Utah State University

DigitalCommons@USU

All Graduate Theses and Dissertations

Graduate Studies

5-2012

Identification of Influential Climate Indicators, Prediction of Long-term Streamflow and Great Salt Lake Elevation Using Machine Learning Approach

Niroj K. Shrestha
Utah State University

Follow this and additional works at: https://digitalcommons.usu.edu/etd

Part of the Civil and Environmental Engineering Commons

Recommended Citation

Shrestha, Niroj K., "Identification of Influential Climate Indicators, Prediction of Long-term Streamflow and Great Salt Lake Elevation Using Machine Learning Approach" (2012). *All Graduate Theses and Dissertations*. 1335.
https://digitalcommons.usu.edu/etd/1335

This Dissertation is brought to you for free and open access by the Graduate Studies at DigitalCommons@USU. It has been accepted for inclusion in All Graduate Theses and Dissertations by an authorized administrator of DigitalCommons@USU. For more information, please contact digitalcommons@usu.edu.

UtahStateUniversity
MERRILL-CAZIER LIBRARY

IDENTIFICATION OF INFLUENTIAL CLIMATE INDICATORS, PREDICTION OF LONG-TERM STREAMFLOW AND GREAT SALT LAKE ELEVATION USING LEARNING MACHINE APPROACH

by

Niroj K. Shrestha

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

of

DOCTOR OF PHILOSOPHY

in

Civil and Environmental Engineering

Approved:

Dr. Gilberto Urroz Major Professor	Dr. Mac McKee Committee Member
Dr. A Bruce Bishop Committee Member	Dr. David K Stevens Committee Member
Dr. YangQuan Chen Committee Member	Dr. Mark R. McLellan Dean of School of Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah

2012

Copyright © Niroj Shrestha 2012
All Rights Reserved

ABSTRACT

Identification of Influential Climate Indicators, Prediction of Long-term Streamflow and Great Salt Lake Elevation Using Machine Learning Approach

by

Niroj K. Shrestha, Doctor of Philosophy

Utah State University, 2012

Major Professor: Dr. Gilberto Urroz
Department: Civil and Environmental Engineering

To meet the surging water demand due to rapid population growth and changing climatic conditions around the world, and to reduce the impact of floods and droughts, comprehensive water management and planning is necessary. Climatic variability, hydrologic uncertainty and variability of hydrologic quantities in time and space are inherent to hydrological modeling. Hydrologic modeling using a physically-based model can be very complex and typically requires detailed knowledge of physical processes. The availability of data is an important issue to justify the use of these models. Data-driven models are an alternative choice. This is a relatively new and efficient approach to modeling. Data-drive models bridge the gap between the classical regression and physically-based models. By using a data-driven model that relies on the machine learning approach, it is possible to produce reasonable predictions from a limited data set and limited knowledge of underlying physical processes of the system by just relating input and output. This dissertation uses the Multivariate Relevance Vector Machine

(MVRVM) and Support Vector Machine (SVM) for predicting a variety of hydrological quantities. These models are used in this dissertation for identifying influential climate indicators, and are used for long-term streamflow prediction for multiple lead times at different locations in Utah. They are also used for prediction of Great Salt Lake (GSL) elevation series. They provide reasonable predictions of hydrological quantities from the available data. The predictions from these models are robust and parsimonious. This research presents the first attempt to identify influential climate indicators and predict long lead-time streamflow in Utah, and to predict lake elevation using machine learning models. The approach presented herein has potential value for water resources planning and management especially for irrigation and flood management.

(194 pages)

PUBLIC ABSTRACT

Identification of Influential Climate Indicators, Prediction of Long-term Streamflow and Great Salt Lake Elevation Using Machine Learning Approach

by

Niroj K. Shrestha, Doctor of Philosophy

Utah State University, 2012

Major Professor: Dr. Gilberto Urroz
Department: Civil and Environmental Engineering

In order to meet rising water demand due to rapid population growth and changing climatic conditions around the world, and to reduce the impact of floods and draughts, a comprehensive water management and planning is necessary. Water resource management requires the prediction of streamflow under climatic variability, and variability of hydrologic quantities that changes in time and space. Prediction of streamflow using physically-based model are usually complex and typically requires detailed knowledge of physical processes. The availability of data is an important issue to justify the use of these models. Using a data-driven model that relies on the machine learning approach, it is possible to produce reasonable predictions from a limited data set and limited knowledge of underlying physical processes of the system by just relating input and output. This dissertation uses the Multivariate Relevance Vector Machine (MVRVM) for identifying influential climate indicators, and uses them for long-term streamflow prediction for multiple lead times at different locations in Utah. Both

MVRVM and Support Vector Machine (SVM) are used for prediction of Great Salt Lake (GSL) elevation series. They provide reasonable predictions of hydrological quantities from the available data. The predictions from these models are robust and parsimonious. The approach presented herein has potential value for water resources planning and management.

This work is dedicated to my family.

ACKNOWLEDGMENTS

I deeply appreciate Dr. Mac McKee for his guidance and expert advice in my research. I would like to thank my major professor, Dr. Gilberto Urroz, for going through my dissertation and correcting it. I thank him for the high degree of freedom that I enjoyed during my research. His friendly behavior encouraged me to work smoothly and comfortably.

I would like to thank Dr. Robert Gillies and Dr. Shih-Yu (Simon) Wang for giving me some time to review the result and giving insightful comments. I would like to thank Dr. YangQuan Chen for giving me insightful comment in one of the chapters. I am grateful to Dr. David Stevens and Dr. A. Bruce Bishop for accepting to be on my committee and giving comments. I appreciate the Utah Water Research Laboratory, Utah State University, for providing me an opportunity for higher education. Thanks to my fellow graduate students for insightful discussion. I would like to thank to Nepalese community in Logan, which helped me to live like home in a foreign land. I am indebted to all who have supported me during my research activity, without whom I would not have reached this level.

I especially thank my family and friends from Nepal for constantly inspiring me throughout my study. Their support meant a lot to me.

Niroj K. Shrestha

CONTENTS

Page

ABSTRACT iii

PUBLIC ABSTRACT v

ACKNOWLEDGMENTS viii

LIST OF TABLES xii

LIST OF FIGURES xiii

CHAPTER

1. INTRODUCTION 2

References 7

2. BAYESIAN MACHINE LEARNING REGRESSION APPROACH FOR IDENTIFICATION OF INFLUENTIAL SEA SURFACE TEMPERATURE LOCATIONS AND PREDICTING STREAMFLOW FOR THE NEXT SIX MONTHS 9

Abstract 9
2.1 Introduction 10
2.2 Study Area 12
2.3 Background 14
2.4 Model Description 15
2.5 Data Collection and Description 20

2.5.1 Streamflow 20
2.5.2 Snow water equivalent 20
2.5.3 Local temperature 21
2.5.4 Sea surface temperature 21

2.6 Model Development 22

2.6.1 Model 1 23
2.6.2 Model 2 24
2.6.3 Performance criteria 25

2.7 Results and Discussion 26

2.7.1 Identification of influential SST location and prediction of monthly mean discharge for the next six months using Model 1 27
2.7.2 Identification of influential SST locations and prediction of the volume of water passing the gage for next six months using Model 2 33
2.7.3 Generalization and robustness 35

2.8 Conclusion 36
References 39

3. LONG LEAD-TIME STREAMFLOW FORECASTING AND IDENTIFICATION OF RELATIVE INFLUENCE OF OCEANIC-ATMOSPHERIC OSCILLATION MODES USING BAYESIAN MACHINE LEARNING REGRESSION APPROACH 59

Abstract 59
3.1 Introduction 60
3.2 Study Area 63
3.3 Background 64
3.4 Model Description 65
3.5 Data Collection and Description 70

3.5.1 Streamflow 70
3.5.2 Pacific Decadal Oscillation (PDO) 70
3.5.3 El-Niño Southern Oscillation (ENSO) 71
3.5.4 Atlantic Multi-decadal Oscillation (AMO) 72
3.5.5 North Atlantic Oscillation (NAO) 72

3.6 Model Development and Performance Criteria 73

3.6.1 Root mean square error (RMSE) 75
3.6.2 Pearson correlation coefficient (*r*) 75
3.6.3 Nash-Sutcliffe efficiency 75

3.7 Results 76

3.7.1 Correlation coefficients 76
3.7.2 Identification of best combination of oscillations and lead time for annual streamflow volume predictions 77
3.7.3 Prediction results from best identified combinations of oscillations for each model types 79
3.7.4 Discussion and relative influence of oscillation indices 81
3.7.5 Comparison with SVM and ANN 88
3.7.6 Generalization and robustness of model 89

3.8 Conclusion 90

References 92

4. MACHINE LEARNING REGRESSION APPROACH FOR PREDICTION OF GREAT SALT LAKE WATER SURFACE ELEVATION 118

Abstract 118
4.1 Introduction 119
4.2 Study Area and Data Collection 120
4.3 Model Description 121

4.3.1 Support Vector Machine 121
4.3.2 Relevance Vector Machine 126

4.4 Model Formulation and Application 131

4.4.1 Estimating time delay and system dimension 132
4.4.2 Performance criteria 135

4.5 Results 137

4.5.1 Parameter estimations 137
4.5.2 Support Vector Machine 138
4.5.3 Effect of embedding dimension in model prediction for SVM 139
4.5.4 Relevance Vector Machine 139
4.5.5 Effect of embedding dimension in model prediction for MVRVM 140
4.5.6 Generalization and robustness of models 141

4.6 Discussion and Conclusion 142
References 144

5. SUMMARY, CONCLUSION, AND RECOMMENDATION 161

5.1 Summary and Conclusion 161
5.2 Recommendation and Future Direction 166
References 167

APPENDIX 169

CURRICULUM VITAE 175

LIST OF TABLES

Table Page

2.1 Geometric characteristic of stream gages 42

2.2 95 percent confidence interval of the median test RMSE for Model 1 43

2.3 Best test statistics for monthly mean discharge prediction (Model 1)...................... 43

2.4 95 percent confidence interval of the median test RMSE for Model 2 44

2.5 Best test statistics for volume of water passing the gage for next six months (Model 2) 44

3.1 Geometric characteristic of stream gages 96

3.2 Best statistics, combination of oscillations, and lead time for Model 1 97

3.3 Best statistics, combination of oscillations, and lead time for Model 2 97

3.4 Test statistics, combination of oscillations, and lead time for Model 3.................... 97

3.5 Test statistics, combination of oscillations, and lead time for Model 4.................... 98

3.6 Relative strength of oscillation modes from Model 2.. 99

3.7 Relative strength of oscillation modes from Model 3.. 99

3.8 Relative strength of oscillation modes from Model 4.. 99

3.9 Combination of oscillations that produces best results for each lead time............. 100

4.1 Construction of multivariate state space out of single measurement using d=5 and τ=8 147

4.2 Results for Figures 6(a)-6(d)... 148

4.3 Results from Figures 9(a)-9(d)... 148

A.1 Observed mean and variance for GSL elevation 172

LIST OF FIGURES

Figure Page

2.1 Location of the stream gages and SnoTel stations in Utah. 45

2.2 The locations for the sea surface temperature (Khalil et al. 2005a). 46

2.3 Flowchart for developing the input file for different combinations of input variables using one individual sea surface temperature at a time for both monthly mean discharge prediction (Model 1) and volumetric prediction (Model 2). The path to each final node shows the set of input variables used for that node. (a) Combination of input variables for Weber River near Oakley, (b) Combination of input variables for Chalk Creek at Coalville, (c) Combination of input variables for Muddy Creek near Emery, and (d) Combination of input variables for Sevier River at Hatch. 47

2.4 Sample flowchart for developing the input file for prediction of volume of water passing through the stream gages (Model 2) using combined SST of different locations of Pacific and Atlantic Oceans. The path to each final node shows the set of input variables used for that node. (a) Combination of input variables for Weber River near Oakley, (b) Combination of input variables for Chalk Creek at Coalville, (c) Combination of input variables for Muddy Creek near Emery, and (d) Combination of input variables for Sevier River at Hatch. 48

2.5 The test RMSE for six individual SST locations at each selected stream gage for monthly mean discharge prediction for next six months. (a) Weber River near Oakley, (b) Chalk Creek at Coalville, (c) Muddy Creek near Emery, (d) Sevier River at Hatch, and (e) Sixth Water Creek near Springville. 49

2.6 Test RMSE for the combinations of SST locations for the monthly mean discharge prediction for next six months at each selected stream gage. (a) Weber River near Oakley, (b) Chalk Creek at Coalville, (c) Muddy Creek near Emery, (d) Sevier River at Hatch, and (e) Sixth Water Creek near Springville. 49

2.7 Locations of the sea surface temperature that develops the best test result for the monthly mean discharge prediction at selected streamflow gages for next six months. 50

2.8 Streamflow prediction for next six months at each selected stream gage. For (a) to (e), first column is for the training phase, second column is for test phase, third column shows the plot of predicted discharge versus actual discharge for training phase, and fourth column shows the similar plot for the test phase. (a) Weber River near Oakley, (b) Chalk Creek at Coalville, (c) Muddy Creek near Emery, (d) Sevier River at Hatch, (e) Sixth Water Creek near Springville, (f) 90 percent confidence interval of prediction in test phase for each selected gage for (a) to (e), and (g) Residual plots for (a) to (e). 51

2.9 The test RMSE of six individual SST locations at each selected stream gage for volume of water passing the gage for next six months. (a) Weber River near Oakley, (b) Chalk Creek at Coalville, (c) Muddy Creek near Emery, (d) Sevier River at Hatch, and (e) Sixth Water Creek near Springville. 53

2.10 Test RMSE for the combinations of SST locations for the volumetric prediction at each selected stream gage. (a) Weber River near Oakley, (b) Chalk Creek at Coalville, (c) Muddy Creek near Emery, (d) Sevier River at Hatch, and (e) Sixth Water Creek near near Springville. 53

2.11 Locations of the sea surface temperature that developed the best test statistics for volume of water passing the stream gage for next six months. 54

2.12 The prediction for volume of water passing through the stream gages for next six months. For (a) to (e), first column is the training phase, second column is test phase, third column is the plot of predicted volume versus actual volume for training phase, and fourth column is similar plot for the test phase. The results are in the order of (a) Weber River near Oakley, (b) Chalk Creek at Coalville, (c) Muddy Creek near Emery, (d) Sevier River at Hatch, (e) Sixth Water Creek near Springville, (f) 90 percent confidence interval of prediction for (a) to (e), and (g) Residual plots of test phase for (a) to (e). 55

2.13 The bootstrap analysis for the best model at each stream gage for monthly mean discharge prediction for next six months. The figures presented are for the RMSE values for (a) Weber River near Oakley, (b) Chalk Creek at Coalville, (c) Muddy Creek near Emery, (d) Sevier River at Hatch, and (e) Sixth Water Creek near Springville. 57

2.14 The bootstrap analysis for the best model for each stream gage for monthly mean discharge prediction. The figures presented are for the Nash-Sutcliffe efficiency for (a) Weber River near Oakley, (b) Chalk Creek at Coalville, (c) Muddy Creek near Emery, (d) Sevier River at Hatch, and (e) Sixth Water Creek near Springville. 57

2.15 The bootstrap analysis of the best model for volumetric prediction at each stream gage. The figures presented are for the RMSE values for (a) Weber River near Oakley, (b) Chalk Creek at Coalville, (c) Muddy Creek near Emery, (d) Sevier River at Hatch, and (e) Sixth Water Creek near Springville. 58

2.16 The bootstrap analysis for the best model for volumetric prediction at each selected stream gages. The figures presented are for the Nash-Sutcliffe efficiency for (a) Weber River near Oakley, (b) Chalk Creek at Coalville, (c) Muddy Creek near Emery, (d) Sevier River at Hatch, and (e) Sixth Water Creek near Springville. 58

3.1 Location of the stream gages. 101

3.2 Ocean Atmospheric Oscillation indices (a) PDO, (b) ENSO, (c) AMO, and (d) NAO. 102

3.3 Absolute value of correlation coefficient between the annual oscillation modes and annual flow volume for (a) PDO, (b) ENSO, (c) AMO, and (d) NAO. 103

3.4 Test RMSE at 1 to 5 year lead for annual streamflow volume prediction for Model 1. (a) Weber River near Oakley, (b) Chalk Creek at Coalville, (c) Sevier River at Hatch, and (d) Muddy Creek near Emery. 104

3.5 The test RMSE at 1 to 5 year lead for annual streamflow volume prediction for Model 2. (a) Weber River near Oakley, (b) Chalk Creek at Coalville, (c) Sevier River at Hatch , and (d) Muddy Creek near Emery. 105

3.6 The test RMSE at 1 to 5 year lead for annual streamflow volume prediction for Model 3. (a) Weber River near Oakley, (b) Chalk Creek at Coalville, (c) Sevier River at Hatch, and (d) Muddy Creek near Emery. 106

3.7 The test RMSE at 1 to 5 year lead for annual streamflow volume prediction for Model 4. (a) Weber River near Oakley, (b) Chalk Creek at Coalville, (c) Sevier River at Hatch, and (d) Muddy Creek near Emery. 107

3.8 The plot of actual versus predicted annual flow volume for Model 1. (a) Weber River near Oakley, (b) Chalk Creek at Coalville, (c) Sevier River at Hatch, (d) Muddy Creek near Emery, (e) 90% confidence interval of prediction in test phase for all gages, and (f) Residual plots of test phase for (a) to (d). 108

3.9 The plot of actual versus predicted annual flow volume for Model 2. (a) Weber River near Oakley, (b) Chalk Creek at Coalville, (c) Sevier River at Hatch, (d) Muddy Creek near Emery, (e) 90% confidence interval of prediction in test phase for all gages, and (f) Residual plots of test phase for (a) to (d). 110

3.10 The plot of actual versus predicted annual flow volume for Model 3. (a) Weber River near Oakley, (b) Chalk Creek at Coalville, (c) Sevier River at Hatch, (d) Muddy Creek near Emery, (e) 90% confidence interval of prediction in test phase for all gages, and (f) Residual plots of test phase for (a) to (d). 112

3.11 The plot of actual versus predicted annual flow volume for Model 4. (a) Weber River near Oakley, (b) Chalk Creek at Coalville, (c) Sevier River at Hatch, (d) Muddy Creek near Emery, (e) 90% confidence interval of prediction in test phase for all gages, and (f) Residual plots of test phase for (a) to (d). 114

3.12 Comparison between MVRVM, SVM and ANN based on RMSE in test phase. 116

3.13 The bootstrap analysis for the best models. First column is for Weber River near Oakley, second column is for Chalk Creek at Coalville, third column is for Sevier River at Hatch, and fourth column is for Muddy Creek. Similarly first row is for Model 1, second row is for Model 2, third row is for Model 3, and fourth row is for Model respectively. 117

4.1 The Great Salt Lake. 149

4.2 ε- sensitive loss function in Support Vector Machine. 150

4.3 Average Mutual Information versus Time lag for GSL water level. 151

4.4 Estimating embedding dimension from False Nearest Neighbor method. 151

4.5 The prediction of GSL water level from SVM. For (a)-(d), first column shows training phase, second column shows testing phase, third column shows predicted versus actual elevation for training phase, and fourth column shows similar plot for test phase. (a) Southern arm of the lake for 1982 to 1987, (b) Southern arm of lake for 1991 to 2008, (c) Northern arm of the lake for 1982 to 1987, (d) Northern arm of lake for 1991 to 2008, and (e) Residual plots for (a) to (d). 152

4.6 RMSE versus embedding dimension for southern arm of lake for (a) 1982 to 1987, (b) 1991 to 2008, and northern arm of the lake for (c) 1982 to 1987, and (d) 1991 to 2008. 153

4.7 Efficiency versus embedding dimension for southern arm of the lake for (a) 1982 to 1987, (b) 1991 to 2008, and northern arm of the lake for (c) 1982 to 1987, (d) 1991 to 2008. 154

4.8 The prediction of GSL water level from MVRVM. For (a)-(d), first column shows training phase, second column shows testing phase, third column shows predicted versus actual elevation for training phase, and fourth column shows similar plot for test phase. (a) Southern arm of the lake for 1982 to 1987, (b) Southern arm of lake for 1991 to 2008, (c) Northern arm of the lake for 1982 to 1987, (d) Northern arm of lake for 1991 to 2008, (e) 90 percent confidence interval for (a) to (d), and (f) Residual plots for (a) to (d). 155

4.9 RMSE versus embedding dimension for southern arm of lake for (a) 1982 to 1987, (b) 1991 to 2008, and northern arm of the lake for (c) 1982 to 1987, and (d) 1991 to 2008. 157

4.10 Efficiency versus embedding dimension for southern arm of the lake using MVRVM model for (a) 1982 to 1987, (b) 1991 to 2008, and northern arm of the lake for (c) 1982 to 1987, (b) 1991 to 2008. 158

4.11 Bootstrap analysis for RMSE and Efficiency for SVM model for 1991-2008 (a) southern arm of the lake, and (b) northern arm of the lake. 159

4.12 Bootstrap analysis for RMSE and Efficiency using MVRVM for 1991-2008 (a) southern arm of the lake, and (b) northern arm of the lake. 160

A.1 GSL elevation for south and north arm for 1982-1987. 170

A.2 GSL elevation for south and north arm for 1991-2008. 171

A.3 Autocorrelation and partial autocorrelation function for GSL elevation for each arm of the lake for two time periods: 1982-1987 and 1991-2008. 172

A.4 Variance trend for GSL elevation time series, (a) North arm for 1982-1987, (b) North arm 1991-2008, (c) South arm 1982-1987, and (d) South arm 1991-2008. 173

A.5 Histogram of GSL elevation for 1982-1987: a) southern arm, b) northern arm, and for 1991-2008: c) southern arm, d) northern arm. 174

CHAPTER 1

INTRODUCTION

Water demand is increasing in the global scale with the rapid growth of population that the World is undergoing. Some places are already experiencing mounting water stress (UNHDR 2006). There is increasing pressure to meet the water demands from the available resources for now and in the future. This issue necessitates more a comprehensive approach to water management in order to meet the surging demand for this limited resource. Long lead-time prediction of streamflow may provide information about future water availability which could, in turn, help water managers to plan effectively in order to maximize the efficiency of water use.

Predicting streamflow for long lead-time is a challenging task for hydrologists. Multitude of factors influences the flow in the stream. The states of the basin, local and regional climatic conditions are major controlling factors. Streamflow is a part of the hydrological cycle which is highly controlled by the climatic variability in the local region and around the world. The climatic variability is further connected to the oceanic-atmospheric interaction. The annual, inter-annual and inter-decadal climatic variability characteristic of climate makes the analysis even more complex. The teleconnection between climate and the ocean-atmospheric interaction/oscillations is the scientific basis of long lead-time streamflow prediction. Their correlation provides the forecast opportunity. This dissertation identifies influential climate indicators for a number of selected stream gages in Utah and develops a long lead-time streamflow prediction model that uses those indicators in order to accurately predict the seasonal and annual streamflow at each selected gage. Use of the correct climate indicator for the given stream gage appropriately captures the effect of ocean-atmospheric interaction for that gage, a fact that eventually improves the predictive ability of the model.

In addition to streamflow prediction, this dissertation also develops a model to predict the water surface elevation of the Great Salt Lake (GSL) for multiple lead-times using past water surface elevation data. This information can be helpful for water managers and other GSL stakeholders for planning purposes in order to reduce the impact of rising lake elevation in the surrounding area. Each model is ensured to be robust and well generalized for future change in data trends, thus making the models reliable for long-term predictions.

Physically-based models are based on the understanding underlying physics of hydrologic processes. These models apply the principles of physics in the form of mathematical equations to specific hydrologic situations. The physically-based model approach has obvious limitations: the physiographic and geomorphic characteristics of most hydrologic systems are so complicated and variable, and the degree of uncertainty in the boundary conditions are so large, that the solutions are feasible only for certain highly simplified situations (Brutsaert 2005). There are quite a few physically-based models developed to understand the behavior of water resources systems. The complexities in these models and difficulties associated with the data acquisitions and corresponding expenses that these models would require has limited the application of such models.

To overcome the limitation of physically based models, data-driven models based on the machine-learning approach is used as an alternative model. These models are gaining increasing popularity in the hydrological modeling community. They bridge the gap between the physically-based models and classical regression models. The classical regression model typically assumes a specific mathematical form with a certain number of parameters. Obtaining the required parameters is based on minimizing the discrepancy between the observed value and model prediction. The assumption of a specific mathematical form may not always correctly represent the input-output relationship in the hydrological model, which is a major drawback of classical regression model. This limitation is overcome by data-driven models that use the machine learning approach. Similar to the human brain, data-driven models are capable of learning from previous experiences. They are characterized by their ability to quickly capture the behavior of the system by relating input and output. They are robust and are capable of making reasonable prediction using historical data (Khalil et al. 2006). They provide potentially valuable methods for reducing the cost of data collection and modeling complex river basin systems in support of water management needs without losing accuracy (Velickov and Solmantine 2000). Use of the machine learning approach also eliminates knowledge acquisition time that would be required for the development of physically based models.

Artificial Neural Network (ANN), Support Vector Machine (SVM), and Relevance Vector Machine (RVM) are some popular machine learning models. The ANN model is capable of understanding the complex nonlinear relationship between inputs and outputs. They usually perform well even if the training data contains noise (Hammerstrom 1993), however, they are not free of limitations. An incorrect network definition may lead to over-fitting. The optimization may converge to local optima rather than global optima (Asefa 2004). Some of the limitations of ANN are overcome by SVM. These are very specific class of algorithms, characterized by usage of kernels, absence of local minima, and sparseness of the solution (Vapnik 1995, 1998). SVM presents the solution by means of a small subset of training points which gives enormous computational advantages over ANNs. However, the number of support vectors typically grows linearly with the size of training data, which may require a large amount of computer memory storage. The prediction from SVM is not probabilistic and optimizing the model parameters needs more data and time for cross validation. RVM is based on sparse Bayesian learning. This is a model of identical function form to SVM. RVM makes prediction using only a small number of relevant data points which are automatically selected from large initial set. RVM does not suffer from any of the above limitations of SVM (Tipping 2001). In last few years, RVM has been widely used in modeling water resources management problems. This is a parsimonious and robust model capable of reasonably accurate predictions from small data sets (Khalil et al. 2006; Ticlavilca 2010). RVM is also capable of estimating the uncertainty of prediction (Ghosh and Mujumdar 2008; Khalil et al. 2006), which is a major advantage over other machine learning models, such as ANN and SVM (Tipping 2000, 2001).

This dissertation uses the Multivariate Relevance Vector Machine (MVRVM) and SVM model for predicting hydrological quantities in basin scale. The MVRVM was developed by Thayananthan (2005) as an extension of the RVM algorithm developed by Tipping and Faul (2003). It retains all properties of conventional RVM, such as sparse modeling, high predictive accuracy, and estimation of uncertainty in the prediction.

This dissertation contains a total of five chapters including an Introduction and a Summary chapter. Two chapters consist of using MVRVM to develop the prediction model, while one chapter consists of using both MVRVM and SVM. These are described briefly below.

Chapter 2 presents the application of the MVRVM model on identifying the influential locations of sea surface temperature (SST) for each selected stream gage in the state of Utah, and predicting the streamflow for next six months using appropriate sea surface temperature locations and other local inputs. The effect of regional meteorological condition in streamflow is incorporated through the use of SST data. SST also represents an atmospheric circulation indicator. The local inputs used in the model consists of past streamflow data, snowpack in the mountains, and local meteorological conditions. The stream gages are selected in such a way that they spatially cover the entire state of Utah from North to South. The streamflow at each gage is predicted in the form of monthly average discharge as well as total volume of water passing the gage for next six months. The results show that the MVRVM model is capable of learning from the existing input-output relationship and predict accurately for a new set of inputs. The uncertainty of the prediction is estimated and shown by confidence intervals in a test phase. Bootstrap analysis is used to test the robustness of the model. This analysis presents the estimate of measure of variability of test statistics with the change in training data. Narrow confidence bound indicates the model is robust over the variability in the input data. The results show the model is robust and well generalized.

In Chapter 3, the MVRVM model is used to predict annual streamflow volume using oceanic-atmospheric oscillation indices at four unimpaired stream gages in Utah that spatially covers the entire state from North to South. The oscillation indices are connected to climatic variability in the region around the globe which is eventually connected to the hydrologic cycle. The teleconnection between climate and the oscillation indices is the scientific basis of long lead-time streamflow prediction. The best combination of oscillation indices and the lead time is identified for each selected stream gage which is, then, used to develop the forecast model. The best combinations of oscillations are also identified for each lead time. This information can be useful to improve the accuracy of the prediction. The test-phase results show the model is capable of learning the relationship between inputs and output and make the prediction reasonably well. The bootstrap analysis shows the model is robust and well generalized.

Chapter 4 presents the application of MVRVM and SVM model to predict the water surface elevation of the Great Salt Lake (GSL), Utah, using only past water surface elevation data. This consists of constructing multivariate input space in which the dynamics unfold by creating a vector of multi-dimension out of a single variable (water surface elevation data). The parameters for constructing the state space are estimated for the GSL elevation. This multivariate input space is used to predict the water surface elevation of the lake at bi-weekly time steps. The test results show that both SVM and MVRVM are able to extract the dynamics using only few observed past water surface elevations out of the training examples. The predictions from SVM and MVRVM in their corresponding test phases are fairly accurate and comparable. An optimum combination of reconstruction dimension and time delay is estimated for the model development which may be used as final prediction model for GSL water surface elevation. MVRVM estimates the uncertainty and presents in the form of confidence interval of the prediction while SVM predicts only the mean value. The narrow bound found through a bootstrap analysis shows the model is well generalized (robust).

References

Asefa, T. (2004). "Statistical learning theory: Concepts and application in water resource management." PhD Dissertation, Utah State University, Logan, UT.

Brutsaert, W. (2005). *Hydrology: An introduction,* Cambridge University Press, New York.

Ghosh, S., and Mujumdar, P. P. (2008). "Statistical downscaling of GCM simulations to streamflow using relevance vector machine." *Adv. Water Resour.,* 31(1), 132-146.

Hammerstrom, D. (1993). "Working with neural networks." *Spectrum*, IEEE, 30(7), 46-53.

Khalil, A. F., McKee, M., Kemblowski, M., Asefa, T., and Bastidas, L. (2006). "Multiobjective analysis of chaotic dynamic systems with sparse learning machines." *Adv. Water Resour*., 29(1), 72-88.

Thayananthan, A. (2005). "Template-based pose estimation and tracking of 3D hand motion." PhD Dissertation, University of Cambridge, Cambridge, UK.

Ticlavilca, A. (2010). "Multivariate Bayesian machine learning regression for operation and management of multiple reservoir, irrigation canal, and river systems." PhD Dissertation, Utah State University, Logan, UT.

Tipping, M. (2000). "The Relevance Vector Machine." *Proc., Advances in Neural Information Processing Systems*, The MIT Press, 652-658.

Tipping, M. (2001). "Sparse Bayesian learning and the Relevance Vector Machine." *J. Machine Learning Res.*, 1, 211-244.

Tipping, M. E., and Faul, A. C. (2003). "Fast marginal likelihood maximization for sparse Bayesian models." *Proc., Ninth International Workshop on Artificial Intelligence and Statistics.*

United Nations Human Development Report. (2006). "Beyond scarcity: Power, poverty and the global water crisis." New York.

Vapnik, V. N. (1995). *The nature of statistical learning theory,* Springer Verlag, New York.

Vapnik, V. N. (1998). *The nature of statistical learning theory,* Springer Verlag, New York.

Velickov, S., and Solmantine, D. P. (2000). "Predictive data mining: Practice examples." *Proc., 2nd Joint Workshop on AI methods in Civil Engineering Applications*, 3-19.

CHAPTER 2

BAYESIAN MACHINE LEARNING REGRESSION APPROACH FOR IDENTIFICATION OF INFLUENTIAL SEA SURFACE TEMPERATURE LOCATIONS AND PREDICTING STREAMFLOW FOR THE NEXT SIX MONTHS

Abstract

Sea surface temperature (SST) has significant influence in the hydrological cycle which eventually affects the discharge in streams. SST is an atmospheric circulation indicator which provides the predictive information about the hydrologic variability in regions around the world. Choosing the right location of sea surface temperature for the prediction of streamflow at a specific location of the gage is crucial. The use of the correct location of SST for the selected stream gage appropriately captures the effect of oceanic-atmospheric interaction, which eventually improves the predictive ability of the model. The strength of the effect of SST changes spatially, thus, the influential locations of SST for different locations of the stream gages will be different. This chapter aims on identifying appropriate locations of sea surface temperature at selected stream gage in the state of Utah that spatially covers the state from South to North, and use them for long-term streamflow prediction. Analysis shows the influence of Pacific Ocean SST to be stronger than that of Atlantic Ocean SST in the state of Utah. Using appropriate location of SST, an accurate and reliable long-term streamflow can be predicted, which may play an important role on water resources planning and management in the river basin scale. This information provides how much water will be available in the next season so that the

water managers, stakeholders and farmers can plan accordingly. Predicting future water availability accurately and reliably is a key step for successful water resource management in arid regions. A data-driven model derived from statistical learning theory is used in this chapter. This model relates input/output without trying to understand the underlying physical process. The model used in this chapter is developed in the form of a Multivariate Relevance Vector Machine (MVRVM). Using the best identified SST locations, along with local climatic condition and the current state of basin, both monthly mean discharge and volume of water passing the gage for the next six months are predicted. Monthly mean discharge is usually predicted best by the North Pacific SST for Northern and Central Utah, while the Tropical Pacific SST develops the best model for Southern Utah. For volumetric prediction, the North Pacific SST develops the best model prediction in most of the selected stream gages in the state of Utah. Each model is demonstrated to be robust by the results of a bootstrap analysis.

2.1 Introduction

Monthly and annual streamflow series are strongly related to long-term climate (Sivakumar 2003). Researches in the atmospheric and hydrologic sciences have recently used sea surface temperature (SST) in an attempt to predict streamflow variability. SST represents oceanic-atmospheric circulation which has important consequence on the weather around the globe. SST provides predictive information about the hydrologic variability in regions around the globe (Tootle and Piechota 2006). This has a strong link with the hydrology of individual river basins. The identification of an appropriate location of SST will be helpful to improve the predictive ability of the model developed herein. This research identifies the best locations of SST for the selected locations of four unimpaired stream gages, and one impaired one, and uses them for long lead-time streamflow prediction.

Streamflow is predicted for the next six months using the SST for the best identified locations in the Pacific and Atlantic Oceans, past streamflow data, snow pack in the mountains, and local climatic condition. Selections of other inputs are based on the understanding of the underlying physical processes. Predictions are made for two scenarios. The first one predicts the monthly average discharge for the next six months, while the next one predicts the total volume of water passing the gage for the next six months. Streamflow predicted using the best identified SST location is more accurate and reliable than using other SST locations because use of right location of SST appropriately captures the effect of ocean-atmospheric interaction for the corresponding stream gage. Precise information about quantity of water availability in next season could be quite useful for agricultural planning, watershed management, and other decision making processes. It can benefit the management of water resources, in particular allowing decision on water allocation for irrigations and other purposes. Financial commitment made by the farmers early in the season can result in substantial economic losses if the resulting seasonal flow does not subsequently supply enough irrigation water. Forecast with long-lead time facilitates co-ordination between different system users that may be important in multiple-use water resource systems (Hamlet and Lettenmaier 1999).

In the present study, inputs are transformed into the output (streamflow) using a Bayesian machine-learning regression model. This is used as a simpler, less costly alternative to physically-based models. The complexities in the physically based models and difficulties associate with their data acquisitions and corresponding expenses has limited the application of such models. Machine learning models are good on capturing the underlying physics of the system by relating input and output through robust mathematical relations. Machine-learning models are robust and capable of making reasonable prediction using historical data (Khalil et al. 2006). Artificial Neural Network (ANN), Support Vector Machine (SVM) and Relevance Vector Machine (RVM) are some of the most popular machine learning models. ANN has the disadvantage that it may get stuck in local minima rather than global minima. SVM is a very popular machine learning model, however, it makes unnecessary liberal use of the basis function. In SVM, the number of support vector linearly grows with the size of training data (Tipping 2001) and the prediction is not also probabilistic. Moreover, optimizing more than two model parameters in SVM requires additional data and time for cross validations. RVM is a Bayesian machine learning model. This is sparser than SVM and gives probabilistic output as well. Optimizing model parameter for RVM is relatively easier than for SVM, however, the performance of their predictions is comparable. RVM has been successfully used by many past researches for water resources operation and management, e.g. (Khalil et al. 2005b; Ticlavilca 2010). Multivariate Relevance Vector Machine (MVRVM) is proposed in this chapter, which is developed by Thayananthan (2005), as an extension of the RVM algorithm developed by Tipping and Faul (2003).

2.2 Study Area

Five stream gages are chosen at different locations of Utah that spatially cover the state from its Northern to its Southern region. Certain data assumptions are made for the site selection, namely: (i) site flows are not affected by diversion or regulation; and (ii) several years of systematic record are available. These data assumptions are valid for all sites except for one. Two sites were chosen from the Northern region of Utah, two from the Central region, and one from the Southern region of the state. The station at the Weber River near Oakley and that at Chalk Creek at Coalville are in the Northern region of the state. They both lie in the Weber River Basin, which is composed of a flat, fertile valley east of the Great Salt Lake. The watershed contains approximately 2060 square miles. Average annual precipitation in the Weber River Basin ranges from 12 to 30 inches. Snow accumulation and melt are very significant features in terms of annual hydrologic cycle for this watershed (Perica and Stayner 2004).

The station at the Sevier River at Hatch is chosen for the Southern region of the state. It lies in the Sevier River Basin. The river flows north from its headwaters and then turns southwest 255 miles before reaching Sevier Lake (Berger et al. 2003). This river basin consists of 12.5 percent of the total area of the State of Utah. Average annual precipitation is close to 13 inches. The major source of surface water for Sevier River comes from snowmelt, which is available during the spring and early summer months. The primary use of water in the basin is for irrigation (Berger et al. 2003).

The other two river gages are in the Central region of Utah. They are the station at Muddy Creek near Emery, and the station at Sixth Water Creek above the Syar Tunnel near Springville. Muddy Creek near Emery is in West Colorado River Basin. This creek drains portion of Emery and Wayne Counties in Central Utah. Muddy Creek begins on the eastern slopes of the Wasatch Plateau. It turns southward near the town of Emery, and then flows along the western edge of the San Rafael Swell. It has an estimated length of

20 miles and a drop of 6000 feet before it combines with the Fremont River to form the Dirty Devil River (McCord 1997).

Sixth Water Creek lies in the Utah Lake Basin. It is about 1 mile long. The flow in the Sixth Water Creek near Springville is partly affected by the diversion from Strawberry Divide until 2004. Figure 2.1 shows the location of the selected stream gages in Utah. The geometric characteristic of each stream gage is shown in Table 2.1.

2.3 Background

Accurate and reliable prediction of streamflow is crucial for water resource planning and management. If the appropriate input variables responsible for the generation of streamflow are used in the model, the accuracy of the prediction improves, and uncertainty reduces, even if the model is data-based, rather than physically-based. This research develops a predictive data-based model using precisely identified input data. The input data used in the model are based on the understanding of the physical processes and climatic factors that affect the discharge in the stream. Streamflow depends not only on the distribution of precipitation in time and in space, but also on the type and the state of the basin, which, in turn, depends on the climate condition. Therefore, input of climatic conditions in the model through the use of sea surface temperature (SST) has significant importance.

SST is an important variable that affects long-term streamflow. This is an atmospheric circulation indicator used to represent the effect of regional climatic conditions in local hydrology of a river basin. Use of SST at appropriate locations in the nearby oceans improves the accuracy of the prediction. This research explores the influential locations of SST for the selected stream gages in the state of Utah that covers the state from South to North, and uses them along with other local inputs for predicting streamflow for the next six months. This is useful information for developing an accurate forecast model for a given stream gage location. The long-term streamflow prediction is crucial information for the water managers, farmers and stakeholders of river basins, especially those located in the arid regions. This information helps water users to plan their water allocations appropriately for the upcoming water season. Long-term streamflow predictions also reduce the risk associated with the financial commitment that needs to be made at the beginning of the season by providing accurate water availability beforehand.

A data-driven model based on the learning machine approach was chosen in this research paper. Other researchers have used similar approaches for predicting hydrologic phenomena. For example, Asefa et al. (2006) predicted multi-time scale streamflow using Support Vector Machine. This paper consists of using a Bayesian machine learning regression approach that uses Multivariate Relevance Vector Machine for the non-linear transformation of readily-available input data to predict streamflow for the next six months.

2.4 Model Description

Multivariate Relevance Vector Machine (MVRVM) (Thayananthan 2005) is proposed. This is a supervised learning model based on sparse Bayesian learning. This is a model of identical functional form to the Support Vector Machine developed by Vapnik

(1995, 1998). MVRVM model is an extension of sparse Bayesian model developed by Tipping and Faul (2003).

For the given input-target pair $\{x_n, t_n\}_{n=1}^{N}$ in training data set, the model learns the dependency of the targets on the inputs with the objective of making accurate predictions of the target (*t*) for previously unseen values of input *x* (Tipping 2000, 2001).

The targets are assumed to be samples from the model (*y*) with additive noise (ε). The target can be written as sum of approximation vector, $y = [y(x_1),\text{..........}y(x_N)]^T$ and the error vector $\varepsilon = (\varepsilon_1,\text{........}\varepsilon_N)^T$ which is independent samples from some noise process. The noise is assumed to be mean-zero Gaussian with variance σ^2. The target vector is written as,

$$t = y + \varepsilon,$$

$$= \Phi w + \varepsilon. \tag{2.1}$$

The target vector can be written as, $t = (t_1\text{.......}t_N)^T$. The weight vector (*w*) is expressed as, $w = (w_1,\text{...}w_i\text{.....}w_N)^T$ and Φ is the design matrix of size N*(N+1). This is given by $\Phi = [\phi(x_1)\text{.....}\phi(x_N)]^T$, wherein $\phi(x)$ is basis function. The basis function is given by,

$$\phi(x_n) = [1, K(x_n, x_1), K(x_n, x_2),\text{.....}, K(x_{n,}, x_N)]^T.$$

The target t_n is assumed to be independent so the likelihood of complete dataset is written as,

$$p(t|w,\sigma^2) = (2\pi\sigma^2)^{-N/2}\exp\{\frac{1}{2\sigma^2}\|t - \Phi w\|^2\}. \tag{2.2}$$

Let τ_i be the i^{th} component of the target vector *t*, and w_i be the weight vector for the i^{th} component of the output target vector *t* such that, $w = (w_1, ...w_i .....w_N)^T$. This is Gaussian distribution which can be written as,

$$p(t|w, \sigma^2) = \prod_{i=1}^{n} N(\tau_i | \Phi w_i, \sigma_i{}^2).$$

To avoid overfitting, Tipping (2001) imposed some additional constraints defining an explicit prior probability distribution over them. This prior ultimately leads the sparsity of the model. The prior probability is given by,

$$p(w|\alpha) = \prod_{i=0}^{N} N(w_i | 0, \alpha_i{}^{-1}), \tag{2.3}$$

where $\alpha = (\alpha_0, ...\alpha_N)^T$ is a vector of N+1 hyper-parameters. Each α_i controls the strength of the prior over its associated weight (Tipping and Faul 2003). Bayes' rule is used for obtaining the posterior over the weight. Given the data, the posterior distribution over the weights is Gaussian which is given by (Tipping 2001),

$$p(w|t, \alpha, \sigma^2) = \frac{p(t|w, \sigma^2).p(w|\alpha)}{p(t|\alpha, \sigma^2)},$$

$$= (2\pi)^{-(N+1)/2} |\Sigma \; |^{-1/2} . \exp\{-\frac{1}{2}(w-\mu)^T \Sigma^{-1}(w-\mu)\}, \tag{2.4}$$

$$= \prod_{i=1}^{N} N(w_i | \mu_i, \Sigma_r).$$

The posterior covariance and mean of the weight are $\Sigma = (\sigma^{-2}\Phi^T\Phi + A)^{-1}$ and $\mu = \sigma^{-2}\Sigma\Phi^T t$ respectively where $A = diag(\alpha_0, \alpha_1, ..., \alpha_N)$.

Some approximation is adapted on the hyper-parameter posterior by a delta function at its mode, i.e., at its most probable values $\alpha_{MP}, \sigma^2_{MP}$ (Tipping 2001),

$$\int p(t_* | \alpha, \sigma^2) \delta(\alpha_{MP}, \sigma^2_{MP}) d\alpha d\sigma^2 \approx \int p(t_* | \alpha, \sigma^2) p(\alpha, \sigma^2 | t) d\alpha d\sigma^2 \ . \qquad (2.5)$$

The learning then becomes the search for the hyper-parameter posterior mode, i.e. the maximization of $p(\alpha, \sigma^2 | t) \propto p(t | \alpha, \sigma^2) p(\alpha) p(\sigma^2)$ with respect to α and σ^2. For uniform hyperpriors over logα and log*σ*, $p(\alpha, \sigma^2 | t) \propto p(t | \alpha, \sigma^2)$, which is further given by,

$$p(t | \alpha, \sigma^2) = \int p(t | w, \sigma^2) p(w | \alpha) dw,$$

$$= (2\pi)^{-N/2} \left| \sigma^2 I + \Phi A^{-1} \Phi^T \right|^{-1/2} \exp\{-\frac{1}{2} t^T (\sigma^2 I + \Phi A^{-1} \Phi^T)^{-1} t\}. \qquad (2.6)$$

In Bayesian models, this quantity is known as the marginal likelihood, and its maximization is known as the type-II maximum likelihood method (Berger 1993). Eq. 2.6 is solved by iterative re-estimation which gives,

$$\alpha_i^{new} = \frac{\gamma_i}{\mu_i^2} \ , \qquad (2.7)$$

where $\gamma_i = 1 - \alpha_i N_{ii}$.

The term μ_i is the *i*th posterior mean weight and N is the number of data examples. N_{ii} is the *i*th diagonal element of the posterior weight covariance computed with the current α and σ^2.

The noise variance is re-estimated from,

$$(\sigma^2)^{new} = \frac{\|t - \Phi\mu\|^2}{N - \sum_i \gamma_i}. \tag{2.8}$$

The learning algorithm proceeds by repeated application of (2.7) to (2.8), together with updating the posterior statistics Σ and μ until some specified convergence criteria is satisfied. It is found that value of α_i generally approaches to infinity which implies that $p(w_i|t,\alpha,\sigma^2)$ becomes highly peaked at zero which makes the model sparse. The relatively nonzero weights correspond to the input vectors that form the sparse core of the RVM model. These input vectors are called relevance vectors (RVs). This sparsity is an effective method to control model complexity, avoid over-fitting and control computational characteristics of model performance (Tipping and Faul 2003).

The predictions is made based on the posterior distribution over the weights, conditioned on the maximizing values α_{MP} and σ^2_{MP}. The predictive distribution for a new input x_* is given by,

$$p(t_*|t,\alpha_{MP},\sigma^2_{MP}) = \int p(t_*|w,\sigma^2_{MP})\, p(w|t,\alpha_{MP},\sigma^2_{MP})\, dw. \tag{2.9}$$

This is easily computable because both terms in the integral are Gaussian,

$$p(t_*|t,\alpha_{MP},\sigma^2_{MP}) = N(t_*|y_*,\sigma^2_*), \tag{2.10}$$

with,

$$y_* = \mu^T \phi(x_*),$$

$$\sigma^2_* = \sigma^2_{MP} + \phi(x_*)^T \Sigma \phi(x_*).$$

The total variance consists of sum of the variance of data and uncertainty in estimating weight. Interested readers for Relevance Vector Machine are referred to

Tipping (2000, 2001), Tipping and Faul (2003), Thayananthan (2005), and Thayananthan et al. (2008).

2.5 Data Collection and Description

The input variables are selected based on the underlying physical processes and climatic factors that influence the generation of streamflow and their relevancy is judged subjectively. The variables used in the model are described below.

2.5.1 Streamflow

Streamflow is a consequence of interaction of hydrologic events. Some examples of these events are precipitation, snow melt, evapotranspiration, etc. The historical streamflow data were collected in the form of monthly mean discharge from 1980 to 2009 from the U.S. Geological Survey (USGS).

2.5.2 Snow water equivalent

When the precipitation falls as snow, it settles, compact and melts several months later, and is a prominent source of streamflow (Soukup et al. 2009). Snow serves as storage of water, especially in the western United States, and has major effect on the streamflow in the spring and early summer months. The snow water equivalent (SWE) is defined as the equivalent depth of water when snow completely melts. The SWE data were obtained from the Natural Resources Conservation Service (NRCS) (http: //www.wcc.nrcs.usda.gov/snow). The period of 1980-2009 was used in this study because of the relative completeness of data in the selected basins for these years.

Using SWE measurements from different SnoTel stations improves prediction compared to the one that uses a single station, in some sense incorporating SWE spatial variability (Asefa et al. 2006). The Harris Flat and Midway Valley SnoTel stations are used for the Sevier River gage station at Hatch. The Smith and Morehouse and Chalk#1 SnoTel stations are used for the Weber River gage near Oakley. The Chalk#1 and Chalk#2 SnoTel stations are used for the Chalk Creek gage at Coalville. The Buck Flat and Dill's Camp SnoTel stations are used for the Muddy Creek gage near Emery. Finally, the Strawberry Divide SnoTel station is used for the Sixth Water Creek gage above Syar tunnel near Springville. Although some SnoTel sites are physically outside of the watershed, they are still included in the model due to their strong relationship with the nearby streamflow processes.

2.5.3 Local temperature

Temperature controls the melting rate of snow which consequently affects the discharge in the stream. The high discharge in the spring and early summer month is due to rising temperature provided that there is enough snowpack in the watershed. The temperature data is also collected from the SnoTel stations operated by NRCS. The period of data collection for local temperature is same as that of SWE.

2.5.4 Sea surface temperature

Sea surface temperature is an important input for long-term streamflow forecasting. This is considered in the present study as atmospheric circulation indicator. The use of SST data over long temporal range can be relevant to the study of basin scale water management issues (Khalil et al. 2005a). It provides important predictive information about hydrologic variability in the regions around the world (Tootle and Piechota 2006). This is a regional meteorological indicator appealing to water managers and forecasters.

The Kaplan sea surface temperature anomaly (SSTA) and the Smith and Reynolds SST are used in this paper. The Kaplan SST covers the majority of the world's oceans with a 5° by 5° grid (Kaplan et al. 1998), while Smith and Reynolds SST covers the majority of the world's oceans with a 2° by 2° grid (Smith and Reynolds 2003). The extended reconstructed global SSTs for Smith and Reynolds SST are based on the comprehensive Ocean-Atmosphere data set from 1854 to present. Six locations, identified as North Pacific (NP), Central Pacific (CP), Tropical Pacific (TP), East Atlantic (EA), Middle Atlantic (MA), and Tropical Atlantic (TA), are selected from the Pacific and Atlantic oceans. Their spatial cover is from Tropical Pacific to North Pacific, and Tropical Atlantic to East Atlantic (Figure 2.2).

2.6 Model Development

The approach applied here in building a model for long-term streamflow prediction is based on a data driven model that uses Multivariate Relevance Vector Machine. This is a Bayesian regression tool extension of the RVM algorithm developed by Tipping and Faul (2003). The model requires the identification of predictor variables (input) and response (output vector). The data needs preprocessing in a way that is suitable for modeling. The model requires the selection of kernel and kernel parameter. The selection of a kernel is heuristic and the Gaussian kernel is used in all combination of data set in order to make the uniform comparison. The right value of kernel width for the given input-output set is obtained from an optimization process, which is actually done by testing the kernel width over a wide range for each combination of input set.

Two models are proposed. Model 1 consists of predicting the monthly mean discharge for the next six months, while Model 2 consists of predicting the volume of water passing the stream gage for the next six months. Inputs to the model consists of past streamflow data, snow water equivalent, local temperature, and sea surface temperature. SST is collected from six different locations of Pacific and Atlantic Ocean. Initially six RVM models are developed using one individual SST at a time. Figure 2.3 shows the combinations of input variables that are initially used to create input file for each stream gage.

In order to improve the test statistics, SST of different locations are combined with a one that develops the best test statistics when using one individual SST at a time. For example, if NP SST produces the best test result among other individual SST locations, the combination of SSTs is then developed with NP SST. This process is repeated for each selected stream gage. The test statistics are computed for each combination of input set and best location of SST is identified for the given stream gage by comparing their test statistics. Figure 2.4 shows example flowcharts of inputs for Model 2. Similar flowcharts are prepared for Model 1.

2.6.1 Model 1

This model predicts the monthly mean discharge for the next six months. The input to the model consists of past stream discharge, SWE, local temperature, and sea surface temperatures. SWE and local temperature input are in the form of monthly mean values. Similarly, monthly values of sea surface temperature are used in the model. The model can be mathematically expressed as,

$$Q_t = f(Q_{t-6}, \overline{S}_{t-12}, \overline{T}_{t-12}, \overline{SST}_{t-12}), \tag{2.11}$$

where Q_{t-6} is the monthly mean discharge (cfs) passing through a stream gage six months prior to time *t*, $\overline{S}_{t-12}$ and $\overline{T}_{t-12}$ are the monthly average SWE (in) and local temperature (ºC) respectively twelve months prior to time *t*, $\overline{SST}_{t-12}$ represents monthly average sea surface temperature twelve month prior to time *t*. The function f is a nonlinear RVM transformation of inputs to output. The output is monthly mean discharge predicted at time *t* which is six months ahead monthly mean discharge. If the local temperature is not included in the model, it is expressed as,

$$Q_t = f(Q_{t-6}, \overline{S}_{t-12}, \overline{SST}_{t-12}). \tag{2.12}$$

2.6.2 Model 2

This model predicts the volume of water passing the stream gage for the next six months. The model for the volumetric prediction is similar to that of Model 1, however, the variable notations have different standings. The input to the model consists of past streamflow volume, SWE, local temperature, and sea surface temperature. Streamflow input in the model is in the form of total volume of water passing through the stream gage in the last six months. The monthly mean discharge obtained from USGS is converted into total volume using appropriate conversion factor. SWE and local temperature input are in the form of average of monthly mean values of last 12 months. The sea surface temperature input to the model is the average of monthly average values of last 12 months. The model can be mathematically expressed as,

$$Q_{t+6} = f(Q_{t-6}, \overline{S}_{t-12}, \overline{T}_{t-12}, \overline{SST}_{t-12}) \tag{2.13}$$

where Q_{t-6} is a total volume of water flowing through the gage in the last six months, $\overline{S}_{t-12}$ and $\overline{T}_{t-12}$ is the average SWE and local temperature computed over the last twelve months, $\overline{SST}_{t-12}$ represents average sea surface temperature value of last 12 months. The output Q_{t+6} is the volume of water passing the stream gage for the next six months. If the local temperature is not included in the model, it can be written as,

$$Q_{t+6} = f(Q_{t-6}, \overline{S}_{t-12}, \overline{SST}_{t-12}). \tag{2.14}$$

2.6.3 Performance Criteria

The statistical measures that are used in this paper for the performance evaluation of the model are root mean square error (RMSE) and Nash-Sutcliffe efficiency.

Root Mean Square Error (RMSE)

The smaller the RMSE value, the better the prediction result is. The ideal value of RMSE is zero. Mathematically this is expressed as,

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(t_i - t_i^*)^2}{n}}, \tag{2.15}$$

where t_i is observed value, t_i^* is prediction from the model and n is sample size.

Nash-Sutcliffe Efficiency

The Nash-Sutcliffe efficiency, *E*, is a popular index to evaluate the performance of hydrological models. It is used to measure the predictive power of the hydrological models (Nash and Sutcliffe 1970). Mathematically it is expressed as,

$$E = 1 - \frac{\sum_{i=1}^{n} (t_i - t_i^*)^2}{\sum_{i=1}^{n} (t_i - \bar{t})^2}, \tag{2.16}$$

where $\bar{t}$ is mean observed value. Nash-Sutcliffe efficiencies range from negative infinity to one. An efficiency of one corresponds to a perfect match of model prediction to observed data. An efficiency of zero indicates that the model prediction is as accurate as the mean of the observed data. A negative efficiency indicates that the observed mean is a better predictor than the model.

2.7 Results and Discussion

Both the Smith and Reynolds SST (Smith and Reynolds 2003) and the Kaplan SSTA were initially used in the model. Based on statistical measures, it was found that, in general, the model predictions using Smith and Reynolds SST data were better than that of using Kaplan SSTA. The present research, therefore, uses Smith and Reynolds SST only. It was also found that the use of local temperature improved the model prediction. Therefore, the model that did not include the local temperature was discarded.

2.7.1 Identification of influential SST location and prediction of monthly mean discharge for the next six months using Model 1

The model was trained for the period 1980-2001 and tested on the period 2002-2009 for Weber River near Oakley, Chalk Creek at Coalville, and Muddy Creek near Emery. The Sevier River at Hatch was trained for the period 1982-2001 and tested for the period 2002-2009, while the Sixth Water Creek near Springville was trained for the period 2000 to 2006, and tested for the period 2007 to 2009. The test RMSE of each prediction model, when one individual SST is used at a time, is shown in Figure 2.5 for each stream gage. Similarly, the test RMSE for the combined SSTs for each selected stream gage is shown in Figure 2.6.

Figure 2.7 shows the best locations of the sea surface temperature for each selected stream gage for monthly mean discharge prediction. This is developed by comparing the test statistics (RMSE and/or efficiency) of individual and combined SSTs of different locations (Figure 2.5 and 2.6). It is noticed that the effect of Pacific Ocean SST is more dominant than that of the Atlantic Ocean, which is consistent with other previous studies (Ting and Wang 1997; Wang and Ting 2000). The streamflow in the Northern and Central Utah are best predicted by the sea surface temperature of North Pacific, and in some cases, Central Pacific SST. However in Southern Utah, streamflow is best predicted by the sea surface temperature of the Tropical Pacific, and Tropical Atlantic (when used in combination of SSTs).

A significance test was used to confirm if the test statistics from the best identified SST locations for each selected gage were significantly better than the test statistics of other SST locations. The 95% confidence interval for the median RMSE

(Table 2.2) was computed. The test RMSE value using best-identified SST locations is outside the boundary of 95% confidence interval. This is therefore, said to be significantly better. The summary result of the best model and corresponding SST locations are shown in Table 2.3.

The best identified SST locations are used along with SWE, local temperature, and past streamflow data to develop forecast model for each selected stream gage. The prediction results are shown in Figure 2.8. Results show the predicted streamflow is quite accurate for all selected unregulated stream gage, while it is reasonably accurate for regulated gage. The predicted discharge shows good agreement with the actual discharge. The uncertainty of prediction is captured by the confidence interval in the test phase for each selected gage. High flows are perfectly captured while there is little discripency in low flows. The ground water flow (base flow) is responsible for the generation of low flow in the stream. Since the input representing ground water flow is not included in the model, this level of discrepancy is obvious. The overall prediction quality, however, remains good. Residuals are higher for low flow conditions, but shows randomness for other flows. This is persuasive evidence that the model has no serious modeling problems. The overall performance of the model shows its ability to forecast streamflow accurately for the next six months. This is crucial for reliable water resource planning and management works, especially in arid regions.

The illustration about the selection of best location of SST for each selected locations of stream gage is as follows: When monthly data are used, the SST data consists of seasonal, annual, inter-annual to inter-decadal components. The effect of seasonal component is stronger than other components. The seasonality may be explained by the

El Niño Southern Oscillation (ENSO) effect. The seasonal cross correlation between ENSO index and hydrological record is demonstrated by Poveda et al. (2001). The ENSO may be characterized by the Southern Oscillation Index (SOI) and sea surface temperature in the region: Niño 1+2 (0°-10°S, 80°W-90°W), Niño 3 (5°N-5°S, 90°W-150°W), Niño 4 (5°N-5°S, 60°E-150°W), and Niño 3.4 (5°N-5°S, 120°W-170°W) (Poveda et al. 2001). Niño 3.4 gives overall representation of ENSO (Soukup et al. 2009). Tropical Pacific SST information may be a useful predictor for the U.S. Precipitation for ENSO period (Wang and Ting 2000). Since the streamflow is a consequence of precipitation, it may be also useful predictor for streamflow prediction. The TP SST station chosen in this research (Figure 2.3) lies in ENSO region (Niño 3.4), therefore, TP SST is most responsive input variable where the effect of ENSO is received.

Utah lies in the boundary of ENSO effect (Wang and Ting 2000). The influence of ENSO is dominant in the southern region of Utah. Therefore, TP SST develops best model prediction for stream gage in Southern Utah. However, the streamflow in the Northern and Central Utah are not much influenced by the ENSO effect because of relatively weak ENSO signal in these regions. But, it is more influenced by annual or interannual and low frequency components. The North Pacific atmosphere-ocean climate system has prominent timescales that range from interannual to decadal (Nakamura et al. 1997). NP SST is associated with low frequency variability and has interannual to decadal component. Therefore it is more responsive to streamflow sites in Central and Northern Utah. Also, the principal moisture source for Central and Northern Utah is the Pacific Ocean (Pope and Brough 1996). This moisture is usually moving from west to east toward United States. The latitude of Northern and Central region of the state is close to NP SST station than any other SST station in Pacific Ocean. Therefore, NP SST usually develops best model prediction in most of the stream gages in the central and northern region of the state.

The prediction result for each streamflow site is discussed individually. For Weber River near Oakley, the best prediction model is obtained from CP SST when one individual SST is used at a time. The CP SST is then combined with the other SST locations and best prediction is obtained from the combination of CP and NP SST. It is noticed that the combination of CP with the SST of northern locations improved the prediction result, however, the combination of CP with the SST of the southern locations deteriorate it. This stream gage is in the Northern Utah, therefore, the effect of ENSO is limited. However it is more influenced by annual and interannual to interdecadal components which are best represented by the SST of North Pacific Ocean. Thus, the NP SST has major influence in the streamflow predictions for the Weber River near Oakley. The principal moisture source of this area is Pacific Ocean. The latitude of this streamflow site is similar to the NP and CP SST stations. Considering those factors, it is therefore very obvious to have the best prediction obtained from NP and CP SST for Weber River near Oakley.

Chalk Creek at Coalville lies in the northern Utah. The best model prediction is obtained at EA SST when individual SSTs are used at a time, and combined EA and NP when used in combination. It is found that combination of EA and MA SST did not improve the model prediction but made it worse. The combination of EA and TA SST made the prediction poor again. Similar result is obtained for the combination of EA SST, MA SST and TA SST. The reason behind the poor prediction is due to the addition of irrelevant input variables in learning machine model. This means these SSTs do not have responsive influence and gives insight not to use TP, TA, MA or other south located SSTs for the prediction of streamflow in Chalk Creek at Coalville. This stream gage is out of the range of the strong ENSO effect so there is no strong effect of Tropical Pacific SST. Similarly, the Middle and Tropical Atlantic SSTs are irrelevant for this streamflow site. The site is more influenced by annual, interannual to interdecadal components which are best described by the SST of North Pacific Ocean. The major source of moisture for this area is Pacific Ocean as mentioned earlier. This justifies the combination of NP and EA SST for the best model prediction.

Muddy Creek near Emery lies in Central Utah. The best model prediction is obtained at NP SST for all cases. The Figure 2.6(c) shows the model developed using the combinations with the SST of Northern locations produces a comparable result to that of using NP SST alone, however, the model developed with the combination of SST of southern locations deteriorate the performance of the model. Since this stream gage is outside of ENSO dominant region, there is no strong effect of TP in this site. The streamflow at Muddy Creek near Emery, therefore, is affected more by the low frequency component, and NP comes into strong position in this case. The major source of moisture coming to this area is again from the Pacific Ocean, as indicated above. Therefore NP SST produced best model prediction for this stream gage.

Sevier River at Hatch lies in the Southern Utah. The best prediction is obtained from TP SST when one individual SST is used at a time. Interestingly enough, the Sevier River at Hatch is in the region of known ENSO influence. Therefore, TP SST is most responsive input variable than that of any other SST locations. The combination of TP SST with the CP and NP SST did not improve the model prediction, but made it worse. The combinations of TP with southern locations improved the model prediction, while the combinations with the central and northern locations of SST deteriorate it. This shows the northern and central SST's does not have strong influence for the discharge prediction in Sevier River at Hatch. This is because NP and CP region has low frequency and its effect is insignificant as compare to the seasonal component influenced by ENSO. The combination of TP and TA SST produced the best result. Since both TP SST and TA SST are located on south, their combination performed best for Sevier River at Hatch.

Sixth Water Creek is in Central Utah. The best model prediction is obtained again at NP SST. The explanation is similar as that of the Muddy Creek near Emery, because both sites are located in central region of the state, and are spatially close to each other. This particular site is partly affected by the diversion from the Strawberry Lake until 2004. The input corresponding to the diversion or regulation is not incorporated in the model for this gage, therefore the model prediction for this site is not very accurate as it is in the other unregulated stream gages (Figure 2.8e).

The identification of appropriate location of SST location is crucial for developing accurate long-term forecast of streamflow, which eventually increases the benefit and reduces the risk associate with the future shortage of water.

2.7.2 Identification of influential SST locations and prediction of the volume of water passing the gage for next six months using Model 2

The training and testing period for each stream gage were same as that used in monthly mean discharge prediction. The test statistics (RMSE and efficiency) are computed for each individual SST for volume of water passing through selected stream gage for next six months. The test RMSE for each stream gage when one individual SST is used at a time is shown in Figure 2.9. The test statistics for the combined SSTs for each gage are shown in Figure 2.10. The SST locations that produce best test statistics are shown in Figure 2.11.

Significance tests for the volumetric prediction were conducted in a similar manner as performed for the monthly stream discharge predictions. The 95% confidence interval for the median RMSE is shown in Table 2.4. The test RMSE from the best identified SST locations is outside of the 95% confidence interval. This indicates the test RMSE from the best chosen SST location is significantly better than the test RMSE from other SST locations. The summary result from using appropriate SST locations for each stream gage is shown in Table 2.5.

Using the best identified SST locations, the volume of water passing through each selected stream gages was predicted (Figure 2.12). The results show the model predictions are quite accurate. A good match between actual and predicted flow volume is obtained. The plot of predicted versus actual flow volume shows point saturation around the 45 degree line, which shows the model is good for use as a forecast model for streamflow volume prediction. The accuracy of the prediction is high for the unimpaired gages, while it is relatively less for impaired gage (Sixth Water Creek near Springville).

The input corresponding to the diversion is not incorporated in the model for Sixth Water Creek, which resulted in reduced accuracy of the predictions. In all cases, the model has perfectly captured the high flow, but the low flow is not captured accurately. Since the inputs representing the ground water flow are not included in the model, this level of discrepancy is obvious. Residual plots of six-month streamflow volume predictions are random, as indicated in Figure 2.12g though the residuals are relatively higher for low flow conditions. This is the evidence that the model has no serious modeling problems. The overall prediction shows the model is good and can be used for predicting streamflow volume six months ahead. The uncertainty of prediction is captured by confidence interval in test phase for each gage.

The selection of best SST locations for each selected gage is justified from the reasons explained herein. Since the variables are either cumulative or averaged over time, the seasonal climatic component gets eliminated in the model. The leftover components are annual, internannual to interdecadal components, which are low frequency components. NP SST has low frequency variability which has annual, interannual to decadal components. NP SST therefore, has a stronger influence than other SST locations for most of the streamflow sites in Utah in terms of the volumetric prediction. This includes Chalk Creek at Coalville, Muddy Creek near Emery, Sixth Water Creek near Springville, and Sevier River at Hatch. In the case of monthly mean streamflow predictions, the best prediction was obtained from TP SST for Sevier River at Hatch. When prediction is made for the volume of water passing through this gage, the variables are averaged or accumulated over the time. The seasonality effect is thus eliminated leaving annual, interannual to interdecadal components. These components are best represented by the NP region. Therefore, the best prediction is obtained from NP SST. In short, TP is replaced by NP in this site. This result is consistent with result obtained by Asefa et al. (2006).

For the Weber River near Oakley, the best prediction result is obtained at CP SST when one individual SST is used at a time. The combination of CP, NP, and TP develops the best model prediction when used in combination, however, these predictions are very close to predictions from the combination of NP and CP SST. The principal moisture source of this area is the Pacific Ocean. In addition, this stream gage is outside of the ENSO dominance region. There is no strong seasonality component therefore NP and CP SST appeared as important input.

2.7.3 Generalization and robustness

The bootstrap analysis is a data-based simulation method for statistical inference (Efron and Tibshirani 1998). This gives the estimate of measure of variability of test statistics with the change in training data. This analysis shows how robust the model is and how well it will generalize. The concept is to randomly draw a large number of 'resamples' of size *n* from the original sample, with replacement. Although each resample has the same number of elements as the original sample, it may include some of the original data points more than once, and some are not included. This process forming the training set is random and it is treated as independent sets (Duda et al. 2000). Therefore, each of these resamples will randomly depart from the original sample. From each bootstrap set, the bootstrap test statistic is computed in exactly the way as the real sample is used (Davidson and MacKinnon 2001). Since the elements in these resamples vary slightly, the statistics calculated from these resamples takes on slightly different values. Having computed statistics each time, a histogram is prepared which gives the variability of the test statistics.

For each stream gage, bootstrap analysis is performed for the best model. The test statistics are computed for each bootstrap sample and a histogram is prepared. Figure 2.13 and 2.14 shows the result of bootstrapping for the best identified model for monthly mean discharge prediction. Figure 2.15 and 2.16 shows the histogram of bootstrap analysis for each selected stream gage for volume of water passing the gage for next six months. The narrow bound in the resulting histograms shows that the model is robust. The variability on these test statistics (RMSE and Efficiency) is consistent. The dotted red line in the Figure 2.13 through Figure 2.16 shows the 2.5^{th} percentile and 97.5^{th} percentile values of test statistics. These bootstrap plots confirm that the model is robust and is good enough to use it as a long-term streamflow prediction model.

2.8 Conclusion

A major aspect of variability in streamflow is the variability of regional climate, which, in turn, is related to larger scale phenomena occurring in the oceans and the atmosphere (Koch and Fisher 2000). The regional meteorological effect is represented by the sea surface temperature of Pacific and Atlantic Oceans. Identification of right location of SST location for given spatial location of stream gage is crucial in order to make accurate and reliable prediction. Along with the regional meteorological inputs, local meteorological inputs are also used to predict the streamflow for the next six months at each of five selected gages in the State of Utah. The local meteorological conditions are incorporated using local temperature and snowpack in the mountains. Thus, the inputs to the model are past streamflow data, snowpack in the mountain, local temperature, and sea surface temperature at various locations in the Pacific and Atlantic Oceans. The input variables are integrated into a machine-learning framework to develop a useful model for the long-term streamflow forecast. The Multivariate Relevance Vector Machine successfully transformed the input variables into reasonably accurate forecasting of outputs. For each gage, the best location of SST was identified by comparing the test statistics among all SST locations. It was found that the sea surface temperature in the Pacific Ocean predicted better than that of the Atlantic Ocean. It is so because this region represents the majority of Ocean-atmosphere climate influencing the Western U.S. (Ting and Wang 1997; Wang and Ting 2000). For the stream gages located in the northern and Central Utah, usually North Pacific and sometimes Central Pacific SST produced the best model predictions. For the stream gages located in the Southern Utah, Tropical Pacific SST produced best predictions for monthly mean discharge for the next six months. However, NP SST produced best prediction for most of the stream gages in Utah for the volume of water passing the gage for the next six months.

Using the best identified SST locations, the streamflow was predicted for the next six months at each selected stream gage that spatially covers the state of Utah from North to South. Predictions are made for (i) Monthly mean discharges for the next six months, and (ii) Volume of water passing through the gage for the next six months. The performance of the model is evaluated based on RMSE and Nash-Sutcliffe efficiency in the testing phase. The model prediction has good agreement with the observed flow value. The prediction result is very accurate for unimpaired stream gages while the accuracy is reasonable for one impaired gage. Since human induced effects are not included in the model for impaired gage, less efficiency of prediction is obvious. The model predicts the streamflow very well for high flow, in general, but predictions for low are not captured perfectly. Since input representing ground water flow is not included in the model, a certain level of discrepancy is to be expected for low flows. The overall prediction is, however, accurate and has good agreement with observed streamflow values. The uncertainty of the prediction is also captured and presented by the confidence intervals of the predictions. The reliability and robustness of the model is tested by using a bootstrap analysis. This analysis confirms the good predictability and robustness of the model.

This paper has demonstrated that with the use of appropriate input, Multivariate Relevance Vector Machine (MVRVM) can be utilized for the successful forecast of long-term streamflow. Accurate and reliable long-term streamflow prediction is crucial for the management of water resources at the basin scale. This information could help the water managers and stakeholders for planning and decision making. This will ultimately reduce the financial risk associated with future water shortages.

The Northern and Central regions of Utah are affected by the annual, and interannual to interdecadal climatic signal. Using those climatic signals, forecast may be extended for longer lead-time than what is demonstrated in this paper. This can be a future direction of research on hydrologic modeling using learning machines.

References

Asefa, T., Kemblowski, M., McKee, M., and Khalil, A. (2006). "Multi-time scale stream flow predictions: The support vector machines approach." *J. Hydrol.*, 318(1-4), 7-16.

Berger, B., Hansen, R., and Jensen, R. (2003). "Sevier River Basin system description." Sevier River Water Users Association, Delta, UT.

Berger, J. O. (1993). *Statistical decision theory and Bayesian analysis/ second edition*, Springer-Verlag, New York.

Davidson, R., and MacKinnon, J., G. (2001). "Bootstrap tests: How many bootstraps?", Queen's University, Department of Economics.

Duda, R. O., Hart, P. E., and Stork, D. G. (2000). *Pattern classification*, Wiley Interscience, Second Edition, New York.

Efron, B., and Tibshirani, R. J. (1998). A*n introduction of the bootstrap, Monographs on statistics and applied probability*, CRC Press LLC, Boca Raton, FL.

Hamlet, A. F., and Lettenmaier, D. P. (1999). "Columbia River streamflow forecasting based on ENSO and PDO climate signals." *J. Water Resour. Plann. Manage.*, 125(6), 333-341.

Kaplan, A., Cane, M. A., Kushnir, Y., Clement, A. C., Blumenthal, M. B., and Rajagopalan, B. (1998). "Analyses of global sea surface temperature 1856-1991." *J. Geophys. Res*., 103(C9), 18567-18589.

Khalil, A. F., McKee, M., Kemblowski, M., and Asefa, T. (2005a). "Basin scale water management and forecasting using Artificial Neural Networks." *JAWRA Journal of the American Water Resources Association*, 41(1), 195-208.

Khalil, A. F., McKee, M., Kemblowski, M., and Asefa, T. (2005b). "Sparse Bayesian learning machine for real-time management of reservoir releases." *Water Resour. Res*., 41(11), W11401.

Khalil, A. F., McKee, M., Kemblowski, M., Asefa, T., and Bastidas, L. (2006). "Multiobjective analysis of chaotic dynamic systems with sparse learning machines." *Adv. Water Res*., 29(1), 72-88.

Koch, R. W., and Fisher, A. R. (2000). "Effects of inter-annual and decadal-scale climate variability on winter and spring streamflow in Western Oregon and Washington." *Proc., Western Snow Conference*, 1-11.

McCord, M. W. (1997). "Southwest Paddler, Outdoor recreation guide for Utah ", <http://southwestpaddler.com/docs/muddyut.html>. (Accessed 11 May 2010).

Nakamura, H., Lin, G., and Yamagata, T. (1997). "Decadal climate variability in the North Pacific during the recent decades." *Bull. Amer. Meteor. Soc.*, 78(10), 2215-2225.

Nash, J. E., and Sutcliffe, I. V. (1970). River flow forecasting through conceptual models.

"Natural Resources Conservation Service." <http: //www.wcc.nrcs.usda.gov/snow>. (Accessed 13 August 2010).

Perica, S., and Stayner, M. (2004). "Regional flood frequency analysis for selected basins in Utah ", Utah Department of Transportation Research and Development Division Salt Lake City, UT.

Pope, D., and Brough, C. (1996). *Utah's Weather and Climate*, Publishers Press, Salt Lake City, UT.

Poveda, G., Jaramillo, A., Gil, M. M., Quiceno, N., and Mantilla, R. I. (2001). "Seasonally in ENSO-related precipitation, river discharges, soil moisture, and vegetation index in Colombia." *Water Resour. Res.*, 37(8), 2169-2178.

Sivakumar, B. (2003). "Forecasting monthly streamflow dynamics in the western United States: a nonlinear dynamical approach." *Environ. Model. Software*, 18(8-9), 721-728.

Smith, T. M., and Reynolds, R. W. (2003). "Extended reconstruction of global sea surface temperatures based on COADS data (1854–1997)." *J. Climate*, 16(10), 1495-1510.

Soukup, T. L., Aziz, O. A., Tootle, G. A., Piechota, T. C., and Wulff, S. S. (2009). "Long lead-time streamflow forecasting of the North Platte River incorporating oceanic–atmospheric climate variability." *J. Hydrol.*, 368(1-4), 131-142.

Thayananthan, A. (2005). "Template-based pose estimation and tracking of 3D hand motion." PhD Dissertation, University of Cambridge, Cambridge, UK.

Thayananthan, A., Navaratnam, R., Stenger, B., Torr, P., and Cipolla, R. (2008). "Pose estimation and tracking using multivariate regression." *Pattern Recognition Lett.*, 29(9), 1302-1310.

Ticlavilca, A. (2010). "Multivariate Bayesian machine learning regression for operation and management of multiple reservoir, irrigation canal, and river systems." PhD Dissertation, Utah State University, Logan, UT.

Ting, M., and Wang, H. (1997). "Summertime U.S. Precipitation Variability and Its Relation toPacific Sea Surface Temperature." *J. Climate*, 10(8), 1853-1873.

Tipping, M. (2000). "The Relevance Vector Machine." *Proc., Advances in Neural Information Processing Systems*, The MIT Press, 652-658.

Tipping, M. (2001). "Sparse Bayesian Learning and the Relevance Vector Machine." *J. Machine Learning Res*., 1, 211-244.

Tipping, M. E., and Faul, A. C. (2003). "Fast marginal likelihood maximization for sparse Bayesian models." *Proc., Ninth International Workshop on Artificial Intelligence and Statistics*.

Tootle, G. A., and Piechota, T. C. (2006). "Relationships between Pacific and Atlantic ocean sea surface temperatures and U.S. streamflow variability." *Water Resour. Res*., 42(7), W07411.

"Utah Center for Climate and Weather." <http://utahweather.org/>. (Accessed 15 November 2010).

Vapnik, V. N. (1995). *The nature of statistical learning theory*, Springer Verlag, New York.

Vapnik, V. N. (1998). *The nature of statistical learning theory*, Springer Verlag, New York.

Wang, H., and Ting, M. (2000). "Covariabilities of winter U.S. precipitation and Pacific sea surface temperatures." *J. Climate*, 13(20), 3711-3719.

Table 2.1 Geometric characteristic of stream gages

Site ID	Name	Basin	Stream		Gage	
		Area (mi^2)	Length (mi)	Slope	Latitude (°)	Longitude (°)
10128500	Weber River near Oakley	162.1	25.3	0.020	40.737	-111.247
10131000	Chalk Creek at Coalville	248.3	37.5	0.010	40.921	-111.401
10174500	Sevier River at Hatch	340	31.1	0.007	37.651	-112.430
09330500	Muddy Creek near Emery	105	20.1	0.004	38.982	-111.249
10149000	Sixth Water Creek near Springville	15	1	0.048	40.118	-111.314

Table 2.2 95 percent confidence interval of the median test RMSE for Model 1

Stream gages	95% confidence interval		Best RMSE	Remark
	Lower	Upper	(cfs)	
Weber River near Oakley	8.87	9.54	8.54	NP and CP
Chalk Creek at Coalville	5.85	6.47	5.57	NP and EA
Muddy Creek near Emery	4.80	5.46	3.56	NP
Sevier River at Hatch	11.76	13.27	11.74	TP and TA
Sixth Water Creek	4.99	6.16	3.08	NP

Table 2.3 Best test statistics for monthly mean discharge prediction (Model 1)

Stream gages	Test RMSE (cfs)	Efficiency	Combination of SST locations
Weber River near Oakley	8.54	0.999	NP and CP
Chalk Creek at Coalville	5.57	0.995	NP and EA
Muddy Creek near Emery	3.56	0.995	NP
Sevier River at Hatch	11.74	0.995	TP and TA
Sixth Water Creek near Springville	3.08	0.816	NP

Table 2.4 95 percent confidence interval of the median test RMSE for Model 2

Streamflow sites	95% confidence interval		Best RMSE (1000 ac-ft)	Remark
	Lower	Upper		
Weber River near Oakley	8.66	11.45	8.31	NP, CP and TP
Chalk Creek at Coalville	2.75	4.46	2.65	NP
Muddy Creek near Emery	2.52	4.11	2.44	NP
Sevier River at Hatch	5.36	7.72	5.04	NP
Sixth Water Creek near Springville	0.88	1.32	0.73	NP

Table 2.5 Best test statistics for volume of water passing the gage for next six months (Model 2)

Stream site	Test RMSE (1000 ac-ft)	Efficiency	Best combination of SST locations
Weber River near Oakley	8.307	0.965	NP, CP and TP
Chalk Creek at Coalville	2.653	0.968	NP
Muddy Creek near Emery	2.438	0.951	NP
Sevier River at Hatch	5.042	0.987	NP
Sixth Water Creek near Springville	0.732	0.739	NP

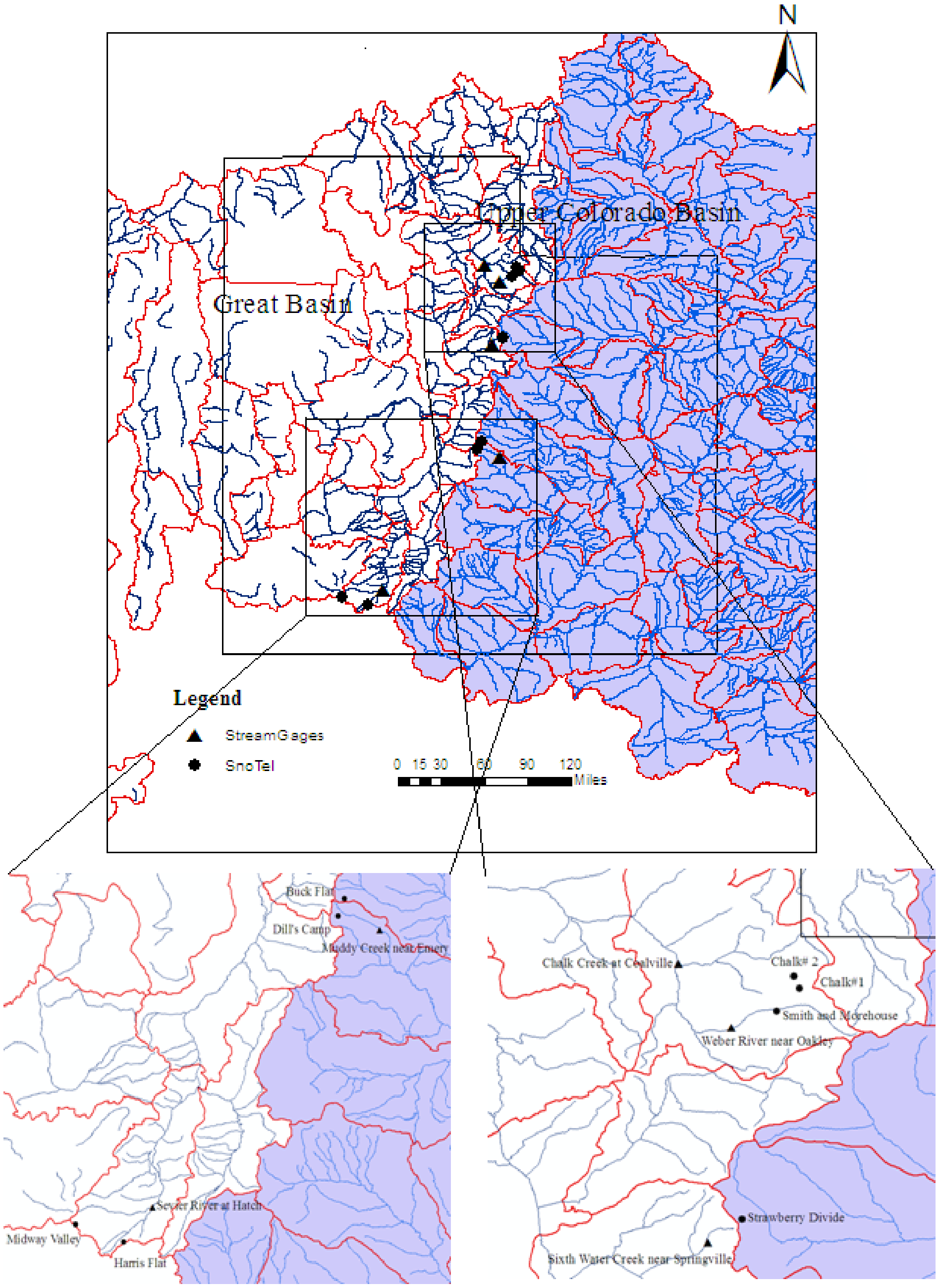


Figure 2.1 Location of the stream gages and SnoTel stations in Utah.

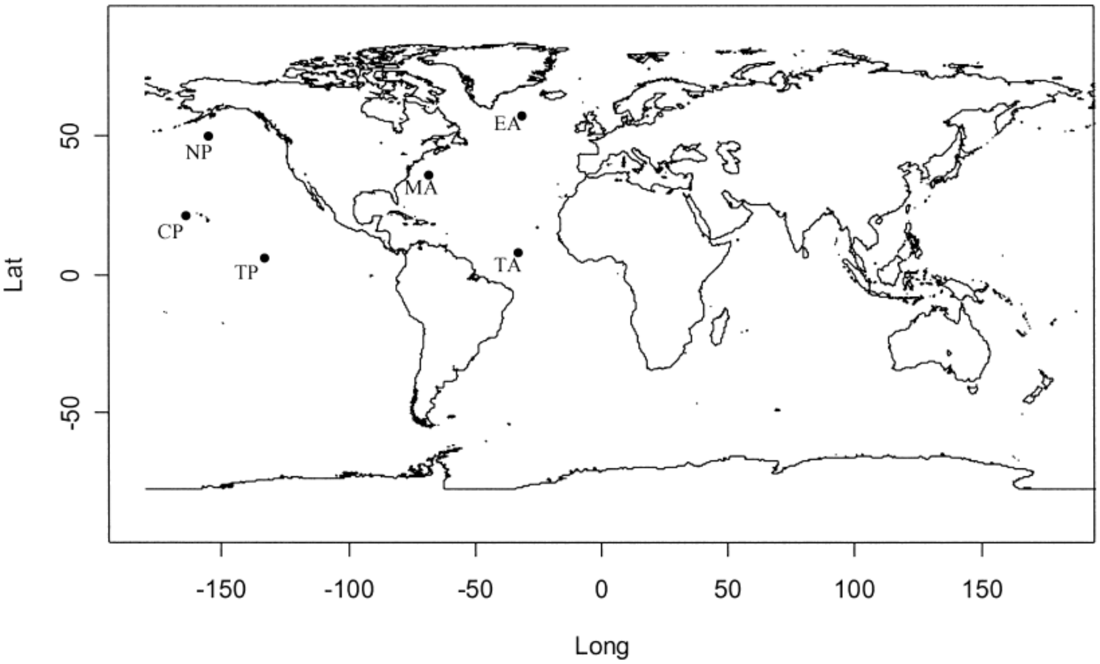


Figure 2.2 The locations for the sea surface temperature (Khalil et al. 2005a).

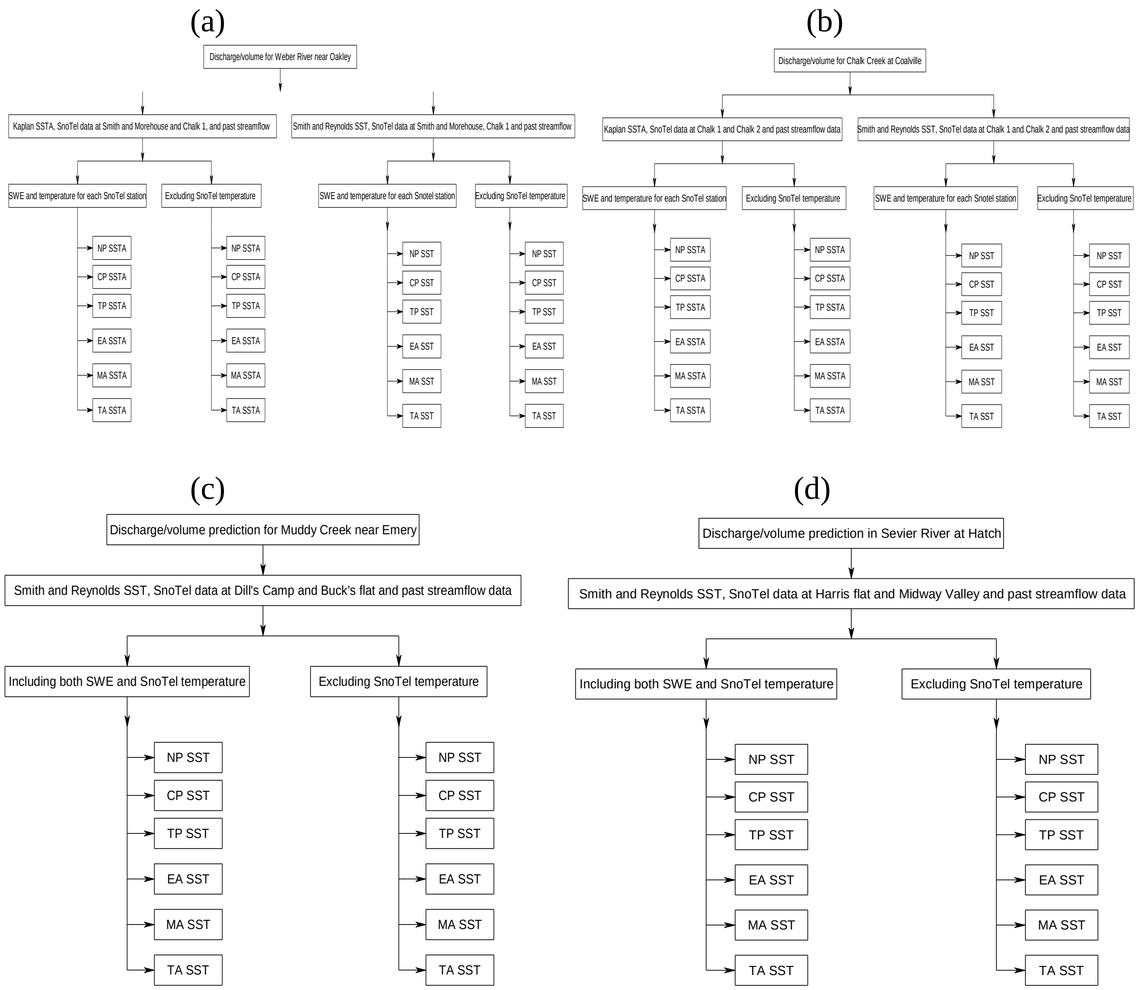

(a)
Discharge/volume for Weber River near Oakley
Kaplan SSTA, SnoTel data at Smith and Morehouse and Chalk 1, and past streamflow
Smith and Reynolds SST, SnoTel data at Smith and Morehouse, Chalk 1 and past streamflow
SWE and temperature for each SnoTel station
Excluding SnoTel temperature
NP SSTA
CP SSTA
TP SSTA
EA SSTA
MA SSTA
TA SSTA
NP SSTA
CP SSTA
TP SSTA
EA SSTA
MA SSTA
TA SSTA
SWE and temperature for each Snotel station
Excluding SnoTel temperature
NP SST
CP SST
TP SST
EA SST
MA SST
TA SST
NP SST
CP SST
TP SST
EA SST
MA SST
TA SST
(b)
Discharge/volume for Chalk Creek at Coalville
Kaplan SSTA, SnoTel data at Chalk 1 and Chalk 2 and past streamflow data
Smith and Reynolds SST, SnoTel data at Chalk 1 and Chalk 2 and past streamflow data
SWE and temperature for each SnoTel station
Excluding SnoTel temperature
NP SSTA
CP SSTA
TP SSTA
EA SSTA
MA SSTA
TA SSTA
NP SSTA
CP SSTA
TP SSTA
EA SSTA
MA SSTA
TA SSTA
SWE and temperature for each Snotel station
Excluding SnoTel temperature
NP SST
CP SST
TP SST
EA SST
MA SST
TA SST
NP SST
CP SST
TP SST
EA SST
MA SST
TA SST
(c)
Discharge/volume prediction for Muddy Creek near Emery
Smith and Reynolds SST, SnoTel data at Dill's Camp and Buck's flat and past streamflow data
Including both SWE and SnoTel temperature
Excluding SnoTel temperature
NP SST
CP SST
TP SST
EA SST
MA SST
TA SST
NP SST
CP SST
TP SST
EA SST
MA SST
TA SST
(d)
Discharge/volume prediction in Sevier River at Hatch
Smith and Reynolds SST, SnoTel data at Harris flat and Midway Valley and past streamflow data
Including both SWE and SnoTel temperature
Excluding SnoTel temperature
NP SST
CP SST
TP SST
EA SST
MA SST
TA SST
NP SST
CP SST
TP SST
EA SST
MA SST
TA SST

Figure 2.3 Flowchart for developing the input file for different combinations of input variables using one individual sea surface temperature at a time for both monthly mean discharge prediction (Model 1) and volumetric prediction (Model 2). The path to each final node shows the set of input variables used for that node. (a) Combination of input variables for Weber River near Oakley, (b) Combination of input variables for Chalk Creek at Coalville, (c) Combination of input variables for Muddy Creek near Emery, and (d) Combination of input variables for Sevier River at Hatch.

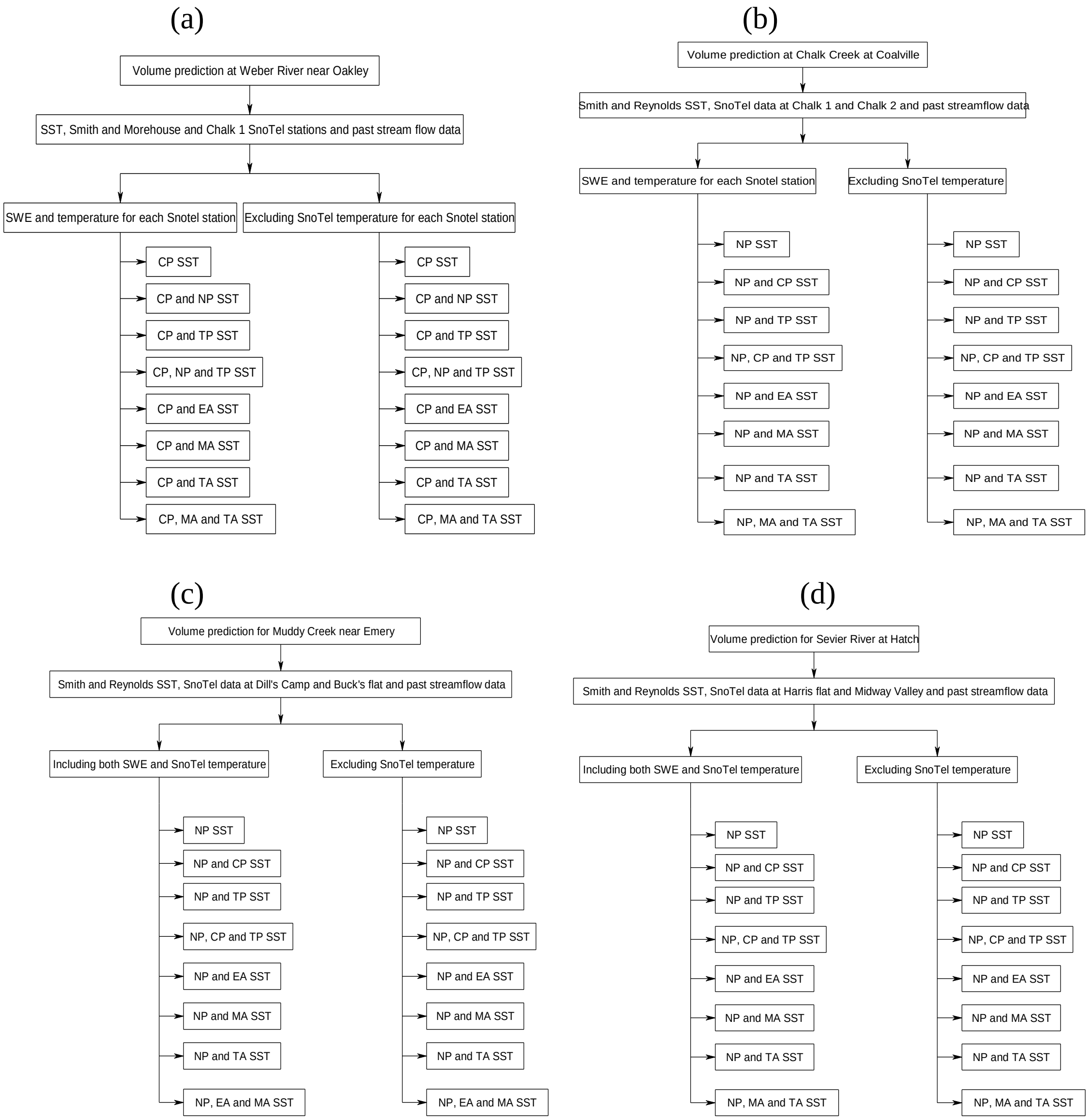


Figure 2.4 Sample flowchart for developing the input file for prediction of volume of water passing through the stream gages (Model 2) using combined SST of different locations of Pacific and Atlantic Oceans. The path to each final node shows the set of input variables used for that node. (a) Combination of input variables for Weber River near Oakley, (b) Combination of input variables for Chalk Creek at Coalville, (c) Combination of input variables for Muddy Creek near Emery, and (d) Combination of input variables for Sevier River at Hatch.

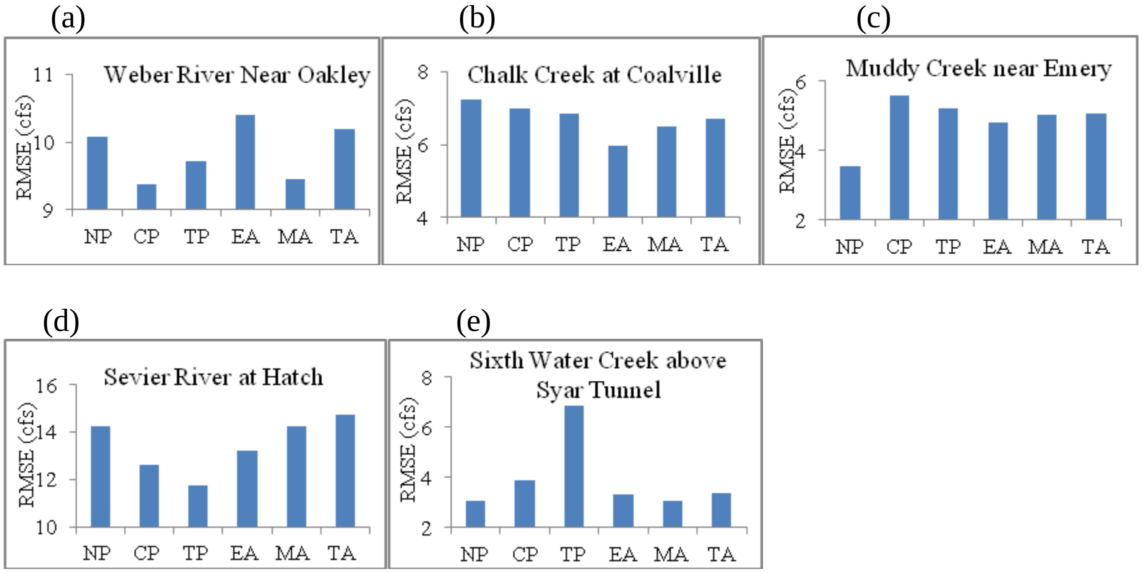


Figure 2.5 The test RMSE for six individual SST locations at each selected stream gage for monthly mean discharge prediction for next six months. (a) Weber River near Oakley, (b) Chalk Creek at Coalville, (c) Muddy Creek near Emery, (d) Sevier River at Hatch, and (e) Sixth Water Creek near Springville.

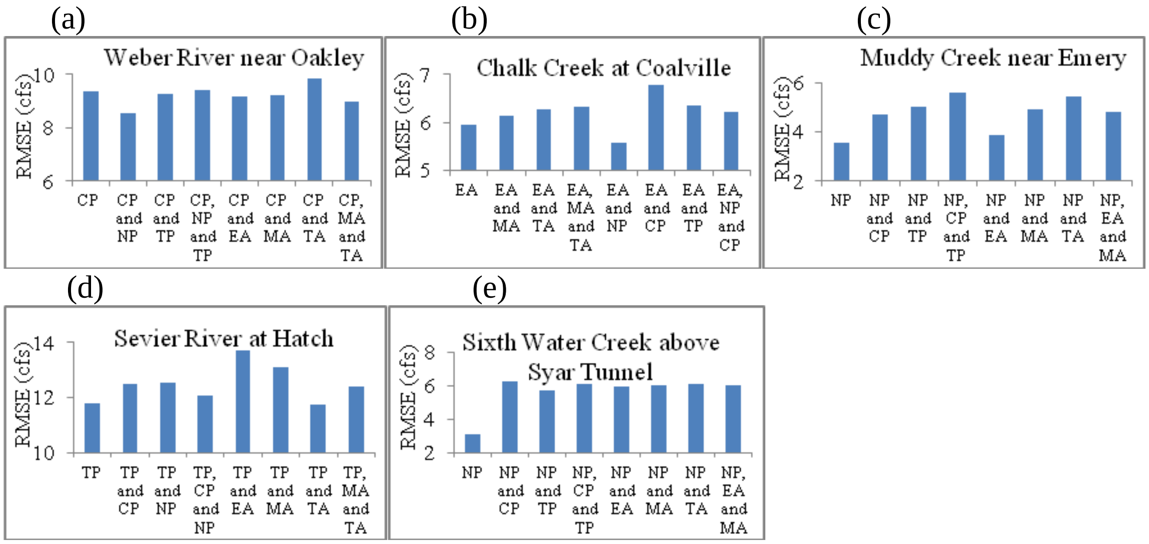


Figure 2.6 Test RMSE for the combinations of SST locations for the monthly mean discharge prediction for next six months at each selected stream gage. (a) Weber River near Oakley, (b) Chalk Creek at Coalville, (c) Muddy Creek near Emery, (d) Sevier River at Hatch, and (e) Sixth Water Creek near Springville.

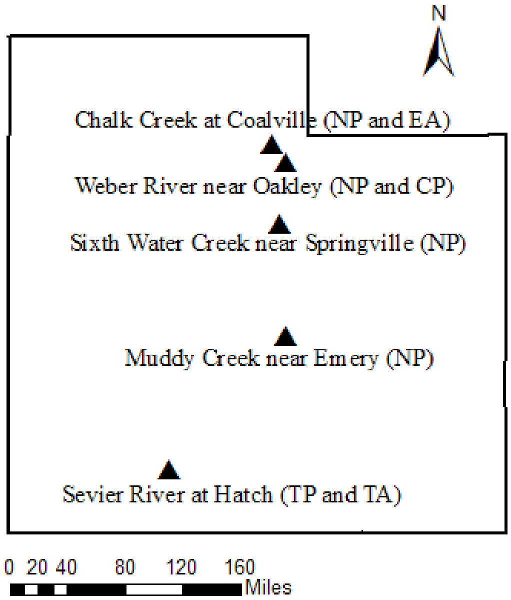

N
Chalk Creek at Coalville (NP and EA)
Weber River near Oakley (NP and CP)
Sixth Water Creek near Springville (NP)
Muddy Creek near Emery (NP)
Sevier River at Hatch (TP and TA)
0 20 40 80 120 160
Miles

Figure 2.7 Locations of the sea surface temperature that develops the best test result for the monthly mean discharge prediction at selected streamflow gages for next six months.

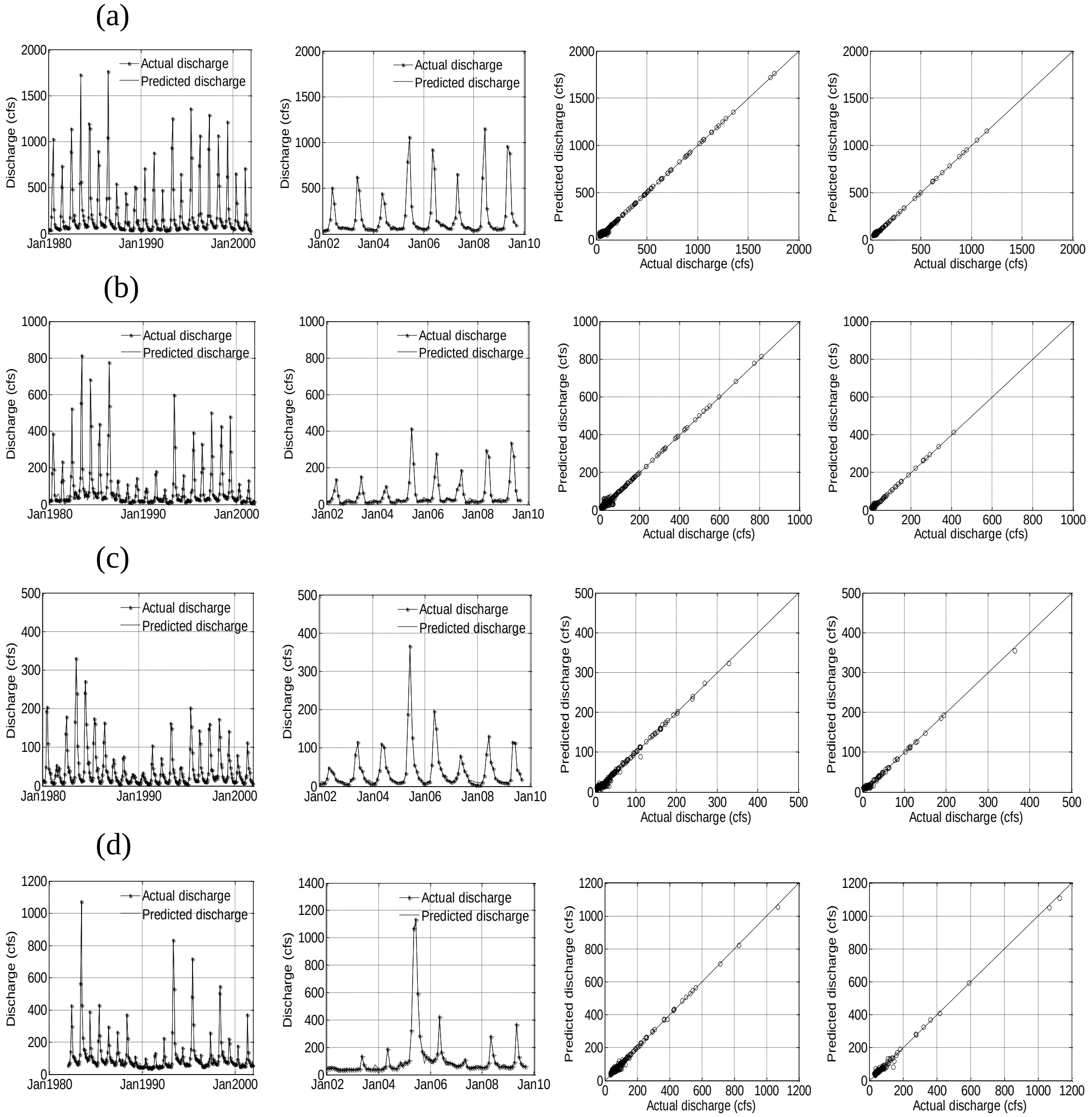


Figure 2.8 Streamflow prediction for next six months at each selected stream gage. For (a) to (e), first column is for the training phase, second column is for test phase, third column shows the plot of predicted discharge versus actual discharge for training phase, and fourth column shows the similar plot for the test phase. (a) Weber River near Oakley, (b) Chalk Creek at Coalville, (c) Muddy Creek near Emery, (d) Sevier River at Hatch, (e) Sixth Water Creek near Springville, (f) 90 percent confidence interval of prediction in test phase for each selected gage for (a) to (e), and (g) Residual plots for (a) to (e).

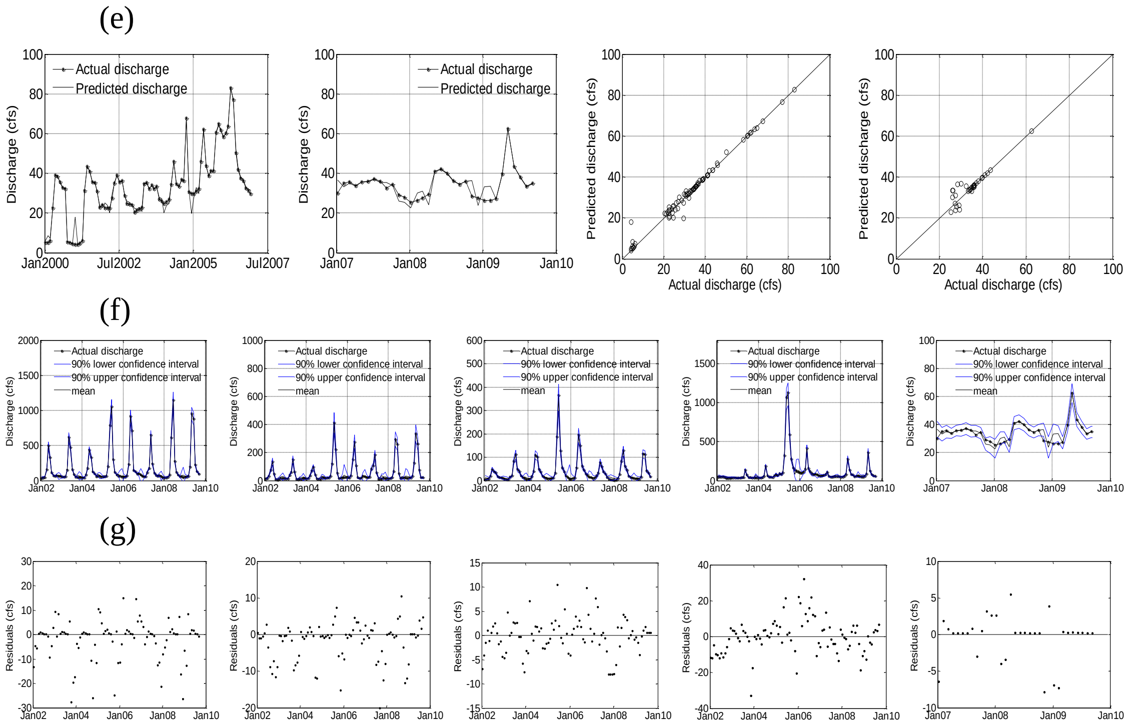


Figure 2.8 Cont.

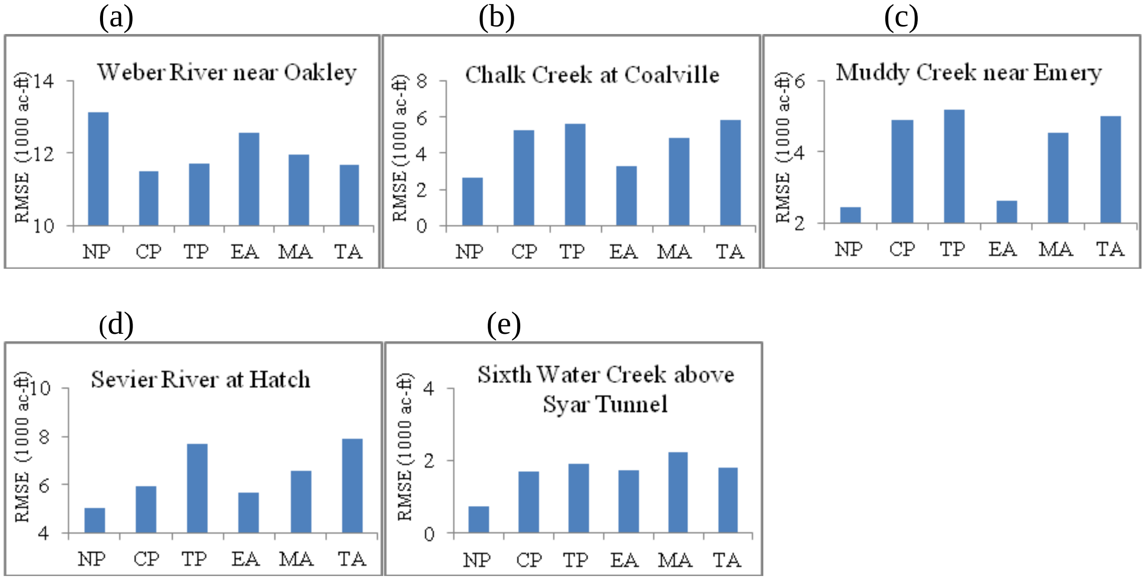


Figure 2.9 The test RMSE of six individual SST locations at each selected stream gage for volume of water passing the gage for next six months. (a) Weber River near Oakley, (b) Chalk Creek at Coalville, (c) Muddy Creek near Emery, (d) Sevier River at Hatch, and (e) Sixth Water Creek near Springville.

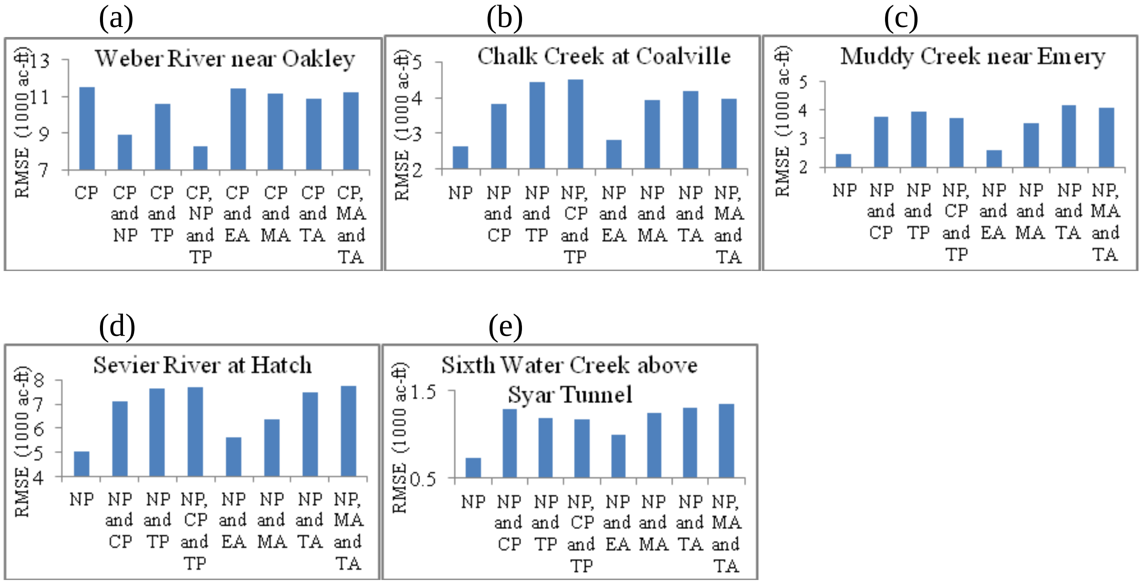


Figure 2.10 Test RMSE for the combinations of SST locations for the volumetric prediction at each selected stream gage. (a) Weber River near Oakley, (b) Chalk Creek at Coalville, (c) Muddy Creek near Emery, (d) Sevier River at Hatch, and (e) Sixth Water Creek near near Springville.

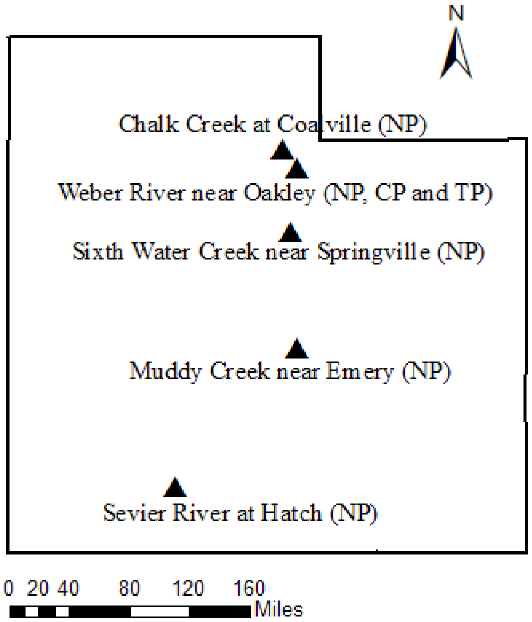

N
Chalk Creek at Coalville (NP)
Weber River near Oakley (NP, CP and TP)
Sixth Water Creek near Springville (NP)
Muddy Creek near Emery (NP)
Sevier River at Hatch (NP)
0 20 40 80 120 160
Miles

Figure 2.11 Locations of the sea surface temperature that developed the best test statistics for volume of water passing the stream gage for next six months.

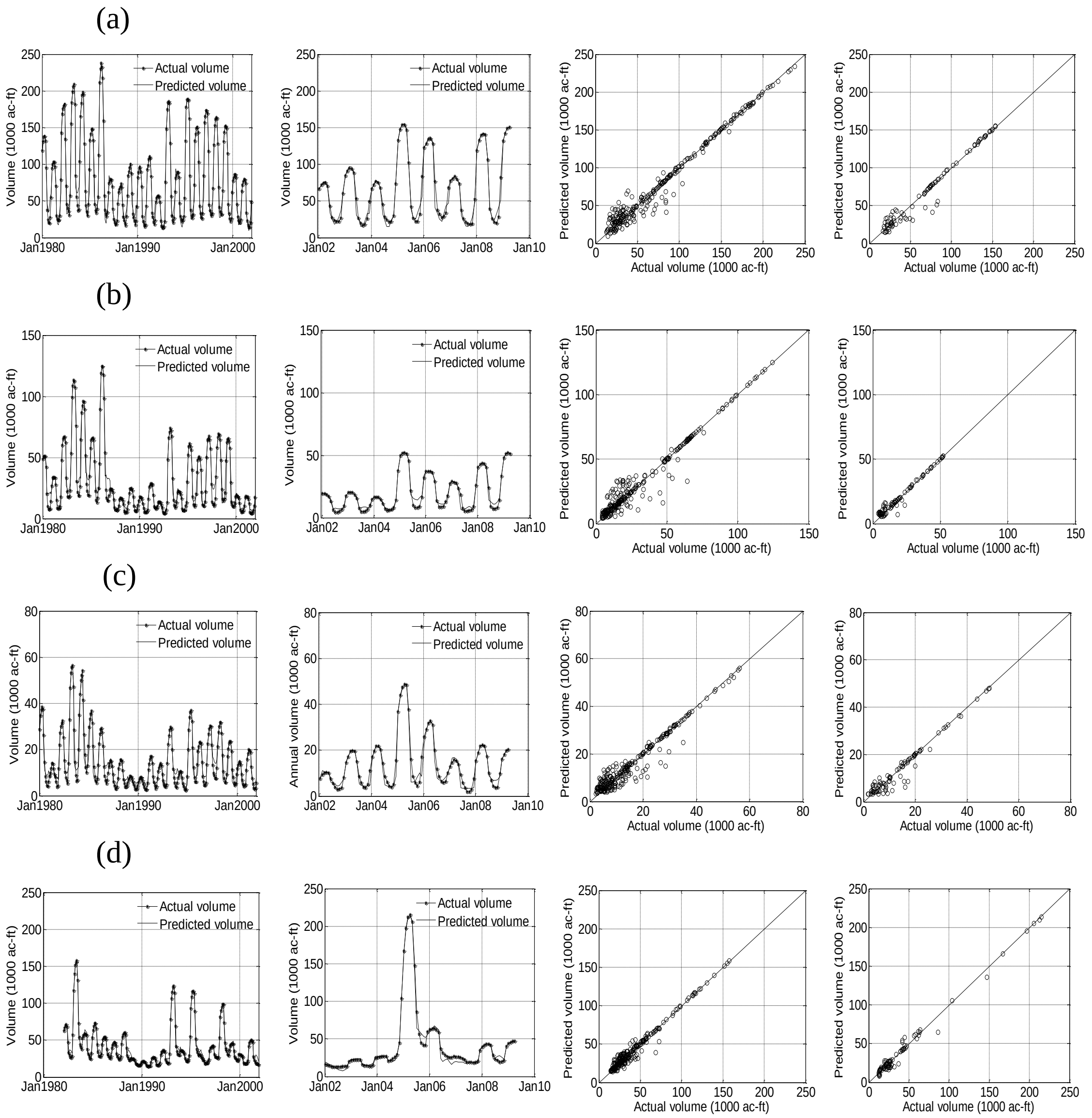


Figure 2.12 The prediction for volume of water passing through the stream gages for next six months. For (a) to (e), first column is the training phase, second column is test phase, third column is the plot of predicted volume versus actual volume for training phase, and fourth column is similar plot for the test phase. The results are in the order of (a) Weber River near Oakley, (b) Chalk Creek at Coalville, (c) Muddy Creek near Emery, (d) Sevier River at Hatch, (e) Sixth Water Creek near Springville, (f) 90 percent confidence interval of prediction for (a) to (e), and (g) Residual plots of test phase for (a) to (e).

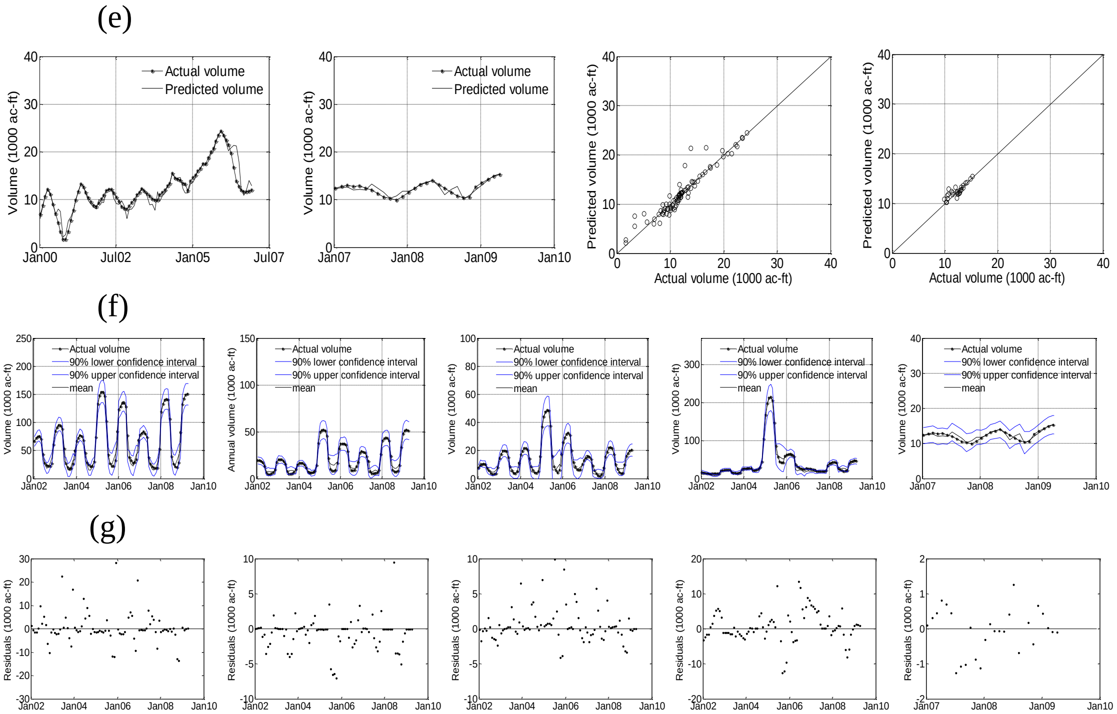


Figure 2.12 Cont.

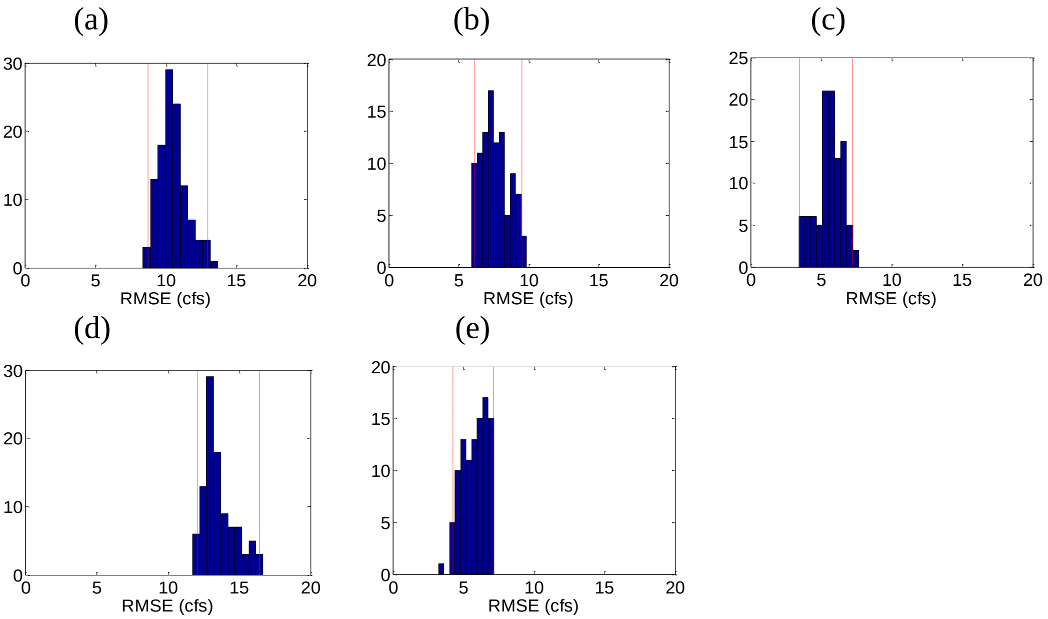


Figure 2.13 The bootstrap analysis for the best model at each stream gage for monthly mean discharge prediction for next six months. The figures presented are for the RMSE values for (a) Weber River near Oakley, (b) Chalk Creek at Coalville, (c) Muddy Creek near Emery, (d) Sevier River at Hatch, and (e) Sixth Water Creek near Springville.

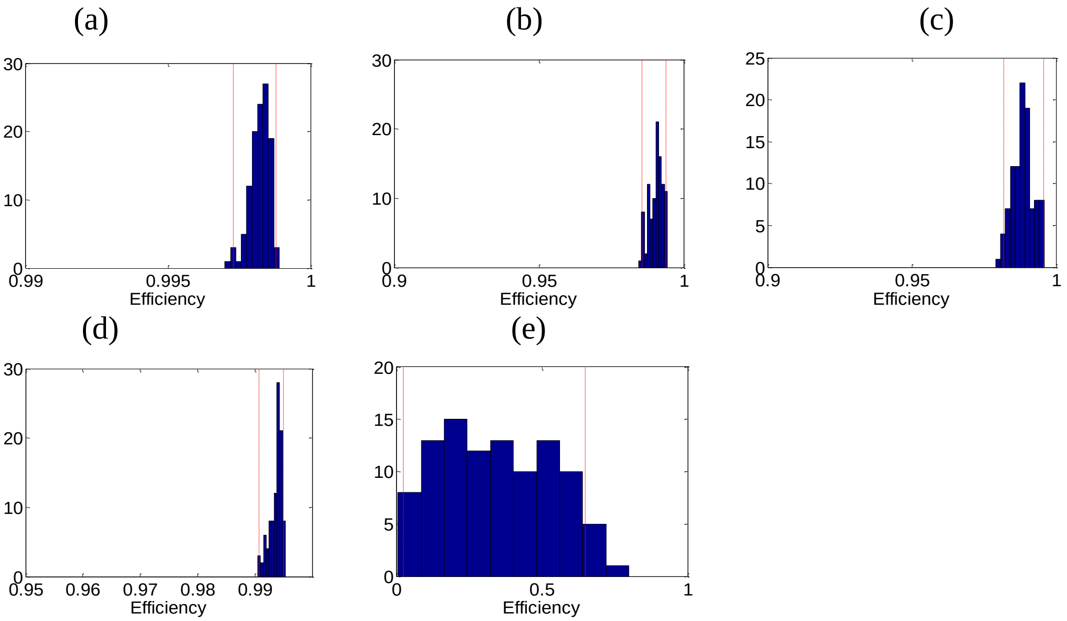


Figure 2.14 The bootstrap analysis for the best model for each stream gage for monthly mean discharge prediction. The figures presented are for the Nash-Sutcliffe efficiency for (a) Weber River near Oakley, (b) Chalk Creek at Coalville, (c) Muddy Creek near Emery, (d) Sevier River at Hatch, and (e) Sixth Water Creek near Springville.

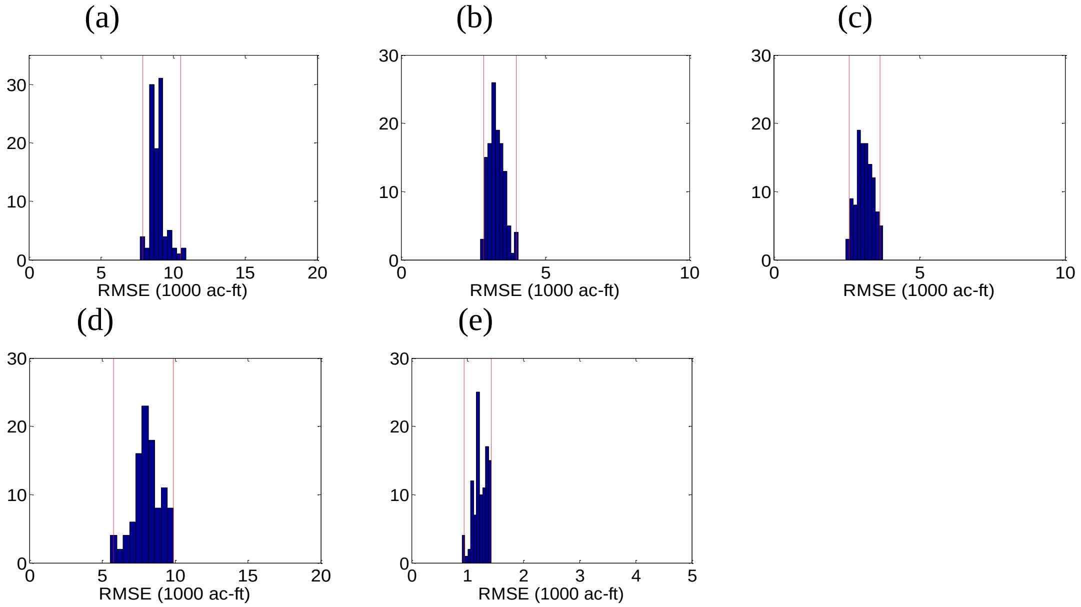


Figure 2.15 The bootstrap analysis of the best model for volumetric prediction at each stream gage. The figures presented are for the RMSE values for (a) Weber River near Oakley, (b) Chalk Creek at Coalville, (c) Muddy Creek near Emery, (d) Sevier River at Hatch, and (e) Sixth Water Creek near Springville.

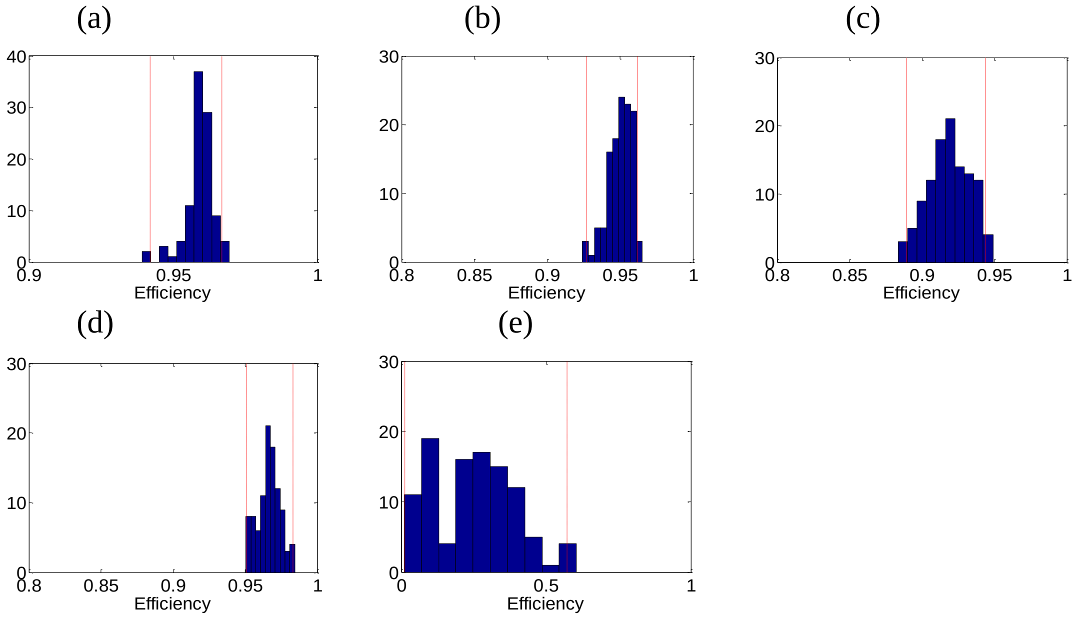


Figure 2.16 The bootstrap analysis for the best model for volumetric prediction at each selected stream gages. The figures presented are for the Nash-Sutcliffe efficiency for (a) Weber River near Oakley, (b) Chalk Creek at Coalville, (c) Muddy Creek near Emery, (d) Sevier River at Hatch, and (e) Sixth Water Creek near Springville.

CHAPTER 3

LONG LEAD-TIME STREAMFLOW FORECASTING AND IDENTIFICATION OF RELATIVE INFLUENCE OF OCEANIC-ATMOSPHERIC OSCILLATION MODES USING BAYESIAN MACHINE LEARNING REGRESSION APPROACH

Abstract

Climatic variability influences the hydrological cycle that subsequently affects the discharge in streams. The variability in the climate can be represented by ocean-atmospheric oscillations which provide an input to forecasting streamflow. Four popular ocean-atmospheric modes are used in this chapter for annual streamflow volume prediction in selected stream gages in Utah. These modes are the Pacific Decadal Oscillation (PDO), the El-Niño Southern Oscillation (ENSO), the Atlantic Multidecadal Oscillation (AMO), and the North Atlantic Oscillation (NAO). Multivariate Relevance Vector Machine (MVRVM), a data-driven model based on a Bayesian learning approach, is used for the streamflow prediction. This is a sparse model and provides probabilistic output. The model is applied at four unimpaired stream gages in Utah that spatially cover the state from North to South. Different model types are developed based on the combination of input oscillation modes. A total of 60 years (1950-2009) of data are used for the analysis. The Model is trained on 50 years of data (1950-1999) and tested on 10 years of data (2000-2009). The accuracy of the prediction is evaluated based on the root mean square error (RMSE), efficiency, and correlation coefficient for the test phase data. An appropriate combination of oscillation modes and lead-time is chosen for each

selected gage, based on test results. These combinations are used to develop final forecast model for annual streamflow volume prediction for next few years. The model prediction has reasonable agreement with the actual annual streamflow volume. Such predictions constitute valuable information to water managers for effective planning and management of water resources. The sensitivity analysis shows that PDO and ENSO have relatively stronger signals than other oscillation modes in influencing streamflow predictions. The prediction results from the MVRVM are compared with Support Vector Machine (SVM) and Artificial Neural Network (ANN) results. MVRVM performs better than other two models using relatively fewer data points in training. Bootstrap analysis confirms the robustness of the model.

3.1 Introduction

The ocean-atmospheric modes are connected to climatic variability around the globe. The precipitation in any region is influenced by the climatic variability that subsequently affects the streamflow. Floods and draughts are also consequences of climatic variability. The streamflow in the western United States is no doubt influenced by it. The teleconnection between climate and ocean/atmospheric modes (oscillation indices) is the scientific basis of long lead-time streamflow prediction. Their correlation provides the ability to reliably forecast streamflow. The Pacific Decadal Oscillation (PDO), the El-Niño Southern Oscillation (ENSO), the Atlantic Multi-decadal Oscillation (AMO), and the North Atlantic Oscillation (NAO) are popular oceanic-atmospheric oscillation indices used in hydrologic prediction. The climatic variation in decadal-scale over the Pacific Ocean and its surrounding are strongly related to PDO, which is coherent with wintertime climate over North America (Mochizuki et al. 2010). ENSO has been linked to climate anomalies throughout the world (Diaz and Markgraf 2000; Philander 1990). Strong ENSO signal exists in mid-latitude United States that affects the flow in rivers and streams (Kahya and Dracup 1993). Many prominent examples of regional multidecadal climate variability have been related to AMO. It affects air temperature and rainfall, and river flow over much of the Northern Hemisphere, in particular, North America and Europe (Enfield et al. 2001; McCabe et al. 2004; Sutton and Hodson 2005). NAO is the dominant mode of winter climate variability in the North Atlantic region ranging from Central North America to Europe and much into Northern Asia. There are several past studies for the long lead time streamflow prediction using ocean-atmospheric oscillation indices. Streamflow responses to individual as well as coupled ocean-atmospheric indices of PDO, ENSO, AMO, and NAO over the United States are well established influencing signals (Hamlet and Lettenmaier 1999; Piechota et al. 1997). Chiew and McMahon (2002) used ENSO-streamflow relationship to forecast streamflow successfully. Soukup et al. (2009) used PDO, ENSO and AMO for seasonal streamflow prediction for the North Platte River. Kalra and Ahmad (2009) used those oscillation modes to predict long lead-time streamflow in the Colorado River basin.

The signal strength of oscillation indices varies spatially in the regions around the world. It is thus important to elucidate the influential oscillation indices, or their combinations, and corresponding lead time that produces reasonable long lead-time annual streamflow volume prediction for a given location of stream gages. This research paper uses an optimum combination of oscillation modes to predict long-lead time annual streamflow volume accurately and reliably at four unimpaired stream gages in Utah that spatially cover the state from North to South. The combination of oscillations that develops the best prediction for each lead time is also identified. This information can be useful to enhance the predictive ability of the streamflow model. Accurate prediction of long-lead time streamflow can benefit the management of water resources at the basin scale (Asefa et al. 2006). This is crucial information for water managers, farmers, and stakeholders, especially in arid regions. Such prediction helps decision making process to maximize the returns from available water resources and ensures a reliable supply. Forecast with long-lead time also facilitates co-ordination between different system users, which may be important in multiple-use water resource systems (Hamlet and Lettenmaier 1999).

There are quite a few physically based models developed to understand the behavior of water resources systems. The complexities in these models and difficulties associated with the data acquisition and corresponding expenses that these models would require has limited in the application. To overcome these limitations, data driven models are often used as an alternative to physically based models. They are characterized by their ability to quickly capture the underlying physics of the system by relating input and output. They are robust and are capable of making reasonable prediction using historical data (Khalil et al. 2005b, 2006).

Artificial Neural Network (ANN), Support Vector Machine (SVM) and Relevance Vector Machine (RVM) are data driven models. ANN model has the ability to implicitly detect complex nonlinear relationships between response and predictor. It performs well even if the data contains noise. However, it has a number of disadvantages. For example, ANN models may get stuck in local minima rather than global minima.

Also, an incorrect network definition may cause over-fitting of the model. SVM is widely used machine learning model. It however makes unnecessary liberal use of the basis function, and the number of support vector linearly increases with the size of the training dataset (Tipping 2001). The prediction is not probabilistic. Moreover, optimizing more than two model parameters in SVM needs more time and data for cross validation. RVM is sparser than SVM and gives probabilistic output as well. Optimizing model parameter for RVM is easier than SVM, however, the performance is comparable. RVM is therefore proposed for long lead-time annual streamflow volume prediction in this paper. This paper uses Multivariate Relevance Vector Machine (MVRVM) (Thayananthan 2005), which is an extension of the RVM algorithm developed by Tipping and Faul (2003). It retains all properties of conventional RVM like sparse modeling, high predictive accuracy, and estimation of uncertainty in the prediction.

3.2 Study Area

Four stream gages were chosen in Utah that spatially covers the state from the northern to the southern region (Figure 3.1). Each gage meets following data assumptions: (i) site flows are not affected by diversion or regulation, and (ii) several years of systematic streamflow records are available. Two gages are chosen from Northern Utah, and one each from Central and Southern Utah. Stream gages at Weber River near Oakley and Chalk Creek at Coalville are chosen from Northern Utah. They both lie in Weber River Basin, a watershed that is composed of a flat, fertile valley east of the Great Salt Lake. The watershed contains approximately 2,060 square miles. The gage at Sevier River at Hatch is chosen from Southern Utah. The river flows North from the headwater and then turns southwest 255 miles before reaching Sevier Lake (Berger et al. 2003). This river basin consists of 12.5 percent of state's total area. The gage at Muddy Creek near Emery is selected from Central Utah which lies in the West Colorado River Basin. It drains portion of Emery and Wayne Counties. The creek begins on the eastern slopes of the Wasatch Plateau. It turns southward near the town of Emery, and then flows along the western edge of the San Rafael Swell. It has an estimated length of 20 miles and a drop of 6000 feet before it combines with the Fremont River to form Dirty Devil River (McCord 1997). The geometric characteristic of watershed and stream for each selected gage are presented in Table 3.1.

3.3 Background

Streamflow depends on the distribution of the precipitation in time and in space as well as in the type and the state of the basin, which, in turn, depends on the climatic conditions (Sivakumar 2003). Annual streamflow is strongly related to long-term climate, therefore, streamflow at this scale may be forecasted using long-term climate information. The inputs representing those climatic conditions are incorporated in the model through popular oceanic-atmospheric oscillation indices. These oscillations have longer persistence, therefore, they are useful for long lead-time annual streamflow volume prediction. These are PDO, ENSO, AMO, and NAO. Use of appropriate oscillation modes is essential to develop a reasonable forecast. Such information is useful to improve the predictive accuracy of a streamflow model. A good forecast of streamflow provides accurate quantity of future water availability that could help water managers for planning and managing in order to maximize the efficiency of water use.

Predicting long lead-time streamflow with a physically based model is complex and is often limited by the extensive requirement of data. Therefore, a data driven model based on the machine learning approach, and using limited amount of data, is proposed here. Asefa et al. (2006) predicted multi-time scale streamflow using Support Vector Machine, Khalil et al. (2005a) predicted streamflow using Artificial Neural Network. This paper uses Multivariate Relevance Vector Machine (MVRVM) to predict annual streamflow volume for next few years in selected streams in Utah.

3.4 Model Description

The approach used for building a model for long-term streamflow prediction is based on a data driven model that uses Multivariate Relevance Vector Machine. This is a model of identical functional form to the Support Vector Machine developed by Vapnik (1995, 1998). The software to develop the model was obtained from Thayananthan (2005), University of Cambridge. This is an extension of the RVM algorithm developed by Tipping and Faul (2003). It retains all properties of conventional RVM, such as sparse modeling, high predictive accuracy, and estimation of uncertainty in the prediction.

For a given input-target pair $\{x_n, t_n\}_{n=1}^{N}$, in the training data set, the model learns the dependency of targets on the inputs with the objective of making accurate predictions of the target (*t*) for previously unseen values of input *x* (Tipping 2000; Tipping 2001).

The targets are assumed to be samples from the model (*y*) with additive noise (ε). The target can be written as sum of approximation vector $y = [y(x_1),..........y(x_N)]^T$ and the error vector $\varepsilon = (\varepsilon_1,........\varepsilon_N)^T$ which is independent samples from some noise process and it is further assumed to be mean-zero Gaussian with variance σ^2. The target vector is given by,

$$t = y + \varepsilon,$$

$$= \Phi w + \varepsilon. \qquad (3.1)$$

where $t = (t_1 ....t_N)^T$, $w = (w_1,...w_i .....w_N)^T$, $\Phi = [\phi(x_1).....\phi(x_N)]^T$, wherein $\phi(x)$ is basis function. The basis function is expressed with a kernel as parameterized by the training vectors. The basis function is thus given by,

$$\phi(x_n) = [1, K(x_n, x_1), K(x_n, x_2),....., K(x_{n,}, x_N)]^T.$$

The target t_n is assumed to be independent so the likelihood of complete dataset is written as,

$$p(t|w,\sigma^2) = (2\pi\sigma^2)^{-N/2} \exp\{\frac{1}{2\sigma^2}\|t - \Phi w\|^2\}. \qquad (3.2)$$

Let τ_i be the *i*th component of the target vector *t*, and w_i be the weight vector for the *i*th component of the output target vector *t*. This is Gaussian distribution which can also be written as,

$$p(t|w,\sigma^2) = \prod_{i=1}^{n} N(\tau_i|\Phi w_i, \sigma_i{}^2).$$

To avoid overfitting, Tipping (2001) imposed some additional constraints on the parameters. The smoother function is made by imposing zero-mean Gaussian prior distribution over *w*. This prior ultimately leads to the sparsity of the model. The prior probability is given by,

$$p(w|\alpha) = \prod_{i=0}^{N} N(w_i|0, \alpha_i{}^{-1}), \qquad (3.3)$$

where $\alpha = (\alpha_0,............\alpha_N)^T$ is a vector of N+1 hyper-parameters. Each α_i controls the strength of the prior over its associated weight (Tipping and Faul 2003).

Bayes' rule is used to compute the posterior over all unknowns given the data,

$$p(w,\alpha,\sigma^2|t) = \frac{p(t|w,\alpha,\sigma^2)p(w,\alpha,\sigma^2)}{p(t)}. \tag{3.4}$$

This term can't be computed in fully analytical form, therefore, an approximation is used. Thus posterior term is decomposed as,

$$p(w,\alpha,\sigma^2|t) = p(w|t,\alpha,\sigma^2)p(\alpha,\sigma^2|t). \tag{3.5}$$

Given the data, the posterior distribution over the weights is Gaussian which is given by (Tipping 2001),

$$\begin{aligned} p(w|t,\alpha,\sigma^2) &= \frac{p(t|w,\sigma^2).p(w|\alpha)}{p(t|\alpha,\sigma^2)}, \\ &= (2\pi)^{-(N+1)/2}|\Sigma|^{-1/2}.\exp\{-\frac{1}{2}(w-\mu)^T\Sigma^{-1}(w-\mu)\}, \qquad (3.6) \\ &= \prod_{i=1}^{N} N(w_i|\mu_i,\Sigma_r). \end{aligned}$$

The posterior covariance and mean of the weight are $\Sigma = (\sigma^{-2}\Phi^T\Phi + A)^{-1}$ and $\mu = \sigma^{-2}\Sigma\Phi^T t$ respectively, where $A = diag(\alpha_0,\alpha_1,........,\alpha_N)$. The key point is that if any $\alpha_m = \infty$, the corresponding $\mu_m = 0$ (Thayananthan et al. 2008).

Some approximation is adapted on the hyper-parameter posterior by the delta function at its mode, i.e., at its most probable values $\alpha_{MP},\sigma^2_{MP}$ (Tipping 2001),

$$\int p(t_*|\alpha,\sigma^2)\delta(\alpha_{MP},\sigma^2_{MP})d\alpha d\sigma^2 \approx \int p(t_*|\alpha,\sigma^2)p(\alpha,\sigma^2|t)d\alpha d\sigma^2. \tag{3.7}$$

The learning then becomes the search for the hyper-parameter posterior mode, i.e., the maximization of $p(\alpha,\sigma^2|t) \propto p(t|\alpha,\sigma^2)p(\alpha)p(\sigma^2)$ with respect to α and σ^2. For uniform hyperpriors over logα and logσ, $p(\alpha,\sigma^2|t) \propto p(t|\alpha,\sigma^2)$. In this case, one needs to maximize only $p(t|\alpha,\sigma^2)$,

$$p(t|\alpha,\sigma^2) = \int p(t|w,\sigma^2)p(w|\alpha)dw,$$

$$= (2\pi)^{-N/2}\left|\sigma^2 I + \Phi A^{-1}\Phi^T \right|^{-1/2} \exp\{-\frac{1}{2}t^T(\sigma^2 I + \Phi A^{-1}\Phi^T)^{-1}t\}. \qquad (3.8)$$

Its maximization is known as type-II marginal likelihood method (Berger 1993). There is no direct solution to estimate the value of α and σ^2 that maximizes Eq. 3.8. Their value is computed by iterative re-estimation which yields,

$$\alpha_i^{new} = \frac{\gamma_i}{\mu_i^2}, \qquad (3.9)$$

where $\gamma_i = 1 - \alpha_i N_{ii}$.

The term μ_i is the *i*th posterior mean weight and N is the number of data points. N_{ii} is the *i*th diagonal element of the posterior weight covariance computed with the current α and σ^2.

The noise variance is re-estimated from,

$$(\sigma^2)^{new} = \frac{\|t - \Phi\mu\|^2}{N - \sum_i \gamma_i}. \qquad (3.10)$$

The learning algorithm proceeds by repeated application of (3.9) to (3.10), together with updating the posterior statistics Σ and μ until some specified convergence criterion is met. It is found that value of α_i generally approaches to infinity, which implies that $p(w_i \mid t, \alpha, \sigma^2)$ becomes highly peaked at zero. This makes the model sparse (Tipping 2001). The relatively nonzero weights correspond to the input vectors that form the sparse core of the RVM model. These input vectors are called relevance vectors (RVs). This sparsity is an effective method to control model complexity, avoid over-fitting and control computational characteristics of model performance (Tipping and Faul 2003).

The predictions are made based on the posterior distribution over the weights, conditioned on the maximizing values α_{MP} and σ_{MP}^2. The predictive distribution for a new input x_* is given by,

$$p(t_* \mid t, \alpha_{MP}, \sigma_{MP}^2) = \int p(t_* \mid w, \sigma_{MP}^2)\, p(w \mid t, \alpha_{MP}, \sigma_{MP}^2)\, dw. \tag{3.11}$$

This is easily computable because both terms in the integral are Gaussian,

$$p(t_* \mid t, \alpha_{MP}, \sigma_{MP}^2) = N(t_* \mid y_*, \sigma_*^2), \tag{3.12}$$

with,

$$y_* = \mu^T \phi(x_*)\,,$$

$$\sigma_*^2 = \sigma_{MP}^2 + \phi(x_*)^T \Sigma \phi(x_*)\,.$$

The total variance consists of sum of the variance of data and uncertainty in estimating weight. Interested readers for Relevance Vector Machine are referred to Tipping (2000), Tipping (2001), Tipping and Faul (2003), Thayananthan (2005), and Thayananthan et al. (2008).

3.5 Data Collection and Description

3.5.1 Streamflow

Unimpaired streamflow data were obtained for Weber River near Oakley, Chalk Creek at Coalville, Sevier River at Hatch, and Muddy Creek near Emery. Monthly average discharges for 1950- 2009 were collected from the U.S. Geological Survey (USGS). The values were then converted to annual flow volume using appropriate conversion factors.

3.5.2 Pacific Decadal Oscillation (PDO)

Pacific Decadal Oscillation (PDO) is a climate phenomenon associated with persistent, bi-modal climate patterns in the North Pacific Ocean. It is an interannual climate index which can be used as an integrator of overall winter climate condition in the North Pacific. The PDO also refers to a numerical climate index based on sea surface temperatures in a particular region of the North Pacific which has an interannual signature (Mantua and Hare 2002). The pattern of PDO is similar to Pacific climate variability of ENSO however it has longer persistence. The warm phase of PDO has similar effects as those of the warm phase of ENSO, and the cold phase PDO has similar effects as those of the cold phase of ENSO. PDO usually persists for 20 to 30 years (a particular phase of PDO typically persists for 25 years). Both indices have similar spatial climate fingerprints, but they have different behavior in time (www.jisao.washington.edu/pdo). The climatic fingerprint of PDO is most visible in the North Pacific region and a secondary signature exists in the tropics. This is opposite for ENSO. Monthly PDO data were obtained from the Joint Institute Study of Atmosphere and Ocean, University of Washington (www.jisao.washington.edu/pdo), and annual averages were computed for 1945-2009 (Figure 3.2a).

3.5.3 El-Niño Southern Oscillation (ENSO)

The El-Niño Southern Oscillation is a complex ocean/atmospheric interaction that causes cyclical patterns of warming and cooling of the sea surface in the tropical Pacific with pronounced global climatic teleconnection (Daly 2008). El-Niño is a warm-phase, and La Niña is a cold phase. ENSO has characteristic return frequency of 4 to 6 years, and usually persists for 1 to 2 years. The Southern Oscillation is the oscillation of surface air pressure between the eastern and western tropical Pacific. When the surface pressure is high in the eastern tropical Pacific, it is low in the western tropical Pacific, and vice-versa. El-Niño is responsible for flooding in some regions, while at the same time producing droughts in other regions. Several studies show it is associated with the streamflow variability in the western United States (Piechota et al. 1997). Not all El-Niño events are of the same intensity nor does the atmosphere always react in the same way from one El-Niño to another. There are several ways ENSO may be represented. Southern Oscillation index (SOI) is one way to represent it (Poveda et al. 2001), which is used in this chapter. The SOI is computed from the monthly fluctuation in air pressure difference between Tahiti and Darwin, Australia. The monthly SOI values were collected from www.cdc.noaa.gov/ENSO/ for 1945-2009. Annual averages were computed from the monthly value for the entire analysis period (Figure 3.2b).

3.5.4 Atlantic Multi-decadal Oscillation (AMO)

The AMO index was introduced by Enfield et al. (2001) as a simple basin average of North Atlantic Ocean (0-70^{o}) sea surface temperature (SST) anomaly. It consists of detrended SST anomalies for the previously defined Atlantic Ocean region. It is a near-global scale mode of observed multi-decadal climate variability with alternating warm and cool phase over large parts of the Northern Atlantic Ocean, with cool and warm phases that may last for 20 to 40 years at a time and a difference of about 1°F between extremes. Many prominent examples of regional multidecadal climate variability have been related to AMO. It affects air temperature and rainfall and river discharge over much of the Northern Hemisphere, in particular, North America and Europe (Enfield et al. 2001; McCabe et al. 2004; Sutton and Hodson 2005). When the AMO is in its warm phase, droughts tend to be more frequent and severe and vice-versa for negative AMO for North America. The unsmoothened monthly AMO data were obtained from www.cdc.noaa.gov/ClimateIndices/List/. The annual average of AMO was computed for 1945-2009 (Figure 3.2c).

3.5.5 North Atlantic Oscillation (NAO)

NAO is a dominant mode of winter climate variability in the North Atlantic region ranging from Central North America to Europe and much into Northern Asia (http://www.ldeo.columbia.edu/res/pi/NAO/). This is a large scale see-saw in atmospheric mass between the subtropical high and polar low. The positive NAO means below normal pressure across the high latitudes of the North Atlantic, and above normal pressure over the Central North Atlantic, Eastern United States, and Western Europe.

This is opposite for its negative phase. The positive phase of NAO is associated with above-average temperature in the Eastern United States and across Northern Europe and below average temperature in Greenland and Europe. It has pronounced effect in regional changes in precipitation patterns (Dai et al. 1997; Hurrell 1995). The NAO index varies from year to year, but also exhibits a tendency to remain in one phase for intervals lasting several years. Monthly average NAO data were obtained from the National Center for Atmospheric Research (www.cgd.ucar.edu/cas/jhurrell/indices.html), and its annual averages were computed for 1945-2009 (Figure 3.2d).

3.6 Model Development and Performance Criteria

Inputs and output to the model are preprocessed according to the model requirements. The input consists of different combinations of annualized ocean-atmospheric oscillation indices and output is annualized streamflow volume. The oscillation indices at time step t is used to predict annual streamflow volume at time step $t + i$, where i= 1, 2, …5, in years. The data is divided into two parts: Training and Testing. The period 1950 to 1999 is used for training the model, and the period 2000 to 2009 is used for testing. The model parameter is optimized in the training phase and the performance of the model is measured based on root mean square error, correlation coefficient, and efficiency in the test phase. A Gaussian kernel is used in all model types. This is a widely used kernel function in learning machines. The input is feed to the Multivariate Relevance Vector Machine and annualized streamflow volume is predicted at specified lead time ($t + i$).

Different model types are developed based on the different combination of oscillation modes in the input. Model 1 consists of using all four oscillation indices (PDO, ENSO, AMO, and NAO). This produces one model run for each lead-time. Model 2 consists of dropping one oscillation index and using the remaining three oscillations to develop the model. This results in four model runs for each lead-time. Model 3 consists of dropping two oscillation indices and using remaining pair to develop the model. This results in a total of six model runs for a given lead-time. Model 4 consists of using only one oscillation index at a time. This results in four model runs for each lead-time. Model 1 is a base case, while Model 2 to Model 4 gives the relative influence of ocean-atmospheric oscillation indices for annual streamflow volume prediction for each selected gage. For each model type, the combination of oscillation indices and lead time corresponding to the best test result is identified. This optimal combination is used to develop forecast model for long lead-time annual streamflow volume prediction. For each lead time, the combination of oscillations that develops the best prediction is also determined. This shows the relative influence of oscillation for each lead time for each selected gage. The sensitivity analysis is performed to categorize the signal strength of each oscillation index for each selected stream gage. The prediction result from Multivariate Relevance Vector Machine is compared with Artificial Neural Network and Support Vector Machine. The comparison shows the relative performance of MVRVM to SVM and ANN.

The performance of the model is evaluated based on RMSE, correlation coefficient and efficiency in test phase.

3.6.1 Root mean square error (RMSE)

RMSE is a commonly used measure for model accuracy. Smaller the RMSE value, better the prediction result is. The best value of RMSE is zero.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(t_i - t_i^*)^2}{n}} \tag{3.13}$$

where t_i observed value, t_i^* is prediction from the model and *n* is sample size.

3.6.2 Pearson correlation coefficient (*r*)

The correlation coefficient (*r*) measures the linear dependence between two variables. It may take any value between -1 and 1. The correlation coefficient close to ±1 indicates strong correlation and close to zero indicates weak correlation. The correlation coefficient is given by,

$$r = \frac{\sum_{i=1}^{n}(t_i - \overline{t})(t_i^* - \overline{t^*})}{\sqrt{\sum_{i=1}^{n}(t_i - \overline{t})^2 \sum_{i=1}^{n}(t_i^* - \overline{t^*})^2}}, \tag{3.14}$$

where $\bar{t}$ and $\overline{t^*}$ are mean observed value and mean modeled value, respectively.

3.6.3 Nash-Sutcliffe efficiency

The Nash-Sutcliffe efficiency is commonly used to evaluate the performance of hydrologic models. It is mathematically expressed as (Nash and Sutcliffe 1970),

$$E = 1 - \frac{\sum_{i=1}^{n}(t_i - t_i^*)^2}{\sum_{i=1}^{n}(t_i - \bar{t})^2}. \tag{3.15}$$

Nash-Sutcliffe efficiency ranges from negative infinity to 1. An efficiency of 1 corresponds to a perfect match of model prediction to observed data. An efficiency of zero indicates that the model prediction is as accurate as the mean of the observed data. The negative efficiency indicates the observed mean is a better predictor than the model.

3.7 Results

3.7.1 Correlation coefficients

The absolute value of correlation coefficient between ocean-atmospheric oscillation indices and annual streamflow volume for 1 to 5 year lead is shown in Figure 3.3. The strength of the observed correlations suggests that large-scale climate indices may be useful predictors for future annual streamflow volume prediction. PDO has the highest correlation coefficient at a 3-year lead for the gage at Weber River near Oakley and that of Chalk Creek at Coalville, however, it is at a 1-year lead for the gage at Sevier River at Hatch and that of Muddy Creek near Emery. ENSO has highest correlation coefficient at a 1-year lead for Sevier River at Hatch and Muddy Creek near Emery, this is however, comparable to 3 and 4 years lead respectively. For Weber River near Oakley and Chalk Creek at Coalville, relatively higher correlation coefficient is obtained at the 4-year lead. AMO has relatively higher correlation coefficient at 4- year lead for Sevier River at Hatch and Muddy Creek near Emery, however, relatively higher correlation coefficients are obtained at 1- and 2-year lead for the remaining stream gages. Comparison among longer lead time (3, 4 and 5 years) shows that higher correlation coefficients are obtained at 4 year lead. For NAO, the correlation coefficients increase from 1 to 2 year lead and then drops at 3 and 4 year lead except for Chalk Creek at Coalville, where the correlation coefficient continuously increases from 1 to 5 year lead.

Overall results show that relatively stronger correlation is obtained at 3- and 4 -year lead for most of the cases. The tradeoff between the correlation coefficient and lead-time shows a 3 or 4 year lead may be the optimal lead time for developing prediction models because it gives longer lead of forecast and performs better.

3.7.2 Identification of best combination of oscillations and lead time for annual streamflow volume predictions

Different models are developed based on the combination of oscillation indices in the input. The performance of the model for the different combinations is evaluated based on test phase RMSE. Figure 3.4 through Figure 3.7 shows test phase RMSE for 1- to 5-year lead for Model 1 to Model 4 respectively. Since relatively higher correlation between individual oscillation index and annual streamflow volume is observed at 3- and 4-year lead, it is quite obvious to obtain better results at similar lead times.

Model 1

Model 1 is a base case, where all four ocean-atmospheric oscillations are used to predict annul streamflow volume. For each gage, annual flow volume is predicted at 1 to 5-year lead. Figure 3.4 shows the plot of test RMSE versus lead time for annual streamflow volume prediction. The smallest test RMSE is obtained at 4-year lead time for Weber River near Oakley and Muddy Creek near Emery. It is however obtained at 1-year lead for Chalk Creek at Coalville. The second and third best test RMSE are obtained at 3- and 4-year lead respectively. Since 1-year lead prediction is not very long lead prediction, 3 and 4-year lead may be used to develop the prediction model. For Sevier River at Hatch, the test RMSE is small at 3- and 5-year lead.

Model 2

Model 2 consist of dropping one oscillation index and using remaining three to develop prediction model. This consists of four model runs for each lead time. Figure 3.5 shows the test RMSE for Model 2 at 1 to 5-years lead for each gage. The smallest test RMSE was obtained at the 4-year lead for Weber River near Oakley. This input corresponds to dropping NAO and using remaining three oscillation modes. Smallest test RMSE was again obtained at the 4-year lead for Chalk Creek at Coalville by dropping AMO. Dropping PDO at 4-year lead produces similar test RMSE. For Sevier River at Hatch, 3-year lead produces reasonable model prediction. This corresponds to dropping NAO and using remaining oscillation indices. The best test RMSE, however, is obtained at 2-year lead, where the input corresponds to dropping AMO. Comparable result is obtained by dropping PDO at 5-year lead. For Muddy Creek near Emery, 3- and 4-year lead produces relatively better result than other ones.

Model 3

Model 3 is developed by using a pair of ocean atmospheric oscillation modes at a time. This results in six model runs for each lead time. Figure 3.6 shows the test RMSE for Model 3 for 1- to 5-years lead. The best test RMSE is obtained from the pair of PDO+ENSO at 4-year lead for Weber River near Oakley. Three-year lead develops the best test RMSE for Chalk Creek at Coalville from ENSO+AMO pair, however, comparable result is obtained at 4-year lead. For Sevier River at Hatch, a pair of

PDO+NAO develops the best test RMSE at 2-year lead. PDO+ENSO also develop reasonable result at 3-year lead. PDO+NAO develop the best test RMSE at 2-year lead for Muddy Creek near Emery. Out of 6 combinations, 3 combinations results poor test RMSE at 2-year lead. The test RMSE at 4-year lead is relatively better than that of 3- and 5-year lead, which corresponds to ENSO+AMO for input variables.

Model 4

Model 4 consist of using only one oscillation index at a time. This results in four model runs for each gage for each lead-time. Figure 3.7 shows the test RMSE for Model 4 for each gage. ENSO develops the best model at 4 year lead for Weber River near Oakley. Comparable results are obtained from AMO at same lead time. For Chalk Creek at Coalville, AMO produces relatively smaller test RMSE than other oscillation indices. ENSO at 4 year lead develops comparable result. For Sevier River at Hatch, PDO produces the best test RMSE at 2-year lead. Next to it, ENSO develops the best result at 4-year lead. For Muddy Creek near Emery, ENSO, and PDO produces relatively better test RMSE at 1- and 2-year lead but when compared among 3-, 4-, and 5-year lead, ENSO and AMO predicts relatively better at 4-year lead.

3.7.3 Prediction results from best identified combinations of oscillations for each model types

The best model for each gage for each model type is presented in Tables 3.2 through 3.5. They are evaluated based on RMSE, correlation coefficient, and efficiency in test phase. The table also presents the combination of the oscillation indices and lead-time for the best model. In general, 4-year lead produces the best test results, which develops reasonable model prediction and gives the long-lead forecast as well. This lead-time is consistent with the correlation analysis performed in the earlier section.

Using the best combinations of oscillations and corresponding lead time as shown in Table 3.2 through 3.5, annual streamflow volume is predicted for each gage for Model 1 to Model 4 respectively. Prediction plots for each selected gage for Model 1 to Model 4 are presented in Figures 3.8 through 3.11, respectively. For rows (a) to (d), the first and second columns show the plot for training and test phases, respectively. The third column shows the actual versus predicted annual flow volume for training phase, and the fourth column shows similar plot for the test phase. The model prediction is said to have good agreement with actual flow volume if the points saturates about the 45° line. The other line is a trend line. The first, second, third, and fourth rows corresponds to Weber River near Oakley, Chalk Creek at Coalville, Sevier River at Hatch, and Muddy Creek near Emery respectively. The fifth row shows 90 percent confidence interval of the mean prediction and sixth row is the residual plot for test phase.

The results show the model has predicted annual flow volume reasonably well using ocean-atmospheric oscillation modes. A reasonable agreement is obtained between the actual volume and predicted volume. The plot of predicted versus actual streamflow volume shows the points are saturated about the 45-degree line, except for extreme flows. This is because the oscillation modes do not fully represent the underlying physical processes responsible for generation of streamflow. The residuals are relatively high for low flow but shows randomness in other flows in most of the plots, which is an indication that the model does not contain serious modeling problem. In general, this prediction gives an idea about the future water availability which could be useful for planning and management of water for future in basin scale.

3.7.4 Discussion and relative influence of oscillation indices

Ocean-atmospheric oscillation indices carry important information about climate, hence, the hydrology of river basins in the many regions of the world can be correlated to those indices. These oscillation indices have long-term persistence, thus they can be used for long lead-time streamflow prediction. It is important to identify the influential and effective oscillation indices for a given stream gage location in order to predict the long lead-time streamflow reasonably well. Different model types were developed based on the combinations of those indices through a MVRVM model. Model 1 is a base case where all oscillations are used, while Models 2 through 4 used different combinations of oscillations. Results from the models show that the long lead-time streamflow is predicted satisfactorily for each of the selected gages. Comparing Model 2 to Model 4 with the base case the relative influence of each oscillation index for each selected gage is estimated subjectively. They are categorized into weak, marginal and strong for each gage. The effect is said to be weak if the oscillation index doesn’t improve the prediction results compared to the base case (Model 1). If a particular index marginally improves the prediction results compared to the base case, the signal is said to have marginal strength. Finally, the signal is strong if it significantly improves the prediction results. For the comparison, a 4-year lead time was chosen for all stream gages, except Sevier River at Hatch, where a 3-year lead time is chosen.

In Model 1, the best model prediction is obtained at a 4-year lead for all stream gages except the one at Sevier River at Hatch where best model prediction is obtained at a 3-year lead. The correlation coefficients between actual and predicted annual streamflow volume for the test phase are 0.53, 0.39, 0.56 and 0.38 for Weber River near Oakley, Chalk Creek at Coalville, Sevier River at Hatch, and Muddy Creek near Emery respectively. Similarly, their corresponding RMSE in test phase are 32.88, 16.76, 58.2, and 9.97 kilo ac-ft respectively. The best prediction result based on correlation coefficient is obtained at Sevier River at Hatch. The second best result is obtained at Weber River near Oakley, and the third best result is obtained at Chalk Creek at Coalville.

In Model 2, the best test RMSE for the gages at Weber River near Oakley, Chalk Creek at Coalville, Sevier River at Hatch, and Muddy Creek near Emery are 29.44, 13.85, 57.22 and 9.26 kilo ac-ft, respectively. Their corresponding correlation coefficients are 0.67, 0.45, 0.62 and 0.58. In Model 2, best correlation is obtained at Weber River near Oakley. The second best correlation coefficient is obtained at Sevier River at Hatch. If 2-year lead is considered, the best test correlation coefficient (0.78) is obtained by dropping AMO for Sevier River at Hatch.

For comparing Model 2 over Model 1, previously specified lead times were used. Based on the test RMSE, the model prediction showed good improvement over Model 1 when NAO was dropped for the gage at Weber River near Oakley. Reasonable improvement was obtained when PDO was dropped. The model prediction for Model 2 marginally deteriorates when AMO was dropped. However, the prediction deteriorated significantly when ENSO was dropped. For Chalk Creek at Coalville, significant improvement was obtained in model prediction compared to Model 1 by dropping AMO, and PDO. Marginal improvement was obtained by dropping NAO, and ENSO. For Sevier River at Hatch, the prediction improved by dropping NAO. The result marginally deteriorated by dropping PDO, and it deteriorated significantly by dropping ENSO. For Muddy Creek near Emery, the prediction result marginally deteriorated by dropping PDO, and NAO individually. The result, however, significantly deteriorated when ENSO and AMO were dropped. If a 3-year lead is considered for the comparison, the model result significantly improved by dropping AMO compared to Model 1. The prediction marginally improved by dropping PDO and NAO, however, the result marginally deteriorated by dropping ENSO.

Summarizing:

- ⇨ In the learning machine approach, the model prediction deteriorates by the use of trivial predictor variables. Since the prediction result improved by dropping NAO compared to Model 1 for Weber River near Oakley, NAO is not an influential ocean-atmospheric oscillation mode for annual streamflow volume prediction at this location. PDO and AMO have marginal influence, while ENSO has strong influence because the prediction results significantly deteriorated compared to Model 1 when it was dropped.
- ⇨ For Chalk Creek at Coalville, AMO and PDO are not influential oscillation indices because prediction results improved, compared to Model 1, when they were dropped. ENSO and NAO, however, had marginal influence as prediction results marginally improved compared to Model 1 when they were dropped.
- ⇨ NAO is not an influential oscillation index for Sevier River at Hatch because the prediction result improved when it was dropped. Since the result marginally

deteriorated by dropping PDO, it may have marginal influence on annual flow volume prediction. The prediction results, however, deteriorated significantly by dropping ENSO. Therefore, ENSO has a relatively stronger signal for annual flow volume prediction in the Sevier River at Hatch.

- ⇨ For Muddy Creek near Emery, PDO, and NAO has a marginal influence, and ENSO has a relatively stronger influence for annual flow volume prediction.

Based on the correlation coefficient between actual and predicted volumes in test phase, the overall result of Model 3 improved over Model 1. The correlation coefficient in the test phase for the gages at Weber River near Oakley, Chalk Creek at Coalville, Sevier River at Hatch, and Muddy Creek near Emery are 0.72, 0.61, 0.87, and 0.82 respectively. Similarly, corresponding best test RMSE for those stream gages are 33.95, 19.01, 41.68, and 7.13 kilo ac-ft. The best prediction was obtained from the combination of PDO and ENSO for Weber River near Oakley. For Chalk Creek at Coalville, ENSO and AMO produced the first best model, while PDO and ENSO produced the second best model predictions. For Sevier River at Hatch and Muddy Creek near Emery, PDO and NAO produced the best model prediction. Comparing, as before, at previously specified lead time, the combination of PDO and ENSO produced similar prediction results as those of Model 1, however, other pair deteriorated prediction results for Weber River near Oakley (Figure 3.7). For Chalk Creek at Coalville, prediction results marginally improved while using the ENSO and AMO combination. Other pairs of oscillation modes deteriorated the prediction result, in comparison. For Sevier River at Hatch, results significantly improved by using the PDO and ENSO pair, compared to Model 1. The prediction result from the pair of ENSO and AMO was marginally different from prediction results from Model 1. Other combinations of oscillation indices deteriorated the prediction for this particular location. For Muddy Creek near Emery, the combination of ENSO and AMO significantly improved the model prediction compared to Model 1. Next to it, the combination of PDO and AMO produced comparable results, however, this model was marginally poorer than Model 1. Other combinations significantly deteriorated the prediction result for this location.

Summarizing:

- For Weber River near Oakley, the combination of PDO+ENSO developed similar model predictions as that of Model 1. Therefore, they may be considered as influential ocean-atmospheric oscillation indices for this location. The combination of PDO with NAO, and PDO with AMO deteriorated the model prediction. The combination of ENSO with NAO, and ENSO with AMO also deteriorated the model prediction. This shows that NAO and AMO do not have strong influence on annual streamflow volume prediction at Weber River near Oakley. However, PDO and ENSO have relatively strong influence in this location.
- For Chalk Creek at Coalville, marginal improvement was obtained by using the ENSO+AMO pair over the base case, Model 1. The prediction marginally deteriorated from the combination of PDO+ENSO. Other combination pairs significantly deteriorated the prediction results. This shows that ENSO and PDO have a marginal influence, while other indices have a weak influence on annual streamflow prediction for Chalk Creek at Coalville.

- ⇨ For Sevier River at Hatch, the combination of PDO+ENSO improved the model prediction compared to Model 1. Other combination pairs deteriorated the prediction results. Some pairs marginally deteriorated the prediction results, while other pairs did so significantly. These results show that PDO and ENSO have a relatively strong influence on annual flow volume prediction for Sevier River at Hatch, while other indices do not have such influence.
- ⇨ For Muddy Creek near Emery, the combination of ENSO and AMO improved the model prediction. Next to it, a combination of PDO and AMO developed the second best model prediction. ENSO and AMO, therefore, have relatively stronger influence and PDO has marginal influence.

In Model 4, the best correlation coefficients in test phase for Weber River near Oakley and Chalk Creek at Coalville are 0.40 and 0.30, respectively. For Sevier River at Hatch, the correlation coefficient is 0.60 for the second best model, and 0.84 for the best model. For Muddy Creek near Emery, the correlation coefficient is 0.84 for the best model, and 0.35 for the second best model. The best RMSE in the test phase for Weber River near Oakley, Chalk Creek at Coalville, Sevier River at Hatch and Muddy Creek near Emery are 36.86, 19.72, 57.09, and 9.13 kilo ac-ft, respectively. Again, a previously specified lead time was used to compare Model 4 with the base case.

For the gage at Weber River near Oakley, Model 4 did not improve the prediction results, but deteriorated them compared to Model 1. The predictions by using ENSO only, in this gage, were relatively better than using other oscillation modes. For Chalk Creek at Coalville, the result did not improve, however, the ENSO and AMO indices, individually, perform relatively better than other individual oscillation modes. For Sevier

River at Hatch, ENSO significantly improved the model prediction compared to Model 1 for a 4-year lead. ENSO and PDO, individually, marginally deteriorated the results, while AMO and NAO significantly deteriorated the results for a 3-year lead. For Muddy Creek near Emery, there is no significant improvement in model prediction, compared to Model 1, by using any of the oscillation indices individually. However, for this location, the prediction from ENSO and AMO, used individually, performed relatively better than other individual oscillation indices.

Summarizing:

- ⇨ For Weber River near Oakley, the prediction from Model 4 did not improve but deteriorated the results compared to Model 1. However, the model predictions by ENSO were relatively better than prediction from other oscillation modes. Therefore ENSO is said to have a relatively stronger signal than other oscillation modes for annual streamflow volume prediction in this location.
- ⇨ For Chalk Creek at Coalville, the prediction results did not improve compared to base case. However, ENSO, and AMO performed relatively better than other oscillation modes. Thus ENSO and AMO have marginal influences in this location, while PDO and NAO have weak signals.
- ⇨ For Sevier River at Hatch, ENSO, and PDO marginally deteriorated model predictions while AMO and NAO significantly deteriorated them. These results show that ENSO and PDO have marginal influence, while NAO and AMO do not have an influential signal in this location.
- ⇨ For Muddy Creek near Emery, ENSO and AMO performed marginally better than other oscillation modes. ENSO and AMO, thus, have relatively influential signals,

while the remaining two indices do not have influential signals for Muddy Creek near Emery.

The strength of oceanic-atmospheric oscillations indices at each gage for Model 2 to Model 4 are shown in Table 3.6 through Table 3.8. The results show that PDO and ENSO have relatively stronger signals than other oscillations, in general. PDO and ENSO possess strong to marginal influence for most of the stream gages. NAO and AMO, however, have weak to marginal signals for most of stream gages.

In addition to fixing the lead time and finding the combinations of oscillation indices that produce the best predictions, the best combination of oscillations are also identified for each lead time in the range of 1- to 5-years (Table 3.9). The best predictions for each lead time resulted from different combinations of input indices at different locations of stream gages. This analysis shows that various combinations of oscillation indices can be used to enhance the predictions for different lead time. ENSO and PDO, however, frequently appeared than other oscillation indices in developing best model for long lead-time annual streamflow volume prediction.

3.7.5 Comparison with SVM and ANN

The prediction results of streamflow from MVRVM were compared to corresponding SVM and ANN in each model type (Model 1 to Model 4) for each selected gage. In general, MVRVM has predicted relatively better than SVM and ANN, however, the pattern of prediction are similar in each model type. The software to develop SVM model was obtained from SVM and Kernel Methods Matlab Toolbox (http://asi.insa-rouen.fr/enseignants/~arakotom/toolbox/index.html). The software to develop ANN model was obtained from Aston University Engineering and Applied Science (http://www1.aston.ac.uk/eas/research/groups/ncrg/resources/netlab/downloads/). This ANN model uses a Bayesian approach. Figure 3.12 shows the comparison of MVRVM results to SVM and ANN based on RMSE on test phase.

- For Weber River near Oakley, MVRVM outperforms ANN and SVM in Model 3 and Model 4, while ANN and SVM performs relatively better in Model 1. For Model 2, MVRVM performed better than ANN, but slightly poorer than SVM.
- For Chalk Creek at Coalville, MVRVM outperforms ANN and SVM for Model 2, while SVM outforms others for Model 1. For rest of model types, prediction from MVRVM is better than ANN, however, prediction result for SVM and MVRVM are very similar.
- For Sevier River at Hatch, MVRVM outperforms ANN and SVM for all model types (Model 1- Model 4).
- For Muddy Creek near Emery, the prediction result is not very different among three model types, but MVRVM performs relatively betters than ANN and SVM in all model types.

3.7.6 Generalization and robustness of model

Bootstrap analysis gives the estimate of variability of test statistics with the change in training data. This analysis shows how robust the model is and how well it generalizes. It is a data-based simulation method for statistical inference (Efron and Tibshirani 1998). The idea is to randomly draw a large number of ‘resamples’ of size *n* from the original sample, with replacement. Although each resample has the same number of elements as the original sample, it may include some of the original data points more than once, and some not included. The process forming the training set is random and the resulting data sets are treated as independent sets (Duda et al. 2000). Each of these resamples randomly departs from the original sample. From each bootstrap set, the bootstrap test statistic is computed in exactly the way as the real sample is used (Davidson and MacKinnon 2001). Since the elements in these resample vary slightly, the statistics calculated from these resample takes on slightly different value. A histogram of computed statistics, e.g., test RMSE is prepared. The width of the histogram is a measure of the robustness of the model. In this paper, bootstrap analysis was used for the best identified model for each selected gage for each model type. A total of 500 bootstrap runs were performed to construct the histograms. 2.5th percentile and 97.5th percentile values of test RMSE were computed. They are shown by the red dotted lines in Figure 3.13. The narrow bound of the resulting histograms shows that the model is robust. The test RMSE of the actual model also lies in between the two red dotted lines. This shows that the developed model is robust and consistent enough to use as a long lead-time streamflow prediction model.

3.8 Conclusion

The relationship between streamflow and climatic variability represented by ocean-atmospheric oscillation indices is a key point for annual streamflow volume prediction. This chapter identifies the best combination of oscillations and lead time for each of four selected stream gages in the state of Utah, and use them for the annual streamflow volume prediction. This chapter also presents the relative influence of each oscillation index at each selected stream gage. The streamflow is predicted at 1- to 5-year leads using Multivariate Relevance Vector Machine (MVRVM), and the prediction results were refined using the optimal combination of oscillations and corresponding lead time. The model prediction showed satisfactorly results. Four Model types were developed. Model 1 is a base case where all four oscillation indices (PDO, ENSO, AMO and NAO) were used. Model 2, 3 and 4 were developed from individual or different combinations of oscillation indices. They may be used to evaluate the relative influence of oscillation indices for annual streamflow volume prediction. The best model prediction was usually obtained at the 4-year lead time. Although relatively better predictions were obtained at a 2- and 3-year lead time in some gages, the 4-year lead time produced comparable results. ENSO and PDO generally predicted better than AMO and NAO for all gages. For the fixed lead time used in this paper (4 year, except for the gage at Sevier River at Hatch), ENSO and PDO showed strong to marginal influence, while AMO and NAO had weak signals for most of the cases, and marginal influences in some cases. The influencial oscillations can be useful to develop the accurate forecast model in the specified location of gages. In addition to this, combination of oscillations that predicts the best results for each lead time were also obtained. Different combinations of oscillations developed best predictions at different lead-time. This information could be used to enhance the model prediction. In general, the model has predicted reasonably well from oceanic-atmospheric oscillation modes. The model however, did not perform well on capturing the extreme events. This shows the oscillation indices used in this research are not enough to represent the physical process associated with the generation of streamflow. The bootstrap analysis was used in order to test the robustness and generalization capability of the model. The narrow bound of the resulting statistics histograms shows that the model is robust. Also, the actual test statistics lies in between 2.5th and 97.5th percentile values, which indicates the model prediction is consistent and well generalized. The predictions from MVRVM were then compared to results from other statistical learning approaches, namely, ANN and SVM. The prediction results showed that MVRVM outperforms ANN and SVM. The pattern of prediction, however, remained the same in all machine learning models.

References

Asefa, T., Kemblowski, M., McKee, M., and Khalil, A. (2006). "Multi-time scale stream flow predictions: The support vector machines approach." *J. Hydrol.*, 318(1-4), 7-16.

Berger, B., Hansen, R., and Jensen, R. (2003). "Sevier River Basin system description." Sevier River Water Users Association, Delta, UT.

Berger, J. O. (1993). *Statistical decision theory and Bayesian analysis/ second edition*, Springer-Verlag, New York.

Chiew, F. H. S., and McMahon, T. A. (2002). "Global ENSO-streamflow teleconnection, streamflow forecasting and interannual variability." *Hydrological Sciences Journal*, 47(3), 505-522.

Dai, A., Fung, I. Y., and Del Genio, A. D. (1997). "Surface observed global land precipitation variations during 1900–88." *J. Climate*, 10(11), 2943-2962.

Daly, J. L. (2008). "The El-Niño Southern Oscillation (ENSO)." <http://www.john-daly.com/elnino.htm>. (Accessed April 17, 2011).

Davidson, R., and MacKinnon, J., G. (2001). "Bootstrap Tests: How many bootstraps?", Queen's University, Department of Economics.

Diaz, H. F., and Markgraf, V. (2000). *El Niño and the Southern Oscillation: Multiscale variability and global and regional impacts*, Cambridge University Press New York.

Duda, R. O., Hart, P. E., and Stork, D. G. (2000). *Pattern Classification*, Wiley Interscience, Second Edition, New York.

Efron, B., and Tibshirani, R. J. (1998). *An introduction of the bootstrap, Monographs on Statistics and Applied Probability*, CRC Press LLC, Boca Raton, FL.

Enfield, D. B., Mestas, Nuñez, A. M., and Trimble, P. J. (2001). "The Atlantic Multidecadal Oscillation and its relation to rainfall and river flows in the continental U.S." *Geophys. Res. Lett*., 28(10), 2077-2080.

Hamlet, A. F., and Lettenmaier, D. P. (1999). "Columbia River streamflow forecasting based on ENSO and PDO climate signals." *J. Water Resour. Plann Manage*., 125(6), 333-341.

Hurrell, J. W. (1995). "Transient Eddy Forcing of the Rotational Flow during Northern Winter." *J. Atmospheric Sciences*, 52(12), 2286-2301.

Kahya, E., and Dracup, J. A. (1993). "U.S. streamflow patterns in relation to the El Niño/Southern Oscillation." *Water Resour. Res*., 29(8), 2491-2503.

Kalra, A., and Ahmad, S. (2009). "Using oceanic-atmospheric oscillations for long lead time streamflow forecasting." *Water Resour. Res*., 45(3), W03413.

Khalil, A. F., McKee, M., Kemblowski, M., and Asefa, T. (2005a). "Basin scale water management and forecasting using Artificial Neural Networks." *JAWRA Journal of the American Water Resources Association*, 41(1), 195-208.

Khalil, A. F., McKee, M., Kemblowski, M., and Asefa, T. (2005b). "Sparse Bayesian learning machine for real-time management of reservoir releases." *Water Resour. Res*., 41(11), W11401.

Khalil, A. F., McKee, M., Kemblowski, M., Asefa, T., and Bastidas, L. (2006). "Multiobjective analysis of chaotic dynamic systems with sparse learning machines." *Adv. Water Res*., 29(1), 72-88.

Mantua, N. J., and Hare, S. R. (2002). "The Pacific Decadal Oscillation." *J. Oceanography*, 58(1), 35-44.

McCabe, G., Palecki, M., Betancourt, J., and Fung, I. (2004). "Pacific and Atlantic Ocean influences on multidecadal drought frequency in the United States." *Proc., National Academy of Sciences of the United States of America*, 101(12), 4136-4141.

McCord, M. W. (1997). "Southwest Paddler, Outdoor recreation guide for Utah ", <http://southwestpaddler.com/docs/muddyut.html>. (Accessed 11 May 2010).

Mochizuki, T., Ishii, M., Kimoto, M., Chikamoto, Y., Watanabe, M., Nozawa, T., Sakamoto, T., Shiogama, H., Awaji, T., Sugiura, N., Toyoda, T., Yasunaka, S., Tatebe, H., and Mori, M. (2010). "Pacific decadal oscillation hindcasts relevant to near-term climate prediction." *Proc., National Academy of Sciences*, 107(5), 1833-1837.

Nash, J. E., and Sutcliffe, I. V. (1970). River flow forecasting through conceptual models.

Philander, S. G. (1990). *El Niño, La Niña, and the Southern Oscillation*, Academic Press, San Diego, California.

Piechota, T. C., Dracup, J. A., and Fovell, R. G. (1997). "Western US streamflow and atmospheric circulation patterns during El Niño-Southern Oscillation." *J. Hydrol.*, 201(1-4), 249-271.

Poveda, G., Jaramillo, A., Gil, M. M., Quiceno, N., and Mantilla, R. I. (2001). "Seasonally in ENSO-related precipitation, river discharges, soil moisture, and vegetation index in Colombia." *Water Resour. Res.*, 37(8), 2169-2178.

Sivakumar, B. (2003). "Forecasting monthly streamflow dynamics in the western United States: a nonlinear dynamical approach." *Environ. Model. Software*, 18(8-9), 721-728.

Soukup, T. L., Aziz, O. A., Tootle, G. A., Piechota, T. C., and Wulff, S. S. (2009). "Long lead-time streamflow forecasting of the North Platte River incorporating oceanic–atmospheric climate variability." *J. Hydrol.*, 368(1-4), 131-142.

Sutton, R. T., and Hodson, D. L. R. (2005). "Atlantic Ocean forcing of North American and european summer climate." *Science*, 309(5731), 115-118.

"SVM and Kernel Methods Matlab toolbox." <http://asi.insa-rouen.fr/enseignants/~arakotom/toolbox/index.html>. (Accessed 15 May 2008).

Thayananthan, A. (2005). "Template-based Pose Estimation and Tracking of 3D Hand Motion." PhD Dissertation, University of Cambridge, Cambridge, UK.

Thayananthan, A., Navaratnam, R., Stenger, B., Torr, P., and Cipolla, R. (2008). "Pose estimation and tracking using multivariate regression." *Pattern Recognition Lett.*, 29(9), 1302-1310.

Tipping, M. (2000). "The Relevance Vector Machine." *Proc., Advances in Neural Information Processing Systems*, The MIT Press, 652-658.

Tipping, M. (2001). "Sparse Bayesian Learning and the Relevance Vector Machine." *J. Machine Learning Res.*, 1, 211-244.

Tipping, M. E., and Faul, A. C. (2003). "Fast marginal likelihood maximization for sparse Bayesian models." *Proc., Ninth International Workshop on Artificial Intelligence and Statistics*.

Vapnik, V. N. (1995). *The nature of statistical learning theory,* Springer Verlag, New York.

Vapnik, V. N. (1998). *The nature of statistical learning theory*, Springer Verlag, New York.

Table 3.1 Geometric characteristic of stream gages

Site ID	Name	Basin	Stream		Gage locations	
		Area (mi^2)	Length (mi)	Slope	Lat (°)	Long (°)
10128500	Weber River near Oakley	162.1	25.3	0.020	40.737	-111.247
10131000	Chalk Creek at Coalville	248.3	37.5	0.010	40.921	-111.401
10174500	Sevier River at Hatch	340	31.1	0.007	37.651	-112.430
09330500	Muddy Creek near Emery	105	20.1	0.004	38.982	-111.249

Table 3.2 Best statistics, combination of oscillations, and lead time for Model 1

Stream gage	Correlation		Test RMSE	Efficiency	Lead	Combination of indices
	Train	Test	(1000 ac-ft)		(year)	
Weber River near Oakley	0.43	0.53	32.88	0.08	4	All
Chalk Creek at Coalville	0.48	0.39	16.76	-	4	All
Sevier River at Hatch	0.34	0.56	58.2	0.22	3	All
Muddy Creek near Emery	0.37	0.38	9.97	0.23	4	All

Table 3.3 Best statistics, combination of oscillations, and lead time for Model 2

Stream gage	Correlation		Test RMSE	Efficiency	Lead	Combination of indices
	Train	Test	(1000 ac-ft)		(year)	
Weber River near Oakley	0.39	0.67	29.42	0.261	4	Dropping NAO
Chalk Creek at Coalville	0.94	0.45	13.85	0.202	4	Dropping AMO
Sevier River at Hatch	0.62	0.62	57.22	0.246	3	Dropping NAO
Muddy Creek near Emery	0.58	0.51	9.26	0.332	3	Dropping AMO
Sevier River at Hatch	0.73	0.82	47.87	0.473	2	Dropping AMO*

Table 3.4 Test statistics, combination of oscillations, and lead time for Model 3

Stream gage	Correlation		Test RMSE	Efficiency	Lead	Combination of indices
	Train	Test	(1000 ac-ft)		(year)	
Weber River near Oakley	0.49	0.72	33.95	0.02	4	PDO+ENSO
Chalk Creek at Coalville	0.57	0.61	19.01	-	4	PDO+ENSO
Sevier River at Hatch	0.62	0.87	41.68	0.60	2	PDO+NAO
Muddy Creek near Emery	0.90	0.82	7.13	0.61	2	PDO+NAO

Table 3.5 Test statistics, combination of oscillations, and lead time for Model 4

Stream gage	Correlation		Test RMSE	Efficiency	Lead	Combination of indices
	Train	Test	(1000 ac-ft)		(year)	
Weber River near Oakley	0.41	0.40	36.86	-	4	ENSO
Chalk Creek at Coalville	0.45	0.30	19.72	-	4	ENSO
Sevier River at Hatch	0.49	0.60	58.26	0.218	4	ENSO
Muddy Creek near Emery	0.14	0.35	10.15	0.199	4	AMO
Sevier River at Hatch	0.36	0.84	57.09	0.249	2	PDO*
Muddy Creek near Emery	0.47	0.84	9.13	0.351	2	PDO*

*Second model

Table 3.6 Relative strength of oscillation modes from Model 2

Stream gage	Lead time (year)	Signal strength		
		Stronger	Marginal	Weak
Weber River near Oakley	4	ENSO	PDO, AMO	NAO
Chalk Creek at Coalville	4	-	ENSO, NAO	AMO, PDO
Sevier River at Hatch	3	ENSO	PDO	NAO
Muddy Creek near Emery	4	ENSO	PDO, NAO	AMO

Table 3.7 Relative strength of oscillation modes from Model 3

Stream gage	Lead time (year)	Signal strength		
		Stronger	Marginal	Weak
Weber River near Oakley	4	PDO, ENSO	-	AMO, NAO
Chalk Creek at Coalville	4	-	PDO, ENSO	AMO, NAO
Sevier River at Hatch	3	ENSO	PDO	AMO, NAO
Muddy Creek near Emery	4	ENSO, AMO	PDO	NAO

Table 3.8 Relative strength of oscillation modes from Model 4

Stream gage	Lead time (year)	Signal strength		
		Stronger	Marginal	Weak
Weber River near Oakley	4	-	ENSO	PDO, AMO, NAO
Chalk Creek at Coalville	4	-	ENSO, AMO	PDO, NAO
Sevier River at Hatch	3	-	ENSO, PDO	AMO, NAO
Muddy Creek near Emery	4	-	ENSO, AMO	PDO, NAO

Table 3.9 Combination of oscillations that produces best results for each lead time

Lead time (year)	Weber River near Oakley	Chalk Creek at Coalville	Sevier River at Hatch	Muddy Creek near Emery
1	ENSO and AMO	PDO, ENSO, AMO, and NAO	ENSO, AMO, and NAO	ENSO, AMO, and NAO
2	ENSO,AMO, and NAO	PDO and ENSO	PDO and NAO	PDO and NAO
3	ENSO,AMO, and NAO	PDO, ENSO, and NAO	PDO, ENSO, and AMO	PDO, ENSO, and NAO
4	PDO, ENSO and AMO	PDO, ENSO, and NAO	PDO, ENSO, and AMO	ENSO and AMO
5	ENSO and AMO	ENSO, AMO, and NAO	ENSO and AMO	AMO

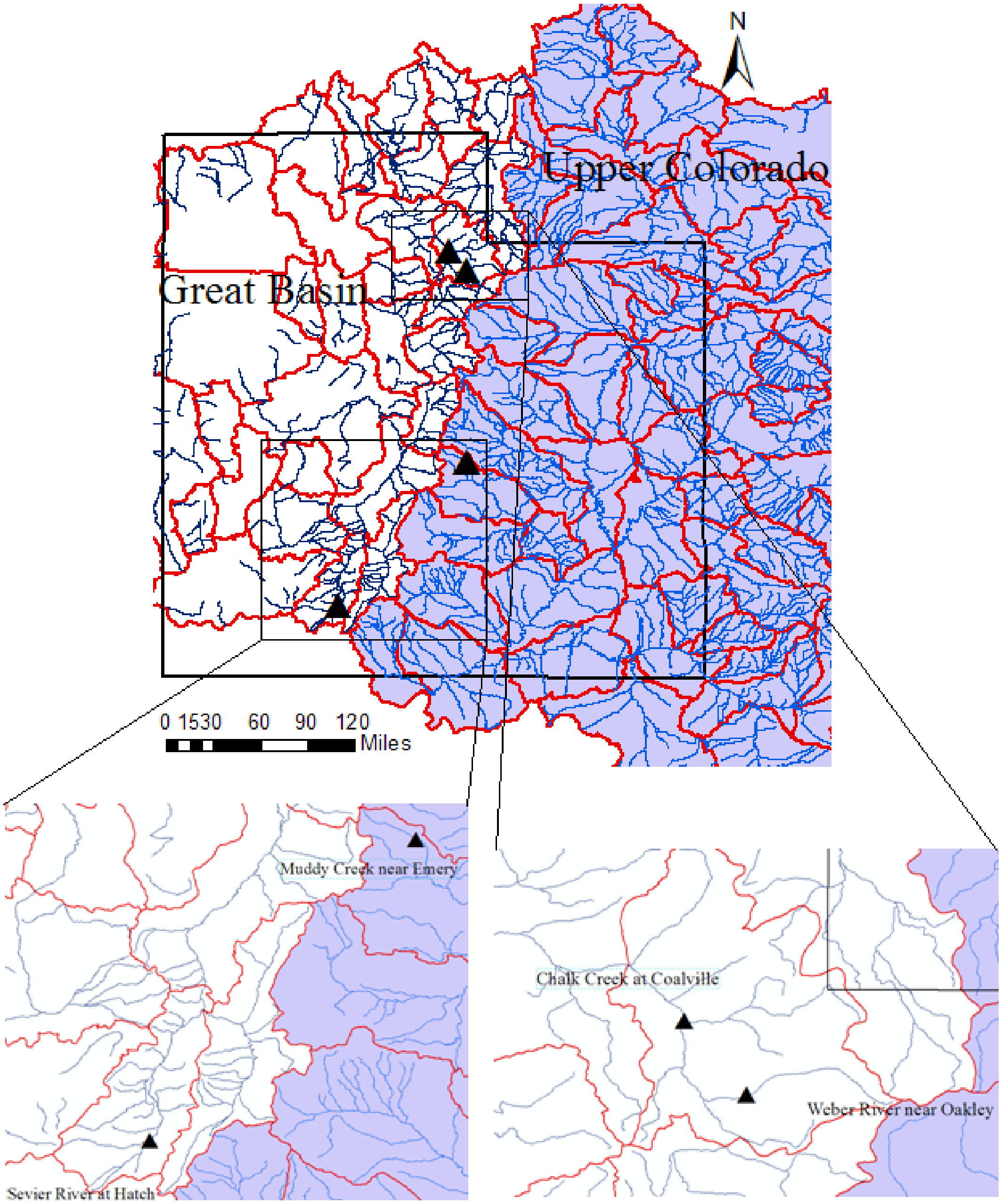


Figure 3.1 Location of the stream gages.

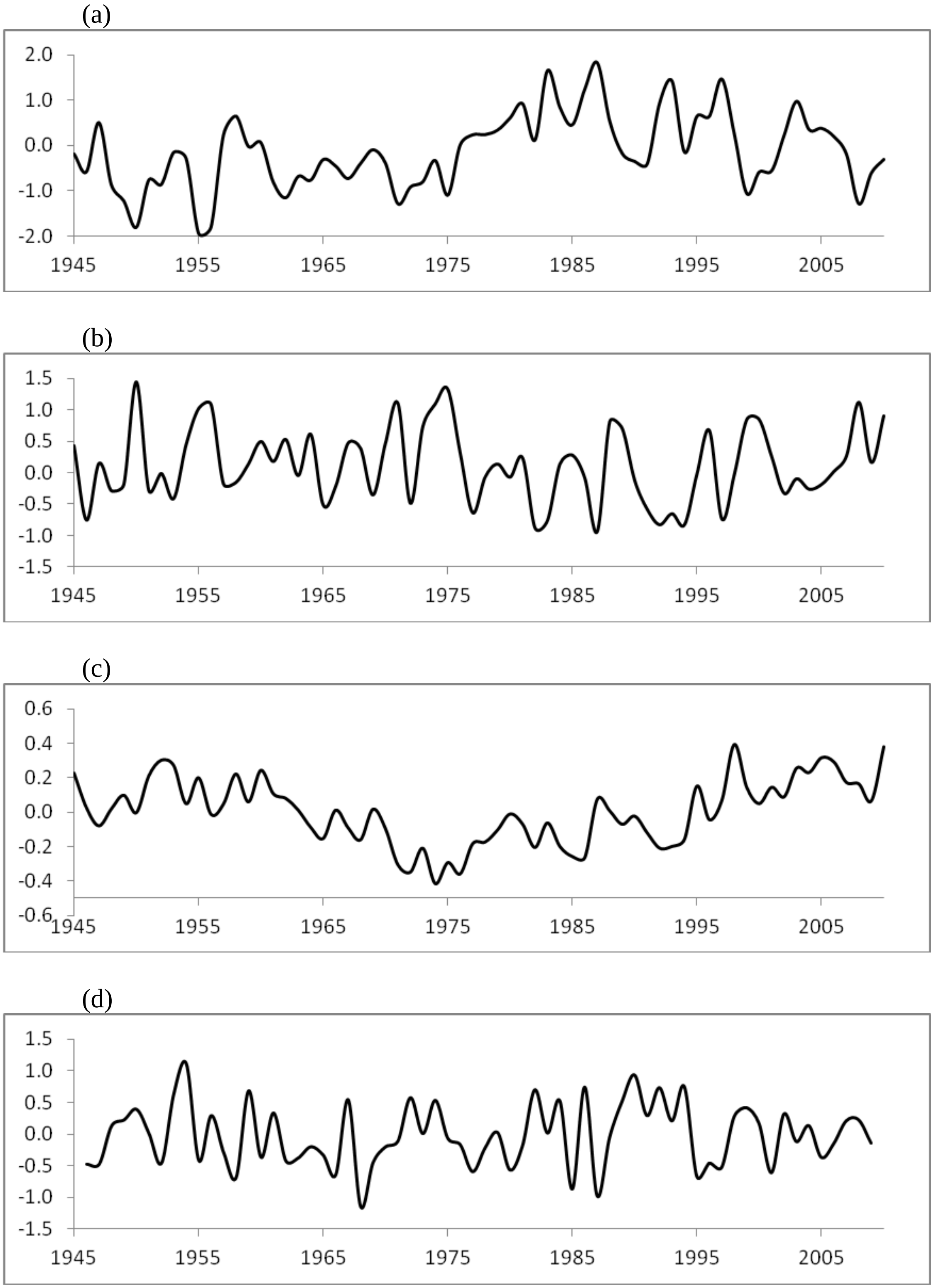


Figure 3.2 Ocean Atmospheric Oscillation indices (a) PDO, (b) ENSO, (c) AMO, and (d) NAO.

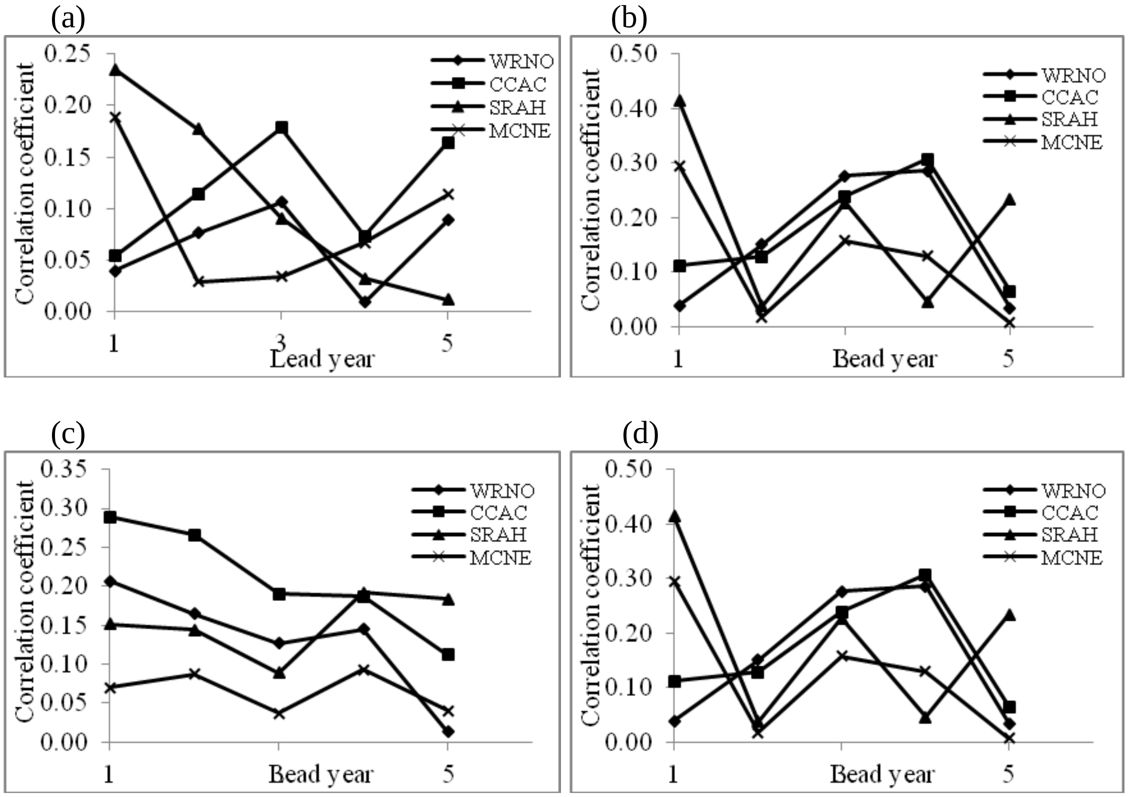


Figure 3.3 Absolute value of correlation coefficient between the annual oscillation modes and annual flow volume for (a) PDO, (b) ENSO, (c) AMO, and (d) NAO.

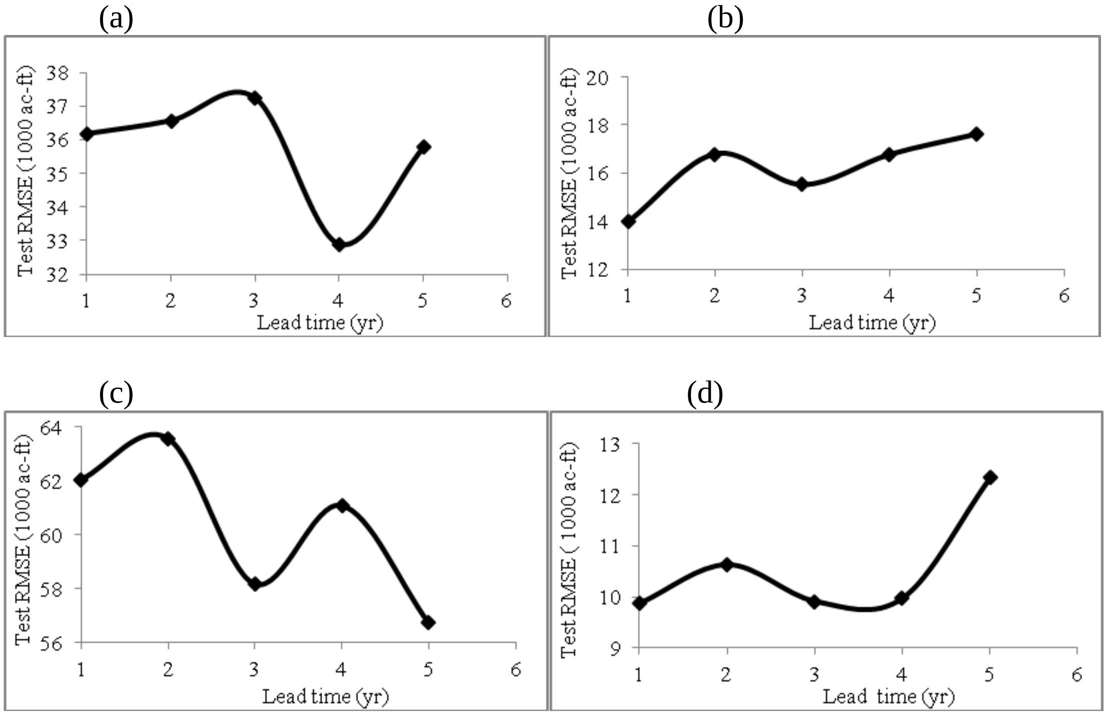


Figure 3.4 Test RMSE at 1 to 5 year lead for annual streamflow volume prediction for Model 1. (a) Weber River near Oakley, (b) Chalk Creek at Coalville, (c) Sevier River at Hatch, and (d) Muddy Creek near Emery.

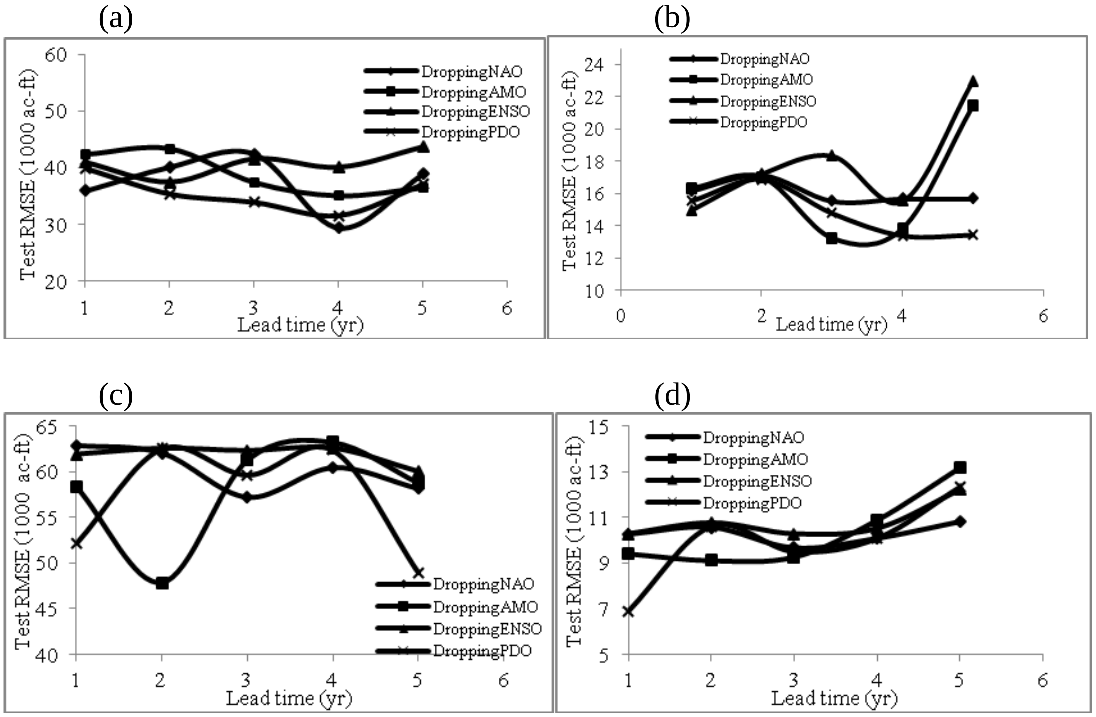


Figure 3.5 The test RMSE at 1 to 5 year lead for annual streamflow volume prediction for Model 2. (a) Weber River near Oakley, (b) Chalk Creek at Coalville, (c) Sevier River at Hatch , and (d) Muddy Creek near Emery.

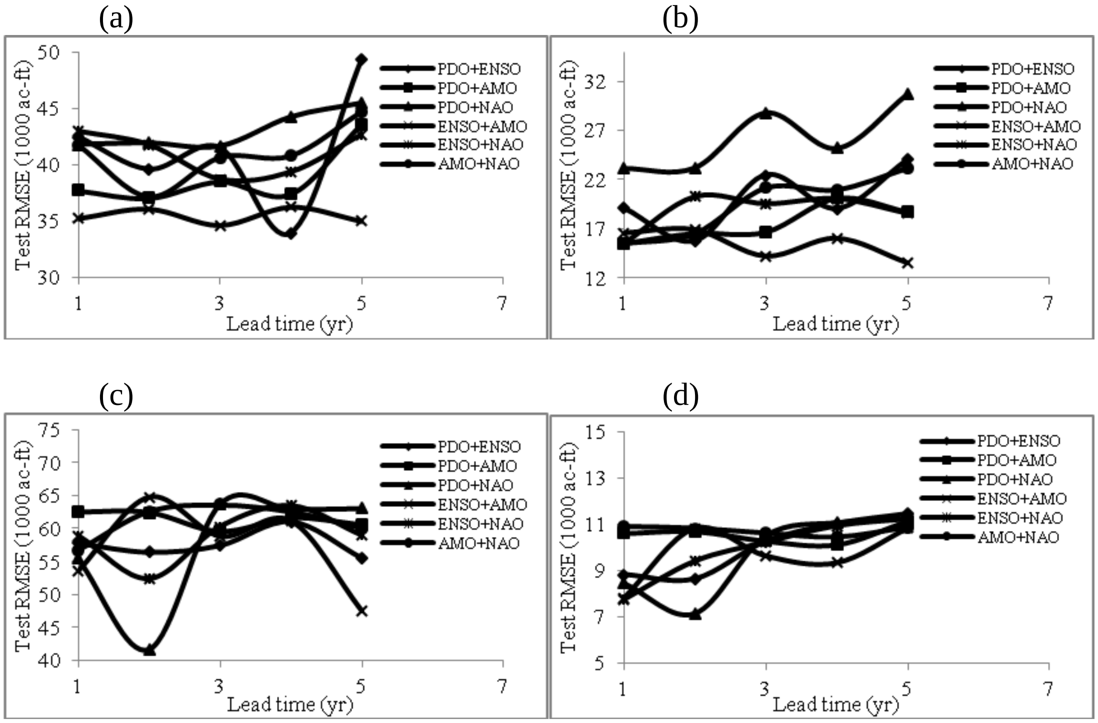


Figure 3.6 The test RMSE at 1 to 5 year lead for annual streamflow volume prediction for Model 3. (a) Weber River near Oakley, (b) Chalk Creek at Coalville, (c) Sevier River at Hatch, and (d) Muddy Creek near Emery.

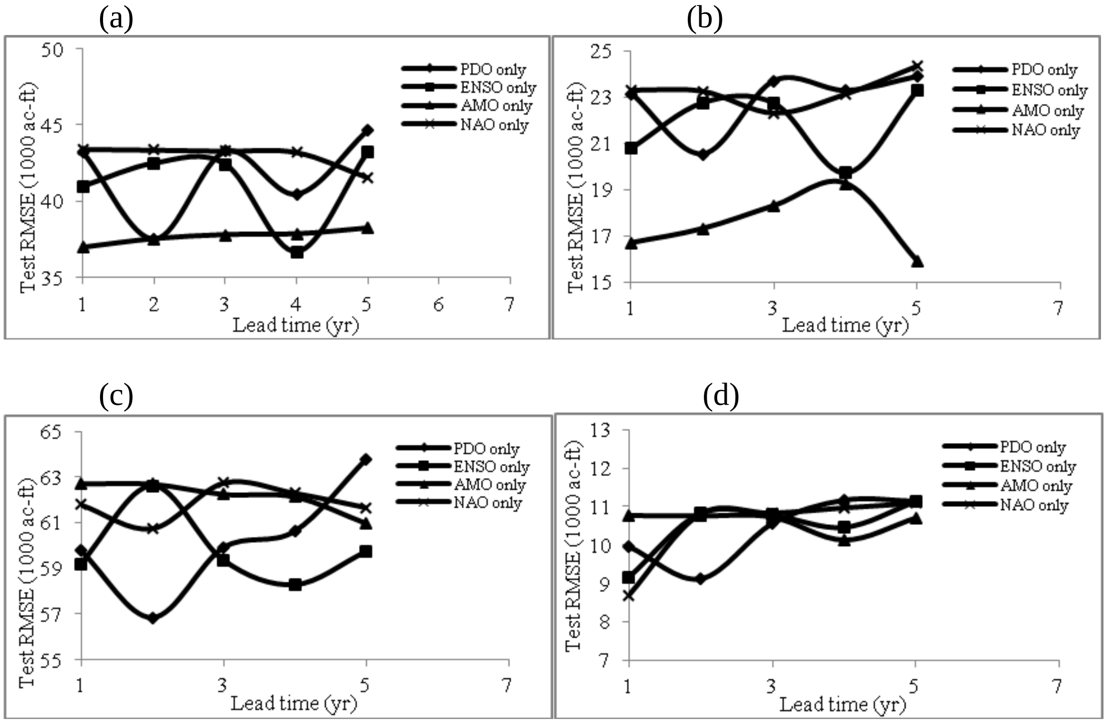


Figure 3.7 The test RMSE at 1 to 5 year lead for annual streamflow volume prediction for Model 4. (a) Weber River near Oakley, (b) Chalk Creek at Coalville, (c) Sevier River at Hatch, and (d) Muddy Creek near Emery.

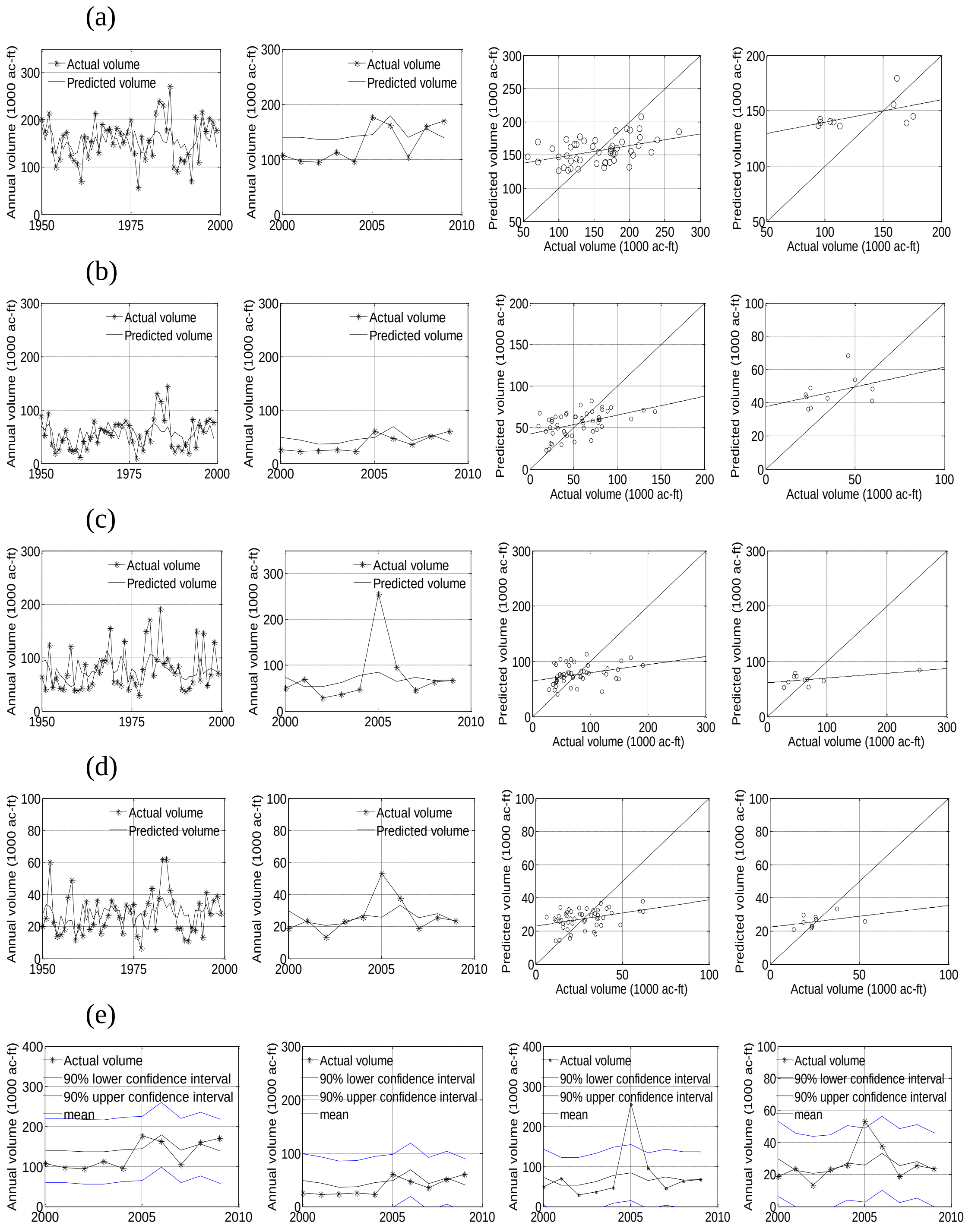


Figure 3.8 The plot of actual versus predicted annual flow volume for Model 1. (a) Weber River near Oakley, (b) Chalk Creek at Coalville, (c) Sevier River at Hatch, (d) Muddy Creek near Emery, (e) 90% confidence interval of prediction in test phase for all gages, and (f) Residual plots of test phase for (a) to (d).

(f)

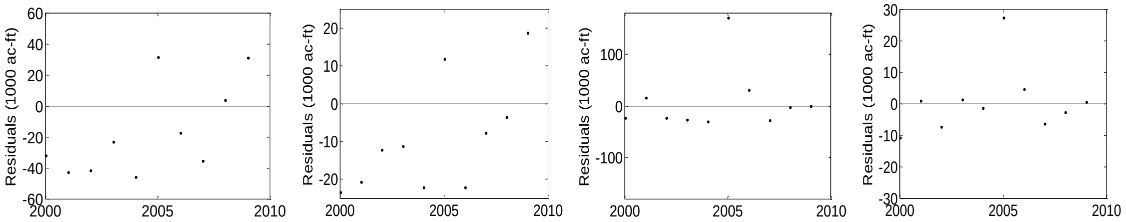


Figure 3.8 Cont.

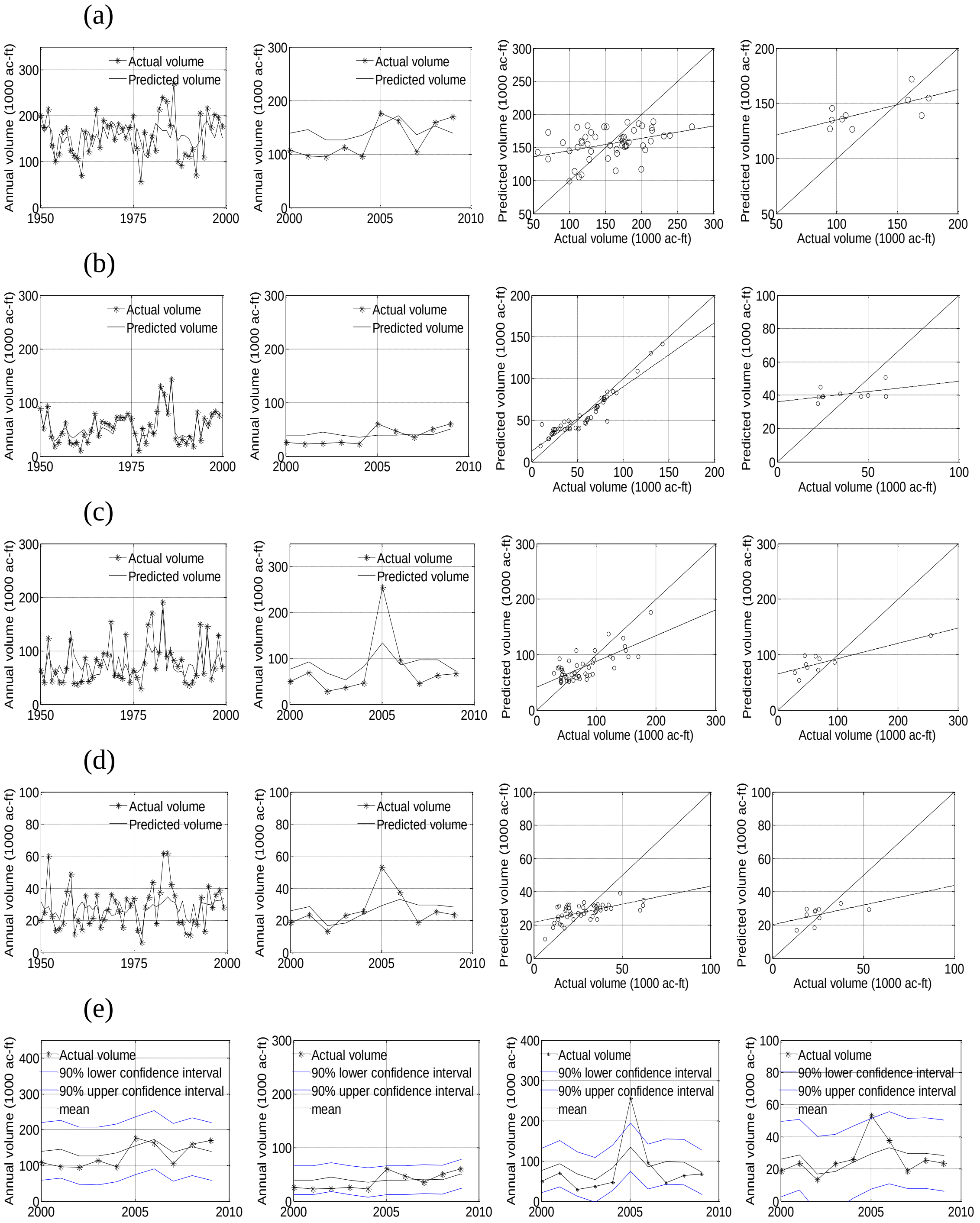


Figure 3.9 The plot of actual versus predicted annual flow volume for Model 2. (a) Weber River near Oakley, (b) Chalk Creek at Coalville, (c) Sevier River at Hatch, (d) Muddy Creek near Emery, (e) 90% confidence interval of prediction in test phase for all gages, and (f) Residual plots of test phase for (a) to (d).

(f)

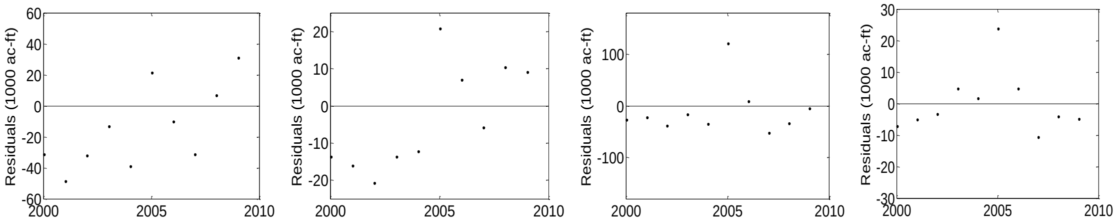


Figure 3.9 Cont.

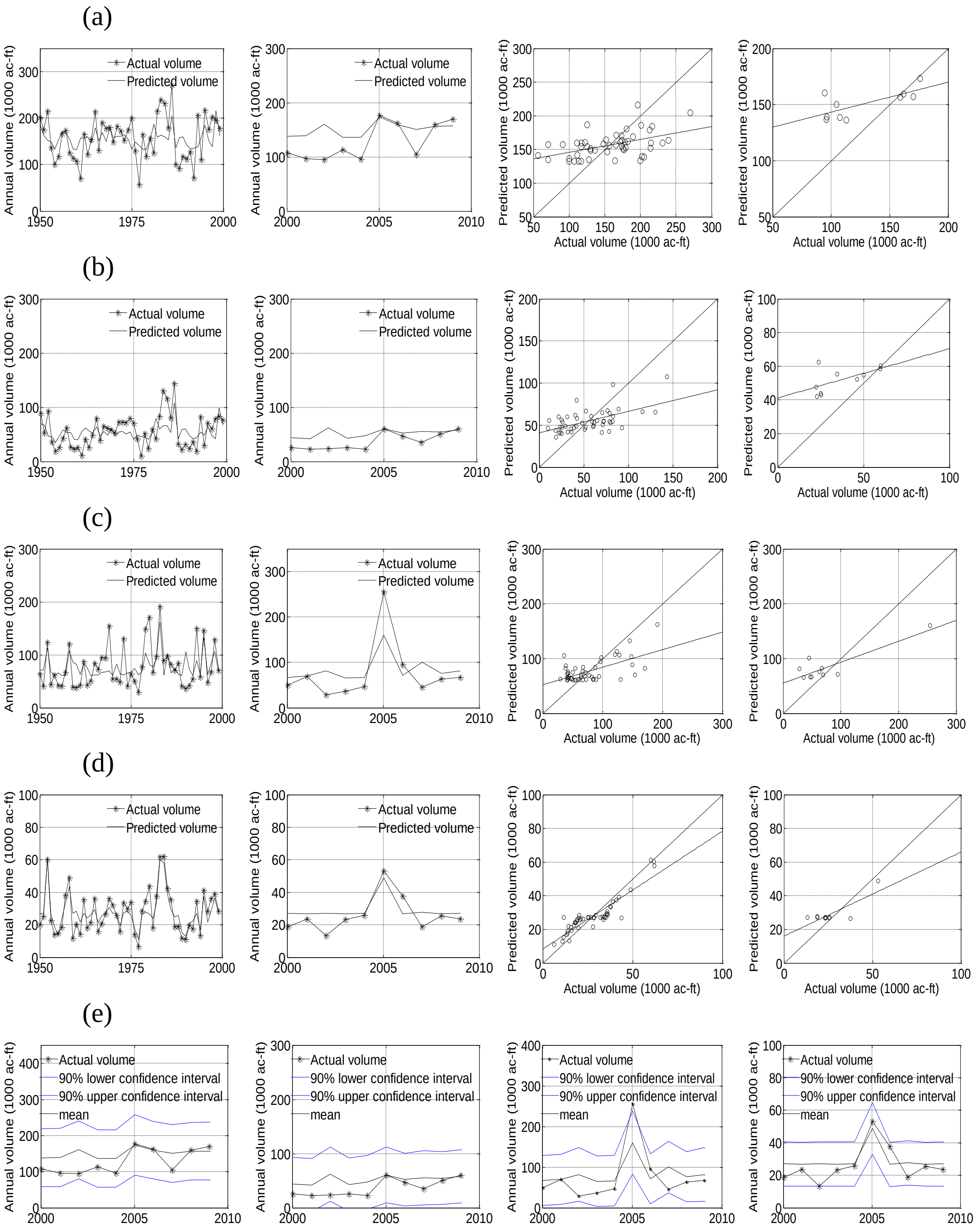


Figure 3.10 The plot of actual versus predicted annual flow volume for Model 3. (a) Weber River near Oakley, (b) Chalk Creek at Coalville, (c) Sevier River at Hatch, (d) Muddy Creek near Emery, (e) 90% confidence interval of prediction in test phase for all gages, and (f) Residual plots of test phase for (a) to (d).

(f)

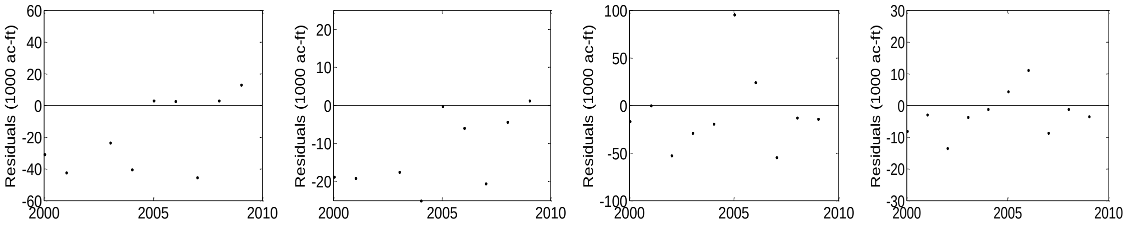


Figure 3.10 Cont.

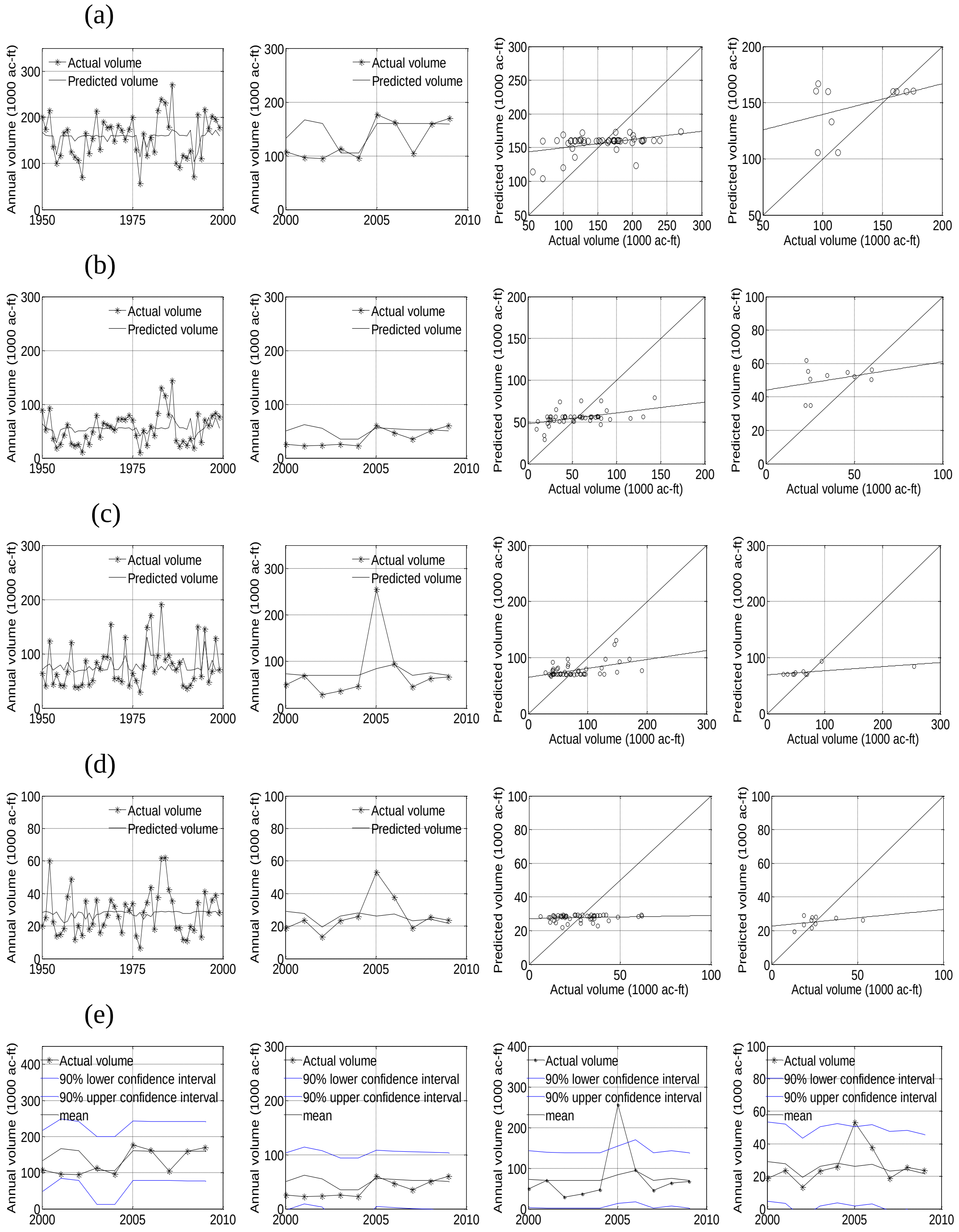


Figure 3.11 The plot of actual versus predicted annual flow volume for Model 4. (a) Weber River near Oakley, (b) Chalk Creek at Coalville, (c) Sevier River at Hatch, (d) Muddy Creek near Emery, (e) 90% confidence interval of prediction in test phase for all gages, and (f) Residual plots of test phase for (a) to (d).

(f)

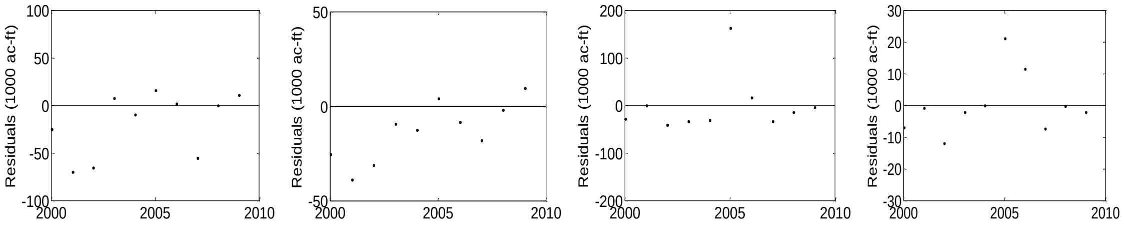


Figure 3.11 Cont.

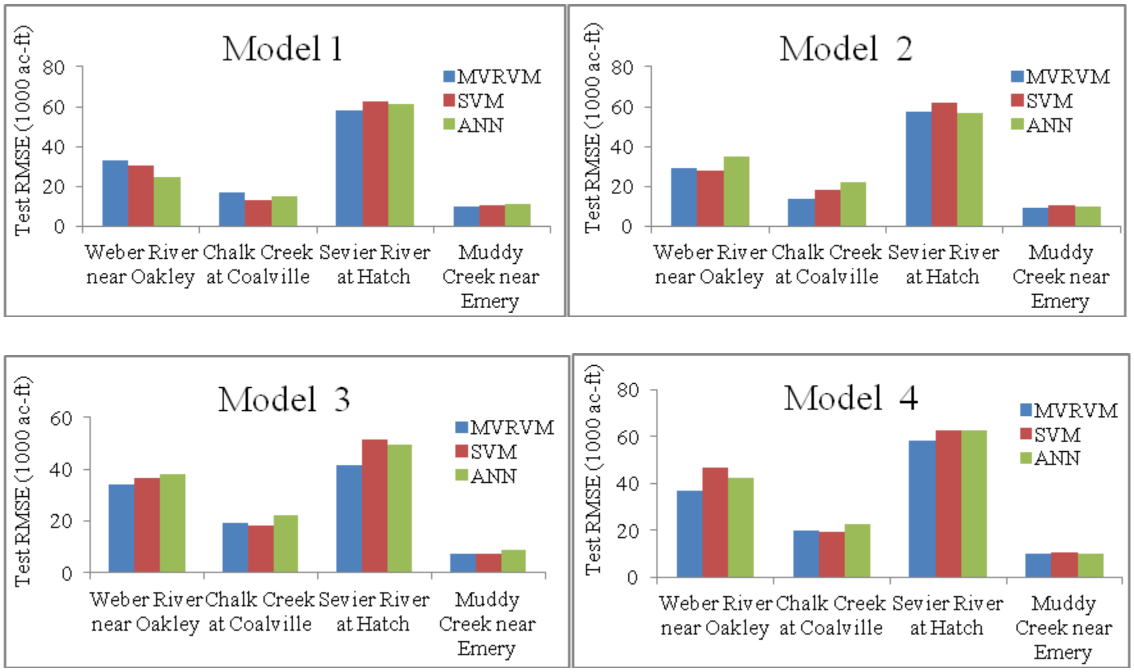


Figure 3.12 Comparison between MVRVM, SVM and ANN based on RMSE in test phase.

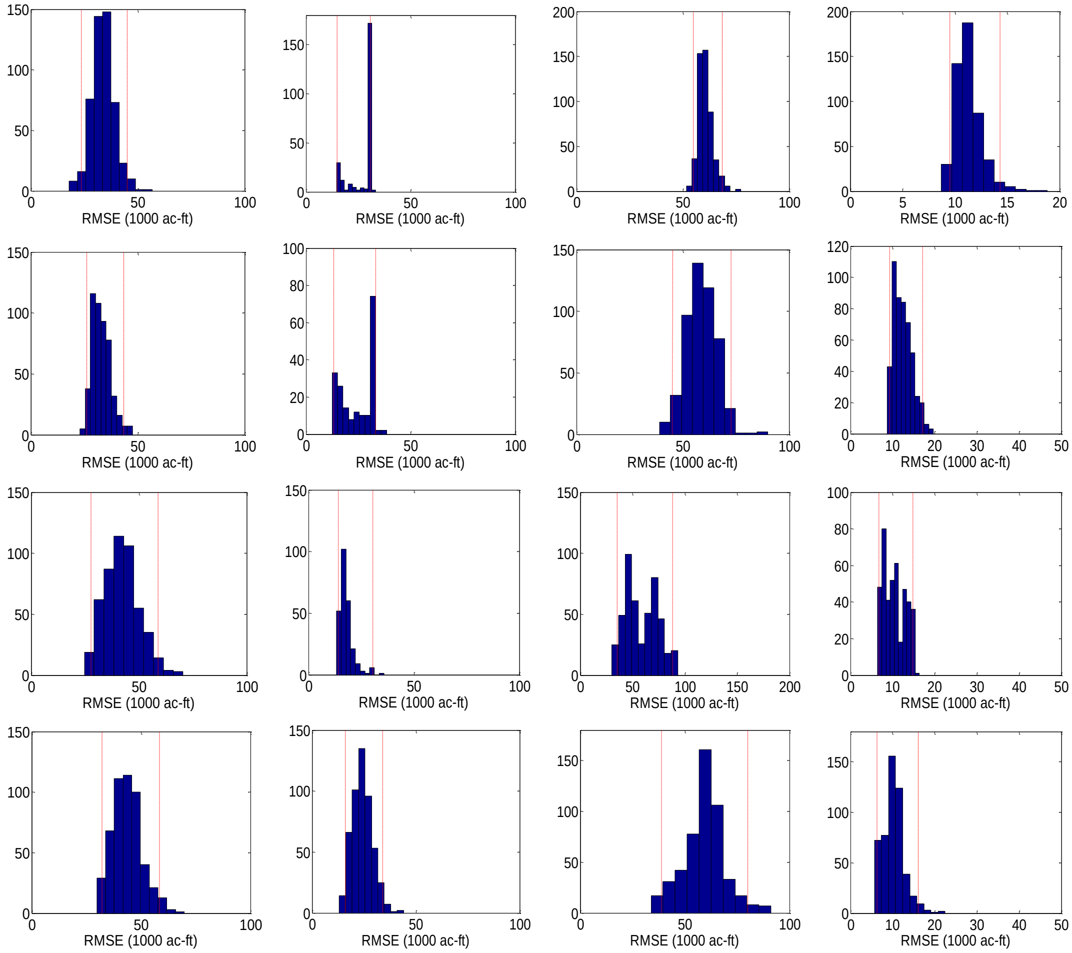


Figure 3.13 The bootstrap analysis for the best models. First column is for Weber River near Oakley, second column is for Chalk Creek at Coalville, third column is for Sevier River at Hatch, and fourth column is for Muddy Creek. Similarly first row is for Model 1, second row is for Model 2, third row is for Model 3, and fourth row is for Model respectively.

CHAPTER 4

MACHINE LEARNING REGRESSION APPROACH FOR PREDICTION OF GREAT SALT LAKE WATER SURFACE ELEVATION

Abstract

A data-driven model based on machine learning approach is used to predict the water surface elevation time series for the Great Salt Lake (GSL) at bi-weekly time step. For data-driven models, even if the data are scarce and the underlying processes are poorly understood, it is still possible to develop a model that produces reasonable predictions as demonstrated herein for the GSL. Support Vector Machines (SVM) and Relevance Vector Machines (RVM), popular data-driven models based on a machine-learning approach, are used in this paper. The concept of phase construction is used to represent the underlying dynamics of the process, i.e., the reconstruction of a single dimensional series into a multi-dimensional phase space using two parameters, ‘Embedding Dimension’ and ‘Time Delay’, which are estimated for GSL elevation series. The model is able to extract the dynamics of the system by using only a few observed data points for the training phase. The reliability of the algorithm in learning and forecasting the dynamics of the system is measured in the test phase. The GSL is divided into two arms by a rock-filled causeway which results in significant differences in water level between them. The water surface elevation is, therefore, predicted for both arms of the lake for two time periods; 1982 to 1987, and 1991 to 2008. The period of 1982 to 1987 is used to test the model performance for a dramatic rise of GSL water surface elevation, while the period of 1991 to 2008 is used to test the performance of the

model for the normal rise-fall of lake elevation. Results indicate that the predicted lake level is in good agreement with the actual lake level measured. The bootstrap analysis shows the model is robust and well generalized.

4.1 Introduction

Record breaking rises of the Great Salt Lake (GSL) water levels were observed between the years 1982 to 1987. These lake level rises resulted in severe economic impact to the State of Utah because the resulting floods damaged highways, railways, recreation facilities and industries located in the exposed lake bed. More precise predictions of GSL level may provide crucial information for planning and decision making processes in order to reduce the impact of the rising lake level. This process necessitates the development of a model capable of predicting the lake elevation accurately well ahead of the time. Lall et al. (1996) predicted GSL volume series in a short-term basis. The present research predicts the GSL elevation at biweekly time step for next few months.

Lorenz (1963) stated that time series of chaotic systems carry enough information about the system’s behavior in order to predict its future behavior. Chaotic systems are nonlinear, dynamic, highly sensitive to initial conditions, fully deterministic, and can be modeled using state-space reconstruction according to the time-delay embedding theorem (Koutsoyiannis and Pachakis 1996). Many hydrologic systems have been observed to be chaotic, therefore, the analysis of chaotic systems is nowadays an important tool in hydrology.

In this paper, one-dimensional time series of the lake elevation is used to develop a multi-dimensional phase space using the embedding dimension and time delay parameters. This reconstruction is a way of approximating the unknown function that describes the state evolution of the chaotic system (Abarbanel 1996). The multi-dimensional phase space is used to predict the lake elevation through a data-driven model based on the machine learning approach.

Prediction of lake elevation using physically-based model requires modeling complex physical processes. The complexities inherent in these models, and the difficulties associated with the corresponding data acquisition, limit the applicability of such models. Usually, the underlying process may not be fully understood, resulting in the use of a simplified approach. The data driven model is hence used for the GSL elevation prediction. Support Vector Machines (SVM) and Multivariate Relevance Vector Machines (MVRVM) are used to predict the GSL time series using reconstructed multi-dimensional phase space. Both of these models have strong regularization capability, ability to quickly capture the underlying physics of the system by relating input and output and provide accurate predictions of system behavior. Using these models, the GSL water level is predicted for two time periods: 1982 to 1987, and 1987 to 2008. The GSL is divided into two arms by a rock-filled causeway which results in significant differences in water level between them. The lake level, therefore, is predicted on both arms of the GSL.

4.2 Study Area and Data Collection

The Great Salt Lake is the largest U.S. Lake West of the Mississippi River and is the world’s fourth largest terminal lake. It is about 75 mile long and 28 mile wide. It has maximum depth of about 35ft. It is a remnant of Lake Bonneville, a prehistoric freshwater lake that was 10 times larger than the current GSL size. The GSL drains water from three states: Utah, Idaho and Wyoming. The drainage area of the GSL is 90,000 km^2. The lake has three major rivers draining into it: the Bear River, the Weber River, and the Jordan River. The lake is divided into two arms by a rock-fill causeway: the northern arm and the southern arm. There exists an elevation difference between two arms due to unequal rate of inflow and evaporation loss (water balance) from each arm of the lake. The difference in elevation started building up, and became significant, from the mid-1980s on. The U.S. Geological Survey (USGS) operates gages that collect water-surface elevation data in the southern arm of the lake at the Boat Harbor Gage (USGS station 10010000), and on the northern arm of the lake at the Saline Gage (10010100) (http://ut.water.usgs.gov/greatsaltlake/). Water surface elevation data was collected from both stations for the training and prediction time periods used in the present study. The details of GSL elevation data and its characteristics are shown in the Appendix. A map of the GSL is shown in Figure 4.1.

4.3 Model Description

4.3.1 Support Vector Machine

Support Vector Machine is a supervised machine learning model used for regression in this research. Vapnik and his co-workers developed SVMs regression (Vapnik 1995), which is the extension of SVM classification developed in 1990's. SVMs are very specific class of algorithms, characterized by usage of kernels, absence of local minima, and sparseness of the solution. SVM utilizes a small subset of training points which gives enormous computational advantages. The use of epsilon-insensitive loss function ensures the existence of a global minimum and the optimization of reliable generalization at the same time. In SVMs, a non-linear function is produced by a linear learning-machine mapping into a high-dimensional kernel-induced feature space. The basic requirement of the kernel is that it must satisfy the Mercer's theorem (Vapnik 1995, 1998). A global optimum is ensured through the formulation of a quadratic optimization problem which makes SVMs superior to traditional learning machine algorithms. The capacity of the system is controlled by parameters that do not depend on the dimensionality of the feature space.

SVM has been successfully used in a variety of hydrological problems. They have been used from multi-time scale streamflow prediction to groundwater head observation networks design (Asefa et al. 2004, 2006). A description of the SVM approach is given next.

For a given data set $\{(x_1, y_1), \ldots\ldots\ldots\ldots\ldots\ldots (x_l, y_l)\} \subset \mathrm{X}$, where X denotes the space of the input patterns, the goal is to find a functional dependency $f(x)$ between inputs x and target y taken from the set of independent and identically distributed observations. The SVM regression is formulated through the minimization of the following objective function:

$$\text{Minimize } \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{L}(\xi_i + \xi_i^{\,*}) \tag{4.1}$$

Subject to,

$$y_i - \langle w, x_i \rangle - b \leq \varepsilon + \xi_i$$

$$\langle w, x_i \rangle + b - y_i \leq \varepsilon + \xi_i^*$$

$$\xi_i \xi_i^* \geq 0$$

where $f(x) = \langle w, x \rangle + b$, $\langle w, x \rangle$ denotes the dot product of *w* and *x, x* is the input vector, *w* is the weights vector norm, ε is Vapnik's insensitive loss function, C is the cost parameter, and b is the bias.

The first term of Eq. 4.1 is a regularization term which avoids the ill-posedness of the estimation problem (Gill et al. 2006; Tychonoff and Arsenin 1977). The second term is the epsilon-insensitive loss function, which represents a discrepancy between the actual measurement and estimated values. SVM performs the regression using ε-insensitive loss functions and, at the same time, tries to reduce model complexity by minimizing $\|w\|^2$. The loss function can be described by introducing (non-negative) slack variables ξ_i, ξ_i^* $i = 1, \ldots n$ to measure the deviation of training samples outside the ε-insensitive zone (Figure 4.2). The slack variables determine the degree to which samples with error greater than ε are penalized. The formulation imposes sparseness in the solution as errors that are less than ε are ignored (Figure 4.2). For any error smaller than ε, ξ_i =0 and ξ_i^* =0. The corresponding point, thus, does not enter into the objective function. This makes the model sparse. The procedure has computational advantages and important implications in hydrological applications.

The Lagrange function (L) is constructed from the objective function and the corresponding constraints by introducing a dual set of variables (Vapnik 1995) (Eq. 4.2),

$$L(w, \xi, \xi^*, \alpha, \alpha^*, \eta, \eta^*, b) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{L}(\xi_i + \xi_i^*) - \sum_{i=1}^{L}\alpha_i(\varepsilon + \xi - y_i + \sum_{j=1}^{K} w_j x_{ji} + b)$$

$$-\sum_{i=1}^{L}\alpha_i^{*}(\varepsilon+\xi^{*}+y_i-\sum_{j=1}^{K}w_j x_{ji}-b)-\sum_{i=1}^{L}(\eta_i\xi_i+\eta_i^{*}\xi_i^{*})\,, \quad (4.2)$$

where α , α^*, η, η^* are Lagrange multipliers. The saddle-point condition states that the partial derivative of L with respect to primal variables (w, b, ξ, ξ^*) have to vanish for optimality. Differentiating L with respect to primal variables and substituting, we can obtain the minimum value of L. The resulting minimum-L is then maximized with respect to the dual variables. The Lagrange multipliers α and α^* are found by maximizing the dual functional subject to constraints. The dual maximization problem then can be written as follows: find α and α^* to

$$\text{Max } w(\alpha^*,\alpha)=-\varepsilon\sum_{i=1}^{L}(\alpha_i+\alpha_i^{*})+\sum_{i=1}^{L}y_i(\alpha_i-\alpha_i^{*})-\frac{1}{2}\sum_{i=1}^{L}\sum_{j=1}^{L}(\alpha_i-\alpha_i^{*})(\alpha_j-\alpha_j^{*})<x_i,x_j>, \quad (4.3)$$

Subject to,

$$\sum_{i=1}^{L}(\alpha_i^{*}-\alpha_i)=0\,, \qquad \alpha_i,\alpha_i^{*}\in[0,C]\,.$$

The approximating function is then written as,

$$f(x)=\sum_{i=1}^{N}(\alpha_i^{*}-\alpha_i)\langle x,x_i\rangle+b\,, \qquad (4.4)$$

where x_i 's are support vectors. The number of support vectors (N) is much smaller than the total number of data points in training (L).

Non-linearity is introduced by preprocessing the training data into a higher dimension through a kernel function. By replacing the inner product through an appropriately chosen kernel function, one can implicitly perform a nonlinear mapping to a high dimension feature space without increasing the number of tunable parameters.

The approximating function can be upgraded to,

$$f(x)=\sum_{i=1}^{N}(\alpha_i^{*}-\alpha_i)k(x,x_i)+b, \tag{4.5}$$

where $k(x,x_i)$ is the kernel function that replaces the dot product of the input data. It approximates the transformation of input data into high dimension feature space and corresponding dot product in feature space. From the Kuhn-Tucker condition, the product between the dual variable and constraints should vanish for optimality. This shows that the Lagrange multiplier is zero inside the ε-tube and non zero outside of it. The sample points with non-vanishing coefficients fit the data, which are called support vectors. The number of support vectors is small relative to the size of data set, yielding a sparse solution. The support vectors carry all the information necessary to determine the optimal solution.

There are mainly three model parameters in SVM. They are the cost parameter (C), the insensitive parameter (ε), and the kernel parameter used in the kernel function. Parameter *C* determines the trade off between minimizing the regularization and minimizing the loss function. Increasing the cost parameter increases the cost of error and forces the creation of a more accurate model, however, this may not generalize well. Epsilon (ε) controls the width of insensitive zone in training data set. The data points with error values less than epsilon are ignored hence don’t enter into the objective function. This ensures the sparseness solution that leads computational advantage over other models. Higher values of epsilon produce a few support vectors resulting in flat estimation while small value of epsilon produces large number of support vectors. Optimum combinations of model parameters are estimated from the grid search method of cross validation. Selecting a particular kernel function and corresponding kernel parameter is usually based on knowledge of the application domain. The choice of kernel function is usually heuristic; however, it may be selected by comparing the test result from different kernel types. In hydrological problems, the Gaussian kernel is commonly used (Tripathi and Govindaraju 2006). Interested readers for SVM are referred to Vapnik (1995, 1998). The software to develop SVM model was obtained from SVM and Kernel Methods Matlab Toolbox (http://asi.insa-rouen.fr/enseignants/~arakotom/toolbox/index.html).

4.3.2 Relevance Vector Machine

Relevance Vector Machine is a supervised learning model based on Bayesian learning. RVM has been successfully applied to resolve water resource management problems (Khalil et al. 2005a, 2005b; Ticlavilca 2010). The software to develop the model was obtained from Thayananthan (2005), at the University of Cambridge, England. This is an extension of the sparse Bayesian model developed by Tipping and Faul (2003).

For the given input-target pair $\{x_n, t_n\}_{n=1}^{N}$ in a training data set, the model learns the dependency of the targets on the inputs with the objective of making accurate predictions of the target *t* for previously unseen values of input *x* (Tipping 2000, 2001).

The targets are assumed to be samples from the model (*y*) with additive noise (ε). The target can be written as sum of an approximation vector $y = [y(x_1),\text{.........}y(x_N)]^T$ and the error vector $\varepsilon = (\varepsilon_1,\text{........}\varepsilon_N)^T$. The errors are independent samples from some noise process, assumed to be mean-zero Gaussian with variance σ^2. The target vector can be written as, $t = (t_1.......t_N)^T$,

where $t = y + \varepsilon,$

$$= \Phi w + \varepsilon. \tag{4.6}$$

The weight vector (w) is expressed as $w = (w_1,...w_i.....w_N)^T$, and Φ is the design matrix of size N× (N+1). The design matrix is expressed as $\Phi = [\phi(x_1).....\phi(x_N)]^T$, wherein $\phi(x)$ is a basis function which is expressed with kernel as parameterized by the training vectors. The basis function is thus given by, $\phi(x_n) = [1, K(x_n, x_1), K(x_n, x_2),....., K(x_{n,}, x_N)]^T$.

The target t_n is assumed to be independent, therefore, the likelihood of the complete data is written as,

$$p(t|w,\sigma^2) = (2\pi\sigma^2)^{-N/2} \exp\{\frac{1}{2\sigma^2}\|t - \Phi w\|^2\}. \tag{4.7}$$

Let τ_i be the i^{th} component of the target vector *t*, and w_i be the weight vector for the i^{th} component of the output target vector *t* such that $w = (w_1,...w_i.....w_N)^T$. This is a Gaussian distribution which can also be written as,

$$p(t|w,\sigma^2) = \prod_{i=1}^{n} N(\tau_i|\Phi w_i, \sigma_i^{\ 2})..$$

The model has roughly as many parameters as in the training set. This causes the maximum likelihood estimation of w and σ^2 of the model to be severely over-fitted. To avoid this, Tipping (2001) imposed an explicit prior probability distribution over them, which ultimately leads the sparsity of the model. The prior probability is given by,

$$p(w|\alpha) = \prod_{i=0}^{N} N(w_i|0, \alpha_i^{-1}), \tag{4.8}$$

where $\alpha = (\alpha_0, ...\alpha_N)^T$ is a vector of N+1 hyper-parameters. Each α_i controls the strength of the prior over its associated weight (Tipping and Faul 2003). Bayes' rule is used to compute the posterior over all unknowns given the data,

$$p(w, \alpha, \sigma^2|t) = \frac{p(t|w, \alpha, \sigma^2)\, p(w, \alpha, \sigma^2)}{p(t)}. \tag{4.9}$$

This term can not be computed in fully analytical form, therefore, some approximation is used. The posterior term is decomposed as:

$$p(w, \alpha, \sigma^2|t) = p(w|t, \alpha, \sigma^2)\, p(\alpha, \sigma^2|t). \tag{4.10}$$

Given the data, the posterior distribution over the weights is given by (Tipping 2001),

$$p(w|t, \alpha, \sigma^2) = \frac{p(t|w, \sigma^2).p(w|\alpha)}{p(t|\alpha, \sigma^2)},$$

$$= (2\pi)^{-(N+1)/2}|\Sigma|^{-1/2}.\exp\{-\frac{1}{2}(w-\mu)^T \Sigma^{-1}(w-\mu)\}, \tag{4.11}$$

$$= \prod_{i=1}^{N} N(w_i|\mu_i, \Sigma_r).$$

The posterior covariance and mean of the weight are $\Sigma = (\sigma^{-2}\Phi^T\Phi + A)^{-1}$ and $\mu = \sigma^{-2}\Sigma\Phi^T t$ respectively, where $A = diag(\alpha_0, \alpha_1, ..., \alpha_N)$. The key point is that if any $\alpha_m = \infty$, the corresponding $\mu_m = 0$.

An approximation for the hyper-parameter posterior is adapted by using a delta function at its mode, i.e., at its most probable values $\alpha_{MP}, \sigma^2_{MP}$ (Tipping 2001),

$$\int p(t_* | \alpha, \sigma^2)\delta(\alpha_{MP}, \sigma^2_{MP}) d\alpha d\sigma^2 \approx \int p(t_* | \alpha, \sigma^2) p(\alpha, \sigma^2 | t) d\alpha d\sigma^2 \ . \tag{4.12}$$

The learning process then becomes the search for the hyper-parameter posterior mode, i.e. the maximization of $p(\alpha, \sigma^2 | t) \propto p(t | \alpha, \sigma^2) p(\alpha) p(\sigma^2)$ with respect to α and σ^2. For uniform hyperpriors over logα and log*σ*, $p(\alpha, \sigma^2 | t) \propto p(t | \alpha, \sigma^2)$, which is further given by,

$$p(t | \alpha, \sigma^2) = \int p(t | w, \sigma^2) p(w | \alpha) dw,$$

$$= (2\pi)^{-N/2} \left| \sigma^2 I + \Phi A^{-1} \Phi^T \right|^{-1/2} \exp\{-\frac{1}{2} t^T (\sigma^2 I + \Phi A^{-1} \Phi^T)^{-1} t\}. \tag{4.13}$$

This quantity is known as marginal likelihood and its maximization is known as the type-II maximum likelihood method (Berger 1993). Eq. 4.13 is solved by iterative re-estimation which gives,

$$\alpha_i^{new} = \frac{\gamma_i}{\mu_i^2} , \tag{4.14}$$

where $\gamma_i = 1 - \alpha_i N_{ii}$.

The term μ_i is the *i*th posterior mean weight and N is the number of data examples. N_{ii} is the *i*th diagonal element of the posterior weight covariance computed with the current α and σ^2. The noise variance is re-estimated from,

$$(\sigma^2)^{new} = \frac{\| t - \Phi\mu \|^2}{N - \sum_i \gamma_i} . \tag{4.15}$$

The learning algorithm proceeds by repeated application of (4.14) to (4.15), together with updating of the posterior statistics Σ and μ until specified convergence criteria are met. The value of α_i generally approaches to infinity which implies that $p(w_i | t, \alpha, \sigma^2)$ becomes highly peaked at zero that makes the model sparse. The relatively nonzero weights correspond to the input vectors that form the sparse core of the RVM model. These input vectors are called relevance vectors (RVs). This sparsity is an effective method to control model complexity, avoid over-fitting and control computational characteristics of model performance (Tipping and Faul 2003).

The predictions of output are made based on the posterior distribution over the weights, conditioned on the maximizing values α_{MP} and σ^2_{MP}. The target (t_*) for new input x_*,

$$p(t_* | t, \alpha_{MP}, \sigma^2_{MP}) = \int p(t_* | w, \sigma^2_{MP}) \, p(w | t, \alpha_{MP}, \sigma^2_{MP}) dw$$

This is readily computed because both terms in the integral are Gaussian,

$$p(t_* | t, \alpha_{MP}, \sigma^2_{MP}) = N(t_* | y_*, \sigma^2_*), \tag{4.16}$$

with, $y_* = \mu^T \phi(x_*)$,

$$\sigma^2_* = \sigma^2_{MP} + \phi(x_*)^T \Sigma \phi(x_*) .$$

The total variance consists of sum of the variance of data and uncertainty in estimating the weight.

Interested readers for Relevance Vector Machine are referred to Tipping (2000; 2001), Tipping and Faul (2003), Thayananthan (2005), and Thayananthan et al. (2008).

4.4 Model Formulation and Application

The time series of a chaotic system itself carries enough information about the behavior of system in order to make predictions (Lorenz 1963). They can be modeled using state-space reconstruction via a time-delay embedding theorem (Koutsoyiannis and Pachakis 1996). This theorem states that, given a recognized state-space representation of chaotic time series, a full knowledge about the system behavior can be obtained through estimation of the time delay and the embedding dimension (Takens 1981). Traditional linear time series analysis models are insufficient to adequately describe the dramatic rise and fall of GSL elevation/volume (Lall et al. 1996) that occurred in the period 1982-1987. One reason for such inadequacy may be the fact that there is not enough information prior to this event within just one-dimensional time series output of the system. This opens up the possibility of investigating whether there is a set of differential equations that are responsible for generation of the single time series hence investigation of chaos by unfolding the dynamics through representation of the data in multi dimensional state space (Khalil et al. 2006).

The underlying physical and other processes responsible for the evolution of GSL elevation dynamics are not considered in this paper. The observed past water surface elevations of the lake are used as input. From the one dimensional variable $y(t)$, a multivariate state space is constructed in which the dynamics unfolds by creating vectors of dimension d at time delay (τ). This is called reconstruction space where the dynamics of the original chaotic system can be reconstructed, i.e.,

$$y(t+1) = F(y(t)).$$

This implies,

$$y_{t+\tau} = f(y_t, y_{t-\tau}....., y_{t-(d-1)\tau}), \tag{4.17}$$

where d is embedding dimension, which is total number of time delay co-ordinates required to develop phase construction, and τ is the time delay. The mapping function f is estimated by minimizing the regularized risk functional. Both SVM and MVRVM have strong regularization capability. The objective of the learning machine is to estimate an unknown real valued function, $y_{t+\tau}$, that is capable of making accurate predictions of output for previously unseen value of input. Using this model, the GSL water surface elevation can be predicted at biweekly time steps, $t+\tau$ in future.

4.4.1 Estimating time delay and system dimension

This paper uses the Average Mutual Information (AMI) function and the False Nearest Neighbor method for the estimation of time delay and dimension of chaotic system respectively. In Average Mutual Information, the regular measurements and time lagged measurements is easy to evaluate directly from the time series and easy to interpret (Abarbanel 1996). False Nearest Neighbor method is commonly used and matured method for estimating system dimension. This is simple and fast method.

Average Mutual Information

The mutual information between measurement a_i drawn from a set A= $\{a_i\}$ and measurement b_j drawn from a set B= $\{b_j\}$ is the amount learned by the measurement of a_i about the measurement of b_j. In bits, this information is measured as

$$\log_2\left[\frac{P_{AB}(a_i,b_j)}{P_A(a_i)P_B(b_j)}\right],$$

where $P_{AB}(a,b)$ is the joint probability density for measurements A and B resulting in values *a* and *b*. $P_A(a)$ and $P_B(b)$ are the individual probability densities for the measurements of A and B, respectively. If the measurement of a value from A resulting in a_i is completely independent of measurement of a value from B resulting in b_j, then $P_{AB}(a,b) = P_A(a)\ P_B(b)$ and amount of information between the measurements, the mutual information is zero. The average over all measurements of this information statistic, called the average mutual information between A measurements and B measurement, is

$$I_{AB} = \sum_{a_i}\sum_{b_j} P_{AB}(a_i,b_j)\log_2\left[\frac{P_{AB}(a_i,b_j)}{P_A(a_i)P_B(b_j)}\right].$$

This is a theoretic idea which connects two sets of measurements with each other and establishes a criterion for their mutual dependence based on the notion of information connection between them. In our case, the average mutual information between measurement $y(t)$ at time *t* are connected in an information-theoretic fashion to measurements $y(t+\tau)$ at time $t+\tau$ by

$$I(\tau) = \sum_{y(t)}\sum_{y(t+\tau)} p(y(t),y(t+\tau))\log_2\left[\frac{p(y(t),y(t+\tau))}{p(y(t))p(y(t+\tau))}\right]. \tag{4.18}$$

By general arguments, $I(\tau) \geq 0$ (Gallager 1968). When τ becomes large, the chaotic behavior of the signal makes the measurements $y(t)$ and $y(t+\tau)$ become independent in a practical sense, and $I(\tau)$ will tend to zero. The τ must be large enough that independent information about the system is in each component of vector, however, it must not be too large that the components of the vectors $y(t)$ are independent enough that they will not contain any new information (Abarbanel 1996). The practical way of choosing τ is when the average mutual information has its first minimum (Fraser 1989; Fraser and Swinney 1986).

False Nearest Neighbor

The global embedding dimension, or actual system dimension, d, is the minimum number of time delay coordinates needed so that the trajectories $y(t)$ do not intersect in d dimension. In dimension less than d, trajectories can intersect because they are projected down into too few dimensions. Subsequent calculations, such as predictions, may then be corrupted. When embedding dimension is large ($>>d$), noise might occupy the embedding space (Khalil et al. 2006) which eventually may deteriorate the prediction.

A false nearest neighbor is a widely used method to estimate the optimum embedding dimension for phase space reconstruction. This method increases the embedding dimension by one at each step (from embedding dimension d to d+1), and counts the percentage of points for which its nearest neighbor falls apart with the addition of a new component and, therefore, these points are called false nearest neighbors. This means the points apparently lying close together due to projection are separated in higher embedding dimensions. The estimated d is the one that first gives the insignificant percentage of false nearest neighbors. We state this criterion by designating as a false nearest neighbor, any neighbor for which the following is valid (Kennel et al. 1992),

$$\left[\frac{R_{d+1}^2(t,r)-R_d^2(t,r)}{R_d^2(t,r)}\right]^2=\frac{|y(t+\tau)-y(t_r+\tau)|}{R_d(t+\tau)}>R_{tot}, \tag{4.19}$$

where *t* and t_r are the times corresponding to the neighbor and reference point, respectively. R_d denotes the distance in phase space with embedding dimension d, and R_{tot} is the tolerance threshold.

The output produced by the function is the percentage of FNN versus increasing d. This function has monotonic decreasing graph. The optimal d usually can be found near the crossing of the 10 to 20 percent threshold (Kennel et al. 1992). The “fractal” package in R is implemented to estimate the system dimension of Great Salt Lake water surface elevation.

4.4.2 Performance Criteria

Model validation is defined to mean “substantiation that a computerized model within its domain of applicability possesses a satisfactory range of accuracy consistent with the intended application of the model” (Schlesinger 1979). Statistical measures, which are objective in nature, can be employed for evaluating the performance of the model in testing phase, hence validating the model. Among the various statistical measures, the Nash-Sutcliffe efficiency, the root mean square error, and the bias are used in this paper.

Root Mean Square Error (RMSE)

Mathematically, the root mean square error (RMSE), is expressed as,

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(t_i - t_i^*)^2}{n}} \tag{4.20}$$

where t_i is observed value, t_i^* is prediction from the model and n is number of sample size. The smaller the value of RMSE, the better the prediction result is. The ideal value of RMSE is zero.

Bias

Bias is a mean difference between the actual and predicted values, i.e.,

$$\text{Bias}=\frac{1}{n}\sum_{i=1}^{n}(t_i - t_i^*). \tag{4.21}$$

The ideal value of the bias is zero.

Nash-Sutcliffe efficiency

The Nash-Sutcliffe efficiency index is used to evaluate the predictive power of the hydrologic models. Mathematically it is expressed as,

$$E = 1 - \frac{\sum_{i=1}^{n}(t_i - t_i^*)^2}{\sum_{i=1}^{n}(t_i - \bar{t})^2}. \tag{4.22}$$

where $\bar{t}$ is mean observed value. Nash-Sutcliffe efficiency ranges from negative infinity to 1. An efficiency of 1 corresponds to a perfect match of model prediction to observed data. An efficiency of zero indicates that the model prediction is as accurate as the mean of the observed data. The negative efficiency indicates that the observed mean is a better predictor than the model.

4.5 Results

4.5.1 Parameter estimations

Average Mutual Information function is used for the estimation of τ. Figure 4.3 shows that the AMI function hits its first minimum at τ =10 for north arm. Therefore τ is estimated to be approximately 10 for GSL elevation series. For the sake of GSL elevation prediction, a range between 8 and 14 is, therefore, used.

Past researchers (Abarbanel 1996; Sangoyomi 1993; Sangoyomi et al. 1996) estimated system dimension of GSL volume series approximately 4. Using False nearest neighbor method, the actual system dimension of GSL elevation is estimated to be approximately 3 (Figure 4.4).

Usually, the use of an embedding dimension smaller than 3 deteriorates the model prediction, while an embedding dimension larger than 3 improves. The embedding dimension of 2d+1 is sufficient to unfold the trajectories (Sangoyomi et al. 1996). The embedding dimension is selected in order to sufficiently describe the evolution of the system. When the embedding dimension is too small, the state-space is said to be not fully unfolded, and when it is large, noise might occupy the embedding space (Khalil et al. 2006) which eventually deteriorates the prediction.

Table 4.1 shows the example for multivariate state space construction from a single dimensional series consisting 40 observations using d=5 and τ=8. The first four columns are inputs and the last column is output. These data are further divided into training and a testing set. In the example problem, the first five rows are used for training and the last three rows are used for testing.

The parameters d and τ for the GSL elevation time series are estimated to be 3 and 10, respectively. Therefore the values of embedding dimension (d) used is between 3 and 9 and values of time delay (τ) is used in between 8 and 14. Using the combinations of d and τ, a multi-dimensional phase space is constructed from the single dimensional series of lake elevation. The value τ =8 predicts the lake elevation for the next 4 months at biweekly time step, while τ =14 predicts the lake elevation for the next 7 months. The model is trained and its performance is evaluated in the test phase based on the Nash-Sutcliffe efficiency, the RMSE, and the bias.

4.5.2 Support Vector Machine

GSL elevation was predicted for multiple lead times using different combinations of time delay and embedding dimension. The combination of d=3 and τ=8 predicted better, which corresponds to prediction at biweekly time step for next 4 months. The result is shown in Figure 4.5. We can see that SVM predicts the GSL water level fairly accurate. The predicted water surface elevation has reasonable agreement with observed water surface elevation of the lake. The model developed, therefore, has appropriately captured the evolution of lake elevation. The plot of predicted versus actual elevation are tightly grouped around the 45° line, which shows their values are fairly similar. Residual plots (Figure 4.5e) appear to be random, which is persuasive evidence that the model has no serious deficiencies. Table 4.2 summarizes the training and testing periods, resulting bias, RMSE, and efficiency for the results presented in Figure 4.5. *South* and *North* refer to the southern and northern arms of the GSL. In all the cases, high values of efficiency

(>0.85) were obtained, and the test results were consistently good. This supports the idea that the SVM model developed is consistently accurate and reliable.

4.5.3 Effect of embedding dimension in model prediction for SVM

Figure 4.6 shows test RMSE versus embedding dimension for SVM for each time delay. The plot shows that the best test RMSE is usually obtained at d=3. This dimension is equal to the actual dimension of the GSL elevation series. With the increase of d to 5, the model result did not improve. This may be because the dynamic of the system is not fully unfolded. When d is increased to 7, which correspond to $2d$+1, where the dynamics of the system fully unfold, the test result improved.

Figure 4.7 shows the efficiency versus embedding dimension for the GSL elevation prediction. Good predictions were normally obtained using τ =8 and 10, where 10 is the estimated time delay for GSL elevation series. The prediction using τ =13 and 14 are poor. When τ is large, the chaotic behavior of the signal makes the measurements $y(t)$ and $y(t+\tau)$ independent in a practical sense. Therefore, GSL water level prediction using τ =13 and 14 are not used for the analysis.

As before, the efficiency is high at d= 3. With the increase of d to 5, the test results slightly deteriorated. The test result is improved at d=7. Further increase in d deteriorated the prediction because of intrusion of noise in the system.

4.5.4 Relevance Vector Machine

MVRVM model was again used to predict GSL elevation for multiple lead times using different combination of time delay and embedding dimension. Some results are shown in Figure 4.8, where lake level was predicted at biweekly time step for next 4 months. MVRVM predictions of the GSL water level are reasonably accurate. The predicted water surface elevations have reasonable agreement with observed water surface elevation of the lake. The model has appropriately captured the evolution of lake elevation. The plot of predicted versus actual elevation are tightly saturated around the 45° line, which indicates their values are fairly similar. Uncertainty is captured through confidence interval of predictions. The residual plots are random, which indicates there is no serious problem in the modeling lake elevations. Table 4.3 summarizes the training and testing periods, resulting bias, RMSE, and efficiency for the results presented in Figure 4.8. In all the cases, high values of efficiency were obtained (>0.80). This shows the MVRVM model is consistently accurate and reliable for predicting water surface elevation of the lake.

4.5.5 Effect of embedding dimension in model prediction for MVRVM

Figure 4.9 shows test RMSE versus embedding dimension for each time delay for GSL elevation prediction using MVRVM. The plot shows the best test RMSE is usually obtained at d=3. With the increase of d to 5, the test RMSE did not improve but slightly deteriorated. When d was increased to 7, the test RMSE improved relatively because the dynamic of the system is fully unfolded. When d was further increased, the results eventually deteriorated, which may be because of noise. Better results were usually obtained at τ =8 and 10. Relatively poorer results were obtained for higher τ (say 14) which may be because the GSL elevations are practically independent at higher τ .

Figure 4.10 shows the efficiency versus embedding dimension for the GSL elevation prediction. As before, better results were obtained at d=3. With the increase of d, the prediction result did not improve until the dynamics of the system is fully unfolded. When the embedding dimension is 7, a relatively higher value of efficiency was obtained. With the further increase in the embedding dimension, the efficiency of the prediction results deteriorated because of intrusion of noise in the system. Good predictions were normally obtained using τ = 8 and 10. τ =14 deteriorated the prediction because of weak correlation between water surface elevations.

4.5.6 Generalization and robustness of models

The bootstrap method was used to estimate the measure of variability of test statistics with the change in nature of input data. This method measures the robustness and generalization capability of the model. This is done by randomly drawing a large number of "resamples" of size *n* from the original sample, with replacement. Although each resample will have the same number of elements as the original sample, it could include some of the original data points more than once, and some points will not be included. This process for forming the training set is random and the resulting sets are treated as independent sets which depart from the original sample (Duda et al. 2000). The statistics calculated from these resampling processes takes on slightly different values for the different samples. Having computed statistics each time, a histogram of test statistics of GSL water elevation prediction is prepared. The narrow bounds shown in the histograms indicate that the model is robust. This result indicates there was not much variability in prediction results with the change in the nature of the input data. Figure

4.11 shows the histogram of bootstrap analysis from the SVM model for each arm of the lake for 1991-2008. The horizontal axis of histogram shows the value of test statistics while the vertical axis shows frequency. The histogram of test statistics has narrow bounds which indicate the model is robust. The test statistics of the original model lies in between the 2.5th and the 97.5th percentiles for the bootstrap results. This result shows the model is consistent and well generalized. Figure 4.12 shows the bootstrap analysis of MVRVM model for each arm of the lake for 1991-2008. The test statistics for this case also shows narrow bounds in the histograms. The actual model test statistics also lies within the 2.5th and 97.5th percentiles of the bootstrap results (shown by dotted line in Figure 4.12). These results show that the proposed model is robust and consistent, hence it can be used reliably as a prediction model for Great Salt Lake water surface elevation.

4.6 Discussion and Conclusion

The one-dimensional time series of the Great Salt Lake (GSL) elevation was used to develop a multi-dimensional phase space using the concept of phase construction to represent the underlying dynamics of the system. This reconstruction is a way of approximating the unknown function that describes the state evolution of the chaotic system. The actual system dimension and time delay were estimated for GSL elevation series. Based on that, different combinations of embedding dimension and time delay were used to develop multi-dimensional phase space, which was used to predict the lake elevation using data driven model that uses machine learning approach. Support Vector Machine (SVM) and Multivariate Relevance Vector Machine (MVRVM) were used in this paper. Relatively better prediction of GSL water surface elevations was obtained at time delay (τ) 8 and 10. Similarly the embedding dimension (d) of 3 and 7 produced relatively better predictions. The optimal combination is used to develop the final prediction model for GSL water surface elevation. The data were analyzed on both arms of the lake at two time periods: 1982 to 1987, when a dramatic rise of the GSL was observed, and 1991 to 2008, when the normal rise/fall of the lake level was observed. The model parameters were optimized in the training phase and its performance was evaluated in the test phase based on Bias, RMSE and efficiency. Both SVM and MVRVM models were able to extract the dynamics of the system using only few past observed data points out of the training samples. The results show good agreement between the actual and predicted values of GSL water surface elevation. This good performance in the testing phase shows that the model has good predictive abilities. The GSL elevation predictions from the past researchers are concentrated for the lake volume. The USGS website has a hypsographic curve which translates the GSL volume to the elevation of southern arm, however, there is no official hypsographic curve for the northern arm of the lake. This research independently predicts the lake elevation for both arms. The previous research estimates the embedding dimension and time delay parameter for the total lake volume, based on which predictions are made. This paper estimates those two parameters for the lake elevation, which may be used for predicting lake elevation of two arms independently for well ahead of time. Bootstrap analysis was used to test the reliability and robustness of the model. The narrow bound of test statistics of water surface elevation prediction in the histograms shows that the model is robust and well generalized. The test statistics from the original model lies within the 2.5th and the

97.5th percentile for the bootstrap test statistics. This proves the model is robust and good enough to use as a forecast model for GSL water surface elevation. The prediction results of both the SVM and the MVRVM models were comparable. The MVRVM model was able to capture the uncertainty in both data and model in the form of confidence intervals of GSL water surface elevation prediction in test phase; however the SVM model was able to accurately predict only the mean value of GSL water surface elevation.

References

Abarbanel, H. D. I. (1996). *Analysis of observed chaotic data*, Springer-Verlag, New York.

Asefa, T., Kemblowski, M., McKee, M., and Khalil, A. (2006). "Multi-time scale stream flow predictions: The support vector machines approach." *J. Hydrol.*, 318(1-4), 7-16.

Asefa, T., Kemblowski, M. W., Urroz, G., McKee, M., and Khalil, A. (2004). "Support vectors-based groundwater head observation networks design." *Water Resour. Res.*, 40(11), W11509.

Berger, J. O. (1993). *Statistical decision theory and Bayesian analysis/ Second Edition*, Springer-Verlag, New York.

Duda, R. O., Hart, P. E., and Stork, D. G. (2000). *Pattern classification*, Wiley Interscience, Second Edition, New York.

Fraser, A. M. (1989). "Information Theory and Strange Attractors." PhD dissertation, University of Texas, Austin.

Fraser, A. M., and Swinney, H. L. (1986). "Independent coordinates for strange attractors from mutual information." *Phys. Rev. A*, 33(2), 1134-1140.

Gallager, R. G. (1968). *Information theory and reliable communication*, John Wiley & Sons, Inc. , New York.

Gill, M. K., Kaheil, Y. H., Khalil, A., McKee, M., and Bastidas, L. (2006). "Multiobjective particle swarm optimization for parameter estimation in hydrology." *Water Resour. Res.*, 42(7), W07417.

Kennel, M. B., Brown, R., and Abarbanel, H. D. I. (1992). "Determining embedding dimension for phase-space reconstruction using a geometrical construction." *Phys. Rev. A*, 45(6), 3403-3411.

Khalil, A. F., McKee, M., Kemblowski, M., and Asefa, T. (2005a). "Basin scale water management and forecasting using artificial neural networks." *JAWRA Journal of the American Water Resources Association*, 41(1), 195-208.

Khalil, A. F., McKee, M., Kemblowski, M., and Asefa, T. (2005b). "Sparse Bayesian learning machine for real-time management of reservoir releases." *Water Resour. Res.*, 41(11), W11401.

Khalil, A. F., McKee, M., Kemblowski, M., Asefa, T., and Bastidas, L. (2006). "Multiobjective analysis of chaotic dynamic systems with sparse learning machines." *Adv. Water Res.*, 29(1), 72-88.

Koutsoyiannis, D., and Pachakis, D. (1996). "Deterministic chaos versus stochasticity in analysis and modeling of point rainfall series." *J. Geophys. Res.*, 101(D21), 26441-26451.

Lall, U., Sangoyomi, T., and Abarbanel, H. D. I. (1996). "Nonlinear dynamics of the Great Salt Lake: Nonparametric short-term forecasting." *Water Resour. Res.*, 32(4), 975-985.

Lorenz, E. N. (1963). "Deterministic Nonperiodic Flow." *J Atmospheric sciences*, 20, 131-141.

Sangoyomi, T. (1993). "Climatic variability and dynamics of Great Salt Lake hydrology." PhD Dissertation, Utah State University, Logan UT.

Sangoyomi, T. B., Lall, U., and Abarbanel, H. D. I. (1996). "Nonlinear Dynamics of the Great Salt Lake: Dimension Estimation." *Water Resour. Res.*, 32(1), 149-159.

Schlesinger, S. (1979). "Terminology for model credibility." *Simulation*, 32(3), 103-104.

Takens, F. (1981). "*Detecting strange attractors in turbulence*." Dynamical Systems and Turbulence, Warwick 1980, D. Rand, and L.-S. Young, eds., Springer Berlin / Heidelberg, 366-381.

Thayananthan, A. (2005). "Template-based pose estimation and tracking of 3D hand motion." PhD Dissertation, University of Cambridge, Cambridge, UK.

Thayananthan, A., Navaratnam, R., Stenger, B., Torr, P., and Cipolla, R. (2008). "Pose estimation and tracking using multivariate regression." *Pattern Recognition Lett.*, 29(9), 1302-1310.

Ticlavilca, A. (2010). "Multivariate Bayesian machine learning regression for operation and management of multiple reservoir, irrigation canal, and river systems." PhD Dissertation, Utah State University, Logan, UT.

Tipping, M. (2000). "The Relevance Vector Machine." *Proc., Advances in Neural Information Processing Systems*, The MIT Press, 652-658.

Tipping, M. (2001). "Sparse Bayesian learning and the Relevance Vector Machine." *J. Machine Learning Res*., 1, 211-244.

Tipping, M. E., and Faul, A. C. (2003). "Fast marginal likelihood maximization for sparse Bayesian models." *Proc., Ninth International Workshop on Artificial Intelligence and Statistics*.

Tripathi, S., and Govindaraju, R. (2006). "On selection of kernel parameters in Relevance Vector Machines for hydrologic applications." *Stoch. Eviron. Res. Risk Asses*., 21, 747-764.

Tychonoff, A. N., and Arsenin, V. Y. (1977). *Solutions of ill posed problems*, Winston & Sons, Washington.

Vapnik, V. N. (1995). *The nature of statistical learning theory*, Springer Verlag, New York.

Vapnik, V. N. (1998). *The nature of statistical learning theory*, Springer Verlag, New York.

Table 4.1 Construction of multivariate state space out of single measurement using d=5 and τ =8

	Inputs				Output
Training	x1	x9	x17	x25	x33
	x2	x10	x18	x26	x34
	x3	x11	x19	x27	x35
	x4	x12	x20	x28	x36
	x5	x13	x21	x29	x37
Test	x6	x14	x22	x30	x38
	x7	x15	x23	x31	x39
	x8	x16	x24	x32	x40

Table 4.2 Results for Figures 6(a)-6(d)

Figure number	Training Period	Test Period	Bias	RMSE	Efficiency
6(a) South	09/01/1982 09/01/1985	09/15/1985 06/15/1987	0.034 m (0.11 ft)	0.113 m (0.37 ft)	0.89
6(b) South	09/01/1991 09/01/2004	09/15/2004 12/15/2008	0.006 m (0.02 ft)	0.095 m (0.31 ft)	0.92
6(c) North	09/01/1981 09/01/1985	09/15/1985 12/15/1987	0.006 m (0.02 ft)	0.104 m (0.34 ft)	0.90
6(d) North	09/01/1991 09/01/2004	09/15/2004 12/15/2008	0.012 m (0.04 ft)	0.098 m (0.32 ft)	0.91

Table 4.3 Results from Figures 9(a)-9(d)

Figure number	Training Period	Test Period	Bias	RMSE	Efficiency
9(a) South	09/01/1982 09/01/1985	09/15/1985 06/15/1987	0.076 m (0.25 ft)	0.149 m (0.49 ft)	0.81
9(b) South	09/01/1991 09/01/2004	09/15/2004 12/15/2008	0.009 m (0.03 ft)	0.091 m (0.30 ft)	0.92
9(c) North	09/01/1981 09/01/1985	09/15/1985 12/15/1987	0.012 m (0.04 ft)	0.082 m (0.27 ft)	0.94
9(d) North	09/01/1991 09/01/2004	09/15/2004 12/15/2008	0.013 m (0.04 ft)	0.082 m (0.27 ft)	0.93

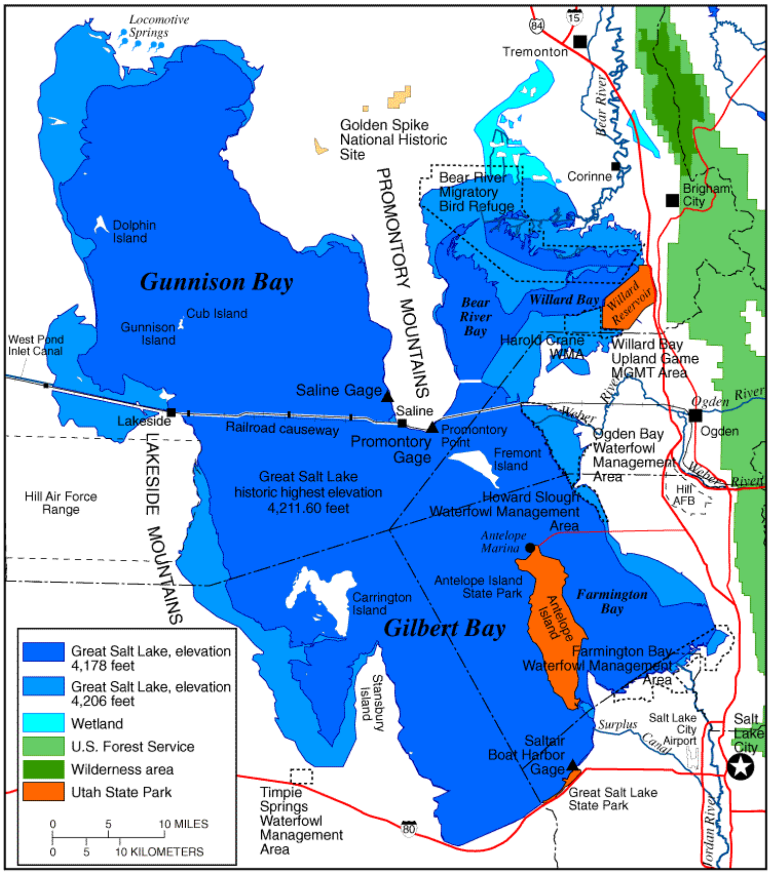


Figure 4.1 The Great Salt Lake.

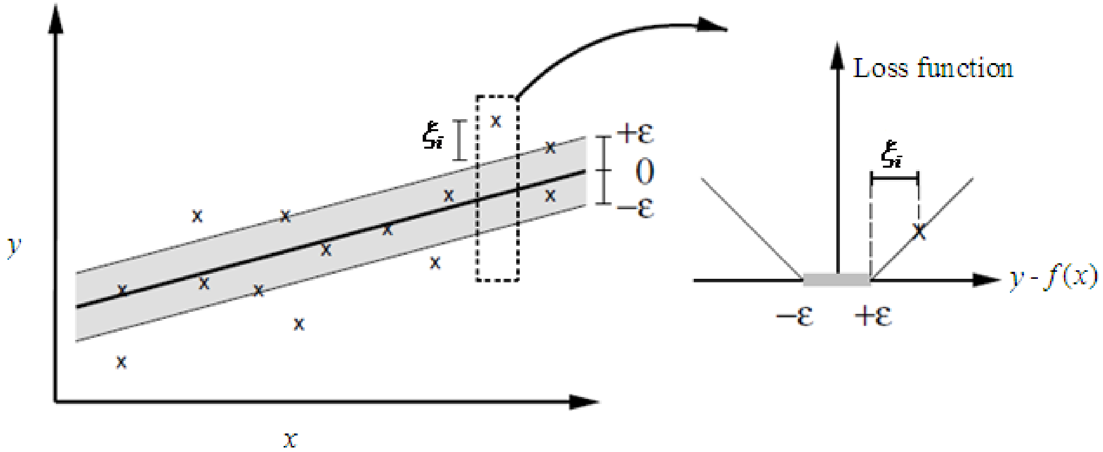


Figure 4.2 ε - sensitive loss function in Support Vector Machine.

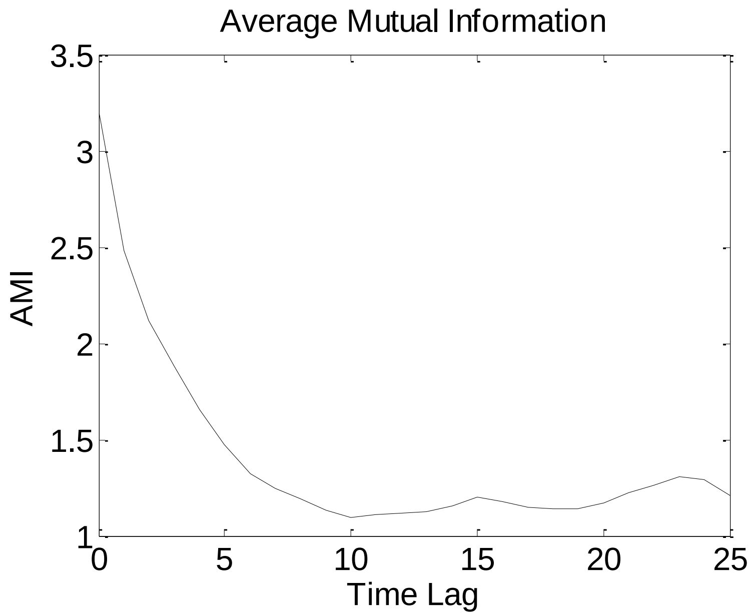


Figure 4.3 Average Mutual Information versus Time lag for GSL water level.

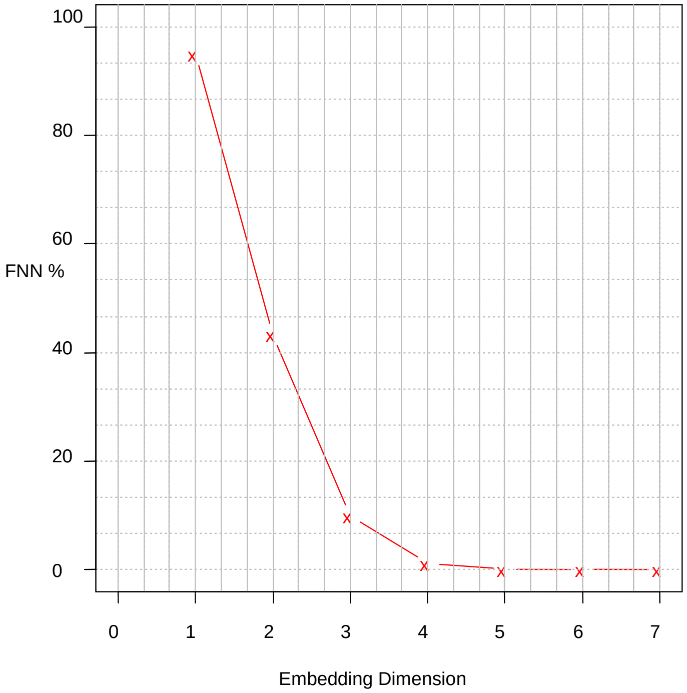


Figure 4.4 Estimating embedding dimension from False Nearest Neighbor method.

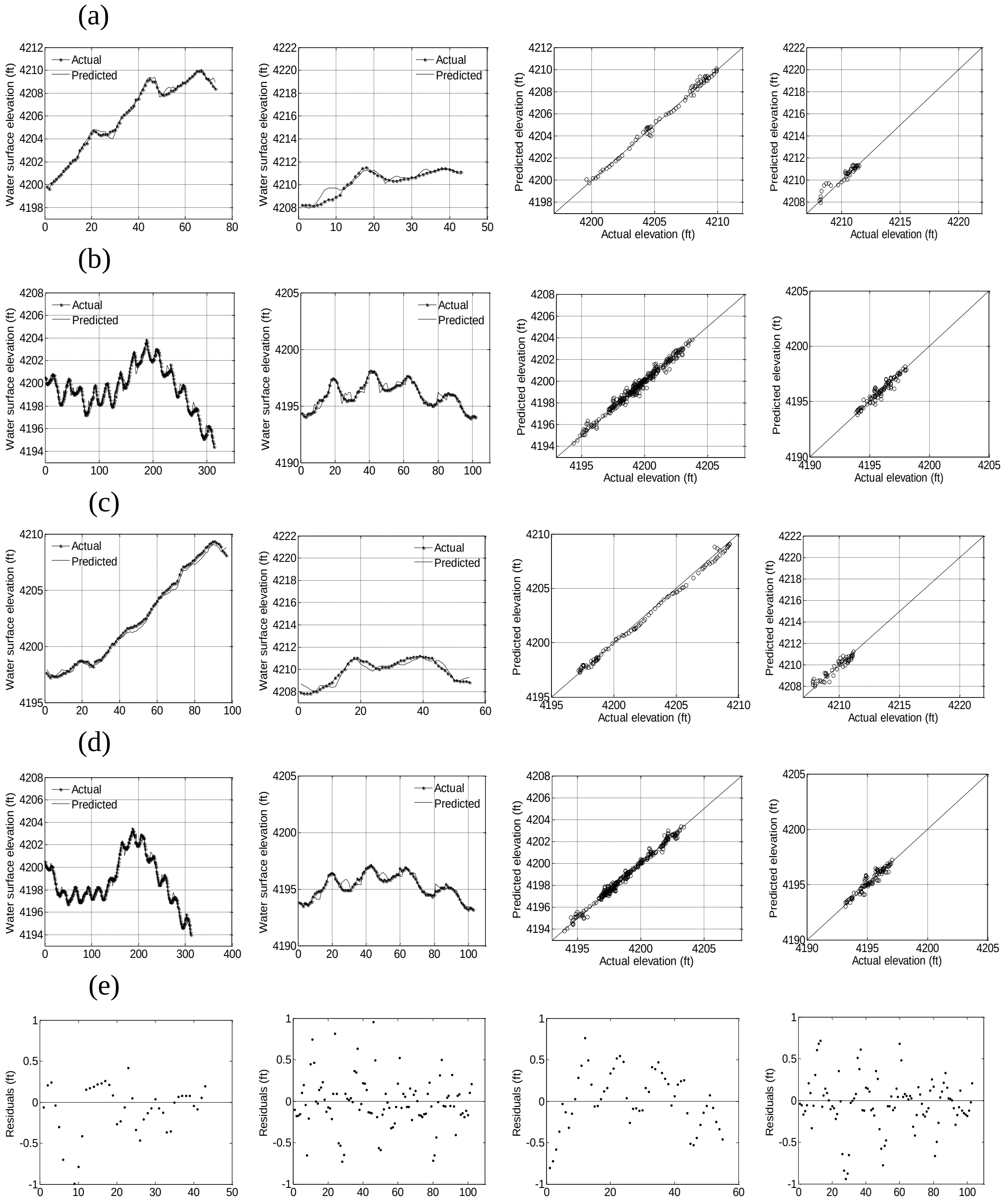


Figure 4.5 The prediction of GSL water level from SVM. For (a)-(d), first column shows training phase, second column shows testing phase, third column shows predicted versus actual elevation for training phase, and fourth column shows similar plot for test phase. (a) Southern arm of the lake for 1982 to 1987, (b) Southern arm of lake for 1991 to 2008, (c) Northern arm of the lake for 1982 to 1987, (d) Northern arm of lake for 1991 to 2008, and (e) Residual plots for (a) to (d).

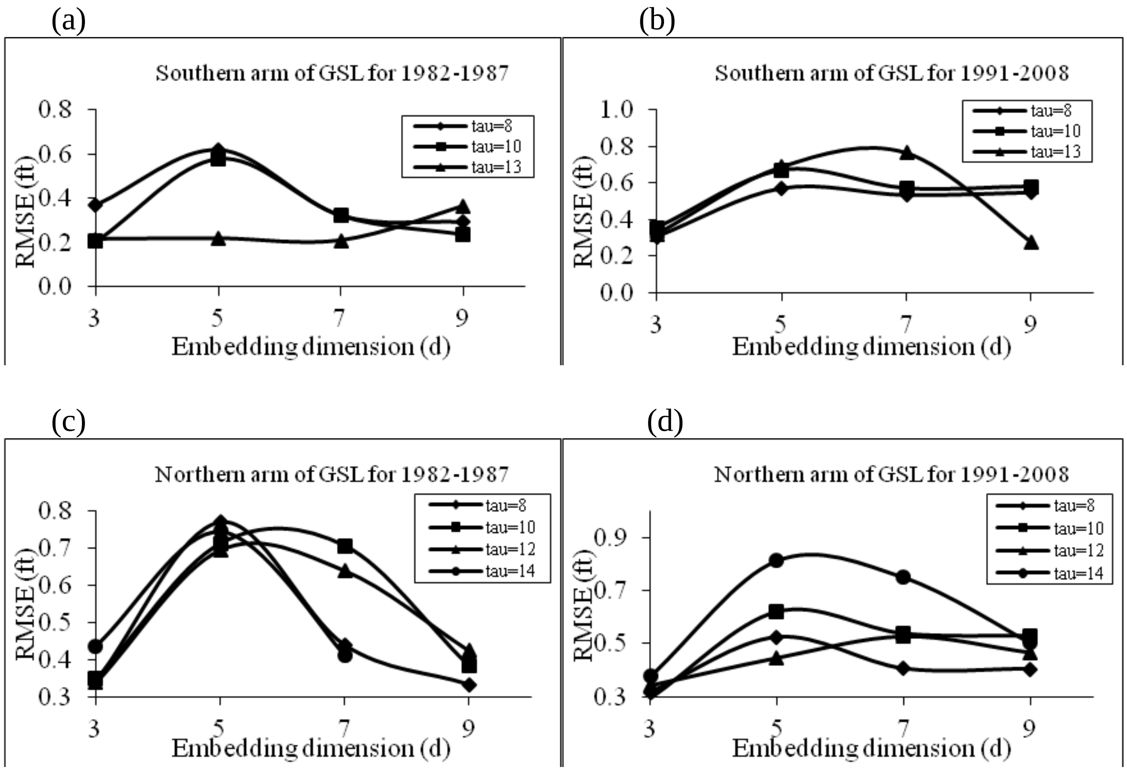


Figure 4.6 RMSE versus embedding dimension for southern arm of lake for (a) 1982 to 1987, (b) 1991 to 2008, and northern arm of the lake for (c) 1982 to 1987, and (d) 1991 to 2008.

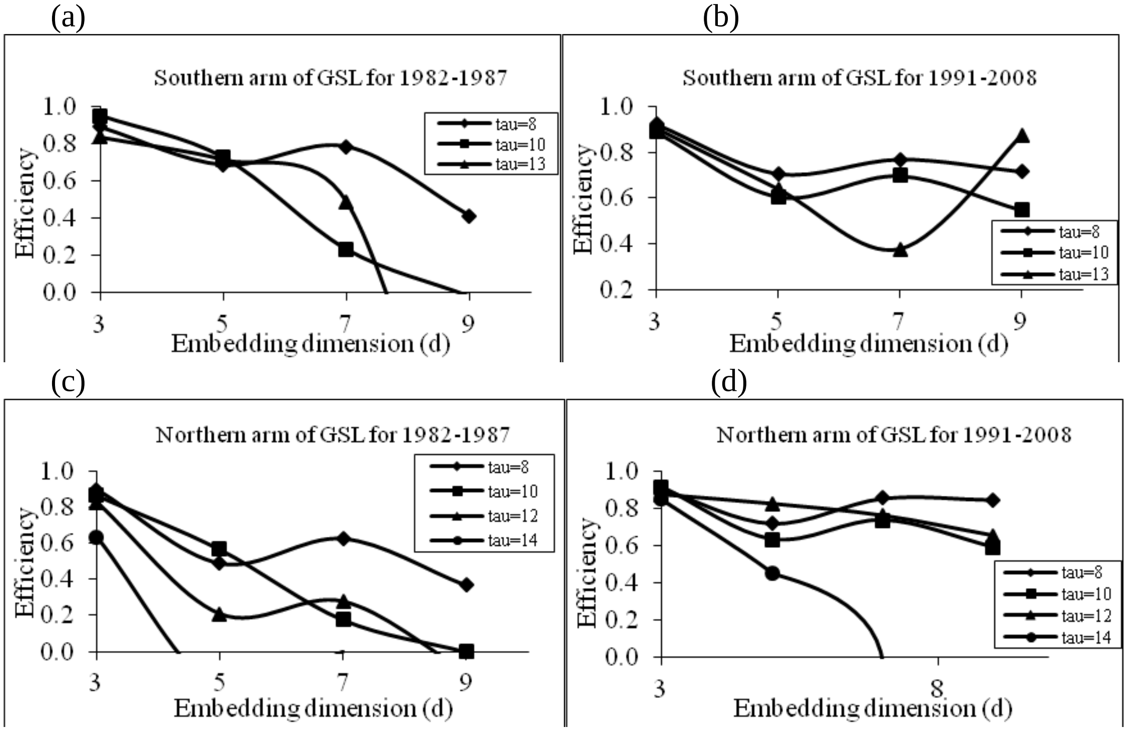


Figure 4.7 Efficiency versus embedding dimension for southern arm of the lake for (a) 1982 to 1987, (b) 1991 to 2008, and northern arm of the lake for (c) 1982 to 1987, (d) 1991 to 2008.

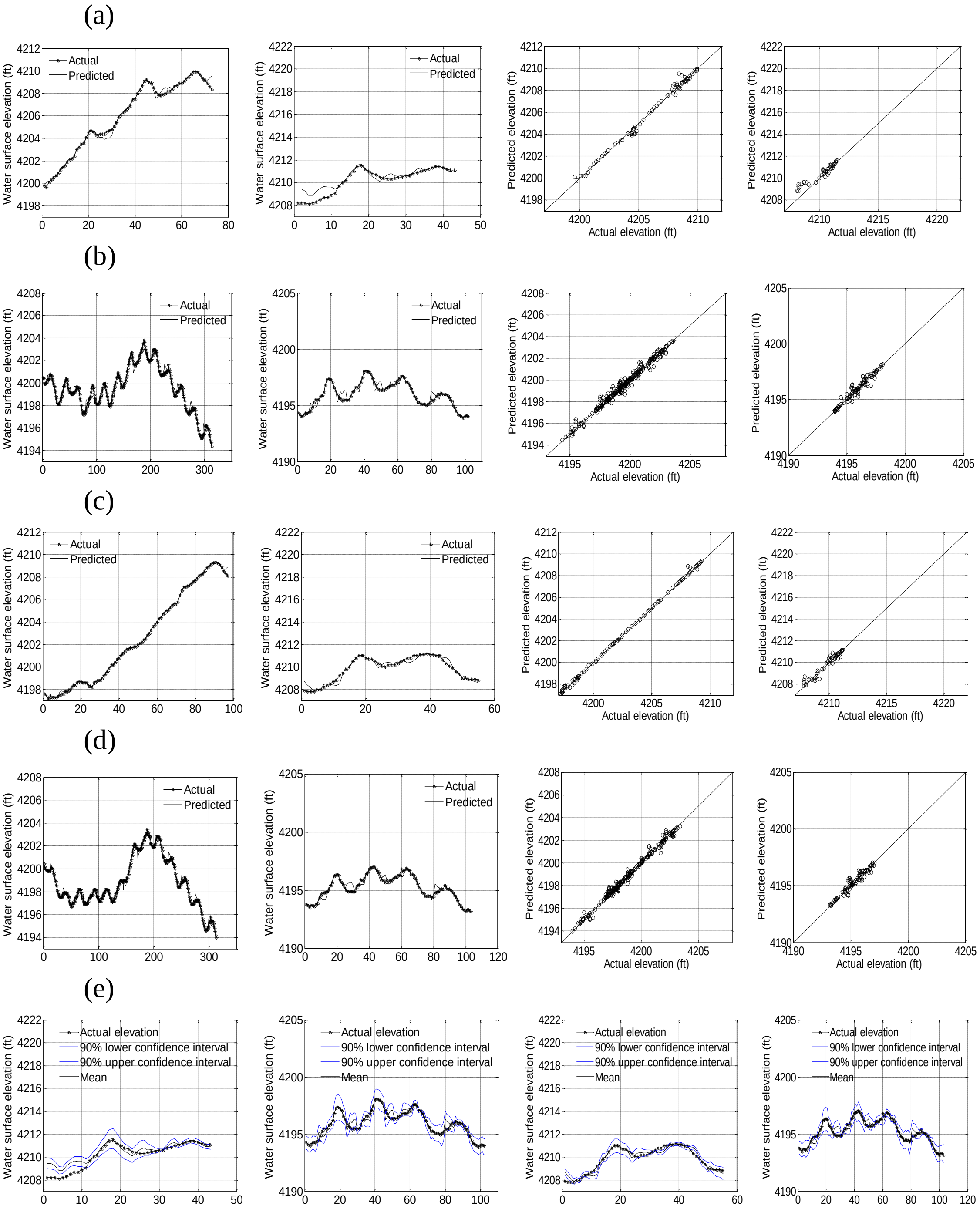


Figure 4.8 The prediction of GSL water level from MVRVM. For (a)-(d), first column shows training phase, second column shows testing phase, third column shows predicted versus actual elevation for training phase, and fourth column shows similar plot for test phase. (a) Southern arm of the lake for 1982 to 1987, (b) Southern arm of lake for 1991 to 2008, (c) Northern arm of the lake for 1982 to 1987, (d) Northern arm of lake for 1991 to 2008, (e) 90 percent confidence interval for (a) to (d), and (f) Residual plots for (a) to (d).

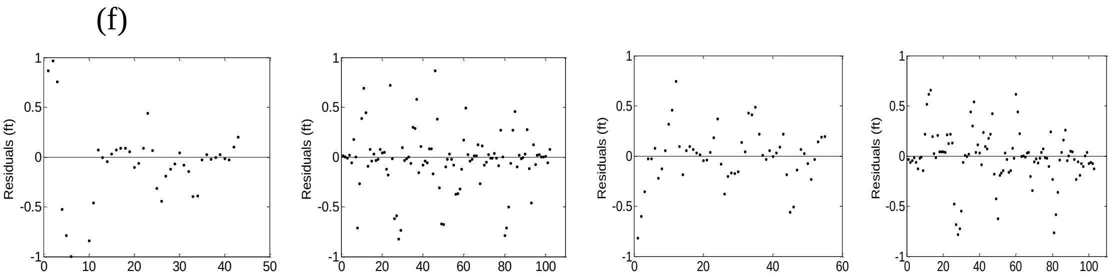


Figure 4.8 Cont.

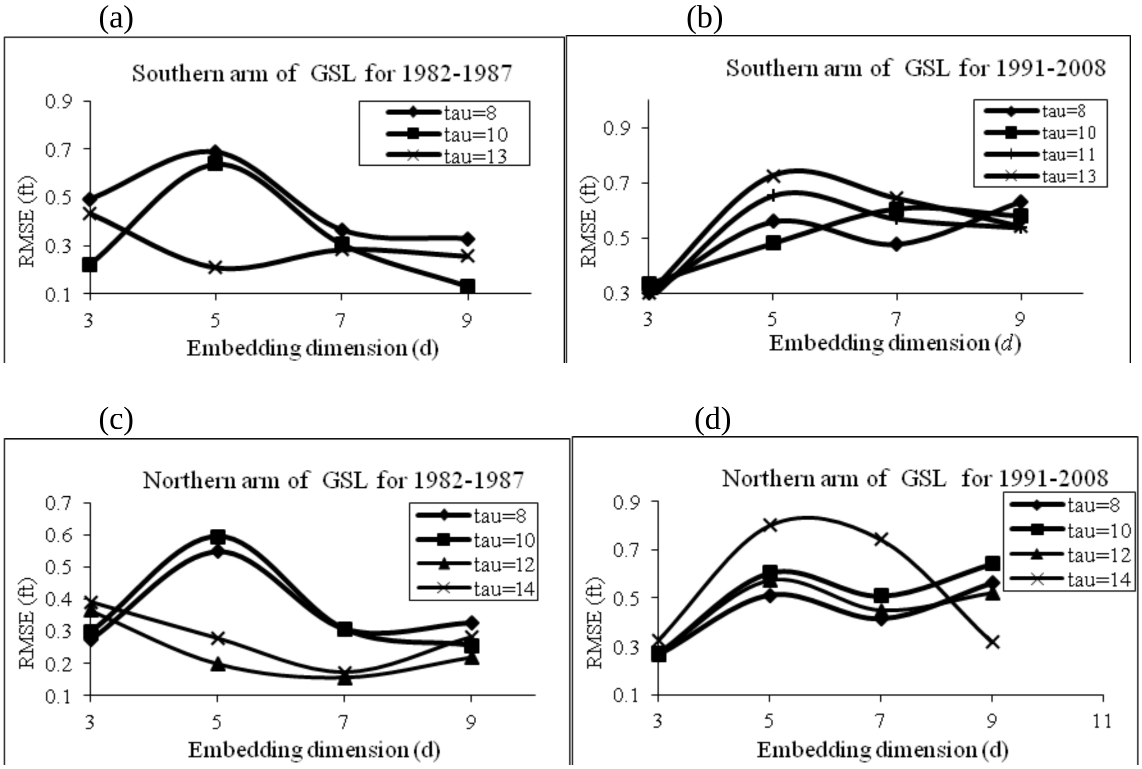


Figure 4.9 RMSE versus embedding dimension for southern arm of lake for (a) 1982 to 1987, (b) 1991 to 2008, and northern arm of the lake for (c) 1982 to 1987, and (d) 1991 to 2008.

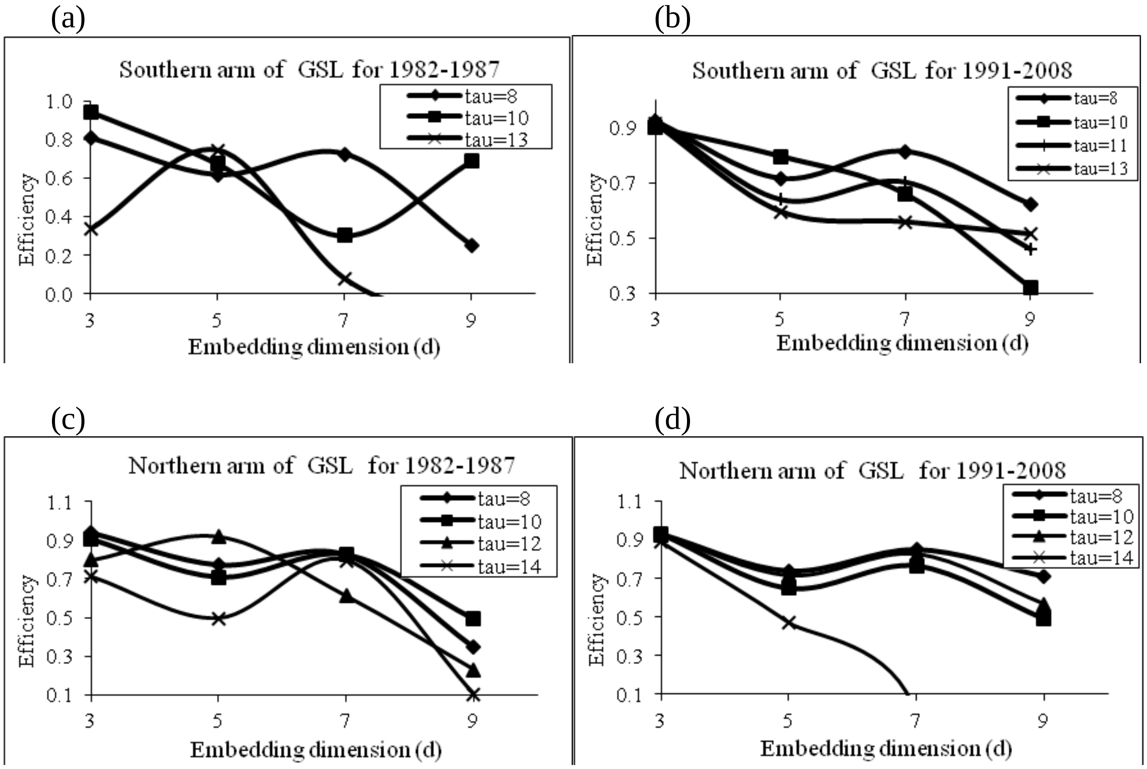


Figure 4.10 Efficiency versus embedding dimension for southern arm of the lake using MVRVM model for (a) 1982 to 1987, (b) 1991 to 2008, and northern arm of the lake for (c) 1982 to 1987, (b) 1991 to 2008.

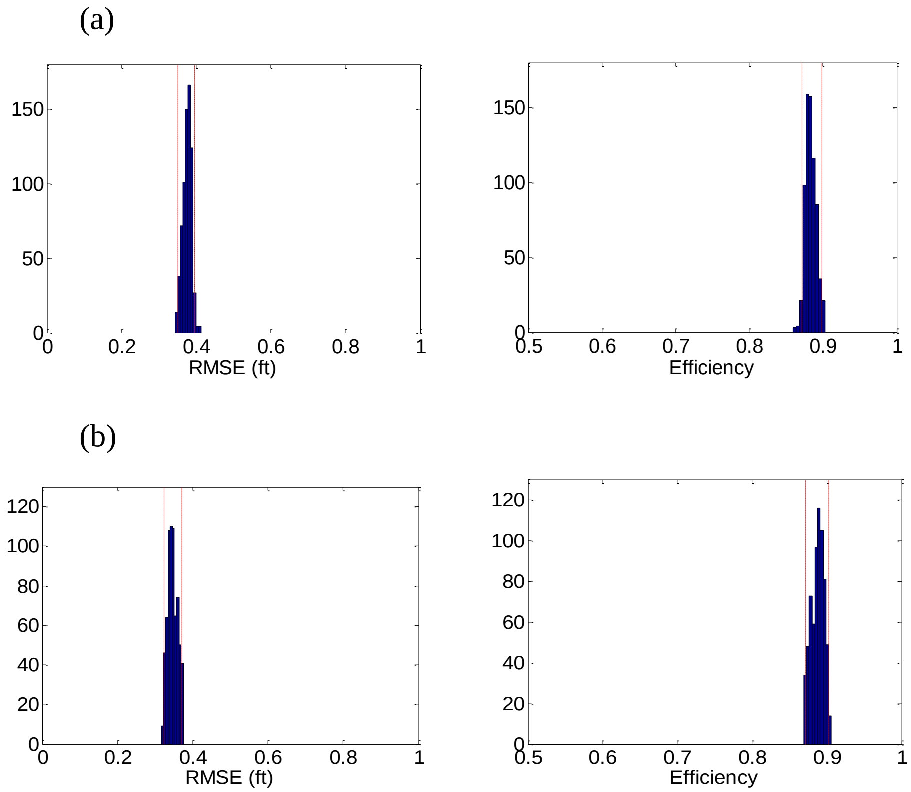


Figure 4.11 Bootstrap analysis for RMSE and Efficiency for SVM model for 1991-2008 (a) southern arm of the lake, and (b) northern arm of the lake.

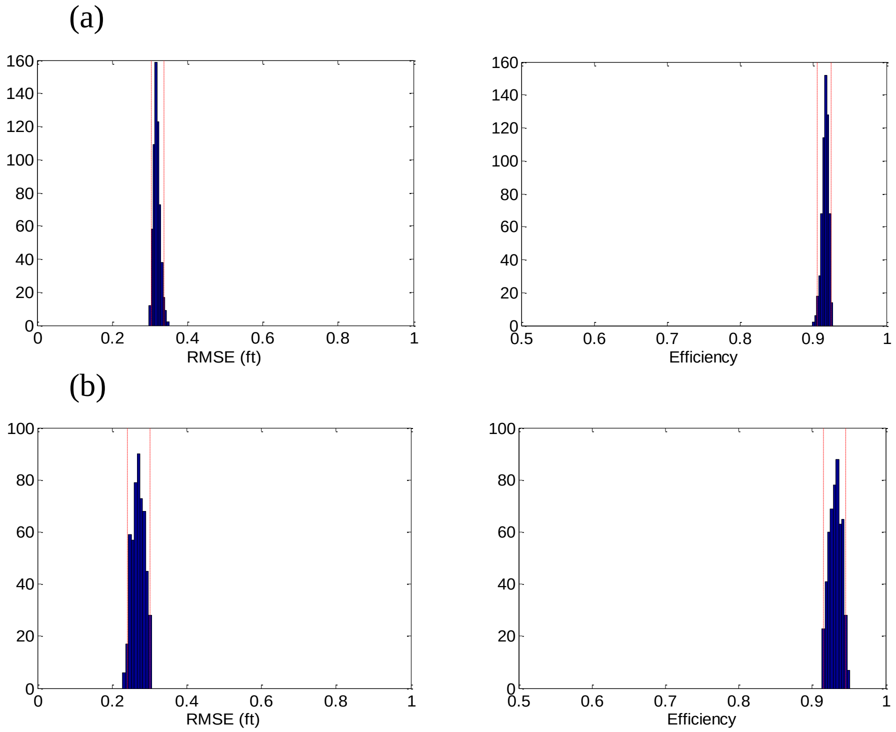


Figure 4.12 Bootstrap analysis for RMSE and Efficiency using MVRVM for 1991-2008 (a) southern arm of the lake, and (b) northern arm of the lake.

CHAPTER 5

SUMMARY, CONCLUSION, AND RECOMMENDATION

5.1 Summary and Conclusion

This dissertation has shown the applicability of machine learning models for identifying the influential climate indicators and forecasting different hydrological variables in river basin scale as an alternative to physically based models. The research is conducted in three different areas where data-driven models based on machine learning approach are used to choose appropriate climate indicators and develop the prediction model in order to solve the water resource planning and management problems at the basin scale. The models showed promising results. They are accurate and robust. They are capable of providing valuable information about the future water availability and future state of water resource system in the basin scale.

Support Vector Machine (SVM) and Multivariate Relevance Vector Machine (MVRVM) were used in this dissertation. The MVRVM model (Thayananthan 2005) is an extension of traditional Relevance Vector Machine developed by Tipping and Faul (2003). It retains all properties of conventional RVM such as accuracy, robustness, and sparseness. Using the Bayesian approach, the uncertainty in both data and model was captured and presented in the form of confidence interval of prediction. The MVRVM model was applied for monthly, seasonal and annual streamflow prediction at each selected stream gages of Utah that spatially covers the state from southern to northern region. The input variables (climate indicators) that produce the best test statistics were identified for each selected gage, and then used them to develop the final forecast model.

The MVRVM model was also used for multiple lead-time prediction of water surface elevation of Great Salt Lake along with SVM, which provides the opportunity to compare the results.

In Chapter 2, influential locations of circulation indicators (sea surface temperature) were identified for five stream gages in Utah that spatially covers the state from North to South. Using the sea surface temperature (SST) of selected locations along with other local inputs, monthly average discharge and total volume of water passing the stream gage was predicted for next six months. The local inputs to the model were represented by past streamflow data, snowpack in the mountain, and local meteorological condition while the regional climatic condition was represented by sea surface temperatures in the Pacific and Atlantic oceans. The input variables were integrated into the machine learning framework to develop a streamflow forecast model. The MVRVM successfully transformed the input variables into reasonably accurate forecasting of outputs. The performance of the model was evaluated based on RMSE and the Nash-Sutcliffe efficiency in the test phase. The result shows the sea surface temperature of Pacific Ocean predicts better than that of Atlantic Ocean. Since the Pacific Ocean represents the majority of ocean-atmosphere climate influence in Western U.S. (Ting and Wang 1997; Wang and Ting 2000), the results make reasonable sense. Although the physical processes responsible for the streamflow generation are not represented, machine learning model predicted accurately from the available inputs by learning the relationship between input and output in a training phase. The accurate and reliable predictions of streamflow are crucial information for farmers and water managers, therefore, the models developed in this dissertation can be useful for those stakeholders.

The reliability and robustness of the model was evaluated from the bootstrap analysis. The narrow bound of histograms resulting from bootstrapping confirms the model is consistent and robust. The successful application of machine learning models in hydrological modeling shows they can be alternatives to expensive and cumbersome physically-based models.

Chapter 3 presents the long lead-time annual streamflow volume prediction at four selected unimpaired stream gages in Utah using oceanic-atmospheric oscillation modes. The correlation between streamflow and climatic variability represented by oceanic-atmospheric oscillation indices is the key point for the prediction. Popular oscillation modes are used as input variables. They are the Pacific Decadal Oscillation (PDO), the El-Nino Southern Oscillation (ENSO), the Atlantic Multi-decadal Oscillation (AMO), and the North Atlantic Oscillation (NAO). Different combinations of oscillation modes are developed and best combinations and corresponding lead-times are identified for each selected gage, which were used to develop the prediction model. PDO and ENSO predicted relatively better than other oscillation indices. The best combinations of oscillations were also identified for each lead time. The performance of the model was evaluated based on the RMSE, efficiency, and correlation coefficient in the test phase. The MVRVM model predicted annual flow volume reasonably well from the oscillation indices. Due to long persistence of these oscillation indices, it is possible to predict the streamflow for long lead-times. The model, however, is not good enough in capturing the extreme events. This shows the oscillation indices used in this paper are not enough to represent the physical processes associated with the generation of streamflow. The bootstrap analysis was used to test the generalization capability of the model. The narrow bound of resulting histograms shows the model is well generalized (i.e. robust). The RMSE of actual model prediction lies in between 2.5th percentile and 97.5th percentile values of the prediction. This shows prediction is good enough for practical use. The comparison of MVRVM to SVM and ANN shows MVRVM outperforms other machine learning models. The pattern of prediction, however, is similar in all machine learning models.

In chapter 4, SVM and MVRVM were used to predict water surface elevation of the Great Salt Lake using past water surface elevation data. The actual system dimension and time delay parameters of GSL water surface elevation series were estimated. Using those parameters, a multivariate input space was constructed from the single variable which unfolds the dynamics of the system. The output was predicted in the form of future water surface elevation of the lake using reconstructed input space. The model was applied for two time periods. One represents the dramatic rise of GSL elevation (1982-1987) while other period represents the normal rise-fall of lake elevation (1991-2008). The test result shows both SVM and MVRVM are able to extract the dynamics of the system using only few observed past water surface elevations from training phase. The prediction results from both SVM and MVRVM were accurate and comparable. The optimum combination of embedding dimension and time delay was also estimated, which may be used to refine the forecast model. The advantage of MVRVM over SVM is that it estimates the uncertainty of the prediction in the form of confidence intervals. The narrow bound in the histograms resulting from the bootstrap analysis shows the model is robust and well generalized.

This dissertation shows the successful application of learning-machines approach in water resource planning and management. Even with the limited knowledge of physical processes, one can come up with a reasonable model using data driven model based on learning machine approach. These models are easy to use and provides accurate and efficient forecast. Since physically-based models are complex and acquires huge amount of data, data-driven models are being used as an alternative to physically based models. Data-driven models are capable of learning dynamic behavior of complex system while accounting for uncertainties (Khalil et al. 2005). The nonlinearity of dynamics of system is learned in the training phase, where the model parameters are optimized, and performance is evaluated in the test phase. The models were tested in a wide range of problems: Monthly and seasonal streamflow prediction, annual streamflow volume prediction, and water surface elevation prediction of the Great Salt Lake. The accuracy of the models was evaluated based on the RMSE and efficiency in the test phase. In all those diverse problem types, MVRVM performed accurately. In Chapter 2 and 4, high accuracy of the prediction was obtained while reasonable accuracy was obtained in Chapter 3. MVRVM also computes the uncertainty of both model and data for the predicted result. In Chapter 3, the results from MVRVM were compared to the results from SVM and ANN. Comparison shows MVRVM outperformed both SVM and ANN. SVM was also used for the GSL water surface elevation prediction. The prediction results of SVM were reasonably accurate and comparable to that of MVRVM.

Developing an accurate model in a complex water system is important, and making it robust is also equally important. The robustness of the model is evaluated from the bootstrap analysis. The narrow bound of the resulting histogram means the model is robust. This implies the model prediction will not change much with the changes in nature of input data. In all chapters, bootstrap analysis shows the narrow bound of resulting histogram where the actual test statistics lies in between 2.5th percentile and 97.5th percentile values. This confirms the models herein developed were robust and consistent.

The dissertation shows MVRVM and SVM are applicable in a wide range of problems in hydrology, and are capable of making accurate and robust predictions. The data driven model based on machine-learning approach are, therefore, useful in water resource planning and management.

5.2 Recommendation and Future Direction

The selection of the kernel function is heuristic in machine learning models. More scientific ways of selecting kernel function are preferred. There are at least two model parameters in SVM. Optimizing the model parameters need cross validation, which requires considerable data and time. Auto search of optimal model parameters is highly preferred, and is recommended. Both SVM and RVM make predictions using only few data points in training phase. They are called Support Vectors (SV’s) and Relevance Vectors (RV’s) in SVM and RVM, respectively. The physical meaning of these points is not yet fully explained. A detail research behind the selection of SV’s and RV’s is recommended. This may lead to new approaches in data collection, since only few points are necessary to explain the dynamic of the system, and hence to make predictions. This research promises to cut the expenses of data collection that would not be required to explain the input-output relationship of the system. Future research will show the optimal spatial and temporal location of the data to be collected. In addition, this will reduce model run time, hence saving computer analysis time too.

Autoregressive moving average (ARMA) models are suitable for short range dependence processes. This model also requires considered time series to be stationary. The autocorrelation function (ACF) for the Great Salt Lake elevation decays gradually, which means the time series is non-stationary (Appendix). The estimate of Hurst parameter indicates GSL elevation series has long range dependence (LRD). The decay of autocorrelation function for LRD process is slower than exponential decay, and area under the autocorrelation function is infinite (Sheng and Chen 2011). Fractional autoregressive integrated moving average (FARIMA) time-series model is capable of capturing long range dependence (LRD) as well as SRD, and forecast the process. Non-convergence of variance of GSL elevation series indicates GSL elevation series is non-Gaussian. Therefore, FARIMA with stable innovations model is also suggested as a means of modeling GSL elevation. This is capable of modeling time series which has infinite variance, long-range dependence characteristics, and non-Gaussian signals (exhibit sharp spikes or occasional bursts of outlying observations than Gaussian distribution signals).

References

Khalil, A. F., McKee, M., Kemblowski, M., and Asefa, T. (2005). "Sparse Bayesian learning machine for real-time management of reservoir releases." *Water Resour. Res*., 41(11), W11401.

Sheng, H., and Chen, Y. (2011). "FARIMA with stable innovations model of Great Salt Lake elevation time series." *Signal Processing*, 91(3), 553-561.

Thayananthan, A. (2005). "Template-based pose estimation and tracking of 3D hand motion." PhD Dissertation, University of Cambridge, Cambridge, UK.

Ting, M., and Wang, H. (1997). "Summertime U.S. precipitation variability and its relation to Pacific sea surface temperature." *J. Climate*, 10(8), 1853-1873.

Tipping, M. E., and Faul, A. C. (2003). "Fast marginal likelihood maximization for sparse Bayesian models." Proc., *Ninth International Workshop on Artificial Intelligence and Statistics*.

Wang, H., and Ting, M. (2000). "Covariabilities of winter U.S. precipitation and Pacific sea surface temperatures." *J. Climate*, 13(20), 3711-3719.

APPENDIX

The Great Salt Lake elevation data

The Great Salt Lake (GSL) is divided into two arms by rock-fill causeway. There exists an elevation difference between two arms of GSL due to unequal rate of inflow and evaporation loss from each arm. The U.S. Geological Survey (USGS) operates gages that collect water-surface elevation data in the southern arm of the lake at the Boat Harbor Gage (USGS station 10010000), and on the northern arm of the lake at the Saline Gage (10010100) (http://ut.water.usgs.gov/greatsaltlake/). USGS collected data at biweekly time step before 10/01/1989 and at daily time step after that. In order to make the analysis compatible, biweekly time step is used over entire analysis period in this paper. Water surface elevation data was collected from both stations in the present study. Two time periods are considered for the analysis. They are: 1982-1987, which represents the dramatic rise of GSL elevation, and 1991-2008, which represents the normal rise-fall of the lake elevation. To reduce the effect of flooding, West desert pumping was launched from the April 1987 to June 1989. Therefore, the data before 1987 and after 1991 is used to avoid the disturbance of pumping in the lake elevation.

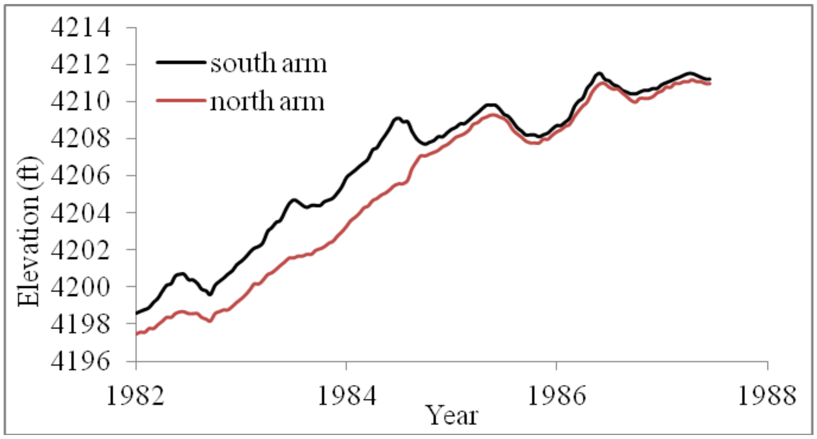


Figure A.1 GSL elevation for south and north arm for 1982-1987.

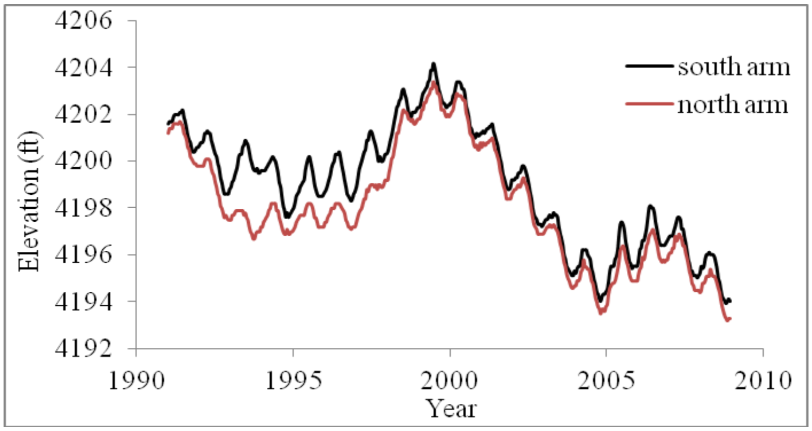


Figure A.2 GSL elevation for south and north arm for 1991-2008.

The GSL elevations for 1982-1987, and 1991-2008 are shown in Figure A.1 and Figure A.2 respectively. Their mean values and variances are shown in Table A.1. The autocorrelation function and partial autocorrelation function for each arm for each time period are shown in Figure A.3. The decay of autocorrelation function (ACF) for GSL elevation time series is gradual, which indicates the GSL elevation time series is non-stationary. For non-stationary process, time dependence exists and matters, unlike stationary process where the effects of the shocks are temporary and the time series reverts to it long-run level (Sheng and Chen 2011). The aggregated variance method and absolute variance method are used to estimate the Hurst parameter for GSL elevation series. The Hurst parameters from two methods for North arm for 1982-2008 are 0.9981 and 0.9866 respectively. This is 0.9981 and 0.9889 for South arm from above two methods respectively. This indicates ($0 < H < 1$) the GSL elevation series has LRD characteristic. The variance of GSL elevation series for each arm is shown in Figure A.4. Non-convergence of variance show the GSL elevation series has non-Gaussian stable distribution. Histogram of GSL elevation for each arm for each time period is shown in Figure A.5.

Table A.1 Observed mean and variance for GSL elevation

	1982-1987		1991-2008	
	south arm	north arm	south arm	north arm
mean (ft)	4206.61	4205.33	4198.98	4198.05
variance (ft^2)	15.98	21.28	6.42	6.26

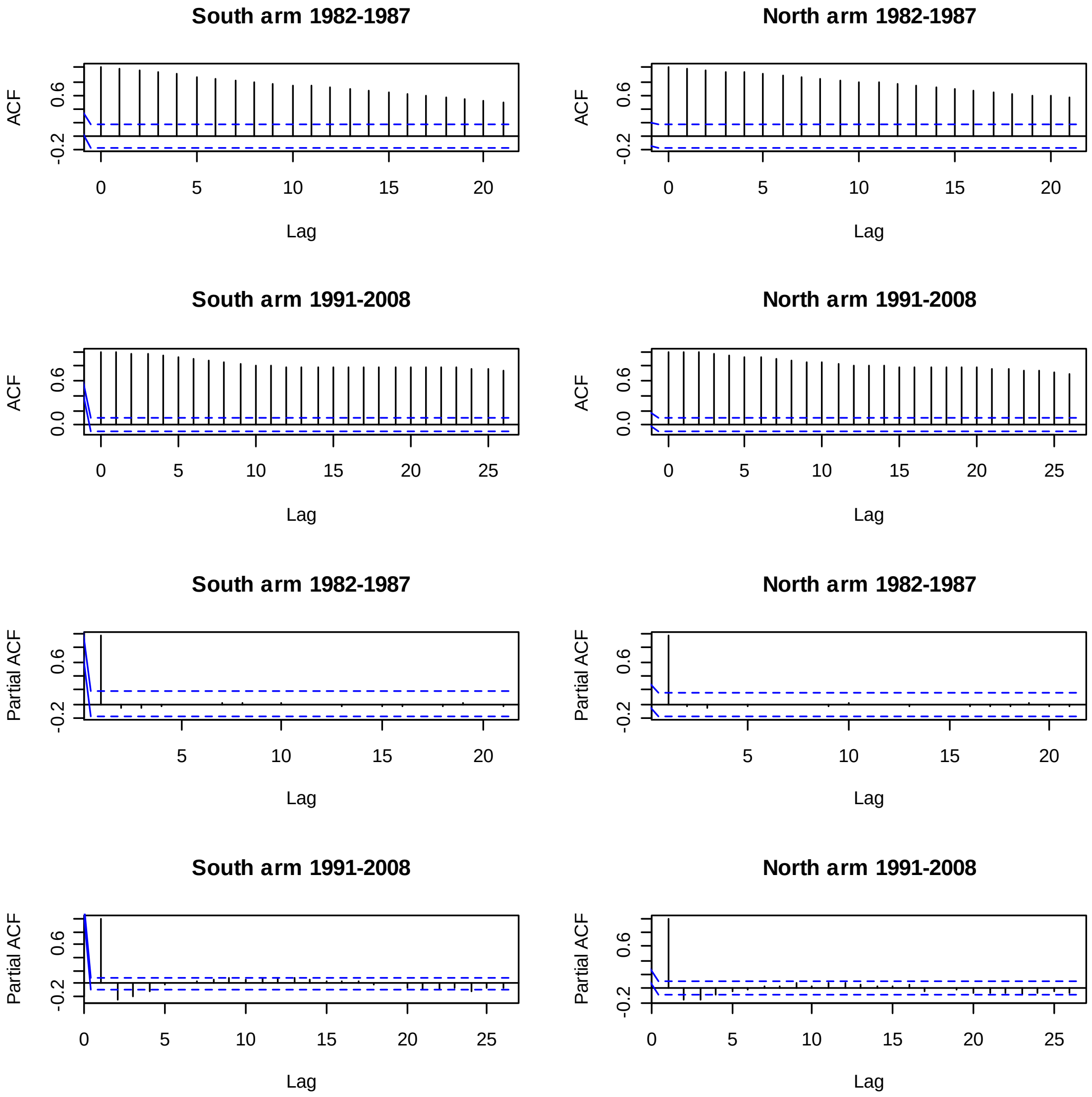


Figure A.3 Autocorrelation and partial autocorrelation function for GSL elevation for each arm of the lake for two time periods: 1982-1987 and 1991-2008.

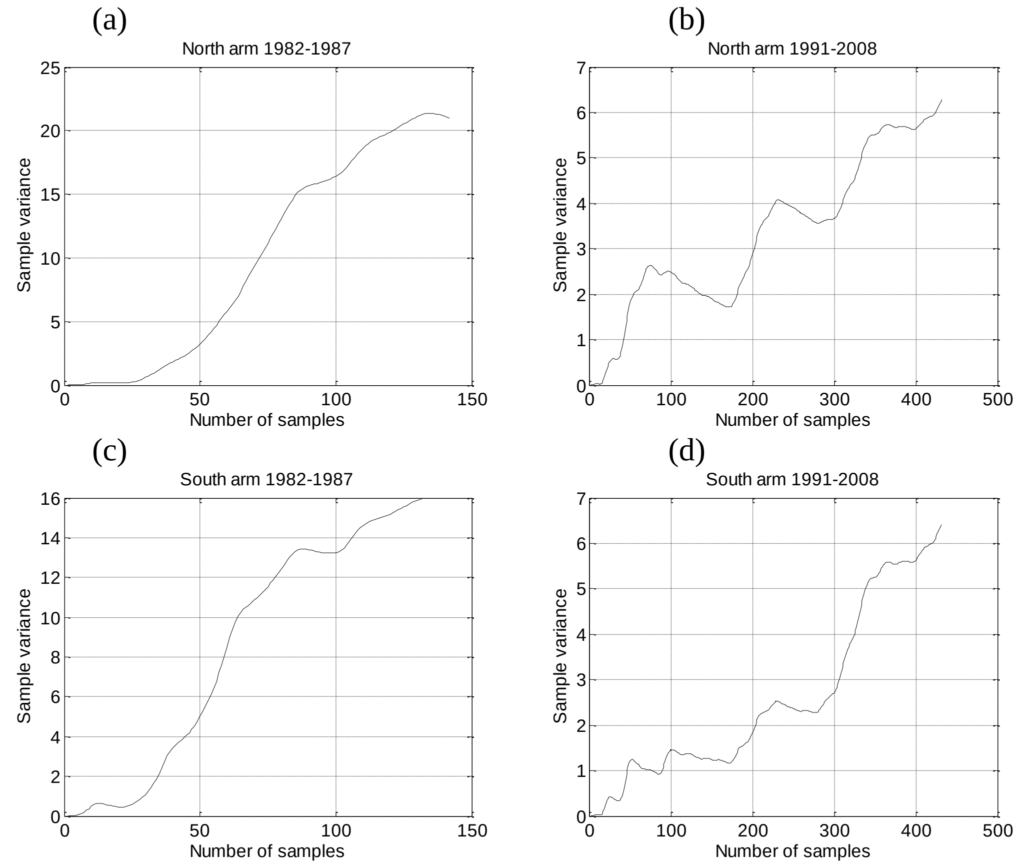
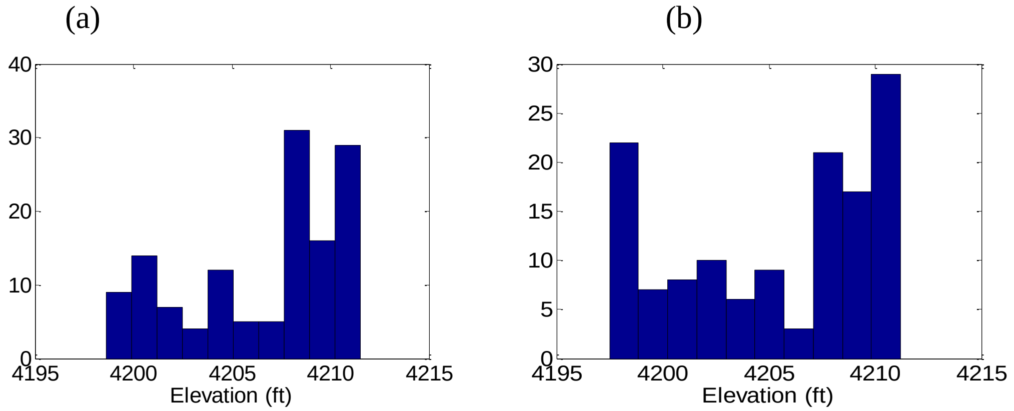


Figure A.4 Variance trend for GSL elevation time series, (a) North arm for 1982-1987, (b) North arm 1991-2008, (c) South arm 1982-1987, and (d) South arm 1991-2008.



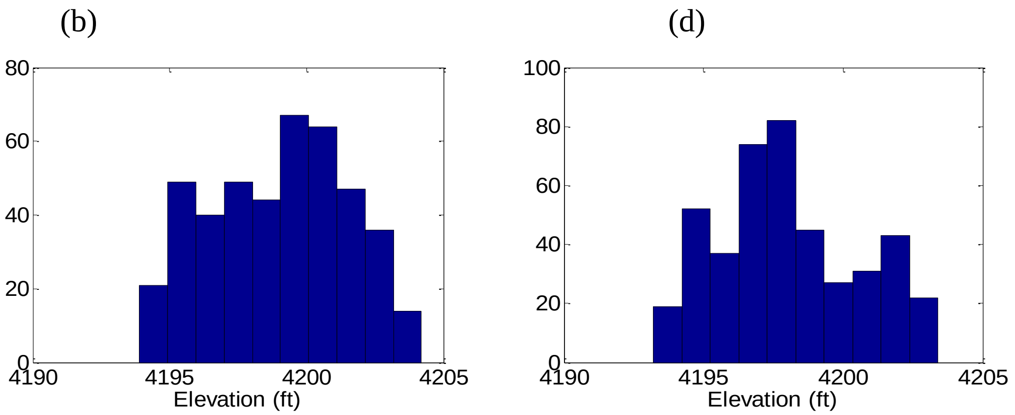


Figure A.5 Histogram of GSL elevation for 1982-1987: a) southern arm, b) northern arm, and for 1991-2008: c) southern arm, d) northern arm.

CURRICULUM VITAE

Niroj K Shrestha

EDUCATION

Ph.D.	Civil and Environmental Engineering Utah State University, Logan, Utah	2012
M.Sc.	Water Resource Engineering Katholieke University, Leuven/ Vrij University, Brussels, Belgium	2006
B.E.	Civil Engineering Tribhuvan University, Kathmandu, Nepal	2003

RESEARCH

Ph.D.	Title: Identification of Influential Climate Indicators and Prediction of Long-term Streamflow and Chaotic Great Salt Lake Elevation Using Machine Learning Approach. This dissertation applies the data driven modeling in water resource problems in basin scale. Machine learning models are used to identify influential climate indicators and predict seasonal and annual streamflow for multiple lead-times, and prediction of chaotic time series of lake elevation.
M.Sc.	Title: Evaluation of Calculation of New FAO Dynamic Crop Water Productivity Model 'AquaCrop'. This dissertation evaluates the new calculation procedure of the AquaCrop software with respect to other corresponding approaches for canopy development, soil evaporation, and crop transpiration component.
B.E. (Project)	Title: The Feasibility Study of Hugdi-Khola Hydropower Project, Gulmi. This project consist of site selection, study of hydrologic behavior of the stream, surveying to develop longitudinal profile of the site and topographical map for the intake and powerhouse, designing damn, conveyance canal, surge tank, penstock pipe, turbine and powerhouse.

RELATED EXPERIENCE

Graduate Research Assistant 2007-Present
Utah Water Research Laboratory Logan, Utah

- Developed machine learning based data driven models for prediction of hydrologic quantities: streamflow and lake elevation.
- Identified important circulation indicator and oscillation modes for streamflow prediction in Utah.
- Simulated stream runoff using GIS based WetSpa model.
- Used several hydrologic models for different projects: HEC-HMS, HEC-RAS, DAMBRK, SWMM, HEC-ResSim.
- Developed a model to design the volume of reservoir.

Graduate Student 2004-2006
Katholieke University and Vrij University Leuven and Brussels, Belgium

- Water quality modeling using WQNCAL model.
- Estimated the water demand of crops for irrigation purpose and preparing detail irrigation schedule for optimal irrigation.
- Used irrigation models for project works: BUDGE, ET0CALCULATOR, RAINBOW, UPFLOW.
- Used WETSPRO model for separating streamflow into different flow components and evaluate the performance of simulated models.
- Conducted Environmental Impact Assessment of gold mine project.
- Optimized hydrologic systems using linear as well as dynamic programming.

TEACHING EXPERIENCE

Instructor
Engineering State Program, challenge session, Utah State University 2011

- Water power: Designing turbine and testing efficiency. Logan, Utah

Instructor 2006-2007
Kantipur Engineering College Kathmandu, Nepal

- Taught fluid mechanics and hydrology.
- Graded homework and class projects.

Instructor 2003-2004
Advance College of Engineering and Management Kathmandu, Nepal

- Taught structural analysis, hydrology and survey to undergraduate students.
- Graded homework and class projects.
- Conducted survey camp for students of group 60.

PRESENTATIONS AND POSTER

Shrestha, N.K., Urroz, G. (Dec 2009). Prediction of Great Salt Lake Water Surface Elevation using Support Vector Machine. Poster presented in the AGU Fall Meeting. Hydrology Section, San Francisco, California.

Shrestha, N.K., Urroz, G., (April 2010). Prediction of Water Surface Level of Great Salt Lake Using Machine Learning Models. USU Spring Runoff Conference and Western Snow Conference, Utah State University, Logan, Utah.

Shrestha, N.K., McKee, M., (March 2011). Spatial Variability of Sea Surface Temperature Effect in Utah and Long-term Streamflow Forecasting Using Relevance Vector Machine. Annual Spring-Runoff Conference, Utah State University, Logan, Utah

COMPUTER SKILLS

- Water Resource Engineering Applications: HEC-HMS, HEC-RAS, DAMBRK, SWMM, HEC-ResSim, Wetspa.
- Irrigation Application: BUDGE, ET0CALCULATOR, RAINBOW, UPFLOW.
- Spatial Analysis Applications: ARCGIS, EDRAS.
- Programming Languages: Matlab, R, and ForTran.
- Design Software: AutoCad.
- Office Application: Microsoft Office.

AWARD AND HONORS

- Vlaamse Interuniversitaire Raad scholarhip, 2004-2006.
- Honored for best grade, Pulchowk Campus, 2000-2003.
- Merit student scholarship, Pulchowk Campus, 1999-2003.
- Fadindra memorial scholarship, 1999.

AFFILATIONS

- American Society of Civil Engineers, American Geophysical Union, Nepal Engineer’s Council