IMPROVING MEDIUM- AND LONG-RANGE HYDROLOGICAL FORECASTS

IMPROVING MEDIUM- AND LONG-RANGE HYDROLOGICAL FORECASTS WITH ENSEMBLE METEOROLOGICAL FORECASTS AND CLIMATIC INFORMATION

By

GETNET YAYEH MULUYE, B.Sc., M.Tech.

A Thesis

Submitted to the School of Graduate Studies

in Partial Fulfillment of the Requirements

for the Degree

Doctor of Philosophy

McMaster University

DOCTOR OF PHILOSOPHY (2010) McMaster University
(Civil Engineering) Hamilton, Ontario

TITLE: Improving Medium- and Long-Range Hydrological Forecasts with Ensemble Meteorological Forecasts and Climatic Information

AUTHOR: Getnet Yayeh Muluye
B.Sc. in Hydraulic Engineering
(Arba Minch University, formerly AWTI)
M.Tech. in Water Resources Development
(Indian Institute of Technology Roorkee)

SUPERVISORS: Dr. Brian Baetz and Dr. Sarah Dickson
NUMBER OF PAGES: xvi, 228

ABSTRACT

The ability to provide reliable and accurate medium- and long-range hydrological forecasts is fundamental for the effective operation and management of water resources systems. The principal objectives of this thesis are (i) to develop a framework for advancing the long-range forecasting skills of hydrological models by coupling pertinent and leading climate information with regional hydro-meteorological variables; and (ii) to develop effective mechanisms for integrating meteorological ensemble systems in a hydrologic prediction system, which would be useful for risk analysis by policy makers for operating both large-scale as well as small-scale water resources systems. This research constitutes three principal components: long-range forecasts, downscaling, and medium-range forecasts.

For long-range hydrological forecasting, four data-driven models, including multilayer perceptron (MLP), time-lagged feedforward network (TLFN), Bayesian neural network (BNN) and recurrent multilayer perceptron (RMLP) were designed by incorporating low-frequency climatic indices to forecast seasonal reservoir inflows. The results indicated that the incorporation of modes of climatic indices in a hydrologic forecasting model resulted in a considerable improvement in the seasonal forecast accuracy. Furthermore, the extended Kalman filter approach was used to train the recurrent multilayer perceptron for capturing the complexity associated with the long-range streamflow forecasting. Results showed that the proposed methodology was able to provide a robust modeling framework capable of capturing the complex dynamics of the hydrologic system.

Different statistical methods were developed and evaluated for downscaling local-scale information of precipitation and temperature from the numerical weather prediction model output. Three different methods were considered: (i) hybrids; (ii) neural networks; and (iii) nearest neighbor-based approaches. The findings revealed that the skills in the downscaled temperature forecasts were superior to those in the downscaled precipitation forecasts. In particular, for downscaling daily precipitation, the artificial neural network-

logistic regression (ANN-Logst), partial least squares (PLS) regression and recurrent multilayer perceptron trained with the extended Kalman filter (EKF) models yielded greater skill values, and the conditional resampling method (SDSM) and *K*-nearest neighbor (KNN) based models showed potential for characterizing the variability in daily precipitation.

For the case of medium-range hydrological forecasting, the downscaled and the raw numerical model outputs were forced into an HBV hydrologic model in order to generate an ensemble of reservoir inflows. The simulation results indicated that the downscaled-based flows had greater skill values, and yielded more accurate forecasts than the raw-based flows. The potential economic values of flow forecasts were further assessed based on a simple optimal decision-making, cost-loss analysis technique. The principal outcomes emerging from the analyses included: (i) the economic benefits associated with probabilistic flow forecasts were more useful than their deterministic counterparts; and (ii) the downscaled-based flow forecasts offered greater benefits, which are applicable to a much wider range of users, than the raw-based flow forecasts.

ACKNOWLEDGEMENTS

I would like to take this opportunity to thank all of those who helped me directly or indirectly throughout the years of my Ph.D. studies. The almighty God, I thank you for all.

My sincere gratitude goes first and foremost to my supervisors, Dr. Brian Baetz (Professor and Chair, Department of Civil Engineering and Director, Programs of Engineering & Society and Engineering & International Studies) and Dr. Sarah Dickson, who have supported me in the preparation of this thesis. Without their guidance, help and, in particular, keen interest in me, this thesis would have not been simply "possible."

I would like to extend my gratitude to my supervisory committee members, Dr. Yiping Guo and Dr. Sue Becker, for their time, valuable comments and constructive discussions. I also would like to thank Dr. Gary Purdy (Chair of Ph.D. Oral Exam) and Dr. Momcilo Markus (External Examiner, University of Illinois at Urbana Champaigne).

I am indebted to Graduate Studies at McMaster University, which has given me the opportunity with financial support to complete my PhD studies. In particular, I would like to thank Dr. Mark Hatton, former Acting Associate Dean of Graduate Studies for the Faculty of Engineering; Dr. Doug Welch, Associate Dean of Graduate Studies for the Faculty of Science; and Shelly Lancaster, McMaster Ombudsperson. Many thanks also go to members of all Faculty, staff and students of Civil Engineering Department, in particular to Carol Robinson (Administrative Secretary) and Tatiana Dobrovolska (Administrator).

I am also grateful to Dr. Ashok Mishra, Dr. Adam Gobena, Dr. Misgana Muleta and Mr. Sadik Ahmed for their valuable comments on the portion of this thesis. Many thanks go to Dr. Martha Dagnew, for her invaluable support, Dr. Yonas Dibike, for his useful advices, Dr. Sitotaw Yirdaw-Zeleke, for his encouragement, and Mr. Biruk Habtemariam, for his valuable time.

I would like to thank my Dad and Mam, Yayeh and Zewudie, my brother and sisters for their continued support and moral. Above all, I owe special thanks to my

beloved wife, Hibret, for her constant encouragement and unwavering support throughout the years of my Ph.D. studies, in both good and bad times. She has been part of the ordeal and was always with me during those unforgettable times.

PREFACE

This thesis is a sandwich-style thesis, and includes five chapters that have been published or submitted for publication. The titles of each chapter along with the corresponding author(s) and contributions are described below.

Chapter 2: Seasonal reservoir inflow forecasting with low-frequency climatic indices: a comparison of data-driven methods

Published in: Hydrological Science Journal

Authors: Getnet Muluye and Paulin Coulibaly

Contributions: Getnet Muluye performed the experiment, analyzed the results, and prepared the first draft. Dr. Paulin Coulibaly conceived the idea of using data-driven models, and contributed at various stages of the manuscript, including discussing the results, editing the manuscript and responding the reviewers' comments.

Chapter 3: Improving long-term streamflow prediction with extended Kalman filters

Submitted to: Hydrological Science Journal

Author: Getnet Muluye

Contributions: Getnet Muluye performed the experiment, analyzed the results and prepared the draft.

Chapter 4: Comparison of statistical methods for downscaling daily precipitation

Submitted to: Journal of Hydrology

Author: Getnet Muluye

Contributions: Getnet Muluye performed the experiment, analyzed the results and prepared the draft.

Chapter 5: Deriving hydrological variables from numerical weather prediction model output: a nearest neighbor approach

Submitted to: Water Resources Research

Author: Getnet Muluye

Contributions: Getnet Muluye performed the experiment, analyzed the results and prepared the draft.

Chapter 6

Implications of medium-range numerical weather model output in hydrologic application: assessment of skill and economic value

Submitted to: Journal of Hydrology

Author: Getnet Muluye

Contributions: Getnet Muluye performed the experiment, analyzed the results and prepared the draft.

Dr. Brian Baetz and Dr. Sarah Dickson offered guidance on the structure of the thesis, and provided editorial and review comments on all thesis chapters with the exception of Chapter 2.

TABLE OF CONTENTS

Abstract iii
Acknowledgements..... v
Preface..... vii
Table of Contents..... ix
List of Figures..... x
List of Tables xv

Chapter 1
General Introduction 1
Chapter 2
Seasonal reservoir inflow forecasting with low-frequency climatic indices: a comparison of data-driven methods..... 8
Chapter 3
Improving long-term streamflow prediction with extended Kalman filter..... 35
Chapter 4
Comparison of statistical methods for downscaling daily precipitation 69
Chapter 5
Deriving hydrological variables from numerical weather prediction model output: a nearest neighbor approach 111
Chapter 6
Implications of medium-range numerical weather model output in hydrologic application: assessment of skill and economic value..... 162
Chapter 7
General conclusions and recommendations..... 225

LIST OF FIGURES

CHAPTER 2

Fig. 1 Location of study basin in northeastern Canada ("black dot" denotes reservoir location within the Churchill Falls watershed). 30

Fig. 2 Time series plots of (a) reservoir inflows and (b) ENSO index series from January 1943 to December 2001. 31

Fig. 3 Focused time-lagged feed-forward network (TLFN). 32

Fig. 4 Partial autocorrelation function (PACF): (a) flow series; (b) ENSO series. 33

Fig. 5 Observed and four-month-ahead predicted inflows for: (a) the MLP and BNN model, (b) the TLFN and BNN model, and (c) the RMLP and BNN model (in each case for a portion of the testing data set from January 1988–December 1990). 34

CHAPTER 3

Figure 1. Distribution-based streamflow forecasting comparative diagrams. The left figures represent the MLP model output and the right figures represent the RMLP model output for a) 4-month ahead forecasts, 1993-2006, b) 8-month ahead forecasts, 1993-2006, and c) 12-month ahead forecasts, 1993-2006 66

Figure 2. Potential EVs of streamflow forecasts, in excess of the 70^{th} percentile, for (a) 4-month, (b) 8-month, and (c) 12-month ahead, 1993-2006. The solid and the dotted lines represent the EV curves associated with the RMLP and MLP, respectively. 67

Figure 3. Potential EVs of streamflow forecasts, in excess of the 85^{th} percentile, for (a) 4-month, (b) 8-month, and (c) 12-month ahead, 1993-2006. The solid and the dotted lines represent the EV curves associated with the RMLP and MLP, respectively. 68

CHAPTER 4

Figure 1. Location map of study area (Source: Dibike and Coulibaly, 2005). 102

Figure 2. Comparison of downscaled precipitation derived from numerical forecast model: (a) mean precipitation totals, and (b) variance of daily precipitation, 1997-2001. 103

Figure 3. Mean length of wet-spells derived from numerical forecast model for forecast range (FR) (a) 3, (b) 7, and (c) 10, 1997–2001. 104

Figure 4. Reduction of variance (RV) with forecast range (FR) as downscaled by various models, 1997-2001. 105

Figure 5. Quantile-quantile (q-q) plots of the quantiles of the observed precipitation (mm) against the quantiles of the simulated precipitation (mm) as downscaled by the different models, Jan 1997 to Dec 2001 for forecast range 7 for a) winter (JFM), b) spring (AMJ), c) summer (JAS), and d) fall (OND) 110

CHAPTER 5

Figure 1. Location map of study area (Source: Dibike and Coulibaly, 2005) 150

Figure 2. Comparison of downscaled precipitation derived from numerical weather prediction model output: (a) mean precipitation totals, and (b) variance of daily precipitation, 1997-2001. 151

Figure 3. Mean length of wet-spells derived from numerical weather prediction model output for forecast ranges (FR) of (a) 3, (b) 7, and (c) 10 days, 1997–2001. 152

Figure 4. Reduction of variance (RV) with forecast range (FR) as downscaled by the different models for (a) precipitation, (b) maximum temperature, and (c) minimum temperature, 1997-2001. 153

Figure 5. Quantile-quantile (q-q) plots of the quantiles of the observed precipitation (mm) against the quantiles of the simulated precipitation (mm) as downscaled by the different models, Jan 1997 to Dec 2001 for forecast range 7 for a) winter (JFM), b) spring (AMJ), c) summer (JAS), and d) fall (OND). 157

Figure 6. Comparison of the downscaled daily precipitation data derived from the numerical weather prediction model output versus the raw model output using reliability diagram, discrimination, and relative operating characteristics (ROC). The graph on the left represent the raw model output and the graph on the right represent the downscaled output for forecast ranges a) 0, b) 3, and c) 7, 1997-2001. 161

CHAPTER 6

Figure 1. Location map of study area (Source: Dibike and Coulibaly, 2005) 199

Figure 2. Schematic diagram illustrating ensemble streamflow prediction system. 200

Figure 3. Box plots of biases as a function of FRs in downscaling total daily precipitation. The Box plots were constructed from ensemble of downscaled outputs resulting from PLS, KNN and RAW, for the test period Jan 97 to Dec 01. The box-and-whisker symbols represent the 10^{th}, 25^{th}, 50^{th}, 75^{th} and 90^{th} percentiles of the forecasts. 201

Figure 4. Box plots of biases as a function of FRs in downscaling average daily temperature. The Box plots were constructed from ensemble of downscaled outputs resulting from PLS, KNN and RAW, for the test period Jan 97 to Dec 01. The box-and-whisker symbols represent the 10^{th}, 25^{th}, 50^{th}, 75^{th} and 90^{th} percentiles of the forecasts. 202

Figure 5. Box plots of RMSE as a function of FRs in downscaling total daily precipitation. The Box plots were constructed from ensemble of downscaled outputs resulting from PLS, KNN and RAW, for the test period Jan 97 to Dec 01. 203

Figure 6. Box plots of RMSE as a function of FRs in downscaling average daily temperature. The Box plots were constructed from ensemble of downscaled outputs resulting from PLS, KNN and RAW, for the test period Jan 97 to Dec 01. 204

Figure 7. Probabilistic diagnostic statistics: Brier skill score (BSS) and ranked probability skill score (RPSS) as a function of FRs in downscaling total daily precipitation (first column) and average daily temperature (second column). These statistics were constructed from ensemble of downscaled outputs resulting from PLS, KNN and RAW, for the test period Jan 97 to Dec 01. 205

Figure 8. Ensemble Prediction System for a FR of 3 days obtained by forcing KNN-based output into an HBV model, for the portion of test period Jan 97 to Dec 99. The light gray lines represent flows corresponding to each of 15 ensemble members, whereas the thick dark gray line and the dark line represent mean of ensemble flows and observed flows, respectively. 206

Figure 9. Flow hydrographs of (a) observed, simulated and KNN-based ensemble mean. The thick gray line represents the KNN-based ensemble mean, whereas the light gray line

and the dark line represent observed and simulated flows based on observed temperature and precipitation, respectively, for the test period Jan 97 to Dec 01; and (b) mean of ensemble flows for a FR of 3 days obtained by forcing PLS, KNN and RAW-based outputs into an HBV model, for the portion of the test period Jan 97 to Dec 99. 207

Figure 10. Deterministic diagnostic statistics as a function of FRs. The deterministic hydrologic forecasts were obtained when the HBV model was (i) forced with the mean of ensemble flow output (column 1), (ii) forced with the mean of downscaled output derived from predictors of individual members (column 2), and (iii) forced with the downscaled outputs derived from predictors of ensemble mean (column 3) for the test period from Jan 97 through Dec 01. The forcings were based on PLS, KNN and RAW outputs. 208

Figure 11. Box plots of biases as a function of FRs. The Box plots were constructed from ensemble of flows generated by forcing PLS, KNN and RAW-based outputs in an HBV model, for the test period Jan 97 to Dec 01. 209

Figure 12. Box plots of RMSE as a function of FRs. The Box plots were constructed from ensemble of flows generated by forcing PLS, KNN and RAW-based outputs in an HBV model, for the test period Jan 97 to Dec 01. 210

Figure 13. Probabilistic diagnostic statistics as a function of FRs for flow forecasts (a) Brier skill score (BSS), and (b) ranked probability skill score (RPSS). These statistics were constructed from ensemble of flows generated by forcing PLS, KNN and RAW outputs in an HBV model, for the test period Jan 97 to Dec 01. 211

Figure 14. Reliability diagram for FRs of 3 and 7 days. The ensemble flows used to construct the reliability diagram originated from the PLS-based output, for the test period from Jan 97 to Dec 01. The 45^{o} line represents perfectly reliable forecasts. 212

Figure 15. Reliability diagram for FRs of 3 and 7 days. The ensemble flows used to construct the reliability diagram originated from the KNN-based output, for the test period from Jan 97 to Dec 01. The 45^{o} line represents perfectly reliable forecasts. 213

Figure 16. Reliability diagram for FRs of 3 and 7 days. The ensemble flows used to construct the reliability diagram originated from the RAW-based output, for the test period from Jan 97 to Dec 01. The 45^{o} line represents perfectly reliable forecasts. 214

Figure 17. Discrimination diagram for a FR of 7 days. The ensemble flows used to construct the discrimination diagram originated from the PLS-based output, for the test period from Jan 97 to Dec 01. Discriminatory forecasts tend to separate among low-, medium- and high-flow portions of the observed time series, and avoid overlapping. 215
Figure 18. Discrimination diagram for a FR of 7 days. The ensemble flows used to construct the discrimination diagram originated from the KNN-based output, for the test period from Jan 97 to Dec 01. Discriminatory forecasts tend to separate among low-, medium- and high-flow portions of the observed time series, and avoid overlapping. 216
Figure 19. Discrimination diagram for a FR of 7 days. The ensemble flows used to construct the discrimination diagram originated from the RAW-based output, for the test period from Jan 97 to Dec 01. Discriminatory forecasts tend to separate among low-, medium- and high-flow portions of the observed time series, and avoid overlapping. 217
Figure 20. Potential EV of deterministic flow forecasts in excess of the 90^{th} percentile for a FR of 7 days, for the test period Jan 97 to Dec 01. Flow forecasts generated from the mean of downscaled members were used to construct the EV curves: PLS (solid line), KNN (pecked line), and RAW (dotted line). 218
Figure 21. Potential EV of deterministic flow forecasts in excess of the 90^{th} percentile for a FR of 7 days, for the test period Jan 97 to Dec 01. Flow forecasts generated from the mean of ensemble flows were used to construct the EV curves: PLS (solid line), KNN (pecked line), and RAW (dotted line). 219
Figure 22. Potential EV of deterministic flow forecasts in excess of the 90^{th} percentile for a FR of 7 days, for the test period Jan 97 to Dec 01. Flow forecasts generated from the downscaled mean predictor were used to construct the EV curves: PLS (solid line), KNN (pecked line), and RAW (dotted line). 220
Figure 23. Potential EVs of flow forecasts using KNN, in excess of the 90^{th} percentile of the observed flows for a FR of 7 days, for the test period Jan 97 to Dec 01. The thin curves represent EV for various probability thresholds *pt*, for each of the ensemble members, and the envelope of these curves (heavy solid line) represents the overall EV of

the probabilistic forecast system, obtained by picking the optimal value of *pt*, corresponding to each cost-loss ratio. 221

Figure 24. Potential EVs of the various flow forecast systems, in excess of the 90th percentile of the observed flows for a FR of 7 days, for the test period Jan 97 to Dec 01. Deterministic (DET) forecasts: PLS (dashed line), KNN (solid line), and RAW (pecked line); Ensemble Prediction System (EPS) probabilistic forecasts: PLS (heavy solid line), KNN (heavy dash-dotted line), RAW (heavy dotted line). 222

LIST OF TABLES

CHAPTER 2

Table 1 Basic descriptive statistics of variables used in modeling 26

Table 2 Optimal network architecture and parameters 27

Table 3 Model performance statistics of seasonal predictions: (a) using the MLP and BNN and (b) using TLFN and RMLP. 28

CHAPTER 3

Table 1. Traditional and distribution-based comparative model performance statistics 61

Table 2. Test for position (i.e. median) and variance using Mann-Whitney and Levene, respectively, for the test period Jan 1993 to Dec 2006. Note that the values in the table represent *p*-values. 62

CHAPTER 4

Table 1. Meteorological stations of Saguenay watershed 97

Table 2. NOAA reforecast ensemble variable fields 98

Table 3. Comparative performance statistics for downscaling daily precipitation. The bold numeric statistics represent the best statistics corresponding to each forecast range. 99

Table 4. Test for position (i.e. median) and variance for downscaling daily precipitation corresponding to the first member and zero forecast range using Mann-Whitney and Levene, respectively, for the test period Jan 97 to Dec 01. Note that the values in the table represent *p*-values. 101

CHAPTER 5

Table 1. Meteorological stations of Saguenay watershed 143
Table 2. NOAA reforecast ensemble variable fields 144
Table 3. Comparative performance statistics for downscaling daily precipitation 145
Table 4. Comparative performance statistics for downscaling daily maximum temperature 146
Table 5. Comparative performance statistics for downscaling daily minimum temperature 147
Table 6. Test for position (i.e. median) and variance for downscaling daily precipitation corresponding to the first member and zero forecast range using Mann-Whitney and Levene, respectively, for the test period Jan 97 to Dec 01. Note that the values in the table represent *p*-values. 148
Table 7. Comparative performance statistics using conventional and distribution-based diagnostic measures. The downscaled daily precipitation is compared to both the raw model output and climatology. The bold statistics represent improvements over the raw due to downscaling 149

CHAPTER 6

Table 1. Performance statistics in downscaling daily precipitation and temperature. Bold statistics represent the best performance corresponding to each forecast range (FR). 223
Table 2. Test for position (i.e. median) and variance for downscaling daily precipitation associated with the mean of ensemble predictors using Mann-Whitney and Levene, respectively, for the test period Jan 97 to Dec 01. Note that the values in the table represent *p*-values. The downscaled output derived from predictors of ensemble mean dataset was used to conduct statistical tests. 224

CHAPTER 1

General Introduction

There is a growing concern among the research community in providing accurate and reliable hydrological forecasts in space and time. The ability to provide timely and accurate short-range hydrological forecasts is crucial for effective early flood warning and defense systems. Medium-range hydrological forecasts up to two weeks are highly beneficial for optimal reservoir operation and scheduling. Likewise, long-range hydrological forecasts for more than one month are particularly important for long-term water resources planning and management. The physical system of the catchment that takes precipitation as an input and runoff as a response is a highly nonlinear and complex system, and it is difficult to account for all sources of uncertainty. Hydrologic predictions contain substantial uncertainty from different sources including uncertainties in initial conditions, model inputs, model structure, and model parameters, ultimately resulting in uncertainties in model states and outputs.

There have been diverse hydrologic models developed for the purpose of forecasting hydrological variables. In a broader perspective, hydrologic models can be classified into: conceptual models, physically-based models and data-driven models. In conceptual hydrologic models, the catchment is considered a single unit where the model parameters do not vary spatially within the basin. Physically-based hydrologic models consider the spatial and temporal variability of hydrologic processes, watershed inputs, boundary conditions and watershed characteristics. Data-driven hydrologic models, on the other hand, are based on transfer functions that relate the input variables to the output variables of the catchment. In some cases this includes the description, even if in a simplified way, of the basic processes involved in runoff formation and development, without the need for fully understanding the physical mechanisms governing the hydrologic processes (Singh, 1995).

Real-time and short- to medium-range hydrological variables are commonly related to numerous local factors as well as initial and boundary conditions of the river basin; whereas the long-range hydrological variables are generally associated with some

pertinent regional or global climatic states and atmospheric circulation patterns. The relationship between hydrological processes and climate fluctuations have been extensively studied over a wide range of geographical regions (e.g., Piechota et al., 1998; Whitaker et al., 2001; Tootle et al., 2007; Sveinsson et al., 2008). In particular, climatic variability, such as El Nino-Southern Oscillation (ENSO) has shown a profound world wide impact on the variability of streamflows and the distribution of surface and groundwater resources. New developments in climate forecasting can provide useful information on large-scale ocean-atmosphere circulation patterns. The use of information related to large-scale circulation patterns, therefore, offers opportunity for improving the predictability of long-range hydrological variables, which can ultimately benefit water managers in decision making. Though incorporating climate fluctuations in the hydrologic forecast system has proven to be useful, the challenge is to identify pertinent large-scale circulation patterns along with adequate hydrologic models that can effectively capture the complex dynamics of the large-scale ocean-atmosphere interaction.

There has been significant progress in the skill of short-term global-scale atmospheric forecasts over the past several years, and it is likely that output from these forecasts, when used as input for hydrologic models, may provide improved hydrological forecasts (Clark and Hay, 2004). The skills of the hydrologic forecast obtained when forcing a distributed-hydrologic model with the National Centers for Environmental Prediction (NCEP) Medium-Range Forecast Model (MRF) output have been successfully demonstrated (e.g., Clark and Hay, 2004; Werner et al., 2005). Nevertheless, the inability of large-scale weather forecasts to characterize local-scale hydrological variables has become increasingly recognized as a drawback in the analysis and modeling of hydrological processes. Direct application of current output from large-scale models is often found to be inadequate to resolve grid and sub-grid scale features of the watershed at which assessment of water resources is required (Murphy, 1999; Beck et al., 2004; Spak et al., 2007). Therefore, some form of downscaling is considered necessary to produce higher temporal and spatial resolutions than the current raw numerical weather

prediction model output, so as to meet the requirements of most hydrologic models (Diaz-Nieto and Wilby, 2005).

In order to ease the above challenges and integrate rapid advances made in the recent past in meteorology, an international Hydrological Ensemble Prediction Experiment (HEPEX) has been established with the intention of advancing reliable hydrological ensemble predictions that can be used with confidence by emergency management and water resources sectors to make decisions that have consequences for the economy, and public health and safety (Schaake et al., 2006). The key scientific issues for HEPEX include (Schaake et al., 2006): (i) adaptations required for coupling meteorological ensemble systems and hydrological ensemble systems; (ii) generation of hydrological ensembles that reflect the total uncertainty; (iii) use of climate information, such as low-frequency climatic indices in hydrological prediction systems; and (iv) best practice for analyzing and visualizing uncertainty.

This research is designed to address the key challenges of HEPEX by coupling meteorology and hydrology. The objectives of the research are to: (i) provide improved long-range hydrological forecasts through coupling pertinent climate information, such as low-frequency climatic indices with regional hydrological variables; (ii) develop and examine diverse downscaling techniques in order to produce skilful and reliable ensemble forcing for use in hydrological applications using ensemble atmospheric forecasts; and (iii) demonstrate a framework for using ensemble atmospheric forecasts in a hydrologic ensemble prediction system, so that the final product can be used with the appropriate degree of confidence for effective water resources planning and decision making.

This is a sandwich-style thesis, and includes five chapters that have been published, or submitted for publication. The published article and manuscripts have been included in this thesis with the permission of journal publishers.

Chapter 2: Seasonal reservoir inflow forecasting with low frequency climatic indices: a comparison of data-driven methods (published in Hydrological Science Journal)

Chapter 3: Improving long-term streamflow prediction with extended Kalman filters (submitted to Hydrological Science Journal)

Chapter 4: Comparison of statistical methods for downscaling daily precipitation (submitted to Journal of Hydrology)

Chapter 5: Deriving hydrological variables from numerical weather prediction model output: a nearest neighbor approach (submitted to Water Resources Research)

Chapter 6: Implications of medium-range numerical weather model output in hydrologic application: assessment of skill and economic value (submitted to Journal of Hydrology)

The performance of a hydrological forecast model varies from basin to basin. This variation is primarily caused by the difference in sources of uncertainty in hydrological forecast systems (Clark and Hay, 2004; Wang et. al., 2009; Gobena and Gan, 2010). For this reason, comparison of model results from different basins on a similar or on a different model is not established. Instead, first, a state-of-the-art model is identified based on literature review and preliminary analysis. Second, the identified model is used as a benchmark and applied to the study basin. Third, model results from the benchmark model are compared against model results from a model which has either been modified or applied for the first time in the present study.

Improving hydrologic prediction is the focus of this thesis, which is presented in three sections: long-range hydrologic prediction, downscaling, and medium-range hydrologic prediction.

Long-range hydrologic prediction is covered in Chapters 2 and 3, which describe the development and evaluation of hydrologic models to forecast long-range hydrological variables. In Chapter 2, four data-driven hydrologic models, namely multilayer perceptron (MLP), time-lagged feedforward network (TLFN), Bayesian neural network (BNN) and recurrent multilayer perceptron (RMLP) are investigated to forecast seasonal reservoir inflows of the Churchill Falls watershed in northeastern Canada. El Nino-Southern Oscillation (ENSO) is incorporated as additional information to the historical reservoir inflows in order to improve seasonal forecasts. The relative successes of each

model in capturing the dynamics of ENSO to achieve improved seasonal forecasts are discussed. In Chapter 3, the potential for using an extended Kalman filter approach to perform supervised training of a recurrent multilayer perceptron is investigated. The method is applied to characterize long-term horizon streamflows in western Alberta (Canada). The performance of the proposed model is compared against the conventional multilayer perceptron through suites of traditional and distribution-based diagnostic measures.

Downscaling is treated in Chapters 4 and 5, which focus on generating local-scale hydrological variables from the numerical weather prediction model output. The general framework is demonstrated on the Chute-du-Diable sub-basin located in northeastern Canada. In Chapter 4, different statistical methods are developed and evaluated focused on downscaling local-scale information of precipitation from the numerical weather prediction model output. Three different methods are considered: (i) hybrids; (ii) neural networks; and (iii) nearest neighbor-based approaches. The performances of each model are compared and discussed in terms of reproducing daily precipitation statistics and skill value. In Chapter 5, variations of nearest neighbor resampler algorithms are investigated for downscaling station daily precipitation, and minimum and maximum temperature fields. On the basis of a measure of closeness and a sampling strategy, six nearest neighbor-based models are designed. The relative performances of each model are examined and suitable downscaling models are suggested for the study basin. Furthermore, the skills of the downscaled precipitation and the raw numerical model outputs are assessed through suites of conventional and distributed-based diagnostic measures.

Medium-range hydrologic prediction is presented in Chapter 6, which demonstrates a framework for using ensemble atmospheric forecasts in a hydrologic prediction system. In this chapter, downscaled ensemble temperature and precipitation fields are forced into an HBV hydrologic model, in order to produce an ensemble of streamflow hydrographs. The ultimate goal of ensemble forecasts is to generate probabilistic streamflow values that can be used with confidence for effective water

resources planning and decision making processes. The hydrologic ensemble prediction system is compared against the deterministic counterpart to confirm the significance of the approach.

Finally, Chapter 8 presents general conclusions of the research and a few remarks on hydrologic prediction.

Given that this thesis is a sandwich-style composed of chapters representing published and submitted journal articles, some material which exists in some chapters, such as model description and performance measures, is partially reproduced and appears in some other chapters in order to make each chapter complete and stand-alone.

References

Beck, A., Ahrens, B., and Stadlbacher, K. (2004) Impact of nesting strategies in dynamical downscaling of reanalysis data. Geophys. Res. Lett., 31 (19), L19101, doi: 10.1029/ 2004GL020115.

Clark, M.P., and Hay, L.E. (2004) Use of medium-range numerical weather prediction model output to produce forecasts of streamflow. J. Hydrometeorol., 5(1), 15–32.

Diaz-Nieto, J., and Wilby, R.L. (2005) A comparison of statistical downscaling and climate change factor methods: impacts on lowflows in the River thames, united kingdom. Climatic change, 69, 245–268.

Gobena, A.K. and Gan, T.Y. (2010) Incorporation of seasonal climate forecasts in the ensemble streamflow prediction system. J. Hydrol., 385 (1-4), 336-352.

Murphy, J.M. (1999) An evaluation of statistical and dynamical techniques for downscaling local climate J. Clim., 12, 2256– 2284.

Piechota, T.C., Chiew, F.H. S., and Dracup, J.A. (1998) Seasonal streamflow forecasting in eastern Australia and the El Nino-Southern Oscillation. Water Resour. Res., 34, 3035–3044.

Schaake, J., Franz, K., Bradley, V., and Buizza, R. (2006) The Hydrologic Ensemble Prediction EXperiment (HEPEX). Hydrol. Earth Syst. Sci., 3, 3321–3332.

Singh, V.P. (1995) Computer Models of Watershed Hydrology. Water Resources Publications, Highlands Ranch, Co.

Spak, S., Holloway, T., Lynn, B., and Goldberg, R. (2007) A comparison of statistical and dynamical downscaling for surface temperature in North America. J. Geophys. Res., 112, D08101, doi:10.1029/2005JD006712.

Sveinsson, O.G.B., Lall, U., Fortin, V., Perrault, L., Gaudet, J., Zebiak, S., and Kushnir, Y. (2008) Forecasting spring reservoir inflows in Churchill Falls Basin in Quebec, Canada. J. Hydrol. Eng., 13(6), 426-437.

Tootle, G.A, Singh, A.K., Piechota, T.C., and Farnham, I. (2007) Long Lead-Time Forecasting of U.S. Streamflow Using Partial Least Squares Regression. J. Hydrol. Eng., 12(5), 442-451.

Wang, W.C., Chau, K.W., Cheng, C.T., and Qiu, L. (2009) A comparison of performance of several artificial intelligence methods for forecasting monthly discharge time series. J. Hydrol., 374, 294–306.

Werner, K., Brandon, D., Clark, M., and Gangopadhyay, S. (2005) Incorporating Medium-Range Numerical Weather Model Output into the Ensemble Streamflow Prediction System of the National Weather Service. J. Hydrometeorol., 6(2), 101-114.

Whitaker, D.W., Wasimi, S.A., and Islam, S. (2001) The El Nino-Southern Oscillation and long-range forecasting of flows in the Ganges. Int. J. Climatol., 21, 77–87.

CHAPTER 2

Seasonal reservoir inflow forecasting with low-frequency climatic indices: a comparison of data-driven methods

GETNET Y. MULUYE & PAULIN COULIBALY

International Association of Hydrological Sciences
Hydrological Science Journal
Special issue: Hydroinformatics
52(3), 508-522, 2007

CHAPTER 2

Seasonal reservoir inflow forecasting with low-frequency climatic indices: a comparison of data-driven methods

This chapter investigates the potential of using data-driven models for forecasting seasonal reservoir inflows using a climate variability indicator, called the El Nino-Southern Oscillation (ENSO), as additional information to the historical reservoir inflows. The relative successes of each model in capturing the dynamics of ENSO to achieve improved seasonal forecasts are evaluated and discussed. The advantages and limitations of the different data-driven methods used in the study are also discussed. This chapter contains a paper entitled "Seasonal reservoir inflow forecasting with low-frequency climatic indices: a comparison of data-driven methods" published in the Hydrological Science Journal in 2007. This forms the first task on long-range hydrologic forecasting.

Seasonal reservoir inflow forecasting with low-frequency climatic indices: a comparison of data-driven methods

Abstract This paper investigates the potential of using data-driven methods, namely Bayesian neural networks (BNN), recurrent multi-layer perceptrons (RMLP), time-lagged feed-forward networks (TLFN), and conventional multi-layer perceptrons (MLP) to forecast seasonal reservoir inflows of the Churchill Falls watershed in northeastern Canada. A climate variability indicator (the El Niño-Southern Oscillation, ENSO) is used as additional information to historical inflow time series in order to predict seasonal reservoir inflows. The prediction results showed that the Bayesian neural network model was best able to capture the additional information provided by the ENSO series, and provided improved predictions in spring and summer seasons relative to the same model using only reservoir inflows. Similarly, time-lagged feed-forward networks and recurrent multi-layer perceptrons showed some improved forecast skill in spring when the ENSO index series are used but generally provided superior performance overall. The conventional multi-layer perceptron appears unable to capture relevant information from the ENSO series regardless of the season. However, when only historical flow series are used, all the selected data-driven methods provide very competitive forecast performances.

Key words data-driven methods; Bayesian neural networks; recurrent networks; time-lagged networks; Churchill Falls basin; ENSO; seasonal forecast; reservoir inflows

INTRODUCTION

The importance and validity of incorporating low frequency climatic indices (e.g. El Niño-Southern Oscillation, ENSO) to improve predictions of medium to long-range hydrological variables, and specifically streamflows, are well documented. Examples of studies include, to name a few, the Nile River (Eltahir, 1996; Wang & Eltahir, 1999), the Ganges River (Whitaker *et al.*, 2001), the River Murray (Simpson *et. al.*, 1993), other

rivers across the southern USA and Australia (Piechota *et al.*, 1998; Piechota & Dracup, 1999, and more recently Tootle *et al.*, 2005). In the study area of concern, similar studies were reported for flow forecasting using climatic indices (Coulibaly *et al.*, 2000a,b; Sveinsson, 2003a,b). Sveinsson (2003a,b) conducted an extensive study by incorporating hydroclimatic information and measures of atmospheric circulation to make seasonal forecasts of streamflow for 15 basins in the Québec-Labrador region. He showed that the use of variables related to low-frequency climatic variability can provide improved forecasts for most lead times and can be an effective alternative to traditional forecasting methods such as the use of basin hydro-meteorological information (e.g. streamflow, precipitation, temperature). Tootle *et al.* (2005) investigated the coupled response of large-scale ocean–atmosphere phenomena with ENSO to evaluate the influence of hydrological variability in regions of continental USA.

To extract relevant information from low-frequency climatic indices for improved long-range forecasts of hydrological variables, different modelling approaches have been investigated. Linear and multiple regression models based on canonical correlation analysis (CCA) have been investigated for long-term rainfall forecasting using ENSO indicators (Barnston *et al.*, 1996; Shabbar & Barnston, 1996). Probabilistic models also have been applied to forecast streamflow using ENSO indices (McKerchar *et al.*, 1996; Piechota *et al.*, 1998). Similarly, Sveinsson (2003a,b) used autoregressive moving average (ARMA) and autoregressive with exogenous input (ARX) models for forecasting aggregated spring (May–July) streamflows in northeastern Canada. The relationship between ENSO and regional hydrological regimes are typically modeled by a multiple regression framework, or some variant thereof. However, the traditional multivariate linear regression models have been found inappropriate for forecasting streamflow using ENSO indices in some studies (Piechota *et al.*, 1998). To incorporate the nonlinear dynamics of climatic mode oscillations in long-term regional streamflow prediction, hydrologists may resort to robust nonlinear data mining approaches. In recent years, artificial neural networks have become popular in hydrological systems modelling (e.g. Coulibaly *et al.*, 2000a; Giustolisi & Laucelli, 2005). The suitability of artificial neural

networks in rainfall–runoff modelling in particular, and in hydrology at large, has been extensively reviewed first by Coulibaly *et al.* (1999), and then by the ASCE Task Committee on the Application of Artificial Neural Networks in Hydrology (ASCE, 2000a,b) and Dawson & Wilby (2001). The main conclusions of these studies are that artificial neural networks can be considered as a robust modelling alternative to conventional hydrological models. Thus, investigating variations of neural networks for long-term reservoir inflow forecasting with low-frequency climatic indices is appropriate.

The main objective of this study is to investigate the potential of different data driven methods to forecast seasonal reservoir inflows using a climate variability indicator (ENSO index). This includes a Bayesian neural network (BNN), the recurrent multi-layer perceptron (RMLP), the time-lagged feed-forward network (TLFN), and the conventional multi-layer perceptron (MLP). A recent study has shown that a Bayesian neural network approach can be an effective basis for rainfall–runoff modelling (Khan & Coulibaly, 2006). Other selected models have been successfully used in hydrological modelling and forecasting (see ASCE, 2000a,b). The primary purpose of this study is to examine the ability of the selected data-driven methods to capture relationships between indices of climate variability and seasonal reservoir inflows in northeastern Canada. The remainder of the paper is organized as follows. The study area and the hydro-climatic data used are presented first. Secondly, each data-driven method is presented. Third, results from the forecasting experiments are reported. Finally, some conclusions are drawn.

DATA AND STUDY AREA

Study area

The study area, the Churchill Falls watershed, is located in northeastern Canada (Fig. 1), at approximately 52°N to 55°N and 62°W to 67°W. This watershed contains a large hydropower station and reservoir managed by Hydro-Quebec. Time series of monthly inflows (1943 to 2001) to this reservoir are available for the present study. These time series records have been analysed independently and provided by the Hydro-Quebec Prediction Department that routinely records and maintains hydrological data and related

information. This database is updated periodically and assumed to be adequate for reliable statistical predictions. The Churchill Falls watershed contributes more than 20% of the annual energy inflow of northeastern Canada (Coulibaly *et al.*, 2000a).

The major dataset used to develop the relationships between low-frequency climatic mode indices and the Churchill Fall reservoir inflows are time series records of reservoir inflows for the Churchill Falls watershed, and the time series of low-frequency climatic indices from the period January 1943 to December 2001 (Fig. 2(a) and (b)). Among the several climatic patterns that can influence hydrological variables in the Northern Hemisphere, such as the North Atlantic Oscillation (NAO), Arctic Oscillation (AO), El Niño-Southern Oscillation (ENSO) and Pacific/North American (PNA), only the ENSO is selected in the present study based on preliminary analysis and previous studies (Coulibaly & Burn, 2004, 2005). It has been shown that the variability of seasonal streamflows in eastern Canada is dominated by the ENSO pattern, especially in spring-summer and winter seasons (Coulibaly & Burn, 2005). The ENSO index used in this study is the monthly mean sea surface temperature (SST) anomalies over the Niño-3 region (5^0N-5^0S; 90^0W-15^0W) (Rasmusson & Carpenter, 1982), and monthly index data were obtained from the National Weather Service (NWS) Climate Prediction Center (CPC) (http://www.cpc.ncep.noaa.gov/data/ indices/).

METHODOLOGY

Bayesian neural network

Despite the rapid growth and use of artificial neural network based models, the conventional neural network training approach suffers from various limitations (Coulibaly *et al.*, 2001a; Khan & Coulibaly, 2006). One of the main limitations is that the network is trained by maximizing a likelihood function of the parameters, or equivalently, minimizing the error function in order to obtain the best set of parameters starting with an initial random set of parameters. Sometimes a regularization term with an error function is used to prevent over-fitting. In the conventional method, a complex model can fit the training data well but it does not necessarily mean that it will provide

smaller errors with respect to new data. This happens when uncertainty about the model parameters or uncertainty about the relationship between input and output mapped by the network during training is not taken into account. The Bayesian approach overcomes such difficulties as the uncertainty about the relationship between input–output is represented by a probability density function of the parameters. Before observing or collecting the data, the parameters are described by a prior probability density function, which is typically very broad to reflect the fact that we have little idea ("vague belief") of what values the parameters should take. Once the data are observed or collected, the corresponding *posterior* probability density function can be derived using Bayes' theorem. The Bayesian neural network (BNN) has been used in various sectors both for regression and classification problems. Neal (1996) used a BNN approach to study the effects of air pollution on housing prices in Boston; Lampinen & Vehtari (2001) demonstrated its application in three areas such as in predicting the quality of concrete in the concrete manufacturing process, in recognizing tree trunks in forest scenes, and in a topographic image reconstruction. The results of the study showed that the BNN model performed better than the standard neural network methods and other statistical models. More recently, the effectiveness of the BNN model in rainfall–runoff modelling has been investigated by Khan & Coulibaly (2006). Its performance was found to be competitive when compared to a widely-used conceptual hydrological model (integrated hydrological modelling system, IHMS HBV-96) and superior to a conventional or standard multi-layer perceptron. Thus, it appears appropriate to investigate the potential of the BNN approach for long-term (seasonal) reservoir inflow forecasting with low frequency climatic indices.

In Bayesian learning approach, the process starts with a suitable prior distribution, $p(\mathbf{w})$, for the network parameters (weights and biases). Once the data, D, are observed, Bayes' theorem is used for writing an expression of the posterior probability distribution for the weights, $p(\mathbf{w} \mid D)$, as:

$$p(\mathbf{w} \mid D) = \frac{p(D \mid \mathbf{w}) p(\mathbf{w})}{p(D)} \tag{1}$$

where, $p(D|\mathbf{w})$ is the dataset likelihood function, and the denominator, $p(D)$, is a normalization factor, which can be obtained by integrating over the weight space as follows:

$$p(D) = \int p(D|\mathbf{w})p(\mathbf{w})d\mathbf{w} \tag{2}$$

This ensures that the left-hand side of (1) gives unity when integrated over the entire weight space. Once the posterior has been calculated, every type of inference is made by integrating over that distribution. Therefore, in implementing Bayesian method, expressions for the prior distribution, $p(\mathbf{w})$ and the likelihood function, $p(D|\mathbf{w})$ are needed. The prior distribution, $p(\mathbf{w})$, which is not related with data, can be expressed in terms of weight-decay regularizer, $E_W = \frac{1}{2}\sum_{i=1}^{W} w_i^2$, where, W is the total number of weights and biases in the network. Similarly, the likelihood function in Bayes' theorem (1), which is dependent on data, can be expressed in terms of error function, $E_D = \frac{1}{2}\sum_{n=1}^{N}\left(y^n(\mathbf{x}^n;\mathbf{w}) - t^n\right)^2$, where, N is the total number of examples (patterns) in the training set, n is the nth example, $\mathbf{x}$ is the input vector, t is the target value and $y(\mathbf{x};\mathbf{w})$ is the network output. Upon deriving the expressions for the prior and likelihood functions, and using those expressions in (1), the posterior distribution of weights can be obtained. The objective function in the Bayesian method corresponds to the inference of the posterior distribution of the network parameters. After defining the posterior distribution (objective function), the network is trained with a suitable optimization algorithm to maximize the posterior distribution, $p(\mathbf{w}|D)$. Thus the most probable value for the weight vector $\mathbf{w}_{MP}$ corresponds to the maximum of the posterior probability. Using the rules of conditional probability, the distribution of outputs, for a given input vector, $\mathbf{x}$, can be written in the form:

$$p(t|\mathbf{x},D) = \int p(t|\mathbf{x},\mathbf{w})p(\mathbf{w}|D)d\mathbf{w} \tag{3}$$

where $p(t|\mathbf{x},\mathbf{w})$ is simply the model for the distribution of noise on the target data for a fixed value of the weight vector $\mathbf{w}_{MP}$, and $p(\mathbf{w}|D)$ is the posterior distribution of the

weights. The posterior distribution over network weights provides a distribution about the outputs of the network. If a single-valued prediction is needed, the mean of the distribution is used, and while the uncertainty about the prediction is needed, the full predictive distribution is used to represent the range of uncertainty about the prediction. For a more detailed description of the BNN approach, the readers are referred to Khan & Coulibaly (2006).

Multilayer perceptron

The multilayer perceptron (MLP) is one of the most widely applied neural networks (ASCE Task Committee, 2000a,b). These networks take in a set of inputs, x_i, and from them compute one or more output values, $f_k(x)$. For a one hidden layer MLP with a linear output neuron the equations are:

$$f_k(x) = b_k + \sum_j w_{jk} h_j(x) \tag{4}$$

$$h_j(x) = \varphi(a_j + \sum_i w_{ij} x_i) \tag{5}$$

where $\varphi(.)$ is the activation function of neuron j, w_{ij} is the weight on the connection from input unit i to hidden unit j; similarly, w_{jk} is the weight on the connection from hidden unit j to output unit k. The a_j and b_k are the biases of the hidden and output units, respectively. These weights and biases are the parameters of the network. Each output value, $f_k(x)$, is just a weighted sum of hidden unit values plus a bias. Each hidden unit computes a similar weighted sum of input values and then passes it through a nonlinear activation function. The activation function can be sigmoid or hyperbolic tangent function. The weights and biases in a MLP network are learned based on a set of training cases, $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$, giving examples of inputs, $x^{(i)}$, and associated targets, $y^{(i)}$. Standard neural network training procedures adjust the weights and biases in the network so as to minimize a measure of "error" in the training cases, most commonly the sum of the squared differences between the network outputs and the targets. Finding the weights and biases that minimize the chosen error function is commonly done using some

gradient-based optimization method, using derivatives of the error with respect to the weights and biases calculated by back-propagation. The detailed theory and derivation of the back-propagation algorithm can be found in Haykin (1999).

Time-lagged feed-forward networks (TLFN)

A time-lagged feed-forward neural network (TLFN) is a neural network that can be formulated by replacing the neurons in the input layer of a MLP with a memory structure, which is sometimes called a tap delay line. The size of the memory layer (the tap delay) depends on the number of past samples that are needed to describe the input characteristics in time and must be determined on a case-by-case basis. A TLFN uses delay-line processing elements which implement memory by simply holding past samples of the input signal as shown in Fig. 3. Given an input signal consisting of the present value $x(n)$ and the p past values $x(n\text{-}1),\ldots, x(n\text{-}p)$ stored in a delay line memory of order p, the free parameters of the network are adjusted to minimize the mean-squared error between the output of the network and the desired (or target) response. At time n, the "temporal pattern" applied to the input layer of the network is the signal vector:

$$x(n)=\left[x(n),x(n-1),...,x(n-p)\right]^T \tag{6}$$

which may be viewed as a description of the state of the nonlinear filter at time n. The output of the nonlinear filter, assuming a single layer as shown in Fig. 3, can be given by:

$$y(n)=\sum_{j=1}^{m_1} w_j y_j(n) =\sum_{j=1}^{m_1} w_j \varphi\left(\sum_{l=0}^{p} w_j(l)x(n-l)+b_j\right)+b_0 \tag{7}$$

where $\varphi(\cdot)$ is the activation function of neuron j, the $w_j(l)$ are its synaptic weights. The output neuron in the TLFN is assumed to be linear. The synaptic weights of the output neuron are denoted by the set $\{w_j\}_{j=1}^{m_1}$, where m_1 is the size of the hidden layer, and the bias is denoted by b_o.

An interesting feature of the TLFN is that the tap delay line of the inputs does not have any free parameters; therefore, the network can still be trained with the classic back-propagation algorithm. The TLFN topology has been successfully used in many

hydrological applications (e.g. Dibike *et al.*, 1999; Coulibaly *et al.*, 2001a). A major advantage of TLFN is that it is less complex than the conventional time delay and recurrent networks and has similar temporal pattern processing capabilities.

Recurrent multi-layer perceptron (RMLP)

The recurrent multi-layer perceptron (RMLP) can be considered as a feed-forward network augmented by recurrent (or feedback) connections. Generally, a RMLP can have one or more recurrent layers, and some layers can receive recurrent inputs from themselves and/or from other layers. Here a simple RMLP with a single hidden layer is used. Let $\mathbf{x}_I(n)$ denote the output of the hidden layer, $\mathbf{x}_o(n)$ be the output of the output layer and $\mathbf{u}(n)$ denote the input vector. The operational principles of the RMLP can then be mathematically expressed by the following equations (Haykin, 1999):

$$\mathbf{x}_I(n+1) = \varphi_I\left(\mathbf{w}_I \cdot \begin{bmatrix} \mathbf{x}_I(n) \\ \mathbf{u}(n) \end{bmatrix}\right) \tag{8}$$

$$\mathbf{x}_o(n+1) = \varphi_o\left(\mathbf{w}_o \cdot \begin{bmatrix} \mathbf{x}_o(n) \\ \mathbf{x}_I(n+1) \end{bmatrix}\right) \tag{9}$$

where $\varphi_I(.,.)$ and $\varphi_o(.,.)$ are the activation functions of the hidden layer and output layer, respectively; and $\mathbf{w}_I$ and $\mathbf{w}_o$ denote the weight matrices of the hidden layer and output layer, respectively. The RMLP can be trained with the backpropatation through time (BPTT) or real-time recurrent learning (RTRL) algorithms. The RMLP used in this study was trained with a BPTT algorithm (Principe *et al.*, 2000).

NEURAL NETWORK DESIGN FOR PREDICTION

Selection of predictors

The reservoir inflow and the ENSO series were analysed using a partial autocorrelation function (PACF) to identify significant time lags. The PACF plots (Fig. 4) show significant lags (i.e. with values above 0.2). Thus, time lags (t-1), (t-2), (t-9), (t-11), (t-12) and (t-24) for flow series, and (t-1) and (t-2) for ENSO series are identified as model

inputs. Once the significant predictors are identified, all the selected neural networks are investigated to predict seasonal reservoir inflows for four months ahead (April), eight months ahead (August) and twelve months ahead (December) for two cases of model inputs: (a) using only flow series, and (b) using both flow and ENSO series.

Model identification

In this study, the BNN approach proposed by Khan & Coulibaly (2006) is used. For modeling with the MLP, RMLP and TLFN, the NeuroSolutions software (NeuroDimension, Inc., 2002, Gainesville, Florida) is used. From the total data available (i.e. January 1943–December 2001), a subset of data from January 1943 to December 1984 was used for model calibration, and a subset from January 1985 to December 2001 was used for testing the performance of the models. All models are investigated based on the same training and testing period. The basic descriptive statistics of the variables (flows an ENSO) used in the modelling are provided in Table 1.

In the search for an optimal network architecture, each model was trained using the same input variables and selected time lags as specified earlier. The best model parameters were noted based on model performance statistics (such as mean squared error (MSE), coefficient of correlation (r), minimum and maximum absolute error). The optimized network is then used to predict seasonal reservoir inflows for four months ahead (April), eight months ahead (August) and twelve months ahead (December) for both cases (i.e. flow only, and flow plus ENSO). The optimal network architecture and the parameters used for each model are presented in Table 2. For all the models, one hidden layer, a tangent hyperbolic function (for hidden neurons), and a linear function (for output neurons) are used. Only the number of hidden neurons identified for each model is different (see Table 2).

The search for an optimal network involved using a simple network having one hidden layer and varying the hidden neurons between 2 to 35. Combinations of different learning rules (such as Delta-Bar-Delta, conjugate gradient and momentum), as well as transfer functions (i.e. tangent hyperbolic, sigmoid and linear) in both the hidden layer

and the output layer, were also investigated in the search for an optimal network. Experimentation showed that the optimal network contained a hyperbolic tangent and linear activation function in the hidden and output layers, respectively, and a learning rule with momentum.

RESULTS

For water resource planners and managers who deal with the total amount of water received by the reservoir in a specific season or period, the volume under the hydrograph is of major concern. Therefore, appropriate model performance criteria in this case are hydrograph plots of observed and predicted inflows along with some common performance statistics such as the mean squared error (MSE), the normalized mean squared error (NMSE = MSE/variance of target), the minimum and maximum absolute error and the linear correlation coefficient (r).

Comparative model performance statistics for the test period (January 1985–December 2001) are presented in Table 3(a) and (b). These tables show that all data-driven models provide adequate performance and are very competitive to each other for four-month-ahead predictions using only historical flow. For eight-month-ahead predictions, the BNN model performed slightly better than the MLP model but was inferior to the RMLP and TLFN in some of the test statistics such as NMSE, MAE and r. However, the BNN model was competitive with respect to the RMLP and TLFN for twelve-month-ahead predictions and the MLP only lags behind slightly in terms of performance. It is interesting to see that the maximum absolute error statistic for the test data set for the BNN model was lower than for the other models. This indicates that the magnitude of the under- or over-estimation of the peak flow by the BNN model may be lower than that of the other models.

The performance of the data-driven models that used both the flow and the ENSO series is also presented in Table 3(a) and (b). The skill scores (1 – MSE_(Flow and ENSO)/MSE_Flow) are computed with respect to case one (using flows only) to help in model comparison. A positive skill score indicates an improvement of model prediction

results due to the use of ENSO while a negative skill score indicates a deterioration of model performance with the use of ENSO. The prediction results suggest that the BNN model is able to capture relevant information using the ENSO data because the skill score is consistently positive for all lead times, and provides improved predictions in spring (i.e. four month-ahead forecast) and summer seasons (i.e. eight month-ahead forecast), the TLFN and RMLP show improvement only in spring season. On the other hand, the MLP model was unable to capture relevant information from the ENSO series for all seasons and yields poor (or deteriorating) results with the addition of this extra input. The BNN forecast improvements shown in the spring and summer are consistent with the strong ENSO streamflow patterns described by Coulibaly & Burn (2005), especially in the spring-summer in eastern Canada. Such results may suggest that the BNN model has the potential to capture relevant information from the climatic oscillation indices to provide improved seasonal forecasts. However, further investigation is needed to assess this potential. The performance of the MLP with respect to the RMLP and TLFN was inferior, as expected, especially for long-term predictions. This is because the latter models can handle time dynamics through their memory structures, whereas the MLP cannot. Otherwise, both the RMLP and TLFN models are equivalent despite the marginal superiority of the latter over the former in the present study.

To further assess the comparative performance of the models, the observed and four-month-ahead predicted reservoir inflows (for the second case, using flows and ENSO inputs) are presented in Fig. 5(a)–(c) which shows the comparative hydrograph plots of observed inflows and four-month-ahead predicted inflows of the BNN against the MLP, TLFN and RMLP, respectively, for a portion of the testing dataset (January 1988–December 1990). In general the figures show that the models are all able to make reasonable predictions of low and medium reservoir inflows but then either under- or over-predict the peaks. The ability of these models to provide reasonable forecasts of low and medium reservoir inflows could be useful for assessing the amount of inflow to the reservoir in advance. In general, all the models are comparable but further improvement is still needed, especially for peak flow forecasting.

CONCLUSIONS

This paper investigates the potential of different data-driven methods to forecast seasonal reservoir inflows of the Churchill Falls watershed (northeastern Canada) using the ENSO index series. All data-driven models developed in this paper showed comparable performance with the RMLP and TLFN showing the best performance as measured by global goodness of fit measures. However, when considering the addition of ENSO data, the prediction results showed that the BNN model was able to capture this additional information and provide improved predictions in spring and summer seasons relative to the model using only reservoir inflows. The RMLP and TLFN showed improvements only in the spring season while the MLP model, in contrast, was unable to capture the relevant information from the ENSO index series for any season and yielded poor (or even deteriorating) results relative to the model using only reservoir inflows. However, all data-driven models considered in the present study failed to accurately model the peak flows, either under- or over-estimating the peaks, and therefore require further improvement. The ability of the BNN to capture additional information from the ENSO index series is a good indication of its potential for improved seasonal inflow forecasting. However, this potential still needs further investigation, including consideration of other low-frequency climatic indices.

Acknowledgements This work was made possible through a grant from the Natural Sciences and Engineering Research Council of Canada (NSERC) to the second author. The authors gratefully acknowledge the Hydro-Quebec Prediction Department for providing the experimental data. The authors also greatly appreciated the valuable comments and suggestions from the two anonymous reviewers and the associate editor.

REFERENCES

ASCE Task Committee on Application of Artificial Neural Network in Hydrology (2000a) Artificial neural network in hydrology. I: Preliminary concepts. *J. Hydrol. Engng* **5**, 115–123.

ASCE Task Committee on Application of Artificial Neural Network in Hydrology (2000b) Artificial neural network in hydrology. II: Hydrologic applications. *J. Hydrol. Engng* **5**, 124–137.

Barnston, A. G., Thiao, W. & Kumar, V. (1996) Long-lead forecast of seasonal precipitation in Africa using CCA. *Weather Forecasting* **11**, 506–520.

Coulibaly, P. & Burn, D. H. (2004) Wavelet analysis of the variability in annual Canadian streamflows. *Water Resour. Res.* **40**, W03105, doi:10, 1029/ 2003 WR002667.

Coulibaly, P. & Burn, D. H. (2005) Spatial and temporal variability of Canadian seasonal streamflows. *J. Climate* **18**(1), 191–210.

Coulibaly, P., Anctil, F. & Bobée, B. (1999) Hydrological forecasting using artificial neural networks: the state of the art (in French). *Can. J. Civil Engng* **26**, 293–304.

Coulibaly, P., Anctil, F., Rasmussen, P. & Bobée, B. (2000a) A recurrent neural networks approach using indices of low-frequency climatic variability to forecast regional annual runoff. *Hydrol. Processes* **14**, 2755–2777.

Coulibaly, P., Anctil, F. Rasmussen, P. & Bobée, B. (2000b) Neural network-based long-term hydropower forecasting system. *Computer-Aided Civil and Infrastructure Engng* **15**, 355–364.

Coulibaly, P., Anctil, F. & Bobée, B. (2001a) Multivariate reservoir inflow forecasting using temporal neural networks. *J. Hydrol. Engng ASCE* **6**, 367–376.

Coulibaly, P., Bobée, B. & Anctil, F. (2001b) Improving extreme hydrologic events forecasting using a new criterion for artificial neural network selection. *Hydrol. Processes* **15**, 1533–1536.

Dawson, C. W. & Wilby, R. L. (2001) Hydrological modeling using artificial neural networks. *Progr. Phys. Geogr.* **25**, 80–108.

Dibike, Y. B., Solomatine, D. & Abbott, M. B. (1999) On the encapsulation of numerical-hydraulic models on artificial neural network. *J. Hydraul. Res.* **37**, 147–161.

Eltahir, E. A. B. (1996) El Nino and the natural variability in the flow of the Nile River. *Water Resour. Res.* **32**, 132–137.

Giustolisi, O. & Laucelli, D. (2005) Improving generalization of artificial neural networks in rainfall–runoff modelling. *Hydrol. Sci. J.* **50**(3), 439–457.

Haykin, S. (1999) *Neural Networks: A Comprehensive Foundation* (second edn). Prentice Hall, Upper Saddle River, New Jersey, USA.

Khan, M. S. & Coulibaly, P. (2006) Bayesian neural network for rainfall-runoff modeling. *Water Resour. Res.* **42**, W07409, doi:10.1029/2005WR003971.

Lampinen, J. & Vehtari, A. (2001) Bayesian approach for neural networks – review and case studies. *Neural Networks* **14**, 257–274.

McKerchar, A. I., Pearson, C. P. & Moss, M. E. (1996) Prediction of summer inflows to lakes in the Southern Alps, New Zealand, using the spring Southern Oscillation Index. *J. Hydrol.* **184**, 175–187.

Neal, R. M. (1996) *Bayesian Learning for Neural Networks*. Springer-Verlag, New York, USA.

Piechota, T. C. & Dracup, J. A. (1999) Long-range streamflow forecasting using El Nino-Southern Oscillation indicators. *J. Hydrol. Engng* **4**, 144–151.

Piechota, T. C., Chiew, F. H. S. & Dracup, J. A. (1998) Seasonal streamflow forecasting in eastern Australia and the El Nino-Southern Oscillation. *Water Resour. Res.* **34**, 3035–3044.

Principe J., Euliano, N. & Lefebvre, C. (2000) *Neural and Adaptive Systems: Fundamentals Through Simulation*. John Wiley, Chichester, UK.

Rasmusson, E. M. & Carpenter, T. H. (1982) Variations in tropical sea surface temperature and surface wind fields associated with the Southern Oscillation/El Nino. *Mon. Weather Rev.* **110**, 354–384.

Shabbar, A. & Barnston, A. G. (1996) Skill of seasonal climate forecasts in Canada using canonical correlation analysis. *Mon. Weather Rev.* **124**, 2370–2385.

Simpson, H. J., Cane, M. A., Herczeg, A. L,. Zebiak, S. E. & Simpson, J. H. (1993) Annual river discharge in southeastern Australia related to El Nino-Southern Oscillation forecasts of sea surface temperatures. *Water Resour. Res.* **29**(11), 3671–3680.

Sveinsson, O. G. B. (2003a) Using climate information for forecasting flows in Churchill Falls basin, Québec Canada. Part 1: Analysis of climatic states and circulation to Québec spring streamflows. Technical report for Ouranos and Hydro- Québec, Canada, International Research Institute for Climate Prediction, Columbia University, Palisades, New York.

Sveinsson, O. G. B. (2003b) Using climate information for forecasting flows in Churchill Falls basin, Québec Canada. Part 2: Forecast of spring inflow. Technical report for Hydro-Québec and Ouranos, Canada, International Research Institute for Climate Prediction, Columbia University, Palisades, New York, USA.

Tootle, G. A., Piechota, T. C. & Singh, A. (2005) Coupled oceanic–atmospheric variability and US streamflow. *Water Resour. Res.* **41**, W12408, doi: 10, 1029/ 2005 WR004381.

Wang, G. L. & Eltahir, E. A. B. (1999) Use of ENSO information in the medium- to long-range forecasting of the Nile flood. *J. Clim.* **12**, 1726–37.

Whitaker, D. W., Wasimi, S. A. & Islam, S. (2001) The El Nino-Southern Oscillation and long-range forecasting of flows in the Ganges. *Int. J. Climatol.* **21**, 77–87.

Table 1 Basic descriptive statistics of variables used in modeling

Data	Variable	Count	Mean	Variance	Minimum	Median	Maximum	Skewness
Training	Flow	504	1431	1520782	195	1025	6272	1.77
	ENSO	504	-0.013	0.627	-1.740	-0.092	3.444	0.822
Testing	Flow	168	1256	870245	203	1013	4871	1.51
	ENSO	168	0.341	1.024	-1.693	0.150	3.725	0.963
Total data	Flow	708	1387	1349063	195	1027	6272	1.78
	ENSO	708	0.052	0.729	-1.740	-0.067	3.725	1.021

Table 2 Optimal network architecture and parameters

Optimized Variable or parameter	**Optimal network architecture and parameters**							
	Flows				Flows and ENSO			
	RMLP	TLFN	MLP	BNN	RMLP	TLFN	MLP	BNN
Number of neurons	8	5	8	7	9	7	10	10
Number of epochs	2000	2000	2000	2000	2000	2000	2000	2000
Number of hidden layer	1	1	1	1	1	1	1	1
Hidden layer transfer function	Tangent hyberbolic							
Output layer transfer function	Linear							

Table 3 Model performance statistics of seasonal predictions: (a) using the MLP and BNN and (b) using TLFN and RMLP.

(a)

Model performance statistics	MLP					BNN				
	Flows		*Flows and ENSO*		Skill (%)	*Flows*		*Flows and ENSO*		Skill (%)
	Training	Testing	Training	Testing		Training	Testing	Training	Testing	
4-month ahead prediction										
MSE	368952	320378	347229	340239	-6	447710	408902	302572	353203	14
NMSE	0.24	0.35	0.23	0.37		0.30	0.45	0.20	0.39	
MAE	379	386	369	414		466	492	357	399	
Min Abs Error	0.39	0.55	0.55	0.59		0.01	2.28	0.27	0.24	
Max Abs Error	3287	2259	3267	2050		3256	1872	2968	2037	
r	0.87	0.83	0.88	0.81		0.84	0.80	0.89	0.82	
8-month ahead prediction										
MSE	696683	737258	651567	740760	0	647783	726365	666300	614964	15
NMSE	0.48	0.79	0.45	0.80		0.44	0.78	0.46	0.66	
MAE	544	649	507	625		509	615	526	585	
Min Abs Error	0.70	17.26	0.09	1.25		1.10	0.46	0.45	13.38	
Max Abs Error	4545	2697	4686	3422		4842	3214	4640	2316	
r	0.72	0.62	0.74	0.66		0.75	0.66	0.74	0.71	
12-month ahead prediction										
MSE	255866	385044	242109	396668	-3	400920	343367	405000	338108	2
NMSE	0.18	0.41	0.17	0.43		0.27	0.37	0.28	0.36	
MAE	335	405	327	414		406	400	401	390	
Min Abs Error	0.10	3.51	1.54	0.67		0.35	3.41	0.03	2.70	
Max Abs Error	3229	2538	2992	2263		3147	2171	3199	2191	
r	0.91	0.79	0.91	0.77		0.85	0.79	0.85	0.80	

(b)

Model performance statistics	RMLP					TLFN				
	Flows		*Flows and ENSO*		Skill (%)	*Flows*		*Flows and ENSO*		Skill (%)
	Training	Testing	Training	Testing		Training	Testing	Training	Testing	
4-month ahead prediction										
MSE	400378	452617	404430	401819	11	219988	346815	200075	302843	13
NMSE	0.26	0.50	0.27	0.44		0.15	0.38	0.13	0.33	
MAE	436	519	435	476		320	387	316	374	
Min Abs Error	0.23	3.36	1.67	5.98		0.54	3.90	0.66	7.90	
Max Abs Error	3219	2721	3284	2177		2460	2811	2246	2486	
r	0.86	0.77	0.86	0.77		0.92	0.81	0.93	0.83	
8-month ahead prediction										
MSE	277397	306123	283055	305243	0	219897	276796	198261	318133	-15
NMSE	0.19	0.33	0.19	0.33		0.15	0.30	0.14	0.34	
MAE	352	375	351	377		315	352	302	405	
Min Abs Error	1.27	2.25	0.82	1.19		4.81	0.02	0.03	0.14	
Max Abs Error	3182	2503	3291	2561		2355	2460	2252	2425	
r	0.90	0.82	0.90	0.82		0.92	0.84	0.93	0.83	
12-month ahead prediction										
MSE	280137	274721	281051	303108	-10	225616	306464	231561	314543	-3
NMSE	0.19	0.29	0.19	0.33		0.15	0.33	0.16	0.34	
MAE	348	361	354	396		317	386	325	380	
Min Abs Error	1.60	2.56	0.18	11.76		0.10	5.41	0.15	2.41	
Max Abs Error	3174	2182	3021	2673		2365	2089	2373	3205	
r	0.90	0.84	0.90	0.82		0.92	0.84	0.92	0.82	

MSE = mean squared error; NMSE = normalized mean squared error; MAE = mean absolute error; Max abs error = maximum absolute error; Min abs error = minimum absolute error; r = linear correlation coefficient.

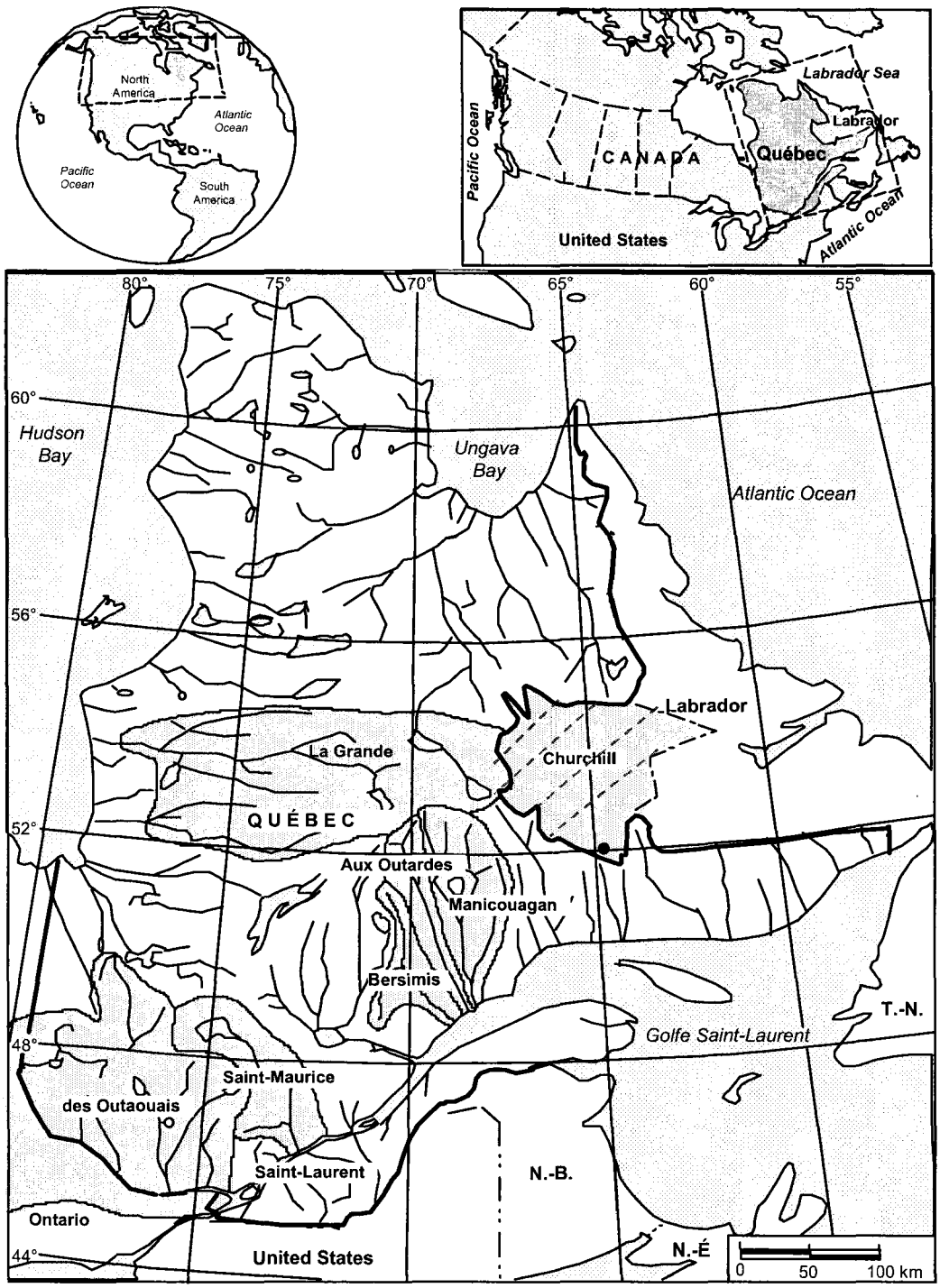


Fig. 1 Location of study basin in northeastern Canada ("black dot" denotes reservoir location within the Churchill Falls watershed).

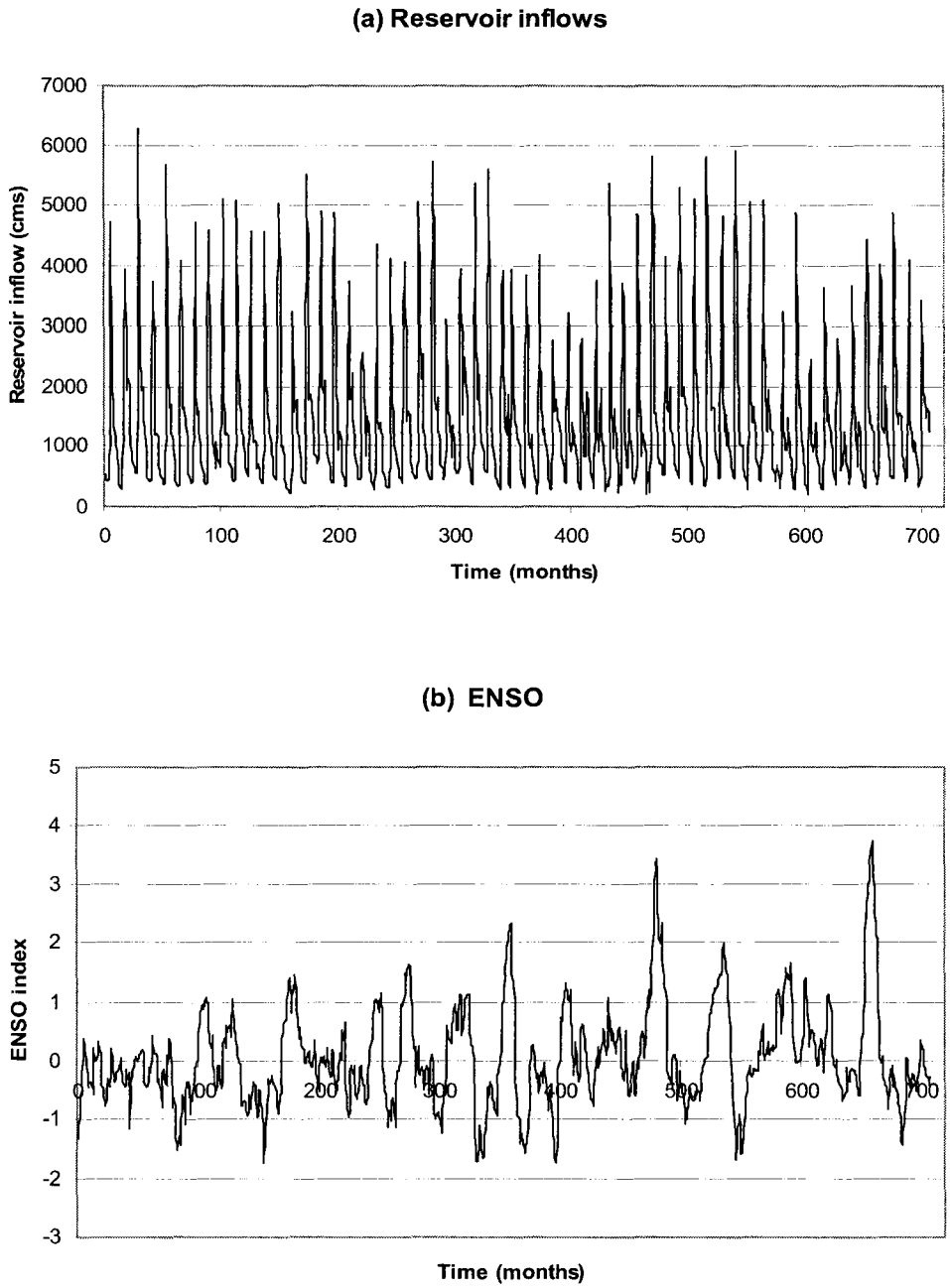


Fig. 2 Time series plots of (a) reservoir inflows and (b) ENSO index series from January 1943 to December 2001.

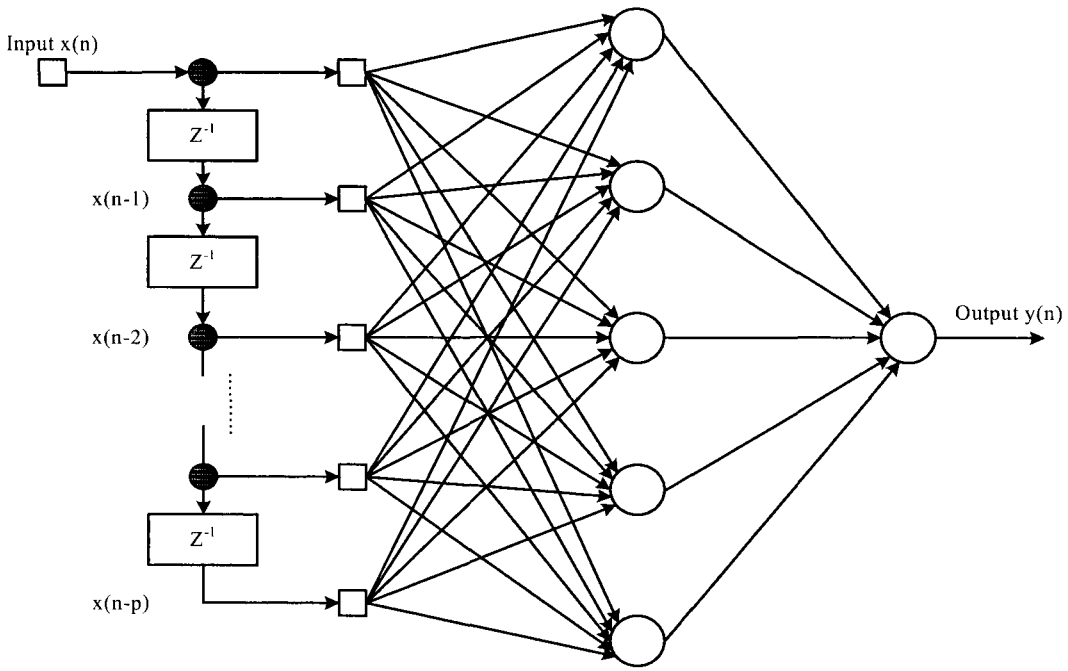


Fig. 3 Focused time-lagged feed-forward network (TLFN).

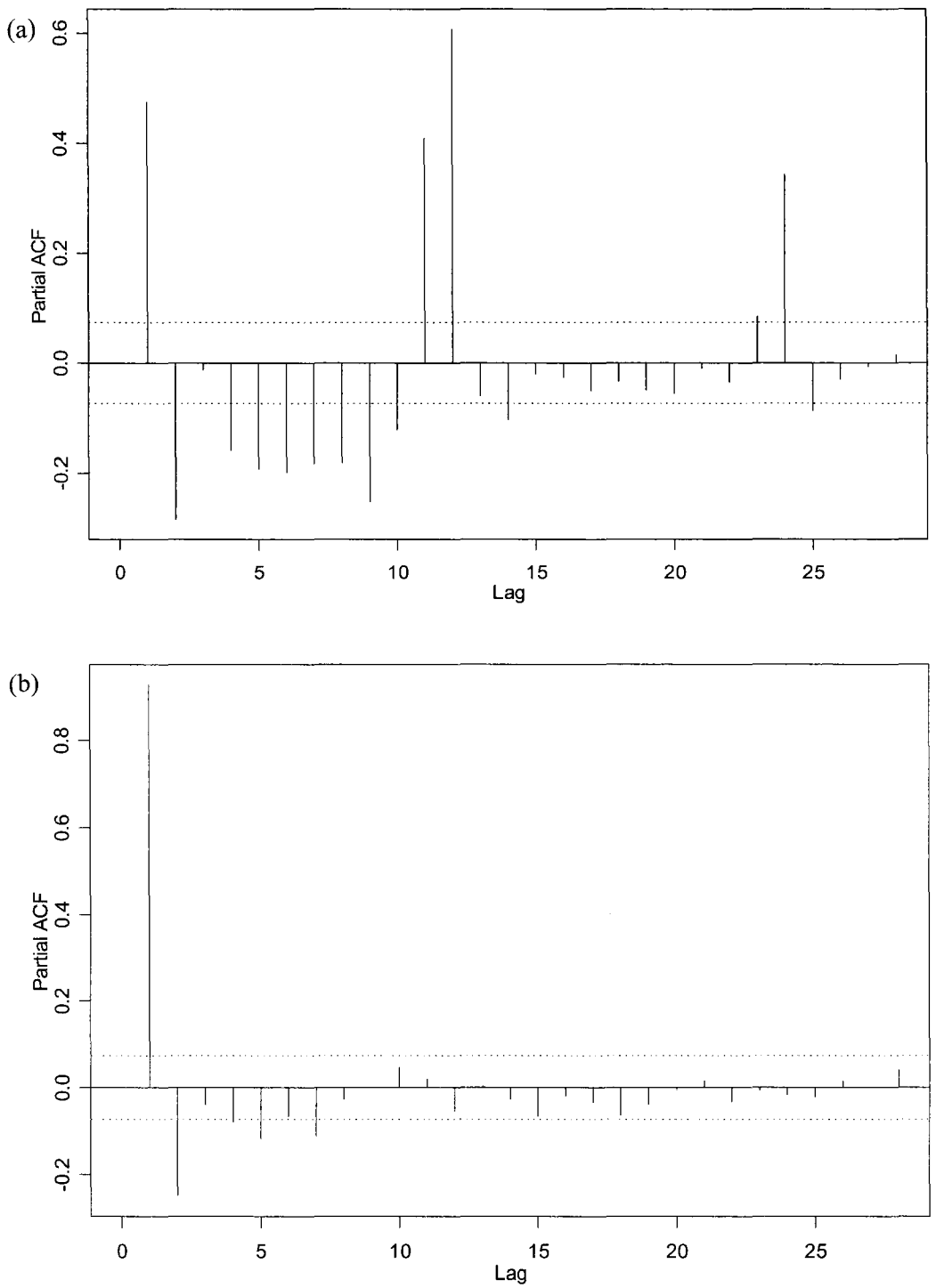


Fig. 4 Partial autocorrelation function (PACF): (a) flow series; (b) ENSO series.

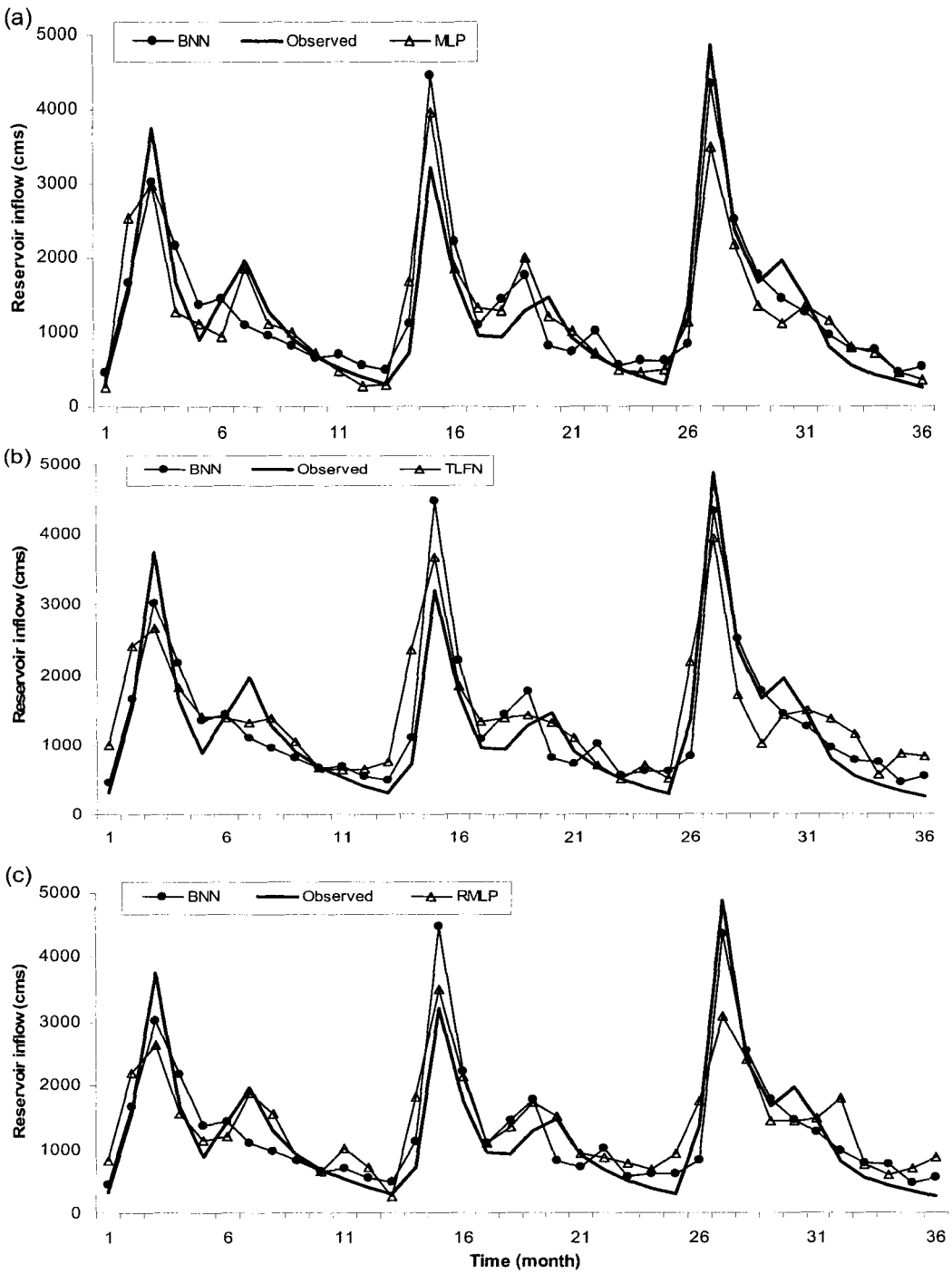


Fig. 5 Observed and four-month-ahead predicted inflows for: (a) the MLP and BNN model, (b) the TLFN and BNN model, and (c) the RMLP and BNN model (in each case for a portion of the testing data set from January 1988–December 1990).

CHAPTER 3

Improving long-range streamflow forecasts with extended Kalman filter

This chapter is devoted to the study of an extended Kalman filter algorithm, which is perhaps the most elegant of all learning algorithms, to perform supervised training of a recurrent multilayer perceptron. The method is applied to characterize long-term horizon streamflows in western Alberta (Canada). The performance of the proposed model is compared against the conventional multilayer perceptron through suites of traditional and distribution-based diagnostic measures. The findings of the study are presented in a paper form and submitted to Hydrological Science Journal for possible publication. This forms the second task on long-range hydrologic forecasting.

Improving long-range streamflow forecasts with extended Kalman filter

Abstract

There is a continuing effort to advance the skill of long-range hydrologic forecasts to support water resources decision making. The present study investigates the potential of an extended Kalman filter approach to perform supervised training of a recurrent multilayer perceptron (RMLP) to forecast long-range horizon streamflows in western Alberta, Canada. The performance of the RMLP was compared with the conventional multilayer perceptron (MLP) using suites of traditional and distribution-based diagnostic measures. The results of the forecasting experiment showed that the RMLP model was able to provide a robust modeling framework capable of describing complex dynamics of the hydrologic processes, thereby yielding more accurate forecasts than the MLP model. The performance of the method in the present study is very promising; however, further investigation is required to ascertain the versatility of the approach in characterizing different water resources and environmental problems.

Keywords: Alberta; forecast; Kalman filters; recurrent multilayer perceptron; streamflow

1. Introduction

Hydrologic forecasting constitutes an essential component of a decision support system for water resources management. Timely and reliable hydrologic forecasts present a potential for considerable economic, health and safety, and other benefits associated with power generation, flood control, inland navigation, recreation, irrigation, water supply, water quality, fish and wildlife enhancement. The accuracy of a hydrologic forecast model, designed to characterize flow processes, largely depends on the extent and the complexity in river flow dynamics. There have been diverse hydrologic models proposed for the purpose of forecasting hydrological variables, including conceptual, physically-based and data-driven approaches. However, none of these models can be

considered as distinct and superior (ASCE, 2000a). For details of the different hydrologic models see, for example, Singh (1995).

The understanding in many areas (including hydrologic processes) is far from perfect, such that empiricism plays an important role in modeling studies (ASCE, 2000b). Linear and multiple regression-based models, such as principal component analysis (PCA), partial least squares (PLS) regression, and canonical correlation analyses (CCA) have been examined for forecasting long-range hydrological variables (e.g., Barnston et al., 1996; Tootle et al., 2007; Sveinsson et al., 2008). Probabilistic models have also been considered to forecast long-range horizon streamflows (Piechota et al., 1998). Likewise, autoregressive-based models have been extensively investigated to characterize a wide spectrum of water resources and environmental problems. These traditional models, however, have been considered to be inadequate or inappropriate in some studies for forecasting streamflows (Piechota et al., 1998; Khan and Coulibaly, 2006; Amenu et al., 2007; Kisi and Cigizoglu, 2007). In order to advance the skill of characterizing complex and highly nonlinear river basin systems, the research community has resorted to neural network approaches (e.g., Pan and Wang, 2005; Giustolisi and Laucelli, 2005; Muluye and Coulibaly, 2007; Carcano et al., 2008). In particular, the demonstrated capability of artificial neural networks (ANNs) in characterizing the non-linearity of dynamic systems and its flexibility to integrate several factors has made it a more feasible and attractive modeling tool for describing complex nonlinear hydrologic processes (Hsu et al., 1995). Extensive reviews on the use of artificial neural networks in hydrology-related applications have been reported in ASCE (2000a; 2000b) and Dawson and Wilby (2001). These and several other studies have suggested that artificial neural networks can be an effective substitute to the traditional modeling approaches for characterizing complex hydrologic systems.

Gobena and Gan (2010) applied the Sacramento Soil Moisture Accounting (SACSMA) model for forecasting seasonal streamflows for the Bow and Castle rivers in the South Saskatchewan River basin, Alberta. The ranked probability skill score (RPSS) values associated with the Bow River show no improvement over the climatology for

forecasts issued prior to Dec 1, and 33% and 59% improvements for forecasts issued on Dec 1 and Apr 1, respectively. For the Castle River, the RPSS values were slightly higher than zero for forecasts issued prior to January 1, and 22% and 36.7% for forecasts issued on January 1 and February 1, respectively. Kalra and Ahmad (2009) used a support vector machine (SVM) and the conventional ANN to predict three step ahead annual streamflow volumes in the Upper Colorado River basin. In their study, Lees Ferry, Cisco and Green river streamflow gages were considered. The reported Nash-Sutcliffe efficiency coefficient (CE) values for the testing period were 0.47 for Lees Ferry, 0.67 for Cisco and 0.46 for Green river when the SVM was used, and 0.28 for Lees Ferry, 0.36 for Cisco and 0.41 for Green river when the conventional ANN was used. Pramanik and Panda (2009) compared the performance of the conventional ANNs and an adaptive neurofuzzy inference system (ANFIS) to predict the barrage outflow in the Mahanadi River, India. Their investigation shows that the ANFIS was slightly better than the conventional ANN for predicting one day lead outflow. The reported CE values were 0.71 and 0.69 for the ANFIS and the conventional ANN, respectively. Adamowski (2008) developed wavelet forecasting method (WT) and compared its performance with a multiple linear regression (MLR), an autoregressive integrated moving average (ARIMA), and the conventional ANN for forecasting daily stream flows with lead times equal to 1, 2 and 6 days for the Rideau River watershed in Ontario, Canada. The CE values for the 6 days lead time were 0.44 for the WT, 0.55 for the ANN and 0.34 for the MLR. The WT generally performed better for 1 and 2 days lead time only.

Artificial neural networks, in particular, a multilayer perceptron (MLP) has demonstrated its potential as a universal function approximation to describe diverse water resources problems (ASCE, 2000a,b). Multilayer perceptron together with time-lag feedforward network architectures have been applied to describe both spatial and temporal dynamics of the hydrologic processes (e.g., Muluye and Coulibaly, 2007). However, it has been shown that a neural network containing a state feedback is more effective, possesses more computational advantages, and is less vulnerable to external noises than input-output models (Palma et al., 2001). Furthermore, the state-space models

employ fewer numbers of parameters, and are capable of describing a large-class of dynamic systems than common input-output models (Palma et al., 2001; Haykin, 2008). One way of effective characterization of state feedback is through the use of recurrent neural networks. There are several algorithms available to perform supervised training of recurrent neural networks. The most widely used algorithms are backpropagation through time (BPTT), real-time recurrent learning (RTRL), and extended Kalman filters (EKF). While the BPTT is the most commonly used method, the RTRL is mathematically straight forward, and the EKF is arguably the technique that performs the best (Jaeger, 2002). Among others, Jonsson and Palsson (1994) and Puskorius and Feldkamp (1994) are the pioneer investigators of the EKF application, who showed that a state filter based on recurrent neural networks can converge to the minimum variance filter. Since then the EKF has found applications in several nonlinear dynamic systems to perform supervised training of recurrent neural networks (Haykin, 2008).

The use of recurrent neural networks in the context of hydrological applications is rather limited and has received modest attention when compared with multilayer perceptron models (e.g., Chang et al., 2004; Pan and Wang, 2005; Muluye and Coulibaly, 2007; Carcano et al., 2008). Furthermore, none of these studies has investigated the EKF to perform training of recurrent neural networks for use in hydrology-related problems. The present study, therefore, attempts to address this gap in the framework of improving long-range hydrologic forecasts in a catchment where the hydrology is dominated by glaciers and snow melt. The specific goals of this study are to: (i) investigate the potential of the EKF approach to perform supervised training of the recurrent multilayer perceptron (RMLP) to forecast long-range horizon streamflows; and (ii) evaluate the forecasting skill of the proposed model over the conventional neural network, namely, MLP using suites of conventional and distribution-based diagnostic measures. The MLP model is selected in the present study based on preliminary analysis and literature review. Other neural network architectures and regression-based models have also been investigated; however, their forecasting results did not show significant improvement over the MLP for the study basin. Furthermore, given that the MLP model is the most

widely used and a well-tested architecture, its use as a basis for comparison is considered appropriate (Pramanik and Panda, 2009; Kalra and Ahmad, 2009; Adamowski, 2008; ASCE, 2000a,b). Therefore, only the MLP model is included in the comparison. The remainder of the material is organized as follows. Section 2 provides description of the forecasting models used in the study. The study area, data and application of the models are presented in section 3. Section 4 discusses the results from the forecasting experiments. Finally, section 5 draws some conclusions and presents recommendations in view of the forecasting results.

2. Model Description

This section describes recurrent multilayer perceptron and its training techniques. The multilayer perceptron, which is selected as a benchmark for model comparison, is also briefly described.

2.1. Multilayer perceptron

Artificial neural networks have witnessed increased application due to the development of more sophisticated algorithms and the emergence of powerful computational tools (ASCE, 2000a). The multilayer perceptron, which is also referred to as feedforward multilayer perceptron, is the most widely used neural network (e.g., Haykin, 2008; ASCE, 2000a,b). In the present study, a simple MLP with a single hidden layer was used. Let $\mathbf{x}_I(n)$ represent the output of the hidden layer, $\mathbf{x}_o(n)$ represent the output of the output layer, and $\mathbf{u}(n)$ represent the input vector, then the operational principles of the MLP can be mathematically expressed by the following equations (Haykin, 2008):

$$\mathbf{x}_I(n+1) = \varphi_I\left(\mathbf{w}_I.\mathbf{u}(n)\right) \tag{1}$$

$$\mathbf{x}_o(n+1) = \varphi_o\left(\mathbf{w}_o.\mathbf{x}_I(n+1)\right) \tag{2}$$

where $\varphi_I(.,.)$ and $\varphi_o(.,.)$ are the activation functions of the hidden layer and output layer, respectively; and $\mathbf{w}_I$ and $\mathbf{w}_o$ represent the weight matrices of the hidden layer and output layer, respectively. The network biases were integrated with the network weights in order

to simplify the computation of network parameters. The activation function of the network could be a linear, sigmoid or hyperbolic tangent function, depending on the complexity of the problem. The network parameters were estimated by presenting a training example ($\mathbf{u}(n)$, $\mathbf{x}_o(n)$), with the input vector $\mathbf{u}(n)$ applied to the input layer, and the desired response $\mathbf{x}_o(n)$ presented to the output layer of the computational neuron. The training of the network was then conducted using a gradient-based optimization technique, which utilizes derivatives of the error with respect to the network parameters calculated by the back-propagation algorithm (Muluye and Coulibaly, 2007). The detailed theory and derivation of the back-propagation algorithm can be found in Haykin (2008).

2.2. Recurrent multilayer perceptron

The applications of artificial neural networks in the modeling of nonlinear and dynamic processes are primarily dominated by static architectures (e.g., MLP) trained by a gradient descent algorithm. An effective way of representing temporal and sequential information is through the use of recurrent network architectures, where feedback connections exist between nodes of the same layer or to nodes of preceding layers (Puskorius and Feldkamp, 1994). The present study used a recurrent neural network with a single hidden layer. The operational principles of the network can be mathematically expressed by (Haykin, 2008):

$$\mathbf{x}_I(n+1) = \varphi_I\left(\mathbf{w}_I \cdot \begin{bmatrix} \mathbf{x}_I(n) \\ \mathbf{u}(n) \end{bmatrix}\right) \tag{3}$$

$$\mathbf{x}_o(n+1) = \varphi_o\left(\mathbf{w}_o \cdot \begin{bmatrix} \mathbf{x}_o(n) \\ \mathbf{x}_I(n+1) \end{bmatrix}\right) \tag{4}$$

Although representation and processing of temporal information are intrinsic capabilities of recurrent neural networks, their applications are primarily restricted by two inter-related difficulties (Puskorius and Feldkamp, 1994): (i) the computation of dynamic derivatives of the RMLP outputs with respect to its weights by the RTRL algorithm is computationally intensive; and (ii) the training of the RMLP with pure gradient descent

methods is typically slow and ineffective. The first issue can be eased by approximate methods such as a truncated BPTT algorithm, whereas the latter can be partially addressed by using second-order training algorithms such as linearized recursive least squares or extended Kalman filtering.

2.2.1. Extended Kalman filter

An extended Kalman filter is a modified and nonlinear version of the Kalman filter. This algorithm has found applications in a number of problems including learning the weights of neural networks (Puskorius and Feldkamp, 1994; Choi et al., 2005). The complete coverage of the subject can be found in one of many sources (e.g., Haykin, 2008). The basic framework for the EKF involves estimation of the parameters by re-writing a new state-space representation of equations (3) and (4) above as (Haykin, 2008):

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \omega(n) \tag{5}$$

$$\mathbf{x}_o(n) = \mathbf{c}(\mathbf{w}(n), \mathbf{u}(n), \mathbf{v}(n)) + v(n) \tag{6}$$

where the parameter $\mathbf{w}(n)$ corresponds to stationary connection weights for the entire network (i.e. $\mathbf{w}(n)$ is the aggregation of weight vectors $\mathbf{w}_I$ and $\mathbf{w}_o$), driven by an artificial process noise $\omega(n)$, the variance of which is zero and determines convergence. $\mathbf{v}(n)$ is the vector containing all recurrent inputs, $v(n)$ is the measurement noise vector, $\mathbf{u}$(n) is the network input vector as defined earlier, and the nonlinear map or transfer function $\mathbf{c}(\cdot)$ is parameterized by the vector $\mathbf{w}$.

The nonlinearity in the transfer function prevents the direct application of a classical Kalman filter approach for learning network parameters. This serious limitation is resolved by applying the extended Kalman filter as a sub-optimal filter. In the implementation of the EKF, the nonlinear measurement terms are linearized using a Taylor series. By simplifying and considering only the first-order linear terms we obtain (Haykin, 2008):

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \omega(n) \tag{7}$$

$$\hat{d}(n)=\mathbf{C}(n)\mathbf{w}(n)+v(n) \tag{8}$$

where $\hat{d}(n)$ is the first order approximation of $\mathbf{x}_o(n)$, and $\mathbf{C}(n)$ is the p-by-W measurement matrix of the linearized model. The Jacobian matrix $\mathbf{C}(n)$ is computed as (Haykin, 2008):

$$\mathbf{C}(n)=\frac{\partial \mathbf{c}(\hat{\mathbf{w}}(n),\mathbf{v}(n),\mathbf{u}(n))}{\partial \mathbf{w}}$$

$$=\begin{bmatrix} \frac{\partial c_1}{\partial w_1} & \frac{\partial c_1}{\partial w_2} & \cdots & \frac{\partial c_1}{\partial w_W} \\ \frac{\partial c_2}{\partial w_1} & \frac{\partial c_2}{\partial w_2} & \cdots & \frac{\partial c_2}{\partial w_W} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial c_p}{\partial w_1} & \frac{\partial c_p}{\partial w_2} & \cdots & \frac{\partial c_p}{\partial w_W} \end{bmatrix} \tag{9}$$

where $c_p(\cdot) = [c_1, c_2 \ldots c_p]$ are the p outputs of the network, and $\mathbf{w} = [w_1, w_2 \ldots w_W]$ are ordered w_W weights of the network. The current estimate of the state $\hat{\mathbf{w}}(n)$ is used in the computation of the derivatives. Obviously, the computation of the Jacobian makes the main distinction in the application of the EKF training to the MLP and the RMLP architectures. Finally, the network weights are updated by the following EKF recursion (Haykin, 2008):

$$\Gamma(n)=\left[\sum_{i=1}^{g}\mathbf{C}_i(n)\mathbf{K}_i(n,n-1)\mathbf{C}_i^T(n)+\mathbf{R}(n)\right]^{-1} \tag{10}$$

$$\mathbf{G}_i(n)=\mathbf{K}_i(n,n-1)\mathbf{C}_i^T(n)\Gamma(n) \tag{11}$$

$$\alpha(n)=\mathbf{d}(n)-\hat{\mathbf{d}}(n/n-1) \tag{12}$$

$$\mathbf{w}_i(n+1/n)=\hat{\mathbf{w}}_i(n/n-1)+\mathbf{G}_i\alpha(n) \tag{13}$$

$$\mathbf{K}_i(n+1,n)=\mathbf{K}_i(n,n-1)-\mathbf{G}_i(n)\mathbf{C}_i(n)\mathbf{K}_i(n,n-1)+\mathbf{Q}_i(n) \tag{14}$$

where $i = 1, 2, \ldots, g$, and g is the number of groups. $\Gamma(n)$ = p-by-p matrix, representing the global conversion factor for the entire network. $\mathbf{G}_i(n)$ = W_i- by-p matrix, representing the Kalman gain group i of neurons. $\alpha(n)$ = p-by-1 vector, representing the innovations

defined as the difference between the desired response $\mathbf{d}(n)$ for the linearized system and its estimate $\hat{\mathbf{d}}(n/n-1)$ based on input data available at time n-1; the estimate $\hat{\mathbf{d}}(n/n-1)$ is represented by the actual output vector $\mathbf{y}(n)$ of the network residing in state $\{\mathbf{w}_i(n/n\text{-}1)\}$, which is produced in response to the input $\mathbf{u}(n)$. $\hat{\mathbf{w}}_i(n/n-1)$ = W- by-1 matrix, representing the estimate of the weight vector $\mathbf{w}_i(n)$ for group i at time n, given the observed data up to time n-1. $\mathbf{K}_i(n,n-1)$ = W-by-W matrix, representing the error covariance matrix for group i of neurons. $\mathbf{R}(n)$ is the diagonal covariance matrix for the measurement noise vector $\nu(n)$. $\mathbf{Q}(n)$ is a diagonal covariance matrix for the artificial process noise that provides a mechanism by which the effect of the artificial process noise vector $\omega(n)$ is included in the Kalman recursion. This algorithm is known as the decoupled extended Kalman filter (DEKF). In the limit of a single weight group (g = 1), the DEKF algorithm reduces exactly to the global extended Kalman filter (GEKF). The computational complexity and the storage requirements for the DEKF can be considerably less than the GEKF, but at the expense of a slightly reduced accuracy (Puskorius and Feldkamp, 1994).

2.2.2. Backpropagation through time

In order to compute the Jacobian matrix $\mathbf{C}(n)$, an efficient gradient algorithm, such as the BPTT or the RTRL is typically used. The present study employed the BPTT algorithm. The solution of the BPTT approach is to unfold the RMLP in time into a multilayer feedforward network by stacking identical copies of the RMLP, and then redirecting connections within the network to obtain connections between subsequent copies, which are amenable to the backpropagation algorithm (Jaeger, 2002). This method of unfolding the RMLP leads to the generation of a network with an infinite number of layers. For practical application, the truncated version of the original BPTT algorithm is used. In the truncated BPTT algorithm, the past network states are saved up to some truncation depth h, and information beyond this depth is not considered. The detailed description of the BPTT algorithm can be found in Haykin (2008).

3. Case Study: Streamflow Forecasting

3.1. Catchment description and data used

The study area selected for evaluation and investigation of the EKF approach to perform supervised training of the recurrent multilayer perceptron is the Athabasca River basin in western Canada. The Athabasca River is part of the Mackenzie hydrologic zone, with a gross drainage area of approximately 9765 square kilometers. The Athabasca River originates from the Colombia Glacier at an altitude of about 1600m. The river flows along various landscapes including ice fields and gorges, which, in turn, create favorable conditions for flora and fauna to flourish. These habitats are protected by several national and provincial parks, including the Jasper National Park, part of a spectacular World Heritage Site. Several communities are on the banks of the river, and Hinton is one of them (Schindler et al., 2007). The grounds for selecting this particular river basin are three fold: (i) reliable and accurate flow forecasts in this river basin could have great consequences to the tourism industry, in particular, and to the local community, in general; (ii) the hydrology of the basin is largely driven by glacier and snow melt, which, together with wide-ranging landscape and land use patterns, pose great challenges to the reliability of hydrologic forecast models. Investigating the forecasting skill of the RMLP model under such a hydrologic system is imperative; and (iii) the basin contains reliable hydrometric data for a long period, which is a fundamental requirement for the design and successful implementation of artificial neural networks. The experimental data used in the present study are collected and maintained by Environment Canada (Hydat). Monthly streamflow data for a period of 47 years (Jan 1960 to Dec 2006) are compiled for the Athabasca River, at Hinton (07AD002), Alberta (Canada). The geographical location of the hydrometric station is approximately at 53.41°N and -117.58°W. Out of 47 years of monthly streamflow data, a subset of data from Jan-1960 to Dec-1993 were used for model calibration, and the remaining data, from Jan-1994 to Dec-2006, were employed for model evaluation.

3.2. Network topology and training

The flow time series was analyzed using partial autocorrelation to identify significant time lags. Consequently, lags 1 to 12 were identified as significant inputs to the forecasting model. Once the model inputs were identified, both the MLP and the RMLP models were investigated for long-range horizon streamflow forecasting. The flow forecasts considered include: 4-month (April), 8-month (August) and 12-month (December) ahead, where the month of January corresponds to a one-month ahead forecast. The selection of the 4-, 8- and 12-month ahead forecasts were done intentionally to represent forecasts from spring, summer and winter seasons, respectively. Prior to the forecasting experiment, appropriate network architectures were designed and the various parameters were optimized. For modeling with the MLP, the search for an optimal network began with a simple network having one hidden layer, keeping the number of neurons to vary from 2 to 30. Different learning rules (such as variable learning rate, conjugate gradient, Levenberg-Marquardt, and Bayesian regularization with Levenberg-Marquardt) and transfer functions (such as tangent hyperbolic, sigmoid and linear) in both the hidden layer and output layer were investigated in the search for an optimal network (Muluye and Coulibaly, 2007). For the case of modeling with the MLP, one hidden layer with 16 neurons provided the best performing network. A hyperbolic tangent and linear activation function in the hidden layer and output layer, respectively, and learning rule with Bayesian regularization with Levenberg-Marquardt provided the optimal network. Similarly, for the case of modeling with the recurrent multilayer perceptron, one hidden layer with 12 neurons provided the best performing network. While the MLP model was trained with the conventional back-propagation algorithm, the RMLP model was trained with the decoupled version of an extended Kalman filter. In the process of training the RMLP model, the covariance matrix of artificial noise $\mathbf{Q}(n)$ was linearly annealed from 10^{-2} to 10^{-6} in order to avoid numerical divergence and poor local minima. The diagonal entry of the covariance matrix of the measurement noise $\mathbf{R}(n)$ was also decreased linearly from 100 to 5 (Haykin, 2008). For the computation of the transfer

function $\mathbf{C}(n)$, the truncated BPTT algorithm was used. A truncation depth of three was retained as there was no significant difference in network performance beyond this depth.

4. Results and discussions

This section compares and discusses the MLP and the RMLP models for forecasting the 4-, 8- and 12-month ahead streamflows. Suites of conventional and distribution-based diagnostic statistics were computed for the validation period to aid model evaluation. The use of such diagnostic measures provides useful information to forecasters and water resources managers about the nature of the forecast system.

4.1. Conventional diagnostic measures

The traditional verification measures provided key statistics including bias (%), root mean squared error (RMSE), Pearson correlation coefficient (r) and deterministic skill score (SS). The SS explains the relative improvement of the forecast over some reference forecast. In the present study, forecasts from climatology and MLP were used as reference forecasts. When the mean squared error (MSE) is used as a score in the SS formulation, the resulting statistic is called the reduction of variance (Stanski et al., 1989). The SS is given by (Wilks, 1995):

$$\mathrm{SS} = 1 - \frac{MSE_{for}}{MSE_{ref}} \tag{15}$$

where MSE_{for} is the mean squared error for the forecast, and MSE_{ref} is the mean squared error for the reference forecast. The SS ranges from $-\infty$ to 1, for which a score of zero represents no improvement over the reference forecast, a score of one represents the perfect forecast, and a negative score suggests the reference forecast is better than the forecast.

Table 1 presents the biases associated with the 4-, 8- and 12-month ahead forecasts. These biases were expressed as a percentage of the mean climatology. The biases associated with the RMLP forecasts were much smaller than those associated with the MLP forecasts for all forecast horizons, signifying more accurate forecasts. With the

exception of the 4- month ahead forecast (made by the MLP model), in which the observed flows were over-forecasted (positive bias) by a relatively higher percentage (10.6%), both models generally yielded smaller biases (less than about 2%). The Pearson correlation coefficients between forecasted and observed streamflows indicate that the RMLP had an advantage over the MLP, in all forecast horizons considered. This view was further supported by the RMSE and the SS statistics: the RMLP model provided smaller RMSE and higher SS. The positive SS in each case suggests that the forecasts made by both models were superior to climatology. When the MSE of MLP was used as a reference forecast, the improvements in MSE due to RMLP (SS_RMLP) were 34.5%, 35.2% and 35.1%, respectively, for 4-, 8- and 12-month ahead forecasts. This clearly suggests the superiority of the RMLP over the MLP for all forecast ranges.

The RMLP model generally yielded better performance statistics than the different hydrologic models reported in the introduction section. Since a hydrologic model performs differently to different basins, comparison of the RMLP model with other models on a different basin is not valid. For this reason such comparison was not established.

4.2. Brier skill score

The Brier score (BS) is a common scalar accuracy measure typically employed to verify forecasts of dichotomous events. The BS is essentially the mean-squared probability error between the observations and forecasts, and is given by (Brier, 1950):

$$\mathrm{BS} = \frac{1}{n}\sum_{i=1}^{n}(f_i - o_i)^2 \tag{16}$$

where f_i is the forecast probability of occurrence of the event, o_i is one if the event occurred and zero if it did not, and n is a set of forecasts of a two category (dichotomous) element. The BS is a negatively oriented score ranging from 0 to 1, with zero representing a perfect score and one representing the worst possible score (Stanski et al., 1989).

The application of the Brier score is more important for dichotomous events, particularly when the focus is to verify whether the event has happened or not. In order to compute the score, it is necessary to define the event for verification. For this purpose, a threshold value of the 75^{th} percentile of the observed flows was considered. The information gained by the Brier score is relevant only for that event, and may not be the appropriate diagnostic measure if the forecast system is made beyond a simple percent "yes" and percent "no" for the particular event being verified (Stanski et al., 1989).

A single value, such as the Brier score, may not offer the actual skill of a forecast system. Thus it is necessary to compare the forecast of interest with some reference forecast, such as persistence forecasts, historical operational forecasts, or climatology forecasts (Wilks, 1995; Franz et al., 2003). In order to assess the relative skill of forecasts, the Brier skill score (BSS) was used (Stanski et al., 1989):

$$\mathrm{BSS} = 1 - \frac{BS_{for}}{BS_{ref}} \tag{17}$$

where BS_{for} is the Brier score for the forecast, and BS_{ref} is the Brier score for the reference forecast. It is a common practice to use climatology as a standard or a reference forecast, but any unskilled or even a skilled forecast can be used instead. The BSS explains the percentage improvement over the reference forecast, and ranges from $-\infty$ to 1. A positive BSS suggests that the forecast system has skill over the reference forecast, unity representing a perfect forecast.

A summary of statistics in Table 1 indicates significant differences when the BS statistics from the RMLP model were compared with those from the MLP model. The BS statistics associated with the RMLP forecasts were smaller than those with the MLP forecasts, for all forecast horizons, indicating more accurate forecasts. When the BS for MLP statistics were used as a reference forecast, the improvements in BS due to RMLP (BSS_RMLP) were 19.8%, 12.9% and 10.7%, respectively, for 4-, 8- and 12-month ahead forecasts. Despite the inferior performance exhibited by the MLP model when compared with the RMLP model, the MLP model still demonstrated greater skill than climatology. Further examination of the BS statistics indicated that the degree of

improvements shown by the RMLP over the MLP deteriorated with increased forecast lead time. It should be noted that the analysis in the present study used a threshold value of the 75th percentile of the observed flows. Thus, the results found here may not be representative if a different threshold is used instead. This is in fact the case as the BSS has a tendency to oscillate through wide variations when it is applied to forecasts of rare events (Stanski et al., 1989).

4.3. Ranked probability skill score

The ranked probability score (RPS) is typically used to evaluate the overall forecast performance of probabilistic forecasts (Wilks, 1995). The RPS for a single forecast is the sum of the squared differences of the cumulative distributions of forecasts and observations (Franz et al., 2003):

$$\mathrm{RPS} = \sum_{k=1}^{J}\left(\sum_{i=1}^{k} f_i - \sum_{i=1}^{k} o_i\right)^2, k = 1, \ldots, J \tag{18}$$

where f_i and o_i are the relative frequency of the forecast and the observed traces, respectively, and J is the number of forecast categories. The computation of the RPS is similar to the Brier score except that the Brier score focuses on a single category, while the RPS considers multiple observation and forecast categories to be examined at once (Franz et al., 2003). The RPS is preferable over the Brier score when the prime interest is on the overall forecast quality rather than on a particular forecast category. The technique, discussed by Gangopadhyay et al. (2005), was applied to compute the RPS. Primarily, the observed time series were used to distinguish the possible categories (e.g., 0 to 10, 10 to 20, …, 90 to 100th percentile) for forecasts of streamflows. The cumulative probabilities were computed for each forecast-observation pair in each category. The RPS was then computed as the sum of the squared differences between the observed and the forecasted cumulative probabilities over the whole categories.

Similar to the Brier score, a single value, such as the ranked probability score, may not offer the actual quality of a forecast system. Thus, it is necessary to compare the

forecast of interest to a reference forecast. The ranked probability skill score (RPSS) is used to evaluate the relative skill of forecasts (Gangopadhyay et al., 2005):

$$\mathrm{RPSS} = 1 - \frac{\overline{RPS}_{for}}{\overline{RPS}_{ref}} \tag{19}$$

where $\overline{RPS}_{for}$ is the average RPS of the forecasts for a particular forecast period, and $\overline{RPS}_{ref}$ is the average RPS of the reference forecasts for the same period. The RPSS expresses the degree of improvement over the reference, and ranges from $-\infty$ to 1. A positive score suggests that the forecasts had skill over the reference forecast, with a score of one representing a perfect forecast.

Table 1 presents the RPSS statistics resulting from the long-range streamflow forecasting using the RMLP and the MLP models. The RPSS statistics associated with the RMLP forecasts were much greater than those associated with the MLP forecasts, indicative of more accurate forecasts. When the RPS statistics associated with the MLP model were used as a reference forecast, the improvements in RPS due to the RMLP model (RPSS_RMLP) were 24.9%, 5.7% and 23.5%, respectively, for 4-, 8- and 12-month ahead forecasts. This indicates that the RMLP model has better characterizing capability than the MLP model, for all forecast ranges. It should be noted that although the RPSS for MLP showed inferior performance when compared with the RMLP, the MLP still had greater skill than climatology.

4.4. Discrimination and reliability

The conditional verification measures offer useful information about the performance of forecasts, or forecast probabilities, given a certain event or condition, and are typically examined through forecast reliability and discrimination (Gangopadhyay et al., 2005). These tools provide better insight into the different aspects of the forecast performance at various levels of the historical distribution of streamflows, including low and peak flows (Franz et al., 2003).

For a particular forecast system under consideration, the conditional distribution of forecasts given the observation is expressed as $p(f|o)$. When the value of $p(f|o)$ for all

possible observation categories is equal to zero except for a single observation category then the forecast system is said to be perfectly discriminatory for forecasts of that observation (Murphy and Winkler, 1987; Franz et al., 2003). Figures 1(a)-(c) show the discrimination diagram that displays the conditional probability distributions of each possible flow category as a function of forecast probability. The flow categories examined in the present study were the lowest (0- 33%), middle (33-67%), and highest (67-100%) of the historical distribution and are referred to as low-, middle-, and high flow categories, respectively (Franz et al., 2003). In order to represent the magnitude of probabilities given to each of the three categories, a range of forecast probability categories (i.e., 0-5%, 6-10%, . . . , 95-100%) were considered. The discrimination diagrams in Figures 1(a)-(c) show that the forecasts made by both models were unable to discriminate between low- and middle-flow categories, particularly at lower probabilities for all forecast horizons, the MLP being more problematic. Some difficulties were also observed in discriminating between the middle- and the high-flow categories by the MLP model at higher probabilities, particularly for 4- and 12-month ahead forecasts. In principle, for forecasts that are highly discriminatory, the different flow categories should not overlap on the discrimination diagram (Murphy et al., 1989). Even though both models demonstrated a lack of effectiveness at lower probabilities, the RMLP model showed better discrimination overall.

Forecast reliability, on the other hand, summarizes the information contained in the conditional distribution $[p(o \mid f)]$ and describes how often an observation occurred given a particular forecast (Franz et al., 2003). For a perfect forecast system, the conditional probability of a forecast is expressed as $[p(o = 1 \mid f) = f)]$(Murphy and Winkler, 1987). In other words, for a set of forecast systems for which a forecast probability value, f is given to a particular observation, o, then the forecast is considered to be perfectly reliable when the relative frequency of the observation and the forecast probability is equal (Murphy and Winkler, 1987). Figures 1(a)-(c) show the reliability diagram for all forecast horizons. The flow categories examined in the present study were the highest portion (95-100%) of the historical flow series. Forecast probability

categories (i.e., 0-10%, 10-20%, . . . , 90-100%) were used to represent the magnitude of the probabilities given to the highest portion category. The forecast reliability in Figures 1(a)-(c) is indicated by the proximity of the plotted curve to the 45 degree line, which is representative of a perfectly reliable forecast. The occurrence of the curve below this line signifies overforecasting (probabilities too high), whereas, above this line signifies underforecasting (Stanski et al., 1989). As shown in Figures 1(a)-(c), the reliability diagrams associated with both forecast models were reasonably close to the diagonal line, suggesting reliable forecasts. There were no indications which suggest overforecasts and underforecasts, at any level of probability. Overall, flow forecasts in either case were considered reliable.

4.5. Relative operating characteristics

The relative operating characteristic (ROC) curve is a highly flexible method for evaluating the quality of dichotomous, categorical, continuous, and probabilistic forecasts (Mason and Graham, 1999). The ROC can be represented by plotting the hit rate and the false-alarm rate. Both ratios can be readily computed from the contingency table. The hit and the false-alarm rates are, respectively, defined as the proportion of events for which a warning was provided correctly, and the proportion of nonevents for which a warning was provided incorrectly (Mason and Graham, 1999).

In the present study, the area under the ROC curve was computed as follows. For each forecast probability category (the highest 30% of the historical distribution was considered), the hit and false-alarm rates were computed. Next, the hit rate was plotted against the false-alarm rate. The area under the ROC curve was then integrated to estimate the ROC area. The area under the ROC curve is a useful summary measure of forecast skill. A perfect forecast system has ROC area equal to unity whereas the no skill forecast system has a value of 0.5. The relative performance of a forecast system was evaluated using the ROC skill score (ROCSS = 2 (ROC area -0.5)). Figures 1(a)-(c) and Table 1 indicate that the ROC areas associated with the RMLP forecasts were greater than those with the MLP forecasts, for all forecast horizons. In terms of the ROCSS

statistics, the RMLP showed greater skill over the MLP, for all forecast horizons. The ROC statistics higher than a value of 0.5 suggest that both models had greater skill, compared to climatology. Graphically, the ROC curves which are in the upper-left triangle had positive skills over climatology, for which forecasts that correspond to a point close to the top-left corner represent more accurate forecasts (Mason and Graham, 1999). The diagonal line in the ROC curve represents a climatological forecast (no skill).

4.6. Economic value of forecasts

The potential economic value of streamflow forecasts was further examined with a simple cost-loss analysis technique (Murphy, 1977; Wilks, 2001; Richardson, 2000). The economic value (EV) of a particular forecast system is typically represented graphically, as a function of cost-loss ratio (C/L). Figure 2 illustrates the EV curves associated with the MLP and the RMLP models, for the event exceeding the 70^{th} percentile of the observed flows, for (a) 4-, (b) 8-, and (c) 12-month ahead forecasts. The solid and the dotted lines represent the EV curves associated with the RMLP and the MLP, respectively. The economic value ranges from minus infinity to one, for which one represents a perfect forecast, and zero represents a climatology forecast. EV less than zero suggest that the forecast system has insufficient skill, thus the preferred strategy is to follow climatological information (Richardson, 2000). As shown in Figure 2, both forecast models provided positive and greater benefits in terms of the overall economic value and the range of end-users that can benefit from the respective forecast systems. The potential economic benefits incurred in both cases range between a C/L ratio of 0.05 and 0.85. In comparison, the RMLP had a slight edge over the MLP, particularly for a 12-month ahead forecast. The largest economic benefit (about 80%) was attained when the C/L ratio was equal to the climatologic frequency of an event, which in this case was 0.3 (Richardson, 2000; Wilks, 2001). Similar EV curves are presented in Figure 3, for the event exceeding the 85^{th} percentile of the observed flows. In this case, with the exception of a 4-month ahead forecast, all forecasts associated with the MLP model yielded virtually no economic benefit, and hence the decision maker would prefer climatology.

The RMLP model on the other hand, provided greater benefits for a wide range of users, with C/L values between 0.05 and 0.75 for all forecast horizons considered. Thus the value of the RMLP model is more pronounced when the event for the forecast interest becomes rare.

4.7. Statistical tests

Statistical tests were performed between the observed and the simulated streamflows to determine if there is evidence of difference in the population locations and variances without assuming a parametric model for the distributions. The first statistical test investigated was the Mann-Whitney test (Lehmann, 1975) of the equality of two population medians. The results of the Mann-Whitney tests on the two dataset (i.e. simulated and observed on test data) are presented in Table 2. The values in this table represent *p*-values. A significance level of 5% is chosen for the purpose of comparison. The computed *p*-value greater than the chosen significance level suggest that the simulated and the observed streamflow medians are statistically equal. Table 2 shows that the computed *p*-values associated with the MLP model were less than the chosen significance level of 5% for all forecast horizons, suggesting the observed and the simulated streamflow medians are statistically different. The computed *p*-values associated with the RMLP model were, however, greater than the 5% significance level for all forecast horizons, suggesting the observed and the simulated streamflow medians are statistically equal. The equality of variances between the simulated and the observed streamflows was conducted using the Levene's test (Levene, 1960). The results of the Levene's test in Table 2 show that the reported *p*-value (0.965) associated with the MLP model for the 4-month ahead forecast is greater than the rejection level of significance, indicating the observed and the simulated streamflow variances are statistically equal. For the 8- and 12-month ahead forecasts, the observed and the simulated variances simulated by the MLP model are significantly different (*p*-values less than the rejection level of significance). The reported *p*-values associated with the RMLP model were all greater than the rejection level of significance for all forecast horizons, suggesting the observed

and the simulated variances are statistically equal. The two statistical tests confirm that the RMLP model is superior to the MLP model in terms of reproducing the observed streamflow median and variance.

5. Summary and conclusions

This study presented the findings of the extended Kalman filter approach to perform supervised training of the recurrent multilayer perceptron. The general framework was implemented in western Alberta, Canada, in order to characterize long-range horizon streamflow forecasts. Out of 47 years (Jan 1960 to Dec 2006) of monthly streamflow data, a subset of data, from Jan-1960 to Dec-1993, were used for model calibration, and the remaining data, from Jan-1994 to Dec-2006, were used for model evaluation. Suites of conventional and distribution-based diagnostic measures were employed to evaluate the skill and economic value of the RMLP over the MLP and climatology.

Results from the forecasting experiment showed that the RMLP model yielded better statistics in each forecast horizon, and outperformed the MLP model by a wide margin. Furthermore, the extended Kalman filter algorithm demonstrated stronger learning capability, better convergence properties, and faster training speed than the standard gradient descent algorithm. Some of the major features that demonstrated improved performance include: the ability of the RMLP model to capture spatio-temporal information from the flow series, and the optimal filtering capability of the extended Kalman filter while training the network. While the present study concludes that the RMLP model outperformed the MLP model in all performance measures, further investigation is recommended on different variables and study areas to ascertain the overall skill and versatility of the approach.

Acknowledgements

This research was supported by the School of Graduate Studies at McMaster University. Environment Canada is gratefully acknowledged for providing the experimental data. Minitab Statistical Software is used to conduct statistical tests.

References

Adamowski, J.F., 2008. Development of a short-term river flood forecasting method for snowmelt driven floods based on wavelet and cross-wavelet analysis. J. hydrol., 353, 247– 266.

Amenu, G.G., Markus, M., Kumar, P., and Demissie, M., 2007. Hydrologic Applications of MRAN Algorithm. J. Hydrol. Eng., 12(1), 124-129.

ASCE Task Committee on Application of Artificial Neural Network in Hydrology, 2000a. Artificial neural network in hydrology. I: Preliminary concepts. J. Hydrol. Eng., 5, 115-123.

ASCE Task Committee on Application of Artificial Neural Network in Hydrology, 2000b. Artificial neural network in hydrology. II: Hydrologic applications. J. Hydrol. Eng., 5, 124-137.

Barnston, A.G., Thiao, W., and Kumar, V., 1996. Long-lead forecast of seasonal precipitation in Africa using CCA. Wea. Forecasting, 11, 506-520.

Brier, G. W., 1950. Verification of forecasts expressed in terms of probability. Mon. Wea. Rev., 78, 1–3.

Carcano, E.C., Bartolini, P., Muselli, M., and Piroddi, L., 2008. Jordan recurrent neural network versus IHACRES in modeling daily streamflows. J. Hydr., 362 (3-4), 291-307.

Chang, L.C., Chang, F.J., and Chiang, Y.M., 2004. A two-step-ahead recurrent neural network for stream-flow forecasting. Hydrol. Processes, 18, 81–92.

Choi, J., Yeap, T.H., and Bouchard, M., 2005. Online State–Space Modeling Using Recurrent Multilayer Perceptrons with Unscented Kalman Filter. Neural Processing Letters, 22, 69–84, DOI 10.1007/s11063-005-2157-2.

Dawson, C.W., and Wilby, R.L., 2001. Hydrological modeling using artificial neural networks. Progress in Physical Geography, 25, 80-108.

Franz, K.J., Hartmann, H.H., Sorooshian, S., and Bales, R., 2003. Verification of National Weather Service Ensemble Streamflow Predictions for Water Supply Forecasting in the Colorado River Basin. American Meteorological Society, 4, 1105-1118.

Gangopadhyay, S., Clark, M., and Rajagopalan, B., 2005. Statistical downscaling using K-nearest neighbors. Water Resour. Res., 41 (2), W02024, doi:10.1029/2004 WR 003444.

Giustolisi, O., and Laucelli, D., 2005. Improving generalization of artificial neural networks in rainfall–runoff modelling. Hydrol. Science J., 50(3), 439-457.

Gobena, A.K., and Gan, T.Y., 2010. Incorporation of seasonal climate forecasts in the ensemble streamflow prediction system. J. Hydrol., 385 (1-4), 336-352.

Haykin, S., 2008. Neural Networks and Learning Machines, 3rd edition, Prentice-Hall, Upper Saddle River, N.J.

Hsu, K.L., Gupta, H.V., and Sorooshian, S., 1995. Artificial neural network modeling in rainfall-runoff process. Water Resour. Res., 31 (10), 2517-2530.

Jaeger, H., 2002. A tutorial on training recurrent neural networks, covering BPPT, RTRL, EKF and the "echo state network" approach Fraunhofer Institute for Autonomous Intelligent Systems (AIS). Fraunhofer Institute for Autonomous Intelligent Systems (AIS).

Jonsson, G., and Palsson, O.P., 1994. An application of extended Kalman filtering to heat exchanger models. ASME J. Dyn. Syst., Measurement, Contr., vol. 116, 257–264.

Khan, M.S., and Coulibaly, P., 2006. Bayesian neural network for rainfall-runoff modeling. Water Resour. Res., 42, W07409, doi:10.1029/2005WR003971.

Khan, M.S., and Coulibaly, P., 2006. Bayesian neural network for rainfall-runoff modeling. Water Resour. Res., 42, W07409, doi:10.1029/2005WR003971.

Kisi, O., and Cigizoglu, H.K., 2007. Comparison of different ANN techniques in river flow prediction. Civil Engineering and Environmental Systems, 24(3), 211-231.

Lehmann, E.L., 1975. Nonparametrics: Statistical Methods Based on Ranks, Holden and Day, San Francisco.

Levene, H., 1960. Contributions to Probability and Statistics, Stanford University Press.

Mason, S.J., and Graham, N.E., 1999. Conditional Probabilities, Relative Operating Characteristics, and Relative Operating Levels. American Meteorological Society, 14, 719-725.

Muluye, G.Y., and Coulibaly, P., 2007. Seasonal reservoir inflow forecasting with low frequency climatic indices: a comparison of data-driven methods. Hydrol. Science J., 52(3), 508-522.

Murphy, A.H., 1977. The value of climatological, categorical and probabilistic forecasts in the cost-loss ratio situation. Mon. Weather Rev., 105, 803-816.

Murphy, A.H., and Winkler, R.L., 1987. A general framework for forecast verification. Mon. Wea. Rev., 115, 1330–1338.

Murphy, A.H., Brown, B.G., and Chen, Y., 1989. Diagnostic verification of temperature forecasts. Wea. Forecasting, 4, 485–501.

Palma, L., Gil, P., Henriques, J., Dourado, A., and Duarte-Ramos, H., 2001. Application of an Extended Kalman Filter for On-line Identification with Recurrent Neural Networks. Proc. Of the 7th Jornadas Hispano-Lusas de Ingenieria Electrica, Madrid: Spain.

Pan, T.Y., and Wang, R.Y., 2005. Using recurrent neural networks to construct rainfall-runoff processes. Hydrol. Processes, 19(18),3603-3619.

Piechota, T.C., Chiew, F.H.S., and Dracup, J.A., 1998. Seasonal streamflow forecasting in eastern Australia and the El Nin˜o-Southern Oscillation. Water Resour. Res., 34, 3035– 3044.

Pramanik, N., and Panda, R.K., 2009. Application of neural network and adaptive neuro-fuzzy inference systems for river flow prediction. Hydr. Sc. J., 54(2), 247-260.

Puskorius, G.V., and Feldkamp, L.A., 1994. Neurocontrol of nonlinear Dynamical Systems with Kalman Filter Trained Recurrent Networks. IEEE Transactions on Neural Networks, 5(2), 279-297.

Richardson, D.S., 2000. Skill and economic value of the ECMWF Ensemble Prediction System. Quarterly Journal of the Royal Meteorological Society, 126, 649–668.

Schindler, D.W., Donahue, W.F., and Thompson, J.P., 2007. Running out of Steam? Oil Sands Development and Water Use in the Athabasca River-Watershed: Science and Market based Solutions. Section 1: Future Water Flows and Human Withdrawals in the Athabasca River. Environmental Research and Studies Centre, University of Alberta, Canada.

Singh, V.P., 1995. Computer Models of Watershed Hydrology. Water Resources Publications, Highlands Ranch, CO.

Stanski, H.R., Wilson, L.J., and Burrows, W.R., 1989. Survey of common verification methods in meteorology. World Weather Watch Tech. Rept. No. 8, WMO/TD No.358, WMO, Geneva, 114 pp.

Sveinsson, O.G.B., Lall, U., Fortin, V., Perrault, L., Gaudet, J., Zebiak, S., and Kushnir, Y., 2008. Forecasting spring reservoir inflows in Churchill Falls Basin in Quebec, Canada. J. Hydrol. Eng., 13(6), 426-437.

Tootle, G.A, Singh, A.K, Piechota T.C., and Farnham, I., 2007. Long lead-time forecasting of US streamflow using partial least squares regression. J. Hydrol. Eng., 12 (5), 442-451.

Wilks, D.S., 1995. Statisistical Methods in the Atmospheric Sciences, Academic Press, San Diego, California.

Wilks, D.S., 2001. A skill score based on economic value for probability forecasts. Meteorol. Appl., 8, 209– 219.

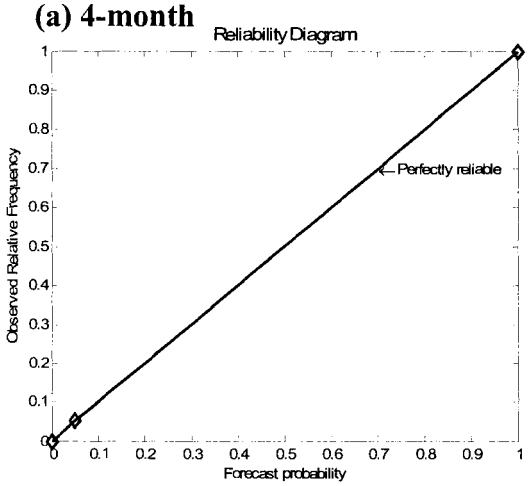
Table 1. Traditional and distribution-based comparative model performance statistics

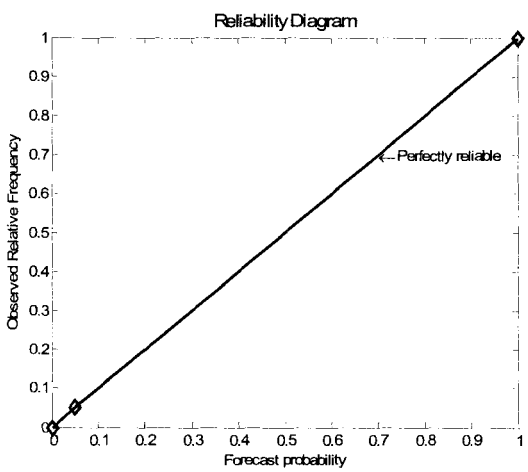
Diagnostic measure	Forecast lead (month)					
	4		8		12	
	MLP	RMLP	MLP	RMLP	MLP	RMLP
Bias (%)	10.6	-0.1	1.9	1.1	-2.2	-0.8
RMSE	220	178	227	182	223	180
r	0.80	0.87	0.79	0.86	0.80	0.87
SS (%)	62.0	75.0	60.0	74.0	61.0	75.0
SS_RMLP (%)	–	**34.5**	–	**35.2**	–	**35.1**
BS	0.13	0.10	0.14	0.12	0.12	0.11
BSS (%)	48.7	59.0	43.6	51.3	51.28	56.4
BSS_RMLP (%)	–	**19.8**	–	**12.9**	–	**10.7**
RPS	1.47	1.10	1.47	1.39	1.72	1.31
RPSS (%)	45.9	59.3	45.6	48.7	36.6	51.5
RPSS_RMLP (%)	–	**24.9**	–	**5.7**	–	**23.5**
ROC	0.865	0.880	0.846	0.875	0.764	0.865
ROCSS	0.731	0.760	0.692	0.750	0.529	0.731

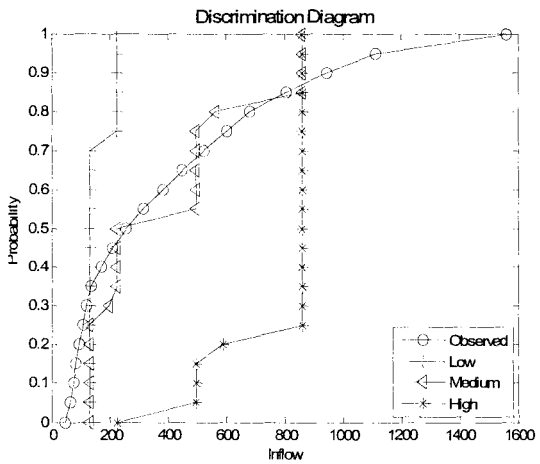
Table 2. Test for position (i.e. median) and variance using Mann-Whitney and Levene, respectively, for the test period Jan 1993 to Dec 2006.

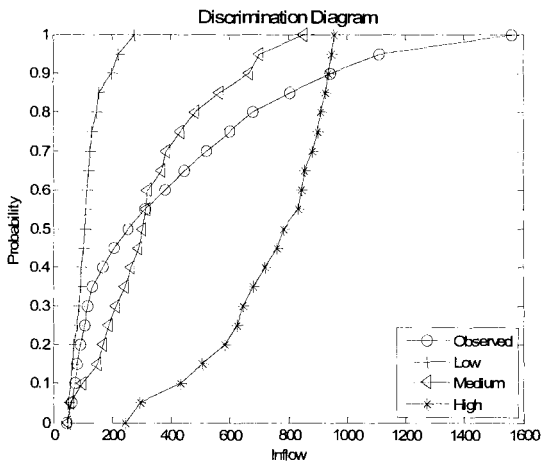
Forecast lead (month)	Test for position (Mann-Whitney)		Test for Variance (Levene)	
	MLP	RMLP	MLP	RMLP
4	0.008	0.500	0.965	0.439
8	0.012	0.866	0.001	0.964
12	0.016	0.263	0.000	0.300

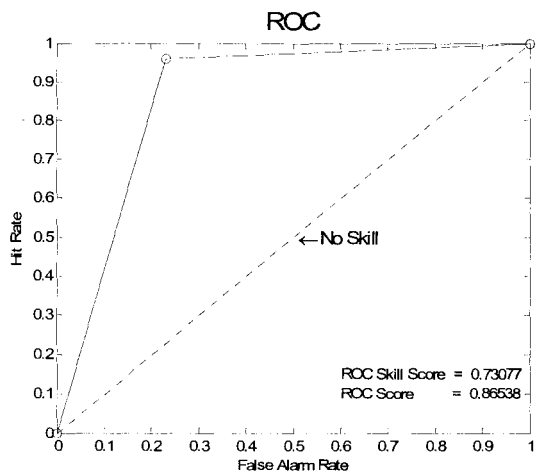
(a) 4-month

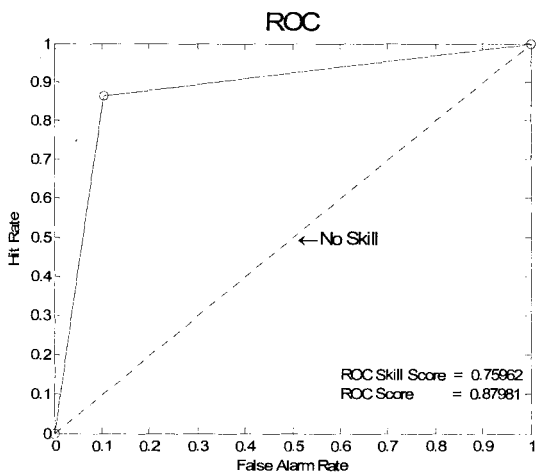

Reliability Diagram
Observed Relative Frequency
Forecast probability
Perfectly reliable


Reliability Diagram
Observed Relative Frequency
Forecast probability
Perfectly reliable

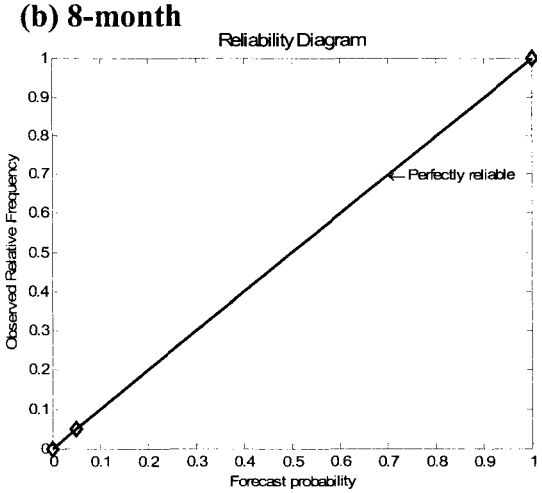

Discrimination Diagram
Probability
Inflow
Observed
Low
Medium
High

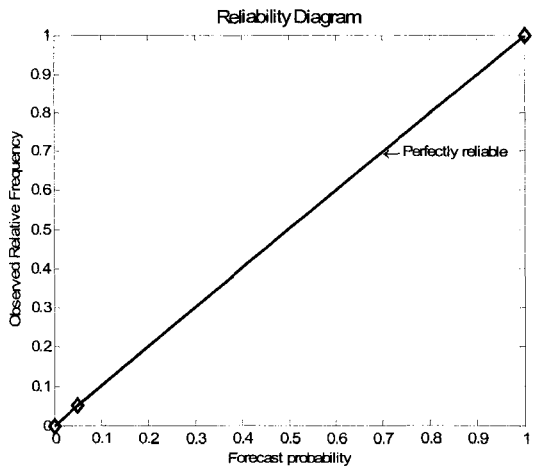

Discrimination Diagram
Probability
Inflow
Observed
Low
Medium
High

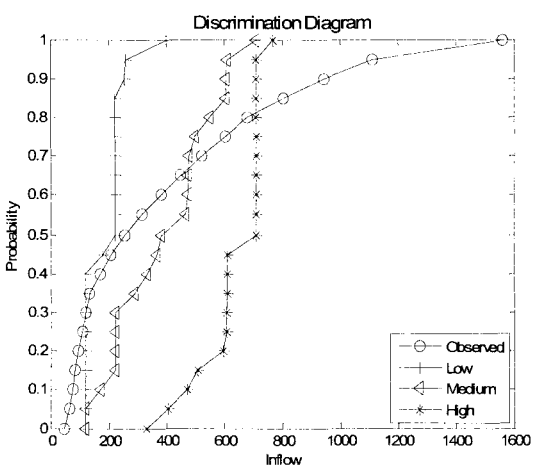

ROC
Hit Rate
False Alarm Rate
No Skill
ROC Skill Score = 0.73077
ROC Score = 0.86538

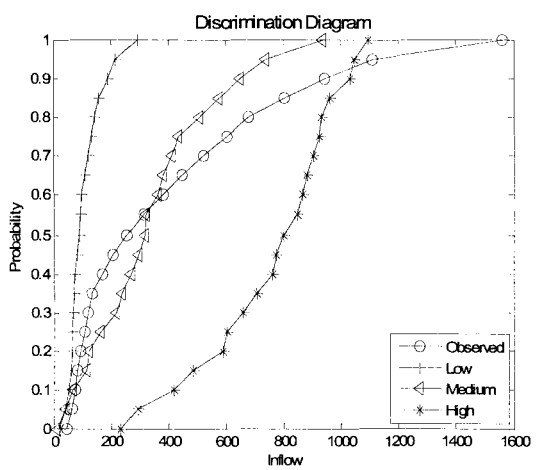

ROC
Hit Rate
False Alarm Rate
No Skill
ROC Skill Score = 0.75962
ROC Score = 0.87981

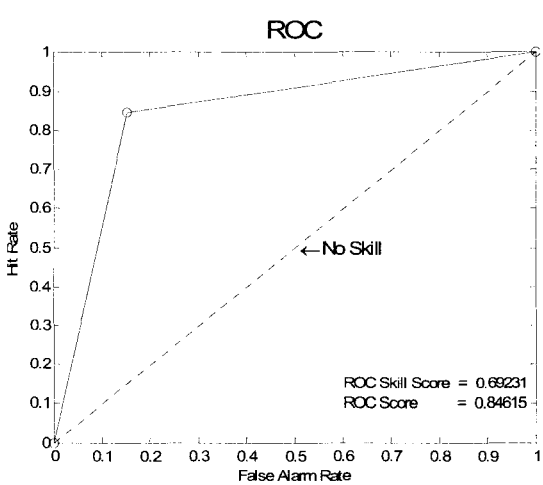
(b) 8-month

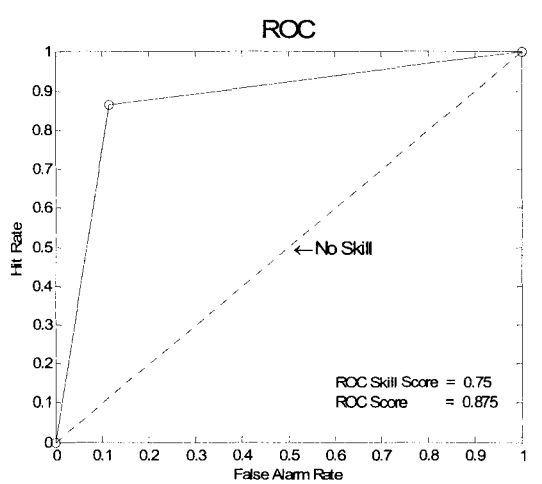

Reliability Diagram
Observed Relative Frequency
Forecast probability
Perfectly reliable


Reliability Diagram
Observed Relative Frequency
Forecast probability
Perfectly reliable

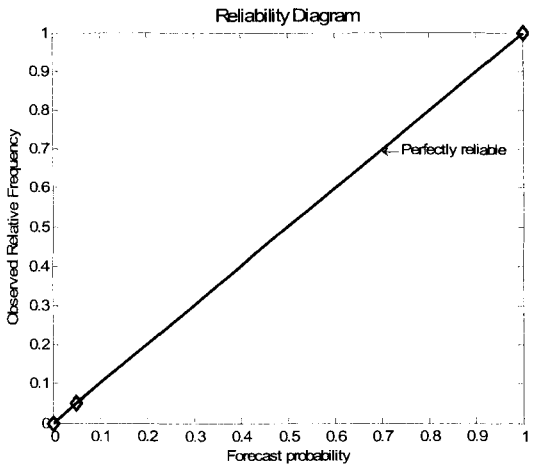

Discrimination Diagram
Probability
Inflow
Observed
Low
Medium
High

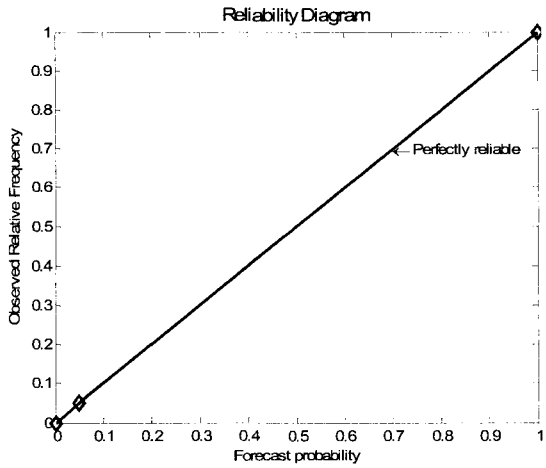

Discrimination Diagram
Probability
Inflow
Observed
Low
Medium
High

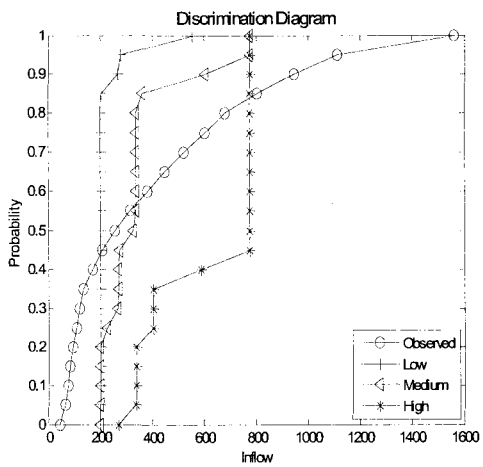

ROC
Hit Rate
False Alarm Rate
No Skill
ROC Skill Score = 0.69231
ROC Score = 0.84615

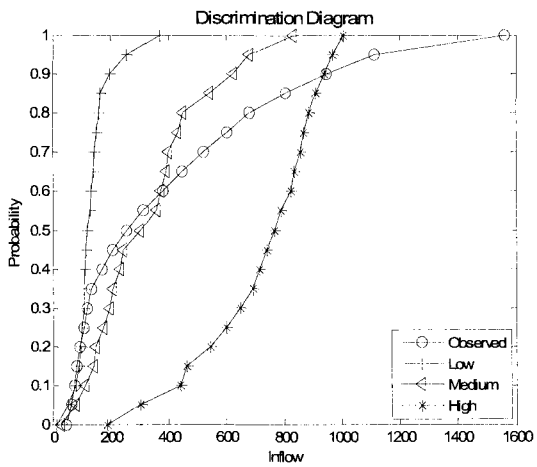

ROC
Hit Rate
False Alarm Rate
No Skill
ROC Skill Score = 0.75
ROC Score = 0.875

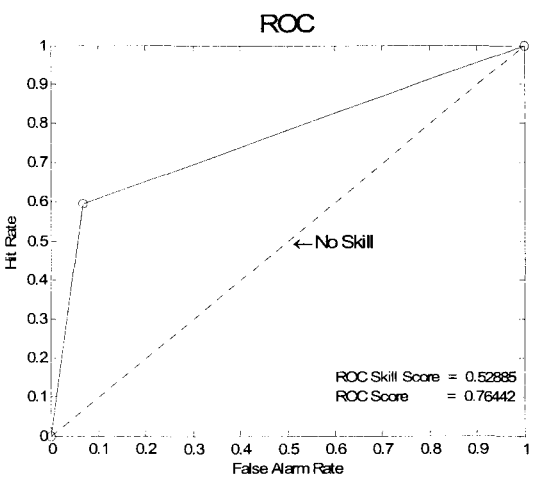
(c) 12-month


Reliability Diagram
Observed Relative Frequency
Forecast probability
Perfectly reliable


Reliability Diagram
Observed Relative Frequency
Forecast probability
Perfectly reliable


Discrimination Diagram
Probability
Inflow
Observed
Low
Medium
High


Discrimination Diagram
Probability
Inflow
Observed
Low
Medium
High


ROC
Hit Rate
False Alarm Rate
No Skill
ROC Skill Score = 0.52885
ROC Score = 0.76442

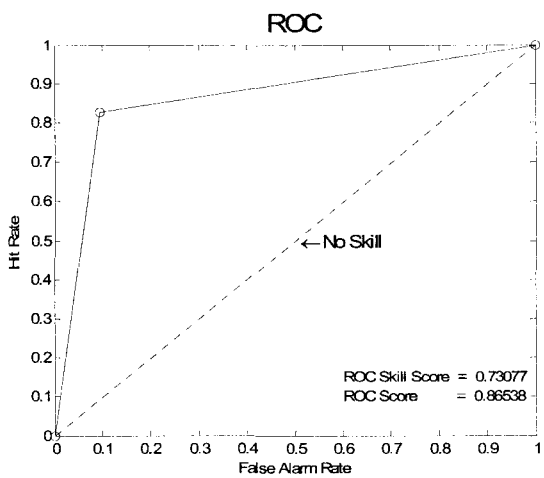

ROC
Hit Rate
False Alarm Rate
No Skill
ROC Skill Score = 0.73077
ROC Score = 0.86538

Figure 1. Distribution-based streamflow forecasting comparative diagrams. The left figures represent the MLP model output and the right figures represent the RMLP model output for a) 4-month ahead forecasts, 1993-2006, b) 8-month ahead forecasts, 1993-2006, and c) 12-month ahead forecasts, 1993-2006

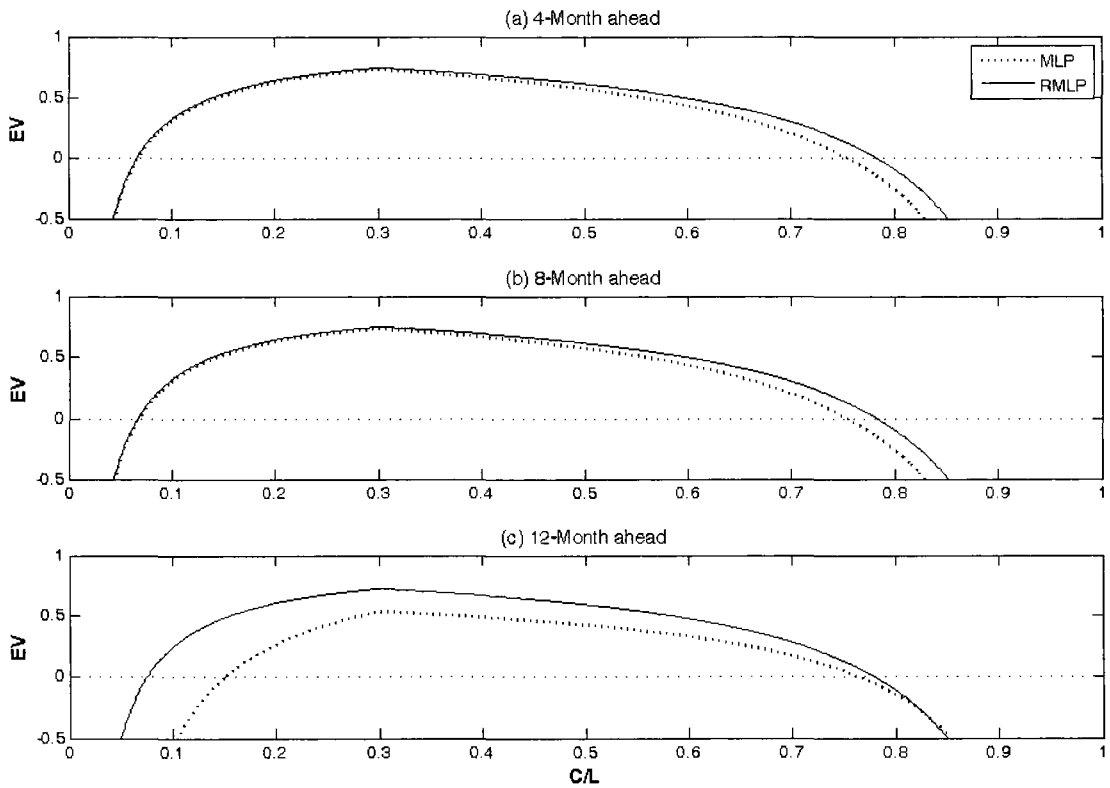


Figure 2. Potential EVs of streamflow forecasts, in excess of the 70th percentile, for (a) 4-month, (b) 8-month, and (c) 12-month ahead, 1993-2006. The solid and the dotted lines represent the EV curves associated with the RMLP and MLP, respectively.

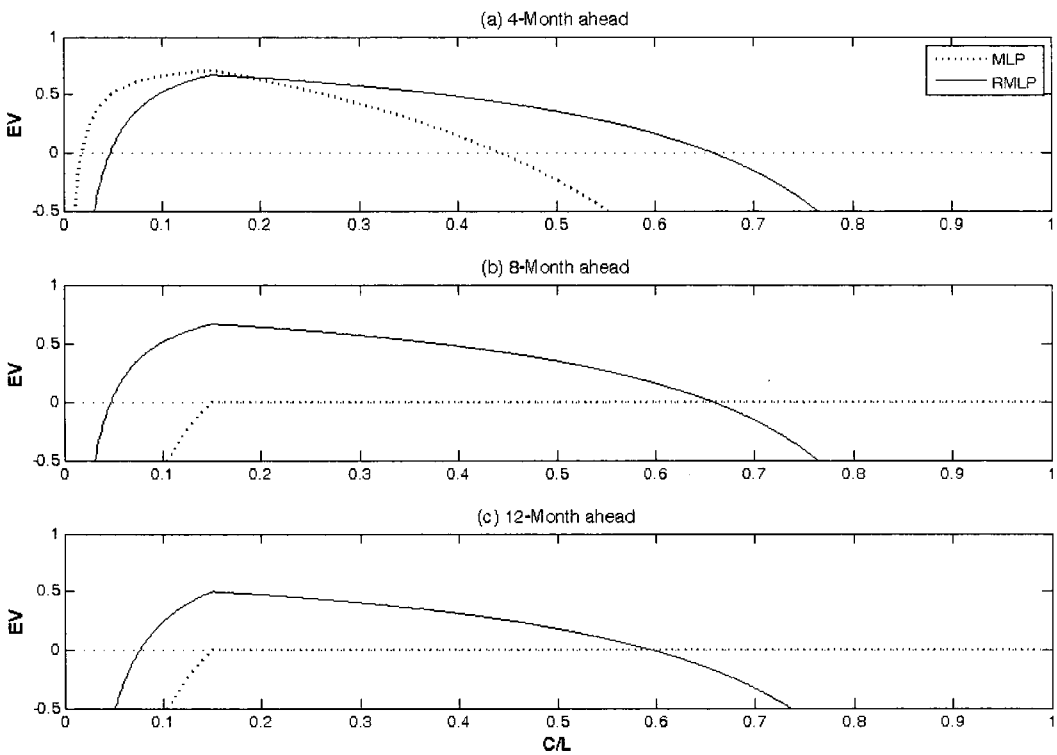


Figure 3. Potential EVs of streamflow forecasts, in excess of the 85th percentile, for (a) 4-month, (b) 8-month, and (c) 12-month ahead, 1993-2006. The solid and the dotted lines represent the EV curves associated with the RMLP and MLP, respectively.

CHAPTER 4

Comparison of statistical methods for downscaling daily precipitation

This broadly based chapter is devoted to downscaling of hydrological variables from numerical weather prediction model output, which is at the core of hydrologic forecast system. This chapter is of profound practical importance. First, it describes the development of the different downscaling models used in the study. Second, it compares the downscaled results and identifies an adequate downscaling model for the study basin, laying down the groundwork for hydrologic forecasting. The findings of the study are presented in a paper form and submitted to Journal of Hydrology for possible publication.

Comparison of statistical methods for downscaling daily precipitation

Abstract

This paper compares different statistical methods for downscaling daily precipitation from numerical weather prediction model output for the Chute-du-Diable weather station located in northeastern Canada. Three different methods were considered: (i) hybrids; (ii) neural networks; and (iii) nearest neighbor-based approaches. Various diagnostic measures were employed to evaluate and inter-compare the downscaling results. Although results of the downscaling experiment showed mixed performances, clear patterns emerged with respect to the reproduction of variation in daily precipitation and skill. Artificial neural network-logistic regression (ANN-Logst), partial least squares (PLS) regression and recurrent multilayer perceptron (RMLP) models showed greater skill values, and conditional resampling method (SDSM) and *K*-nearest neighbor (KNN)-based models showed the potential to characterize the variability in daily precipitation.

Keywords: Chute-du-Diable; nearest neighbor; partial least squares; precipitation; statistical downscaling

1. Introduction

The accuracy of a river forecast system largely depends on the quality of hydrologic inputs. The lack of reliable precipitation forecasts is one of the many challenges confronting hydrologists, particularly for use in short- to medium- range hydrological forecasts. Recent studies have shown that the use of information from meteorological forecasts and climate outlooks can help provide improved hydrological forecasts [e.g., *Schaake et al.*, 2006; *Roulin*, 2007; *Shi et al.*, 2008; *Li et al.*, 2009]. *Clark and Hay* [2004] investigated atmospheric forecasts over the contiguous United States from the National Centers for Environmental Prediction (NCEP) reanalysis project to assess the possibilities for using Medium-Range Forecast (MRF) model output for

prediction of streamflow. The *Clark and Hay* [2004] findings indicated that forcing MRF output to a distributed-hydrologic model can yield useful results.

Large-scale model outputs possess substantial skill at global and regional scales [e.g., *Harpham and Wilby*, 2005; *Schmidli et al.*, 2007]. Nevertheless, their usefulness for local hydrological studies are greatly limited owing to coarse spatial resolution as well as an inability to resolve important sub-grid scale features such as clouds and topography [*Wilby et al.* 2002]. Bridging the gap between the resolution of large-scale model output and local-scale hydrological processes signifies a considerable difficulty for effective assessment of grid and sub-grid scale hydrological responses [*Fowler et al.*, 2007].

There exist two broad approaches to translate information from large-scale model outputs to the local-scale. One way to achieve this is through dynamical downscaling, in which a regional climate model (RCM) uses large-scale model output as initial and lateral boundary conditions for much more spatially detailed climatological simulations over a region of interest [*Hay et al.*, 2002]. The skill of such regional models in deriving local-scale variables from large-scale model output has been successfully demonstrated in various regions [e.g., *Mearns et al.*, 1999; *Hay and Clark*, 2003; *Schmidli et al.*, 2007; *Spak et al.*, 2007]. Although dynamical downscaling techniques are based on strong physical realism, they have the following limitations [*Mearns et al.*, 1999]: (1) The modeling is computationally expensive; (2) the control run simulations at regional scale still suffer from inaccuracies; (3) the quality of the regional model control run is dependent on the quality of the global model run providing the boundary conditions; and (4) nested models usually require much tuning of parameterizations when applied to new regions. An alternative approach to dynamical downscaling is to use statistical methods, in which an empirical relationship is established between large-scale model output and local-scale variables [e.g., *von Storch et al.*, 1993; *Wilby et al.*, 2002; *Hay and Clark*, 2003; *Harpham and Wilby*, 2005]. *Wilby and Wigley* [1997] further classify statistical downscaling techniques into regression methods, weather pattern-based approaches and stochastic weather generators. The general theory, applications, advantages and

shortcomings of common downscaling methods are described in the literature [*Wilby and Wigley*, 1997; *Xu*, 1999; *Yarnal et al.*, 2001; *Fowler et al.*, 2007].

Gangopadhyay et al. [2005] used a *K*-nearest neighbor model for downscaling daily precipitation and temperature in the contiguous United States from the NCEP output. The reported ranked probability skill score (RPSS) values were approximately 10% and 30% for downscaling daily precipitation and temperature for a seven day forecast, respectively. *Hay and Clark* [2003] downscaled output from the NCEP using a regional climate model (RCM) in three snowmelt-dominated basins in the western United States. The three basins considered in the *Hay and Clark* [2003] study were: (i) Animas River at Durango, Colorado (Animas); (ii) East Fork of the Carson River near Gardnerville, Nevada (Carson); and (iii) Cle Elum River near Roslyn, Washington (Cle Elum). The Nash-Sutcliffe efficiency coefficient (CE) values reported in the *Hay and Clark* [2003] study were 0.3, 0.4 and 0.35 for downscaling daily precipitation, 0.65, 0.6 and 0.7 for downscaling minimum daily temperature, and 0.72, 0.71 and 0.7 for downscaling maximum daily temperature for the Animas, Carson and Cle Elun basins, respectively. *Wilby et al.* [2000] simulated daily rainfall and surface temperatures for the Animas River basin, Colorado using dynamically and statistically downscaled output from the NCEP-National Centers for Atmospheric Research (NCAR) reanalysis. The reported Pearson correlation coefficient (r) values were: 0.5 and 0.42 for the downscaled precipitation, 0.82 and 0.93 for the downscaled minimum temperature, and 0.83 and 0.88 for the downscaled maximum temperature, using the dynamical and statistical downscaling models, respectively.

Statistical downscaling methods are becoming increasingly popular due to their (1) simplicity in design and implementation, (2) computational efficiency, and (3) comparable or superior performance when compared with dynamical downscaling. Given a pool of statistical downscaling models available in the scientific literature, a limited number of articles are devoted to rigorous inter-comparison of downscaling hydrological variables [e.g., *Schoof and Pryor*, 2001; *Mehrotra and Sharma*, 2005; *Dibike and Coulibaly*, 2005; *Wilby and Harris*, 2006; *Wetterhall et al.*, 2006; *Liu et al.*, 2007].

Furthermore, even fewer studies have reported downscaling of station daily precipitation from the NCEP MRF numerical weather prediction model output. Another important consideration is the demonstrated failure of current downscaling methods to effectively describe precipitation characteristics. Outputs from these downscaling models generally yield modest or poor performance when forced into a distributed-hydrologic model [*Clark and Hay*, 2004]. This clearly underscores the need for novel and innovative downscaling methods, particularly for daily precipitation. Therefore, the main objectives of this paper are: (i) to develop promising statistical methods for downscaling daily station precipitation from the NCEP MRF numerical weather prediction model output for the Chute-du-Diable weather station in northern Quebec, Canada; (ii) to assess the skill of these downscaling models for long-lead time forecasts (i.e., 1 day to 14 day); and (iii) to compare the newly developed models with the commonly used statistical downscaling models. The scope of model comparison in the present study is limited to statistical downscaling models, and as such, model comparison with dynamical downscaling models will not be performed as RCM results are not available for the study basin. The commonly used models, such as the Statistical DownScaling Model (SDSM), are used as benchmark for comparison. The selection of benchmark models are based on preliminary analysis and literature review.

The remainder of the material in this chapter is organized as follows. Section 2 provides a brief description of statistical downscaling techniques used in the downscaling experiment. Experimental setup and calibration of the downscaling methods are presented in section 3. Section 4 compares the performance of the developed methods with that of the commonly used methods and discusses the results of the downscaling experiment. Finally, conclusions and recommendations are given in section 5 based on results of the downscaling exercise.

3. Model Description

This section presents a brief description of the statistical downscaling techniques selected for evaluation and inter-comparison of downscaling daily precipitation fields.

The methods considered are: (i) conditional resampling method, (ii) partial least squares regression, (iii) hybrids, (iv) nearest neighbor-based models, and (v) a family of artificial neural networks.

2.1. Conditional resampling method

The most common regression-based technique used to map global climate models to individual sites or localities is the Statistical DownScaling Model (*Wilby et al.*, 2002). The SDSM is best described as a hybrid of the stochastic weather generator and regression-based method. In this model, local-scale weather generator parameters, such as daily precipitation occurrence and intensity are linearly conditioned on the basis of large scale circulation patterns and atmospheric moisture variables. The variance of the downscaled daily time series is synthetically inflated by means of stochastic methods to reasonably represent the observed time series. The downscaling algorithm of SDSM has been extensively applied to a wide spectrum of meteorological, hydrological and environmental assessments across the world including Africa, Europe, North America and Asia [*Wilby and Dawson*, 2007]. Technical details of the SDSM are found in *Wilby and Dawson* [2007].

2.2. Partial least squares regression

Multiple linear regression (MLR) is a common statistical method used to link large-scale model output to station-scale variables [*Clark et al.*, 2004; *Harpham and Wilby*, 2005; *Liu et al.*, 2007]. The major limitation of multiple linear regression is that when the predictor variables are not independent and are collinear, the model predictions can be poor. Because of this collinearity issues, principal component regression (PCR) is preferred over the MLR method. Partial least squares (PLS) regression, a technique that generalizes and combines good features of the principal component analysis and the multiple linear regression method [*Wold et al.*, 1987; *Lorber et al.*, 1987], is becoming popular. PLS regression attempts to identify factors (called latent variables) that maximize the amount of variation explained in **X** (predictors) that are relevant for

predicting **Y** (predictands). This is an advantage when compared to the PCR regression where the factors (called Principal Components) are selected solely based on the amount of variation that they explain in **X** [*Wold et al.*, 1987]. As a result, the PLS regression is considered a better alternative to both the multiple linear regression and the PCR methods in terms of producing improved forecast skill.

There are several ways of computing the PLS model parameters. The most widely used algorithms are [*de Jong,* 1993]: NIPALS (Non-Iterative Partial Least Squares) and SIMPLS. The latter algorithm was used in the present study. A detailed description and working procedure of SIMPLS can be found in *de Jong* [1993].

2.3. Hybrid models

Statistical models such as multiple linear regression and artificial neural networks (ANN) generally have a tendency to overestimate precipitation occurrences and underestimate precipitation amounts [*Clark et al.*, 2004]. It is not unusual to acquire negative precipitation forecasts from such models. To address this serious weakness, hybrid models are proposed in this study. In hybrid models, precipitation is modeled in a two stage process: Logistic regression is used to identify the occurrence of wet days, and PLS or ANN is used to model the amount of precipitation. A similar two stage procedure has been employed for modeling with SDSM; however, the procedure is different from the approach used in this study [e.g. *Wilby et al.*, 2002].

The intermittent and skewed nature of daily precipitation data requires some preprocessing prior to developing the downscaling models. In order to model the occurrence of wet days, the site precipitation time series $\{d_1, d_2, \ldots, d_n\}$ is converted into a binary time series $\{y_1, y_2, \ldots, y_n\}$, with 0 representing dry days and 1 representing wet days. Given a data set $\{(x_1^t, x_2^t, \ldots, x_m^t)/t = 1,2,\ldots,n\}$ of m large-scale predictor variables x_i, $i = 1, 2, \ldots, m$, the probability of occurrence of wet days $\{\hat{p}_t / t = 1,2,\ldots, n\}$ can be modeled by performing the logistic regression:

$$\hat{p}_t = \frac{1}{1+\exp(-(\alpha + \sum_{i=1}^{m} \beta_i x_i^t))} \tag{1}$$

where α and β are model parameters of the logistic regression. The time series $\{\hat{p}_t / t = 1,2,...,n\}$ is further processed and transformed into a binary time series $\{\hat{y}_t / t = 1,2,...,n\}$ using the relation:

$$\hat{y}_t = \begin{cases} 1, & if \ \hat{p}_t - p \geq 0; \\ 0, & otherwise \end{cases} \tag{2}$$

The parameter p is a threshold of probability according to which $\{\hat{p}_t / t = 1,2,...,n\}$ is transformed into binary time series, $\{\hat{y}_t / t = 1,2,...,n\}$, and this transformation makes mismatches between $\{y_t / t = 1,2,...,n\}$ and $\{\hat{y}_t / t = 1,2,...,n\}$ minimal [*Hua and Zhang*, 2006]. The value of p can be set at some fixed number or a random number from a uniform distribution ranging between 0 and 1 [e.g., *Clark et al.*, 2004]. In the present study, threshold values from 0.1 to 0.9 were considered and the value of p = 0.5 provided the best result. The result was inter-compared against p values assigned from a uniformly generated random number, and the threshold value of p = 0.5 was consistently superior. Intuitively, when the probability of precipitation occurrence is over 50%, then there is a greater chance that precipitation will occur, and vise versa. Thus, the selected p value makes practical sense.

Once wet and dry days are identified using logistic regression, any kind of regression or neural network model can be used to model precipitation amounts. Two hybrid models were developed: (i) PLS-Logst – in which the logistic regression was used to model precipitation occurrence and the PLS was used to develop models for precipitation amounts; and (ii) ANN-Logst – in which the logistic regression was used to model precipitation occurrence and the conventional multilayer perceptron (simply ANN) was used to develop models for precipitation amounts. It should be noted that it is possible to develop several hybrid models from such coupling.

2.4. *K*-nearest neighbors

The nearest neighbor approaches work on the principle of classic bootstrapping techniques [e.g., *Yakowitz*, 1993; *Yates et al.*, 2003; *Gangodhyay et al.*, 2005]. The *K*-

nearest neighbor (KNN) algorithm searches for analogs of a feature vector on the basis of similarity criteria in the observed time series. Observed station precipitation is considered as analog days to be selected on the basis of large-scale numerical model output [*Gangodhyay et al.*, 2005]. KNN-based models have been successfully applied to generating synthetic weather data [e.g., *Rajagopalan and Lall*, 1999; *Buishand and Brandsma*, 2001; *Yates et al.*, 2003; *Sharif and Burn*, 2006; *Mehrotra and Sharma*, 2006].

Among others, one of the key issues that dictate effectiveness of nearest neighbor-based models is the type of distance measure or nearness used in the modeling. The problem is more apparent particularly when the feature vector consists of multiple variables. The measure of closeness is typically quantified either using the Euclidean or Mahalanobis distance. The Mahalanobis distance formulation has the following advantages over the traditional resampling models which utilize the Euclidean distance [*Mehrotra et al.*, 2004]: (i) the use of Mahalanobis distance obviates the standardizing of the predictor variables; and (ii) the Mahalanobis distance measure considers the existing dependence amongst the predictor variables. The Euclidean distance formulation discussed in *Gangopadhyay et al.* [2005] overcomes the major limitation in (ii) by considering less significant, correlated or redundant predictors through the use of principal component analysis (PCA), and the subsequent PCs are then weighted according to their eigenvalues. But both distance metrics ignore the dependence between the predictors and predictands. Because of their inherent strengths and weaknesses, both distance metrics are considered in the present study. A more detailed description of the model is presented in *Muluye* [2010]. The present study employs methods as discussed in *Gangopadhyay et al.* [2005] and *Yates et al.* [2003], respectively, for modeling with Euclidian and Mahalanobis distance formulation. Thus, the two KNN-based models considered are: (i) KNN model based on Mahalanobis distance (KNN-M), and (ii) KNN model based on the Euclidean distance (KNN-E).

2.5. Artificial neural networks

The utility of artificial neural networks in various hydrology-related areas, including precipitation forecasting, has been extensively studied. There exists diverse families of neural networks in the scientific literature, among which, feedforward neural networks or multilayer perceptron (MLP) is the most widely used architecture [e.g., *ASCE*, 2000; *Haykin*, 2008]. Pure MLP along with time-lagged feedforward network architectures have been applied to represent the spatial and temporal information of dynamic systems. But it has been shown that a neural network containing a state feedback possesses more computational advantages and uses fewer numbers of parameters than input-output models [*Palma et al.*, 2001]. One way of effective characterization of a state feedback is through the use of a recurrent multilayer perceptron (RMLP). The application of feedback enables the RMLP to acquire state representations, which makes the architecture suitable to describe a large class of nonlinear dynamic systems [*Choi et al.*, 2005].

The two most widely used algorithms to perform supervised training of the recurrent neural networks are back-propagation through time and real-time recurrent learning. Both algorithms are based on the gradient method which utilizes first-order derivative information. The real-time recurrent learning algorithm that uses continuous learning based on gradient descent has the following drawbacks [*Haykin*, 2008]: (i) it is typically slow due to its reliance on instantaneous estimates of gradients, compared with a learning algorithm that uses second-order derivative information; and (ii) it suffers from the vanishing gradient problem. These serious weaknesses can be overcome through the use of second-order training algorithms such as a linearized recursive least squares or an extended Kalman filter. The algorithms applied in the present study to perform supervised training of the MLP and the RMLP models are, respectively, the classical back-propagation algorithm and the decoupled version of the extended Kalman filter. For detailed treatment of the subject, the reader is referred to one of many sources [e.g., *Puskorius and Feldkamp*, 1994; *Haykin*, 2008].

3. Experimental setup and data

3.1. Study area and data

The study area selected for the evaluation and investigation of downscaling models was the Saguenay-Lac-Saint-Jean (SLSJ) hydrologic system in northern Quebec, Canada (Figure 1). The drainage basin covers an area of approximately 73,800 square kilometers extending roughly between 47.3° to 52.2° N and 70.5° to 74.3° W. Alcan Company operates and maintains a large dataset from 13 meteorological and 11 hydrometric stations primarily for the purpose of hydroelectric power generation (Table 1). The Chute-du-Diable meteorological station, located at 48.75° N and 71.7° W was considered for the present study. Station daily precipitation and minimum and maximum air temperature data were collected from Alcan hydro-meteorological database from 1979 through 2001.

All atmospheric predictor variables used for the downscaling experiment were a "reforecast" data set generated by the Climate Diagnostics Center [e.g., *Hamill et al.*, 2004]. A T62 resolution version of the NCEP MRF model was used to generate an ensemble of 15-day forecasts over a 23-year period from 1979 to 2001. The NCEP MRF model output ensembles are defined on a global lat-lon grid with 2.5° resolutions both in longitude and latitude (144×73 grid points). The data used in this analysis consists of daily values from a 23-year period from January 1979 through December 2001. Large-scale model outputs corresponding to the first ensemble member (i.e. out of 15) and all 15-forecast ranges were considered (Table 2). The large-scale model outputs were retrieved from the nearest grid to the Chute-du-Diable meteorological station.

3.2. Application and calibration

Two data types were available for the present study: (i) local-scale predictands such as daily precipitation as well as minimum and maximum temperature, and (ii) eight large-scale model output predictor variables (Table 2). These data sets were further divided into two parts: the first half of the data (1979-1996) were used to estimate the

statistical parameters of the models considered and the remaining data (1997-2001) were used to evaluate the models. The experimental data were prepared as follows. The numerical model output was processed to form a data matrix consisting of 8401 rows (corresponding to the number of days from 1 January 1979 to 31 December 2001) and 8 variable columns (corresponding to each of 15 ensemble members and 15 forecast ranges or lead times). The present study considered a data matrix corresponding to the first member and all the 15-forecast ranges [*Muluye*, 2010].

The first model investigated in the downscaling experiment was the SDSM. A multiple linear regression model was developed between large-scale predictor variables (Table 2) and the local-scale predictand (precipitation). The parameters of the regression model were estimated using ordinary least squares. Precipitation was modeled as a conditional process, in which, large-scale circulation patterns and atmospheric moisture variables were used to linearly condition local-scale weather generator parameters such as precipitation occurrence and intensity. The downscaled daily time series was then adjusted for its mean and variance via bias correction and variance inflation factors, respectively, to better accord with observations. For downscaling experiments with a conditional resampling method, a Statistical DownScaling Model developed by *Wilby and Dawson* [2007] was used.

For the case of modeling with partial least squares regression, the SIMPLS algorithm discussed in *de Jong* [1993] was used. The performance of the model was studied over a range of latent variables (factors) that should be retained in order to maximize the amount of variation explained in large-scale predictors in relevant for predicting station precipitation. Five latent variables provided adequate results in the present study. Similarly, the performances of hybrids in downscaling station precipitation were studied. The sensitivity of models to various choices of threshold probability (p) for the occurrence of wet and dry days were analyzed, and the value of p=0.5 was found to offer adequate results for both PLS-Logst and ANN-Logst models.

The other models considered in the present study were K-nearest neighbors. For modeling with KNN, days similar to each of the 8401 days in the archive were identified

using the KNN algorithm as discussed in *Gangopadhyay et al.* [2005] and *Yates et al.* [2003]. The parameters associated with the KNN model were also examined. Two model parameters were considered: moving window (w), and optimal number of nearest neighbors (K). After analyzing the sensitivity of the KNN model to different choices of width of moving window, a value of $w = 14$ days was chosen for use in the downscaling experiment. Similarly, an analysis was performed to find out the optimal value of K, and a value of $K = 19$ was found to offer adequate results in the present study [*Muluye*, 2010].

To perform modeling using multilayer perceptron, appropriate network architecture was designed and the various parameters were optimized. The search for an optimal network begins with a simple network having one hidden layer, keeping the number of hidden neurons between 2 and 30. The different learning rules as well as transfer functions were investigated in both the hidden layer and output layer in the search for the optimal network [*Muluye and Coulibaly*, 2007]. For the case of modeling with the MLP, one hidden layer with 14 neurons provided the best performing network. Hyperbolic tangent and linear activation function in the hidden layer and output layer, respectively, and the learning rule with a conjugate gradient yielded the optimal network. Similarly, for the case of modeling with the recurrent multilayer perceptron, one hidden layer with 12 neurons provided the best performing network. In the method presented here, the MLP model was trained with the conventional back-propagation algorithm, whereas the RMLP model was trained with the back-propagation through time algorithm via the decoupled version of an extended Kalman filter.

4. Comparison of downscaling results

4.1. Diagnostic measures

Diagnostic measures use logistical metrics in order to evaluate the quality of the forecast system. Two types of diagnostic tools were used: generic and categorical. Wet- and dry-spell length, quantile-quantile (q-q) plots of observed and downscaled daily precipitation amounts, and statistical tests were also used. The general statistics employed here provided key statistics such as bias (%), root mean squared error (RMSE), Pearson

correlation coefficient (r) and reduction of variance (RV). On the other hand, categorical statistics provided the skill of downscaling models on one of two key features of precipitation (i.e. occurrence) using frequency bias (BiasF), probability of detection (POD) and false alarm ratio (FAR). The contingency table was used to determine the different types of errors made in the downscaled precipitation. A perfect forecast system produces only hits and correct rejections, and no misses or false alarms [*Stanski et al.*, 1989]. Several categorical statistics can be computed from the elements in the contingency table. In this paper, only the major score statistics are computed and discussed. All metrics were evaluated with a 0.3 mm threshold for dry and wet day occurrences [e.g., *Wilby and Harris*, 2006].

The bias score (frequency bias) explains how the forecast frequency of "yes" events compared to the observed frequency of "yes" events. The range of the score is between zero and infinity, for which a score of one represents a perfect forecast. Basically, the bias score signifies whether the forecast system has a tendency to under-represent (BiasF<1) or over-represent (BiasF>1) occurrences but does not quantify how well the forecasts correspond to observations; i.e., BiasF only measures relative frequencies [*Stanski et al.*, 1989]. The Bias score is given by [*Wilks*, 1995]:

$$\text{BiasF} = \frac{(N_{11} + N_{10})}{(N_{11} + N_{01})} \tag{3}$$

where N_{11} is the number of correct wet-days, N_{10} is modelled wet- and observed dry-days, and N_{01} is modelled dry- and observed wet-days.

The probability of detection explains the fraction of the observed "yes" events with correct forecasts. The range of the score is between 0 and one, with a score of one representing a perfect forecast. The POD is sensitive to hits, but ignores false-alarms. For this reason, POD is usually used in conjunction with the false-alarm ratio [*Stanski et al.*, 1989]. The score is given by [*Wilks*, 1995]:

$$\text{POD} = \frac{(N_{11})}{(N_{11} + N_{01})} \tag{4}$$

Conversely, the FAR explains the fraction of the predicted "yes" events which did not occur. The range of the score is between 0 and one, with a score of zero representing a perfect forecast. The FAR is sensitive to false alarms, but ignores hits, and is given by [*Wilks*, 1995]:

$$\mathrm{FAR} = \frac{(N_{10})}{(N_{11} + N_{10})} \tag{5}$$

The downscaled precipitation was further assessed using a deterministic skill score (SS). The SS represents the improvement in the downscaled precipitation over some reference forecast. Climatology was employed as the reference forecast in this work. The mean squared error (MSE) was used as a metric score to construct the skill score. When the MSE is used as a score in the SS computation, the resulting statistic is called the reduction of variance (RV) and is given by [*Stanski et al.*, 1989]:

$$\mathrm{RV} = 1 - \frac{\mathrm{MSE}_{\mathrm{forecast}}}{\mathrm{MSE}_{\mathrm{reference}}} \tag{6}$$

An RV of zero indicates no improvement over the reference forecast, one indicates a perfect forecast, and a negative value indicates that the reference forecast is better than a forecast. A more detailed description of model performance statistics is provided by *Stanski et al.* [1989] and *Wilks* [1995].

4.2. Discussion of results

Table 3 shows the error metrics associated with the various models in downscaling daily precipitation (Prec). In most of the cases, the downscaling models tended to under-represent (negative bias) the mean precipitation amounts. The smallest mean bias (-6.18%) over the 15 forecast ranges was simulated by the RMLP model. The ANN-Logst and PLS-Logst models under-represented the mean precipitation amounts during the first five forecast ranges, and over-represented significantly thereafter. The Pearson correlation coefficients between the downscaled and the observed daily precipitation indicated that the RMLP had an advantage over the other downscaling techniques in the vast majority of cases although the ANN-Logst showed competency

beyond the five day forecast. This observation was further supported by the overall RMSE and RV statistics. On the other hand, while the SDSM model performed poorly in terms of these statistics, the KNN-E and KNN-M models provided the worst statistics. In comparison, the KNN-M performed relatively better, with smaller RMSE and greater Pearson correlation coefficient values than the KNN-E, in spite of relatively larger biases. Note that the bold numeric statistics in Tables 3 represent the best statistics associated with each forecast range.

Figure 4 presents RV plots of the various downscaling models against forecast ranges in downscaling daily precipitation. The comparative RV plots show that the skill of precipitation forecasts was generally poor and inadequate. In particular, the SDSM, KNN-M and KNN-E models performed poorly when compared with the other competing downscaling models. Further analysis of Figure 4 indicates that there is a decreasing trend in skill with forecast lead time. This clearly demonstrates the great difficulties associated with the long-range precipitation downscaling. In general, the RV statistics associated with the RMLP, ANN-Logst and PLS models yielded relatively better skills for all forecast ranges.

Table 3 shows the frequency biases simulated by the various models. The biases of wet periods simulated by the SDSM, KNN-E and KNN-M models were reasonably accurate (BiasF close to one). The PLS, MLP and RMLP models consistently over-represented the wet periods, by more than 90% on average over the 15 forecast ranges. It is interesting to note that the coupled models (PLS-Logst and ANN-Logst) appeared to perform better than their counterparts (PLS and MLP) in terms of representing the occurrence of wet periods. On average, only about 13% of the wet periods were over-represented by the coupled models, over the 15 forecast ranges. Model performance was further evaluated through probability of detection (hit rate) and false-alarm ratio. The ability of the RMLP, MLP and PLS models to capture the POD statistics were generally adequate (POD close to unity), however about 50% of the cases were incorrectly forecasted (Table 3). In comparison, the smallest false-alarm ratio (on average 42%) was

attained by the PLS-Logst and the ANN-Logst models although some difficulties were clearly noticed in simulating the hit rates (on average 65%).

While the performances of the RMLP, MLP and PLS models were reasonably adequate in representing hit rates, very poor performances were observed in representing frequency of wet days. This can be explained in part by the technique used in modeling the precipitation process. In the case of modeling with the SDSM, PLS-Logst and ANN-Logst, precipitation was modeled as a conditional process, in which, local precipitation amounts were correlated with the occurrences of wet-days and, which in turn, were correlated with large-scale atmospheric predictors; and the KNN was modeled on the principle of random sampling techniques. Such methods enabled these models to preserve the stochasticity of daily precipitation, thus permitting the occurrence of both dry and wet days. Conversely, the MLP, RMLP and PLS models performed poorly in this regard, due to direct linkage or conditioning of large-scale atmospheric predictors with precipitation amounts and occurrences; leaving less or no room for dry days to occur. This phenomenon is clearly evident from the relative superior performances demonstrated by their counterparts, particularly on BiasF and FAR statistics. Furthermore, as discussed in the subsequent sections, these models (i.e. MLP, RMLP and PLS) have great difficulties in characterizing key precipitation features, such as dry-spell, wet-spell and variance statistics in downscaling daily precipitation fields.

The various downscaling models are further inter-compared using more general statistics such as mean and variability of daily precipitation amounts. Such generic tests provide the overall performance of models over first and second order moments. Figure 2 compares observed and downscaled daily precipitation amounts for each forecast range. The statistics associated with the observed and the downscaled precipitation values were computed annually. Figure 2a shows that all the downscaling models (with the exception of ANN-Logst and PLS-Logst) yielded comparable performances in terms of representing the mean observed precipitation. Also, these models tended to slightly under-represent the mean observed local precipitation. On the other hand, the ANN-Logst and the PLS-Logst models over-represented significantly the mean observed local

precipitation, beyond the five day forecast. Similarly, comparative plots of the variances of the observed and the downscaled daily precipitation amounts are shown in Figure 2b. Despite a slight and a consistent under-representation of the observed precipitation, the KNN-E and the KNN-M models demonstrated the best performance in representing the variability in daily precipitation. The SDSM model also demonstrated a reasonably adequate performance. The remaining downscaling models were, however, unable to capture the variability in daily precipitation and even showed further deterioration with increased forecast ranges. The relative poor performance of these models can be explained in part by the technique used in modeling the precipitation process (i.e. the method has a tendency to smoothen and deflate the variance of daily precipitation). The failure to capture the variability in daily precipitation is one of the acknowledged limitations of most statistical models [e.g., *Clark and Hay*, 2004; *Harpham and Wilby*, 2005].

The average lengths of dry- and wet-spells in each month are important diagnostic measures commonly used for evaluating the accuracy of the downscaled precipitation. In the present study, dry days were considered dry when days had precipitation intensities of 0.3 mm or less; and dry-spell length in a given month was computed as the maximum number of consecutive dry days in that month [*Khan et al.*, 2006]. In this study, results are presented only for mean wet-spell lengths, for forecast ranges of 3, 7 and 10. The comparative plots of the average wet-spell lengths for each month of observed and downscaled precipitation are shown in Figure 3. The MLP, RMLP and PLS-based downscaled precipitation data appeared to over-represent the mean wet-spells significantly. The other competing models reproduced the portion of wet-spells reasonably, for all months, in spite of a slight under-representation of observed precipitation. The coupled models (PLS-Logst and ANN-Logst) represented the mean wet-spells reasonably for the first and the last few forecast ranges, and significantly over-represented the remaining forecast ranges. Similar but opposite model performances were observed for mean dry-spell characteristics (results not shown).

Figures 5(a)-(d) depict quantile-quantile plots of observed and downscaled daily precipitation, for each season (only results of forecast range 7 are shown). The q-q plots were constructed using an empirical relationship between the quantiles of the observed precipitation and the quantiles of the downscaled precipitation. The purpose of the q-q plot was to determine whether the samples of the observed and the downscaled precipitation came from populations with a common distribution. If the samples came from the same distribution then the points of the q-q plot should fall approximately along some reference line [*Muluye*, 2010]. The overall results emerging from the various downscaling models suggested a relatively good performance for the winter and fall seasons. In comparison, performance was poor for summer and worst for spring, which could presumably be associated with the failure of the downscaling models to adequately describe isolated convective storms and more dynamic circulation patterns. This finding is consistent with results from earlier studies [e.g., *Harpham and Wilby*, 2005; *Hundecha and Bardossy*, 2007]. The SDSM, KNN-E and KNN-M models exhibited superior performance when compared with the other competing models, the latter two representing the best performance overall. Regardless of the seasons under consideration, all models (with the exception of SDSM, KNN-E and KNN-M) over-represented dry and light precipitation events, and under-represented significantly heavy events. Performance was particularly poor for heavy events. The demonstrated poor performance could be associated with the smoothening of precipitation events in the process of modeling, which presumably did not leave any, or enough space for dry days to occur at the early stage of the q-q plot and therefore deflated the variability of daily precipitation at the latter stage of the q-q plot [*Muluye*, 2010].

In addition to the above diagnostic measures, statistical tests were also performed between the observed and the downscaled precipitation to determine if there is evidence of difference in the population locations and variances without assuming a parametric model for the distributions. Two statistical tests were conducted: (i) Mann-Whitney test [*Lehmann*, 1975] of the equality of two population medians; and (ii) Levene's test [*Levene*, 1960] of the equality of two population variances. Table 4 presents the results of

the statistical tests conducted. The values in this table represent *p*-values. A significance level of 5% is chosen for the purpose of comparison. A *p*-value greater than the chosen significance level suggests the downscaled and the observed precipitation medians or variances are statistically equal. The two statistical tests were conducted on five of the seven downscaling models investigated in the present study. The ANN-Logst and the PLS models were not considered as no substantial differences in performance were observed when compared to their associate downscaling models.

The results of the Mann-Whitney test (Table 4) confirm that the SDSM, KNN-M and KNN-E-based precipitation medians and the observed precipitation medians are statistically equal for most of the months (reported *p*-values greater than the rejection level of significance). With regard to the other downscaling models, except for few months of fall and winter, the observed and the downscaled precipitation medians are statistically different. Similarly, the results of the Levene's test show that the reported *p*-values associated with the SDSM, KNN-E and PLS-Logst models were greater than the rejection level of significance for most of the months suggesting the variances of the observed and the downscaled precipitation are statistically equal (Table 4). The RMLP model generally reproduced the observed variances accurately except for September and December where the downscaled and the observed precipitation yielded statistically different variances. The *p*-values associated with the KNN-M model were, however, less than the chosen level of significance for most of the months (except for June, August and September) indicating the observed and the downscaled precipitation variances are statistically different. The relative performance of the different downscaling models in terms of reproducing the observed precipitation median and variance can be noticed in Table 4.

5. Summary and conclusions

The present study undertook a rigorous inter-comparison of daily precipitation statistics as downscaled by eight different statistical downscaling models for the Chute-du-Diable sub-basin located in northeastern Canada. A "reforecast" data set generated by

the Climate Diagnostics Center with a T62 resolution version of the NCEP MRF model was used in the present study. An ensemble of 15-day forecasts over a 23-year period from 1979 through 2001 was available for analysis. Eight model output variables corresponding to the first ensemble member (i.e. out of 15) and all 15-forecast ranges were considered in the downscaling experiment. Concurrent observed station daily precipitation was collected from Alcan Company. The data set from 1979-1996 was used to calibrate the models, and the remaining data set from 1997-2001 was used to evaluate the performance of downscaling models. A range of standard suites of diagnostic measures were employed to evaluate and inter-compare the downscaling results.

Multiple linear regression and principal component regression are common statistical methods used to link the large-scale model output to station-scale variables. Partial least squares regression was, however, used in the present study as it generalizes and combines good features from both models. In order to properly characterize precipitation occurrences by the PLS and the neural network models, hybrid models were proposed. In these hybrid models, precipitation was modeled in a two stage process: logistic regression was used to identify occurrence of wet days, and the PLS and the ANN models were used to model the amounts. A similar but different two-stage approach was also used for modeling with the SDSM. The hybrid models generally showed improvements in representing precipitation occurrence, but their overall skills were not superior to their counterparts.

To link the dynamics of the large scale predictors to station scale precipitation, two neural network models were investigated. The overall performance of the MLP and the RMLP models in representing daily precipitation characteristics were fairly satisfactory. In comparison, the RMLP model outperformed the MLP model in the vast majority of model evaluation measures. The demonstrated performance of the RMLP was not a coincidence given its strong mathematical foundation. The RMLP model used in the present study utilized an extended Kalman filter approach to perform supervised training. The improved statistics shown in this study could be attributed in part to the optimal filtering capability of the model.

The other models considered in downscaling daily precipitation fields were nearest neighbors. Two types of KNN models were designed on the basis of distance formulation. In general, the KNN model which used Euclidean distance yielded less skill and showed poor statistics, when compared with the KNN model which used Mahalanobis distance. Compared to the other competing models, both KNN models exhibited the best performance in representing the variability in daily precipitation.

The present study considered a number of models in downscaling daily precipitation in order to provide better insights into the nature of the problem. In view of the downscaling results, the following general conclusions and recommendations have been outlined.

(i) The skill values in the downscaled daily precipitation were inadequate.

(ii) Large differences were observed in performance from season-to-season. Performance was better in winter followed by fall. On average, spring was the season with the lowest skill values.

(iii) Considerable distinction was observed among models in reproducing the variability in daily precipitation. The KNN and SDSM models showed a clear advantage over their counterparts in reproducing the variability in daily precipitation.

(iv) Most models showed a decreasing trend in skill with increased forecast lead time.

(v) None of the models examined in this work will always be superior in downscaling daily precipitation. Nevertheless, a clear pattern emerged with respect to the reproduction of variations in daily precipitation and skill. The PLS, PLS-Logst, ANN-Logst, MLP and RMLP models showed modest skill, whereas the SDSM and KNN models showed considerable potential to capture the variability in daily precipitation.

(vi) The performance of models may be quite different from season to season and from region to region. Thus, care should be exercised while models are inter-compared.

(vii) The present study was carried out using a data matrix corresponding to the first member and the 15 forecast ranges (there are 15 members and 15 forecast ranges).

The performance may be different when different data matrices from other members are considered.

(viii) Most downscaling models used in the present study generally provided better performance statistics than the different downscaling models reported in the introduction section. It should be noted that the different performance measures reported here cannot be employed to compare the performance of the different downscaling models which are applied on the study basin against the performance of the different downscaling models which are applied on a different basin, as all of the downscaling models considered were not applied on the same basin.

Acknowledgements

This research was supported by the School of Graduate Studies at McMaster University. The author is grateful to Dr. Paulin Coulibaly and to Alcan Company for making the experiment data available. The author is also grateful to Dr. Noel Evora for providing the pre-processed ensemble weather predictors for the study area. The ensemble reforecast data is made available by NOAA at: http://www.cdc.noaa.gov/reforecast/. Minitab Statistical Software is used to conduct statistical tests.

References

ASCE Task Committee on Application of Artificial Neural Network in Hydrology (2000), Artificial neural network in hydrology. I: Preliminary concepts, *J. Hydrol. Eng.*, *5*, 115-123.

Buishand, T. A., and T. Brandsma (2001), Multisite simulation of daily precipitation and temperature in the Rhine basin by nearest-neighbor resampling, *Water Resour. Res.*, *37*(11), 2761– 2776.

Choi, J., T.H. Yeap, and M. Bouchard (2005), Online State–Space Modeling Using Recurrent Multilayer Perceptrons with Unscented Kalman Filter, *Neural Processing Letters*, *22*, 69–84, DOI 10.1007/s11063-005-2157-2.

Clark, M. P., and L.E. Hay (2004), Use of medium-range numerical weather prediction model output to produce forecasts of streamflow, *J. Hydrometeorol.*, *5*(1), 15– 32.

Clark, M., S. Gangopadhyay, L. Hay, B. Rajagopalan, and R. Wilby (2004), The Schaake shuffle: A method for reconstructing space-time variability in forecasted precipitation and temperature fields, *J. Hydrometeorol.*, *5*(1), 243-262.

de Jong, S. (1993), SIMPLS: an alternative approach to partial least squares regression, *Chemom. Intell. Lab. Syst.*, *18*, 251-263.

Dibike, Y.B., and P. Coulibaly (2005), Hydrologic impact of climate change in the Saguenay watershed: comparison of downscaling methods and hydrologic models, *J. Hydrol.*, *307* (1-4), 145-163.

Fowler, H.J., S. Blenkinsopa, and C. Tebaldib (2007), Review Linking climate change modelling to impacts studies: recent advances in downscaling techniques for hydrological modeling, *Int. J. Climatol.*, *27*, 1547–1578

Gangopadhyay, S., M. Clark, and B. Rajagopalan (2005), Statistical downscaling using K-nearest neighbors, *Water Resour. Res.*, *41* (2), W02024, doi:10.1029/2004 WR 003444.

Hamill, T.M., J.S. Whitaker, and X. Wei (2004), Ensemble reforecasting: mproving medium-range forecast skill using retrospective forecasts, *Mon. Weather Rev.*, *132*, 1434– 1447.

Harpham, C., and R.L. Wilby (2005), Multi-site downscaling of heavy daily precipitation occurrence and amounts, *J. Hydrol*, *312*, 235–255.

Hay, L.E., and M.P. Clark (2003), Use of statistically and dynamically downscaled atmospheric model output for hydrologic simulations in three mountainous basins in the western United States, *J. Hydrol.*, *282* (1-4), 56-75.

Hay, L.E., M.P. Clark, R.L. Wilby, W.J. Gutowski, R.W. Arritt, E.S. Takle, Z. Pan, and G.H. Leavesley (2002), Use of regional climate model output for hydrologic simulations, *J. Hydrometeor.*, 3, 571–590.

Haykin, S. (2008), *Neural Networks and Learning Machines*, 3rd edition, Prentice-Hall, Upper Saddle River, N.J.

Hua, Z., and B. Zhang (2006), A hybrid support vector machines and logistic regression approach for forecasting intermittent demand of spare parts, *Applied Mathematics and Computation*, *18*(2), 1035–1048.

Hundecha, Y., and A. Bardossy (2007), Statistical downscaling of extremes of daily precipitation and temperature and construction of their future scenarios, *Int. J. Climatol*, *28*, 589 – 610.

Kalman, R.E. (1960), A new approach to linear filtering and prediction problems, Transactions of the ASME, *Journal of Basic Engineering*, *82*, 35-45.

Khan, M.S., P. Coulibaly, and Y. Dibike (2006), Uncertainty Analysis of Statistical Downscaling Methods, *J. Hydro.*, *319*(1-4), 357-382.

Lehmann, E.L. (1975), *Nonparametrics: Statistical Methods Based on Ranks*, Holden and Day, San Francisco.

Levene, H. (1960), *Contributions to Probability and Statistics*, Stanford University Press.

Li, H., L. Luo, E.F. Wood, and J. Schaake (2009), The role of initial conditions and forcing uncertainties in seasonal hydrologic forecasting, *J. Geophys. Res.*, *114*, D04114, doi:10.1029/2008JD010969.

Liu, X., P. Coulibaly, and N. Evora (2007), Comparison of data-driven methods for downscaling ensemble weather forecasts, *Hydrol. Earth Syst. Sci. Discuss.*, *4*, 189–210.

Lorber, A., L.E. Wangen, and B.R. Kowalski (1987), A Theoretical Foundation for the PLS Algorithm, *J. Chemometrics*, 1(19).

Mearns, L.O., I. Bogardi, F. Giorgi, I. Matyasovszky, and M. Palecki (1999), Comparison of climate change scenarios generated from regional climate model experiments and statistical downscaling, *J. Geophys. Res.*, *104*, 6603– 6621.

Mehrotra, R., A. Sharma, and I. Cordery (2004), Comparison of two approaches for downscaling synoptic atmospheric patterns to multisite precipitation occurrence, *J. Geophys. Res –Atmospheres*, *109*, D14107, doi:10.1029/2004JD004823109.

Mehrotra, R., and A. Sharma (2005), A nonparametric nonhomogeneous hidden Markov model for downscaling of multi-site daily rainfall occurrences, *J. Geophys. Res.*, *110*, D16108, doi:10.1029/2004JD005677.

Mehrotra, R., and A. Sharma (2006), Conditional resampling of hydrologic time series using multiple predictor variables: A k-nearest neighbour approach, *Advances in Water Resources*, *29*, 987–999.

Muluye, G.Y. (2010), Deriving hydrological variables from numerical weather prediction model output: a nearest neighbor approach, submitted to *Water Resources Research*, Manuscript No. 2010WR009750.

Muluye, G.Y., and P. Coulibaly (2007), Seasonal reservoir inflow forecasting with low frequency climatic indices: a comparison of data-driven methods, *Hydrological Science Journal*, *52*(3), 508-522.

Palma, L., P. Gil, J. Henriques, A. Dourado, and H. Duarte-Ramos (2001), Application of an Extended Kalman Filter for On-line Identification with Recurrent Neural Networks, *Proc. Of the 7th Jornadas Hispano-Lusas de Ingenieria Electrica*, Madrid: Spain.

Puskorius, G.V., and L.A. Feldkamp (1994), Neurocontrol of nonlinear Dynamical Systems with Kalman Filter Trained Recurrent Networks, *IEEE Transactions on Neural Networks*, *5*(2), 279-297.

Rajagopalan, B., and U. Lall (1999), A k-nearest neighbour simulator for daily precipitation and other weather variables, *Water Resour. Res.*, *35*, 3089–3101.

Roulin, E. (2007), Skill and relative economic value of medium-range hydrological ensemble predictions, *Hydrol. Earth Syst. Sci.*, *11*, 725–737.

Schaake, J., K. Franz, V. Bradley, and R. Buizza (2006), The Hydrologic Ensemble Prediction EXperiment (HEPEX), *Hydrol. Earth Syst. Sci.*, *3*, 3321-3332.

Schmidli J., C.M. Goodess, C. Frei, M.R. Haylock, Y. Hundecha, J. Ribalaygua, and T. Schmith (2007), Statistical And Dynamical Downscaling Of Precipitation: An Evaluation And Comparison Of Scenarios For The European Alps, *J. Geophy. Res.*, *112*, 10.1029/2005JD007026.

Schoof, J.T., and S.C. Pryor (2001), Downscaling temperature and precipitation: a comparison of regression-based methods and artificial neural networks, *Int. J. Climatol.*, *21*, 773–790.

Sharif, M, and D. Burn (2006), Simulating climate change scenarios using an improved K-nearest neighbor model, *J. Hydrol.*, *325* (2006) 179–196

Shi, X., A.W. Wood, and D.P. Lettenmaier (2008), How Essential is Hydrologic Model Calibration to Seasonal Streamflow Forecasting, *J. Hydrometeorol.*, *9*, 1350-1363.

Spak, S., T. Holloway, B. Lynn, and R. Goldberg (2007), A comparison of statistical and dynamical downscaling for surface temperature in North America, *J. Geophys. Res.*, *112*, D08101, doi:10.1029/2005JD006712.

Stanski, H.R., L.J. Wilson, and W.R. Burrows (1989), Survey of Common Verification Methods in Meteorology. *WMO World Weather Watch Tech. Report No. 8, WMOI TD No. 358*, 114 pp.

von Storch, H., E. Zorita, and U. Cubash (1993), Downscaling of global climate change estimates to regional scales: an application to Iberian rainfall in wintertime, *J. Clim.*, *6*, 1161–71.

Wetterhall, F., A. Bardossy, D. Chen, S. Halldin, and C.Y. Xu (2006), Daily precipitation-downscaling techniques in three Chinese regions. *Water Resour. Res.*, *42*, W11423, doi:10.1029/2005WR004573.

Wilby, R.L., and C.W. Dawson (2007), SDSM 4.1 – a decision support tool for the assessment of regional climate change impacts, *User Manual,* UK.

Wilby, R.L., and I. Harris (2006), A framework for assessing uncertainties in climate change impacts: Low-flow scenarios for the river Thames, UK, *Water Resour. Res.*, *42*, W02419, doi:10.1029/2005WR004065.

Wilby, R.L., C.W. Dawson, and E.M. Barrow (2002), SDSM–A decision support tool for the assessment of regional climate impacts, *Environ. Modell. Software*, *17*, 145–157.

Wilby, R.L., L.E. Hay, W.J. Gutowski, and R.W. Arritt (2000), Hydrological responsesto dynamicallya nd statistically downscaled climate model output, *Geophysical Res. Letters.*, *27*(8), 1199-1202.

Wilby, R.L., T.M.L. Wigley (1997), Downscaling general circulation model output: a review of methods and limitations, *Progress in Physical Geography*, *21*, 530–548.

Wilks, D.S. (1995), Statisistical Methods in the Atmospheric Sciences, Academic Press, San Diego, California.

Wold, S., P. Geladi, K. Esbensen, and J. Ohman (1987), Multi-Way Principal Components and PLS-Analysis, *J. Chemometrics*, *1*, 41-56.

Xu, C.Y. (1999), From GCMs to river flow: a review of downscaling methods and hydrologic modeling approaches, *Progress in Physical Geography, 23*(2), 229–249.

Yakowitz, S. (1993), Nearest neighbor regression estimation for null-recurrent Markov time series, *Stochastic Processes Their Appl.*, *48*, 311–318.

Yarnal, B., A.C. Comrie, B. Frakes, and D.P. Brown (2001), Developments and prospects in synoptic climatology, *Int. J. Climatol.*, *21*, 1923–1950.

Yates, D., S. Gangopadhyay, B. Rajagopalan, and K. Strzepek (2003), A technique for generating regional climate scenarios using a nearest-neighbor algorithm, *Water Resour. Res.*, *39*(7), 1199, doi:10.1029/2002WR001769.

Table 1. Meteorological stations of Saguenay watershed

Meteorological station name	Station location(in degrees)		Altitude (m)	Nearest grid
	Latitude	Longitude		
Bagot	48.3	71	159	(70,47.5)
Benoit	51.53	71.1	549	(70,52.5)
Bonard	50.7	71	506	(70,50)
Chute-Du-Diable	48.75	71.7	174	**(72.5,50)**
Chute-Des-Passes	49.9	71.25	399	(70,50)
Chiba	49.8	74.5	387	(75,50)
Cygnes	49.9	72.9	405	(72.5,50)
Long	50.5	72.95	468	(72.5,50)
Machisque	50.9	71.8	543	(72.5,50)
Metabetchouan	48.4	71.96	220	(72.5,47.5)
Mistassibi2	49.4	71.9	183	(72.5,50)
Normandin	48.83	72.55	137	(72.5,50)
Roberval	48.5	72.27	179	(72.5,47.5)

Table 2. NOAA reforecast ensemble variable fields

Variable Field	Description	Surface level (mb)	Grid
1. *apcp*	Accumulated precipitation (mm)	Surface	Latlon
2. *heating*	Vertically integrated diabatic heating (K/s/mb)	Vertical average	Latlon
3. *pwat*	Precipitable water	Surface	Latlon
4. *prmsl*	Pressure reduced to mean sea-level (Pa)	Surface	Latlon
5. *t2m*	Temperature at 2 meters (K)	Surface	Latlon
6. *rhum*	Relative humidity (%)	700 mb	Latlon
7. *u10m*	Zonal wind at 10 meters (m/s)	Surface	Latlon
8. *v10m*	Meridional wind at 10 meters (m/s)	Surface	Latlon

Table 3. Comparative performance statistics for downscaling daily precipitation. The bold numeric statistics represent the best statistics corresponding to each forecast range.

Model	Forecast range														
	Diagnostic	**F0**	**F1**	**F2**	**F3**	**F4**	**F5**	**F6**	**F7**	**F8**	**F9**	**F10**	**F11**	**F12**	**F13**	**F14**
	Bias (%)	-9.71	-7.25	-13.44	-11.26	-8.70	-9.71	-11.60	-11.73	**-0.41**	**-5.36**	-12.23	-6.03	-8.95	**-3.46**	-8.89
	RMSE	5.73	6.11	6.33	6.55	6.82	5.73	6.74	6.87	7.23	7.22	7.05	7.06	6.92	6.99	7.20
	r	0.40	0.27	0.18	0.14	0.07	0.40	0.05	0.07	0.05	0.00	0.00	0.02	0.01	0.04	-0.01
SDSM	RV	-0.04	-0.18	-0.27	-0.36	-0.47	-0.04	-0.44	-0.49	-0.65	-0.65	-0.57	-0.58	-0.52	-0.55	-0.64
	POD	0.62	0.59	0.59	0.58	0.56	0.62	0.51	0.49	0.52	0.52	0.52	0.54	0.51	0.53	0.5
	FAR	0.37	0.42	0.41	0.44	0.45	0.37	0.49	0.48	0.49	0.5	0.49	0.49	0.5	0.5	0.49
	BiasF	**0.99**	**1.02**	**1**	**1.03**	**1.01**	**0.99**	**1.01**	**0.94**	**1.01**	**1.04**	**1**	**1.07**	**1.02**	**1.06**	**0.98**
	Bias (%)	-12.01	-10.49	-10.82	-11.28	-10.38	-8.82	-8.87	-9.33	-9.61	-9.46	-9.39	-8.96	-8.64	-8.93	-8.3
	RMSE	4.54	4.87	5.19	5.3	5.43	5.55	5.57	5.6	5.6	5.6	5.6	5.6	5.59	5.59	5.58
	r	0.61	0.51	0.39	0.35	0.27	0.16	0.14	0.1	0.1	0.09	0.1	0.09	0.11	0.12	0.12
PLS	RV	0.35	0.25	0.15	0.11	0.07	0.02	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	POD	**0.93**	**0.96**	**0.98**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
	FAR	0.44	0.46	0.49	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
	BiasF	1.66	1.78	1.9	1.99	1.99	1.99	1.99	1.99	1.99	1.99	1.99	1.99	1.99	1.99	1.99
	Bias (%)	-8.45	-5.73	-3.17	-6.56	-3.33	**2.4**	16.97	22.42	16.83	6.95	25.58	7.24	17.37	21.82	38.07
	RMSE	4.53	4.88	5.26	5.41	5.62	5.83	5.9	5.95	6.01	6.04	5.97	5.98	5.95	5.9	5.95
	r	0.59	0.5	0.39	0.33	0.25	0.16	0.13	0.1	0.08	0.07	0.09	0.1	0.11	0.12	0.1
PLS-Logst	RV	0.35	0.25	0.12	0.07	0	-0.07	-0.1	-0.12	-0.14	-0.16	-0.13	-0.13	-0.12	-0.1	-0.12
	POD	0.69	0.68	0.67	0.61	0.61	0.61	0.67	0.69	0.63	0.57	0.67	0.56	0.63	0.66	0.74
	FAR	**0.22**	**0.26**	**0.3**	**0.36**	**0.39**	**0.43**	**0.45**	**0.46**	**0.48**	**0.48**	**0.49**	**0.48**	**0.48**	**0.48**	**0.49**
	BiasF	0.89	0.92	0.95	0.95	0.99	1.06	1.23	1.29	1.21	1.09	1.31	1.09	1.21	1.28	1.45
	Bias (%)	-14.29	-8.13	-8.91	**0.41**	5.15	-4.21	-14.35	-7.89	-6.67	-13.89	-11.73	-12.66	-7.2	-7.99	-17.46
	RMSE	7.13	7.63	7.35	7.21	7.97	7.68	7.08	7.58	7.58	7.07	7.31	7.31	7.6	7.55	7.07
	r	0.03	-0.01	0.03	0.07	0.05	-0.01	0.03	0.02	0.01	0.03	0.03	0.01	0	0.03	0.02
KNN-E	RV	-0.63	-0.87	-0.73	-0.66	-1.04	-0.89	-0.6	-0.84	-0.84	-0.6	-0.71	-0.71	-0.85	-0.83	-0.6
	POD	0.5	0.5	0.52	0.54	0.52	0.53	0.55	0.53	0.5	0.5	0.52	0.49	0.5	0.51	0.47
	FAR	0.49	0.5	0.48	0.48	0.5	0.51	0.46	0.49	0.5	0.5	0.47	0.52	0.49	0.48	0.5
	BiasF	**0.98**	**1.02**	**1.01**	**1.05**	**1.03**	**1.07**	**1**	**1.03**	**1**	**0.99**	**0.98**	**1.02**	**0.98**	**0.98**	**0.94**
	Bias (%)	-13.58	-17.02	-14.2	-8.79	-15.07	-6.54	-12.47	-7.39	-9.65	-9.45	-15.44	-10.58	-8.48	-8.23	-5.31
	RMSE	6.33	6.35	6.61	7.02	7.5	7.75	7.25	7.54	7.39	7.32	7.17	7.33	7.4	7.27	7.56
	r	0.27	0.24	0.17	0.17	0.03	0.03	0.03	-0.01	0.06	0	0.03	0.01	0.02	0.06	0.01
KNN-M	RV	-0.28	-0.29	-0.4	-0.58	-0.8	-0.92	-0.68	-0.82	-0.75	-0.71	-0.65	-0.72	-0.76	-0.69	-0.83
	POD	0.58	0.61	0.56	0.58	0.53	0.53	0.53	0.5	0.53	0.53	0.51	0.53	0.52	0.53	0.5
	FAR	0.42	0.39	0.44	0.43	0.46	0.48	0.48	0.51	0.48	0.49	0.48	0.5	0.47	0.48	0.51
	BiasF	**0.99**	**1**	**1**	**1.02**	**0.99**	**1.01**	**1.02**	**1.02**	**1.02**	**1.04**	**0.98**	**1.07**	**0.99**	**1.03**	**1.02**
	Bias (%)	-10.57	**-5.7**	**-2.05**	-7.14	**-3.16**	3.31	14.98	21.94	15.51	6.94	23.34	8.1	17.69	21.76	36.47
ANN-	RMSE	4.56	4.86	5.26	5.4	5.59	5.84	5.88	5.95	6.02	6.06	5.96	5.99	5.97	5.92	5.93
Logst	*r*	0.59	0.51	0.39	0.33	0.26	0.16	0.13	0.1	0.07	0.06	0.09	0.1	0.11	0.12	0.11
	RV	0.34	0.25	0.13	0.08	0.01	-0.08	-0.1	-0.12	-0.15	-0.16	-0.13	-0.14	-0.13	-0.11	-0.11

Model	Forecast range														
	Diagnostic	**F0**	**F1**	**F2**	**F3**	**F4**	**F5**	**F6**	**F7**	**F8**	**F9**	**F10**	**F11**	**F12**	**F13**	**F14**
	POD	0.69	0.68	0.67	0.61	0.61	0.61	0.67	0.69	0.63	0.56	0.67	0.56	0.63	0.66	0.74
	FAR	**0.22**	**0.26**	**0.3**	**0.36**	**0.39**	**0.43**	**0.45**	**0.46**	**0.48**	**0.48**	**0.49**	**0.48**	**0.48**	**0.48**	**0.49**
	BiasF	0.89	0.92	0.95	0.95	0.99	1.06	1.23	1.29	1.21	1.09	1.31	1.09	1.21	1.28	1.45
	Bias (%)	-10.37	-12.61	-12.18	-10.99	-11.13	-8.61	-8.12	-9.67	-10.39	-10.63	-10.02	-8.39	-8.21	-8.84	-4.13
	RMSE	4.58	4.86	5.21	5.32	5.44	5.55	5.57	5.58	5.64	5.63	5.6	5.6	5.58	5.61	5.68
	r	0.59	0.52	0.38	0.33	0.25	0.16	0.13	0.12	0.06	0.06	0.09	0.09	0.12	0.07	0.04
MLP	RV	0.34	0.25	0.14	0.1	0.06	0.02	0.02	0.01	-0.01	0	**0.01**	**0.01**	**0.01**	0	-0.02
	POD	**0.97**	**0.98**	**1**	**1**	**0.98**	**0.99**	**0.99**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**0.97**
	FAR	0.47	0.49	0.49	0.5	0.49	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
	BiasF	1.83	1.91	1.97	1.98	1.93	1.97	1.97	1.99	1.98	1.98	1.99	1.99	1.99	1.99	1.94
	Bias (%)	**-7.81**	-7.68	-7.59	-10.23	-8.58	-5.61	**-5.29**	**-5.44**	**-4.96**	-6.43	**-5.24**	**-4.24**	**-4.39**	-5.3	**-3.92**
	RMSE	**4.39**	**4.77**	**5.11**	**5.26**	**5.41**	**5.52**	**5.56**	**5.59**	**5.58**	**5.59**	**5.59**	**5.59**	**5.59**	**5.59**	**5.58**
	r	**0.63**	**0.53**	**0.42**	**0.36**	**0.27**	**0.19**	**0.14**	**0.11**	**0.11**	**0.1**	**0.1**	**0.11**	**0.11**	**0.11**	**0.12**
RMLP	RV	**0.39**	**0.28**	**0.17**	**0.12**	**0.07**	**0.03**	**0.02**	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**
	POD	**0.92**	**0.94**	**0.93**	**0.97**	**0.99**	**0.99**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
	FAR	0.42	0.43	0.46	0.48	0.49	0.49	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
	BiasF	1.6	1.64	1.71	1.85	1.94	1.96	1.99	1.99	1.99	1.99	1.99	1.99	1.99	1.99	1.99

Table 4. Test for position (i.e. median) and variance for downscaling daily precipitation corresponding to the first member and zero forecast range using Mann-Whitney and Levene, respectively, for the test period Jan 97 to Dec 01. Note that the values in the table represent *p*-values.

Month	Test for Position (Mann-Whitney)					Test for Variance (Levene)				
	SDSM	RMLP	PLS-Logst	KNN-M	KNN-E	SDSM	RMLP	PLS-Logst	KNN-M	KNN-E
Jan	0.219	0.158	0.375	0.020	0.950	0.357	0.517	0.483	0.000	0.986
Feb	0.156	0.226	0.067	0.971	0.026	0.221	0.147	0.128	0.003	0.237
Mar	0.249	0.000	0.605	0.264	0.467	0.578	0.291	0.327	0.001	0.635
Apr	0.059	0.000	0.522	0.939	0.010	0.917	0.973	0.423	0.000	0.651
May	0.071	0.000	0.028	0.542	0.231	0.080	0.903	0.260	0.001	0.016
Jun	0.411	0.000	0.004	0.535	0.115	0.855	0.896	0.576	0.141	0.394
Jul	0.703	0.000	0.004	0.852	0.128	0.244	0.191	0.505	0.013	0.851
Aug	0.156	0.226	0.067	0.982	0.290	0.956	0.014	0.363	0.250	0.951
Sep	0.315	0.038	0.732	0.826	0.224	0.851	0.035	0.193	0.571	0.982
Oct	0.854	0.001	0.641	0.501	0.336	0.934	0.101	0.503	0.472	0.734
Nov	0.921	0.756	0.067	0.476	0.873	0.650	0.065	0.141	0.000	0.971
Dec	0.604	0.538	0.262	0.541	0.860	0.480	0.027	0.089	0.000	0.813

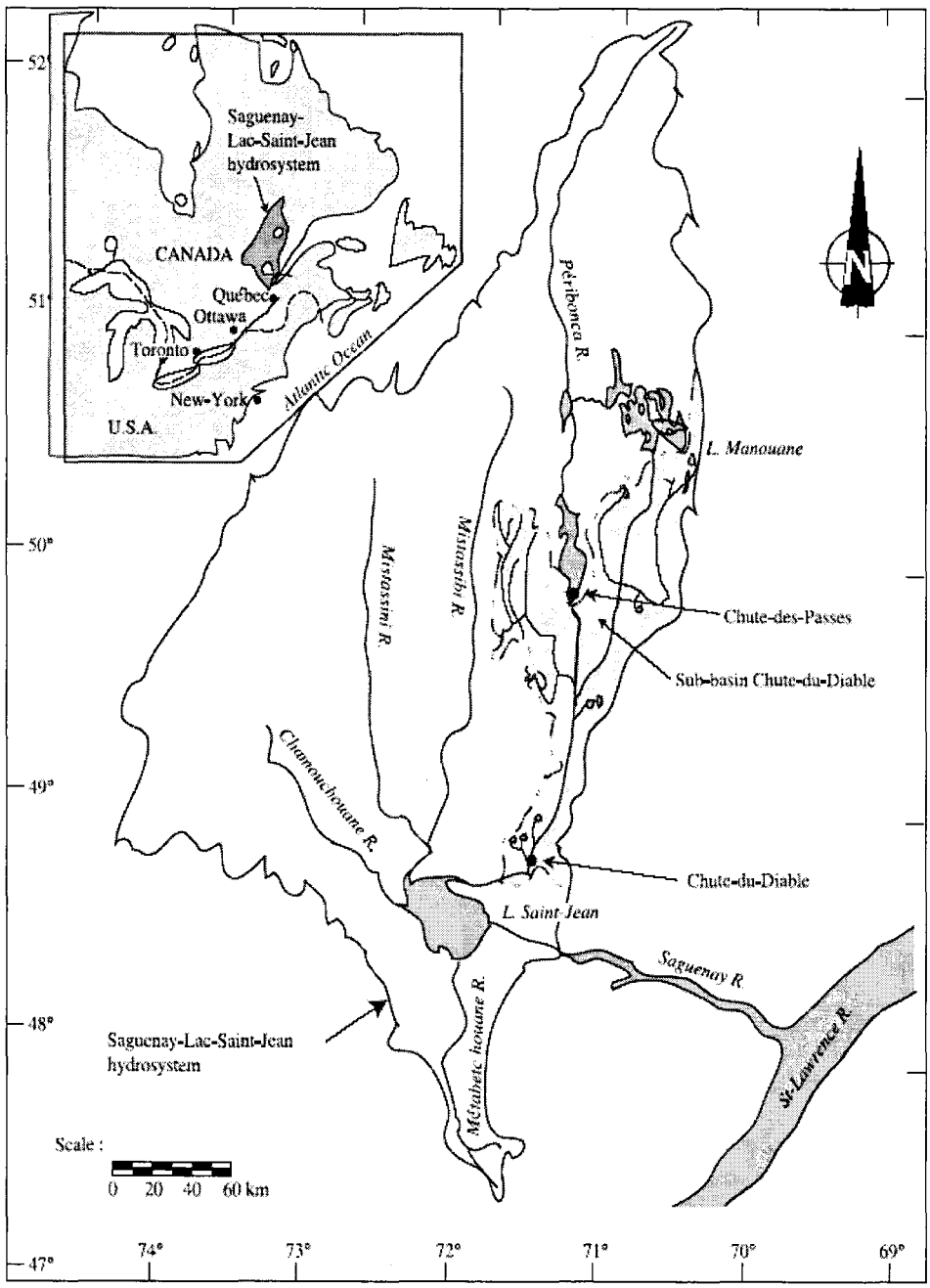


Figure 1. Location map of study area (Source: Dibike and Coulibaly, 2005).

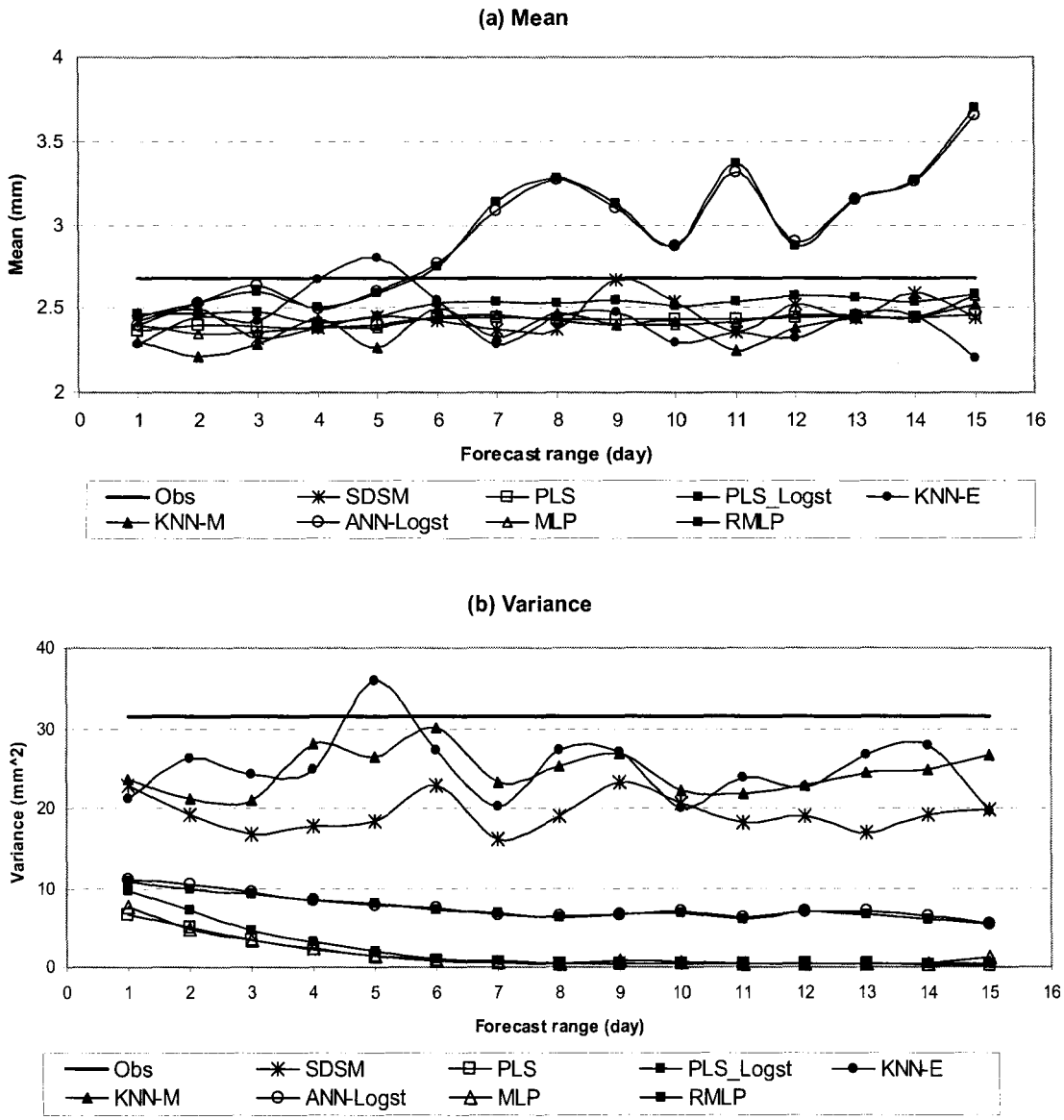


Figure 2. Comparison of downscaled precipitation derived from numerical forecast model: (a) mean precipitation totals, and (b) variance of daily precipitation, 1997-2001.

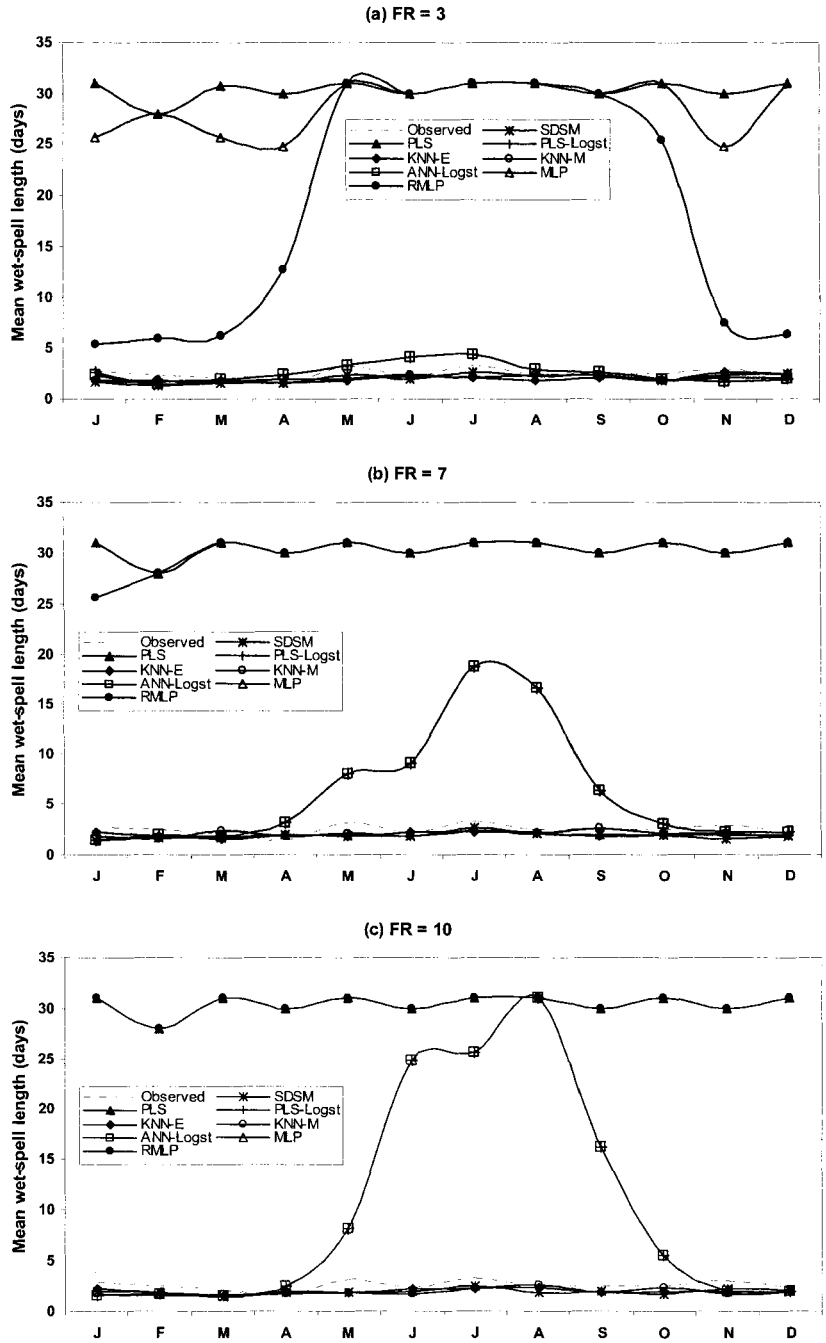


Figure 3. Mean length of wet-spells derived from numerical forecast model for forecast range (FR) (a) 3, (b) 7, and (c) 10, 1997–2001.

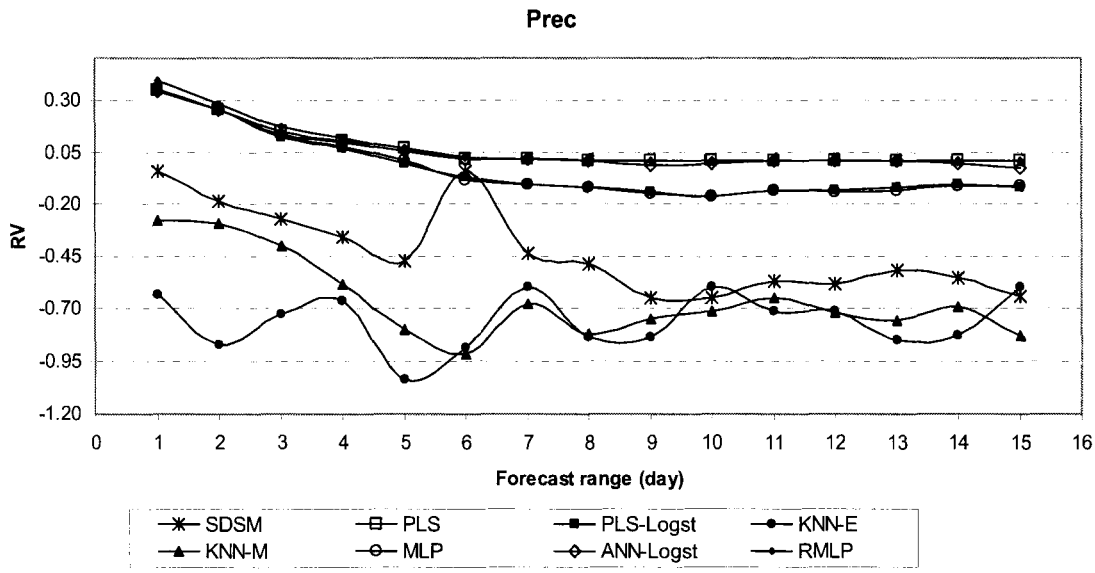
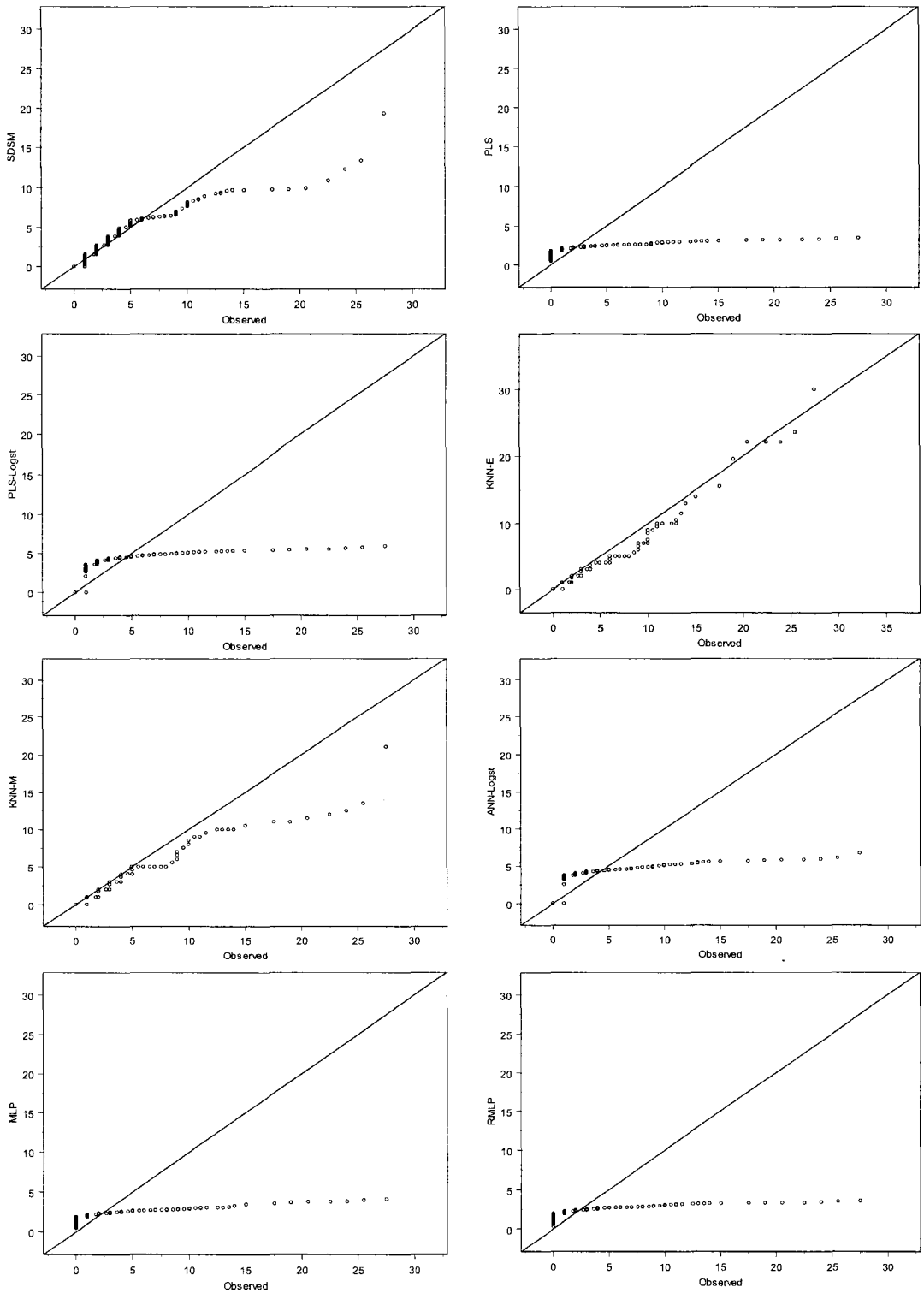
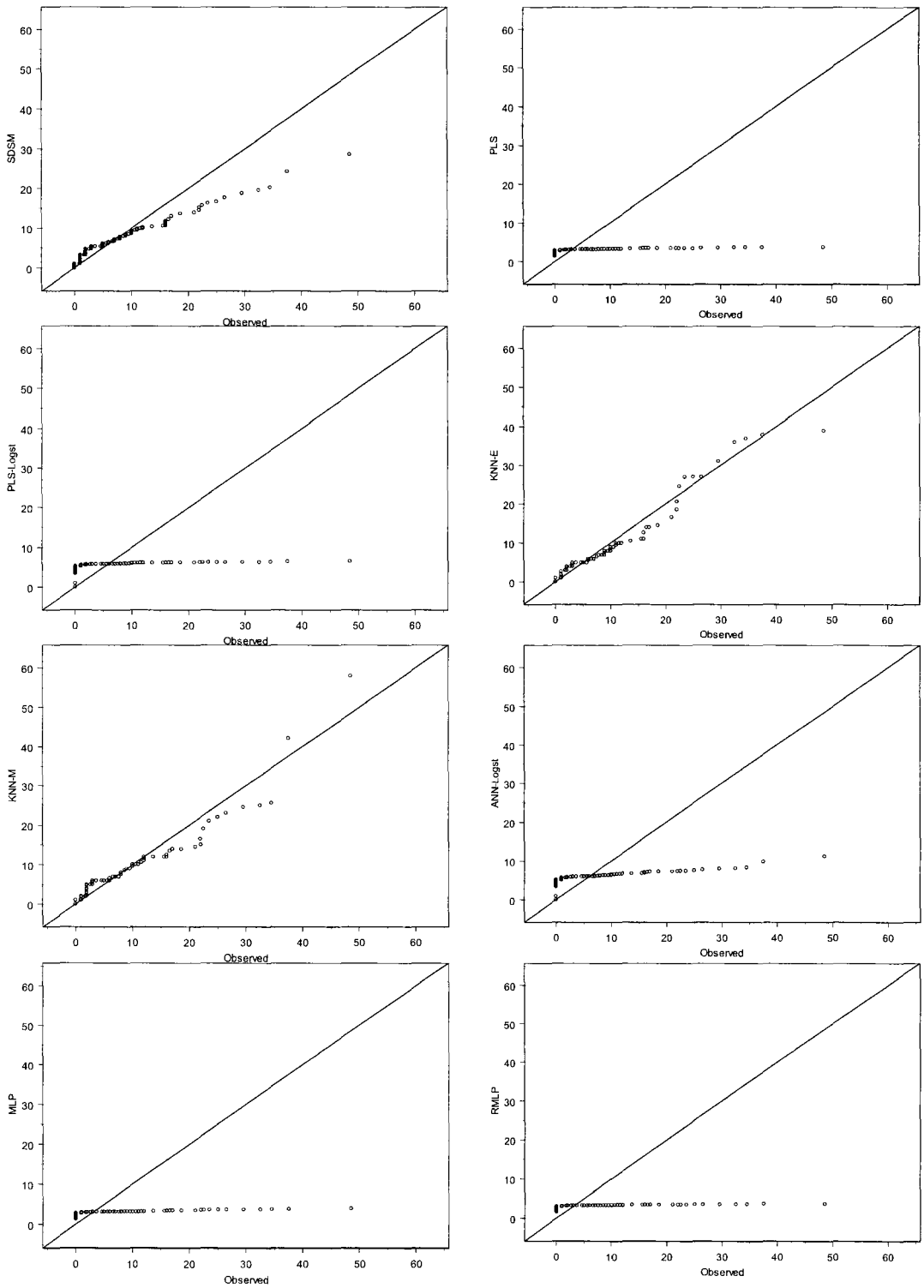


Figure 4. Reduction of variance (RV) with forecast range (FR) as downscaled by various models, 1997-2001.

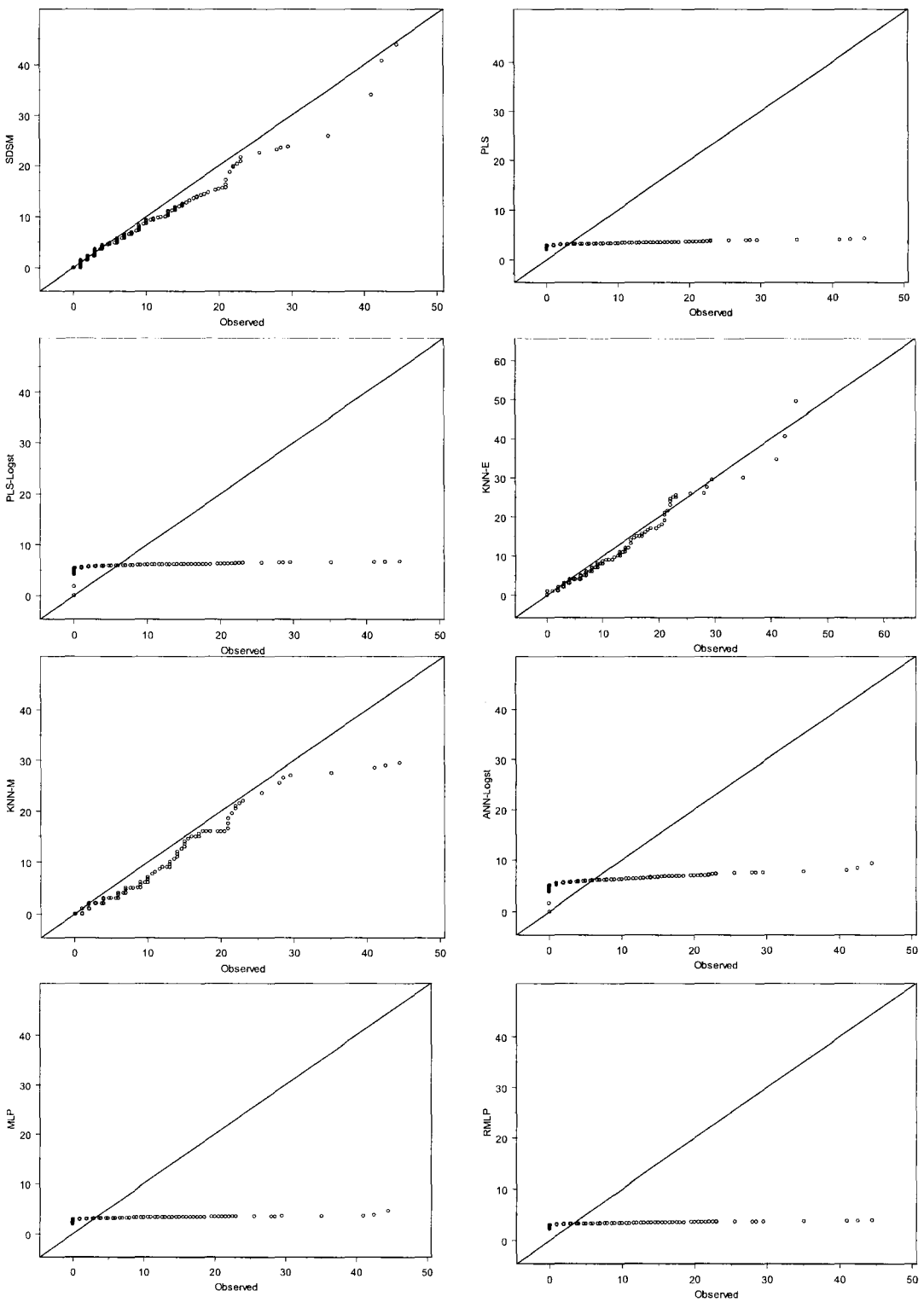
(a) Winter



(b) Spring


SDSM
PLS
PLS-Logsl
KNN-E
KNN-M
ANN-Logsl
MLP
RMLP
Observed
0
10
20
30
40
50
60

(c) Summer



(d) Fall

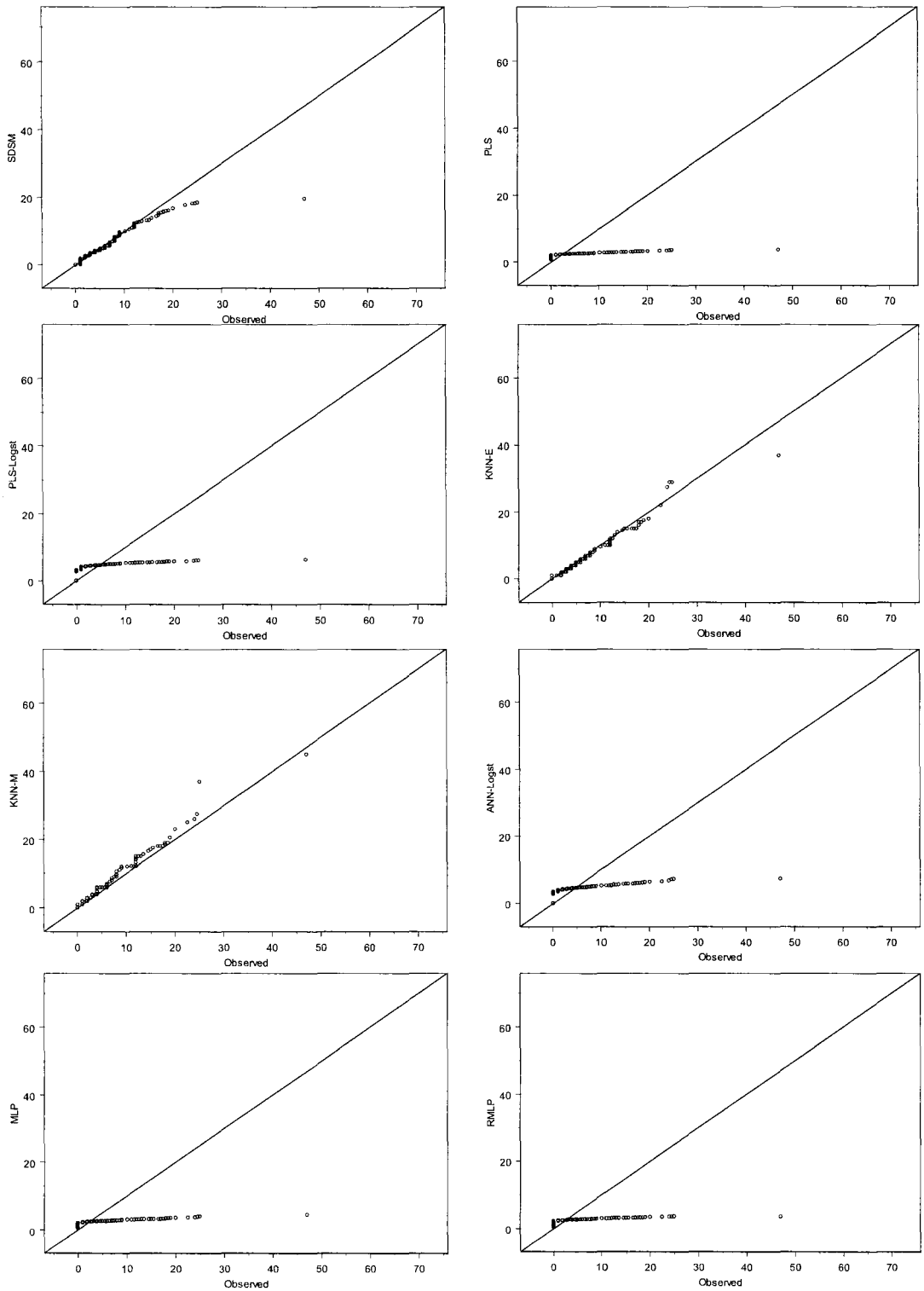

SDSM
PLS
PLS-Logst
KNN-E
KNN-M
ANN-Logst
MLP
RMLP
Observed
0
10
20
30
40
50
60
70

Figure 5. Quantile-quantile (q-q) plots of the quantiles of the observed precipitation (mm) against the quantiles of the simulated precipitation (mm) as downscaled by the different models, Jan 1997 to Dec 2001 for forecast range 7 for a) winter (JFM), b) spring (AMJ), c) summer (JAS), and d) fall (OND)

CHAPTER 5

Deriving hydrological variables from numerical weather prediction model output: a nearest neighbor approach

Chapter 5 is a continuation of chapter 4. In this chapter, various matters relating to the performance of the *K*-nearest neighbors for downscaling station daily precipitation and minimum and maximum temperature fields are studied. Two issues are discussed: a measure of closeness and a sampling strategy. Six different *K*-nearest neighbor-based models are developed and the results are compared. The skill values of the downscaled precipitation and the raw numerical model outputs are also assessed. The findings of the study are presented in a paper form and submitted to Water Resources Research for possible publication.

Deriving hydrological variables from numerical weather prediction model output: a nearest neighbor approach

Abstract

This paper presents the application of variations in a nearest neighbor resampler approach for deriving local-scale hydrological variables from numerical weather prediction model output. On the basis of measure of closeness and sampling strategy, six nearest neighbor models were designed. The proposed models were applied to downscale station daily precipitation, and minimum and maximum temperature fields in northeastern Canada. Suites of deterministic diagnostic measures were employed for evaluating individual models as well as for inter-comparison among the downscaling models. The comparative results indicated that nearest neighbor models employing Mahalanobis distance as a measure of closeness had greater skill values than the models using Euclidean distance, and the models weighting the *K*-nearest neighbors in proportion to their probabilities yielded better performance.

On the basis of inter-comparison among models a relatively better nearest neighbor resampler was identified and the subsequent model was further investigated with a focus on downscaling daily precipitation. Suites of conventional and distribution-based diagnostic measures were employed for evaluating the skill of the downscaled precipitation over the raw numerical model output. The comparative results showed that the downscaled precipitation had greater skill values based on different performance measures which include median bias, Brier skill score, ranked probability skill score, discrimination, reliability, and relative operating characteristics.

Keywords: Chute-du-Diable; nearest neighbor; precipitation; statistical downscaling

1. Introduction

The demands for accurate hydrological forecasts are becoming more crucial than ever to reduce the risk as well as to increase the efficiency of water management

decisions. Consequently, considerable efforts are underway to improve the quality of these forecasts. There has been significant progress in the skill of short-term global-scale atmospheric forecasts over the past several years, and it is likely that outputs from these forecasts, when used as input to hydrologic models, may provide improved hydrological forecasts (*Clark and Hay*, 2004). The skill of the hydrologic forecast obtained when forcing a distributed-hydrologic model with the National Centers for Environmental Prediction (NCEP) Medium-Range Forecast Model (MRF) output have been successfully demonstrated (e.g., *Clark and Hay*, 2004; *Werner et al.*, 2005). Nevertheless, the failure of large-scale weather forecasts to adequately characterize local-scale hydrological variables has become increasingly recognized as a drawback in the analysis and modeling of hydrological processes. Direct application of current output from large-scale models is often found to be inadequate to provide reliable sub-grid scale hydrological variables [*Murphy*, 1999; *Beck et al.*, 2004; *Spak et al.*, 2007]. To circumvent the scale mismatch, techniques are developed to extract local-scale hydrological variables from trends in large-scale fields related to these variables [*Wilby and Wigley*, 1997].

Downscaling is a method for obtaining local-scale information from large-scale model output. There are several downscaling models currently in use, which generally fall into two categories: dynamical and statistical downscaling. Dynamical downscaling uses physical relationships describing the interactions of the Earth's energy and moisture states to simulate finer-scale physical processes [e.g., *Mearns et al.*, 1999; *Murphy*, 1999; *Beck et al.*, 2004; *Spak et al.*, 2007]. Statistical downscaling uses statistical relationships to establish links between large-scale model outputs and surface variables [e.g., *von Storch et al.*, 1993; *Wilby et al.*, 2002; *Hay and Clark*, 2003; *Harpham and Wilby*, 2005; *Mehrotra and Sharma*, 2005]. The broad theory, applications, advantages and shortcomings of common downscaling methods are documented in the scientific literature [e.g., *Wilby and Wigley*, 1997; *Xu*, 1999; *Yarnal et al.*, 2001; *Fowler et al.*, 2007].

Gangopadhyay et al. [2005] used a *K*-nearest neighbor model for downscaling daily precipitation and temperature in the contiguous United States from the NCEP

output. The reported ranked probability skill score (RPSS) values were approximately 10% and 30% for downscaling daily precipitation and temperature for a seven day forecast, respectively. Similarly, *Werner et al.* [2005] used a multivariate linear regression approach to downscaled mean areal temperature (MAT) and mean areal precipitation (MAP) in the upper Colorado River basin from the NCEP output. For downscaling the MAT, the RPSS values reported showed substantial improvement (up to 70%) over climatological MATs during the first week, and for downscaling the MAP, the RPSS values showed slight improvement (0.2) for the first 3–4 days only after which no skill is added to the climatological MAPs. *Clark et al.* [2004] used a stepwise multiple linear regression approach to downscale precipitation and temperature for four stations located in the United States. The RPSS values reported in the *Clark et al.* [2004] study were approximately 10% and 20% for downscaling daily precipitation and temperature for a seven day forecast, respectively. *Wilby et al.* [2000] simulated daily rainfall and surface temperatures for the Animas River basin, Colorado using dynamically and statistically downscaled output from the NCEP-National Centers for Atmospheric Research (NCAR) reanalysis. The reported Pearson correlation coefficient (r) values were: 0.5 and 0.42 for the downscaled precipitation, 0.82 and 0.93 for the downscaled minimum temperature, and 0.83 and 0.88 for the downscaled maximum temperature, using the dynamical and statistical downscaling models, respectively.

Because of the efficiency advantages of statistical downscaling over dynamical downscaling, this paper exclusively focuses on statistical downscaling techniques. *K*-nearest neighbor-based models are selected on the basis of preliminary analysis [*Muluye*, 2010] and literature review. The *K*-nearest neighbor-based models are becoming popular and proven to be superior downscaling models when compared with other statistical downscaling models such as multiple linear regression-based models [e.g., *Gangopadhyay et al.*, 2005; *Mehrotra and Sharma*, 2006]. The main objectives of this paper are: (i) to investigate the potential of nearest neighbor approaches in downscaling minimum and maximum temperature as well as precipitation on a daily time scale from the NCEP MRF numerical weather prediction model output for the Chute-du-Diable sub-

basin located in northeastern Canada; (ii) to assess the skill of these downscaling models for long-lead time forecasts (i.e. out to 14 days); (iii) to decide a suitable downscaling method by comparing among the nearest neighbor models developed for the study basin; and (iv) to evaluate the skill of the selected downscaling model over the raw model output using suites of conventional and distribution-based diagnostic measures. The structure of this paper is organized as follows: The study area and data used are presented in section 2. Section 3 describes design and application of the nearest neighbor approach. Inter-comparison and discussion of the results of the downscaling experiments are presented in section 4. Finally, section 5 draws some conclusions and recommendations based on the downscaling results.

4. Study area and data

2.1. Study area and station data

The study area selected for the investigation and inter-comparison of downscaling models is the Saguenay-Lac-Saint-Jean (SLSJ) hydrologic system in northern Quebec, Canada (Figure 1). The basin is located approximately between 47.3° to 52.2° N and 70.5° to 74.3° W and drains a total catchment area of about 73,800 square kilometers. There are 13 meteorological and 11 hydrometric stations, and a number of reservoirs managed by the Alcan Company (Table 1). These reservoirs are typically used for hydropower generation. The meteorological station selected in the present study is the Chute-du-Diable, located approximately at 48.75° N and 71.7° W. For calibration and validation of the downscaling models, daily total precipitation as well as daily maximum and minimum temperature records are collected from Alcan hydro-meteorological stations for the period 1979 to 2001.

2.2. Reforecast dataset

The advances in a "reforecast" (retrospective forecast) system have fostered interest among various forecast users. The "reforecast" dataset are generated by the Climate Diagnostics Center which is readily available for download at http://www.cdc.noaa.gov/reforcast. A T62 resolution version of the NCEP MRF model

was used to generate an ensemble of 15-day forecasts over a 23-year period from 1979 to 2001 [*Hamill et al.*, 2004]. A 15-member ensemble was produced for every day of the 23 year dataset and the model outputs were saved at 0000 UT and 1200 UT with 0000 UTC initial conditions. The ensemble initial conditions consisted of a control initialized with the NCEP-National Centers for Atmospheric Research (NCAR) reanalysis [*Kalnay et al.*, 1996] and a set of 7 bred pairs of initial conditions [*Toth and Kalnay*, 1993] re-centered each day on the reanalysis initial condition [*Hamill et al.*, 2004]. An archive of data from 1 January 1979 to 31 December 2001 for a period of 23-years was used in the present study. Eight model output variables corresponding to 15-member ensemble and 15-forecast ranges were considered in the downscaling experiments (Table 2). The nearest grid to the Chute-du-Diable station was used to retrieve the large-scale model output.

5. Methodology

In this section, a brief description of a statistical downscaling model, namely, *K*-nearest neighbor is presented. Six versions of *K*-nearest neighbor-based models are developed on the basis of sampling strategy and measure of closeness. Finally, a summary of experimental setup and model calibration is presented.

3.1. Nearest neighbors

Stochastic weather generators have been used in a wide spectrum of water resources and environmental problems [e.g., *Oelschliigel*, 1995; *Wilks and Wilby*, 1999; *Dubrovsky et al.*, 2000; *Yates et al.*, 2003]. In a broader view, weather generators are classified into parametric and nonparametric approaches. Parametric weather generator methods are mainly based on stochastic weather generator models such as WGEN [*Richardson*, 1981] and LARS-WG [*Semenov and Barrow*, 1997]. These models are basically single-site models and hence cannot simultaneously simulate weather data at multiple sites. In addition, they necessitate specification of model parameters and generally have difficulties in reproducing the annual variability in monthly means of the variables [e.g., *Sharif and Burn*, 2006]. Nonparametric approaches on the other hand,

offer a different route by specifying the downscaling model based solely on observations, thus avoiding the need to estimate parameters for the downscaling to proceed. In general, nonparametric stochastic approaches are grouped into [*Mehrotra and Sharma*, 2006]: kernel density estimation [e.g., *Sharma*, 2000; *Sharma and O'Neill*, 2002] and nearest neighbor resampling [e.g., *Lall and Sharma*, 1996; *Mehrotra et al.*, 2004; *Gangodhyay et al.*, 2005]. The capabilities of these techniques have been demonstrated on a variety of water resources and environmental problems [e.g., *Rajagopalan and Lall*, 1999; *Buishand and Brandsma*, 2001; *Yates et al.*, 2003; *Mehrotra and Sharma*, 2006].

Nearest neighbor resampling methods are generally based on classic bootstrapping techniques [e.g., *Yakowitz*, 1993; *Yates et al.*, 2003; *Gangodhyay et al.*, 2005]. The *K*-nearest neighbor (KNN) algorithm essentially involves the search for similar feature vectors which exist in the observed time series on the principle of similarity criteria (i.e. *K* unique days are identified for the current day's weather, and one of these days is then randomly sampled and used as the next day's weather forecast) [*Gangodhyay et al.*, 2005]. In the present study, an observed station variable such as precipitation and temperature are considered as analog days to be identified on the basis of large-scale model output. Even though multiple linear regression (MLR) and other regression-based models are commonly in use [e.g., *Wilby et al.*, 2002; *Clark et al.*, 2004; *Dibike and Coulibaly*, 2005], KNN-based models have the following advantages [*Gangodhyay et al.*, 2005]: (i) KNN is a data-driven model and as such no prior assumptions are made on the distributions of variables; (ii) the KNN-based models intrinsically preserve the spatial covariability and consistency of the downscaled climate fields; (iii) ensemble MRF runs can be readily utilized in the downscaling process; (iv) the ensemble spread information from MRF runs can be utilized to develop spread-skill relationships; and (v) weather data at multiple sites can be simulated simultaneously [*Mehrotra et al.*, 2004].

The nearest neighbor resampling techniques developed in the present study are extensions of previous works by *Yates et al.* [2003] and *Gangodhyay et al.* [2005]. The primary focus of this paper is to examine the two modeling issues of nearest neighbors,

namely, the measure of closeness and the sampling strategy employed to design the KNN model. While the developments which have been made on KNN are briefly presented, the detailed theory and procedures may be found in *Yates et al.* [2003] and *Gangodhyay et al.* [2005].

1. Distance formulation: One of the key issues that dictates the effectiveness of nearest neighbor based models is the type of distance measure or nearness used in the modeling. The effect is more pronounced when multiple predictor variables exist in the feature vector. The nearness is usually quantified through a distance formulation. Euclidean and Mahalanobis distance metrics are normally used for this purpose. The Euclidean distance formulation views all predictors equally and as such the estimated distance metric may be biased. On the other hand, the Mahalanobis distance formulation has the following advantages over the traditional resampling models that utilize the Euclidean distance [*Mehrotra and Sharma*, 2006]: (i) the use of Mahalanobis distance obviates standardizing of the predictor variables; and (ii) the Mahalanobis distance measure considers the existing dependence amongst the predictor variables. The Euclidean distance formulation presented in *Gangopadhyay et al.* [2005] circumvents the traditional limitation in (ii) and considers the significance of each predictor. In this approach, correlated or redundant predictor variables are systematically screened through principal component analysis (PCA) and the resulting PCs selected are weighted according to their eigenvalues. Both distance measures, however, ignore the dependence between the predictors and predictands. Owing to their inherent advantages and limitations, both distance metrics have been considered in the present study. The technique which is discussed in *Gangopadhyay et al.* [2005] was used for modeling the Euclidean distance, and the technique which is discussed in *Yates et al.* [2003] was used for modeling the Mahalanobis distance.

2. Sampling strategy: Another important issue in nearest neighbor resampling is the strategy used to sample amongst the selected K-neighbors. Three sampling strategies have been considered in the present study.

(i) First-nearest neighbor: It is natural that once the distances computed by either Euclidean or Mahalanobis are sorted out, the one which ranked first is the most probable to be resampled. The sampling strategy applied here basically takes a day as the next analog (forecast) corresponding to the shortest distance (the first sorted K). Though the method is attractive as it circumvents the computation of the parameter K, the first-nearest neighbor does suffer from its occurrence by chance. This is indeed one of the acknowledged drawbacks of the method in problems associated with classification; but it is less likely to be the case in prediction tasks such as downscaling. In practice, the first-nearest neighbor is often used for efficiency.

(ii) K-nearest neighbor: This is the most common resampling technique generally employed in real applications [e.g., *Rajagopalan and Lall*, 1999; *Yates et al.*, 2003; *Gangopadhyay et al.*, 2005; *Mehrotra and Sharma*, 2006]. First, a probability metric with a weight w_i $(0 < w_i < 1)$ is assigned to each of the K- nearest neighbors for all $i = 1$ to K using the bisquare weight function based on distance $d_{(i)}$ [e.g., *Gangopadhyay et al.*, 2005]:

$$w_i = \frac{\left[1-\left(\frac{d_{(i)}}{d_{(K)}}\right)^2\right]^2}{\sum_{i=1}^{K}\left[1-\left(\frac{d_{(i)}}{d_{(K)}}\right)^2\right]^2} \tag{1}$$

The neighbor with the smallest distance is assigned the highest weight, while the neighbor with the largest distance (i.e., the Kth nearest neighbor) is assigned the least weight. Second, a uniform random number, $u \sim U[0, 1]$, is generated, and if $u \geq w_1$, then the day which corresponds to distance $d_{(1)}$ is sampled, or else if $u \leq w_K$, then the day which corresponds to distance $d_{(K)}$ is sampled, otherwise (i.e. $w_1 < u < w_K$) the day which corresponds to distance $d_{(i)}$ is sampled for which u is closer to w_i. It should be noted that other types of kernel functions can be used to compute the probability metric.

(iii) Weighted K-nearest neighbors: In this approach, the selected K-neighbor candidates are weighted in accordance with their probabilities as:

$$x_{next} = \sum_{i=1}^{K} w_i x_i \tag{2}$$

where x_{next} is the weighted value of all K-nearest neighbors, x_i is the value corresponding to each of the K- nearest neighbors for all i = 1 to K, and the weight w_i is as defined above.

On the basis of (i) distance formulation, and (ii) sampling strategy employed, six KNN models were developed. These include: (1) KNN model based on the Mahalanobis distance as a measure of closeness and the first-nearest neighbor as a resampling method (1NN-M), (2) KNN model based on the Mahalanobis distance as a measure of closeness and the K-nearest neighbor as a resampling method (KNN-M), (3) KNN model based on the Mahalanobis distance as a measure of closeness and the weighted K-nearest neighbor as a resampling method (KNN-MW), (4) KNN model based on the Euclidean distance as a measure of closeness and the first-nearest neighbor as a resampling method (1NN-E), (5) KNN model based on the Euclidean distance as a measure of closeness and the K-nearest neighbor as a resampling method (KNN-E), and (6) KNN model based on the Euclidean distance as a measure of closeness and the weighted K-nearest neighbor as a resampling method (KNN-EW).

3.2. Application

Two types of data are available for the present study: (i) local-scale predictands such as daily precipitation and minimum and maximum temperature, and (ii) eight large-scale model output predictor variables. The KNN algorithm is applied as follows. First, the numerical model output was processed to form a data matrix consisting of 8401 rows (corresponding to the number of days from 1 January 1979 to 31 December 2001) and 8 variable columns (corresponding to each of 15 ensemble members and 15 forecast ranges or lead times). Second, days similar to each of the 8401 days in the archive were identified using the KNN algorithm. While the first half of the data (1979-1996) was used to derive local-scale variables, the remaining data (1997-2001) was used to evaluate the performance of downscaling models. Suites of traditional and distribution-based

diagnostic measures were computed to evaluate the performance of the downscaling models.

The parameters associated with the KNN model were also examined. Two model parameters were considered: moving window (w), and optimal number of nearest neighbors (K). After analyzing the sensitivity of the KNN model to different choices of the width of moving window, a value of w = 14 days was chosen for use in the downscaling experiment. Similarly, an analysis was performed to find out the optimal value of K, and a value of K = 19 was found to offer adequate results in the present study. Out of the six variations of KNN models identified, the weighted KNN was less sensitive to the choice of the value K, and no analysis was performed for the case of 1NN as it always takes the shortest distance.

4. Results and discussion

In this section, the downscaling models described in the previous section are inter-compared and discussed. These models include: 1NN-E, 1NN-M, KNN-E, KNN-M, KNN-EW and KNN-MW. The discussion of results is presented in two sections. The first section is focused on inter-comparing six-nearest neighbor models in downscaling daily precipitation and minimum and maximum temperature. The data matrix used for this purpose corresponds to the first member (out of 15 ensemble members) and all of the 15 forecast ranges. Suites of deterministic diagnostic measures were computed to evaluate and inter-compare these downscaling models. The second section is focused on evaluating the skill of downscaled daily precipitation over the raw numerical model output, using the comparatively better nearest neighbor resampler identified in the first section. Suites of both conventional and distribution-based diagnostic measures were employed. The data matrix used in this experiment consisted of all 15 ensemble members and 15 forecast ranges. The accumulated precipitation over a 24-hour period (*apcp*) is considered to be the raw numerical model output (Table 2).

4.1. Comparison of downscaling models

The deterministic diagnostic measures used for comparing the downscaling results include general and categorical statistics. The general statistics employed here provided key statistics such as bias (%), root mean squared error (RMSE), Pearson correlation coefficient (r) and reduction of variance (RV). The categorical statistics provided the skill of downscaling models on one of the two key features of precipitation (i.e. occurrence) via frequency bias (BiasF), probability of detection (POD) and false alarm ratio (FAR). In addition, wet- and dry-spell length, and quantile-quantile (q-q) plots of observed and downscaled daily precipitation amounts were used. It should be noted that the discussion in this paper is largely focused on downscaling daily precipitation.

4.1.1. Error metrics

Forecast verification is an important part of a forecast process which compares the downscaled values against the corresponding observations. The deterministic error metrics are computed from the downscaled and the observed data. The error metrics used in the present study include bias, RMSE, r and RV statistics. Tables 3, 4 and 5 show the error metrics for daily precipitation (Prec), maximum temperature (Tmax) and minimum temperature (Tmin), respectively. In most of the cases, the downscaling models generally under-represented the mean precipitation amounts, hence the negative bias. The smallest mean bias (-5.3%) was produced by the 1NN-E over the 15 forecast ranges. In the case of downscaling Tmax and Tmin, the smallest mean bias of -9.1% and 9.8%, respectively, over the 15 forecast ranges were produced by the 1NN-M. The correlation coefficients between the downscaled and the observed precipitation indicated that the KNN-MW had an advantage over the others in the vast majority of cases, although the KNN-EW showed competency beyond the nine day forecast. This observation was further supported by the overall RMSE and RV statistics. On the other hand, for downscaling daily minimum and maximum temperature, the KNN-MW model showed unarguably outstanding performance followed by the KNN-EW (Tables 4 and 5). The other competing models

also demonstrated reasonably adequate performance. Bold numeric statistics in Tables 3, 4 and 5 represent the best statistics associated with each forecast range.

4.1.2. Precipitation occurrence

Categorical metrics were used to compare the occurrence processes for the different models considered. For this purpose, the contingency table was used to identify the types of errors made in the downscaled precipitation. In principle, a perfect forecast system would generate only hits and correct rejections, and no misses or false alarms [*Stanski et al.*, 1989]. From the elements in the contingency table, several categorical statistics can be generated. In this paper, only the major score statistics were computed and discussed. All metrics were evaluated with a 0.3 mm threshold for dry and wet day occurrences [e.g., *Wilby and Harris*, 2006]. The first categorical statistic considered was the bias score. The bias score (frequency bias) explains how the forecast frequency of "yes" events compared to the observed frequency of "yes" events. The range of the score is between 0 and infinity, for which a score of one represents a perfect forecast. Basically, the bias score signifies whether the forecast system has a tendency to under-represent (BiasF<1) or over-represent (BiasF>1) occurrences, but does not quantify how well the forecasts correspond to observations; i.e., BiasF only measures relative frequencies [*Stanski et al.*, 1989]. Table 3 shows the frequency biases associated with the various models. The frequency bias of wet periods simulated by all nearest neighbors, with the exception of the KNN-EW and KNN-MW models, were reasonably accurate (BiasF close to one). In contrast, the KNN-EW and KNN-MW models over-represented the occurrence of wet days consistently by more than 90% on average over the 15 forecast ranges.

Model performance was further evaluated using probability of detection (hit rate) and false alarm ratio [*Stanski et al.*, 1989; *Wilks*, 1995]. The POD explains the fraction of the observed "yes" events with correct forecasts. The range of the score is between zero and one, with a score of one representing a perfect forecast. The POD is sensitive to hits, but ignores false alarms; for this reason, the POD is normally used in conjunction with

the FAR. The FAR explains the fraction of the predicted "yes" events which did not occur. The range of the score is between 0 and one, with a score of zero representing a perfect forecast. This score is sensitive to false alarms, but ignores misses [*Stanski et al.*, 1989]. The ability of the KNN-EW and the KNN-MW models to capture the POD statistics were generally adequate (POD close to unity). However, approximately 50% of the cases were incorrectly forecasted (Table 3). In comparison, a relatively fewer false alarm ratios (on average 45%) were achieved by the 1NN-M, although some difficulties were observed in representing hit rates 54% on average.

While the performances of the KNN-EW and the KNN-MW models were fairly adequate in characterizing hit rates, very poor performances were observed in characterizing the frequency of wet days. This can be explained in part by the technique used in modeling the precipitation processes. In the case of modeling with the first-nearest neighbor and the *K*-nearest neighbor-based models, station daily precipitation which corresponds to either the first-nearest neighbor or the *K*-nearest neighbor was identified on the principle of random sampling techniques. This approach permits both dry and wet days to occur. Conversely, the KNN-EW and the KNN-MW models performed poorly in this regard due to the smoothening of all *K*-nearest neighbors, which, in turn, generated far too many wet days, leaving fewer or no room for dry days to occur. Furthermore, as will be discussed in the subsequent sections, the KNN-EW and the KNN-MW models have difficulties in characterizing key features such as dry-spell, wet-spell, and variance statistics in downscaling daily precipitation.

4.1.3. Mean and variability of precipitation amounts

The downscaling models are further compared using more general statistics such as mean and variability of daily precipitation amounts. Such generic tests provide the overall performance of models over first and second order moments. Figure 2 compares observed and downscaled daily precipitation amounts for each forecast range. The statistics of the observed and downscaled data were computed annually. The comparative results presented in Figure 2a show that all models demonstrated competitive

performances overall. In most of the cases, the mean observed local precipitation appeared to be under-represented slightly and consistently. The comparison in Figure 2b between the variance of observed and downscaled daily precipitation amounts provides useful insight into the performance of each downscaling models. The KNN-EW and KNN-MW models consistently and significantly under-represented the variability in daily precipitation, whereas the other competing models reasonably reproduced this key statistic, albeit with a slight under-representation. The relative poor performance of the KNN-EW and KNN-MW models can be explained in part by the technique used in modeling the precipitation process (i.e. the weighting of *K*-nearest neighbors has a tendency to smoothen and deflate the variance of daily precipitation).

4.1.4. Mean dry- and wet-spell length

The average lengths of dry- and wet-spells in each month can be a useful diagnostic tool for evaluating the performance of models in downscaling daily precipitation. In the present study, precipitation intensities of 0.3 mm or less were considered as dry days; and the dry-spell length in a given month was computed as the maximum number of consecutive dry days in that month [*Khan et al.*, 2006]. Experimental results are presented only for average wet-spell length, for forecast ranges of 3, 7 and 10 days. Figure 3 compares plots of the average wet-spell length for each month of observed and downscaled precipitation. The overall results indicated that the downscaled precipitation associated with the KNN-EW and KNN-MW models were less accurate and over-represented the wet-spell lengths consistently and significantly. Conversely, the other downscaling models were more accurate and reproduced the wet-spells reasonably for all months, despite a slight under-representation of observed precipitation. Similar but opposite model performances were observed for mean dry-spell characteristics (results not shown).

4.1.5. Forecast skill

The deterministic skill of the downscaled precipitation was assessed through a deterministic skill score (SS). This metric describes the relative improvement of forecasts over some reference forecast. When the MSE is used as a score in the SS formulation then the resulting statistic is called a reduction of variance [*Stanski et al.*, 1989]. An RV value of zero indicates no improvement over the reference forecast, one indicates a perfect forecast, and a negative value indicates the reference forecast is better than the forecast. Climatology is typically used as a reference forecast. Figure 4 presents RV plots of the various downscaling models with forecast ranges for downscaling daily precipitation, and maximum and minimum temperature. The comparative RV plots in Figure 4a show that the skill of precipitation forecasts were generally inadequate. The KNN-MW and KNN-EW models performed substantially better than the other KNN-based models, however, positive skills were observed only for the first three forecast ranges. Figure 4b and Figure 4c present similar RV plots for maximum and minimum temperature respectively. It is evident from these plots that these temperature forecasts are far superior than the precipitation forecasts, and downscaling minimum temperature appeared slightly more problematic than maximum temperature. Further analysis of Figures 4b and 4c exhibit a decreasing trend in skill with increased forecast ranges for models using the Mahalanobis distance as a measure of closeness (i.e. 1NN-M, KNN-M and KNN-MW). However, virtually no clear patterns were observed for those models using the Euclidean distance. Albeit Mahalanobis based models exhibited modest skill for the first four forecast ranges, their performances deteriorated rapidly afterwards. Overall, the RV statistics associated with the weighting scheme-based models, namely, KNN-EM and KNN-MW, had the greatest skill values in terms of reproducing daily minimum and maximum temperature, with the latter being the best for all forecast ranges and variables.

4.1.6. Quantile-quantile plots

Figures 5(a)-(d) illustrate quantile-quantile plots of observed and downscaled daily precipitation data for each season (only results of forecast range 7 are shown). The

q-q plots were constructed using empirical relationships between the quantiles of observed dataset versus the quantiles of downscaled dataset. The purpose of the q-q plot is to determine whether the samples of the observed and the downscaled data sets come from populations with a common distribution. If the samples come from the same distribution then the points of the q-q plot should fall approximately along some reference line. The q-q plot is a useful explanatory data analysis tool that offers a better insight about the nature of the difference than the various statistical techniques commonly in use.

In general, the q-q plots indicated that all KNN-based models offered reasonably adequate forecasts in Winter followed by Fall. Performance was poor for Summer and even worst for Spring, presumably due to the failure of downscaling models to effectively represent dynamic circulation patterns and recurrent storms. This is consistent with earlier findings from other related studies [e.g., *Harpham and Wilby*, 2005; *Hundecha and Bardossy*, 2007]. The q-q plots associated with both types of distance measures (i.e. Euclidean and Mahalanobis) were quite competitive. In comparison, the first-nearest neighbor (i.e. 1NN) sampling strategy showed better performance when compared with random sampling KNN and, both sampling approaches were, in turn, superior to the weighting scheme KNN (i.e. KNN-W). In most cases, with the exception of the summer season, the Euclidean and the Mahalanobis distance measures, respectively, showed a tendency to over-forecast and under-forecast the storms. Both distance measures significantly under-forecasted the storms in the KNN weighting scheme (i.e. KNN-EW and KNN-MW). In addition, regardless of the season under consideration, the KNN-MW and KNN-EW models over-represented dry and light events. Overall, both models (i.e. KNN-EW and KNN-MW) showed poor performance particularly for heavy events. Such poor performance could be associated with the smoothing of precipitation events while modeling, which likely left less room for the occurrence of dry days at the early stages of the q-q plot and deflated the variability of precipitation at the latter stages of the q-q plot.

4.1.7. Statistical tests

Statistical tests were performed between the observed and the downscaled precipitation to determine if there is evidence of difference in the population locations and variances without assuming a parametric model for the distributions. Statistical tests were conducted for the 1NN-M and the KNN-WM downscaling models. The preliminary evaluation results indicate that the 1NN-M and the KNN-WM models yielded better performance statistics in terms of reproducing precipitation variability and skill, respectively. The first statistical test investigated was the Mann-Whitney test (Lehmann 1975) of the equality of two population medians. The results of the Mann-Whitney test are presented in Table 6. The values in this table represent *p*-values. A significance level of 5% is chosen for the purpose of comparison. If the computed *p*-value is greater than the chosen significance level, then the simulated and the observed medians are statistically equal. Table 6 shows that the *p*-values associated with the 1NN-M model were greater than the chosen significance level of 5% for all 12-months, suggesting the observed and the downscaled medians are statistically equal. For the case of downscaling with the KNN-WM model, the computed *p*-values were less than the rejection level of significance (except for the month of January) indicating the observed and the downscaled medians are statistically different. The other non-parametric test conducted to check the equality of variances between the downscaled and the observed precipitation is Levene's test (Levene 1960). The Levene's test results show that the variances of all 12-months downscaled by the 1NN-M model yielded higher *p*-values, suggesting the observed and the downscaled variances are statistically equal (Table 6). The *p*-values associated with the KNN-WM model were, however, less than the chosen level of significance for most of the months (except for January, March and April) indicating the observed and the downscaled precipitation variances are statistically different. The relative performance of the two downscaling models in terms of reproducing the median and variance of the observed precipitation can be noticed in Table 6.

4.2. Evaluating the skill of downscaled precipitation over the raw model output

On the basis of the detailed downscaling model comparison conducted in the previous section, the first-nearest neighbor (1NN-M) was identified as the best alternative for further analysis. The 1NN-M was then used to downscale daily precipitation and the subsequent output was compared against the raw model output. Forecast ranges 0, 3 and 7 days were considered for this purpose. Suites of conventional and distribution-based diagnostic statistics were computed for the validation period to help for model evaluation. The conventional statistic employed was the median bias of all members, and the distribution-based diagnostic statistics employed were the Brier skill score, a ranked probability skill score, discrimination, reliability, and relative operating characteristics.

4.2.1. Bias

Bias is generally defined as the deviation of the expected value of a variable from the true value. The bias used in this section is the median bias (MB_l), whose general formulation is basically identical to the deterministic bias discussed earlier. The median bias for each forecast lead time (l) and forecast length (n) was estimated by [*Gangopadhyay et al.*, 2005]:

$$MB_l = Median\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\left(f_i^l\right)^e - o_i^l\right]; e=1,...,m\right\} \quad (3)$$

where, n is the total number of days in the time series for a given forecast period; o_i^l is the observation for day i and lead time l; and $\left(f_i^l\right)^e$ is the downscaled variable value for day i, lead time l, and ensemble member e. The bias is then computed for each ensemble member, and the biases for all ensemble members (m = 15) are further processed to compute the median bias, MB_l, for the desired lead time l.

Table 6 shows the median biases for forecast range 0, 3 and 7 days. The letters D and R in Table 6 represent the downscaled and the raw output scores, respectively, and RD represents the skill of the downscaled scores over the raw scores. The median biases were then expressed as a percentage of the mean climatology to yield better insights. While the raw daily precipitation under-forecasted (negative bias) the observed station

daily precipitation, the downscaled daily precipitation over-forecasted (positive bias) the observed station daily precipitation. In comparison, the downscaled median biases (Bias_med_D) for the three forecast ranges were much smaller than the raw median biases (Bias_med_R). Using the raw bias statistics as a reference forecast, the improvements in median bias due to downscaling (Bias_med_RD) were 56%, 76% and 81% for forecast ranges 0, 3 and 7 days respectively. This clearly demonstrates that the downscaled precipitation forecast were more accurate than the raw data forecasts, particularly for longer lead times.

4.2.2. Brier skill score

The Brier score (BS) is a common scalar accuracy measure which is typically employed to verify forecasts of dichotomous events. The BS is essentially the mean-squared probability error between the observations and forecasts, and is given by [*Brier*, 1950]:

$$\mathrm{BS} = \frac{1}{n}\sum_{i=1}^{n}(f_i - o_i)^2 \tag{4}$$

where f_i is the forecast probability of occurrence of the event, o_i is one if the event actually occurred and zero if it did not, and n is a set of forecasts of a two category (dichotomous) element. The BS is a negatively oriented score which ranges between zero and one, with a score of zero representing a perfect forecast and a score of one representing the worst possible forecast [*Stanski et al.*, 1989].

The application of the Brier score is more important to dichotomous events where the focus is to verify the occurrence of the event. To compute the score, it is necessary to define the event for verification. The Brier score may not be the appropriate diagnostic measure if the forecast system is made beyond a simple percent "yes" and percent "no" for the particular event being verified [*Stanski et al.*, 1989]. In the present study, a threshold value of 0.3 mm was used to classify dry and wet day occurrences [e.g., *Wilby and Harris*, 2006].

A single value, such as the Brier score, may not offer the actual value of a forecast system. Thus, it is necessary to compare the forecast of interest to a reference forecast, such as persistence forecasts, historical operational forecasts, or forecasts based on the historical distribution of observed values (climatology) [*Wilks*, 1995]. The Brier skill score (BSS) was used to assess the relative skill score [*Brier*, 1950]:

$$\mathrm{BSS} = 1 - \frac{BS_{for}}{BS_{ref}} \tag{5}$$

where BS_{for} is the Brier score for the forecast, and BS_{ref} is the Brier score for the reference forecast. It is a common practice to use climatology as a standard or reference forecast, however, any unskilled or even a skilled forecast can be used instead. The BSS describes the percentage improvement over the reference forecast, and ranges between $-\infty$ and 1. In this case, a positive score indicates that the ensemble forecast system showed skill over the reference forecast, for which a score of one represents a perfect forecast [*Stanski et al.*, 1989].

Table 6 indicates that the BS statistics associated with the downscaled daily precipitation (BS_D) yielded smaller errors, compared to the BS for the raw output (BS_R). When the raw BS statistics were used as a reference forecast, the improvements in BS due to downscaling (BSS_RD) were 11%, 26% and 26% for forecast ranges 0, 3 and 7 days respectively. However, despite its inferior performance when compared to downscaled BSS (BSS_D), the raw BSS (BSS_R) still had greater skill than climatology. The BS statistics associated with the raw and the downscaled precipitation suggest that the latter had a greater advantage over the former with increased forecast range.

4.2.3. Ranked probability skill score

The ranked probability score (RPS) is used to evaluate the overall forecast performance of probabilistic forecasts (Wilks, 1995). The RPS for a single forecast is the sum of the squared differences of the cumulative distributions of forecasts and observations [*Franz et al.*, 2003]:

$$\text{RPS} = \sum_{k=1}^{J}\left(\sum_{i=1}^{k} f_i - \sum_{i=1}^{k} o_i\right)^2, k = 1, \ldots, J \tag{6}$$

where f_i and o_i are the relative frequency of the forecast and observed traces respectively, and J is the number of forecast categories. The computation of RPS is similar to the Brier score except that the RPS considers multiple observation and forecast categories to be examined at once while the Brier score focuses on a single category [*Franz et al.*, 2003]. The RPS is preferable over the Brier score when the interest is on the overall forecast quality rather than a particular forecast category. The technique discussed in *Gangopadhyay et al.* [2005] was applied to compute the RPS. Primarily, the observed time series was used to distinguish the possible categories (e.g., 0 to 10, 10 to 20, . . ., 90 to 100th percentile) for forecasts of precipitation. Cumulative probabilities were computed for each forecast-observation pair in each category. The RPS was then computed as the sum of the squared differences between the observed and forecast cumulative probabilities over the whole categories.

Similar to the Brier score, a single value, such as the ranked probability score may not provide the actual quality of a forecast system. Thus, it is necessary to compare the forecast of interest with a reference forecast. The ranked probability skill score (RPSS) is used to evaluate the relative skill of forecasts [*Gangopadhyay et al.*, 2005]:

$$\text{RPSS} = 1 - \frac{\overline{RPS}_{for}}{\overline{RPS}_{ref}} \tag{7}$$

where $\overline{RPS}_{for}$ is the average RPS of the forecasts for a particular forecast period, and $\overline{RPS}_{ref}$ is the average RPS of the reference forecasts for the same period. The RPSS expresses the percentage improvement of the forecast over the reference forecast, and ranges from $-\infty$ to 1. A positive score indicates that the ensemble forecasts show skill over the reference forecast, with a value of one indicating a perfect forecast, and a negative value indicating that the reference forecast is better than the forecast.

Table 6 shows the RPSS statistics for the raw and the downscaled daily precipitation. The downscaled RPSS (RPSS_D) for the three forecast ranges were greater than the raw RPSS (RPSS_R). When the raw RPS statistics were used as a reference

forecast, the improvements in RPS due to downscaling (RPSS_RD) were 3%, 14% and 37% for forecast ranges 0, 3 and 7 days respectively. The RPSS statistics of raw and downscaled precipitation indicated that the latter had an advantage over the former with increasing forecast range. Even though the raw RPSS showed inferior performance when compared with the downscaled RPSS, the raw RPSS still showed greater skill than climatology.

4.2.4. Discrimination and reliability

One of the key issues in a forecast verification system is the method used to analyze the joint probability distribution of forecasts and observations [Wilks, 1995; *Gangopadhyay et al.*, 2005]. Forecast reliability and discrimination are the two most common methods employed to examine the overall conditional distribution of a forecast system. These methods are typically useful to evaluate the forecast performance of events at various levels of the historical distribution (e.g., heavy precipitation).

For a particular forecast system, the conditional distribution of forecasts given the observation is expressed as $p(f|o)$ [*Wilks*, 1995]. When the value of $p(f|o)$ for all possible observation categories is equal to zero, except for a single observation category, then the forecast system is perfectly discriminatory for forecasts of that particular observation [*Murphy and Winkler*, 1987; *Franz et al.*, 2003]. Figures 6(a)-(c) show the discrimination diagram that displays the conditional probability distributions of each possible precipitation category as a function of forecast probability. The precipitation categories examined in the present study were the lowest (0- 33%), middle (33-67%), and highest (67-100%) of the historical distributions and are referred to as low-, middle-, and high precipitation categories. Forecast probability categories (i.e., 0-5%, 5-10%, . . . , 95-100%) were used to represent the magnitude of the probabilities given to each of the three categories. It is evident from Figures 6(a)-(c) that both the raw and the downscaled precipitation were unable to discriminate low (or light) and high (or heavy) categories for all forecast ranges. The problem became more pronounced with increased forecast ranges. If forecasts were discriminatory, low- and high-precipitation categories could

have not been overlapped to a greater degree on the discrimination diagram [*Murphy et al.*, 1989]. Therefore, forecasts made by such models lack confidence, though, to some degree, precipitation forecasts in the middle-category showed modest discrimination.

Forecast reliability summarizes the information contained in the conditional distribution [$p(o \mid f)$] and describes how often an observation occurred given a particular forecast [*Franz et al.*, 2003]. For a perfect forecast system, the conditional probability of a forecast will have [$p(o = 1 \mid f) = f$)][*Murphy and Winkler* 1987]. In other words, for a set of forecasts for which a forecast probability value, f, is given to a particular observation, o, then the forecast is considered to be perfectly reliable when the relative frequency of the observation and the forecast probability is equal [*Murphy and Winkler* 1992; *Wilks*, 1995; *Franz et al.*, 2003]. Figures 6(a)-(c) depict the reliability diagram resulting from the different downscaling models. The precipitation categories examined in the present study were the highest portion (95-100%) of the historical time series. Forecast probability categories (i.e., 0-10%, 10-20%, . . . , 90-100%) were used to represent the magnitude of the probabilities given to the category for the highest portion. The overall reliability diagram indicates that the downscaled precipitation had an advantage over the raw precipitation forecast. In comparison to the raw probabilities, the downscaled probabilities are in reasonable accord with the perfect reliability line (i.e. 45^0 line). Even though the observed precipitation was under-represented at the early stages of forecast probability (i.e. forecasts fall above perfect reliability), it was significantly over-represented at higher probabilities (i.e. forecasts fall below perfect reliability). Generally, precipitation forecasts in either case (i.e., raw and downscaled) yielded poor performances.

4.2.5. Relative operating characteristics

The relative operating characteristic (ROC) curve has seen increased acceptance as a tool for verifying the various forecast systems. The ROC is a highly flexible method for evaluating the quality of dichotomous, categorical, continuous, and probabilistic forecasts [*Mason and Graham*, 1999]. The ROC measures the hit rate of the forecast

system against the false-alarm rate. These ratios can be readily computed from the contingency table.

The area under the ROC curve is the most widely used summary skill measure. In order to construct the ROC curve, the highest 30% of the historical distribution was considered and a set of hit and false-alarm rates were computed. The resulting hit rates were plotted against the corresponding false-alarm rates to construct the ROC curve. The ROC curve that lies along the diagonal line indicates no skill, a curve that is far towards the upper left corner indicates the highest skill, and a curve that lies below the diagonal line indicates negative skill [*Mason and Graham*, 1999]. The ROC is usually summarized through the integrated area under the ROC curve. A perfect forecast system has ROC area equal to one and the no skill forecast system has a value of 0.5. The relative performance of precipitation forecasts was also evaluated using the ROC skill score (ROCSS = 2 (ROC area -0.5)). As shown in Figures 6(a)-(c) and Table 6, the ROC area for the raw precipitation was greater for forecast range 0, equal to the downscaled precipitation for forecast range 3, and less than the downscaled precipitation for forecast range 7. Further analysis of Figures 6(a)-(c) and Table 6 reveals that the skills of precipitation forecasts under both methods deteriorated with increased forecast ranges. However, these forecasts still demonstrated greater skill than climatology.

5. Summary, conclusions and recommendations

The present study undertook an inter-comparison of daily precipitation, and minimum and maximum temperature statistics as downscaled by six nearest neighbor resampling techniques for the Chute-du-Diable sub-basin located in northeastern Canada. A "reforecast" data set generated by the Climate Diagnostics Center with a T62 resolution version of the NCEP MRF model was used in the present study. An ensemble of 15-day forecasts over a 23-year period from 1979 to 2001 was available for analysis. Eight model output variables corresponding to 15-ensemble members and 15-forecast ranges were used in the downscaling experiment. Concurrent observed station daily precipitation and minimum and maximum temperature data were collected from Alcan

company. While the first half of the data (1979-1996) were used to derive local-scale variables, the remaining data (1997-2001) were used to evaluate the performance of downscaling models. Standard suites of diagnostic measures were computed for the validation period for model evaluation.

To link the dynamics of large-scale predictors to station-scale hydrological variables, the nearest neighbor resampling downscaling technique was investigated. On the basis of measure of closeness and sampling strategy, six KNN-based models were developed. In light of the downscaling results and model comparisons, the following conclusions were drawn:

(i) The skills in the temperature forecasts were superior to those in the precipitation forecasts although downscaling minimum temperature appeared less reliable than downscaling maximum temperature. This illustrates the difficulties associated with downscaling daily precipitation. In general, all models exhibited less skill in downscaling daily precipitation.

(ii) In downscaling daily precipitation, large differences were observed in performance from season-to-season. Performance was better in winter followed by fall. On average, the lowest skill was observed in spring.

(iii) Significant differences were found with respect to the reproduction of daily precipitation statistics, especially in representing variance. The first-nearest neighbor (1NN) and stochastic nearest neighbor (KNN) based models showed superior performance than their associate weighting scheme (KNN-W)-based models in terms of reproducing the variability in daily precipitation.

(iv) A decreasing trend in skill with forecast lead time was noticed for the models utilizing the Mahalanobis distance as a measure of closeness (i.e. 1NN-M, KNN-M and KNN-MW); however, no clear trend was observed for the models employing the Euclidean distance as a measure of closeness. Overall the Mahalanobis-based models outperformed the Euclidean-based models.

(v) The weighting scheme-based nearest neighbor models (i.e. KNN-EW and KNN-MW) showed greater skill values compared to their counterparts for both temperature and precipitation forecasts.

(vi) The downscaled precipitation (which used 1NN-M) showed greater skill values than the raw model output.

(vii) Although the downscaled precipitation yielded better performance when compared to the raw model output, the model performance was not sufficient for downscaling daily precipitation. Poor performance of the model was evident from (1) the reliability diagram (i.e. forecast probabilities fall far away from perfect reliability line), (2) the discrimination diagram (i.e. forecasts lack confidence in discriminating among forecast categories), and (3) the relative operating characteristics (i.e. the area under the ROC curves were not all close enough to unity).

(viii) None of the models examined in this work will always be superior in downscaling daily precipitation. There exists a potential to inter-compare statistical–statistical, dynamical–dynamical, even statistical–dynamical models. It is important that the best identified model be compared with the raw model output to confirm its overall skill.

(ix) The KNN-based downscaling models used in the present study generally provided better performance statistics than the different downscaling models reported in the introduction section. It should be noted that the different performance measures reported here cannot be employed to compare the performance of the KNN-based downscaling models which are applied on the study basin against the performance of the different downscaling models which are applied on a different basin, as all of the downscaling models considered were not applied on the same basin.

As a final remark, although the use of the nearest neighbor resampling technique has been proven to be successful to solve a wide spectrum of problems associated with water resources, environmental and other challenges, the method is still emerging and has the potential to be improved further. The fundamental assumption of the KNN model is

that similar feature vectors respond to identical output. For that matter, extraction of feature vectors in relevance to the variable to be resampled could help provide improved results. For instance, two of the many ways which may possibly improve the performance of KNN via improved feature extraction are: (i) the use of latent variables from a partial least squares (PLS) regression model, and (ii) the use of outputs from hidden layer artificial neural networks.

Acknowledgements

This research was supported by the School of Graduate Studies at McMaster University. The author is grateful to Dr. Paulin Coulibaly and to Alcan Company for making the experiment data available. The author is also grateful to Dr. Noel Evora for providing the pre-processed ensemble weather predictors for the study area. The ensemble reforecast data is made available by NOAA at: http://www.cdc.noaa.gov/reforecast/. Minitab Statistical Software is used to conduct statistical tests.

References

Beck, A., B. Ahrens, and K. Stadlbacher (2004), Impact of nesting strategies in dynamical downscaling of reanalysis data, *Geophys. Res. Lett.*, *31* (19), L19101, doi: 10.1029/ 2004GL020115.

Brier, G.W. (1950), Verification of forecasts expressed in terms of probability, *Mon. Wea. Rev.*, *78*, 1–3.

Buishand, T.A., and T. Brandsma (2001), Multisite simulation of daily precipitation and temperature in the Rhine basin by nearest-neighbor resampling, *Water Resour. Res.*, *37*(11), 2761– 2776.

Clark, M., S. Gangopadhyay, L. Hay, B. Rajagopalan, and R. Wilby (2004), The Schaake shuffle: A method for reconstructing space-time variability in forecasted precipitation and temperature fields, *J. Hydrometeorol.*, *5*(1), 243-262.

Clark, M.P., and L.E. Hay (2004), Use of medium-range numerical weather prediction model output to produce forecasts of streamflow, *J. Hydrometeorol.*, *5*(1), 15– 32.

Dibike, Y.B., and P. Coulibaly (2005), Hydrologic impact of climate change in the Saguenay watershed: comparison of downscaling methods and hydrologic models, *J. Hydrol.*, *307* (1-4), 145-163.

Dubrovsky, M., Z. Zalud, and M. Stastna (2000), Sensitivity of CERES-Maize yields to statistical structure of daily weather series, *Clim. Change*, *46*, 447– 472.

Fowler, H.J., S. Blenkinsopa, and C. Tebaldib (2007), Review Linking climate change modelling to impacts studies: recent advances in downscaling techniques for hydrological modeling, *Int. J. Climatol.*, *27*, 1547–1578

Franz, K.J., H.H. Hartmann, S. Sorooshian, and R. Bales (2003), Verification of National Weather Service Ensemble Streamflow Predictions for Water Supply Forecasting in the Colorado River Basin, *American Meteorological Society*, *4*, 1105-1118.

Gangopadhyay, S., M. Clark, and B. Rajagopalan (2005), Statistical downscaling using K-nearest neighbors, *Water Resour. Res.*, *41* (2), W02024, doi:10.1029/2004 WR 003444.

Hamill, T.M., J.S. Whitaker, and X. Wei (2004), Ensemble reforecasting: mproving medium-range forecast skill using retrospective forecasts, *Mon. Weather Rev.*, *132*, 1434– 1447.

Harpham, C., and R.L. Wilby (2005), Multi-site downscaling of heavy daily precipitation occurrence and amounts, *J. Hydrol*, *312*, 235–255.

Hay, L.E., and M.P. Clark (2003), Use of statistically and dynamically downscaled atmospheric model output for hydrologic simulations in three mountainous basins in the western United States, *J. Hydrol.*, *282* (1-4), 56-75.

Hundecha, Y., and A. Bardossy (2007), Statistical downscaling of extremes of daily precipitation and temperature and construction of their future scenarios, *Int. J. Climatol*, *28*, 589 – 610.

Kalnay, E., et al. (1996), The NCEP/NCAR 40-year reanalysis project, *Bull. Am. Meteorol. Soc.*, *77*(3), 437– 471.

Khan, M.S., P. Coulibaly, and Y. Dibike (2006), Uncertainty Analysis of Statistical Downscaling Methods, *J. Hydro.*, *319*(1-4), 357-382.

Lall, U., and A. Sharma (1996), A nearest neighbor bootstrap for time series resampling, *Water Resour. Res.*, *32*, 679– 693.

Lehmann, E.L. (1975), *Nonparametrics: Statistical Methods Based on Ranks*, Holden and Day, San Francisco.

Levene, H. (1960), *Contributions to Probability and Statistics*, Stanford University Press.

Mearns, L.O., I. Bogardi, F. Giorgi, I. Matyasovszky, and M. Palecki (1999), Comparison of climate change scenarios generated from regional climate model experiments and statistical downscaling, *J. Geophys. Res.*, *104*, 6603– 6621.

Mehrotra, R., A. Sharma, and I. Cordery (2004), Comparison of two approaches for downscaling synoptic atmospheric patterns to multisite precipitation occurrence, *J. Geophys. Res –Atmospheres*, *109*, D14107, doi:10.1029/2004JD004823109.

Mehrotra, R., and A. Sharma (2005), A nonparametric nonhomogeneous hidden Markov model for downscaling of multi-site daily rainfall occurrences, *J. Geophys. Res.*, *110*, D16108, doi:10.1029/2004JD005677.

Mehrotra, R., and A. Sharma (2006), Conditional resampling of hydrologic time series using multiple predictor variables: A k-nearest neighbour approach, *Advances in Water Resources*, *29*, 987–999.

Muluye, G.Y. (2010), Comparison of statistical methods for downscaling daily precipitation, submitted to *Journal of Hydrology*, Manuscript No. HYDROL10158.

Murphy, A.H., and R.L. Winkler (1987), A general framework for forecast verification, *Mon. Wea. Rev.*, *115*, 1330–1338.

Murphy, A.H., and R.L. Winkler (1992), Diagnostic verification of probability forecasts, *Int. J. Forecasting*, *7*, 435–455.

Murphy, A.H., B.G. Brown, and Y. Chen (1989), Diagnostic verification of temperature forecasts, *Wea. Forecasting*, *4*, 485–501.

Murphy, J.M. (1999), An evaluation of statistical and dynamical techniques for downscaling local climate, *J. Clim.*, *12*, 2256– 2284.

Oelschliigel, B. (1995), A method for downscaling global climate model calculations by a statistical weather generator, *Ecological Modelling*, *82* , 199-204.

Rajagopalan, B., and U. Lall (1999), A k-nearest neighbour simulator for daily precipitation and other weather variables, *Water Resour. Res.*, *35*, 3089–3101.

Richardson, C.W. (1981), Stochastic simulation of daily precipitation, temperature and solar radiation, *Water Resour. Res.*, *17*, 182–90.

Schmidli J., C.M. Goodess, C. Frei, M.R. Haylock, Y. Hundecha, J. Ribalaygua, and T. Schmith (2007), Statistical And Dynamical Downscaling Of Precipitation: An Evaluation And Comparison Of Scenarios For The European Alps, *J. Geophy. Res.*, *112*, 10.1029/2005JD007026.

Semenov, M.A., and E.M. Barrow (1997), Use of a stochastic weather generator in the development of climate change scenarios, *Climatic Change*, *35*, 397–414.

Sharif, M, and D. Burn (2006), Simulating climate change scenarios using an improved K-nearest neighbor model, *J. Hydrol.*, *325*, 179–196.

Sharma, A. (2000), Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: part 3-A nonparametric probabilistic forecast model, *J. Hydrol.*, *239*, 249– 258.

Sharma, A., and R. O'Neill (2002), A nonparametric approach for representing interannual dependence in monthly streamflow sequences, *Water Resour. Res.*, *38*(7), 1100, doi:10.1029/2001WR000953.

Spak, S., T. Holloway, B. Lynn, and R. Goldberg (2007), A comparison of statistical and dynamical downscaling for surface temperature in North America, *J. Geophys. Res.*, *112*, D08101, doi:10.1029/2005JD006712.

Stanski, H.R., L.J. Wilson, and W.R. Burrows (1989), Survey of Common Verification Methods in Meteorology. *WMO World Weather Watch Tech. Report No. 8, WMOI TD No. 358*, 114 pp.

Toth, Z., and E. Kalnay (1993), Ensemble forecasting at NMC: The generation of perturbations, Bull. *Am. Meteorol. Soc.*, *74*, 2317–2330.

von Storch, H., E. Zorita, and U. Cubash (1993), Downscaling of global climate change estimates to regional scales: an application to Iberian rainfall in wintertime, *J. Clim.*, *6*, 1161–71.

Werner, K., D. Brandon, M. Clark, and, S. Gangopadhyay (2005), Incorporating Medium-Range Numerical Weather Model Output into the Ensemble Streamflow Prediction System of the National Weather Service, *J. Hydrometeorol.*, *6*(2), 101-114.

Wilby, R. L., and I. Harris (2006), A framework for assessing uncertainties in climate change impacts: Low-flow scenarios for the river Thames, UK, *Water Resour. Res.*, *42*, W02419, doi:10.1029/2005WR004065.

Wilby, R.L., C.W. Dawson, and E.M. Barrow (2002), SDSM–A decision support tool for the assessment of regional climate impacts, *Environ. Modell. Software*, *17*, 145–157.

Wilby, R.L., L.E. Hay, W.J. Gutowski, and R.W. Arritt (2000), Hydrological responsesto dynamicallya nd statistically downscaled climate model output, *Geophysical Res. Letters.*, *27*(8), 1199-1202.

Wilby, R.L., T.M.L. Wigley (1997), Downscaling general circulation model output: a review of methods and limitations, *Progress in Physical Geography*, *21*, 530–548.

Wilks, D.S. (1995), *Statisistical Methods in the Atmospheric Sciences*, Academic Press, San Diego, California.

Wilks, D.S., and R.L. Wilby (1999), The weather generation game: a review of stochastic weather models, *Progress in Physical Geography*, *23*, 329–357.

Xu, C.Y. (1999), From GCMs to river flow: a review of downscaling methods and hydrologic modeling approaches, *Progress in Physical Geography, 23*(2), 229–249.

Yakowitz, S. (1993), Nearest neighbor regression estimation for null-recurrent Markov time series, *Stochastic Processes Their Appl.*, *48*, 311–318.

Yarnal, B., A.C. Comrie, B. Frakes, and D.P. Brown (2001), Developments and prospects in synoptic climatology, *Int. J. Climatol., 21*, 1923–1950.

Yates, D., S. Gangopadhyay, B. Rajagopalan, and K. Strzepek (2003), A technique for generating regional climate scenarios using a nearest-neighbor algorithm, *Water Resour. Res.*, *39*(7), 1199, doi:10.1029/2002WR001769.

Table 1. Meteorological stations of Saguenay watershed

Meteorological station name	Station location (in degrees)		Altitude (m)	Nearest grid
	Latitude	Longitude		
Bagot	48.3	71	159	(70,47.5)
Benoit	51.53	71.1	549	(70,52.5)
Bonard	50.7	71	506	(70,50)
Chute-Du-Diable	48.75	71.7	174	**(72.5,50)**
Chute-Des-Passes	49.9	71.25	399	(70,50)
Chiba	49.8	74.5	387	(75,50)
Cygnes	49.9	72.9	405	(72.5,50)
Long	50.5	72.95	468	(72.5,50)
Machisque	50.9	71.8	543	(72.5,50)
Metabetchouan	48.4	71.96	220	(72.5,47.5)
Mistassibi2	49.4	71.9	183	(72.5,50)
Normandin	48.83	72.55	137	(72.5,50)
Roberval	48.5	72.27	179	(72.5,47.5)

Table 2. NOAA reforecast ensemble variable fields

Variable Field	Description	Surface level (mb)	Grid
1. *apcp*	Accumulated precipitation (mm)	Surface	Latlon
2. *heating*	Vertically integrated diabatic heating (K/s/mb)	Vertical average	Latlon
3. *pwat*	Precipitable water	Surface	Latlon
4. *prmsl*	Pressure reduced to mean sea-level (Pa)	Surface	Latlon
5. *t2m*	Temperature at 2 meters (K)	Surface	Latlon
6. *rhum*	Relative humidity (%)	700 mb	Latlon
7. *u10m*	Zonal wind at 10 meters (m/s)	Surface	Latlon
8. *v10m*	Meridional wind at 10 meters (m/s)	Surface	Latlon

Table 3. Comparative performance statistics for downscaling daily precipitation

Model	Diagnostic	Forecast range (days)														
		F0	**F1**	**F2**	**F3**	**F4**	**F5**	**F6**	**F7**	**F8**	**F9**	**F10**	**F11**	**F12**	**F13**	**F14**
	Bias (%)	**-6.69**	-10.35	-7.1	-2.45	-3.25	**-2.72**	-14.6	**-5.58**	11.2	**-0.58**	**-8.13**	-12.43	**3.91**	-12.9	-8.5
	RMSE	7.46	7.21	7.38	7.65	7.79	7.44	6.9	7.55	7.21	7.63	7.24	7.46	8	6.97	7.53
	r	0.08	0.03	0.06	0	0.03	0	0.03	0	0.05	0.03	0.04	0	0.01	0.08	0.01
1NN-E	RV	-0.78	-0.67	-0.74	-0.88	-0.95	-0.77	-0.52	-0.82	-0.66	-0.87	-0.68	-0.78	-1.05	-0.56	-0.81
	POD	0.51	0.53	0.52	0.52	0.49	0.55	0.51	0.51	0.5	0.55	0.5	0.5	0.53	0.51	0.52
	FAR	0.48	0.47	0.48	0.5	0.51	0.49	0.5	0.49	0.49	0.49	0.49	0.49	0.48	0.49	0.5
	BiasF	0.98	1.01	**1**	1.06	0.99	1.07	1.03	**1**	**0.99**	1.07	**0.99**	**1**	1.02	**0.99**	1.03
	Bias (%)	-10.13	-16.08	**-1.75**	-10.52	-22.07	-8.48	**-0.78**	-8.95	-8.52	-4.73	-16.86	-7.27	-6.22	**-0.25**	-12.3
	RMSE	6.32	6.78	7.04	6.87	6.82	7.53	7.67	7.31	7.42	7.77	7.12	7.38	7.54	7.64	7.48
	r	0.33	0.22	0.16	0.17	0.06	0.06	0.04	0.05	0.03	-0.02	0.03	0.05	0.03	0.03	0
1NN-M	RV	-0.28	-0.47	-0.58	-0.51	-0.49	-0.82	-0.88	-0.71	-0.77	-0.94	-0.62	-0.75	-0.82	-0.87	-0.79
	POD	0.62	0.6	0.59	0.59	0.55	0.52	0.54	0.5	0.53	0.52	0.51	0.52	0.54	0.53	0.52
	FAR	**0.37**	**0.36**	**0.43**	**0.41**	**0.45**	**0.47**	**0.48**	**0.49**	**0.47**	**0.49**	**0.49**	**0.48**	**0.46**	**0.47**	**0.49**
	BiasF	0.97	0.94	1.04	**1.01**	**1**	**1**	1.04	0.98	**1.01**	1.02	0.98	**1.01**	**0.99**	**1.02**	**1.01**
	Bias (%)	-14.29	-8.13	-8.91	**0.41**	5.15	-4.21	-14.35	-7.89	**-6.67**	-13.89	-11.73	-12.66	-7.2	-7.99	-17.46
	RMSE	7.13	7.63	7.35	7.21	7.97	7.68	7.08	7.58	7.58	7.07	7.31	7.31	7.6	7.55	7.07
	r	0.03	-0.01	0.03	0.07	0.05	-0.01	0.03	0.02	0.01	0.03	0.03	0.01	0	0.03	0.02
KNN-E	RV	-0.63	-0.87	-0.73	-0.66	-1.04	-0.89	-0.6	-0.84	-0.84	-0.6	-0.71	-0.71	-0.85	-0.83	-0.6
	POD	0.5	0.5	0.52	0.54	0.52	0.53	0.55	0.53	0.5	0.5	0.52	0.49	0.5	0.51	0.47
	FAR	0.49	0.5	0.48	0.48	0.5	0.51	0.46	0.49	0.5	0.5	0.47	0.52	0.49	0.48	0.5
	BiasF	0.98	1.02	1.01	1.05	1.03	1.07	**1**	1.03	**1**	**0.99**	0.98	1.02	0.98	0.98	0.94
	Bias (%)	-13.58	-17.02	-14.2	-8.79	-15.07	-6.54	-12.47	-7.39	-9.65	-9.45	-15.44	-10.58	-8.48	-8.23	**-5.31**
	RMSE	6.33	6.35	6.61	7.02	7.5	7.75	7.25	7.54	7.39	7.32	7.17	7.33	7.4	7.27	7.56
	r	0.27	0.24	0.17	0.17	0.03	0.03	0.03	-0.01	0.06	0	0.03	0.01	0.02	0.06	0.01
KNN-M	RV	-0.28	-0.29	-0.4	-0.58	-0.8	-0.92	-0.68	-0.82	-0.75	-0.71	-0.65	-0.72	-0.76	-0.69	-0.83
	POD	0.58	0.61	0.56	0.58	0.53	0.53	0.53	0.5	0.53	0.53	0.51	0.53	0.52	0.53	0.5
	FAR	0.42	0.39	**0.44**	0.43	**0.46**	**0.48**	**0.48**	0.51	**0.48**	**0.49**	**0.48**	**0.5**	**0.47**	**0.48**	0.51
	BiasF	**0.99**	**1**	**1**	1.02	0.99	1.01	1.02	1.02	1.02	1.04	0.98	1.07	**0.99**	1.03	**1.02**
	Bias (%)	-12.62	**-5.91**	-7.46	-2.33	**1.31**	-6.77	-9.34	-7.02	-9.67	-16.27	-11.78	-9.3	-9.29	-3.46	-8.03
	RMSE	5.85	5.92	5.9	5.83	5.92	5.82	5.83	5.8	**5.77**	**5.74**	**5.79**	**5.76**	**5.76**	**5.76**	5.91
	r	0.03	0.04	0.04	0.05	0.07	0.07	0.04	0.07	0.06	0.06	0.05	0.07	0.06	0.1	0.01
KNN-EW	RV	-0.1	-0.12	-0.12	-0.09	-0.12	2.48	-0.09	-0.08	**-0.06**	**-0.05**	**-0.07**	**-0.06**	**-0.06**	**-0.06**	**-0.12**
	POD	**0.98**	**0.98**	**0.99**	**0.99**	**0.98**	0.98	**0.99**	**0.99**	**0.99**	0.98	**0.99**	**0.99**	**0.99**	**1**	0.98
	FAR	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
	BiasF	1.96	1.94	1.97	1.98	1.97	1.98	1.97	1.99	1.98	1.96	1.98	1.98	1.98	1.99	1.97
	Bias (%)	-19.55	-19.16	-12.97	-14	-16.26	-9.78	-6.88	-10.23	-10.23	-8.28	-11.95	**-7.06**	-7.89	-5.35	-6.89
	RMSE	**4.83**	**5.1**	**5.34**	**5.41**	**5.61**	**5.74**	**5.83**	**5.73**	**5.79**	**5.72**	**5.8**	**5.8**	**5.78**	**5.84**	**5.86**
	r	**0.51**	**0.42**	**0.32**	**0.29**	**0.17**	**0.12**	**0.07**	**0.1**	**0.07**	**0.02**	**0.05**	**0.07**	**0.08**	**0.06**	**0.02**
KNN-MW	RV	**0.25**	**0.17**	**0.09**	**0.06**	**-0.01**	**-0.06**	**-0.09**	**-0.05**	**-0.07**	**-0.11**	**-0.08**	**-0.08**	**-0.07**	**-0.09**	**-0.1**
	POD	0.95	0.95	0.96	0.97	**0.98**	0.98	**0.99**	**0.99**	0.98	**0.99**	**0.99**	**0.99**	**0.99**	0.99	**0.99**
	FAR	0.46	0.46	0.48	0.49	0.49	0.5	0.5	0.49	0.5	0.5	0.5	0.5	0.5	0.5	0.5
	BiasF	1.76	1.77	1.84	1.9	1.93	1.95	1.97	1.95	1.97	1.98	1.97	1.97	1.98	1.98	1.97

Table 4. Comparative performance statistics for downscaling daily maximum temperature

Model	Forecast range (days)														
	Diagnostic	**F0**	**F1**	**F2**	**F3**	**F4**	**F5**	**F6**	**F7**	**F8**	**F9**	**F10**	**F11**	**F12**	**F13**	**F14**
1NN-E	Bias (%)	-11.58	-9.32	-14.09	-13.92	-9.72	-10.27	-9.2	-10.51	-12.08	-10.04	-12.48	-14.67	-14.36	-12.6	-15.3
	RMSE	7.54	7.79	7.47	7.43	7.56	7.71	7.72	7.69	7.78	7.46	7.6	7.76	7.84	7.55	7.83
	r	0.83	0.82	0.83	0.83	0.83	0.83	0.82	0.83	0.82	0.83	0.83	0.83	0.82	0.83	0.83
	RV	0.66	0.63	0.66	0.67	0.65	0.64	0.64	0.64	0.64	0.66	0.65	0.64	0.63	0.66	0.66
1NN-M	Bias (%)	**-6.41**	**-6.41**	**-7.33**	**-8.34**	**-9.23**	**-10.06**	**-8.69**	**-9.03**	-10.76	-7.91	-11.24	-13.04	-10.15	-9.4	**-9.15**
	RMSE	4.11	4.11	4.72	5.5	6	6.4	6.73	6.88	7.11	7.35	7.28	7.35	7.53	7.3	7.91
	r	0.95	0.95	0.93	0.91	0.89	0.88	0.86	0.86	0.85	0.84	0.84	0.84	0.83	0.84	0.83
	RV	0.9	0.9	0.87	0.82	0.78	0.75	0.73	0.71	0.7	0.67	0.68	0.67	0.66	0.68	0.65
KNN-E	Bias (%)	-10.92	-10.5	-10.78	-12.91	-9.17	-10.72	-9.5	-10.66	-9.22	-11.68	-3.18	-14.4	-14.06	-15.57	-12.25
	RMSE	7.53	7.48	7.94	7.61	7.58	7.52	7.6	7.68	7.46	7.71	7.55	7.43	7.7	7.67	7.56
	r	0.83	0.83	0.81	0.83	0.83	0.83	0.83	0.82	0.83	0.83	0.83	0.81	0.83	0.83	0.83
	RV	0.66	0.66	0.62	0.65	0.65	0.66	0.65	0.64	0.66	0.64	0.66	0.67	0.64	0.64	0.66
KNN-M	Bias (%)	-10.08	-8.17	-8.39	-10.42	-10.39	-7.93	-10	-9.552	-9.11	**-7**	-11.71	**-9.81**	**-7.41**	**-8.15**	-11.17
	RMSE	4.63	4.72	5.02	5.71	6.13	6.48	6.88	6.95	7.21	7.35	7.47	7.47	7.4	7.51	7.67
	r	0.94	0.93	0.92	0.9	0.89	0.87	0.86	0.85	0.84	0.84	0.83	0.83	0.83	0.83	0.83
	RV	0.87	0.87	0.85	0.8	0.77	0.75	0.71	0.71	0.69	0.67	0.66	0.66	0.67	0.66	0.65
KNN-EW	Bias (%)	-10.99	-11.61	-11.48	-11.16	-10.66	-12.73	-11.82	-9.84	**-9.95**	-11.08	-6.82	-14.05	-12.04	-11.18	-11.33
	RMSE	5.77	5.79	5.7	5.81	5.73	5.63	5.67	5.75	5.52	5.66	5.65	5.678	**5.63**	5.56	**5.63**
	r	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.91	0.9	0.9	**0.9**	**0.9**	**0.9**	**0.9**
	RV	0.8	0.8	0.8	0.8	0.8	0.81	0.81	0.8	0.82	0.81	0.81	**0.81**	**0.81**	**0.81**	**0.81**
KNN-MW	Bias (%)	-8.1	-8.5	-8.99	-9.85	-10.08	-10.1	-10.51	-10.31	-10.78	-9	-10.76	-11.46	-9.54	-9.95	-10.12
	RMSE	**3.1**	**3.23**	**3.55**	**4.04**	**4.5**	**4.81**	**5.11**	**5.16**	**5.27**	**5.45**	**5.47**	**5.58**	**5.63**	**5.52**	**5.65**
	r	**0.97**	**0.97**	**0.96**	**0.95**	**0.94**	**0.93**	**0.92**	**0.92**	**0.91**	**0.91**	**0.91**	**0.9**	**0.9**	**0.91**	**0.9**
	RV	**0.94**	**0.94**	**0.92**	**0.9**	**0.88**	**0.86**	**0.84**	**0.84**	**0.83**	**0.82**	**0.82**	**0.81**	**0.81**	**0.82**	**0.81**

Table 5. Comparative performance statistics for downscaling daily minimum temperature

Model	Forecast range (days)														
	Diagnostic	**F0**	**F1**	**F2**	**F3**	**F4**	**F5**	**F6**	**F7**	**F8**	**F9**	**F10**	**F11**	**F12**	**F13**	**F14**
1NN-E	Bias (%)	16.07	9.41	31.31	19.8	12.97	10.62	15.37	19.97	16.88	8.47	23.17	30.62	19.48	17.2	25.4
	RMSE	8.12	8.46	8.35	8.11	8.45	8.64	8.81	9.02	8.59	8.41	8.51	8.68	8.75	8.55	8.61
	r	0.82	0.8	0.81	0.82	0.8	0.8	0.79	0.78	0.8	0.8	0.8	0.8	0.79	0.8	0.79
	RV	0.61	0.58	0.59	0.61	0.58	0.56	0.54	0.52	0.56	0.58	0.57	0.55	0.55	0.57	0.56
1NN-M	Bias (%)	**3.08**	**3.08**	6.82	**9.9**	8.98	20.9	**5.99**	7.8	12.07	**0.36**	23.71	16.69	15.24	9.66	**2.38**
	RMSE	4.91	4.91	5.3	5.86	6.72	7.21	7.6	7.44	8.1	8.26	8.11	8.14	8.73	8.26	8.52
	r	0.93	0.93	0.92	0.91	0.87	0.86	0.84	0.81	0.82	0.81	0.82	0.82	0.79	0.81	0.8
	RV	0.86	0.86	0.83	0.8	0.73	0.69	0.66	0.67	0.61	0.6	0.61	0.61	0.55	0.6	0.57
KNN-E	Bias (%)	21.78	7.54	7.79	16.18	2.99	7.04	14.6	**6.94**	11.36	19.54	19.76	20.68	29.05	37.45	12.86
	RMSE	8.96	8.48	8.64	8.68	9.08	8.28	8.49	8.43	8.52	8.51	8.59	8.25	8.86	8.69	8.4
	r	0.78	0.8	0.79	0.79	0.78	0.8	0.8	0.8	0.8	0.8	0.79	0.82	0.8	0.81	0.8
	RV	0.52	0.57	0.56	0.55	0.51	0.59	0.57	0.58	0.57	0.57	0.56	0.6	0.54	0.55	0.58
KNN-M	Bias (%)	12.3	9.99	**6.59**	14.36	18.04	**3.00**	15.95	16.81	**5.62**	2.39	24.88	**3.27**	**4.51**	**2.29**	13.68
	RMSE	5.51	5.51	5.67	6.17	6.71	7.24	7.83	8.02	7.8	8.13	8.19	8.44	8.36	8.37	8.5
	r	0.91	0.92	0.91	0.9	0.88	0.85	0.83	0.82	0.83	0.82	0.81	0.8	0.8	0.8	0.8
	RV	0.82	0.82	0.81	0.77	0.73	0.69	0.64	0.62	0.64	0.61	0.6	0.58	0.59	0.59	0.57
KNN-EW	Bias (%)	16.7	7.1	9.91	14.58	**3.48**	10.5	15.02	7.11	15.94	12.46	**-5.2**	16.94	23.39	19.68	12.27
	RMSE	6.27	6.26	6.18	6.31	6.29	6.02	6.17	6.21	6.14	6.14	6.1	6.07	6.3	6.14	6.16
	r	0.88	0.88	0.88	0.88	0.88	0.89	0.88	0.88	0.89	**0.89**	**0.88**	**0.89**	**0.88**	**0.89**	**0.88**
	RV	0.77	0.77	0.77	0.76	0.77	0.79	0.77	0.77	0.78	**0.78**	**0.78**	**0.78**	**0.76**	**0.78**	**0.78**
KNN-MW	Bias (%)	8.9	11.37	12.08	14.47	11.62	14.72	11.22	12.76	11.8	8.18	17.05	13.99	9.97	7.63	8.63
	RMSE	**3.59**	**3.62**	**3.87**	**4.27**	**4.71**	**5.27**	**5.41**	**5.5**	**5.74**	**5.94**	**5.93**	**6.02**	**6.13**	**6.06**	**6.14**
	r	**0.96**	**0.96**	**0.96**	**0.95**	**0.93**	**0.92**	**0.91**	**0.91**	**0.9**	**0.89**	**0.89**	**0.89**	**0.88**	**0.89**	**0.88**
	RV	**0.92**	**0.92**	**0.91**	**0.89**	**0.87**	**0.84**	**0.83**	**0.82**	**0.8**	**0.79**	**0.79**	**0.79**	**0.78**	**0.78**	**0.78**

Table 6. Test for position (i.e. median) and variance for downscaling daily precipitation corresponding to the first member and zero forecast range using Mann-Whitney and Levene, respectively, for the test period Jan 97 to Dec 01. Note that the values in the table represent *p*-values.

Month	Test for Position (Mann-Whitney)		Test for Variance (Levene)	
	1NN-M	KNN-WM	1NN-M	KNN-WM
Jan	0.164	0.064	0.839	0.052
Feb	0.361	0.001	0.970	0.011
Mar	0.813	0.036	0.278	0.318
Apr	0.845	0.000	0.469	0.200
May	0.562	0.000	0.684	0.058
Jun	0.982	0.000	0.591	0.043
Jul	0.636	0.000	0.602	0.033
Aug	0.361	0.001	0.051	0.001
Sep	0.819	0.002	0.567	0.004
Oct	0.537	0.000	0.166	0.004
Nov	0.295	0.038	0.597	0.015
Dec	0.783	0.001	0.811	0.017

Table 7. Comparative performance statistics using conventional and distribution-based diagnostic measures. The downscaled daily precipitation is compared to both the raw model output and climatology. The bold statistics represent improvements over the raw due to downscaling

Diagnostic Measure	Forecast range		
	F0	F3	F7
Bias_med_R (%)	-28.0	-40.7	-44.8
Bias_med_D (%)	12.3	9.7	8.6
Bias_med_RD (%)	**56%**	**76%**	**81%**
BS_R	0.28	0.31	0.34
BS_D	0.25	0.23	0.25
BSS_R	0.44	0.37	0.32
BSS_D	0.50	0.54	0.49
BSS_RD (%)	**11%**	**26%**	**26%**
RPS_R	0.94	1.06	1.19
RPS_D	0.91	0.91	0.75
RPSS_R	0.53	0.47	0.4
RPSS_D	0.54	0.54	0.5
RPSS_RD (%)	**3%**	**14%**	**37%**
ROCS_R	0.81	0.71	0.54
ROCS_D	0.77	0.71	0.55
ROCSS_R	0.62	0.42	0.07
ROCSS_D	0.54	0.41	0.11

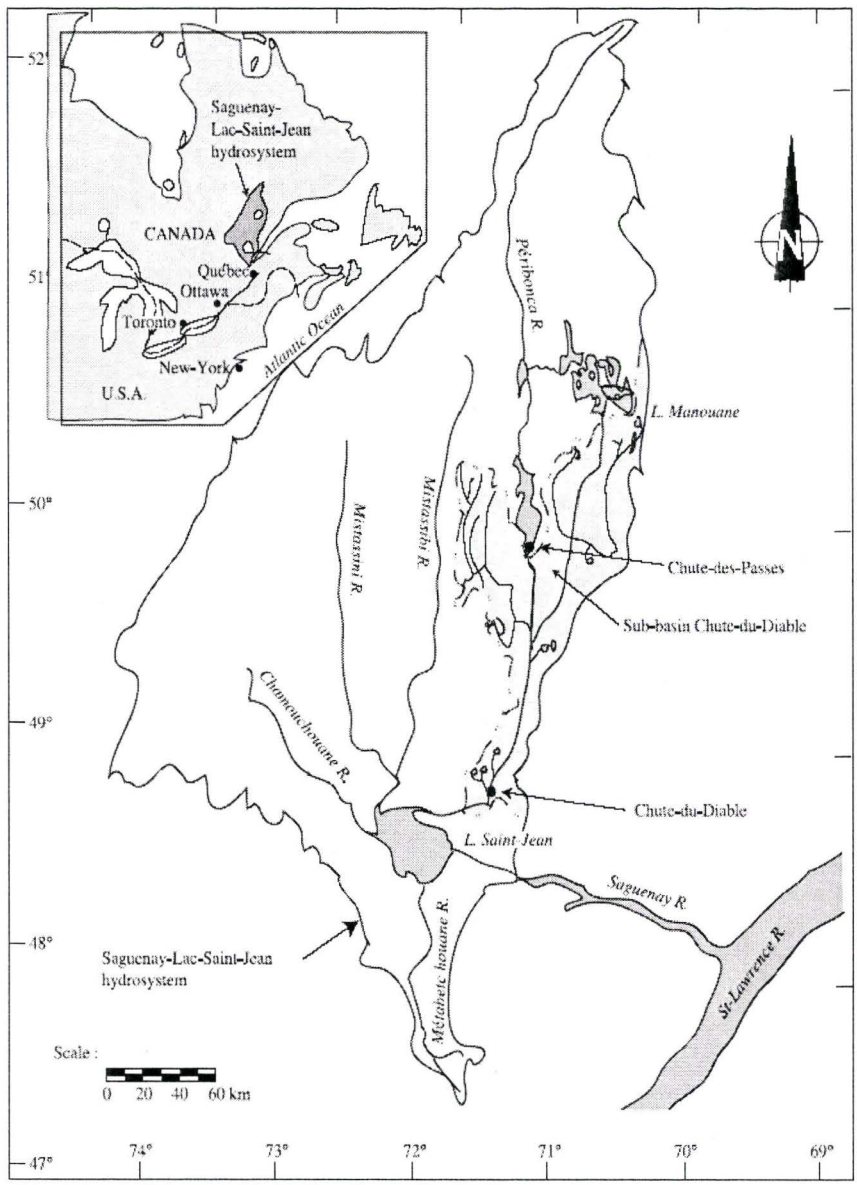


Figure 1. Location map of study area (Source: Dibike and Coulibaly, 2005)

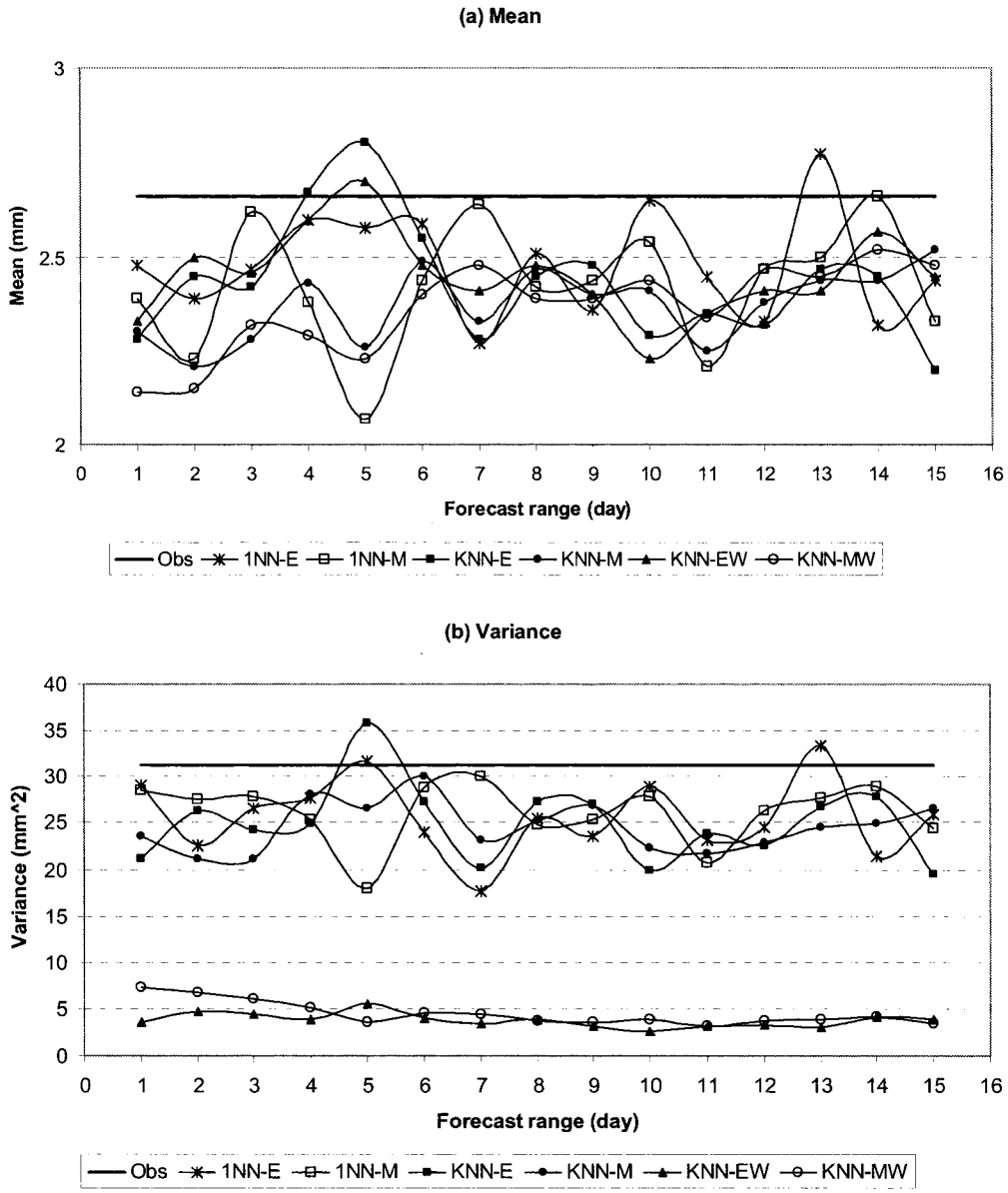


Figure 2. Comparison of downscaled precipitation derived from numerical weather prediction model output: (a) mean precipitation totals, and (b) variance of daily precipitation, 1997-2001.

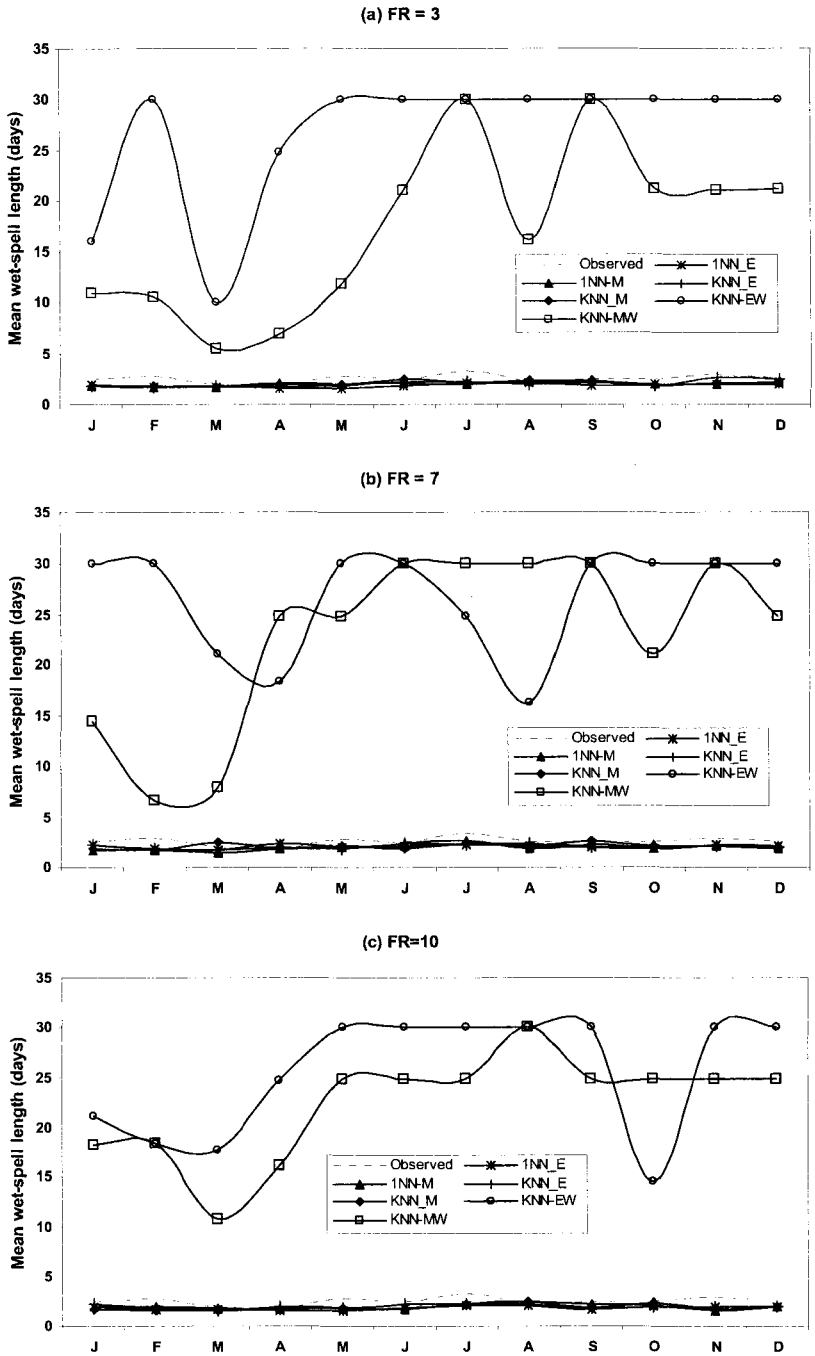


Figure 3. Mean length of wet-spells derived from numerical weather prediction model output for forecast ranges (FR) of (a) 3, (b) 7, and (c) 10 days, 1997–2001.

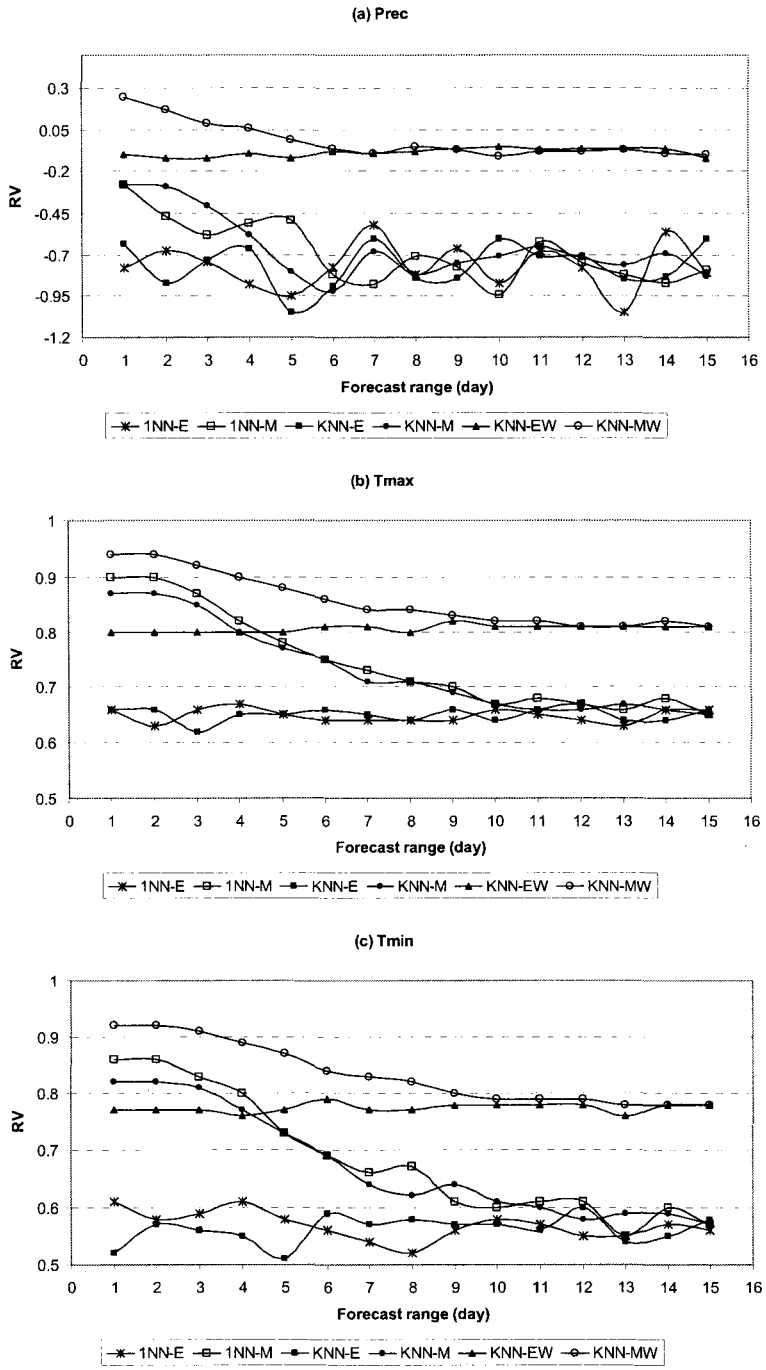
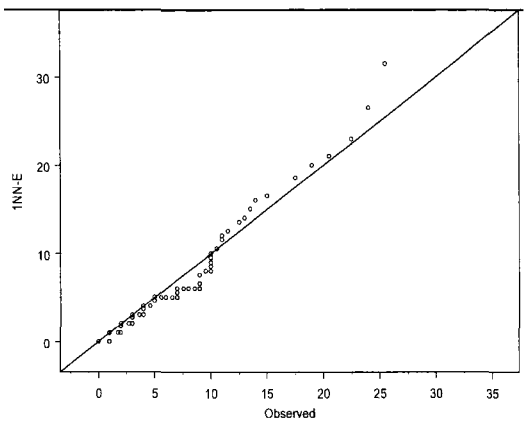
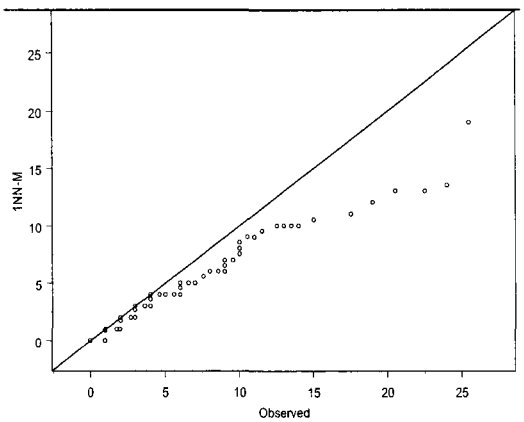
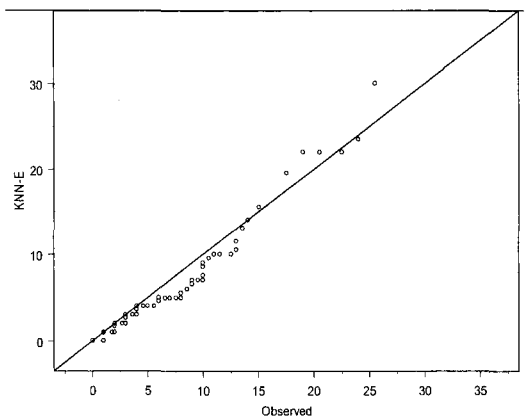


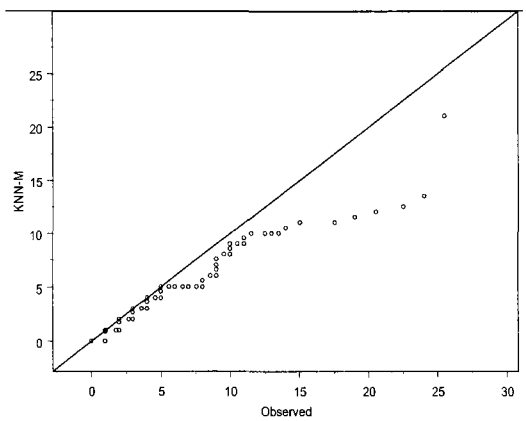
Figure 4. Reduction of variance (RV) with forecast range (FR) as downscaled by the different models for (a) precipitation, (b) maximum temperature, and (c) minimum temperature, 1997-2001.

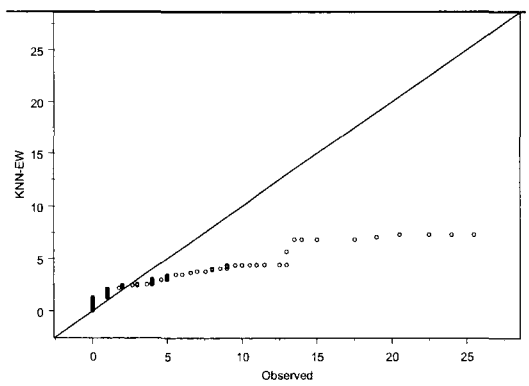
(a) Winter

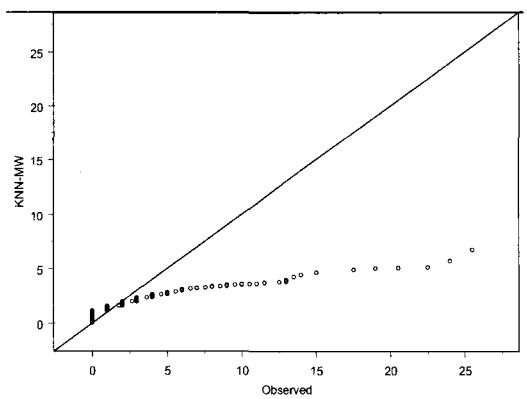

1NN-E
Observed


1NN-M
Observed

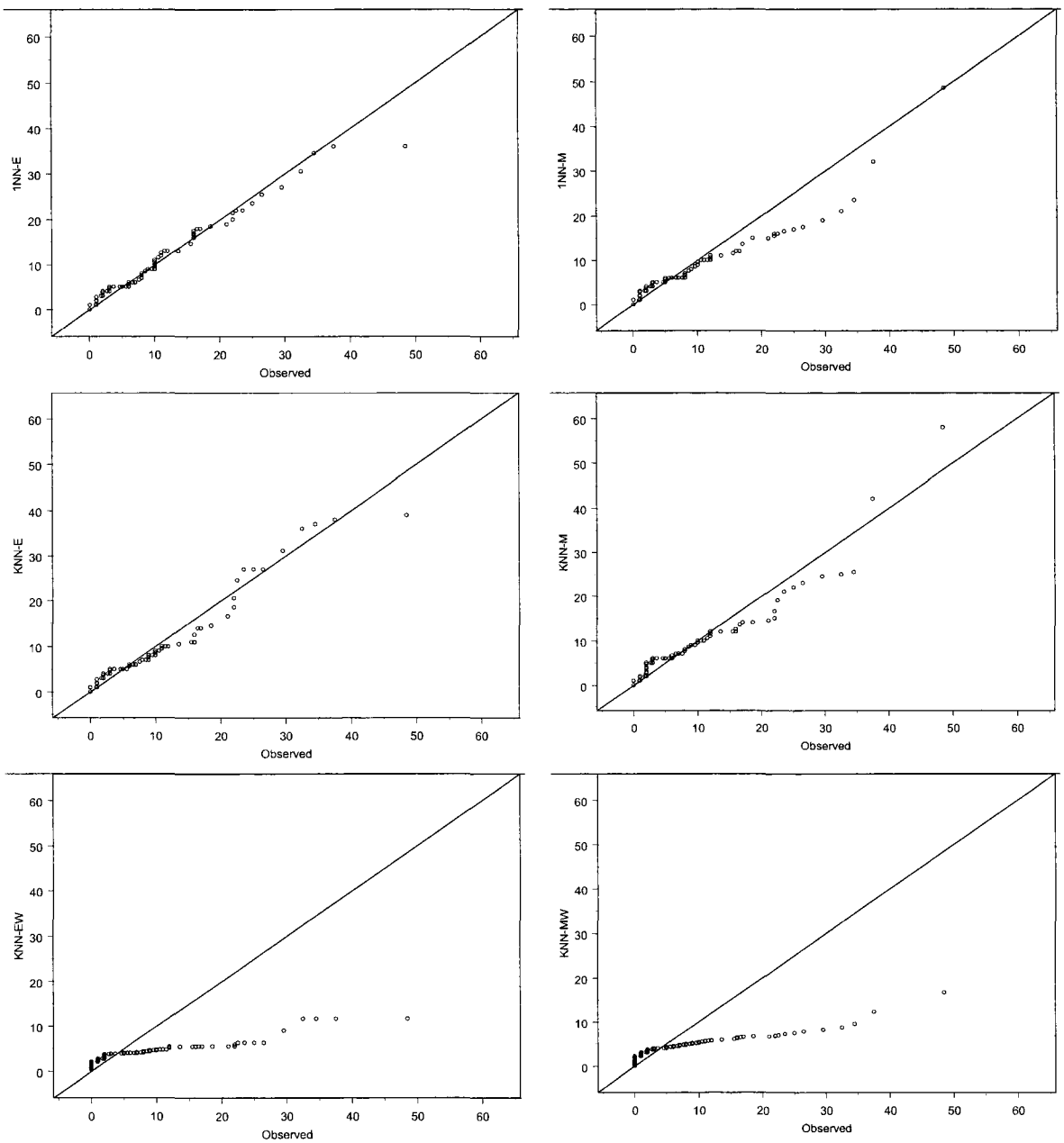

KNN-E
Observed


KNN-M
Observed

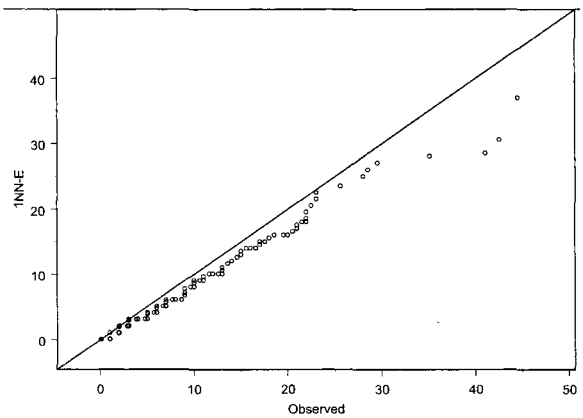

KNN-EW
Observed

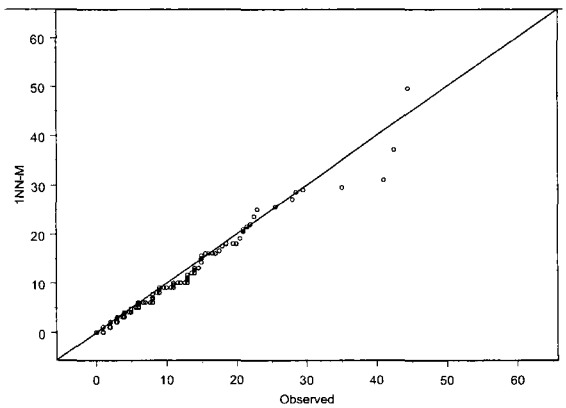

KNN-MW
Observed

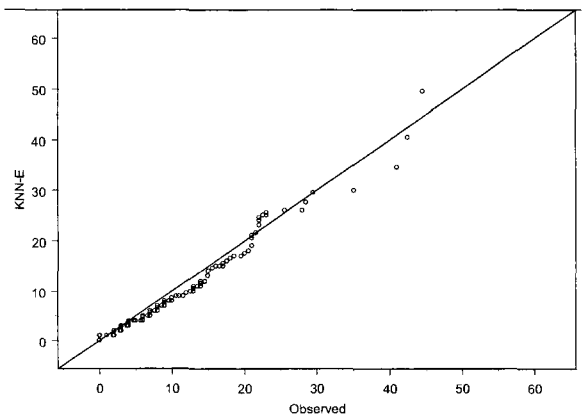
(b) Spring

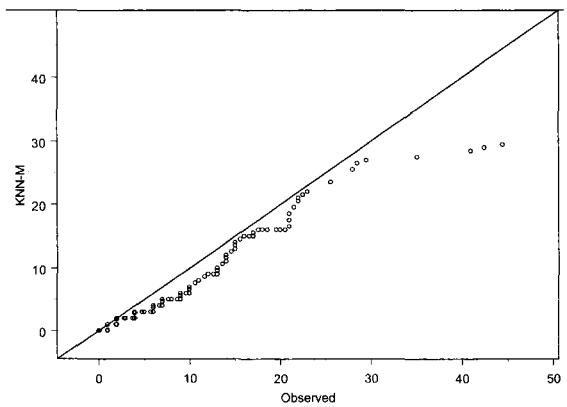


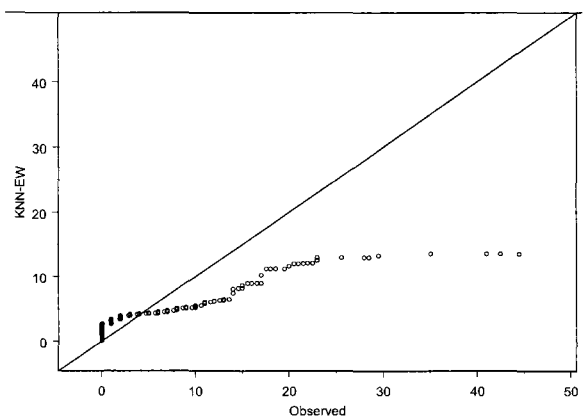
(c) Summer

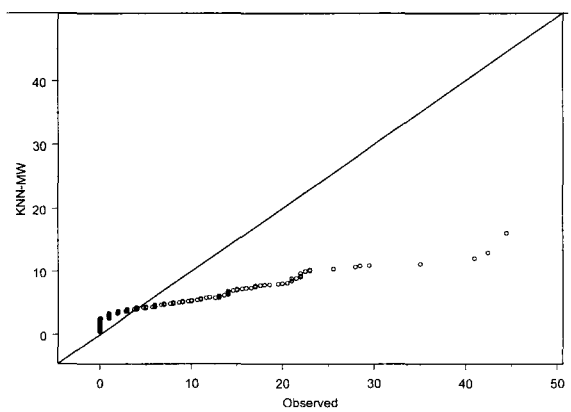

1NN-E
Observed
0
10
20
30
40
50


1NN-M
Observed
0
10
20
30
40
50
60


KNN-E
Observed
0
10
20
30
40
50
60


KNN-M
Observed
0
10
20
30
40
50


KNN-EW
Observed
0
10
20
30
40
50


KNN-MW
Observed
0
10
20
30
40
50

(d) Fall

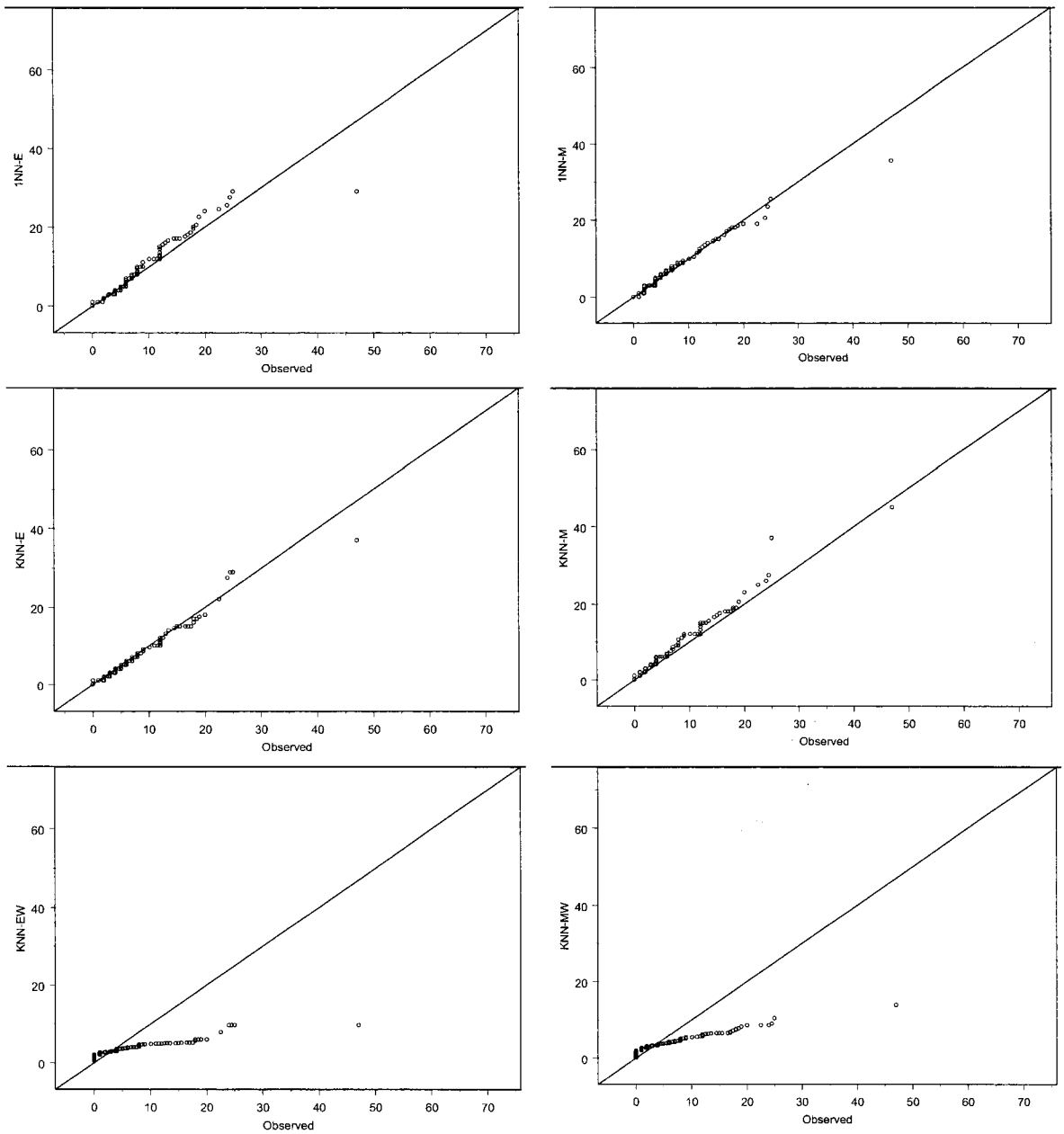
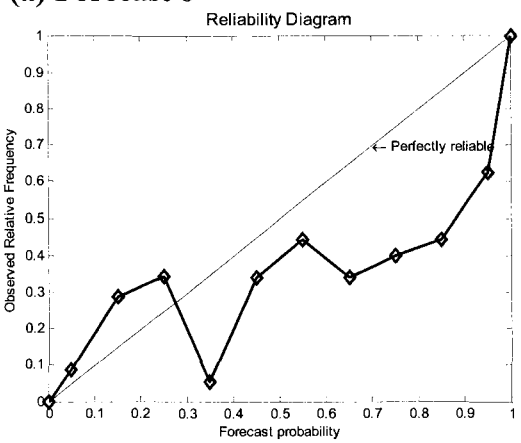
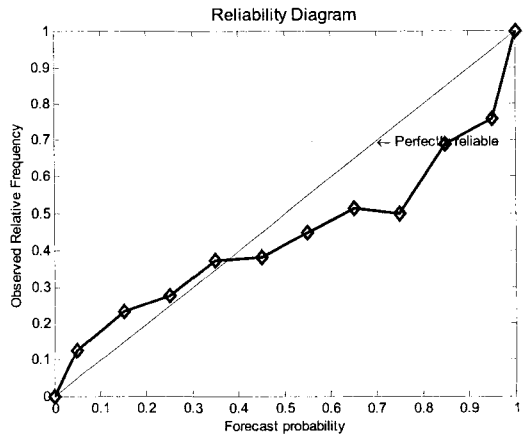
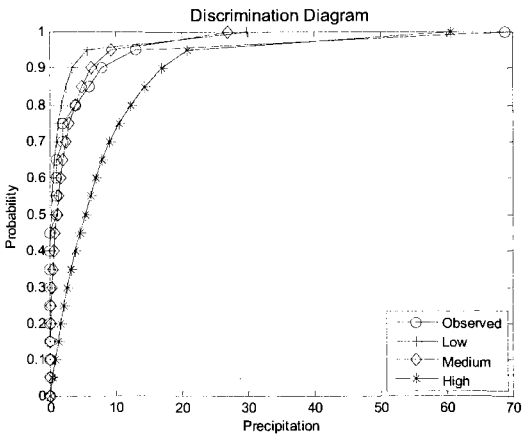


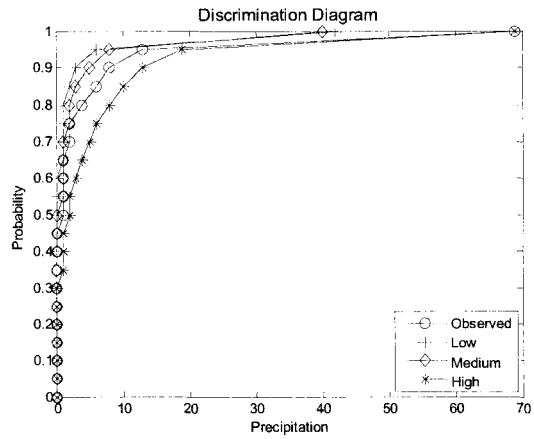
Figure 5. Quantile-quantile (q-q) plots of the quantiles of the observed precipitation (mm) against the quantiles of the simulated precipitation (mm) as downscaled by the different models, Jan 1997 to Dec 2001 for forecast range 7 for a) winter (JFM), b) spring (AMJ), c) summer (JAS), and d) fall (OND).

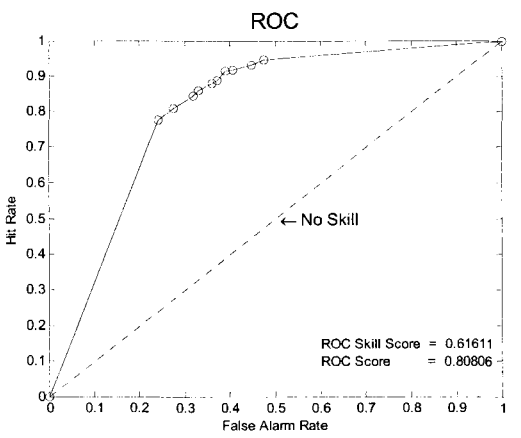
(a) Forecast 0

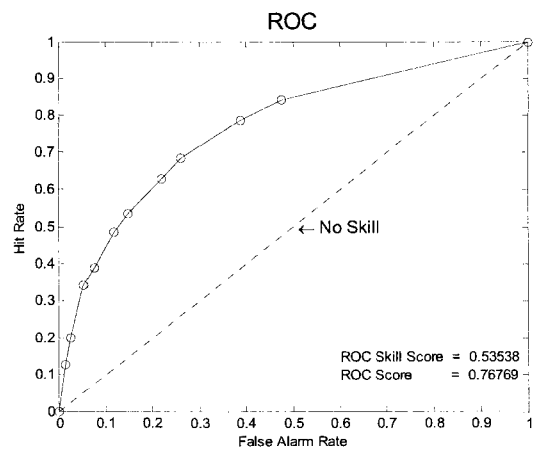

Reliability Diagram
Observed Relative Frequency
Forecast probability
← Perfectly reliable


Reliability Diagram
Observed Relative Frequency
Forecast probability
← Perfectly reliable

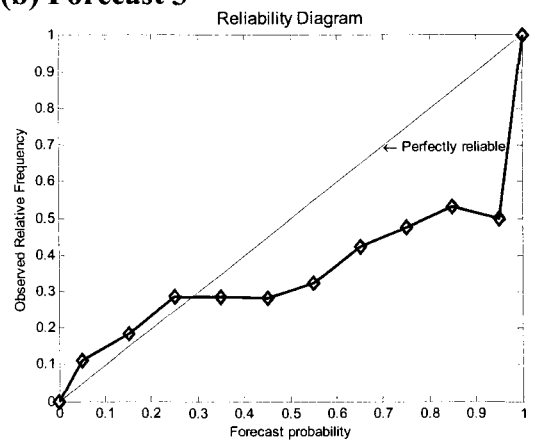

Discrimination Diagram
Probability
Precipitation
Observed
Low
Medium
High

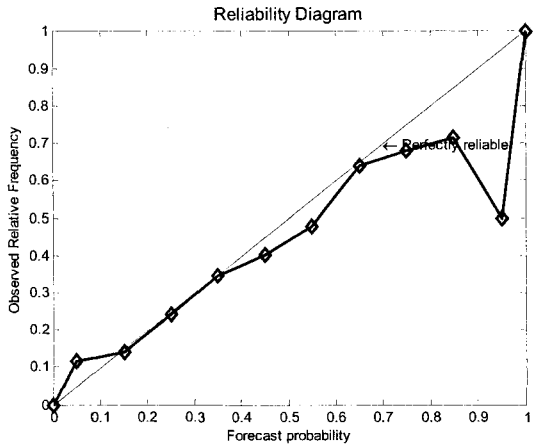

Discrimination Diagram
Probability
Precipitation
Observed
Low
Medium
High

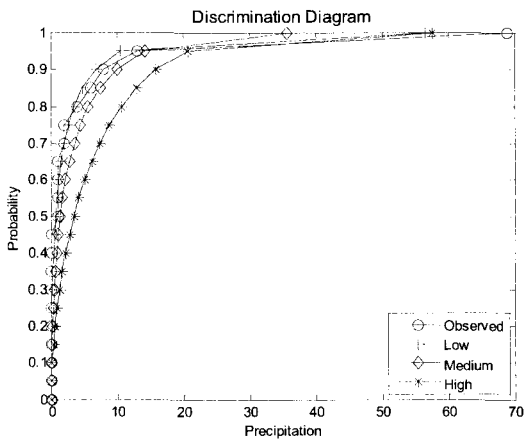

ROC
Hit Rate
False Alarm Rate
← No Skill
ROC Skill Score = 0.61611
ROC Score = 0.80806

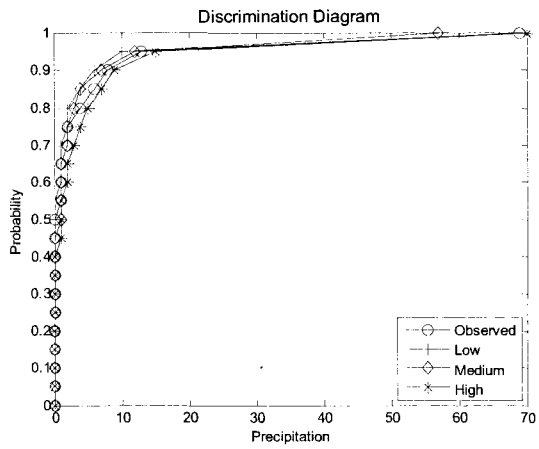

ROC
Hit Rate
False Alarm Rate
← No Skill
ROC Skill Score = 0.53538
ROC Score = 0.76769

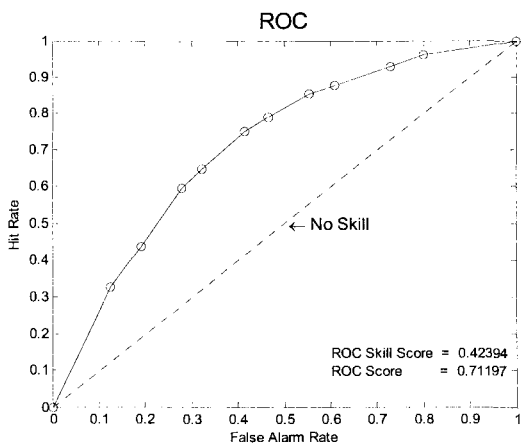
(b) Forecast 3

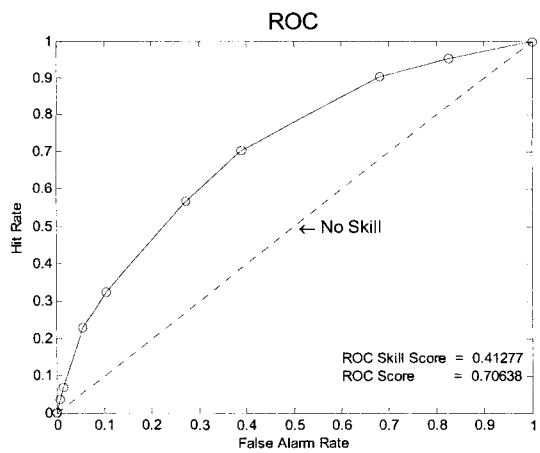

Reliability Diagram
Observed Relative Frequency
Forecast probability
← Perfectly reliable


Reliability Diagram
Observed Relative Frequency
Forecast probability
← Perfectly reliable

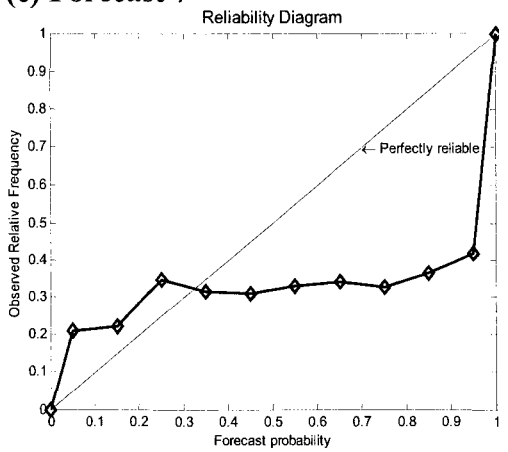

Discrimination Diagram
Probability
Precipitation
Observed
Low
Medium
High

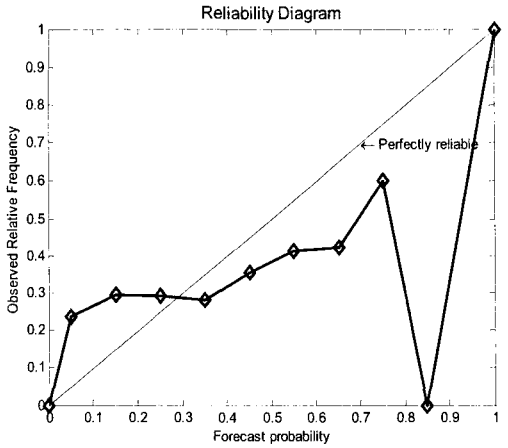

Discrimination Diagram
Probability
Precipitation
Observed
Low
Medium
High

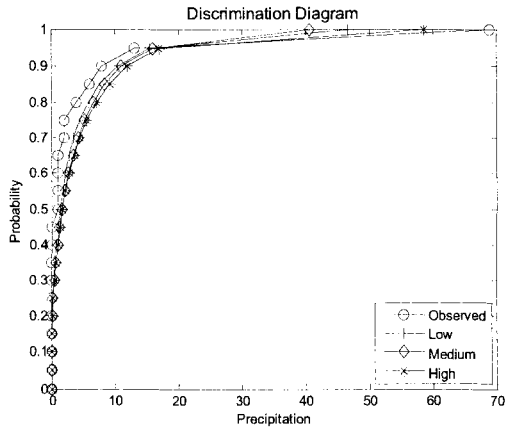

ROC
Hit Rate
False Alarm Rate
← No Skill
ROC Skill Score = 0.42394
ROC Score = 0.71197

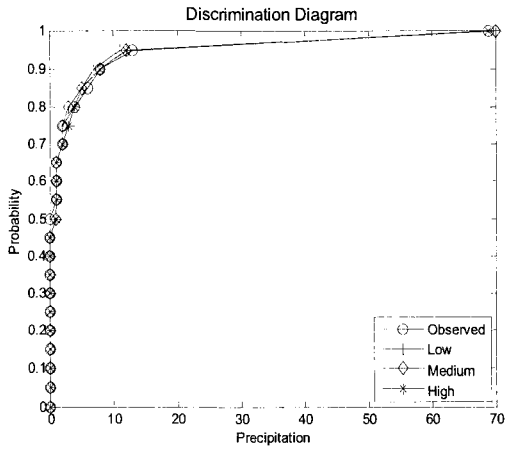

ROC
Hit Rate
False Alarm Rate
← No Skill
ROC Skill Score = 0.41277
ROC Score = 0.70638

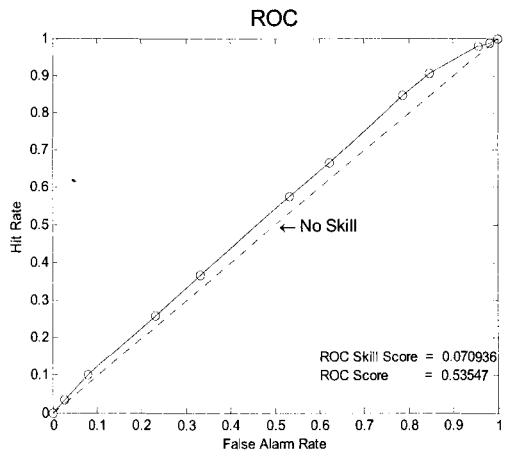
(c) Forecast 7


Reliability Diagram
Observed Relative Frequency
Forecast probability
Perfectly reliable


Reliability Diagram
Observed Relative Frequency
Forecast probability
Perfectly reliable


Discrimination Diagram
Probability
Precipitation
Observed
Low
Medium
High


Discrimination Diagram
Probability
Precipitation
Observed
Low
Medium
High


ROC
Hit Rate
False Alarm Rate
No Skill
ROC Skill Score = 0.070936
ROC Score = 0.53547

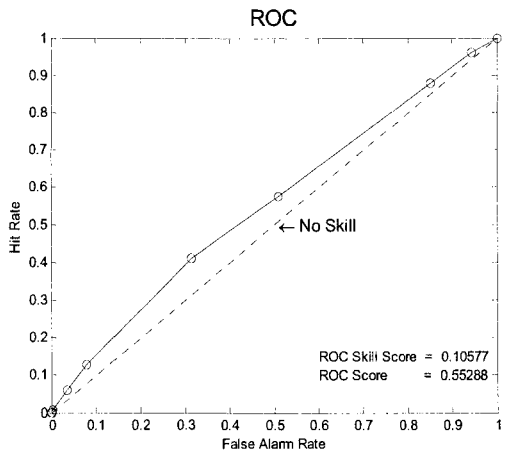

ROC
Hit Rate
False Alarm Rate
No Skill
ROC Skill Score = 0.10577
ROC Score = 0.55288

Figure 6. Comparison of the downscaled daily precipitation data derived from the numerical weather prediction model output versus the raw model output using reliability diagram, discrimination, and relative operating characteristics (ROC). The graph on the left represent the raw model output and the graph on the right represent the downscaled output for forecast ranges a) 0, b) 3, and c) 7, 1997-2001.

CHAPTER 6

Implications of Medium-Range Numerical Weather Model Output in Hydrologic Applications: Assessment of Skill and Economic Value

Chapters 4 and 5 identified suitable and adequate downscaling models for the study basin. The downscaled precipitation and temperature values are forced into an HBV hydrologic model in order to generate an ensemble of reservoir inflows. The skill values of downscaled-based flows are compared against the raw-based flows. The economic values associated with probabilistic flow forecasts are also compared to the deterministic flow forecasts in terms of the overall economic value and the range of end-users that can benefit from the respective forecast systems. The findings of the study are presented in a paper form and submitted to Journal of Hydrology for possible publication. This forms the final task on hydrologic forecasting and demonstrates how proper use of ensemble atmospheric forecasts in a hydrologic forecast system improves the reliability of hydrologic forecasts.

Implications of Medium-Range Numerical Weather Model Output in Hydrologic Applications: Assessment of Skill and Economic Value

Abstract

Integrating medium range numerical weather model output into hydrologic models has been shown to yield useful information. The emphases in this study were to (i) develop effective mechanisms for integrating meteorological ensemble systems into hydrologic forecasts; and (ii) evaluate the skill and economic values of the subsequent hydrologic forecasts. The general framework was demonstrated under a number of conditioning paradigms for the Chute-du-Diable watershed located in Quebec, Canada.

For this purpose, two downscaling models were employed to generate station total daily precipitation data and average daily temperature data. The performances of the downscaled outputs were compared against the RAW model output using suites of diagnostic measures. The comparative results indicated that the downscaled outputs yielded more accurate forecasts than the RAW model output. These outputs were then forced into an HBV hydrologic model in order to generate a 14 day forecast. The approach effectively generated deterministic and probabilistic flows. The subsequent simulation results revealed that the downscale-based flows yielded greater skill than the RAW-based flows, which, in several cases, were relatively consistent with the results shown in the downscaling experiment.

The potential economic values of flow forecasts were assessed based on a simple optimal decision-making, cost-loss analysis technique. The principal outcomes emerging from the analyses include: (i) the economic benefits associated with probabilistic flow forecasts were more useful than their deterministic counterparts; and (ii) the downscale-based flow forecasts offered greater benefits, which are applicable to a much wider range of users, than the RAW-based flow forecasts. The findings of the present study clearly illustrate the potential added value that may be obtained as a result of adequate downscaling, as opposed to using the RAW model output, for hydrologic applications.

Keywords: downscaling; economic value; HBV; MRF; probabilistic forecast

1. Introduction

Precipitation is the most important component of a hydrologic system governing the generation of surface and subsurface runoff. The ability to provide reliable and accurate precipitation forecasts in space and time may greatly increase the significance of hydrologic forecasts. The advances in numerical weather prediction (NWP) over the last several years have generated considerable research activity, fostering the development of novel approaches for improving hydrologic forecasts. Several studies have shown that integrating meteorological forecasts and climate outlooks in a hydrologic prediction system can substantially improve the skill of streamflow forecasts (Clark and Hay, 2004; Werner et al., 2005; Roulin, 2007; Tucci et al., 2008; Shi et al., 2008; Li et al., 2009). For example, Clark and Hay (2004) forced statistically downscaled temperature and precipitation from the National Centers for Environmental Prediction (NCEP) Medium Range Forecast Model (MRF) output into a distributed hydrologic model for four basins across the United States. The subsequent streamflow forecasts (Clark and Hay, 2004) were compared against forecasts from the traditional climatic Ensemble Streamflow Prediction (ESP) procedure (Day, 1985). The results of Clark and Hay (2004) study indicated that while streamflow forecasts using downscaled estimates offered increased skill in the snowmelt dominated river basins, performance was essentially comparable with the ESP forecasts in the rainfall dominated basins. Consistent with these findings, Werner et al. (2005) demonstrated that streamflow forecasts based on MRF-driven temperature and precipitation data were generally superior to the control reforecast, although there were instances where the downscaled MRF output in fact degraded the flow forecasts. In general, the Werner et al. (2005) study showed a maximum ranked probability skill score (RPSS) value near 0.5, indicating a 50% improvement over the historical forecast and these values generally occur near lead times of about one week for most years.

Integrating MRF model output in a hydrologic prediction system has been shown to yield useful information. There are, however, several difficulties that can greatly limit its effectiveness, including, among others, identification of adequate downscaling models, proper forcing of meteorological forecasts, and forecast verification methods (Schaake, 2006). In order to ease these concerns and integrate rapid advances made in the recent past in research meteorologists, an international Hydrological Ensemble Prediction Experiment (HEPEX) was established in 2004. The goal of HEPEX is to advance reliable probabilistic hydrologic forecast techniques by establishing test-bed projects so as to evaluate the skill and economic values associated with the respective forecast systems (see HEPEX web page at http://hydis8.eng.uci.edu/hepex/).

The purpose of the present study is primarily focused on developing effective mechanisms for integrating meteorological ensemble systems into a hydrologic prediction system, that are amenable and suited to risk analysis and decision-making processes. The general framework is demonstrated under a number of conditioning paradigms for the Chute-du-Diable watershed located in northeastern Canada. In particular, the following key research questions were considered. 1) How can the MRF model output be best utilized in a hydrologic prediction system? 2) Is there any significant skill present in the downscaled MRF model output as opposed to forcing the RAW model output for use in hydrologic applications? 3) Is a probabilistic hydrologic prediction system, in fact, more skillful than its deterministic counterpart? 4) What is the potential economic gain, if any, from a probabilistic forecast relative to a deterministic forecast?

The remainder of this paper is organized as follows. Section 2 presents pertinent features of the downscaling models, characteristics of the application region, a description of the data used, and the experimental results and discussion. Section 3 presents the main features of the hydrologic model used, the model set-up and calibration techniques, and the results and discussions of the deterministic and probabilistic hydrologic forecasts. Section 4 discusses the potential economic values resulting from the

various hydrologic prediction systems considered, and section 5 summarizes the findings in light of the reported experimental results, and outlines future research directions.

2. Downscaling of meteorological forecasts

2.1. Downscaling models

Downscaling is a technique used to translate large-scale model output to a finer resolution. There are two broad fundamental downscaling approaches: (i) dynamical downscaling, and (ii) statistical downscaling. Dynamical downscaling uses fine spatial-scale numerical atmospheric models or regional climate models (RCMs) to simulate finer-scale physical processes (e.g., Mearns et al., 1999; Murphy, 1999; Spak et al., 2007). Statistical downscaling, on the other hand, uses a statistical relationship between large-scale model output and local-scale variables (e.g., von Storch et al., 1993; Hay and Clark, 2003; Harpham and Wilby, 2005). The broad theory, applications, advantages and shortcomings of common downscaling techniques are well documented in the scientific literature (e.g., Wilby and Wigley, 1997; Xu, 1999; Yarnal et al., 2001; Fowler et al., 2007). The present study focuses on two statistical downscaling techniques, namely, partial least squares regression (PLS) and *K*-nearest neighbor (KNN). These models are selected based on preliminary analysis (Muluye, 2010) and literature review (Yates et al., 2003; Gangodhyay et al., 2005) and are briefly described below.

Partial least squares regression

Partial least squares regression is a multivariate data analysis technique that generalizes and combines good features from multiple linear regression and principal component analysis (Wold et al., 1987). The PLS regression obviates some important restrictions that are normally imposed when modeling with traditional multivariate methods, such as principal component regression (PCR), discriminant analysis and canonical correlation. The PLS regression technique is particularly important when (i) the number of predictors is comparable to or greater than the number of data points; (ii) there exists multicollinearity in predictor data; and (iii) the data involve substantial random

noise. The PLS regression technique attempts to find factors (called latent variables) that maximize the amount of variation explained in predictors (relevant for predicting predictands), which is in contrast to PCR, where the factors (called Principal Components) are selected solely based on the amount of variation that they explain in predictors (e.g., Wold et al., 1987; Lorber et al., 1987). This enables the PLS regression to produce more stable and highly predictive models, particularly where the use of traditional multivariate methods are severely restricted. In-depth treatment of the subject and the working procedures of the method can be found in de Jong (1993).

The majority of statistical models including multiple linear regression and artificial neural networks experience great difficulties in identifying correct wet and dry days, and thus overestimate precipitation occurrences and underestimate precipitation intensities. In order to overcome such difficulties, a hybrid model has been developed. In this hybrid model, precipitation is modeled in a two stage process, where logistic regression is used to identify the occurrence of wet days and PLS is used to model the intensities. For downscaling total daily precipitation fields, the hybrid version of the PLS model was used, and for downscaling average daily temperature, the original PLS model was employed.

_K_-nearest neighbor

The nearest neighbor resampling methods are generally based on classic bootstrapping techniques (e.g., Yakowitz, 1993; Yates et al., 2003; Gangodhyay et al., 2005). The *K*-nearest neighbor algorithm essentially involves the search for similar feature vectors that exist in the observed time series on the principle of similarity criteria (i.e. *K* unique days are identified for the current day's weather, and one of these days is then randomly sampled and used as the next day's weather forecast) (Gangodhyay et al., 2005). In the present study, observed station variables such as total daily precipitation are considered as analog days to be identified on the basis of large-scale model output. The two key issues in the design of KNN-based modeling are the measure of closeness and the sampling strategy employed. To help identify the best model, several KNN-based

models have been developed and tested including Mahalanobis (Yates et al., 2003) and Euclidean (Gangodhyay et al., 2005) distance metrics (all results are not reported here). The present study used a KNN model based on Mahalanobis distance as described by Yates et al. (2003). A detailed investigation of the various sampling strategy options indicated that the resampled variables which correspond to the shortest distance yielded optimal performance.

2.2. Application

The study region selected for the application and investigation of the downscaling experiment was the Saguenay-Lac-Saint-Jean (SLSJ) hydrologic system in northern Quebec, Canada (Figure 23). The basin is approximately located between 47.3° to 52.2° N and 70.5° to 74.3° W with a total area of about 73,800 square kilometers. The basin contains 13 meteorological and 11 hydrometric stations and a number of reservoirs managed by the Alcan Company. These reservoirs are typically used for hydropower generation. The Chute-du-Diable meteorological station, located approximately at 48.75° N and 71.7° W, was used for this purpose. Total daily precipitation and average daily temperature were collected from Alcan hydro-meteorological stations for the period of 1979 through 2001.

The atmospheric predictor variables employed in the present study were from a T62 resolution version of NCEP MRF model output generated from an ensemble of 15-day forecasts over a 23-year period from 1979 through 2001 (Hamill et al., 2004). The NCEP MRF model output ensembles are defined on a global lat-lon grid with 2.5° resolutions both in longitude and latitude (144×73 grid points). The archive contains 8 daily predictors (model output variables corresponding to each ensemble member and forecast range), from 1 January 1979 to 31 December 2001 for a period of 23-years. These predictor variables include: 24-h accumulated precipitation (apcp), vertically integrated diabatic heating (heating), total column precipitable water (pwat), pressure reduced to mean sea-level (prmsl), temperature at 2 meters (t2m), relative humidity at 700 mb (rhum), zonal wind at 10 meters (u10m) and meridional wind at 10 meters

(v10m). Except for the heating variable all predictors have previously been identified as being useful for downscaling daily precipitation and temperature in the contiguous United States (Clark and Hay, 2004; Werner et al., 2005; Gangodhyay et al., 2005). The two downscaling models employed here are capable of extracting relevant features, irrespective of the number of predictors which exist in the feature vector, thus obviating the need for identifying significant predictors. The nearest grid to the Chute-du-Diable station was used to retrieve these predictors. Figure 24 shows a schematic diagram illustrating the different steps involved in a hydrologic ensemble prediction system.

Thus, the two data sets available for the downscaling experiment include: (i) local-scale predictands such as station daily total precipitation and average daily temperature, and (ii) eight large-scale model output predictor variables described above. These data sets were further divided into two parts: the first half of the data (1979-1996) were used to estimate the statistical parameters of the downscaling models considered and the rest of the data (1997-2001) were used to evaluate the performance of these models. The experimental data were prepared as follows. The numerical model output were processed to form a data matrix consists of 8401 rows (corresponding to the number of days from 1 January 1979 to 31 December 2001) and 8 variable columns (corresponding to each of 15 ensemble members and 15 forecast ranges or lead times). PLS and KNN downscaling models were then applied to generate estimates of daily temperature and precipitation values as described above.

2.3. Comparison of downscaling results

The downscaling experiments were implemented under two conditioning paradigms, which are reported and discussed in the subsequent sections. Suites of deterministic and probabilistic diagnostic measures were used to evaluate the downscaled outputs. The deterministic diagnostic measures include: bias (%), root mean squared error (RMSE), Pearson coefficient of correlation (r), Nash-Sutcliffe efficiency coefficient (CE), Kuipers score (KS), and mean and variance of the downscaled outputs. The probabilistic diagnostic measures to assess skill values include: median bias, median

RMSE, Brier skill score (BSS), ranked probability skill score (RPSS), reliability and discrimination diagrams. Furthermore, line, bar and box plots were used to assess the performances of the downscaled outputs.

2.3.1. Downscaled output derived from predictors of ensemble mean

In this component of the experiment, mean predictors taken over 15 members were used to condition PLS and KNN downscaling models. This resulted in a single series of downscaled daily precipitation and temperature values corresponding to each of the 15 forecast ranges and variables. It should be noted that the numerical model outputs of precip and t2m were considered as the RAW model output for all purposes in the present study. Table 1 shows the various model performance statistics for the test period. In the case of downscaling daily precipitation, the biases associated with the PLS model were generally small, indicative of more accurate precipitation forecasts. The RAW model outputs consistently and significantly overestimated (positive bias) precipitation intensities for all forecast ranges considered. The values of the RAW biases varied from 27% at forecast range (FR) 0 to 50% at FR 10, indicative of poor forecast accuracy with increased forecast ranges. The KNN model, on the other hand, showed a tendency to underestimate (negative bias) precipitation intensities for all forecast ranges. In the case of the PLS model, precipitation intensities were overestimated for FR of 0 and 3 days, and underestimated for FR of 7 and 10 days. In terms of the RMSE statistics, the PLS model yielded the smallest error for all FR. The RMSE associated with the KNN and RAW precipitation forecasts were quite competitive and yielded small errors. In general, precipitation forecasts were inadequate under all the three methods. This fact was clearly reflected and supported by the CE statistics. With the exception of FR of 0 days, where PLS and RAW outputs exhibited about 35% and 24% skill, respectively, performance was generally poor and decreased with increasing FR.

The variances of daily precipitation associated with the KNN model were in good agreement with the observed values, albeit with a slight under-representation. The PLS and the RAW forecasts, on the other hand, tended to under-represent these key statistics

significantly and consistently. The failure to adequately reproduce the variability in daily precipitation is one of the acknowledged limitations of most statistical downscaling approaches (e.g., Clark and Hay, 2004). The accuracy of precipitation forecasts was further assessed with the KS statistic. The KS indicates the degree of conformity between forecast precipitation occurrences (or non-occurrences) and observed precipitation occurrences (or non-occurrences) (Wilks, 1995). The KS statistics associated with the PLS model were more skillful when compared with the KNN and the RAW models; however, they tended to deteriorate rapidly with increased FR.

In the case of downscaled temperature, the observed daily temperature was slightly and consistently underrepresented by all the three forecast methods. In comparison, the PLS model provided the smallest bias and RMSE, for all FR. In general, the demonstrated skills of the downscaled temperature under all forecast methods were quite satisfactory (CE values from 70% to 96%). Further analysis into the CE statistics in Table 1 indicates that with increased FR, the accuracy of temperature forecasts tended to decrease, with the greater loss of accuracy being exhibited by the KNN model. As expected, downscaling daily precipitation was more problematic than downscaling daily temperature.

In addition to the above diagnostic measures, statistical tests were also performed between the observed and the downscaled precipitation to determine if there is evidence of difference in the population locations and variances without assuming a parametric model for the distributions. The first statistical test investigated was the Mann-Whitney test (Lehmann 1975) of the equality of two population medians. The results of the Mann-Whitney test between the downscaled and the observed dataset are presented in Table 2. The values in this table represent *p*-values. A significance level of 5% is chosen for the purpose of comparison. The computed *p*-value greater than the chosen significance level indicates the simulated and the observed medians are statistically equal. Table 2 shows that the *p*-values associated with the KNN model were greater than the chosen significance level of 5% for all forecast ranges, suggesting the observed and the downscaled precipitation medians are statistically same. For the case of the PLS and the

RAW-based precipitation, the computed *p*-values were less than the rejection level of significance, indicating the observed and the downscaled medians are statistically different (except for forecast ranges of 0 and 3 days, respectively). The other non-parametric test conducted to check the equality of variances between the downscaled and the observed precipitation is Levene's test (Levene 1960). The computed *p*-values associated with the KNN and the PLS models were greater than the rejection level of significance for forecast ranges of 0, 7 and 10 days, and 0 and 3 days, respectively, suggesting that the observed and the downscaled precipitation variances are statistically equal (Table 2). The RAW-based precipitation variances were, however, significantly different from the observed precipitation variances for all forecast ranges considered.

2.3.2. Downscaled output derived from predictors of individual members

In this experiment, predictors from individual members were used to condition PLS and KNN downscaling models for generating daily precipitation and temperature fields, which, in turn, resulted in 15 downscaled values corresponding to each of 15 forecast ranges and variables. The subsequent downscaled outputs were then evaluated using deterministic and probabilistic approaches.

2.3.2.1. Deterministic approach

In this approach, ensembles of downscaled temperature and precipitation values were averaged over 15 members, resulting in a series of single forecast values corresponding to each of the 15 forecast ranges. The biases in Table 1 indicate that while precipitation forecasts using the KNN model showed a tendency to underestimate the intensities, the RAW forecasts showed the opposite. The PLS model, on the other hand, underestimated the precipitation intensities at FR of 0 and FR of 5 days, and then overestimated at FR of 7 and FR of 10 days. In general, the PLS model yielded more accurate precipitation forecasts, followed by the KNN model, although great difficulties were exhibited in adequately representing the variability in daily precipitation. Better performance statistics in terms of RMSE, *r* and KS supports the relative superiority of

PLS over the others. As expected, the skill of downscaled daily precipitation deteriorated quickly with increased forecast ranges.

In the case of downscaling average daily temperature, with the exception of the RAW-based forecasts at a FR of 10 days, all three forecast methods tended to slightly underestimate the observed temperature. On the whole, there were no consistent and significant distinctions which can be drawn in terms of performance statistics among the three temperature forecast schemes. The CE statistics in Table 1 show that the skill of temperature forecasts were quite satisfactory (CE values from 82% to 96%), for all forecast ranges, despite a gradual loss of skill with increased forecast ranges.

2.3.2.2. Probabilistic approach

In this section, the downscaled ensemble temperature and precipitation forecasts were assessed using suites of probabilistic diagnostic measures. Figure 1 depicts the biases associated with downscaling total daily precipitation. The box plots represent the spread of precipitation forecasts corresponding to each forecast range. While forecasts of precipitation using the KNN model showed a tendency to overestimate the intensities, the RAW forecasts showed a tendency to underestimate the intensities. The PLS model overestimated the intensities at FRs of 0 and 5 days and underestimated the intensities at FRs of 7 and 10 days. In general, the biases associated with the RAW precipitation forecasts were larger (about 1 to 1.5mm) when compared with the PLS forecasts (close to zero) and the KNN forecasts (under 0.5mm). Similar box plots are shown in Figure 2 for the downscaling of daily temperature. With the exception of the RAW forecasts at a FR of 10 days, all three forecast methods tended to slightly underestimate the observed temperature. The temperature biases were quite small and were within 0.5°C for the PLS and KNN forecasts, and about 0.8°C for the RAW forecasts.

Figures 3 and 4 illustrate box plots of RMSE as a function of forecast range for downscaled daily precipitation and temperature, respectively. The RMSE statistics associated with the PLS model provided the smallest error, and yielded more accurate precipitation forecasts when compared with the KNN and the RAW-based forecasts. In

general, the RMSE increased with increased forecast ranges, suggesting the difficulties associated with long lead time precipitation forecasts. In the case of downscaling average daily temperature, the PLS-based forecasts yielded the smallest RMSE, for all forecast ranges, when compared with the KNN and the RAW-based forecasts. Similar to precipitation forecasts, the skill of temperature forecasts deteriorated with increased forecast ranges.

The accuracy of probabilistic precipitation forecasts was further evaluated using the Brier skill score (Murphy, 1997; Wilks, 1995). The BSS is a scalar measure of the association between the forecasts and observations and measures the improvement of the forecasts relative to the reference forecasts (Murphy and Winkler, 1987; Wilks, 2001). Figure 5 (a) illustrates a BSS bar graph summarizing probabilistic precipitation forecasts. A threshold of 0.3mm was used to classify the occurrence of wet and dry days, and the subsequent values were used to construct BSS statistics. With the exception of PLS and KNN-based forecasts, where 10% and 5% positive skills were apparent respectively, the forecast performance was generally inferior to climatology. Such inconsistent results should not be surprising, given the fact that squeezing together a high dimensional verification problem into a single scalar value normally tends to obscure some important information (Murphy and Winkler, 1987; Wilks, 2001). In comparison, the PLS-based forecasts appeared to be better than the KNN and the RAW-based forecasts. For the case of downscaling average daily temperature (Figure 5(b)), the BSS statistics were generally well above 70%, signifying more accurate temperature forecasts. In comparison, the skill of the KNN forecasts were slightly superior to the PLS and the RAW-based temperature forecasts, which is inconsistent with the results obtained when verified using deterministic approaches.

The ranked probability skill score is one of the most common verification techniques normally used to evaluate the skill of probabilistic forecasts (Stanski et al., 1989). In contrast to the BSS, the RPSS considers the whole distribution of the forecasting system, and thus provides better insight about the overall skill of the forecasts under consideration. Figure 5 (c) summarizes the RPSS of probabilistic precipitation

forecasts, for the top 90th percentile of observations. These bar plots clearly indicate that precipitation forecasts were generally skillful, with the exception of the PLS model for which the forecasts were inferior to climatology at FRs of 7 and 10 days. In comparison, the KNN-based forecasts demonstrated quite superior performance when compared with the PLS and the RAW-based forecasts. The improvements in relative skill were more pronounced with increased forecast ranges. In the case of downscaling average daily temperature (Figure 5 (d)), performance was quite satisfactory, with over 60% skill. In comparison, the KNN-based temperature forecasts were marginally superior to the PLS and the RAW-based forecasts, which is inconsistent with the earlier results obtained while verified using deterministic approaches.

2.3.3. Discussion

Two conditioning paradigms were investigated to identify the best practice for utilizing the MRF model output for the purpose of downscaling daily precipitation and temperature. Furthermore, the RAW model output was inter-compared against outputs from two downscaling models to observe any evidence of forecast improvement as a result of downscaling. The experimental results revealed that there were virtually no significant distinctions in performance between the two conditioning paradigms. The PLS and the RAW-based forecasts showed a slight edge in performance when predictors obtained from ensemble means were used. The KNN-based forecasts, however, yielded more accurate precipitation and temperature forecasts when the mean downscaled values of 15 members were considered. However, such modest improvements were attained at the slight expense of deflated variability in the daily precipitation, where the KNN model is normally known for its strength. Furthermore, the performances of individual downscaled members were also examined. The overall performances of each of the 15 members were, in fact, identical, and did not exhibit substantial differences compared to those results found in downscaling predictors of ensemble means (results not reported here). Therefore, the choice of conditioning paradigm largely depends on the intended purpose of the downscaled output (i.e. for deterministic or probabilistic uses). For

instance, users interested in deterministic applications may prefer to downscale predictors of ensemble means, thus saving substantial resources. In general, if the ultimate purpose of downscaling is for use in hydrologic applications, then the relative success of each case should be examined individually, which is the subject of the subsequent sections.

When the downscaled and RAW model outputs were compared, inconsistent and mixed performances were observed based on the deterministic and probabilistic diagnostic measures. The PLS model appeared to offer more accurate forecasts when deterministic metrics, such as bias, RMSE, CE and KS statistics were considered, while the KNN model exhibited modest forecast accuracy. The KNN model, on the other hand, reproduced the variability in daily precipitation reasonably well, whereas the PLS and the RAW model outputs significantly and consistently under-represented these key statistics. The failure to adequately represent the variability in daily precipitation is, in fact, one of the major weaknesses of most statistical approaches (Clark and Hay, 2004). In the case of modeling with the PLS model, the large-scale model outputs were directly conditioned to estimate the statistical parameters of the model. This approach emphasizes more on the expected value, thus tending to deflate the variability in daily precipitation. In the case of modeling with the KNN model, the observed time series was randomly sampled on the principle of similarity criteria, which helped preserve the variability in daily precipitation from being deflated.

When probabilistic verification measures were considered, the KNN model appeared to perform better in terms of BSS and RPSS statistics, for both temperature and precipitation. Conversely, the PLS model provided less accurate precipitation forecasts, yielding inferior forecasts than climatology when verified using RPSS.

It is difficult to draw a clear distinction among the different forecast approaches considered. In general, the downscaled outputs yielded more accurate forecasts than the RAW output, from which downscaling with the KNN model appeared to be the best choice. The evaluation measures employed here may not necessarily reflect the actual skill of the downscaled output when forced into a hydrologic model. Thus, the downscaled outputs obtained should be investigated further to identify any skill transfer

between downscaled outputs and hydrologic forecasts. The ability to understand and recognize leading weather forecast verification metrics, which, in turn, indicate some clues about the potential skill that may be present in a hydrologic forecast system, is imperative.

3. Hydrologic forecasting

3.1. Hydrologic model

The hydrologic model selected for the investigation of flows obtained by forcing MRF model output was an HBV hydrologic model, developed at the Swedish Meteorological and Hydrological Institute (Bergstrom, 1995). The HBV model can best be described as a semi-distributed conceptual hydrologic model, and was originally developed for use in Scandinavian catchments in the 1970's (Lindstrom et al., 1997). In Scandinavia, the HBV model is the standard operational runoff forecasting tool for nearly 200 basins. To date, it has found operational or scientific applications in more than 50 countries around the world with a wide range of climatic conditions and geographical locations (Bergstrom, 2002). The HBV model has been successfully implemented in numerous water resources and environmental problems including flood forecasting, water resources evaluation, short-term and long-term reservoir inflow forecasting, dam safety studies, simulation of stream-flow in ungauged rivers, pre-feasibility studies, and climate change studies (Bergstrom et al., 2001; Menzel and Burger, 2002; Bergestrom et al., 2002; Dibike and Coulibaly, 2005, Wang et al., 2006; Sorman et al., 2009). The HBV model uses sub-basins as the primary units for runoff generation and provides options for variations in area-elevation zones and basin land-use patterns (forest, open areas and lakes). The model comprises subroutines including snow accumulation and melt, soil moisture accounting procedures, routines for runoff generation and a simple routing procedure (Bergstrom, 1995). The HBV model is normally run on daily values of precipitation, and in areas with snow, mean daily air temperature and/or monthly estimates of evaporation are required (Singh, 1995). The grounds for the selection of the HBV model in the present study include: (i) there exists some hydrologic similarities

between the study basin and the origin of the hydrologic model. Both basins are located in cold regions where the hydrology (particularly spring flow) is largely driven by snowmelt; (ii) the HBV model can be implemented with limited data and still maintains better accuracy under diverse topographic and hydrologic conditions; and (iii) the HBV model has found several applications in forecasting short-term and long-term reservoir inflows and has been shown to offer adequate performances (Bergstrom, 2002).

The calibration of the HBV model was done as follows. The 9700 km^2 study basin was subdivided into six area-elevation zones and basin land-use patterns (forested and open areas). Then the model was forced with observed total daily precipitation and average daily temperature. The purpose of this experiment was to adequately model reservoir inflows in the Chute-du-Diable sub-basin. Daily reservoir inflows (simply flows) for the period from 1 January 1979 through 31 December 2001, which is a sufficient time period to cover both dry and wet years, were collected from Alcan Company for calibration and evaluation of the HBV model. For adequate implementation of the HBV model, the user manual recommends 5 to 10 years of good quality calibration data. Therefore, 16 years of data (from Jan 1979 to Dec 1996) were used for model calibration and the remaining 5 years of data (from Jan 1997 to Dec 2001) were used for evaluating the performance of the calibrated model. The various model parameters were adjusted until the simulated flows were in good agreement with the observed flows. During the calibration process, the adequacy of model performance was primarily assessed using the Nash-Sutcliffe efficiency coefficient, as suggested in the HBV user manual. Once the optimal model parameters were found, the model was saved and the calibrated model was then used to forecast to a 14 day lead using the downscaled and the RAW outputs.

To illustrate the skill of the HBV model in modeling flows, observed and simulated flow hydrographs are shown in Figure 7a for the test period. The light gray line and the dark line represent observed and simulated flow hydrographs respectively. For the most part, the simulated flows were in good agreement with the observed flows. The Nash-Sutcliffe efficiency coefficient of 0.8, for the test period, indicates that the flow

simulation was fairly adequate in spite of the underlying hydrologic and topographic difficulties existing in the study basin. In particular, the HBV model was able to adequately represent the peaks which occurred during the spring season. The ability to provide reliable and accurate reservoir inflows, especially during the spring season, has enormous economic consequences, where hydropower production is the ultimate and the foremost objective of such a forecast system. The Chute-du-Diable basin contains three hydropower generating stations managed by the Alcan Company. The basin is sparsely populated and most of the watershed is covered with forests. The elevation of the basin ranges from 700m to 100m. The average annual temperature of the basin is 2°C and the annual precipitation amount is 935 mm, from which 27% is in the form of snow. Nearly 35% of the annual stream flows occur in the spring season due to snow melt. The mean annual flow of the basin is about 214 m^3/s, whereas the mean monthly low and peak flows are approximately 15 m^3/s and 1221 m^3/s respectively, and occur during the winter (in March) and spring (in May) seasons. The flow coefficient of variation is 1.0 and skewness is 2.5. The existence of such strong nonlinearity in daily flows, coupled with a wide-ranging landscape, poses great difficulties to adequately model the reservoir inflow of the Chute-du-Diable sub-basin. Given the scale and magnitude of these challenges, the simulation results obtained from the HBV model are deemed to be adequate.

3.2. Deterministic flow forecasts

3.2.1. Case 1: Hydrologic forecasts forced with downscaled output derived from predictors of ensemble mean

In this experiment, flow forecasts were made by forcing downscaled temperature and precipitation derived from mean predictors over 15 members. The mean of the RAW model outputs of temperature and precipitation were also used. Figure 8 shows the various flow forecast performance statistics against forecast ranges. The last column in Figure 8(c) represents flow forecast statistics resulting from forcing downscaled temperature and precipitation data derived from mean predictors. The biases in this diagram show that PLS and KNN-driven flows tended to under-represent the observed

flows, whereas those derived from the RAW-driven flows appeared to over-represent the observed flows. It is interesting to note that with increased forecast ranges, the biases associated with the PLS and the KNN-based flows appeared to diminish, and the opposite was exhibited for the RAW-based flows. In comparison, the PLS-based flows produced the smallest biases (within 20 m^3/s) when compared with the KNN (within 40 m^3/s) and the RAW (within 50 m^3/s) for all forecast ranges. In terms of RMSE statistics, the PLS and the KNN-based flows resulted in smaller errors, indicative of more accurate forecasts relative to the RAW-based flow forecasts. The RMSE associated with the RAW-based flows were approximately twice as large as the PLS and the KNN-based RMSE. The linear correlation coefficient, between the observations and simulations, r, appeared satisfactory for all forecast methods, albeit with a slight deterioration with increased forecast ranges. The entire forecast set attained over 0.7, from which the PLS-based flows exhibited marginal advantage for the first few forecast ranges. Higher values of r do not necessarily translate into better performance skill. The corresponding CE statistics, in fact, supports this assertion. With the exception of the first few forecast ranges where 5% to 10% skill was evident, flows associated with the RAW forecasts exhibited inferior performances when compared with climatology. Conversely, the PLS and the KNN-based flow forecasts demonstrated approximately 70% skill at the early forecast ranges, and diminished to 50% skill at the latter forecast ranges. The benefits associated with the downscaled-based flows were more apparent with increased forecast ranges.

3.2.2. Case 2: Hydrologic forecasts forced with mean of downscaled output derived from predictors of individual members

In this component of the experiment, the deterministic flow forecasts were obtained as follows. First, each of the 15 members was downscaled to generate 15 values of temperature and precipitation. The downscaled outputs were then averaged over 15 members. The subsequent mean values were forced into the HBV model to produce deterministic flow forecasts. Figure 8 (b) illustrates the performance of flow forecasts using suites of deterministic diagnostic metrics (the middle column). The performance

statistics indicate that there were no significant distinctions in performance between Case 1 and Case 2 experiments. The PLS and the KNN-based flows still provided better statistics in terms of bias, RMSE, *r* and CE, compared with the RAW-based flows. The skill of KNN-based flows showed slight improvements in this set of experiment, which is quite consistent with the modest improvements noticed earlier while downscaling with the KNN. The PLS-based flows also showed some improvements, but the gains were quite marginal. In general, in terms of the CE statistics, the skills of flow forecasts were approximately 72% for the early forecast ranges and diminished to 60% for the long-range forecasts. Note that the performance statistics associated with the RAW-based flows in both cases (Case 1 and Case 2) were identical as the same model inputs were considered.

3.2.3. Case 3. Hydrologic forecasts considering mean of ensemble flow output

In this set of experiments, the deterministic flow forecasts were obtained as follows. The individual downscaled members were first forced into the hydrologic model to generate an ensemble of flows. Similarly, the RAW model outputs from each of the 15 members were forced into the hydrologic model to generate ensembles of flows. The deterministic flow forecasts were then obtained by averaging over the 15 flows, corresponding to each of the forecast ranges. This is, in fact, the most common way of analyzing ensemble flows using a deterministic approach. The first column in Figure 8 illustrates the various deterministic metrics employed for assessing flow forecasts. The biases associated with the KNN and the RAW-based flows showed a tendency to underestimate and overestimate the observed flows respectively. The biases associated with the RAW-based flows (within 50 m^3/s) were much larger than those associated with the KNN-based flows (within 25 m^3/s). In comparison, the PLS-based flows produced the smallest biases (within 20 m^3/s). In terms of the RMSE statistics, the KNN-based flows consistently produced the smallest errors. The accuracy of PLS-based flows was as good as KNN-based flows in the first few forecast ranges. In comparison, the RMSE associated with the RAW-based flows were about twice as large as the PLS-based errors.

In this case of the experiment, the KNN-based flows showed a slight advantage in skill over Cases 1 and 2. The RAW-based flows also exhibited marginal improvements in the early forecast ranges, but no improvements were observed for the PLS-based flows. In general, the KNN-based flows had an edge over the PLS, particularly for long range forecasts. In terms of the CE statistics, the skills associated with the KNN and the PLS-based flows were about 70% in the early forecast ranges, and, decreased to about 60% in the latter forecast ranges. The RAW-based flows, on the other hand, showed some skills (5% to 20%) during the first few forecast ranges and, tended to deteriorate rapidly with increased forecast ranges.

As an example of graphical representation, Figure 7b illustrates hydrographs of observed and simulated flows for a FR of 3 days, for the portion of test period from January 1997 through December 1999. These flows were obtained by forcing the HBV model with the mean of ensemble flows resulting from PLS, KNN and RAW-based outputs. Careful observation of Figure 7b indicates that the flow hydrographs associated with KNN and PLS were generally in good agreement with the observed flows, compared with the RAW-based flows.

3.3. Probabilistic flow forecasts

In this section, individual downscaled members and the RAW model outputs were forced into the hydrologic model in order to generate an ensemble of flows. The subsequent flows obtained were evaluated using suites of probabilistic approaches. Figure 9 shows the biases corresponding to each of the forecast ranges. The biases associated with the KNN-based flows showed a tendency to overestimate the observed flows, whereas the RAW-based flows showed the opposite. The PLS-based flows slightly overestimated the first few forecast ranges and then tended to underestimate the latter forecast ranges. The median biases associated with the PLS-based flows were the smallest (within 40 m^3/s); followed by the KNN-based flows (within 55 m^3/s). The median biases associated with the RAW-based flows were quite large, and increased with increasing forecast range (varies from 50 to 105 m^3/s). These results were reasonably

consistent with the results shown in the downscaling experiments. In terms of RMSE statistics, the errors associated with PLS-based flows were generally small, indicative of more accurate flow forecasts (Figure 10). Conversely, the errors produced by the RAW-based flows were quite large. On the whole, the RMSE associated with all flow forecast methods exhibited a tendency to increase with increasing forecast range.

The Brier skill score against the forecast range is presented in Figure 11(a). These bar graphs show that PLS and KNN-based flows had greater skills, whereas the RAW-based flows demonstrated no skill, and thus yielded inferior performance when compared with climatology. The skill associated with the KNN-based flows (about 50% to 65%) was greater than that of the PLS-based flows (approximately 35% to 55%). Figure 11(b) illustrates similar bar plots for RPSS statistics. The flow forecasts made by all the three methods appeared to yield substantial skills for all forecast ranges. In particular, the RAW-based flows showed a good performance although the KNN-based flows still demonstrated superior performance. The KNN-based flows yielded approximately 30% to 35% skill, whereas the RAW-based flows yielded approximately 20% to 30%, and those of the PLS-based flows yielded approximately 10% to 30%. No performance trend was observed by all forecast methods with increased forecast ranges. It is noteworthy to mention that the comparative flow forecast results shown here were generally inconsistent with those of the downscaled results observed earlier when verified under BSS and RPSS.

The accuracy of ensemble flows was further assessed using reliability and discrimination diagrams. The reliability diagram is a useful diagnostic technique which provides valuable information about the distribution of forecasts given the observations. The discrimination diagram provides some insight about the distribution of observations given the forecasts (Franz et al., 2003). Figures 12, 13 and 14 depict the reliability diagrams for FRs of 3 and 7 days resulting from the PLS, KNN and RAW-based flows respectively. The reliability diagrams were constructed for those flow series exceeding a threshold corresponding to the 95^{th} percentile of the observed flows. The reliability diagram in Figure 13 associated with the KNN-based flows was in reasonably close

proximity to the perfect reliability line (45^0), indicative of accurate and reliable flow forecasts. In particular, forecasts made for 7 days lead time were more reliable than 3 days lead time. In general, the KNN-based forecasts slightly over-forecasted (forecasts below the 45^o line) the observations at lower probabilities and under-forecasted at higher probabilities. In the case of the RAW-based flows, the forecasts were, in general, not reliable (Figure 14). The RAW-based flows showed a tendency to consistently and significantly under-represent the observed probabilities. In comparison, the reliability diagrams associated with the PLS-based forecasts were superior to the RAW-based forecasts, but provided inferior performances when compared to the KNN-based forecasts (Figure 12).

Figures 15, 16 and 17 illustrate the discrimination diagrams for the FR of 7 days, resulting from the PLS, KNN and RAW-based flows respectively. The discrimination diagram was constructed using the low-flow (below 33 percentile), medium-flow (between 33 and 67 percentile) and high-flow (above 67 percentile) portions of the observed flows. If the forecasts are discriminatory, low-, medium-, and high-flow portions tend to separate one from the other and will not overlap (Murphy et al., 1989). In comparison, the forecasts associated with the KNN-based flows (Figure 16) were the most discriminatory of all forecasts considered. The RAW-based forecasts (Figure 17), on the other hand, experienced great difficulties in discriminating between low- and medium-flow portions, particularly at higher probabilities (from 600 to 1500 m^3/s). Similar problems were also observed in Figure 15 for flows of PLS, in discriminating low- and medium-flow portions, at higher probabilities (from 400 to 600 m^3/s). In general, all flow forecasts yielded good discriminatory skills, the KNN-based flows being the most discriminatory, followed by the PLS-based flows.

As an example of graphical representation, Figure 6 illustrates the Ensemble Prediction System for a FR of 3 days obtained by forcing the KNN-based output into the HBV model, for the portion of the test period from January 1997 through December 1999. The light gray lines represent flows corresponding to each of the 15 ensemble members, whereas the thick dark gray line and the dark line represent the mean of

ensemble flows and the observed flows respectively. Careful observation of Figure 6 reveals that the observed hydrograph was contained within the ensemble in most of the cases. In particular, the hydrograph associated with the mean of ensemble flows was in good agreement with the observed flows.

3.4. Discussion

The deterministic flow forecast skills were assessed under three cases. The simulation results revealed virtually no significant distinctions in the skill of flow forecasts when the hydrologic model was forced with (i) the downscaled output derived from predictors of the ensemble mean; and (ii) the mean of the downscaled output derived from predictors of individual members. In comparison, the PLS-based flows had a slight advantage over the KNN-based flows, whereas the RAW-based flows generally had no skill and yielded inferior performance to climatology. As described in the previous section, the hydrologic model was calibrated once and the calibrated model was used to generate flows by feeding the downscaled and the RAW outputs. When the hydrologic model was independently calibrated using the RAW outputs, the generated RAW-based flows in fact provided greater skill values, compared to those results presented in this study (CE about 0.5 for FR3; results not reported). These skill values were, however, inferior to those flows originating from the downscaled-based flows. In general, the best performance associated with the deterministic flow forecasts was shown when the mean individual flows were considered. The improvements obtained were quite modest and appeared only for the KNN and the RAW-based forecasts.

The KNN-based flows generally yielded superior performances when inter-compared using probabilistic approaches. These results were generally consistent with results from the downscaling experiment presented earlier. There were, however, some instances where mixed and inconsistent results were observed. For example, in the case of downscaling daily precipitation, the BSS statistics showed a lack of skill for almost all forecast methods considered; whereas, in the case of flow forecasts, the BSS statistics showed greater skills for the PLS and the KNN-based flows and a lack of skill for the

RAW-based flows. Similarly, when the RPSS statistics were used, the PLS-based forecasts showed a lack of skill for longer lead precipitation forecasts and modest skill for the KNN and the RAW-based forecasts. Inconsistent with these observations, the RPSS associated with the different flow forecasts showed greater skill for all flow forecasts including PLS-based flows. In fact, there were some clear indications shown when the deterministic forecast measures were used. The PLS and the KNN-based flows provided small RMSE and slightly under-represented the observed flows, whereas the RAW-based flows showed just the reverse. These results were reasonably consistent with the observed relative performances obtained in downscaling daily precipitation. Conversely, when the probabilistic diagnostic measures were considered, the median biases associated with the PLS and the KNN tended to over-represent the observed precipitation, for the majority of forecast cases. These observations were consistent with the observed relative performances shown on the flow forecasts. The RAW-based precipitation and flow forecasts exhibited just the opposite. The temperature forecasts used to feed the HBV model were reasonably accurate. Even the RAW-based temperature forecasts were generally as accurate as the PLS and the KNN-based temperature forecasts, but generated flows which were inferior to climatology. The KNN-based flows, on the other hand, provided higher skill flow values when compared with its counterparts. This could partly be attributed to the intrinsic ability of the KNN model to adequately represent the variability in daily precipitation.

In a region where the hydrology is largely dominated by snowmelt, reliable and accurate temperature forecasts may provide adequate flow forecasts (Clark and Hay, 2004). Given the more accurate RAW temperature forecasts coupled with the distinct nature of the study basin, the RAW-based flows were expected to perform better than what had been observed in this study. In particular, the RAW-based flows performed poorly in the spring season by significantly over-forecasting the peaks. This poor performance is clearly evident from the reliability and discrimination diagram. The RAW-based flows consistently and significantly over-represented the whole forecast probabilities and demonstrated great difficulties in discriminating among low- and

medium-flow portions. The different forces that led to these inconsistencies should be investigated further. In addition, indicators of appropriate precipitation forecast diagnostic measures should be clearly understood, and identified, for use in hydrologic applications.

4. Economic value of hydrologic forecasts

The potential economic value of flow forecasts was assessed based on a simple optimal decision-making, cost-loss analysis technique (Murphy, 1977; Wilks, 2001; Richardson, 2000). There are several courses of action in which a decision maker can choose from and, each action has associated costs which lead to either economic benefit or loss depending on the final outcome of the forecast (Richardson, 2000). For a certain event Y, the decision-maker has two options: take action or do nothing, depending on the belief that the event Y will occur or not. Taking protective measures against the event Y incurs a cost C, whereas occurrence of the event Y without protection incurs a loss L. Thus the goal is to decide a suitable and optimal course of action that minimizes the expected loss.

The economic value of forecasts is simply defined as the difference in expected expenses between some baseline (often climatology) information and the expected expenses given the forecasts under consideration (Wilks, 2001). The potential economic value (EV) of forecasts is given by (Richardson, 2000):

$$\mathrm{EV} = \frac{\min(k, Pc) - F(1-Pc)k + HPc(1-k) - Pc}{\min(k, Pc) - Pck} \tag{1}$$

where k = C/L, the cost-loss ratio; H and F are the hit rate and the false alarm rate, respectively; and Pc is the climatologic frequency of an event. Equation (1) shows that the relative economic value of a particular forecast system depends on parameters k and Pc, which are external to the system, and H and F, which are model-dependent (Richardson, 2000). The EV ranges between minus infinity and one, for which one represents a perfect forecast, zero represents the climatology forecast, and a negative value suggests the forecasting system has insufficient skill, and the preferred strategy is

to follow climatological information. For the same forecast system under consideration, the potential EV can be different depending on the choice of cost-loss ratio. For that matter the EV is typically represented graphically as a function of C/L. For a perfectly reliable forecast system, the EV is shown to be optimal when $k = Pc$ (Richardson, 2000; Wilks, 2001). The computation of EV is straight forward for a deterministic forecast system, where the H and F values can readily be computed from the contingency table. But in the case of probabilistic forecasts, a probability threshold, Pt, is first chosen so that the probabilistic forecast system is converted into a deterministic forecast system. Next, for each value of Pt, the corresponding values of H and F are computed. The EV values can then be generated using equation (1), similar to the deterministic case. For a perfectly reliable probabilistic forecast system, the optimal Pt is shown to be equal to the cost to loss ratio (Richardson, 2000).

Figures 18, 19 and 20 show the potential economic values of deterministic flow forecasts obtained from PLS, KNN and RAW-based flows, for the event exceeding the 90th percentile of the observed flows. The EV curves were constructed for three flow scenarios: (i) flows generated from the mean of downscaled output derived from predictors of individual members (MDM); (ii) flows generated from downscaled output derived from predictors of the ensemble mean (PEM); and (iii) flows obtained from the mean of ensemble flow output (MEF). The EV curves associated with the MDM are shown in Figure 18 and their usefulness was highly dependent on the cost-loss ratio. In particular, the KNN-based flows showed greater skill in the range of $0.04 < C/L < 0.75$, signifying the range of users that can benefit from this forecast system. For a C/L ratio outside of this range, the forecast lacks value, and therefore decision makers will be better off with climatology information. Similarly, the decision makers benefit from the PLS and the RAW-based flows in the range of $0.04 < C/L < 0.61$, and $0.03 < C/L < 0.38$ respectively. In comparison, the KNN-based flows exhibited positive economic values for a wide range of C/L, followed by the PLS-based flows. Thus a much wider range of users can be benefited from the KNN-based flows. It was apparent from the EV curves

that the largest economic benefits (about 70%) were attained approximately at C/L = *Pc*, which, in this case was 0.1 (Richardson, 2000; Wilks, 2001).

Figure 19 and 20 illustrate similar EV plots for PEM and MEF, respectively. On the whole, the performances of PEM and MEF were quite identical to the MDM, with the exception of some added economic benefits exemplified by the KNN and the PLS-based flows, particularly for higher C/L ratios. For the MEF scenario, the decision makers will be better off with the KNN and the PLS-based flows, having C/L ratios between $0.04 < C/L < 0.84$, and $0.03 < C/L < 0.68$ respectively. The RAW-based flows, however, did not show substantial improvement in value when the MEF was used as opposed to either the PEM or the MDM. In general, the MEF demonstrated wider range of economic benefits relative to its counterparts, offering greater opportunities for decision makers to benefit at higher C/L ratios.

In the case of probabilistic forecasts, a set of EV curves were generated corresponding to various threshold probability values, *Pt*. As an example, Figure 21 depicts EV curves resulting from the KNN-based ensemble flows for a FR of 7 days. The thin lines represent EV curves associated with the various probability thresholds and, the envelope of these curves (heavy solid line) represents the overall relative value of probabilistic forecasts. It is apparent from Figure 21 that the probabilistic forecasts exhibited greater values for most C/L ratios although the potential benefits varied considerably amongst decision makers with different cost-loss ratios (Richardson, 2000). The relative value of forecasts is very sensitive to the choice of *Pt*. For a small C/L (large potential losses), decision makers can still benefit even if low forecast probability exists, whereas for higher C/L, decision makers can benefit if the forecast probability is higher (Richardson, 2000). As can be observed from Figure 21, the overall EV of probabilistic forecasts were much greater than the deterministic forecasts, signifying the information content of a probabilistic forecast system is higher than the information content of a deterministic forecast system (Buizza, 2001).

Figure 22 shows EV curves comparing deterministic and probabilistic forecasts for FR7. The probabilistic EV curves were compared against the EV curve obtained from

the best deterministic forecast scenario (MEF). The following points were observed from Figure 22: (i) the probabilistic forecasts associated with the KNN, PLS and RAW-based flows were more useful than their best deterministic counterparts; (ii) the probabilistic forecasts associated with the KNN-based flows were more useful than any other forecast systems considered; (iii) the deterministic KNN-based flows were more useful than the PLS-based probabilistic forecasts, and the deterministic PLS-based forecasts were, in turn, more useful than the RAW-based probabilistic forecasts. It is interesting to note that while probabilistic forecasts possess greater economic benefits when compared with deterministic forecasts generated from the same model and forcings, these forecasts can be quite inferior to deterministic forecasts resulting from different forcings and models. Therefore, identification of an adequate hydrologic model coupled with a proper forcing technique could substantially improve the skill and economic value of hydrologic forecasts.

5. Summary, conclusions and recommendations

The present study reported the experimental findings obtained when forcing the NCEP MRF based outputs into a hydrologic model. The general framework was implemented in the Chute-du-Diable watershed located in northeastern Canada. The primary motivations in this paper included: identification of proper forcing methods in a hydrologic model, and evaluation of the skill values and economic benefits associated with the various hydrologic forecast systems. The principal findings emerging from the material are presented now in the following three sections.

1. Downscaling: PLS and KNN models were applied to extract daily total precipitation and daily average temperature from the NCEP MRF model output under two conditioning paradigms: (i) mean predictors over 15 members; and (ii) predictors from individual members. The subsequent downscaled outputs were compared against the corresponding RAW model output through deterministic and probabilistic diagnostic measures. The experimental results indicated that while precipitation forecasts were generally inadequate, temperature forecasts were fairly satisfactory. These findings

were indicative of the difficulties posed in downscaling daily precipitation. When downscaled outputs were compared through deterministic and probabilistic approaches, inconsistent and mixed performances were observed. The PLS model appeared to perform better when evaluated using deterministic metrics, whereas the KNN model exhibited greater skill when evaluated using probabilistic approaches. The RAW-based forecasts, however, yielded inferior performance overall using both diagnostic measures.

2. Hydrologic forecasting: The downscaled and the RAW model outputs were forced into an HBV hydrologic model in order to predict 14 day lead reservoir inflows. Three deterministic and one probabilistic hydrologic forecast cases were considered: (a) deterministic hydrologic forecasts obtained when (i) forced with downscaled outputs derived from predictors of the ensemble mean; (ii) forced with the mean of downscaled output derived from predictors of individual members; (iii) considering the mean of ensemble flow output; and (b) probabilistic hydrologic forecasts obtained from 15 flow series corresponding to 15 members. The performances of hydrologic forecasts shown under these conditioning scenarios were, in general, consistent with the results shown in the downscaling experiments. When evaluated under deterministic diagnostic measures, the skill of PLS-based flows showed an edge over its counterparts and, when the probabilistic diagnostic approaches were considered, the KNN-based flows yielded greater skill values. In comparison, hydrologic forecasts based on (ii) above provided relatively more accurate forecasts. The deterministic forecast skills (i.e. CE) associated with the PLS and the KNN-based flows were approximately 72%, for the short lead forecast ranges, and 60%, for the long lead forecast ranges. The RAW-based flows, on the other hand, provided inferior performances when compared with climatology, with the exception of the first few forecast ranges, where quite modest skill values were observed. In the case of probabilistic flow forecasts, the relative performance of flow forecasts was fairly consistent with the downscaled results overall. In comparison, the KNN-based flows

yielded greater skill values, followed by the PLS, under the vast majority of diagnostic measures, including BSS, discrimination and reliability diagram.

3. Economic value of hydrologic forecasts: The potential economic values of flow forecasts were assessed based on a simple optimal decision-making, cost-loss analysis technique. The potential economic values associated with the various hydrologic forecast systems considered were quantified and inter-compared. In the case of deterministic forecasts, the mean of ensemble flows (i.e. from (ii) above) provided greater benefits, signifying the potential for much wider users. The range of potential benefits included: $0.04 < C/L < 0.75$ for the KNN-based flows; $0.04 < C/L < 0.61$ for the PLS-based flows; and $0.03 < C/L < 0.38$ for the RAW-based flows. In general, the potential economic benefits of probabilistic forecasts associated with the KNN, PLS and RAW-based flows were greater than their best deterministic counterparts. Furthermore, the KNN-based flows (deterministic/probabilistic) were more useful than the PLS-based flows (deterministic/probabilistic), and the PLS-based flows were, in turn, more useful than the RAW-based flows (deterministic/probabilistic). These findings strongly suggest the potential added value which can be incurred as a result of adequate downscaling, as opposed to using the RAW model output, for hydrologic applications.

There has been substantial progress in the past several years towards advancing the accuracy of hydrologic forecasts. However, several difficulties still remain which greatly concern the research community. Below are a few remarks on some key aspects in the framework of improving the accuracy of hydrologic forecasts. (i) The spatial resolution of the current reforecast system is too coarse to adequately resolve important sub-grid scale features such as clouds and topography. Improving the resolution of the current reforecast system may offer great opportunities for better characterization of local-scale hydrological variables. (ii) Direct use of the RAW reforecasts for hydrologic applications has been found to yield less skill (Clark and Hay, 2004). In order to advance the skill of hydrologic forecasts, downscaled outputs have been used. The failure of current downscaling approaches to adequately reproduce daily precipitation poses further

challenges, signifying the need for novel and promising techniques. (iii) In addition to the various forcing methods investigated in the present study, further research is warranted in exploring proper forcing techniques for different (similar) hydrologic applications. (iv) Evaluating the performance of forecasts can be quite challenging, particularly for daily total precipitation. Different diagnostic measures offer different (or inconsistent) insights about the nature of the forecast system (e.g., among deterministic-deterministic, probabilistic-probabilistic and deterministic-probabilistic). For instance, a precipitation forecast system which is identified to be superior on the basis of certain diagnostic metrics, may not, in fact, be the one which provides superior flow forecasts. Thus the ability to understand how the various precipitation forecast metrics translate into skillful flow forecasts is quite imperative. (v) Hydrologic forecast uncertainties may arise from model inputs, model structure and model parameters. Ensemble flow forecast systems, which use a single hydrologic model, could capture some input uncertainties, but not a significant portion. Super-ensemble forecast systems, which utilize multi-model flow outputs, may effectively capture most of the hydrologic forecast uncertainties; thereby offering improved forecast skill and value. (vi) Model-data assimilation is another promising method recently emerged in hydrologic applications (e.g., Lohmann et al., 2004; Rudiger, 2007; Barrett et al., 2008). The method provides an effective mechanism to combine observations with physical models in order to advance the skill of hydrologic forecasts through estimation of optimal model state variables.

Acknowledgements

This research was supported by the School of Graduate Studies at McMaster University. The author is grateful to Dr. Paulin Coulibaly and to Alcan Company for making the experiment data available. The author is also grateful to Dr. Noel Evora for providing the pre-processed ensemble weather predictors for the study area. The ensemble reforecast data is made available by NOAA at: http://www.cdc.noaa.gov/reforecast/. The author is also grateful to the Swedish Meteorological and Hydrological Institute (SMHI) for providing the HBV model. Minitab Statistical Software is used to conduct statistical tests.

References

Barrett, D.J., Kuzmin, V.A., Walker, J.P., McVicar, T.R., and Draper, C. (2008) Improving stream flow forecasting by integrating satellite observations, in situ data and catchment models using model-data assimilation methods. eWater Technical Report. eWater Cooperative Research Centre, Canberra, available from http://ewatercrc.com.au/ reports/ Barrett_et_al-2008-Flow_Forecasting.pdf.

Bergstrom, S. (1995) The HBV model, Chapter 13 of Computer models of watershed hydrology. Water Resources Publications, 443-476.

Bergstrom, S. (2002) The HBV Model—past, present and the future. In XXII Nordic Hydrological Conference, 4–7 August 2002, Roros.

Bergstrom, S., Carlsson, B., Gardelin, M., Lindstrom, G., Pettersson, A., and Rummukainen, M. (2001) Climate change impacts on the runoff in Sweden - assessments by global climate models, dynamical downscaling and hydrological modeling. Climate Research, 16, 101-112.

Bergstrom, S., Lindstrom, G., and Pettersson, A. (2002) Multi-variable parameter estimation to increase confidence in hydrological modeling. Hydrological Processes, 16, 413–421.

Buizza, R. (2001) Accuracy and economic value of categorical and probabilistic forecasts of discrete events. Monthly Weather Review, 129, 2329–2345.

Clark, M.P., and Hay, L.E. (2004) Use of medium-range numerical weather prediction model output to produce forecasts of streamflow. J. Hydrometeorol., 5(1), 15– 32.

Day, G.N. (1985) Extended streamflow forecasting using NWSRFS. ASCE J. Water Res. Plann. Manage., 111, 157–170.

de Jong, S. (1993) SIMPLS: an alternative approach to partial least squares regression. Chemom. Intell. Lab. Syst., 18, 251-263.

Dibike, Y.B., and Coulibaly, P. (2005) Hydrologic impact of climate change in the Saguenay watershed: comparison of downscaling methods and hydrologic models. J. Hydrol., 307 (1-4), 145-163.

Fowler, H.J., Blenkinsopa, S., and Tebaldib, C. (2007) Review Linking climate change modelling to impacts studies: recent advances in downscaling techniques for hydrological modeling. Int. J. Climatol., 27, 1547–1578

Franz, K.J., Hartmann, H.H., Sorooshian, S., and Bales, R. (2003) Verification of National Weather Service Ensemble Streamflow Predictions for Water Supply Forecasting in the Colorado River Basin. American Meteorological Society, 4, 1105-1118.

Gangopadhyay, S., Clark, M., and Rajagopalan, B. (2005) Statistical downscaling using K-nearest neighbors. Water Resour. Res., 41 (2), W02024, doi:10.1029/2004 WR 003444.

Hamill, T.M., Whitaker, J.S., and Wei, X. (2004) Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. Mon. Weather Rev., 132, 1434– 1447.

Harpham, C., and Wilby, R.L. (2005) Multi-site downscaling of heavy daily precipitation occurrence and amounts. J. Hydrol, 312, 235–255.

Hay, L.E., and Clark, M.P. (2003) Use of statistically and dynamically downscaled atmospheric model output for hydrologic simulations in three mountainous basins in the western United States. J. Hydrol., 282 (1-4), 56-75.

Lehmann, E.L. (1975). Nonparametrics: Statistical Methods Based on Ranks, Holden and Day, San Francisco.

Levene, H. (1960). Contributions to Probability and Statistics, Stanford University Press.

Li, H., Luo, L., Wood, E.F., and Schaake, J. (2009) The role of initial conditions and forcing uncertainties in seasonal hydrologic forecasting. J. Geophys. Res., 114, D04114, doi:10.1029/2008JD010969.

Lindstrom, G., Johansson, B., Persson, M., Gardelin, M., and Bergstrom, S. (1997) Development and test of the distributed HBV- 96 model. J. Hydrol., 201, 272-288.

Lohmann, D., Mitchell, K.E., Houser, P.R., Wood, E.F., Schaake, J.C., Robock, A., Cosgrove, B.A., Sheffield, J., Duan, Q., Luo, L., Higgins, R.W., Pinker, R.T., and Tarpley, J.D. (2004) Streamflow and water balance intercomparisons of four land

surface models in the North American Land Data Assimilation System project. Journal of Geophysical Research, 109: D07S91. doi:10.1029/2003JD003517.

Lorber, A., Wangen, L.E., and Kowalski, B.R. (1987) A Theoretical Foundation for the PLS Algorithm. J. Chemometrics, 1, 19-31.

Mearns, L.O., Bogardi, I., Giorgi, F., Matyasovszky, I., and Palecki, M. (1999) Comparison of climate change scenarios generated from regional climate model experiments and statistical downscaling. J. Geophys. Res., 104, 6603– 6621.

Menzel, L., and Burger, G. (2002) Climate change scenarios and runoff response in the Mulde catchment (Southern Elbem Germany). J. Hydrol., 267, 53–64.

Muluye, G.Y. (2010) Comparison of statistical methods for downscaling daily precipitation, Journal of Hydrology, Manuscript No. HYDROL10158.

Murphy, A.H. (1977) The value of climatological, categorical and probabilistic forecasts in the cost-loss ratio situation. Mon. Weather Rev., 105, 803-816.

Murphy, A.H. (1997) Forecast Verification. In Economic Value of Weather and Climate Forecasts, R. W. Katz and A. H. Murphy, eds., Cambridge University Press, 19–74.

Murphy, A.H., and Winkler, R.L. (1987) A general framework for forecast verification. Mon. Wea. Rev., 115, 1330–1338.

Murphy, A.H., Brown, B. G., and Chen, Y. (1989) Diagnostic verification of temperature forecasts. Wea. Forecasting, 4, 485–501.

Murphy, J.M. (1999) An evaluation of statistical and dynamical techniques for downscaling local climate. J. Clim., 12, 2256–2284.

Richardson, D.S. (2000) Skill and economic value of the ECMWF Ensemble Prediction System. Quarterly Journal of the Royal Meteorological Society, 126, 649–668.

Roulin, E. (2007) Skill and relative economic value of medium-range hydrological ensemble predictions, Hydrol. Earth Syst. Sci., 11, 725–737.

Rudiger, C. (2007) Streamflow data assimilation for soil moisture prediction. University of Melbourne, PhD Dissertation.

Schaake, J., Franz, K., Bradley, V., and Buizza, R. (2006) The Hydrologic Ensemble Prediction EXperiment (HEPEX). Hydrol. Earth Syst. Sci., 3, 3321-3332.

Shi, X., Wood, A.W., and Lettenmaier, D.P. (2008) How Essential is Hydrologic Model Calibration to Seasonal Streamflow Forecasting. J. Hydrometeorol., 9, 1350-1363.

Singh, V.P. (1995) Computer Models of Watershed Hydrology. Water Resources Publications, Highlands Ranch, CO.

Sorman, A.A., Sensoy, A., Tekeli, A.E., Sorman, A.U., and Akyurek, Z. (2009) Modelling and forecasting snowmelt runoff process using the HBV model in the eastern part of Turkey. Hydrol. Process., 23, 1031–1040.

Spak, S., Holloway, T., Lynn, B., and Goldberg, R. (2007) A comparison of statistical and dynamical downscaling for surface temperature in North America. J. Geophys. Res., 112, D08101, doi:10.1029/2005JD006712.

Stanski, H. R., Wilson, L. J., and Burrows, W. R. (1989) Survey of Common Verification Methods in Meteorology. WMO World Weather Watch Tech. Report No. 8, WMOI TD No. 358, 114 pp.

Tocci, C.E.M., Collischonn, W., Clarke, R.T., Paz, A.R., and Allasia, D. (2008) Short- and long-term flow forecasting in the Rio Grande watershed (Brazil). Atmos. Sci. Let., 9, 53–56.

von Storch, H., Zorita, E., and Cubash, U. (1993) Downscaling of global climate change estimates to regional scales: an application to Iberian rainfall in wintertime. J. Clim., 6, 1161–71.

Wang, S., McGrath, R., Semmler, T., and Nolan, P. (2006) The impact of the climate change on discharge of Suir River Catchment (Ireland) under different climate scenarios. Nat. Hazards Earth Syst. Sci., 6, 387–395.

Werner, K., Brandon, D., Clark, M., and Gangopadhyay, S. (2005) Incorporating Medium-Range Numerical Weather Model Output into the Ensemble Streamflow Prediction System of the National Weather Service. J. Hydrometeorol., 6(2), 101-114.

Wilby, R.L., and Wigley, T.M.L. (1997) Downscaling general circulation model output: a review of methods and limitations. Progress in Physical Geography, 21, 530–548.

Wilks, D.S. (1995) Statisistical Methods in the Atmospheric Sciences, Academic Press, San Diego, California.

Wilks, D.S. (2001) A skill score based on economic value for probability forecasts. Meteorol. Appl., 8, 209– 219.

Wold, S., Geladi, P., Esbensen, K., and Ohman, J. (1987) Multi-Way Principal Components and PLS-Analysis. J. Chemometrics, 1, 41-56.

Xu, C.Y. (1999) From GCMs to river flow: a review of downscaling methods and hydrologic modeling approaches. Progress in Physical Geography, 23(2), 229–249.

Yakowitz, S. (1993) Nearest neighbor regression estimation for null-recurrent Markov time series. Stochastic Processes Their Appl., 48, 311–318.

Yarnal, B., Comrie, A.C., Frakes, B., and Brown, D.P. (2001) Developments and prospects in synoptic climatology. Int. J. Climatol., 21, 1923–1950.

Yates, D., Gangopadhyay, S., Rajagopalan, B., and Strzepek, K. (2003) A technique for generating regional climate scenarios using a nearest-neighbor algorithm. Water Resour. Res., 39(7), 1199, doi:10.1029/2002WR001769.

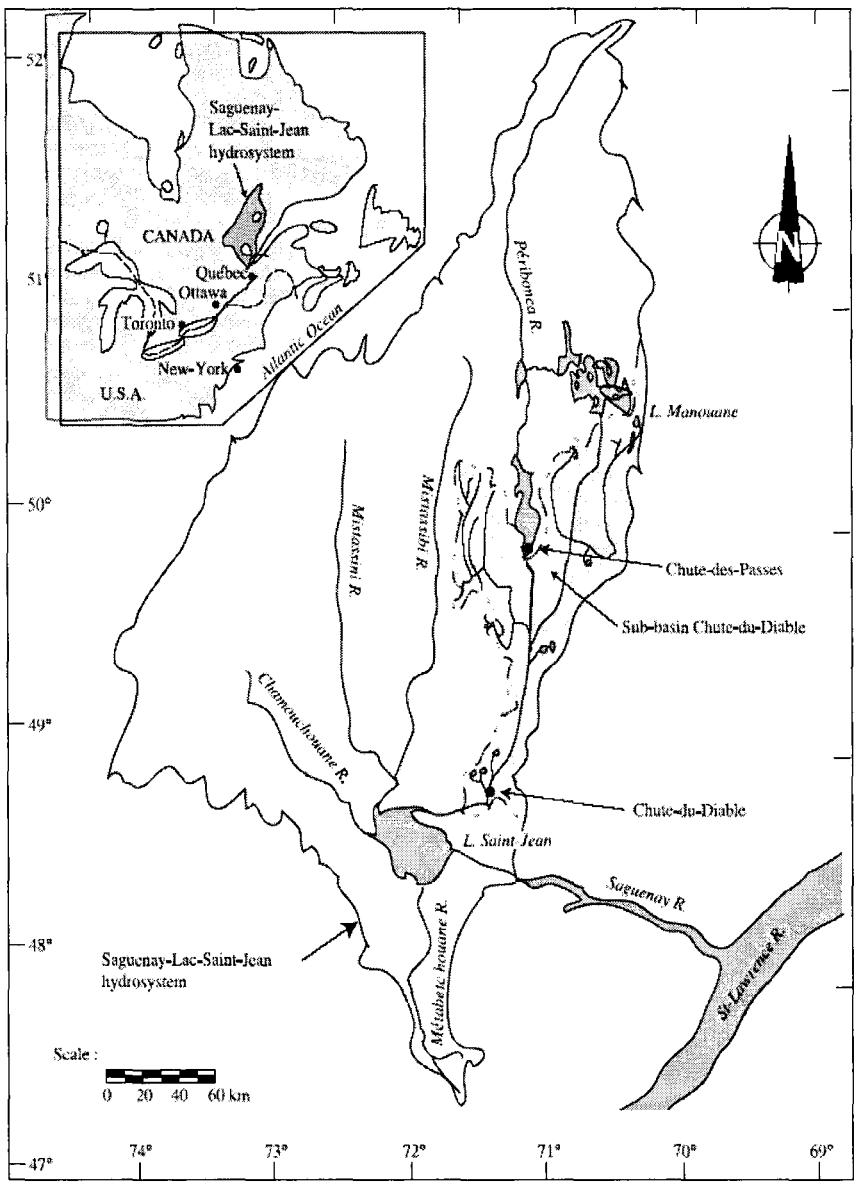


Figure 1. Location map of study area (Source: Dibike and Coulibaly, 2005)

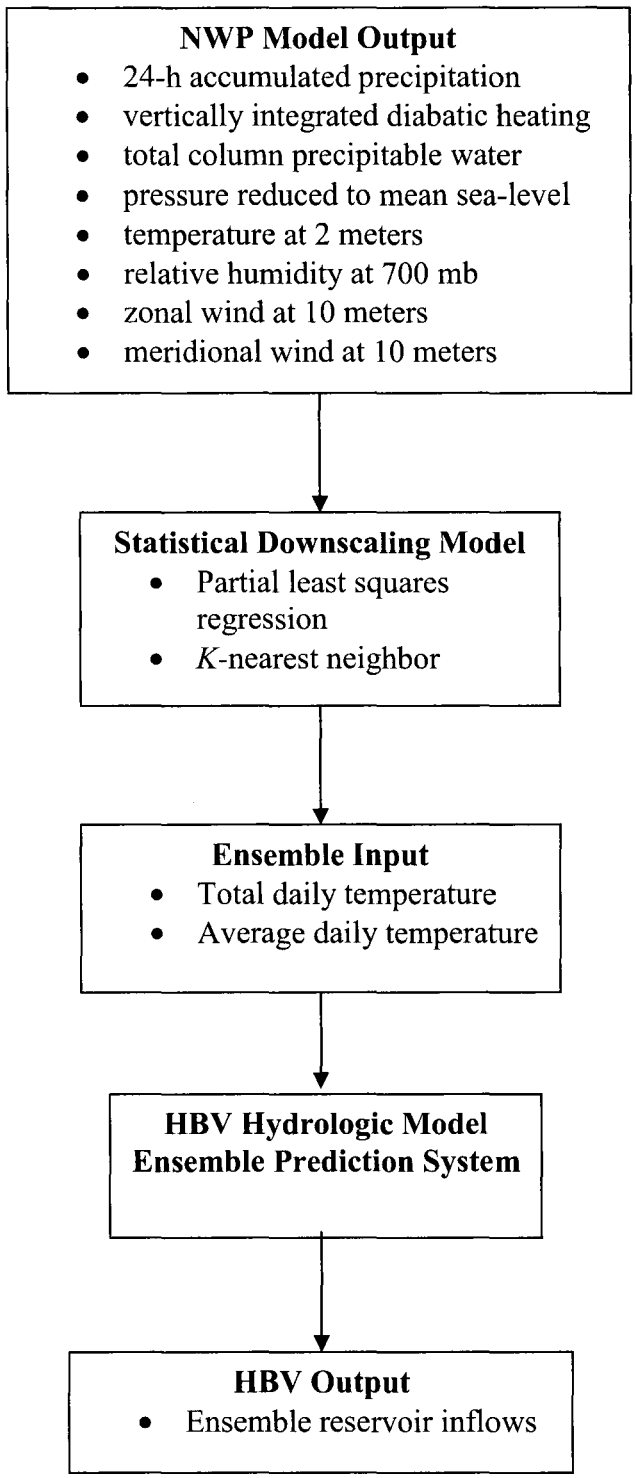


Figure 2. Schematic diagram illustrating ensemble streamflow prediction system.

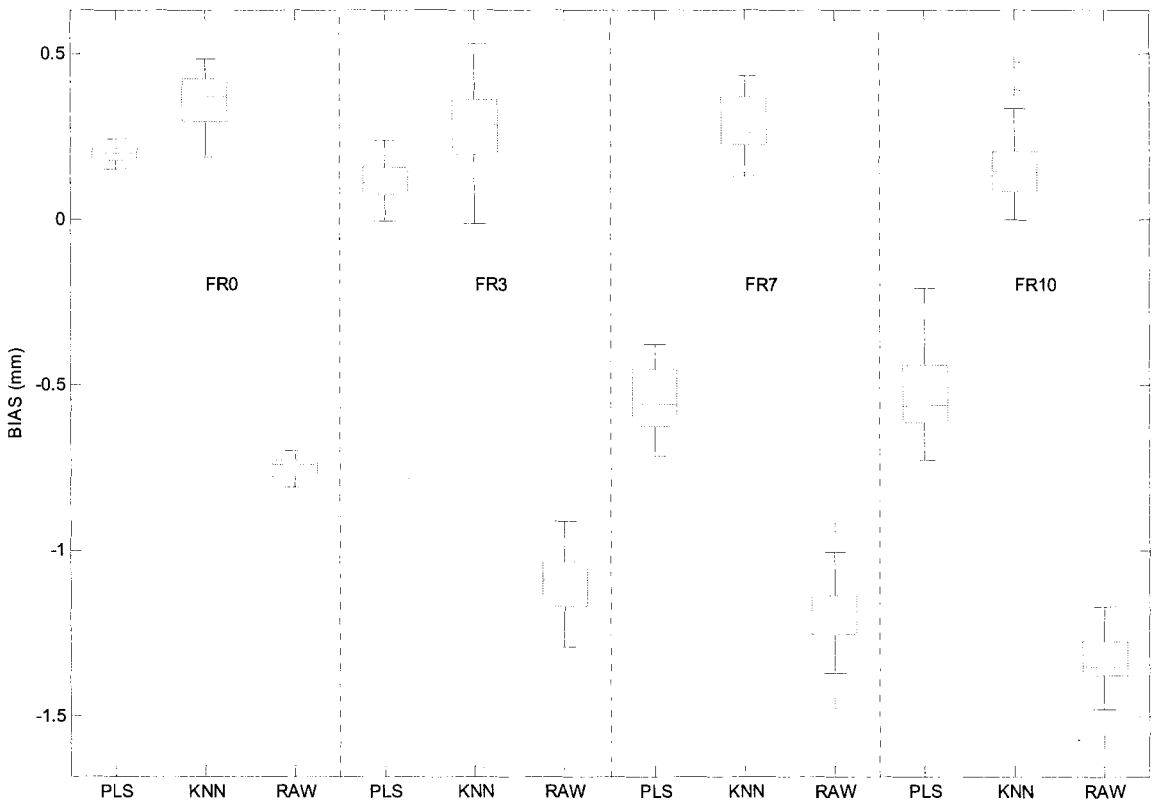


Figure 3. Box plots of biases as a function of FRs in downscaling total daily precipitation. The Box plots were constructed from ensemble of downscaled outputs resulting from PLS, KNN and RAW, for the test period Jan 97 to Dec 01. The box-and-whisker symbols represent the 10^{th}, 25^{th}, 50^{th}, 75^{th} and 90^{th} percentiles of the forecasts.

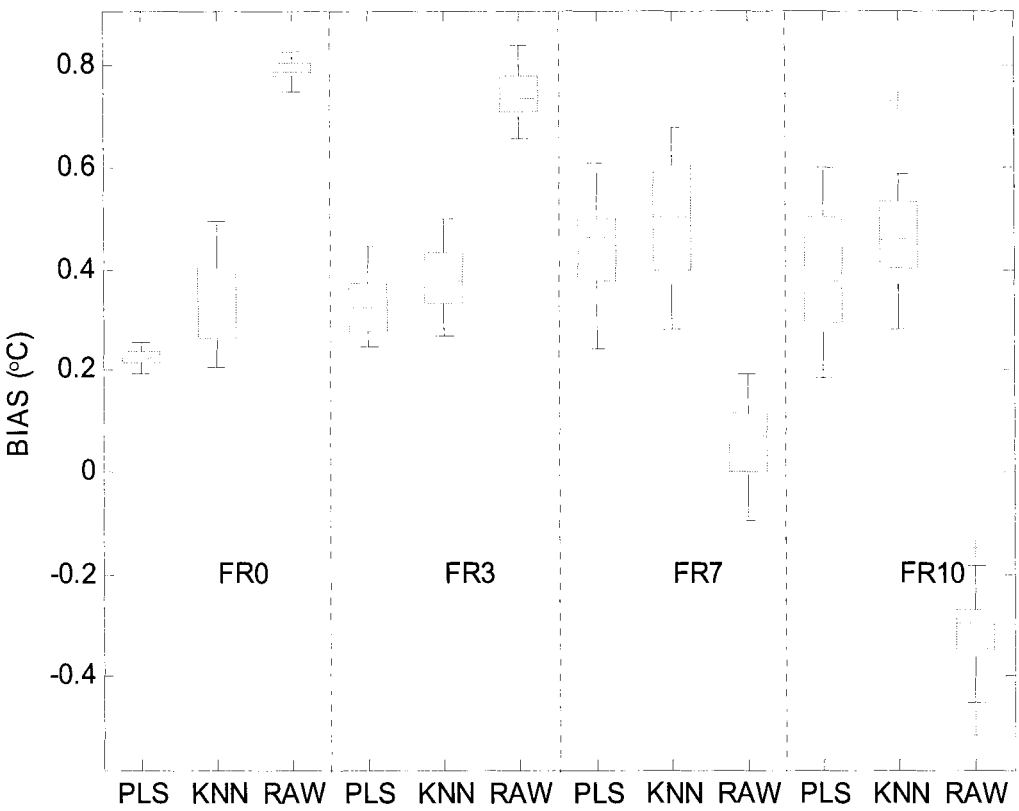


Figure 4. Box plots of biases as a function of FRs in downscaling average daily temperature. The Box plots were constructed from ensemble of downscaled outputs resulting from PLS, KNN and RAW, for the test period Jan 97 to Dec 01. The box-and-whisker symbols represent the 10^{th}, 25^{th}, 50^{th}, 75^{th} and 90^{th} percentiles of the forecasts.

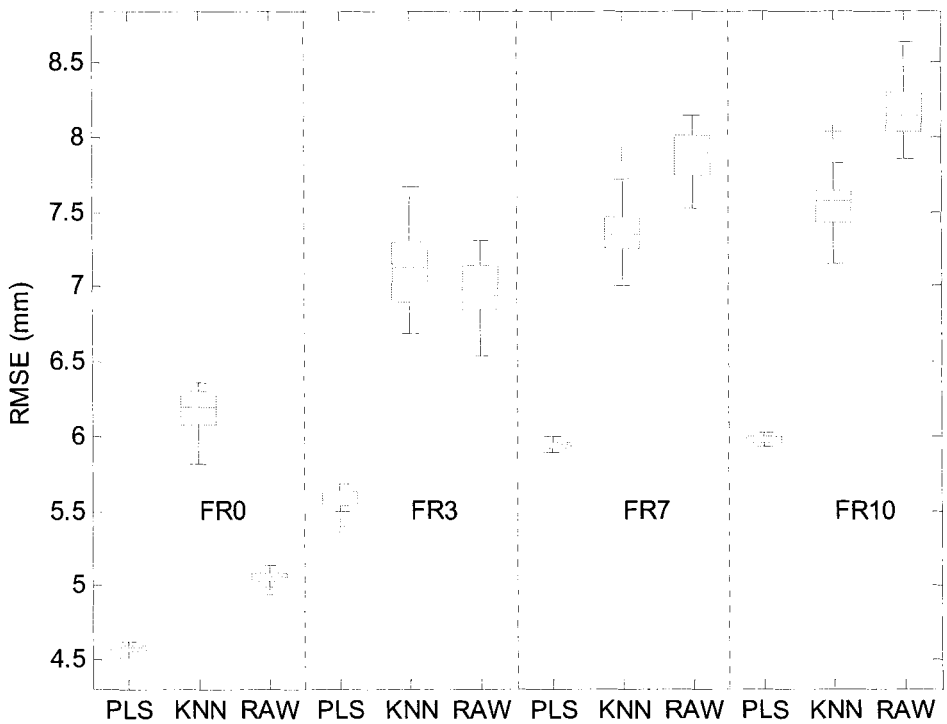


Figure 5. Box plots of RMSE as a function of FRs in downscaling total daily precipitation. The Box plots were constructed from ensemble of downscaled outputs resulting from PLS, KNN and RAW, for the test period Jan 97 to Dec 01.

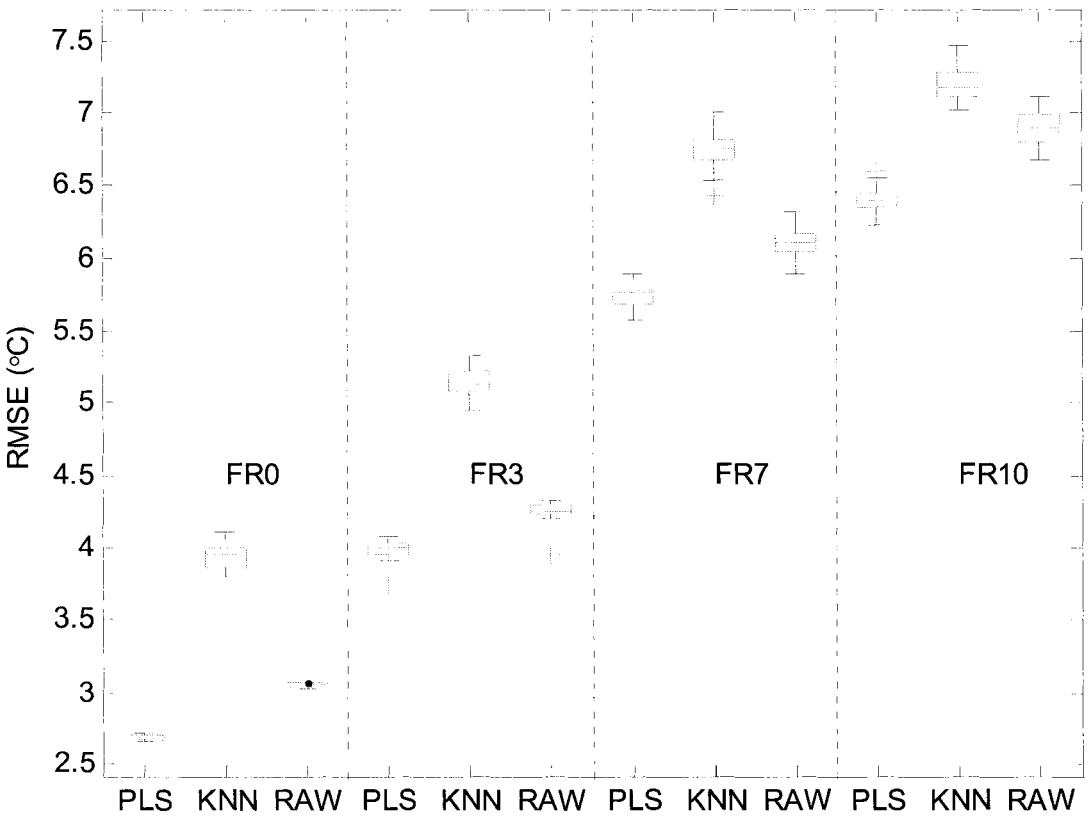


Figure 6. Box plots of RMSE as a function of FRs in downscaling average daily temperature. The Box plots were constructed from ensemble of downscaled outputs resulting from PLS, KNN and RAW, for the test period Jan 97 to Dec 01.

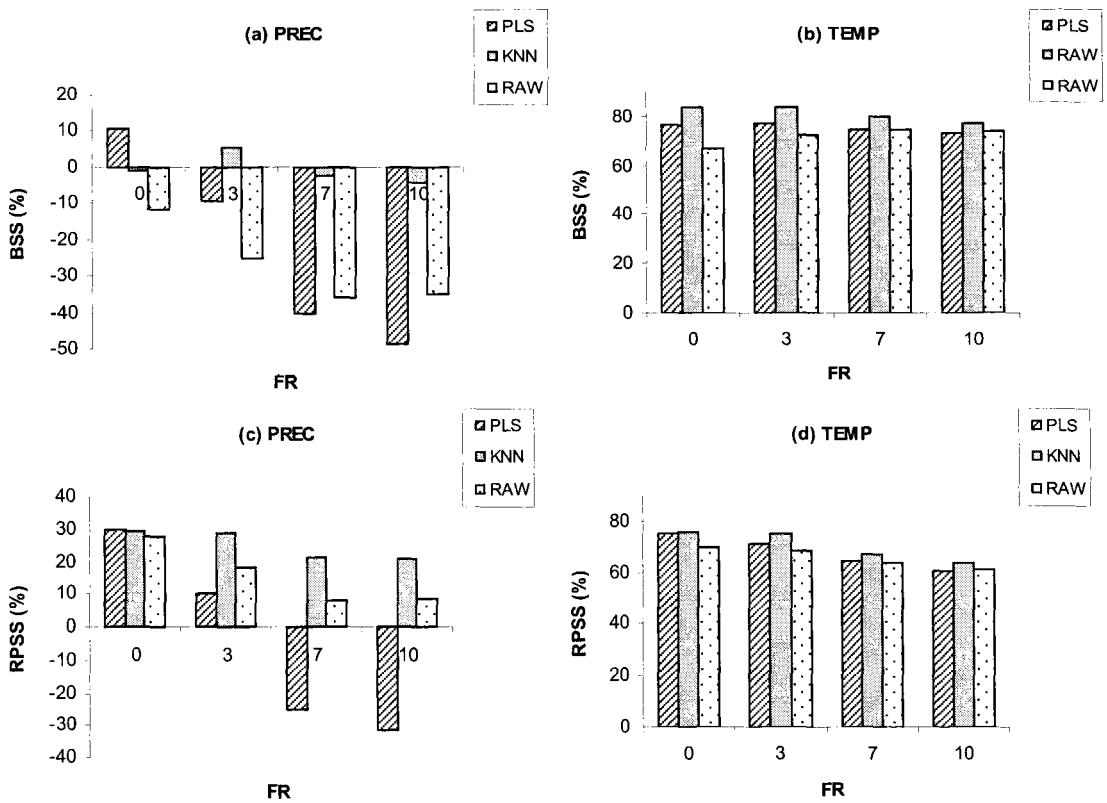


Figure 7. Probabilistic diagnostic statistics: Brier skill score (BSS) and ranked probability skill score (RPSS) as a function of FRs in downscaling total daily precipitation (first column) and average daily temperature (second column). These statistics were constructed from ensemble of downscaled outputs resulting from PLS, KNN and RAW, for the test period Jan 97 to Dec 01.

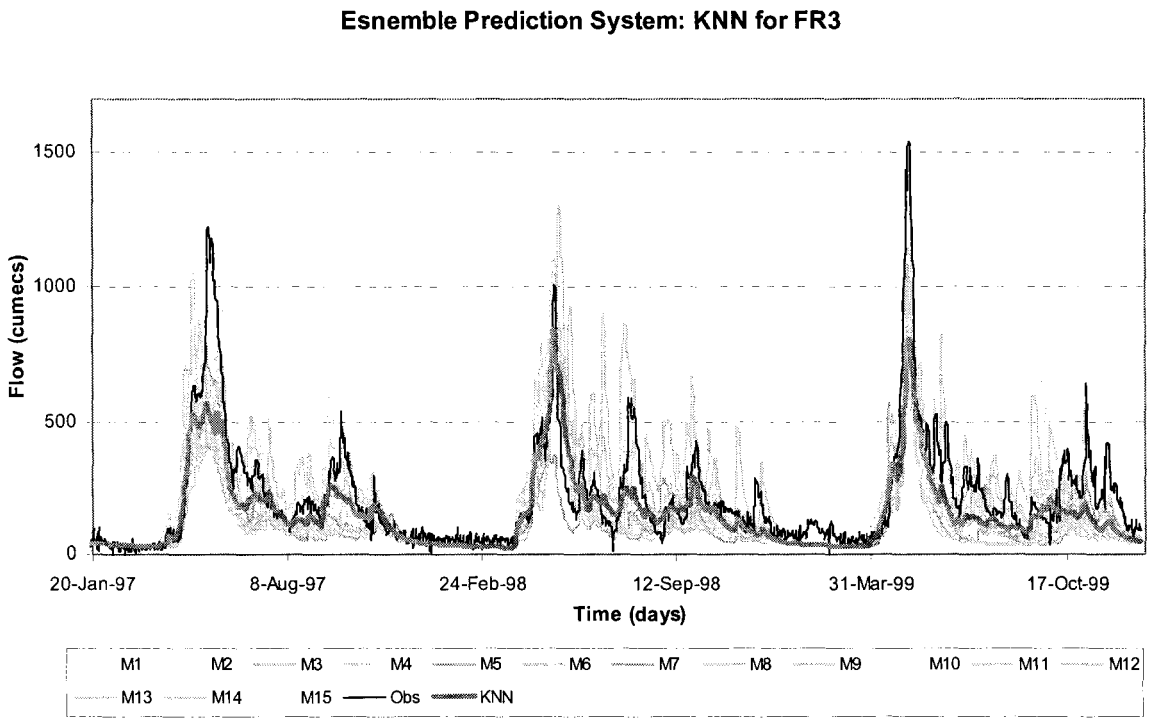


Figure 8. Ensemble Prediction System for a FR of 3 days obtained by forcing KNN-based output into an HBV model, for the portion of test period Jan 97 to Dec 99. The light gray lines represent flows corresponding to each of 15 ensemble members, whereas the thick dark gray line and the dark line represent mean of ensemble flows and observed flows, respectively.

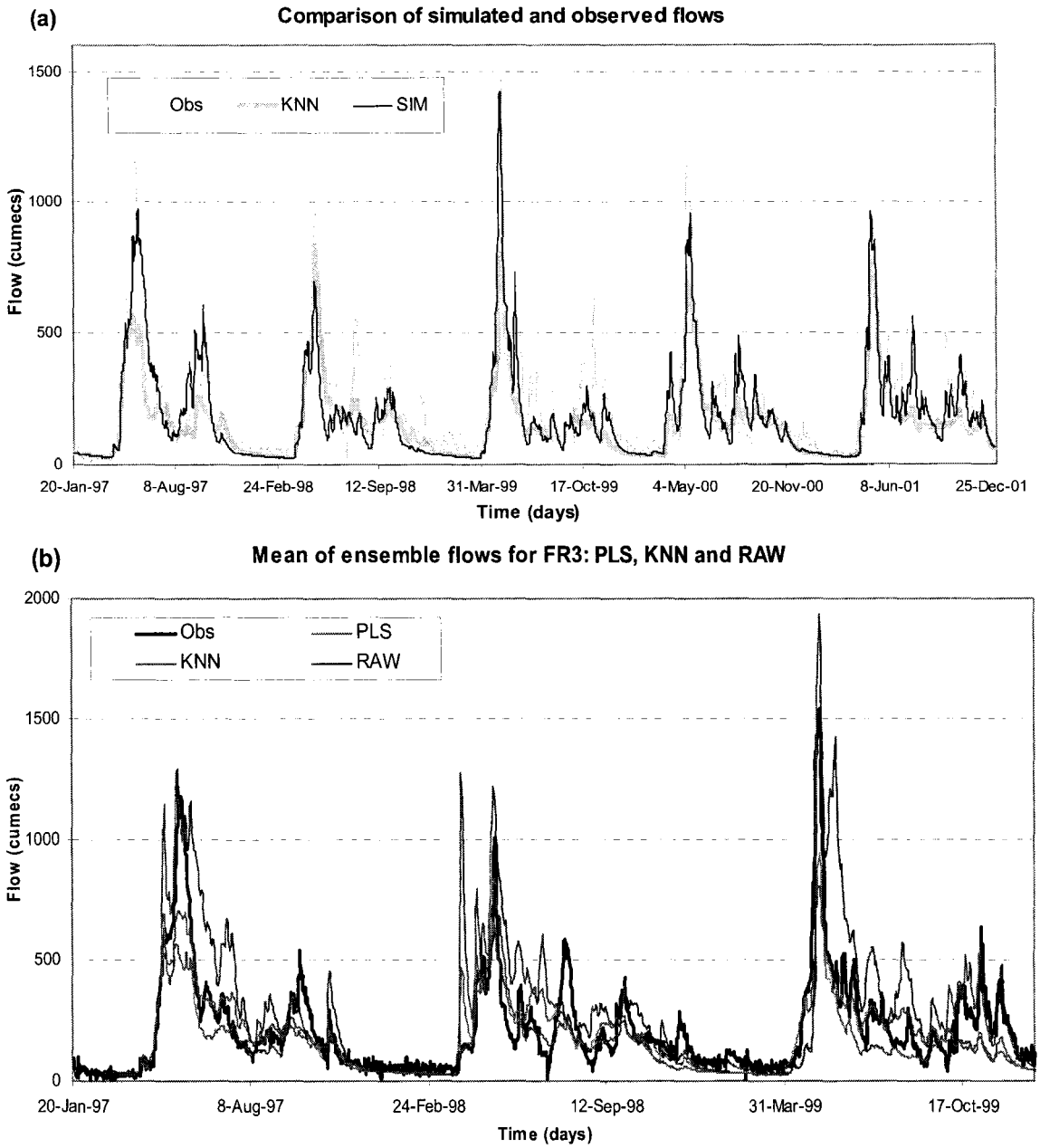


Figure 9. Flow hydrographs of (a) observed, simulated and KNN-based ensemble mean. The thick gray line represents the KNN-based ensemble mean, whereas the light gray line and the dark line represent observed and simulated flows based on observed temperature and precipitation, respectively, for the test period Jan 97 to Dec 01; and (b) mean of ensemble flows for a FR of 3 days obtained by forcing PLS, KNN and RAW-based outputs into an HBV model, for the portion of the test period Jan 97 to Dec 99.

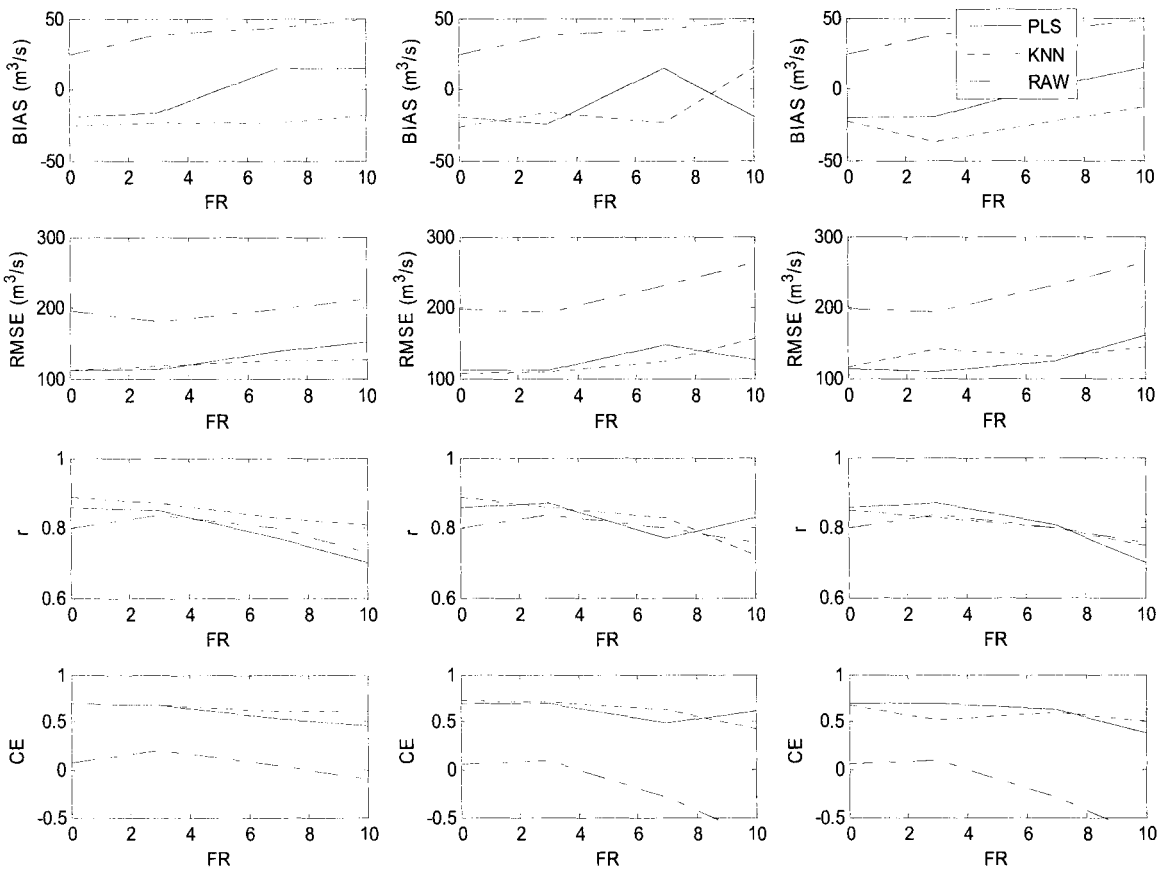


Figure 10. Deterministic diagnostic statistics as a function of FRs. The deterministic hydrologic forecasts were obtained when the HBV model was (i) forced with the mean of ensemble flow output (column 1), (ii) forced with the mean of downscaled output derived from predictors of individual members (column 2), and (iii) forced with the downscaled outputs derived from predictors of ensemble mean (column 3) for the test period from Jan 97 through Dec 01. The forcings were based on PLS, KNN and RAW outputs.

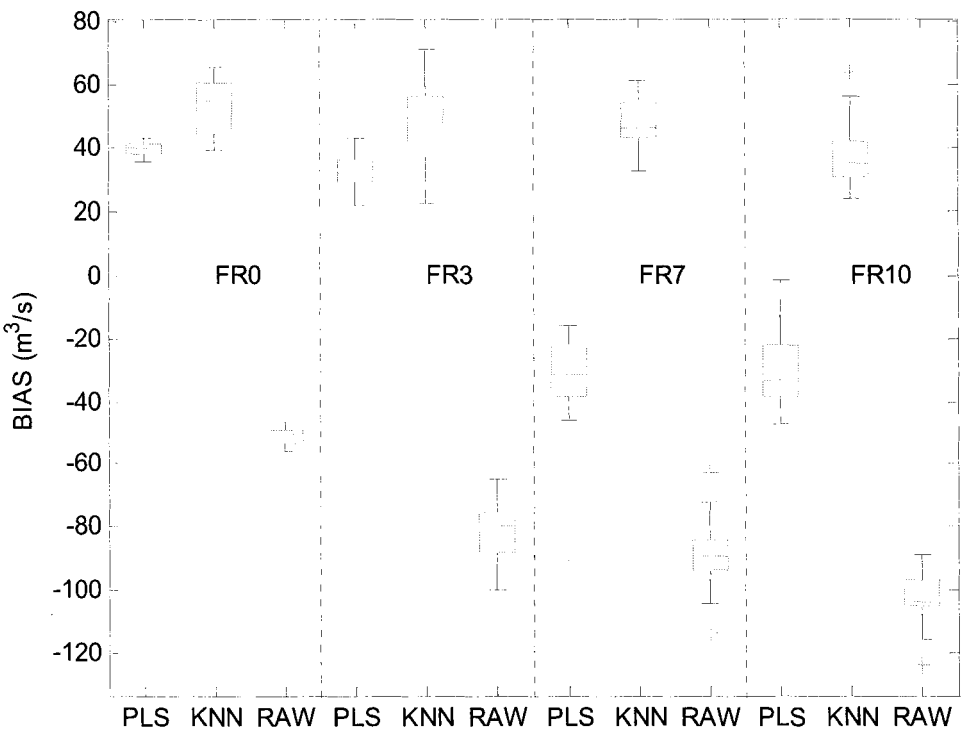


Figure 11. Box plots of biases as a function of FRs. The Box plots were constructed from ensemble of flows generated by forcing PLS, KNN and RAW-based outputs in an HBV model, for the test period Jan 97 to Dec 01.

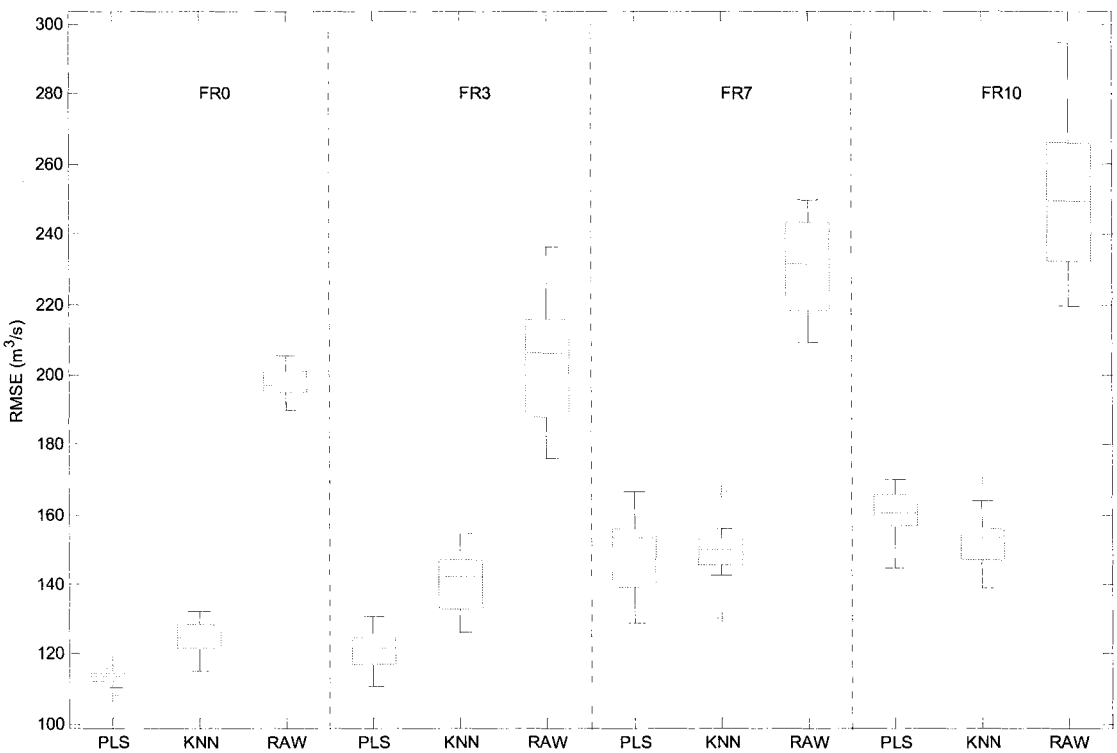


Figure 12. Box plots of RMSE as a function of FRs. The Box plots were constructed from ensemble of flows generated by forcing PLS, KNN and RAW-based outputs in an HBV model, for the test period Jan 97 to Dec 01.

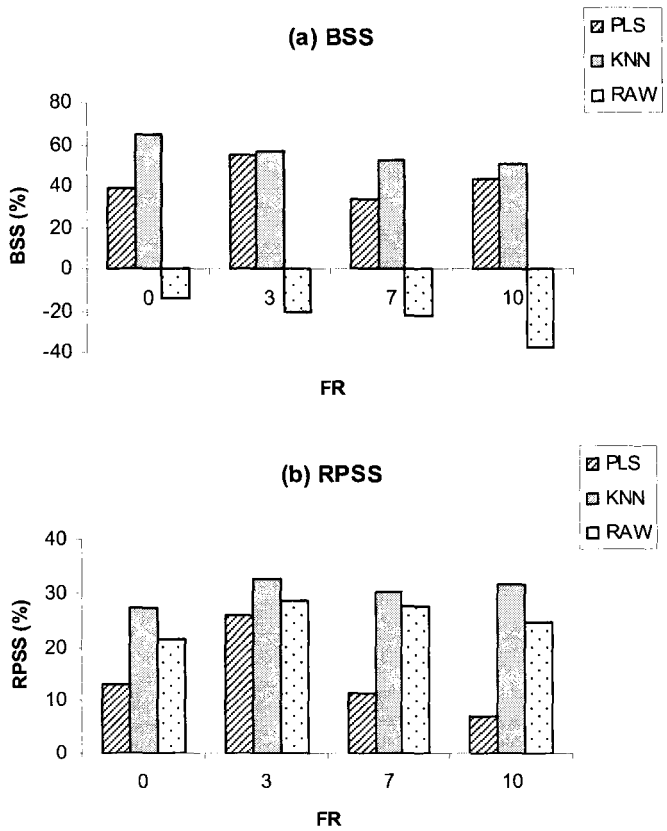


Figure 13. Probabilistic diagnostic statistics as a function of FRs for flow forecasts (a) Brier skill score (BSS), and (b) ranked probability skill score (RPSS). These statistics were constructed from ensemble of flows generated by forcing PLS, KNN and RAW outputs in an HBV model, for the test period Jan 97 to Dec 01.

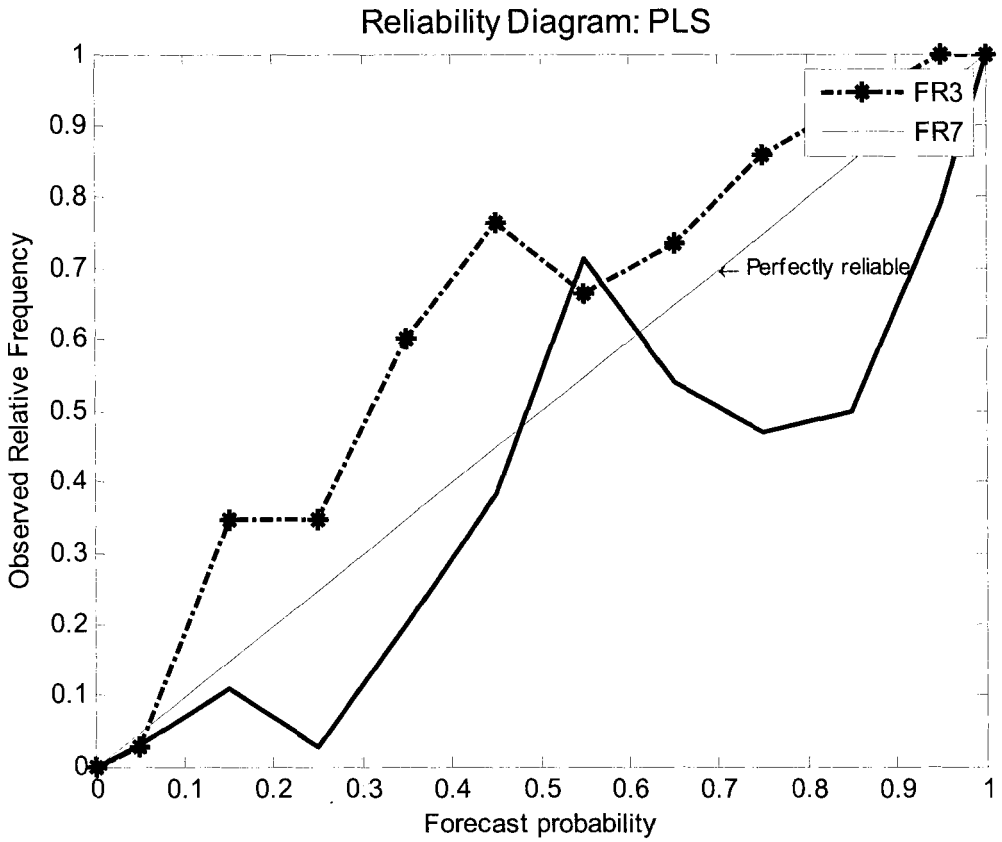


Figure 14. Reliability diagram for FRs of 3 and 7 days. The ensemble flows used to construct the reliability diagram originated from the PLS-based output, for the test period from Jan 97 to Dec 01. The 45° line represents perfectly reliable forecasts.

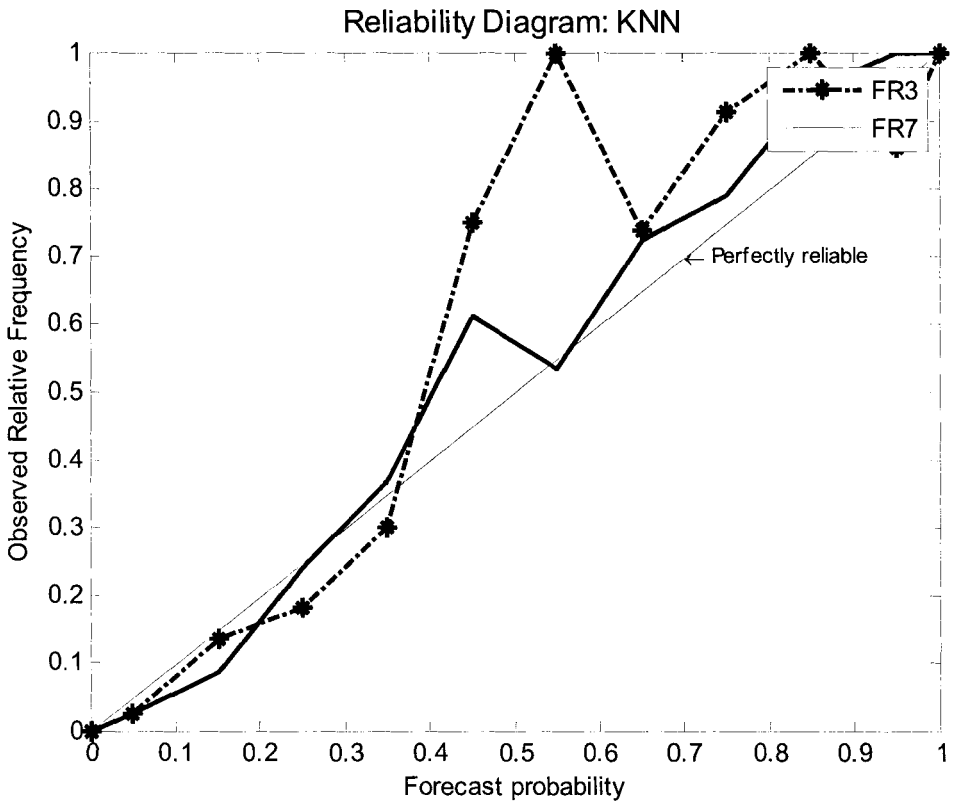


Figure 15. Reliability diagram for FRs of 3 and 7 days. The ensemble flows used to construct the reliability diagram originated from the KNN-based output, for the test period from Jan 97 to Dec 01. The 45° line represents perfectly reliable forecasts.

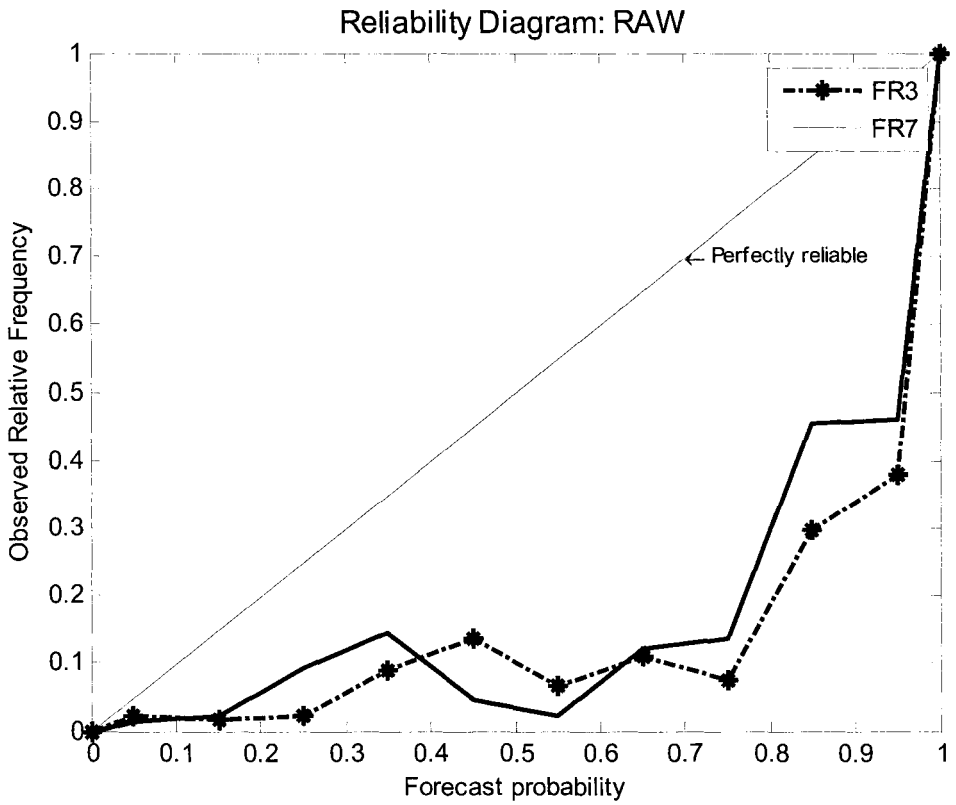


Figure 16. Reliability diagram for FRs of 3 and 7 days. The ensemble flows used to construct the reliability diagram originated from the RAW-based output, for the test period from Jan 97 to Dec 01. The 45° line represents perfectly reliable forecasts.

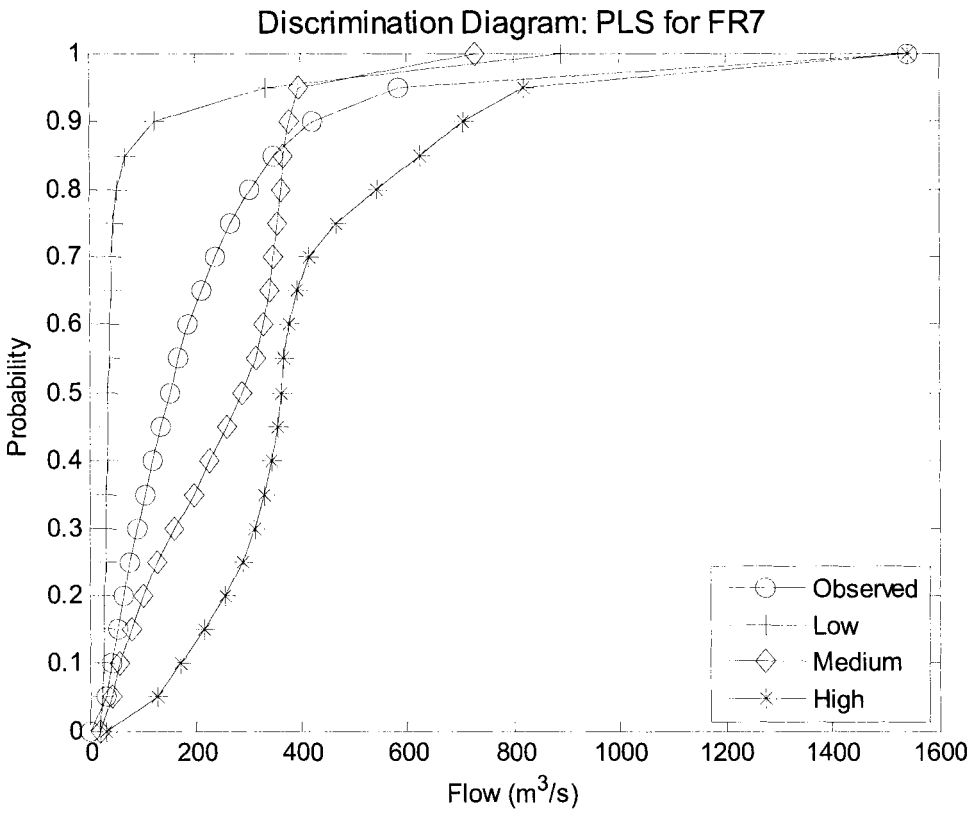


Figure 17. Discrimination diagram for a FR of 7 days. The ensemble flows used to construct the discrimination diagram originated from the PLS-based output, for the test period from Jan 97 to Dec 01. Discriminatory forecasts tend to separate among low-, medium- and high-flow portions of the observed time series, and avoid overlapping.

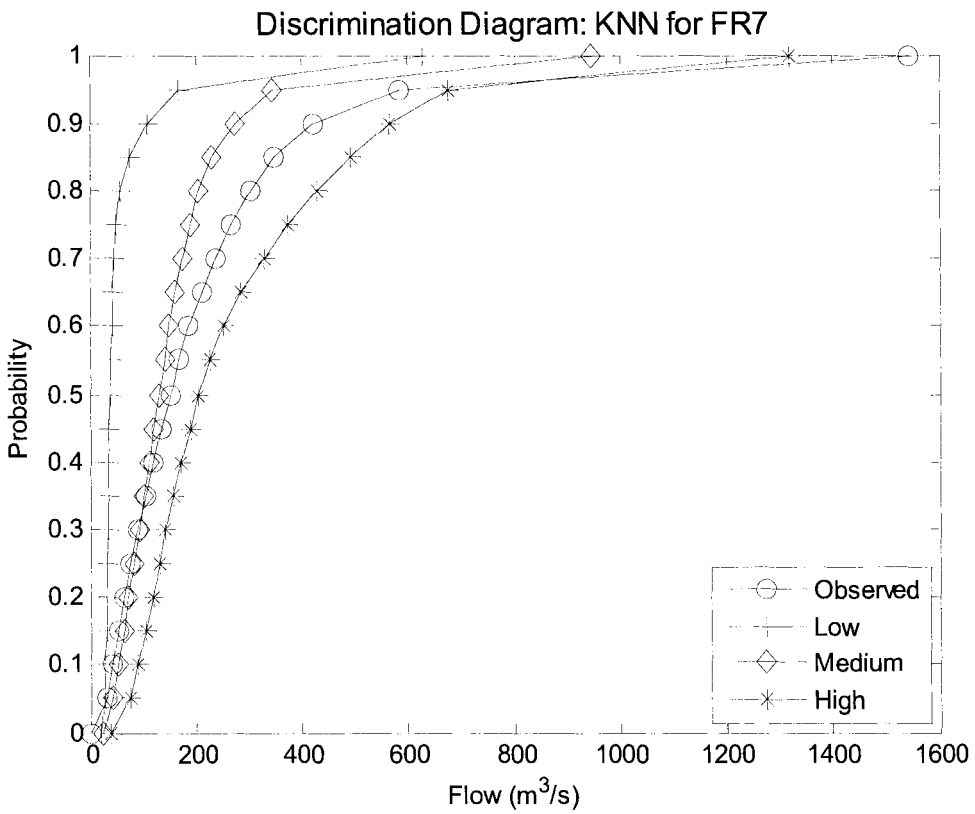


Figure 18. Discrimination diagram for a FR of 7 days. The ensemble flows used to construct the discrimination diagram originated from the KNN-based output, for the test period from Jan 97 to Dec 01. Discriminatory forecasts tend to separate among low-, medium- and high-flow portions of the observed time series, and avoid overlapping.

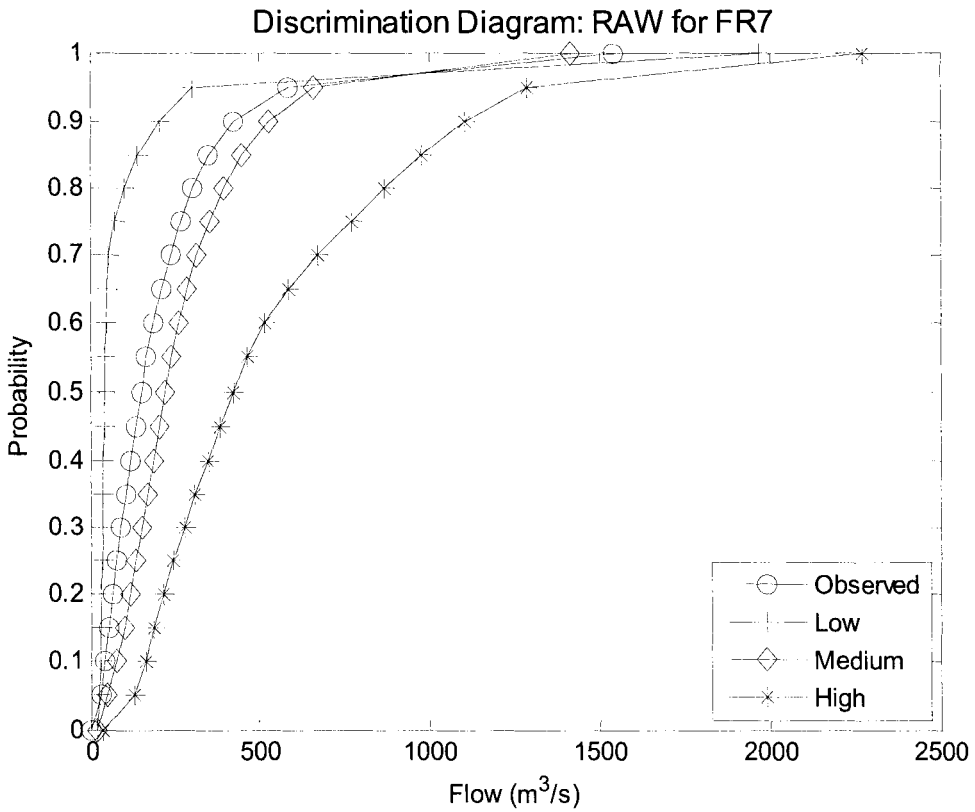


Figure 19. Discrimination diagram for a FR of 7 days. The ensemble flows used to construct the discrimination diagram originated from the RAW-based output, for the test period from Jan 97 to Dec 01. Discriminatory forecasts tend to separate among low-, medium- and high-flow portions of the observed time series, and avoid overlapping.

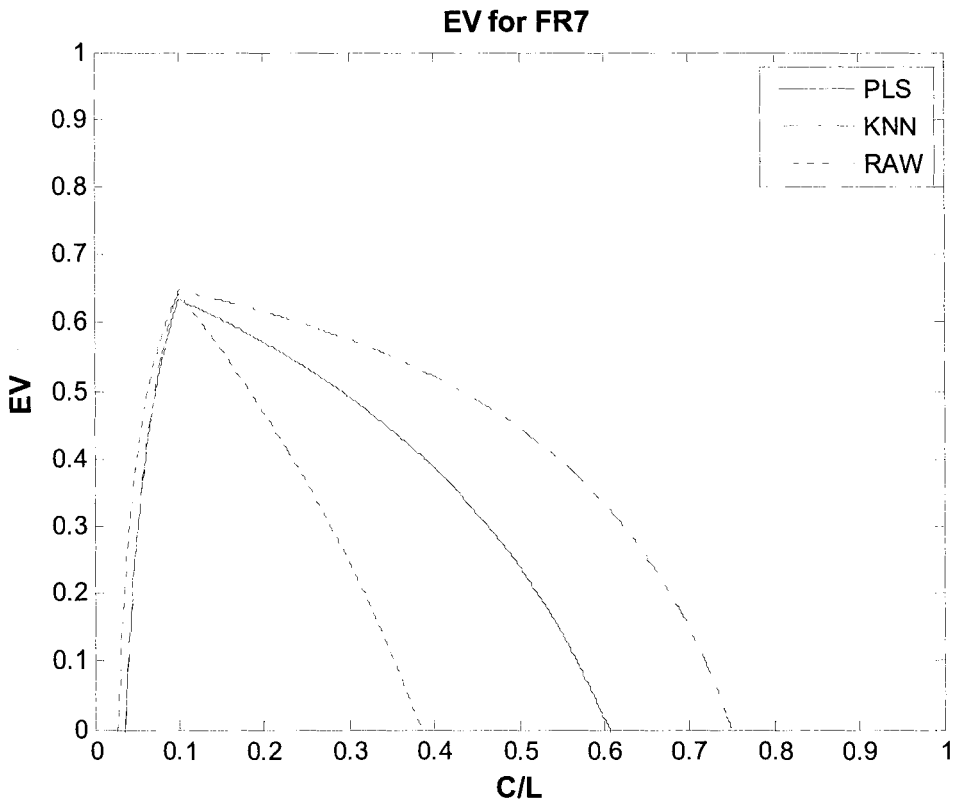


Figure 20. Potential EV of deterministic flow forecasts in excess of the 90th percentile for a FR of 7 days, for the test period Jan 97 to Dec 01. Flow forecasts generated from the mean of downscaled members were used to construct the EV curves: PLS (solid line), KNN (pecked line), and RAW (dotted line).

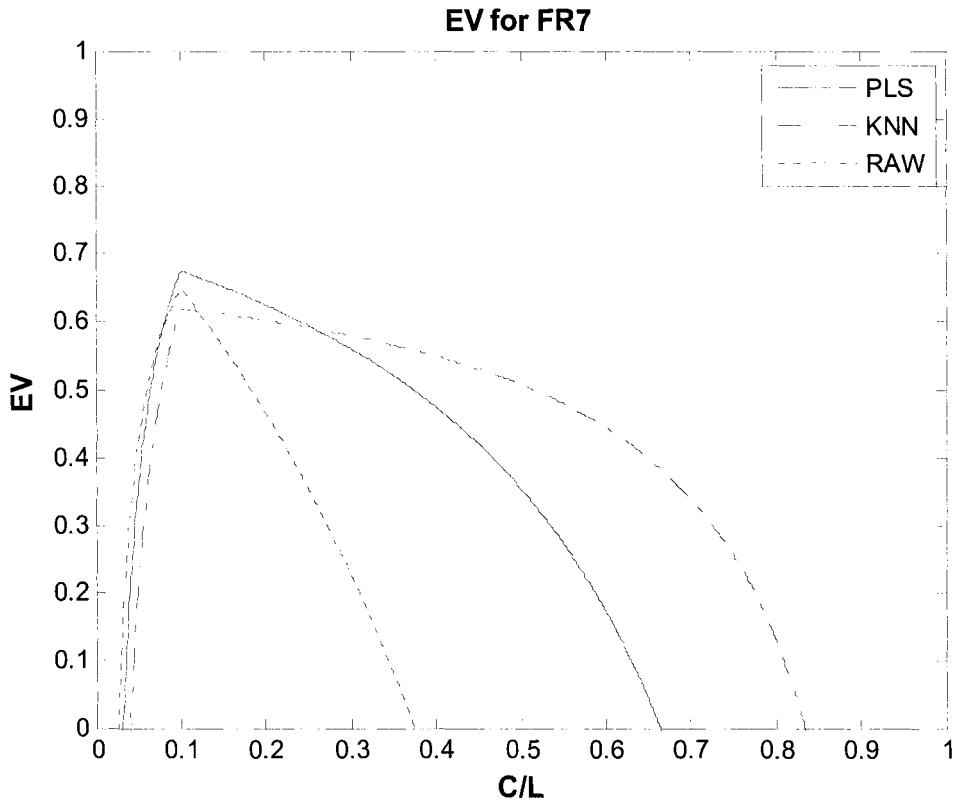


Figure 21. Potential EV of deterministic flow forecasts in excess of the 90th percentile for a FR of 7 days, for the test period Jan 97 to Dec 01. Flow forecasts generated from the mean of ensemble flows were used to construct the EV curves: PLS (solid line), KNN (pecked line), and RAW (dotted line).

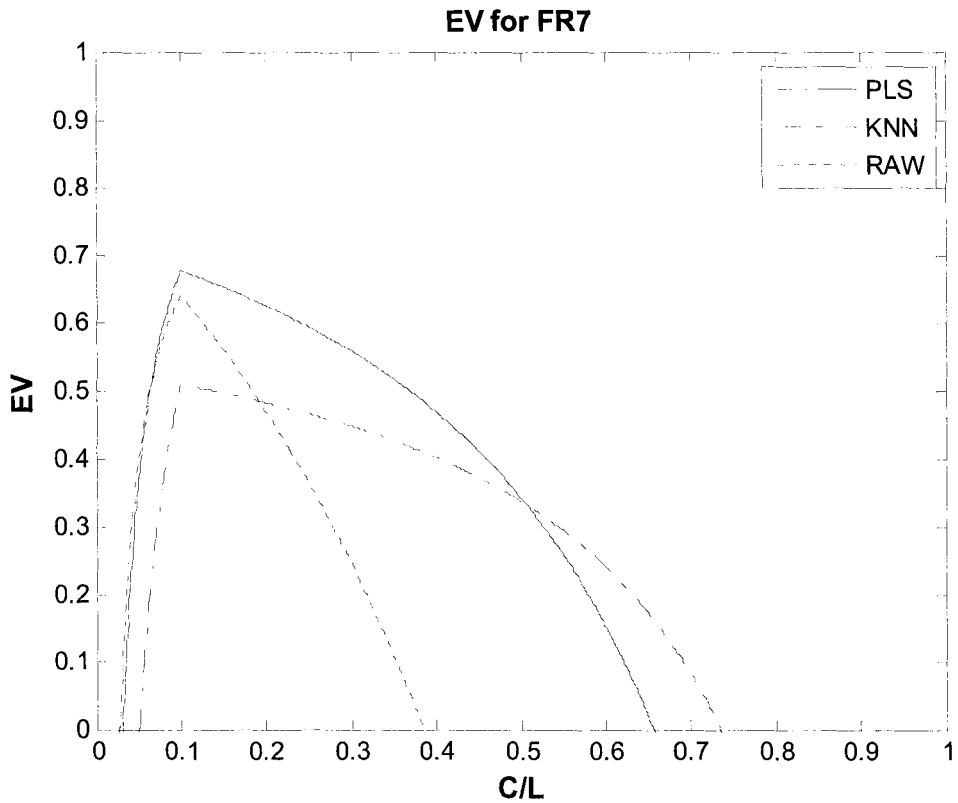


Figure 22. Potential EV of deterministic flow forecasts in excess of the 90th percentile for a FR of 7 days, for the test period Jan 97 to Dec 01. Flow forecasts generated from the downscaled mean predictor were used to construct the EV curves: PLS (solid line), KNN (pecked line), and RAW (dotted line).

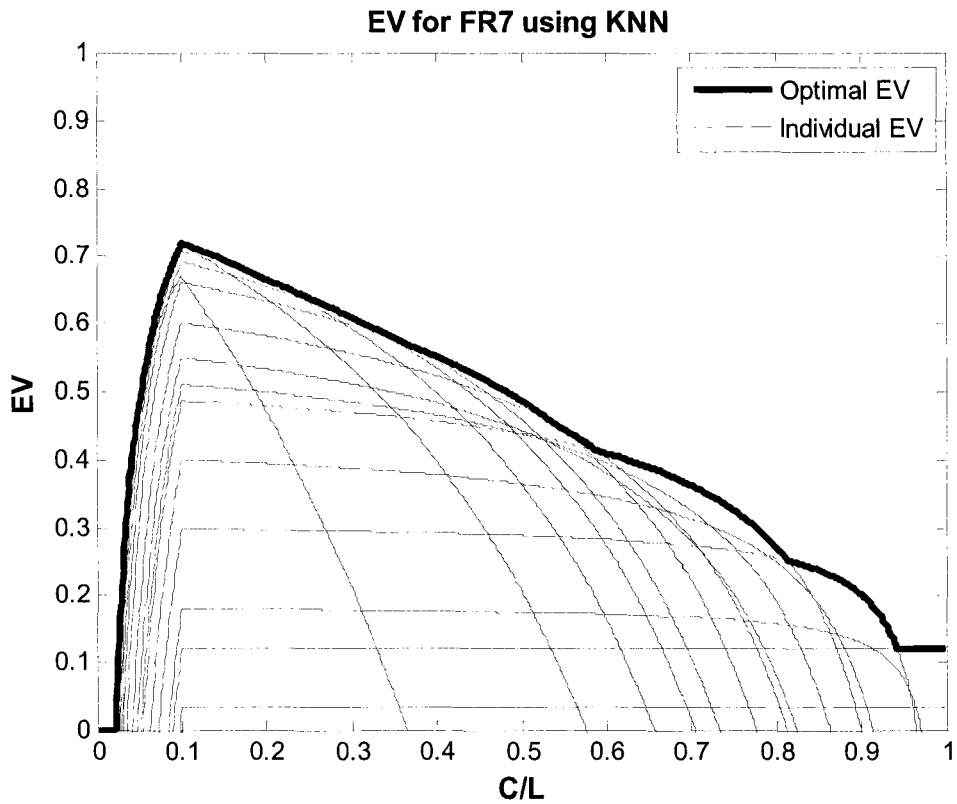


Figure 23. Potential EVs of flow forecasts using KNN, in excess of the 90th percentile of the observed flows for a FR of 7 days, for the test period Jan 97 to Dec 01. The thin curves represent EV for various probability thresholds *pt*, for each of the ensemble members, and the envelope of these curves (heavy solid line) represents the overall EV of the probabilistic forecast system, obtained by picking the optimal value of *pt*, corresponding to each cost-loss ratio.

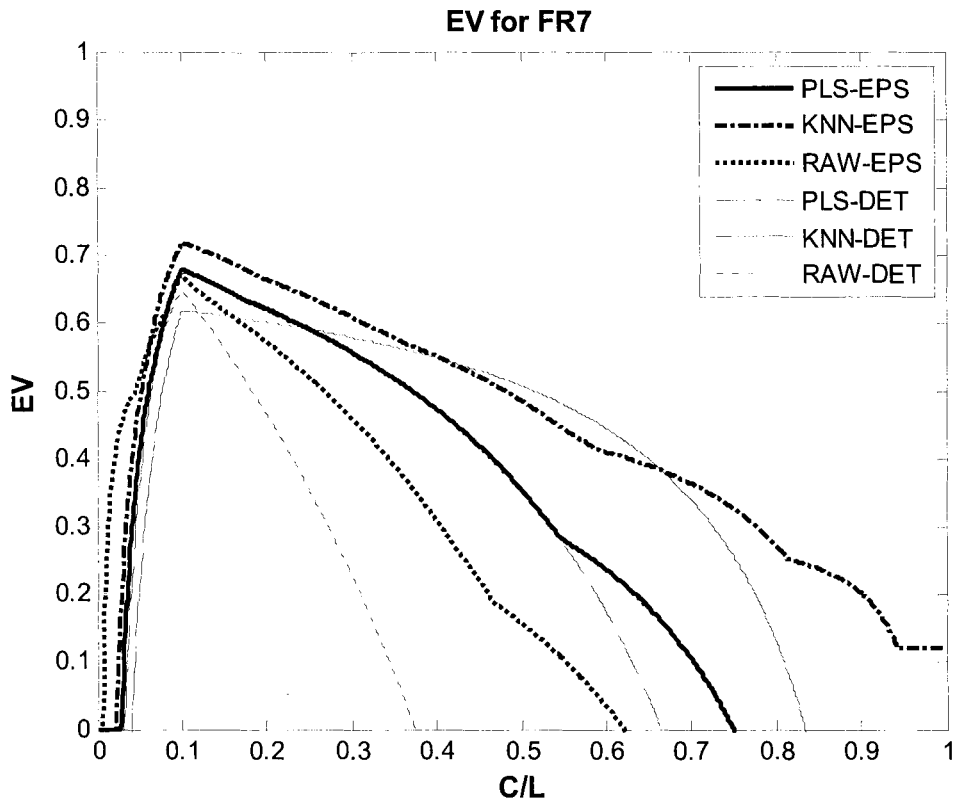


Figure 24. Potential EVs of the various flow forecast systems, in excess of the 90th percentile of the observed flows for a FR of 7 days, for the test period Jan 97 to Dec 01. Deterministic (DET) forecasts: PLS (dashed line), KNN (solid line), and RAW (pecked line); Ensemble Prediction System (EPS) probabilistic forecasts: PLS (heavy solid line), KNN (heavy dash-dotted line), RAW (heavy dotted line).

Table 1. Performance statistics in downscaling daily precipitation and temperature. Bold statistics represent the best performance corresponding to each forecast range (FR).

Downscaling using mean of ensemble predictors												
Prec												
	FR0			**FR3**			**FR7**			**FR10**		
	PLS	KNN	RAW	PLS	KNN	RAW	PLS	KNN	RAW	PLS	KNN	RAW
Bias (%)	**-7.6**	-9.5	27.2	**-6.9**	-22.1	40.3	11.4	**-9.3**	44.6	23.6	**-1.3**	50.4
RMSE	**4.5**	6.3	4.9	**5.4**	6.8	6.0	**6.0**	7.3	6.2	**6.0**	7.7	6.2
r	**0.6**	0.33	**0.6**	**0.35**	0.1	0.34	**0.09**	0.05	0.08	**0.1**	0.02	0
CE	**0.35**	-0.27	0.24	**0.09**	-0.45	-0.13	-0.13	-0.71	-0.21	-0.12	-0.86	-0.22
MeanObs	2.7	2.7	2.7	2.7	2.7	2.7	2.7	2.7	2.7	2.7	2.7	2.7
MeanSim	**2.5**	2.4	3.4	**2.5**	2.1	3.8	**3.0**	**2.4**	3.9	3.3	**2.7**	4.0
VarObs	31.6	31.6	31.6	31.6	31.6	31.6	31.6	31.6	31.6	31.6	31.6	31.6
VarSim	11.0	**28.5**	26.7	8.7	18.6	**19.9**	6.9	**25.1**	7.9	6.3	**28.7**	5.1
KS	**0.49**	0.26	0.35	**0.3**	0.15	0.11	**0.1**	0.02	0.01	**0.08**	-0.01	0
Temp												
Bias (%)	**-7.9**	-8.2	-27.8	**-11.3**	-12.9	-26.1	-13.0	-17.0	**-2.0**	**-10.8**	-12.5	11.5
RMSE	**2.6**	3.8	3.0	**3.5**	4.8	3.7	**4.7**	6.5	4.9	**5.1**	6.9	5.4
r	**0.98**	0.96	0.97	**0.96**	0.93	**0.96**	**0.93**	0.87	**0.93**	**0.92**	0.85	0.91
CE	**0.96**	0.91	0.94	**0.92**	0.86	0.91	**0.86**	0.74	0.85	**0.84**	0.7	0.82
MeanObs	2.8	2.8	2.8	2.8	2.8	2.8	2.8	2.8	2.8	2.8	2.8	2.8
MeanSim	**2.6**	**2.6**	2.0	**2.5**	**2.5**	2.1	2.5	2.4	**2.8**	**2.5**	**2.5**	3.2
VarObs	162	162	162	162	162	162	162	162	162	162	162	162
VarSim	153	**164**	138	150	**167**	141	136	**166**	116	135	**165**	107
Mean of 15 downscaled members												
Prec												
	FR0			**FR3**			**FR7**			**FR10**		
	PLS	KNN	RAW	PLS	KNN	RAW	PLS	KNN	RAW	PLS	KNN	RAW
Bias (%)	**-7.5**	-13.8	27.2	**-4.6**	-10.7	40.3	19.7	**-10.6**	44.6	19.5	**-6.3**	50.4
RMSE	**4.5**	5.1	4.9	**5.3**	5.5	6.0	**5.7**	**5.7**	6.2	**5.8**	**5.8**	6.2
r	**0.6**	0.46	**0.6**	0.33	0.24	**0.34**	**0.13**	0.06	0.08	**0.12**	0.04	0
CE	**0.36**	0.17	0.24	**0.1**	0.05	-0.13	-0.04	-0.04	-0.21	-0.05	-0.05	-0.22
MeanObs	2.7	2.7	2.7	2.7	2.7	2.7	2.7	2.7	2.7	2.7	2.7	2.7
MeanSim	**2.5**	2.3	3.4	**2.6**	2.4	3.8	3.2	**2.4**	3.9	3.2	**2.5**	4.0
VarObs	31.6	31.6	31.6	31.6	31.6	31.6	31.6	31.6	31.6	31.6	31.6	31.6
VarSim	10.2	12.9	**26.7**	5.6	3.9	**19.9**	3.9	2.2	**7.9**	4.0	2.3	**5.1**
KS	**0.48**	0.21	0.35	**0.2**	0.04	0.11	**0.04**	0.01	0.01	**0.01**	-0.01	0
Temp												
Bias (%)	**-7.9**	-11.7	-27.8	**-11.4**	-13.6	-26.1	-15.8	-17.3	**-2.0**	**-13.6**	-16.5	11.5
RMSE	**2.6**	3.0	3.0	3.5	**3.4**	3.7	4.8	**4.6**	4.9	5.3	**5.0**	5.4
r	**0.98**	0.97	0.97	**0.96**	**0.96**	**0.96**	**0.93**	**0.93**	**0.93**	0.91	**0.92**	0.91
CE	**0.96**	0.94	0.94	0.92	**0.93**	0.91	0.86	**0.87**	0.85	0.83	**0.84**	0.82
MeanObs	2.8	2.8	2.8	2.8	2.8	2.8	2.8	2.8	2.8	2.8	2.8	2.8
MeanSim	**2.6**	2.5	2.0	**2.5**	2.4	2.1	2.4	2.3	**2.8**	**2.4**	**2.4**	**3.2**
VarObs	162	162	162	162	162	162	162	162	162	162	162	162
VarSim	153	**159**	138	143	**152**	141	117	**142**	116	108	**141**	107

Table 2. Test for position (i.e. median) and variance for downscaling daily precipitation associated with the mean of ensemble predictors using Mann-Whitney and Levene, respectively, for the test period Jan 97 to Dec 01. Note that the values in the table represent *p*-values. The downscaled output derived from predictors of ensemble mean dataset was used to conduct statistical tests.

Downscaling model	Forecast Range							
	Test for position (Mann-Whitney)				Test for Variance (Levene)			
	0	3	7	10	0	3	7	10
PLS	0.299	0.015	0.000	0.000	0.174	0.201	0.018	0.000
KNN	0.203	0.195	0.791	0.302	0.157	0.000	0.112	0.685
RAW	0.016	0.081	0.000	0.000	0.015	0.000	0.000	0.000

CHAPTER 7

General conclusions and recommendations

The material in this chapter is structured in two sections. The first section outlines the main conclusions of the research described in the thesis; and the second section presents some recommendations for future work on hydrologic prediction.

7.1. Conclusions

The main conclusions emerging from the research include:

- The incorporation of the low-frequency climatic indices information in addition to the reservoir inflow information can significantly improve the seasonal hydrologic forecast skill;
- More accurate spring flow forecasts can be obtained using BNN, TLFN and RMLP as early as the beginning of January preceding the runoff year;
- The extended Kalman filter method is able to provide a robust modeling framework capable of capturing complex dynamics of the hydrologic system, and can yield more accurate long-range streamflow forecasts. This method is particularly effective in characterizing rare events such as peaks;
- There are no specific models which will always be superior in downscaling daily precipitation. The PLS, PLS-Logst, ANN-Logst, MLP and RMLP models have modest skill, and the SDSM and KNN models have the potential to capture the variability in daily precipitation;
- The nearest neighbor models which use the Mahalanobis distance as a measure of closeness outperform the models which use the Euclidean distance;
- The downscaled precipitation has greater skill, compared to the raw numerical model output;
- The hydrologic model conditioned on the statistically downscaled temperature and precipitation forecasts has greater skill, and yields more accurate flow forecasts than the hydrologic model conditioned on the raw numerical model output;

- The probabilistic flow forecasts possess greater economic benefits when compared to the deterministic flow forecasts in terms of the overall economic value and the range of end-users that can benefit from the respective forecast systems; and
- The downscaled-based flow forecasts offer greater benefits, which are applicable to a much wider range of users, than the raw-based flow forecasts. The added value which can be incurred as a result of adequate downscaling, in lieu of using the raw model output, is significant.

7.2. Recommendations for future research

There is a growing interest among the research community in fostering the accuracy of medium- and long-range hydrological forecasts. In the last two decades alone, research on hydrologic modeling has seen significant advances in computational methods. The increased computational resources coupled with the emergence of novel and efficient algorithms greatly further effective description of the physical systems governing the hydrologic processes. Nevertheless, accurate characterization of the various hydrologic systems is still in its infancy and is far from perfect. This section briefly discusses a few of the several avenues which can potentially improve the forecast accuracy in the context of hydrologic modeling.

1. Multi-model super-ensemble technique: hydrologic forecast uncertainties may arise from uncertainty in model inputs, model structure, model parameters, and observed discharge. A hydrologic forecast system which produces a single flow estimate could not incorporate the various uncertainties associated with hydrologic forecasting system. Ensemble hydrologic forecast systems, which make use of deterministic hydrologic models, may capture some of the input uncertainties, but not a significant portion. Conversely, super-ensemble forecast systems, which utilize multi-model approaches, could effectively describe most of the hydrologic forecast uncertainties. For example, in the case of long-range hydrologic forecasting, several hydrologic models having different strengths can be forced with a range of pertinent and leading climate information and measures of atmospheric circulation, such as climatic indices, gridded NCEP reanalysis,

NCEP/NCAR reanalysis, etc. The resulting flows obtained can then be used with the appropriate degree of confidence, particularly for long-term water resources planning and management including drought. Likewise, in the case of medium-range hydrologic forecasting, traces of ensemble flows can be generated from a number of carefully identified and optimally combined hydrologic models in order to construct super-ensemble flows. The resulting flows from such forecast systems could, in general, be amenable and suited to risk based decision-making processes, particularly for optimal hydropower operation and scheduling.

2. Developing innovative downscaling models: the spatial resolution of the current reforecast system is too coarse to adequately represent important sub-grid scale features such as clouds and topography. Reforecasts are large-scale model outputs generated by imperfect numerical models. Statistical downscaling techniques can be used if meaningful association exists between local-scale predictands and large-scale predictors. Therefore, improvements in the resolution of the current reforecast system may offer great opportunity for better characterization of local-scale hydrological variables. Furthermore, it has been found that direct use of the current numerical model output for hydrologic applications are shown to yield modest skill and in some cases lacks skill. In order to enhance the skill of such model output, downscaled estimates have been used. But several studies have shown that even dynamical models, which are built on higher levels of sophistication and physical realism, are ineffective in representing important features of precipitation. The failure of the current downscaling models to adequately describe daily precipitation, poses further challenges, signifying a call for novel and promising techniques.

3. Model-data assimilation: model-data assimilation, a recent and emerging state-of-the-art technique, is becoming more popular in hydrologic applications. The method provides an effective mechanism by combining observations with physical models to advance the skill of hydrologic forecasts through estimation of optimal model state variables. The most up-to-date data assimilation technique coupled with adequately downscaled numerical weather model output can yield more accurate and reliable

hydrologic forecasts that are useful for risk analysis by policy makers for operating both large-scale as well as small-scale water resources systems.