables
$\mathbf{p}_{\mathbf{h'}}$	Row vector of loadings for X block
$\mathbf{Q'}$	The matrix of Y loadings
$\mathbf{q}_{\mathbf{h'}}$	Row vector of loadings for Y block, factor h
$R_{i,j}$	The *j*th predicted residual for the model with *i* extracted components
r_{ij}	The correlation coefficient of the precipitation of month *j* and SNOTEL site *i* with the seasonal flow volume;
SWE_t	Snow water equivalents on the first day of month *t*
$SWE_{t'}$	Snow water equivalents on the first day of month *t'*

t'	First month of season t
$\mathbf{T}$	The matrix of X scores, (n × a)
$\mathbf{t_h}$	Column vector of scores for X block
$\mathbf{U}$	The matrix of Y scores
$\mathbf{u_h}$	Column vector of scores for Y block , factor h
w_{ji}	A weight in the hidden layer connecting the i^{th} neuron in the input layer and the j^{th} neuron in the hidden layer
w_{jo}	The bias for the j^{th} hidden neuron
w_j	A weight in the output layer connecting the j^{th} neuron in the hidden layer
w_o	The bias for the output
$\mathbf{X}$	The matrix of predictor variables, (n × m)
X_t	Explanatory variable
x_{pi}	A value of the i^{th} input for pattern p
$\mathbf{Y}$	The matrix of dependent variables, (n × p)
$\hat{y}_p(n)$	The actual response of the network at the n^{th} iteration for pattern p
$y_p(n)$	The desired target responses of the network at the n^{th} iteration for pattern p
$\hat{y}_j(i)$	Predicted values for extracted factor i for jth observation
y_t	Deseasonalized flow series for season or month t
$\hat{y}_t$	Forecasted deseasonalized seasonal or monthly flow at season or month t

Y_t Observed flow at season or month *t*

$\overline{Y}_t$ Sample average of the observed flow for season or month *t*

$Y_{t,\ forecasted}$ Seasonal or monthly flow forecasts using TFN model with PRCP input at season or month *t*

$Y_{t,\ modified}$ Seasonal or monthly flow forecasts after the forecast modification at season or month *t*

Z_i The *i*th principal components

$\hat{\sigma}_t$ Sample standard deviation of the observed flow for season and month t

φ, θ, ω, ψ Coefficients of forecast equations

1 INTRODUCTION

1.1 Problem Statement

Water resources play a crucial role in the economic development of the southwestern United States. The region's explosive population growth and resulting new demands on limited water resources require efficient management of existing water resources rather than building new facilities to meet the challenge. In the water management communities, it is well known that to combat water shortage issues, maximizing water management efficiency based on streamflow forecasting is crucial.

Streamflow forecasting is of vital importance to flood mitigation and water resources management and planning. While short-term forecasting such as hourly or daily forecasting is crucial for flood warning and defense, long-term forecasting based on monthly, seasonal or annual time scales is very useful in reservoir operations and irrigation management decisions such as scheduling releases, allocating water to downstream users, drought mitigation and managing river treaties or implementing compact compliance. Particularly, the seasonal volumetric streamflow represents an important hydrologic parameter for water supply purposes in the southwestern United States, since it represents spring-summer snowmelt runoff which accounts for a large portion of annual runoff. Hence, seasonal snowmelt runoff forecasting is particularly important in improving water management efficiency and benefiting various water use needs such as irrigation, hydropower generation, recreation, and environmental protection.

The high-quality streamflow forecasts and efficient use of these forecasts in water management can result in considerable economic and social benefits. There are many studies regarding the evaluation of value of streamflow forecasting (e.g., Burges and Hoshi, 1978; Dong et al., 2006; Hamlet et al., 2002; Kim and Palmer, 1997; Parker et al., 2005; Yeh et al., 1982). The quality of streamflow forecasting can be evaluated in terms of lead time and accuracy. Lead time refers to the time interval between the forecast issuing date and the occurrence of the forecasted flow event (Lettenmaier and Wood, 1993). Previous studies confirm the increased benefits of high-quality streamflow forecasting for flood defense and water management through improved reservoir operation. These benefits include improving water quality and navigation conditions, protecting wildlife and environmental restoration, reduction of expenses in flood mitigation and drought management, and protecting human life and property.

Due to the importance of hydrologic forecasting, a considerable number of forecasting models and methodologies have been developed and applied in streamflow forecasting. These streamflow forecasting models can be categorized as process-driven methods and data-driven methods (Wang, 2006). The process-based modeling approach is a knowledge-driven modeling process that explains the underlying process. Various forms of rainfall-runoff models such as lumped, semi-distributed and distributed, and snowmelt-runoff models are in this category. Data-driven models, on the other hand, are based on a limited knowledge of the internal physical mechanism of the watershed system and rely on data describing input and output characteristics. They are essentially black-box models that characterize the relationships between inputs and outputs without a consideration of the details or explicit simulation of the underlying physical process. They may include regression models, time series models, artificial neural networks (ANN) models and non-parametric models such as K-nearest neighbor method. Recently, data-driven modeling has become quite popular in streamflow forecasting, due to the increase in data availability from metering stations, real-time data retrieval, and increasing computational capability with the development of more robust methods and computer techniques (Wang, 2006).

Streamflow forecasting is challenging because of the complexity of hydrologic systems. Improving the quality of streamflow forecasting has always been an important task for researchers and hydrologic forecasters. There is no single streamflow forecasting method that provides optimum forecast results under all circumstances. No single forecasting model is powerful and general enough to outperform others for all types of catchments and under all circumstances or even one catchment with different behavioral phases (Shamsheldin, 2004). It is expected that better forecasting models can be developed for a specific basin due to increasing data availability, computational power, sophistication of modeling theory and software development.

There are several issues in improving seasonal streamflow forecasting. First, more robust multivariate regression methods may be introduced into seasonal streamflow forecasting that could efficiently deal with the multi-collinearity of predictor variables. Multivariate regression methods that are currently used in the hydrologic community, such as principal components regression and Z-score regression, can effectively eliminate the collinearity problem. There are also other robust methods such as canonical correlation analysis (CCA), partial least squares regression (PLSR) that can be applicable to seasonal streamflow forecasting to deal with collinearity issues. However, the application of these alternative methods in seasonal streamflow forecasting is only at an early stage. More research efforts are needed to introduce them to the field of operational streamflow forecasting. Second, the primary operational method of seasonal streamflow forecasting in the western United States is still focused on regression based methods. Risley et al. (2005) explored the application of artificial neural networks on seasonal streamflow forecasts. Although the goal of improved forecast accuracy by using neural networks was not conclusive in their study, it provided an impetus for the application of more complex methods in seasonal streamflow forecasting. More research work is needed in developing more robust and complex modeling methods, including hybrid modeling approaches in seasonal streamflow forecasting. Third, there are hundreds of input variables available for seasonal streamflow forecasting equation development. The selection of important variables and obtaining more reliable forecast equations are always challenging. Apart from searching for optimal or near-optimal variable combinations as proposed by Garen (1992), other possible data pre-processing approaches leading to improved forecast skills, such as developing composite indices and using them as inputs for multivariate regression equation, also need to be tested in seasonal streamflow forecasting equation development.

Monthly streamflow forecasting is also as important as the seasonal streamflow forecasting in water resources allocation and management. In particular, the monthly reservoir net inflow forecasting is of great significance to reservoir management as it is an indication of water availability from a reservoir. The time series models, including ARIMA and TFN, have long been used in the modeling of monthly streamflow processes. The challenging problem in monthly streamflow forecasting models may be the inclusion of snowpack information in time series models for spring-summer season monthly flow forecasting. This may be due to the seasonal presence of snowfall in a year and the timing of snowmelt runoff, which makes it difficult to get equally spaced snow water equivalent time series and systematic cross correlation relationship between monthly snow water equivalent and monthly streamflow for entire year in the building of TFN model. Further, the separate performance evaluation of monthly time series models for each month of the year is crucial because the better performance of monthly models evaluated for the entire year do not mean the models are necessarily applicable for every month of a year. Sometimes the observed mean will be better than the time series model forecasts for some months of a year.

Finally, the operational capability of streamflow forecasting models should be paid more attention in the development of forecasting models. Operational limitation of streamflow forecasting models is the limited application of models in real life forecasting environment due to on-time data availability and/or complexity of model itself. Recently developed high-quality spatially distributed hydrological modeling for improved streamflow forecasting using remote sensing (RS) and geographic information system (GIS) technologies has showed a promising capability of improving forecast accuracy. However, the distributed models are often not convenient and timely because of limited satellite data availability in the operational forecasting environment (Pagano, 2005). In addition to distributed models, some models that have been developed using weather station data as predictor variables may also not be operationally robust since the weather station data from climate networks are not readily available on the first day of a month (Pagano, personal communication, September 20, 2007). Therefore, the operational capability may also be used as an index in evaluating the robustness of the streamflow forecasting models.

1.2 Objective

To address the above issues in seasonal and monthly streamflow forecasting, the objectives of this research are the following:

1) To investigate more robust multivariate regression techniques, namely partial least squares regression, in improving seasonal volume streamflow forecasting and to examine the utilization of composite indices in obtaining better forecasting skills.

2) To investigate the application of hybrid modeling approaches such as the combination of time series and neural networks modeling methods, and the combination of principal components analysis and neural networks methods in seasonal streamflow forecasting.
3) To study the improvement of the performance of monthly streamflow forecasting time series models with the inclusion of basin snowpack information and El Niño Southern Oscillation (ENSO) signals in the modeling and the performance evaluation issues of monthly forecasting models.
4) To enhance the operational capability of streamflow forecasting models by using readily available Snow Telemetry (SNOTEL) data and simpler model structures; evaluate the performance of models that are developed using only SNOTEL data as inputs.
5) To make specific recommendations for forecasting target seasonal and monthly flows, and present the most functional models for those forecasts.

1.3 Scope and Limitations

To achieve the objectives described above, two time scales, seasonal and monthly streamflow volume, and two hydrologic variables, naturalized streamflow volume at Del Norte Gaging Station, Rio Grande, Colorado and reservoir net inflow for Elephant Butte Reservoir, Rio Grande, New Mexico, were selected for modeling in this study. The modeling procedures included autoregressive integrated moving average (ARIMA) models, transfer function-noise (TFN) models with SNOTEL precipitation input, artificial neural networks (ANN) models, principal components regression (PCR), partial least squares regression (PLSR). In addition, the application of two hybrid modeling approaches including forecast modification using ANN and a combination of principal components analysis (PCA) and ANN in streamflow forecasting was also investigated in the study.

To build the various models, the predictability of different hydrologic variables such as snow water equivalent, SNOTEL precipitation, SNOTEL temperature and El Niño Southern Oscillation Index (ENSO) on streamflow processes was analyzed using cross correlation between those variables and the streamflow. The potential variables and their lag relationships for the model inputs were identified in monthly and seasonal time scales for the study sites. The forecast performance comparisons of the proposed models were performed and potential capability of hybrid modeling approaches were analyzed using various statistical indices for model performance evaluation. Based on the analysis of model performance, the final models that could be used for operational streamflow forecasting in the study basins were suggested for both seasonal and monthly time scales.

To develop improved seasonal and monthly streamflow forecasting models and enhance the operational capability of the proposed models, the following considerations and procedures were used in the study:

1) In order to reduce forecast error, all the predictor variables used in the modeling were observed values; no predicted/forecasted values were used in any of the modeling procedures.

2) The ANN models including single or hybrid models were developed for only spring-summer seasons in the selected basins to include snowpack information in the models.
3) The predictor variables including snow water equivalent, precipitation, and temperature used in the study were from NRCS automatic SNOTEL stations. Weather station data were used to extend the period of SNOTEL data.
4) The easily accessible real time data were used in the model development; all data used as model inputs were available online from Natural Resources Conservation Center (NRCS), United States Geology Service (USGS), and National Weather Service (NWS) websites.
5) The autoregressive integrated moving average (ARIMA) models were used as a benchmark for the comparison of the model performance. Additionally, the forecasting results were also compared to the NRCS official forecasts to evaluate the performance of the proposed models in the study relative to current practice.

1.4 Study Sites and Data Used

Two Rio Grande watersheds were used as the study basins: the Rio Grande Headwaters above Del Norte Gaging Station, Colorado and Rio Grande Basin above Elephant Butte Reservoir, New Mexico. The main reason that the two watersheds were used in the study is their importance in the Rio Grande Compact compliance and water management in the region. Two different hydrologic variables, namely Del

Norte natural flow and Elephant Butte net inflow, and two time scales including monthly and seasonal flow volumes have been modeled in this study. The hydrologic settings of the basins, the basic data including snow water equivalent (SWE), precipitation (PRCP), temperature (TEMP), El Niño Southern Oscillation Index (SOI) and some approaches that used in the data preparation are described in the following sections.

1.4.1 Rio Grande Headwaters Basin above Del Norte Gaging Station

The Rio Grande near Del Norte Gaging Station, Colorado was selected for modeling site for following reasons:

1) The Del Norte Gaging Station is one of the main index stations for determining water delivery obligations from Colorado to New Mexico based on the Rio Grande Compact.
2) The flow is relatively less regulated and located in the uppermost part of the Rio Grande.
3) Relatively longer natural flow, snow water equivalent, precipitation and temperature data are available in the Basin.

1.4.1.1 Site Description

The Rio Grande near Del Norte, Colorado (USGS Station no. 08220000), is located at latitude 37°41'22"N, longitude 106°27'38"W, Rio Grande County, Colorado. The drainage area is about 1320 square miles; the elevation of the site is

7980 ft; and the period of record dates back to 1889. The natural flow of stream is affected by storage reservoirs, transmountain diversions from Colorado River Basin, diversions for irrigation and municipal use, ground water withdrawals, return flows from irrigated areas, and effluent flows from sewage treatment plants. Flow has been regulated by Beaver Creek Reservoir since 1910, Santa Maria Reservoir since 1912, Rio Grande Reservoir since 1912, and Continental Reservoir since 1925, with a combined capacity of 126,100 acre-ft, and by several smaller reservoirs.

The National Water and Climate Center of the Natural Resources Conservation Service (NRCS) operates and maintains automatic Snow Telemetry (SNOTEL) measurement sites in the Basin (Figure 1.1 and Table 1.1) that provide continuous measured time series data from the 1980s, and most SNOTEL sites have estimated snow water equivalent time series data back to the 1960s.

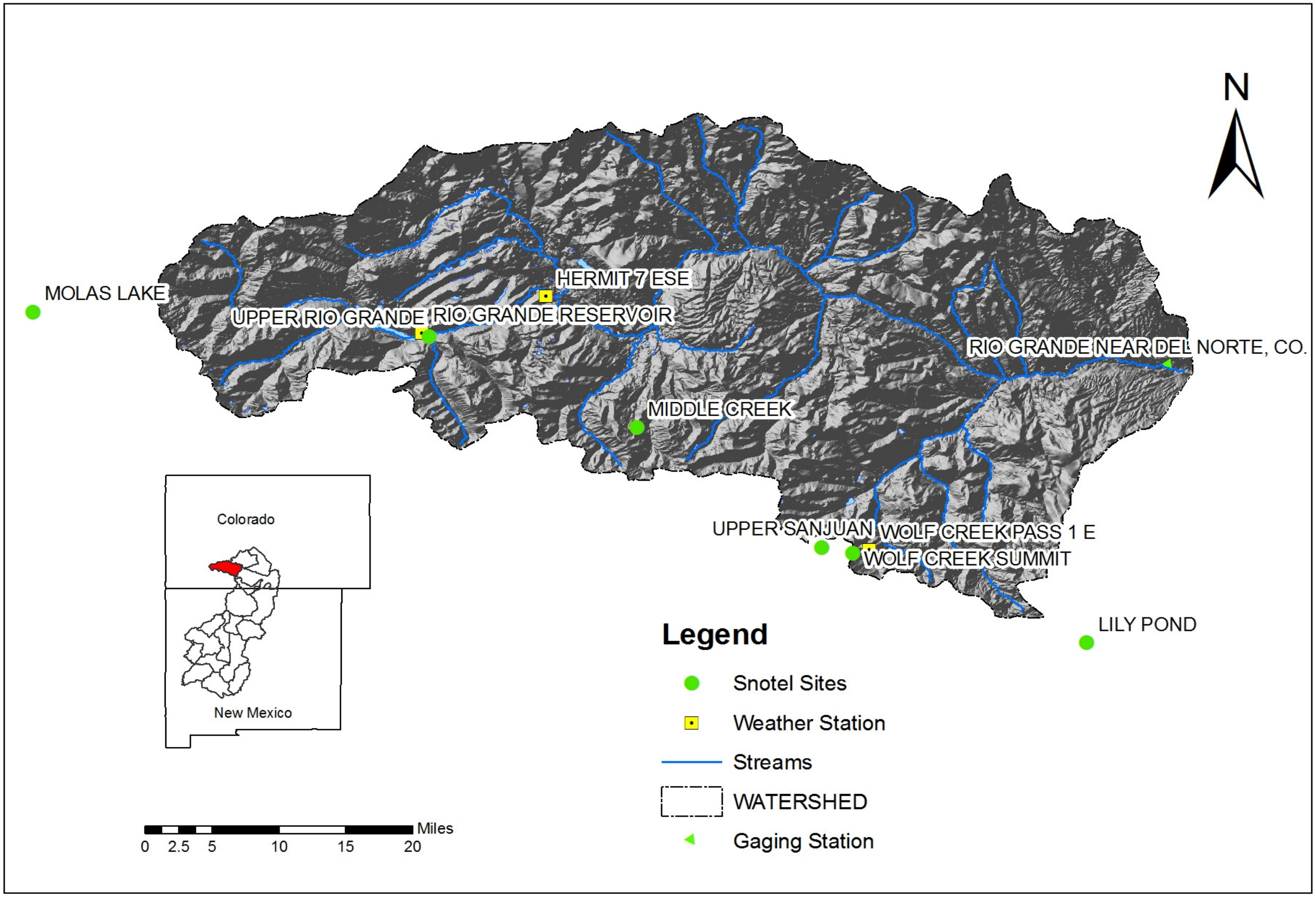

Figure 1.1 Rio Grande Headwaters Subbasin above Del Norte Gaging Station.

1.4.1.2 Streamflow Data

The monthly natural flow data of Del Norte Gaging Station was obtained from NRCS and was used as the dependent variable in the modeling. The data period from 1961 to 2007 was used in this study because of the natural flow and snow water equivalent data availability in the Rio Grande Headwater Basin (the natural flow data for 2006 and 2007 are provisional data). A monthly streamflow time series, monthly averages and standard deviation for the period of 1961-2007 are plotted in Figure 1.2 and Figure 1.3. As can be seen, the monthly flow has strong seasonality of order 12. May-June flow accounts for more than half of the annual runoff, while the April-September runoff accounts for almost 90% of the annual total runoff. This illustrates that the Basin is snow-dominated, since the large portion of the runoff is contributed by the snowmelt in the Basin. The spring-summer runoff months have the highest standard deviations making the need for developing more accurate forecasting models.

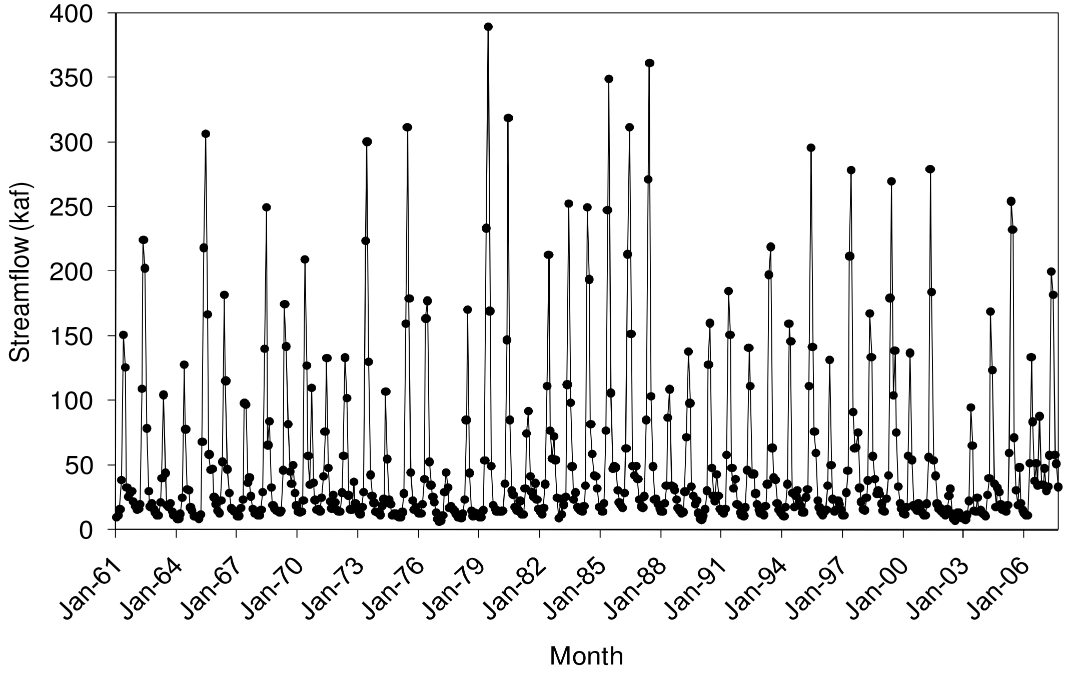

Figure 1.2 Monthly natural flow time series at Del Norte Gaging Station, Rio Grande (1961-2007)

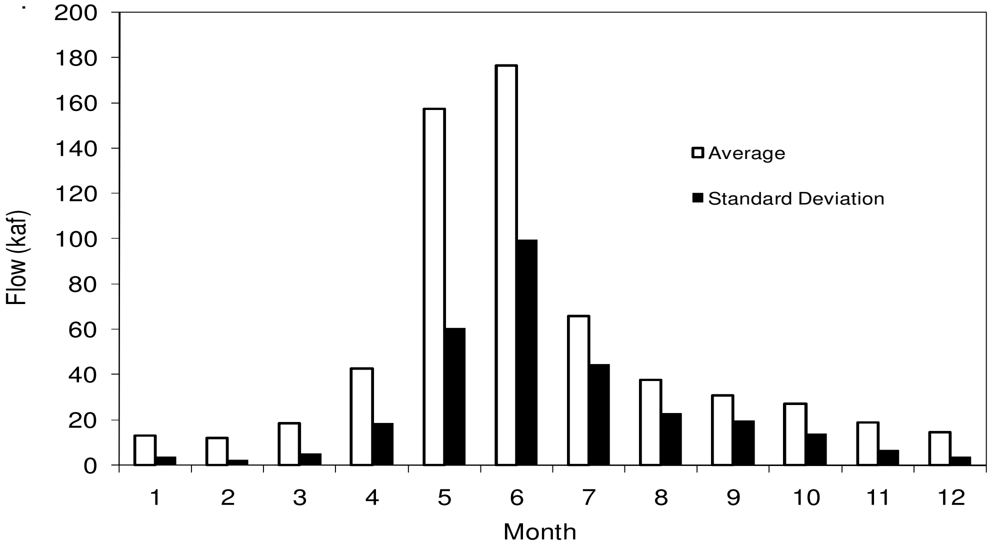

Figure 1.3 Averages and standard deviation of monthly natural flow of Del Norte Gaging Station, Rio Grande (average of 1961-2007)

For modeling of spring-summer runoff volume, the April-September natural seasonal streamflow volume at Del Norte Gaging Station, Rio Grande was selected as the forecast target volume so as to be consistent with the NRCS forecast dates and volume at the Station and to compare the modeling results with the NRCS median forecasts (the 50% percent exceedance probability) for the same period natural runoff volume. The NRCS provides seasonal streamflow volume forecasts at the same Station on the first day of January, February, March, April, May and June. The forecasts consist of the volumes corresponding to each of the 10%, 30%, 50%, 70% and 90% exceedance probabilities.

The April-September seasonal volume is calculated as the sum of monthly natural flow volume from April to September of a year. The data period from 1981 to 2007 was used for the seasonal runoff volume modeling in this study because of the natural flow, real time snow water equivalent, precipitation and temperature data availability from automatic SNOTEL sites in the Basin. To be consistent with the NRCS forecast dates and volume, the April-September, May-September and June-September volumes were used as the dependent variables in the modeling.

1.4.1.3 Snow Water Equivalent

The snow water equivalent data were obtained from The Natural Resources Conservation Service (NRCS) website: http://www.wcc.nrcs.usda.gov/snow/. The NRCS maintains numerous automatic SNOTEL sites in the Basin. Although the Molas Lake and Lily Pond SNOTEL sites are physically outside of the Basin, they were still included in the calculation of the study due to their relationship with streamflow at the Del Norte Gaging Station and data availability. To ensure the operational ability of the forecasting models, only the snow water equivalent (SWE) data from Natural Resources Conservation Center (NRCS) automatic SNOTEL sites were used in the study. The Beartown SNOTEL site was not used in the analysis because of the limited data availability. The detailed information of these SNOTEL sites is shown in Table 1.1 and Figure 1.1.

Table 1.1 SNOTEL sites used in the study of Rio Grande Headwaters Basin above Del Norte Gaging Station

Snotel sites	Location		Elevation (feet)	Available data period		
	West Longitude	North Latitude		SWE	PRCP	TEMP
LILY POND	106.55	37.38	11000	49-present	81-present	83-present
MIDDLE CREEK	107.03	37.62	11250	79-present	81-present	83-present
MOLAS LAKE	107.69	37.75	10500	51-present	87-present	83-present
UPPER RIO GRANDE	107.26	37.72	9400	61-present	87-present	83-present
UPPER SANJUAN	106.84	37.49	10200	36-present	79-present	83-present
WOLF CREEK SUMMIT	106.80	37.48	11000	61-present	87-present	83-present

The Basin average snow water equivalent index was used in the time series and neural networks modeling of monthly and defined seasonal flows. The six SNOTEL sites located in the Basin (as shown in the Table 1.1) were used for calculating the average SWE index of the Rio Grande Headwater Subbasin above the Del Norte Gaging Station. The simple average was used calculate the average SWE index. The Middle Creek SWEs were not included in the averaging before 1979, because no records or estimated values were available at this site for that period. The time series of the average SWE index for the Basin was plotted in Figure 1.4.

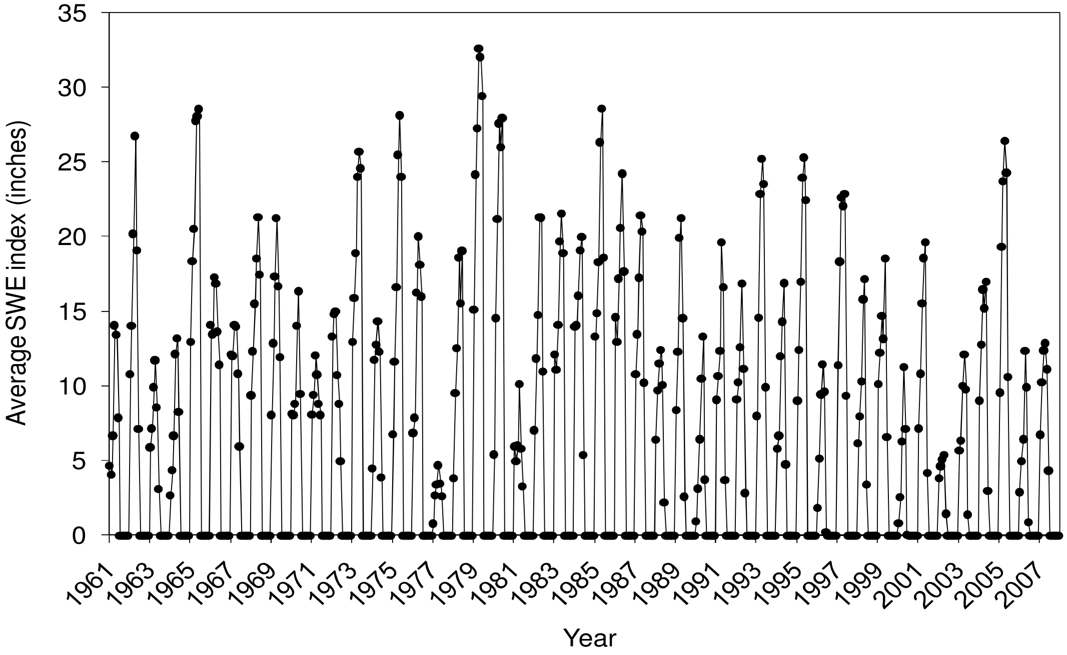

Figure 1.4 Average snow water equivalent index in Rio Grande Basin above Del Norte Gaging Station (1961-2007)

1.4.1.4 SNOTEL Precipitation

The precipitation measured at the SNOTEL sites selected was used for analysis. This is also due to operational forecasting practices, since weather station data from other climate networks are not readily available on the first day of a month (Pagano, personal communication, September 20, 2007). However, the real-time precipitation data of the SNOTEL sites have a short record of measurement (some SNOTEL sites started only after 1987, as shown in Table 1.1). To ensure the same data covering period with other hydrologic variables, two precipitation data preparation methods were employed in the study:

1) For the April-September seasonal flow regression equation development, the precipitation of some single sites was extended back to 1981 using linear regression with precipitation of nearest weather stations. As indicated in the Table 1.1, the Molas Lake, Upper Rio Grande and Wolf Creek Summit SNOTEL sites were needed to be extended from 1987 back to 1981.
2) For time series and neural networks modeling of monthly and defined seasonal flows, a basin average precipitation index was developed for the period covering 1961 to 2007 using linear regression with average precipitation of the weather stations located near to the SNOTEL sites used in the study.

The following method was used to extend the average SNOTEL precipitation back to 1961:

1) Locate weather stations that had a longer precipitation record (beyond 1961) and were located within a 15 mile buffer of the SNOTEL sites;
2) Calculate the simple average precipitation of the SNOTEL sites using the data period of 1981-2007. Some sites that did not have records back to 1981 were excluded from these calculations;
3) Calculate the average precipitation measured in the selected weather stations (as shown in Table 1.2) for the period of 1961-2007;

4) Develop a linear regression equation between monthly average precipitation of selected weather stations and average SNOTEL precipitation index;

5) Extend the SNOTEL precipitation data through the use of the regression equation forced with the weather station precipitation data from 1961-1979.

The resulting extended average SNOTEL precipitation index covering the period of 1961-2007 is shown in Figure 1.5. The SNOTEL precipitation data were obtained from NRCS website: http://www.wcc.nrcs.usda.gov/snow/. The weather stations used for extending the precipitation data were shown in Table 1.2 and Figure 1.1. The monthly precipitation data of weather stations were obtained from the Western Region Climate Center (WRCC) website: http://www.wrcc.dri.edu/.

Table 1.2 Weather stations use in the study of Rio Grande Headwaters Basin above Del Norte Gaging Station

Station Name	North Latitude	West Longitude	COOP	NWS	State	County	Elevation(ft)
HERMIT 7 ESE	37.77	107.13	53951	HERC2	CO	MINERAL	9000
RIO GRANDE RESERVOIR	37.73	107.27	57050	CRRC2	CO	HINSDALE	9455
WOLF CREEK PASS 1 E	37.48	106.78	59181	SFKC2	CO	MINERAL	10640

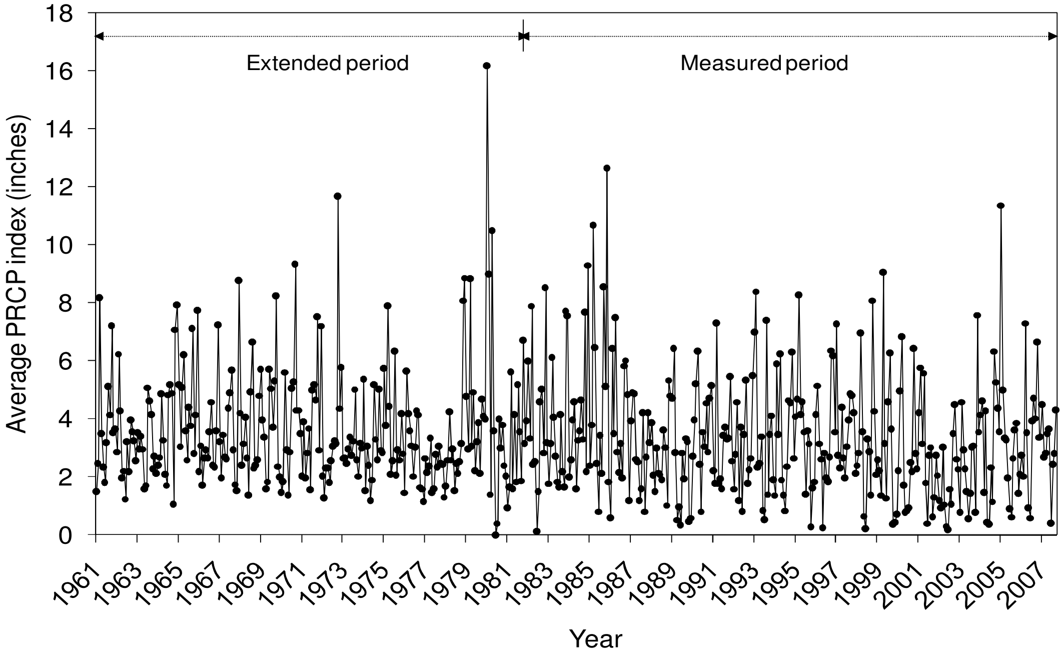

Figure 1.5 Extended monthly average SNOTEL precipitation index in Rio Grande Basin above Del Norte Gaging Station (1961-2007)

1.4.1.5 SNOTEL Temperature

The real time measured temperature data of the SNOTEL sites started from 1983. Some sites even started measuring temperature data from 1987 in the Basin. The interannual variability of temperature data displays strong regional coherence and therefore a basin-average temperature index was calculated from the individual station data. The basin average SNOTEL temperature index was calculated as follows:

1) Calculate the average monthly temperature of six SNOTEL sites (as shown in Table 1.1) to obtain an SNOTEL average temperature index for the period of 1983-2007;

2) Calculate the average temperature measured in the weather stations located near to the SNOTEL sites (as shown Table 1.2) in the Basin for the period of 1981-2007;
3) Develop a regression equation between monthly average temperature of selected weather stations and SNOTEL average temperature index;
4) Extend the SNOTEL average temperature index through the use of the regression equation forced with the weather station average temperature from 1981-1983.

The resulting extended monthly SNOTEL average temperature index covering the period of 1980-2007 is shown in Figure 1.6. The SNOTEL temperature data were obtained from NRCS website: http://www.wcc.nrcs.usda.gov/snow/. The monthly average temperature data of weather stations were obtained from the Western Region Climate Center (WRCC) website: http://www.wrcc.dri.edu/.

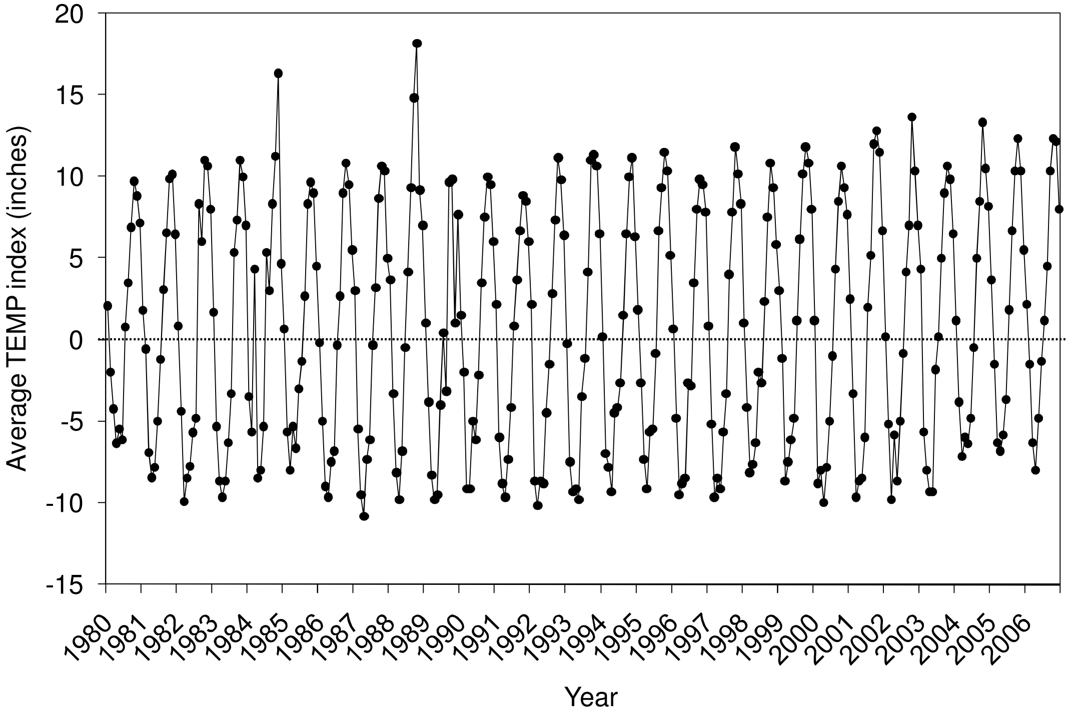

Figure 1.6 Extended monthly average SNOTEL temperature index in Rio Grande Basin above Del Norte Gaging Station

1.4.2 Rio Grande Basin above Elephant Butte Reservoir

Elephant Butte Reservoir net inflow is an important hydrologic variable that is used for determining water delivery from New Mexico to Texas according to Rio Grande Compact (Rio Grande Compact Commission, 2006). In this study, the monthly and seasonal net inflow of Elephant Butte Reservoir, Rio Grande, New Mexico was selected for modeling using different data-driven modeling approaches aiming at providing a decision basis for the operation of Elephant Butte Reservoir for water management and compact compliance purposes in the region.

Net inflow into a reservoir can be defined as the metered and unmetered surface water flows entering the reservoir, direct precipitation, and groundwater exfiltration entering the reservoir minus releases from the reservoir, evaporation, seepage, and any other reservoir losses. It is computed as the sum of the change in storage volume within the reservoir and the volume of all measured flow releases and spills for a given period. Although the reservoir net inflows are usually difficult to forecast due to heavy upstream regulation and human development effects, reservoir net inflow forecasting is of great importance to reservoir management as it is an indication of water availability from a reservoir.

1.4.2.1 Site Description

The study area, shown in Figure 1.7, is the Rio Grande watershed in Colorado and New Mexico upstream from Elephant Butte Dam, New Mexico. The primary reservoir inflow from the Rio Grande is metered at San Marcial, New Mexico at the upstream end of the reservoir. Two US Geological Survey gaging stations are used to quantify the flow at San Marcial: the Rio Grande Floodway at San Marcial, NM (USGS Station no. 08358400), which is the river channel, and the Rio Grande Conveyance Channel at San Marcial, NM (USGS Station no. 08358300), a constructed channel that parallels the Rio Grande. The sum of these two gaging stations is the total flow at San Marcial used in this study.   The gaging station data at the Rio Grande below Elephant Butte Dam, NM (USGS Station no. 08361000) were used as the reservoir outflow. All the water released from Elephant Butte Reservoir including irrigation use, hydropower generation, and reservoir spill is measured at this site. For both San Marcial and below Elephant Butte, the historical daily average flow rate in cfs were retrieved and then converted into monthly total flow in acre-ft. The Elephant Butte storage data were obtained from U.S. Bureau of Reclamation as daily storage volume in acre-ft, and then the monthly change in storage was calculated as the difference of volume stored at the end and the beginning of each month. Stage-Storage relationships for Elephant Butte reservoir are developed from bathymetric surveys that are updated roughly every ten years.

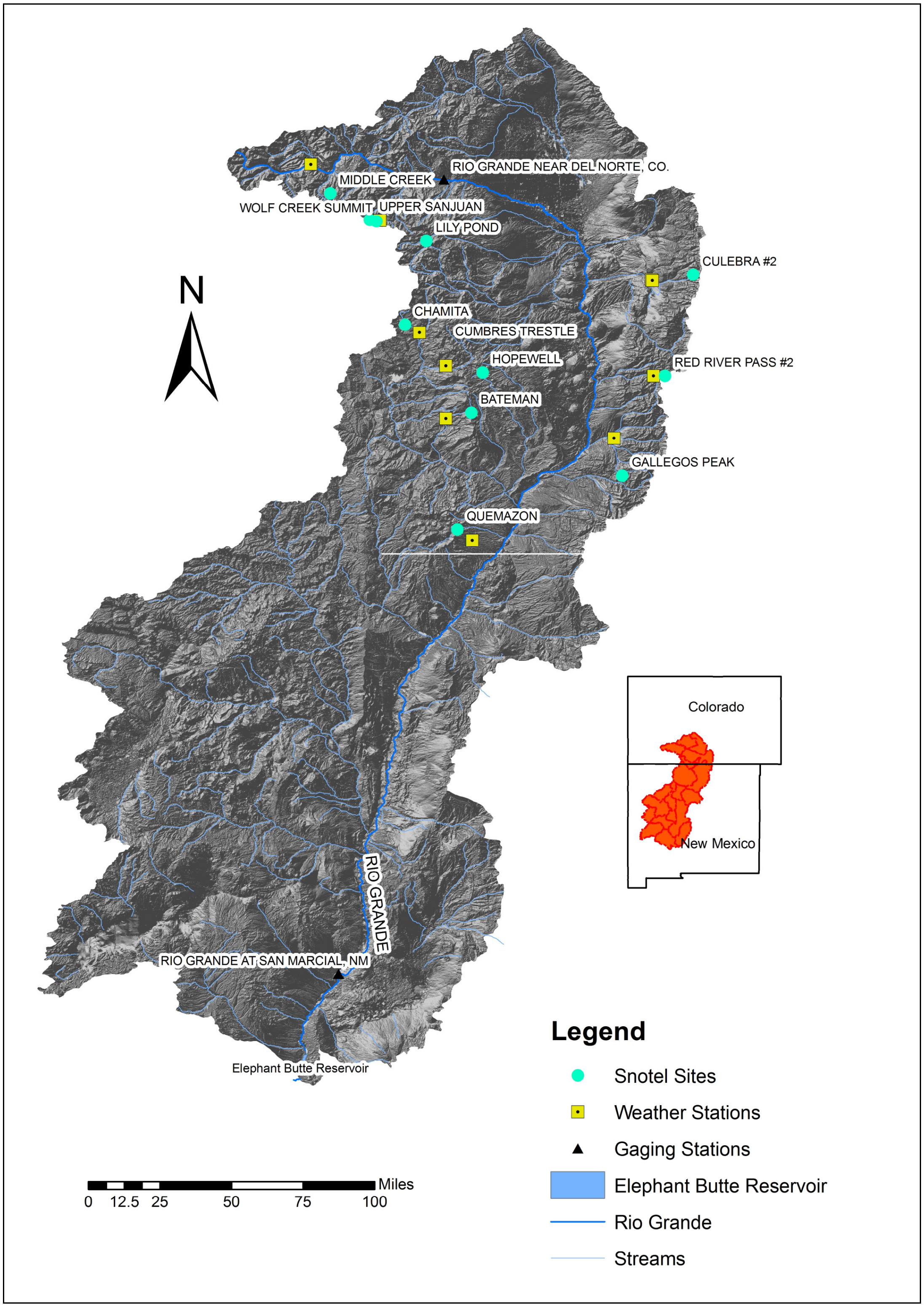

Figure 1.7 Rio Grande Basin above Elephant Butte Reservoir

1.4.2.2 Reservoir Net Inflow

The monthly Elephant Butte Reservoir net inflow was calculated as the sum of monthly releases measured below Elephant Butte Dam and the monthly change in storage of the reservoir. The calculated monthly net inflow time series for the period of 1961- 2007 is shown in Figure 1.8. Unlike the Del Norte monthly natural flow, the Elephant Butte monthly net inflow is heavily regulated and has been highly variable over the years. The standard deviations of monthly net inflow (as shown in Figure 1.9) also suggest the highly variable features of the net inflow, particularly in the spring-summer season, starting from April through October. The summer net inflows of July, August, September and October are of particularly high variability.

The Figure 1.9 also shows the comparison of the average monthly flow at the San Marcial Gaging Station and the average Elephant Butte Reservoir monthly net inflow. In all months, the average Elephant Butte Reservoir net inflow is smaller than the average flow that is measured at the San Marcial Gaging Station, but the two are highly correlated (correlation coefficients of 0.98). This indicated that the main contribution to the Elephant Butte Reservoir net inflow comes from the Rio Grande streamflow although there are other factors, such as reservoir evaporation, seepage and infiltration, unmetered tributaries, and operational practices that affect reservoir net inflow. The data period from 1961 to 2007 was used in this study for developing monthly and defined seasonal net inflow time series and neural networks models so as to keep the same time period of snow water equivalent data available from SNOTEL sites in the Basin.

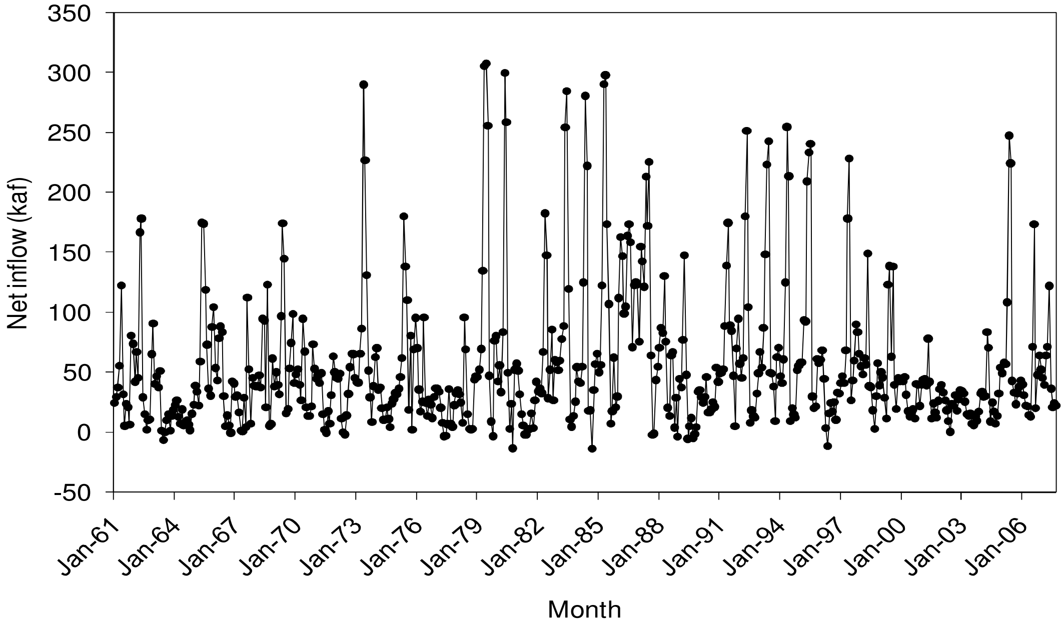

Figure 1.8 Monthly net inflow time series Elephant Butte Reservoir, Rio Grande (1961-2007)

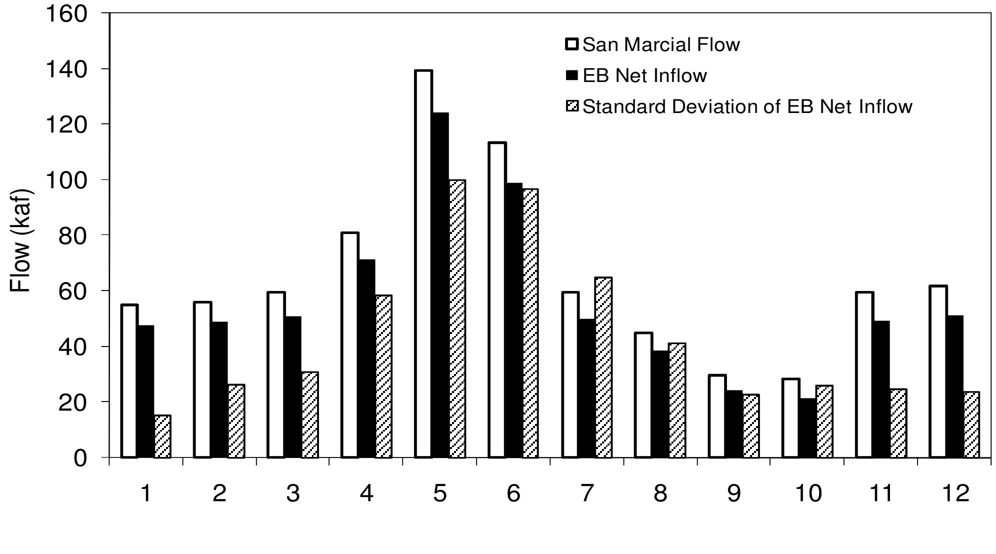

Figure 1.9 Comparison of monthly average San Marcial measured flow and Elephant Butte Reservoir net inflow (average of 1961-2007)

For the modeling of spring-summer seasonal net inflow volume, the March-July seasonal net inflow volume of Elephant Butte Reservoir was selected as the forecast target volume. The March-July net inflow was calculated as the sum of monthly net inflow from March to July. The data period from 1981 to 2007 was used to develop forecast regression equation in the Basin because of the net inflow flow and real time snow water equivalent, precipitation and temperature data availability in the Basin.

The NRCS does not provide seasonal net inflow volume forecasts at Elephant Butte Reservoir, but does provide March-July San Marcial natural flow volume forecasts at San Marcial Gaging Station on the first day of January, February, March, April, and May. The forecasts consist of the volume corresponding to each of the 10%, 30%, 50%, 70% and 90% exceedance probabilities. To be consistent with the NRCS forecast dates and volume, the March-July, April-July and May-July volumes were used as the dependent variables in March-July seasonal volume modeling in the study.

1.4.2.3 Snow Water Equivalent

The snow water equivalent data from the NRCS automatic SNOTEL sites were used due to their readily availability on the first day of a month. Eighteen SNOTEL sites that are located in the Rio Grande Basin above Elephant Butte Dam were initially selected for analysis. To use the SNOTEL sites that are spatially representative for the Basin and highly correlated with Elephant Butte Reservoir net inflow, the following procedure was carried out to select the final SNOTEL sites for use in the study:

1) Calculate the correlation coefficients of April 1st SWE of eighteen SNOTEL sites with Elephant Butte Reservoir April-July seasonal net inflow;
2) Select the SNOTEL sites that have correlation coefficients above 0.7 as the first group candidate sites;
3) Select several more SNOTEL sites that represent different regions of the basin even if the correlation coefficients are little lower.

As a result, twelve SNOTEL sites were selected for this study (Figure 1.7 and Table 1.3). The correlation analysis showed that the SNOTEL sites located in the Rio Grande west region in Colorado and New Mexico have a stronger correlation than the SNOTEL sites located in the Rio Grande east region in New Mexico. Three SNOTEL sites in the Rio Grande east region in New Mexico (Culebra #2, Gallegos Peak and Red River Pass #2) still were selected in the study considering the spatial representation of the whole basin.

Table 1.3 SNOTEL sites used in the study of Rio Grande Basin above Elephant Butte Reservoir, New Mexico.

Snotel sites	Location			Elevation (feet)	Available data period	
	West Longitude	North Latitude	Region		SWE	PRCP
BATEMAN	106.32	36.51	Rio Grande west in New Mexico	9300	61-present	80-present
CHAMITA	106.66	36.96	Rio Grande west in New Mexico	8400	61-present	80-present
CULEBRA #2	105.20	37.21	Rio Grande east in New Mexico	10500	61-present	80-present
CUMBRES TRESTLE	106.45	37.02	Rio Grande west in Colorado	10040	61-present	81-present
GALLEGOS PEAK	105.56	36.19	Rio Grande east in New Mexico	9800	78-present	81-present
HOPEWELL	106.26	36.72	Rio Grande west in New Mexico	10000	72-present	80-present
LILY POND	106.55	37.38	Rio Grande west in Colorado	11000	49-present	81-present
MIDDLE CREEK	107.03	37.62	Rio Grande west in Colorado	11250	79-present	81-present
QUEMAZON	106.39	35.92	Rio Grande west in New Mexico	9500	61-present	81-present
RED RIVER PASS #2	105.34	36.70	Rio Grande east in New Mexico	9850	61-present	80-present
UPPER SANJUAN	106.84	37.49	Rio Grande west in Colorado	10200	36-present	79-present
WOLF CREEK SUMMIT	106.80	37.48	Rio Grande west in Colorado	11000	61-present	87-present

For time series and neural networks modeling of monthly and defined seasonal flows, the basin average snow water equivalent index was used. The selected twelve SNOTEL sites (as shown in Table 1.3) were used for calculation of the average SWE index of the Rio Grande Basin above Elephant Butte Reservoir. The calculated basin average SWE index that covering the period of 1961-2007 is shown in Figure 1.10.

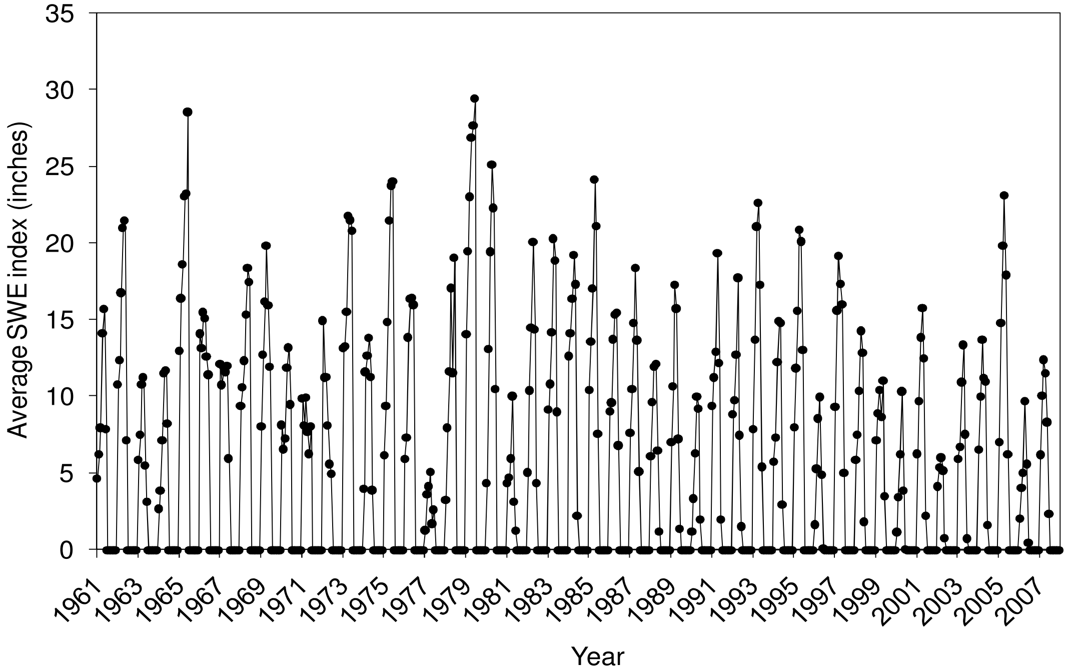

Figure 1.10 Average snow water equivalent index of Rio Grande Basin above Elephant Butte Reservoir (1961-2007)

1.4.2.4 SNOTEL Precipitation

The precipitation data from the same SNOTEL sites selected were also used for analysis due to operational forecasting practices. For March-July seasonal flow regression equation development, the precipitation at some single sites was extended back to 1981 using linear regression with precipitation at the nearest weather stations (as shown in the Table 1.4). Only the Wolf Creek Summit SNOTEL site needed to be extended from 1987 back to 1981. The other SNOTEL sites have precipitation measurement records back to 1981 (as shown in Table 1.3).

For time series and neural networks modeling of monthly and defined seasonal flows, a basin average precipitation index was calculated for the covering period of 1961 to 2007 by applying the same approach that was used in section 1.4.1.4. The extended average monthly SNOTEL precipitation index that covers the period 1961-2007 is shown in Figure 1.11. The SNOTEL precipitation data were obtained from NRCS website: http://www.wcc.nrcs.usda.gov/snow/. The weather stations used for extending the precipitation data were shown in Table 1.4 and Figure 1.7. The monthly precipitation data of weather stations were obtained from the Western Region Climate Center (WRCC) website: http://www.wrcc.dri.edu/.

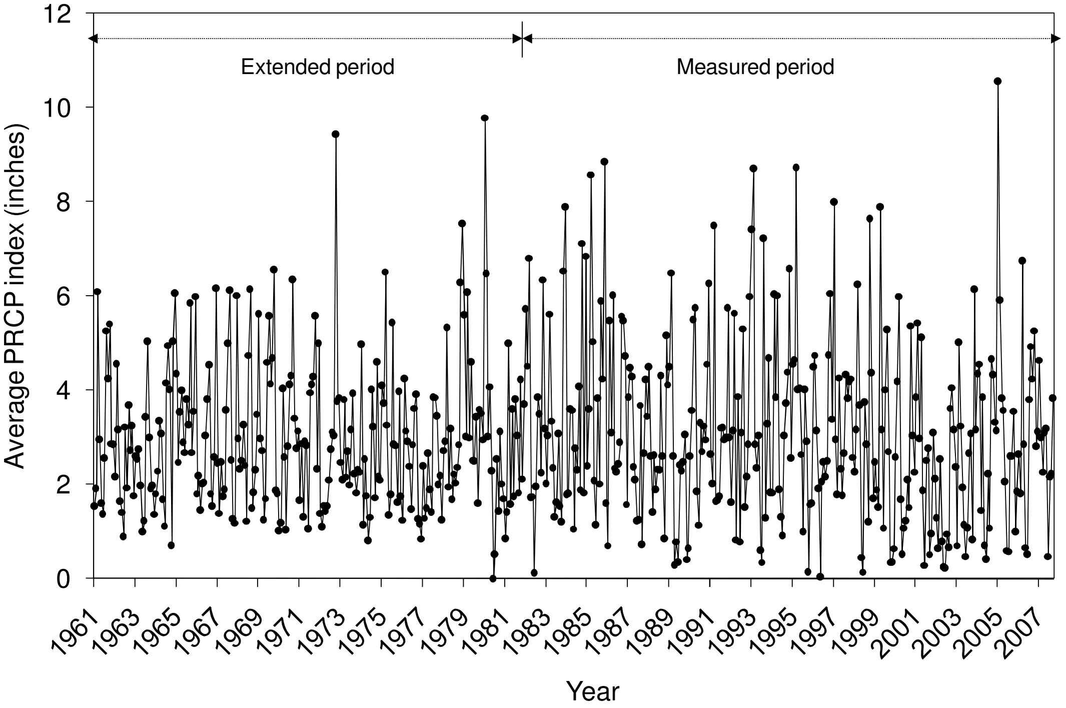

Figure 1.11 Extended monthly average SNOTEL precipitation index of Rio Grande Basin above Elephant Butte Reservoir (1961-2007)

Table 1.4 Weather stations used in the study of Rio Grande Basin above Elephant Butte Reservoir

Station Name	North Latitude	West Longitude	COOP	NWS	State	County	Elevation(ft)
HERMIT 7 ESE	37.77	107.13	53951	HERC2	CO	MINERAL	9000
SAN LUIS 2 SE	37.18	105.41	57430	SLSC2	CO	COSTILLA	8033
SAN LUIS 3 SE	37.18	105.41	57428	SANC2	CO	COSTILLA	8017
WOLF CREEK PASS 1 E	37.48	106.78	59181	SFKC2	CO	MINERAL	10640
BRAZOS LODGE	36.75	106.45	291180	CMAN5	NM	RIO ARRIBA	8009
CANJILON R S	36.48	106.45	291389	CJLN5	NM	RIO ARRIBA	7828
CHAMA	36.92	106.58	291664	CHMN5	NM	RIO ARRIBA	7850
LOS ALAMOS	35.87	106.32	295084	LOAN5	NM	LOS ALAMOS	7424
RED RIVER	36.70	105.40	297323	REDN5	NM	TAOS	8676
TAOS	36.38	105.60	298668	E23	NM	TAOS	6965

1.4.2.5 SNOTEL Temperature

As described in the previous sections, only March-July seasonal net inflow modeling included the monthly basin average SNOTEL temperature index as one of the potential input variables in the regression equation development. The monthly basin average temperature index was calculated using the same approach described in section 1.4.1.5.The resulting monthly extended SNOTEL average temperature index covering the period of 1981-2007 is of similar magnitude and pattern as the average monthly SNOTEL temperature index for the Rio Grande Headwaters Basin above Del Norte Gaging Station. The weather stations used for extending the data are shown in Table 1.4 and Figure 1.7. The SNOTEL temperature data were obtained from NRCS website: http://www.wcc.nrcs.usda.gov/snow/. The monthly average temperature data of the weather stations were obtained from the Western Region Climate Center (WRCC) website: http://www.wrcc.dri.edu/.

1.4.3 El Niño-Southern Oscillation Index

To analyze the correlations between large-scale climate indices and Rio Grande streamflow, the monthly Southern Oscillation Index (SOI) data from 1961 to 2007 were used in the study. The El Niño-Southern Oscillation (ENSO) phenomenon is associated with anomalous sea level pressure, surface winds and sea surface temperature near the equatorial Pacific. The signature of an ENSO event is in the sea-level pressure gradient currently measured between Darwin, Australia and Tahiti. This gradient is the primary variable used to measure the magnitude of an ENSO event in the form of the Southern Oscillation Index (SOI). As a measure of the state of the ENSO, the SOI is computed as the normalized difference in standardized sea-level pressure anomalies between Tahiti and Darwin relative to its root mean square. The monthly standardized SOI data used in this study was retrieved from the Climate Prediction Center (http://www.cpc.ncep.noaa.gov/data/indices/soi) and is shown in Figure 1.12.

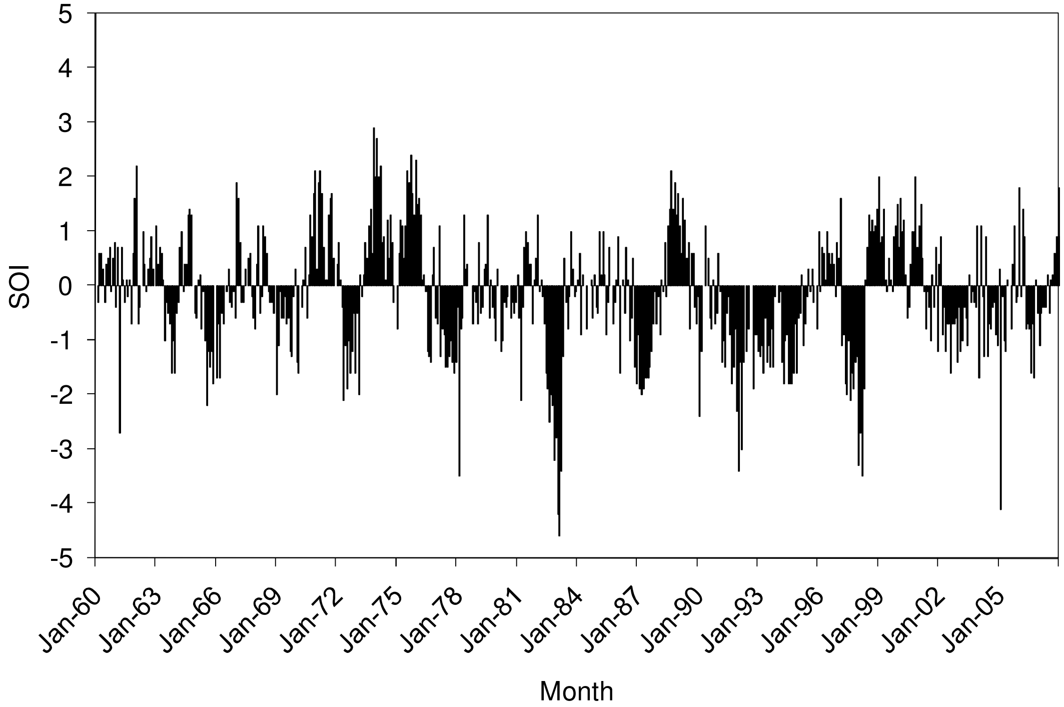

Figure 1.12 Monthly Southern Oscillation Index (SOI) for the period 1960-2007

According to the criteria proposed by Redmond and Koch (1991) and Cayan et al. (1999), when the previous calendar year averaged SOI from June to November is -0.5 or less, then the present water year is designated as El Niño . If it is greater than or equal to +0.5, then present water year is designated as La Nina. If it is between them, then present water year is neutral. Based on this criteria, the data period years used in the study, including calibration years (1961-1999) and forecasting years (2000-2007) were categorized as El Niño years, La Nina years and neutral years (as shown in Figure 1.13).

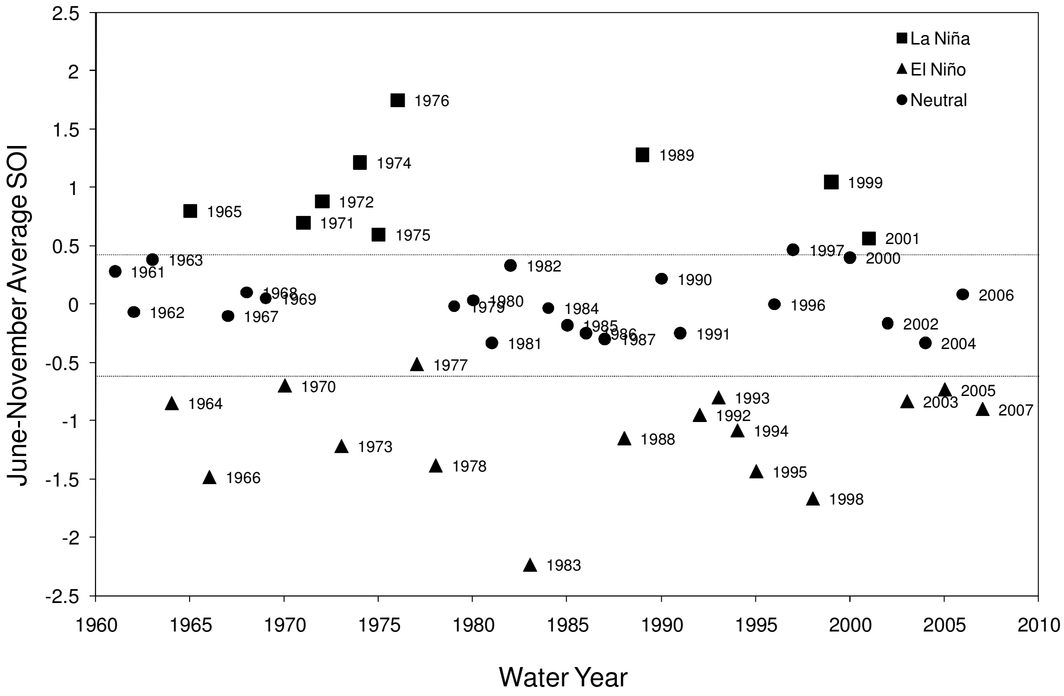

Figure 1.13 Designation of ENSO phases for the data period (1961-2007) used in the study (Adapted from Lee, 2004)

1.5 Organization of the Dissertation

This dissertation is composed of six chapters. The general organization of the dissertation and the focus of each chapter can be described as follows:

Chapter 1 introduces current issues existing in seasonal and monthly streamflow forecasting, and covers the scope and objectives of the study. The study sites, data used and some data preparation procedures used in the study are also presented in this chapter.

Chapter 2 conducts a comprehensive literature review regarding seasonal and monthly streamflow forecasting, and current issues concerning the application of hybrid modeling methods in streamflow forecasting. Methodologies that are used in the study, including algorithms of partial least squares regression (PLSR), principal components regression (PCR), autoregressive integrated moving average (ARIMA) models, transfer function-noise (TFN) models, artificial neural networks (ANN) are described in detail. In addition, the possible categorization of hybrid modeling methods and the hybrid models that are used in the study are also presented in this chapter.

Chapter 3 investigates the application of partial least squares regression (PLSR) in seasonal streamflow forecasting. The chapter focuses on the development of PLSR and principal components regression (PCR) models for seasonal streamflow volume forecasts using snow water equivalent, precipitation, temperature, and previous flow conditions as input variables. The selection of an optimal number of components using the jackknife cross validation scheme and variable selection approach using PLSR are discussed in detail. The performance of the PLSR and PCR models in seasonal streamflow forecasting are compared to each other and to NRCS official forecasts. The final regression equations and conclusions are presented at the end of the chapter.

Chapter 4 proposes an application of hybrid modeling approaches in seasonal streamflow forecasting. Two hybrid modeling approaches, a forecast modification using a combination of transfer function-noise (TFN) model with artificial neural networks (ANN), and the combination of principal components analysis (PCA) with ANN, are investigated for the purpose of improving seasonal streamflow forecasts.

To perform time series modeling of seasonal flow, different seasons are defined for the two basins used in the study. The forecast performances of two hybrid modeling approaches are compared to the different single modeling techniques such as ARIMA, TFN and ANN. Finally, some general discussions and conclusions are summarized at the end of the chapter.

Chapter 5 investigates the response of monthly streamflow processes to basin precipitation, snow water equivalent, El Niño Southern Oscillation (ENSO) using cross correlation analysis. Several statistical models including ARIMA, TFN, and ANN were built for monthly streamflows in the study sites. Then, one-month-ahead forecasts of those models for spring-summer season were modified using snow water equivalents and ENSO signals using ANN technique. The performances of different modeling approaches are compared with each other and some general discussion and conclusions are presented at the end of the chapter.

Chapter 6 summarizes the results of the previous chapters. The capabilities and limitations of different modeling methods are discussed in both monthly and seasonal time scales. Some suggestions and recommendations for future research work are also proposed.

Relevant figures and tables are included and are numbered sequentially within each chapter. The SAS codes used for PLSR model development including model calibration, jackknife cross validation, and one-step- ahead rolling forward forecasting are also included in the Appendices that are located at the end of the dissertation.

2 LITERATURE REVIEW

2.1 General Classes of Forecasting Models

Streamflow forecasting is of great importance to water resources management and planning. Particularly, the long-range forecasting such as monthly, seasonal, or annual time scales is very useful in reservoir operations and irrigation management decisions such as scheduling releases, allocating water to downstream users, and managing river treaties or Compact compliances. Due to their importance, a large number of forecasting models have been developed and applied in the streamflow forecasting practices for the last several decades.

Streamflow forecasting models may fall into two categories in general: process-driven methods and data-driven methods (Wang, 2006). The process-driven modeling is a knowledge-driven modeling method that tries to explain the underlying physical processes of the watershed system. The low flow recession models, conceptual rainfall-runoff models, and snowmelt-runoff models are in this category. Data-driven models, on the contrary, are based on a limited knowledge of the internal physical mechanism of the watershed system and rely on the data describing input and output characteristics. They are essentially black-box models that characterize the relationships between inputs and outputs without detailed consideration of the details or explicit simulation of the underlying physical process. They may include regression models, time series models, artificial neural networks (ANN) models and non-parametric models such as K-nearest neighbor method. Recently, the application of some hybrid models, which combine the features of different type models have been reported in literature (e.g., Abrahart and See, 2002; Jain and Kumar, 2007; Karamouz and Zahraie, 2004; Kişi, 2008; Srinivas and Sirinivasan, 2001; Wang et al., 2005b).

Recently, data-driven modeling has become quite popular in streamflow forecasting due to the increase in data availability from metering stations and real-time data retrieval. Increasing computational power and sophistication of modeling theory and software have also fueled the popularity of data-driven modeling. Data-driven models have the added advantage of predicting the relationship between inputs and outputs without requiring detailed conceptualization of the extremely complex and often poorly understood physical processes which, in reality, cause the input-output behavior (Wang, 2006).

Among the various data-driven models in streamflow forecasting, the time series models, including different types of autoregressive integrated moving average (ARIMA) models and transfer function-noise (TFN) models, have been widely used in the last few decades. Recently developed artificial neural networks (ANN) models and non-parametric models such as modified K-nearest neighbor and kernel density estimator, have been applied to streamflow forecasting due to their ability to extract nonlinear relationships without any prior assumptions (Wang, 2006). Multivariate regression has been used in the seasonal and annual streamflow forecasting because of its ability to deal with collinearity more effectively than multiple linear regression. Among various forms of multivariate regression methods, the principal components regression (PCR) is the most frequently used method because of its simplicity and systematic way of developing the regression equation. Recently, the partial least squares regression (PLSR) has been introduced into hydrologic forecasting after it had been used in chemometrics for several decades (Tootle et al., 2007; Wold, 1966).

The application of various data-driven models in streamflow forecasting is largely dependent upon the time scale of the dependent variable. The regression methods are used mostly for the larger time scales; such as seasonal or annual, since the longer time scale hydrologic variables are usually characterized by linear relationship tendency between predictors and dependent variables. In addition, no significant autocorrelation may be detected in large time scales such as annual streamflows and spring-summer streamflows in consequent years. The hydrologic variables with shorter time scales, such as monthly, daily and hourly usually show a strong nonlinear relationship between input and output. Therefore, they are usually modeled using ANN models, non-parametric models or time series models through appropriate transformations. In the next sections of this chapter, the streamflow forecast modeling in two time scales including seasonal and monthly will be discussed and analyzed.

2.2 Seasonal Flow Forecasts

2.2.1 Current Practices in Western United States

Forecasting seasonal volume of river flow is important for making decisions related to economic management, flood mitigation, and environmental consideration of the water resources system. Seasonal volumetric streamflow represents an important hydrologic parameter for water supply purposes. Hence, several agencies, namely the National Weather Service (NWS) River Forecast Centers (RFCs), the National Water and Climate Center (NWCC) of the United States Department of Agriculture (USDA), Natural Resources Conservation Service (NRCS), Bureau of Reclamation (BOR) and some local cooperating agencies, such as the Salt River Project (SRP) in Arizona, issue seasonal streamflow forecasts at various forecasting points for numerous rivers in the continental United States (Pagano et al., 2004).

In the Western United States, the *Water Supply Outlooks* (WSOs) are issued jointly by the NWS River Forecast Centers (RFCs) and Natural Resources Conservation Service (NRCS). These forecasts are available in print “Basin Outlook Report” publications or on the Internet at http://www.wcc.nrcs.usda.gov/wsf. The *Water Supply Outlooks* have been used by water managers for almost 70 years. They are critical components in effective water management and are utilized by a broad spectrum of users for a variety of purposes, ranging from irrigated agriculture, flood control, municipal water supply, endangered species protection, power generation and recreation (Pagano, 2005; Pagano et al., 2004).

The primary operational method of seasonal streamflow forecasting in the western United States is the regression of seasonal streamflow volume on indicator variables, primarily point observations of snow-water equivalent (Wood and Lettenmaier, 2006). The multiple linear regression was used by NRCS for many years until the early 1990s, when Garen (1992) proposed and facilitated the use of principal components regression in streamflow volume forecasting. Since then, the principal components regression has been the standard methodology used by NRCS. In 2006, NRCS employed an Excel® spreadsheet based water supply forecasting software called Visual Interactive Prediction and Estimation Routines (VIPER) with data retrieval, visualization, and forecast calibration and execution functions. The VIPER supports both principal components regression as well as another method, Z-score regression. Other statistical techniques can also be performed with VIPER, including searching for optimum combinations of independent variables, searching for optimum time periods covered by selected independent variables, and jackknife testing of models (NRCS, 2007).

The NWS and NRCS are two U.S. Federal agencies with primary responsibility for seasonal streamflow forecasting in the western United States. While regression based seasonal streamflow forecasts still form the basis of the NWS and NRCS operational systems, a number of different methods have been developed and tested to improve the robustness of operational forecast methods (Wood and Lettenmaier, 2006). The National Weather Service (NWS) is developing Advanced Hydrologic Prediction System (AHPS) to provide long-lead predictions of peak flows and low flows. The AHPS Ensemble Prediction System (ESP) involves the calibration of a hydrologic simulation model, model initialization using current watershed states, and forcing based on a number of observed historical meteorological traces. The NRCS's NWCC is also actively developing a similar capability, including an advanced spatially distributed hydrologic simulation models (Pagano and Garen, 2006; Pagano, 2005).

2.2.2 Collinearity Issue in Seasonal Flow Forecasts

The predictor variables used in seasonal water supply forecasting are usually highly intercorrelated. For example, the snow water equivalent, precipitation data of different Snow Telemetry (SNOTEL) sites and different months are highly correlated with each other. If a multiple linear regression model were built using these variables in one equation, the model would fit data very well, but would produce worse predictions on the new data. A large number of intercorrelated predictor variables are often referred to as a multicollinearity issue, which causes irrational coefficients in the regression equation and does not provide reliable predictions over time. However, there are several solutions for the problem in seasonal flow forecasting. For instance, 1) the elimination of predictor variables that are associated with irrational coefficients; 2) elimination of predictor variables using stepwise regression method; 3) constructing composite indices that used as predictor variables; 4) using an orthogonal transformation of the correlation matrix to restructure a set of intercorrelated variables into an equal number of uncorrelated variables (Garen, 1992; McCuen, 1985).

The elimination of predictor variables might not provide a better solution because of the heavy reliance on information from very few sites, which may not represent spatially variable information of a basin. Although constructing composite indices can remove the major source of intercorrelations of the predictor variables, these indices are usually determined without consideration of regression, and may not be statistically optimal for forecasting (Garen, 1992). The Z-score regression that is currently used by NRCS could be a better solution for this problem because the weightings used in the Z-score method are calculated based on correlations with the dependent variable. However, the weightings have no knowledge of the intercorrelations among the independent variables (NRCS, 2007; Pagano, personal communication, March 3, 2008).

Transforming the original variables into a number of uncorrelated (orthogonal) variables and then performing a regression on them may be a better solution due to its rigorous way of dealing with multicollinearity problems. Several methods such as principal components regression (PCR) and reduced rank regression (RR) were developed for this purpose and have been applied in various fields. The PCR extracts factors to explain as much predictor variation as possible, but may not be associated with the variations of response variable. In contrast, the RR extracts factors to explain as much response variable variation as possible, but predictions may not be accurate (Tobias, 1995). To balance the two objectives of explaining response variables variation and explaining predictor variation, the partial least squares regression (PLSR) was developed and used extensively in the chemometrics the last few decades and has provided a better solution for multicollinearity problems. Yeniay and Göktaş (2002) compared the performance of the three regression methods using economic data and found that PLSR performed better in terms of predictive ability compare to the other two methods. However, compared to principal components regression, there are limited applications of PLSR and RR in seasonal streamflow volume forecasting. Only recently, has PLSR gained importance in the hydrologic community because of its attractive feature of dealing with highly intercorrelated variables that could result in improved forecast skills compared to other methods.

2.2.3 Possible Improvement of Seasonal Flow Forecasts

As described in section 2.2.1, the principal components regression (PCR) is the main regression method used by the NRCS for seasonal flow forecasts at present. The PCR is typically utilized to account for collinearity issues and has been successfully applied to seasonal streamflow forecasting. Garen (1992) introduced principal components regression and methodology of searching for optimal and near optimal combination of variables to the NRCS forecasting practices. Since then, the principal components regression has become one of the main forecasting techniques at NRCS because substantial improvement in forecast accuracy has been achieved with PCR compared to a multiple linear regression. Eldaw et al. (2003) applied the PCR to the seasonal Nile River streamflow forecasting based on sea surface temperatures (SST) and the previous year of Guinea precipitation. In their study, the PCR streamflow forecast models showed significant improvement over the multiple-regression models for several long lead times.

Z-score regression, a comparatively new, effective method in dealing with collinearity and missing data issues, has been introduced and is being used currently by NRCS together with PCR. The weightings used in the Z-score method are based on correlations with the dependent variable, whereas principal component weightings have no knowledge of the dependent variable. The Z-score method is particularly useful when dealing with sets of independent variables that are not serially complete (i.e., have missing values) or have varying periods of record (NRCS, 2007; Pagano personal communication, March 4, 2008).

Improving the seasonal volume forecasts has always been an important issue for researchers and hydrologic forecasters. As mentioned previously, efforts are under way to develop high-quality spatially distributed hydrological modeling for improved streamflow forecasting using remote sensing (RS), geographic information system (GIS) technology, and data gathered in field experiments. However, the distributed models are not convenient and timely because of the satellite data availability in the operational forecasting environment (Pagano, 2005). Hence, researchers are looking at two main approaches to improve seasonal volume forecasts.

First, the forecasts could be improved if the model contains new input variables that represent antecedent and interannual flow and climate conditions. These variables could be the state of regional groundwater system, streamflows from previous 1 or 2 years, El Niño Southern Oscillation (ENSO) and Pacific Decadal Oscillation (PDO). Many research papers look at the relationship between various climate variables and streamflow processes (e.g., Cayan et al. 1999; Hamlet and

Lettenmaier, 1999; Hsieh et al., 2003, Lee et al., 2004; Pagano and Garen, 2005; Pagano and Garen, 2006; Redmond and Koch, 1991; Tootle and Piechota, 2006).

The second possible alternative to improve forecast models is the application of more robust modeling techniques, such as using principal components regression (PCR), partial least squares regression (PLSR), artificial neural networks (ANN), genetic algorithms, non-parametric methods and hybrid modeling approaches. Some of these methods, such as Z-score regression and PLSR, could improve model forecasts by dealing with multicollinearity issues, while other methods including ANN, genetic algorithm and nonparametric methods, could improve forecast accuracy by modeling nonlinear relationship between input and output variables efficiently without any normality and linearity assumption of the streamflow processes.

Many research results were reported in the literature regarding the issue of improved seasonal and annual streamflow forecasts using more robust approaches. Hsieh et al. (2003) applied multiple linear regressions (MLR) and feed-forward neural network models using principal components of large-scale climatic indices to predict the seasonal volume of Columbia River in British Columbia. Their results showed that the neural network and MLR predictions were essentially identical. This was because the detectable relationships in the small sample size might have been linear. Nonparametric methods, such as kernel density estimator and K-nearest neighbor method have been successfully applied to seasonal streamflow forecasting and simulation. Unlike parametric models, the nonparametric methods are assumption free and no parameter estimation and data transformation are necessary. Piechota and Dracup (1999) applied a kernel density based non-parametric method for long-lead time forecasting of seasonal streamflow. Their results showed that 3- to 7- month lead time forecasts of spring-summer runoff using El Niño Southern Oscillation indicators had better forecast skills than climatology. Successful application of other nonparametric methods including K-nearest neighbor, modified K-nearest neighbor method and hybrid parametric/nonparametric models can be found in the literature (e.g., Grantz, 2003; Lall and Sharma, 1996; Piechota et al., 1998; Prairie et al., 2006; Rajagopalan and Lall, 1999; Shamseldin and O'Connor, 1996; Sharma et al., 1997; Souza and Lall, 2003; Srinivas and Srinivasan, 2001; Tootle and Piechota, 2004).

In addition to the above mentioned models, a partial least squares regression (PLSR) can be applied to streamflow forecasting. Although PLSR is relatively new to hydrologic applications, it has been used in chemometric studies since its development by Herman Wold in the 1960s for use in econometrics (Wold, 1966). The attractive feature of PLSR is that the regression is based on the principal components of both predictors and response variables. The application of PLSR in the seasonal streamflow forecasting is starting to draw the attention of hydrologists. Tootle et al. (2007) applied the PLSR in long lead-time seasonal streamflow forecasting for the first time using Pacific and Atlantic sea surface temperatures (SSTs). Their research suggested that the PLSR could provide strong forecast skill in long lead-time seasonal runoff volume forecasts.

In PCR, the intercorrelations among the predictors are the basis, whereas weightings have no knowledge of the response variable. On the contrary, the Z-score regression is based on correlations with the response variable, whereas the Z-score weightings have no knowledge of the intercorrelations among the predictors. In contrast to PCR and Z-score regression, the PLSR balances the information in both predictors (i.e., precipitation and snow water equivalent) and response variables (i.e., seasonal streamflow volume) by focusing on the covariance between them, and reduces the impact of large, but irrelevant predictor variations. Partial least squares regression is similar to canonical correlation analysis (CCA). Canonical correlation analysis is also a well-known technique for feature extraction from two sets of multidimensional variables. However, unlike PLSR, the CCA is not a prediction technique, but rather a technique for describing the relationship between two sets of multivariate data. The fundamental difference between CCA and PLSR is that CCA maximizes the correlation while PLS maximizes the covariance (Sun et al., 2009). Both PLSR and CCA are the various multivariate extensions of the multiple linear regression models. However, the CCA extends multiple linear regression that imposes restrictions, such that factors underlying the response and predictor variables are extracted from predictor variation and response variation, respectively, and never from covariation involving both predictor and response variables. In PLSR, prediction functions are represented by factors extracted based on the covariance between predictor and response variables. This is probably the least restrictive multivariate technique. This flexibility allows it to be used in situations where the use of traditional multivariate methods is limited, such as when there are fewer observations than predictor variables (StatSoft Inc., 2008). In the next sections of this chapter the methodologies of PCR and PLSR will be discussed in detail.

2.2.4 Principal Components Regression (PCR)

Principal components regression is a standard multivariate regression that deals with highly intercorrelated independent variables by using principal components as regressors in the regression. It is an alternative regression solution for the multiple linear regression when there are a large number of predictor variables that are correlated with each other (NRCS, 2007). When the predictor variables are not correlated, the multiple linear regression would be the first choice in regression. The matrix form of the multiple linear regression models is expressed as follows (Geladi and Kowalski, 1986):

$$\mathbf{Y=XB+E^*} \qquad (2.1)$$

Where

Y - the matrix of dependent variables, (n × p)

X -the matrix of predictor variables, (n × m)

$\mathbf{E}^*$ - the residual matrix, (n × p)

B-the matrix of coefficients, (m × p)

n- the number of observations in calibration set

p- the number of dependent variables, (p=1 in this study, it is April-September runoff volume at Rio Grande near Del Norte Gaging Station or March-July Elephant

Butte net inflow volume, Rio Grande).

m- the number of independent variables

The least squares solution is:

$$\hat{B} = (\mathbf{X'X})^{-1}\,\mathbf{X'Y} \tag{2.2}$$

In principal components regression, the multicollinearity that existed in the predictor variables can be eliminated by extracting a group of orthogonal principal components from predictors through principal components analysis (PCA) on **X**, and then performing multiple linear regressions on **Y** using principal components of **X.** Principal Components Analysis (PCA) of the matrix (**X)** decomposes (**X)** into a score matrix (**T)** times a loading matrix (**P)** and a residual (i.e., error) matrix (**E)** (Abdi, 2003; Tootle et al., 2007). It is possible to let the score matrix, **T**, represent the predictor matrix, **X**:

$$\mathbf{X=TP'+E} \quad \textbf{and } \mathbf{T=XP} \tag{2.3}$$

Where

T – the matrix of **X** scores, (n × a)

P' – the matrix of **X** loadings, (a × m)

E – a residual matrix of **X**

a – the number of factor used in the regression

Then, the multiple linear regression formula can be written as following by replacing **X** with **T**:

$$\mathbf{Y=TB+E^*} \qquad (2.4)$$

The solution is:

$$\hat{B} = \mathbf{(T'T)^{-1}T'Y} \qquad (2.5)$$

The graphical description is shown in Figure 2.1.

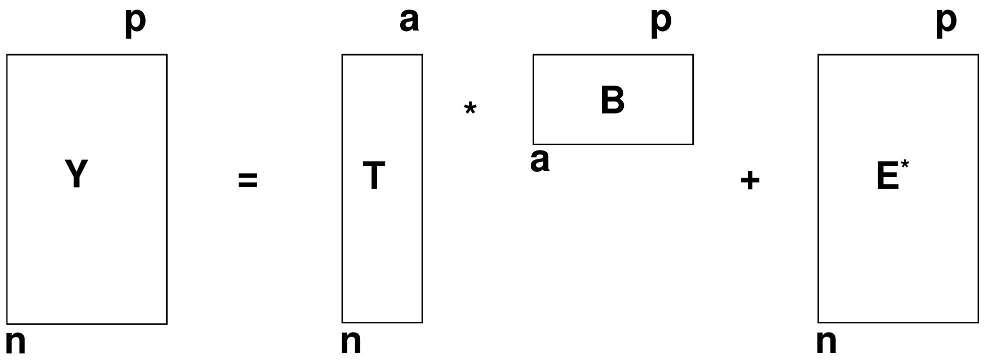

Figure 2.1 Graphical representation of principal components regression algorithm (adapted from Geladi and Kowalski, 1986)

The determination of which and how many principal components should be retained is a key issue in the principal components regression. Garen (1992) provided a detailed description for the selection of principal components in seasonal streamflow forecasting. The description of detailed methodology of principal components analysis and principal components regression can also be found in Geladi and Kowalski (1986), McCuen (1985), McCuen and Snyder (1986), and NRCS (2007).

2.2.5 Partial Least Squares Regression (PLSR)

In the principal components regression, the principal components of **X** explain the variability in **X** rather than **Y**, hence a PCR may include components that are irrelevant to the prediction of **Y** (Abdi, 2003). In contrast, PLSR is developed based on both principal components of **X** and **Y**. Specifically, PLSR searches for a set of components (also called latent vectors) that explains as much of the covariance between **X** and **Y** as possible by performing simultaneous decomposition of both **X** and **Y** (Abdi, 2003).

Partial least squares regression (PLSR) is a combination of individual outer relations of **X** and **Y**, and an inner relation of linking both **X** and **Y** matrices (as shown in Figure 2.2).

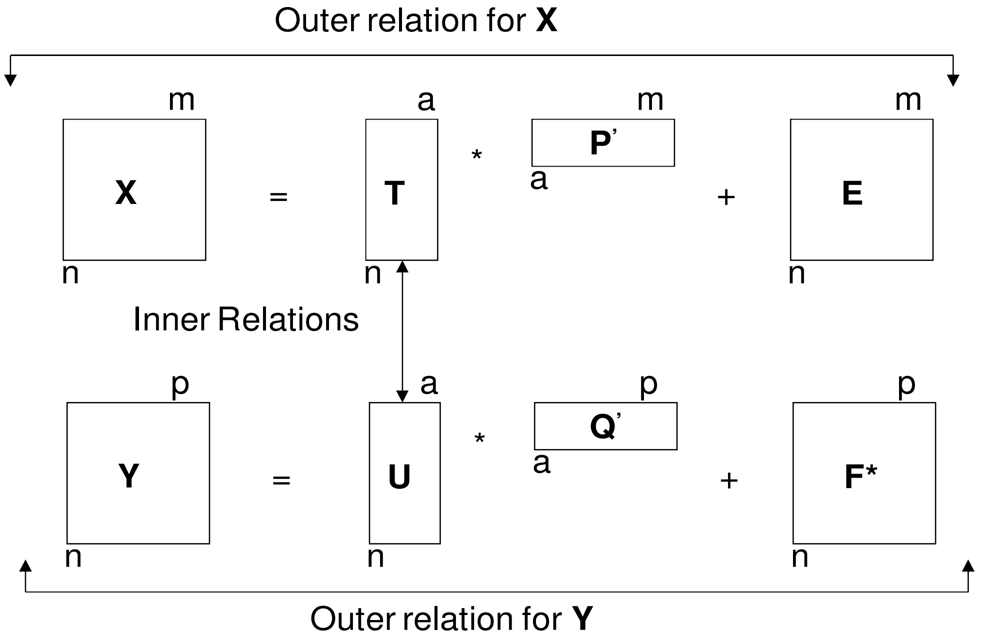

Figure 2.2 Algorithm of PLSR showing an inner and outer relationship of **X** and **Y** matrices (Tootle et al., 2007)

The outer relation for the **X** matrix, which is similar decomposition as the principal components analysis, can be expressed as (Geladi and Kowalski, 1986):

$$\mathbf{X=TP'+E} \;\; = \sum \mathbf{t_h p_{h'}} + \mathbf{E} \qquad (2.6)$$

The outer relation for the Y matrix can be expressed in the same way:

$$\mathbf{Y=UQ'+F^*} \;\; = \;\; \sum \mathbf{u_h q_{h'}} + \mathbf{F^*} \qquad (2.7)$$

Where,

$\mathbf{t_h}$ - a column vector of scores for X block

$\mathbf{p_h}$**'-** a row vector of loadings for X block

U - the matrix of **Y** scores

Q'- the matrix of **Y** loadings

$\mathbf{F}^*$ – a residual matrix of **Y**

$\mathbf{u_h}$ – a column vector of scores for **Y** block , factor h

$\mathbf{q_h}$**'**- a row vector of loadings for **Y** block, factor h

The inner relation of **X** and **Y** can be expressed by the regression of **Y** block score, u, against **X** block score, t , for every component. The simplest model for this relation is linear one (Geladi and Kowalski, 1986):

$$\mathbf{\hat{u}_h} = b_h \mathbf{t_h} \qquad (2.8)$$

where $b_h = \mathbf{u_h{}'t_h/t_h{}'t_h}$. The b_h is equivalent to the regression coefficients. This simple model (Equation 2.8) is not the best one, because the principal components of **X** and **Y** are calculated separately so that they have a weak relation to each other. The inner relation can be improved by exchanging scores between the **X** and **Y** blocks in the iterative process. Considering the outer and inner relation of **X** and **Y** blocks, the following mixed relation can be given to **Y** where the error, **F** , is minimized:

$$\mathbf{Y}=\mathbf{TBQ}' + \mathbf{F} \qquad (2.9)$$

There are several algorithms available for obtaining partial least squares estimators, such as nonlinear iterative partial least squares (NIPALS), singular value decomposition (SVD), and SIMPLS method of de Jong (1993). Geladi and Kowalski (1986) provided a detailed tutorial on the PLSR method. More detailed information about PLSR can be found in Abdi (2003), Tootle et al. (2007), Wold (1966) and Wold (1994).

2.3 Monthly Flow Forecasts

2.3.1 Modeling of Monthly Streamflow Processes

There is a considerable amount of literature on the modeling of monthly streamflow process due to its importance in water resources management and planning. Many research studies have covered application of different kinds of modeling approaches to monthly streamflow forecasting and simulation, including conceptual models to time series analysis, artificial neural networks, non-parametric modeling methods and, more recently, the hybrid modeling approaches. Monthly streamflow processes usually show strong seasonality and nonlinearity. The application of various data-driven models in streamflow forecasting is largely dependent on the time scale of the dependent variable.

Time series analysis is one of the most popular forms of data-driven modeling for streamflow forecasting. The technique has been widely used in recent decades because of its forecasting capability, simple and readily available data needs, and more systematic way of building models by three modeling stages (identification, estimation, and diagnostic check) which had been standardized by Box and Jenkins (1976). The application of time series modeling in streamflow forecasting includes univariate models which deal with only one time series and more complex multivariate models (dynamic regression models, also called transfer function-noise models). The univariate models are based on past streamflows and do not take into account the effects of other time series variables, such as precipitation, snowmelt and temperature. In contrast, the transfer function-noise (TFN) models can incorporate exogenous time series variables in addition to past streamflows. Because more information is used for making forecasts, usually transfer function-noise (TFN) models can make better forecasts than the univariate autoregressive integrated moving average (ARIMA) models (Wang, 2006; Wang et al., 2005b).

The univariate time series models including ARIMA and its derivatives such as seasonal ARIMA, periodic ARIMA, deseasonalized ARMA have long been applied in streamflow forecasting, particularly in the modeling of monthly streamflow (e.g., Abrahart and See, 2000; Bender and Simonovic, 1994; Hipel and McLeod, 1994; McKerchar and Delleur, 1974; Noakes et al., 1985; Salas, 1992; Tesfaye et al., 2005; Yürekli et al., 2005). The application of transfer function-noise (TFN) models with exogenous variables in streamflow forecasting can also be found in the literature. Thompstone et al. (1985) compared deseasonalized autoregressive integrated moving average (ARIMA), periodic autoregressive (PAR), and TFN models with rainfall and snowmelt inputs, and a conceptual model. They found that the TFN model performed better than other models when forecasting quarter-monthly streamflow. Awadallah and Rousselle (2000) used El Niňo Southern Oscillation (ENSO) sea-surface temperature signals as exogenous input variables to develop a TFN model to forecast summer runoff of the Nile River. Their TFN model suggested that the ENSO input explained 63% of the variability of Nile summer runoff. Mondal and Wasimi (2005) proposed a periodic TFN model and applied it to monthly forecasts of the Ganges River flow using monthly rainfall data of northern India as the predictor. The results suggested that the methodology has the potential capability of capturing the seasonally varying dynamic relationship between monthly rainfall and streamflow processes.

The ARIMA and TFN models used in the streamflow forecasting process were generally linear models. They were built under the assumption that the process follows normal distribution. But most streamflow processes are commonly accepted as nonlinear (Wang, 2006). Moreover, the normal distribution assumption is frequently violated in streamflow processes. Hence, the recently developed machine learning techniques, artificial neural networks (ANN), have gained more and more popularity for hydrological forecasting because of their capability of identifying complex non-linear relationships between input and output data sets without the necessity of understanding the nature of the phenomena and without making any underlying assumptions regarding linearity or normality.

Previous studies have concluded that ANNs are useful for forecasting streamflows. Markus et al. (1995) predicted monthly flow at Rio Grande near Del Norte in southern Colorado using neural network models and compared the results with the periodic transfer function model. The study showed that the ANN models provided slightly better results than periodic TFN models using standardized monthly flow data. Hsu et al. (1995) concluded that the ANN model is an effective alternative to ARMAX (autoregressive moving average with exogenous inputs). Huang et al. (2004) compared the ANN and ARIMA for daily, monthly, quarterly and yearly flow forecasting and concluded that the ANN provide better forecasting accuracy than ARIMA model. Many of the following studies have confirmed the superiority or comparableness of the ANN models over the traditional statistical and/or conceptual techniques in modeling the hydrological process (e.g., Abrahart et al., 2004; Birikundavyi et al., 2002; Coulibaly et al., 2000; Dibike and Solomatine, 2001; Govindaraju and Rao, 2000; Raman and Sunilkumar, 1995; Salas et al., 2000; Shamseldin, 1997; Tokar and Markus, 2000).

The identification and inclusion of exogenous variables is a critical step in building TFN and ANN models for monthly streamflow forecasting. There are many factors that may affect the streamflow process. These may include local factors such as discharges at the upstream gauging stations, precipitation and temperature, snowpack information in the watershed, and evaporation, as well as larger-scale phenomena characterized by geophysical indices, such as El Niño Southern Oscillation (ENSO) and Pacific Decadal Oscillation (PDO) (Mantua et al., 1997). In addition to local factors, the linkage between streamflow processes and geophysical indices have been intensively studied in the last few decades (e.g., Eltahia, 1996; Hamlet and Lettenmaier, 1999; Piechota et al., 1998; Piechota et al., 2001; Piechota and Dracup, 1999; Whitaker et al., 2001). Most of the previous studies were based on regression type models that established connections between climatic indices and streamflows. Longer time scales, such as seasonal and annual were reported in the previous studies when using geophysical indices as predictors of streamflow forecasting. On the other hand, the inclusion of ENSO and PDO signals in forecasting monthly streamflow processes have rarely been reported. This may be because the long-range streamflows are commonly related to some remote geophysical quantities, while real-time and short- to medium-range discharges are associated with local factors and initial conditions of the watershed.

In conclusion, the inclusion of various exogenous variables in building monthly TFN and ANN models is a very difficult task which requires understanding of hydrologic characteristics of the specific basin under study. Complete pre-modeling analysis to identify the magnitude and patterns of relationships existing between modeled hydrologic variables and predictor variables on a monthly time scale should be examined. For example, the ENSO signals and snow budget information in a basin may have effects on the streamflow processes of some months, while streamflows in other months may not have any correlation with the ENSO signals and snow budgets in the basin. This could result in difficulties including monthly snowpack information in the TFN modeling, since the snow only exists in winter and spring seasons in most of the basins. Hence, the pre-modeling analysis could provide modelers with an insight to choose appropriate forms of models to be used, and the approach that could be used to include this information in the modeling processes.

2.3.2 Autoregressive Integrated Moving Average (ARIMA) model

Several types of ARIMA modeling methods and their derivatives could be used in the modeling seasonal time series, such as monthly streamflow time series. They are seasonal ARIMA, periodic ARIMA and deseasonalized ARMA model. The deseasonalized ARMA type of modeling strategy was adopted in this study due to its simplicity and effectiveness of modeling. The general form of ARIMA model is expressed as (Vandaele,1983) :

$$\varphi(B)y_t = \theta(B)a_t \qquad (2.10)$$

Where,

$y_t = (1-B)^d Y_t$ - stationary series after differencing

$\varphi(B) = 1 - \varphi_1 B - \varphi_2 B^2 - .... - \varphi_p B^p$ - nonseasonal autoregressive polynomial

$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - .... - \theta_q B^q$ - nonseasonal moving average polynomial

a_t = white noise process

Y_t = dependent variable

B is the backward shift operator defined as $BX_t = X_{t-1}$

Examination of the autocorrelation function (ACF) and partial autocorrelation function (PACF) provides a thorough basis for analyzing the system behavior under time dependence, and will suggest the appropriate parameters to include in the model. The Box and Jenkins (1976) three-stage standard modeling procedure (identification, estimation, and diagnostic check) can be used to develop ARIMA models.

2.3.3 Transfer Function-Noise (TFN) Model

The transfer function-noise model is a time series model that incorporates more than one time series and introduces explicitly the dynamic characteristics of the system (Vandaele, 1983). It can be generally written as:

$$Y_t = \nu(B)X_t + N_t \tag{2.11}$$

Where

X_t = explanatory variable

$\nu(B) = \nu_0 + \nu_1 B + \nu_2 B^2........$ - referred to as impulse response polynomial. It can be approximated by a ratio of two finite polynomials.

N_t = a stochastic disturbance (noise) term that may or may not be correlated, but must be independent of input series. Assume that N_t can be modeled by ARIMA models through proper differencing.

Then Equation 2.11 can be written as :

$$Y_t = \frac{\omega(B)}{\delta(B)} X_{t-b} + \frac{\theta(B)}{\varphi(B)(1-B)^d} a_t \qquad (2.12)$$

If the disturbance term needs differencing to induce stationarity, the same differencing should be applied to the dependent and explanatory variables. In reality, stationary dependent and explanatory variables are used in building a TFN model. Therefore, there is no need to use same difference operator for each variable. Furthermore, it is then possible that after these stationarity transformations the noise term (N_t) is a simple ARIMA process, not involving any difference operator (Vandaele, 1983). As a result, the transfer function can more generally be written as:

$$y_t = \frac{\omega(B)}{\delta(B)} x_{t-b} + \frac{\theta(B)}{\varphi(B)} a_t \qquad (2.13)$$

With $y_t = (1-B^s)^{d'}(1-B)^{d'} Y_t$

$x_t = (1-B^s)^d (1-B)^d X_t$

Where

$\omega(B) = \omega_0 - \omega_1 B - .... - \omega_l B^l$ - referred to as numerator polynomial

$\delta(B) = \delta_0 - \delta_1 B - .... - \delta_r B^r$ - referred to as denominator polynomial,

where δ_0 =1 without loss of generality

d'= order of consecutive differencing of the dependent variable Y_t until getting stationary series y_t .

d = order of consecutive differencing of the input variable X_t until getting stationary series x_t .

b = delay parameter or dead time, is defined as $B^b x_t = x_{t-b}$, indicates the number of periods it takes before input variable starts influencing the dependent variable.

The single input TFN model as shown in Equation 2.13 can easily be extended to multiple-input TFN model which includes several input variables to the following form:

$$y_t = \sum_{i=1}^{m} \frac{\omega_i(B)}{\delta_i(B)} x_{i,t-b_i} + \frac{\theta(B)}{\varphi(B)} a_t \tag{2.14}$$

Where, *m* = number of input variables included in the model. Box and Jenkins (1976) provided a comprehensive procedure for TFN modeling using prewhitened input series and filtered output series. The detailed algorithm is also described by Vandaele (1983). The Statistical Analysis Software ® (SAS) version 9.1 can be used for time series model development.

2.3.4 Artificial Neural Networks (ANN)

Artificial neural networks are flexible mathematical structures that are capable of identifying complex non-linear relationships between input and output data sets. The motivation for the development of neural network technology stemmed from the desire to develop an artificial system that could perform "intelligent" tasks similar to those performed by the human brain.

The most commonly used type of ANN is a feed-forward network termed the multilayer perceptron (MLP). In this type of network, the artificial neurons, or processing units, are arranged in a layered configuration containing an input layer, usually one hidden layer, and an output layer. Units in the hidden and output layers are connected to all of the units in the preceding layer. Each connection carries a weighting factor. The weighted sum of all inputs to a processing unit is calculated and compared to a threshold value. The activation signal then is passed through a mathematical transfer function to create an output signal that is sent to processing units in the next layer. Kim and Valdes (2003) described three-layered feed forward neural networks (FFNN) and provided a general framework for representing nonlinear functional mapping between a set of input and output variables.

Three-layered FFNNs are based on a linear combination of the input variables, which are transformed by a nonlinear activation function. The explicit expression for an output value of FFNN for one output neuron is given by:

$$\hat{y}_p = f_0\left[\sum_{j=1}^{M} w_j f_h\left(\sum_{i=1}^{N} w_{ji} x_{pi} + w_{j0}\right) + w_0\right] \qquad (2.15)$$

Where

w_{ji} is a weight in the hidden layer connecting the ith neuron in the input layer and the jth neuron in the hidden layer

w_{jo} is the bias for the jth hidden neuron

f_h is the activation function of the hidden neuron

w_j is a weight in the output layer connecting the jth neuron in the hidden layer

w_o is the bias for the output

x_{pi} a value of the ith input for pattern p

f_0 is the activation function for the output neuron

The weights are different in the hidden and output layer, and their values can be changed during the process of network training. The relationship of the available input variables and output variables is generated by the training process. The process of training ANNs is accomplished by a backpropagation algorithm, which has been applied successfully to solve difficult and diverse problems. This algorithm is based on the error-correction learning rule. Basically, the error-propagation process consists of two passes through the different layers of the network as shown in Fig. 2. In the forward pass, an input vector is applied to the neurons of the network, and its effect propagates through the network layer by layer. A set of output is produced as the actual response of the network. During the backward pass, on the other hand, the weights are all adjusted in accordance with the error-correction rule. The error signal is then propagated backward through the network. The weights are adjusted so as to make the actual response of the network closer to the desired response.

The objective of the backpropagation training process is to adjust the weights of the network to minimize the sum of squared errors of the network, which approximates the model outputs to the target values with a selected error goal:

$$E(n) = \frac{1}{2}\sum_{p=1}^{n}\left[y_p(n) - \hat{y}_p(n) \right]^2 \qquad (2.16)$$

Where

n is the number of observations

$y_p(n)$ is the desired target responses

$\hat{y}_p(n)$ is the actual response of the network at the n^{th} iteration

The detailed description of the algorithm is provided in many studies (e.g., Tokar and Markus, 2000; Coulibaly et al., 2000; Dibike and Solomatine, 2001; Kim and Valdes, 2003). The NeuroSolutions$^{\text{TM}}$ version 5.1 software, a neural network development environment (NeuroDimension Inc., 2009) was used in neural network modeling in this study.

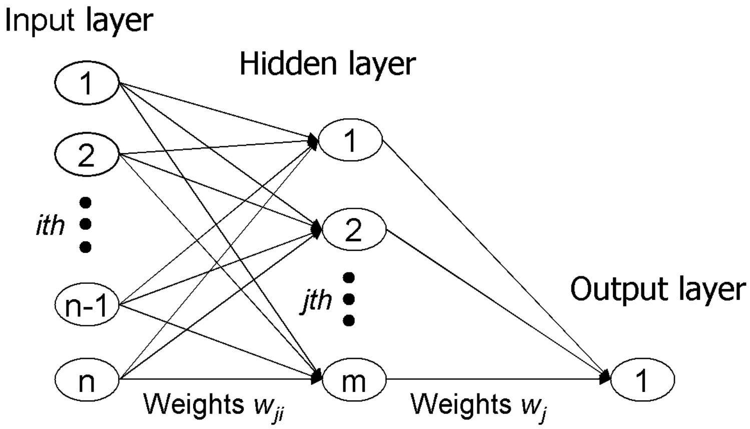

Figure 2.3 A typical three-layered feedforward neural network structure with one output neuron

2.4 Hybrid Modeling of Streamflow Processes

2.4.1 Application of Hybrid Modeling in Streamflow Forecasts

Streamflow forecasting is a challenging task because of the complexity of the hydrologic system. There is no individual streamflow forecasting model that provides better forecast results under all circumstances with respect to alternative competing models (Shamsheldin, 2004). No single forecasting model is powerful and general enough to outperform the others for all types of catchments and under all circumstances or even one catchment with different behavioral phases. Every model has some degree of uncertainty, including structure and parameter uncertainty (Shamsheldin et al, 1997). Therefore, the reliance of a single model in streamflow forecasting may result in a considerable risk in water management if that model failed to provide reliable forecasts. A possible method to overcome this deficiency could be the application of a hybrid modeling approach which includes the combination of forecasts from different individual models and integration of different models that may provide better forecasting solutions than a single model.

Although relatively new to hydrological forecasting, the hybrid modeling approach has long been applied in diverse fields such as economics, business, and meteorology (Clemen, 1989; Shamsheldin, 2004). The early work of McLeod et al. (1987) that combined quarter-monthly river flow forecasts from different time series models laid a foundation for the application of hybrid modeling of streamflow forecasting. Since then, hybrid modeling has found wide application in the field of streamflow forecasting and many research studies have been reported in the literature

(e.g., Abrahart and See, 2002; Jain and Kumar, 2007; Kişi, 2008; See and Abrahart, 2001; See and Openshaw, 1999; See and Openshaw, 2000; Shamsheldin et al, 1997; Shamsheldin et al., 2002; Srinivas and Sirinivasan, 2001; Wang et al., 2005b)

The recent work of Shamsheldin (2004) and Wang (2006) provide systematic analysis and summary on the application of hybrid modeling in streamflow forecasting and the underlying theory behind modeling procedures. According to their work, hybrid modeling as the integration of different models by definition may be divided into two approaches, namely a non-modular approach and a modular approach. The non-modular approach is essentially a forecast combination method that combines the individual forecast outputs from each different model. The modular approach, on the other hand, uses a divide-and-conquer principle to divide a complex forecasting problem into several simpler modeling subtasks, each of which is modeled by a different appropriate modeling method. Finally, the forecast results are integrated to yield a hybrid forecast.

Previous studies have shown the advantage of combining forecasts in streamflow forecasting. By combining the forecasts from several models, one can obtain a more reliable and accurate output than would be obtained by selecting a single model. The research results of Coulibaly et al. (2005) showed that using a weighted average method to combine three dynamically different models (artificial neural network, conceptual model and nearest neighbor model) could significantly improve the accuracy of the daily reservoir inflow forecast for up to four days ahead. Shamseldin et al. (1997) examined three different combination methods in the context of flood forecasting; namely, the simple average method, the weighted-average method and the neural network method. See and Openshaw (2000) used four different approaches such as simple average, a Bayesian approach, and two fuzzy logic models, to combine the river level forecasts of three models (i.e., a hybrid neural network, an autoregressive moving average model, and a simple fuzzy rule-based model), and found that the addition of fuzzy logic to the crisp Bayesian approach yielded overall results that were superior to the other individual and integrated approaches. Wang (2006) employed four combination techniques, (i.e., simple average method, rollingly-updated weighted average method, semi-fixed weighted average method, and modular semi-fixed weighted average), and used them to combine the daily streamflow forecasts. The results showed that simple average method could improve the accuracy of forecasts with a four to five day lead time, and generally performed best among four competitive combination methods.

The modular approach of hybrid modeling can facilitate integration of conventional hydrological models such as time series analysis, conceptual models with those newly developed modeling techniques such as artificial neural networks, fuzzy logic and non-parametric modeling. Hence, it can exploit the strength of different modeling techniques to produce a better forecasting solution (Shamsheldin, 2004). Recent research results reported in the literature justified the robustness of the methodology in streamflow forecasting. Wang et al. (2005b) presented the application of three forms of hybrid artificial neural networks (ANNs); namely, the threshold-based ANN, the cluster-based ANN, and the periodic ANN in daily streamflow forecasting. For the purpose of comparing forecasting efficiency, the normal multi-layer perceptron (MLP) form of ANN was selected as the baseline ANN model. Compared to the MLP which is fitted to the deseasonalized data, the periodic ANN, which is based on the soft seasonal partitioning, performed better for short lead times (3 days). Srinivas and Sirinivasan (2001) presented a hybrid model for stochastic simulation of multi-season streamflows which involves partial prewhitening of the streamflows using a parsimonious linear periodic parametric model, followed by resampling the resulting residuals using the moving block bootstrap method. See and Openshaw (1999) developed a modular neural network (MNN) river flow forecasting model for the River Ouse in the United Kingdom. In their study, they divided the hydrographs into four sections: rising flow limb, peak flow, falling limb and low flows. Different neural network models were developed for each section. The produced results for each section by individual neural network models was then integrated using a sophisticated fuzzy rule-based approach. The results showed that the proposed hybrid approach provided a well-performing and low-cost solution compared to the existing methods.

There are two main issues in the application of modular hybrid modeling in streamflow forecasting. The first step is to break down the complex forecasting problem into simpler components based on the conditions of the hydrologic processes. Secondly, integrate individual modeling results of each component and assess the forecast uncertainty with respect to hybrid modeling results. Various approaches were reported in the literature with regards to the problem. Zhang and

Govindaraju (2000) divided the flow into low, medium and high flow events and modeled each flow category with different neural network models and finally used linear function to combine the results of individual inputs. Hsu et al. (2002) used Self-Organizing Feature Map (as developed by Kohenon, 1984) to partition the input domain of neural networks into several regions such as base flow, increasing rainfall, and peaking hydrograph, then used a set of piecewise linear equations to combine the results of neural networks developed for each region. Parasuraman and Elshorbagy (2007) investigated the performance of a cluster-based neural network model trained using a genetic algorithm on two distinct case studies. The input data was clustered using K-means algorithm. Each cluster was then trained by an individual neural network module. The result of the study showed that the cluster-based neural network performed better than its counterparts in predicting the chaotic time series. Other methodologies such as fuzzy logic based partitioning of neural network model inputs (See and Openshaw, 1999), clustering inputs using self organizing map (Abrahart and See, 2000), range-dependent neural networks that partition input data as low, medium and high flow sets (Hu et al., 2001), and spiking modular neural networks (Parasuraman et al., 2006) have also been applied in the hybrid neural networks modeling of hydrologic forecasting.

Apart from modular and non-modular hybrid modeling approaches, another possible category of hybrid modeling techniques may be categorized as ‘complementary’ modeling. It may be a combined application of physically-based models and data-driven models, or the combination of different modeling techniques on the same forecasting problem. Solomatine and Price (2004) proposed the idea of complementary models in categorizing hybrid modeling methods. They pointed out that the complementary modeling methods can focus on the mismatch between physically-based models and observations. A data-driven model could be used as the secondary model to estimate the measured mismatch and to update the results of original models.

Complementary modeling is essentially a modeling of errors by combined application of several models. For example, a first model, such as time series model, may be used to generate the first forecasts. Then the second model, such as neural networks, would be used to model the forecast errors from the first model and combine them to generate final forecasts. Some application of such models in econometrics has been reported in the literature. Tseng et al. (2002) proposed a hybrid forecasting model that combined seasonal ARIMA and neural network backpropagation models for forecasting total production value for the Taiwan machinery industry and the soft drink time series. Zhang (2003) used a combination of ARIMA and neural networks in forecasting sun spot data, Canadian lynx data and the British pound/ US dollar exchange rate data. The results demonstrated the effectiveness of the hybrid methodology over any single modeling approach. The use of the hybrid modeling methodology and its role in improving forecasting accuracy was also reported in other literature (Aryal and Wang, 2004; Aslanargun et al., 2005; Joy and Jones, 2005).

In addition to error modeling, the assimilation of data pre-processing and forecast modification into forecast modeling methods can also be conceived as the ‘complementary’ modeling technique since they exploit advantages of both for improving streamflow forecasting accuracy. The idea behind such a modeling methodology is that it may be possible to improve the performance of neural network models in streamflow forecasting by removing the long-term trend and seasonal variations that exist in the raw data (Jain and Kumar, 2007). Recently, a number of studies on the combination of data pre-processing and neural networks have been reported in the literature. Kişi (2008) proposed a neuro-wavelet technique for modeling monthly streamflows. The combination of discrete wavelet transform and multi-layer perceptron were tested for one-month-ahead streamflow forecasting and the results revealed that the methodology improved the forecast accuracy compared to single neural network, multiple linear regression and autoregressive models. Mehdicani et al. (2006) presented an approach similar to that used in Kişi (2008). It was a conjunctive nonlinear model using wavelet transforms and artificial neural network. They applied the methodology in 15-day and monthly reservoir inflow and found that the forecasting performance of the methodology was better than conventional neural network prediction models. Jain and Kumar (2007) proposed a new hybrid time series neural network to exploit the strengths of both traditional time series and artificial neural networks. Two data pre-processing approaches were employed for the input data of neural networks: de-trended data, de-trended and deseasonalized data. The de-trending was performed by removing long term trends by subtracting the annual average flow from the original time series. The deseasonalization was performed using Fourier mean approach. The results of the study suggested that the hybrid modeling method was a robust modeling technique that would be capable of capturing the non-linear nature of the complex streamflow time series.

Forecast modification can be another rigorous method applicable to streamflow forecasting. Karamouz and Zahraie (2004) presented a seasonal streamflow forecast modification approach using fuzzy rules based on the snow budget over a watershed and El Niño Southern Oscillation climate signals. The seasonal streamflow volume forecasted by ARIMA was modified by using proposed algorithm in their study. The results indicated that the proposed methodology has shown improvement in the statistical forecasts of Salt River Basin in Arizona. The application of various methodologies in forecast modification, and innovative use of various techniques in streamflow forecasting, may contribute significantly to improving forecast accuracy. Therefore, further studies may be needed to substantiate practical application of the methodology in the hybrid modeling of streamflow processes.

The introduction of soft modeling techniques such as neural networks and fuzzy logic models facilitated the application of various hybrid modeling approaches in modeling complex streamflow processes. The number of publications on hybrid modeling of streamflow forecasting is increasing rapidly in recent years due to its ability to improve forecast accuracy. However, there is still a need for further research that applies the hybrid models in streamflow forecasting; particularly when dealing with the issue of forecast uncertainty evaluation of hybrid modeling approaches. The literature review has presented the possible categorization and brief evaluation of hybrid modeling techniques in hydrologic modeling. It is hoped that future hydrological forecasting research efforts will also exploit the potential capabilities of hybrid modeling in achieving increased forecast accuracies in streamflow forecasting.

2.4.2 Hybrid Modeling Approaches Used in the Study

Based on the literature review in the previous sections, the following two hybrid modeling approaches have been investigated for the purpose of improving seasonal and monthly streamflow forecasting performance in this study:

1) Forecast modification using neural networks.
2) The combination of principal components analysis (PCA) and artificial neural networks (ANN).

As discussed, the forecasting modification may fall into a ‘complementary’ hybrid modeling category which is the combination of two models; one used to generate the first forecasts, and then the other to modify the forecasts of the first model by modeling errors. In this study, instead of modeling errors, the forecasts of the first model were used as inputs for the second model. A detailed description of methodology is given in chapter 4.

The second approach, a combination of PCA and ANN, is essentially a combination of data pre-processing and neural networks models. It is also conceived as a ‘complementary’ hybrid modeling approach. The principal components analysis is a statistical technique that deals with highly intercorrelated predictor variables by extracting an equal number of uncorrelated variables. A brief discussion on the collinearity issue in streamflow processes was given in section 2.2.2. The highly intercorrelated variables may affect performance of neural networks by easily over training the network and giving very low performance in the new prediction data. Hence, the combination of PCA and ANN may facilitate the effective neural network modeling because the network may converge fast and easy due to the orthogonal features of principal components using fewer PCs than original variables as inputs. A more detailed description of the approach is given in chapter 4.

2.5 Model Comparison and Forecast Evaluation Measures

Among many factors such as model structure, complexity, and computational requirements, forecast accuracy may be the most important factor used to compare various models and evaluate model performance. There are various forms of model comparison measures in the streamflow forecasting publications. To name a few, Thompstone et al. (1985) used standard error, mean absolute percentage errors (MAPE), mean absolute median errors, and significance test of differences in mean squared errors (MSE). The Pitman test and non-parametric Wilcoxon rank-sum test were applied in their study to compare the performance of the different models. Wang

(2006) utilized the coefficient of efficiency, or model efficiency, which was proposed by Nash and Sutcliffe (1970). Additionally, they used seasonally-adjusted coefficient of efficiency, and root mean squared error in the model performance comparison. Abrahart and See (2000) applied a multicriteria assessment in performance comparison of different models which is the combination of five different global evaluation measures and two event-specific evaluation measures. The five global indices included mean absolute error (MAE), RMSE, mean higher order error function which emphasizes peak flow prediction, model efficiency, and percentage of predictions grouped according to degree of error. The two event-specific indices included average difference in peak prediction over all flood events and percentage of early, on-time, or late occurrence for prediction of individual peaks.

The most commonly used performance evaluation indices in the literature are the mean squared error (MSE) or its variants, such as root mean squared error (RMSE), sum of squared error (SSE), and mean relative error (MRE) (Elshorobagy et al., 2000). Karunanithi et al. (1994) suggested that the two measures can provide different information about predictive ability of the model. For example, the MSE is a good measure for indicating goodness of fit at the high flows, while the MRE provides a more balanced perspective of the goodness of fit at moderate flows. Another derivative of RMSE, the normalized root mean squared error (NRMSE), is a unitless index that can be used to compare the forecasts of different models without considering the magnitude of the flow modeled. The NRMSE is also a measure of residual variance, and is indicative of the model's predictive uncertainty. The low value of NRMSE implies that the model is able to forecast the flows with reasonable accuracy (Jain et al., 2004). Additionally, the coefficient of determination and/or the correlation coefficients, may be another index used extensively for model comparison due to its unitless feature. There are many applications of correlation coefficients in model performance evaluation (Coulibaly et al., 2005; Jain et al., 2004; Jain and Kumar, 2007; Parasuraman and Elshorbagy, 2007; Shamseldin et al., 1997). Some other performance evaluation measures can also be seen in the literature. For example, the normalized mean bias error (Jain et al., 2004), pooled mean squared error - which is combination of MSE and MRE (Elshorobagy et al., 2000), threshold statistics (Jain and Kumar, 2007; Parasuraman and Elshorbagy, 2007), and relative bias (Coulibaly et al., 2005).

In order to compare the forecasting accuracy of the different models, a multicriterion performance evaluation procedure was used in this study. The following indices were used to evaluate the performance of the models:

1. Coefficient of determination (R^2):

$$R^2 = \left[\frac{\sum_{i=1}^{n} (Y_i - \overline{Y})(F_i - \overline{F})}{\sqrt{\sum_{i=1}^{n} (Y_i - \overline{Y})^2 \sum_{i=1}^{n} (F_i - \overline{F})^2}} \right]^2 \qquad (2.17)$$

2. Mean Absolute Error (*MAE*) :

$$MAE = \frac{1}{n} \sum_{i=1}^{n} | Y_i - F_i | \qquad (2.18)$$

3. Mean Absolute Percentage Error (*MAPE*):

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i - F_i|}{Y_i}\times 100 \qquad (2.19)$$

4. Root Mean Squared Error (*RMSE*):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - F_i\right)^2} \qquad (2.20)$$

5. Normalized Root Mean Squared Error (*NRMSE*)

$$NRMSE = \frac{RMSE}{\hat{\sigma}} \qquad (2.21)$$

6. Model Efficiency (*E*):

$$E = 1 - \frac{\sum (Y_i - F_i)^2}{\sum (Y_i - \overline{Y})^2} \qquad (2.22)$$

Where

Y_i = the observed flow;

F_i = the forecasted flow;

$\overline{Y}$ = the mean of observed flow;

$\overline{F}$ = mean of forecasted flow;

$\hat{\sigma}$ = standard deviation of observed flow;

The model efficiency, *E*, is a model evaluation criterion proposed by Nash and Sutcliffe (1970). A model efficiency of 90% and above indicates very satisfactory performance. A value in the range of 80–90% indicates fairly good performance. A value below 80% indicates a questionable fit.

3 SEASONAL FLOW FORECASTING

The primary goal of this chapter is to develop partial least squares regression (PLSR) and principal components regression (PCR) models for seasonal streamflow volume forecasts using snow water equivalent, precipitation, temperature, and previous flow conditions as input variables. The selection of an optimal number of components using jackknife cross validation scheme and variable selection approach with PLSR have been discussed. The performance of PLSR and PCR models in seasonal streamflow forecasting were compared to each other and to NRCS official forecasts. Two subbasins in the Rio Grande and two hydrologic variables, river flow and reservoir net inflow, were used for seasonal flow forecasting model development. Statistical Analysis Software ® (SAS) version 9.1 was used in model development and forecasting.

3.1 Model formulation

The two multivariate regression approaches, partial least squares regression (PLSR) and principal components regression (PCR), were utilized in seasonal volume forecast equation development. Model formulation of the methods is similar. However, variable selection procedure was performed by PLSR methodology and PCR used the same variables that were selected in the PLSR procedure for regression equation development.

3.1.1 Cross Validation

The selection of optimal numbers of extracted components (factors) is a key issue in developing PLSR and PCR, particularly when the models are used for prediction (Tootle et al., 2007). All regression methods, including PCR and PLSR, approach multiple linear regressions (MLR) as more components are extracted. However, when there are many predictors, MLR can over-fit the observed data; biased regression methods with fewer extracted components can provide better predictability of future observations (SAS Institute, 2008).

The number of extracted components must be chosen on the basis of how well the model fits observations not involved in the modeling procedure itself. One method of choosing the number of extracted components is to fit the model to only part of the available data and measure how well models with different numbers of extracted components fit the other part of the data. This is called test set validation. Because of data availability, the cross validation technique is usually used for the selection of significant components in PLSR and PCR. There are several different types of cross validation methods: 1) one-at-a-time cross validation, also known as "leave one out" or jackknifing, fits the model on *n-1* observations and uses the one left out for validation; 2) block cross validation, which is to hold out successive blocks of observations as test sets; 3) split-sample cross validation, in which successive groups of widely separated observations are held out as the test set; 4) random sample cross validation, in which the test sets can be randomly selected from the observed data. Among them, the test set validation approach is preferred when there is enough data to make a division into a sizable training set and test set that represent the predictive population well (SAS Institute, 2008). However, when the sample size is small, the jackknife cross validation method would be the appropriate solution.

Usually, the prediction residual sum of squares (PRESS) statistic is used to determine the minimum number of components required in PLSR (Geladi and Kowalski, 1986). For only one response variable, the PRESS statistic for each of the extracted factors *i,* can be calculated by the following expression:

$$PRESS(i) = \sum_{j=1}^{n} (y_j - \hat{y}_j(i))^2 \tag{3.1}$$

Where

y_j = actual values

$\hat{y}_j(i)$ = predicted values for extracted factor *i* for *j*th observation

n= number of samples

The PRESS statistic is also expressed as the root mean prediction residual sum of squares (RMPRESS) and normalized root mean prediction residual sum of squares (NRMPRESS). They can be given by the following expressions:

$$RMPRESS(i) = \sqrt{\frac{PRESS(i)}{n-1}} \tag{3.2}$$

$$NRMPRESS(i) = \frac{RMPRESS(i)}{\hat{\sigma}} \tag{3.3}$$

Where

$\hat{\sigma}$ = standard deviation of observed values

The fitted models are tested using the cross validation data set, and the predicted values are compared with observed values using PRESS to assess the predictive ability of the model.

3.1.2 Selection of Number of Components to Retain

There are several methods to determine the optimal number of components used in the PLSR and PCR using the PRESS statistic. The combination of two methods was employed for the selection of an optimal number of components in this study:

1) The number of components based on the minimum PRESS statistic value;
2) The number of components using van der Voet's significance test (van der Voet, 1994).

In the first method, the number of components chosen is generally the one that minimizes the PRESS. However, often models with fewer components have PRESS statistics that are only slightly larger than the absolute minimum. To address this, van der Voet (1994) proposed a statistical test for comparing the predicted residuals from different models. When the van der Voet's test is applied, the number of components chosen is the fewest with residuals that are not significantly larger than the residuals of the model with minimum PRESS (SAS Institute, 2008).

The van der Voet's test is configured as follows: Let $R_{i,j}$ be the jth predicted residual for the model with i extracted components; the PRESS statistic then is $\sum_j R^2{}_{i,j}$. Also let $i_{\min}$ be the number of components for which PRESS is minimized. The critical value for van der Voet's test is based on the differences between squared predicted residuals

$$D_{i,j} = R^2{}_{i,j} - R^2{}_{i_{\min},j} \tag{3.4}$$

One alternative for the critical value is $C_i = \sum_j D_{i,j}$, which is just the difference between the PRESS statistics for i and i_{min} components. Virtually, the significance level for van der Voet's test is obtained by comparing C_i with the distribution of values that result from randomly exchanging $R^2{}_{i,j}$ and $R^2{}_{i_{\min},j}$. In practice, a Monte Carlo sample of such values is simulated and the significance level is approximated as the proportion of simulated critical values that are greater than C_i . Usually, the number of extracted components chosen is the smallest number with an approximate significance level that is greater than 0.10.

In addition to the above two methods, two other approaches, including t-test of significance and testing rationality of coefficients, were utilized to determine the number of components that should be retained in PLSR and PCR. The rationality of coefficients includes examination of both sign and magnitude of coefficients of the predictor variables in the final equation (McCuen et al., 1979). Garen (1992) provided a detailed discussion on the selection of principal components to retain for PCR equation development using a combination of t-test of significance of components and sign test of coefficient of original variables. The t-test of significance is a standard t-test used in the stepwise variable selection to determine the significance of the regression coefficients for the variable (component) in multiple linear regressions.

Garen (1992) suggested that the t-test is adequate for the determination of the components to be kept in PCR. However, if all the significant components are used in the regression equation development based on the t-test results, it does not guarantee that the regression coefficients will have the same algebraic sign as the correlation coefficients of the predictor variables with the dependent variables. If a predictor variable that has positive correlation with the dependent variable has a negative regression coefficient in the final equation, it would suggest that there is intercorrelation among the independent variables; the negative sign would indicate that this variable is trying to compensate for the some of the effect of another independent variable with which it is highly correlated (Garen, 1992; McCuen, 1985; McCuen et al., 1979). As the main goal of principal components regression is to deal with the highly intercorrelated predictor variables, it would imply that the intercorrelations are reintroduced in the equation if the regression coefficient of any variable has an opposite sign with the correlation with dependent variable. Hence, the components of PLSR and PCR should be chosen such that the regression coefficients of all variables in the final equation have the same algebraic signs as the correlation coefficients with the dependent variables.

Together with the approach proposed by Garen (1992), the combination of minimum PRESS and van der Voet's test in determining the number of components to retain in both PCR and PLSR was used in this study. The method is described as follows: Let N be the number of components that has minimum PRESS, M be the number of components with p>0.1 in van der Voet's test, P be the number components that passed sign test in sequence, and A be the final selected number of components, one may fall across four different scenarios:

1) N=M=1
2) N=M>1
3) N>M=1
4) N>M>1

1) Scenario (1) is the most straightforward and frequently used in seasonal streamflow forecasting due to high intercorrelations among the predictor variables. In this case, the A=1 is the most appropriate selection;
2) In scenario (2), check the sign test for A=M, if passed, then A=M. Otherwise, A=P, where 1≤ P< M;
3) In scenario (3), perform sign test in sequence starting from M until P (M≤ P< N). Temporarily select A=P, then perform t-test of significance in PCR for P components. If it passes, keep all P components. If not, select the largest number of components that passed the t-test. This number should be smaller than P. For example, if the second component failed in a significance test, but the third component passes a significant test, still keep all three components as long as they pass the sign test.

4) In scenario (4), check the sign test for A=M, if it passes goes to scenario (3). Otherwise, check the sign test starting from 1 to P, select A=P, where $1 \leq P < M$.

In the selection procedures, the components were kept in sequence even though some of them failed in the significance test for the stepwise variable selection using principal components as the regressors. The result of t-test significance in PCR was used as the reference for PLSR. In addition, the magnitude of the coefficients in the final equation was also checked for rationality by examining the magnitude coefficients in standardized form when the selected components were more than one. The combination of four methods seems complicated, but in most of the seasonal volume forecasting problems, usually the first and second components meet the criteria mentioned above. In this study, it was observed that the number of components that had both minimum PRESS and passed van der Voet's test usually passed the sign test, which made the selection procedure much easier.

3.2 Del Norte Natural Flow

3.2.1 Data Description

3.2.1.1 Seasonal Runoff Volume

April-September natural seasonal streamflow volume at Del Norte Gaging Station, Rio Grande (Figure 1.1) was selected as the forecast target volume. The NRCS provides April-September runoff volume forecasts at the same Station on the first day of January, February, March, April, May and June. To be consistent with the NRCS forecast dates and volume, the April-September, May-September and June-September volumes were used as the dependent variables in the modeling so that the results could be comparable to the NRCS median forecasts (the 50% percent exceedance probability) for the same period.

The data period from 1981 to 2007 was used in this study because of the natural flow and real time snow water equivalent, precipitation and temperature data availability in the Basin (The 2006 and 2007 data are provisional data). To calibrate PLSR and PCR regression equations, the total data period was divided into two data sets: the calibration data set (1981-2002) and test data set (2003-2007). Since the data set for the calibration phase was very short (only 22 years of data), the jackknife (leave-one-out) cross validation procedure was used to validate the equation. Finally, the five years of data was used to test the equations and compare with NRCS official forecasts for 2003-2007.

Two statistical features of the data, normality and autocorrelation, were examined to see if data were normally distributed and if autocorrelation existed in the data. Three dependent variables, April-September, May-September and June-September natural runoff volume data, and the data period of 1981-2007 were used to conduct these tests. The Shapiro-Wilk normality test and Ljung-Box white noise test up to six lags were performed to test the normality and autocorrelations. The test results are shown in Table 3.1. It is suggested that the normality distribution assumption of seasonal flows be accepted at 0.05 significance level. However, the white noise test is rejected at 0.05 significance level. The white noise test for such a short data time series (only 27 samples) may not be conclusive. Hence, another white noise test was performed using data period of 1961-2007, and the results showed that the white noise process hypothesis is accepted at 0.1 significance level. Therefore, it can be concluded that the seasonal flows are essentially a white noise process and comply with the normality distribution assumption. The normality plot of the data (Figures 3.1-3.3) also indicated that no transformation was needed for the modeling.

Table 3.1 Normality and autocorrelation tests for Del Norte seasonal flow data (1981-2007)

Flow	Shapiro-Wilk test of normality		Ljung-Box white noise test		
	W Statistic	*p*-value	To Lag	Chi-Square statistic	*p*-value
April-September	0.973	0.703	6	13.97	0.018
May-September	0.965	0.481	6	14.55	0.015
June-September	0.944	0.155	6	13.60	0.027

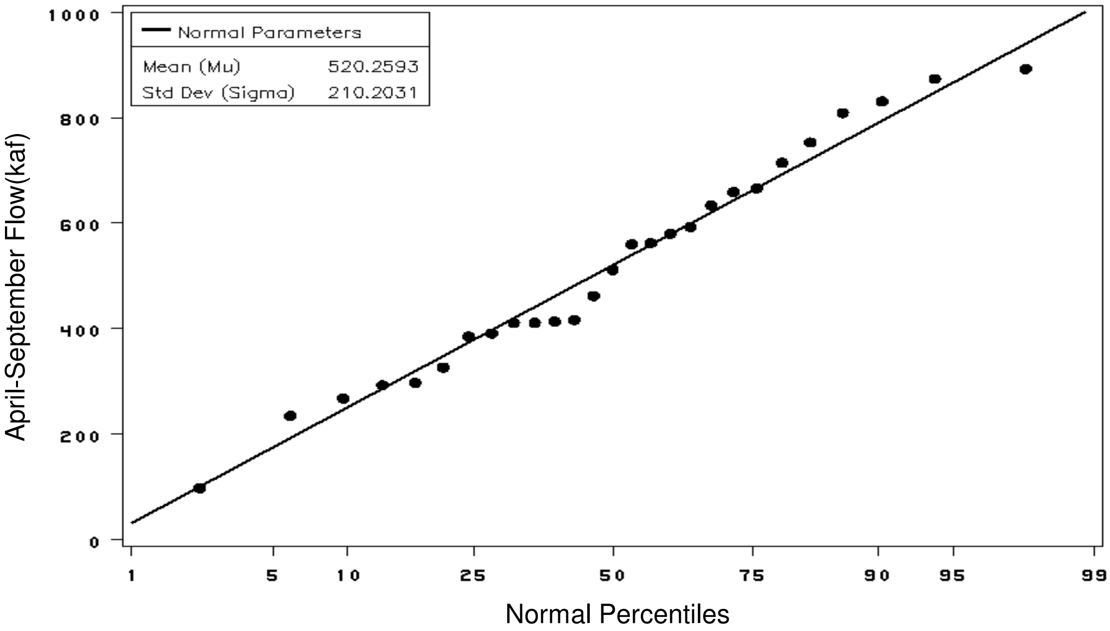

Figure 3.1 Normality plot of April-September natural flow at Del Norte Gaging Station (1981-2007)

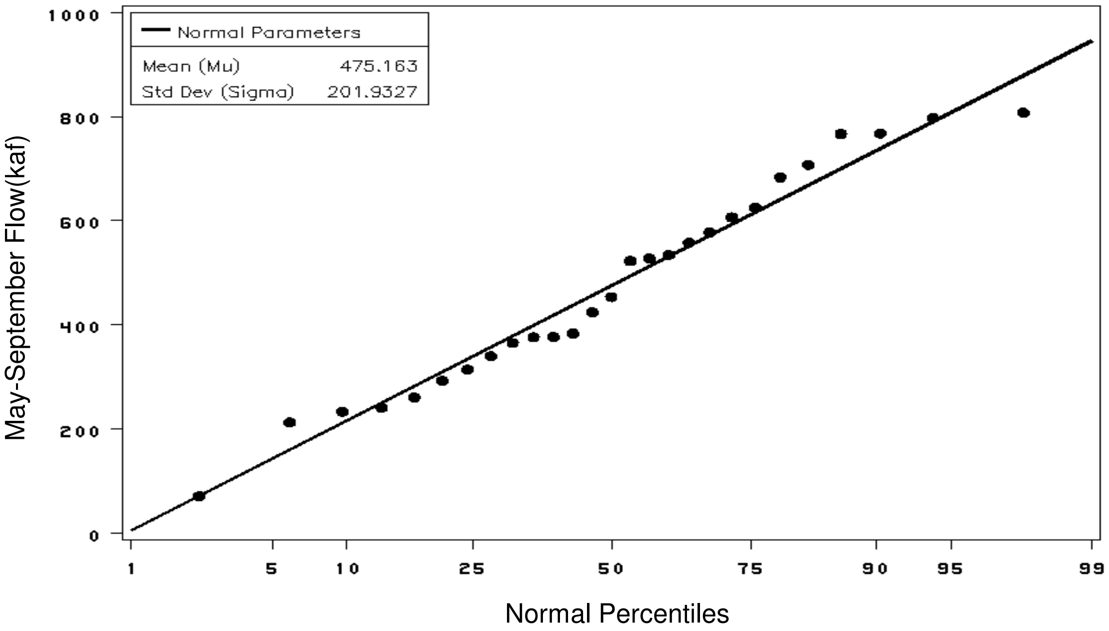

Figure 3.2 Normality plot of May-September natural flow at Del Norte Gaging Station (1981-2007)

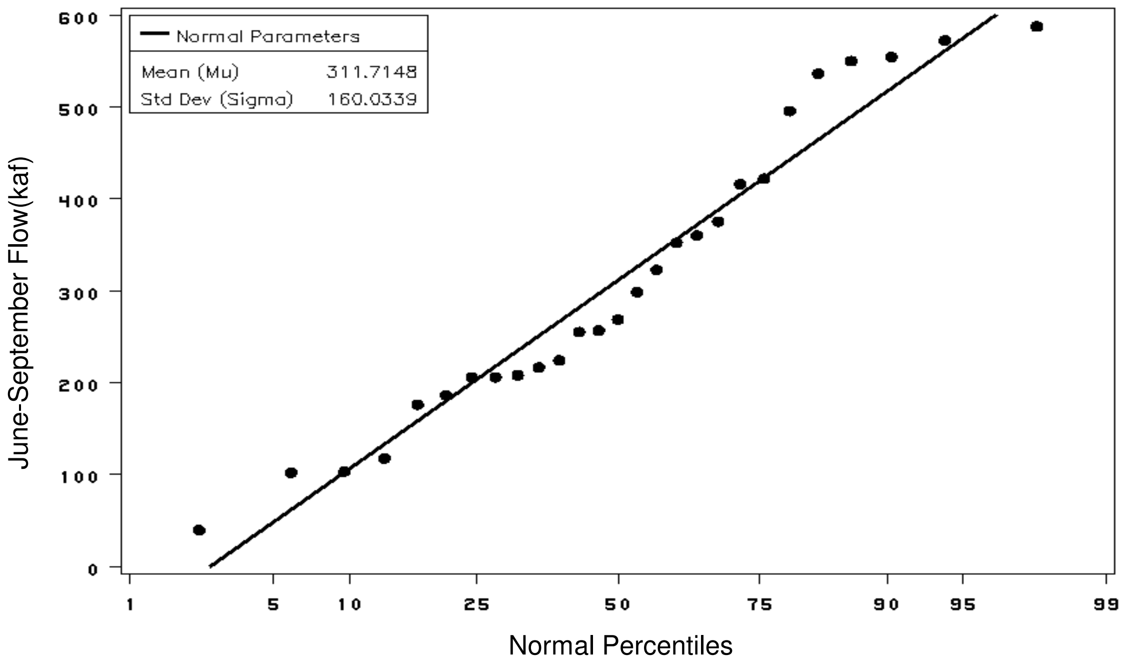

Figure 3.3 Normality plot of May-September natural flow at Del Norte Gaging Station (1981-2007)

3.2.1.2 Other Data

Five different categories of data were used in PLSR and PCR model development. They included SNOTEL snow water equivalent that was measured on the first day of each month from January to June, the monthly SNOTEL precipitation from October to May of a water year, monthly SNOTEL average temperature index from October to May of a water year, monthly flow from October to May of a water year, and previous year averaged October-December El Niño-Southern Oscillation Index (SOI). The data period covered from 1981 to 2007. The detailed data description was given in section 1.4.3. In addition, the NRCS official forecast data was used for comparison with the PLSR and PCR equations developed in this study. The NRCS historical forecast data were obtained from NRCS (Pagano, personal communication, April 10, 2008). The April-September official seasonal volume forecasts for 2003-2007 at Del Norte Gaging Station issued by NRCS were shown in Table 3.2.

Table 3.2 April-September flow NRCS official forecasts for 2003-2007 at Del Norte Gaging Station, Rio Grande, Colorado (Units: 1000acre-feet)

Year	JAN 1st	FEB 1st	MAR 1st	APR 1st	MAY 1st	JUN1st	Observed
2003	372	290	315	285	230	255	235
2004	600	570	495	460	460	445	417
2005	640	770	770	770	785	795	666
2006	395	315	280	355	355	350	391
2007	555	555	490	415	385	450	580

3.2.2 Model Development

3.2.2.1 Input Variables Used in the Regression

The input variables used for forecast equation development were selected on the basis of variables that were used in NRCS forecast equations at the same site and some previous research results. The equations can be generally expressed as:

$$F_{APR-SEP} = \sum_{i=1}^{n'} \varphi_i SWE_i + \sum_{i=1}^{n'} \sum_{j=1}^{m'} \theta_{ij} PRCP_{ij} + C \qquad (3.5)$$

Where

$F_{APR\text{-}SEP}$ = Forecasted April-September runoff volume

SWE_i = snow water equivalent at SNOTEL site i

$PRCP_{ij}$ = the monthly total precipitation for month j and SNOTEL site i

n'= number of SNOTEL sites used in the forecast equation

m' = number of months from October to forecast date of a water year

φ_i, θ_{ij} - are coefficients

C = constant parameter

Table 3.3 shows the variables and coefficients of typical forecast equation being used currently by NRCS. This is a Jan 1st equation for the April-September flow at Rio Grande near Del Norte Gaging Station that has been developed by Z-score regression method using 41 years of data. The variables include snow water equivalent and October to forecast date monthly precipitation from six SNOTEL sites in the Basin. To compare the developed equations with NRCS forecasting equations, the same SNOTEL sites were used for forecast equation development.

Table 3.3 The variables and coefficients of January 1st NRCS forecast equation for April-September runoff volume for Rio Grande near Del Norte Gaging Station

Rio Grande nr Del Norte (2) ztp05r.713n41mx914			PUB DATE = 01/01/2008
ELEMENT	MONTH	SITE NAME	COEFFIENTS
SWE	Jan 1st	Upper Rio Grande	0.02217
SWE	Jan 1st	Upper San Juan	0.00889
SWE	Jan 1st	Middle Creek	0.02098
SWE	Jan 1st	Wolf Creek Summit	0.00839
SWE	Jan 1st	Molas Lake	0.00873
SWE	Jan 1st	Lily Pond	0.01788
PRCP	OCT	Upper Rio Grande	0.03479
PRCP	NOV	Upper Rio Grande	0.03479
PRCP	DEC	Upper Rio Grande	0.03479
PRCP	OCT	Upper San Juan	0.01352
PRCP	NOV	Upper San Juan	0.01352
PRCP	DEC	Upper San Juan	0.01352
PRCP	OCT	Middle Creek	0.01754
PRCP	NOV	Middle Creek	0.01754
PRCP	DEC	Middle Creek	0.01754
PRCP	OCT	Wolf Creek Summit	0.01306
PRCP	NOV	Wolf Creek Summit	0.01306
PRCP	DEC	Wolf Creek Summit	0.01306
PRCP	OCT	Molas Lake	0.02280
PRCP	NOV	Molas Lake	0.02280
PRCP	DEC	Molas Lake	0.02280
PRCP	OCT	Lily Pond	0.01925
PRCP	NOV	Lily Pond	0.01925
PRCP	DEC	Lily Pond	0.01925
C		INTERCEPT	6.04949

Two variable combinations were used for the development of PLSR and PCR models: variable combination-I and variable combination-II. The combination-I included the forecast date SNOTEL snow water equivalents, October to forecast date monthly SNOTEL precipitation, October to forecast date monthly SNOTEL average temperature index (TEMP), October to forecast date previous monthly flow of a water year, and previous year averaged October-December El Niño-Southern

Oscillation Index (SOI). Each data type except for SOI was measured from the six SNOTEL sites located in the Basin (See Figure 1.1). The monthly average basin temperature index was used as a temperature variable instead of using the temperature of each SNOTEL site because of data availability. The variables used in the variable combination-I are shown in Table 3.4.

Table 3.4 List of variables in variable combination-I

Variables	**Notation**	**Description**	**Number of variables used for monthly equation**					
			Jan	Feb	Mar	Apr	May	Jun
Snow water equivalent	SWE	Measured on the first day of a month	6	6	6	6	6	6
Precipitation Index	PRCP	October to forecast date precipitation for each SNOTEL site	18	24	30	36	42	48
Temperature Index	TEMP INDEX	October to forecast date monthly average temperature index	3	4	5	6	7	8
Previous flow	FLOW	October to forecast date monthly flow	3	4	5	6	7	8
Southern Oscillation Index	SOI	October to December average of previous year	1	1	1	1	1	1
Total number of variables used for variable selection in monthly equations			31	39	47	55	63	71

The variable combination-II is essentially the same as variable combination-I except the measured monthly precipitation for each month from October to forecast date was not used. Instead, the October to forecast date composite precipitation index for each SNOTEL Station was used as inputs for precipitation information. The composite PRCP index was calculated as a weighted average of monthly precipitation from October to the forecast date based on the correlation coefficients of the precipitation of a specific month with the April-September seasonal flow volume. The PRCP index for a SNOTEL site can be calculated by the following equation:

$$(PRCPINDEX)_i = \frac{\sum_{j=1}^{m'} r_{ij} PRCP_{ij}}{\sum_{j=1}^{m'} r_{ij}} \tag{3.6}$$

Where

(PRCPINDEX)$_i$ = the composite precipitation index for SNOTEL site *i;*

PRCP$_{ij}$ = the monthly total precipitation for month *j* and SNOTEL site *i;*

r_{ij} = the correlation coefficient of the precipitation of month *j* and SNOTEL site *i* with the seasonal flow volume;

m' = number of months from October to forecast date of a water year.

The calculated correlation coefficients between monthly precipitation from October to July, month-specific weighted average October-July precipitation index, and equal-weighted average October-July precipitation index of the SNOTEL sites in the Basin and the April-September seasonal flow are presented in Table 3.5. The composite PRCP index for each SNOTEL site can be calculated using the correlation coefficients and Equation 3.5, and then used as the inputs for each forecast equation development. It was observed that the higher correlations occurred in fall (October, November) and spring (April, May) than in winter season. This may be because the fall precipitation is especially important for setting up soil moisture and the spring precipitation may have a non-linear effect of having high runoff efficiency due to saturated soils from snowmelt (Pagano, personal communication, August 29, 2008). As can be seen, the correlation coefficient of month-specific weighted average precipitation with April-September flow were higher than the correlation coefficients of simple averaged precipitation with the flow for most of the SNOTEL sites. This indicates that a little more forecast skill could be obtained by developing a composite PRCP index and using it as the inputs in the regression (Pagano, personal communication, May 5, 2008). In addition, the number of input variables can be reduced dramatically when the information from more SNOTEL sites is included in the forecast equation development. The initial variables used in the variable combination-II are described in Table 3.6.

Table 3.5 Correlation coefficients of monthly and October-July average precipitation of SNOTEL sites with April-September natural flow at Del Norte Gaging Station, Rio Grande (1981-2002)

Snotel Site	OCT	NOV	DEC	JAN	FEB	MAR	APR	MAY	JUN	JUL	OCT-JUL	
											Weighted average	Simple average
Lily Pond	0.79	0.46	0.27	0.24	0.18	0.18	0.57	0.36	0.13	-0.2	0.91	0.88
Middle Creek	0.70	0.43	0.27	0.19	0.31	0.06	0.59	0.39	0.05	-0.1	0.91	0.91
Molas Lake	0.62	0.39	0.06	0.21	0.11	0.04	0.40	0.49	0.34	0.31	0.72	0.69
Upper SanJuan	0.71	0.45	0.34	0.20	0.11	0.13	0.58	0.45	0.27	0.42	0.77	0.79
Upper Rio Grande	0.70	0.50	0.18	0.21	0.27	0.00	0.43	0.19	0.06	0.02	0.89	0.87
Wolf Creek	0.73	0.35	0.06	0.23	0.05	0.16	0.40	0.65	0.27	0.41	0.82	0.70
Average	0.71	0.43	0.20	0.21	0.17	0.10	0.49	0.42	0.19	0.16	0.84	0.81

Table 3.6 List of variables in variable combination-II

Variables	Notation	Description	Number of variables used for monthly equation					
			Jan	Feb	Mar	Apr	May	Jun
Snow water equivalent	SWE	Measured on the first day of a month	6	6	6	6	6	6
Precipitation Index	PRCP INDEX	October to forecast date precipitation index for each SNOTEL site	6	6	6	6	6	6
Temperature Index	TEMP INDEX	October to forecast date monthly average temperature index	3	4	5	6	7	8
Previous flow	FLOW	October to forecast date monthly flow	3	4	5	6	7	8
Southern Oscillation Index	SOI	October to December average of previous year	1	1	1	1	1	1
Total number of variables used for variable selection in monthly equations			19	21	23	25	27	29

3.2.2.2 Variable Selection Procedure in PLSR

After formulating the procedures to determine the components to retain in PLSR (section 3.1.2) and preparing the variables that were used for model development in the previous section, the initial PLSR equations can be developed using all the variables in combination-I and combination-II. However, not all the predictor variables are important in the regression equations. Some variables can be dropped from the equation because of their insignificant relationship with the dependent variable and/or high collinearity with other independent variables to ensure optimality or near optimality of the regression equations. The significance of individual variables in PLSR could be examined by the magnitude of standardized coefficients. Wold (1994) proposed a technique in variable selection in PLSR using a variable influence on projection (VIP). The VIP is a weighted sum of squares of the partial least squares weights with the weights calculated from the amount of dependent variable variance of each partial least squares component. This statistic shows the contribution of each independent variable to the model and represents the value of each predictor in fitting the PLSR model for both predictors and responses. For a selected number of components and the initial variables, the VIP values of each predictor variable can be calculated and used to examine the strength of the relationship, irregularities, and the contribution of the independent variables in the model. Therefore, it can be used to select the most important variables.

The quality of the model can be evaluated by examining the residuals for both the response and the predictor variables for any possible outliers. To determine which predictor to be eliminated from the model, the regression coefficient and the variable importance for the projection (VIP) of each predictor should be analyzed. The regression coefficients represent the importance each predictor has in the prediction of the response. An independent variable may have a small coefficient value, but may have a large VIP, which implies that this independent variable is important and contributes significantly to the prediction and therefore, has to be kept in the model. Wold (1994) suggested that a VIP value of less than 0.8 is small. If a predictor has a relatively small coefficient (in absolute value) and a small value of VIP (less than 0.8), then it is a prime candidate for deletion (Umetrics, Inc., 1995).

In this study, the variable selection was carried out by analyzing the VIP of each predictor and considering the forecast consistency of the regression equations from month to month. The forecast inconsistency, an unexpected variability from month to month, could be caused by the usage of different variables and coefficients in month to month regression models. Acceptable forecast change should occur due to hydrologic reasons, not statistical noise (Garen, 1992). To ensure the forecast consistency from month to month, highly similar predictor variables in the regression models for different months can be used without losing forecast accuracy. To achieve this goal, the following procedure was used in this study:

1) Calculate VIP values of all the predictor variables for the PLSR regression equations from January 1st to June 1st;
2) Select the variables with VIP values greater than 0.8 for all the forecast equations from January 1st to June 1st ;
3) Calculate VIP values again for the selected variables for each regression equation starting from January 1st and finalize the variables for the January 1st equation. Once the variables for the January 1st equation are selected, they will be kept in the all regression equations through June 1st, even some of the variables have VIP values smaller than 0.8;
4) Based on the VIP values of variables for the February 1st equation, new variables could be added. The added variables will be kept in the following month's equation even though they have VIP values smaller than 0.8, and so on. For example, the January equation variables will be kept in the February

equation, and all the February equation variables will be kept in March equation, and so on.

3.2.2.3 Variable Selection Procedure in PCR

McCuen (1985) discussed the excess variable elimination procedure in principal components regression through examining the magnitudes of the eigenvector values for each variable. But the methodology requires subjective judgment and a considerable amount of work when there are many input variables. To overcome these drawbacks, Garen (1992) proposed a method of systematic searching for optimal variable combinations in PCR. This methodology has been used by NRCS since then to create seasonal streamflow forecasting equations using a semi-automated approach from a pool of available input variables (Risley et al., 2005). The search algorithm can be performed with VIPER, an Excel-based program used by NRCS, including searching for optimum combinations of independent variables, searching for optimum time periods covered by selected independent variables, and jackknife testing of models (NRCS, 2007). In this study, the selected variables in PLSR were directly used in PCR. The final PLSR and PCR equations were developed using the same variables as to compare the performance of both methods in seasonal streamflow forecasting.

3.2.3 Model Testing and Comparison

3.2.3.1 Comparison of Models for Two Variable Combinations

Based on the procedures described in the previous sections, forecasting equations were calibrated using data from 1981 to 2002 (22 years) with the jackknife cross validation scheme. For each variable combination, PLSR and PCR equations were developed. A variable selection was carried out by examining the VIP of each variable and considering forecast consistency from month to month in PLSR. No variable selection was performed in PCR during the calibration since the same variables selected for PLSR were used in PCR equation development.

Model adequacy was tested by analyzing the residuals of calibration periods for all equations. The residuals can provide information about the prediction capability of models over the range of dependent variable. It also can be used to identify an improperly formulated model. The student's t-test was used to test if the residuals had zero mean, the Shapiro-Wilk normality test was used to test if residuals were normally distributed and the Ljung-Box white noise test was applied to test any autocorrelations existing in the residuals or if the residuals were a white noise process. These statistical tests were used to examine the residuals of all forecast equations for two variable combinations. The results showed that residuals had zero mean, were normally distributed and were a white noise process statistically at 0.05 significance level. This suggests that all the models were adequate with accepted predictive ability.

Consider the April 1st PLSR equation with variable combination-II as an example. The *p*-value of student's t-test for zero mean was 0.996, the *p*-value of the Shapiro-Wilk normality test was 0.946, the *p*-value of the Ljung-Box white noise test up to lag 6 was 0.726. The normal probability plot of residuals for this model is shown in Figure 3.4. All the statistical tests and normality plots suggested that no modelable information was left in the residuals. Similar results were obtained for all other forecast equations in the study.

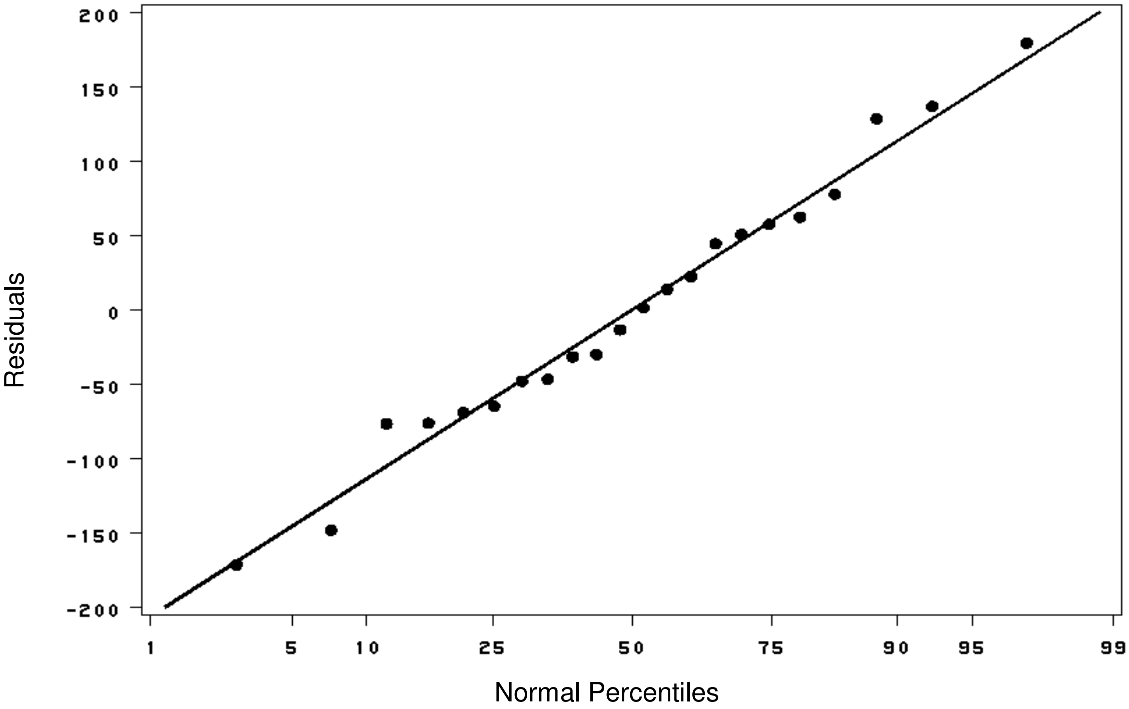

Figure 3.4 Normality plot of residuals of April 1st PLSR equation with variable combination-II (1981-2002)

The calibrated PLSR and PCR forecast equations for two variable combinations can be expressed in following general forms:

For variable combination-I:

$$F_{APR-SEP} = \sum_{i=1}^{n'} \varphi_i SWE_i + \sum_{i=1}^{n'} \sum_{j=1}^{m'} \theta_{ij} PRCP_{ij} + \sum_{j=1}^{m'} (\omega_j FLOW_j + \psi_j TEMPINDEX_j) + C \qquad (3.7)$$

For variable combination-II:

$$F_{APR-SEP} = \sum_{i=1}^{n'} \varphi_i SWE_i + \sum_{i=1}^{n'} \theta_i PRCPINDEX_i + \sum_{j=1}^{m'} (\omega_j FLOW_j + \psi_j TEMPINDEX_j) + C \qquad (3.8)$$

Where

$FLOW_j$ = monthly flow at month j

$TEMPINDEX_j$ = monthly temperature index at month j

θ_i ,ω_i,ψ_i - are coefficients

The coefficients of variables of developed forecast equations are shown in Table 3.7 - Table 3.10. The final maximum number of variables of forecast equations with variable combination-I (June 1st forecast equation) was 24, and with variable combination-II (May 1st equation) was 18. It can be observed that there is no significant difference in the magnitude of variable coefficients of PLSR and PCR equations. The results indicated that the developed forecast equations could very likely improve forecast accuracy compared to NRCS current forecasting equations at the study site. As shown in Table 3.9, the calibration coefficient of determination (R^2) of January 1st to April1st equations developed in this study (calibration period 22 years) were higher than calibration of R^2 of NRCS equations (calibration period 41 years). For example, the calibration R^2 of January 1st NRCS equation at Del Norte

Gaging Station is 0.51, while PLSR equation is 0.77. Similar results were obtained for other forecast equations developed by PCR.

The application of effective precipitation index also facilitated the developing of more parsimonious regression model with fewer input variables as compared to NRCS current forecasting equations at the study site. Compared to the same site NRCS equation (as shown in Table 3.3 and Eq. 3.5), the PLSR equation with variable combination-II (Table 3.9) has 14 variables including SWE, PRCP index, temperature index and previous flow, while NRCS equation consist of 24 variables such as SWE and monthly PRCP. This feature is more evident for the later forecast dates without losing forecast accuracy. For example, the current NRCS April 1st forecast equation at Del Norte Gaging Station had 42 predictor variables with the calibration R^2 of 0.73. In comparison, the number of variables in the proposed April 1st PLSR forecast equation was 17 and calibration R^2 was 0.84 (Table 3.9).

Table 3.7 Forecast equations from Jan 1st to June 1st for Del Norte Gaging Station developed by PLSR using variable combination-I

VARIABLES	**CATEGORY**	**Coefficients**					
		JAN 1st	**FEB 1st**	**MAR 1st**	**APR 1st**	**MAY 1st**	**JUN 1st**
Constant	C	75.05	42.74	-74.70	-146.70	-149.47	-144.67
Lily Pond	SWE	4.57	4.12	2.72	2.28	3.02	*
Middle Creek	SWE	4.65	4.42	4.00	3.68	3.29	2.37
Upper SanJuan	SWE	2.39	2.08	1.69	1.57	2.00	1.67
Wolf Creek	SWE	2.89	2.67	2.22	1.85	1.94	1.44
Upper Rio Grande	SWE	8.45	6.45	5.54	4.67	13.16	*
Molas Lake	SWE	4.80	3.48	2.80	2.46	2.39	2.39
Middle Creek-10	PRCP	5.91	6.02	6.35	6.13	1.54	1.99
Upper Rio Grande-10	PRCP	9.26	9.43	9.94	9.60	4.47	3.91
Wolf Creek-10	PRCP	6.02	6.14	6.47	6.24	2.46	3.15
Upper SanJuan-10	PRCP	5.85	5.96	6.28	6.07	2.16	2.07
Lily Pond-10	PRCP	11.00	11.21	11.82	11.41	7.34	5.53
Molas Lake-10	PRCP	9.13	9.30	9.80	9.46	0.59	3.43
Temperature-10	TEMP INDEX	-10.80	-11.01	-11.60	-11.20	-5.51	-4.63
Flow-11	FLOW	2.25	2.30	2.42	2.34	1.55	1.55
Flow-12	FLOW	3.39	3.46	3.64	3.52	2.65	2.17
Flow-2	FLOW			6.45	6.23	7.99	6.50
Flow-3	FLOW				2.97	2.51	2.49
Middle Creek-4	PRCP					3.80	3.38
Upper SanJuan-4	PRCP					1.48	1.46
Lily Pond-4	PRCP					2.84	3.55
Flow-4	FLOW					0.76	0.20
Middle Creek-5	PRCP						9.59
Upper SanJuan-5	PRCP						8.02
Lily Pond-5	PRCP						12.74
Flow-5	FLOW						0.16
Number of components used		1	1	1	1	2	2
Calibration R^2		0.81	0.82	0.85	0.86	0.93	0.93
Jackknife cross validation R^2		0.76	0.77	0.81	0.82	0.86	0.87
Cross validation RMSE (kaf)		106	104	94	91	77	61
Cross validation NRMSE		0.48	0.47	0.43	0.42	0.36	0.36
Number of years for calibration		22	22	22	22	22	22

** Snow water equivalent is zero at that site on the specific month, or variables are not included.*

Table 3.8 Forecast equations from Jan 1st to June 1st for Del Norte Gaging Station developed by PCR using variable combination-I

VARIABLES	CATEGORY	Coefficients					
		JAN 1st	FEB 1st	MAR 1st	APR 1st	MAY 1st	JUN 1st
Constant	C	73.74	51.52	-47.07	-128.19	-115.55	-114.59
Lily Pond	SWE	4.23	3.90	2.68	2.17	2.93	*
Middle Creek	SWE	4.49	4.20	3.75	3.45	2.47	2.16
Upper SanJuan	SWE	2.12	1.89	1.59	1.41	1.85	1.45
Wolf Creek	SWE	2.70	2.46	2.06	1.82	2.03	1.18
Upper Rio Grande	SWE	8.42	6.60	5.75	4.39	8.82	*
Molas Lake	SWE	4.73	3.10	2.54	2.09	2.59	2.42
Middle Creek-10	PRCP	6.48	6.81	7.18	7.01	2.72	3.59
Upper Rio Grande-10	PRCP	9.58	10.19	10.79	9.84	3.37	4.07
Wolf Creek-10	PRCP	6.17	6.52	6.86	6.55	3.05	3.51
Upper SanJuan-10	PRCP	6.14	6.28	6.53	6.34	2.38	3.54
Lily Pond-10	PRCP	11.34	11.33	11.71	11.48	5.34	6.69
Molas Lake-10	PRCP	9.94	10.71	11.12	10.18	4.22	4.10
Temperature-10	TEMP INDEX	-9.46	-9.00	-9.61	-10.31	-10.07	-7.37
Flow-11	FLOW	2.22	2.02	2.16	2.39	1.91	1.77
Flow-12	FLOW	3.54	3.28	3.55	3.71	1.98	2.51
Flow-2	FLOW			4.71	5.20	6.14	4.90
Flow-3	FLOW				2.96	2.88	2.04
Middle Creek-4	PRCP					2.41	1.82
Upper SanJuan-4	PRCP					2.83	1.31
Lily Pond-4	PRCP					3.59	2.52
Flow-4	FLOW					0.66	0.39
Middle Creek-5	PRCP						8.71
Upper SanJuan-5	PRCP						7.43
Lily Pond-5	PRCP						11.47
Flow-5	FLOW						0.09
Number of components used		1	1	1	1	1	3
Calibration R^2		0.80	0.80	0.84	0.85	0.92	0.92
Jackknife cross validation R^2		0.76	0.77	0.77	0.81	0.83	0.89
Cross validation RMSE (kaf)		106	104	103	95	95	69
Cross validation NRMSE		0.48	0.48	0.47	0.43	0.43	0.33
Number of years for calibration		22	22	22	22	22	22

** Snow water equivalent is zero at that site on the specific month, or variables are not included.*

Table 3.9 Forecast equations from Jan 1st to June 1st for Del Norte Gaging Station developed by PLSR using variable combination-II

VARIABLES	**CATEGORY**	**Coefficients**					
		JAN 1st	**FEB 1st**	**MAR 1st**	**APR 1st**	**MAY 1st**	**JUN1st**
Constant	C	52.19	19.37	-122.9	-209.6	-169.9	-170.37
Lily Pond	SWE	3.99	0.79	2.32	1.97	2.08	n/a
Middle Creek	SWE	4.06	4.77	3.41	3.19	2.24	1.96
Upper SanJuan	SWE	2.08	0.87	1.44	1.36	1.27	1.15
Wolf Creek	SWE	2.52	2.21	1.90	1.60	1.43	1.09
Upper Rio Grande	SWE	7.37	1.78	4.73	4.05	7.32	n/a
Molas Lake	SWE	4.19	2.22	2.39	2.13	1.65	1.59
Lily Pond	PRCP INDEX	13.06	18.64	18.06	18.57	17.09	17.57
Middle Creek	PRCP INDEX	9.70	9.08	13.64	13.99	12.81	13.07
Molas Lake	PRCP INDEX	12.27	2.09	15.56	16.17	13.77	14.10
Upper SanJuan	PRCP INDEX	8.75	11.44	11.21	11.55	17.28	10.95
Upper Rio Grande	PRCP INDEX	15.30	16.01	19.61	19.24	8.14	16.93
Wolf Creek	PRCP INDEX	7.05	5.05	8.52	9.16	10.50	9.05
Temperature-10	TEMP INDEX	-9.42	-33.99	-9.90	-9.71	-8.64	-7.84
Flow-11	FLOW	1.96	5.09	2.06	2.03	1.74	1.61
Flow-12	FLOW	2.96	5.66	3.11	3.05	2.60	2.36
Flow-2	FLOW			5.51	5.40	4.59	4.13
Flow-3	FLOW				2.58	2.23	2.06
Flow-4	FLOW					0.49	n/a
Flow-5	FLOW						0.15
Number of components used		1	2	1	1	1	1
Number of components by min PRESS		1	2	2	1	1	1
Calibration R^2		0.77	0.85	0.82	0.84	0.92	0.92
NRCS Equation Calibration R^2 (n=41)		***0.51***	***0.57***	***0.60***	***0.73***	***-***	***-***
Jackknife cross validation R^2		0.71	0.74	0.76	0.79	0.90	0.90
Cross validation RMSE (kaf)		117	111	105	98	66	52
Cross validation NRMSE		0.53	0.51	0.48	0.45	0.31	0.31
Number of years for calibration		22	22	22	22	22	22

** Snow water equivalent is zero at that site on the specific month, or variables are not included.*

Table 3.10 Forecast equations from Jan 1^{st} to June 1^{st} for Del Norte Gaging Station developed by PCR using variable combination-II

VARIABLES	CATEGORY	Coefficients					
		JAN 1^{st}	FEB 1^{st}	MAR 1^{st}	APR 1^{st}	MAY 1^{st}	JUN1^{st}
Constant	C	48.89	-52.85	-222.52	-207.66	-170.53	-176.49
Lily Pond	SWE	4.28	1.85	1.06	2.29	2.04	n/a
Middle Creek	SWE	3.93	3.53	3.02	3.00	2.10	1.87
Upper SanJuan	SWE	2.14	0.63	0.74	1.47	1.26	1.06
Wolf Creek	SWE	2.61	1.86	1.41	1.72	1.43	1.03
Upper Rio Grande	SWE	7.39	5.10	4.24	4.21	6.29	n/a
Molas Lake	SWE	4.65	1.56	1.62	2.35	1.71	1.49
Lily Pond	PRCP INDEX	13.27	16.79	18.44	18.00	16.92	17.21
Middle Creek	PRCP INDEX	9.79	13.32	14.82	13.92	12.76	13.01
Molas Lake	PRCP INDEX	12.69	12.43	13.95	16.08	14.61	14.54
Upper SanJuan	PRCP INDEX	8.70	9.45	10.65	11.07	16.95	10.70
Upper Rio Grande	PRCP INDEX	14.61	20.14	21.28	17.90	8.69	17.26
Wolf Creek	PRCP INDEX	6.85	8.17	8.45	9.26	10.50	9.39
Temperature-10	TEMP INDEX	-7.01	-12.85	-10.76	-7.57	-8.76	-8.24
Flow-11	FLOW	1.78	4.17	3.74	1.84	1.73	1.64
Flow-12	FLOW	2.84	7.26	6.66	2.87	2.57	2.50
Flow-2	FLOW			11.34	4.62	4.34	4.08
Flow-3	FLOW				2.82	2.50	2.35
Flow-4	FLOW					0.46	n/a
Flow-5	FLOW						0.16
Number of components used		1	2	2	1	1	1
Number of components by min PRESS		1	5	4	1	1	1
Calibration R^2		0.76	0.81	0.84	0.83	0.92	0.92
NRCS Equation Calibration R^2 (n=41)		***0.51***	***0.57***	***0.60***	***0.73***	***-***	***-***
Jackknife cross validation R^2		0.71	0.73	0.78	0.79	0.90	0.90
Cross validation RMSE (kaf)		117	112	101	97	65	51
Cross validation NRMSE		0.53	0.51	0.46	0.44	0.31	0.30
Number of years for calibration		22	22	22	22	22	22

** Snow water equivalent is zero at that site on the specific month, or variables are not included.*

PLSR and PCR models were compared with each other using their performance in calibration phase and jackknife cross validation scheme for two variable combinations. Some performance statistics such as coefficient of determination for calibration and cross validation, RMSE, normalized RMSE were also tabulated in Tables 3.7 through 3.10. It is evident that, there was no significant difference in performance of the forecast equations using different variable combinations. The cross validation RMSE of forecast equations for January 1st, February 1st , March 1st and April 1st with variable combination-I were somewhat smaller than that of variable combination-II. On the other hand, the May 1st and June 1st forecast equations with combination-I have larger cross validation RMSE than that of combination-II. However, the forecast equations calibrated using variable combination-II has, at least, fewer variables in the equations. This approach is preferred because of the parsimonious feature of models by reducing the input variables considerably using composite PRCP index without loss of accuracy. This is particularly important for the forecast equation development in larger basins where more information is available from numerous SNOTEL sites. The comparison of forecasting performance of PLSR and PCR using variable combination-II is further discussed by applying the models for new test data set in the following sections of the chapter.

In general, there is a reduction in model errors as the period in time between the forecast date and the actual forecast period is narrowed. As indicated in Tables 3.7 through 3.10, both PLSR and PCR have the same general behavior for different variable combinations. The cross validation RMSE decreased from highest errors in January 1st forecast equations to lowest errors in June 1st equations. For instance, the normalized RMSE of January 1st PLSR equation decreased from 0.53 to 0.31 for the June 1st PLSR equation for variable combination-II. Similar trends were observed in PCR forecasting equations.

3.2.3.2 Comparison of PLSR and PCR Methods

The performances of PLSR and PCR using two variable combinations are shown in Tables 3.7 and 3.10. Based on the performance statistics both for calibration phase and jackknife cross validation, no significant difference in model performance could be observed between PLSR and PCR. For some months, the PCR performed slightly better than PLSR in terms of cross validation coefficient of determination and RMSE. However, it can be observed that there was a difference in the number of components used in PLSR and PCR. The explained variation (coefficient of determination for calibration) of the dependent variable by PLSR is higher than PCR for most cases. This was the result of the unique feature of PLSR that extracts components based on the covariance between predictor and dependent variables, which can explain more variations of dependent variables than PCR does.

The results also showed that the PLSR reaches its minimal prediction error with a smaller number of components than PCR. This is a unique feature of PLSR compared to PCR when developing the regression equation. Tables 3.9 and 3.10 illustrate the difference of PLSR and PCR forecast equations in extracting optimal components based on minimum prediction error using jackknife cross validation. There is a difference in the number of components that has minimal PRESS in PLSR and PCR. The PLSR can reach minimum prediction error using 2 components for February 1st and March 1st equations, whereas the PCR can reach minimum prediction errors with 5 and 4 components respectively. Figure 3.5 illustrates the difference of April 1st PLSR and PCR forecast equations in extracting optimal components based on minimum prediction error and van der Voet's test using jackknife cross validation for variable combination-I. It can be seen that while the number of components suggested by van der Voet's test for both methods is 1, the number of components that has minimal PRESS in PLSR and PCR are not same. The PLSR can reach minimum prediction error using 6 components, whereas the PCR can reach minimum prediction error with 13 components. This may be also because the PLSR can extract components based on the covariance between predictor and dependent variables, so PLSR to have a stronger power in extracting components compared to PCR. This feature was also reported in the literature (Yeniay and Göktaş, 2002; SAS Institute, 2008). In general, the PLSR method is more powerful than PCR in extracting components that deal with the collinearity issue.

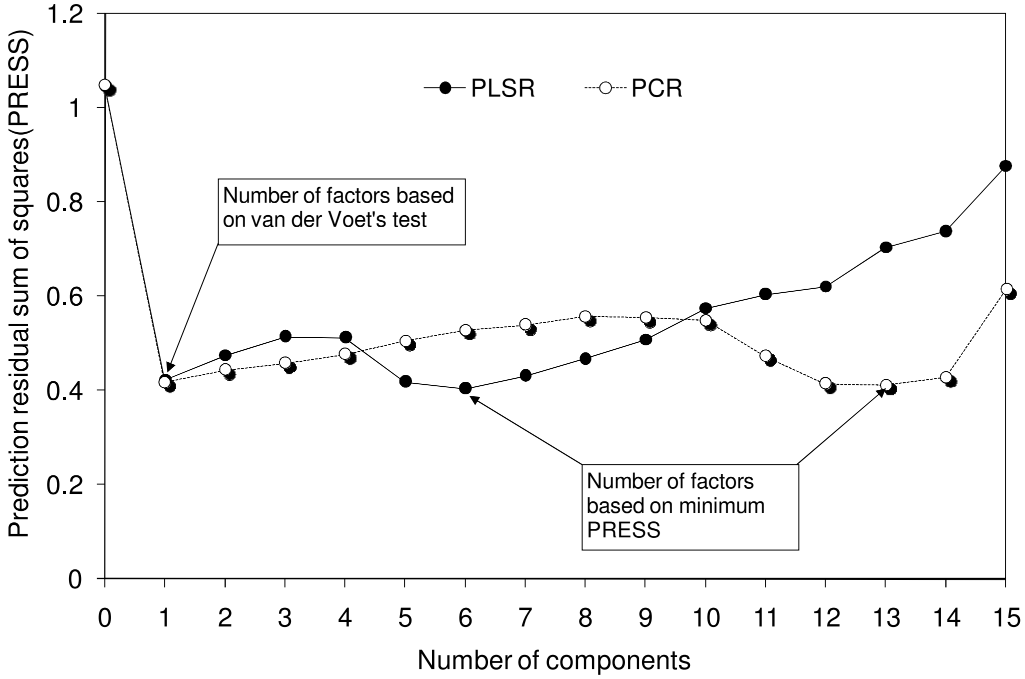

Figure 3.5 Optimum number of components extracted by PLSR and PCR for April1st forecasting equation development for Del Norte Gaging Station

3.2.3.3 Comparison of Models for Testing Data

The models were tested for new data from 2003 to 2007 to examine how well the forecasts of the PLSR and PCR models performed on new test data compared to NRCS official forecasts. Since the test data set was very short, conclusive results on the comparison of the models could not be obtained, but at least it gave an insight between the performance of PLSR and PCR, and effects of composite precipitation index inputs on forecasting accuracy. In order to compare the performance of models to NRCS official forecasts, the April-September flow volume forecast for all forecast dates was used. To compute the April-September forecast volume by May 1st and

June 1st equations, the observed April flow was added to May-September volume forecasts of May 1st equations; the observed April and May flows were added to the June-September forecasts of June 1st forecasting equations.

Model performances were tested by examining the model forecasts made at each forecast date from 2003 and 2007. Figure 3.6 shows how the forecasts evolved through forecast dates. For 2003, the NRCS forecasts were better for all forecast dates compared to both PLSR and PCR. For 2004 and 2007, the PLSR and PCR forecasts were better or equivalent compared to NRCS forecasts for all the forecast dates. The comparison results of the model forecasts with NRCS official forecasts was encouraging since the PLSR has showed some potential ability in both modeling procedure and improving forecast accuracy.

When comparing the PLSR and PCR methods, they were usually of similar performance and no significant difference was observed between PLSR and PCR forecasts. The forecasts of PLSR and PCR equations for all forecast dates were of similar performance, although the PLSR equations tended to do slightly better than PCR in some cases. As discussed earlier, no significant difference in model performance was observed between PLSR and PCR based on the jackknife cross validation performance statistics. This implies that although the PLSR method has higher calibration R^2 than PCR and is more powerful than PCR in extracting components that deal with the collinearity issue, yet it may not necessarily guarantee that PLSR would be better than PCR in terms of forecasting accuracy in seasonal streamflow forecasting within the scope of this study.

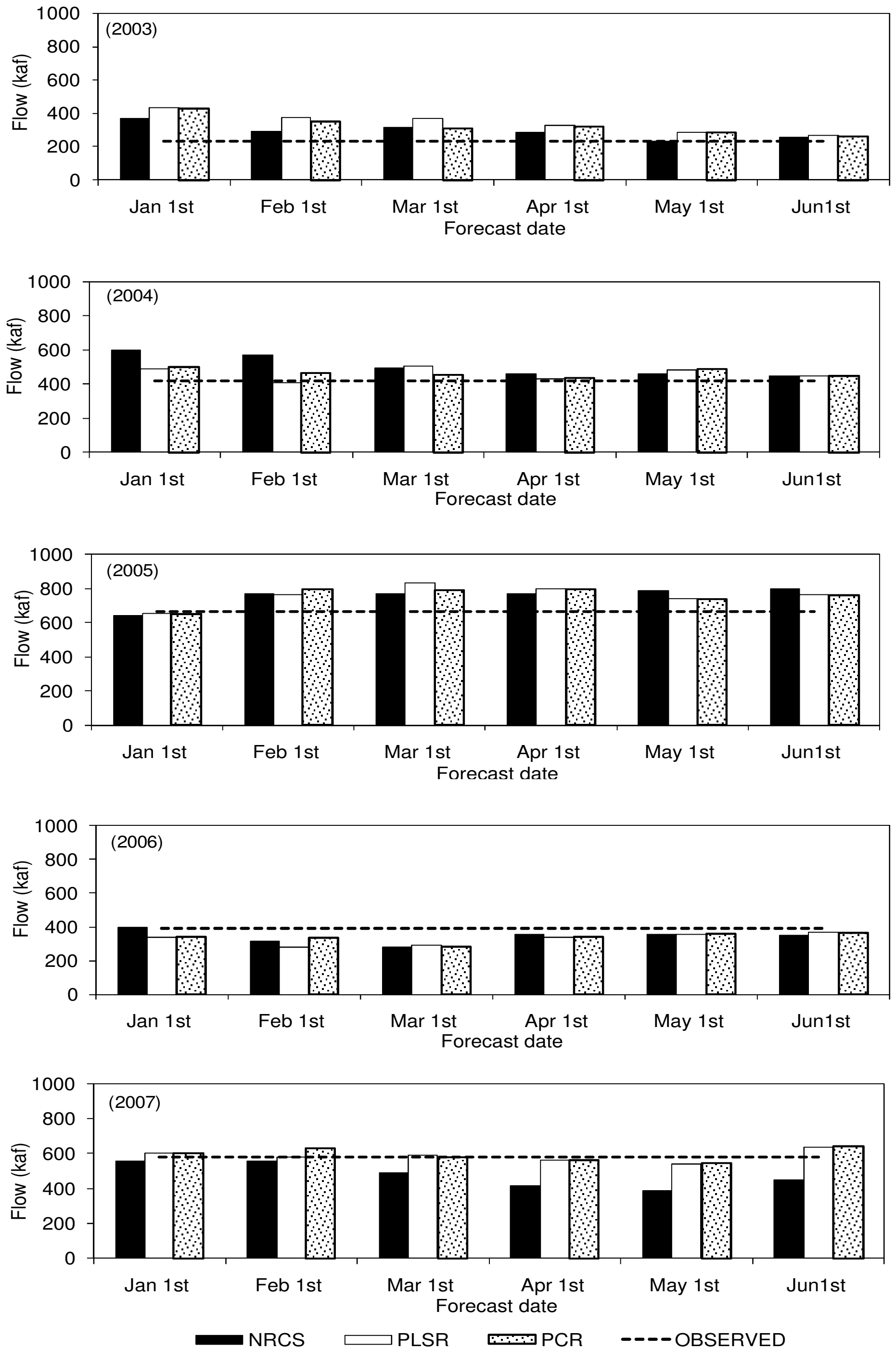

Figure 3.6 Comparison of April-September PLSR, PCR model forecasts with NRCS official forecasts at Del Norte Gaging Station for 2003-2007

3.2.4 Final Model

The discussions in the previous sections with respect to the comparison of different models and two variable combinations suggested that the performance of partial least squares regression is similar to principal components regression. However, it was also observed that PLSR reaches its minimum prediction error with a smaller number of components and can account for more variation of dependent variables with the same components compared to PCR. This is a unique feature of PLSR compared to PCR when developing regression equations. Based on the previous analysis and purpose of this dissertation study, it is believed that a final forecast model should be developed using the best-performing variable combination, methodology, and longest available calibration data period, so that the model could be used in practical applications.

Considering all the factors that have been discussed in the previous sections, the following final model was developed with variable combination-II (composite precipitation index as inputs) using data from 1981 to 2007 (27 years) with the partial least squares regression modeling approach. The final forecast equation is shown in Table 3.11. The model performance was evaluated using the jackknife cross validation scheme and the results were also shown in Table 3.11. It is hoped that the proposed final model can be used for April-September seasonal natural flow forecasting at Del Norte Gaging Station, Rio Grande, Colorado.

Table 3.11 Final PLSR forecasting equation for Del Norte Gaging Station, Rio Grande, Colorado

VARIABLES	CATEGORY	Coefficients					
		JAN 1st	FEB 1st	MAR 1st	APR 1st	MAY 1st	JUN1st
Constant	IC	29.2	-1.3	-114.5	-204.9	-177.2	-169.7
Lily Pond	SWE	4.21	3.59	2.41	2.09	1.98	*
Middle Creek	SWE	4.06	3.57	3.16	3.24	2.23	2.01
Upper SanJuan	SWE	2.03	1.54	1.18	1.21	1.19	1.09
Wolf Creek	SWE	2.57	2.09	1.66	1.58	1.44	1.06
Upper Rio Grande	SWE	7.80	5.43	5.02	4.07	7.58	*
Molas Lake	SWE	4.33	2.93	2.37	2.11	1.65	1.67
Lily Pond	PRCP INDEX	13.82	15.77	18.05	19.27	17.64	17.73
Middle Creek	PRCP INDEX	10.21	11.06	13.63	14.20	13.36	13.16
Molas Lake	PRCP INDEX	11.25	12.88	14.33	15.08	13.25	13.38
Upper SanJuan	PRCP INDEX	8.84	9.62	10.29	11.10	17.82	10.79
Upper Rio Grande	PRCP INDEX	15.99	16.95	19.84	19.65	7.94	16.70
Wolf Creek	PRCP INDEX	6.73	7.17	7.53	8.33	10.70	8.49
Temperature-10	TEMP INDEX	-7.68	-7.70	-7.84	-7.77	-7.03	-6.52
Flow-11	FLOW	2.07	2.08	2.11	2.09	1.82	1.68
Flow-12	FLOW	3.26	3.27	3.33	3.29	2.85	2.59
Flow-2	FLOW			5.48	5.43	4.71	4.21
Flow-3	FLOW				2.08	1.81	1.59
Flow-4	FLOW					0.53	*
Flow-5	FLOW						0.15
Number of components used		1	1	1	1	1	1
Calibration R^2		0.76	0.76	0.79	0.84	0.92	0.91
Jackknife cross validation R^2		0.70	0.70	0.74	0.80	0.90	0.89
Cross validation RMSE (kaf)		114	113	106	93	63	52
Cross validation NRMSE		0.54	0.54	0.50	0.44	0.31	0.32
Number of years for calibration		27	27	27	27	27	27

** Snow water equivalent is zero at that site on the specific month, or variables are not included.*

3.3 Elephant Butte Net Inflow

3.3.1 Data Description

3.3.1.1 Seasonal Net Inflow Volume

March-July seasonal Elephant Butte Reservoir net inflow, Rio Grande, New Mexico was selected as the forecast target volume. The NRCS does not provide seasonal net inflow forecasting for Elephant Butte Reservoir, but NRCS does issue March-July natural seasonal volume forecasts at San Marcial Gaging Station, Rio Grande, which is located at the entrance of Elephant Butte Reservoir. The San Marcial seasonal volume forecasts are very important and may be comparable to the Elephant Butte Reservoir net inflow, since it is the main inflow to the reservoir. The correlation coefficient of San Marcial March-July natural flow and Elephant Butte Reservoir March-July measured net inflow is 0.98 (calculated for the period of 1961-2000), which indicates the importance of NRCS forecasts at the site. To be consistent with the NRCS forecast dates and volume at San Marcial Gaging Station, the March-July, April-July and May-July volumes were used as the dependent variables in the seasonal net inflow modeling.

The data period from January, 1981 to September, 2007 was used in this study. To calibrate the PLSR and PCR equations, the total period was divided into two data sets: the calibration data set (1981-2002) and test data set (2003-2007). Since the data set for the calibration phase was very short (only 22 years of data), the jackknife (leave-one-out) cross validation procedure was used to validate the equation. The test data was used for testing the forecast equations. Two statistical features of the seasonal net inflow data, the normality and autocorrelation, were examined to see if data is normally distributed and if there is autocorrelation existing in the data. Three dependent variables, the March-July, April-July and May-July net inflow volume, and data period of 1981-2007 were used to conduct these tests. The normality test was performed using Shapiro-Wilk normality test, the autocorrelation test was conducted using the Ljung-Box white noise test up to six lags. The test results are shown in Table 3.12. It is suggested that the normality distribution of seasonal flows be accepted except for May-July flow and there were no significant autocorrelations existing in all the seasonal flows at 0.05 significance level. They were essentially a white noise process. The normality plot of the data (Figures 3.7-3.9) also indicated that no data transformation would be needed for modeling.

Table 3.12 Normality and autocorrelation tests for Elephant Butte Reservoir seasonal net inflow (1981-2007)

Flow	Shapiro-Wilk test of normality		Ljung-Box white noise test		
	W Statistic	*p*-value	To Lag	Chi-Square statistic	*p*-value
March-July	0.937	0.102	6	9.090	0.169
April-July	0.930	0.071	6	8.450	0.207
May-July	0.905	0.018	6	8.710	0.190

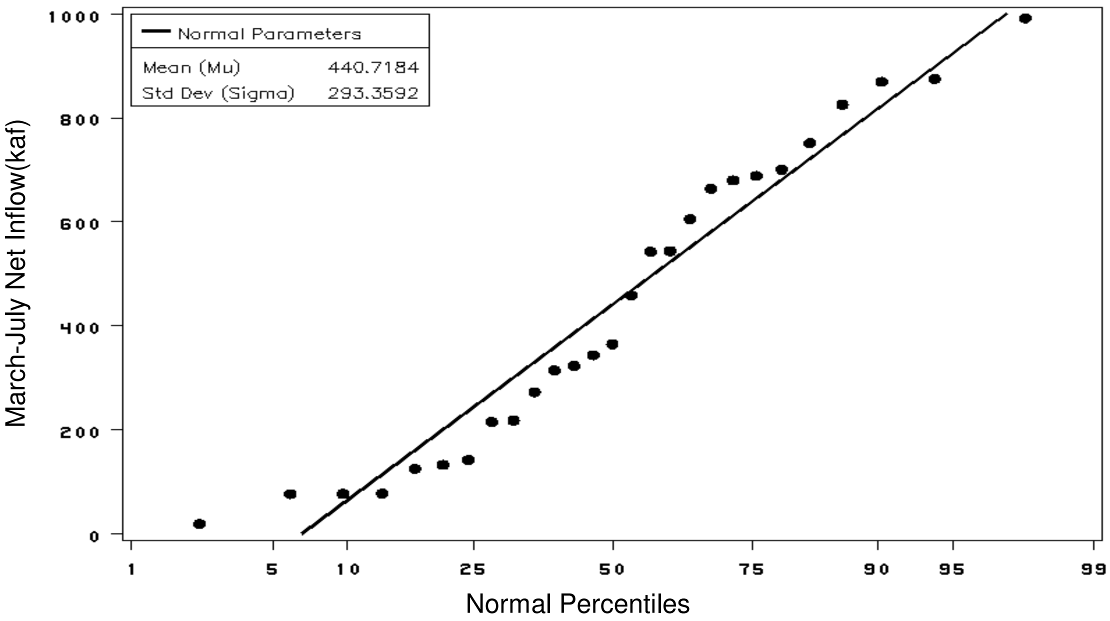

Figure 3.7 Normality plot of March-July net inflow of Elephant Butte Reservoir, Rio Grande (1981-2007)

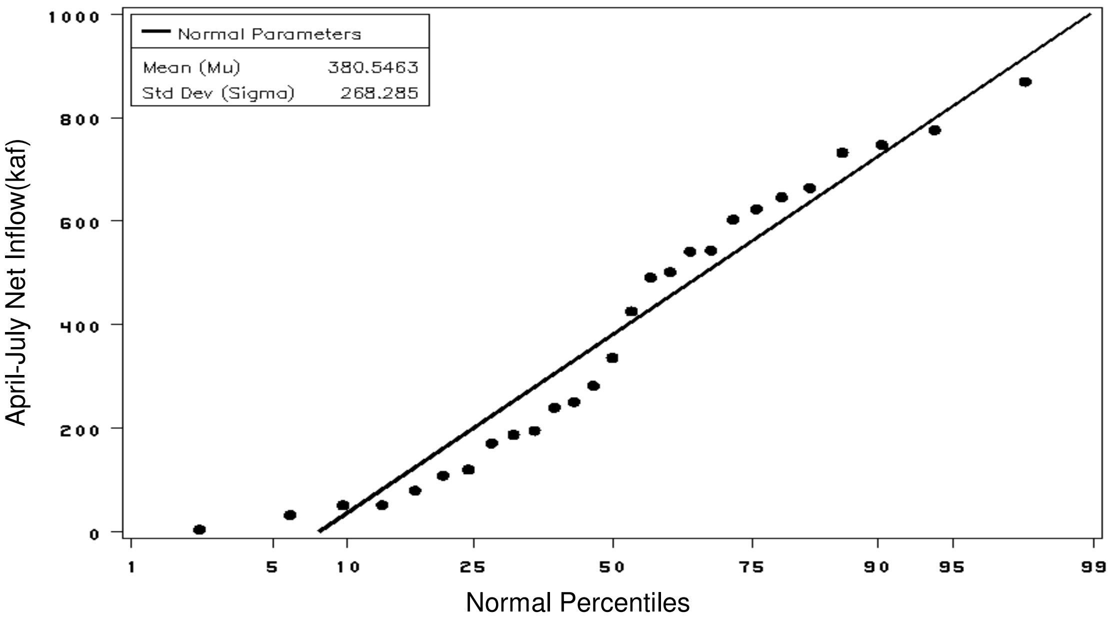

Figure 3.8 Normality plot of April-July net inflow of Elephant Butte Reservoir, Rio Grande (1981-2007)

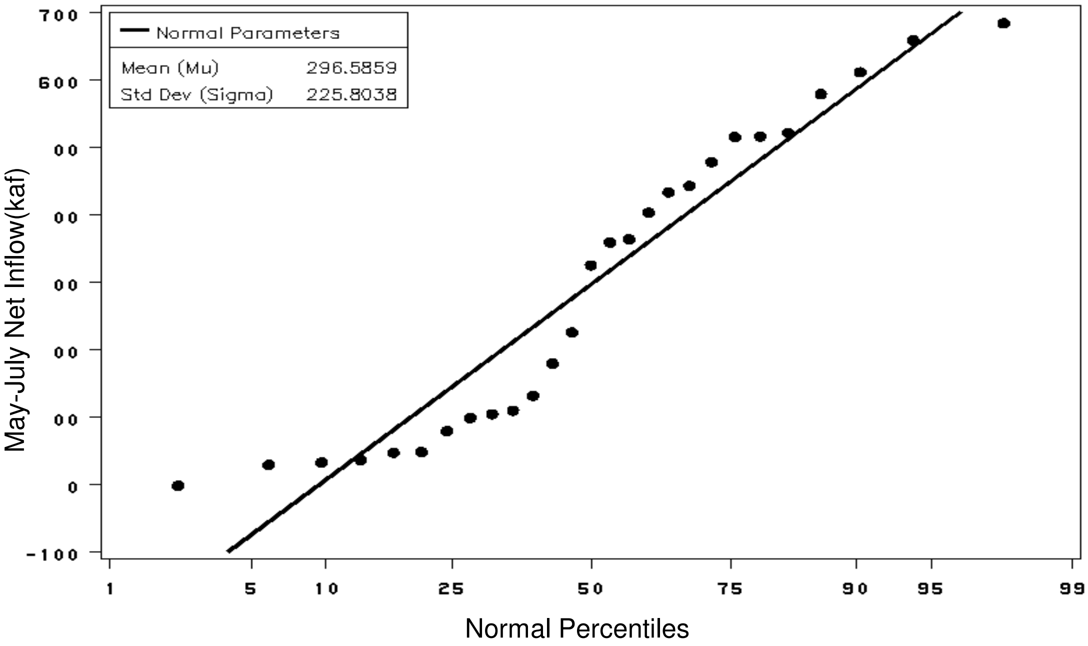

Figure 3.9 Normality plot of May-July net inflow of Elephant Butte Reservoir, Rio Grande (1981-2007)

3.3.1.2 Other Data

Similar to the seasonal flow model at Del Norte Gaging Station, five categories of data were used in PLSR and PCR model development. They included SNOTEL snow water equivalent that was measured on the first day of each month from January to May, the monthly SNOTEL precipitation from October to April of a water year, monthly SNOTEL average temperature index from October to April of a water year, monthly flow from October to April of a water year, and previous year averaged October-December El Niño-Southern Oscillation Index (SOI). The data period covers from 1981 to 2007. The data description has been given in section 1.4.3. Additionally, NRCS official forecast data for San Marcial Gaging Station, Rio

Grande were obtained from the NRCS (Pagano, personal communication, April 10, 2008) and used for comparison with PLSR and PCR equations. The March-July official seasonal volume forecasts for 2003-2007 at San Marcial Gaging Station issued by NRCS are shown in Table 3.13.

Table 3.13 March-July flow NRCS official forecasts at different forecast dates for 2003-2007 at San Marcial Gaging Station, Rio Grande (Units: 1000acre-feet)

Year	JAN 1st	FEB 1st	MAR 1st	APR 1st	MAY 1st
2003	370	260	305	305	245
2004	455	470	420	385	400
2005	400	675	760	930	950
2006	255	145	57	100	65
2007	520	570	440	265	255

3.3.2 Model Development

3.3.2.1 Input Variables Used in the Regression

Based on the variables used in current NRCS forecast equation at San Marcial Gaging Station, modeling results of Del Norte Gaging Station and the size of the Basin, the variable combination-II (using October to forecast date composite precipitation index as precipitation input variables) was adopted in modeling of Elephant Butte Reservoir net inflow. The October to forecast date composite precipitation index of the Basin was calculated as a weighted average of monthly precipitation from October to forecast date based on the correlation coefficients of the precipitation of a specific month with the March-July seasonal net inflow volume. The calculated correlation coefficients between monthly precipitation from October to April, month-specific weighted average October-April precipitation index, and equal-weighted average October-April precipitation index of the SNOTEL sites with the March-July seasonal net inflow are shown in Table 3.14. The composite PRCP index for each SNOTEL site can be calculated by Equation 3.5, and then used as the inputs for each forecast equation development. Again, the results confirmed the potential forecast improvement by using month-specific weighted average precipitation index in the forecast equations. The variables used in the net inflow modeling are described in Table 3.15.

Table 3.14 Correlation coefficients of monthly and October-April average precipitation of SNOTEL sites with March-July Elephant Butte Reservoir net inflow (1981-2002)

Snotel Site	OCT	NOV	DEC	JAN	FEB	MAR	APR	OCT-APR	
								Weighted average	Simple average
Bateman	0.32	0.54	0.49	0.28	0.46	0.29	0.26	0.88	0.84
Chamita	0.52	0.52	0.51	0.16	0.45	0.40	0.25	0.90	0.88
Culebra #2	0.25	0.55	0.29	0.09	0.36	0.27	0.34	0.84	0.79
Cumbres Trestle	0.41	0.56	0.48	0.09	0.37	0.40	0.30	0.86	0.80
Gallegos Peak	0.35	0.68	0.38	0.14	0.18	0.36	0.46	0.82	0.83
Hopewell	0.48	0.60	0.40	0.15	0.33	0.37	0.29	0.87	0.83
Lily Pond	0.42	0.53	0.48	0.10	0.09	0.36	0.26	0.78	0.76
Middle Creek	0.32	0.51	0.44	0.01	0.19	0.15	0.25	0.74	0.67
Quemazon	0.41	0.54	0.55	0.26	0.46	0.00	0.17	0.82	0.73
Red River Pass #2	0.29	0.61	0.23	0.14	0.13	0.30	0.39	0.69	0.68
Upper San Juan	0.37	0.52	0.51	0.11	0.17	0.33	0.24	0.81	0.76
Wolf Creek Summit	0.43	0.47	0.26	0.11	0.10	0.28	0.11	0.62	0.51
Average	0.38	0.55	0.42	0.14	0.27	0.29	0.28	0.80	0.76

Table 3.15 List of variables used in regression equation development for Elephant Butte Reservoir net inflow

Variables	**Notation**	**Description**	**Number of variables used for monthly equation**				
			Jan	**Feb**	**Mar**	**Apr**	**May**
Snow Water Equivalent	SWE	Measured on the first day of a month	12	12	12	12	12
Precipitation Index	PRCP INDEX	October to forecast date precipitation index for each SNOTEL site	12	12	12	12	12
Temperature Index	TEMP INDEX	October to forecast date monthly average temperature index	3	4	5	6	7
Previous Flow	FLOW	October to forecast date monthly flow	3	4	5	6	7
Southern Oscillation Index	SOI	October to December average of previous year	1	1	1	1	1
Total number of variables used for variable selection in monthly equations			31	33	35	37	39

3.3.2.1 Variable Selection Procedure

The same variable selection procedure for the PLSR and PCR equation development that was described in section 3.2.2.2 was used in the variable selection in the seasonal net inflow forecast equation development. The variable selection was carried out using PLSR by analyzing predictor variables using variable importance for the projection (VIP) of each predictor and considering the forecast consistency of the regression equations from month to month. The same selected variables were used for principal components regression modeling.

3.3.3 Model Testing and Comparison

3.3.3.1 Comparison of Models for Cross Validation Phase

Based on the procedures described in the previous sections, the PLSR and PCR forecasting equations were calibrated using data from 1981to 2002 (22 years) with the jackknife cross validation scheme. The variable selection was performed in calibrating PLSR model by analyzing VIP and forecast consistency. Principal components regression was calibrated using the same variables used in PLSR, and no variable selection was performed for PCR during the calibration.

Model adequacy was tested by analyzing the residuals of calibration period for all the equations. The residuals can provide information about the adequacy and prediction capability of models over the range of dependent variables. The student's t-test was used to test if the residuals had zero mean, Shapiro-Wilk normality test was used to test if residuals were normally distributed and Ljung-Box white noise test was used to check whether any autocorrelations exist in the residuals or if the residuals were white noise process. The test results showed that all the model residuals were statistically of a zero mean, normally distributed and white noise process. This suggested that all the models were adequate with acceptable predictive ability. Taking the March 1st PLSR equation as an example, the *p*-value of student's t-test for zero mean was 0.996, *p*-value of the Shapiro-Wilk normality test was 0.994, the *p*-value of Ljung-Box white noise test up to lag 6 was 0.774. The normal probability plot of residuals for this model is plotted in Figure 3.10. All the statistical tests and normality plots suggested that there was not any modelable information left in the residuals. Similar results were obtained for all other forecasting equations for Elephant Butte seasonal net inflow.

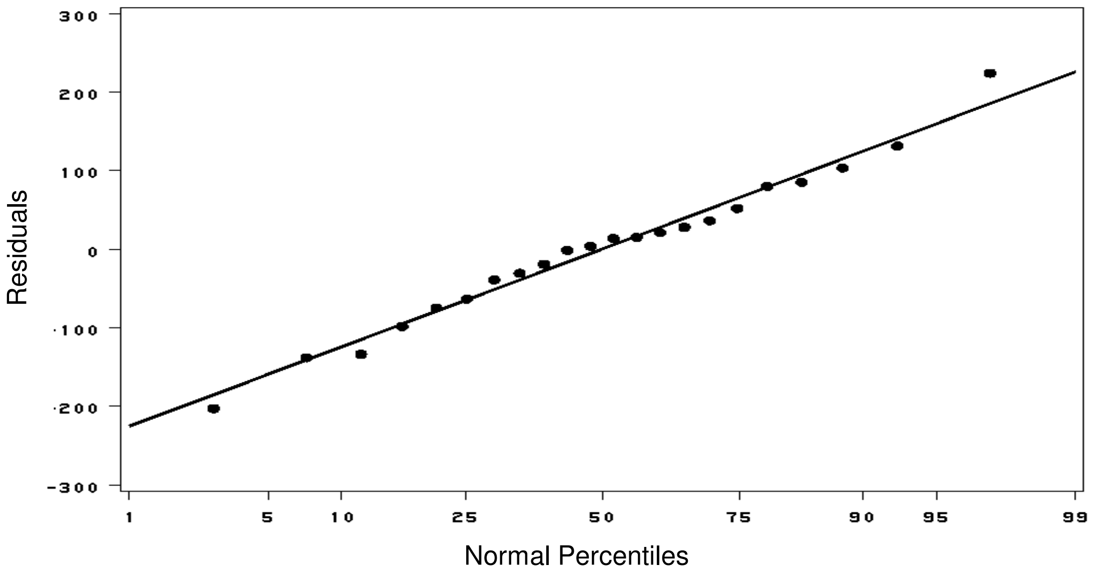

Figure 3.10 Normality plot of residuals of March 1st PLSR equation (1981-2002)

The coefficients of variables of the calibrated PLSR and PCR forecast equations are shown in Tables 3.16 and 3.17. The maximum number of variables of forecast equations (April 1st forecast equation) was 31. Similar to the results of Del Norte flow modeling, there was no significant difference in the magnitude of variable coefficients of PLSR and PCR equations. To compare the forecasting performance of these models, several performance statistics of PLSR and PCR models, such as coefficient of determination for calibration and cross validation, RMSE, normalized RMSE, are also described in Tables 3.16 and 3.17. In general, there is a reduction in model error as the period in time between forecast date and the actual forecast period is narrowed. The same trend was also observed in Del Norte forecasting equations. The cross validation RMSE decreased from highest errors at January 1st forecast equations to lowest errors at May 1st equations. For instance, the normalized RMSE of January 1st PLSR equation decreased from 0.62 to 0.39 for the May 1st PLSR equation. Similar trends were also observed for PCR equations.

As in the results of Del Norte modeling, the developed forecast equations for Elephant Butte Reservoir net inflow could very likely improve forecast accuracy compared to NRCS current forecasting equations at San Marcial Gaging Station. Table 3.16 suggests that the calibration coefficient of determination (R^2) of January 1st to April 1st PLSR forecasting equations (calibration period 22 years) were higher than calibration of R^2 of NRCS equations (calibration period 25 years). For example, the calibration R^2 of March 1st NRCS equation at San Marcial Gaging Station was 0.57, while PLSR equation was 0.75. Similar results were obtained for other forecast equations developed by PCR. The application of effective precipitation index also facilitated the developing of more parsimonious regression model with fewer input variables as compared to NRCS current forecasting equations. This feature is more evident for the later forecast dates without losing forecast accuracy. For example, the current NRCS April 1st forecast equation at San Marcial Gaging Station had 43 predictor variables from six SNOTEL sites with the calibration R^2 of 0.70. In comparison, the number of variables in the proposed April 1st PLSR forecast equation was 33 from twelve SNOTEL sites and calibration R^2 was 0.86 (Table 3.16).

Table 3.16 Forecast equations from Jan 1st to May 1st for Elephant Butte Reservoir net inflow developed using PLSR

VARIABLES	CATEGORY	Coefficients				
		JAN 1st	FEB 1st	MAR 1st	APR 1st	MAY 1st
Constant	C	-151.58	-288.20	-375.53	-461.53	-254.24
Bateman	SWE	3.42	3.45	3.50	2.62	2.97
Chamita	SWE	3.31	2.94	2.34	2.14	15.62
Culebra #2	SWE	4.31	4.11	3.12	2.23	2.10
Cumbres Trestle	SWE	2.18	1.77	1.44	1.38	0.75
Gallegos Peak	SWE	3.97	4.54	4.03	2.65	2.72
Hopewell	SWE	3.13	2.41	2.28	2.08	1.52
Lily Pond	SWE	3.49	2.93	1.79	1.65	1.05
Middle Creek	SWE	2.35	2.05	1.79	1.62	0.39
Quemazon	SWE	4.47	3.82	3.28	1.89	*
Red River Pass #2	SWE	6.87	7.78	3.87	2.51	9.55
Upper San Juan	SWE	1.71	1.36	1.06	1.13	0.49
Wolf Creek Summit	SWE	1.84	1.64	1.37	1.13	0.39
Bateman	PRCP INDEX	12.17	14.78	19.31	20.24	18.82
Chamita	PRCP INDEX	12.98	14.92	17.17	17.54	17.53
Culebra #2	PRCP INDEX	15.06	17.01	18.22	18.13	18.89
Cumbres Trestle	PRCP INDEX	8.22	9.08	10.02	9.87	6.26
Gallegos Peak	PRCP INDEX	13.16	16.00	17.58	15.62	13.87
Hopewell	PRCP INDEX	10.05	11.69	13.51	13.52	13.31
Lily Pond	PRCP INDEX	8.82	9.82	10.34	10.52	5.53
Middle Creek	PRCP INDEX	7.26	7.59	8.80	8.95	2.24
Quemazon	PRCP INDEX	14.53	15.59	17.21	15.50	15.94
Red River Pass #2	PRCP INDEX	16.12	19.50	20.47	20.77	8.27
Upper San Juan	PRCP INDEX	5.86	6.48	7.20	7.68	3.46
Wolf Creek Summit	PRCP INDEX	4.89	5.17	5.43	6.25	4.10
Temperature-10	TEMP INDEX	-6.85	-7.09	-6.95	-6.26	-4.07
Temperature-11	TEMP INDEX	-3.50	-3.62	-3.55	-3.13	*
Flow-12	FLOW	0.28	0.29	0.29	0.26	0.95
Southern Oscillation Index	SOI	-7.69	-7.96	-7.80	-7.10	-19.22
Flow-1	FLOW		0.33	0.33	0.24	*
Flow-2	FLOW			0.19	0.14	*
Flow-3	FLOW				0.27	0.32
Flow-4	FLOW					0.07
Number of components		1	1	1	1	2
Number of components by min PRESS		2	2	2	1	2
Calibration R^2		0.68	0.71	0.75	0.86	0.91
San Marcial NRCS Eq. Calib. R^2 (n=25)		**0.69**	**0.60**	**0.57**	**0.70**	**-**
Jackknife cross validation R^2		0.61	0.65	0.69	0.83	0.85
Cross validation RMSE (kaf)		184	174	162	110	89
Cross validation NRMSE		0.62	0.58	0.54	0.41	0.39
Number of years for calibration		22	22	22	22	22

** Snow water equivalent is zero at that site on the specific month, or the variable is not included*

Table 3.17 Forecast equations from Jan 1st to May 1st for Elephant Butte Reservoir net inflow developed using PCR

VARIABLES	CATEGORY	Coefficients				
		JAN 1st	**FEB 1st**	**MAR 1st**	**APR 1st**	**MAY 1st**
Constant	C	-167.77	-259.27	-344.07	-455.66	-235.77
Bateman	SWE	5.28	3.47	3.23	2.59	2.39
Chamita	SWE	4.58	3.54	2.80	2.29	12.20
Culebra #2	SWE	6.46	4.00	3.30	2.38	2.66
Cumbres Trestle	SWE	1.87	1.95	1.60	1.44	0.95
Gallegos Peak	SWE	6.23	4.32	3.58	2.59	4.15
Hopewell	SWE	3.23	2.68	2.35	2.09	1.41
Lily Pond	SWE	1.26	3.00	2.20	1.82	1.02
Middle Creek	SWE	-0.02	2.19	1.93	1.72	0.44
Quemazon	SWE	8.81	3.48	2.96	2.02	*
Red River Pass #2	SWE	11.06	7.09	4.40	2.71	7.36
Upper San Juan	SWE	0.90	1.47	1.21	1.16	0.70
Wolf Creek Summit	SWE	0.44	1.58	1.43	1.20	0.55
Bateman	PRCP INDEX	17.19	14.79	17.41	19.28	16.43
Chamita	PRCP INDEX	14.47	14.18	15.60	17.24	13.10
Culebra #2	PRCP INDEX	18.39	15.73	17.68	18.52	20.00
Cumbres Trestle	PRCP INDEX	7.02	8.75	9.49	10.03	7.97
Gallegos Peak	PRCP INDEX	15.87	14.85	15.47	15.02	14.53
Hopewell	PRCP INDEX	6.99	11.16	12.30	12.91	11.21
Lily Pond	PRCP INDEX	1.00	9.57	9.82	10.57	5.10
Middle Creek	PRCP INDEX	0.98	7.50	8.60	9.11	2.84
Quemazon	PRCP INDEX	19.25	14.11	15.41	15.11	10.22
Red River Pass #2	PRCP INDEX	23.08	18.43	18.92	19.94	21.06
Upper San Juan	PRCP INDEX	1.76	6.37	6.87	7.59	4.74
Wolf Creek Summit	PRCP INDEX	3.60	5.96	6.21	6.18	4.16
Temperature-10	TEMP INDEX	-7.78	-3.68	-3.39	-4.86	-5.61
Temperature-11	TEMP INDEX	-6.56	-3.11	-3.51	-2.88	*
Flow-12	FLOW	1.00	0.07	0.06	0.17	0.40
Southern Oscillation Index	SOI	-25.75	-3.19	-3.02	-5.85	-15.93
Flow-1	FLOW		0.14	0.20	0.17	*
Flow-2	FLOW			0.10	0.10	*
Flow-3	FLOW				0.23	0.12
Flow-4	FLOW					0.08
Number of components		2	1	1	1	2
Number of components by min PRESS		3	4	6	3	2
Calibration R^2		0.70	0.68	0.70	0.85	0.89
San Marcial NRCS Eq. Calib. R^2 (n=25)		**0.69**	**0.60**	**0.57**	**0.70**	**-**
Jackknife cross validation R^2		0.61	0.62	0.62	0.66	0.83
Cross validation RMSE (kaf)		184	181	180	171	111
Cross validation NRMSE		0.61	0.60	0.57	0.41	0.37
Number of years for calibration		22	22	22	22	22

** Snow water equivalent is zero at that site on the specific month, or the variable is not included*

Based on the performance statistics both for the calibration phase and jackknife cross validation, no significance difference could be found between the performance of PLSR and PCR. Similar results were also obtained in Del Norte seasonal natural flow modeling. However, there was a difference in the number of components used in PLSR and PCR. It can be observed from Tables 3.16 and 3.17 that the optimum number of components extracted by minimum PRESS statistic in PLSR was lower than that of PCR for all forecast date equations. Partial least squares regression reaches its minimal prediction error with a smaller number of components than PCR. Moreover, the explained variation (coefficient of determination for calibration) of dependent variable by PLSR is always higher than PCR for the same number of components extracted.

Figure 3.11 also illustrates the difference of PLSR and PCR in extracting optimal components based on minimum prediction error and van der Voet's test using jackknife cross validation for March 1st forecast equation. It can be seen that the number of components that were suggested by minimum PRESS and van der Voet's test was 2 for PLSR, whereas, they were 6 and 3 for PCR respectively. The PLSR can reach minimum prediction error using 2 components, and the PCR can reach minimum prediction error with 6 components. In general, the PLSR method is more powerful than PCR in extracting components that deal with collinearity issue; but this does not guarantee that PLSR regression will be better than PCR at seasonal streamflow forecasting accuracy. Similar results were obtained in Del Norte Gaging Station PLSR and PCR equation development.

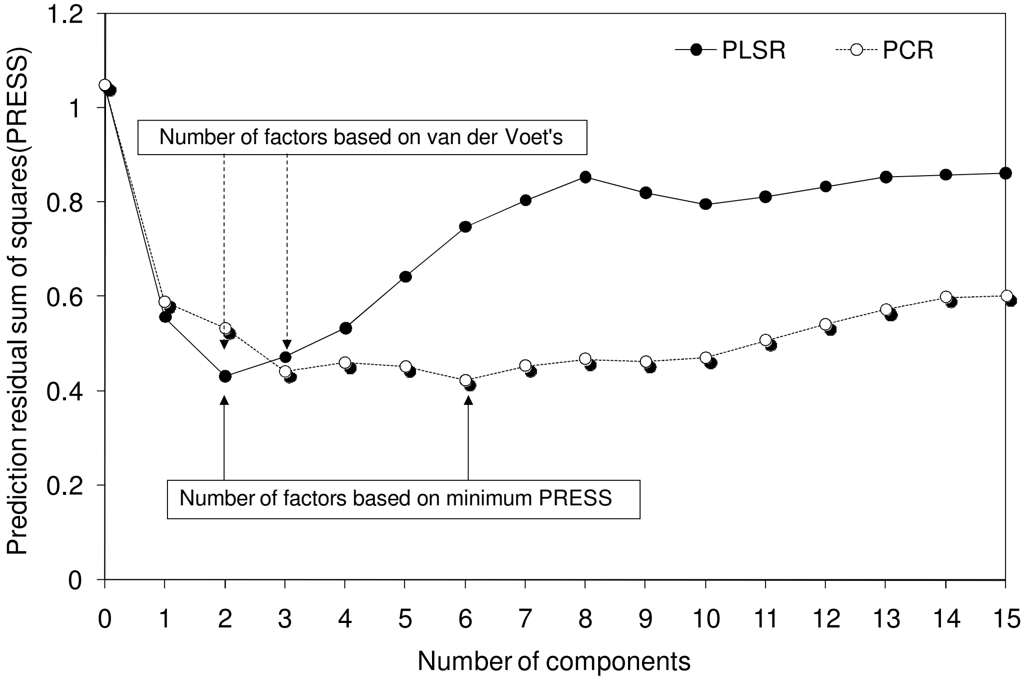

Figure 3.11 Optimum number of components extracted by PLSR and PCR for March 1st forecasting equation

3.3.3.2 Comparison of Models for Testing Data

The models were tested for new data from 2003 to 2007 to examine the forecast performance of PLSR and PCR models. Although NRCS does not provide seasonal net inflow forecasting for Elephant Butte Reservoir, the NRCS does issue March-July natural seasonal volume forecasts for San Marcial Gaging Station, Rio Grande, which is located in the entrance of Elephant Butte Reservoir. To compare the model forecasts with NRCS forecasts, a routing forecast equation was developed using the March-July natural flow of San Marcial Gaging Station and Elephant Butte net inflow. The routing equation for Elephant Butter Reservoir net inflow is as follows:

$$F_{MJ} = 0.728 F_{NRCS} + 36.56 \qquad (R^2=0.96, n=40) \qquad (3.9)$$

Where

F_{MJ} = Elephant Butte March-July net inflow forecasts (kaf);

F_{NRCS} = NRCS forecasts for March-July natural flow at San Marcial Gaging Station (kaf);

In order to compare the performance of models for the data period from 2003 to 2007, the forecasts for all forecast dates were used. The NRCS official forecasts on forecast dates of January 1st to May 1st (as shown in Table 3.13) were used for calculation of routed forecasts of Elephant Butte March-July net inflow using Equation 3.6. To compare the March-July forecast volume by PLSR and PCR equations, the observed March flow was added to April-July volume forecasts of April 1st equations; the observed March and April flows were added to May-July forecasts of May 1st forecasting equations. The comparison of forecasts from 2003 to 2007 are shown in Figure 3.12.

The Figure 3.12 suggests that the forecasts of PLSR and PCR for the all years and forecast dates are similar, although PLSR performed somewhat better than PCR for most of the years except 2006. In general, no significant difference was observed between PLSR and PCR. When comparing the PLSR and PCR forecasts with converted NRCS official forecasts for Elephant Butte Reservoir net inflow, the NRCS forecasts for 2003, 2006 and 2007 were better than PLSR and PCR in general except for some forecast dates. In contrast, PLSR and PCR equations provided better forecasts than NRCS official forecasts for 2004 and 2005. Overall, the single PLSR equation developed in this study showed potential capability in net inflow forecasting of Elephant Butte Reservoir when compared with the NRCS official forecasts that are the result of coordinated work from different agencies using a number of forecasting equations.

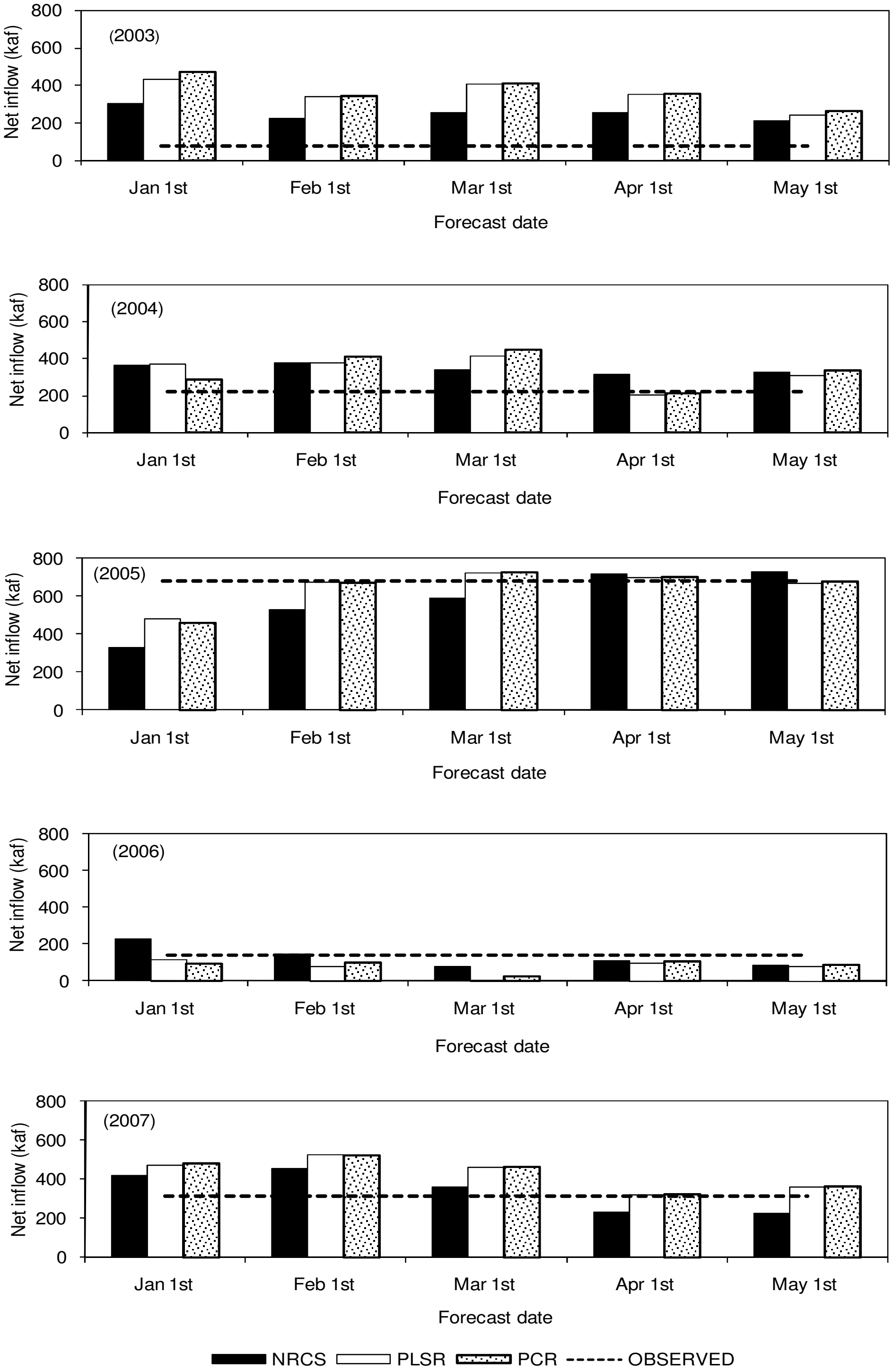

Figure 3.12 Comparison of March-July PLSR, PCR model forecasts with converted NRCS official forecasts for Elephant Butte Reservoir net inflow for 2003-2007

3.3.4 Final Model

The discussions in the previous sections suggested that the performance of partial least squares regression is similar to principal components regression. However, it was also observed that PLSR reaches its minimal prediction error with a smaller number of components and can account for more variation of the dependent variable with the same components, as compared to PCR. This is a unique feature of PLSR compared to PCR when developing regression equations. Considering all those factors, the following final model was developed using data from 1981 to 2007 (27years) and the PLSR modeling approach. The final forecast equation and the model performance results that were evaluated using the jackknife cross validation scheme are shown in Table 3.18. The proposed final model could be used for March-July seasonal net inflow forecasting of Elephant Butte Reservoir, Rio Grande, New Mexico.

Table 3.18 Final PLSR forecasting equation for Elephant Butte Reservoir net inflow, Rio Grande, New Mexico

VARIABLES	CATEGORY	Coefficients				
		JAN 1st	FEB 1st	MAR 1st	APR 1st	MAY 1st
Constant	INTERCEPT	-177.89	-319.18	-383.67	-478.13	-254.09
Bateman	SWE	3.42	3.69	3.42	2.65	3.62
Chamita	SWE	3.46	3.36	2.54	2.05	16.98
Culebra #2	SWE	4.60	4.04	2.93	2.19	1.61
Cumbres Trestle	SWE	2.26	1.89	1.44	1.42	0.70
Gallegos Peak	SWE	4.11	4.68	3.71	2.39	2.65
Hopewell	SWE	3.20	2.65	2.32	2.13	1.49
Lily Pond	SWE	3.67	2.94	1.91	1.78	1.38
Middle Creek	SWE	2.51	2.07	1.78	1.75	0.51
Quemazon	SWE	4.54	3.86	3.13	1.94	*
Red River Pass #2	SWE	5.58	5.74	3.51	2.42	11.44
Upper San Juan	SWE	1.74	1.23	0.93	1.05	0.57
Wolf Creek Summit	SWE	1.94	1.53	1.25	1.14	0.51
Bateman	PRCP INDEX	11.37	14.97	18.01	19.07	13.25
Chamita	PRCP INDEX	12.71	15.22	17.68	17.53	23.40
Culebra #2	PRCP INDEX	15.86	17.78	18.89	17.89	14.98
Cumbres Trestle	PRCP INDEX	8.72	9.75	10.10	10.26	6.16
Gallegos Peak	PRCP INDEX	13.65	16.54	17.33	14.74	12.41
Hopewell	PRCP INDEX	10.07	12.00	13.17	13.25	13.95
Lily Pond	PRCP INDEX	8.83	9.71	10.01	10.84	5.64
Middle Creek	PRCP INDEX	7.36	7.65	8.61	9.13	0.62
Quemazon	PRCP INDEX	14.76	15.83	16.59	15.00	12.47
Red River Pass #2	PRCP INDEX	16.06	19.48	19.84	21.51	1.01
Upper San Juan	PRCP INDEX	6.09	6.66	6.98	8.07	3.18
Wolf Creek Summit	PRCP INDEX	4.73	4.94	5.00	6.14	2.97
Temperature-10	TEMP INDEX	-6.27	-6.46	-6.17	-5.55	-4.04
Temperature-11	TEMP INDEX	-3.32	-3.42	-3.27	-2.90	*
Flow-12	FLOW	0.33	0.34	0.32	0.29	1.22
Southern Oscillation Index	SOI	-7.79	-8.03	-7.68	-7.01	-18.61
Flow-1	FLOW		0.43	0.41	0.32	*
Flow-2	FLOW			0.22	0.18	*
Flow-3	FLOW				0.28	0.42
Flow-4	FLOW					0.09
Number of components used		1	1	1	1	2
Calibration R^2		0.64	0.71	0.72	0.85	0.91
Jackknife cross validation R^2		0.58	0.65	0.67	0.82	0.85
Cross validation RMSE (kaf)		188	171	167	112	95
Cross validation NRMSE		0.64	0.58	0.57	0.42	0.42
Number of years for calibration		27	27	27	27	27

** Snow water equivalent is zero at that site on the specific month, or the variable is not included*

3.4 Conclusions and Remarks

In this chapter, the application of partial least squares regression (PLSR) and principal components regression (PCR) approaches in seasonal streamflow forecasting has been presented. Some issues related to regression equation development such as the selection of an optimal number of components using jackknife cross validation scheme and the variable selection procedure in PLSR have been discussed. Seasonal streamflow forecasting performance of PLSR and PCR models was compared with each other as well as with NRCS official forecasts. Two subbasins of the Rio Grande, the Rio Grande Headwaters above Del Norte Gaging Station and Rio Grande Basin above Elephant Butte Reservoir, and two hydrologic variables, river flow and reservoir net inflow, were used for seasonal flow forecasting model development.

The effective precipitation index was first introduced in this study to capture complex precipitation data in a concise parameter and to examine if better forecast skills could be obtained. Effective precipitation indices were found to be an efficient method of both improving forecast accuracy and developing more parsimonious regression models with fewer input variables. This is particularly important for larger basins where more information is available from numerous SNOTEL sites for the forecast equation development. The algorithm used in this study was limited to using the weighted average of monthly precipitation based on the correlation coefficients with corresponding spring-summer forecast target volume.

The comparison of the performance of PLSR and PCR suggested that there are no significant differences in the forecasting performance of the two methodologies; similar forecast accuracies were obtained for both methods. However, PLSR can reach its minimal prediction error with a smaller number of components than PCR. Moreover, the explained variation of the dependent variable by PLSR is always higher than PCR for the same number of components extracted. In general, the PLSR method is more powerful than PCR in extracting components that deal with collinearity issue, yet it may not necessarily guarantee that PLSR would be better than PCR in terms of forecasting accuracy in seasonal streamflow forecasting.

The proposed forecasting equations in the study using PLSR and PCR were calibrated using only 22 years of data. This calibration period in this study is shorter than that of the NRCS forecasting equations because the data used in the calibration are continuous high quality data that are measured from NRCS automatic SNOTEL sites. Except for several years of SNOTEL precipitation data that were extended back to the 1980s using weather station precipitation data, neither an estimation of missing data was performed, nor was the Snow Course data used in the calibration of the regression equations. The standoff between using a shorter calibration period and using real-time good quality data is often encountered in seasonal streamflow forecasting equation development. However, with the accumulation of real-time measured SNOTEL data through time, the regression equations can be recalibrated every year when the new data become available.

The PLSR and PCR regression equations developed in this study are used to compute the median value of the seasonal water volume forecast distribution. If needed, the ensembles/probabilistic forecasts can be added by analyzing the statistical properties of the model error series (i.e., residuals) that occur in reproducing observed historical streamflow data using jackknife procedure. The results from statistical tests of residuals of all PLSR models developed study showed that they are normally distributed at 0.05 significance level. Based on the normally- distributed errors, the exceedance probability forecasts of PLSR equation can be provided. The width of the probabilistic forecast error bound is proportional to the root mean squared error between these jackknife hindcasts and their respective observations.

In general, the application of PLSR in seasonal streamflow forecasting is promising. Together with PCR and Z-score regression, the PLSR approach can be combined into NRCS's operational forecasting platform to facilitate its application in operational forecasting environment. The selection of numbers of components with PLSR and variable selection procedures in seasonal streamflow forecasting equation development were attempted in this study. However, the variable selection in PLSR is always a challenging task due to the complexity of the hydrologic process. Similar to Garen's (1992) method of variable selection for PCR, the investigation and application of more robust variable selection approaches, such as systematic searching of optimal or near optimal variable combination in PLSR would be desirable in future seasonal streamflow forecasting research studies.

4 HYBRID MODELING OF SEASONAL FLOW

In this chapter, the application of hybrid modeling approaches in seasonal streamflow forecasting is proposed. Two hybrid modeling approaches, a forecast modification using a combination of transfer function - noise (TFN) model with artificial neural networks (ANN), and the combination of principal components analysis (PCA) with ANN, have been investigated for the purpose of improving seasonal streamflow forecasts. To perform time series modeling of seasonal flow, different seasons were defined for two basins used in the study. Several statistical models, including autoregressive integrated moving average (ARIMA), TFN, and ANN, were built for seasonal natural flow at Del Norte Gaging Station, Rio Grande, Colorado and seasonal net inflow of Elephant Butte Reservoir, Rio Grande, New Mexico. The forecast performance of two hybrid modeling approaches was compared to the different single modeling techniques such as ARIMA, TFN and ANN. Finally, some general discussions and conclusions are summarized at the end of the chapter.

4.1 Hybrid Model Formulation

As described in section 2.4.2, two hybrid modeling approaches have been presented for the purpose of improving seasonal and monthly streamflow forecasting performance in this study. The general model formulation of the approaches are described in the following sections.

4.1.1 Forecast Modification Using ANN

Forecast modification is essentially a complementary hybrid modeling approach, which is the combination of two or more models. In this study, the TFN model with precipitation input was used to produce one-step-ahead streamflow forecasts; the forecasts were then modified using the artificial neural networks technique with the inclusion of new information such as snow water equivalent (SWE) and El Niño Southern Oscillation Index (SOI). The configuration of the ANN model for forecast modification is as follows:

$$Y_{t,\ modified} = f\,(SWE_{t'},\ SWE_{t'-1},\ SOI,\ Y_{t,\ forecasted}) \qquad (4.1)$$

Where

$Y_{t,\ modified}$ = seasonal flow forecasts after the forecast modification at season t ;

$Y_{t,\ forecasted}$ = seasonal flow forecasts using TFN model with PRCP input at season t;

$SWE_{t'}$, $SWE_{t'-1}$ = snow water equivalents on the first day of month t', $t'-1$ (t' is the first month of season t);

SOI = the October-December averaged El Niño Southern Oscillation Index of a previous year.

The number of SWE inputs and lag relationships in the ANN models may vary slightly depending on the particular season. But the maximum number of SWE inputs was limited to 2, since they are highly intercorrelated between the consecutive months. In addition, limiting the number of SWE inputs can keep a network size smaller so that it has a better generalization capability. The detailed procedures, including the selection of seasons to be modified and number of SWE inputs, are discussed in the following sections according to the hydrologic characteristics of the specific basin to be modeled.

4.1.2 Combination of PCA and ANN

A combination of PCA and ANN is a complementary hybrid modeling technique that combines the advantages of principal components analysis and neural networks. As discussed in section 2.2.2, the predictor variables used in seasonal water supply forecasting are usually highly intercorrelated. For example, the snow water equivalent, precipitation data of different SNOTEL sites and different months are highly correlated with each other. A large number of intercorrelated predictor variables are often referred to as a multi-collinearity issue, which may affect performance of neural networks by easily overtraining the network and giving very low performance on the prediction of new data.

The principal components analysis is a statistical technique that deals with highly intercorrelated predictor variables by extracting an equal number of uncorrelated variables. Hence, the combination of PCA and ANN may facilitate the effective neural network modeling because the network may converge faster due to the orthogonal features of principal components, and using fewer principal components as inputs than original variables. The graphical representation of the algorithm is described in Figure 4.1. The network consists of four layers including input layer for PCA, PCA outputs (also used as input layer for neural networks), a hidden layer and the output. Basically, in this approach, a new set of input predictor variables for an ANN model were created by transforming the original variables into a fewer number of orthogonal variables using PCA analysis.

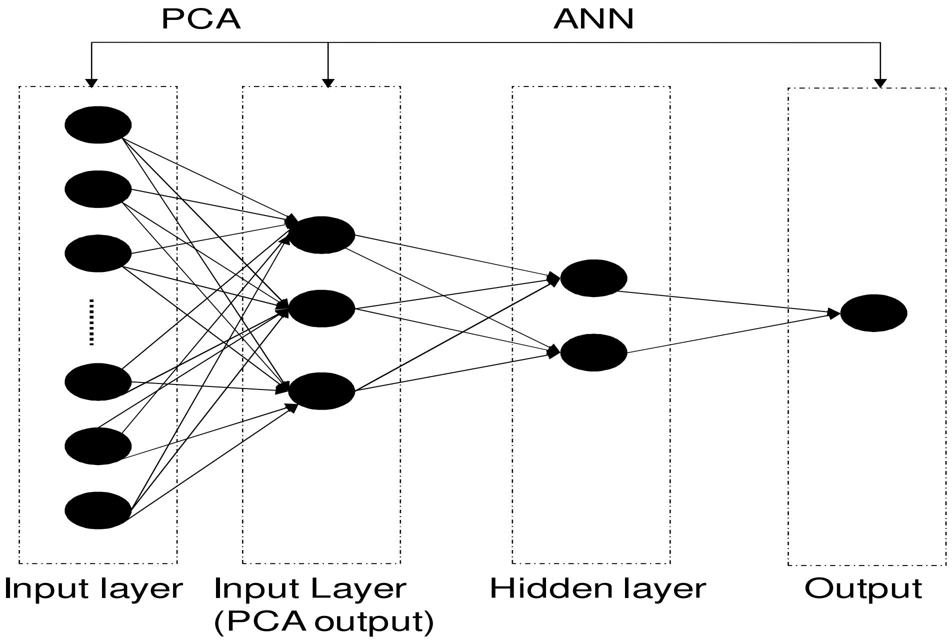

Figure 4.1 Graphical representation of combination of PCA and ANN hybrid modeling approach

The following procedure was applied to select the number of principal components that used as the inputs for ANN model building in this study:

1) Extract the first five principal components (PC) of the input variables; only the first five PCs were used, mainly because of the desirability of keeping smaller inputs and network sizes for better generalization. In addition, most of

the information usually can be explained by the first few PCs of predictor variables in streamflow forecasting (See chapter 3).

2) Perform stepwise variable selection at 0.05 significance level
3) Select the number of principal components until getting the second significant principal component; For example, if the second significant principal component was the fourth PC, then the first, second, third and fourth PCs would be selected as the inputs for the ANN, even if the second and third components were not significant predictors according to the t-test of significance in stepwise variable selection.

The selection of the number of PCs in sequence for the hybrid modeling was adopted from the approach suggested by Garen (1992) for the selection of the number of PCs retained in the principal components regression modeling. The selection of PCs in sequence may prevent an unexpected jumps and drops in model prediction. It was also assumed that there may be some nonlinear relationships existing between PCs and the dependent variable even if the relationships are not linearly significant; and ANN has the capability of mapping the nonlinear relationship patterns between predictors and the dependent variable.

4.2 Del Norte Natural Flow

4.2.1 Definition of Seasons

To develop seasonal flow time series models, different seasons in a year should be defined for the basin. To define the seasons in the Rio Grande Headwaters Basin above Del Norte Gaging Station, following factors were considered:

1) Statistical similarities such as mean, standard deviation, of monthly streamflow;
2) The magnitude of correlation between monthly streamflow and snow water equivalent;
3) NRCS forecasting periods;
4) Generation of an equally spaced continuous seasonal streamflow volume time series.

The Figure 1.3 showed the monthly average flow and standard deviation of Del Norte natural flow for the calculation period of 1961-2007. It can be seen that the May and June flows account for more than half of the annual runoff, while the April-September runoff accounts for almost 90% of the annual total runoff. This suggested that the Basin is snow-dominated, since a large portion of the annual runoff is contributed by the snowmelt. The snowmelt season lasts from April to September, and most of the snowmelt runoff is concentrated in the months of April, May and June. The NRCS forecasting period is also April-September of a calendar year. In order to get equally spaced continuous time series for the modeling of seasonal flow, the entire snowmelt season should be divided into two seasons: April-June and July-September.

To determine the pattern and magnitude of relationship between the monthly flow at Del Norte Gaging Station and monthly SWE, a correlation analysis between the monthly SWE index, which is measured in the first days of January to June, and monthly flow at Del Norte Gaging Station was performed. The results (Table 4.1) indicated that the basin average SWE index is significantly correlated with the March to September monthly flow at 0.01 significant level. Although the correlations between March flow with March 1st and February 1st SWEs are statistically significant at 0.05 significance level, the correlation coefficients are not as high as in other months. The September flow has a significant correlation coefficient with the May 1st SWE only. The monthly streamflow from April to August is significantly correlated with each month's SWE from January 1st to June 1st , since the *p*-values of Pearson correlation coefficients significance test are all smaller than 0.0001.

Considering those factors discussed above, the following seasons have been defined for the Rio Grande Headwaters Basin above Del Norte Gaging Station.

Season 1: January, February and March

Season 2: April, May and June

Season 3: July, August and September

Season 4: October, November and December

Based on the correlation analysis with snow water equivalent and monthly distribution of Del Norte flow, only the second season and third season were of interest in this study. Moreover, the NRCS provides April-September seasonal forecasts at Del Norte Gaging Station, Rio Grande, so they can be compared with the forecasts of second and third seasons that have been produced by the modeling approaches in this study.

Table 4.1 Correlation coefficient significance test between SWE index of the Basin and monthly streamflow at Del Norte Gaging Station (1961-1999)

Month	SWE					
	Jan 1st	Feb 1st	March 1st	April 1st	May 1st	June 1st
Jan	**0.41**					
Feb	**0.36**	0.23				
Mar	**0.42**	**0.32**	**0.35**			
Apr	**0.42**	**0.39**	**0.44**	**0.43**		
May	**0.59**	**0.59**	**0.62**	**0.65**	**0.69**	
Jun	**0.60**	**0.72**	**0.74**	**0.83**	**0.92**	**0.81**
Jul	**0.59**	**0.68**	**0.66**	**0.78**	**0.86**	**0.82**
Aug	**0.46**	**0.45**	**0.40**	**0.38**	**0.54**	**0.37**
Sep	0.28	0.23	0.17	0.21	**0.37**	0.17
Oct	0.32	0.26	0.21	0.18	0.29	0.09
Nov	0.33	0.19	0.15	0.18	0.33	0.07
Dec	**0.40**	0.28	0.22	0.22	**0.38**	0.20

*Note: The **bold** numbers indicate the correlations are significant at 0.01 significance level.*

4.2.2 Data Description

4.2.2.1 Seasonal Flow Time Series

After defining the seasons, the seasonal time series data were calculated by summing up corresponding monthly flows into seasonal volume. To obtain a better approximation of normal distribution for time series modeling, the seasonal flow time series was deseasonalized using seasonal average and standard deviation to remove the seasonal variation of the data. The deseasonalized series is the series after removing the seasonal variation by standardization using following expression:

$$y_t = \frac{Y_t - \overline{Y}_t}{\hat{\sigma}_t} \tag{4.2}$$

Where y_t = deseasonalized series for season t

Y_t = original series for season t

$\overline{Y}_t$ = sample average of the original series for season t

$\hat{\sigma}_t$ = sample standard deviation of the original series for season t

Figures 4.2 and 4.3 show the normal probability plots of original seasonal flow series and deseasonalized seasonal flow series. It can be seen that the deseasonalized transformation improved the normality approximation of the data significantly compared to the original series. The Shapiro-Wilk normality tests that performed to both series showed that normality assumption was not accepted at 0.05 significance level. The *p*-value corresponded to the original series was much smaller than 0.001, while the *p*-value for deseasonalized series was 0.035. However, it was assumed that the normal distribution approximation of deaseasonalized series was sufficient for time series model building.

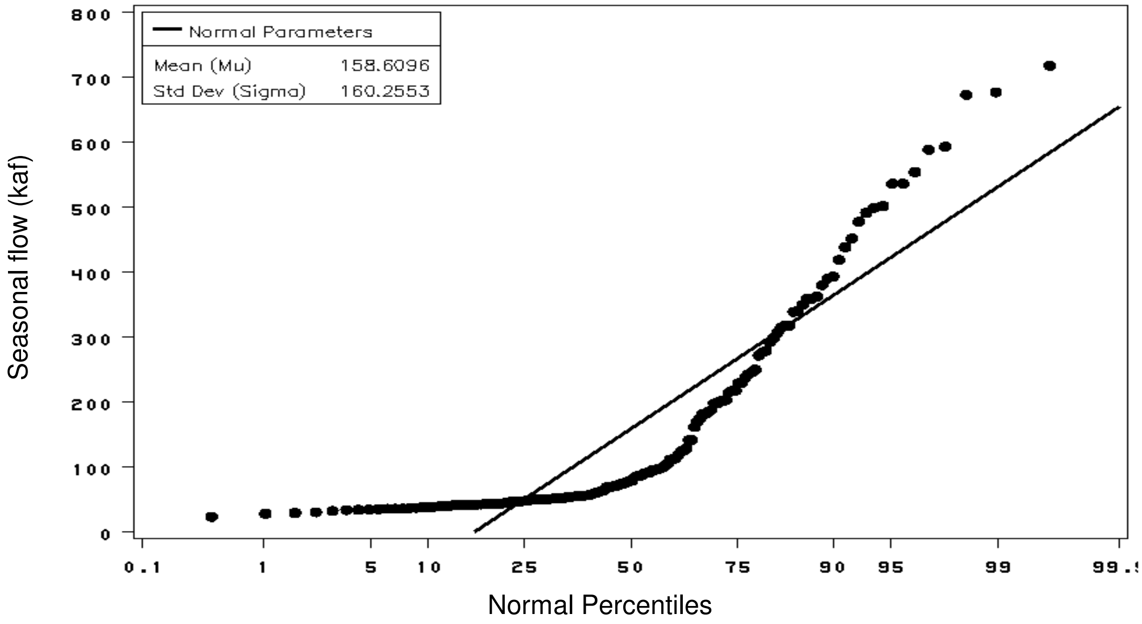

Figure 4.2 Normality plot of the original seasonal flow at Del Norte Gaging Station (1961-1999)

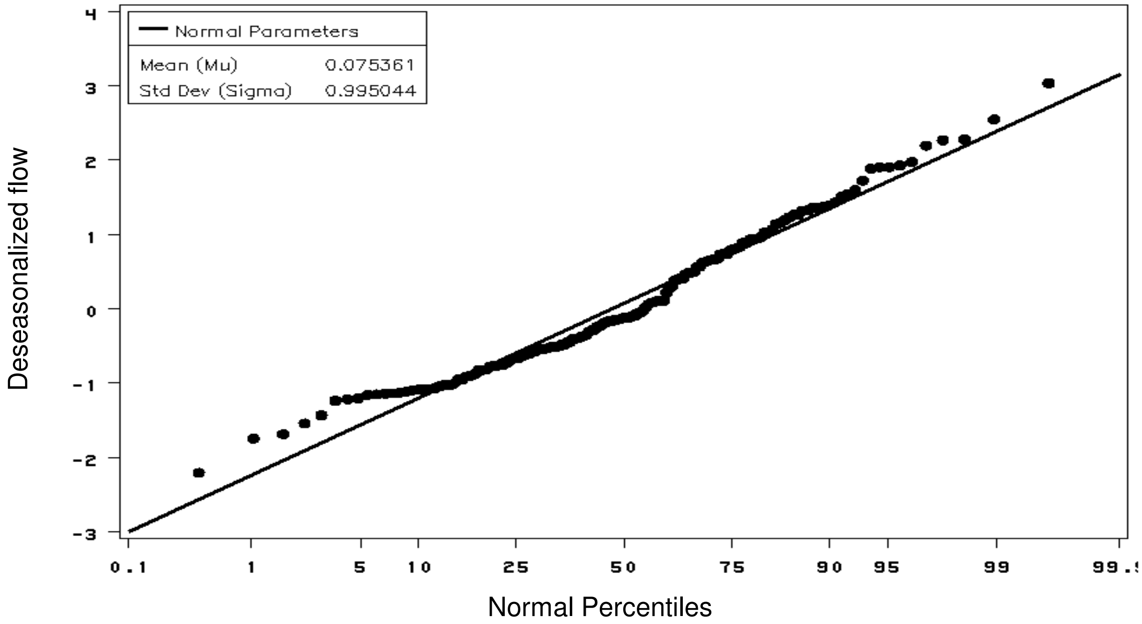

Figure 4.3 Normality plot of deseasonalized seasonal flow series at Del Norte Gaging Station (1961-1999)

4.2.2.2 Other Data

The average seasonal precipitation data from 1961 to 2005 was used in the seasonal time series modeling (the 2006 and 2007 data was not used because it is provisional data). The average seasonal precipitation of the Basin was calculated as the average of the monthly average SNOTEL precipitation index (shown in Figure1.5) for the defined seasons. SWE data used in this chapter was the monthly basin average SWE index data as shown in Figure 1.4 that also covered the period of 1961-2005. The previous year October-December SOI data was also used in the modeling (See Figure 1.12).

4.2.3 Development of Single Models

4.2.3.1 ARIMA

The deseasonalized seasonal flow series was used to fit ARIMA model. The entire data set was divided into a calibration set and a test set, which covered the periods of 1961-1999 and 2000-2005 respectively. The autocorrelation functions (ACF) and partial autocorrelation functions (PACF) suggested a AR(9) model with only lag 1 and lag 9 were significant. The following ARIMA model was developed for the series and it passed all the diagnostic checks:

$$y_t = 0.553 y_{t-1} + 0.188 y_{t-9} - 0.104 y_{t-10} + a_t \qquad (4.3)$$

Where

y_t , y_{t-1} , y_{t-9} , y_{t-10} = deseasonalized seasonal flow series at season *t*, *t-1, t-9, t-10* respectively.

The model explains 75% of variability of the Del Norte seasonal flow. One-season-ahead rolling forward forecasts were made for six years (2000-2005) to evaluate the performance of the model in the next sections. The one-season-ahead rolling forward forecast means the model parameters will be modified for every season forecasts when the new observation data become available.

4.2.3.2 TFN

To build a TFN model, the relationship of seasonal average precipitation and seasonal streamflow time series was investigated using cross correlation analysis. Based on the sample cross correlation functions between the input series and output series, and the impulse response function, an appropriate form of TFN model was suggested. The sample cross correlation between the Del Norte seasonal flow and seasonal precipitation time series was performed by prewhitening the input series and filtering the output series using ARIMA models. The precipitation series was deseasonalized by subtracting means and dividing by seasonal standard deviations, and the autocorrelation function (ACF) showed the resulting series to be white noise. The sample cross correlation function(CCF) between two series is shown in Figure 4.4.

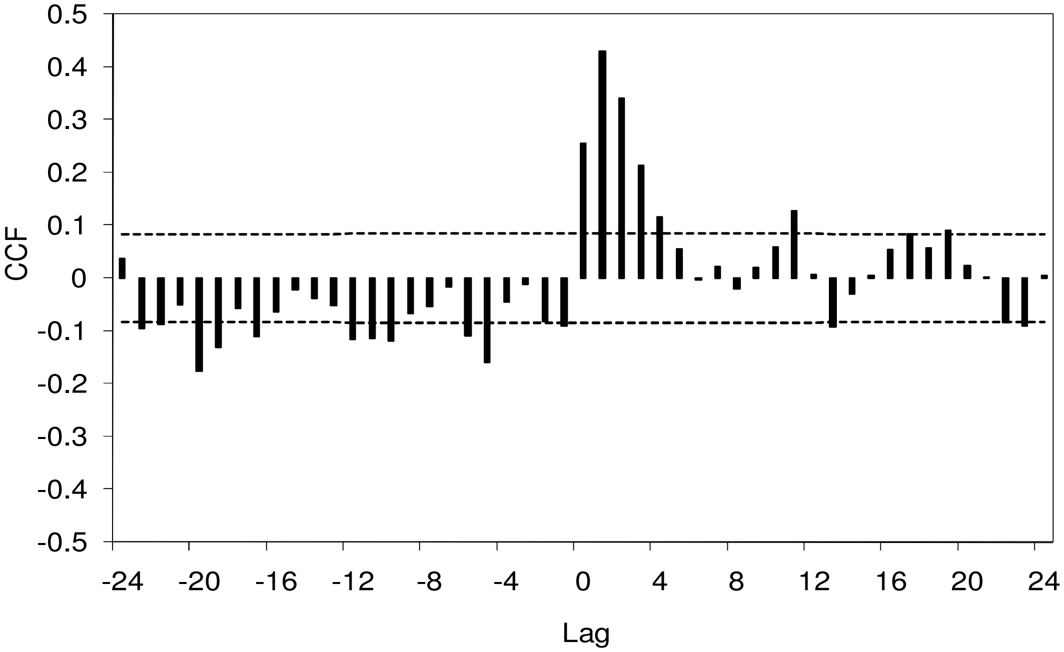

Figure 4.4 Cross correlation function between deseasonalized seasonal precipitation and filtered deseasonalized seasonal flow at Del Norte Gaging Station (1961-1999)

The following TFN model was built for deseasonalized seasonal flow series with the deseasonalized precipitation series as an input according to CCF using calibration set data. The model parameters were estimated using the conditional least squares method and the model passed all the diagnostic check according to the Box and Jenkins (1976) modeling procedure:

$$y_t = \frac{0.440}{(1-0.645B)} PRCP_{t-1} + \frac{1}{(1-0.389B)} a_t \tag{4.4}$$

Equation 4.4 can be written in more straightforward form as follows:

$$\hat{y}_t = 0.389 y_{t-1} + 0.645 \hat{y}_{t-1} - 0.251 y_{t-2} + 0.44 PRCP_{t-1} - 0.171 PRCP_{t-2} \tag{4.5}$$

Where

$PRCP_{t-1}$ = Basin average seasonal SNOTEL precipitation at season *t-1*;

$\hat{y}_t$, $\hat{y}_{t-1}$ = forecasted deseasonalized monthly flow at month *t, t-1.*

The built TFN model can explain 83% percent variability of Del Norte seasonal flow, indicating that the TFN model performed better than the ARIMA model.

The better performance of the TFN model compared to ARIMA can also be explained by the model variance and Akaike Information Criterion (*AIC*) (Akaike, 1974) of both models. *AIC* is a popular statistical measure for model discrimination, and is a mathematical formulation of maximum likelihood estimation with the parsimony principle of model building, which is defined as:

$$AIC(k) = -2\ln L + 2k \qquad (4.6)$$

where *L* is maximum likelihood and *k* is the number of independently adjusted parameters within the model. The best model is given by the model with the lowest *AIC* value. The model performance of the deseasonalized seasonal flow series for calibration phase showed that the model variance and *AIC* of TFN model decreased from 0.673 to 0.598, and 383 to 361 respectively compared to the ARIMA model.

4.2.3.3 ANN

In the previous sections, the time series models were developed for continuous seasonal flow data for all seasons. As discussed earlier, the snow water equivalent (SWE) is only related to the April to September flow of the Basin. Hence, the ANN models were developed only for two seasons in this study: April-June and July-September.

To examine the pattern and magnitude of the relationship between seasonal flow and monthly SWE and Southern Oscillation Index (SOI), a correlation analysis between the monthly SWE index, SOI and seasonal flow at Del Norte Gaging Station was performed. The results (See Table 4.2) indicated that the basin average SWE indices, which were measured in the first days of January to June, were significantly correlated with the April-June and July-September seasonal flows, while January-March and October-December flows showed systematic relationships with SWE in the Basin. The January-March and October-December seasons are significantly related to SWE for some months, but their relationships were very weak (correlation coefficients were less than 0.5).

Previous studies (Redmond and Koch, 1991) suggested that there is a negative correlation between the averaged June-November SOI and average monthly October–March streamflow and precipitation for the southwest United States. NRCS (1997) completed an analysis of the correlation of the Southern Oscillation Index (SOI) with spring and summer volume runoff in the western U.S. and found that the Rio Hondo and Lower Rio Grande March-July flow have the highest correlations coefficients with October-December SOI index. Based on the previous research results in the Rio Grande Basin, the correlation analysis has been performed between the averaged June-November SOI, averaged October-December SOI and Del Norte seasonal flow. The results (Table 4.2) suggested that the SOI did not show any significant relationship with Del Norte seasonal flow. Similar results were reported by NRCS (1997).

Table 4.2 The correlation coefficients significance test between SWEs, SOI and seasonal flow at Del Norte Gaging Station (1961-1999)

Season	**Snow Water Equivalent (SWE)**						**Southern Oscillation Index**	
	Jan1st	**Feb1st**	**Mar1st**	**Apr1st**	**May1st**	**June1st**	**JUN-NOV Average**	**OCT-DEC Average**
JAN-MAR	**0.490**	**0.308**	**0.324**				0.020	0.101
p-value	***0.001***	***0.042***	***0.032***				*0.895*	*0.513*
APR-JUN	**0.663**	**0.740**	**0.770**	**0.836**	**0.893**	**0.657**	0.059	-0.072
p-value	***<.0001***	***<.0001***	***<.0001***	***<.0001***	***<.0001***	***<.0001***	*0.702*	*0.642*
JUL-SEP	**0.588**	**0.624**	**0.579**	**0.657**	**0.804**	**0.668**	0.029	-0.116
p-value	***<.0001***	***<.0001***	***<.0001***	***<.0001***	***<.0001***	***<.0001***	*0.850*	*0.453*
OCT-DEC	**0.354**	0.262	0.206	0.194	**0.331**	0.105	0.094	0.028
p-value	***0.019***	*0.086*	*0.179*	*0.208*	***0.028***	*0.496*	*0.545*	*0.855*

Notes: the ***bold*** *numbers indicate the correlations are significant at 0.05 significance level*

The identification of proper input variables for ANN models is a challenging task. To build ANN models for season 2 and season 3, the following procedure was employed in this study to select input variables for ANN models:

Step 1: Identify the different lag relationships between precipitation and previous flow with the modeled flow based on the structure of developed TFN model and sample cross correlation function between precipitation series and filtered flow series;

Step 2: Identify the lag relationship between SWE, SOI and modeled flow using cross correlation analysis;

Step 3: Select the final inputs based on the magnitude of correlation coefficients of variables proposed by previous steps with modeled flow; Drop out the variables that correlation coefficients smaller than 0.4.

Step 4: Set the maximum numbers of input variables to five in order to keep smaller network size for each model to avoid network overfitting.

The following two ANN models were proposed based on the input variable selection analysis using the procedure described above. The structures of models for second and third seasons were formulated as follow:

For season 2 (April-June):

$$Y_t = f\,(Y_{t-1},\ PRCP_{t-1},\ PRCP_{t-2},\ SWE_{apr1st}) \qquad (4.7)$$

For season 3 (July-September):

$$Y_t = f\,(Y_{t-1},\ PRCP_{t-1},\ PRCP_{t-3},\ SWE_{jun1st}) \qquad (4.8)$$

Where

$PRCP_{t-1}$, $PRCP_{t-2}$, *and* $PRCP_{t-3}$ = average seasonal precipitation indices at seasons *t-1, t-2,and t-3*;

SWE_{apr1st} , SWE_{jun1st} = Snow water equivalent indices measured on April 1st and June 1st ;

To build ANN models, the total data period (1961-2005) was divided into a calibration set (1961-1999) and a testing set (2000-2005); then the calibration set was randomized and divided further into a training set and a cross validation set. Cross validation was used for early stopping of the training so as to protect the network from overtraining, since relatively short data set (39 years) and more predictors (maximum 5 input variables) were used for model training. The cross validation data set accounted for approximately 10 percent of the training data (5 observations), which could be practically acceptable when the data series is stationary (Wang, 2006).

In order to prevent the networks from overtraining and to enhance the generalization capability of networks that trained for such a short period, the following procedure was followed in the ANN model building process:

1) Keep the number of inputs less than 5;
2) Select maximum of two processing elements in the hidden layers so as to keep the number of network weights less than 10;
3) Use cross validation stopping criteria to stop the network if there is no improvement after 100 epochs in cross validation data set;
4) Train the networks multiple times (6-10) and select best network that has the lowest cross validation testing error.

Based on the identified input variables and maximum number of processing elements in the hidden layer, the ANN (4-2-1) model structure was used for both seasons. The training process utilized the hyperbolic tangent function as the activation function in the hidden layer, linear function in output layer, and the momentum learning rule. The training termination criteria employed cross validation techniques that would stop the training when the cross validation error begins to increase. The number of maximum training epochs was set for 1000 and the training was terminated when there is no further improvement in cross validation after 100 epochs. The best weights of the network would be automatically saved at the point when cross validation error reached its lowest point.

4.2.4 Development of Hybrid Models

4.2.4.1 Combination of TFN and ANN

Based on the model formulation in section 4.1.1, two networks were developed for modification of TFN model forecasts using snow water equivalent for seasons 2 and 3. No significant correlation was observed between El Niño Southern Oscillation Index (SOI) and Del Norte Gaging Station flow for all defined seasons (as shown in Table 4.2). Hence, the SOI was not included in the forecast modification for both seasons 2 and 3. It was also observed that the latest SWE information was sufficient for forecast modification. The adding of more SWE information from previous months did not contribute to the modification accuracy. Hence, both ANN models for seasons 2 and 3 have only two inputs, as shown in following expressions:

Season 2 (April-June):

$$Y_{t,\ modified} = f\ (SWE_{apr1st},\ Y_{t,\ forecasted}) \qquad (4.9)$$

Season 3(July-September):

$$Y_{t,\ modified} = f\ (SWE_{jun1st},\ Y_{t,\ forecasted}) \qquad (4.10)$$

The same data period (1961-2005), calibration set (1961-1999), testing set (2000-2005), cross validation stopping criteria, procedures that applied to enhance the generalization capability of networks that were used in the previous section, were also used in the ANN model development for forecast modification. To keep a smaller network size, the maximum processing elements in the hidden layers were selected as 3. The ANN (2-3-1) model structure was used for both seasons. The training process utilized the hyperbolic tangent function as the activation function in the hidden layer, linear function in output layer, and the momentum learning rule.

4.2.4.2 Combination of PCA and ANN

Following the Step 1 and Step 2 of the input variable selection procedure that was discussed in section 4.2.3.3, the initial input variables for the hybrid ANN model were selected based on the TFN model structure, sample cross correlation function between precipitation and seasonal flow, and cross correlation analysis of SWE, SOI and seasonal flow. Seven initial variables were used in building the hybrid ANN models for both season 2 and season 3. They were: average seasonal precipitation of three previous seasons, snow water equivalents of three previous months, and previous seasonal flow.

Based on the procedures described in section 4.1.2, the first 4 principal components for season 2 and the first 2 principal components for season 3 were selected as the inputs for ANN models. The structures of models for second and third seasons are described as follows:

For season 2 (April-June):

$$Y_t = f\,(Z_1, Z_2, Z_3, Z_4\,) \qquad (4.11)$$

For season 3 (July-September):

$$Y_t = f\,(Z_1,\, Z_2\,) \tag{4.12}$$

Where

Z_i = the i th principal components of the initial input variables.

The same data period (1961-2005), calibration set (1961-1999), testing set (2000-2005), model training algorithm, cross validation stopping criteria, procedures that applied to enhance the generalization capability of networks that were used in the previous sections were also used in the ANN model development for the hybrid approach. To keep a smaller network size, the maximum processing elements in the hidden layers were selected as 2 for second season, and 3 for third season. Hence, the model structures of ANN (4-2-1) and ANN (2-3-1) were applied to second season and third season respectively.

4.2.5 Model Diagnostics and Comparison

4.2.5.1 Comparison of Models for Test Data

The performance of models was compared by performing one-season-ahead forecasting of all models for the testing data set which covered from 2000 to 2005. The ARIMA and TFN models were developed using continuous time series data, hence one-season-ahead forecasts for the second (April-June) and third season (July-September) were made with developed ARIMA and TFN models. The ANN models were only developed for these two seasons. The two hybrid approaches, the modification of TFN model forecasts with ANN and combination of PCA and ANN, were also applied to these two seasons. Table 4.3 shows some performance statistics of the models for the testing data from 2000 to 2005.

It was observed from the study that the TFN model forecasts could be improved by ANN method using latest snow water equivalent information for only for the April-June season. The TFN forecasts for July-September season were not improved by modification using ANN with latest snow water equivalent. This may be because the precipitation information from April to July is adequate to reflect the basin snow water equivalent information in the modeling. In this season, the temperature in the basin is rising, the main form of precipitation is rainfall, and the rainfall also contributes to the melting of snowpack. The combination of PCA and ANN also did not improve the forecasts of July-September season. Only the ANN model provided slightly better forecasts than did other models. In general, there was no improvement in the forecasts of July-September season by using hybrid models, since their performance was essentially the same as the ARIMA model.

Table 4.3 Performance comparison of models for second and third seasons for test data set at Del Norte Gaging Station (2000-2005)

Models	APRIL-JUNE			JULY-SEPTEMBER		
	R^2	RMSE(kaf)	NRMSE	R^2	RMSE	NRMSE
ARIMA	0.40	145	0.93	0.13	67	0.93
TFN	0.64	111	0.71	0.04	67	0.93
ANN	0.77	87	0.56	0.23	61	0.85
TFN+ANN	0.84	72	0.47	*	*	*
PCA+ANN	0.85	67	0.43	0.12	64	0.89

** No improvement*

As indicated in the Table 4.3 and Figure 4.5, the model performance improved significantly when hybrid models were applied to April-June season flow forecasts. The best performing model would be the combination of PCA and ANN, and TFN with forecasts modification coming second, indicating that there is a potential capability of hybrid approaches to improve forecast accuracy as compared to the single models. For example, the RMSE of forecasted and observed flow for April-June decreased from 145 acre-ft to 67 acre-ft from the ARIMA model to the combination of PCA and ANN approach. The normalized RMSEs of different model forecasts for July-September (as shown in Figure 4.5) were almost same. Although a slightly smaller NRMSE was reported for the ANN model and the combination of PCA and ANN, there was no significant improvement using any of these models. One-season-ahead forecasts of the TFN model for the first season (January-March) and fourth season (October-December) were better than the ARIMA model with smaller NRMSEs. The ANN and hybrid models were not developed for these seasons, since there was no meaningful relationship existing between SWE and the flow of these seasons (Table 4.1). The TFN with precipitation input was considered to be sufficient for the modeling of the flow of these seasons.

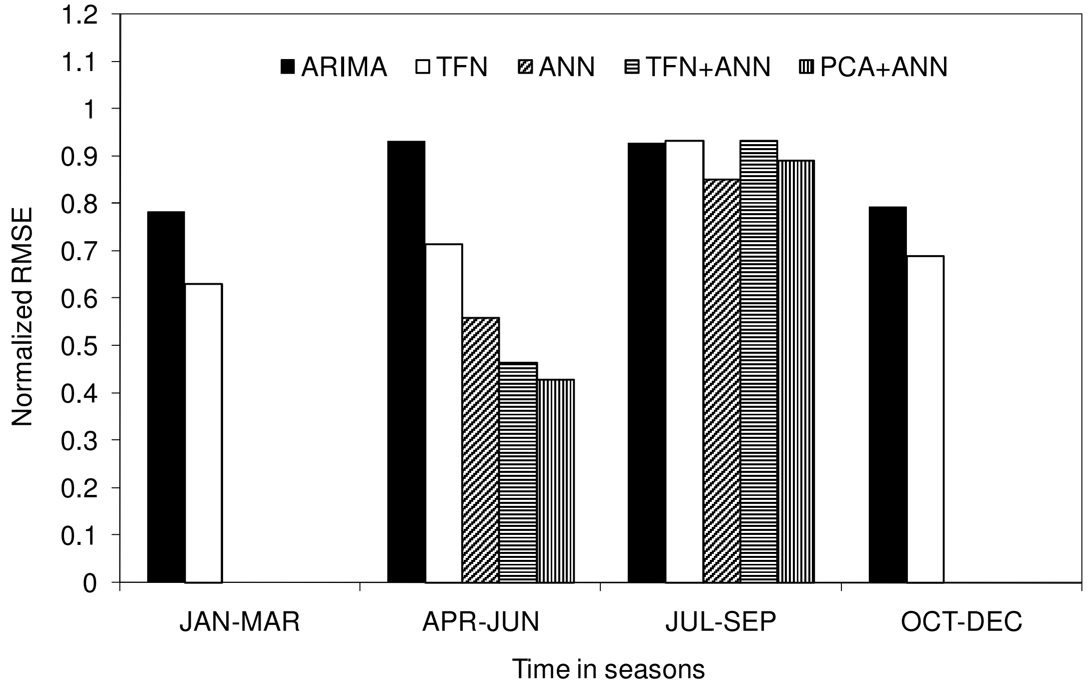

Figure 4.5 Comparison of normalized RMSEs of different model forecasts for all seasons for the test period (2000-2005) at Del Norte Gaging Station

The comparison of forecasted and observed seasonal flow time series from 2000 through 2005 is shown in Figure 4.6. As can be seen, the distinct improvement of forecasts can be observed for season 2 (April-June) using hybrid approaches. The April-June flow forecasts made by TFN model were significantly smoothed out by using hybrid modeling for all years except 2001 and 2005. There was a significant improvement in forecast accuracy on 2000, 2002 and 2004 using hybrid modeling approaches, again showing the effectiveness of the hybrid modeling approach in improving forecast accuracy of seasonal flow compared to single models. Figure 4.6 also suggested the importance of improving the forecast accuracy of the April-June flow since a considerable amount of the annual total flow occurs during this season, as compared to the other three seasons.

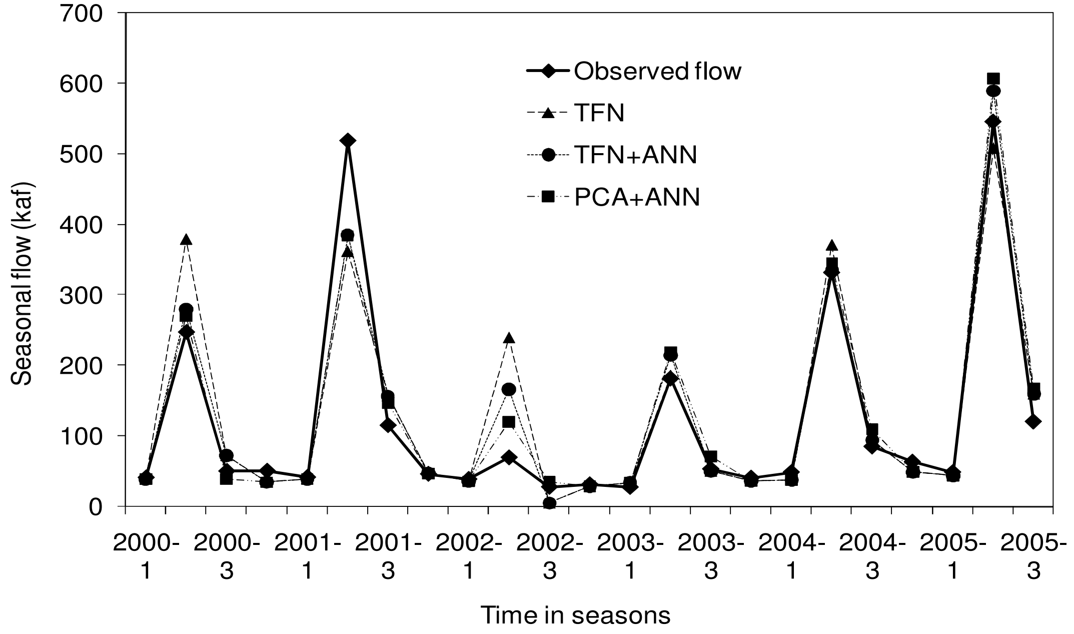

Figure 4.6 Comparison of observed and forecasted seasonal flow using TFN and second and third seasons modified by hybrid models for the testing data at Del Norte Gaging Station (2000-2005)

4.2.5.2 Comparison to NRCS Official Forecasts

To compare the performance of hybrid modeling approaches with NRCS official forecasts at the site for April-September seasonal runoff volume, the following models were developed and forecasts were made for the years from 2000 to 2005.

1) The forecast date was selected as April 1st ;
2) One-season ahead and two-season ahead forecasts were made using the developed TFN model, and then summed up the one-season-ahead forecasts for April-June and two-season-ahead forecasts for July-September to get the April-September flow forecasts on April 1st;
3) An ANN model using April-September flow as the dependent variable and using the same inputs as in Equation 4.6 was developed;
4) A TFN+ANN model using April-September flow as the dependent variable and using the same inputs as in Equation 4.8 was developed;
5) A PCA+ANN model using April-September flow as the dependent variable and same inputs as in Equation 4.10 was developed;

Several performance statistics of April-September seasonal volume runoff forecasts of all models are reported in Table 4.4. As can be seen, the model performance increased from TFN model to the hybrid models. However, the forecasts from the TFN model were not acceptable because of high forecasting errors. This was due to high forecast errors of two-season-ahead forecasting of July-September flow, which accumulated the error of one-season-ahead forecasts for April-June flow. The combination of PCA and ANN performed better compared to all other models and its forecasting performance was comparable to NRCS official forecasts. Considering the fact that the NRCS official forecasts are not the output of single model, instead are the coordinated approach among the several agencies, both hybrid modeling approaches, particularly the combined PCA and ANN approach performed very well for April-September runoff volume forecasting at Del Norte Gaging Station for the testing period of 2000-2005.

Table 4.4 Comparison of April 1st forecasts of different models with NRCS official forecasts for April-September volume of Del Norte Gaging Station for the period of 2000-2005 (n=6)

Models	R^2	MAE (kaf)	MAPE(%)	RMSE (kaf)	NRMSE	E
TFN	0.63	119	62	147	0.68	0.49
ANN	0.84	79	39	90	0.42	0.81
TFN + ANN	0.88	64	29	77	0.36	0.86
PCA + ANN	0.94	64	24	68	0.32	0.89
NRCS	0.97	59	25	65	0.30	0.90

4.3 Elephant Butte Net Inflow

4.3.1 Definition of Seasons

To develop seasonal time series models for Elephant Butte Reservoir net inflow, different seasons needed to be defined for the Rio Grande Basin. The same principles that were used in the definition of seasons in Rio Grande Headwaters Basin above Del Norte Gaging Station were applied in defining seasons for the Rio Grande Basin above Elephant Butte Reservoir.

To determine the pattern and magnitude of relationships between monthly net inflow and monthly SWE, a correlation analysis between the monthly SWE index and Elephant Butte Reservoir monthly net inflow was performed. The results (Table 4.5) indicated that the basin average SWE index which are measured in the first days of January to June are significantly correlated with the March to July Elephant Butte monthly net inflow at a 0.01 significant level. Although the correlations between March net inflow with February and January SWEs are statistically significant at the 0.01 significance level, they are not as high as other months. In general, the net inflow of March to July is largely contributed by the snowpack in the upper Rio Grande Basin including southern Colorado and northern New Mexico.

The NRCS provides March-July natural seasonal volume forecasts for the San Marcial Gaging Station, Rio Grande, which is located at the entrance of Elephant Butte Reservoir. The San Marcial natural seasonal volume forecasts are very important and comparable to the Elephant Butte Reservoir net inflow, since it is the main inflow to the Reservoir. The correlation coefficient of San Marcial March-July natural flow and Elephant Butte March-July measured net inflow (that has been calculated for the period 1961-2000) is 0.98, which indicates the importance of the NRCS forecasts at the site. In addition, Figure 1.9 also indicates that most of the Reservoir net inflow is concentrated in the month of March to July each year. Hence, the spring-summer net inflow forecasting is very important in Elephant Butte Reservoir operation and water management for the region.

Table 4.5 Correlation coefficient significance test between SWE index of the Basin and monthly Elephant Butte Reservoir net inflow (1961-1999)

Month	SWE					
	Jan1st	Feb 1st	Mar 1st	Apr 1st	May 1st	Jun 1st
Jan	0.11					
Feb	0.15	0.17				
Mar	0.35	0.37	**0.42**			
Apr	**0.41**	**0.56**	**0.63**	**0.64**		
May	**0.43**	**0.68**	**0.73**	**0.85**	**0.79**	
Jun	**0.41**	**0.69**	**0.71**	**0.79**	**0.76**	**0.44**
Jul	0.32	**0.44**	**0.42**	**0.53**	**0.56**	**0.49**
Aug	0.14	0.05	-0.05	-0.09	0.04	0.05
Sep	0.09	0.13	0.13	0.09	0.11	0.09
Oct	0.18	0.13	0.06	-0.02	0.02	-0.11
Nov	0.34	0.32	0.25	0.29	**0.44**	0.37
Dec	0.39	**0.41**	**0.40**	**0.48**	**0.57**	**0.43**

*Note. The **bold** numbers indicate the correlations are significant at 0.01significance level.*

Considering all the factors mentioned above and the need for generating equally or approximately equally spaced seasonal volume flow for time series modeling, the following seasons have been defined for Rio Grande Basin above Elephant Butte Reservoir:

Season 1: January, February, March

Season 2: April, May, June, July

Season 3: August, September, October

Season 4: November, December

According to the correlation analysis between snow water equivalent and Elephant Butte Reservoir net inflow, and the NRCS forecasting period, season 2 was considered in the modeling for this study. Moreover, the modeling results could be compared to NRCS forecasts at San Marcial Gaging Station as a reference for evaluating the capability of hybrid approaches used in this study in improving forecast accuracy of seasonal reservoir net inflow forecasting.

4.3.2 Data Description

4.3.2.1 Seasonal Net Inflow Time Series

After defining the seasons, the seasonal net inflow time series was calculated by summing the corresponding monthly net inflow into the seasonal net inflow volume. To obtain better approximation of normal distribution for time series modeling, the seasonal net inflow series was deseasonalized using seasonal average and standard deviation of the data period 1961-2007 as expressed in Equation 4.2. Figures 4.4 and 4.5 show the normal probability plots of the original seasonal net inflow series and deseasonalized seasonal net inflow series. They suggest that the normality approximation of the deseasonalized series improved greatly compared to the original series. But in both cases, the normality assumptions were rejected with small p-values (p-value < 0.001) when using Shapiro-Wilk normality test. The results were similar to natural seasonal flow at Del Norte Gaging Station, but the Del Norte flow approximated a normal distribution better than did the Elephant Butte Reservoir net inflow after deseasonalization transformation of the original data.

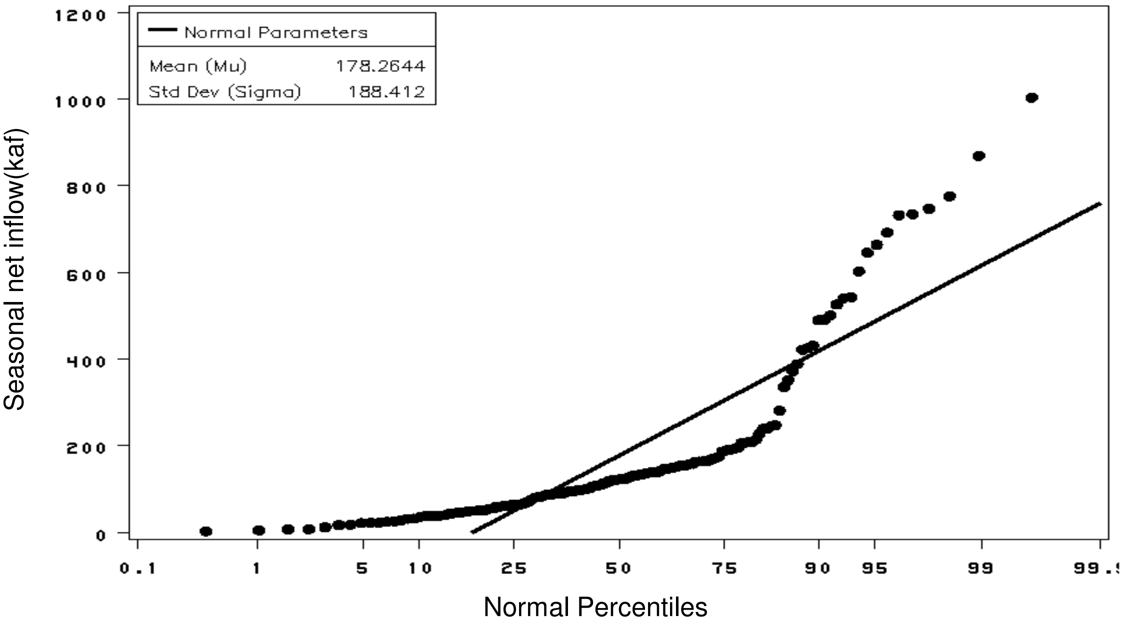

Figure 4.7 Normality plot of original seasonal net inflow series of Elephant Butte Reservoir, Rio Grande (1961-1999)

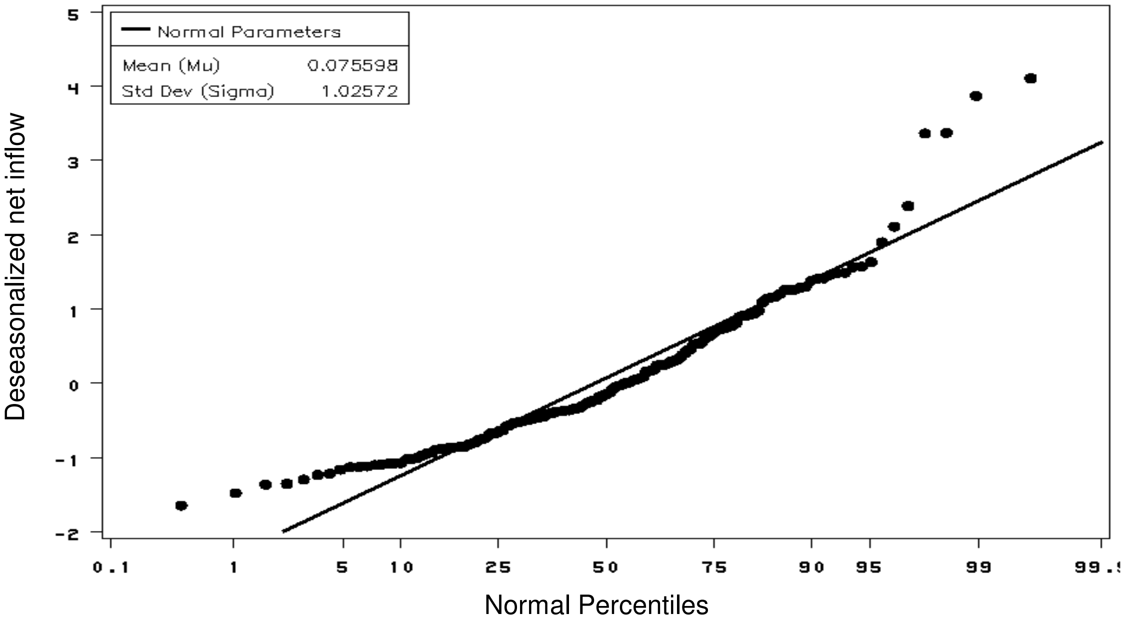

Figure 4.8 Normality plot of deseasonalized seasonal net inflow series of Elephant Butte Reservoir, Rio Grande (1961-1999)

4.3.2.2 Other Data

The average seasonal precipitation data from 1961 to 2007 was used in the seasonal time series modeling. The average seasonal precipitation of the Basin is calculated as the average of monthly average SNOTEL precipitation index (as shown in Figure 1.11) for a defined season. The SWE data used in this chapter were the monthly basin average SWE index data as shown in Figure 1.10 that also covered the period 1961-2007. The previous year October-December SOI data was also used in the modeling (Figure 1.12).

4.3.3 Development of Single Models

4.3.3.1 ARIMA

The deseasonalized seasonal net inflow series was used to fit the ARIMA model. The entire data set was divided into a calibration set and a test set, which covered the periods 1961-1999 and 2000-2007, respectively. The autocorrelation functions (ACF) and partial autocorrelation functions (PACF) suggested the use of AR(2) model. The following ARIMA model was developed for the deseasonalized series and it passed all the diagnostic checks:

$$y_t = 0.272 y_{t-1} + 0.25 y_{t-2} + a_t \tag{4.13}$$

Where

y_t , y_{t-1} , y_{t-2} = deseasonalized seasonal net inflow series at season *t*, *t-1, t-2* respectively.

The model explained only 45% of the variability of the Elephant Butte Reservoir net inflow. To evaluate the performance of the models, the one-season-ahead rolling forward forecasts were made for 2000-2007 and are discussed further in the following sections.

4.3.3.2 TFN

The sample cross correlation between the Elephant Butte Reservoir net inflow and basin average seasonal precipitation time series was performed by prewhitening the input series and filtering the output series using ARIMA models. The precipitation series was deseasonalized by subtracting means and dividing by seasonal standard deviations, and the autocorrelation function (ACF) showed the resulting series which suggested that the AR (6) (only lag 6 is significant) model was sufficient to prewhiten the precipitation series. The sample cross correlation function between two series is shown in Figure 4.9.

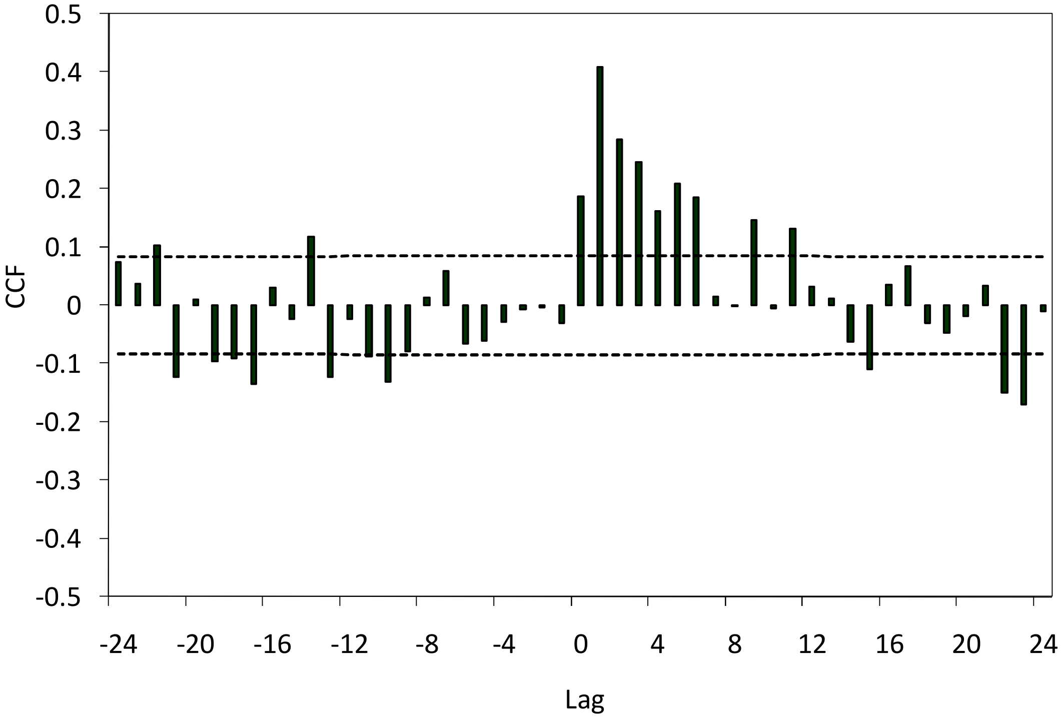

Figure 4.9 Cross correlation function between prewhitened deseasonalized seasonal precipitation and filtered deseasonalized seasonal Elephant Butte net inflow (1961-1999)

The CCF suggests that the following TFN model can be built for a deseasonalized seasonal flow series with the deseasonalized precipitation series as an input. The model parameters are estimated using conditional least squares method and the model has passed all the diagnostic check according to Box and Jenkins (1976) modeling procedure:

$$y_t = \frac{0.494}{(1-0.701B)} PRCP_{t-1} + \frac{1}{(1-0.213B^2-0.218B^7)} a_t \tag{4.14}$$

Or can be written as:

$$\begin{aligned}\hat{y}_t &= 0.701\hat{y}_{t-1} + 0.213y_{t-2} - 0.149y_{t-3} + 0.218y_{t-7} - 0.153y_{t-8} \\ &+ 0.494PRCP_{t-1} - 0.105PRCP_{t-3} - 0.108PRCP_{t-8} + a_t\end{aligned} \tag{ 4.15}$$

The built TFN model can explain 74% of the variability of the Elephant Butte Reservoir net inflow, indicating that the TFN model performed much better than the ARIMA model. The model performance of the deseasonalized seasonal flow series for calibration phase showed that the model variance and *AIC* of TFN model decreased from 0.868 to 0.600, and 423 to 362 respectively compared to the ARIMA model. Again, the TFN model showed significant improvement compared to ARIMA model.

4.3.3.3 ANN

In contrast to the time series models that were developed for continuous seasonal net inflow for all seasons, the ANN model was built for only season 2 (April-July) since the snow water equivalent (SWE) is significantly correlated with the April-July reservoir net inflow. Table 4.6 shows the correlation coefficient and its significance at the 0.05 level between the monthly SWE index, SOI and seasonal net inflow of Elephant Butte Reservoir. The results indicated that the basin average SWE indices, which are measured in the first days of January to June, are significantly correlated with the April-July seasonal net inflow, while the net inflow of the other seasons of a year showed systematic relationships with SWE in the Basin. The November-December season is also significantly related to SWE for some months, but relationships are weak and not meaningful.

According to previous research results in the Rio Grande Basin, the correlation analysis has been performed between the averaged June-November SOI, averaged October-December SOI and Elephant Butte Reservoir seasonal net inflow. The results (Table 4.6) suggested that the October-December SOI has a significant negative relationship with April-July net inflow of Elephant Butte Reservoir. The correlation coefficient is -0.40, which indicates that there tends to be a higher than average net inflow during El Niño years (when the SOI is negative), and lower than average net inflow during a La Niña (when the SOI is positive). This is consistent with the results reported by NRCS (1997).

The input variables were identified for the building of the ANN model for April-July season using the same procedure described in the ANN model building for Del Norte seasonal flow (section 4.2.3). The following ANN model was proposed based on the input variable selection analysis. The structure of model was formulated as follows:

$$Y_t = f\,(Y_{t-1},\ PRCP_{t-1},\ PRCP_{t-2},\ SOI,\ SWE_{apr1st}\ ) \qquad (4.16)$$

To train the ANN model, the total data period (1961-2007) was divided into a calibration set (1961-1999) and a testing set (2000-2007); then the calibration set was randomized and divided further into a training set and a cross validation set. The same cross validation stopping criteria, procedures that applied to enhance the generalization capability of networks that were used in Del Norte seasonal flow modeling were used in the ANN model development for April-July seasonal net inflow modeling. To keep a smaller network size, the maximum processing elements in the hidden layers was selected as 2 and a model structure of ANN (5-2-1) was used for training and testing. The training process utilized the hyperbolic tangent function as the activation function in the hidden layer, linear function in output layer, and the momentum learning rule.

Table 4.6 The correlation coefficients significance test between SWEs, SOI and Elephant Butte Reservoir seasonal net inflow (1961-2007)

Season	**Snow Water Equivalent (SWE)**						**Southern Oscillation Index**	
	Jan1st	**Feb1st**	**Mar1st**	**Apr1st**	**May1st**	**Jun1st**	**JUN-NOV Average**	**OCT-DEC Average**
JAN-MAR	0.245	0.271	**0.324**				-0.211	-0.206
p-value	*0.097*	*0.065*	***0.027***				*0.155*	*0.165*
APR-JUL	**0.466**	**0.711**	**0.749**	**0.852**	**0.786**	**0.432**	-0.279	**-0.401**
p-value	***0.001***	***<.0001***	***<.0001***	***<.0001***	***<.0001***	***0.002***	*0.057*	***0.005***
AUG-OCT	0.179	0.126	0.036	-0.031	0.067	0.017	0.216	0.127
p-value	*0.230*	*0.400*	*0.808*	*0.836*	*0.656*	*0.912*	*0.145*	*0.395*
NOV-DEC	**0.392**	**0.392**	**0.348**	**0.409**	**0.543**	**0.433**	0.082	-0.027
p-value	***0.007***	***0.006***	***0.017***	***0.004***	***<.0001***	***0.002***	*0.582*	*0.858*

Note: the ***bold*** *numbers indicate the correlations are significant at 0.05 significance level*

4.3.4 Development of Hybrid Models

4.3.4.1 Combination of TFN and ANN

Based on the model formulation explained in section 4.1.1, an ANN model was developed for modification of TFN model forecasts using latest snow water equivalent information and October-December averaged El Niño Southern Oscillation Index of a previous calendar year. It was also observed that the latest SWE information is adequate for forecast modification. Therefore, the adding of more SWE information from previous months does not contribute to the modification accuracy. Hence, the final ANN model for April-July net inflow has only three inputs. The structure of the network was shown as follows:

$$Y_{t,\ modified} = f\,(SWE_{apr1st},\ SOI,\ Y_{t,\ forecasted}) \qquad (4.17)$$

The same data period (1961-2007), calibration set (1861-1999), testing set (2000-2007), cross validation stopping criteria, procedures that applied to enhance the generalization capability of networks that were used in the previous section were utilized in the ANN model development. To keep a smaller network size, the maximum processing elements in the hidden layers was selected as 2. The ANN(3-2-1) model structure was used for training. The training process utilized the hyperbolic tangent function as the activation function in the hidden layer, linear function in output layer, and the momentum learning rule.

4.3.4.2 Combination of PCA and ANN

The initial input variables for the hybrid ANN model were selected based on the TFN model structure, sample cross correlation function between precipitation and seasonal flow, cross correlation analysis of SWE, and seasonal net inflow based on the Step 1 and Step 2 of variable selection procedure discussed in the section 4.2.3. In addition, the October-December averaged El Niño Southern Oscillation Index of a previous calendar year was also included in the inputs for the network. Together, nine initial variables were used for April-July net inflow hybrid ANN model building. They are average seasonal precipitation of three previous seasons, snow water equivalents of three previous months, SOI and two previous seasonal net inflows.

Based on the procedures described in section 4.1.2, the first 3 principal components were selected as the inputs for ANN model. The structure of model was described as follows:

$$Y_t = f\,(Z_1, Z_2, Z_3\,) \qquad (4.18)$$

Where

Z_i = the i th principal components of the input variables.

The same data period (1961-2007), calibration set (1961-1999), testing set (2000-2007), model training algorithm, cross validation stopping criteria, procedures that applied to enhance the generalization capability of networks that were used in the previous sections were used in the ANN model development. To keep a smaller network size, the maximum processing elements in the hidden layers were selected as 2. Hence, the model structure of ANN (3-2-1) was applied for training.

4.3.5 Model Diagnostics and Comparison

4.3.5.1 Comparison of Models for Test Data

The performance of models was compared by performing one-season-ahead forecasts of all models for the testing data set which was from 2000 to 2007. The ARIMA and TFN models were developed using continuous time series data, hence one-season-ahead forecasts for April-July season were made with developed ARIMA and TFN models. The ANN model and the two hybrid approaches, the modification of TFN models with ANN and combination of PCA and ANN, were applied only to the April-July net inflow.

Table 4.7 shows some performance statistics of one-season-ahead forecasts of the models for testing period. The two hybrid approaches performed well compared to single models. Particularly, the modification of TFN forecasts with ANN improved forecast accuracy significantly compared to TFN models. The RMSE of the observed and forecasted net inflow by the TFN+ANN hybrid approach was only a half of the RMSE of TFN model forecasts. In contrast to the performance in Del Norte seasonal flow modeling, the PCA+ANN hybrid approach did not perform significantly better than single ANN model in the Elephant Butte Reservoir net inflow modeling. This may be because the number of principal components included in the ANN model was three, since several other significant principal components such as the 8th and 9th components were not included in the model in order to keep the network size smaller. However, the network size of PCA+ANN was much smaller than the single ANN model, meaning that it may have a better generalization ability as compared to the single ANN model.

Table 4.7 Performance comparison of models for Elephant Butte Reservoir April-July net inflow for test data set (2000-2007)

Models	R^2	RMSE (kaf)	NRMSE
ARIMA	0.34	171	0.62
TFN	0.85	100	0.36
ANN	0.90	66	0.24
TFN+ANN	0.97	52	0.19
PCA+ANN	0.93	63	0.23

In general, the forecast accuracy was improved significantly when hybrid models were applied to April-July net inflow forecasts. The best-performing model was the TFN with forecast combination, and the combination of PCA and ANN came in second, again, indicating the potential capability of hybrid approaches in improving forecast accuracy as compared to the single models. For example, the normalized RMSE of forecasted and observed net inflow for April-July decreased from 0.36 to 0.19 from the TFN model to the TFN with forecast modification. The ANN and hybrid models were not developed for other seasons in a year, since no meaningful relationship exists between SWE and the flow of these seasons (Table 4.6).

Figure 4.10 describes the performance of the models for all seasons in a year by plotting the normalized RMSEs between observed and forecasted net inflow for the testing period. As mentioned earlier, only the April-July seasonal net inflow was modeled by those different models including ANN and hybrid approaches, while other seasons were modeled by ARIMA and TFN with precipitation input only. According to the definition of NRMSE, if NRMSE is greater than 1, it implies that the forecasts are not better than average. It can be seen that the April-July net inflow can be modeled by hybrid models with high forecast accuracy (the minimum RMSE was 0.2 only). While for the August-October net inflow, the historical average may be the best estimation because of high forecasting errors of ARIMA and TFN. This may be due to the fact that the August-October net inflow is highly affected by the monsoon season precipitation in the basin, for which both ARIMA and TFN models developed in this study are not accounted. As mentioned in the previous sections, the precipitation input for TFN model is a basin average SNOTEL precipitation index that represents the higher elevation regions of a basin.

As far as January-March and November-December net inflows are considered, the TFN model with precipitation input provided better forecasts than ARIMA with smaller NRMSE. Since the net inflows of both seasons are not related to basin SWE and SOI, it is difficult to find other readily available input variables for improving the model performance. In both seasons, the NRMSEs for TFN model forecasts were smaller than 0.5, particularly the NRMSE for November-December season as 0.36, indicating that the TFN model could be used for net inflow forecasting of these seasons.

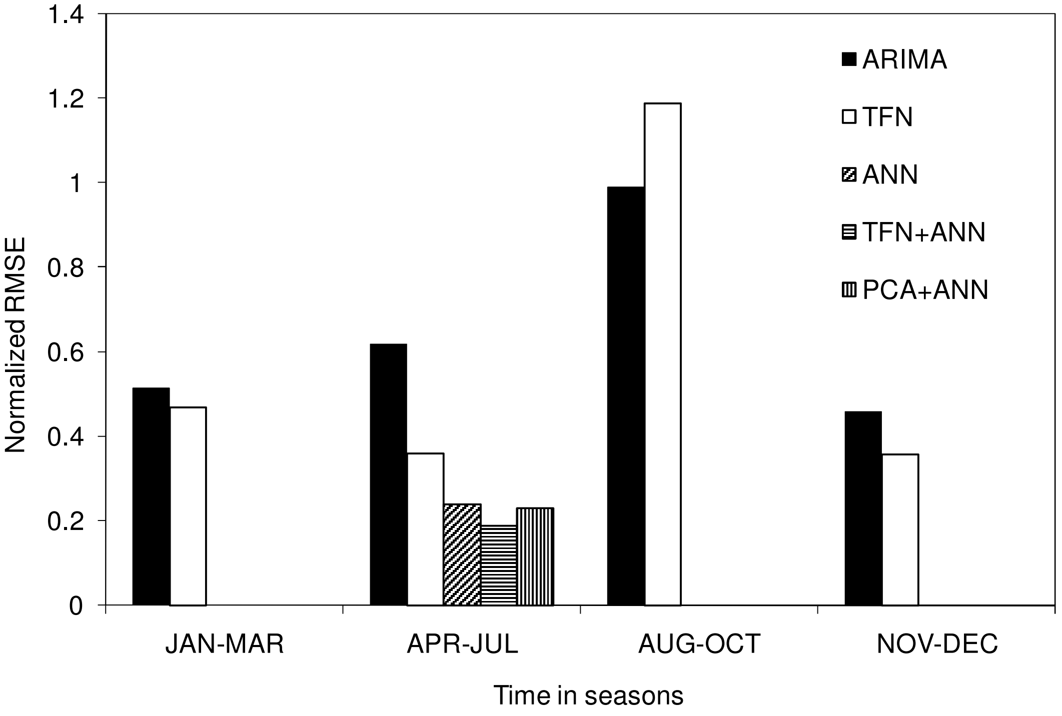

Figure 4.10 Comparison of normalized RMSEs of different model forecasts for all seasons for the test period (2000-2007) for Elephant Butte seasonal net inflow

To visualize the performance of different models for all seasons for the testing period, the comparison of forecasted and observed seasonal flow time series from 2000 through 2007 is plotted in Figure 4.11. As can be seen, the April-September flow forecasts made by TFN model were significantly smoothed out using hybrid modeling for all years except 2005. These significant improvements achieved by using hybrid modeling approaches showed the effectiveness of the hybrid modeling in improving forecast accuracy of seasonal flow compared to single models. It also suggested that the most of the unacceptable forecasts occurred in season 3 (August-October), especially during 2002 and 2006. This may be due to the exceptionally high rainfall that occurred in August 2006 in the lower part in the Basin, which made the Elephant Butte Reservoir net inflow as high as three times more than the historical averages. The possible reason for low forecasts of August-October net inflow of 2002 may be attributed to the very low observed precipitation in April-July season of this year (only one fourth of the historical average), which made one-season-ahead forecast of August-October flow worse. In general, the forecasting of August-October net inflow is associated with very high uncertainties mainly because of the randomness of monsoon precipitation of this season and omission of this information in the TFN models constructed in this study. Hence, none of the models used in this study was recommended for the forecasting of Elephant Butte Reservoir net inflow for August-October season.

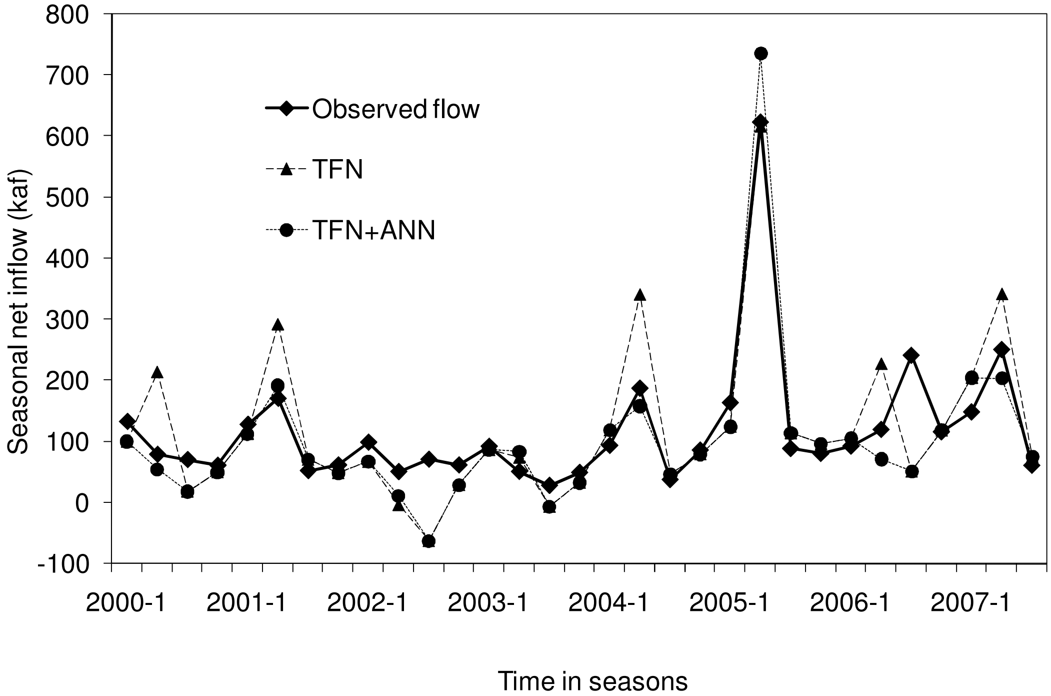

Figure 4.11 Comparison of observed and forecasted net inflow using TFN and April-July modified forecasts by ANN for the test data (2000-2007)

4.3.5.2 Comparison to NRCS official forecasts

To further examine the performance of the models, a comparison was made to NRCS official forecasts at San Marcial Gaging Station as a reference. The NRCS March-July volume official forecasts on April 1st from 2000 to 2007 were obtained from NRCS and converted into Elephant Butte March-July net inflow forecasts using the developed routing equation (Equation 3.6) proposed in section 3.3.3. To compare the April-July net inflow forecasts with converted NRCS March-July official forecasts, the observed March net inflow was added to all April-July net inflow forecasts of the different models to produce April 1st forecasts of March-July net inflow volume. Although it was not a direct comparison, the NRCS converted forecasts for Elephant Butte net inflow by using routing equation may provide some insights to examine the performance of hybrid models developed in this study.

The comparison of different models to the NRCS converted forecasts is shown in Table 4.8 using several performance statistics that calculated for the March-July net inflow forecasts of testing period from 2000-2007. Among all models the TFN with forecast combination performed best in terms of NRMSE and coefficient of determination. The model efficiency (Nash and Sutcliffe, 1970) of the approach is 0.92, which indicates that the hybrid approach showed very satisfactory performance. The combination of PCA and ANN approach performed slightly better compared to the single ANN model. But its performance was not as good as TFN with forecast modification. The converted NRCS forecasts for March-July net inflow were not as good as any model forecasts. For example, the NRMSE of converted forecasts is almost twice of the NRMSE of TFN with forecast modification. This indicates that the routed forecast solution using NRCS official forecasts may not be applicable for April 1st forecasts of March-July Elephant Butte Reservoir seasonal net inflow.

Table 4.8 Comparison of April 1st forecasts of different models with converted NRCS official forecasts for Elephant Butter Reservoir March-July net inflow for the period of 2000-2007

Models	R^2	MAE (kaf)	MAPE(%)	RMSE (kaf)	NRMSE	E
TFN	0.86	86	55	100	0.34	0.71
ANN	0.91	55	36	66	0.22	0.87
TFN+ANN	0.97	44	25	52	0.18	0.92
PCA+ANN	0.94	55	40	63	0.21	0.88
NRCS Converted	0.77	85	61	105	0.36	0.68

4.4 Final Models and Summary

The application of two hybrid modeling approaches including a forecast modification using a combination of transfer function - noise (TFN) with artificial neural networks (ANN), and the combination of principal components analysis (PCA) with ANN has been analyzed and discussed in the previous sections. Two hydrologic variables, the seasonal streamflow volume at Del Norte Gaging Station, and seasonal net inflow at Elephant Butte Reservoir, Rio Grande, were used for modeling through a detailed analysis of the relationships among the different variables at a seasonal time scale. Based on the modeling results and discussions, the following models were suggested for use in the operational forecasting of the streamflow in the study basins.

For the modeling of seasonal natural streamflow at Del Norte Gaging Station, the combination of PCA and ANN (Equation 4.10) performed better for April-June and April-September streamflow volume forecasting. The performance of the approach is comparable to the NRCS official forecasts. Hence, the approach could be used for April 1st forecasting of two seasonal volumes in the Basin. The TFN with forecast combination (Equation 4.8) could also be used as an alternative model to forecast April-June and April-September streamflow volume due to its similar performance compared to PCA+ANN approach and NRCS official forecasts. However, none of the models performed with reasonable accuracy for the forecasting of July-September streamflow since all the normalized RMSEs of the forecasted and observed streamflow were greater than 0.85 and close to 1.0 (Figure 4.5), which indicated that there was no significant difference, compared to using the historical average as the forecasts. Therefore, it is not recommended to use any of the models developed in the study to forecast July-September streamflow because of high forecast errors.

The January-March and October-December seasonal streamflows at Del Norte Gaging Station have been forecasted reasonably well using TFN model with precipitation input. Although the streamflow volumes for these seasons seemed not as important as spring-summer volume for water management in the basin due to their smaller contribution to the annual runoff, forecasting of streamflows for these seasons may provide an indication of magnitude of spring-summer runoff volume early and they can be forecasted using time series models with reasonably accuracy because of the smaller interannual variation (Figure 1.3). In this study, the TFN model with SNOTEL precipitation input provided reasonable forecasts for both seasons with the

NRMSEs smaller than 0.7 (Figure 4.5). Hence, the TFN model could be used for the forecasting of streamflow volume for these seasons in the basin.

For the modeling of Elephant Butte Reservoir seasonal net inflow, the best-performing model for April-July seasonal net inflow was the forecast modification of TFN with ANN approach. The performance of the single ANN model and combination of PCA and ANN approach were also comparable to the TFN with forecast modification approach. However, the latter is recommended for use in April-July net inflow volume forecasting due to its simpler structure and higher forecast accuracy. The converted forecasts made by the routing equation using NRCS official forecasts at San Marcial Gaging Station is not preferred due to the higher forecast errors compared to other modeling approaches. This may be because the relationship between the input variables and the Elephant Butte Reservoir seasonal net inflow is not linear due to human intervention and complex features of the net inflow that is affected by many factors, such as reservoir evaporation, seepage and the contribution of low-altitude rainfall to the net inflow. Most of the operational forecasts issued by NRCS are based on the linear regression equations.

The August-October seasonal net inflow of the Elephant Butte Reservoir is of highly variability from year to year, therefore it is difficult to forecast with reasonable accuracy using the models presented in this study. It was observed in this study that the August-October flow was not significantly related to any of the SNOTEL information, including SWE and SNOTEL precipitation. As indicated in Figure 4.10 the one-season-ahead forecasts of the ARIMA and TFN models were not better than the historical average. The normalized RMSEs of the forecasted and observed August-October net inflow were close to 1.0 or even higher, which indicates that the forecasts are not acceptable. Therefore, it is recommended not to use any of the models proposed in this study for forecasting of August-October Elephant Butte net inflow because of high forecast errors.

Again, similar to winter streamflow at the Del Norte Gaging Station, the January-March and November-December seasonal net inflow of Elephant Butte Reservoir can be forecasted using time series models with reasonable forecast accuracy. As shown in Figure 4.10, the normalized RMSEs of forecasted and observed seasonal net inflow for both seasons were smaller than 0.5; particularly the NRMSE for November-October net inflow forecasts by TFN model is 0.36. This suggests that the TFN model with SNOTEL precipitation as input is sufficient for forecasting seasonal net inflow volume for these seasons. This may be due to the smaller interannual variation of the net inflow in the winter seasons (as shown in Figure 1.9).

In conclusion, both hybrid modeling approaches used in the study showed a potential capability of improving forecast accuracy in seasonal streamflow modeling compared to single models. The results were consistent with the previous research reported in literature (e.g., Abrahart and See, 2002; Kişi, 2008; Jain and Kumar, 2007; See and Abrahart, 2001; See and Openshaw, 1999; See and Openshaw, 2000; Shamsheldin et al., 1997; Shamsheldin et al., 2002; Srinivas and Sirinivasan, 2001; Wang et al., 2005b). However, due to the limitation of this dissertation study, the forecast uncertainty evaluation of these hybrid modeling approaches is not presented in this chapter. It is hoped that the future hydrological forecasting research efforts will also exploit the potential capabilities of hybrid modeling in achieving increased forecast accuracy in streamflow forecasting.

5 MONTHLY FLOW FORECASTING

Monthly streamflow forecasting is crucial for water resources allocation and management. Particularly, the monthly reservoir net inflow forecasting is of great importance to reservoir management as it is an indication of water availability from a reservoir; it can provide a basis for decisions concerning reservoir operation, and water management, and legal and institutional compliance purposes. In this chapter, the response of monthly streamflow processes to basin precipitation, snow water equivalent, El Niño Southern Oscillation (ENSO) was investigated using cross correlation analysis. Several statistical models including ARIMA, TFN, and ANN were built for monthly natural flow at Del Norte Gaging Station, Rio Grande, Colorado and reservoir net inflow at Elephant Butte Reservoir, Rio Grande, New Mexico. Then, one-month-ahead forecasts of those models for spring-summer season were modified used snow water equivalents and ENSO signals using ANN technique. The performance of different modeling approaches was compared with each other. Finally some general discussions and conclusions are presented at the end of the chapter.

5.1 Del Norte Natural Flow

5.1.1 Data Description

The monthly natural flow time series from 1961 to 2007 at the Del Norte Gaging Station was shown in Figure 1.2 (the 2006 and 2007 data were not used in the monthly modeling because they are not final data). To develop time series models for the monthly flow, a deseasonalization transformation was performed to original data using monthly average and standard deviation to produce deseasonalized time series. The Equation 4.2 was used for the deseasonalization of monthly flow by replacing the seasonal time scale with monthly time scale. The normality plots of original monthly time series and deseasonalized monthly time series are shown in Figures 5.1 and 5.2. The normality plots show that the normality approximation of deseasonalized series improved significantly compared to the original series. Although the normality assumption for both series was rejected at the 0.05 significance level using the Shapiro-Wilk normality test, the deseasonalized series was used for time series modeling due to its better approximation of the normal distribution. Some research results also indicated that deseasonalization is an effective data pre-processing procedure for model development (Jain and Kumar, 2007; Wang et al., 2005a). Another reason for using deseasonalized time series models in this study is that it requires less model parameters as compared to seasonal autoregressive integrated moving average (SARIMA) models. Particularly in the building of TFN model, not only does it use fewer parameters, but also it is much easier to identify a model structure and calibrate models.

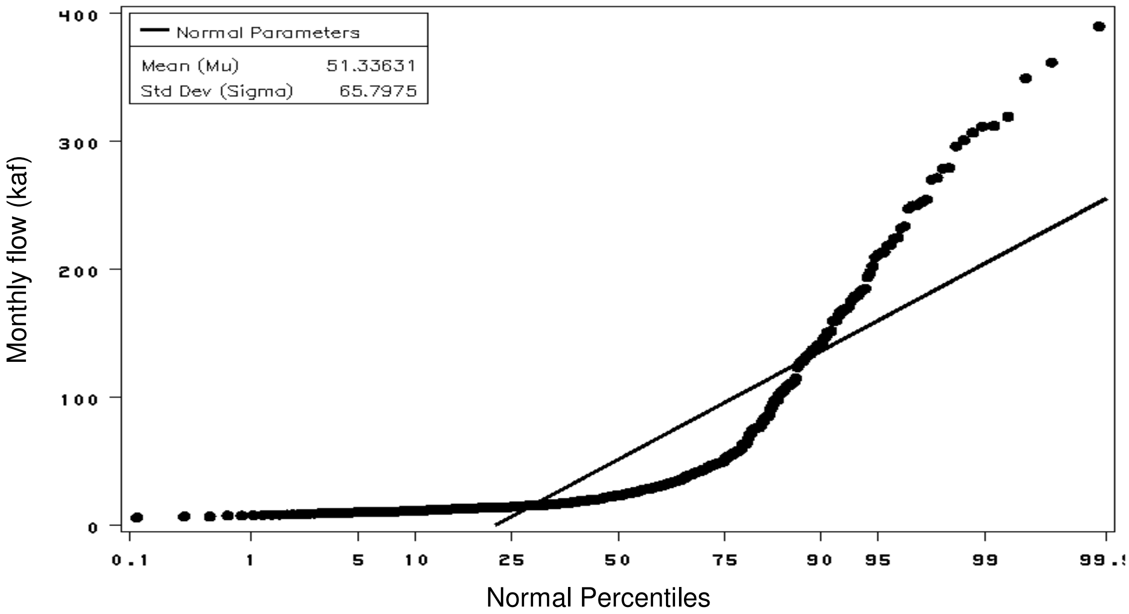

Figure 5.1 Normality plot of the original monthly flow at Del Norte Gaging Station (1961-1999)

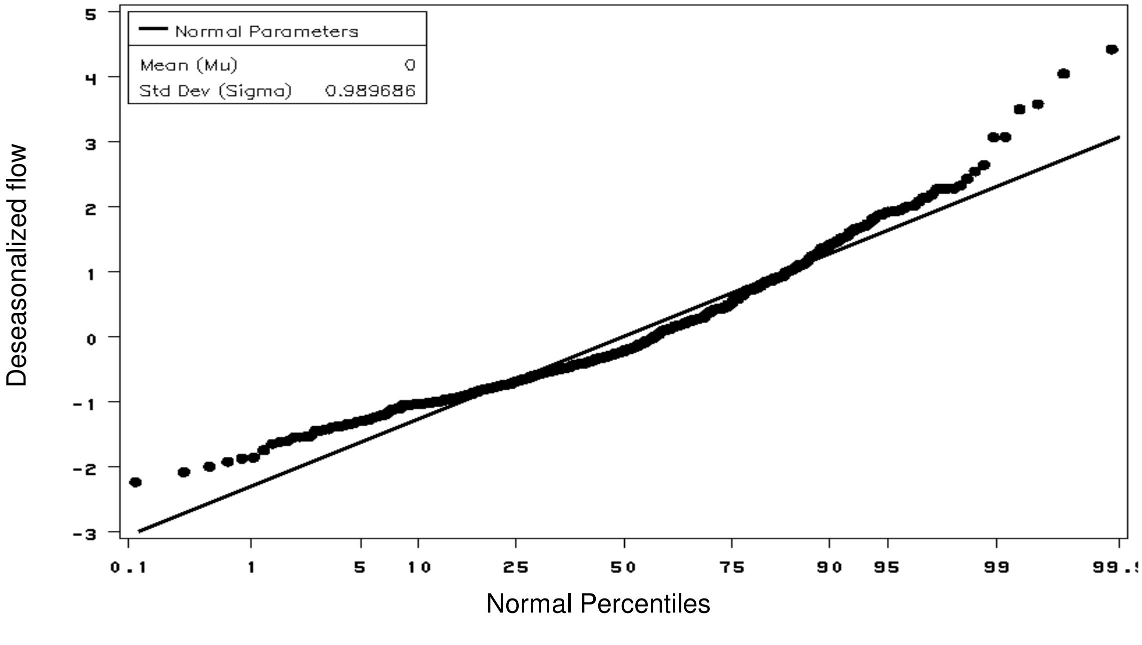

Figure 5.2 Normality plot of the deseasonalized monthly flow at Del Norte Gaging Station (1961-1999)

In addition to monthly flow data, the monthly average SNOTEL precipitation index of the Basin from 1961 to 2005 was used in monthly time series modeling. The calculation of monthly SNOTEL precipitation index was given in detail in section 1.4.1 and the data were shown in Figure 1.5. The SWE data used in this chapter were the monthly basin average SWE index data as shown in Figure 1.4 that also covers the period of 1961-2005. The same period SOI data (Figure 1.12) was also used in the correlation analysis.

5.1.2 Model Formulation and Development

5.1.2.1 Correlation Analysis between SOI, SWE and Monthly Streamflow

Based on the previous research results in the Rio Grande Basin (Redmond and Koch, 1991; NRCS, 1997), correlation analysis has been performed between the averaged June-November SOI, averaged October-December SOI and Del Norte monthly flow to determine if SOI could be a predictor for any specific month. The results (Table 5.1) suggested that both the averaged October-December SOI and the averaged June-November SOI do not have significant correlations with Del Norte monthly flow (the p-values for Pearson correlations significance test were greater than 0.08 for all the months). The results were somewhat similar to the results of NRCS (1997) in that there is no significant relationship between the SOI and streamflows in the Upper Rio Grande Basin. Hence, the SOI was not considered as a predictor in monthly streamflow modeling of Del Norte Gaging Station, Rio Grande.

Table 5.1 The correlation coefficients significance test between June-November average SOI, October-November average SOI and Del Norte monthly flow (1961-1999)

Monthly flow	Jun-Nov average SOI		Oct-Dec average SOI	
	Correlation coefficients	*p*-value	Correlation coefficients	*p*-value
Jan	-0.20	*0.19*	-0.16	*0.29*
Feb	0.05	*0.76*	0.09	*0.58*
Mar	0.15	*0.33*	0.25	*0.10*
Apr	0.26	*0.09*	0.27	*0.08*
May	0.10	*0.53*	0.04	*0.78*
Jun	-0.01	*0.93*	-0.19	*0.22*
Jul	0.00	*0.99*	-0.16	*0.31*
Aug	0.09	*0.54*	-0.03	*0.83*
Sep	0.00	*0.99*	-0.04	*0.81*
Oct	0.17	*0.28*	0.08	*0.59*
Nov	0.05	*0.75*	0.01	*0.95*
Dec	-0.14	*0.37*	-0.16	*0.29*

To determine the pattern and magnitude of the relationship between monthly flow at Del Norte Gaging Station and monthly SWE, a correlation analysis between the monthly SWE index and monthly flow at Del Norte Gaging Station was performed in section 4.2.1. The results (Table 4.1) indicated that the basin average SWE index which is measured on the first days of January to June is significantly correlated with the March to September monthly flow at 0.05 significant level. The correlations of March flow with March 1st SWE and September flow with May 1st SWE were 0.35 and 0.37 respectively, indicating that the correlations are very weak though they are statistically significant at the 0.05 significance level. The correlation coefficients between monthly streamflow from April to August and monthly SWE are significantly high, with the highest correlation coefficients of 0.92. Therefore, the basin SWE information is particularly useful in forecasting monthly streamflow from April to August of a year in the Rio Grande Headwaters Basin above Del Norte Gaging Station.

5.1.2.2 ARIMA

To fit an ARIMA model, the monthly deseasonalized flow series was divided into calibration set and test set, which covered the periods of 1961-1999 and 2000-2005 respectively. The calibration set data was used for the calibration and diagnostics of the model, and the test set data was used to test the model forecasting performance for new data. The autocorrelation functions (ACF) and partial autocorrelation functions (PACF) suggest AR (1) model could be fitted to deseasonalized series. The following ARIMA model was developed for the deseasonalized series and it passed all the diagnostic checks.

$$y_t = 0.643 y_{t-1} + a_t \tag{5.1}$$

Where

y_t, y_{t-1} = deseasonalized flow series at month *t* and *t-1* respectively.

The model explains 82% of variability of the Del Norte monthly flow. One-month-ahead rolling forward forecasts were made for 6 years (2000-2005) to evaluate the performance of the model. The one-month-ahead rolling forward forecast means the model parameters will be modified for every month forecasts when the new observations are available. For example, the January 2000 forecast is made by the model that was calibrated using the data from January 1961 to December 1999; the February, 2000 forecast is made by the model that was calibrated for the period of January 1961-January 2000; and so on. The ARIMA forecasts for April to September months were used for analysis in this Basin.

5.1.2.3 TFN

The cross correlation between monthly precipitation time series and Del Norte monthly flow was performed by prewhitening the input series and filtering the output series using ARIMA models. The precipitation series was deseasonalized by subtracting means and dividing by seasonal standard deviations, and the AR (1) model was fitted to the deseasonalized precipitation series based on the ACF and PACF of the series. The monthly PRCP series was prewhitened and the monthly flow series was filtered by using the model. The sample cross correlation function between two series is shown in Figure 5.3.

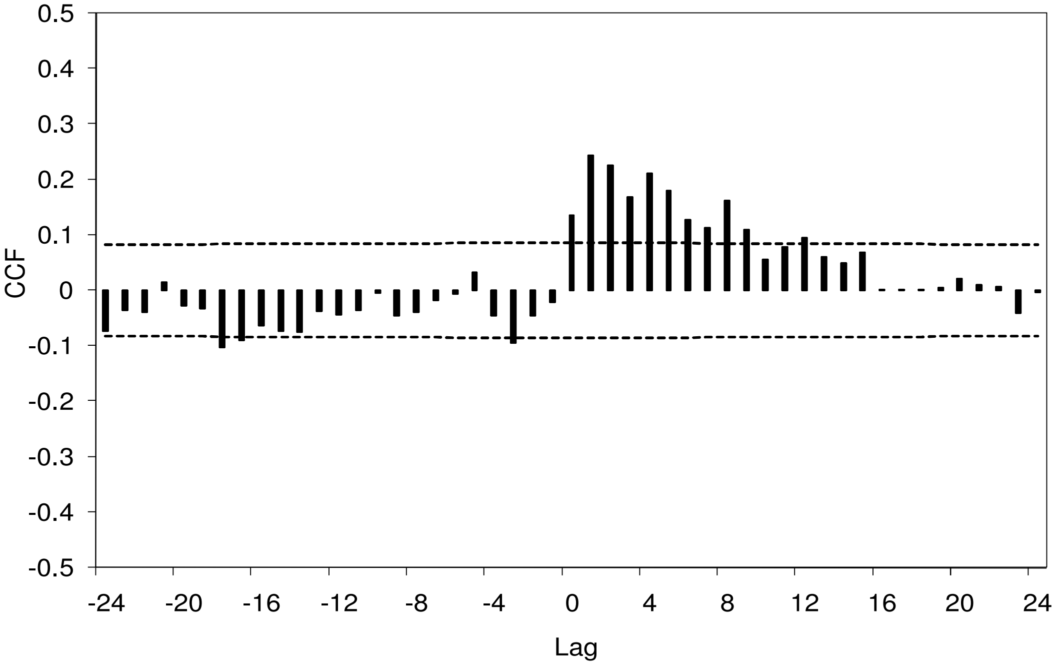

Figure 5.3 Cross correlation function (CCF) between prewhitened deseasonalized SNOTEL precipitation and filtered deseasonalized monthly flow at Del Norte Gaging Station (1961-1999)

Based on the structure of CCF in Figure 5.3, the following TFN model was built for deseasonalized monthly flow series with the deseasonalized monthly precipitation series as an input using calibration set data. The model parameters were estimated using the conditional least squares method and the model passed all the diagnostic checks according to the Box and Jenkins (1976) modeling procedure:

$$y_t = \frac{0.175}{(1-0.867B)} PRCP_{t-1} + \frac{1}{(1-0.49B-0.103B^2)} a_t \tag{5.2}$$

Equation 5.2 can also be written as:

$$\begin{aligned}\hat{y}_t &= 0.49y_{t-1} + 0.867\hat{y}_{t-1} - 0.322y_{t-2} - 0.089y_{t-3} + 0.175PRCP_{t-1} \\ &- 0.086PRCP_{t-2} - 0.018PRCP_{t-3}\end{aligned} \tag{5.3}$$

Where

$PRCP_{t-1}$, $PRCP_{t-2}$, and $PRCP_{t-3}$ = Basin average SNOTEL precipitation at month *t-1, t-2* and *t-3;*

$\hat{y}_t$, $\hat{y}_{t-1}$ = forecasted deseasonalized monthly flow at month *t, t-1.*

The proposed TFN model explained 87% percent variability of Del Norte monthly flow and indicated the better performance of the TFN model compared to ARIMA model. The model performance of the deseasonalized monthly flow series for calibration phase showed that the model variance and *AIC* of TFN model decreased from 0.590 to 0.546, and 1082 to 1044 respectively compared to the ARIMA model. The one-month-ahead rolling forward forecasts were made for all months of 2000-2005 using the TFN model. Again, only the TFN forecasts for April to September were compared and analyzed.

5.1.2.4 ANN

The correlation analysis between SWE index and monthly flow (as shown in Table 4.1) indicated that the correlations between SWE index and monthly flow are significant only for certain months of a year. Therefore, the ANN models were built for only the months that the flow has significant relationships with SWE, which are the months of April to September each year in this study. Figure 1.3 suggested that the large portion of the annual runoff occurs in the months of April to September for the Del Norte Gaging Station. The NRCS also provides April-September seasonal volume runoff forecasts on the first days of January to June of a year.

To build monthly ANN models, the input variables should be identified for each monthly model at the first step. The input variable selection procedure that was used for seasonal ANN model development in section 4.2.3 was utilized in the selection of input variables for the monthly ANN models in this study. The sample cross correlation function between prewhitened precipitation series and filtered flow series (Figure 5.3) suggested that the flow is significantly correlated with the Basin SNOTEL precipitation index back to four months prior. The TFN structure also indicated that the current month flow is strongly related to the flows of the previous two months. The results of the cross correlation analysis between SWE index and monthly flow suggested that up to three previous months of SWEs could be potential input variables for ANN, but the lag relationship could be different from month to month. Southern Oscillation Index (SOI) was not used as input variable because no significant correlation existed between SOI and Del Norte monthly flow.

According to input variable selection procedures and analysis of the relationships between predictors and dependent variables, a number of ANN models were separately developed for the months of April to September using SWE, PRCP and flow of the previous months as inputs to the models. The general structure of the monthly ANN models was formulated as follows:

$$Y_t = f\,(Y_{t-1},\ Y_{t-2},\ PRCP_{t-1},\ PRCP_{t-2},\ PRCP_{t-3},\ PRCP_{t-4},\ SWE_t,\ SWE_{t-1},\ SWE_{t-2}) \quad (5.4)$$

Where

Y_t ,Y_{t-1} ,Y_{t-2} = streamflow at month *t, t-1, t-2;*

SWE_t, SWE_{t-1} ,SWE_{t-2} = snow water equivalents on the first day of month *t, t-1, t-2.*

After initial variables were selected (as in Equation 5.4), a simple correlation analysis was carried out for these potential variables to keep important variables for each month's ANN model. Variables that had correlation coefficients smaller than 0.4 were dropped from the input variable list. The remaining variables after the screening were used as the inputs for each ANN model. However, the maximum number of input variables was set to five in order to keep smaller network sizes for each model to avoid network overfitting. Depending on the correlations between monthly flows and candidate input variables, the number of previous SWE, PRCP, and flow inputs and lags were not completely same for ANN models for different months. The final monthly ANN models developed for April to September using SWE, PRCP and previous flow as inputs are tabulated in Table 5.2.

Table 5.2 The ANN models developed for monthly flows at Del Norte Gaging Station, Rio Grande

Months	Model structure	Model configuration
April	ANN (4-2-1)	$Y_{apr} = f\ (Y_{mar}, PRCP_{feb}\ , SWE_{apr1st}, SWE_{mar1st})$
May	ANN (5-2-1)	$Y_{may} = f\ (Y_{apr}, PRCP_{apr}\ , PRCP_{feb}\ ,SWE_{may1st}, SWE_{apr1st})$
June	ANN (5-2-1)	$Y_{jun} = f\ (Y_{may}, PRCP_{may}\ , PRCP_{apr}\ ,SWE_{jun1st}, SWE_{may1st})$
July	ANN (5-2-1)	$Y_{jul} = f\ (Y_{jun}, PRCP_{may}\ , PRCP_{apr}\ ,SWE_{jun1st}, SWE_{may1st})$
August	ANN (5-2-1)	$Y_{aug} = f\ (Y_{jul}, PRCP_{may}\ , PRCP_{apr}\ ,SWE_{jun1st}, SWE_{may1st})$
September	ANN (3-2-1)	$Y_{sep} = f\ (Y_{aug}, PRCP_{aug}\ , SWE_{may1st})$

To build ANN models, the total data period (1961-2005) was divided into a calibration set (1961-1999) and a testing set (2000-2005); then the calibration set was randomized and divided further into a training set and a cross validation set. Cross validation was used for early stopping of the training to protect the network from overtraining. One hidden layer was used in all ANN network types in this study. Previous research results (Coulibaly et al., 2000; Zhang et al., 1998) indicated that one hidden layer may be enough for most forecasting problems. The nodes in the hidden layer were decided based on the number of inputs for each month's ANN model. The maximum number of the nodes was selected as 3, so as to keep the number of network weights less than 10. The detailed procedure that was used to prevent the networks from overtraining was described in section 4.2.3 for seasonal ANN model building. The same ANN model building procedures that were used in chapter 4, such as model structure, training algorithm, transfer function selection, number of epochs for training and cross validation, were used in the building of monthly ANN models.

5.1.2.5 TFN with Forecast Modification

The correlation analysis between the monthly flow and basin SWE index (as shown in Table 4.1) showed that the April to August monthly flow are highly correlated with basin SWE index. It is apparent that the SWE should be included in the monthly flow forecasts from April to August each year in the Basin. However, it is difficult to build monthly streamflow forecasting TFN models with snowpack information as an input. A limited number of studies have been reported in the inclusion of snowpack information in TFN model for monthly flow forecasting. Thompstone et al. (1985) built a TFN model using calculated snowmelt series as an input for quarter monthly flow and obtained better results compared to univariate ARIMA models.

The direct inclusion of snow water equivalent information in the monthly TFN models is challenging. This may be because of following reasons. First, the SWE information is only available in winter and spring seasons. There is no SWE information in the summer months of a year, and it could be zero or near zero from July to November in the Upper Rio Grande Basin. This makes it difficult to get equally spaced, continuous SWE time series that could be included in the TFN time series modeling. Secondly, because of the seasonality of SWE information, it is less likely to get systematic cross correlation relationship between monthly SWE and monthly streamflow for entire year in the building of TFN model for continuous monthly time series. For example, all monthly flows from May to September at Del Norte Gaging Station are highly correlated with May 1st SWE only, indicating that there is no systematic lag relationship existed between the monthly flow time series and the monthly SWE index in the Basin. Even if TFN models with SWE index as input were assumed to have been built, the calibrated model parameters would not be sound and /or the best estimates for some months of the year.

The inclusion of SWE information in the forecasting of April to September monthly flow is of vital importance since the April-September monthly streamflows are mainly related to the SWE in the Basin. To address this, a hybrid modeling procedure proposed in chapter 4, the forecast modification of TFN model with PRCP input using SWE information, was performed using the artificial neural networks (ANN) method for the months of April to September in the Basin. The general formulation of TFN with modification approach was described using Equation 4.1 in section 4.1.1. The only difference was the monthly time scale was used here instead of seasons.

Based on the results of cross correlation analysis and SWE data availability in the Basin, a number of monthly ANN models were developed for modification of TFN model forecasts for each month. The final monthly ANN models developed for April to September using previous SWE and TFN forecasts as inputs are shown in Table 5.3.

Table 5.3 The monthly ANN models developed for forecast modification at Del Norte Gaging Station, Rio Grande

Months	Model structure	Model configuration
April	ANN (3-2-1)	$Y_{apr,\ modified} = f\ (SWE_{mar1st},\ SWE_{apr1st},\ Y_{apr,\ forecasted})$
May	ANN (3-2-1)	$Y_{may,\ modified} = f\ (SWE_{apr1st},\ SWE_{may1st},\ Y_{may,\ forecasted})$
June	ANN (3-2-1)	$Y_{jun,\ modified} = f\ (SWE_{may1st},\ SWE_{jun1st},\ Y_{jun,\ forecasted})$
July	ANN (3-2-1)	$Y_{jul,\ modified} = f\ (SWE_{may1st},\ SWE_{jun1st},\ Y_{jul,\ forecasted})$
August	ANN (3-2-1)	$Y_{aug,\ modified} = f\ (SWE_{may1st},\ SWE_{jun1st},\ Y_{aug,\ forecasted})$
September	ANN (2-3-1)	$Y_{sep,\ modified} = f\ (SWE_{may1st},\ Y_{sep,\ forecasted})$

Notes: $Y_{t,\ modified}$ = *monthly flow forecasts after the forecast modification in month t*
$Y_{t,\ forecasted}$ = *monthly flow forecasts of TFN model with PRCP input in month t*

The same data partitioning, training and cross validation procedure, and cross validation stopping criteria that were used in section 5.1.2.4 were utilized in ANN model building in this section. The significant difference is that the maximum number of inputs for ANN models in this procedure was limited to three, so as to keep a much smaller network size compared to previous section and enhance the generalization ability of the ANN with smaller sample size.

5.1.3 Model Diagnostics and Comparison

Using the ARIMA and TFN models developed in section 5.1.2, the one-month-ahead forecasts were performed for January 2000 to September 2005. Then, April to September flows from 2000 to 2005 were forecasted using ANN models, and the forecast modification for the TFN forecasts for the same period was made by using the hybrid modeling approach described in chapter 4. Table 5.4 summarized the one-month-ahead forecasting performance of the TFN model for monthly flow for both the calibration and the testing period. The R^2, RMSE and NRMSE were tabulated for each month, April to September, and the whole year. It can be seen that the testing phase coefficient of determination for the whole year was 0.85, which suggests fairly good performance of the model. However, the coefficient of determination of some months were very low, such as those for March, April, and September, which indicates that model forecasts for these months may not be acceptable even though the forecasting performance indices for the whole year was good. These results suggested that the forecasting performance of the model should be evaluated for each individual month when developing a monthly streamflow model.

Table 5.4 One-month-ahead forecasting performance of TFN model for Del Norte monthly flow

Month	Calibration period (1961-1999)			Forecasting period (2000-2005)		
	R^2	RMSE (kaf)	NRMSE	R^2	RMSE (kaf)	NRMSE
JAN	0.30	3.0	0.87	0.67	1.5	0.43
FEB	0.46	1.7	0.73	0.85	0.9	0.38
MAR	0.35	4.2	0.81	0.20	4.6	0.90
APR	0.25	16.3	0.88	0.48	11.8	0.63
MAY	0.41	42.2	0.70	0.79	55.9	0.92
JUN	0.65	60.5	0.61	0.89	51.7	0.52
JUL	0.83	21.4	0.48	0.97	5.5	0.12
AUG	0.29	19.4	0.85	0.85	4.9	0.21
SEP	0.33	16.5	0.85	0.00	11.9	0.61
OCT	0.34	11.5	0.84	0.63	3.5	0.26
NOV	0.79	3.4	0.51	0.50	2.3	0.35
DEC	0.68	2.1	0.59	0.90	1.1	0.30
APR-SEP	0.82	33.7	0.43	0.82	31.9	0.40
YEAR	0.87	24.1	0.37	0.85	23.2	0.35

As mentioned earlier, the spring-summer snowmelt runoff of the Rio Grande Headwaters Basin above Del Norte occurs from April to September. Hence, the monthly ANN models and forecast modification models were developed for April-September months and their performance were analyzed and compared with one another. Table 5.5 illustrated the performance of the different models for one-month-ahead forecasting of monthly flow from April through September for the period of

2000-2005. The results showed that forecast performance improved significantly from simple ARIMA model to TFN model with precipitation input, to ANN models that were calibrated for each month using previous SWE, PRCP, flow as the inputs, and to the TFN with forecast modification using SWE information. It can be seen that the TFN model with forecast modification performed better than any other modeling method. There was a significant improvement in forecast performance compared to the simple ARIMA model and the TFN model with precipitation input. The overall forecast performance of ANN models that were calibrated for each month was not as good as the TFN with forecast modification. Similar results were obtained for the seasonal flow forecasting in chapter 4.

Table 5.5 Forecast performance of different models for the April to September of 2000-2005 at Del Norte Gaging Station

Models	R^2	MAE (kaf)	MAPE(%)	RMSE (kaf)	NRMSE	E
ARIMA	0.75	27.8	62	41.5	0.52	0.67
TFN	0.82	19.5	38	31.9	0.40	0.80
ANN	0.87	15.3	36	25.5	0.32	0.87
TFN with modification	0.92	12.5	25	21.1	0.27	0.91

Figure 5.4 shows the improvement of forecast accuracy using TFN with forecast modification method compared to ARIMA, TFN and ANN models by

plotting forecasted and observed April to September monthly flows for the period of 2000-2005. As can be seen, the low to medium-high flow forecasts which deviated from the observed flow using ARIMA and TFN models were successfully smoothed using TFN with forecast modification. The correlation coefficients between forecasted and observed monthly flow for the April to September of the testing period was increased from 0.86 to 0.96 from simple ARIMA model to TFN with forecast modification. Although the TFN with forecast modification was not very effective in modification of high flows, it performed well in forecasting low to medium-high flows.

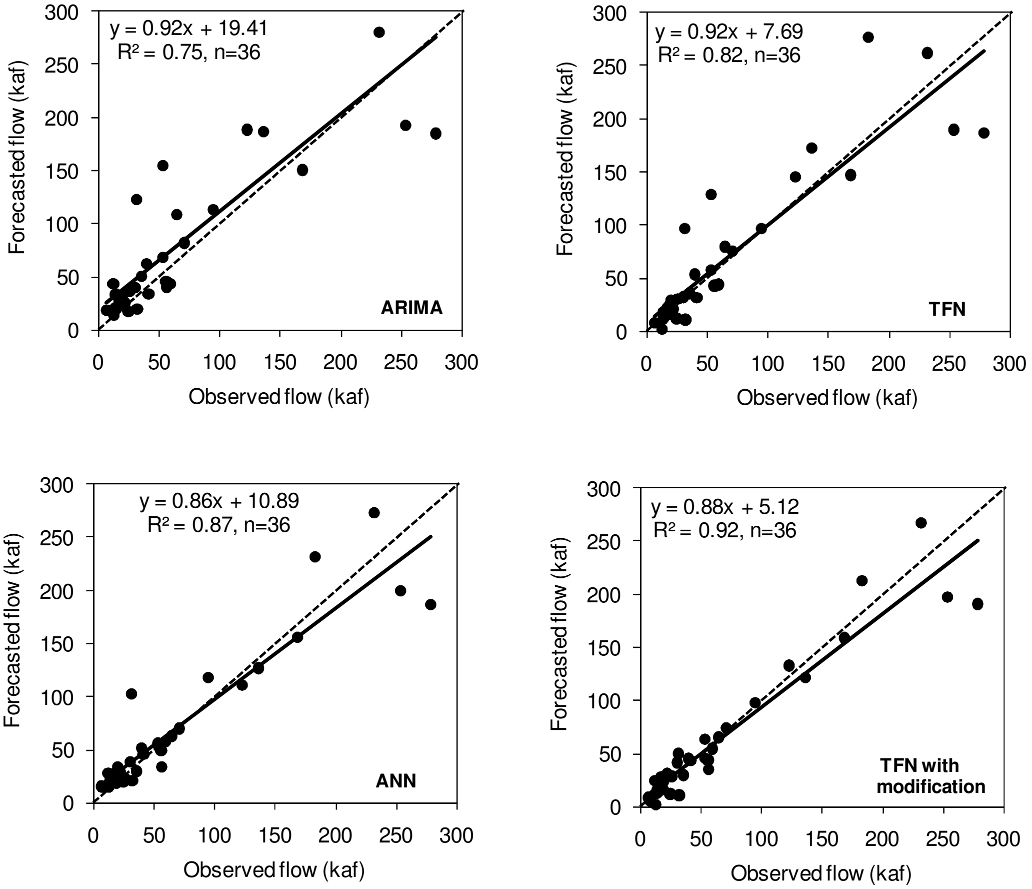

Figure 5.4 Scatter plots of observed and forecasted monthly Del Norte flow using ARIMA, TFN, ANN and TFN with modification for the April to September months of 2000-2005

As shown in Table 5.5, the TFN with forecast combination had a coefficient of determination 0.92, a model efficiency of 0.91, and a normalized root mean squared error of 0.27 for the whole spring-summer season of April to September at Del Norte Gaging Station, which indicated a very good one-month-ahead forecast performance and a satisfactory model. However, the results vary for model performance for each month from April to September. Table 5.6 and Figure 5.5 illustrate the forecast performance of different models for individual months from April to September. It can be inferred that no forecast improvement was found for September after using TFN forecast combination and ANN modeling technique using SWE information as the inputs. Rather, the ARIMA model was the best-performing model with the lowest normalized root mean squared error of 0.45 for September flow .

There was substantial improvement in forecasting performance by using ANN or TFN with forecast modification for the months of April to August. When comparing the forecast accuracy of different months, the May, June, July and August flows were forecasted with high accuracy, while the April flow forecasts were fairly acceptable. For example, the forecast RMSEs of TFN with modification for June and July were 20120 and 4520 acre-ft, and were less than one third of the highest forecasting RMSEs (75890 and 13650 acre-ft) by using the simple ARIMA model. Overall, the forecast modification with SWE information using ANN method showed potential capability of improving monthly streamflow forecasting accuracy in Del Norte Gaging Station, Rio Grande.

Table 5.6 Forecast performance of modeling methods for different months from April September at Del Norte Gaging Station (2000-2005)

Models	**ARIMA**		**TFN**		**ANN**		**TFN with modification**	
Month	**R^2**	**NRMSE**	**R^2**	**NRMSE**	**R^2**	**NRMSE**	**R^2**	**NRMSE**
April	0.17	0.79	0.48	0.63	0.47	0.62	0.52	0.61
May	0.65	1.05	0.79	0.92	0.94	0.89	0.96	0.73
June	0.89	0.76	0.89	0.52	0.97	0.27	0.99	0.20
July	0.98	0.31	0.97	0.12	0.96	0.12	0.97	0.10
August	0.72	0.38	0.85	0.21	0.97	0.29	0.88	0.27
September	0.01	0.45	0.00	0.61	0.01	0.46	0.00	0.61

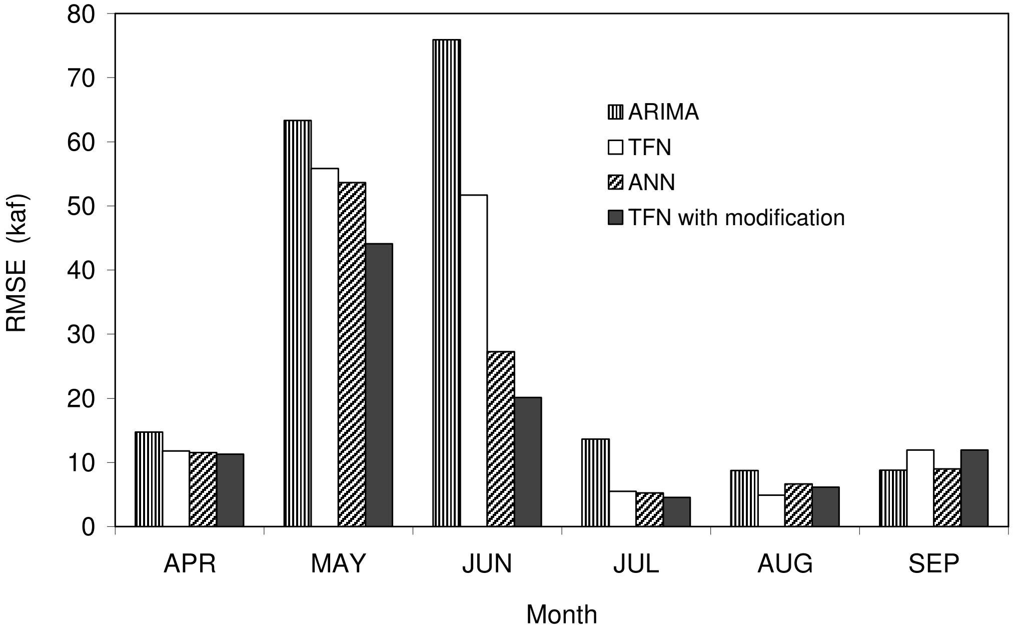

Figure 5.5 Comparison of April to September forecast RMSEs of different modeling methods for Del Norte monthly flow (2000-2005)

5.1.4 Final Models

Based on the comparison and performance analysis of different models in the previous sections, the forecasting models that can be used for one-month-ahead flow forecasting at Del Norte Gaging Station can be summarized according to the performance of models for each month of a year. The TFN with forecast modification can be used for one-month-ahead flow forecasting of April, May, June, July and August of a year. For the other months of a year, except September, the TFN model with precipitation input may be used to provide reasonable monthly forecasts. The final TFN model that could be used for one-month-ahead flow forecasting at Del Norte Gaging Station has been calibrated using the data period of 1961-2005 and is given as follows:

$$y_t = \frac{0.182}{(1-0.857B)} PRCP_{t-1} + \frac{1}{(1-0.483B-0.1B^2)} a_t \qquad (5.5)$$

Equation 5.5 can also be written as:

$$\begin{aligned} y_t &= 0.483y_{t-1} + 0.857\hat{y}_{t-1} - 0.314y_{t-2} - 0.086y_{t-3} + 0.182PRCP_{t-1} \\ &- 0.088PRCP_{t-2} - 0.018PRCP_{t-3} + a_t \end{aligned} \qquad (5.6)$$

The final models that can be used to provide Del Norte monthly flow one-month-ahead forecasts for each month of a year were proposed in Table 5.7. However, some conclusions presented in the table were based on the testing performance of the models for only 2000 to 2005. More data may be needed for further testing of the model performance to reach more accurate conclusions.

Table 5.7 Final models that could be used for one-month-ahead forecasting of Del Norte monthly flow

Month	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
Deseasonalization equation	$y_t = \frac{Y_t - \overline{Y}_t}{\hat{\sigma}_t}$											
One-month-ahead forecast equation	$\hat{y}_t = 0.483 y_{t-1} + 0.857 \hat{y}_{t-1} - 0.314 y_{t-2} - 0.086 y_{t-3} + 0.182 PRCP_{t-1}$ $- 0.088 PRCP_{t-2} - 0.018 PRCP_{t-3}$											
Forecast modification using SWE	NO	NO	NO	YES	YES	YES	YES	YES	YES	NO	NO	NO
Improvement by modification	-	-	-	YES	YES	YES	YES	YES	NO	-	-	-
Backtransforming	$Y_t = \overline{Y}_t + \hat{\sigma}_t y_t$											

Notes: y_t = *deseasonalized series for month t*
Y_t = *original series for month t*
$\overline{Y}_t$ = *sample average of the original series for month t*
$\hat{\sigma}_t$ = *sample standard deviation of the original series for month t*

5.2 Elephant Butte Reservoir Net Inflow

In the previous sections, different statistical models were developed to improve one-month-ahead forecast accuracy for the monthly natural flow at Del Norte Gaging Station, Rio Grande, Colorado. The time series models such as ARIMA, TFN and ANN models were proposed and their one-month-ahead forecasting performance was analyzed and compared. The proposed methodologies showed favorable results in forecasting monthly natural flow in the spring-summer season in the Basin. The monthly Elephant Butter Reservoir net inflow, which is defined as the sum of monthly releases measured below Elephant Butte Reservoir and the monthly change in storage of the Reservoir, is heavily regulated and liable to human intervention. As a result, the correlation of some predictors with net inflow has been weakened due to regulation. This inevitably results in more difficulties in modeling procedure and less forecast accuracy compared to Del Norte natural flow. However, the importance of Elephant Butte Reservoir net inflow forecasting in reservoir operation and water management in the region is one of the main motives to conduct this study. In the following sections of this chapter, the application of the methodologies in the modeling of the monthly Elephant Butter Reservoir net inflow, Rio Grande was examined and discussed.

5.2.1 Data Description

The Elephant Butte Reservoir monthly net inflow time series from 1961 to 2007 is shown in Figure 1.8. To develop time series models for the monthly net inflow, a deseasonalization transformation was performed to original data using monthly average and standard deviation to produce deseasonalized time series. Equation 4.2 was used for the deseasonalization of monthly flow by replacing the season with month. The normality plots of monthly net inflow time series and deseasonalized monthly net inflow time series are shown in Figures 5.6 and 5.7. The normality plots showed the normality approximation of deseasonalized series improved significantly compared to the original data series. Hence, as in Del Norte monthly flow modeling, the deseasonalized series was also used in the Elephant Butte net inflow time series modeling.

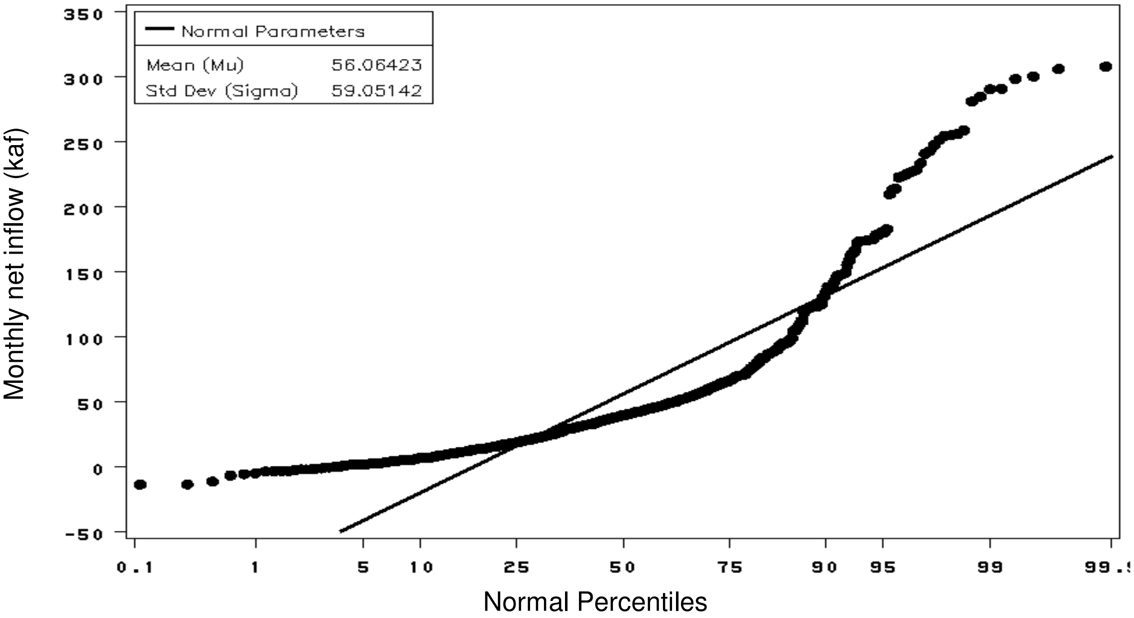

Figure 5.6 Normality plot of the monthly net inflow of Elephant Butte Reservoir (1961-1999)

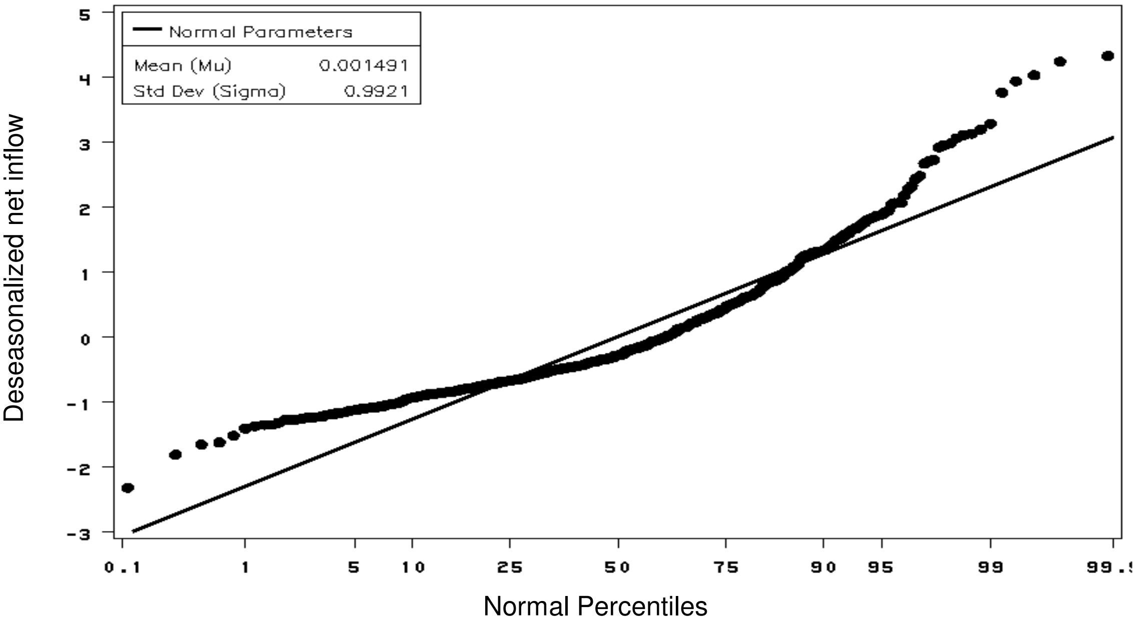

Figure 5.7 Normality plot of the deseasonalized monthly net inflow of Elephant Butte Reservoir (1961-1999)

In addition to monthly net inflow data, the monthly average SNOTEL precipitation index of the Rio Grande Basin above Elephant Butte Reservoir from 1961 to 2007 was used in monthly time series modeling. The calculation of monthly SNOTEL precipitation index was given in detail in section 1.4.1 and the data was shown in Figure 1.11. the SWE data used in this chapter was the monthly basin average SWE index data for the period of 1961-2007 (as shown in Figure 1.10). The previous year June-November and October-December SOI data were also used in the modeling (Figure 1.12).

5.2.2 Model Formulation and Development

5.2.2.1 Correlation Analysis between SOI, SWE and Monthly Net Inflow

A correlation analysis between the monthly SWE index and Elephant Butte Reservoir monthly net inflow was performed in order to determine the pattern and magnitude of relationship between the two variables. The results (Table 4.5) indicate that the basin average SWE index which was measured in the first days of January to June were significantly correlated with the March to July Elephant Butte monthly net inflow at the 0.05 significant level. Although the correlations between March net inflow and February, January SWEs were statistically significant at the 0.05 significance level, they are not as high as in other months. The December and November net inflow were also significantly correlated with May 1st and June 1st SWEs, but these relationships were not realistic because the effect of snowmelt runoff on Elephant Butte Reservoir net inflow is only up to July each year. Hence, the SWEs were not used as the predictors to net inflows from August to December. The monthly SWEs from March to June have been used in the forecast modification for the March to July net inflow.

In addition, a correlation analysis has been performed between the averaged June-November SOI, averaged October-December SOI and Elephant Butte Reservoir monthly net inflow to determine if SOI could be a predictor for any specific month. Previous studies have suggested that there is a negative correlation between the averaged June-November SOI and average monthly October-March streamflow and precipitation for the southwest United States (Redmond and Koch, 1991). The March-

July flow has the highest correlations coefficients with October-December SOI index, and the correlation coefficients are less than -0.35 in some part of the Middle Rio Grande (NRCS, 1997). This indicates that higher than average streamflow may occur during El Niño years (when the SOI is negative), and lower than average streamflow may occur during La Niña (when the SOI is positive). The correlation results (Table 5.8) suggest that the averaged October-December SOI has higher correlation coefficients with May and June net inflow than the June-November SOI, and only the May and June net inflow of a year are significantly related to the October-December SOI with the correlation coefficients less than -0.4. Therefore, the October-December SOI could be used as potential predictors in forecasting Elephant Butte Reservoir net inflow of those specific months.

Table 5.8 The correlation coefficients significance test between June-November average SOI, October-November average SOI and monthly Elephant Butte Reservoir net inflow (1961-2007)

Monthly Net Inflow	Jun-Nov average SOI		Oct-Dec average SOI	
	Correlation coefficients	*p*-value	Correlation coefficients	*p*-value
Jan	-0.062	*0.682*	-0.034	*0.825*
Feb	-0.151	*0.316*	-0.174	*0.247*
Mar	**-0.298**	***0.044***	-0.281	*0.058*
Apr	-0.201	*0.179*	-0.198	*0.188*
May	**-0.296**	***0.046***	**-0.405**	***0.005***
Jun	-0.272	*0.068*	**-0.422**	***0.004***
Jul	-0.175	*0.246*	-0.288	*0.052*
Aug	0.145	*0.337*	0.064	*0.671*
Sep	0.195	*0.193*	0.075	*0.623*
Oct	0.161	*0.286*	0.160	*0.288*
Nov	0.090	*0.552*	0.002	*0.987*
Dec	0.053	*0.727*	-0.062	*0.683*

*Note: the **bold** numbers indicate the correlations are significant at 0.05 significance level*

5.2.2.2 ARIMA

To fit an ARIMA model, the monthly net inflow series was deseasonalized using monthly average and monthly standard deviation of the data period 1961-2007, then the data set was divided into a calibration set and a test set, which covered the periods of 1961-1999 and 2000-2007 respectively. The autocorrelation functions (ACF) and partial autocorrelation functions (PACF) suggested a AR(1) model could be fitted to deseasonalized net inflow series. The following ARIMA model was developed for the series and it passed all the diagnostic checks:

$$y_t = 0.602 y_{t-1} + a_t \qquad (5.7)$$

Where

y_t, y_{t-1} = deseasonalized net inflow series at month *t* and *t-1* respectively.

The model explains 62% of variability of the Elephant Butte Reservoir monthly net inflow. One-month-ahead rolling forward forecasts were made for 8 years (2000-2007) to evaluate the performance of the model. Only the ARIMA model forecasts for March to July were compared and analyzed.

5.2.2.3 TFN

To build a TFN model using basin precipitation as an input, the relationship of two time series was investigated using cross correlation analysis. Based on the sample cross correlation functions between the input series and output series, the appropriate form of TFN model was suggested. The sample cross correlation between the Elephant Butte monthly net inflow and monthly precipitation time series was performed by prewhitening the input series and filtering the output series using ARIMA models. The precipitation series was deseasonalized by subtracting means and dividing by seasonal standard deviations. The AR(1) model was fitted to the deseasonalized precipitation series based on the ACF and PACF of the series. The monthly PRCP series was prewhitened and the monthly net inflow series was filtered by using the same model. The sample cross correlation function between two series is shown in Figure 5.8.

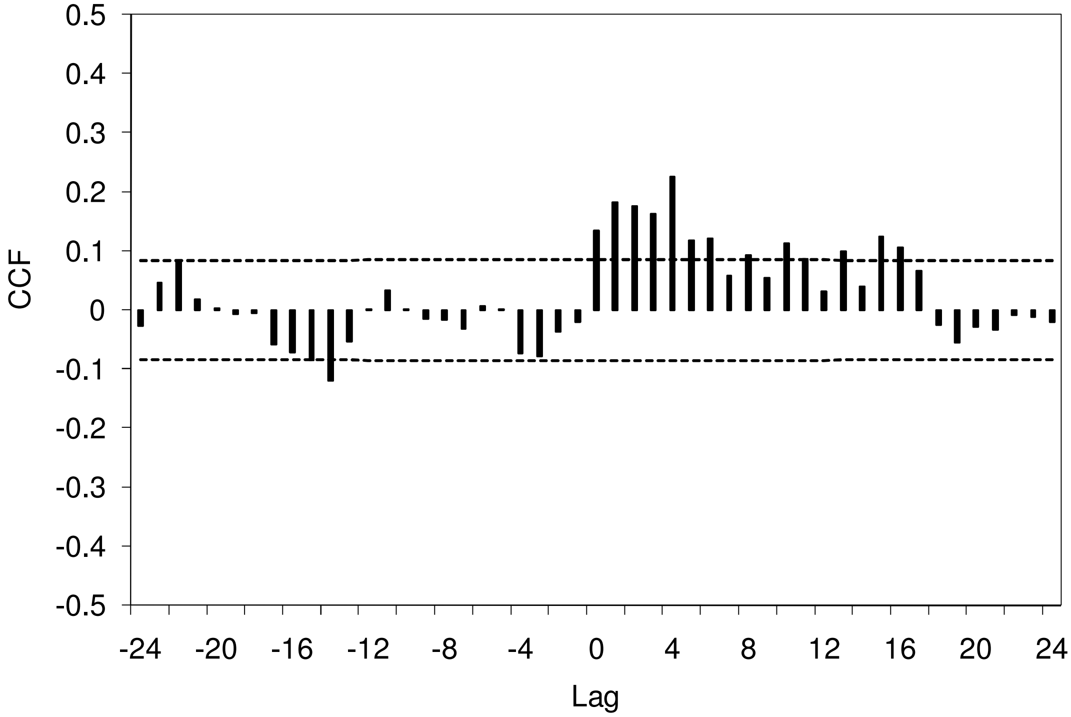

Figure 5.8 Cross correlation function between prewhitened deseasonalized precipitation and filtered deseasonalized net inflow (1961-2007)

The following TFN model was built for the deseasonalized monthly net inflow series with the deseasonalized precipitation series as an input using calibration set data. The model parameters were estimated using conditional least squares method and the model passed all the diagnostic checks according to Box and Jenkins (1976) modeling procedure:

$$y_t = \frac{(0.182 + 0.083B^3)}{(1 - 0.863B)} PRCP_{t-1} + \frac{1}{(1 - 0.474B + 0.097B^2)} a_t \tag{5.8}$$

Equation 5.8 can also be written as:

$$\begin{aligned}\hat{y}_t = {} & 0.377 y_{t-1} + 0.863 \hat{y}_{t-1} - 0.409 y_{t-2} + 0.084 y_{t-3} + 0.182 PRCP_{t-1} - 0.086 PRCP_{t-2} \\ & + 0.018 PRCP_{t-3} + 0.083 PRCP_{t-4} - 0.039 PRCP_{t-5} + 0.008 PRCP_{t-6}\end{aligned} \tag{5.9}$$

The built TFN model can explain 70% of variability of Elephant Butte monthly net inflow, indicating that TFN model performed better than ARIMA model. This was also explained by the model variance and Akaike Information Criterion (*AIC*) (Akaike, 1974) of both models. The model performance of the deseasonalized net inflow series for the calibration phase showed that the model variance and AIC of TFN model decreased from 0.685 to 0.614, and 1152 to 1095 respectively compared to the ARIMA model. The one-month-ahead rolling forward forecasts were made for 2000-2007 using the TFN model for all the months. Again, only the TFN forecasts for March to July were compared and analyzed.

5.2.2.4 ANN

The cross correlation analysis in section 4.3.1 (Table 4.5) showed that the correlations between SWE index and net inflow are significant only during certain months of a year. Therefore, the ANN models were built for only the months that the net inflow has significant relationships with SWE, which were the March to July months of a year in this study. The Figure 1.9 also suggested that the spring-summer net inflow, which is mainly contributed by snowmelt runoff in the Basin, occurs in the months of March to July for the Elephant Butte Reservoir inflow. The NRCS seasonal volume runoff forecasts at San Marcial Gaging Station is also focused on the March-July flow which is very important in the operation of Elephant Butte Reservoir.

To build monthly ANN models, the input variable selection procedure that was used for seasonal ANN model development in section 4.2.3.3 was also utilized in the selection of input variables for the monthly ANN models in this study. The sample cross correlation function between prewhitened precipitation series and filtered net inflow series (Figure 5.8) suggested that the net inflow is significantly correlated with the basin SNOTEL precipitation index prior to four months. The TFN structure also indicated that the current month flow is strongly related to the net inflows of previous two months. The results of cross correlation analysis between SWE, SOI and monthly net inflow in the previous sections suggested that up to three previous months of SWEs could be potential input variables for ANN, but the lag relationship could be different from month to month. The SOI has a stronger relation with the May and June net inflow, while there is no significant correlation existing between SOI and the net inflow for the rest of the months of a year.

Based on the input variable selection procedures discussed above, a number of ANN models were developed for the months of March to July separately using SWE, PRCP and net inflow of the previous months as the inputs. The general structure of the ANN models was formulated as follows:

$$Y_t = f\,(Y_{t-1}, Y_{t-2}, PRCP_{t-1}, PRCP_{t-2}, PRCP_{t-3}, PRCP_{t-4}, SWE_t, SWE_{t-1}, SWE_{t-2}, SOI) \quad (5.10)$$

After initial input variables for ANNs were selected, the same procedure that had been used to select input variables for the Del Norte monthly flow modeling was applied in order to enhance the generalization capability of the networks. Depending on the correlations between monthly net inflows and candidate input variables, the number of previous SWE, PRCP, SOI and net inflow inputs and lags were different for ANN models for different months. The structures of final monthly ANN models developed for March to July are shown in Table 5.9.

Table 5.9 The ANN models developed for monthly Elephant Butte Reservoir net inflow from March to July

Months	Model structure	Model configuration
March	ANN (4-2-1)	$Y_{mar} = f\ (Y_{feb}, PRCP_{nov}\ , PRCP_{feb}\ ,SWE_{mar1st})$
April	ANN (5-2-1)	$Y_{apr} = f\ (Y_{mar}, PRCP_{dec}\ , PRCP_{feb}\ ,SWE_{mar1st}, SWE_{apr1st})$
May	ANN (5-2-1)	$Y_{may} = f\ (Y_{apr}, PRCP_{jan}\ , PRCP_{mar}\ ,SWE_{may1st}, SOI)$
June	ANN (5-2-1)	$Y_{jun} = f\ (Y_{may}, PRCP_{apr}\ , PRCP_{may}\ ,SWE_{apr1st}, SOI)$
July	ANN (4-2-1)	$Y_{jul} = f\ (Y_{jun}, PRCP_{mar}\ , PRCP_{may}\ , SWE_{may1st})$

To build ANN models, the total data period (1961-2007) was divided into a calibration set (1961-1999) and a testing set (2000-2007); then the calibration set was randomized and divided further into a training set and a cross validation set. Cross validation was used for early stopping of the training to protect the network from overtraining. The same ANN model building procedures used in chapter 4 and ANN modeling of Del Norte monthly flow including model structure, training algorithm, transfer function selection, number of epochs for training and cross validation, were used in building of monthly ANN models.

5.2.2.5 TFN with Modification

The results of correlation analysis between basin SWE index, SOI and Elephant Butte monthly net inflow showed that the March to July monthly net inflows were highly correlated with SWEs. The October-December averaged SOI has a significant relationship with the May and June net inflow. As mentioned in previous sections, the inclusion of these variables in the TFN modeling procedure is very difficult since there is no systematic lag relationship between SWE, SOI and the dependent variables. However, the inclusion of these variables, particularly the SWE, in forecasting March to July net inflow is of vital importance since the March to July streamflows are mainly related to the SWE in the Basin. To address this, same hybrid approach as in chapter 4, a modification of TFN model forecasts with PRCP input using SWE and SOI, was performed using the artificial neural networks (ANN) method for the months of March to July.

The general formulation of TFN with modification approach was described as in Equation 4.1 in section 4.1.1. The only difference is that the monthly net inflow was used as the dependent variable instead of seasonal flow. The structures of the final monthly ANN models developed for March to July net inflow are shown in Table 5.10. The same data partitioning, training and cross validation procedure, and cross validation stopping criteria that were used in the previous sections were utilized in the ANN model building. The maximum number of inputs for ANN models in this procedure was limited to 4, so as to keep a much smaller network size and to enhance the generalization ability of the ANN with a smaller sample size.

Table 5.10 The monthly net inflow ANN models developed for forecast modification of Elephant Butte Reservoir, Rio Grande.

Months	Model structure	Model configuration
March	ANN (3-2-1)	$Y_{mar,\ modified} = f\ (SWE_{feb1st}, SWE_{mar1st}, Y_{mar,\ forecasted})$
April	ANN (3-2-1)	$Y_{apr,\ modified} = f\ (SWE_{mar1st}, SWE_{apr1st}, Y_{apr,\ forecasted})$
May	ANN (4-2-1)	$Y_{may,\ modified} = f\ (SWE_{apr1st}, SWE_{may1st}, SOI, Y_{may,\ forecasted})$
June	ANN (3-2-1)	$Y_{jun,\ modified} = f\ (SWE_{apr1st}, SWE_{may1st}, SOI, Y_{jun,\ forecasted})$
July	ANN (3-2-1)	$Y_{jul,\ modified} = f\ (SWE_{apr1st}, SWE_{may1st}, Y_{jul,\ forecasted})$

Notes: $Y_{t,\ modified}$ = *monthly net inflow forecasts after the forecast modification in month t*
$Y_{t,\ forecasted}$ = *monthly net inflow forecasts of TFN model with PRCP input in month t*

5.2.3 Model Diagnostics and Comparison

Wang (2006) discussed an important issue in evaluating forecast performance of seasonal models. The model performance for the whole year is better than most separate seasons when using coefficient of determination (R^2) as an evaluation index. Obtaining high coefficient of determination between observed and forecasted monthly net inflow for the whole testing period using a certain forecasting model does not necessarily mean it is a good model. Although the overall coefficient of determination may be high for the whole year, the model performance for some months may be very poor. Sometimes the observed mean might be a better forecast for some months. For example, Table 5.11 summarized the one-month-ahead forecast performance of TFN model for monthly net inflow for both the calibration and the testing periods. The R^2, RMSE and NRMSE are tabulated in Table 5.11 for each month, for March-July, and whole year, respectively. As can be seen, the coefficient of determination for the whole year and the March-July months were higher than the averages of individual months, which illogically indicates that model performance for the whole year is better than for most of the individual months (Wang, 2006). Hence, when a monthly model is developed, the performance of the model for each month should be evaluated, so that water managers can selectively use the model for each month of a year based on the performance of the model for that specific month.

No better results than in Table 5.11 can be expected for the forecasting of net inflow in the study site using ARIMA and TFN model with precipitation input. This is because the ARIMA model is basically built for the persistence of the runoff, while

TFN model includes only the precipitation information other than persistence. The inclusion of precipitation information may improve model performance for winter and fall months. In contrast, precipitation may not contribute significantly to forecast improvement for the spring-summer months, since the large proportion of spring-summer flow are contributed by snowmelt in the Basin. As shown in Table 4.5, the March to July reservoir net inflow is significantly correlated with SWE in the Basin. The net inflows for some months, such as May and June, have very high correlations coefficients with SWE in the Basin. Moreover, net inflows of these months also have a significant correlation with SOI information. Therefore, the forecasts of March to July could be improved by inputting information from these predictors to the forecasting models.

Table 5.11 One-month-ahead forecasting performance of TFN model for Elephant Butte Reservoir monthly net inflow

Month	**Calibration period (1961-1999)**			**Forecasting period (2000-2007)**		
	R^2	RMSE (kaf)	NRMSE	R^2	RMSE (kaf)	NRMSE
JAN	0.23	14.2	0.93	0.55	4.5	0.29
FEB	0.59	18.5	0.70	0.64	7.5	0.28
MAR	0.68	19.0	0.62	0.70	8.0	0.26
APR	0.55	41.1	0.70	0.46	24.1	0.41
MAY	0.61	63.5	0.63	0.81	39.2	0.39
JUN	0.83	43.3	0.45	0.70	40.3	0.42
JUL	0.55	47.0	0.72	0.09	37.8	0.58
AUG	0.16	37.2	0.90	0.35	48.0	1.16
SEP	0.16	22.8	1.00	0.00	19.9	0.87
OCT	0.30	23.3	0.90	0.07	23.3	0.90
NOV	0.41	19.4	0.79	0.89	8.1	0.33
DEC	0.44	18.5	0.78	0.34	10.6	0.45
APR-SEP	0.71	45.1	0.57	0.65	32.4	0.41
YEAR	0.70	34.0	0.58	0.56	27.3	0.46

As far as the winter and fall net inflow are considered, the persistence and precipitation might be the only inputs that should be included in the net inflow forecasts in the study (as in ARIMA and TFN) since they are not significantly correlated with SWE and SOI. Moreover, from the practical point of view, about two thirds of the Elephant Butte Reservoir annual net inflow is concentrated in the March to July months (Figure 1.9). Hence, the forecasting of spring-summer net inflow with acceptable accuracy is a crucial issue in water management for irrigation, environmental and compact purposes in the Basin, particularly in the lower Rio Grande region. In the following sections, the March to July net inflow forecast performance is mainly discussed and analyzed.

Table 5.12 illustrates the different model performance for the one-month-ahead forecasting of net inflows from March through July for the period of 2000-2007. The results showed that the forecast performance improved significantly from simple the ARIMA model that was built based on the autocorrelation of monthly net inflow itself to the TFN model with precipitation, to ANN models that were calibrated for each month using previous SWE, PRCP, SOI, and net inflow as the inputs, and to the TFN model with forecast modification using SWE and SOI information. It can be seen that the TFN model with forecast modification performed better than any other modeling method. There was a significant improvement in forecast performance compare to the ARIMA model and TFN model with precipitation input.

Table 5.12 Monthly net inflow forecast performance of different models for the March to July months of 2000-2007

Models	R^2	MAE (kaf)	MAPE(%)	RMSE (kaf)	NRMSE	E
ARIMA	0.58	28.5	403	36.2	0.46	0.51
TFN	0.65	24.1	539	32.4	0.41	0.61
ANN	0.87	13.6	96	18.9	0.24	0.87
TFN with modification	0.89	11.6	45	17.0	0.21	0.89

It can also be seen from Table 5.12 that the overall forecast performance of ANN models calibrated for each month was not as good as the TFN with forecast modification. This may be due to the generalization capability and inclusion of information of the neural networks that were used in both methods. As in TFN with forecast modification, fewer inputs were used than ANN models calibrated for each month, which enables the network to have a smaller size and more generalization capability. In addition, the forecasts from TFN with precipitation input have already included persistence, precipitation and stochastic components in one input variable for the ANN models that were used for the modification.

Figure 5.9 illustrates the forecast accuracy improvement of TFN with the forecast modification method compared to other modeling approaches by plotting a scatter plot of forecasted and observed March to July monthly net inflow for the period of 2000-2007. As can be seen from the figure, the forecasted net inflow that deviated largely from the observed net inflow using ARIMA and TFN models were successfully smoothed and lessened using ANN modeling and forecast modification approach for March to July net inflows. The correlation coefficients between observed and forecasted net inflows using TFN models and TFN with modification were 0.81 and 0.95 respectively, indicating that considerable improvement in forecast accuracy was obtained through using the forecast modification approach. This also suggested that the forecast modification with SWE and SOI information for the March to July net inflow using ANN method is practically effective in net inflow forecasting of Elephant Butte Reservoir.

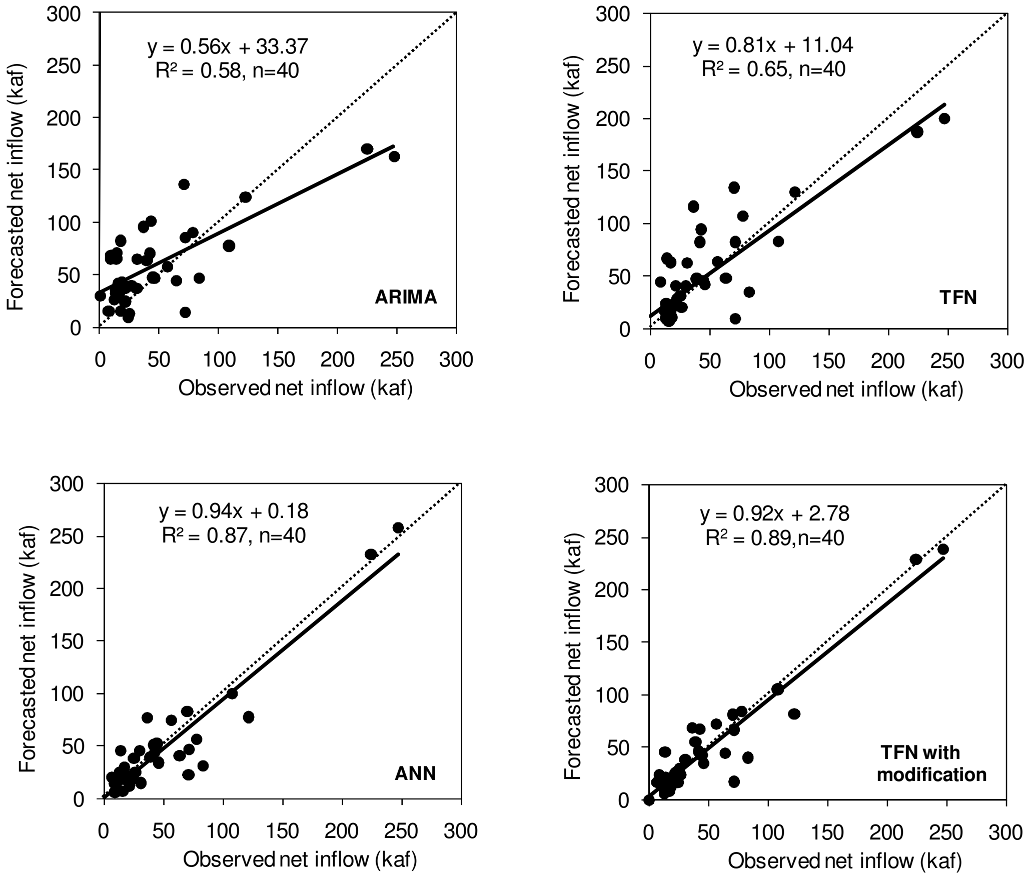

Figure 5.9 Scatter plots of observed and forecasted monthly Elephant Butte Reservoir net inflow using ARIMA, TFN, ANN and TFN with modification for the March to July months of 2000-2007

As mentioned earlier, the well-performed model which is evaluated for whole year and/or season by model performance evaluation indices, such as coefficient of determination for March to July period, does not mean it performs well for each individual month. As shown in Table 5.12, the TFN with forecast combination approach has coefficient of determination 0.89, model efficiency of 0.89, and normalized root mean squared error of 0.21, which indicates that a good one-month-ahead performance, and is a satisfactory model. However, the results vary in the model performance for each month from March to July. Table 5.13 and Figure 5.10 illustrate how each modeling method performed for each month from March to July. No forecast improvement was obtained for March even after using forecast combination and ANN modeling technique using SWE information as the inputs. Rather, the TFN model with precipitation input was a best-performing model for the March net inflow of the Elephant Butte Reservoir with coefficient of determination of 0.7 and RMSE of 0.26. This may be because of the heavy regulation of March net inflow. The regulation of the net inflow may weaken the natural relationship existing between March net inflow and the basin SWE index. For July, none of the models were acceptable in terms of the coefficient of determination although July had relatively smaller normalized RMSEs.

Table 5.13 Forecast performance of modeling methods for different months from March to July for Elephant Butte Reservoir net inflow (2000-2007)

Models	ARIMA		TFN		ANN		TFN with modification	
Month	R^2	NRMSE	R^2	NRMSE	R^2	NRMSE	R^2	NRMSE
March	0.58	0.38	0.70	0.26	0.38	0.43	0.56	0.33
April	0.41	0.46	0.46	0.41	0.73	0.38	0.75	0.29
May	0.80	0.56	0.81	0.39	0.91	0.22	0.93	0.20
June	0.85	0.41	0.70	0.42	0.95	0.18	0.97	0.13
July	0.06	0.47	0.09	0.58	0.14	0.28	0.09	0.34

Figure 5.10 and Table 5.13 indicated that there was substantial improvement in forecasting performance by ANN and TFN with forecast modification for the months of April, May and June. When comparing the forecast accuracies of different months, the May and June net inflows were forecasted with high accuracy, while the April net inflows forecasts were fairly acceptable. For example, the forecast RMSE of TFN with modification for June is 13015 acre-ft, is about one third of the highest forecasting RMSE (39660 acre-ft) of simple ARIMA model. Hence, in order to improve the one-month-ahead forecast accuracy, the different models may be used for the different months of a year for net inflow forecasting at Elephant Butte Reservoir, Rio Grande.

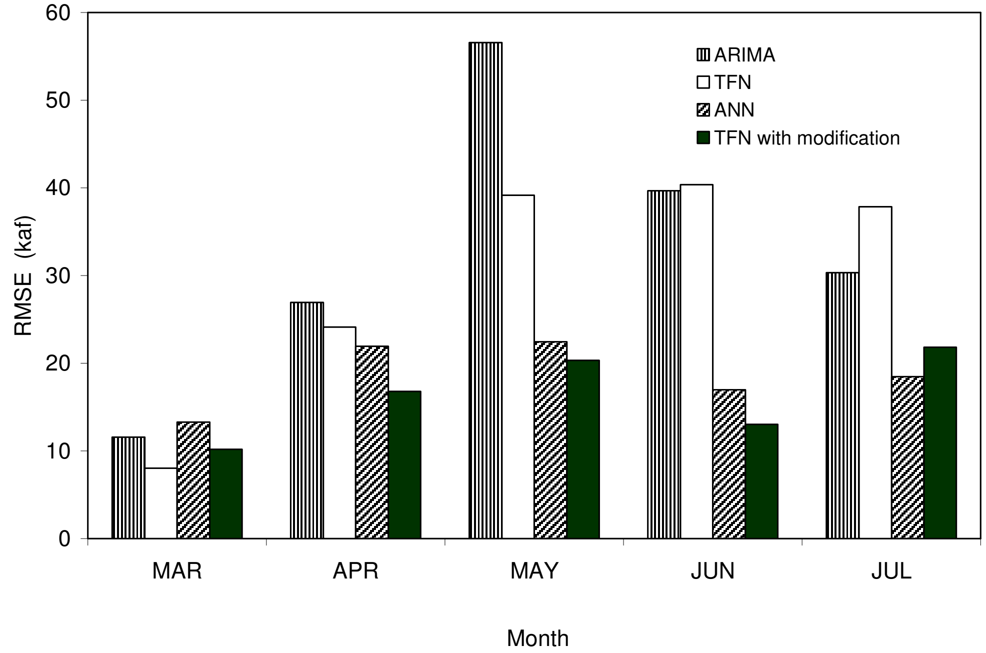

Figure 5.10 Comparison of March to July forecast RMSEs of different modeling methods for Elephant Butte Reservoir monthly net inflow (2000-2007)

5.2.4 Final Models

According to model performance analysis in the previous sections, the TFN model with modification using basin SWE and SOI can be recommended for the months of April, May, June and July of the year for one-month-ahead net inflow forecasting. For fall and winter months such as November, December, January, February and March, the TFN model with SNOTEL precipitation index as input could be used for one-month-ahead net inflow forecasting. For the other months of the year such as August, September and October, the TFN model with precipitation input may be used to provide monthly forecasts, but is not recommended as a basis for reservoir operation because of limited forecast accuracy. The final TFN model that could be used for one-month-ahead net inflow forecasting at Elephant Butter Reservoir has been calibrated using data period of 1961-2007 and is given as follows:

$$y_t = \frac{(0.148 + 0.081B^3)}{(1-0.853B)} PRCP_{t-1} + \frac{1}{(1-0.447B)} a_t \qquad (5.11)$$

or

$$\begin{aligned} y_t &= 0.447\, y_{t-1} + 0.853\, \hat{y}_{t-1} - 0.381\, y_{t-2} + 0.148 PRCP_{t-1} - 0.066 PRCP_{t-2} \\ &+ 0.081 PRCP_{t-4} - 0.036 PRCP_{t-5} + a_t \end{aligned} \qquad (5.12)$$

The final models that can be used for each month of the year for forecasting of Elephant Butte net inflow are proposed in Table 5.14. The recommendation to use different modeling approaches for individual months is also given based on the model forecast performance. It was suggested that the methodologies presented in the study may not be applicable for the forecasting of July to October net inflow due to high forecast error and uncertainty. However, the suggestions given here were based on the performance of models that were calibrated and tested using the very short data period of 47 years. Some other limitations such as the period of calibration, data preparation, selection of input variables, and selected study site for the models may also affect the conclusions.

Table 5.14 Final models that could be used for Elephant Butte Reservoir monthly net inflow forecasting

Month	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
Deseasonalization equation	$y_t = \dfrac{Y_t - \overline{Y}_t}{\hat{\sigma}_t}$											
One-month-ahead forecast equation	$\hat{y}_t = 0.447\, y_{t-1} + 0.853\hat{y}_{t-1} - 0.381\, y_{t-2} + 0.148 PRCP_{t-1} - 0.066 PRCP_{t-2} + 0.081 PRCP_{t-4} - 0.036 PRCP_{t-5}$											
Forecast modification using SWE and SOI	NO	NO	YES	YES	YES	YES	YES	NO	NO	NO	NO	NO
Improvement by modification	-	-	NO	YES	YES	YES	YES	-	-	-	-	-
Backtransforming	$Y_t = \overline{Y}_t + \hat{\sigma}_t y_t$											

Notes: y_t = *deseasonalized net inflow series for month t*

Y_t = *original net inflow series for month t*

$\overline{Y}_t$ = *sample average of the original net inflow for month t*

$\hat{\sigma}_t$ = *sample standard deviation of the original net inflow for month t*

5.3 Summary

In this chapter, the application of autoregressive integrated moving average (ARIMA), transfer function-noise (TFN) modeling and artificial neural network (ANN) techniques in forecasting naturalized streamflow and reservoir net inflow using various input variables has been illustrated and their one-month ahead forecasting performance has been evaluated by using several performance indices, such as coefficient of determination (R^2), root mean squared error (RMSE) and normalized RMSE. The results from this study indicate that the ANN is a useful tool in forecasting monthly streamflow whether it is used for direct modeling of monthly flow or used as a forecast modification technique when there are enough observed data the model development. The TFN modeling could be used for monthly streamflow forecasting in the winter and fall seasons of the year. Furthermore, the TFN modeling procedure could also be used to identify significant input variables for the building of ANN models.

The forecast modification using ANN technique with inputs such as forecast results of TFN with precipitation input, basin snow water equivalent and SOI, was proved to be effective method through this study in one-month-ahead forecasting of Del Norte spring-summer flow and Elephant Butte spring-summer net inflow. This method has showed better generalization capability than the ANN models that specifically calibrated for each month of snowmelt season, and therefore, was able to improve net inflow forecast accuracy. It also implied that the combination of different modeling methods could be a powerful and favorable approach in improving forecast accuracy of one-month-ahead monthly reservoir net inflow.

The comparison of time series models, namely ARIMA and TFN with precipitation input which were calibrated for the whole year, with ANN models that either were directly used for modeling or were used for forecast modification of TFN model forecasts for the months of spring-summer season suggest that the ANN models perform better than time series models in general. This may be because the ANN modeling technique can incorporate various predictors easily into the model and showed more flexibility in modeling procedure compared to the TFN models. In addition, and more importantly, the ANN algorithm can capture a nonlinear relationship between inputs and outputs, while the TFN model is able to deal only with linearity. This is particularly useful in modeling regulated flows, such as reservoir net inflow, where input and output relationships are of a complex and nonlinear nature.

Usually, the model performance for the whole year is better than most separate seasons when using coefficient of determination (R^2) as an evaluation index. Hence, high coefficient of determination between observed and forecasted monthly streamflow for the whole testing period does not necessarily mean it would be a good model. Although the overall coefficient of determination is high for the whole year, the model performance for some months may be very poor, and sometimes the observed mean might actually be a better forecast for some months. Therefore it was suggested that the performance of the model for each month should be evaluated when using a monthly forecasting model, so that water managers can selectively use the model for each month of a year based on the performance of the model for that specific month.

The results from this study may provide an impetus for monthly streamflow forecasting, particularly reservoir net inflow forecasting, by using a combined TFN and ANN approach with various operationally available climatic variables. However, both the TFN models and the ANN models developed in this study are local models that are limited to applications within the study area and with the observed specific conditions. For other study areas, new ANN architectures may be developed using different input variables and lag time relations for forecasting monthly flow and reservoir net inflow based on the specific characteristics of a basin. Further research may be directed to explore the application of more effective modeling methodologies aimed at improving monthly reservoir net inflow forecasting for reservoir operations.

6 SUMMARY AND CONCLUSIONS

Streamflow forecasting is challenging because of the complexity of the hydrologic system. Improving the quality of streamflow forecasting has always been an important task for researchers and hydrologic forecasters. In this dissertation, the improvement of seasonal and monthly streamflow forecasting using various data-driven statistical models was investigated for naturalized streamflow at Del Norte Gaging Station and observed reservoir net inflow of Elephant Butte Reservoir, Rio Grande. The application of partial least squares regression and hybrid models in seasonal streamflow forecasting were investigated with the purpose of improving the quality of seasonal streamflow forecasting. Some issues in monthly streamflow forecasting, such as monthly model performance evaluation, inclusion of snowpack and El Niño Southern Oscillation (ENSO) information in the monthly streamflow forecasting were discussed and analyzed.

Two approaches were presented in forecasting seasonal streamflow volume runoff. They were discussed in chapter 3 and chapter 4 respectively. In chapter 3, the spring-summer seasonal runoff volumes, the April-September streamflow volume at Del Norte Gaging Station and the March-July net inflow volume of Elephant Butte Reservoir, were used for modeling to be consistent with the current forecast practices of NRCS. The multivariate regression methods, such as partial least squares regression (PLSR) and principal components regression (PCR) were applied to develop spring-summer seasonal volume runoff forecasting equations in the selected basins and the forecast performance was compared to the NRCS official forecasts. Some issues related to regression equation development, such as the selection of the optimal number of components using the jackknife cross validation scheme, the variable selection procedures in PLSR, and forecast performance evaluation were analyzed and discussed. In chapter 4, to apply time series models in seasonal streamflow forecasting, four different seasons were defined for each study basin based on the characteristics of streamflow processes of the particular basin. The ARIMA and TFN models were developed for the defined seasonal flow time series. The ANN models and two hybrid modeling approaches, a forecast modification using a combination of transfer function-noise (TFN) with artificial neural networks (ANN) and the combination of principal components analysis (PCA) with ANN, were applied to the snowmelt runoff seasons in both basins to investigate the possible forecast improvement by using these approaches. The one-season-ahead forecasts of proposed models were compared to the same lead time NRCS official forecasts.

Monthly streamflow forecasting was discussed in chapter 5. The application of ARIMA, TFN modeling and ANN techniques in forecasting naturalized streamflow and reservoir net inflow using various input variables has been illustrated. The one-month-ahead forecasting performances of the models have been evaluated by using several performance indices. The inclusion of snowpack and ENSO information in the monthly streamflow forecasting models were carried out using a cross correlation analysis and forecast modification using ANN techniques. Some issues related to the performance of monthly forecasting models were also addressed to make better use of monthly models in the operational streamflow forecasting environment.

6.1 Seasonal Model Capabilities and Limitations

6.1.1 PLSR and PCR Forecasting Equations

The developed forecasting equations for the April-September natural flow at Del Norte Gaging Station and the March-July net inflow of Elephant Butter Reservoir using partial least squares regression showed potential ability in modeling procedure and improving forecast accuracy. Particularly, the PLSR equations using the composite precipitation index as input resulted in better performance with parsimonious structure. The comparative results of the model forecasts with NRCS official forecasts were also encouraging. Moreover, this approach is easily applicable in operational seasonal streamflow forecasting environment. The final forecasting equations proposed for the April-September natural flow forecasting at the Del Norte Gaging Station and the March-July Elephant Butte Reservoir net inflow forecasting can be applied in operational seasonal streamflow forecasting within the study basins.

The composite precipitation index was first introduced in this study to examine if better forecast skills could be obtained. The results indicate that both PLSR and PCR equations using composite index as inputs provided better or equivalent forecasts as compared to those using monthly observed precipitation as the direct inputs. In addition, this approach is more robust because of the parsimonious feature of models that reduces the number of input variables without loss of accuracy.

This is particularly important for larger basins where more information is available from numerous SNOTEL sites for forecast equation development. The algorithm used in this study was limited to using the weighted average of monthly precipitation based on the correlation coefficients with corresponding spring-summer forecast target volume. However, this may be not the best way to drive the composite precipitation indices. In practice, there may be other ways to calculate different composite indices to develop better models.

The comparison of the performance of PLSR and PCR in forecasting both natural flow and measured reservoir net inflow suggest that there are no significant differences in the forecasting performance of the two methodologies. However, it was observed that the PLSR can reach its minimal prediction error with a smaller number of components than PCR. This is a unique feature of PLSR compared to PCR when developing regression equations. Moreover, the explained variation (coefficient of determination for calibration) of the dependent variable by PLSR is always higher than PCR for the same number of components extracted. In general, the PLSR is more powerful than PCR in extracting components that deal with collinearity issue, yet it may not necessarily guarantee that PLSR would be better than PCR in terms of forecasting accuracy in seasonal streamflow forecasting.

The proposed forecasting equations in the study using PLSR and PCR were calibrated using only 22 years of data. This calibration period in this study is shorter than that of the NRCS forecasting equations because the data used in the calibration were of continuous high quality data measured from NRCS automatic SNOTEL sites.

Except for several years of SNOTEL precipitation data that were extended back to the 1980s using weather station precipitation data, neither an estimation of missing data was performed, nor was the Snow Course data used in the calibration of the regression equations. The standoff between using a shorter calibration period and using real-time good quality data is often encountered in seasonal streamflow forecasting equation development. However, with the accumulation of real-time measured SNOTEL data through time, the regression equations can be recalibrated every year when the new data become available. The final forecasting equations proposed in the study were calibrated using 27 years of data, which are all the data available up to the present.

The PLSR and PCR regression equations developed in this study were used to compute the median value of the seasonal water volume forecast distribution. If needed, the ensembles/probabilistic forecasts can be added by analyzing the statistical properties of the model error series (i.e., residuals) that occur in reproducing observed historical streamflow data using a jackknife procedure. The results from statistical tests of residuals of all PLSR models developed in the study showed that they are normally distributed at 0.05 significance level. Based on the normally- distributed errors, the exceedance probability forecasts of PLSR equation can be provided. The width of the probabilistic forecast error bound is proportional to the root mean squared error between these jackknife hindcasts and their respective observations. In general, the application of PLSR in seasonal streamflow forecasting is promising. Together with PCA and Z-score regression, the PLSR approach can be combined into

NRCS's operational forecasting platform to facilitate its application in operational forecasting environment.

6.1.2 Hybrid Modeling Approaches

The application of hybrid modeling approaches in the two study basins showed their potential capability to improve forecast accuracy in seasonal streamflow modeling as compared to single models. For the modeling of seasonal natural streamflow at the Del Norte Gaging Station, the combination of PCA and ANN performed better for April-June and April-September streamflow volume forecasting. The performance of this approach was comparable to NRCS official forecasts. For the modeling of Elephant Butte Reservoir seasonal net inflow, the best-performing model for April-July seasonal net inflow forecasting was the forecast modification of TFN with ANN approach. However, no distinct difference could be observed when comparing the performance of both approaches for either basin. Both approaches performed reasonably well compared to the single models. The number of principal components used as inputs in the ANN models is crucial and may affect the performance of the PCA+ANN modeling approach. Overall, the TFN with forecast modification approach was preferred in this study due to smaller network size and the inclusion of more information using transfer function-noise models.

It was observed that none of the models presented in this study performed with reasonable accuracy in the forecasting of late summer and early fall streamflows. In particular, the July-September flow at the Del Norte Gaging Station and August-

October Elephant Butte Reservoir net inflow. The normalized root mean squared errors of the forecasted and observed streamflow for these seasons was close to 1.0 or higher, indicating that there was no significant difference compared to using the historical average as the forecast. This may be due to the high variability of streamflows in these seasons that are affected by low elevation rainfall in the basins. In addition, there is no significant relationship existing between snowpack information and streamflow processes during these seasons. This is particularly obvious for August-October net inflow of the Elephant Butte Reservoir. Hence, it would not be appropriate to use any of the models developed in this study to forecast streamflows in these seasons due to high forecast errors.

The hybrid modeling approaches were not applied for the late fall and winter streamflow because no meaningful relationship exists between snowpack and streamflow processes during these seasons. However, the streamflows of these seasons have been forecasted reasonably well using the TFN model with precipitation input for both study basins. These seasons are the January-March and October-December seasonal streamflows at Del Norte Gaging Station and the January-March and November-December seasonal net inflow of Elephant Butte Reservoir. Although the runoff for these seasons does not seem not as important as spring-summer runoff for water management in the basin due to their smaller contribution to the annual runoff, forecasting of streamflows for these seasons may provide an earlier indication of magnitude of spring-summer runoff volume. In addition, the streamflow volumes for these seasons can be forecasted using time series models with reasonable accuracy. This is partly due to the smaller interannual variation of the streamflow processes in these seasons. This suggests that the TFN model with SNOTEL precipitation as input is sufficient for forecasting seasonal streamflow volume for these seasons in the study basins.

It was observed in the Elephant Butte Reservoir net inflow modeling that the performance of the single ANN model and combination of PCA and ANN approach were also comparable to the TFN with forecast modification approach. However, the latter is preferable for use in April-July Elephant Butte Reservoir net inflow volume forecasting due to its simpler structure and higher forecast accuracy. The converted forecasts made by routing equation using NRCS official forecasts at San Marcial Gaging Station was not preferred because of the higher forecast errors compared to other modeling approaches. This may be because the relationship between the input variables and the Elephant Butte Reservoir seasonal net inflow is not linear due to human intervention and complex features of the net inflow that are affected by many factors such as reservoir evaporation, seepage and the contribution of low-altitude rainfall to the net inflow. Most of the operational forecasts issued by NRCS are based on linear regression equations.

Hybrid modeling proved to be a robust approach in seasonal streamflow forecasting. The TFN with forecast modification approach using one-season-ahead TFN forecasts shows significant improvement in forecast accuracy. However, this is applicable for one-season-ahead seasonal volume forecasting only. To obtain longer lead time forecasting, the two-season-ahead TFN forecast may be used. Yet, the two-season-ahead forecasting accumulates forecasting errors which inevitably affects the performance of the hybrid approach. It was observed in this study that the two-season-ahead forecasts were not acceptable due to high forecast errors. This limitation results in difficulties in the application of the TFN+ANN approach in longer lead time forecasting.

The definition of hydrological seasons for both study basins is somewhat subjective, requiring personal judgment of the author based on understanding of the specific basin being studied. This is an important step in the application of time series models and hybrid approaches in the seasonal streamflow modeling and may also affect the quality of the modeling and forecasting. More studies are needed on this issue. Further, due to the time limitation of this dissertation study, the forecast uncertainty evaluation of these hybrid modeling approaches was not presented. It is hoped that future hydrological forecasting research efforts will exploit the potential capabilities of hybrid modeling in achieving increased forecast accuracy and perform forecasting uncertainty analysis of hybrid models.

6.2 Monthly Model Capabilities and Limitations

Several statistical methods including ARIMA, TFN, ANN and the forecast modification with ANN were applied to monthly streamflow and net inflow modeling in the study basins. The results indicated that the ANN is a useful tool in forecasting monthly streamflow, whether it is used for direct modeling of monthly flow or used as a forecast modification technique when there are enough observed data for model development. The TFN modeling could be used for monthly streamflow forecasting in the winter and fall seasons of the year. Furthermore, the TFN modeling procedure could also be used to identify significant input variables for the building of ANN models.

The forecast modification approach using ANN technique with inputs such as forecast results of TFN with precipitation input, basin snow water equivalent and SOI, proved to be an effective method throughout this study in one-month-ahead forecasting of Del Norte spring-summer flow and Elephant Butte spring-summer net inflow. This method displayed better generalization capability than the ANN models that were specifically calibrated for each month of snowmelt season and was able to improve forecast accuracy significantly. These findings also implied that the combination of different modeling methods is an advantageous approach in improving forecast accuracy of one-month-ahead monthly streamflow and reservoir net inflow.

The comparison of time series models, namely ARIMA and TFN with precipitation input which were calibrated for the whole year, with ANN models that either were directly used for modeling or were used for forecast modification of TFN model forecasts for the months of spring-summer season, suggested that the ANN models perform better than time series models in general. This may be because the ANN modeling technique can incorporate various predictors easily into the model and showed more flexibility in modeling procedure compared to the TFN models. In addition, and more importantly, the ANN algorithm can capture a nonlinear relationship between inputs and outputs, while the TFN model is able to deal only with linearity. This is particularly useful in modeling regulated flows, such as reservoir net inflow, where input and output relationships are of a complex and nonlinear nature.

In general, the model performance for the whole year is better than for most separate seasons when using the coefficient of determination (R^2) as an evaluation index. Hence, high coefficient of determination between observed and forecasted monthly streamflow for the whole testing period does not mean it is necessarily a good model. Although the overall coefficient of determination is high for the entire year, the model performance for some months may be poor, and sometimes the observed mean might actually be a better forecast for some months. Therefore it was suggested that the performance of the model for each month should be evaluated when using a monthly forecasting model, so that water managers may selectively use the model for each month of a year based on the performance of the model for that specific month.

Similar to seasonal flow modeling, the data period used to calibrate the monthly ANN models and TFN with forecasts modification was short due to the limited availability of snow water equivalent data (starting from 1961). Although the procedures that can protect the network from overtraining were applied in the training of all ANN models in the study, the generalization of ANN is always an issue in model training. Moreover, it was observed from the study that the longer lead time such as two-month-ahead forecasts with proposed monthly models were not adequate in terms of forecast accuracy. Hence, this will limit the application of the monthly models for the early decision-making in the reservoir operation and water management.

The findings of this study may provide an impetus for monthly streamflow forecasting, particularly reservoir net inflow forecasting, by using a combined TFN and ANN approach with various operationally available climatic variables. However, both the TFN models and the ANN models developed in this study are local models that are limited to applications within the study area and within the observed specific conditions. For other study areas, new ANN architectures may be developed using different input variables and lag time relations for forecasting monthly flow and reservoir net inflow based on the specific characteristics of a basin.

6.3 Recommendations and Future Work

The application of partial least squares regression in seasonal streamflow forecasting is promising. The selection of numbers of components with PLSR and variable selection procedures in seasonal streamflow forecasting equation development have been attempted for the first time in this study. However, the variable selection in PLSR is always a challenging task due to the complexity of the hydrologic process. Similar to Garen's (1992) method of variable selection for PCR, the investigation and application of more robust variable selection approaches, such as systematic searching of optimal or near optimal variable combination in PLSR would be desirable in future seasonal streamflow forecasting research studies.

Hybrid modeling is a new and robust approach in streamflow forecasting. The possible categorization of hybrid modeling methods was discussed in the study through a literature review. Two hybrid approaches including the combination of time series and ANN, combination of PCA and ANN were introduced. Their applications in seasonal and monthly streamflow forecasting were investigated in this study. There are many other hybrid approaches that can be applicable in streamflow forecasting that need to be explored in future research. The study of the relationships between the application of the complex hybrid modeling methods and their operational capability in streamflow forecasting environment are potentially useful and are also an interesting research topic in the field of streamflow forecasting.

In order to ensure the operational capabilities of proposed models, only hydrologic variables that are measured in the SNOTEL sites were used as the predictor variables. The low elevation precipitation information was not included in the models for two reasons. First, weather station data from climate networks are not readily available on the first day of the month. Second, the monthly and seasonal streamflow process is highly correlated with low elevation precipitation in the same period or lag. The inclusion of low elevation precipitation in the monthly and seasonal streamflow forecasting models requires that predicted precipitation should be used as an input, which would introduce prediction errors of precipitation into streamflow forecasting, consequently degrading forecast accuracy. As observed in this study, the low elevation precipitation affects significantly the late summer and early fall streamflow processes in the study basins. Therefore, it would be worthwhile to investigate the methods of inclusion of low elevation precipitation in monthly and seasonal streamflow forecasting models to improve the forecasting accuracy of the late summer and early fall streamflow processes.

Due to snow water equivalent and precipitation data availability from SNOTEL sites, a relatively short period of data was used to calibrate the multivariate regression equations, monthly and seasonal ANN models. To obtain more reliable seasonal runoff volume forecasting equations and enhance the generalization ability of ANN models in monthly and seasonal streamflow forecasting, more appropriate procedures for data extension and data pre-processing are worth investigating. At the same time, it would be useful to investigate the application of more robust generalization methods such as Bayesian regularization and reducing the weights of networks by data pre-processing, to enhance the generalization capability of ANN models.

Another limitation of the present study is that PLSR, PCR, ARIMA, TFN, ANN models and hybrid approaches proposed in this study are of local models that are limited to applications within the study areas. Only two subbasins of the Rio Grande and two hydrologic variables with seasonal and monthly time scales were investigated in detail. To obtain general conclusions about seasonal and monthly operational streamflow forecasting, the application of the methodologies in more streamflow processes in different river basins should be investigated.

APPENDICES

APPENDIX A

SAS CODES FOR PLSR CALIBRATION

```
*STEP1-PLSR CALIBRATION;

data dat ;
input year x1 - x25 y ;
n=_n_;
cards;

1981    8.4     14.8    16.6    20.6    3.3     7.3     2.2     2.5     0.8     3.7     1.0
        1.1     2.1     -2.0    -4.2    17.2    14.7    14.9    -6.4    22.1 -5.511.6   -6.1
        11.5    -0.4    292.6
1982    22.3    21.6    33.5    45.6    3.7     22.7    4.0     4.9     3.5     6.6     2.0
        6.2     1.8     -0.6    -6.9    35.7    23.0    15.9    -8.5    13.8    -7.8    11.6
        -5.0    16.4    0.0     562.5
1983    15.7    16.1    42.0    38.5    6.0     19.5    3.1     3.6     3.0     6.2     1.4
        4.7     0.8     -4.4    -9.9    53.8    24.5    8.9     -8.5    23.6    -7.8    11.6
        -5.7    19.2    -2.7    559.8
1984    16.3    17.6    37.4    31.6    5.9     24.9    3.4     3.5     2.6     5.7     1.5
        3.8     1.7     -5.3    -8.7    28.4    17.4    15.8    -9.7    14.2    -8.7    13.8
        -6.3    18.0    0.0     659.3
1985    27.2    29.5    48.2    48.3    8.1     23.1    4.8     5.1     3.1     8.0     1.6
        5.4     -3.5    -5.7    4.3     41.2    31.6    17.4    26.7    17.1    -8.0    13.5
        -5.3    20.0    -0.3    873.8
1986    17.6    25.3    32.4    40.5    7.4     21.0    4.6     5.5     3.0     7.4     2.3
        4.6     0.7     -5.7    -8.0    48.0    30.4    20.8    -5.3    18.8    -6.7    16.8
        -3.0    28.3    -0.3    830.9
1987    15.0    28.6    34.5    37.9    12.8    21.3    3.9     5.6     3.6     6.0     3.4
        6.7     -0.2    -5.0    -9.0    49.3    39.1    23.0    -9.7    17.7    -7.5    17.0
        -6.8    25.6    -0.9    892.5
1988    10.6    12.3    21.8    23.6    2.7     16.1    2.0     3.2     2.9     4.8     1.6
        4.3     3.0     -5.5    -9.5    23.4    19.7    15.9    -10.8   13.9    -7.3    13.8
        -6.2    20.0    -0.5    326.4
1989    21.7    18.1    37.4    37.9    8.2     25.5    3.3     4.2     2.3     4.8     1.5
        4.3     3.7     -3.3    -8.2    23.1    16.6    14.2    -9.8    12.7    -6.8    13.3
        -0.5    29.2    1.5     385.2
1990    10.2    13.0    16.6    19.0    3.3     11.4    1.6     2.3     2.2     3.7     1.0
        4.1     1.0     -3.8    -8.3    22.6    12.7    8.2     -9.8    7.5     -9.5    10.4
        -4.0    15.6    -0.2    413.4
1991    19.0    19.8    32.6    37.1    5.1     23.8    3.8     3.4     3.0     5.5     1.4
        4.8     1.5     -2.0    -9.2    42.5    25.9    15.9    -9.2    13.8    -5.0    12.1
        -6.2    16.0    -0.4    511.3
1992    15.0    18.5    35.2    27.8    7.8     13.9    3.0     3.7     2.6     5.4     1.7
        5.1     2.2     -6.0    -8.8    19.6    18.7    13.1    -9.7    10.9    -7.3    10.5
        -4.2    17.0    -1.5    411.3
1993    26.6    24.2    46.2    41.6    7.6     30.3    3.0     3.9     3.4     5.9     1.6
        4.9     2.2     -8.7    -10.2   19.3    15.3    12.6    -8.7    12.6    -8.8    11.1
        -4.5    18.1    -1.2    592.9
1994    12.6    13.2    25.9    26.9    3.6     18.1    2.7     3.2     3.4     4.7     1.8
        4.5     -0.3    -7.5    -9.3    20.4    15.4    12.9    -9.2    11.0    -9.8    10.2
        -3.5    17.5    -0.6    411.3
1995    22.8    23.1    44.2    42.6    4.5     30.5    4.4     4.9     3.9     6.7     2.6
        7.3     0.2     -7.0    -7.8    30.5    22.2    18.1    -9.3    13.2    -4.5    13.3
        -4.2    24.9    -1.3    714.6
```

```
1996    9.8     12.2    18.7    20.2    0.6     18.8    1.6     1.9     2.0     2.6     0.6
        2.1     1.8     -2.7    -7.3    22.3    16.9    12.9    -9.2    11.0    -5.7    13.5
        -5.5    15.1    -0.4    267.6
1997    19.2    25.1    41.7    36.1    6.9     25.5    4.1     5.9     5.1     7.4     3.7
        7.7     0.7     -4.8    -9.5    22.4    19.7    15.3    -8.8    10.8    -8.5    10.7
        -2.7    28.4    0.3     753.2
1998    11.0    18.0    25.8    32.1    6.2     17.7    3.2     4.6     3.0     5.6     1.8
        5.8     0.8     -5.2    -9.7    75.2    32.1    21.4    -8.5    15.6    -9.2    14.7
        -5.7    24.4    -1.5    462.1
1999    7.9     18.1    21.1    26.2    0.8     18.2    4.0     4.8     4.2     6.7     3.1
        7.3     1.0     -4.2    -8.2    29.9    27.2    19.9    -7.7    14.4    -6.3    13.5
        -2.0    23.9    1.2     809.0
2000    8.0     13.0    17.8    17.5    5.2     17.5    1.5     1.6     1.2     1.7     0.3
        1.8     3.0     -1.2    -8.7    33.3    20.2    16.0    -7.5    12.1    -6.2    11.6
        -4.8    17.4    1.2     297.5
2001    14.1    24.6    32.6    31.4    9.0     18.4    3.7     5.2     3.5     5.9     3.1
        6.0     1.2     -8.8    -8.0    20.3    16.8    13.4    -10.0   11.0    -7.8    10.8
        -5.0    19.9    1.2     633.8
2002    4.0     5.8     9.5     10.0    0.0     8.4     1.6     1.3     2.2     2.8     0.6
        2.4     2.5     -3.3    -9.7    15.9    16.5    13.7    -8.7    12.3    -8.5    10.6
        -6.0    16.0    -0.3    97.3
2003    8.5     13.5    25.9    21.6    0.0     15.4    2.2     3.1     3.1     5.2     0.9
        5.1     0.2     -5.2    -9.8    13.1    9.6     8.7     -5.8    9.0     -8.7    7.5
        -5.0    11.1    -0.9    234.6
2004    11.4    18.6    27.5    30.8    0.2     18.1    3.2     3.8     2.9     5.0     1.5
        4.0     4.3     -5.7    -8.0    13.7    14.7    12.2    -9.3    11.4    -9.3    10.3
        -1.8    26.9    0.1     416.7
2005    21.2    25.4    53.2    48.3    8.8     28.0    4.0     5.6     4.2     7.8     3.1
        8.2     1.2     -3.8    -7.2    28.8    19.1    15.4    -6.0    16.4    -6.4    13.5
        -4.8    18.7    -0.8    666.2
2006    11.1    15.1    17.5    20.8    4.5     17.6    2.5     3.1     2.6     3.7     1.7
        4.5     3.7     -1.5    -6.3    44.1    16.1    11.9    -6.8    10.6    -5.8    9.3
        -3.7    13.0    0.2     411.6
2007    9.8     16.3    18.8    28.7    3.8     12.7    3.7     4.6     4.4     6.5     2.3
        6.8     2.2     -1.5    -6.3    89.1    23.1    13.5    -8.0    12.5    -4.8    10.7
        -1.3    30.5    -0.7    593.2
;
data dat1;
set dat;
if (n <= 22);
run;

data dat2;
set dat;
if (n > 22);
run;


********************************************************/
/ Set Parameters for Macros /
/********************************************************/;
%global xvars yvars predname resname xscrname yscrname
```

```
num_x num_y lv;
%let xvars= X1 X2 X3      X4         X5 X6   X7         X8         X9 X10 X11 X12X13 X17 X18 X22
X24;

%let yvars=y;
%let ypred=yhat1;
%let yres=yres1;
%let predname=yhat;
%let resname=res;
%let xscrname=xscr;
%let yscrname=yscr;
%let num_y=1;
%let num_x=25;


*/*********************************************************/
/ Fit the PLS model /
/********************************************************/;

proc pls data=dat1 method=pls cv=one cvtest(stat=press) outmodel=est1;
model &yvars = &xvars/solution;
output out=outpls p=yhat1 yresidual=yres1
xresidual=xres1-xres55 xscore=xscr yscore=yscr
stdy=stdy stdx=stdx h=h press=press t2=t2
xqres=xqres yqres=yqres;
run;
%let lv=1; *** Used 1 PLS components ***;

%plot_scr(outpls);

%plotxscr(outpls,max_lv=2);

%get_wts(est1,dsxwts=xwts);

/********************************************************/
*/  Plot the X-weights vs. the frequency on the same axes /*
/********************************************************/;
%pltwtfrq(xwts,plotyvar=w,plotxvar=n,max_lv=&lv,label=Weight);


%plot_wt(xwts,max_lv=2);

%getxload(est1,dsxload=xloads);

/********************************************************/
*/ Plot the X-loadings for each component vs. frequency /
/********************************************************/;
%pltwtfrq(xloads,plotyvar=p,plotxvar=n,max_lv=&lv,
label=Loading);

%pltxload(xloads,max_lv=2);
```

```
%get_bpls(est1,dsout=bpls);

%plt_bpls(bpls);


/*********************************************************/
*/ Get VIP and plot it across frequencies /
/*********************************************************/;
%get_vip(est1,dsvip=vip_data);
%plot_vip(vip_data);


%res_plot(outpls);

%nor_plot(outpls);

run;

data eval;
merge bpls vip_data;
run;

proc print data=eval;
run;


PROC DBLOAD DBMS=xls DATA=eval;
PATH='C:\Documents and Settings\Administrator\Desktop\RESULT\ssssss.XLS';
PUTNAME=yes;
LOAD;
RUN;
QUIT;
```

* Notes: Macros used in the PLSR calibration are available through following links:

25009 - Macros to plot statistics generated by PROC PLS
http://support.sas.com/kb/25/009.html

Examples Using the PLS Procedure
http://support.sas.com/rnd/app/papers/plsex.pdf

APPENDIX B

SAS CODES FOR PLSR JACKKNIFE CROSS VALIDATION

```
* STEP2-PLSR JACKKINFE CROSS VALIDATION;

data dat ;
input year x1 - x25 y ;
n=_n_;
cards;

1981    8.4     14.8    16.6    20.6    3.3     7.3     2.2     2.5     0.8     3.7     1.0
        1.1     2.1     -2.0    -4.2    17.2    14.7    14.9    -6.4    22.1 -5.511.6   -6.1
        11.5    -0.4    292.6
1982    22.3    21.6    33.5    45.6    3.7     22.7    4.0     4.9     3.5     6.6     2.0
        6.2     1.8     -0.6    -6.9    35.7    23.0    15.9    -8.5    13.8    -7.8    11.6
        -5.0    16.4    0.0     562.5
1983    15.7    16.1    42.0    38.5    6.0     19.5    3.1     3.6     3.0     6.2     1.4
        4.7     0.8     -4.4    -9.9    53.8    24.5    8.9     -8.5    23.6    -7.8    11.6
        -5.7    19.2    -2.7    559.8
1984    16.3    17.6    37.4    31.6    5.9     24.9    3.4     3.5     2.6     5.7     1.5
        3.8     1.7     -5.3    -8.7    28.4    17.4    15.8    -9.7    14.2    -8.7    13.8
        -6.3    18.0    0.0     659.3
1985    27.2    29.5    48.2    48.3    8.1     23.1    4.8     5.1     3.1     8.0     1.6
        5.4     -3.5    -5.7    4.3     41.2    31.6    17.4    26.7    17.1    -8.0    13.5
        -5.3    20.0    -0.3    873.8
1986    17.6    25.3    32.4    40.5    7.4     21.0    4.6     5.5     3.0     7.4     2.3
        4.6     0.7     -5.7    -8.0    48.0    30.4    20.8    -5.3    18.8    -6.7    16.8
        -3.0    28.3    -0.3    830.9
1987    15.0    28.6    34.5    37.9    12.8    21.3    3.9     5.6     3.6     6.0     3.4
        6.7     -0.2    -5.0    -9.0    49.3    39.1    23.0    -9.7    17.7    -7.5    17.0
        -6.8    25.6    -0.9    892.5
1988    10.6    12.3    21.8    23.6    2.7     16.1    2.0     3.2     2.9     4.8     1.6
        4.3     3.0     -5.5    -9.5    23.4    19.7    15.9    -10.8   13.9    -7.3    13.8
        -6.2    20.0    -0.5    326.4
1989    21.7    18.1    37.4    37.9    8.2     25.5    3.3     4.2     2.3     4.8     1.5
        4.3     3.7     -3.3    -8.2    23.1    16.6    14.2    -9.8    12.7    -6.8    13.3
        -0.5    29.2    1.5     385.2
1990    10.2    13.0    16.6    19.0    3.3     11.4    1.6     2.3     2.2     3.7     1.0
        4.1     1.0     -3.8    -8.3    22.6    12.7    8.2     -9.8    7.5     -9.5    10.4
        -4.0    15.6    -0.2    413.4
1991    19.0    19.8    32.6    37.1    5.1     23.8    3.8     3.4     3.0     5.5     1.4
        4.8     1.5     -2.0    -9.2    42.5    25.9    15.9    -9.2    13.8    -5.0    12.1
        -6.2    16.0    -0.4    511.3
1992    15.0    18.5    35.2    27.8    7.8     13.9    3.0     3.7     2.6     5.4     1.7
        5.1     2.2     -6.0    -8.8    19.6    18.7    13.1    -9.7    10.9    -7.3    10.5
        -4.2    17.0    -1.5    411.3
1993    26.6    24.2    46.2    41.6    7.6     30.3    3.0     3.9     3.4     5.9     1.6
        4.9     2.2     -8.7    -10.2   19.3    15.3    12.6    -8.7    12.6    -8.8    11.1
        -4.5    18.1    -1.2    592.9
1994    12.6    13.2    25.9    26.9    3.6     18.1    2.7     3.2     3.4     4.7     1.8
        4.5     -0.3    -7.5    -9.3    20.4    15.4    12.9    -9.2    11.0    -9.8    10.2
        -3.5    17.5    -0.6    411.3
1995    22.8    23.1    44.2    42.6    4.5     30.5    4.4     4.9     3.9     6.7     2.6
        7.3     0.2     -7.0    -7.8    30.5    22.2    18.1    -9.3    13.2    -4.5    13.3
        -4.2    24.9    -1.3    714.6
```

```
1996    9.8     12.2    18.7    20.2    0.6     18.8    1.6     1.9     2.0     2.6     0.6
        2.1     1.8     -2.7    -7.3    22.3    16.9    12.9    -9.2    11.0    -5.7    13.5
        -5.5    15.1    -0.4    267.6
1997    19.2    25.1    41.7    36.1    6.9     25.5    4.1     5.9     5.1     7.4     3.7
        7.7     0.7     -4.8    -9.5    22.4    19.7    15.3    -8.8    10.8    -8.5    10.7
        -2.7    28.4    0.3     753.2
1998    11.0    18.0    25.8    32.1    6.2     17.7    3.2     4.6     3.0     5.6     1.8
        5.8     0.8     -5.2    -9.7    75.2    32.1    21.4    -8.5    15.6    -9.2    14.7
        -5.7    24.4    -1.5    462.1
1999    7.9     18.1    21.1    26.2    0.8     18.2    4.0     4.8     4.2     6.7     3.1
        7.3     1.0     -4.2    -8.2    29.9    27.2    19.9    -7.7    14.4    -6.3    13.5
        -2.0    23.9    1.2     809.0
2000    8.0     13.0    17.8    17.5    5.2     17.5    1.5     1.6     1.2     1.7     0.3
        1.8     3.0     -1.2    -8.7    33.3    20.2    16.0    -7.5    12.1    -6.2    11.6
        -4.8    17.4    1.2     297.5
2001    14.1    24.6    32.6    31.4    9.0     18.4    3.7     5.2     3.5     5.9     3.1
        6.0     1.2     -8.8    -8.0    20.3    16.8    13.4    -10.0   11.0    -7.8    10.8
        -5.0    19.9    1.2     633.8
2002    4.0     5.8     9.5     10.0    0.0     8.4     1.6     1.3     2.2     2.8     0.6
        2.4     2.5     -3.3    -9.7    15.9    16.5    13.7    -8.7    12.3    -8.5    10.6
        -6.0    16.0    -0.3    97.3
2003    8.5     13.5    25.9    21.6    0.0     15.4    2.2     3.1     3.1     5.2     0.9
        5.1     0.2     -5.2    -9.8    13.1    9.6     8.7     -5.8    9.0     -8.7    7.5
        -5.0    11.1    -0.9    234.6
2004    11.4    18.6    27.5    30.8    0.2     18.1    3.2     3.8     2.9     5.0     1.5
        4.0     4.3     -5.7    -8.0    13.7    14.7    12.2    -9.3    11.4    -9.3    10.3
        -1.8    26.9    0.1     416.7
2005    21.2    25.4    53.2    48.3    8.8     28.0    4.0     5.6     4.2     7.8     3.1
        8.2     1.2     -3.8    -7.2    28.8    19.1    15.4    -6.0    16.4    -6.4    13.5
        -4.8    18.7    -0.8    666.2
2006    11.1    15.1    17.5    20.8    4.5     17.6    2.5     3.1     2.6     3.7     1.7
        4.5     3.7     -1.5    -6.3    44.1    16.1    11.9    -6.8    10.6    -5.8    9.3
        -3.7    13.0    0.2     411.6
2007    9.8     16.3    18.8    28.7    3.8     12.7    3.7     4.6     4.4     6.5     2.3
        6.8     2.2     -1.5    -6.3    89.1    23.1    13.5    -8.0    12.5    -4.8    10.7
        -1.3    30.5    -0.7    593.2
;
options ls =75 formdlim = '_';

data dat1;
set dat;
if (n <= 22);
run;

*  start macro ;

data all;
pred = .;

%macro rena(nn);

%do i = 1 %to &nn;
```

```
data dat2;
set dat1;
if _n_ = &i  then y = .;

proc pls data = dat2 nfac=1 method=pls;
model y = X1 X2 X3        X4        X5        X6        X7        X8        X9        X10       X11       X12
          X13 X17 X18  X22 X24 /;
output out = out2  p = pred ;

data out2 ;
set out2(keep = pred) ;
if _n_ = &i ;

data all;
set all out2 ;

%end;
;
dm 'clear log;';
dm 'clear output;';

%mend ;

*  end of macro ;

* call the macro once for each observation ;

options mprint mlogic;
%rena( 22);

data all;
set all;
if _n_ > 1 ;

* join the Y vector with the pred vector ;

data all_y ;
set dat1( keep = y ) ;

data allpred ;
merge all_y all ;

*  print the actual y vector and
jackknife prediction vector ;

proc print data = allpred ;
run;

*Export the final jackknife prediction to Excel;

PROC DBLOAD DBMS=xls DATA=ALLPRED;
PATH='C:\Documents and Settings\Administrator\Desktop\RESULT\pred.XLS';
```

```
PUTNAME=yes;
LOAD;
RUN;
QUIT;
```

APPENDIX C

SAS CODES FOR PLSR PREDICTION OF NEW DATA

```
* STEP3-1-PLSR PREDICTION FOR NEW DATA;

data dat ;
input yr x1 - x25 y ;
n=_n_;
cards;

1981    8.4     14.8    16.6    20.6    3.3     7.3     2.2     2.5     0.8     3.7     1.0
        1.1     2.1     -2.0    -4.2    17.2    14.7    14.9    -6.4    22.1 -5.511.6   -6.1
        11.5    -0.4    292.6
1982    22.3    21.6    33.5    45.6    3.7     22.7    4.0     4.9     3.5     6.6     2.0
        6.2     1.8     -0.6    -6.9    35.7    23.0    15.9    -8.5    13.8    -7.8    11.6
        -5.0    16.4    0.0     562.5
1983    15.7    16.1    42.0    38.5    6.0     19.5    3.1     3.6     3.0     6.2     1.4
        4.7     0.8     -4.4    -9.9    53.8    24.5    8.9     -8.5    23.6    -7.8    11.6
        -5.7    19.2    -2.7    559.8
1984    16.3    17.6    37.4    31.6    5.9     24.9    3.4     3.5     2.6     5.7     1.5
        3.8     1.7     -5.3    -8.7    28.4    17.4    15.8    -9.7    14.2    -8.7    13.8
        -6.3    18.0    0.0     659.3
1985    27.2    29.5    48.2    48.3    8.1     23.1    4.8     5.1     3.1     8.0     1.6
        5.4     -3.5    -5.7    4.3     41.2    31.6    17.4    26.7    17.1    -8.0    13.5
        -5.3    20.0    -0.3    873.8
1986    17.6    25.3    32.4    40.5    7.4     21.0    4.6     5.5     3.0     7.4     2.3
        4.6     0.7     -5.7    -8.0    48.0    30.4    20.8    -5.3    18.8    -6.7    16.8
        -3.0    28.3    -0.3    830.9
1987    15.0    28.6    34.5    37.9    12.8    21.3    3.9     5.6     3.6     6.0     3.4
        6.7     -0.2    -5.0    -9.0    49.3    39.1    23.0    -9.7    17.7    -7.5    17.0
        -6.8    25.6    -0.9    892.5
1988    10.6    12.3    21.8    23.6    2.7     16.1    2.0     3.2     2.9     4.8     1.6
        4.3     3.0     -5.5    -9.5    23.4    19.7    15.9    -10.8   13.9    -7.3    13.8
        -6.2    20.0    -0.5    326.4
1989    21.7    18.1    37.4    37.9    8.2     25.5    3.3     4.2     2.3     4.8     1.5
        4.3     3.7     -3.3    -8.2    23.1    16.6    14.2    -9.8    12.7    -6.8    13.3
        -0.5    29.2    1.5     385.2
1990    10.2    13.0    16.6    19.0    3.3     11.4    1.6     2.3     2.2     3.7     1.0
        4.1     1.0     -3.8    -8.3    22.6    12.7    8.2     -9.8    7.5     -9.5    10.4
        -4.0    15.6    -0.2    413.4
1991    19.0    19.8    32.6    37.1    5.1     23.8    3.8     3.4     3.0     5.5     1.4
        4.8     1.5     -2.0    -9.2    42.5    25.9    15.9    -9.2    13.8    -5.0    12.1
        -6.2    16.0    -0.4    511.3
1992    15.0    18.5    35.2    27.8    7.8     13.9    3.0     3.7     2.6     5.4     1.7
        5.1     2.2     -6.0    -8.8    19.6    18.7    13.1    -9.7    10.9    -7.3    10.5
        -4.2    17.0    -1.5    411.3
1993    26.6    24.2    46.2    41.6    7.6     30.3    3.0     3.9     3.4     5.9     1.6
        4.9     2.2     -8.7    -10.2   19.3    15.3    12.6    -8.7    12.6    -8.8    11.1
        -4.5    18.1    -1.2    592.9
1994    12.6    13.2    25.9    26.9    3.6     18.1    2.7     3.2     3.4     4.7     1.8
        4.5     -0.3    -7.5    -9.3    20.4    15.4    12.9    -9.2    11.0    -9.8    10.2
        -3.5    17.5    -0.6    411.3
1995    22.8    23.1    44.2    42.6    4.5     30.5    4.4     4.9     3.9     6.7     2.6
        7.3     0.2     -7.0    -7.8    30.5    22.2    18.1    -9.3    13.2    -4.5    13.3
        -4.2    24.9    -1.3    714.6
```

```
1996    9.8     12.2    18.7    20.2    0.6     18.8    1.6     1.9     2.0     2.6     0.6
        2.1     1.8     -2.7    -7.3    22.3    16.9    12.9    -9.2    11.0    -5.7    13.5
        -5.5    15.1    -0.4    267.6
1997    19.2    25.1    41.7    36.1    6.9     25.5    4.1     5.9     5.1     7.4     3.7
        7.7     0.7     -4.8    -9.5    22.4    19.7    15.3    -8.8    10.8    -8.5    10.7
        -2.7    28.4    0.3     753.2
1998    11.0    18.0    25.8    32.1    6.2     17.7    3.2     4.6     3.0     5.6     1.8
        5.8     0.8     -5.2    -9.7    75.2    32.1    21.4    -8.5    15.6    -9.2    14.7
        -5.7    24.4    -1.5    462.1
1999    7.9     18.1    21.1    26.2    0.8     18.2    4.0     4.8     4.2     6.7     3.1
        7.3     1.0     -4.2    -8.2    29.9    27.2    19.9    -7.7    14.4    -6.3    13.5
        -2.0    23.9    1.2     809.0
2000    8.0     13.0    17.8    17.5    5.2     17.5    1.5     1.6     1.2     1.7     0.3
        1.8     3.0     -1.2    -8.7    33.3    20.2    16.0    -7.5    12.1    -6.2    11.6
        -4.8    17.4    1.2     297.5
2001    14.1    24.6    32.6    31.4    9.0     18.4    3.7     5.2     3.5     5.9     3.1
        6.0     1.2     -8.8    -8.0    20.3    16.8    13.4    -10.0   11.0    -7.8    10.8
        -5.0    19.9    1.2     633.8
2002    4.0     5.8     9.5     10.0    0.0     8.4     1.6     1.3     2.2     2.8     0.6
        2.4     2.5     -3.3    -9.7    15.9    16.5    13.7    -8.7    12.3    -8.5    10.6
        -6.0    16.0    -0.3    97.3
2003    8.5     13.5    25.9    21.6    0.0     15.4    2.2     3.1     3.1     5.2     0.9
        5.1     0.2     -5.2    -9.8    13.1    9.6     8.7     -5.8    9.0     -8.7    7.5
        -5.0    11.1    -0.9    234.6
2004    11.4    18.6    27.5    30.8    0.2     18.1    3.2     3.8     2.9     5.0     1.5
        4.0     4.3     -5.7    -8.0    13.7    14.7    12.2    -9.3    11.4    -9.3    10.3
        -1.8    26.9    0.1     416.7
2005    21.2    25.4    53.2    48.3    8.8     28.0    4.0     5.6     4.2     7.8     3.1
        8.2     1.2     -3.8    -7.2    28.8    19.1    15.4    -6.0    16.4    -6.4    13.5
        -4.8    18.7    -0.8    666.2
2006    11.1    15.1    17.5    20.8    4.5     17.6    2.5     3.1     2.6     3.7     1.7
        4.5     3.7     -1.5    -6.3    44.1    16.1    11.9    -6.8    10.6    -5.8    9.3
        -3.7    13.0    0.2     411.6
2007    9.8     16.3    18.8    28.7    3.8     12.7    3.7     4.6     4.4     6.5     2.3
        6.8     2.2     -1.5    -6.3    89.1    23.1    13.5    -8.0    12.5    -4.8    10.7
        -1.3    30.5    -0.7    593.2
;

data dat1;
set dat;
if (n <= 22);
run;

data dat2;
set dat;
if (n > 22) ;
if n > 22 then y = .;
run;

data all;
set dat1 dat2;
```

```
proc pls data=all nfac=1 details method=PLS;
model y=X1      X2      X3      X4      X5      X6      X7      X8      X9 X10 X11      X12
        X13 X17 X18 X22 X24/solution;
output out = salam
     press= ssum
      predicted=yhat
                 yresidual=yres;
                 run;

proc print data=salam;
where n>22;
var y yhat yres ssum;
RUN;

PROC DBLOAD DBMS=xls DATA=salam;
PATH='C:\Documents and Settings\Administrator\Desktop\RESULT\PRED.XLS';
PUTNAME=yes;
LOAD;
RUN;
QUIT;
```

APPENDIX D

SAS CODES FOR PLSR
ROLLING-FORWARD PREDICTION OF NEW DATA

```
* STEP3-2-PLSR ROLLING-FORWARD PREDICTION FOR NEW DATA;

data dat ;
input yr x1 - x25 y ;
n=_n_;
cards;

1981    8.4     14.8    16.6    20.6    3.3     7.3     2.2     2.5     0.8     3.7     1.0
        1.1     2.1     -2.0    -4.2    17.2    14.7    14.9    -6.4    22.1 -5.511.6   -6.1
        11.5    -0.4    292.6
1982    22.3    21.6    33.5    45.6    3.7     22.7    4.0     4.9     3.5     6.6     2.0
        6.2     1.8     -0.6    -6.9    35.7    23.0    15.9    -8.5    13.8    -7.8    11.6
        -5.0    16.4    0.0     562.5
1983    15.7    16.1    42.0    38.5    6.0     19.5    3.1     3.6     3.0     6.2     1.4
        4.7     0.8     -4.4    -9.9    53.8    24.5    8.9     -8.5    23.6    -7.8    11.6
        -5.7    19.2    -2.7    559.8
1984    16.3    17.6    37.4    31.6    5.9     24.9    3.4     3.5     2.6     5.7     1.5
        3.8     1.7     -5.3    -8.7    28.4    17.4    15.8    -9.7    14.2    -8.7    13.8
        -6.3    18.0    0.0     659.3
1985    27.2    29.5    48.2    48.3    8.1     23.1    4.8     5.1     3.1     8.0     1.6
        5.4     -3.5    -5.7    4.3     41.2    31.6    17.4    26.7    17.1    -8.0    13.5
        -5.3    20.0    -0.3    873.8
1986    17.6    25.3    32.4    40.5    7.4     21.0    4.6     5.5     3.0     7.4     2.3
        4.6     0.7     -5.7    -8.0    48.0    30.4    20.8    -5.3    18.8    -6.7    16.8
        -3.0    28.3    -0.3    830.9
1987    15.0    28.6    34.5    37.9    12.8    21.3    3.9     5.6     3.6     6.0     3.4
        6.7     -0.2    -5.0    -9.0    49.3    39.1    23.0    -9.7    17.7    -7.5    17.0
        -6.8    25.6    -0.9    892.5
1988    10.6    12.3    21.8    23.6    2.7     16.1    2.0     3.2     2.9     4.8     1.6
        4.3     3.0     -5.5    -9.5    23.4    19.7    15.9    -10.8   13.9    -7.3    13.8
        -6.2    20.0    -0.5    326.4
1989    21.7    18.1    37.4    37.9    8.2     25.5    3.3     4.2     2.3     4.8     1.5
        4.3     3.7     -3.3    -8.2    23.1    16.6    14.2    -9.8    12.7    -6.8    13.3
        -0.5    29.2    1.5     385.2
1990    10.2    13.0    16.6    19.0    3.3     11.4    1.6     2.3     2.2     3.7     1.0
        4.1     1.0     -3.8    -8.3    22.6    12.7    8.2     -9.8    7.5     -9.5    10.4
        -4.0    15.6    -0.2    413.4
1991    19.0    19.8    32.6    37.1    5.1     23.8    3.8     3.4     3.0     5.5     1.4
        4.8     1.5     -2.0    -9.2    42.5    25.9    15.9    -9.2    13.8    -5.0    12.1
        -6.2    16.0    -0.4    511.3
1992    15.0    18.5    35.2    27.8    7.8     13.9    3.0     3.7     2.6     5.4     1.7
        5.1     2.2     -6.0    -8.8    19.6    18.7    13.1    -9.7    10.9    -7.3    10.5
        -4.2    17.0    -1.5    411.3
1993    26.6    24.2    46.2    41.6    7.6     30.3    3.0     3.9     3.4     5.9     1.6
        4.9     2.2     -8.7    -10.2   19.3    15.3    12.6    -8.7    12.6    -8.8    11.1
        -4.5    18.1    -1.2    592.9
1994    12.6    13.2    25.9    26.9    3.6     18.1    2.7     3.2     3.4     4.7     1.8
        4.5     -0.3    -7.5    -9.3    20.4    15.4    12.9    -9.2    11.0    -9.8    10.2
        -3.5    17.5    -0.6    411.3
1995    22.8    23.1    44.2    42.6    4.5     30.5    4.4     4.9     3.9     6.7     2.6
        7.3     0.2     -7.0    -7.8    30.5    22.2    18.1    -9.3    13.2    -4.5    13.3
        -4.2    24.9    -1.3    714.6
```

```
1996    9.8     12.2    18.7    20.2    0.6     18.8    1.6     1.9     2.0     2.6     0.6
        2.1     1.8     -2.7    -7.3    22.3    16.9    12.9    -9.2    11.0    -5.7    13.5
        -5.5    15.1    -0.4    267.6
1997    19.2    25.1    41.7    36.1    6.9     25.5    4.1     5.9     5.1     7.4     3.7
        7.7     0.7     -4.8    -9.5    22.4    19.7    15.3    -8.8    10.8    -8.5    10.7
        -2.7    28.4    0.3     753.2
1998    11.0    18.0    25.8    32.1    6.2     17.7    3.2     4.6     3.0     5.6     1.8
        5.8     0.8     -5.2    -9.7    75.2    32.1    21.4    -8.5    15.6    -9.2    14.7
        -5.7    24.4    -1.5    462.1
1999    7.9     18.1    21.1    26.2    0.8     18.2    4.0     4.8     4.2     6.7     3.1
        7.3     1.0     -4.2    -8.2    29.9    27.2    19.9    -7.7    14.4    -6.3    13.5
        -2.0    23.9    1.2     809.0
2000    8.0     13.0    17.8    17.5    5.2     17.5    1.5     1.6     1.2     1.7     0.3
        1.8     3.0     -1.2    -8.7    33.3    20.2    16.0    -7.5    12.1    -6.2    11.6
        -4.8    17.4    1.2     297.5
2001    14.1    24.6    32.6    31.4    9.0     18.4    3.7     5.2     3.5     5.9     3.1
        6.0     1.2     -8.8    -8.0    20.3    16.8    13.4    -10.0   11.0    -7.8    10.8
        -5.0    19.9    1.2     633.8
2002    4.0     5.8     9.5     10.0    0.0     8.4     1.6     1.3     2.2     2.8     0.6
        2.4     2.5     -3.3    -9.7    15.9    16.5    13.7    -8.7    12.3    -8.5    10.6
        -6.0    16.0    -0.3    97.3
2003    8.5     13.5    25.9    21.6    0.0     15.4    2.2     3.1     3.1     5.2     0.9
        5.1     0.2     -5.2    -9.8    13.1    9.6     8.7     -5.8    9.0     -8.7    7.5
        -5.0    11.1    -0.9    234.6
2004    11.4    18.6    27.5    30.8    0.2     18.1    3.2     3.8     2.9     5.0     1.5
        4.0     4.3     -5.7    -8.0    13.7    14.7    12.2    -9.3    11.4    -9.3    10.3
        -1.8    26.9    0.1     416.7
2005    21.2    25.4    53.2    48.3    8.8     28.0    4.0     5.6     4.2     7.8     3.1
        8.2     1.2     -3.8    -7.2    28.8    19.1    15.4    -6.0    16.4    -6.4    13.5
        -4.8    18.7    -0.8    666.2
2006    11.1    15.1    17.5    20.8    4.5     17.6    2.5     3.1     2.6     3.7     1.7
        4.5     3.7     -1.5    -6.3    44.1    16.1    11.9    -6.8    10.6    -5.8    9.3
        -3.7    13.0    0.2     411.6
2007    9.8     16.3    18.8    28.7    3.8     12.7    3.7     4.6     4.4     6.5     2.3
        6.8     2.2     -1.5    -6.3    89.1    23.1    13.5    -8.0    12.5    -4.8    10.7
        -1.3    30.5    -0.7    593.2
;

data dat1;
set dat;
if (n <= 22);
run;

data dat2;
set dat;
if (n > 22);
   run;

data all;
pred = .;

*start macro;
```

```
%macro combine( n );
%do  i = 1 %to &n;
data dat3;
set dat2;
if _n_<=&i ;
if _n_ = &i  then y = .;

proc print data=dat3;

data dat4;
set dat1 dat3;

proc print data=dat4;


proc pls data= dat4  nfac=1 details method=pls ;
model y=X1       X2       X3       X4       X5       X6       X7       X8       X9       X10      X11
             X12      X13 X17 X18  X22         X24;
output out = pred1   p=pred;


data dat5;
  set pred1(keep=pred);
             if _n_= &i+22;

  proc print data=dat5;
  title1 ' print of forecasts and resuduals ' ;

  data all;
   set all dat5;
proc print data=all;

%end;
dm 'clear log;';
dm 'clear output;';

%mend;
*end of macro ;

* call the macro n times ;

options mprint mlogic;
%combine(5);
run;

data all;
set all;
if _n_ > 1 ;

proc print data=all;
run;
```

```
data dat2 ;
set dat2( keep = y ) ;

proc print data=dat2;
run;

data allpred ;
merge dat2 all ;

*  print the actual y vector and
jack knife prediction vector ;

proc print data = allpred ;
run;

PROC DBLOAD DBMS=xls DATA=ALLPRED;
PATH='C:\Documents and Settings\Administrator\Desktop\RESULT\ROLLPRED.XLS';
PUTNAME=yes;
LOAD;
RUN;
QUIT;
```

APPENDIX E

SAS CODES FOR STEPWISE VARIABLE SELECTION ON PRINCIPAL COMPONENTS

```
* PRINCIPCAL COMPONENTS STEPWISE VARIABLE SELECTION;

data dat ;
input year x1 - x25 y ;
n=_n_;
cards;

1981    8.4     14.8    16.6    20.6    3.3     7.3     2.2     2.5     0.8     3.7     1.0
        1.1     2.1     -2.0    -4.2    17.2    14.7    14.9    -6.4    22.1 -5.511.6   -6.1
        11.5    -0.4    292.6
1982    22.3    21.6    33.5    45.6    3.7     22.7    4.0     4.9     3.5     6.6     2.0
        6.2     1.8     -0.6    -6.9    35.7    23.0    15.9    -8.5    13.8    -7.8    11.6
        -5.0    16.4    0.0     562.5
1983    15.7    16.1    42.0    38.5    6.0     19.5    3.1     3.6     3.0     6.2     1.4
        4.7     0.8     -4.4    -9.9    53.8    24.5    8.9     -8.5    23.6    -7.8    11.6
        -5.7    19.2    -2.7    559.8
1984    16.3    17.6    37.4    31.6    5.9     24.9    3.4     3.5     2.6     5.7     1.5
        3.8     1.7     -5.3    -8.7    28.4    17.4    15.8    -9.7    14.2    -8.7    13.8
        -6.3    18.0    0.0     659.3
1985    27.2    29.5    48.2    48.3    8.1     23.1    4.8     5.1     3.1     8.0     1.6
        5.4     -3.5    -5.7    4.3     41.2    31.6    17.4    26.7    17.1    -8.0    13.5
        -5.3    20.0    -0.3    873.8
1986    17.6    25.3    32.4    40.5    7.4     21.0    4.6     5.5     3.0     7.4     2.3
        4.6     0.7     -5.7    -8.0    48.0    30.4    20.8    -5.3    18.8    -6.7    16.8
        -3.0    28.3    -0.3    830.9
1987    15.0    28.6    34.5    37.9    12.8    21.3    3.9     5.6     3.6     6.0     3.4
        6.7     -0.2    -5.0    -9.0    49.3    39.1    23.0    -9.7    17.7    -7.5    17.0
        -6.8    25.6    -0.9    892.5
1988    10.6    12.3    21.8    23.6    2.7     16.1    2.0     3.2     2.9     4.8     1.6
        4.3     3.0     -5.5    -9.5    23.4    19.7    15.9    -10.8   13.9    -7.3    13.8
        -6.2    20.0    -0.5    326.4
1989    21.7    18.1    37.4    37.9    8.2     25.5    3.3     4.2     2.3     4.8     1.5
        4.3     3.7     -3.3    -8.2    23.1    16.6    14.2    -9.8    12.7    -6.8    13.3
        -0.5    29.2    1.5     385.2
1990    10.2    13.0    16.6    19.0    3.3     11.4    1.6     2.3     2.2     3.7     1.0
        4.1     1.0     -3.8    -8.3    22.6    12.7    8.2     -9.8    7.5     -9.5    10.4
        -4.0    15.6    -0.2    413.4
1991    19.0    19.8    32.6    37.1    5.1     23.8    3.8     3.4     3.0     5.5     1.4
        4.8     1.5     -2.0    -9.2    42.5    25.9    15.9    -9.2    13.8    -5.0    12.1
        -6.2    16.0    -0.4    511.3
1992    15.0    18.5    35.2    27.8    7.8     13.9    3.0     3.7     2.6     5.4     1.7
        5.1     2.2     -6.0    -8.8    19.6    18.7    13.1    -9.7    10.9    -7.3    10.5
        -4.2    17.0    -1.5    411.3
1993    26.6    24.2    46.2    41.6    7.6     30.3    3.0     3.9     3.4     5.9     1.6
        4.9     2.2     -8.7    -10.2   19.3    15.3    12.6    -8.7    12.6    -8.8    11.1
        -4.5    18.1    -1.2    592.9
1994    12.6    13.2    25.9    26.9    3.6     18.1    2.7     3.2     3.4     4.7     1.8
        4.5     -0.3    -7.5    -9.3    20.4    15.4    12.9    -9.2    11.0    -9.8    10.2
        -3.5    17.5    -0.6    411.3
1995    22.8    23.1    44.2    42.6    4.5     30.5    4.4     4.9     3.9     6.7     2.6
        7.3     0.2     -7.0    -7.8    30.5    22.2    18.1    -9.3    13.2    -4.5    13.3
        -4.2    24.9    -1.3    714.6
```

```
1996    9.8     12.2    18.7    20.2    0.6     18.8    1.6     1.9     2.0     2.6     0.6
        2.1     1.8     -2.7    -7.3    22.3    16.9    12.9    -9.2    11.0    -5.7    13.5
        -5.5    15.1    -0.4    267.6
1997    19.2    25.1    41.7    36.1    6.9     25.5    4.1     5.9     5.1     7.4     3.7
        7.7     0.7     -4.8    -9.5    22.4    19.7    15.3    -8.8    10.8    -8.5    10.7
        -2.7    28.4    0.3     753.2
1998    11.0    18.0    25.8    32.1    6.2     17.7    3.2     4.6     3.0     5.6     1.8
        5.8     0.8     -5.2    -9.7    75.2    32.1    21.4    -8.5    15.6    -9.2    14.7
        -5.7    24.4    -1.5    462.1
1999    7.9     18.1    21.1    26.2    0.8     18.2    4.0     4.8     4.2     6.7     3.1
        7.3     1.0     -4.2    -8.2    29.9    27.2    19.9    -7.7    14.4    -6.3    13.5
        -2.0    23.9    1.2     809.0
2000    8.0     13.0    17.8    17.5    5.2     17.5    1.5     1.6     1.2     1.7     0.3
        1.8     3.0     -1.2    -8.7    33.3    20.2    16.0    -7.5    12.1    -6.2    11.6
        -4.8    17.4    1.2     297.5
2001    14.1    24.6    32.6    31.4    9.0     18.4    3.7     5.2     3.5     5.9     3.1
        6.0     1.2     -8.8    -8.0    20.3    16.8    13.4    -10.0   11.0    -7.8    10.8
        -5.0    19.9    1.2     633.8
2002    4.0     5.8     9.5     10.0    0.0     8.4     1.6     1.3     2.2     2.8     0.6
        2.4     2.5     -3.3    -9.7    15.9    16.5    13.7    -8.7    12.3    -8.5    10.6
        -6.0    16.0    -0.3    97.3
2003    8.5     13.5    25.9    21.6    0.0     15.4    2.2     3.1     3.1     5.2     0.9
        5.1     0.2     -5.2    -9.8    13.1    9.6     8.7     -5.8    9.0     -8.7    7.5
        -5.0    11.1    -0.9    234.6
2004    11.4    18.6    27.5    30.8    0.2     18.1    3.2     3.8     2.9     5.0     1.5
        4.0     4.3     -5.7    -8.0    13.7    14.7    12.2    -9.3    11.4    -9.3    10.3
        -1.8    26.9    0.1     416.7
2005    21.2    25.4    53.2    48.3    8.8     28.0    4.0     5.6     4.2     7.8     3.1
        8.2     1.2     -3.8    -7.2    28.8    19.1    15.4    -6.0    16.4    -6.4    13.5
        -4.8    18.7    -0.8    666.2
2006    11.1    15.1    17.5    20.8    4.5     17.6    2.5     3.1     2.6     3.7     1.7
        4.5     3.7     -1.5    -6.3    44.1    16.1    11.9    -6.8    10.6    -5.8    9.3
        -3.7    13.0    0.2     411.6
2007    9.8     16.3    18.8    28.7    3.8     12.7    3.7     4.6     4.4     6.5     2.3
        6.8     2.2     -1.5    -6.3    89.1    23.1    13.5    -8.0    12.5    -4.8    10.7
        -1.3    30.5    -0.7    593.2
;

data dat1;
set dat;
if (n <= 22);
run;

proc princomp data=dat1 OUT=Result1  PREFIX=Z OUTSTAT = Result2;
var X1  X2      X3      X4      X5      X6      X7      X8      X9      X10     X11     X12
        X13  X17 X18 X22 X24;
run;

proc print data=result1;
run;

proc print data=result2;
```

```
run;

/* Use a macro to build the list Z[i] of the first “num” Z’s; call the list “depends.” */
%MACRO ADINA (n);
%LOCAL units;
%DO units=1 %TO &n;
Z&units
%END;
%MEND depends;

*STEPWISE VARIABLE SELECTION AT 0.1 SIGNIFICANCE LEVEL;

PROC REG DATA=Result1;
MODEL y= %ADINA(5)/vif collinoint p r selection=stepwise slentry=0.1 slstay=0.1;
output out = dat2
predicted=yhat
r= res;

proc print data=dat2;
run;

* NORMALITY CHECK FOR RESIDUALS;

proc univariate data=dat2 normal plot;
var res;
run;

*EXPORT RESULTS TO EXCEL;

PROC DBLOAD DBMS=xls DATA=DAT2;
PATH='C:\Documents and Settings\Administrator\Desktop\RESULT\ssssss.XLS';
PUTNAME=yes;
LOAD;
RUN;
QUIT;
```

REFERENCES

Abdi, H. (2003). "Partial least squares regression (PLS-regression)." In Michael S. Lewis-Beck, Alan Bryman, and Tim Futing Liao (Eds.). *The Sage* e*ncyclopedia of social sciences research methods.* Sage Publications, Thousand Oaks, CA, 1–17.

Abrahart, R. J., and See, L. (2000). "Comparing neural network and autoregressive moving average techniques for the provision of continuous river flow forecasts in two contrasting catchments." *Hydrological Processes*, 14, 2157-2172.

Abrahart, R. J., and See, L. (2002). "Multi-model data fusion for river flow forecasting: an evaluation of six alternative methods based on two contrasting catchments." *Hydrology and Earth Systems Sciences*, 6(4), 655-670.

Abrahart, R. J., Kneale, P. E., and See, L. M. (2004). *Neural networks for hydrological modeling*. A. A. Balkema, Leiden.

Akaike, H. (1974). "A new look at statistical model identification." *IEEE Transactions on Automatic Control*, 19(6), 716–722.

Aryal, D. R., and Wang, Y. (2004). "Time-series analysis with hybrid Box-Jenkins ARIMA and neural network model." *Journal of Harbin Institute of Technology (English ed.)*, 11(4), 413-421.

Aslanargun, A., Mammadov, M., Yazici, B., and Yolacan, S. (2007). "Comparison of ARIMA, neural networks and hybrid models in time series: tourist arrival forecasting." *Journal of Statistical Computation and Simulation*, 77(1), 29-53.

Awadallah, A. G., and Rousselle, J. (2000). "Improving forecasts of Nile flood using SST inputs in TFN model." *Journal of Hydrological Engineering*, 5(4), 371-379.

Bender, M., and Simonovic, S. (1994). "Time-series modeling for long-range stream-flow forecasting." *Journal of Water Resources Planning and Management,* 120(6), 857-870.

Birikundavyi, S., Labib, R., Trung, H. T., and Rousselle, J. (2002). "Performance of neural networks in daily streamflow forecasting." *Journal of Hydrologic Engineering*, 7(5), 392-398.

Box, G. E. P., and Jenkins, G. M. (1976). *Time series analysis: Forecasting and control*. Holden-Day, San Francisco.

Burges, S. J., and Hoshi, K. (1978). "Incorporation of forecasted seasonal runoff volumes into reservoir management." *Water Resources Series Technical Report No. 58*, Department of Civil and Environmental Engineering, University of Washington, Seattle, Washington, USA. <http://www.ce.washington.edu/pub/WRS/WRS058> (July 30, 2008)

Cayan, D. R., Redmond, K. T., and Riddle, L. G. (1999). "ENSO and hydrological extremes in the western United States." *Journal of Climate,* 12, 2881–2893.

Clemen, R. T. (1989). "Combining forecasts: A review and annotated bibliography." *International Journal of Forecasting*, 5, 559-583.

Coulibaly, P., Anctil, F., and Bobée, B. (2000). "Daily reservoir inflow forecasting using artificial neural networks with stopped training approach." *Journal of Hydrology*, 230(3/4), 244–257.

Coulibaly, P., Hache, M., Fortin, V., and Bobee, B. (2005). "Improving daily reservoir inflow forecasts with model combination." *Journal of Hydrologic Engineering,* 10(2), 91-99.

de Jong, S. (1993). "SIMPLS: an alternative approach to partial least squares regression." *Chemometrics and Intelligent Laboratory Systems*, 18, 251-263.

Dibike,Y. B., and Solomatine D. P. (2001). "River flow forecasting using artificial neural networks." *Journal of Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere,* 26(1), 1-8.

Dong, X., Dohmen-Janssen, C. M., Booij, M., and S. Hulscher, S. (2006). "Effect of flow forecasting quality on benefits of reservoir operation – a case study for the Geheyan reservoir (China)." *Hydrology and Earth System Sciences Discussions,* 3, 3771–3814. < http://www.hydrol-earth-syst-sci-discuss.net/3/3771/2006/hessd-3-3771-2006-print.pdf> (July 18, 2008)

Eldaw, A. K., Salas, J. D., and Garcia, L. A. (2003). "Long-range forecasting of the Nile River flows using climatic forcing." *Journal of Applied Meteorology,* 42(7), 890-904.

Elshorbagy, A., Simonovic, S. P., and Panu, U. S. (2000). "Performance evaluation of artificial neural networks for runoff prediction." *Journal of Hydrologic Engineering*, 5(4), 424–427.

Eltahia, E. A. B. (1996). “El Niño and the natural variability in the flow of the Nile River. *Water Resources Research,* 32(1), 131-137.

Garen, D. C. (1992). “Improved Techniques in Regression-Based Streamflow Volume Forecasting.” *Journal of Water Resources Planning and Management*, 118(6), 654-670.

Geladi, P., and Kowalski, B. (1986). “Partial least squares regression: A tutorial.” *Analytica Chimica Acta*, 185, 1–17.

Govindaraju, R. S., and Rao, A. R. (Eds.). (2000). *Artificial neural networks in hydrology*. Kluwer Academic Publishers, Dordrecht.

Grantz, K. A. (2003). *Using large-scale climate information to forecast seasonal streamflow in the Truckee and Carson Rivers*. M.S. Thesis, University of Colorado, Boulder, CO. <http://cadswes.colorado.edu/PDF/Theses-PhD/Grantz_MS_thesis2003.pdf> (May 14, 2007)

Hamlet, A. F., and Lettenmaier, D. P. (1999). “Columbia River streamflow forecasting based on ENSO and PDO climate signals.” *Journal of Water Resources Planning and Management,* 125(6), 333-341.

Hamlet, A. F., Huppert, D., and Lettenmaier, D. P. (2002). “Economic value of long-lead streamflow forecasts for Columbia River hydropower.” *Journal of Water Resources Planning and Management*, 128(2), 91-101.

Hipel, K. W., and McLeod, A. I. (1994). *Time series modeling of water resources and environmental systems*. Elsevier Science, Amsterdam.

Hsieh, W. W., Yuval, Li, J., Shabbar, A., and Smith, S. (2003). “Seasonal prediction with error estimation of Columbia River streamflow in British Columbia.” *Journal of Water Resources Planning and Management*, 129(2), 146-149.

Hsu, K., Gupta, H. V., Gao, X., Sorooshian, S., and Imam, B. (2002). “Self-organizing linear output map (SOLO): An artificial neural network suitable for hydrologic modeling and analysis.” *Water Resources Research*, *38*(12), 1302.

Hsu, K., Gupta, H.V., and Sorooshian, S. (1995). “Artificial neural network modeling of the rainfall-runoff process.” *Water Resources Research*, 31(10), 2517-2530.

Hu, T. S., Lam, K. C., and Ng, S. T. (2001). “River flow time series prediction with a range-dependent neural network.” *Hydrological Sciences Journal,* 46(5), 729-745.

Huang, W., Xu, B., and Chan-Hilton, A. (2004). “Forecasting flows in Apalachicola River using neural networks.” *Hydrological Processes*, 18(13), 2545-2564.

Jain, A., and Kumar, A. M. (2007). “Hybrid neural network models for hydrologic time series forecasting.” *Applied Soft Computing*, 7(2), 585-592.

Jain, A., Sudheer, K. P., and Srinivasulu, S. (2004). “Identification of physical processes inherent in artificial neural network rainfall–runoff models.” *Hydrological Processes*, 18(3), 571–581.

Joy, M. P., and Jones, S. (2005). “Predicting bed demand in a hospital using neural networks and ARIMA models: a hybrid approach.” In Michel Verleysen (Ed.). *ESANN' 2005:13th European Symposium on Artificial Neural Networks, Bruges, Belgium, April 27-28-29, 2005: proceedings.* d-side Publ., Evere, Belgium, 127-132. <http://www.dice.ucl.ac.be/Proceedings/esann/esannpdf/es2005-39.pdf> (May 15, 2009).

Karamouz, M., and Zahraie, B. (2004). “Seasonal streamflow forecasting using snow budget and El Niño Southern Oscillation climate signals: Application to the Salt River Basin in Arizona.” *Journal of Hydrologic Engineering,* 9(6), 523-533.

Karunanithi, N., Grenney, W. J., Whitley, D., and Bovee, K. (1994). “Neural networks for river flow prediction.” *Journal of Computing in Civil Engineering,* 8, 2, 201-220.

Kim, T., and Valdes, J. B. (2003). Nonlinear model for drought forecasting based on a conjunction of wavelet transforms and neural networks. *Journal of Hydrologic Engineering,* 8(6), 319-328.

Kim, Y. O. and Palmer, R. N. (1997). “The value of seasonal flow forecasts in Bayesian stochastic programming.” *Journal of Water Resources Planning and Management*, 123(6), 327-335.

Kişi, Ö. (2008). “Stream flow forecasting using neuro-wavelet technique.” *Hydrological Processes*, 22, 4142-4152. <http://www3.interscience.wiley.com/cgi-bin/fulltext/117927378/PDFSTART (August 18, 2008).

Kohenon, T. (1984). *Self-organization and associative memory.* Springer-Verlag, Berlin.

Lall, U. B., and Sharma, A. (1996).“A nearest neighbor bootstrap for resampling hydrologic time series.” *Water Resources Research*, 32(3), 679-693.

Lee, S. W. (2004). *Investigation of techniques for improvement of seasonal streamflow forecasts in the Upper Rio Grande Basin*. Ph.D. dissertation, Texas A&M University, College Station, TX. <http://handle.tamu.edu/1969.1 /2764 > (August 13, 2008).

Lee, S. W., Klein, A., and Over, T. (2004). “Effects of the El Niño - Southern Oscillation on temperature, precipitation, snow water equivalent and resulting streamflow in the Upper Rio Grande River Basin. *Hydrological Processes,* 18(6), 1053-1071. <http://www3.interscience.wiley.com/cgi-bin/fulltext/107640166/PDFSTART> (August 18, 2008).

Lettenmaier, D. P., and Wood, E. F. (1993). “Hydrological forecasting.” In D.R. Maidment (Eds.), *Handbook of Hydrology*. McGraw-Hill, New York, 26.1-26.30.

Mantua, N. J., Hare, S. R., Zhang, Y., Wallace, J. M., and Francis, R. C. (1997). “A Pacific interdecadal climate oscillation with impacts on salmon production.” *Bulletin of the American Meteorological Society*, 78, 1069-1079.

Markus, M., Salas, J. D., and Shin, H. (1995). “Predicting streamflows based on neural networks.” *Proceedings of the First International Conference on Water Resources Engineering.* American Society of Civil Engineers., New York, 1641-1646.

McCuen, R. H. (1985). *Statistical methods for engineers*. Prentice Hall, inc., Englewood Cliffs, New Jersey.

McCuen, R. H., and Snyder, W. M. (1986). *Hydrologic modeling: Statistical methods and applications*. Prentice Hall, inc., Englewood Cliffs, New Jersey.

McCuen, R. H., Rawls, W. J., and Whaley, B. L. (1979). “Comparative evaluation of statistical methods for water supply forecasting.” *Water Resources Bulletin*, 15(4), 935-947.

McKerchar, A. I., and Delleur, J. W. (1974). “Application of seasonal parametric linear stochastic models to monthly flow data.” *Water Resources Research,* 10, 246-255.

McLeod, A. I., Noakes, D. J., Hipel, K. W., and Thompstone, R. M. (1987). "Combining hydrologic forecasts." *Journal of Water Resources Planning and Management*, 113(1), 29-41.

Mehdikhani, H., Abrishamchi, A., and Khodaei, H. (2006). "Developing a conjunctive nonlinear model for inflow prediction using wavelet transforms and artificial neural networks: A case study of Dez Reservoir Dam, Iran." In Darell D. Zimbelman and Werner Loehlein (Eds.). *Operating reservoirs in changing conditions; Proceedings of the Operation Management 2006 Conference*, *August, 2006, Sacramento, California*. American Society of Civil Engineers and Environmental and Water Resources Institute (EWRI), Reston, VA, 69-78.

Mondal, M. S., and Wasimi, S. A. (2005). "Periodic transfer function-noise model for forecasting." *Journal of Hydrologic Engineering*, 10(5), 353-362.

Nash, J. E., and Sutcliffe, J. V. (1970). "River flow forecasting through conceptual models: 1. A discussion of principles." *Journal of Hydrology,* 10, 282–290.

Natural Resources Conservation Service (NRCS) (1997). *Southern oscillation index statistical correlation with spring runoff in the western US.* <http://www.wrcc.dri.edu/enso/soiwsf2.pdf> (June 16, 2008).

Natural Resources Conservation Service (NRCS) (2007). "Statistical techniques used in the VIPER water supply forecasting software." *NRCS Technical Note* 210-2. <http://directives.sc.egov.usda.gov/tn_sswsf_2_a.pdf> (April 2, 2008).

NeuroDimension, inc. (2009). *NeuroSolutions Getting Started Manual Version 5.* NeuroDimension, inc., Gainesville, FL. <http://www.neurosolutions.com/downloads/documentation.html> (January 12, 2009).

Noakes, D. J., McLeod, A. I., and Hipel, K.W. (1985). "Forecasting monthly riverflow time series." *International Journal of Forecasting*, 1, 179-190.

Pagano, T. (2005). *The role of climate variability in operational water supply forecasting for the western United States*. Ph.D. dissertation. University of Arizona, Tucson, AZ.

Pagano, T., and Garen, D. (2005). "A recent increase in western U.S. streamflow variability and persistence." *Journal of Hydrometeorology,* 6(2)**,** 173–179.

Pagano, T., and Garen, D. (2006). “Integration of climate information and forecasts into Western US water supply forecasts.” In J. D. Garbrecht and T. C. Piechota, (Eds.). *Climate variations, climate change, and water resources engineering*. American Society of Civil Engineers, Reston, VA, 86-102.

Pagano, T., Garen, D., and Sorooshian, S. (2004). “Evaluation of official western U.S. seasonal water supply outlooks, 1922–2002.” *Journal of Hydrometeorology,* 5(5), 896–909.

Parasuraman, K., and Elshorbagy, A. (2007). “Cluster-based hydrologic prediction using genetic-algorithm-trained neural networks.” *Journal of Hydrologic Engineering,* 12(1), 52-62.

Parasuraman, K., Elshorbagy, A., and Carey, S. K. (2006). “Spiking modular neural networks: A neural network modeling approach for hydrological processes.” W*ater Resources Research*, *42(2),* W05411. <http://http-server.carleton.ca/~scarey/2005WR004317.pdf > (August 15, 2008)

Parker, D., Tunstall S., and Wilson, T. (2005). "Socio-economic benefits of flood forecasting and warning.” In *Proceedings of the International Conference on Innovation, Advances and Implementation of Flood Forecasting Technology, October 17-19, 2005, Norway, Tromso*. < http://www.actif-ec.net/conference2005/proceedings/PDF%20docs/Session_08_Flood_warning/Parker_Dennis.pdf> (August 18, 2008)

Piechota, T. C., and Dracup, J. A. (1999). “Long-range streamflow forecasting using El–Nino southern oscillation indicators.” *Journal of Hydrologic Engineering,* 4(2), 144-151.

Piechota, T. C., Chiew, F. H. S., Dracup, J. A., and McMahon, T. A. (1998). “Seasonal streamflow forecasting in eastern Australia and the El Niño–Southern Oscillation.” *Water Resources Research*, 34(11), 3035– 3044.

Piechota, T. C., Chiew, F. H. S., Dracup, J. A., and McMahon, T. A. (2001). “Development of exceedance probability streamflow forecast.” *Journal of Hydrologic Engineering*, 6(1), 20–28.

Prairie, J. R., Rajagopalan, B., Fulp, T. J., and Zagona, E. A. (2006). “Modified K-NNmodel for stochastic streamflow simulation.” *Journal of Hydrologic Engineering*, 11(4), 371–378.

Rajagopalan, B., and Lall, U. (1999). “A nearest neighbor bootstrap resampling scheme for resampling daily precipitation and other weather variables.” *Water Resources Research,* 35(10), 3089–3101.

Raman, H., and Sunilkumar, N. (1995). “Multivariate modeling of water resources time series using artificial neural network.” *Hydrological Sciences Journal,* 40(2), 145–163.

Redmond, K. T., and Koch, R. W. (1991). “Surface climate and streamflow variability in the western United States and their relationship to large-scale circulation indices.” *Water Resources Research,* 27(9), 2381-2399.

Rio Grande Compact Commission (2006). *Report of the Rio Grande Compact Commission 2005.* New Mexico Office of the State Engineer, Santa Fe, New Mexico.

Risley, J. C., Gannett, M. W., Lea, J. K. and Roehl, E. A. (2005). “An analysis statistical methods for seasonal flow forecasting in the Upper Klamath River Basin of Oregon and California.” *Scientific Investigations Report of the U.S. Geological Survey, 2005-5177.* < http://pubs.usgs.gov/sir/2005/5177/> (March 21, 2008)

Salas, J. D. (1992). “Analysis and modeling of hydrologic time series.” In D. R. Maidment. (Ed.). *Handbook of hydrology*. McGraw-Hill, New York, 19.1-19.72.

Salas, J. D., Markus, M., and Tokar, A. S. (2000). “Streamflow forecasting based on artificial neural networks.” In R. S. Govindaraju and A. R. Rao (Eds.). *Artificial neural networks in hydrology.* Kluwer Academic Publishers, Dordrecht, the Netherlands, 23-51.

SAS Institute (2008) *SAS/STAT 9.2 user's guide. The PLS procedure*. SAS Institute Inc., Cary, NC. <http://www.technion.ac.il/docs/sas/stat/chap51/index.htm> (May 21, 2008)

See, L. M., and Abrahart, R. J. (2001).” Multi-model data fusion for hydrological forecasting.” *Computers and Geosciences*, 27(8), 987-994.

See, L., and Openshaw, S. (1999). “Applying soft computing approaches to river level forecasting.” *Hydrological Sciences Journal*, 44(5), 763 -778.

See, L., and Openshaw, S. (2000). “A hybrid multi-model approach to river level forecasting.” *Hydrological Sciences Journal*, 45(4), 523 –536.

Shamseldin, A. Y. (1997). “Application of a neural network technique to rainfall–runoff modeling.” *Journal of Hydrology,* 199, 272–294.

Shamseldin, A. Y. (2004). “Hybrid neural network modelling solutions.” In R. J. Abrahart, P. E. Kneale, and L. M. See (Eds.). *Neural networks for hydrological modeling*. A.A. Balkema Publishers, Leiden, 61-79.

Shamseldin, A. Y., and O’Connor, K. M. (1996). “A nearest neighbor linear perturbation model for river flow forecasting.” *Journal of Hydrology,* 179, 353–375.

Shamseldin, A. Y., O’Connor, K. M., and Liang, G. C. (1997). “Methods for combining the output of different rainfall-runoff models.” *Journal of Hydrology,* 197, 203–229.

Shamseldin, A. Y., Nasr, A. E., and O’Connor, K. M. (2002). “Comparison of different forms of the multi-layer feed-forward neural network method used for river flow forecasting.” *Hydrology and Earth System Sciences*, 6, 671-684.

Sharma, A., Tarboton, D. G., and Lall, U. (1997). “Streamflow simulation: A nonparametric approach.” *Water Resources Research*, 33(2), 291-308.

Solomatine, D. P., and Price, R. K. (2004). “Innovative approaches to flood forecasting using data driven and hybrid modeling.” *Proceedings of the 6th International Conference on Hydroinformatics, Singapore, 21-24 June 2004*. World Scientific Publishing Co., Singapore. <http://www.hi.ihe.nl/hi/sol/papers/HIC04_196_InnovApprFlood_Price_Sol.pdf> (September 9, 2008)

Souza, F. A., and Lall, U. B. (2003). “Seasonal to interannual ensemble streamflow forecasts for Ceara, Brazil: Applications of a multivariate, semi-parametric algorithm.” *Water Resources Research*, 39, 1307–1320.

Srinivas, V. V., and Srinivasan, K. (2001). “A hybrid stochastic model for multiseason streamflow simulation”. *Water Resources Research*, 37(10), 2537-2549.

StatSoft , Inc. (2008). *Partial Least Squares (PLS).* (Electronic Textbook). <http://statsoft.eu/uk/textbook/stpls.html> (May 2, 2009)

Sun, L., Ji, S., Yu, S., and Ye, J. (2009). “On the Equivalence between Canonical Correlation Analysis and Orthonormalized Partial Least Squares.” *Proceedings of the Twenty-First International Joint Conference on Artificial*

Intelligence, 2009. Springer Verlag, Berlin. <http://www.public.asu.edu/~lsun27/Publications/IJCAI_2009.pdf> (May 2, 2009)

Tesfaye, Y. G., Meerschaert, M. M., and Anderson, P. L. (2005). *Parsimonious PARMA models and their application to modeling river flows.* <www.maths.otago.ac.nz/~mcubed/FourierPARMA.pdf> (May 8, 2006)

Thompstone, R.M., Hipel, K.W., and Mcleod, A. I. (1985). "Forecasting quarter-monthly river flow." *Water Resources Bulletin*, 21(5), 731-741.

Tobias, R. (1995). "An introduction to partial least squares regression." In Neil Howard (Ed.). *SUGI 20, Orlando: proceedings of the twentieth annual SAS Users Group International Conference, Orlando, Florida, April 2-5, 1995.* SAS Institute, Inc., Cary, NC, 1250-1257.

Tokar, A. S., and Markus, M. (2000). "Precipitation-runoff modeling using artificial neural network and conceptual models." *Journal of Hydrologic Engineering*, 5(2), 156–161.

Tootle, G. A., and Piechota, T. C. (2004). "Suwannee River long range streamflow forecasts based on seasonal climate predictors." *Journal of the American Water Resources Association,* 40(2), 523–532.

Tootle, G. A., and Piechota, T. C. (2006). "The relationships between Pacific and Atlantic Ocean sea surface temperatures, and U.S. streamflow variability." *Water Resources Research*, 42(7), W07411. <http://www.agu.org/journals/wr/wr0607/2005WR004184/2005WR004184.pdf > (August 28, 2008).

Tootle, G. A., Singh, A. K., Piechota, T. C., and Farnham, I. (2007). "Long lead-time forecasting of U.S. streamflow using partial least squares regression." *Journal of Hydrologic Engineering*, 12(5), 442–451.

Tseng, F., Yu, H., and Tzeng, G. (2002). "Combining neural network model with seasonal time series ARIMA model." *Technological Forecasting and Time & Social Change*, 69, 71-87.

Umetrics, Inc. (1995). *Multivariate Analysis* (3-day course), Winchester, MA

van der Voet, H. (1994). "Comparing the predictive accuracy of models using a simple randomization test." *QCAR: Chemometrics and Intelligent Laboratory Systems*, 25, 313-323.

Vandaele, W. (1983). *Applied time series and Box-Jenkins models*. Academic Press, New York.

Wang, W. (2006). *Stochasticity, nonlinearity and forecasting of streamflow processes.* IOS Press, Amsterdam.

Wang, W., van Gelder, P. H. A. J. M., and Vrijling, J. K. (2005a). Constructing prediction interval for monthly streamflow forecasts. In J. K. Vrijling et al. (Eds.) *Proceedings of the 9th International Symposium on Stochastic Hydraulics*, *May 23- 24, 2005, Nijmegen, Netherlands.* International Association of Hydraulic Research, Madrid, Spain. <http://www.bing.com/search?q=Constructing+prediction+interval+for+monthly+streamflow+forecasts.&src=IE-SearchBox>_ (August 25, 2008)

Wang, W., van Gelder, P. H. A. J. M., Vrijling, J. K., and Ma, J. (2005b). "Forecasting daily streamflow using hybrid ANN models." *Journal of Hydrology*, 9, 1-17.

Whitaker, D.W., Wasimi, S.A. and Islam, S.(2001). "The El Niño-Southern Oscillation and long-range forecasting of flows in the Ganges." *International Journal of Climatology*, 21(1), 77-87.

Wold, H. (1966). "Estimation of principal components and related models by iterative least squares." In P. R. Krishnaiah (Ed.). *Multivariate analysis*. New York: Academic Press, New York, 391-420.

Wold, S. (1994). "PLS for multivariate linear modeling." In H. van de Waterbeemd (Ed*.*). *QSAR: Chemometric methods in molecular design, methods and principles in medicinal chemistry*. Verlag-Chemie, Weinheim, Germany,. 195-218.

Wood, A. W., and Lettenmaier, D. P. (2006). "A test bed for new seasonal hydrologic forecasting approaches in the western United States." *Bulletin of the American Meteorological Society*, 87(12), 1699-1712.

Yeh, W W-G., Becker, L., and Zettlemoyer, R. (1982). "Worth of inflow forecast for reservoir operation." *Journal of Water Resources Planning and Management,* 108, 257-269.

Yeniay, Ö., and Göktaş, A. (2002). "A comparison of partial least squares regression with other prediction methods." *Hacettepe Journal of Mathematics and Statistics*, 31, 99-111.

Yürekli, K., Kurunç, A., and Öztürk, F. (2005). “Testing the residuals of an ARIMA model on the Cekerek Stream Watershed in Turkey.” *Turkish Journal of Engineering and Environmental Sciences*, 29, 61-74.

Zhang, B., and Govindaraju, R. S. (2000). “Prediction of watershed runoff using Bayesian concepts and modular neural networks.” *Water Resources Research,* 36(3), 753–762.

Zhang, G. P. (2003). “Time series forecasting using hybrid ARIMA and neural network models.” *Neurocomputing*, 50, 159-175.

Zhang, G., Patuwo, B. E., and Hsu, M.Y. (1998). “Forecasting with artificial neural networks: The state of the art.” *International Journal of Forecasting,* 14, 35-62.