Improving Hydrometeorologic Numerical Weather Prediction Forecast Value via Bias Correction and Ensemble Analysis

by

Douglas M^{c}Collor

B.Sc., The University of British Columbia, 1979
M.Sc., The University of British Columbia, 1982
Dip. Met., Dalhousie University, 1985

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

in

The Faculty of Graduate Studies

(Atmospheric Science)

The University Of British Columbia

July, 2008

Abstract

This dissertation describes research designed to enhance hydrometeorological forecasts. The objective of the research is to deliver an optimal methodology to produce reliable, skillful and economically valuable probabilistic temperature and precipitation forecasts.

Weather plays a dominant role for energy companies relying on forecasts of watershed precipitation and temperature to drive reservoir models, and forecasts of temperatures to meet energy demand requirements. Extraordinary precipitation events and temperature extremes involve consequential water- and power-management decisions.

This research compared weighted-average, recursive, and model output statistics bias-correction methods and determined optimal window-length to calibrate temperature and precipitation forecasts. The research evaluated seven different methods for daily maximum and minimum temperature forecasts, and three different methods for daily quantitative precipitation forecasts, within a region of complex terrain in southwestern British Columbia, Canada.

This research then examined ensemble prediction system design by assessing a three-model suite of multi-resolution limited area mesoscale models. The research employed two different economic models to investigate the ensemble design that produced the highest-quality, most valuable forecasts.

The best post-processing methods for temperature forecasts included moving-weighted average methods and a Kalman filter method. The optimal window-length proved to be 14 days. The best post-processing methods for achieving mass balance in quantitative precipitation forecasts were a moving-average method and the best easy systematic estimator method. The optimal window-length for moving-average quantitative precipitation forecasts was 40 days. The best ensemble configuration incorporated all resolution members from all three models.

A cost/loss model adapted specifically for the hydro-electric energy sector indicated that operators managing rainfall-dominated, high-head reservoirs should lower their reservoir with relatively low probabilities of forecast precipitation. A reservoir-operation model based on decision theory and variable energy pricing showed that applying an ensemble-average or full-ensemble precipitation forecast provided a much greater profit than using only a single deterministic high-resolution forecast.

Finally, a bias-corrected super-ensemble prediction system was designed to produce probabilistic temperature forecasts for ten cities in western North America. The system exhibited skill and value nine days into the future when using the ensemble average, and 12 days into the future when employing the full ensemble forecast.

Table of Contents

Abstract . . . ii

Table of Contents . . . iv

List of Tables . . . vii

List of Figures . . . viii

Acknowledgements . . . xii

Dedication . . . xiv

Statement of Co-Authorship . . . xv

1 Background . . . 1
- 1.1 Forecast Evaluation . . . 4
- 1.2 Statistical Post-processing . . . 7
- 1.3 Ensemble Forecasting . . . 10
  - 1.3.1 Historical Perspective . . . 10
  - 1.3.2 Recent Developments . . . 11
- 1.4 Economic Value of Forecasts . . . 16
- 1.5 Summary . . . 17

Bibliography . . . 20

2 Hydrometeorological Accuracy Enhancement via Post-processing . . . 31
- 2.1 Introduction . . . 31
  - 2.1.1 Background . . . 32
  - 2.1.2 Review . . . 33
- 2.2 Data . . . 35

2.2.1 Temperature 35

2.2.2 Precipitation 37

2.3 Results 39

2.3.1 Daily Temperature 39

2.3.2 Daily Precipitation 42

2.4 Summary and Conclusions 45

Bibliography 65

3 Short-Range Ensemble Forecasts. Part I: Meteorological Evaluation . 68

3.1 Introduction 68

3.2 Methodology 69

3.2.1 Location and Observation Data 69

3.2.2 Numerical Models and Study Dates 71

3.3 Verification Procedures 74

3.4 Results 75

3.4.1 Error Performance of Individual Ensemble Members 75

3.4.2 Brier Skill Score: Resolution and Reliability 78

3.4.3 ROC Curves 79

3.4.4 Equal Likelihood 80

3.5 Conclusions 80

Bibliography 103

4 Short-Range Ensemble Forecasts. Part II: Economic Evaluation 107

4.1 Introduction 107

4.2 Forecast Evaluation via Economic Models 108

4.2.1 Background 108

4.2.2 Cost/Loss Model 108

4.2.3 Decision Theory Model 111

4.3 Results 113

4.3.1 Cost/Loss Model 113

4.3.2 Decision Theory Model 116

4.4 Case Studies 116

4.4.1 Cost/Loss Model 116

4.4.2 Decision Theory Model 120

4.5 Conclusion . . . 121

Bibliography . . . 141

5 Medium-Range Ensemble Temperature Forecasts . . . 144
5.1 Introduction . . . 144
5.2 Methodology . . . 145
5.3 Results and Discussion . . . 148
5.3.1 Skill Scores . . . 149
5.3.2 Continuous Ranked Probability Score . . . 149
5.3.3 ROC Area . . . 151
5.3.4 Economic Value . . . 152
5.3.5 Equal Likelihood . . . 153
5.4 Summary and Conclusions . . . 155

Bibliography . . . 172

6 Conclusions . . . 176
6.1 Summary of Methods and Procedures . . . 176
6.2 Summary of Findings . . . 178
6.3 Discussion and Recommendations . . . 179

Bibliography . . . 182

Appendices

A Kalman Filter Bias Correction . . . 183

Bibliography . . . 185

B Statistical Significance Hypothesis Test . . . 186

Bibliography . . . 188

C Equations for Meteorological Statistical Analysis . . . 189

Bibliography . . . 199

List of Tables

1.1 Weather factors that influence hydro-electric planning operations. 19

2.1 Table of post-processing methods and descriptions for temperature and precipitation forecasts. 64

3.1 List of ensemble members for all four different short-range ensemble forecast system configurations. 99

3.2 Values of Relative Operating Characteristic (ROC) area skill scores for all four different short-range ensemble forecast system configurations for forecast days one and two. 100

3.3 Values of flatness δ for all four short-range ensemble forecast system configurations for forecast days one and two. 101

3.4 Values of Reliability Index RI for all four SREF configurations for forecast days one and two. 102

4.1 Cost/loss contingency table of precipitation forecasts and observations. . . . 137

4.2 Reservoir physical parameters for cost/loss calculations. 138

4.3 Cost/loss ratios for the reservoirs in this study. 139

4.4 Cost/loss contingency table summary of actual precipitation forecasts and observations for Jordan River for the 50 mm day^{-1}precipitation threshold. . 140

5.1 Ten cities for which forecast and associated verifying observations were evaluated. 171

C.1 Contingency table for hit rate and false alarm rate calculations. 196

List of Figures

2.1 The area of study in southwestern British Columbia, Canada. 47

2.2 Mean absolute error (°C) vs. window length in days for the linear method of post-processing systematic error correction for daily maximum temperature. 48

2.3 The 24 hour precipitation forecast degree of mass balance as a function of averaging window length for the moving average post-processing method. . 49

2.4 The 24 hour precipitation forecast mean absolute error as a function of averaging window length for the moving average post-processing method. . . . 50

2.5 The daily maximum and minimum temperature mean error from the sample for forecast days one through eight. 51

2.6 The same as Fig. 2.5 but for mean absolute error. 52

2.7 The same as Fig. 2.6 but for mean absolute error skill score (relative to direct model output). 53

2.8 The daily maximum and minimum temperature absolute errors greater than 5°C (in percent) from the sample for forecast days one through eight. . . . 54

2.9 The same as Fig. 2.8 but for temperature absolute errors greater than 10°C (in percent). 55

2.10 The daily maximum and minimum temperature mean error from the UMOS sub-sample for forecast days one and two. 56

2.11 The same as Fig. 2.10 but for temperature mean absolute error. 57

2.12 The daily 24 hour precipitation degree of mass balance from the sample for forecast days one through eight. 58

2.13 A comparison of 24 hour precipitation degree of mass balance values for each of the 19 stations for the day two forecast. 59

2.14 The same as Fig. 2.12 but for precipitation mean absolute error. 60

2.15 The same as Fig. 2.14 but for precipitation mean absolute error skill score (relative to direct model output). 61

2.16 The daily 24 hour precipitation errors greater than 10mm day^{-1} (in percent) from the sample for forecast days one through eight. 62

2.17 The same as Fig. 2.16 but for precipitation errors greater than 25mm day^{-1} (in percent). 63

3.1 Reference map for watersheds in the area of study in southwestern British Columbia, Canada. 83

3.2 Summary hypsometric curve and cumulative total precipitation curve for the stations and watersheds in this study 84

3.3 Box and whisker plots for 11 members of the ensemble for the day-one forecast of 24-hour precipitation amount. The ensemble average and associated observations are also included. 85

3.4 Same as Fig. 3.3 but for the day-two forecast. 86

3.5 The LEPS (linear error in probability space) skill score for ensemble members for the day-one forecasts. 87

3.6 The LEPS skill score for ensemble members for the day-two forecasts. . . . 88

3.7 Error characteristics of individual ensemble members for the full 11-member ensemble for day-one and day-two forecasts. 89

3.8 Brier skill score, relative reliability and relative resolution for the short-range ensemble forecasts for forecast days one and two. The precipitation threshold is 5 mm day^{-1}. 90

3.9 Same as Fig. 3.8 but for the 10 mm day^{-1} precipitation threshold. 91

3.10 Same as Fig. 3.8 but for the 25 mm day^{-1} precipitation threshold. 92

3.11 Same as Fig. 3.8 but for the 50 mm day^{-1} precipitation threshold. 93

3.12 Reliability diagrams from the full 11-member ensemble for the day-one forecasts for precipitation thresholds of 5 mm day^{-1}, 10 mm day^{-1}, 25 mm day^{-1}, and 50 mm day^{-1}. Inset diagrams depict sharpness. 94

3.13 Same as Fig. 3.12 but for the day-two forecasts. 95

3.14 Relative Operating Characteristic (ROC) curves for the full 11-member and high resolution 8-member ensemble prediction systems for precipitation thresholds 5, 10, 25, and 50 mm day^{-1}. The area under the ROC curves provides a measure of discrimination skill. 96

3.15 Same as Fig. 3.14 but for the medium resolution 6-member and very high resolution 5-member ensemble prediction systems. 97

3.16 Rank histogram diagrams for day-one and day-two forecasts for the full 11-member short-range ensemble forecast system configuration. 98

4.1 Photos of four BC Hydro dams in southwestern British Columbia, Canada. 124

4.2 Daily precipitation vs. daily inflow for Jordan River on southwest Vancouver Island. 125

4.3 Forecast value as a function of cost/loss ratio (C/L) for day-one forecasts. . 126

4.4 Forecast value as a function of cost/loss ratio (C/L) for day-two forecasts. . 127

4.5 Forecast value envelope as a function of C/L for day-one forecasts. 128

4.6 Forecast value envelope as a function of C/L for day-two forecasts. 129

4.7 Forecast value as a function of probability threshold for day-one forecasts. . 130

4.8 Forecast value as a function of probability threshold for day-two forecasts. . 131

4.9 Each bar in the figure spans the portion of the C/L range having positive value for different short-range ensemble forecast system configurations and for different precipitation thresholds for the day-one forecast. 132

4.10 Same as Fig. 4.9 except for the day-two forecast. 133

4.11 The relationship between relative operating characteristic (ROC) and forecast value. 134

4.12 The reservoir schematic diagram for the cost/loss economic model. 135

4.13 Profit as a function of the market/contract price ratio for different forecast configurations. 136

5.1 Mean error for ensemble average DMO forecasts and ensemble average post-processed forecasts for daily maximum and daily minimum temperatures for forecast days 1 through 15. 158

5.2 RMSE and correlation skill scores for daily maximum and daily minimum temperature forecasts for forecast days 1 through 15. The ensemble average forecast is compared with a deterministic operational forecast. 159

5.3 Continuous ranked probability skill score for ensemble daily maximum and daily minimum temperature anomaly forecasts. The CRP skill score equals the relative resolution less the relative reliability. 160

5.4 ROC area skill score for ensemble daily maximum temperature forecasts with a +5°C anomaly threshold and daily minimum temperature with a -5°C anomaly threshold. 161

5.5 ROC Area plots for a subset of the forecast period for daily maximum temperature anomaly +5°C threshold. 162

5.6 Forecast value as a function of cost/loss ratio (C/L) for ensemble daily maximum forecasts (+5°C anomaly threshold), comparing the value of the full ensemble to the ensemble mean. 163

5.7 Cost/loss ratio (C/L) maximum value for ensemble daily maximum temperatures with a +5°C anomaly threshold and minimum temperatures with a -5°C anomaly threshold. 164

5.8 Value as a function of probability threshold p_t for cost/loss ratio C/L=0.2, C/L=0.4, C/L=0.6 and C/L=0.8. The forecasts are daily maximum temperature forecasts with a +5°C anomaly threshold. 165

5.9 Rank histograms for the ensemble of daily maximum temperatures. 166

5.10 Rank histograms for the ensemble of daily minimum temperatures. 167

5.11 Rank histogram δ score for ensemble daily maximum and daily minimum temperature forecasts. 168

5.12 A trade-off diagram of CRPSS vs. spread score ($1/\delta$) for daily maximum temperature and daily minimum temperature forecast anomalies. 169

5.13 Similar to Fig. 5.12 but for ROC$_{SS}$ vs. spread score ($1/\delta$). 170

C.1 Left: ROC values transformed to standard normal deviates of hit rate H and false alarm rate F. Right: ROC curves transformed back to probability axes. Upper plots: day-one full 11-member ensemble forecasts with the 5 mm day^{-1} precipitation threshold. Lower plots: the same ensemble but for the rarer 50 mm day^{-1} precipitation threshold. 197

C.2 Same as fig. C.1 except for the highest resolution ensemble vhires5. 198

Acknowledgements

I would like to thank my supervisor, Dr. Roland Stull, for his continuous and enthusiastic support for this project. I would also like to thank my supervisory committee members, Drs. Josh Hacker and Ziad Shawwash, who shared their knowledge and expertise with me and provided invaluable advice.

Additionally, I would like to thank Dr. Josh Hacker, from the NCAR/UCAR Research Applications Laboratory, for particularly valuable comments and suggestions to the manuscript in Chapter 2.

I would like to thank BC Hydro corporation and the Meteorological Service of Canada for providing the temperature and precipitation observations necessary to perform the analyses presented in Chapters 3 and 4. In addition, the Geophysical Disaster Computational Fluid Dynamics Centre (GDCFDC), located in the Department of Earth and Ocean Sciences at the University of British Columbia (UBC), is acknowledged for providing the numerical limited-area-model forecasts used in the short-range ensemble portion of this study. In particular I would like to thank Mr. George Hicks II for his efforts in making the numerical forecast information available for my research. The GDCFDC gratefully acknowledges the following organizations for computer-purchase grants: Canadian Foundation for Innovation, British Columbia Knowledge Development Fund, and UBC, plus additional grant support provided by the Canadian Natural Sciences and Engineering Research Council, and Environment Canada. At UBC the WRF model was run operationally by Henryk Modzelewski, the MM5 by Yongmei Zhou and George Hicks II, and the MC2 by Xingxiu Deng, Yan Shen, and Atoossa Bakhshaii.

I would like to thank the Meteorological Service of Canada and the U.S. National Weather Service for making available the temperature observations necessary to perform the analysis presented in Chapter 5. I would also like to thank the national weather agencies of Canada (the Meteorological Service of Canada), the United States (the National Weather Service), and Mexico (the National Meteorological Service of Mexico) for providing public access to forecasts from the North American Ensemble Forecast System. I would like to thank Uwe Gramman of Mountain Weather Services of Smithers, B.C. for writing the scripts to access the NAEFS forecasts. Additional support was provided by the Canadian Natural Science and Engineering Research Council.

I would like to thank all the anonymous referees who reviewed my work prior to publication (Chapters 2, 3, and 4), providing invaluable advice and suggestions for improvement.

BC Hydro is gratefully acknowledged for providing salary support throughout this project. I would like to thank the following BC Hydro executives and managers for their support of this research project and my dissertation: Ms. Dawn Farrell, Mr. Brian Fast, Ms. Renata Kurschner, Mr. Cam Matheson, Mr. Doug Smith, and Mr. Eric Weiss. I would like to thank my co-workers at BC Hydro who supported me with advice, aid, and encouragement: Dr. Dave Bonser, Dr. Garth Nash, Ms. Stephanie Smith, Dr. Tom Siu, and Mr. Frank Weber.

I would like to thank the following students and staff of the GDCFDC Weather Forecast Research Team, recent and current, for their assistance, support, and friendship: Ms. Atoossa Bakhshaii, Dr. Luca Delle Monache, Mr. George Hicks II, Ms. Rosie Howard, Dr. Henryk Modzelewski, Mr. Thomas Nipen, Dr. John Spagnol, and Mr. Bruce Thomson.

Dedicated to my wife Vivian, and to my children Alex, Michael, and Maylee.

Statement of Co-Authorship

This thesis is based on several manuscripts, resulting from collaboration between two researchers.

The author's specific contributions in these collaborations are detailed in the following paragraph.

The research program was designed jointly based on ongoing discussion between the author and Roland Stull. All major research, including detailed problem specification, model design, performance of analysis and identification of results was performed by the author, with assistance from Roland Stull. All numerical examples and data analyses were performed by the author. The author prepared all manuscripts. Roland Stull edited and helped revise all manuscripts.

Other contributions to this work are detailed below.

The concept of skill vs. spread trade-off diagram (Section 5.3.5) originated with Roland Stull.

The maps of the study areas included in this dissertation, Figs. 2.1 and 3.1, were prepared by Frank Weber of BC Hydro Corporation.

Numerical forecasts analyzed in Chapters 3 and 4 were obtained from weather models run operationally by the following individuals: Xingxiu Deng, Yan Shen, and Atoossa Bakhshaii (Mesoscale Compressible Community model); Yongmei Zhou and George Hicks II (Penn State/NCAR Mesoscale model); and Henryk Modzelewski (Weather Research and Forecast model).

Chapter 1

Background

Energy utility companies must efficiently balance production from limited energy sources with increasing demand from industrial, business, and residential consumers. Hydro-electric companies rely on forecasts of watershed precipitation and temperature to drive hydrologic reservoir inflow models, and forecasts of urban temperatures to anticipate energy demand (Dempsey et al., 1998; Stensrud, 2006). Weather forecasts affect three types of hydro-electric planning— operations, financial, and social/environmental:

1. Operations includes managing water levels in reservoirs, changing the water flow through turbines to generate energy, and controlling river flow downstream of dams. Operations planners rely on weather forecasts for extreme high-inflow (flood) events, for daily operations planning, and for long-term water resource management. Table 1.1 shows how weather factors on various timescales influence the decision making process.

2. Financial planners consider spot market prices for selling and buying power, and for short- and long-term futures markets. On a daily basis, significant precipitation events or temperature extremes involve potential profit or loss decisions in the tens to hundreds of thousands of dollars worth of power generation. These economic factors add to environmental and societal benefits and costs that must be considered as part of a complete decision-making structure.

3. Environmental and social aspects of planning and decision-making may be voluntary on behalf of a power-generation company; they may be in the best economic interests of a company in a competitive energy marketplace where consumers want green energy; or they may be regulated by water-use licenses agreed upon by multiple stakeholders. There are many competing interests constraining a simple maxim of generating power to meet load. For example, government environmental agencies tasked with supporting multi-species fish stocks require reservoir levels and downstream flows that sustain fish and riparian habitat. Municipalities oversee jurisdictional authority

> to ensure public safety during high-flow flooding events. Private businesses that operate docks, marinas, and recreational facilities rely on stable reservoir water levels to sustain their economic viability. Therefore, improvement in hydrometeorological forecasts provides a cascading arena of benefits to companies, government agencies, stakeholders, and the general public.

This dissertation is motivated by the requirement by water resource managers for improved weather forecasts. The approach is to investigate the ability of recent numerical weather prediction (NWP) computer models to provide site-specific detailed forecasts in the one to fifteen day forecast range.

The research will have potential applications for all energy utilities. A recent survey commissioned by the Canadian Electrical Association found that 70% of responding electric utility companies in Canada, the United States, and other worldwide locations did not incorporate NWP model output for hydrologic guidance (Morassutti, 2004). The remaining 30% utilized NWP output but without post-processing (bias correcting) the data prior to inputting into their respective hydrologic models. The report states that:

> "[P]oor hydrologic forecasts are largely the result of deficient forecasts of precipitation and temperature. It would also appear that there is a fair measure of subjective judgement involved when making modifications to the NWP model output before implementation into hydrological models."

The report also states that:

> "The reason why utility companies have not developed more intricate techniques for handling and processing NWP model output for hydrological prediction lies in the fact that the private sector generally lags five to ten years behind methodological developments achieved within the scientific research community."

In addition to published reports, recent workshops and conferences espouse the need for improved hydrometeorological information. A series of workshops organized by HEPEX[1] (the latest in June 2007 in Stresa, Italy) have brought the international hydrological community together with the meteorological community. The objective of the workshops was to develop a method to produce reliable hydrological ensemble forecasts that can be used with

[1] Hydrological Ensemble Prediction EXperiment. HEPEX is an international collaboration of scientists formed specifically to address downscaling and ensemble forecasting issues for the hydrometeorological community.

confidence to assist water resource managers. Water resource managers frequently confront decisions of significant consequence for public health and safety and for the economy.

HEPEX is working toward a state-of-the-art end-to-end hydrological ensemble streamflow prediction system (Schaake et al., 2007). The ultimate goal ensures an automated, skillful, reliable ensemble streamflow forecast product that explicitly accounts for the major sources of uncertainty in the forecasting process. The initial stage in the process classifies an ensemble of atmospheric forecasts forcing a hydrologic model with downscaled, bias-corrected meteorological forecasts of specific hydrometeorological variables.

Krzysztofowicz (2001) targeted future precipitation as the prevalent source of uncertainty in hydrologic forecasting. Krzysztofowicz et al. (1993) acknowledged that precipitation constitutes the principal input into a hydrologic model, and Hoinka and Davies (2007) stated that an accurate quantitative precipitation forecast (QPF) is currently one of the most critical operational forecasting challenges. Droegemeier et al. (2000) claimed that accurate forecasts of precipitation amounts at fine spatial and temporal resolution are a critical input for hydrometeorological flood and river flow forecasting models. As short-range precipitation forecasts generated by mesoscale ensemble forecast systems continue to improve, Yuan et al. (2005) agreed that it is becoming feasible to consider their use to drive hydrologic models for general flood and streamflow forecasting. Yuan et al. (2005) summarized their recent findings by stating:

> "[W]e believe the results of this paper support the notion that the time is ripe to pursue an accelerated development of *ensemble* hydrometeorological prediction systems. Before the goal of a skillful coupled ensemble forecast runoff system can be routinely realized, however, comprehensive calibration studies and multivariate calibration efforts must be performed for atmospheric variables that historically have not been scrutinized. Such studies must be extended to longer forecast projections."

A widespread need exists to research probabilistic downscaled forecasts specifically for hydrometeorologic applications. This dissertation addresses this need by merging the advancements forged in the meteorological research community with the requirements of professional hydrologists to improve water resource management capabilities.

1.1 Forecast Evaluation

The first question that a water resource manager asks when given a weather forecast is: "How good are these forecasts?" Before attempting to answer this question, a suite of defined terms must be agreed upon. Murphy (1993) distinguished three types of "goodness":

> *Consistency* - the degree to which the forecast corresponds to the forecaster's best judgement about the situation, based upon the forecaster's knowledge base.
>
> *Quality* - the degree to which the forecast corresponds to what actually happened.
>
> *Value* - the degree to which the forecast helps a decision-maker realize some incremental economic, environmental, or societal benefit.

Looking more closely at forecast quality, Murphy (ibid.) describes nine attributes that assess the quality of a forecast. These are:

> Bias - the correspondence between the sample mean forecast and the sample mean observation.
>
> Association - the strength of the linear relationship between the forecasts and observations (e.g., the correlation coefficient measures this linear relationship).
>
> Accuracy - the level of agreement between the forecast and the truth (as represented by observations). The difference between the forecast and the observation is the error. The lower the errors, the greater the accuracy.
>
> Skill - the relative accuracy of the forecast over some reference forecast. The reference forecast is generally an unskilled forecast such as random chance, persistence (defined as the most recent set of observations, "persistence" implies no change in condition), or climatology. Skill refers to the increase in accuracy due purely to the capability of the forecast system. Weather forecasts may be more accurate simply because the weather is easier to forecast – skill takes this into account.
>
> Reliability - the average agreement between the forecast values and the observed values. If all forecasts are considered together, then the overall reliability is the same as the bias. If the forecasts are stratified into different ranges or categories, then the reliability is the same as the conditional bias, i.e., it has a different value for each category.

> Resolution - the ability of the forecast to sort or resolve the set of events into subsets with different frequency distributions. This means that the distribution of outcomes when “A” is forecast is different from the distribution of outcomes when “B” is forecast. Even if the forecasts are wrong, the forecast system has resolution if it can successfully separate one type of outcome from another.
>
> Sharpness - the tendency of the forecast to predict extreme values. To use a counterexample, a forecast of “climatology” has no sharpness. Sharpness is a property of the forecasts only, and, like resolution, a forecast can have this attribute even if it’s wrong (in this case it would have poor reliability).
>
> Discrimination - ability of the forecast to discriminate among observations, that is, to have a higher prediction frequency for an outcome whenever that outcome occurs.
>
> Uncertainty - the variability of the observations. The greater the uncertainty, the more difficult the forecast will tend to be. This definition of uncertainty is dependent on the observations only and is not related to predictive uncertainty.

While statistical post-processing can enhance reliability, the same does not apply to resolution and discrimination. Therefore measures of resolution and discrimination assess more directly the inherent quality of a forecasting system.

Traditionally, forecast verification has emphasized accuracy and skill. It is important to note that the other attributes of forecast performance also have a strong influence on the value of the forecast. Forecast value is not the same as forecast quality. A forecast has high quality if it predicts the observed conditions well according to some objective or subjective criteria. It has value if it helps the user to make a better decision.

The “truth” data that we use to verify forecasts generally come from observational data. Typically rain gauge measurements, temperature observations, wind speed and direction, or any other measurable meteorologic quantities constitute an observational dataset. In many cases it is difficult to know the exact truth because there are errors in the observations. Sources of uncertainty include random and systematic errors in the measurements themselves, sampling error, instrument calibration error, and analysis error if the observational data are altered or smoothed to match the scale of the forecast. Most verification studies neglect the errors in observations (or analysis fields), the unwritten argument in each study assuming that observation errors are small compared to errors in the forecasts and therefore have a negligible effect on the results. Saetra et al. (2004) claimed that, in the short-range, where forecast errors are still small, this argument may not be justified.

A single verification metric is inadequate for evaluating all of the above desired attributes of an ensemble forecasting system. Also, in contrast to a single deterministic forecast, it is impossible to objectively assess the quality of a single probabilistic forecast; ensemble systems must be verified over many cases. Therefore the research presented here employed the following suite of verification measures, proposed at a 1999 workshop on short-to-medium range ensemble forecasting (Hamill et al., 2000a), to assess the quality and value of the hydrometeorological ensemble forecasts examined in this dissertation.

The use of these verification measures was confirmed in a report from a November 2007 European Centre for Medium-Range Weather Forecasts workshop on ensemble prediction (available online at http://www.ecmwf.int/newsevents/meetings/workshops/2007/ensemble_prediction/wg3.pdf, accessed April 12, 2008). See Jolliffe and Stephenson (2003) or Wilks (2006) for a complete description of these verification metrics:

1. Probabilistic skill scores, including the Brier Score, Brier Skill Score, Brier Score decomposition (into reliability, resolution, and uncertainty terms), and the Continuous Ranked Probability Score (also decomposed into reliability, resolution, and uncertainty terms), to judge the quality of probabilistic forecasts.

2. Reliability diagrams, with a subset diagram indicating sharpness of the forecasts, to evaluate forecast reliability.

3. The Relative Operating Characteristic (ROC) curve, employed to provide a measure of resolution and discrimination. In addition, a static economic decision model (Richardson, 2000, 2003), uniquely determined from ROC curves, was employed to assess the potential economic value of the forecasts.

4. Rank histograms and trade-off diagrams, included to diagnose the ability of the ensemble to sample the spread of the forecasts in relation to the associated observations.

Even the most thorough examination of ensemble prediction systems cannot escape the tenet that all evaluation methods are contaminated by at least three sources of noise (Hamill et al., 2000a): improper estimates of probabilities from small-sized ensembles, insufficient variety and number of cases in the forecast evaluation, and imperfect observations. Saetra et al. (2004) reported the findings that rank histograms were especially sensitive to the presence of observation errors in their study, while reliability diagrams were far less sensitive in this respect.

Given these noise sources, how can we remove or minimize errors in the NWP forecasts? Systematic errors can be reduced using statistical post-processing, described next. Random errors can be reduced by combining an ensemble of different NWP forecasts, as described in section 1.3. Knowledge of the probability distribution of the remaining errors can be used to suggest the best, most economical decisions in the face of uncertainty, as described in section 1.4.

1.2 Statistical Post-processing

The operational use of NWP model output for hydrologic (and other) applications is hampered by large biases in surface sensible-weather fields (Clark et al., 2004; Eckel and Mass, 2005). Statistical post-processing the output of numerical weather forecasts can be a successful technique to reduce, minimize, correct, or eliminate bias. The bias corrections are needed to account (for example) for differences between the model and actual terrain heights, errors in the depth of the surface layer and near-surface lapse rate, the amount of entrainment from the boundary layer top, the ratio of sensible to latent heat fluxes, net radiation (Stensrud and Yussouf, 2003), and any systematic errors in NWP model initial conditions (ICs) and boundary conditions (BCs).

The earliest method of statistical post-processing that gained widespread use was the perfect prog method (Klein et al., 1959). In the perfect prog method (PPM), regression equations are derived for predictands that were not forecast as a function of NWP analyses predictors (as if the forecasts were perfect). However, forecast model bias is not taken into account in PPM, therefore, this method is usually applied to very short-term forecasts (Brunet et al., 1988; Mohanty and Dimri, 2004).

The method of Model Output Statistics, or MOS, was originally formalized in the seminal paper by Glahn and Lowry (1972), and is the gold standard of NWP model output post-processing (Kalnay, 2003). MOS routines are generally much more computer intensive than PPM routines, requiring a long, stable dataset since MOS forecast variance is both model and forecast-period dependent. Woodcock and Engel (2005) suggested that the importance of direct-model-output (DMO) will increase relative to MOS because MOS cannot easily accommodate new sites, models, and model changes; MOS often employs over a million predictive equations that require 2 to 4 years of stable developmental data for their derivation.

Direct comparison between MOS forecasts and PPM forecasts (Brunet et al., 1988; Carter et al., 1989; Hay and Clark, 2003; Wilson and Vallée, 2003; Clark and Hay, 2004; Hart et al., 2004) has shown that PPM forecasts retain their sharpness (forecast extreme events more often) over the entire range of forecast projections, while MOS forecasts conservatively tend toward climatology with increasing projection. However MOS forecasts retain their reliability (remain well calibrated) while PPM forecasts become less reliable with increasing projection.

Taylor and Leslie (2005) performed a single-station assessment of MOS temperature forecasts within the continental United States, concluding that error in the forecasts is proportional to variance in the observations, and that forecasts for any one station rarely follow a consistent temporal pattern for more than two or three consecutive days. Taylor and Leslie (2005) also demonstrated that MOS forecasts are often inaccurate in the vicinity of strong temperature gradients, for locations affected by shallow cold air masses, or for stations in regions of anomalously warm or cold temperatures.

For hydrometeorological applications that require operational forecasts for many locations and many projection times, the development and maintenance of a MOS system can prove too onerous. Even national weather centers such as the Meteorological Service of Canada (MSC) discontinued their MOS forecast system in the late 1980's because significant model changes had become too frequent and the development cost too great to maintain a statistically stable MOS system (Wilson and Vallée, 2002).

Further research has indicated updateable MOS (UMOS) can lessen the disadvantage of model variance over standard MOS routines. UMOS (Wilson and Vallée, 2002, 2003) can be implemented with briefer training than MOS, but requires an even more extensive database overhead to maintain and update. UMOS forecasts in western Canada are provided by MSC, but only at a limited set of valley-bottom airport sites. Other advancements in the use of MOS are the application of consensus MOS (CMOS) forecasts (an average of MOS over two or more models), and weighted MOS (WMOS) (Baars and Mass, 2005).

Non-linear methodologies, including neural networks (Hall et al., 1999; Koizumi, 1999; Marzban, 2003; Yuval and Hsieh, 2003; Yuan et al., 2007), logistic regression (Vislocky and Young, 1989), and generalized additive modeling (Vislocky and Fritsch, 1995) have been adapted to post-processing weather forecasts in various specific applications. Neural network applications require an extensive dataset and are computationally expensive to implement operationally on a real-time basis. Clustering (Gutierrez et al., 2004) is an analog method requiring a very large dataset. Pattern Typing (Conway et al., 1996) is a dynamic method that may require subjective intervention. An alternative theoretical approach that combines the benefit of MOS and PPM by utilizing the information in reanalysis data has been proposed by Marzban et al. (2006).

The drawbacks to MOS-based techniques have led to the exploration of other approaches to statistically post-process model forecast output that do not require long data archival periods (Stensrud and Yussouf, 2003; Cheng and Steenburgh, 2007). The overall objective of error correction is to minimize the error of the next forecast using measurement estimates of past errors. A short learning period enables an updating system to respond quickly to factors that affect error (for example model changes and changes in weather regime) but increases vulnerability to missing data and/or large errors and other limitations of small sample estimates such as missing extreme events (Woodcock and Engel, 2005).

A simple 7-day moving average error calculation is shown to improve upon raw model point forecasts (Stensrud and Skindlov, 1996). Eckel and Mass (2005) and Jones et al. (2007) successfully implemented a moving average technique as a post-processing method to reduce systematic error in DMO, using a 14-day moving average error calculation to reduce mean error in temperature forecasts. Woodcock and Engel (2005) stated a 15-30 day bias correction window effectively removes bias in DMO forecasts using the best easy systematic estimator (Wonnacott and Wonnacott, 1972).

Kalman filtering (Kalman and Bucy, 1961; Bozic, 1994; Homleid, 1995; Majewski, 1997; Kalnay, 2003; Roeger et al., 2003; Delle Monache, 2005) is a linear method that is adaptive to model changes, using the previous observation-forecast pair to calculate model error. It then predicts the model error to correct the next forecast. This recursive, adaptive method does not need an extensive, static database to be trained.

Exploring a wide array of statistical post-processing methodologies continues to be an active area of research, as documented by several very recent publications on the subject. Hacker and Rife (2007) and Gel (2007) provided analyses for estimating systematic error on (bias-correcting) mesoscale model output grids of near-surface parameters. Hansen (2007) used a fuzzy logic-based analog method to improve standard objective techniques for airport ceiling and visibility forecasts in TAFs (terminal aerodrome forecast).

In summary, statistical post-processing adds value to DMO by objectively reducing systematic error between forecasts and observations by producing site-specific forecasts. In Chapter 2 of this dissertation, several weighted-average post-processing methods, a recursive method (Kalman filtering), and an updateable MOS method are investigated for hydrometeorological applications (specifically temperature and precipitation forecasts) in complex terrain.

1.3 Ensemble Forecasting

1.3.1 Historical Perspective

John M. Lewis (2005) has published a treatise on the roots of ensemble forecasting, tracing the generation of a probabilistic view of dynamical weather prediction back to the early 1950's. Lewis (ibid.) crediteds Eric Eady with first questioning the limits of deterministic prediction (Eady, 1951) and introducing the view that a model's estimate of the atmosphere's initial state is generally erroneous and that the models are imperfect. By the end of the decade, Philip Thompson (Thompson, 1957) and Edward Lorenz explored the predictability limits of deterministic forecasting. By the early 1960's, anchored by the seminal work by Lorenz (1963), the essential nature of non-linear dynamic flows−unstable systems characterized by nonperiodicity (later referred to as chaotic systems)−placed bounds on the predictability of the system. Edward Epstein (1969) is credited with first formulating the method of stochastic-dynamic prediction which, though impractical for all but very low order models, forms the theoretical basis for subsequent ensemble forecasting. However, many of today's authors cite Leith (1974) as laying the foundation stone of operational ensemble forecasting−approximating a forecast probability density function using a finite sample of statistically-generated forecast scenarios [e.g., Du et al. (1997); Hamill and Colucci (1998); Hamill et al. (2000a); Hou et al. (2001); Wandishin et al. (2001); Buizza et al. (2005); Eckel and Mass (2005); Chien et al. (2006)] and establishing that the ensemble mean yields a forecast more skillful than the individual ensemble members, while the ensemble dispersion provides quantitative information on forecast uncertainty (Tracton and Kalnay, 1993).

Buizza et al. (2005) summarized this past half-century of work on probabilistic forecasting in the following encapsulated description:

> "The weather is a chaotic system: small errors in the initial conditions of a forecast grow rapidly and affect predictability. Furthermore, predictability is limited by model errors linked to the approximate simulation of atmospheric processes of the state-of-the-art numerical models. These two sources of uncertainty limit the skill of single, deterministic forecasts in an unpredictable way, with days of high/poor quality forecasts followed by days of poor/high quality forecasts. Ensemble prediction is a feasible way to complement a single, deterministic forecast with an estimate of the probability density function of the forecast states."

1.3.2 Recent Developments

Operational forecasting centers employ several methods to create ensemble members, such as a Monte Carlo-like approach (Palmer et al., 1990), the breeding of growing modes (Toth and Kalnay, 1993), singular vectors (Buizza and Palmer, 1995; Molteni et al., 1996), and by perturbing observations with random errors (Houtekamer et al., 1996; Hamill et al., 2000b). And, since the underlying atmospheric model remains imperfect, forecasters gain a more complete picture of uncertainty by including perturbed model physics, in addition to perturbed initial conditions, in an ensemble prediction system (Andersson et al., 1998; Houtekamer and Mitchell, 1998; Stensrud et al., 2000). For example, Wang and Seaman (1997) found that different convective parameterization schemes embedded in the same model produced different precipitation results, and no one scheme always outperformed the others. Buizza et al. (2005) assembled a thorough description and inter-comparison of perturbation methods adopted at the Canadian, European, and American global ensemble prediction centers, while Kalnay (2003) authored a rich source of information about atmospheric modeling. Readers looking for a short synopsis on the subject of current ensemble forecasting methods are referred to the 2-page article by Gneiting and Raftery (2005). The following paragraphs summarize the use of hydrometeorological ensemble forecasting techniques.

Global Ensemble Forecasting

Global ensemble forecasting has been a mainstay of operational production at both the American National Centers for Environmental Prediction [NCEP; Toth and Kalnay (1993); Tracton and Kalnay (1993); Toth et al. (1997)] and the European Centre for Medium-Range Weather Forecasts [ECMWF; Palmer et al. (1993); Molteni et al. (1996)] since December 1992. The Canadian Meteorological Centre [CMC; Houtekamer et al. (1996); Houtekamer and Lefaivre (1997)] has maintained an operational global ensemble prediction system since January 1996. Each of the centers produces the ensembles using different forecast models and different ensemble construction techniques (Hamill et al., 2000a). The Japan Meteorological Agency (Kobayashi et al., 1996), the US Navy Fleet Numerical and Oceanography Center (Rennick, 1995), and the U.K. Met Office have also joined the ranks of weather centers around the world producing global ensembles.

In the first five years of operational ensemble production, research (Toth and Kalnay, 1993; Tracton and Kalnay, 1993; Molteni et al., 1996; Du et al., 1997) focused on medium-range applications (6−10 day forecasts) and impacts on planetary-scale flow regimes, synoptic wave patterns, and cyclone positions. Analyses in these early days of operational ensemble prediction encompassed mid-level parameters such as geopotential height and temperature.

In the mid 1990's, attention turned to proposals (Mullen and Baumhefner, 1991, 1994; Brooks and Doswell, 1993; Brooks et al., 1995) that ensemble methods could also benefit short-range (1−2 day) forecasts of sensible weather elements such as surface-based temperature and precipitation. Successful experimental testing of short-range ensemble forecasts at NCEP (Hamill and Colucci, 1998; Stensrud et al., 1999; Du and Tracton, 2001; Wandishin et al., 2001) and by a larger community of scientists during the Storm and Mesoscale Ensemble Experiment (SAMEX) of 1998 (Hou et al., 2001) paved the way for other research groups and operations centers to initiate their own regional ensemble modeling systems (Grimit and Mass, 2002).

Sensible-weather-element Ensemble Forecasting with Limited-area Models

The success of global ensemble prediction systems led to researching ensemble forecasting with limited-area models (LAMs) at finer resolutions and shorter time-frames. The LAM ensemble approach proved successful at improving short-range weather forecasts, especially temperature and precipitation (Brooks et al., 1995; Stensrud et al., 1999; Hamill et al., 2000a; Hou et al., 2001; Eckel and Mass, 2005). More recently, Engel and Ebert (2007) have extended the operational consensus forecast (OCF) method of Woodcock and Engel (2005) by blending multiple statistically post-processed, weighted-average models to produce near-surface forecasts of air temperature, dewpoint temperature, relative humidity, MSLP, and wind speed and direction.

Short-range forecasting utilizing LAM ensembles is not without its share of pitfalls, however. Mass et al. (2002) reminded us that the value of short-term mesoscale ensembles remains unproven. These authors discovered diminishing returns as forecast model grid spacing drops below 12 km, when evaluated using standard measures of forecast skill (with the caveat that, for some areas of complex terrain and where orographic flows dominate, very-high resolution models produce far more realistic mesoscale weather features). In addition, the value of resolution may be underestimated when normal verification procedures are applied at fixed observation locations (usually spaced much further apart than the horizontal grid points of the forecast model). The structures of important mesoscale features often become better defined as resolution increases, but objective verification scores are profoundly degraded by even small timing and spatial errors.

Mass et al. (2002) suggested that non-traditional means of verification (e.g., verifying mesoscale structures) would demonstrate the benefits of high resolution. The authors completed their review by stating that the optimal approach to high-resolution NWP may well be a hybrid: a collection of medium-resolution ensemble runs providing probabilistic information about larger-scale evolution and forecast reliability, and a limited number of high-resolution runs to realistically simulate crucial topographic circulations in regions of complex terrain. Chapter 3 in this dissertation probes and evaluates this latter ensemble-forecasting concept.

Temperature Forecasting

Hydrologic models require accurate temperature forecasts so that the model can distinguish between rain and snow, because rain can lead to rapid surface runoff while snow usually adds to longer-term storage. High-resolution temperature forecasts are especially important in mountainous regions where precipitation type is a function of surface elevation, and where rain-on-snow events can produce exceptionally high inflows and flooding events.

Energy load models also require accurate temperature forecasts at hourly intervals in the $0-48$ hour range for generation-scheduling models, and in the one-to-seven day (or more) range for peak power consumption predictions. Urban load-center temperatures drive electrical load and peak power usage, especially during extreme heat waves and arctic outbreaks. Accurate temperature forecasts allow energy planners to ensure generation units are on-line and market purchases are secured to meet expected peak loads.

Fortunately, ensemble prediction research in temperature forecasting has proven positive in many published reports (Eckel and Mass, 2005; Jones et al., 2007). Stensrud and Yussouf (2003) published a study of a short-range ensemble prediction system for post-processed 2-m temperature and dew-point temperature forecasts that specifically cited the energy sector as a primary customer of the forecasts. Yussouf and Stensrud (2007) formulated a promising and inexpensive scheme for a short-range bias-corrected ensemble of post-processed forecasts. In Chapter 5 of this dissertation, I extend this scheme through the medium and long range (1-15 days), motivated by the benefit to energy planners and traders.

Precipitation Forecasting

Probabilistic QPFs serve as critical parameters for weather predictions because of the significant economic and social impacts of accurate precipitation forecasts (Fritsch et al., 1998). Unfortunately, precipitation forecast skill has improved relatively slowly compared to other weather elements (Sanders, 1986; Applequist et al., 2002; Chien and Ben, 2004), a consequence of the initial errors that saturate much faster for precipitation forecasting than for other sensible weather elements (Wandishin et al., 2001). Errors in the model physical parameterization schemes also play a significant role in rapid degradation of precipitation forecasts. Many authors have realized and lamented that an accurate precipitation forecast remains one of the most challenging tasks in meteorology (Olson et al., 1995; Chien and Ben, 2004; Hoinka and Davies, 2007). Therefore, it is extremely desirable to learn how much short-range ensemble forecasting can improve a quantitative precipitation forecast, given accessible model capability and computing power (Du et al., 1997)— research that gives direction to this dissertation. Nonetheless, ensemble precipitation forecast research has proven beneficial and advantageous in many case studies and circumstances [e.g., Du et al. (1997); Hamill and Colucci (1998); Mullen et al. (1999); Mullen and Buizza (2001); Wandishin et al. (2001); Bowler et al. (2006); Chien et al. (2006); Gallus et al. (2007); Jones et al. (2007); Stensrud and Yussouf (2007)].

Ensemble Prediction System Design

A seemingly endless array of choices faces the mesoscale ensemble prediction system designer, who is, ultimately, limited by computer power and real-time forecasting deadlines. Ensemble system designers should be aware of the challenges Eckel and Mass (2005) submitted to the task:

- The small-scale, surface variables of interest are often less predictable than mid-level synoptic fields and their errors may saturate too quickly for an ensemble to be of use.

- Model error, which is poorly understood and difficult to account for, has a larger impact on the surface variables in the short range than mid-level fields in the mid-range.

- The best method for defining the initial conditions is unclear since large-scale error propagation growth is initially linear. Most ensemble initial condition methodologies (e.g., breeding of growing modes) were developed for the medium range in which nonlinear error growth generates a large spread of solutions.

- Use of a limited area model may inhibit ensemble dispersion even when perturbed lateral boundary conditions are applied.

- It may be important to capture variability at small scales using very high resolution LAMs.

A sample of recent research into mesoscale ensemble prediction system design is provided next. Chien et al. (2006) found that a single 15 km resolution mesoscale model run with 16 different combinations of cumulus parameterization schemes and microphysics schemes performed well, though still underforecast rainfall. Chien et al. (2006) also found that the ensemble mean produced better forecasts than any individual member. Du et al. (1997) found, in an ensemble driven by perturbing initial conditions, that large sensitivity to initial condition uncertainty marks ensemble QPFs in special cases of explosive cyclogenesis. Eckel and Mass (2005) evaluated a multi-model (different LAMs generate each ensemble member) vs. a multi-IC (a single LAM is driven with multiple initial condition configurations derived from different global models) approach to mesoscale ensemble forecast production and found that the multi-model system exhibited greater dispersion and superior performance.

Jones et al. (2007) constructed multiple short-range 12 km resolution LAM forecasts specifically to compare an ensemble system that varies internal physics packages (planetary boundary layer schemes and convective parameterizations) to generate ensemble members with a system that varies external initial conditions (from different global driving models) to generate ensemble members; results were mixed depending on cool season vs. warm season and on weather parameter (surface temperature, wind speed, sea-level pressure, and 24-h precipitation). Jones et al. (2007) also incorporated a 14-day running mean bias calibration that added 10%−30% more skill by improving reliability. Stensrud and Yussouf (2003) cited eight studies published between 1996 and 2001 that indicated multi-model ensembles delivered more skill than varied-model ensembles; the most recent of these studies (Wandishin et al., 2001) promoted the advantages gained from a multi-model strategy, stating that even ensemble configurations with as few as five 80-km resolution members can outperform a single 29-km model. Roebber et al. (2004) submitted a comprehensive summary of open questions that remain in the quest for improving operational high-resolution ensembles for forecasters.

A review of the debate into ensemble-member generation methodologies remains incomplete without acknowledging the deleterious effect that the observation-poor Pacific Ocean region imparts on North American west coast forecasts. Two specific studies (Hacker et al., 2003; McMurdie and Mass, 2004) relate numerical forecast failures in the northeast Pacific region stemming from poor model initializations. Kelly et al. (2007) and Buizza et al. (2007) used NWP data denial experiments to show how reduced observational data over the north Pacific cause large forecast errors in days 1–3 downstream over northwestern North America. Optimal ensemble design is regional— some regions may benefit more from improving model physics, others from improving observations or data-assimilation schemes.

Hamill and Colucci (1998) provided some suggestions that guided the research into ensemble prediction testing presented in this dissertation. These authors suggested that ensemble system development should include testing to determine an optimal ensemble size and resolution. Coarser resolution forecasts are generally less precise than finer-resolution forecasts, so an increase in the size of the ensemble typically comes at the expense of somewhat lessened accuracy of each member forecast. Chapter 3 of this dissertation delves into this issue of short-range ensemble size vs. member resolution particularly for precipitation forecasts used as inputs into hydrologic models. Hamill and Colucci (1998) also stated that, though previous work showed some ability of medium-range ensembles to forecast mid-tropospheric forecast skill on the large scale, users should not assume the skill of surface weather effects at specific locations can also be forecast from day to day until rigorous testing confirms this; also, research is needed into designing an ensemble forecast system with different perturbation methodology, changes to model physics, or changes in the post-processing strategy. Chapter 5 of this dissertation explores the skill of medium-range surface temperature forecasts driven by a super-ensemble incorporating initial-condition perturbations, model physics and parameterization perturbations, and site-specific optimal post-processing.

1.4 Economic Value of Forecasts

Additional work (Murphy, 1977; Katz and Murphy, 1997; Richardson, 2000, 2003; Zhu et al., 2002) adds to this meteorological view by attributing value to specific probabilistic forecasts. Proper evaluation of the benefit of a forecast system should involve not only the intrinsic skill of the forecasts, but also knowledge of the weather-sensitivity and decision-making processes of the end users (Hamill et al., 2000a). The most prevalent economic decision model found in applied meteorologic studies is the cost/loss model (Thompson and Brier, 1955; Murphy, 1977; Katz and Murphy, 1997; Richardson, 2000, 2003). It has been incorporated into recent studies focusing on the general economic value of ensemble weather forecasts (Mullen and Buizza, 2002; Stensrud and Yussouf, 2003; Legg and Mylne, 2004, 2007). Roulston and Smith (2004) and Roulston et al. (2006) explored the social value of using probabilistic forecasts based on the cost/loss economic model.

Compared to other economic sectors, such as agriculture and transportation, relatively little work has been published in the meteorological literature defining the economic value for weather forecasts for hydrological applications. Roulin and Vannitsem (2005) incorporated a water balance model into a hydrologic ensemble prediction system for two low-relief Belgian water catchments. The value of this ensemble prediction system was evaluated using a cost/loss decision model. Chapter 4 of this dissertation incorporates the cost/loss model and a decision-theory model, building on the meteorologic evaluation presented in Chapter 3, to express the potential economic value of forecasts intended for hydro-power reservoirs based in high-precipitation watersheds in regions of complex terrain.

Recent studies investigated the skill of bias-corrected ensemble (BCE) forecasts designed specifically to save costs for energy companies (Stensrud et al., 2006). The cost/loss economic model is an integral component of Chapter 5 in this dissertation to communicate the potential value of medium and long range ensemble temperature forecasts to end users in the energy sector.

1.5 Summary

The entire realm of ensemble mesoscale weather forecasting has reached a remarkable level of maturity in less than 15 years. Wilks and Hamill (1995), in their 1995 article espousing the potential economic value of ensemble-based surface weather forecasts, published this statement on the state-of-the-art at that time:

> "It is assumed implicitly that the enhanced information available from ensemble forecasts will yield greater economic value, but to our knowledge this proposition has not yet been investigated formally. A significant impediment to this kind of investigation is that most real-world weather-sensitive decision problems relate to surface weather variables (e.g., maximum temperature or precipitation amount). While ensemble forecasting of large-scale geopotential height fields is presently practiced at major operational centers, to our knowledge no ensemble-based forecast guidance for surface weather elements is either available or under development. Therefore, we resort here to the analysis of hypothetical ensemble-based forecasts of surface weather, which may be broadly representative of the kind of forecast guidance that may eventually become available."

Compare this to the current sentiment on the subject, conveyed in the December 2007 article by Hacker and Rife (2007):

> "The maturity of limited-area, mesoscale modeling systems has occurred with an explosion in the number of users who rely on precise forecasts for regional applications."

People endeavour to forecast the weather not because it's *easy*, but because it's *important*. No less eminent a scientist than Richard Feynman, Nobel Prize laureate and one of the 20th century's greatest physicists, known (among his many remarkable achievements) for expanding the theory of quantum electrodynamics, had this to say about weather forecasting[2]:

> The theory of meteorology has never been satisfactorily worked out by the physicist… If we know the condition of air today, why can't we figure out the condition of the air tomorrow? We do not *really* know what the condition is today; the air is swirling and twisting everywhere. The motion of the air is very sensitive, and even unstable… Even a smooth moving mass of air going over a mountain turns into complex whirlpools and eddies. Nobody in physics has really been able to analyze it in spite of its importance: *circulating or turbulent fluid flow.* We cannot analyze the weather. Quickly, let's leave the subject of weather, and discuss geology!

This dissertation describes research undertaken to enhance hydrometeorological forecasts to benefit decision makers in water management and electric utility sectors. The objective of the research is to deliver an optimal methodology to produce reliable, skillful and economically valuable probabilistic temperature and precipitation forecasts.

[2] From: Feynman, R. P., Leighton, R. B., and Sands, M., 1963: *The Feynman Lectures on Physics, Vol I.* Addison-Wesley, Reading, Mass., Chapter 3 pp.7-8.

Table 1.1: Weather factors that influence hydro-electric planning operations.

Weather and energy impact	Planning timeline
Heavy rainfall resulting in extreme runoff and reservoir inflow	3-hour increments out 48 hours
Real-time generation-load balance	hourly increments out 48 hours
Daily rainfall and snowmelt	One to fourteen days
Daily peak power consumption	One to fourteen days
Monthly precipitation and temperatures	Quarterly and yearly water supply
Yearly precipitation and temperatures affecting reservoir levels and inflow regimes	Five years
Climate change affecting precipitation and temperature patterns	Decades

Bibliography

Andersson, E., J. Haseler, P. Unden, P. Courtier, G. Kelly, D. Vasiljevic, C. Brankovic, C. Cardinali, C. Gaffard, A. Hollingsworth, C. Jakob, P. Janseen, E. Klinker, A. Lanzinger, M. Miller, F. Rabier, A. Simmons, B. Strauss, J.-N. Thépaut, and P. Viterbo, 1998: The ECMWF implementation of three-dimensional variational assimilation (3D-Var). III: Experimental results. *Quart. J. Roy. Meteor. Soc.*, **124**, 1831–1860.

Applequist, S., G. E. Gahrs, R. L. Pfeffer, and X. Niu, 2002: Comparison of methodologies for probabilistic quantitative precipitation forecasting. *Wea. Forecasting*, **17**, 783–799.

Baars, J. A. and C. F. Mass, 2005: Performance of national weather service forecasts compared to operational, consensus, and weighted model output statistics. *Wea. Forecasting*, **20**, 1034–1047.

Bowler, N. E., C. E. Pierce, and A. W. Seed, 2006: STEPS: A probabilistic precipitation forecasting scheme which merges an extratropical nowcast with downscaled NWP. *Quart. J. Roy. Meteor. Soc.*, **132**, 2127–2155.

Bozic, S. M., 1994: *Digital and Kalman Filtering, Second Ed.*. John Wiley and Sons, New York.

Brooks, H., M. S. Tracton, D. J. Stensrud, G. J. DiMego, and Z. Toth, 1995: Short-range ensemble forecasting (SREF): Report from a workshop. *Bull. Amer. Meteor. Soc.*, **76**, 1617–1624.

Brooks, H. E. and C. A. Doswell, 1993: New technology and numerical weather prediction–wasted opportunity? *Weather*, **48**, 173–177.

Brunet, N., R. Verret, and N. Yacower, 1988: An objective comparison of model output statistics and "perfect prog" systems in producing numerical weather element forecasts. *Wea. Forecasting*, **3**, 273–283.

Buizza, R., C. Cardinali, G. Kelly, and J.-N. Thépaut, 2007: The value of observations. II: The value of observations located in singular-vector-based target areas. *Quart. J. Roy. Meteor. Soc.*, **133**, 1817–1832.

Buizza, R., P. L. Houtekamer, Z. Toth, G. Pellerin, M. Wei, and Y. Zhu, 2005: A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Mon. Wea. Rev.*, **133**, 1076–1097.

Buizza, R. and T. N. Palmer, 1995: The singular-vector structure of the atmospheric general circulation. *J. Atmos. Sci.*, **52**, 1434–1456.

Carter, G. M., J. P. Dallavalle, and H. R. Glahn, 1989: Statistical forecasts based on the national meteorological center's numerical weather prediction system. *Wea. Forecasting*, **4**, 401–412.

Cheng, W. Y. Y. and W. J. Steenburgh, 2007: Strengths and weaknesses of MOS, running-mean bias removal, and Kalman filter techniques for improving model forecasts over the western United States. *Wea. Forecasting*, **22**, 1304–1318.

Chien, F. and J. J. Ben, 2004: MM5 ensemble mean precipitation forecasts in the Taiwan area for three early summer convective (Mei-Yu) seasons. *Wea. Forecasting*, **19**, 735–750.

Chien, F., Y.-C. Liu, and J. J. Ben, 2006: MM5 ensemble mean forecasts in the Taiwan area for the 2003 Mei-Yu season. *Wea. Forecasting*, **21**, 1006–1023.

Clark, M., L. Hay, A. Slater, K. Werner, D. Brandon, A. Barrett, S. Gangopadhyay, and B. Rajagopalan, 2004: Ensemble streamflow forecasting in snowmelt dominated river basins. *GEWEX News*, **14**, 4–6.

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

Conway, D., R. L. Wilby, and P. D. Jones, 1996: Precipitation and air flow indices over the British Isles. *Climate Research*, **7**, 169–183.

Delle Monache, L., 2005: *Ensemble-Averaged, Probabilistic, and Kalman-Filtered Regional Ozone Forecasts*. Ph.D. thesis, University of British Columbia.

Dempsey, C. L., K. W. Howard, R. A. Maddox, and D. H. Phillips, 1998: Developing advanced weather technologies for the power industry. *Bull. Amer. Meteor. Soc.*, **79**, 1019–1035.

Droegemeier, K. K., J. D. Smith, S. Businger, C. Doswell, J. Doyle, C. Duffy, E. Foufoula-Georgiou, T. Graziano, L. D. James, V. Krajewski, M. LeMone, D. Lettenmaier, C. Mass, R. Pielke Sr., P. Ray, S. Rutledge, J. Schaake, and E. Zipser, 2000: Hydrological aspects of weather prediction and flood warnings: Report of the ninth prospectus development team of the U.S. weather research program. *Bull. Amer. Meteor. Soc.*, **81**, 2665–2680.

Du, J., S. Mullen, and F. Sanders, 1997: Short-range ensemble forecasting of quantitative precipitation. *Mon. Wea. Rev.*, **125**, 2427–2459.

Du, J. and M. S. Tracton, 2001: Implementation of a real-time short-range ensemble forecasting system at NCEP: An update. *Preprints, Ninth Conf. on Mesoscale Processes*, Amer. Meteor. Soc., Ft. Lauderdale, FL, 355–360.

Eady, E., 1951: The quantitative theory of cyclone development. *Compendium of Meteorology*, T. Malone, ed., Amer. Meteor. Soc., 464–469.

Eckel, F. A. and C. F. Mass, 2005: Aspects of effective mesoscale, short-range ensemble forecasting. *Wea. Forecasting*, **20**, 328–350.

Engel, C. and E. Ebert, 2007: Performance of hourly operational consensus forecasts (OCFs) in the Australian region. *Wea. Forecasting*, **22**, 1345–1359.

Epstein, E., 1969: Stochastic dynamic prediction. *Tellus*, **21**, 739–759.

Fritsch, J. M., R. A. Houze Jr., R. Adler, H. Bluestein, L. Bosart, J. Brown, F. Carr, C. Davis, R. H. Johnson, N. Junker, Y.-H. Kuo, S. Rutledge, J. Smith, Z. Toth, J. W. Wilson, E. Zipser, and D. Zrnic, 1998: Quantitative precipitation forecasting: Report of the eighth prospectus development team, U.S. weather research program. *Bull. Amer. Meteor. Soc.*, **79**, 285–299.

Gallus, W. A. J., M. E. Baldwin, and K. L. Elmore, 2007: Evaluation of probabilistic precipitation forecasts determined from Eta and AVN forecasted amounts. *Wea. Forecasting*, **22**, 207–215.

Gel, Y., 2007: Comparative analysis of the local observation-based (LOB) method and the nonparametric regression-based method for gridded bias correction in mesoscale weather forecasting. *Wea. Forecasting*, **22**, 1243–1256.

Glahn, H. and R. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.*, **11**, 1203–1211.

Gneiting, T. and A. E. Raftery, 2005: Weather forecasting with ensemble methods. *Science*, **310**, 248–249.

Grimit, E. P. and C. F. Mass, 2002: Initial results of a mesoscale short-range ensemble forecasting system over the Pacific northwest. *Wea. Forecasting*, **17**, 192–205.

Gutierrez, J. M., A. Cofio, R. Cano, and M. Rodriguez, 2004: Clustering methods for statistical downscaling in short-range weather forecasts. *Mon. Wea. Rev.*, **132**, 2169–2183.

Hacker, J. P., E. S. Krayenhoff, and R. B. Stull, 2003: Ensemble experiments on numerical weather prediction error and uncertainty for a North Pacific forecast failure. *Wea. Forecasting*, **18**, 12–31.

Hacker, J. P. and D. L. Rife, 2007: A practical approach to sequential estimation of systematic error on near-surface mesoscale grids. *Wea. Forecasting*, **22**, 1257–1273.

Hall, T., H. E. Brooks, and C. A. Doswell, 1999: Precipitation forecasting using a neural network. *Wea. Forecasting*, **14**, 338–345.

Hamill, T. M. and S. J. Colucci, 1998: Verification of Eta-RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.*, **126**, 711–724.

Hamill, T. M., S. L. Mullen, C. Snyder, Z. Toth, and D. P. Baumhefner, 2000a: Ensemble forecasting in the short to medium range: Report from a workshop. *Bull. Amer. Meteor. Soc.*, **81**, 2653–2664.

Hamill, T. M., C. Snyder, and R. E. Morss, 2000b: A comparison of probabilistic forecasts from bred, singular-vector, and perturbed observation ensembles. *Mon. Wea. Rev.*, **128**, 1835–1851.

Hansen, B., 2007: A fuzzy logic-based analog forecasting system for ceiling and visibility. *Wea. Forecasting*, **22**, 1319–1330.

Hart, K. A., W. J. Steenburgh, D. J. Onton, and A. J. Siffert, 2004: An evaluation of mesoscale-model-based output statistics (MOS) during the 2002 Olympic and Paralympic games. *Wea. Forecasting*, **19**, 200–218.

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**, 56–75.

Hoinka, K. P. and H. C. Davies, 2007: Upper-tropospheric flow features and the Alps: An overview. *Quart. J. Roy. Meteor. Soc.*, **133**, 847–865.

Homleid, M., 1995: Diurnal corrections of short-term surface temperature forecasts using the Kalman Filter. *Wea. Forecasting*, **10**, 689–707.

Hou, D., E. Kalnay, and K. K. Droegemeier, 2001: Objective verification of the SAMEX '98 ensemble forecasts. *Mon. Wea. Rev.*, **129**, 73–91.

Houtekamer, P. L. and L. Lefaivre, 1997: Using ensemble forecasts for model validation. *Mon. Wea. Rev.*, **125**, 2416–2426.

Houtekamer, P. L., L. Lefaivre, J. Derome, H. Ritchie, and H. L. Mitchell, 1996: A system simulation approach to ensemble prediction. *Mon. Wea. Rev.*, **124**, 1225–1242.

Houtekamer, P. L. and H. L. Mitchell, 1998: Data assimilation using an ensemble Kalman filter technique. *Mon. Wea. Rev.*, **126**, 796–811.

Jolliffe, I. T. and D. B. Stephenson, 2003: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. J. Wiley, England.

Jones, M. S., B. A. Colle, and J. S. Tongue, 2007: Evaluation of a mesoscale short-range ensemble forecast system over the northeast United States. *Wea. Forecasting*, **22**, 36–55.

Kalman, R. E. and R. S. Bucy, 1961: New results in linear filtering and prediction theory. *Trans. ASME, J. Basic Eng.*, **83**, 95–108.

Kalnay, E., 2003: *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, New York.

Katz, R. W. and A. H. Murphy, 1997: *Economic Value of Weather and Climate Forecasts*. Cambridge University Press, England.

Kelly, G., J.-N. Thépaut, R. Buizza, and C. Cardinali, 2007: The value of observations. I: Data denial experiments for the Atlantic and the Pacific. *Quart. J. Roy. Meteor. Soc.*, **133**, 1803–1815.

Klein, W. H., B. M. Lewis, and I. Enger, 1959: Objective prediction of five-day mean temperatures during winter. *J. Atmos. Sci.*, **16**, 672–682.

Kobayashi, C., K. Yoshimatsu, S. Maeda, and K. Takano, 1996: Dynamical one-month forecasting at JMA. *Preprints, 11th Conf. on Numerical Weather Prediction*, Amer. Meteor. Soc., Norfolk, VA, 13–14.

Koizumi, K., 1999: An objective method to modify numerical model forecasts with newly given weather data using an artificial neural network. *Wea. Forecasting*, **14**, 109–118.

Krzysztofowicz, R., 2001: The case for probabilistic forecasting in hydrology. *J. Hydrol.*, **249**, 2–9.

Krzysztofowicz, R., W. J. Drzal, T. R. Drake, J. C. Weyman, and L. A. Giordano, 1993: Probabilistic quantitative precipitation forecasts for river basins. *Wea. Forecasting*, **8**, 424–439.

Legg, T. P. and K. R. Mylne, 2004: Early warnings of severe weather from ensemble forecast information. *Wea. Forecasting*, **19**, 891–906.

— 2007: Corrigendum. *Wea. Forecasting*, **22**, 216–219.

Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.*, **102**, 409–418.

Lewis, J. M., 2005: Roots of ensemble forecasting. *Mon. Wea. Rev.*, **133**, 1865–1885.

Lorenz, E., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.*, **20**, 130–141.

Majewski, D., 1997: Operational regional prediction. *Meteor. Atmos. Phys.*, **63**, 89–104.

Marzban, C., 2003: Neural networks for postprocessing model output: ARPS. *Mon. Wea. Rev.*, **131**, 1103–1111.

Marzban, C., S. Sandgathe, and E. Kalnay, 2006: MOS, perfect prog, and reanalysis. *Mon. Wea. Rev.*, **134**, 657–663.

Mass, C. F., D. Ovens, K. Westrick, and B. Cole, 2002: Does increasing horizontal resolution produce more skillful forecasts? *Bull. Amer. Meteor. Soc.*, **83**, 407–430.

McMurdie, L. and C. Mass, 2004: Major numerical forecast failures over the Northeast Pacific. *Wea. Forecasting*, **19**, 338–356.

Mohanty, U. C. and P. Dimri, 2004: Location-specific prediction of the probability of occurrence and quantity of precipitation over the western Himalayas. *Wea. Forecasting*, **19**, 520–533.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF ensemble prediction system: methodology and validation. *Quart. J. Roy. Meteor. Soc.*, **122**, 73–119.

Morassutti, M. P., 2004: Evaluation of quantitative precipitation and temperature forecast translation methods for use in hydrologic models. Technical Report Water Management Interest Group Report No. T032700-0402, CEA Technologies, Inc.

Mullen, S. L. and D. P. Baumhefner, 1991: Monte Carlo simulations of explosive cyclogenesis using a low-resolution, global spectral model. *Preprints, Ninth Conf. on Numerical Weather Prediction*, Amer. Meteor. Soc., Denver, CO, 750–751.

— 1994: Monte Carlo simulations of explosive cyclogenesis. *Mon. Wea. Rev.*, **122**, 1548–1567.

Mullen, S. L. and R. Buizza, 2001: Quantitative precipitation forecasts over the United States by the ECMWF ensemble prediction system. *Mon. Wea. Rev.*, **129**, 638–663.

— 2002: The impact of horizontal resolution and ensemble size on probabilistic forecasts of precipitation by the ECMWF ensemble prediction system. *Wea. Forecasting*, **17**, 173–191.

Mullen, S. L., J. Du, and F. Sanders, 1999: The dependence of ensemble dispersion on analysis-forecast systems: Implications to short-range ensemble forecasting of precipitation. *Mon. Wea. Rev.*, **127**, 1674–1686.

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

— 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting*, **8**, 281–293.

Olson, D. A., N. W. Junker, and B. Korty, 1995: Evaluation of 33 years of quantitative precipitation forecasting at the NMC. *Wea. Forecasting*, **10**, 498–511.

Palmer, T. N., F. Molteni, R. Mureau, R. Buizza, P. Chapelet, and J. Tribbia, 1993: Ensemble prediction. *Proc. ECMWF Seminar on validation of models over Europe*, ECMWF, Shinfield Park, Reading, UK, volume 1, 21–66.

Palmer, T. N., R. Mureau, and M. F., 1990: The Monte Carlo forecast. *Weather*, **45**, 198–207.

Rennick, M. A., 1995: The ensemble forecast system (EFS). Technical Report 2-95, Fleet Numerical and Oceanography Center, 19 pp., [Available from Models Department, FLENUMMETOCCEN, 7 Grace Hopper Ave., Monterey, CA, 93943].

Richardson, D. S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **126**, 649–667.

— 2003: Economic value and skill. *Forecast verification: A practitioner's guide in atmospheric science*, I. Jolliffe and D. B. Stephenson, eds., Wiley, 164–187.

Roebber, P. J., D. M. Schultz, B. A. Colle, and D. J. Stensrud, 2004: Toward improved prediction: High-resolution and ensemble modeling systems in operations. *Wea. Forecasting*, **19**, 936–949.

Roeger, C., R. Stull, D. McClung, J. Hacker, X. Deng, and H. Modzelewski, 2003: Verification of mesoscale numerical weather forecasts in mountainous terrain for application to avalanche prediction. *Wea. Forecasting*, **18**, 1140–1160.

Roulin, E. and S. Vannitsem, 2005: Skill of medium-range hydrological ensemble predictions. *J. Hydrometeorology.*, **6**, 729–744.

Roulston, M. S., G. E. Bolton, A. N. Kleit, and A. L. Sears-Collins, 2006: A laboratory study of the benefits of including uncertainty information in weather forecasts. *Wea. Forecasting*, **21**, 116–122.

Roulston, M. S. and L. A. Smith, 2004: The boy who cried wolf revisited: The impact of false alarm intolerance on cost-loss scenarios. *Wea. Forecasting*, **19**, 391–397.

Saetra, O., H. Hersbach, J.-R. Bidlot, and D. S. Richardson, 2004: Effects of observation errors on the statistics for ensemble spread and reliability. *Mon. Wea. Rev.*, **132**, 1487–1501.

Sanders, F., 1986: Trends in skill of Boston forecasts made at MIT, 1966-84. *Bull. Amer. Meteor. Soc.*, **67**, 170–176.

Schaake, J. C., T. M. Hamill, R. Buizza, and M. Clark, 2007: HEPEX: The hydrologic ensemble prediction experiment. *Bull. Amer. Meteor. Soc.*, **88**, 1541–1547.

Stensrud, D. J., 2006: NEHRTP workshop: Improving weather forecast services used by the electric utility industry. *Bull. Amer. Meteor. Soc.*, **87**, 499–501.

Stensrud, D. J., J. Bao, and T. T. Warner, 2000: Using initial conditions and model physics perturbations in short-range ensemble simulations of mesoscale convective systems. *Mon. Wea. Rev.*, **128**, 2077–2107.

Stensrud, D. J., H. E. Brooks, J. Du, M. S. Tracton, and E. Rogers, 1999: Using ensembles for short-range forecasting. *Mon. Wea. Rev.*, **127**, 433–446.

Stensrud, D. J. and J. A. Skindlov, 1996: Gridpoint predictions of high temperature from a mesoscale model. *Wea. Forecasting*, **11**, 103–110.

Stensrud, D. J. and N. Yussouf, 2003: Short-range ensemble predictions of 2-m temperature and dewpoint temperature over New England. *Mon. Wea. Rev.*, **131**, 2510–2524.

— 2007: Reliable probabilistic quantitative precipitation forecasts from a short-range ensemble forecasting system. *Wea. Forecasting*, **22**, 3–17.

Stensrud, D. J., N. Yussouf, M. E. Baldwin, J. T. McQueen, J. Du, B. Zhou, B. Ferrier, G. Manikin, F. M. Ralph, J. M. Wilczak, A. B. White, I. Djlalova, J.-W. Bao, R. J. Zamora, S. G. Benjamin, P. A. Miller, T. L. Smith, T. Smirnova, and M. F. Barth, 2006: The New England high-resolution temperature program. *Bull. Amer. Meteor. Soc.*, **87**, 491–498.

Taylor, A. A. and L. M. Leslie, 2005: A single-station approach to model output statistics temperature forecast error assessment. *Wea. Forecasting*, **20**, 1006–1020.

Thompson, J. C. and G. W. Brier, 1955: The economic utility of weather forecasts. *Mon. Wea. Rev.*, **83**, 249–254.

Thompson, P., 1957: Uncertainty of initial state as a factor in predictability of large-scale atmospheric flow patterns. *Tellus*, **9**, 275–295.

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

Toth, Z., E. Kalnay, M. S. Tracton, R. Wobus, and J. Irwin, 1997: A synoptic evaluation of the NCEP ensemble. *Wea. Forecasting*, **12**, 140–153.

Tracton, M. S. and E. Kalnay, 1993: Operational ensemble prediction at the National Meteorological Center: Practical aspects. *Wea. Forecasting*, **8**, 378–398.

Vislocky, R. L. and J. M. Fritsch, 1995: Generalized additive models versus linear regression in generating probabilistic MOS forecasts of aviation weather parameters. *Wea. Forecasting*, **10**, 669–680.

Vislocky, R. L. and G. Y. Young, 1989: The use of perfect prog forecasts to improve model output statistics forecasts of precipitation probability. *Wea. Forecasting*, **4**, 202–209.

Wandishin, M. S., S. L. Mullen, D. J. Stensrud, and H. E. Brooks, 2001: Evaluation of a short-range multimodel ensemble system. *Mon. Wea. Rev.*, **129**, 729–747.

Wang, W. and N. L. Seaman, 1997: A comparison study of convective parameterization schemes in a mesoscale model. *Mon. Wea. Rev.*, **125**, 252–278.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*. Academic Press, San Diego, 2nd edition.

Wilks, D. S. and T. M. Hamill, 1995: Potential economic value of ensemble-based surface weather forecasts. *Mon. Wea. Rev.*, **123**, 3564–3575.

Wilson, L. J. and M. Vallée, 2002: The Canadian updateable model output statistics (UMOS) system: Design and development tests. *Wea. Forecasting*, **17**, 206–222.

— 2003: The Canadian updateable model output statistics (UMOS) system: Validation against perfect prog. *Wea. Forecasting*, **18**, 288–302.

Wonnacott, T. H. and R. J. Wonnacott, 1972: *Introductory Statistics*. Wiley, New York.

Woodcock, F. and C. Engel, 2005: Operational consensus forecasts. *Wea. Forecasting*, **20**, 101–111.

Yuan, H., X. Gao, S. L. Mullen, S. Sorooshian, J. Du, and H.-M. H. Juang, 2007: Calibration of probabilistic quantitative precipitation forecasts with an artificial neural network. *Wea. Forecasting*, **22**, 1287–1303.

Yuan, H., S. Mullen, X. Gao, S. Sorooshian, J. Du, and H.-M. H. Juang, 2005: Verification of probabilistic quantitative precipitation forecasts over the southwest United States during winter 2002/03 by the RSM ensemble system. *Mon. Wea. Rev.*, **133**, 279–294.

Yussouf, N. and D. J. Stensrud, 2007: Bias-corrected short-range ensemble forecasts of near surface variables during the 2005/2006 cool season. *Wea. Forecasting*, **22**, 1274–1286.

Yuval and W. Hsieh, 2003: An adaptive nonlinear MOS scheme for precipitation forecasts using neural networks. *Wea. Forecasting*, **18**, 303–310.

Zhu, Y., Z. Toth, R. Wobus, D. Richardson, and K. Mylne, 2002: The economic value of ensemble-based weather forecasts. *Bull. Amer. Meteor. Soc.*, **83**, 73–83.

Chapter 2

Hydrometeorological Accuracy Enhancement via Post-processing of Numerical Weather Forecasts in Complex Terrain[3].

2.1 Introduction

Statistical post-processing to calibrate NWP output has a long history in operational weather forecasting. It is firmly established that statistical post-processing of NWP output significantly improves the skill of deterministic forecasts, primarily through the reduction of systematic errors (Hamill et al., 2000). The goal of post-processing was originally to produce point forecasts of sensible weather elements that were not readily available from early low-resolution models, incorporating statistical relationships between large-scale aspects of the flow and specific point observations. More recently, post-processing has been employed to improve the skill of point forecasts of sensible weather elements obtainable directly from higher-resolution models. The objective is to reduce the future systematic error between direct model output (DMO) forecasts and verifying observations by building and employing statistical relationships between past model output and past observations.

Previous studies (Stensrud and Skindlov, 1996; Mao et al., 1999; Eckel and Mass, 2005; Stensrud and Yussouf, 2005) have shown how straightforward moving-average techniques can reduce systematic error in DMO. In the study presented in this paper, moving average and other related post-processing techniques are applied to daily maximum and minimum temperature forecasts and daily quantitative precipitation forecasts (QPFs). These sensible weather element forecasts of temperature and precipitation are extremely important in hy-

[3] A version of this chapter has been published. McCollor, D. and R. Stull, 2008: Hydrometeorological Accuracy Enhancement via Postprocessing of Numerical Weather Forecasts in Complex Terrain. *Wea. Forecasting*, **23**, 131-144

drometeorological applications such as runoff forecasting. The goal of these post-processing techniques is to reduce the error in the current forecast using estimates of past error.

To achieve this goal, a balance must be reached between a short learning period to enable an updating system to respond quickly to changes in error patterns, and a longer training period that increases statistical stability (Woodcock and Engel, 2005). Several moving-average and weighted-average methods are compared in the present paper, along with a Kalman Filter technique, to determine the effect of these straightforward post-processing techniques in reducing systematic error in DMO forecasts. Secondly, the techniques presented here are compared to post-processed forecasts produced by the Canadian Meteorological Centre (CMC), in an effort to test these new techniques against an established standard.

2.1.1 Background

Both systematic error and random error exist in a forecast. In mountainous regions, the systematic component of error, often termed *bias*, is largely due to differences between model topography and actual orography and landuse, and also deficiencies in model representation of physical processes. Subgrid-scale processes parameterized in the model do not adequately represent local dynamic and thermodynamic effects required for detailed point-specific forecasting (Hart et al., 2004). Improvement in model parameterization and representation of topography as well as post-processing DMO can reduce the systematic error component. Random forecast error, on the other hand, is largely associated with the inherent chaotic nature of atmospheric motion and cannot be reduced through post-processing. Associated observations also have inherent and sometimes unknown bias and random error due to equipment limitations such as sampling rate, calibration, and site representation.

Forecast error, then, can be written $\varepsilon^f = x^f - x^t$, where x^t is the unknown truth, and x^f is a forecast. Sample realizations of the real atmosphere can be taken as observations y^o, which also have error $\varepsilon^o = y^o - x^t$. If observation errors are random (not due to equipment mis-calibration), then $\langle y^o \rangle = \langle x^t \rangle$ where $\langle\ \rangle$ is the expected or time-averaged value. A bias is then the expected value of the forecast error, $\langle \varepsilon^f \rangle = \langle x^f - x^t \rangle = \langle x^f - y^o \rangle$. Expected values can be estimated from a sufficiently large sample of errors. The expectation is approximated by the mean of a given sample of errors, where the sample can include the entire climatological record or be limited to (hence characteristic of) the recent synoptic flow regime. The best post-processing techniques will attempt to address both sources of bias in the forecast.

Calibrated forecasts of temperature and precipitation are necessary when used as input into hydrologic streamflow forecast models. Such forecasts pose a unique challenge in areas of complex terrain where meteorological conditions exhibit dramatic spatial variability (Hart et al., 2004). Specifically for hydrometeorological applications, improvements in streamflow forecasts require more accurate local-scale forecasts of precipitation and temperature (Clark and Hay, 2004).

DMO of sensible weather elements, such as temperature and precipitation amount, are often not optimal estimates of local conditions in complex terrain. One reason is large systematic error inherent in the NWP models (Mao et al., 1999; Clark and Hay, 2004). Numerous methods have been devised to adapt NWP forecasts to produce local, or point forecasts for specific applications. The general methodology of post-processing techniques derives a statistical regression between NWP produced values and weather-element observations over a set period of time, and then uses this regression to adjust future NWP forecasts. If successful, the adjusted, or post-processed forecasts should have reduced mean error compared to the original DMO.

In complex mountainous terrain, regression adjustments to local NWP forecasts are highly flow-dependent. All but extremely high-resolution NWP models will have difficulty discerning topographically-induced differences in precipitation patterns, but a properly trained post-processing technique could likely improve the individual forecasts by greatly reducing measured forecast error due to topographic effects. Only post-processing techniques with very short averaging periods would be able to adapt rapidly changing flow-dependent conditional bias correction factors. However, short averaging periods lead to statistical instability. The best post-processing methods must find a balanced trade-off between conditional bias correction and statistical stability.

2.1.2 Review

The earliest method of statistical post-processing that gained widespread use was the perfect prog method (Klein et al., 1959), which derives regression equations between analyzed fields and near concurrent observations. However, since the perfect prog method uses analyses for the predictors, forecast model error is not taken into account in determining the regression coefficients. The method of Model Output Statistics, or MOS (Glahn and Lowry, 1972), uses model forecast fields for predictors hence MOS forecast variance is both model and forecast-period dependent. Therefore, MOS requires forecasts from a model that remains unchanged over at least several seasons (Stensrud and Yussouf, 2005) so that the operational forecast relationship matches the developmental historical relationship. Woodcock and Engel (2005) suggested that the importance of DMO will increase relative to MOS because MOS cannot easily accommodate new sites, models, and model changes; MOS often employs over a million predictive equations that require at least 2 years of stable developmental data for their derivation (Jacks et al., 1990).

For hydrometeorological applications that require operational forecasts for many locations and many projection times, the development and maintenance of a MOS system can prove too onerous. Even national weather centers such as the Meteorological Service of Canada (MSC) discontinued their MOS forecast system in the late 1980s because significant model changes had become too frequent and the development cost too great to maintain a statistically stable MOS system (Wilson and Vallée, 2002).

Further research has indicated updateable MOS (UMOS) can lessen the disadvantage of model variance over standard MOS routines. UMOS (Wilson and Vallée, 2002, 2003) can be implemented with briefer training than MOS, but requires an even more extensive database overhead to maintain and update. UMOS forecasts in western Canada are provided by MSC, but only at a limited set of valley-bottom airport sites, similar to that found by Hart et al. (2004) for MOS forecasts in the western U.S.

The drawbacks to MOS-based techniques have led to other approaches to statistically post-process model forecast output that do not require long data archival periods (Stensrud and Yussouf, 2003). The overall objective of error correction is to minimize the error of the next forecast using measurement estimates of past errors. A short learning period enables an updating system to respond quickly to factors that affect error (for example model changes and changes in weather regime) but increases vulnerability to missing data and/or large errors and other limitations of small sample estimates (Woodcock and Engel, 2005).

Simple 7-day and 12-day moving average error calculations are shown to improve upon raw model point forecasts (Stensrud and Skindlov, 1996; Stensrud and Yussouf, 2003, 2005). Eckel and Mass (2005) successfully implemented a moving average technique as a post-processing method to reduce systematic error in DMO, using a 14-day moving average error calculation to reduce mean error in temperature forecasts. Jones et al. (2007) found a 14-day moving average bias correction improved temperature forecasts better than a 7-day or 21-day moving average. Woodcock and Engel (2005) stated a 15-30 day bias correction window effectively removes bias in DMO forecasts using the best easy systematic estimator.

Kalman filtering (Homleid, 1995; Majewski, 1997; Kalnay, 2003; Roeger et al., 2003; Delle Monache et al., 2006) is adaptive to model changes, using the previous observation-forecast pair to calculate model error. It then predicts the model error to correct the next forecast. This recursive, adaptive method does not need an extensive, static database to be trained.

In summary, statistical post-processing adds value to DMO by objectively reducing systematic error between forecasts and observations by producing site-specific forecasts. In the following sections, several weighted-average post-processing methods, a recursive method, and an updateable MOS method are investigated for hydrometeorological applications (specifically temperature and precipitation forecasts) in complex terrain.

2.2 Data

A meso-network of observations and forecasts from 19 locations in southwestern British Columbia is employed in this study to compare various post-processing techniques (see Fig. 2.1). The meso-network includes 12 observation sites operated as part of a hydrometeorologic data analysis and forecast program in support of reservoir operations for 15 managed watersheds within the region. Additionally, seven sites maintained by MSC were included in this study. Station elevations range from 2 m AMSL at coastal locations to 1920 m AMSL in high mountainous watersheds, to capture significant orographic components of the forecasts and observations. Southwestern British Columbia is a topographically complex region of land/ocean boundaries, fjords, glaciers, and high mountains, and is subject to rapidly changing weather conditions under the influence of land-falling Pacific frontal systems.

2.2.1 Temperature

The time period 1 April 2004 to 31 March 2005 defined the one-year evaluation portion of this study. This period exhibited examples of extreme temperature. The lowest temperature recorded at any of the stations during the period was -24°C, while the highest temperature recorded at any station was $+38$°C. Concurrent forecasts were obtained from NWP forecasts produced by the Canadian Meteorological Centre. Point forecasts for each of the 19 sites were obtained for forecast days one through eight from the operational Global Environmental Multiscale (GEM) model (Côté et al., 1998a,b). Each forecast day was treated separately by applying the techniques independently to a series of forecast-observation pairs valid for each particular forecast day (one through eight) only.

For maximum and minimum temperature forecasts, six different post-processing techniques were applied to the DMO on a daily basis for direct comparison during the period of study. The first method was a seasonal mean error (SNL). In SNL, the average mean error over a constant previous six-month period was subtracted from the current day's forecast. SNL mean error was seasonally adjusted. For forecasts issued for the cool season (1 October through 31 March), a mean error calculated from the previous cool season was calculated and applied. For the warm season (1 April through 30 September), a separate mean error was calculated from the previous warm season and applied. The second method was a moving average of the previous error estimates (MA), with each previous value receiving equal weight. A simple moving average is the unweighted mean of the previous n data points, where n is a pre-determined value. The third method was a linear-weighted average of the previous errors (LIN). For LIN forecasts, recent error estimates are weighted more heavily than previous error, in a linear fashion. The fourth method was a weighted average with a $\cos^2$ weighting function (COS^2) applied. The fifth method uses the best easy systematic estimator (BES) as described in Woodcock and Engel (2005). The BES method is robust with respect to extreme values but represents the bulk of the mean error distribution because it involves quartiles:

$$BES = (Q1 + 2Q2 + Q3)/4 \tag{2.1}$$

where Q1, Q2, and Q3 are the first, second, and third quartiles of the mean error, respectively. The sixth method was a Kalman Filter (KF) applied to the forecast error (see Appendix A for a description of the KF method), which recursively weights recent errors more heavily than past errors.

The concept of using LIN and COS^2 weighting functions is also to weight recent error estimates more heavily than past error estimates in a smoothly-varying form. The objective is to achieve a balance between a short weighting period to include recent changes in error due to weather regime and model changes and a longer weighting period to enhance statistical stability. The post-processing techniques are summarized in Table 2.1.

The equation used to apply the weights is given by:

$$\varepsilon_c^f = \frac{1}{M}\sum_{k=1}^{M} w_k \cdot \varepsilon_k^f \tag{2.2}$$

where

- ε_c^f is the error correction applied to today's DMO by subtracting this value from the current DMO forecast;
- M is the number of prior days errors to be weighted. It is also known as the window length;
- w_k is the weight on the k^{th} day prior to the current day, where $\sum_{k=1}^{M} w_k = 1$; and
- ε_k^f is the error estimate on the k^{th} day prior to the current day, $x_k^f - y_k^o$.

The next issue to address is the length of the sample error correction window, to find a balance between synoptic pattern changes and model changes (requiring a shorter window) and statistical stability (requiring a longer window). Published studies include windows of $M = 7$ days (Stensrud and Skindlov, 1996), 12 days (Stensrud and Yussouf, 2005), 21 days (Mao et al., 1999), and 15-30 days (Woodcock and Engel, 2005). Woodcock and Engel (2005) tested the relationship between number of days in the running error-correction window and improvement in the day one forecast.

A similar test was performed in the study reported here, for forecast days one through eight. It was found that all post-processing methods tested showed a rapid decrease in mean absolute error, reaching an asymptotic value by 14 days (see Fig. 2.2 for the test using maximum temperature forecasts and the LIN method of mean error reduction). In Fig. 2.2, all forecast days show a rapid improvement in MAE initially then reach a long-term value. Longer range forecasts take systematically more time to reach the long-term value. By day 14 all forecast days have reached a steady value. Similar results were obtained (not shown) for the MA, COS^2, and BES methods and for minimum temperatures. Therefore, a common value of 14 days was employed for the MA, LIN, COS^2, and BES methods.

UMOS

To examine a direct comparison between the post-processing methods described in this study and one of the standard methods described in section 2.1, a set of UMOS temperature forecasts was obtained from CMC for the same forecast period: 1 April 2004 through 31 March 2005. UMOS forecasts were available only for a subset of stations consisting of six of the seven MSC stations, and were available only for temperature. Also, only the first two days of the forecast cycle have UMOS forecasts available. The post-processing techniques listed in Table 2.1 were recalculated for this same subset of UMOS-available stations. The UMOS forecasts are included in this study to serve as a direct comparison among the various weighted average techniques described in section 2.2.1.

2.2.2 Precipitation

Forecast verification for precipitation was performed for the six month period 1 October 2004 through 31 March 2005, to encompass the local wet season. The verification period exhibited examples of extreme precipitation: the highest 24-hour precipitation amount reported at a single site was 152 mm.

Observations of 24-hour precipitation amounts were obtained from the observation network described at the beginning of section 2.2. Equivalent forecasts were obtained from NWP forecasts produced by the same CMC GEM model that produced the temperature forecasts described in section 2.2.1. Point forecasts for each of the 19 sites were obtained for forecast days one through eight. Each forecast day was treated separately by applying techniques independently to a series of forecast-observation pairs valid for each particular forecast day (one through eight) only.

Accuracy in precipitation forecasts is a much more challenging goal than for temperature forecasts for reasons discussed later in section 2.3.2. Day-to-day variability in precipitation is much higher than for temperature; therefore, a longer averaging period is needed. In addition, sample mean error correction, as applied to the temperature forecasts, is not appropriate for precipitation because the resulting bias-corrected forecast could be negative. An appropriate error measure for quantitative precipitation forecasts is degree of mass balance (DMB) between DMO and observations. DMB describes the ratio of the predicted to the observed net water mass for a given interval (Grubišić et al., 2005), and is given by

$$DMB_N = \frac{\sum_{k=1}^{N} x_k^f}{\sum_{k=1}^{N} y_k^o} \tag{2.3}$$

where DMB_N is the degree of mass balance for the interval of N days, x_k^f is the 24-hour precipitation forecast for day k and y_k^o is the associated observation for day k.

DMB cannot be applied to a single day's forecast-observation pair because the operation would result in division by zero on days with no observed precipitation. DMB must cover a period of time for which some precipitation has been observed. For the current study, it was found that a period of 21 days was sufficient to ensure precipitation had occurred and that DMB could be evaluated. For the same reason, some of the bias-correction methods employed for temperature foecasts are not suitable for precipitation forecasts because they must be applied on a daily basis. The techniques suitable for DMB-correction are SNL, MA, and BES.

A DMB-correction method should, ideally, correct model over- or under-prediction of precipitation. DMB calculations with varying length of MA window from 21 to 120 days, for forecast days one through eight, are shown in Fig. 2.3. The results indicate that the MA technique corrects the DMB to within about 10% with a DMB-correction window of 30-40 days. By day 40 all forecast days have essentially reached a steady value between 1.0 and 1.1 DMB, indicating the method is producing forecasts that reflect the quantity of precipitation observed. Similar results (not shown) were obtained for the BES method. In Fig. 2.3, the DMB values for forecast day seven drift slightly to lower values at longer range. This drifting away from a steady value for forecast day seven does not appear to affect the results, as shown in the next section.

It was also found that the post-processing methods applied to precipitation showed a decrease in mean absolute error, reaching an asymptotic value by 40 days (see Fig. 2.4 for the test employing the MA method of DMB-correction). The longer range forecasts take longer to stabilize to a steady value than the shorter range forecasts. By day 40 all forecast days have reached a steady MAE value. Therefore a common value of 40 days was employed for the MA and BES methods.

DMB-corrected quantitative precipitation forecasts, using the MA and BES methods, are calculated by

$$QPF^c = \frac{DMO}{DMB^N} \tag{2.4}$$

where

- DMB^N is the DMB-correction error applied to the current DMO forecast;
- QPF^c is the new DMB-corrected QPF forecast.

For 24-hour precipitation forecasts, the SNL method averaged DMB over a constant six month period, and this factor was applied to each day's forecast. SNL was seasonally adjusted, that is, for forecasts issued for the wet season 1 October through 31 March, a constant wet-season DMB was calculated from the same period of the previous year and applied.

2.3 Results

2.3.1 Daily Temperature

All seven post-processing methods tested in this analysis indicated improvement over the DMO maximum and minimum daily temperature forecasts, for all forecast periods day one through eight. Mean error (ME) and mean absolute error (MAE) were calculated to evaluate the forecasts using the following standard equations:

$$ME = \frac{1}{N}\sum_{k=1}^{N}(x_k^f - y_k^o) \tag{2.5}$$

$$MAE = \frac{1}{N}\sum_{k=1}^{N}|x_k^f - y_k^o| \tag{2.6}$$

where N is the number of forecast-observation pairs, and x_k^f and y_k^o are individual forecasts and observations for day k, respectively.

In terms of mean error, the daily maximum temperature DMO forecast errors ranged from -2.5°C to -3.5°C for days one through eight, respectively, indicating a cold bias to DMO maximum temperature forecasts (Fig. 2.5). Daily minimum temperature DMO forecast errors ranged from $+1.3$°C to $+3.0$°C, indicating a warm bias for DMO overnight temperature forecasts. All post-processing methods show a significant reduction in DMO mean error for all forecast days. All methods except SNL reduce the DMO mean error to near zero for all forecast days.

The trend in forecast bias for minimum temperatures was unusual in that bias decreased from $+2.0$°C to $+1.3$°C for days one through three, then increased considerably to $+3.0$°C on day four; bias then remained nearly constant at $+2.8$°C for days five through eight. The increase in DMO minimum-temperature bias between days three and four may be due to the fact that forecast temperatures for days one, two, and the first half of day three come from the higher-resolution regional GEM model (a 48-hour forecast model), while the DMO forecast temperatures for the second half of day three through day eight come from the lower-resolution version of the GEM global model (termed the global or GLB model that runs operationally out to a lead time of at least 10 days). The model grid change between day three and day four may account for some of the sudden increase in minimum temperature bias after day three, since a large part of the bias error is attributable to poor representation of topography in the model. No such change is seen in the maximum temperature forecast bias, however.

All post-processing methods reduced the mean error in the DMO forecasts. The four moving-weighted-average methods (MA, LIN, COS^2, and BES) and the Kalman Filter method (KF) were nearly equal in keeping mean errors under 0.2°C for both maximum and minimum temperature forecasts. These methods respond rapidly to changes in the model and changes in flow regime. The seasonal method of mean error reduction (SNL) maintains the same error-correction factor throughout a complete six-month period (cold or warm season). The SNL method shows gradually increasing negative mean error (for maximum temperature forecasts), with a mean error near -0.8°C for days five through eight, though this method still greatly outperformed DMO. The difference in performance between SNL and the other methods, though slight, may be due to residual impact of model changes or flow regime changes that are not readily incorporated in this method.

Mean absolute error (MAE) was evaluated for all the post-processing methods (Fig. 2.6). All post-processing methods show significant improvement in MAE over the DMO forecasts, especially for short-range daily maximum temperature forecasts. The KF technique performs best by a slight margin. MAE was generally greater for maximum temperature forecasts than for minimum temperature forecasts, and generally increased from days one through eight. In terms of MAE, all six methods (SNL, MA, LIN, COS^2, BES, and KF) performed nearly equally well in keeping MAE between 1.5°C and 3.5°C throughout the forecast period.

To show these results more clearly, an MAE skill score, measured with DMO as the reference forecast, is shown in Fig. 2.7. The MAE skill score is adapted from Wilks (2006) and Mao et al. (1999):

$$MAE\ Skill\ Score = 1 - \frac{MAE_{PP}}{MAE_{DMO}} \tag{2.7}$$

In this formulation MAE_{PP} represents the MAE of the particular post-processing technique, as indicated in the text and in Fig. 2.7. An MAE skill score value of zero indicates no improvement in skill over the associated DMO forecast; the range from zero to one indicates improving skill, with a value of one indicating perfect forecasting skill. A negative MAE skill score indicates the post-processing technique shows less skill than the corresponding DMO forecast.

One trend visible in Fig. 2.7 is that the SNL technique shows improving skill with lead time, relative to the moving-weighted-average and KF methods. As one would expect, early in the forecast period (days one through four) the weighted average techniques hold a slight advantage over SNL because they are weighted by recent error estimates that may be affected by weather regime changes or, infrequently, model changes that the SNL technique does not incorporate. At longer lead times (days five through eight) the inclusion of recent error becomes less effective because overall random error in the forecast increases so much at the longer lead times; the statistical stability of a six-month seasonal error average shows its advantage at the longer lead times.

For some applications of temperature forecasting an error threshold is important, in that the forecasts are expected to remain within a certain critical error threshold. For example, agencies or commercial providers may have guidelines or contractual stipulations that require temperature forecast errors to not exceed a particular threshold for a certain percentage of time without incurring a penalty. The post-processing techniques evaluated here were also tested against critical error thresholds of 5°C and 10°C (see Figs. 2.8 and 2.9). All post-processing techniques show a significant reduction in errors over DMO forecasts at both thresholds, through the eight day forecast period. The KF technique indicates the best results, especially in the longer forecast range.

Daily maximum temperature errors are greater than minimum temperature errors, with forecast errors greater than 5°C occurring less than 10% of the time through day three and less than 25% of the time through day eight. Daily minimum temperature errors greater than 5°C remain below 10% through day six.

Daily minimum temperature errors greater than 10°C are rare even for DMO forecasts, though the methods presented here still achieve improvement in these forecasts. Daily maximum temperature errors greater than 10°C occur less than 1% of the time through day 5 for all methods, increasing to occurrences about 3% of the time by day eight; these methods show much improvement over DMO forecasts, which indicate 10°C errors occur 5% to 10% of the time for forecast days five through eight.

UMOS

The UMOS forecasts were available only for a subset of six stations. The weighted-average methods (listed in Table 2.1) were re-calculated for this subset of stations to allow direct comparison with the UMOS forecasts. The UMOS forecasts were the poorest amongst all methods at reducing mean error, with mean error as great as 1.8°C (Fig. 2.10).

MAE comparisons of UMOS with weighted average techniques (Fig. 2.11) indicate that UMOS forecasts fare worse than the other post-processing methods for both maximum temperature forecasts and minimum temperature forecasts. Stensrud and Yussouf (2003) found similar results with temperature-forecast errors corrected with a simple 7-day running mean that proved competitive with, or better than, MOS forecasts. Results from Stensrud and Yussouf (2005) showed that temperature-forecast errors corrected with a 12-day moving-average have lower error than comparable MOS forecasts.

2.3.2 Daily Precipitation

One of the most long-standing and challenging problems in weather forecasting is the quantitative prediction of precipitation (Mao et al., 2000; Chien and Ben, 2004; Barstad and Smith, 2005; Zhong et al., 2005), and for hydrometeorological applications, precipitation is the most important factor in watershed modeling (Beven, 2001). The challenge of precipitation forecasting is often exacerbated in regions with complex terrain (Westrick and Mass, 2001). This difficulty in forecasting precipitation is unfortunate since forecasting for water resource planning, flash floods, and glacier mass balance in mountainous terrain depends on accurate weather forecast models. Inflow forecasting, based largely on accurate precipitation forecasts from NWP models, is critical for reservoir management (Yao and Georgakakos, 2001).

There are many reasons that quantitative-precipitation forecasting and verification is much more challenging than for daily maximum or minimum temperatures, including the following:

- The basic moisture variable in NWP models is specific humidity while temperature itself is a basic variable in primitive equation models. Instantaneous precipitation rate in NWP is a parameterized variable encapsulating many complex physical processes. QPF is an integrated compilation of precipitation rate. Precipitation requires a finite spin-up time for operational NWP models that encompass a dry start.
- Precipitation is discontinuous in space and time; temperature is a continuous space-time variable.
- Precipitation exhibits a non-normal sampling distribution; maximum and minimum temperatures tend toward a normal distribution.
- Daily precipitation amounts are highly variable from one day to the next; daily temperatures less so.
- Daily temperature falls into a natural 24-hour evaluation window because of diurnal temperature trends. Cool-season stratiform precipitation stems from evolving mid-latitude storm systems that do not have such a natural 24-hour cycle.
- Precipitation occurs in different phases and types.
- Varying rates of horizontal advection and vertical fall speed occur for different sizes and composition of hydrometeors, which affects collection efficiency in rain gauge observations.
- Precipitation in NWP models is calculated as a grid-square average, while direct precipitation observation is by point (radar-derived precipitation totals are areal in nature; however, the derived values depend on subjective algorithms that are difficult to implement in mountainous terrain due to blocking of the radar beam). The difference in scale between a mesoscale model forecasting precipitation on a 10 km grid spacing and a typical precipitation gauge (1 m^2 opening) is eight orders of magnitude.
- Precipitation observation technologies are more subject to errors due to equipment limitations than temperature recording technology (e.g., tipping bucket gauges are

susceptible to upper-rain-rate limits and below-freezing temperatures; rain bucket gauges are susceptible to snow-capping; all gauges are susceptible to misrepresenting hydrometeor fall rates in strong winds).

DMO for QPF shows a slight overforecasting (wet) trend throughout the eight day forecast period, averaged over all 19 stations (see Fig. 2.12). The SNL method of DMB-correction overcompensates by reducing the DMB to less than one for all forecast days. The MA and BES methods perform equally well in reducing DMB error in the DMO forecasts. The resulting QPF-adjusted values from the MA and BES methods result in well balanced DMB values (near one) for all forecast days— the main objective of the post-processing application. A non-parametric statistical significance test was performed to confirm the results. Error bars in Fig. 2.12 represent the 95% confidence limits for SNL, MA, and BES methods. Details of the statistical significance test methodology are given in Appendix B.

Closer inspection, however, shows that the DMO forecasts have a much wider range of error determined station to station compared to SNL, MA and BES. For example, on forecast day two the DMO DMB ranges from 0.57 to 2.3, while SNL ranges from 0.36 to 1.27, MA ranges from just 0.97 to 1.27, and BES ranges from 0.92 to 1.35 (see Fig. 2.13). This diagram shows that the MA and BES methods are similar in correcting DMB values to near one, outperforming the SNL method in a station-by-station comparison. DMO precipitation forecast errors as shown are significant for certain stations in this region of complex terrain and require appropriate post-processing measures to reduce topographically forced systematic errors in the forecasts.

Examination of the output for other forecast days indicates similar results. Therefore even though the average DMB over all stations is fairly comparable for DMO compared to the other methods, either MA or BES proved a good choice for a post-processing method based on station-to-station DMB consideration.

Considering the MAEs of daily precipitation forecasts (Fig. 2.14), all the post-processing techniques show improvement over DMO throughout the eight-day forecast period. All of the techniques show similar improvement for days one through four with SNL showing slightly less MAE (better performance) on days two through four. From forecast days five through eight the SNL method clearly proves the best method (based on MAE), pointing to the value of a longer seasonal training period in improving mid-range DMO precipitation forecasts.

An MAE skill score chart (Fig. 2.15) shows this trend more clearly in a different format. The MA and BES methods, relying on just the most recent 40 days for an error correction window, show very slight MAE skill relative to DMO for the day five through day eight forecasts (the methods show less than 8% improvement in MAE skill over DMO, with no skill for day six). The long seasonal correction window employed by the SNL technique is necessary to improve the mid-range precipitation forecasts (indicating about a 12% MAE skill improvement over DMO forecasts). This MAE skill score diagram shows the results evident in Fig. 2.14 more clearly.

Daily precipitation thresholds of 10 mm and 25 mm were chosen to show the improvement of the forecasts over DMO by using the stated post-processing techniques (see Figs. 2.16 and 2.17). All techniques show slight improvement over DMO at both threshold criteria. Similar to the MAE analysis, the MA and BES methods perform best for the day one forecast, then the SNL method performs best for the remainder of the forecast days two through eight (though by a very slight margin).

2.4 Summary and Conclusions

Daily forecast values of maximum temperature, minimum temperature, and quantitative precipitation are the prime drivers of inflow forecasts for the 15 reservoirs within the area of study described in this paper, with precipitation being the most important factor. Seven statistical post-processing techniques have been compared here for maximum and minimum temperature forecasts, and three methods for 24-hour quantitative precipitation forecasts. For this study in mountainous western Canada, all the techniques for temperature post-processing showed significant improvement over direct model output. The best method was Kalman filtering, followed closely by the four 14-day moving-weighted-average methods (moving average, linear weighted average, cosine squared weighted average, and best easy systematic estimator). A method based on seasonally-averaged error characteristics also showed similar positive error-reduction results, especially in the longer forecast period days five through eight. All the methods performed better than a comparative study against updateable MOS temperature forecasts.

All three post-processing methods for 24-hour QPFs improved error characteristics over DMO forecasts. The best method for QPF, using DMB calibration to unity as a metric, was shared by the moving average method and best easy systematic estimator method (both based on a 40-day averaging period). However the seasonal method had slightly better error reduction characteristics judging by MAE and particular error thresholds in the day two through eight period.

The post-processing methods tested in this study can provide requisite error reduction in DMO for local point forecasts to aid decision makers in hydrometeorological and other economic or regulatory sectors. Specifically, water resource managers rely on weather forecasts of precipitation and temperature to drive hydrologic reservoir inflow models as a major component of their decision-making process. Decision makers rely on current, value-added weather forecasts for daily reservoir operations planning, extreme high-inflow events, and long-term water resource management. Such forecasts present added challenges in regions of complex topography where steep mountains and land-ocean boundaries increase the variability of local weather.

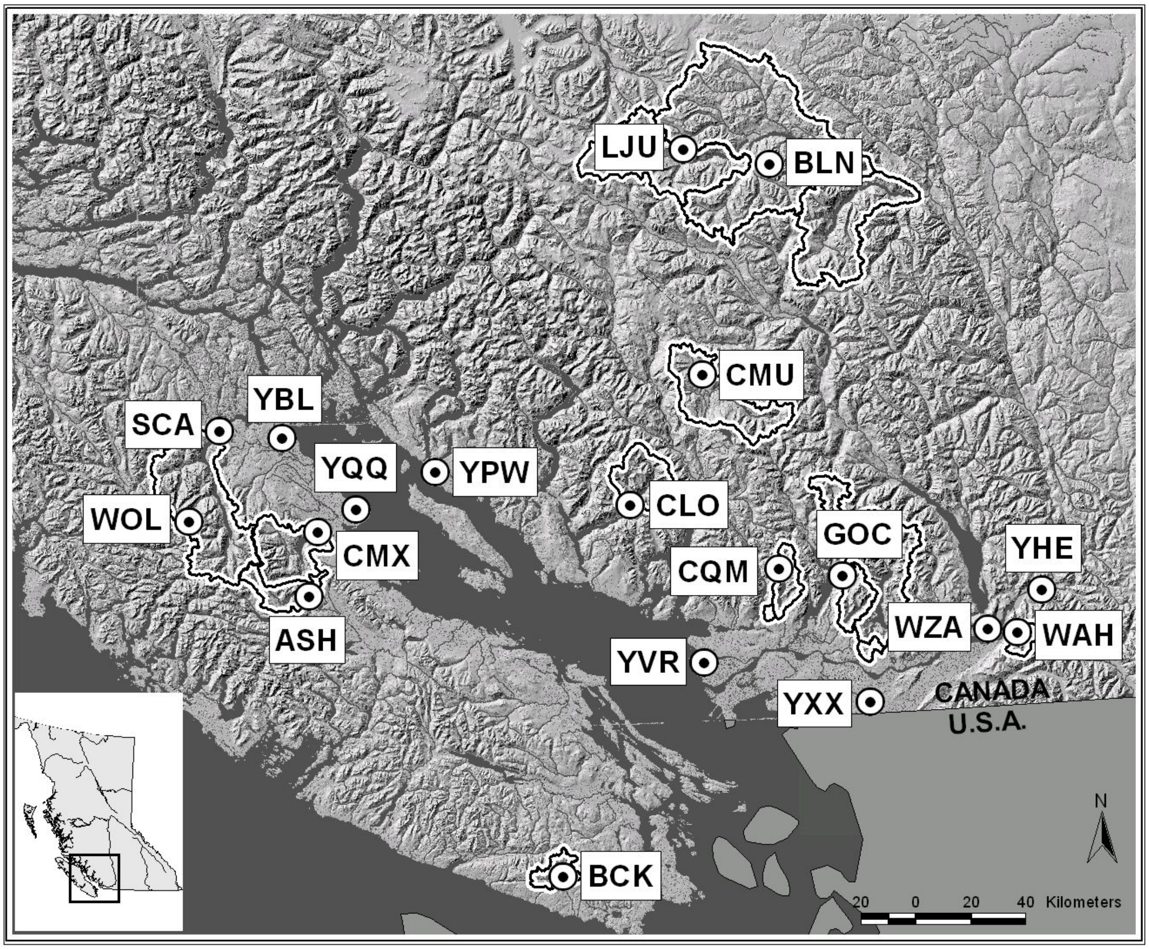

Figure 2.1: The area of study in southwestern British Columbia, Canada. 19 weather-station locations (dots) in 15 watersheds (black lines) are depicted in the figure. Elevations range from 2 m above MSL at YVR to 1920 m above MSL at BLN.

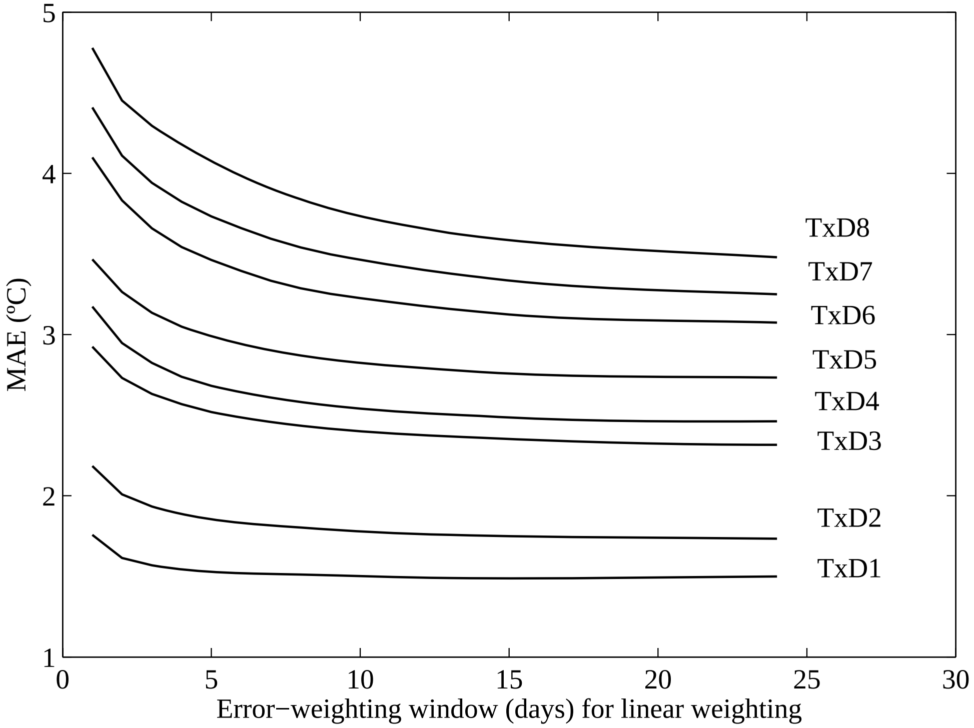

Figure 2.2: Mean absolute error (°C) vs. window length in days for LIN method of post-processing systematic error correction for daily maximum temperature. Plots are for forecast days one (bottom) through eight (top) as shown on the diagram.

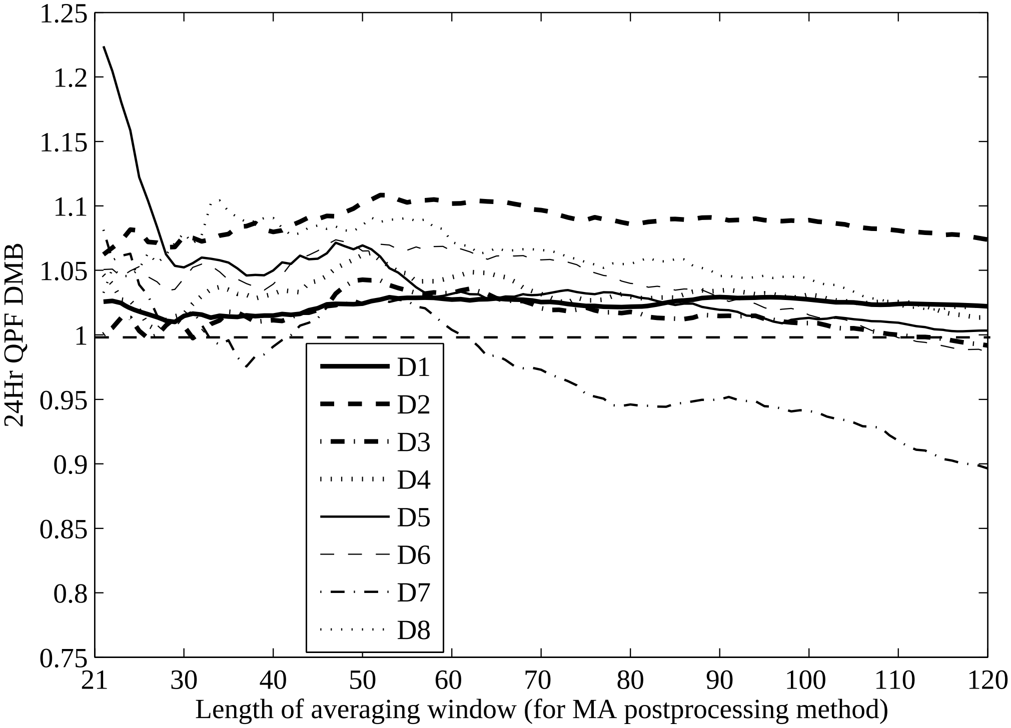

Figure 2.3: The 24 hour precipitation forecast degree of mass balance as a function of averaging window length for the MA post-processing method. Calculations for each of the forecast days one through eight are shown. DMB values near one are better since values greater than one indicate the degree of over-forecasting precipitation and values less than one indicate the degree of underforecasting precipitation.

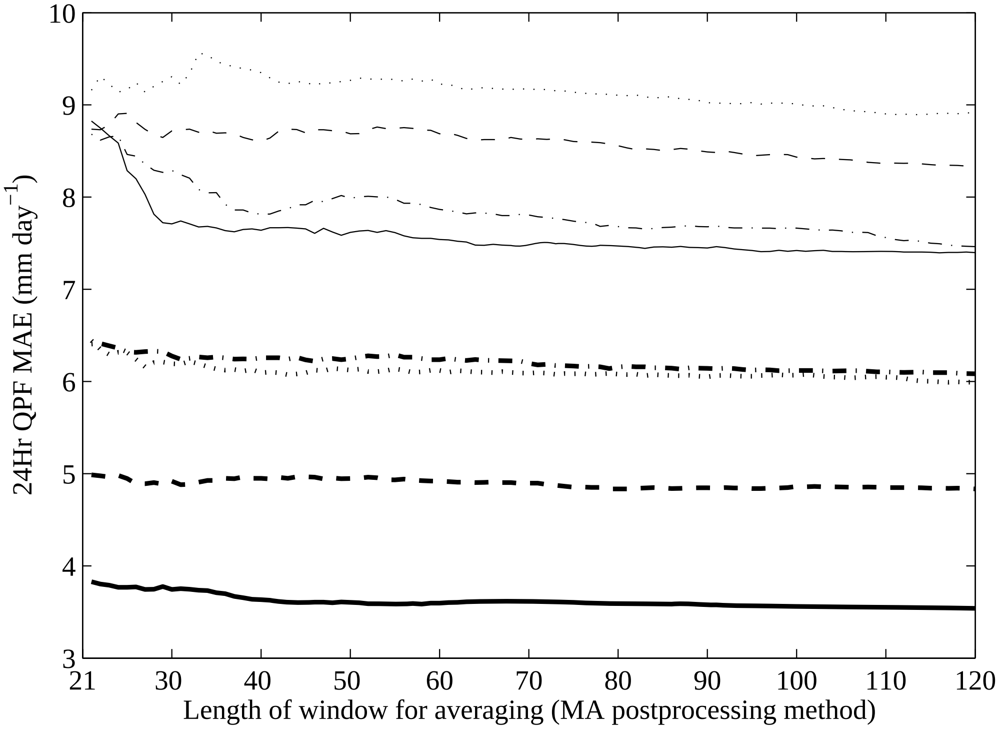

Figure 2.4: The 24 hour precipitation forecast mean absolute error (mm day^{-1}) as a function of averaging window length for the MA post-processing method. Calculations for each of the forecast days one (lowest MAE) through eight (highest MAE) are shown. Lower values are better.

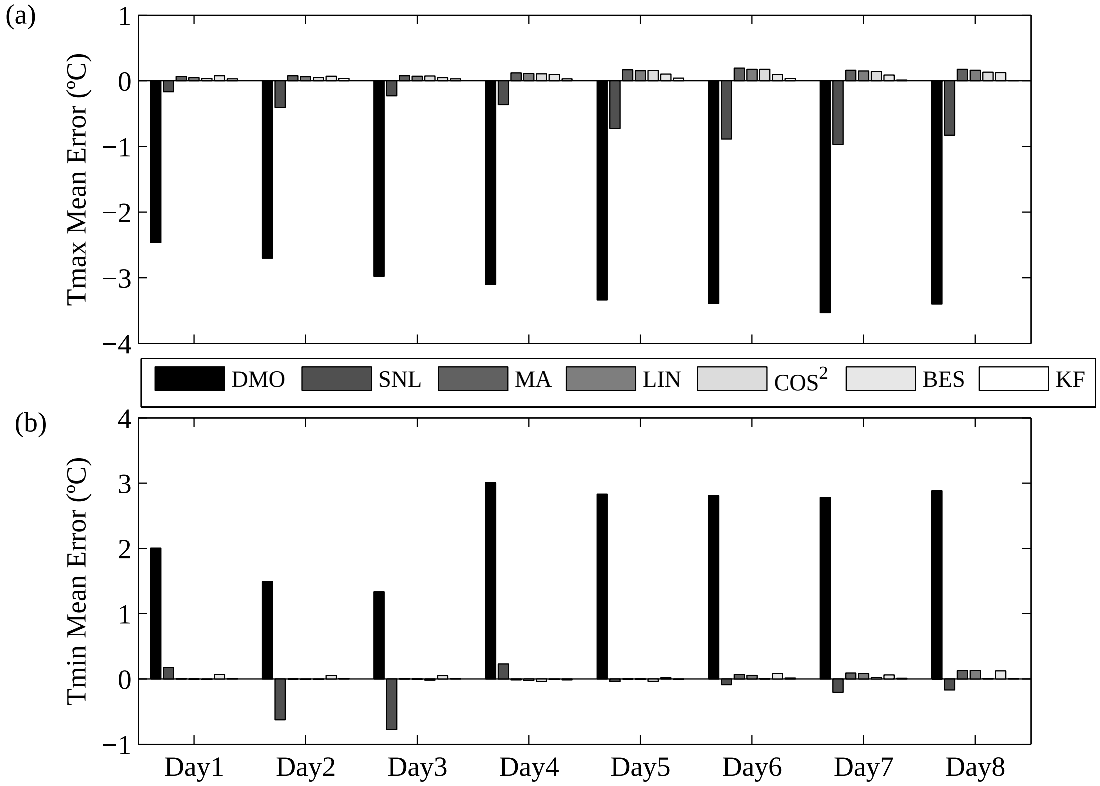

Figure 2.5: The daily maximum (a) and minimum (b) temperature mean error from the sample for forecast days one through eight. The different post-processing techniques, from left (black) to right (white) are DMO, SNL, MA, LIN, COS^2, BES, and KF, as shown in the legend. Zero mean error is best.

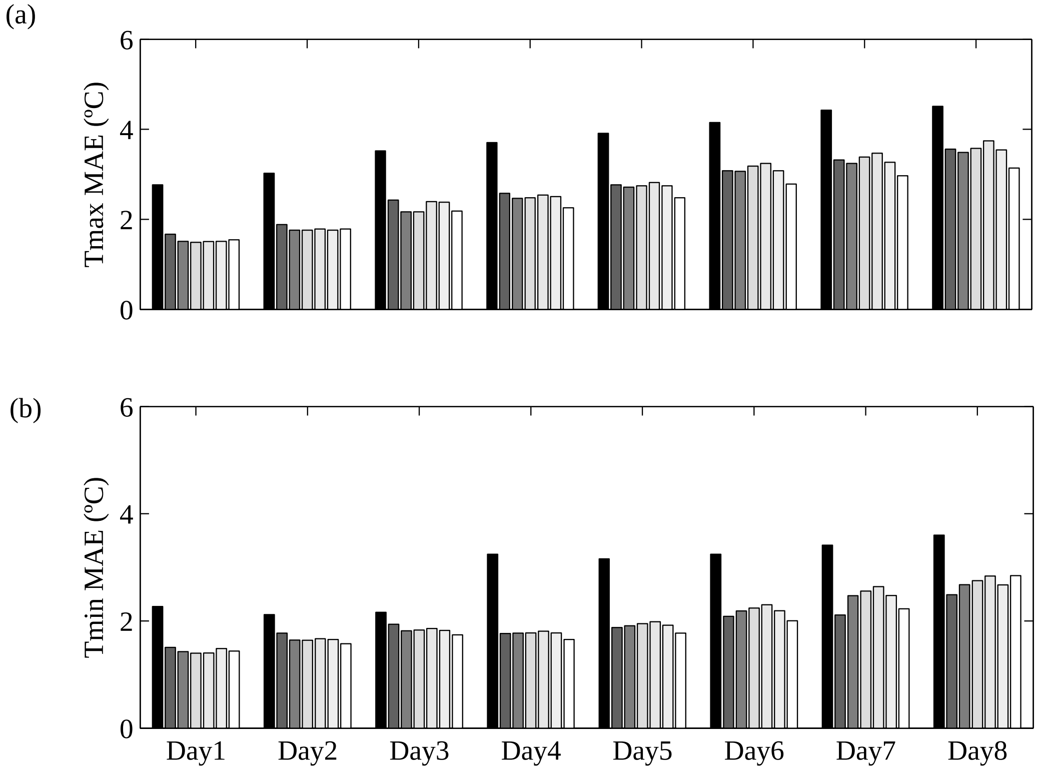

Figure 2.6: The same as Fig. 2.5 but for mean absolute error. Zero MAE is best.

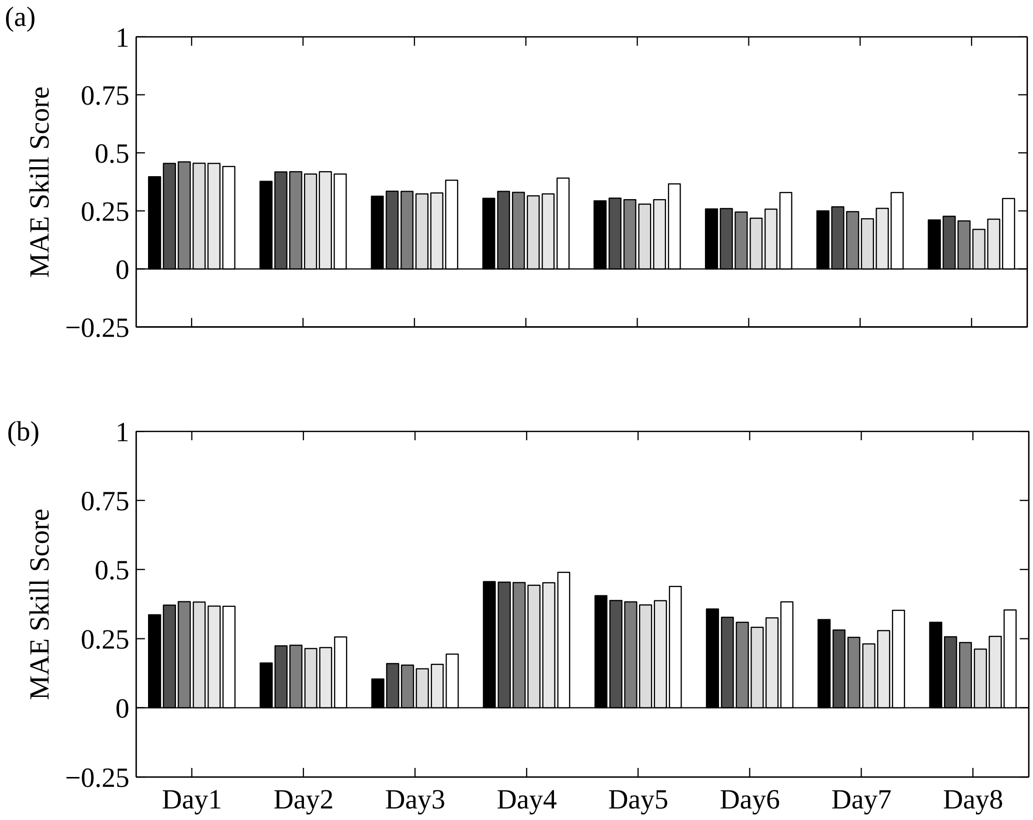

Figure 2.7: The same as Fig. 2.6 but for mean absolute error skill score (relative to DMO). Skill-score values closer to one are better.

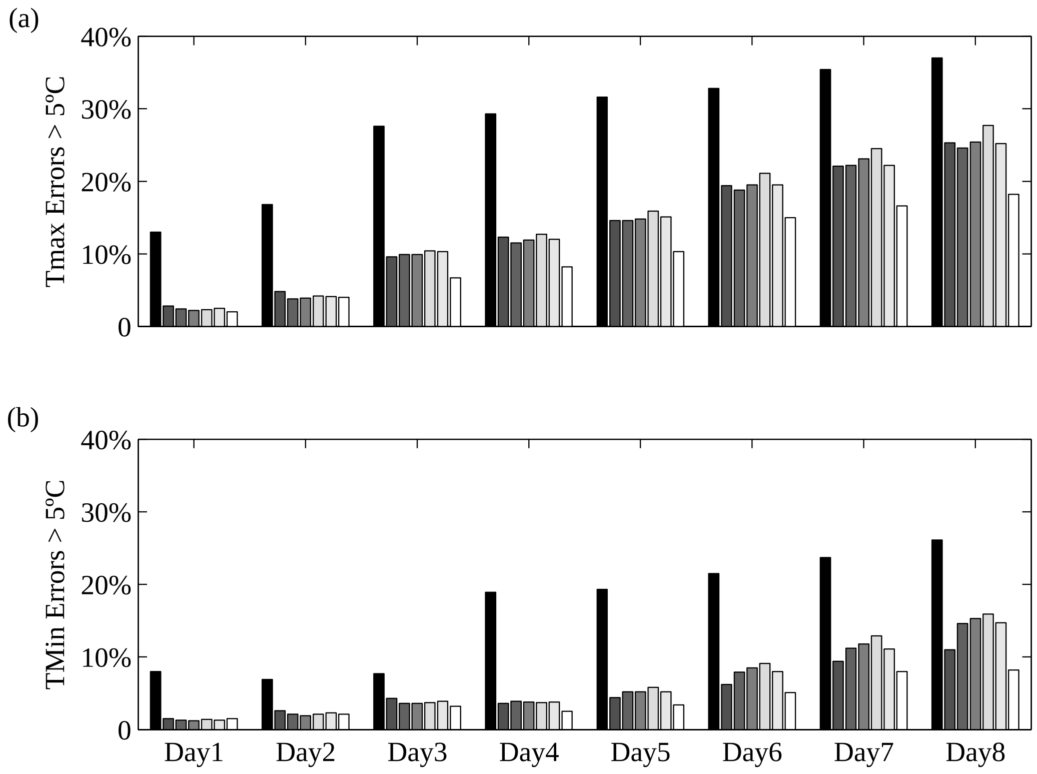

Figure 2.8: The daily maximum (a) and minimum (b) temperature absolute errors greater than 5°C (in percent) from the sample for forecast days one through eight. The different post-processing techniques, from left (black) to right (white) are DMO, SNL, MA, LIN, COS^2, BES, and KF. Values closer to zero are better.

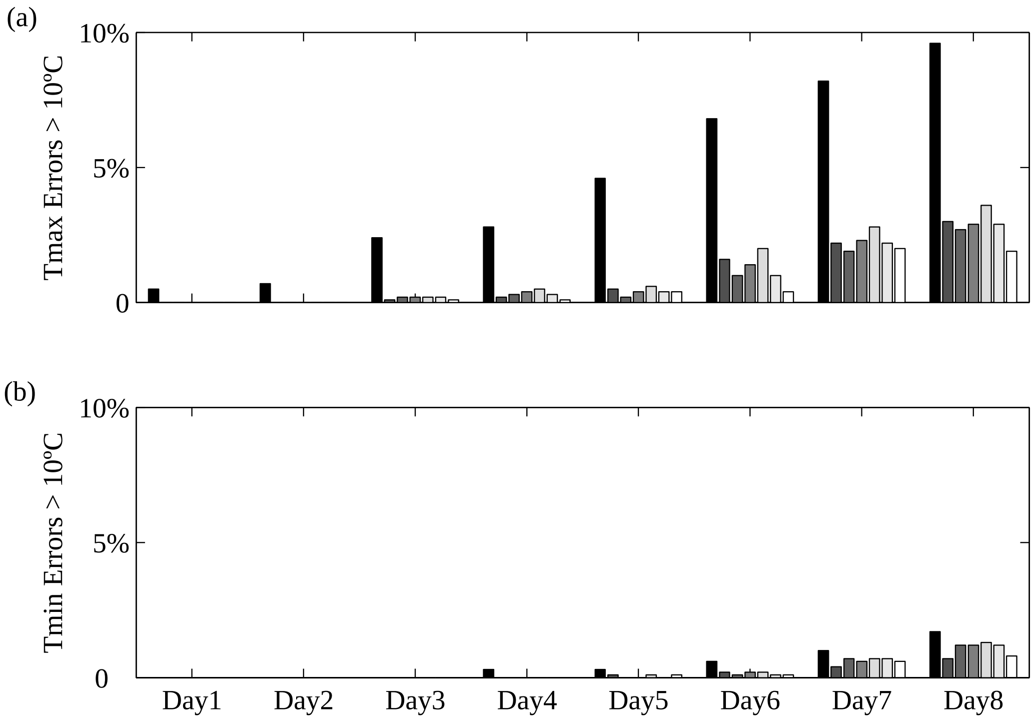

Figure 2.9: The same as Fig. 2.8 but for temperature absolute errors greater than 10°C (in percent).

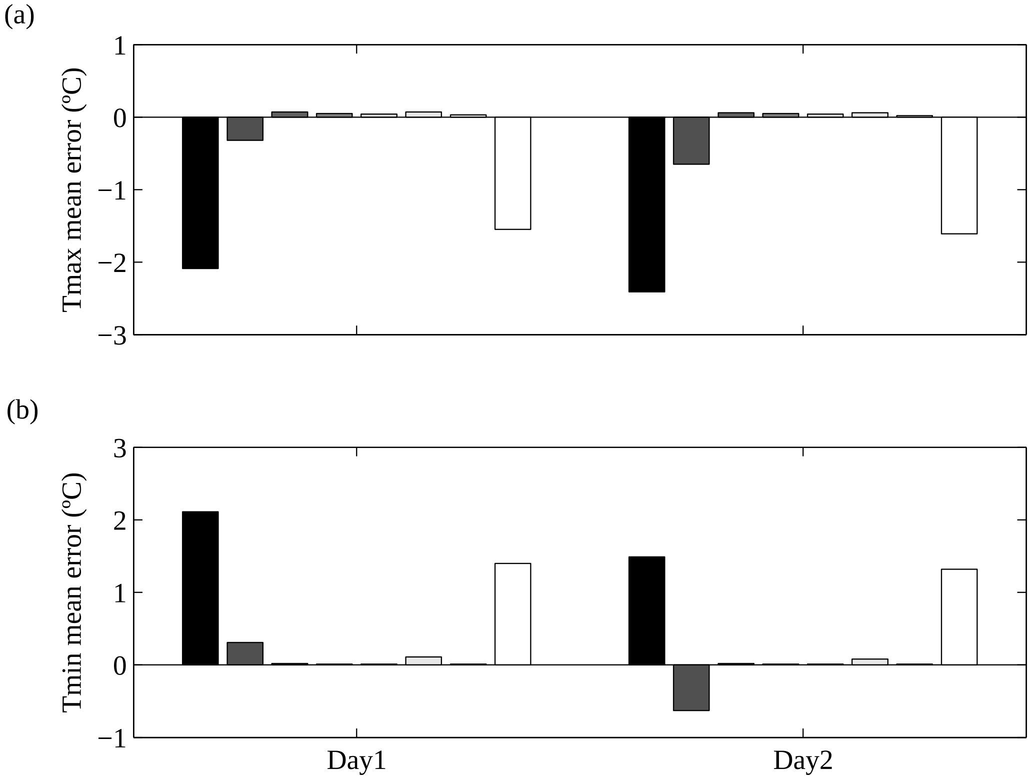

Figure 2.10: The daily maximum (a) and minimum (b) temperature mean error from the UMOS sub-sample for forecast days one and two. The different post-processing techniques, from left (black) to right (white) are DMO, SNL, MA, LIN, COS^2, BES, KF, and UMOS. Zero mean error is best.

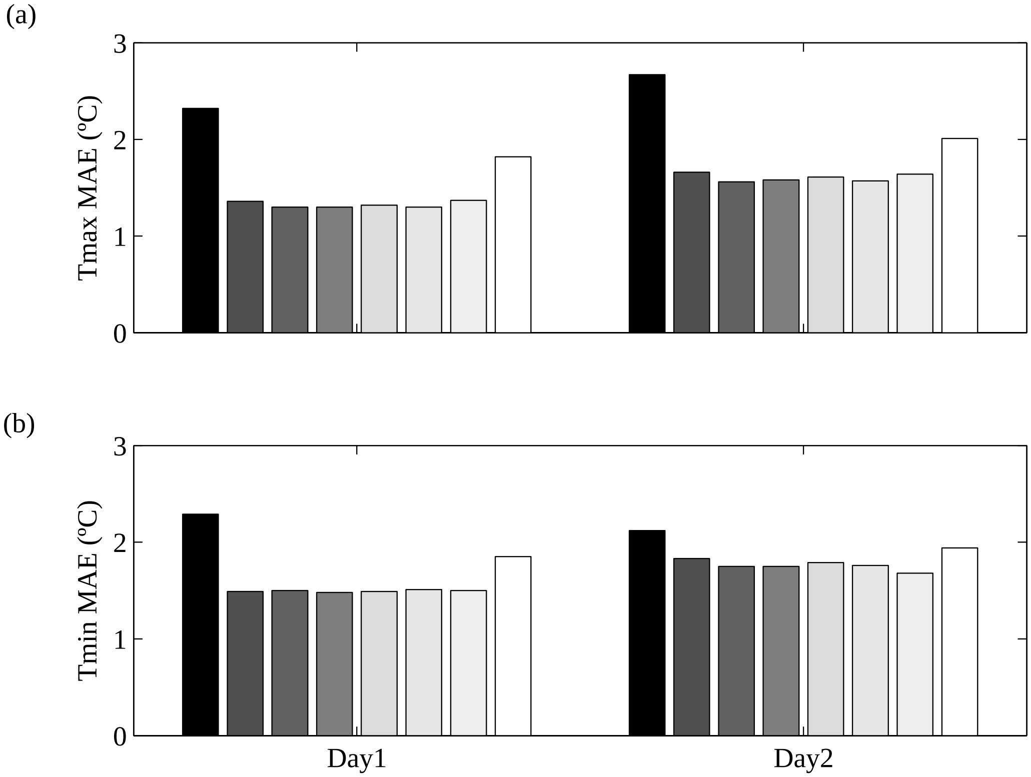

Figure 2.11: The same as Fig. 2.10 but for temperature mean absolute error. Zero MAE is best.

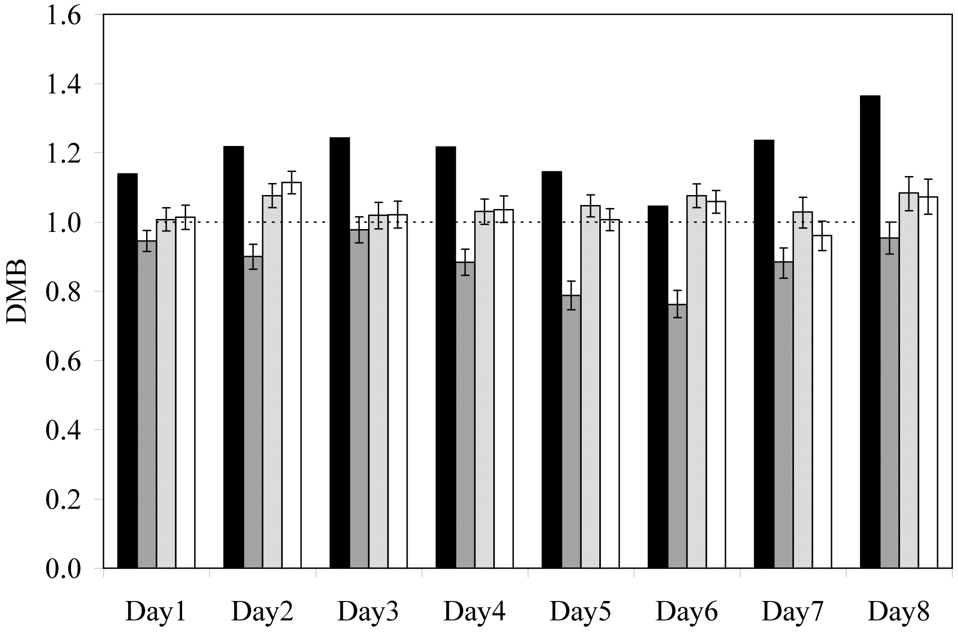

Figure 2.12: The daily 24 hour precipitation DMB (degree of mass balance) from the sample for forecast days one through eight. For precipitation the DMB is calculated as forecasts divided by observations so that balanced forecasts have a value of one. The different post-processing techniques, from left (black) to right (white) are DMO, SNL, MA, and BES. Values closer to one are better. For statistical significance analysis, error bars are included to indicate the 2.5th and 97.5th percentiles of a resampled distribution, referenced to the DMO forecasts.

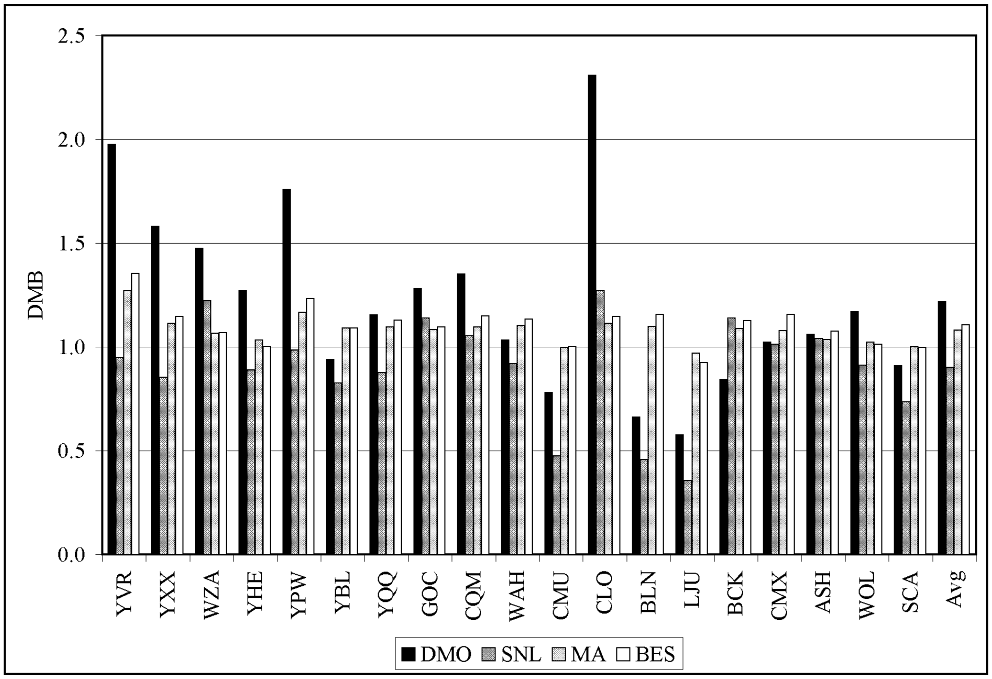

Figure 2.13: A comparison of 24 hour precipitation DMB values for each of the 19 stations for the day two forecast. The different post-processing methods are (as labelled) DMO, SNL, MA, and BES. Values closer to one are better.

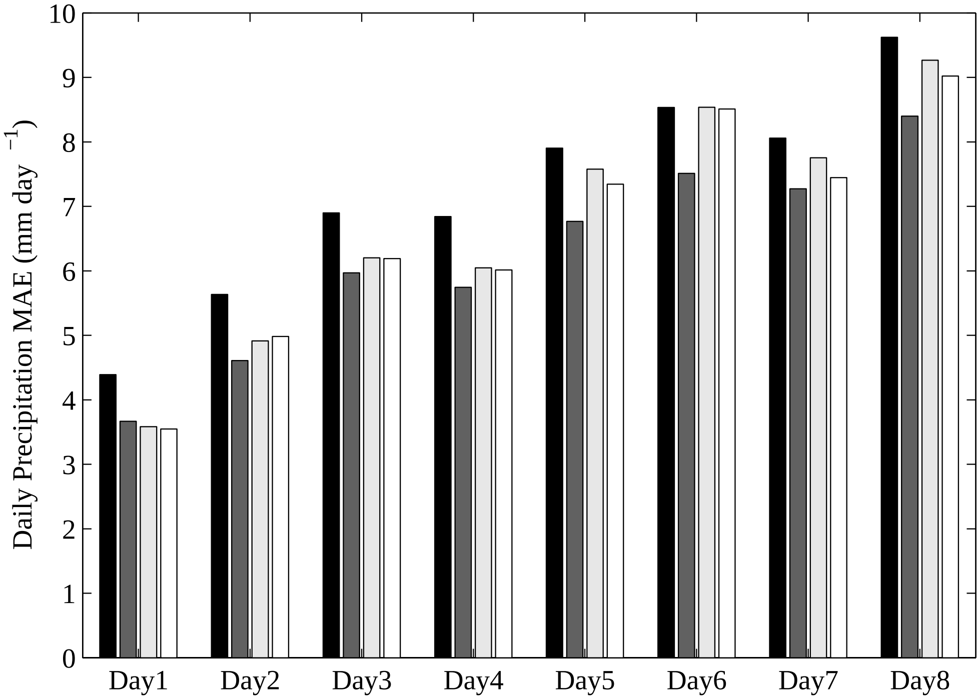

Figure 2.14: The same as Fig. 2.12 but for precipitation mean absolute error. Values closer to zero are better.

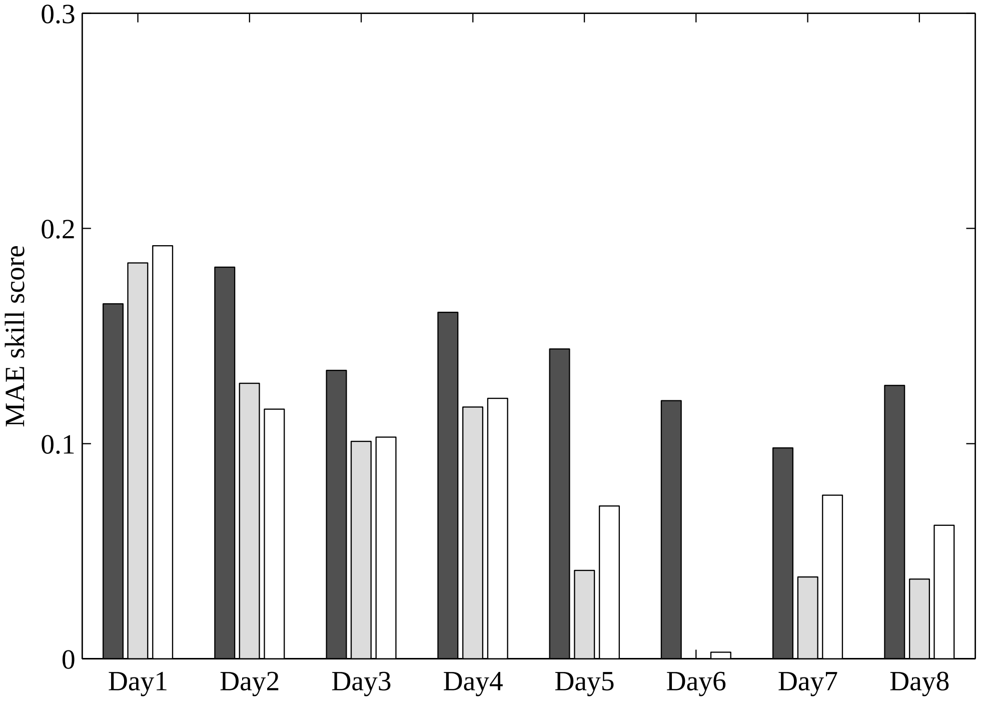

Figure 2.15: The same as Fig. 2.14 but for precipitation mean absolute error skill score (relative to DMO). Values closer to one are better.

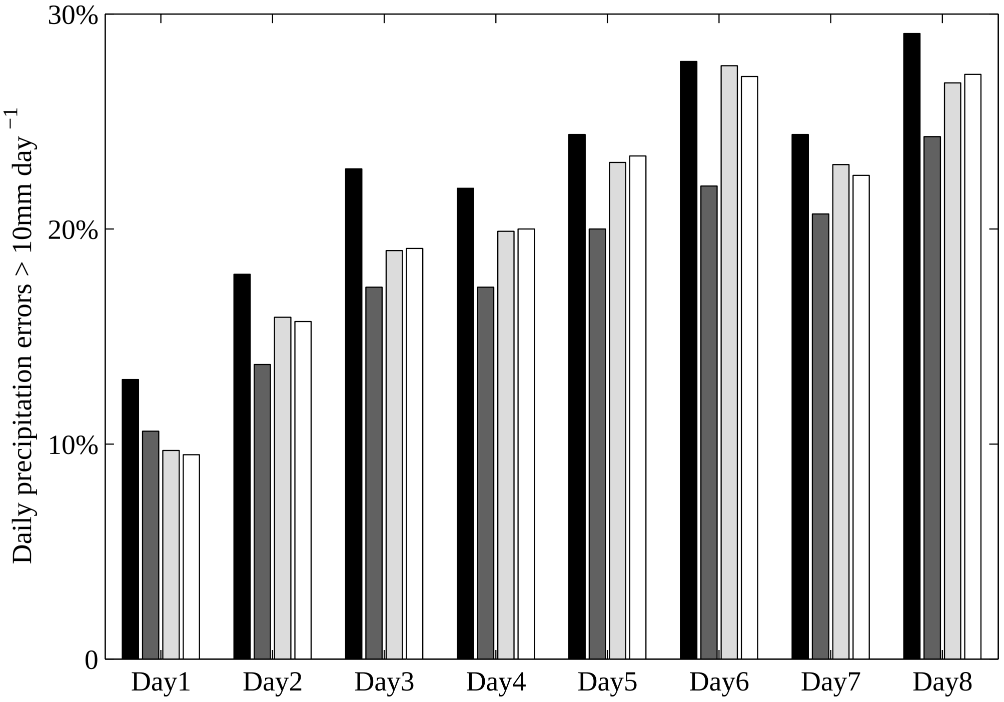

Figure 2.16: The daily 24 hour precipitation errors greater than 10mm day^{-1} (in percent) from the sample for forecast days one through eight. The different post-processing techniques, from left (black) to right (white) are DMO, SNL, MA, and BES.

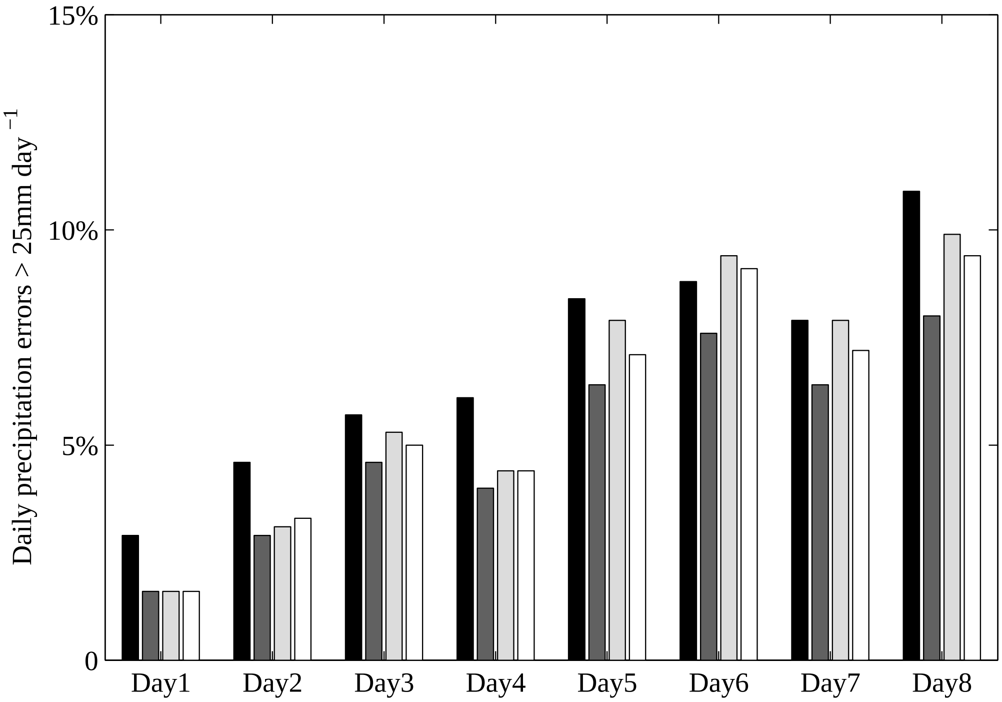

Figure 2.17: The same as Fig. 2.16 but for precipitation errors greater than 25mm day^{-1} (in percent).

Table 2.1: Table of post-processing methods and descriptions for temperature and precipitation forecasts. The post-processing methods are applied to DMO forecasts of daily maximum and minimum temperature for days one through eight. The three methods SNL, MA, and BES are evaluated for 24-hour precipitation forecasts for days one through eight.

Post-processing method	Description
SNL	Six month seasonal average
MA	Moving average
LIN	Linear weighting
COS^2	cos^2 weighting
BES	Best easy systematic estimator
KF	Kalman filter

Bibliography

Barstad, I. and R. B. Smith, 2005: Evaluation of an orographic precipitation model. *J. Hydrometeorology*, **6**, 85–99.

Beven, K. J., 2001: *Rainfall-runoff Modelling: The Primer*. J. Wiley, New York.

Chien, F. and J. J. Ben, 2004: MM5 ensemble mean precipitation forecasts in the Taiwan area for three early summer convective (Mei-Yu) seasons. *Wea. Forecasting*, **19**, 735–750.

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

Côté, J., J.-G. Desmarais, S. Gravel, A. Méthot, A. Patoine, M. Roch, and A. Staniforth, 1998a: The operational CMC-MRB global environmental multiscale (GEM) model: Part I – design considerations and formulation. *Mon. Wea. Rev.*, **126**, 1373–1395.

— 1998b: The operational CMC-MRB global environmental multiscale (GEM) model: Part II – results. *Mon. Wea. Rev.*, **126**, 1397–1418.

Delle Monache, L., T. Nipen, X. Deng, Y. Zhou, and R. Stull, 2006: Ozone ensemble forecasts: 2. A Kalman filter predictor bias correction. *J. Geophys. Res.*, **111**, D05308, doi:10.1029/2005JD006311.

Eckel, F. A. and C. F. Mass, 2005: Aspects of effective mesoscale, short-range ensemble forecasting. *Wea. Forecasting*, **20**, 328–350.

Glahn, H. and R. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.*, **11**, 1203–1211.

Grubišić, V., R. K. Vellore, and A. W. Huggins, 2005: Quantitative precipitation forecasting of wintertime storms in the Sierra Nevada: Sensitivity to the microphysical parameterization and horizontal resolution. *Mon. Wea. Rev.*, **133**, 2834–2859.

Hamill, T. M., S. L. Mullen, C. Snyder, Z. Toth, and D. P. Baumhefner, 2000: Ensemble forecasting in the short to medium range: Report from a workshop. *Bull. Amer. Meteor. Soc.*, **81**, 2653–2664.

Hart, K. A., W. J. Steenburgh, D. J. Onton, and A. J. Siffert, 2004: An evaluation of mesoscale-model-based output statistics (MOS) during the 2002 Olympic and Paralympic games. *Wea. Forecasting*, **19**, 200–218.

Homleid, M., 1995: Diurnal corrections of short-term surface temperature forecasts using the Kalman Filter. *Wea. Forecasting*, **10**, 689–707.

Jacks, E., J. B. Bower, V. J. Dagostaro, J. P. Dallavalle, M. C. Erickson, and J. C. Su, 1990: New NGM-based MOS guidance for maximum/minimum temperature, probability of precipitation, cloud amount, and surface wind. *Wea. Forecasting*, **5**, 128–138.

Jones, M. S., B. A. Colle, and J. S. Tongue, 2007: Evaluation of a mesoscale short-range ensemble forecast system over the northeast United States. *Wea. Forecasting*, **22**, 36–55.

Kalnay, E., 2003: *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, New York.

Klein, W. H., B. M. Lewis, and I. Enger, 1959: Objective prediction of five-day mean temperatures during winter. *J. Atmos. Sci.*, **16**, 672–682.

Majewski, D., 1997: Operational regional prediction. *Meteor. Atmos. Phys.*, **63**, 89–104.

Mao, Q., R. T. McNider, S. F. Mueller, and H. H. Juang, 1999: On optimal model output calibration algorithm suitable for objective temperature forecasting. *Wea. Forecasting*, **14**, 190–202.

Mao, Q., S. F. Mueller, and H. H. Juang, 2000: Quantitative precipitation forecasting for the Tennessee and Cumberland river watersheds using the NCEP regional spectral model. *Wea. Forecasting*, **15**, 29–45.

Roeger, C., R. Stull, D. McClung, J. Hacker, X. Deng, and H. Modzelewski, 2003: Verification of mesoscale numerical weather forecasts in mountainous terrain for application to avalanche prediction. *Wea. Forecasting*, **18**, 1140–1160.

Stensrud, D. J. and J. A. Skindlov, 1996: Gridpoint predictions of high temperature from a mesoscale model. *Wea. Forecasting*, **11**, 103–110.

Stensrud, D. J. and N. Yussouf, 2003: Short-range ensemble predictions of 2-m temperature and dewpoint temperature over New England. *Mon. Wea. Rev.*, **131**, 2510–2524.

— 2005: Bias-corrected short-range ensemble forecasts of near surface variables. *Meteorol. Appl.*, **12**, 217–230.

Westrick, K. J. and C. F. Mass, 2001: An evaluation of a high-resolution hydrometeorological modeling system for prediction of a cool-season flood event in a coastal mountainous watershed. *J. Hydrometeorology*, **2**, 161–180.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*. Academic Press, San Diego, 2nd edition.

Wilson, L. J. and M. Vallée, 2002: The Canadian updateable model output statistics (UMOS) system: Design and development tests. *Wea. Forecasting*, **17**, 206–222.

— 2003: The Canadian updateable model output statistics (UMOS) system: Validation against perfect prog. *Wea. Forecasting*, **18**, 288–302.

Woodcock, F. and C. Engel, 2005: Operational consensus forecasts. *Wea. Forecasting*, **20**, 101–111.

Yao, H. and A. Georgakakos, 2001: Assessment of Folsom Lake response to historical and potential future climate scenarios 2. Reservoir management. *J. Hydrol.*, **249**, 176–196.

Zhong, S., H.-J. In, X. Bian, J. Charney, W. Heilman, and B. Potter, 2005: Evaluation of real-time high-resolution MM5 predictions over the Great Lakes region. *Wea. Forecasting*, **20**, 63–81.

Chapter 3

Short-Range Ensemble Forecasts in Complex Terrain. Part I: Meteorological Evaluation[4]

3.1 Introduction

The effect of horizontal resolution and ensemble size on a regional short-range ensemble forecast (SREF) system is assessed for probabilistic forecasts of 24-h accumulated precipitation in complex terrain. Model error has a larger impact on surface variables in the short range (Stensrud et al., 2000) than on free atmosphere variables in the mid range; therefore model error, being a large source of forecast uncertainty in short-range forecasts, must be accounted for to maximize SREF utility (Eckel and Mass, 2005; Jones et al., 2007), particularly for mesoscale sensible weather elements. Therefore a multi-model SREF is used here rather than a multi-initial-condition SREF.

In a recent SREF study of the surface elements mean sea-level pressure (MSLP), temperature, and wind in the mountainous U.S. Pacific Northwest, Eckel and Mass (2005) found that a multi-model system outperformed a system that used variations of a single model. Eckel and Mass (2005) also found that an ensemble of unequally-likely members can be skillful as long as each member occasionally performs well, and that the inclusion of finer-resolution models led to greater ensemble spread as smaller scales of atmospheric motion were modeled.

One of the keys to skillful quantitative precipitation forecasts (QPF) in complex terrain is high-resolution modeling, to capture the orographic and variable-surface flux components of precipitation distribution. Employing very high-resolution models is crucial to capture variability at small scales (Hamill et al., 2000; Eckel and Mass, 2005).

[4] A version of this chapter has been accepted for publication. McCollor, D. and R. Stull, 2008: Hydrometeorological Short-Range Ensemble Forecasts in Complex Terrain. Part I: Meteorological Evaluation. *Wea. Forecasting.*

In a study of a SREF system in the cool season over complex terrain by Eckel and Mass (2005), the impact of model error remained significant even where forecast uncertainty was largely driven by synoptic-scale errors originating from analysis uncertainty. In complex terrain, many mesoscale weather phenomena, particularly cool season precipitation, are driven by the interaction between the synoptic-scale flow and underlying mesoscale topographic irregularities and boundaries.

A summary of results from Du et al. (1997), who analyzed a 25-member ensemble of QPFs, includes the finding that 90% of the improvement found in using an ensemble average was obtainable using an ensemble size as small as 8-10 members. Further results from that paper report that an ensemble QPF from an 80 km grid model is more accurate than a single forecast from a 40 km grid model, and that SREF techniques can provide increased accuracy in QPF even without further improvement in the forecast model system. Wandishin et al. (2001) found that even ensemble configurations with as few as five members can significantly outperform a higher-resolution deterministic forecast. An analysis of a global ensemble prediction system (EPS) for 24-h QPF (Mullen and Buizza, 2002) found that coarser-resolution, larger-member ensembles can outperform higher-resolution, smaller-member ensembles in terms of the ability to predict rare precipitation events.

This paper addresses the issue of SREFs for hydrometeorologic applications in complex terrain. The principal input into a hydrologic model, which produces forecasts of river stages or reservoir inflow, is the precipitation over the associated watershed (Krzysztofowicz et al., 1993). The objective of the paper is to compare the skill of 24-h accumulated QPF from various SREF configurations of three independent mesoscale NWP models run as multiple-resolution nested systems.

Section 3.2 describes the location, the models used in the SREF, and the dataset of verifying observations. Section 3.3 introduces the verification metrics for probabilistic forecasts, and section 3.4 gives the results. Conclusions are summarized in section 3.5.

3.2 Methodology

3.2.1 Location and Observation Data

Southwestern British Columbia is a region of complex topography characterized by high mountains, glaciers, fjords, and land/ocean boundaries. The region, lying between 48° and 51° North latitude on the west coast of North America, is subject to land-falling Pacific cyclones and frontal systems, especially during the cool season months from October through

March.

Precipitation data for this study were collected from a meso-network of 27 gauges within 15 small watersheds in southwestern British Columbia (see Fig. 3.1 for a reference map). Twenty-four of the stations are part of the hydrometric data collection program in support of reservoir operations for BC Hydro Corporation. The other three stations are operated by the Meteorological Service of Canada. Observations consist of 24-h accumulation of precipitation of all types (where solid and mixed precipitation are measured in liquid form) based on the 12 UTC (Universal Coordinated Time, Z) synoptic hour.

Forecast-system performance can be verified against model-based gridded analyses or against actual point observations. Gridded analyses offer more data for investigation through widespread coverage. However gridded analyses are subject to model-based errors and smoothing errors, both of major concern in regions of complex terrain. Even though point observations are subject to instrumentation error and site-representativeness error, it is preferable to verify forecasts against observations as opposed to model analyses (Wilson, 2000; Hou et al., 2001).

The observation stations used in this study range in elevation from 2 m AMSL at coastal locations to 1969 m AMSL in high mountainous watersheds. Because of the complex terrain in the study region and the fact that hydrologic models are often calibrated with station precipitation records, it was important in this study to verify forecasts against actual observations.

More information about the stations used in this study is shown in the summary hypsometric curve provided in Fig. 3.2 (watershed elevation area bands from all 15 watersheds are combined in this summary hypsometric curve). 67% of the stations are below 500 m elevation (the MSL to 500 m elevation band represents 13% of the area of the watersheds). 18% of the stations are between 500 m and 1500 m elevation, representing 44% of the area of the watersheds. The highest 15% of the stations, located between 1500 m and 1969 m, represent 19% of the area of the watersheds. Elevations above 2000 m, representing the highest 24% of the area of the watersheds, remain ungauged, with the highest peak in the watersheds at 2600 m.

Also shown in Fig. 3.2 is a cumulative total precipitation curve, and average daily station precipitation from each station. This information reflects the nature of the non-uniform and rugged terrain, in that higher elevation stations can be further from the ocean and in the rain-shadow of other mountains, and may lie within mountain ranges more parallel with the predominant flow, therefore do not receive the highest precipitation amounts. Lower elevation stations, because of their open exposure to landfalling Pacific frontal systems and orientation on mountain ranges perpendicular to the moist onshore flow, receive relatively high precipitation amounts.

3.2.2 Numerical Models and Study Dates

The Geophysical Disaster Computational Fluid Dynamics Centre (GDCFDC), in the Department of Earth and Ocean Sciences at the University of British Columbia in Vancouver, runs a real-time suite of three independent nested limited-area high-resolution mesoscale models over the region of interest. For the study presented here, the coarse-resolution (108 km horizontal grid spacing) outer nests of these models were all initialized using the National Centers for Environmental Prediction (NCEP) North American Mesoscale (NAM) model (previously Eta model) at 90 km grid spacing. Time-varying lateral boundary conditions were also extracted from the NAM forecasts.

Ensemble forecasts of 24-h accumulated total precipitation were available from an archive of the three real-time limited-area model (LAM) runs over two consecutive wet seasons, October 2003 through March 2004 and October 2004 through March 2005. The Mesoscale Compressible Community (MC2) model is a fully compressible, semi-implicit, semi-Lagrangian, non-hydrostatic mesoscale model (Benoit et al., 1997). One-way nesting is applied to produce model output at horizontal grid spacings of 108, 36, 12, 4, and 2 km. The Penn State/NCAR Mesoscale model (MM5) is a fully compressible, non-hydrostatic, sigma-coordinate model designed to simulate and predict mesoscale and regional-scale atmospheric circulations (Grell et al., 1994). The MM5 is run for the same five grids, but with 2-way nesting. The Weather Research and Forecast (WRF) model (Skamarock et al., 2005) is a non-hydrostatic mesoscale model, run for three grids (108, 36, and 12 km) with two-way nesting applied.

The LAM gridded precipitation forecast values were interpolated to the precipitation-gauge locations using a 2D cubic spline. The interpolation uses a 16-point (4 by 4) stencil, where the interpolated point is always inside the center square of the stencil.

It is important to incorporate uniquely different models in a SREF design, as opposed to a varied-model technique (a varied-model technique incorporates differing combinations of model physics and sub-grid scale parameterization methods in a single model), to improve ensemble spread characteristics. Eckel and Mass (2005) found that even with extensive variations of a single model, a multi-model SREF design vastly outperformed a varied-model SREF design in representing model uncertainty.

Precipitation duration from individual extratropical weather systems in the cool season can vary from hours to days, so there is no obvious choice for choosing a particular time-window for accumulation totals. Precipitation gauges available for this study record precipitation accumulation at one-hour intervals. However, the choice of verifying 24-h precipitation accumulation totals was made for two reasons. First, water managers often require daily timestep precipitation forecasts for input into hydrologic models. Second, lengthening the accumulation period beyond 1-, 3-, 6-, or 12-h accumulations monotonically and significantly enables clearer evaluation of model QPF skill (Wandishin et al., 2001; Stensrud and Yussouf, 2007) as opposed to timing differences. For example, if a model forecast erred by delaying the start of a 12-h rainstorm by 4 h, then this rainfall would be missed entirely by some of the 1- and 3-h accumulation forecasts. The 6-h and 12-h forecasts would capture part of the storm, while only the 24-h forecast would be considered entirely accurate. Of course there will be instances in which timing of storms will affect even the 24-h accumulation window, but this impact is statistically lessened, and QPF assessment improved, by using the 24-h forecast over a large number of days.

The suite of nested LAMs is designed so that the lower resolution models (108 km) encompass a large enough domain that weather features advected in from the boundaries are incorporated throughout the 60-h forecast window of the GDCFDC models. Successive higher-resolution models (36 km and 12 km) incorporate smaller domains, eventually focussing on the region of interest with the highest resolution models (4 km and 2 km).

In building a SREF for hydrometeorologic forecasting of precipitation, the question arises as to how to best incorporate these forecast models of multiple resolution into a viable EPS. A viable EPS requires enough ensemble members to reflect and characterize the spread of observations so that the probability distribution of the forecasts is, ideally, indistinguishable from the probability distribution of the observations.

Including lower-resolution models that incorporate a much larger domain may benefit the SREF design by increasing spread in the ensemble. Nutter et al. (2004) showed that ensembles that use a larger domain produce greater spread than smaller-domain ensembles because the use of periodically updated, coarse lateral boundary conditions can filter out short waves and reduce the amplitude of nonstationary waves entering from the larger-domain model. Alternately, previous research (Mass et al., 2002) has indicated that forecast skill for precipitation in mountainous terrain improves as grid spacing decreases from 36 to 12 km. But Mass et al. (2002) found that verification scores generally degrade as resolution is further increased from 12 to 4 km as overprediction develops over windward slopes and crests of terrain (though for heavy precipitation amounts on windward slopes, the transition from 12 to 4 km enhances the forecast accuracy).

Lower-resolution models may not have the topographic resolution necessary to adequately resolve terrain-influenced precipitation patterns, but they have proven to contribute to model diversity and hence improve ensemble spread. Extremely high-resolution models may not incorporate accurate enough sub-grid scale parameterizations to characterize individual cloud precipitation processes. In addition, computer-resource limitations prohibit running a large-member, high-resolution-only multi-model system.

Since nested-grid LAMs by design produce forecasts across a widely-varying scale of resolution, this paper investigates the benefits of including multiple resolution LAMs to increase ensemble size (hence improve spread) in a SREF system. This approach of including multi-resolution nested LAMs in a SREF is supported by Wandishin et al. (2001), who concluded that including less skillful members can add skill in a mixed-ensemble system.

Each of the three models was initialized at 00Z and run for 60 hours. The nesting nature of the models means that the 108 km run begins at the 00Z initialization time. For the model with one-way nesting, the next 36 km nested run begins 3 hours later, at 03Z; the next 12 km nested run begins at 06Z; and the 4 km nested run begins at 09Z. Since all models are initialized with a dry start, there will be spin-up errors in the early portion of each run, which lead to a dry bias in the model forecasts by not including precipitation that is occurring at initialization time.

To partially ameliorate the spin-up problem, and because the one-way nesting model relies on a lagged forecast start time for the higher-resolution nested runs as outlined in the previous paragraph, we chose to ignore the first 12 h (00Z to 12Z) of the forecast, and use forecast hours 12 to 36 as “day one”. The forecasts from the period T+36 to T+60 of each model ensemble member were extracted to form the “day-two” forecast.

Not all forecast days were available due to model-run failures; the 2 km runs were especially susceptible to missing forecast days, which led to the decision not to include 2 km forecast runs in the ensemble with the present operational configuration. There remained 5027 forecast-observation pairs for the day-one forecast and 4737 forecast-observation pairs for the day-two forecast in this study.

Different configurations of a SREF system were constructed from the dataset to measure the influence of including lower-resolution and/or higher-resolution LAM ensemble members on the performance of the EPS. The full 11 member suite of ensemble members (all11) included MC2 (108 km, 36 km, 12 km, 4 km), MM5 (108 km, 36 km, 12 km, 4 km), and WRF (108 km, 36 km, 12 km) models. A suite of eight ensemble members (hires8) including all but the lowest (108 km) resolution models was constructed to measure the performance of the higher-resolution models only. Another suite of six medium-resolution ensembles (mres6) included the 36 km and 12 km LAMs only. The mres6 configuration was included to examine the effect of excluding both the lowest (108 km) and highest (4 km) resolution models. The final suite (vhires5) included the five very highest-resolution LAMs only (12 km and 4 km). Table 3.1 provides a summary of the SREF configurations.

3.3 Verification Procedures

A single verification score is generally inadequate for evaluating all of the desired information about the performance of a SREF system (Murphy and Winkler, 1987; Murphy, 1991). Different measures, emphasizing different aspects and attributes of forecast performance, should be employed to assess the statistical reliability, resolution, and discrimination of an EPS. A standardized set of evaluation methods, scores, and diagrams to interpret short to medium-range ensemble forecasts is provided in Hamill et al. (2000) and is incorporated here to assess a SREF system for hydrometeorological forecasts. We use degree of mass balance (DMB), mean error (ME), mean absolute error (MAE), mean square error (MSE), root mean square error (RMSE), Pearson correlation (r), linear error in probability space (LEPS), Brier skill score (BSS) including relative reliability and resolution terms, relative operating characteristic (ROC) curves derived from hit rate (H) and false alarm rate (F), and rank histograms. A description of these evaluation methods is provided in Appendix C. The reader is referred to Toth et al. (2003) or Wilks (2006) for comprehensive descriptions of the verification measures.

It is impossible to objectively assess the probabilistic quality of an individual ensemble forecast. Ensemble forecast systems must be verified over many cases. As a result, the scoring metrics described in Appendix C are susceptible to several sources of noise (Hamill et al., 2000): improper estimates of probabilities arising from small-sized ensembles; insufficient variety and number of cases leading to statistical misrepresentation; and imperfect observations making true forecast evaluation impossible.

A complete evaluation of ensemble mesoscale forecast models involves many different verification methods and scores, observation criteria, interpretation of results, and definition of user needs. In fact, a succinct and satisfactory approach to ensemble mesoscale verification remains elusive (Davis and Carr, 2000; Mass et al., 2002). Traditional objective approaches based on verification at fixed observing locations are greatly influenced by timing and spatial errors, as well as deficiencies of the observing network. The value of high-resolution numerical forecasts is clearly user-dependent, and one verification system cannot address the needs of all forecast users. The verification in this study is designed for a hydrometeorological user needing to forecast water inflow into hydro-electric reservoirs.

3.4 Results

First, the performance of individual ensemble members is described and compared against other members. Reliability and resolution of EPS performance are found next using the Brier skill score. Probabilistic values are assigned to forecasts based on the number of ensembles exceeding a particular threshold, and are used to create ROC curves. Finally, we describe the extent that individual ensemble members are equally likely to verify by employing rank histogram diagrams.

3.4.1 Error Performance of Individual Ensemble Members

An important criterion in an EPS is that both forecasts and observations have similar distributions. Precipitation follows a distinctly asymmetrical distribution that is highly skewed to the right (tail towards high precipitation values) with a physical limit of zero precipitation on the left. Such highly non-normal distributions require analysis methods that are both resistant (not overly influenced by outliers) and robust (independent of the underlying distribution). In this study we employ box-and-whisker diagrams and linear error in probability space, LEPS, (see Appendix C) to meet these requirements.

There are 11 available ensemble members, as defined in Table 3.1. Model and observation box-and-whisker percentiles are seen in Fig. 3.3 for day-one forecasts and in Fig. 3.4 for day-two forecasts. The median of all observations (rain-days and non-rain-days combined) is approximately 2 mm day^{-1} indicating that half of the station-days in the cool-season sample experienced precipitation of at least 2 mm day^{-1}. The boxplots show percentiles only above the median because detail in this portion of the distribution is of most hydrometeorological importance. The ordinate is logarithmic, to adequately display the full range of values.

The ensemble members show that the distributions for all models resemble the distribution for the observations at the 75 percentile, especially for the day-one forecast. As the model distributions move into the climatologically rarer precipitation events (the 90 percentile, 95 percentile, 99 percentile, and the maximum event) a trend emerges showing that as the resolution increases the model distributions more closely resemble the distribution of the observations. The two highest-resolution model distributions, MM5-4 km and MC2-4 km, very closely resemble the distribution of the observations for the day-one forecast. The three 12 km model distributions show reasonable agreement with the distribution of the observations, except for slight differences for the very rare events distinguished by the 99 percentile and maximum value categories. The lower-resolution model distributions, 108 km and 36 km, do not capture the rare events (90th percentile and higher) as well as the high-resolution model distributions for both day-one and day-two forecasts. For the day-two forecast, the lowest-resolution model distributions diverge somewhat from the distribution of the observations even at the 75 percentile.

A LEPS skill score with the climatological median as a reference is shown in Fig. 3.5 for the day-one forecasts and in Fig. 3.6 for the day-two forecasts. In terms of LEPS the 108 km models perform the poorest, followed in general by the 36 km models for both forecast days. The 12 km and 4 km models exhibit the best results.

In addition to LEPS, standard summary measures to examine error performance of individual ensemble members in this study have been calculated, and include degree of mass balance (DMB), Pearson product moment correlation, mean error, mean absolute error, and root mean square error (see Appendix C for a summary of the equations used for the meteorological statistical analysis of the individual ensemble members).

DMB values indicate that all 108 km and 36 km resolution models exhibit an under-forecast tendency (DMB < 1) averaged over all stations in this region of complex terrain, for both forecast days one and two (see Fig. 3.7). DMB values exhibit a clear trend of improving mass balance for the higher-resolution models, so that the MM5-12 km, WRF-12 km, MM5-4 km, and MC2-4 km all exhibit nearly perfect mass balance in the day-one forecasts. Day-two forecasts follow a similar trend as their day-one counterparts in terms of improving mass balance with the higher-resolution models, however even the highest-resolution models show a slight tendency to under-forecast in the day-two timeframe.

Mean error improves dramatically with higher resolution. The finest-resolution models exhibit negligble mean error (less than 0.5 mm day^{-1}) for the day-one forecasts. The mean error shows the same trend towards underforecasting for the coarser-resolution models as exhibited by DMB. Day-one forecasts show improvement in mean error over the day-two forecasts.

The mean absolute error of the forecasts is highly consistent across models and resolutions, between 5 mm day^{-1} and 7.5 mm day^{-1}. Day-two forecasts show consistently higher mean absolute error than corresponding day-one forecasts.

The mean error and mean absolute error results can be explained in light of the definition of mean error as an overall measure of systematic bias in the forecasts (Jolliffe and Stephenson, 2003). NWP model forecasts contain systematic biases due to imperfect model physics, initial conditions, and boundary conditions (Cheng and Steenburgh, 2007). Also, NWP models prognose fields at gridpoint locations that represent a volume average, yet observations are taken at a specific location. Cheng and Steenburgh (2007) also made the point that bias results from differences between model and actual station elevation, plus nearby topographic influences too small to be represented in the model. These systematic biases are model-resolution dependent as different-resolution versions of a model will represent local topography differently. Higher-resolution models should represent local topography better and have less bias than lower-resolution models. This is evident in the results shown in Fig. 3.7. Systematic bias can be addressed and corrected for via post-processing of the forecasts (Cheng and Steenburgh, 2007; Yussouf and Stensrud, 2007; McCollor and Stull, 2008a).

Random error has positive variance but zero mean. Therefore random error is a component of mean absolute error, but not mean error. Random error is a function of model physics and parameterization schemes and is driven by how well the forecast model performs in representing the atmosphere. Random error cannot be improved by post-processing model forecasts. The core model must be improved to diminish random error. Therefore no significant variations in mean absolute error is evident among the different resolutions in our results. Random error can be reduced by making ensemble averages.

The root mean square error of the forecasts is highly consistent across all models and resolutions, between 10 mm day^{-1} and 15 mm day^{-1}. Day-two forecasts exhibit consistently higher rms error than corresponding day-one forecasts.

Correlation between forecasts and observations is highly consistent among models, resolutions, and forecast days. The Pearson product moment correlation is between 0.65 and 0.75 for all models for both day-one and day-two forecasts. The finer-resolution models (12 km and 4 km) tend to show a higher correlation between forecasts and observations than the coarser-resolution models (108 km and 36 km). Day-one forecasts exhibit slightly higher correlation than the corresponding day-two forecasts.

Summarizing the above results, an intercomparison among all models and resolutions indicates a definite trend toward higher similarity among forecast-observation distributions for the finer-resolution models. A trend toward better mass balance and improving mean error statistics are found in the finer-resolution models. Mean absolute error, root mean square error, and correlation show little variation among different-resolution models, but do show improving error statistics moving from the day-two forecast to the day-one forecast.

The differences between the coarser-resolution and finer-resolution models varies across different metrics, indicating that each of the 11 members have the potential for contributing to improving the quality of an ensemble prediction system for 24-h precipitation. As described in section 3.2, the full 11-member ensemble was subdivided into smaller ensembles to test the contribution of the coarse-resolution and fine-resolution members relative to the full suite of ensemble members. The following sections employ verification measures designed for an EPS to examine the contribution of the coarse-resolution and fine-resolution members in building an operational SREF system.

3.4.2 Brier Skill Score: Resolution and Reliability

Brier skill score results (BSS_{∞}, see Appendix C), incorporating an adjustment factor so that the different SREF systems composed of different size ensembles can be effectively compared, and including relative reliability and relative resolution components, are shown in Fig. 3.8 (5 mm day^{-1} precipitation threshold), Fig. 3.9 (10 mm day^{-1} threshold), Fig. 3.10 (25 mm day^{-1} threshold), and Fig. 3.11 (50 mm day^{-1} threshold). The reference system for the Brier skill score is the climatological forecast in which the probability of the event is derived from the average of all observations in the sample.

The full 11-member ensemble shows the best Brier skill score (BSS_{∞}, consisting of best reliability and resolution components) for all four thresholds for the day-one forecast and for the higher precipitation thresholds for the day-two forecast. For the 5 mm day^{-1} threshold day-two forecast, the 11-member ensemble shows similar skill to the fine-resolution 8-member and medium-resolution 6-member SREF systems.

The very fine-resolution 5-member SREF performs the worst in terms of Brier skill score, for all thresholds and for both forecast days, while the fine-resolution 8-member SREF is generally second best and the medium-resolution 6-member SREF is generally third best. Thus, increasing membership size, even at the expense of including coarser-resolution members, is advantageous in terms of reliability and resolution across a wide range of 24-h precipitation thresholds, for this case study. All SREF configurations show skill relative to climatology.

The reliability diagrams for the full 11 member ensemble for different 24-h precipitation thresholds are shown in Fig. 3.12 for the day-one forecasts, and in Fig. 3.13 for the day-two forecasts. The SREF exhibits reasonably good reliability as indicated by the reliability diagrams. The reader is referred to Hamill (1997) or Wilks (2006) for a display of a variety of hypothetical reliability diagrams useful in gauging different degrees of reliability and sharpness. The forecasted probabilities match the observed relative frequencies fairly well, though the 5 mm day^{-1} and 10 mm day^{-1} thresholds show slight overforecasting at the higher probabilities for both forecast days one and two. The zig-zag to the distribution of points in the rarer 25 mm day^{-1} and 50 mm day^{-1} event threshold reliability diagrams indicates that these results exhibit signs of small sample size. Reliability diagrams for higher thresholds, 75 mm day^{-1} and 100 mm day^{-1} (not shown), exhibit a highly erratic pattern due to the small statistical sample realized at these high precipitation thresholds.

Reliability diagrams for the other SREF systems are not shown, but relative reliability for the full 11-member system (Figs. 3.8 through 3.11) tends to be better than for the other lower-member systems.

The SREF system exhibits a high degree of sharpness, especially at the higher precipitation thresholds of 10 mm day^{-1}, 25 mm day^{-1}, and 50 mm day^{-1} (see insert histograms in Figs. 3.12 and 3.13). The high degree of sharpness indicated by the forecast systems ensures that the forecasts do not cluster near the climatological mean.

3.4.3 ROC Curves

The transformed ROC curves for the four SREF systems for different 24-h precipitation thresholds are shown in Figs. 3.14 and 3.15. ROC area skill scores greater than 0.4 indicate reasonable and useful discriminating ability; values 0.6 or higher indicate good discriminating ability; and ROC area skill scores of 0.8 indicate excellent discriminating ability (Buizza et al., 1999; Bright et al., 2005; Yuan et al., 2005; Stensrud and Yussouf, 2007).

Table 3.2 examines the ability of the EPS to discriminate among events by comparing ROC area skill scores for all four different SREF configurations in detail. All four SREF configurations are able to discriminate between events by exhibiting ROC skill scores greater than 0.6 for each threshold, for both forecast days. ROC skill scores increase as the precipitation threshold increases for the day-one forecasts, with little difference among the different SREF configurations. ROC skill scores are lower for the day-two forecasts than for the day-one forecasts, indicating a deterioration in discriminating ability with forecast period. Discriminating ability, as measured by ROC skill scores for the day-two forecast, also increases with precipitation threshold up to the 25 mm day^{-1} threshold. The forecasts for the 50 mm day^{-1} threshold for the day-two forecast indicate a marked decrease in discriminating ability compared to the 25 mm day^{-1} threshold, the opposite trend to the day-one forecasts (also evident from the ROC curves for the day-two forecasts in Figs. 3.14 and Fig. 3.15).

The plotted points of the actual ROC curves tend to cluster more toward the lower left-hand corner of the curve for higher precipitation thresholds, therefore, the full transformed curves (see Appendix C) contain more area and reflect higher skill scores. This does not hold true for the day-two 50 mm day^{-1} threshold forecasts as the SREF systems lose discriminating ability at these high-threshold, rare events in the longer-range day-two timeframe. These results are consistent with an analysis of transformed ROC curves performed by Wilson (2000).

3.4.4 Equal Likelihood

The rank-histogram (Talagrand) diagrams for the day-one and day-two forecasts for the full 11-member ensemble are shown in Fig. 3.16. The measure of the flatness of the histograms, δ, for the different SREF configurations is given in Table 3.3.

The rank histogram for the full 11 member ensemble exhibits a U shape, indicating the spread of the ensemble is under-dispersive for 24-h precipitation forecasts. Eliminating ensemble members, as is done in the other three SREF configurations, can only exacerbate the EPS problem of under-dispersal. Table 3.3 shows that $\delta >> 1$ for all SREF configurations for both forecast days one and two, indicating a lack of spread among all ensembles. The normalized Reliability Index (RI), shown in Table 3.4, also shows that the full 11 member ensemble provides the best reliability and that reliability does, in fact, improve with increasing ensemble size.

Poor ensemble spread (underdispersion) is a well known limitation of an EPS, especially SREFs that forecast precipitation and other surface weather elements. However, ensembles still possess skill equal to or better than a single deterministic forecast run at higher resolution (Wandishin et al., 2001). Eckel and Mass (2005) showed, through a display of rank histograms, that model diversity plays a much greater role for surface sensible weather elements (wind speed and temperature) than for other synoptic variables (50 kPa height and MSLP).

3.5 Conclusions

Forecasts of 24-h precipitation generated by a short-range ensemble forecast system have been analyzed to study the usefulness of such forecasts in a hydrologic ensemble prediction system. Precipitation is the key forecast parameter in hydrologic modeling for the 15 small watersheds included in this study.

In this study, three different limited-area mesoscale models (MC2, MM5, and WRF) are run in a nested fashion, generating forecasts at up to four different telescoping resolutions (horizontal grid spacings of 108 km, 36 km, 12 km, and 4 km). This paper addresses the question of whether it is beneficial to include the low (108 km) and high (4 km) resolution members to increase the size of the ensemble, knowing the limitations of mesoscale models at very low (poor topographic representation) and very high (inadequate sub-grid scale parameterization) resolutions.

The results of this study for precipitation in complex terrain show that it is best to include all available resolution models in the SREF configuration, even at the expense of including coarser-resolution (e.g. 108 km) members that exhibit higher errors and lower correlation with observations. The finer-resolution members (e.g. 4 km) exhibited the lowest error characteristics of the suite of individual members, so should definitely be included in the ensemble. Therefore it appears that the noted weakness of very high resolution models (inadequate sub-grid scale parameterization) does not preclude incorporating the 4 km resolution models into the SREF examined in this research. It is more likely that inadequate sub-grid scale parameterization may begin to adversely affect SREF members with resolutions on the order of 1 km. More research is needed to evaluate the effectiveness of such extremely high-resolution LAMs.

The benefit of including more members includes improved reliability and resolution, as proven by analyzing components of the Brier skill score. Including more coarse-resolution members does not inhibit discrimination, as shown by analyzing ROC scores. In terms of equal likelihood, the full 11-member ensemble exhibits under-dispersion of the ensemble (a trait common amongst other ensemble prediction systems), as did all SREF configurations examined in this paper. However the full 11-member ensemble did exhibit the greatest reliability relative to the other tested configurations.

Day-two forecasts exhibit a definite deterioration in error characteristics, Brier skill score, and discrimination over day-one forecasts. However the variations are generally slight and day-two forecasts remain skillful in terms of the stated probabilistic verification measures.

The evidence gleaned from previous SREF studies and the results presented here indicate that SREF performance, in terms of skill and dispersion, may be increased by incorporating certain model run procedures. The spin-up problem of mesoscale models can be avoided or lessened by increasing the data assimilation/pre-forecast period for the higher-resolution model runs. Including the very-fine (2 km or finer) resolution nested model runs would generate increased ensemble dispersion at the smaller scales, but the error characteristics of these extremely fine-resolution models need to be examined, as noted above.

Precipitation forecasting on the scale of small pluvial watersheds can be a daunting task, especially on the west coast of North America where the upwind Pacific data void introduces greater uncertainty (Hacker et al., 2003; McMurdie and Mass, 2004; Spagnol et al., 2004) into the initial conditions of numerical weather forecasts than for regions in the center and on the eastern side of the continent. Forecast precipitation amounts over small watersheds depend largely on the storm track, and a shift of a few tens of kilometers to the north or south can make a significant difference between a “correct forecast” and a “miss” or “false alarm” of high precipitation that leads to high inflow (Weber et al., 2006).

Support for the development of ensemble hydrometeorological prediction systems is a predecessor to the goal of a skillful coupled precipitation-runoff ensemble forecast system. This paper is part I of a two-part evaluation study that recent investigations (Yuan et al., 2005) confirm must be performed for atmospheric variables that have historically not been scrutinized. Part II concludes this series of papers by incorporating user-dependent economic analyses to ensure a full suite of forecast evaluation methodologies are addressed in determining the usefulness of the SREF system.

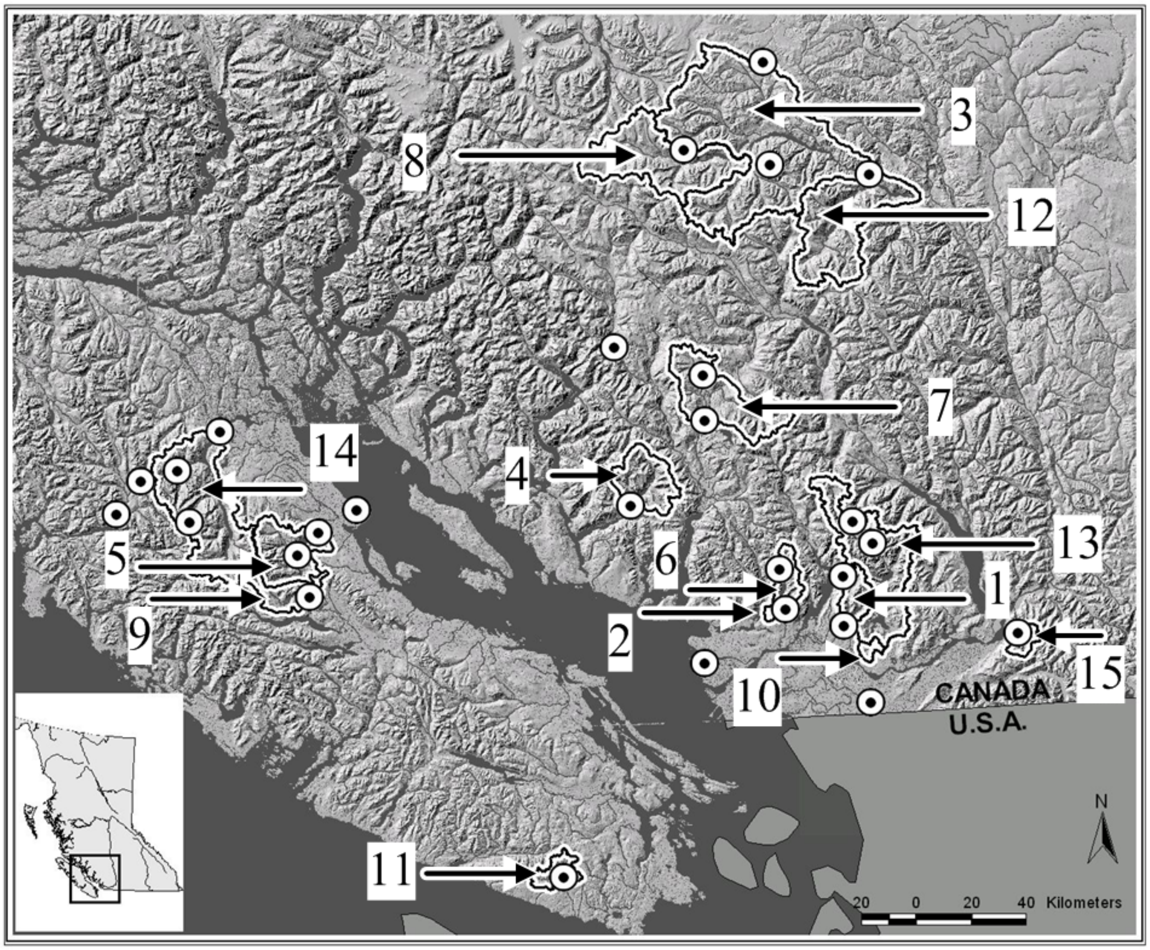

Figure 3.1: Reference map for the area of study in southwestern British Columbia, Canada. 27 weather-station locations (dots) representing 15 watersheds (black lines) are depicted in the figure. Station elevation ranges from 2 m to 1969 m above MSL. The reservoirs are numbered and listed by name in Table 2 of the companion paper (McCollor and Stull, 2008b).

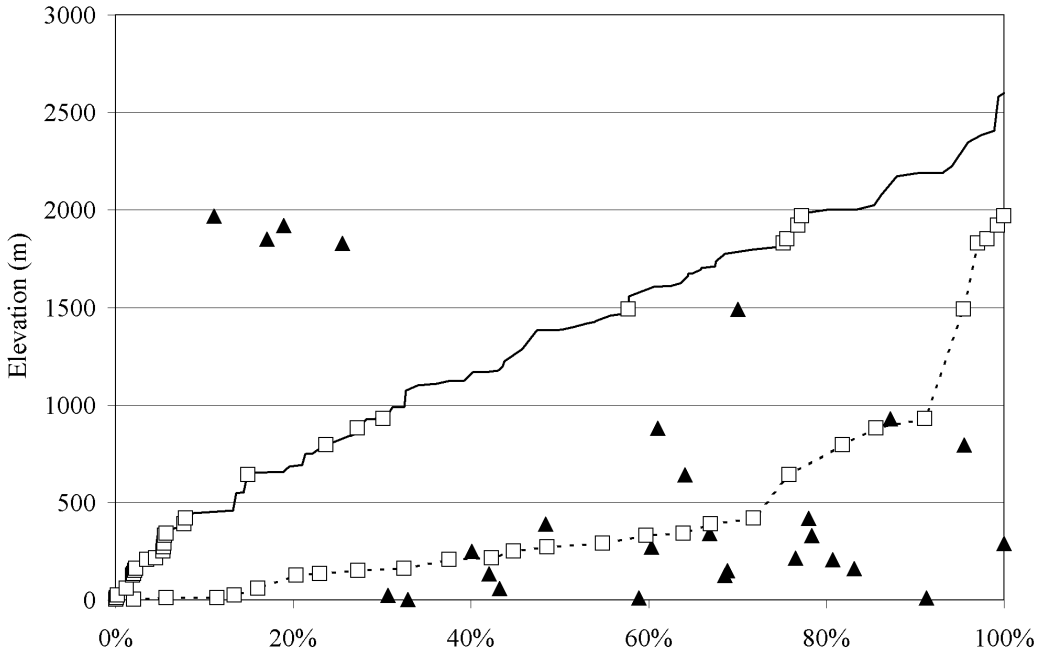

Figure 3.2: Summary hypsometric curve (solid line) and cumulative total precipitation curve (dashed line). Individual stations (squares) are indicated on the curves. The total area of the 15 watersheds in this study is 7899 square km. The total cumulative precipitation for all 27 stations over the study period is 87,324 mm. The triangles indicate the average daily station precipitation as a percentage of the station with the highest average daily precipitation (15 mm day^{-1}).

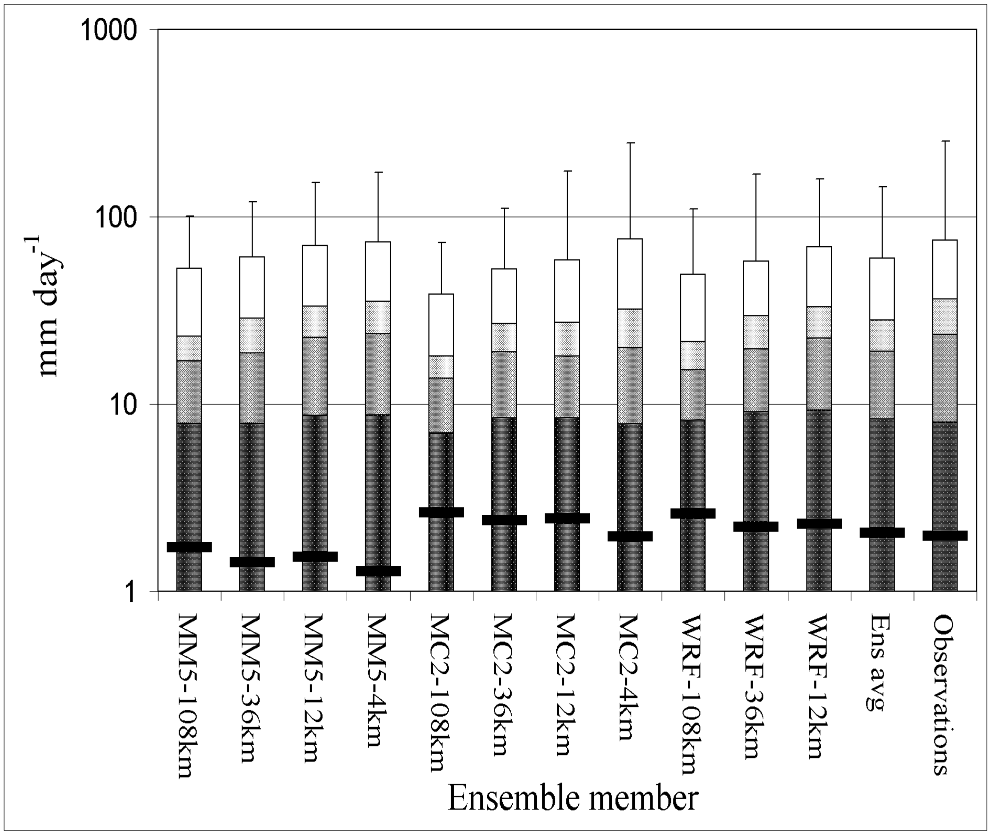

Figure 3.3: Box and whisker plots for 11 members of the ensemble for the day-one forecast of 24-hour precipitation amount. The ensemble average and associated observations are also included. The black horizontal bar indicates the median of each sample (rain-days and non-rain-days combined). Increasing percentiles (lighter colored bars) indicate 75, 90, 95, and 99 percentiles. The topmost "whisker" indicates the maximum value. The plots indicate that the distributions of the higher resolution models more closely resemble the distribution of the observations for events greater than 2 mm day^{-1}.

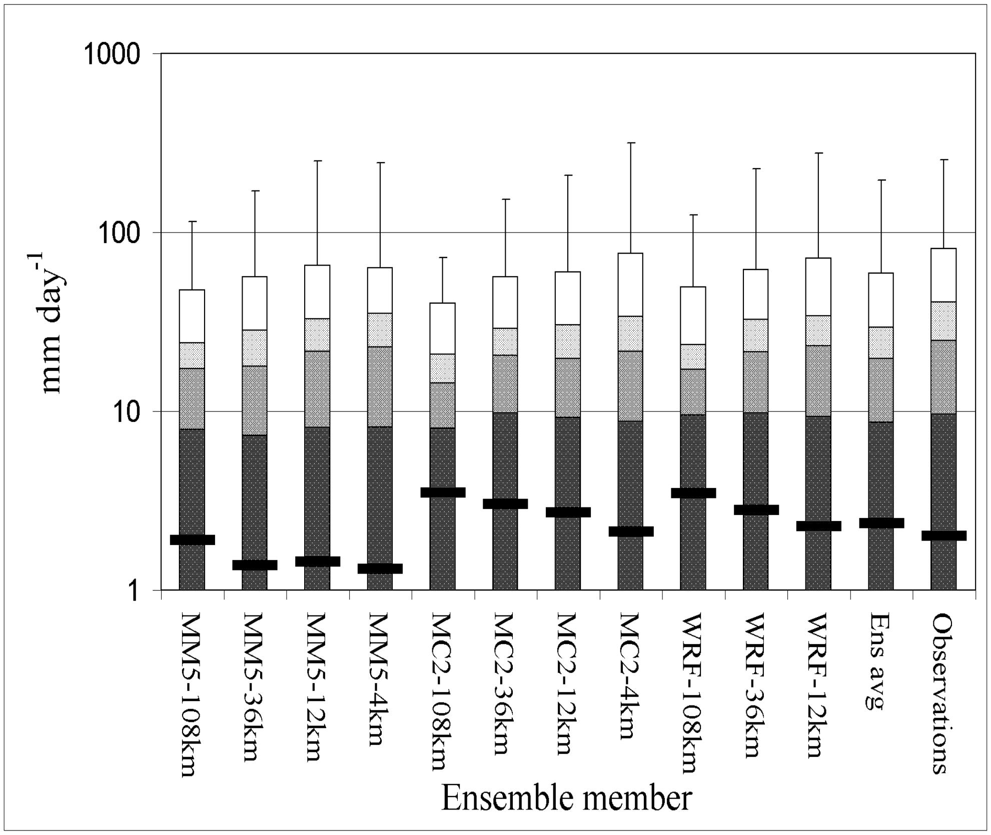

Figure 3.4: Same as Fig. 3.3 but for the day-two forecast.

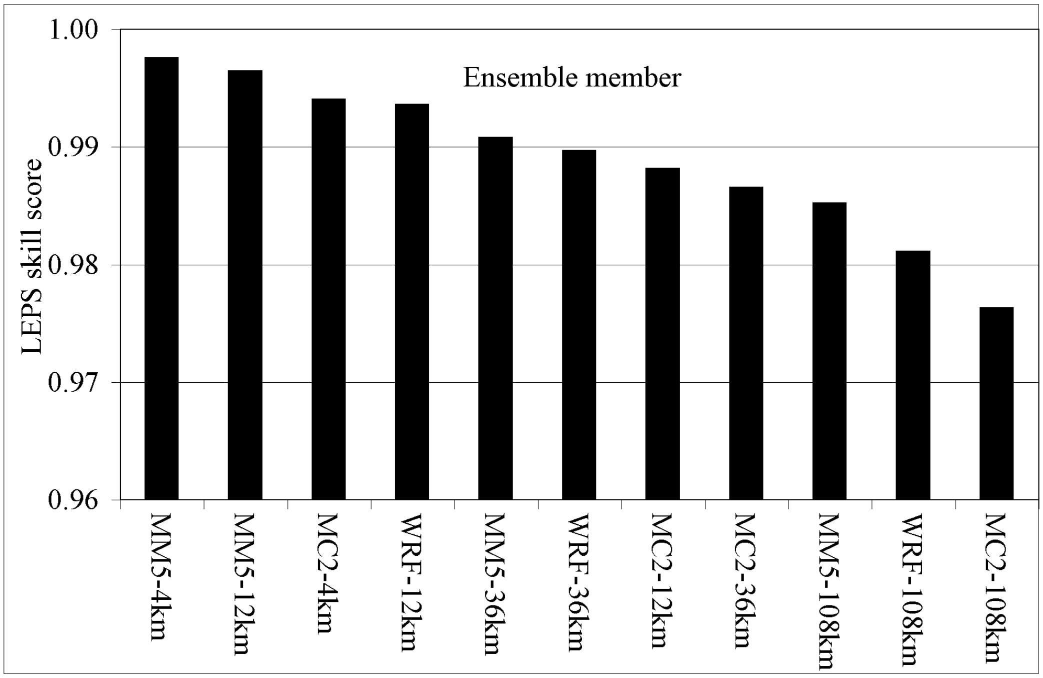

Figure 3.5: The LEPS (linear error in probability space) skill score for ensemble members for the day-one forecasts. Values closer to one are better. The LEPS skill score compares the cumulative distributions of each model ensemble member with the corresponding distribution of the observations. Discrepancies in the distributions near the center of the distributions detract more from the LEPS skill score than discrepancies in the extreme regions that correspond to rarer events. The coarsest-resolution members show a clear deterioration of skill compared to the finer-resolution members.

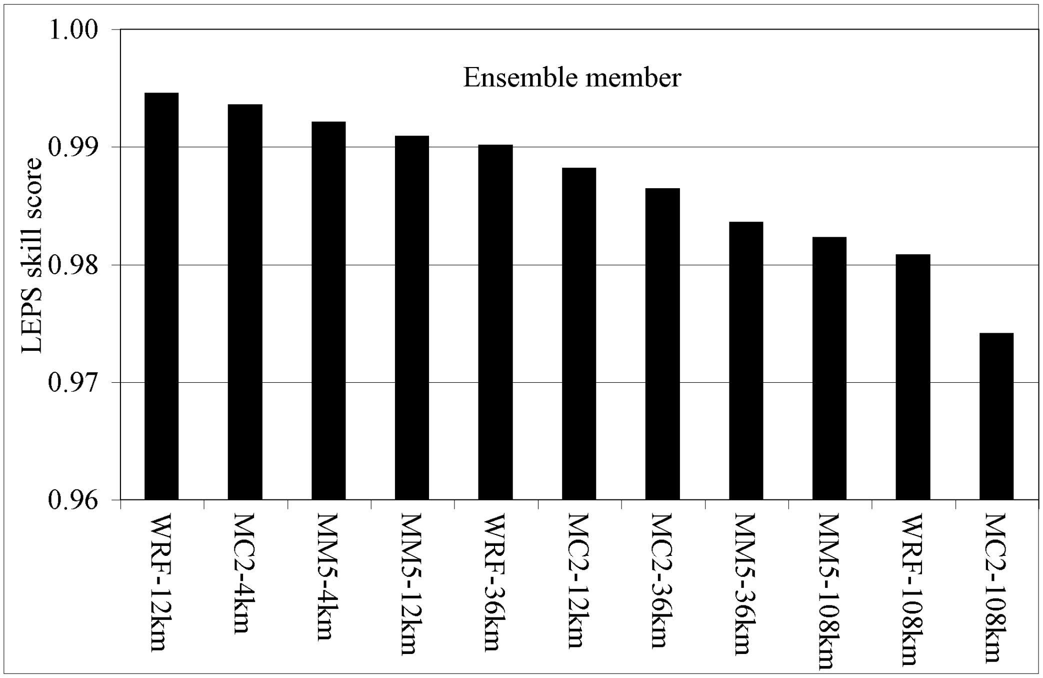

Figure 3.6: The LEPS skill score for ensemble members for the day-two forecasts. Values closer to one are better. The coarsest-resolution members show a clear deterioration of skill compared to the finer-resolution members, as for the day-one forecasts. In general the day-two forecasts are slightly less skillful than the day-one forecasts in terms of the LEPS skill score.

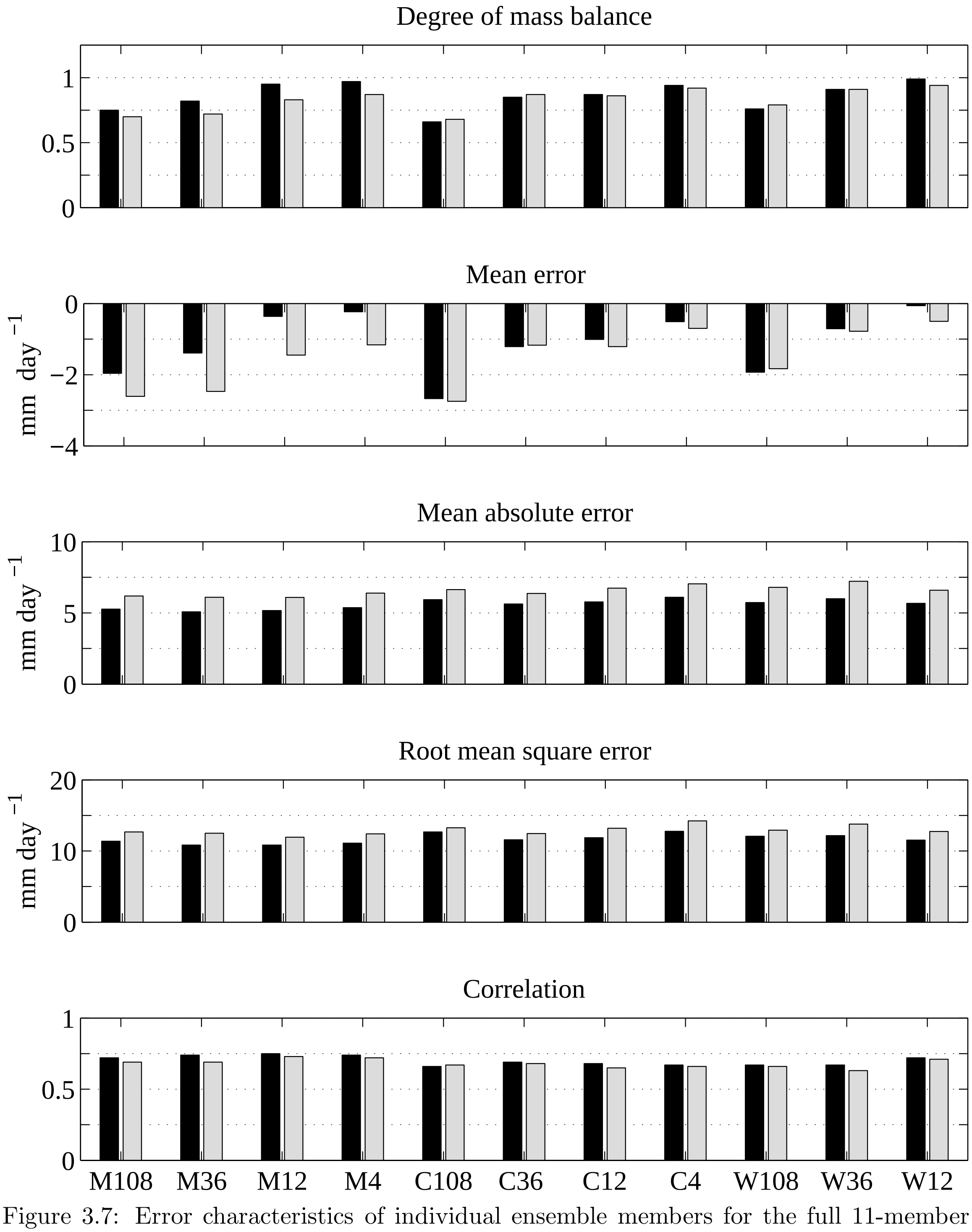

Figure 3.7: Error characteristics of individual ensemble members for the full 11-member ensemble for day-one forecasts (black) and day-two forecasts (grey). M refers to the MM5 model, C the MC2 model, and W the WRF model. The associated number is the model resolution (km).

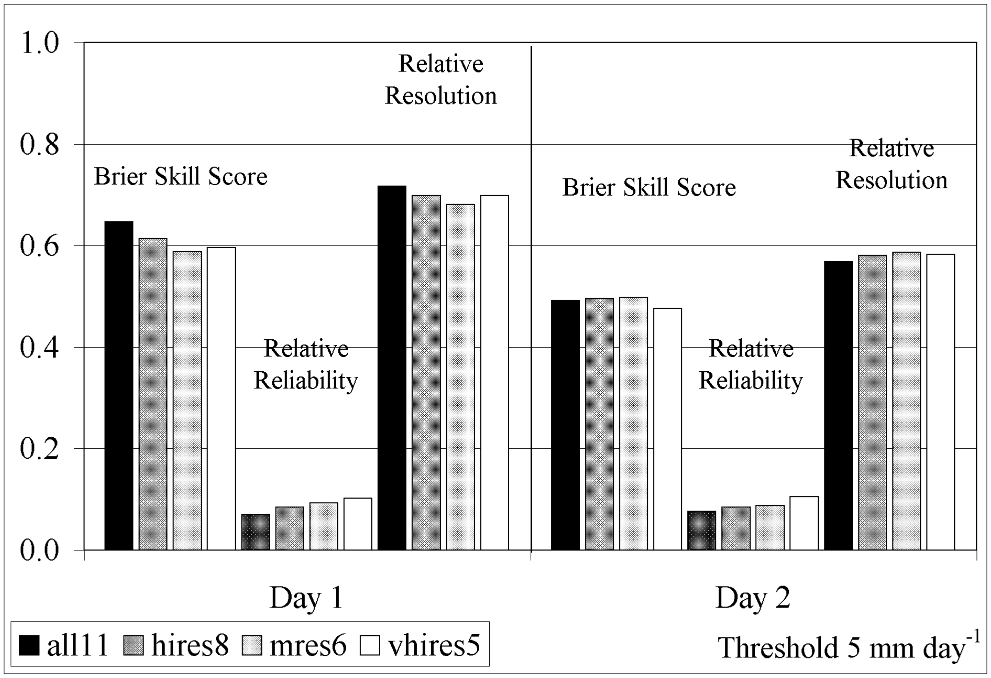

Figure 3.8: Brier skill score (one is perfect), relative reliability (zero is perfect) and relative resolution (one is perfect) for the SREFs for forecast days one and two. The precipitation threshold is 5 mm day^{-1}. The configuration encompassing all 11 ensemble members (from 108 km down to 4 km for all three models) performs better for the day-one forecast than the other configurations with fewer ensemble members. The four SREF configurations show similar Brier skill scores for the day-two forecast. Note that resolution deteriorates in the day-two timeframe while reliability is consistent for both day-one and day-two forecasts.

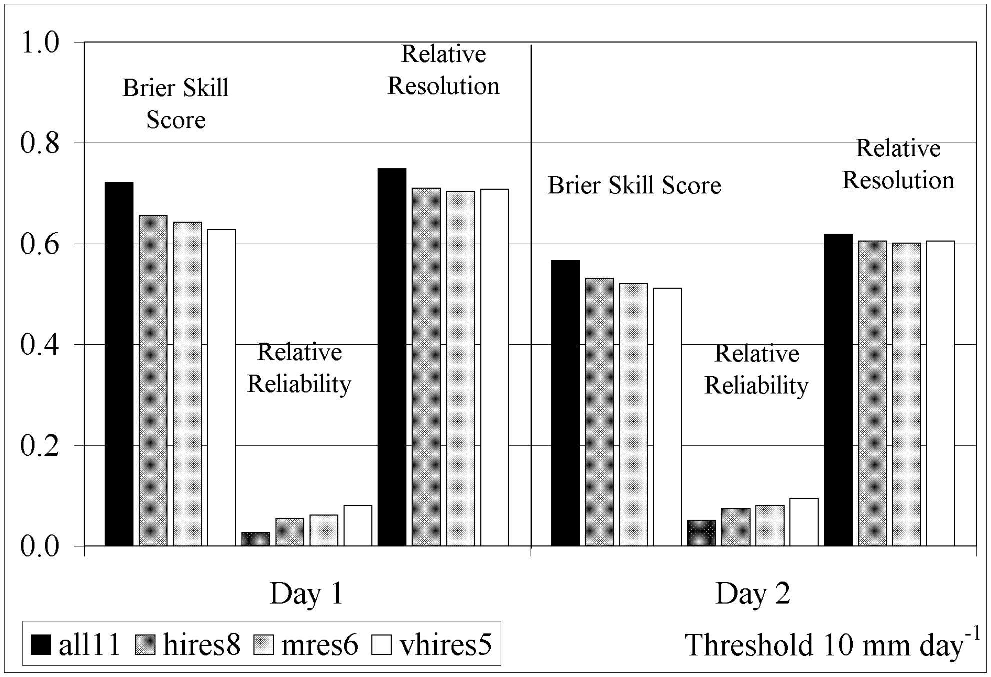

Figure 3.9: Same as Fig. 3.8 but for the 10 mm day^{-1} precipitation threshold. The configuration encompassing all 11 ensemble members performs best for both the day-one and day-two forecasts at this threshold.

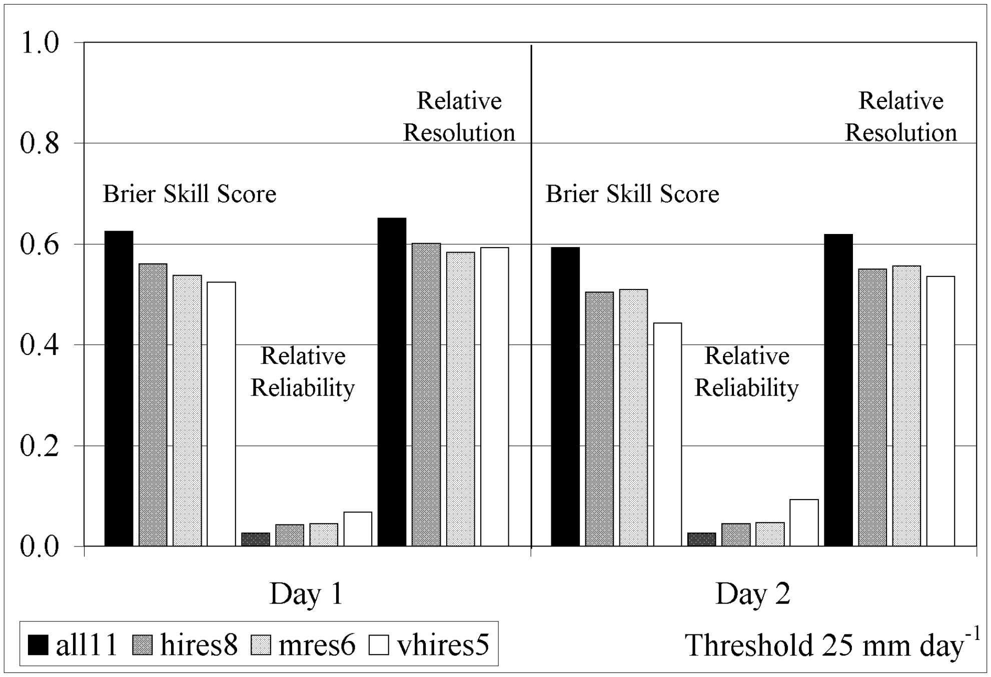

Figure 3.10: Same as Fig. 3.8 but for the 25 mm day^{-1} precipitation threshold. The configuration encompassing all 11 ensemble members performs best for both day-one and day-two forecasts at this threshold.

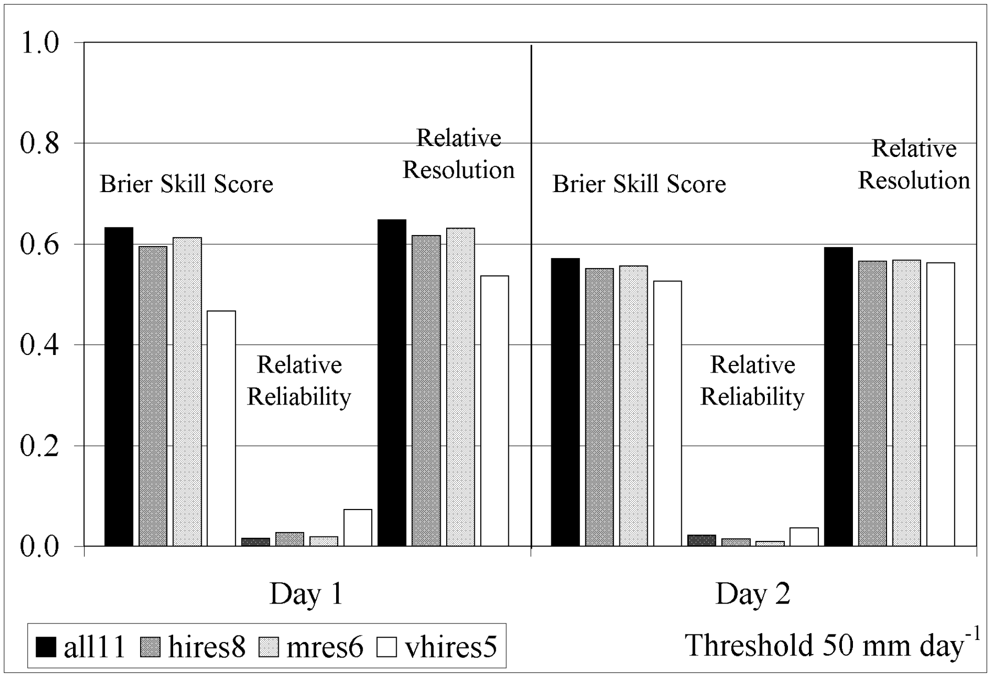

Figure 3.11: Same as Fig. 3.8 but for the 50 mm day^{-1} precipitation threshold. The configuration encompassing all 11 ensemble members performs best at this threshold.

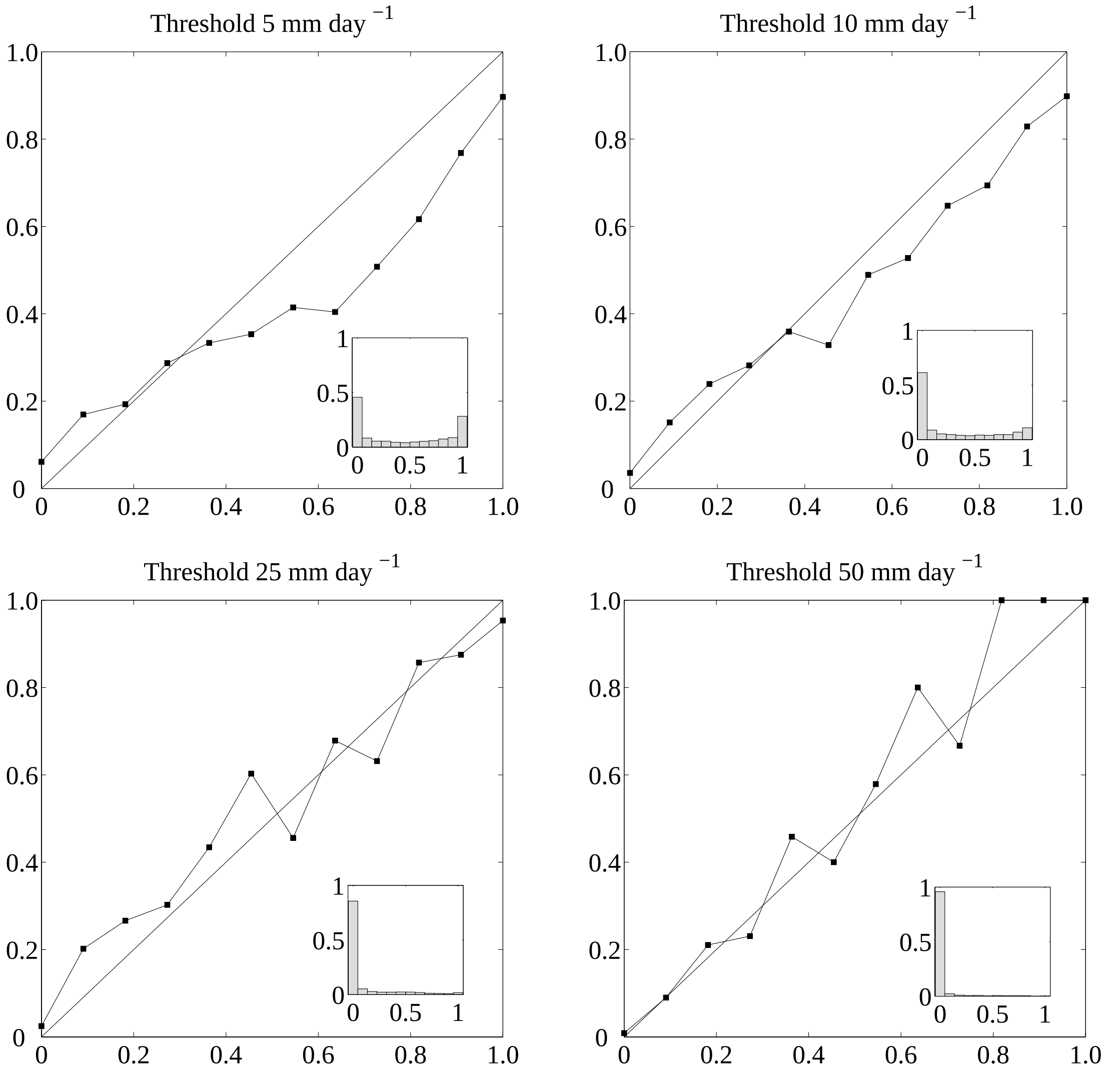

Figure 3.12: Reliability diagrams from the full 11-member ensemble for the day-one forecasts for precipitation thresholds of 5 mm day^{-1}, 10 mm day^{-1}, 25 mm day^{-1}, and 50 mm day^{-1}. The abscissa in these graphs is forecast probability, and the ordinate is observed relative frequency. The straight diagonal line indicates perfect reliability. The inset diagrams depict sharpness, where the abscissa is forecast probability and the ordinate is the relative frequency of use.

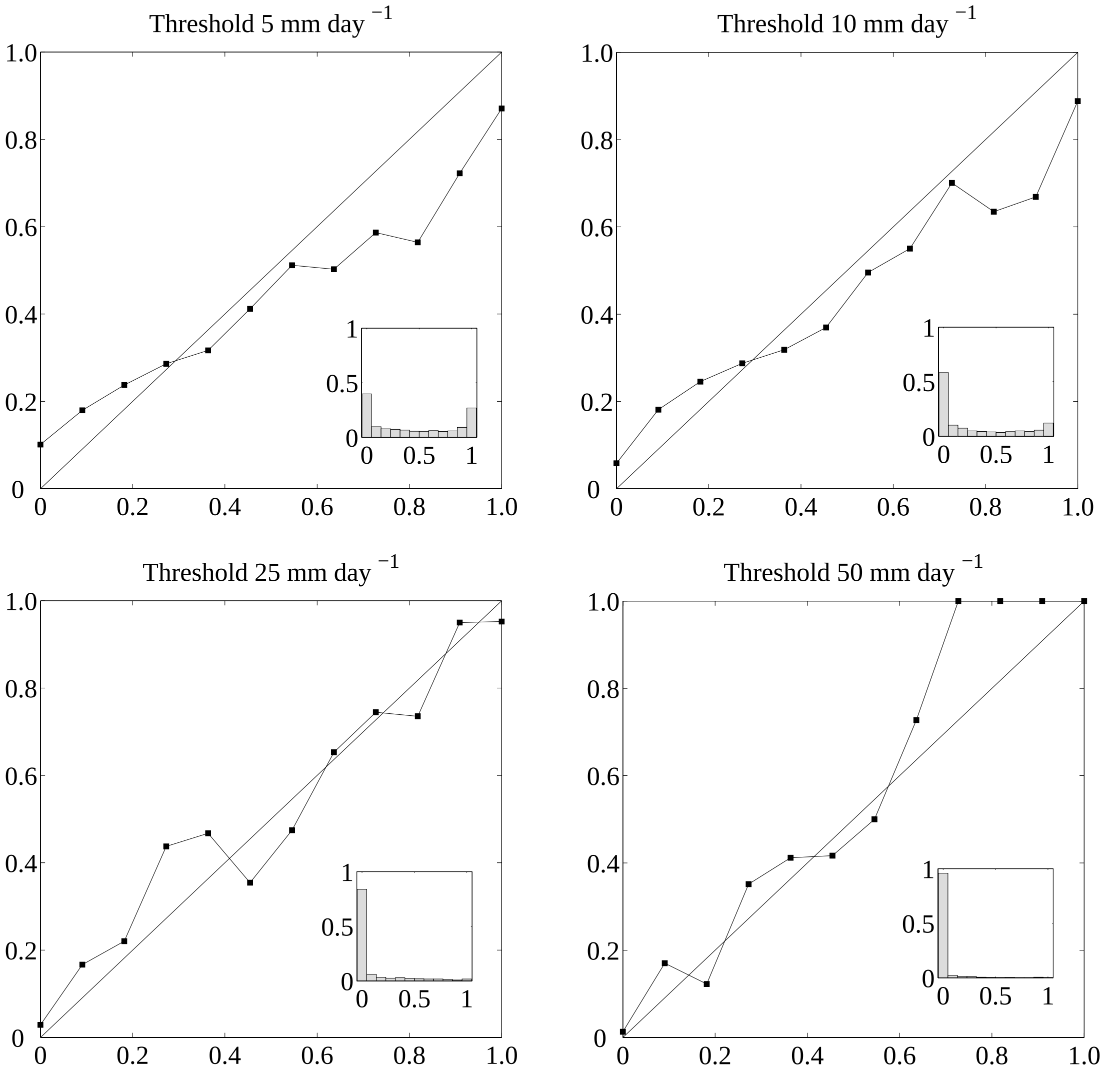

Figure 3.13: Same as Fig. 3.12 but for the day-two forecasts.

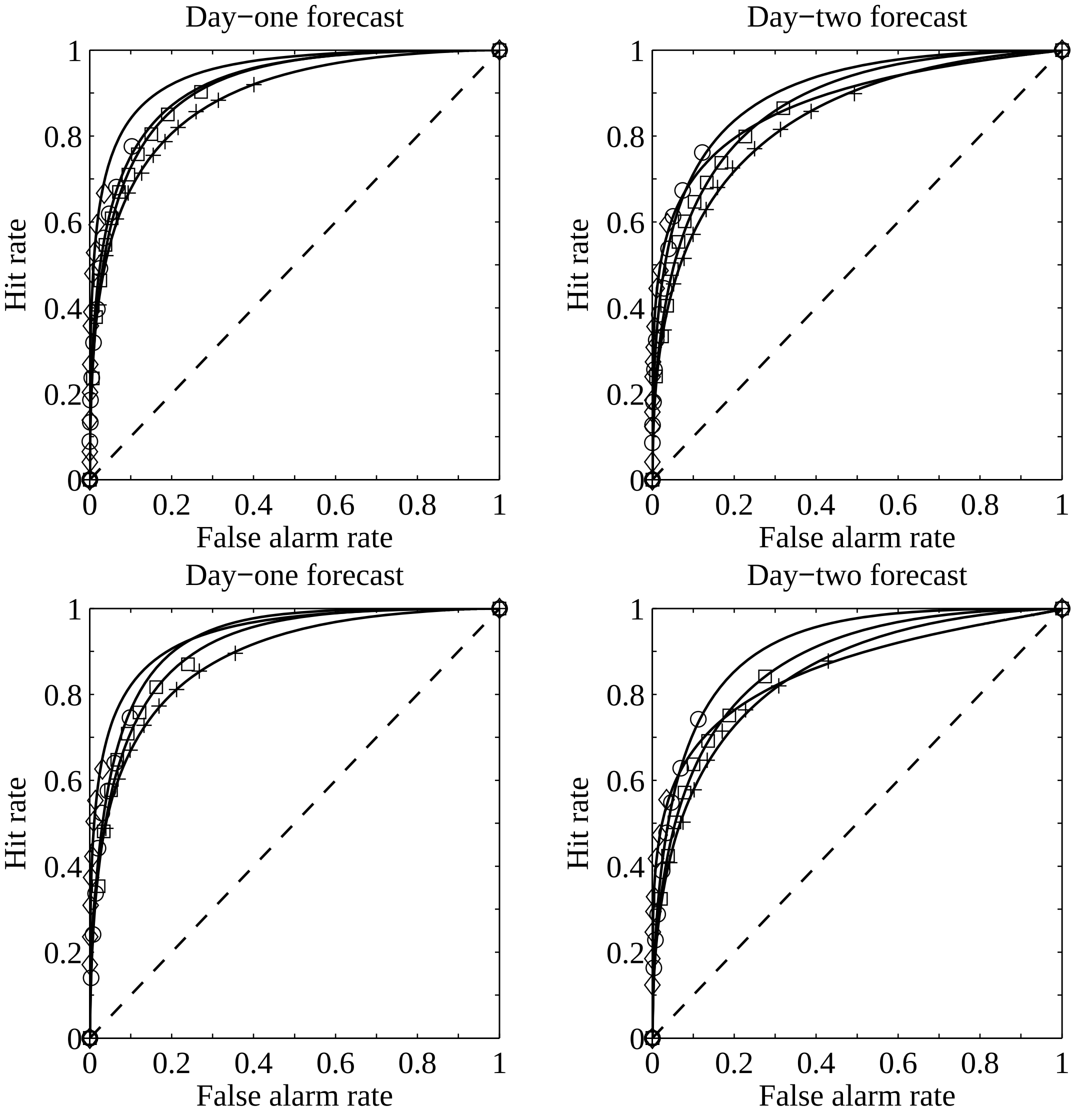

Figure 3.14: ROC curves for the full 11-member (top) and high resolution 8-member (bottom) ensemble prediction systems for precipitation thresholds 5 (crosses), 10 (squares), 25 (circles), and 50 mm day^{-1} (diamonds). The area under the ROC curves provides a measure of discrimination skill. Table 3.2 provides a summary of the ability of the SREF system to discriminate among events for all four different SREF configurations.

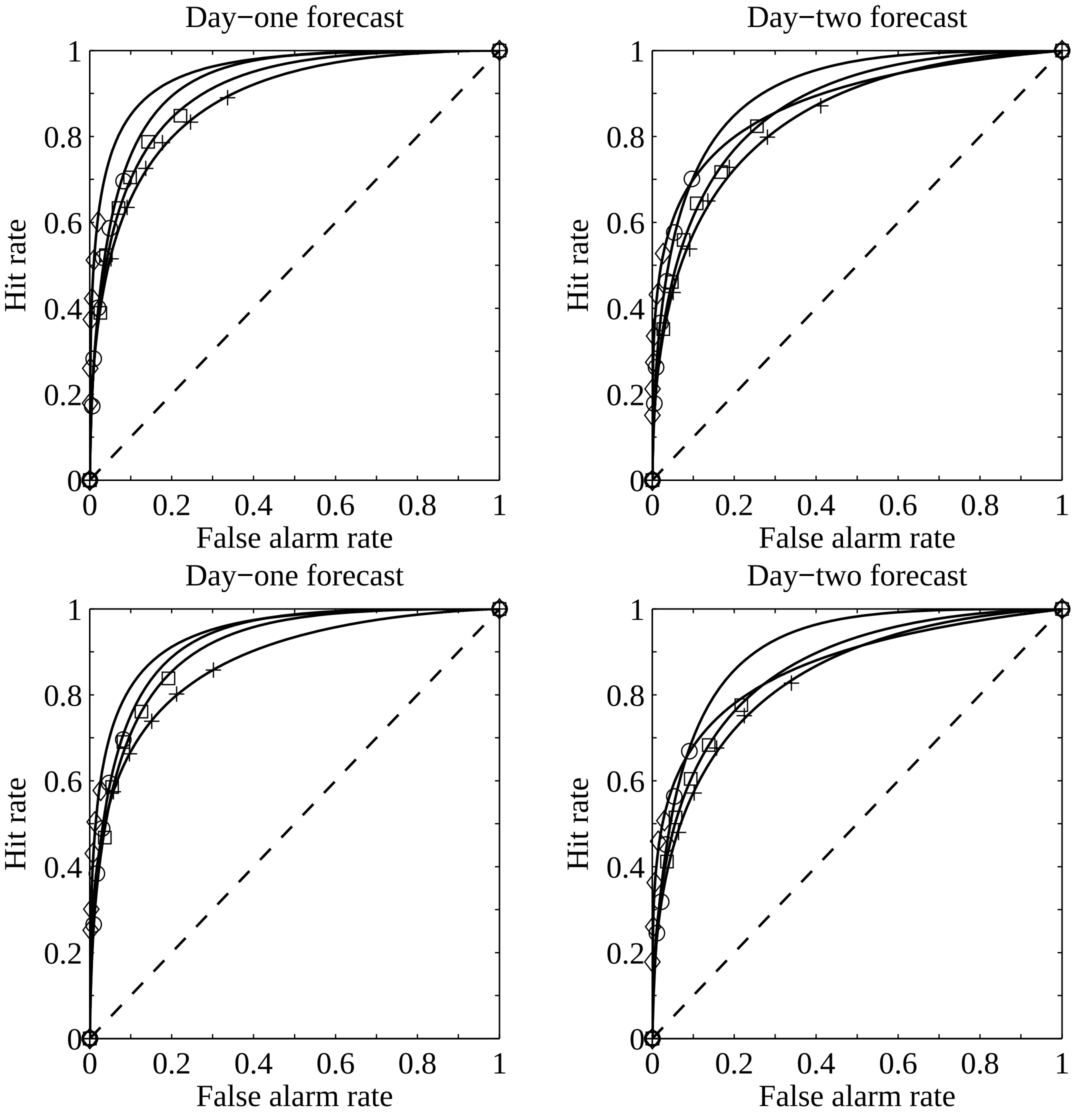

Figure 3.15: Same as Fig. 3.14 but for the medium resolution 6-member (top) and very high resolution 5-member (bottom) ensemble prediction systems.

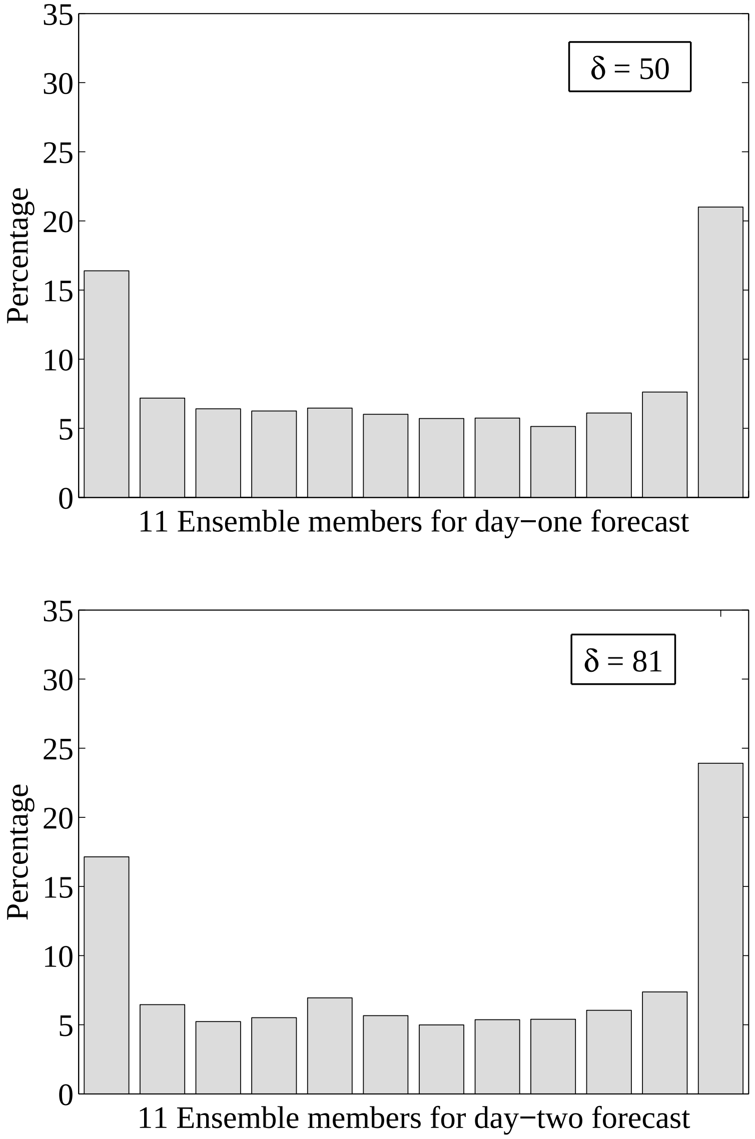

Figure 3.16: Rank histogram diagrams for day-one (top) and day-two (bottom) forecasts for the full 11-member SREF configuration. The values of $\delta >> 1$ for each forecast day indicate that the histogram is not flat.

Table 3.1: List of ensemble members for all four different SREF configurations. The "all11" configuration includes all three mesoscale model forecasts at all available horizontal resolutions. The "hires8" configuration includes all available members except the coarsest (108 km) members. The "mres6" configuration includes the medium-resolution mesoscale models at 12 km and 36 km. The "vhires5" configuration consists of the five highest-resolution models only, at 4 km and 12 km.

Number	Label	Description	all11	hires8	mres6	vhires5
1	MM5-108	MM5 108km resolution	√			
2	MM5-36	MM5 36km resolution	√	√	√	
3	MM5-12	MM5 12km resolution	√	√	√	√
4	MM5-4	MM5 4km resolution	√	√		√
5	MC2-108	MC2 108km resolution	√			
6	MC2-36	MC2 36km resolution	√	√	√	
7	MC2-12	MC2 12km resolution	√	√	√	√
8	MC2-4	MC2 4km resolution	√	√		√
9	WRF-108	WRF 108km resolution	√			
10	WRF-36	WRF 36km resolution	√	√	√	
11	WRF-12	WRF 12km resolution	√	√	√	√

Table 3.2: Values of ROC area skill scores for all four different SREF configurations for forecast days one and two. Values of ROC area skill scores closer to 1 are better. ROC area skill scores greater than 0.6 indicate good discriminating ability; scores 0.8 or higher indicate excellent discriminating ability. All four SREF configurations exhibit good to excellent discriminating ability. Discrimination generally increases with increasing precipitation threshold, except for the day-two 50 mm day^{-1} threshold. Discrimination deteriorates moving from the day-one to day-two forecast.

Day-one forecast

SREF Configuration	5 mm	10 mm	25 mm	50 mm
all11	0.77	0.82	0.84	0.89
hires8	0.77	0.82	0.85	0.88
mres6	0.76	0.81	0.85	0.90
vhires5	0.75	0.82	0.84	0.88

Day-two forecast

SREF Configuration	5 mm	10 mm	25 mm	50 mm
all11	0.68	0.74	0.80	0.75
hires8	0.69	0.74	0.82	0.71
mres6	0.69	0.74	0.81	0.76
vhires5	0.68	0.73	0.82	0.73

Table 3.3: Values of flatness δ for all four SREF configurations for forecast days one and two, where $\delta = 1$ indicates a perfectly flat rank histogram. Values of $\delta >> 1$ indicate under-dispersion in the ensemble forecasts. All four EPS configurations exhibit under-dispersion, a common trait amongst mesoscale SREFs.

SREF Configuration	Day-one forecast	Day-two forecast
all11	50	81
hires8	51	83
mres6	40	66
vhires5	42	76

Table 3.4: Values of Reliability Index RI for all four SREF configurations for forecast days one and two. Lower values of RI are better with a value of zero indicating perfect flatness of the rank histogram. RI is normalized so that ensembles of different size can be compared with one another. The full 11-member ensemble shows the best RI values. Reliability, as measured by this Reliability Index, improves with increasing ensemble size and shorter forecast period.

SREF Configuration	Day-one forecast	Day-two forecast
all11	8.54	10.5
hires8	10.7	12.9
mres6	11.7	13.8
vhires5	13.9	16.7

Bibliography

Benoit, R., M. Desgagne, P. Pellerin, S. Pellerin, and Y. Chartier, 1997: The Canadian MC2: A semi-lagrangian, semi-implicit wideband atmospheric model suited for finescale process studies and simulation. *Mon. Wea. Rev.*, **125**, 2383–2415.

Bright, D. R., M. S. Wandishin, R. E. Jewell, and S. J. Weiss, 2005: A physically based parameter for lightning prediction and its calibration in ensemble forecasts. *Preprints, Conf. on Meteor. Applications of Lightning Data*, AMS, San Diego, CA.

Buizza, R., A. Hollingsworth, F. Lalaurette, and A. Ghelli, 1999: Probabilistic predictions of precipitation using the ECMWF ensemble prediction system. *Wea. Forecasting*, **14**, 168–189.

Cheng, W. Y. Y. and W. J. Steenburgh, 2007: Strengths and weaknesses of MOS, running-mean bias removal, and Kalman filter techniques for improving model forecasts over the western United States. *Wea. Forecasting*, **22**, 1304–1318.

Davis, C. and F. Carr, 2000: Summary of the 1998 workshop on mesoscale model verification. *Bull. Amer. Meteor. Soc.*, **81**, 809–819.

Du, J., S. Mullen, and F. Sanders, 1997: Short-range ensemble forecasting of quantitative precipitation. *Mon. Wea. Rev.*, **125**, 2427–2459.

Eckel, F. A. and C. F. Mass, 2005: Aspects of effective mesoscale, short-range ensemble forecasting. *Wea. Forecasting*, **20**, 328–350.

Grell, G., J. Dudhia, and D. R. Stauffer, 1994: A description of the fifth-generation Penn State/NCAR mesoscale model (MM5). Technical Report TN-398+STR, NCAR, 121 pp.

Hacker, J. P., E. S. Krayenhoff, and R. B. Stull, 2003: Ensemble experiments on numerical weather prediction error and uncertainty for a North Pacific forecast failure. *Wea. Forecasting*, **18**, 12–31.

Hamill, T. M., 1997: Reliability diagrams for multicategory probabilistic forecasts. *Wea. Forecasting*, **12**, 736–741.

Hamill, T. M., S. L. Mullen, C. Snyder, Z. Toth, and D. P. Baumhefner, 2000: Ensemble forecasting in the short to medium range: Report from a workshop. *Bull. Amer. Meteor. Soc.*, **81**, 2653–2664.

Hou, D., E. Kalnay, and K. K. Droegemeier, 2001: Objective verification of the SAMEX '98 ensemble forecasts. *Mon. Wea. Rev.*, **129**, 73–91.

Jolliffe, I. T. and D. B. Stephenson, 2003: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. J. Wiley, England.

Jones, M. S., B. A. Colle, and J. S. Tongue, 2007: Evaluation of a mesoscale short-range ensemble forecast system over the northeast United States. *Wea. Forecasting*, **22**, 36–55.

Krzysztofowicz, R., W. J. Drzal, T. R. Drake, J. C. Weyman, and L. A. Giordano, 1993: Probabilistic quantitative precipitation forecasts for river basins. *Wea. Forecasting*, **8**, 424–439.

Mass, C. F., D. Ovens, K. Westrick, and B. Cole, 2002: Does increasing horizontal resolution produce more skillful forecasts? *Bull. Amer. Meteor. Soc.*, **83**, 407–430.

McCollor, D. and R. Stull, 2008a: Hydrometeorological accuracy enhancement via postprocessing of numerical weather forecasts in complex terrain. *Wea. Forecasting*, **23**, 131–144.

— 2008b: Hydrometeorological short-range ensemble forecasts in complex terrain. Part II: Economic evaluation. *Wea. Forecasting*, (in press).

McMurdie, L. and C. Mass, 2004: Major numerical forecast failures over the Northeast Pacific. *Wea. Forecasting*, **19**, 338–356.

Mullen, S. L. and R. Buizza, 2002: The impact of horizontal resolution and ensemble size on probabilistic forecasts of precipitation by the ECMWF ensemble prediction system. *Wea. Forecasting*, **17**, 173–191.

Murphy, A. H., 1991: Forecast verification: Its complexity and dimensionality. *Mon. Wea. Rev.*, **119**, 1590–1601.

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

Nutter, P., D. Stensrud, and M. Xue, 2004: Effects of coarsely-resolved and temporally-interpolated lateral boundary conditions on the dispersion of limited-area ensemble forecasts. *Mon. Wea. Rev.*, **132**, 2358–2377.

Skamarock, W. C., J. B. Klemp, J. Dudhia, D. O. Gill, D. M. Barker, W. Wang, and J. G. Powers, 2005: A description of the advanced research WRF version 2. Technical Report TN-468+STF, NCAR, 88 pp.

Spagnol, J., C. Readyhough, M. Stull, J. Mundy, R. Stull, S. Green, and G. Schajer, 2004: Rocketsonde buoy system observing system simulation experiments. *20th Conference on weather analysis and forecasting/16th Conference on numerical weather prediction*, AMS.

Stensrud, D. J., J. Bao, and T. T. Warner, 2000: Using initial conditions and model physics perturbations in short-range ensemble simulations of mesoscale convective systems. *Mon. Wea. Rev.*, **128**, 2077–2107.

Stensrud, D. J. and N. Yussouf, 2007: Reliable probabilistic quantitative precipitation forecasts from a short-range ensemble forecasting system. *Wea. Forecasting*, **22**, 3–17.

Toth, Z., O. Talagrand, G. Candille, and Y. Zhu, 2003: Probability and ensemble forecasts. *Forecast verification: A practitioner's guide in atmospheric science*, I. Jolliffe and D. B. Stephenson, eds., Wiley, 137–163.

Wandishin, M. S., S. L. Mullen, D. J. Stensrud, and H. E. Brooks, 2001: Evaluation of a short-range multimodel ensemble system. *Mon. Wea. Rev.*, **129**, 729–747.

Weber, F., L. Perreault, and V. Fortin, 2006: Measuring the performance of hydrological forecasts for hydropower production at BC Hydro and Hydro-Quebec. *18th Conference on climate variability and change*, AMS.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*. Academic Press, San Diego, 2nd edition.

Wilson, L. J., 2000: Comments on “Probabilistic predictions of precipitation using the ECMWF ensemble prediction system”. *Wea. Forecasting*, **15**, 361–364.

Yuan, H., S. Mullen, X. Gao, S. Sorooshian, J. Du, and H.-M. H. Juang, 2005: Verification of probabilistic quantitative precipitation forecasts over the southwest United States during winter 2002/03 by the RSM ensemble system. *Mon. Wea. Rev.*, **133**, 279–294.

Yussouf, N. and D. J. Stensrud, 2007: Bias-corrected short-range ensemble forecasts of near surface variables during the 2005/2006 cool season. *Wea. Forecasting*, **22**, 1274–1286.

Chapter 4

Short-Range Ensemble Forecasts in Complex Terrain. Part II: Economic Evaluation[5]

4.1 Introduction

Part I of this two-part paper (McCollor and Stull, 2008) documents the meteorological evaluation of a short-range ensemble forecast (SREF) system for hydrometeorological applications in complex terrain. A standardized set of evaluation methods, scores, and diagrams (as described in Hamill et al. 2000) was employed to compare and judge the meteorological skill of the SREF system. However, skill and dispersion measures of the meteorology alone do not give a complete assessment of SREF forecast value (Anderson-Berry et al., 2004).

Complete evaluation of the benefit of a forecast system should consider the weather sensitivity and decision-making processes of a particular end user or group of users. Societal and environmental impacts are also important components of weather forecast evaluation, however, in most cases economic impacts are easier to directly identify and quantify.

In a study by Roulston et al. (2006), laboratory experiments documented the ability of non-specialist weather forecast users to make sound decisions. That study showed how users increase their expected economic reward while reducing their exposure to risk by incorporating quantitative estimates of weather forecast uncertainty into their decisions.

We employ economic skill measures to evaluate user-dependent hydrometeorological forecasts. Two economic models adapted for the hydro-electric energy sector are examined in this paper: a cost/loss ratio decision model and a utility-maximizing decision-theory model. The economic models are applied in an operations assessment at Jordan River, a hydro-electric facility on the southwest coast of Vancouver Island, British Columbia, Canada.

[5] A version of this chapter has been accepted for publication. McCollor, D. and R. Stull, 2008: Hydrometeorological Short-Range Ensemble Forecasts in Complex Terrain. Part II: Economic Evaluation. *Wea. Forecasting.*

Figure 4.1 includes a photo of the Jordan River site plus three other reservoir facilities in this region of complex terrain.

Forecast evaluation using economic models applicable to the hydro-electric sector is introduced in section 4.2. Application of the models to probabilistic forecasts from a SREF system is given in section 4.3. A case study application is presented in section 4.4, and conclusions are provided in section 4.5.

4.2 Forecast Evaluation via Economic Models

4.2.1 Background

In rapid runoff response pluvial watersheds (the subject of this study), fore-knowledge of inflow allows optimal reservoir operation under a set of given operating constraints. Operating constraints may include, for example, optimizing the depth (hydraulic pressure head) to maximize electrical generation, avoiding spill under high inflow events, maintaining minimum flow requirements under dry conditions to provide adequate water for fish survival and aquatic habitat, and maintaining stationary reservoir levels under fluctuating inflow conditions to enhance recreational usage.

Uncertainty in future rainfall, hence uncertainty in future reservoir inflow, poses significant problems for reservoir operators tasked with meeting stated operating constraints. Minimizing future inflow uncertainty through the knowledgeable use of probabilistic forecasts can be analyzed with suitable economic models.

4.2.2 Cost/Loss Model

Fore-knowledge of high inflow allows a reservoir operator to draft (i.e., lower the water level of) a reservoir ahead of the inflow event. This allows the operator to obtain full value of the water by routing all reservoir water through the turbines and generators before and during the event. Thus the operator can capture all the precipitation-induced runoff without being forced to release excess water (and by-pass the revenue producing generators) downstream. This action is colloquially known as “digging a hole in the reservoir” to capture subsequent inflow. In the managed watersheds analyzed in this study, the dominant source of higher inflow is rainfall events produced by land-falling Pacific frontal systems impinging on the mountainous coastal terrain of southwest British Columbia (see Fig. 3.1 in the companion paper for a reference map of the region).

There is an economic risk associated with operating a reservoir in the manner outlined above. If no action is taken prior to a high inflow event, the excess water flowing into the reservoir is released or spilled without generating revenue, hence incurring an economic loss from the value of that water. If the operator does in fact generate power in advance of the high inflow event, an economic cost is entailed involving hydraulic head loss from lowering the reservoir elevation, which may not be offset if the forecast inflow event does not occur.

Economic forecast value can be addressed using the concept of a simple (static) cost/loss ratio decision model (Murphy, 1977; Katz and Murphy, 1997; Richardson, 2000; Zhu et al., 2002; Richardson, 2003). Sensible weather forecast examples employing a cost/loss ratio decision model in specific weather forecast evaluations are provided in the literature for road-weather forecasts (Thornes and Stephenson, 2001), air quality forecasts (Pagowski and Grell, 2006), severe weather forecasts (Legg and Mylne, 2004, 2007), precipitation forecasts from a global ensemble prediction system (Mullen and Buizza, 2002), precipitation forecasts over the southwestern United States (Yuan et al., 2005), temperature forecasts for the energy sector (Stensrud and Yussouf, 2003), and medium-range flood predictions for watersheds in low-relief terrain (Roulin, 2007). A description of the cost/loss model as applied here to reservoir operations is provided below.

A reservoir operator has essentially two alternative courses of action to choose from, either draft the reservoir or not. The choice depends partly on the capacity of the reservoir to hold additional water, and largely on the precipitation forecast, since precipitation is highly correlated with inflow in these particular watersheds (see Fig. 4.2 for an example of precipitation-inflow correlation). Each action has an associated cost (from head loss) and leads to an economic benefit (minimize lost revenue due to forced spill) depending on the precipitation (hence inflow) that occurs. The task of the reservoir operator is to choose the appropriate action in each instance that will minimize both the cost and the expected loss over the long term.

The cost/loss model is especially useful in this study because it shows how probabilistic forecasts can be used in the decision-making process, and provides a measure of the benefit of probabilistic forecasts over deterministic or ensemble average forecasts. In the case at hand, the forecast user (the reservoir operator) is sensitive to the specific daily precipitation forecast thresholds analyzed in the companion paper: 5 mm day^{-1}, 10 mm day^{-1}, 25 mm day^{-1}, and 50 mm day^{-1}.

The four possible combinations of action and occurrence (with the net cost depending on what happened and what action was taken) are summarized in Table 4.1. A separate analysis will be performed for each precipitation threshold. In Table 4.1, a is the number of hits (successful event forecasts), b is false alarms (forecast but not observed), c is the number of misses (observed but not forecast), and d refers to correct rejections (neither forecast nor observed). Hits and false alarms both incur a cost of taking preventative action (C), since the operator drafts the reservoir whenever the event is forecast. Misses are associated with a loss due to lack of prevention (L). Correct rejections incur no expenses either way. The reservoir operator acts to minimize the average long term loss by taking appropriate action on each occasion.

The economic value (V) of the forecasts are defined in terms of expenses, E, in an equivalent manner as skill scores for meteorological forecasts (Richardson, 2003):

$$V = \frac{E_{climate} - E_{forecast}}{E_{climate} - E_{perfect}} \tag{4.1}$$

where the mean *climate expense* $E_{climate}$ is the least expensive of two options, (1) always take protective action, thus incurring a constant cost C but never experiencing a loss L, or (2) never take protective action, which involves no cost but will result in total losses equivalent to $\overline{o}L$, where $\overline{o}$ is the climatological base-rate probability of the event occurring. Therefore, $E_{climate} = \min(C, \overline{o}L)$. Given perfect knowledge of future events, the reservoir operator would take action only when the event was going to occur, and therefore would only incur costs at the climatological base rate. The mean expense would then be $E_{perfect} = \overline{o}C$.

The sample mean expense of the forecast is calculated by multiplying the relative frequencies of each of the four possible outcomes listed in Table 4.1 by the corresponding expense, resulting in:

$$E_{forecast} = \frac{a}{n}C + \frac{b}{n}C + \frac{c}{n}L \tag{4.2}$$

where $n = a+b+c+d$ is the total number of forecast/observation pairs. Thus, the economic value equation (4.1) becomes:

$$V = \frac{\min(r, \overline{o}) - \frac{r}{n}(a+b) - \frac{c}{n}}{\min(r, \overline{o}) - \overline{o}r} \tag{4.3}$$

where $r = C/L$ is the specific user's *cost/loss ratio.*

Summarizing the event categorization described by the binary contingency table elements (a, b, c, and d) into hit rate $H = \frac{a}{a+c}$, false alarm rate $F = \frac{b}{b+d}$, and (as already suggested) climatological base rate $\overline{o} = \frac{a+c}{n}$ allows equation (4.3) to be written as:

$$V = \frac{\min(r, \overline{o}) - F(1-\overline{o})r + H\overline{o}(1-r) - \overline{o}}{\min(r, \overline{o}) - \overline{o}r} \tag{4.4}$$

The introduction of hit rate H and false alarm rate F into the evaluation of forecast value shows that potential economic value for the cost/loss decision model is uniquely determined from relative operating characteristic (ROC) curves of H versus F (see Appendix C of the companion paper). Additionally, the introduction of the forecast user into the determination of forecast value shows that different users of the same forecast will derive different potential benefits depending on that user's specific cost/loss ratio.

For a particular SREF system and a given precipitation threshold, the value V of the forecasts is solely determined by the user's cost/loss ratio. By plotting V as a function of cost/loss ratio, r, a water resource manager can determine if the full ensemble forecast system provides more value than a deterministic forecast or the ensemble mean forecast.

For a reliable ensemble forecast system, maximum value for a particular user will be obtained by choosing a forecast probability equal to the user's particular cost/loss ratio, r (Richardson, 2000, 2003). Also, the maximum value of V always occurs when $r = \overline{o}$. Therefore, forecast users whose C/L happens to equal the climatological base-rate frequency $\overline{o}$ for a particular event gain the most value from the probability forecasts.

4.2.3 Decision Theory Model

The cost/loss model described in section 4.2.2 is a binary-decision model. Other techniques that incorporate economic decision-making for continuous input/output variables (Katz and Murphy, 1997; Smith et al., 2001) such as precipitation and power production can also be applied to reservoir planning processes.

In certain situations a reservoir operator may be tasked with maintaining a constant-level reservoir under variable inflow conditions. Optimal use of probabilistic forecasts may be analyzed using a maximum-utility decision theory economic model. In decision-theory models, optimal decisions in uncertain environments are associated with the maximum of expected rewards.

Formal decision theory was originally developed in the first half of the twentieth century by mathematicians aiming at modeling the process by which optimal rational decisions are made (Baldi, 2001). In this model, a decision maker in an uncertain environment is faced with a choice of possible actions. Each action has an associated consequence. The consequences are not known with certainty but are expressed as a probability of occurrence. The optimal decision-making strategy is to maximize the expected value (i.e., maximize gain or minimize loss), where expected value refers to a weighted average taken over all possible probabilistic states of the environment. Thus, decision theory naturally leads to probabilistic determination, and an ensemble forecast is a method to define such weights or probabilities.

Economic decison theory is based on the concept of utility, in which users can assign a value or *utility* "U" to each potential outcome (Katz and Murphy, 1997; Smith et al., 2001). A utility function is a mathematical transformation from *money* to *utility*, reflecting the perceived worth of each outcome (instead of the absolute worth) to the decision maker. Let the user's decision be represented by a continuous variable, x, and the outcome be represented by a continuous variable y. If the probability distribution function (PDF) of y is $p(y)$ then the expected utility for a given decision, $E[U](x)$, is:

$$E[U](x) = \int_y U(x, y)p(y)dy \tag{4.5}$$

The user can make the choice X that maximizes their expected utility:

$$E[U](X) = \max_x(E[U](x)) \tag{4.6}$$

The expected utility hypothesis in economics (Laffont, 1989) states that the utility of a user facing uncertainty is calculated by considering the utility in each possible state and constructing a weighted average, the weights being the user's estimate of the probability of each state. For hydrometeorological analysis, the user must define a utility function that describes the benefit or loss incurred based on a particular reservoir operation decision. The uncertainty in the decision is derived from uncertainty in future rainfall, and the probability of such an uncertain rainfall event is determined from the SREF prediction system.

The utility function for this hydrologic model is adapted from another renewable energy source (wind generation) utility function as decribed by Smith et al. (2001). Short-term rainfall, hence reservoir inflow, is a major factor in determining the capacity to supply electrical energy to the grid. The reservoir operator must be able to forecast generation in advance, subject to maintaining a constant-level reservoir (for the benefit of recreational and other users), since failure to provide energy specified in a contract will lead to purchasing energy at market prices to meet a pre-specified contractual delivery.

The reservoir manager must decide in advance how much energy their operations will supply the grid on the next day, stipulated by a contract energy production K_c. This energy is sold at a price S_c. If actual production, K_a, falls short of the contractual amount, the manager must replace the missing energy at a market price S_m. If the reservoir manager is assumed to be risk neutral, then the applicable utility function is given by:

$$\begin{aligned} U &= S_c \cdot K_c & K_a &\geq K_c \\ U &= S_c \cdot K_a - S_m \cdot (K_c - K_a) & K_a &< K_c \end{aligned} \tag{4.7}$$

The expected utility can be evaluated by assigning a probability p_k to each potential energy production amount K_a:

$$E[U](K_a) = \sum_k p_k \cdot U(K_a, k) \tag{4.8}$$

4.3 Results

4.3.1 Cost/Loss Model

The forecast user (in this case a reservoir manager) is faced with the question of how high the forecast probability of a particular event needs to be before the threat is great enough to warrant taking preventative action. The decision maker needs to set a *threshold probability* p_t and take preventative action only when the forecast probability exceeds p_t. By varying p_t over $0/M, 1/M, 2/M, ...1$, where M is the number of ensemble members employed in the SREF, a sequence of hit H vs. false-alarm F rates can be plotted that traces out the ROC curve for the SREF system (see section 3.4.3). Different value traces V are then plotted as a function of C/L [see equation (4.4) for each distinct point (F,H) corresponding to each particular p_t on the ROC curve]. The value curves for each precipitation threshold are displayed in Fig. 4.3 for the day-one forecasts and in Fig. 4.4 for the day-two forecasts.

In each of the individual V vs. C/L graphs shown in Figs. 4.3 and 4.4, the envelope curve (displayed with a heavy solid line) shows the optimum value of the SREF system, obtained when each decision maker employs the probability threshold p_t that maximizes V for their specific C/L. As seen in Figs. 4.3 and 4.4, the envelope curve is never less than any of the individual p_t curves, indicating that no particular single forecast will benefit different users utilizing different cost/loss ratios. The maximum value (V_{max}) of the envelope curve for V occurs at that value of C/L equal to the climatological base rate $\overline{o}$, and is seen to shift toward smaller cost/loss ratios (lower $\overline{o}$ hence rarer events) with increasing precipitation threshold.

The shape of the V_{max} envelope curves for the day-one forecasts are similar to those for the day-two forecasts, for all four precipitation thresholds. This indicates little change in relative value among different users employing different cost/loss ratios in their decision making. However the V_{max} envelope curve is shifted downward slightly in the day-two forecasts compared to the day-one forecasts, indicating that, as expected, the measured value of the forecasts decreases as the forecast range is extended. If the probabilistic forecasts were extended into the mid-range (3 to 10 day forecasts), the V_{max} envelope curve would necessarily collapse as forecast skill approaches climatology [equation (4.1)]. A similar assessment for mid-range forecasts (a potential avenue for future research), would indicate how far into the future these forecasts maintain positive economic value.

Plotting the envelope V_{max} vs. C/L allows a clearer view of how different forecast users benefit from the probabilistic nature of SREF systems, especially when plots of deterministic forecasts and ensemble-average forecasts are included for comparison. In Figs. 4.5 and 4.6 (for day-one and day-two forecasts, respectively), the maximum potential value curves are plotted along with a best-deterministic forecast and an ensemble-average forecast. The best-deterministic forecast was chosen among the individual SREF members as the one with the best error characteristics (see section 3.4.1), namely the very-high resolution MM5-4 km model. The ensemble-average forecast is the equally-weighted average of all SREF system member forecasts. The ensemble-average forecast is included for comparison, although in practice caution is advised in using this forecast because it may not represent a physically-realizable solution to the model equations (in the manner that the average of the numbers on a six-sided die, 3.5, is not a physically-realizable outcome of rolling the die).

The plots in Figs. 4.5 and 4.6 indicate that the SREF system provides the best economic value for both day-one and day-two forecasts, all precipitation thresholds, and all cost/loss ratios. Also, these figures show positive economic value to all users at the highest precipitation threshold of 50 mm day^{-1}. However for lower precipitation thresholds, reservoir managers employing the very lowest cost/loss ratios progressively lose value as the precipitation threshold lowers from 25 mm day^{-1} to 10 mm day^{-1} to 5 mm day^{-1}. The advantage of a probabilistic forecast system is evident from the fact that the ensemble prediction system (EPS) exhibits a larger range of forecast users that can economically benefit from the forecasts than either the deterministic forecasts or the ensemble-average forecasts (namely, the dotted and dashed curves are narrower and lower than the ensemble envelope curve).

For the higher-precipitation thresholds of 25 mm day^{-1} and 50 mm day^{-1} the ensemble-average forecast outperforms the deterministic forecast in terms of value, however, the probabilistic forecasts are even more advantageous as they extend the valuable range of the forecasts to a broader cast of users. At the lowest precipitation threshold of 5 mm day^{-1} the measured maximum values V_{max} of the SREF system, the deterministic forecast, and the ensemble-average forecast are the same. As the thresholds and their impact on reservoir operations increase, the measured maximum value V_{max} of the SREF system becomes greater than that of the deterministic and ensemble-average forecasts.

Information about choosing the correct probability threshold p_t for each decision maker can be shown by plotting value V as a function of p_t for different C/L (see Fig. 4.7 for the day-one forecasts and Fig. 4.8 for the day-two forecasts). The plots show that reservoir managers with low $C/L = 0.2$ who have low costs but potentially high losses benefit from taking action (pre-generating electricity thus lowering the reservoir in advance of a forecast precipitation event) at a precipitation-threshold-exceedance probability as low as 10%.

As C/L increases to 0.6 for specific reservoirs, the managers should wait until the probability of an event is 30% before taking preventative action. For managers of a reservoir with $C/L = 0.8$, the associated high cost of lowering the reservoir in advance of a forecast inflow event and limited potential losses from forced release or spill mean that they should wait until forecasts are in the 70% to 80% range before pre-generating.

Day-one SREF system forecasts are measureably better than day-two forecasts. The plots in Figs. 4.7 and 4.8 indicate the trend that the probability of an event in the day-two forecast timeframe should be about 10% higher than the probability in the equivalent day-one forecast to ensure the same value ensues from action taken. In other words, reservoir managers operating a reservoir with $C/L = 0.2$ could pre-generate in the day-two timeframe when precipitation event forecast probability reaches 20%. Managers operating a reservoir with $C/L = 0.8$ can take preventative action two days before a forecast event of 5 mm day^{-1} or 10 mm day^{-1}, but the probability of occurrence must be close to 90%. This effect is less evident for 25 mm day^{-1} precipitation events and is non-evident for 50 mm day^{-1} events, likely because the smaller number of cases for these climatologically rarer events introduces sampling errors into the analysis (as shown in section 3.4.2).

Looking more closely at the relation between precipitation threshold and value for different SREF combinations, Figs. 4.9 and 4.10 show this relation for the day-one forecasts and day-two forecasts, respectively. The full ensemble SREF provides value for a greater range of users at all precipitaton thresholds, compared to the deterministic forecast and the ensemble-average forecast. As the precipitation threshold increases, the forecasts provide value to more and more users toward lower C/L values.

There are specific links between the economic-value analysis of the cost/loss model and the ROC analysis presented in section 3.4.3. Figure 4.11 shows the ROC curve for the day-one forecasts with a 10 mm day^{-1} threshold alongside the corresponding maximum value envelope curve. The deterministic high-resolution forecast and ensemble-average forecast are each represented by single points on the ROC diagram. These points are shown connected with straight-line segments to the upper-right and lower-left corners of the ROC diagram.

The deterministic forecast point and the ensemble-average forecast point both lie near the probability forecast ROC curve. This suggests that basic forecast system performance is similar for all three forecast systems. The benefit of the probability forecast depends on the varying needs of different users. As seen on the value curve, all three forecast systems show similar value for a narrow range of users with C/L values between about 0.3 and 0.4. For C/L values outside this range the ensemble forecasts prove most economically valuable.

An example using ensemble precipitation forecasts in a cost/loss model for specific reservoirs is given in section 4.4.1.

4.3.2 Decision Theory Model

The utility functions calculated from equation (4.7) for Jordan River operations indicate the variation of utility with pricing structure. The particular utility function explored here displays increasing penalties as the S_m/S_c ratio increases, and increasing benefits as higher energy contract amounts are selected. The penalties and benefits increase jointly with the precipitation threshold indicated.

An example using ensemble precipitation forecasts to maximize the expected value function for Jordan River is given in section 4.4.2.

4.4 Case Studies

4.4.1 Cost/Loss Model

Model Development

In this section we provide a practical example of the cost/loss model evaluation for specific watersheds in southwest British Columbia. The case study represents a simplification of reservoir management and operational constraints, and is intended to examine the value of probabilistic precipitation forecasts for inflow into reservoirs.

The power produced from the generator of a hydropower facility depends on the net hydraulic head available to the turbine and the flow through the hydraulic conveyance facility (e.g., canal, intake, or penstocks) (see Fig. 4.12), as given by the following equation (Gulliver, 1991):

$$P = \eta\gamma Q_t H \tag{4.9}$$

where:

P = generator power output (W)

η = turbine/generator efficiency as a fraction

γ = specific weight of water = 9807 Nm^{-3} at 5° C

Q_t = flow through the turbine (m^{3}s^{-1})

H = difference in elevation (termed "head") between the reservoir level and turbine (m)

The energy, K (in Joules) produced by the generator operating for time T (in seconds) under constant flow conditions Q_t is given by:

$$K = PT \tag{4.10}$$

The market value, ν ($) of this energy is the quantity of energy K times the selling price per unit value of energy, S ($/J):

$$\nu = KS = PTS \tag{4.11}$$

Electric energy is typically bought and sold in units of either kilowatt-hour (kWh) or megawatt-hour (MWh), and the price of electrical energy is typically quoted in dollars per MWh. The conversion factor between Joules and MWh is 1MWh=3.6x10^{9}J.

We will use the following general variables to describe the hydrometeorological forecast inflow model (we use a time-increment of one day for all measurements, calculations, and forecasts):

h_2 = the difference between the lowered reservoir elevation and the low-level intake (m)

h_3 = the difference between the reservoir at full operating level and the lowered reservoir elevation prior to an expected inflow event (m)

Q_s = the water that spills past the generator (m^3) without producing any power (hence lost revenue)

We will also use the following reservoir-specific constants in describing the model:

h_1 = nominal head [the height difference between the low-level intake and the turbine (m)]

Q_b = the daily base inflow (m^3)

A_r = the surface area of the reservoir (m^2)

G = the space in the reservoir available for storage (m^3) = $(h_2 + h_3) \cdot A_r$.

When the reservoir is full, $H = h_1 + h_2 + h_3$. If the reservoir is lowered an amount h_3 in anticipation of a rainfall-induced inflow event, the overall head H will be reduced to $H = h_1 + h_2$, resulting in less power production.

The cost/loss ratio for a particular reservoir is determined in the following way. The loss that would ensue if the reservoir was full during an inflow event is the value of the water Q_s that spills past the generator without producing any power (hence any revenue).

The amount of loss, L ($), would be:

$$L = P_L TS = \eta\gamma Q_s(h_1 + h_2 + h_3)TS \tag{4.12}$$

where P_L is lost power. Q_s is the amount of water spilled if the reservoir is not lowered a distance h_3:

$$Q_s = A_r \cdot h_3 \tag{4.13}$$

The cost of preventing the reservoir from spilling results from operating the reservoir at a lower head, $h_1 + h_2$, in anticipation of an inflow event, vs. power produced at $h_1 + h_2 + h_3$ (effectively operating at a net head of h_3):

$$C = \eta\gamma Q_t h_3 TS \tag{4.14}$$

The flow through the turbine, Q_t, is at least equal to the baseflow Q_b.

The C/L value is then:

$$C/L = \frac{Q_b h_3}{A_r h_3(h_1 + h_2 + h_3)} = \frac{Q_b}{A_r h_1 + G} \tag{4.15}$$

The C/L value can then be calculated for individual reservoirs. See Table 4.2 for the physical parameters of the reservoirs included in the study, and Table 4.3 for the associated ratios for this basic cost/loss model.

Equation (4.15) shows that small C/L values are associated with reservoirs with large surface areas, high head, large storage, and low baseflow. This type of reservoir is typically found in regions of high mountains and deep valleys where most of the reservoir inflow stems from precipitation-generated surface runoff (climatically high rainfall zones). Conversely, large C/L values are associated with reservoirs with small surface areas, low head, low storage, and high baseflow. This type of reservoir would typically be found in areas of low, rolling relief where much of the reservoir inflow derives from sub-surface aquifers (climatically low rainfall zones). The reservoirs included in this study are of the former sub-type and are characterized by very low C/L values.

We can further refine the C/L values calculated by equation (4.15). The cost of operating a reservoir at a lower head may be realized for multiple days, until the next precipitation event occurs. The length of time until the next precipitation event is estimated by $1/\overline{o}$, the inverse of the climatological frequency of the event. In this manner, C/L is defined:

$$C/L = \frac{Q_b}{\overline{o}(A_r h_1 + G)} \tag{4.16}$$

A further refinement in the cost/loss model is realized by eliminating the constraint of a constant price of energy, S. In real-time energy markets, the price of energy fluctuates hourly so that energy lost at one time and recouped at another time may not have the same value. In the cost/loss model examined here, the energy lost through spill may be valued at a contract price S_c, while the energy lost through head-loss must be replaced at a market price, S_m. In this case, C/L would be:

$$C/L = \frac{S_m}{S_c} \cdot \frac{Q_b}{\overline{o}(A_r h_1 + G)} \tag{4.17}$$

Table 4.3 also lists the C/L values for individual reservoirs (for precipitation threshold 50 mm day^{-1}) incorporating the climatological frequency of the event [equation (4.16)] and the variable price of energy [equation (4.17)].

Example

We can provide an estimate of the actual sample expense of the ensemble forecasts by evaluating the elements of Table 4.1 for a particular set of forecasts for a particular reservoir. In this example we will use the set of two cool season forecasts for Jordan River. The actual sample expense of the forecasts can be calculated in a similar manner as equation (4.2):

$$E_{actual} = a \cdot C + b \cdot C + c \cdot L + d \cdot 0 \tag{4.18}$$

where a, b, c, and d are the elements of Table 4.1 (also called hits, false alarms, misses, and correct rejections, respectively). The summary of the cost/loss table elements for Jordan River and a 50 mm day^{-1} precipitation threshold are given in Table 4.4. Separate calculations are included for the full ensemble, the ensemble average, and the best deterministic forecast. The cost/loss ratio for Jordan River that incorporates the climatological frequency of the 50 mm day^{-1} event and a constant market price for energy is $C/L = 0.068$. This C/L value is used for the calculations. The turbine/generator efficiency, η, is assumed to be 0.90 (Gulliver, 1991), and the cost of energy is assumed constant at \$50/MWh[6].

Dollar value estimates for L can be calculated from equations (4.12) and (4.13). The resulting cost estimates for cost/loss operating decisions at Jordan River based on the 50 mm day^{-1} precipitation threshold are given in Table 4.4. The full ensemble provides a net benefit (lower cost) of \$80,000 (\$40,000 annually) or 8% improvement compared to the deterministic forecasts, and a net benefit of \$545,000 (\$272,500 annually) or 57% improvement compared to the ensemble average forecasts.

4.4.2 Decision Theory Model

In this section we provide a practical example of the decision theory model for a specific watershed in southwest British Columbia. As with the cost/loss model, this case study represents a significant simplification of reservoir operating decisions, and is intended to examine the value of probabilistic precipitation forecasts derived from SREF-generated ensembles.

The energy produced by the reservoir outflow, Q_{out}, is given by equation (4.10). If the constraint to maintain a level reservoir is met, then $Q_{out} = Q_{in}$. Forecasts of day-one 24-hour precipitation and associated daily inflow observations from Jordan River on southwest Vancouver Island were used in this analysis. The linear Pearson correlation coefficient for Jordan River is 0.88 (see Fig. 4.2). Therefore the hydrologic response (inflow) to rainfall can be approximated by a linear function $Q_{in} = \kappa R = Q_{out} = Q_t$.

The energy, K, is evaluated from equations (4.9) and (4.10):

$$K = \eta\gamma\kappa RHT \tag{4.19}$$

The utility function [equation (4.7)] can then be evaluated in terms of the energy produced from different precipitation events, R. The probabilities assigned to forecasting different

[6] Source: U.S. Energy Information Administration, Form EIA-861, Annual Electric Power Industry Report. The price quoted is the latest available average wholesale price of electricity for the United States (2005).

precipitation amounts, p_k, are derived from the SREF system forecasts of 24-hour precipitation for Jordan River, and the resultant expected value is calculated from equation (4.8). The results (using day-one forecasts) are shown in Fig. 4.13 where profit[7] (relative to perfect forecasts) is shown as a function of the market price to contract price ratio S_m/S_c. For comparison purposes, Fig. 4.13 also includes the profit from operating the reservoir using climatology to assign probabilities, as well as the SREF ensemble average and the deterministic forecasts from the MM5-4 km forecast model.

The results show that when the contract price is equivalent to the market price, there is little difference whether using the full ensemble, the ensemble average, or the deterministic single forecast to provide probability estimates. Note that using climatologic representative probabilities results in a loss for all values of S_m/S_c. However, as the spread between the market price and contract price widens and the S_m/S_c ratio increases, a different profit scenario emerges. The ensemble average forecast provides the greatest profit, followed closely by the full 11-member ensemble. The profit realized by using the deterministic high-resolution forecast drops quickly relative to the SREF forecasts.

4.5 Conclusion

Weather forecasts are made and used under the assumption that forecast users can take preventative action, prior to a predicted weather event, that provides more benefit to the user than having no forecast at all. This assumption should be specifically tested to aid weather forecast providers in improving their forecasts, and to assist weather forecast users in assessing the benefit of the forecasts to them.

Forecasts can be tested by assessing both their meteorological skill and their economic value. Forecast skill is generally determined by assessing the accuracy of a forecast-producing system relative to a given standard such as climatology, persistence, or another forecasting system. Forecast skill will vary among different forecast-producing systems, but remains user-independent. The companion paper assessed the meteorological skill of a short-range ensemble forecast system designed to provide one and two-day precipitation forecasts in complex mountainous and coastal terrain.

Forecast value is a separate assessment of a forecasting system that takes into account diverse needs and benefits attributed to different forecast users. This paper comprises

[7] Profit as used in this context is the difference between generation revenue and contract costs for undelivered energy. There are many other costs associated with hydropower generation, such as water license fees and transmission charges, that are not included in the profit calculations here.

an economic value-assessment of the same set of forecasts that provided the basis for the meteorological skill assessment described in the companion paper.

The particular forecast users studied here are water resource managers tasked with making efficient and effective use of rainfall that ultimately drives hydro-electric power in rainfall-fed mountain-coastal reservoirs.

Employing a static cost/loss economic model adapted to reservoir operations for hydro-electric managers, a forecast system that provides a suite of ensemble members provides positive economic value for a broader range of users than either a single fine-resolution forecast model or a single forecast composed of the ensemble average. Secondly, the study showed that different forecast users, characterized by specific cost/loss ratios, require different precipitation probability thresholds to trigger preventative actions. Users with different cost/loss ratios correspond to operators of reservoirs with different head-loss parameters, different sizes of reservoirs, and different base-flows (e.g., lower C/L value applies to reservoirs with low base inflows, high head, large areal reservoirs, and high storage, typical of rainfall-fed mountain watersheds). The larger economic benefit of probabilistic forecasts is achieved by allowing users to match forecast probabilities with cost/loss ratios specific to their particular operational needs.

Decision theory, employing a utility function designed specifically for market selling of hydro-electric energy from rapid-response rainfall-dominated reservoirs, was used to quantify the positive impact (profit generation) from employing an EPS to generate probabilistic forecasts. The forecasts generated by the SREF system analyzed in Part I are superior to a single, high-resolution deterministic forecast in maximizing profit from the particular hydro-electric facility (Jordan River) studied here, with the ensemble-averaged forecasts outperforming the full suite of ensembles by a relatively small margin.

Examining the results from the cost/loss model and the decision theory model provides insight into how different economic models can yield complementary information from numerical weather models. Modes of operations, operational constraints, and economic operating environments will likely vary among different hydro-electric reservoirs; therefore different economic models may be required to assess operating strategy for different reservoir operations. In section 4.4.1, the cost/loss model was applied to reservoir operations in order to minimize spill and head loss, with no constraints on reservoir level (the reservoir could be lowered in anticipation of an inflow event without regard to other concerns or stakeholders). In the cost/loss example, the SREF employing the full ensemble provided the greatest benefit, followed fairly closely by the deterministic model forecast, with the ensemble average forecast far behind. In section 4.4.2, a decision theory model was applied to reservoir operations under an added constraint on reservoir level (the reservoir level must be maintained at a constant level) and a set deliverable contract price (and variable market purchase price) for energy. In this case the ensemble average provided the greatest benefit, followed closely by the full ensemble, and the deterministic forecast lagged far behind. Therefore the system constraints (on reservoir operations and market pricing) may partially dictate what economic model best fits operational requirements, and an economic analysis based on that particular economic model may dictate what numerical weather guidance to choose to maximize economic benefit. Other valid economic models, such as a decision tree model (Ang and Tang, 1984; ReVelle et al., 2004), may prove valuable in future research incorporating many complex constraints and alternative utility functions and responses into a decision-making model.

Employing economic models such as a static binary cost/loss model, or a continuous decision-theory model, is a preliminary step in assessing the value of ensemble-based precipitation forecasts to water resource managers. The ultimate goal of ensemble forecasting for reservoir operating planners is to be able to translate the uncertainty in future weather into uncertainty in runoff and inflow, and ultimately the amount of energy generated at a particular reservoir/hydro-electric facility [related examples can be found in Carpenter and Georgakakos (2001) and Yao and Georgakakos (2001)]. The uncertainty in overall inflow must also incorporate uncertainty in the hydrologic model that translates forecast precipitation and temperature into reservoir inflow. Future research will include a hydrometeorologic model driven by probabilistic meteorological input variables to determine inflows, and ultimately available energy.

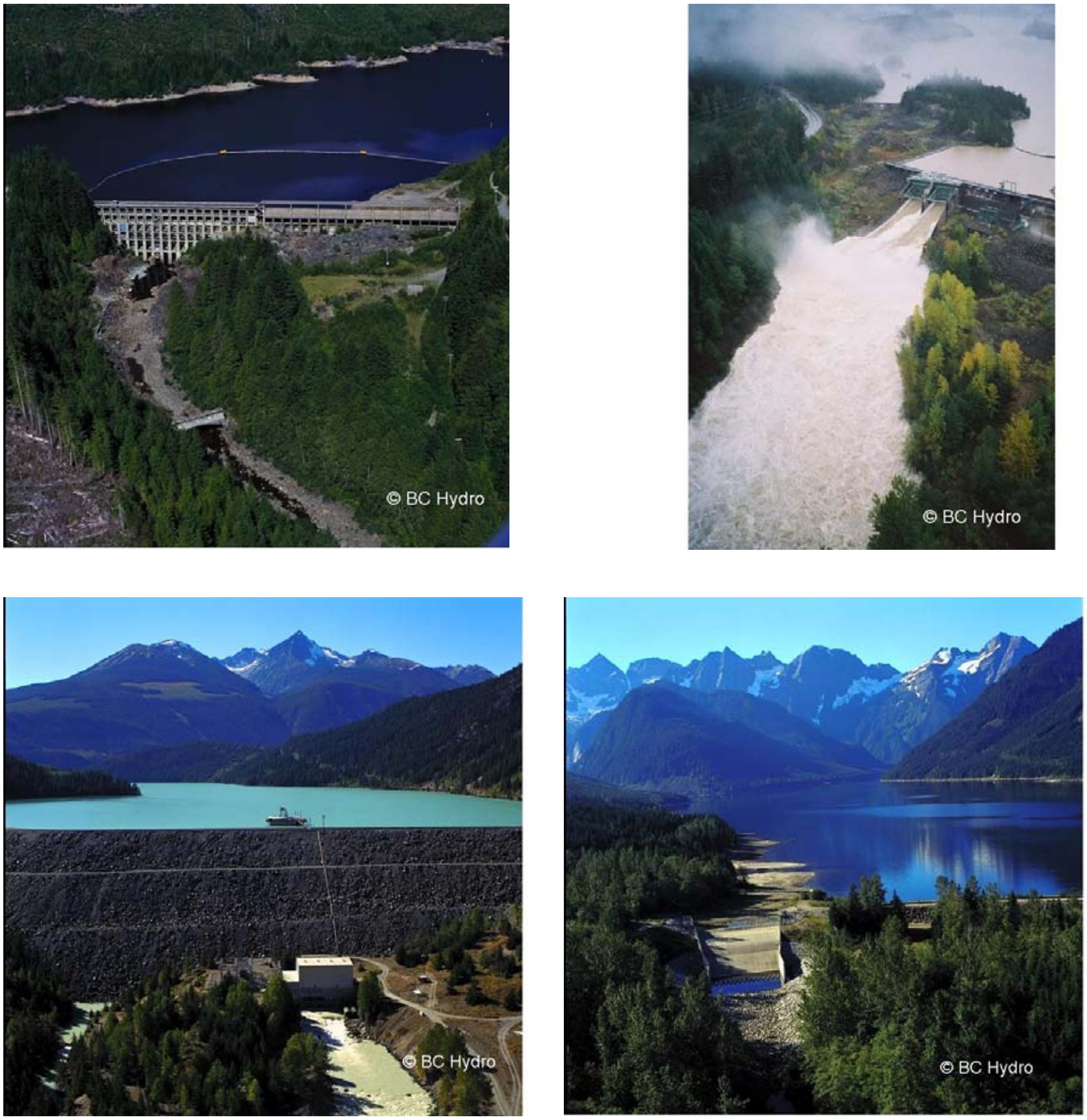

Figure 4.1: Photos of four BC Hydro dams in southwestern British Columbia. Clockwise from upper left: Jordan dam, Daisy dam, Wahleach dam and La Joie dam. Daisy dam is seen spilling 700 m^3s^{-1} after very heavy rain in mid October 2003.

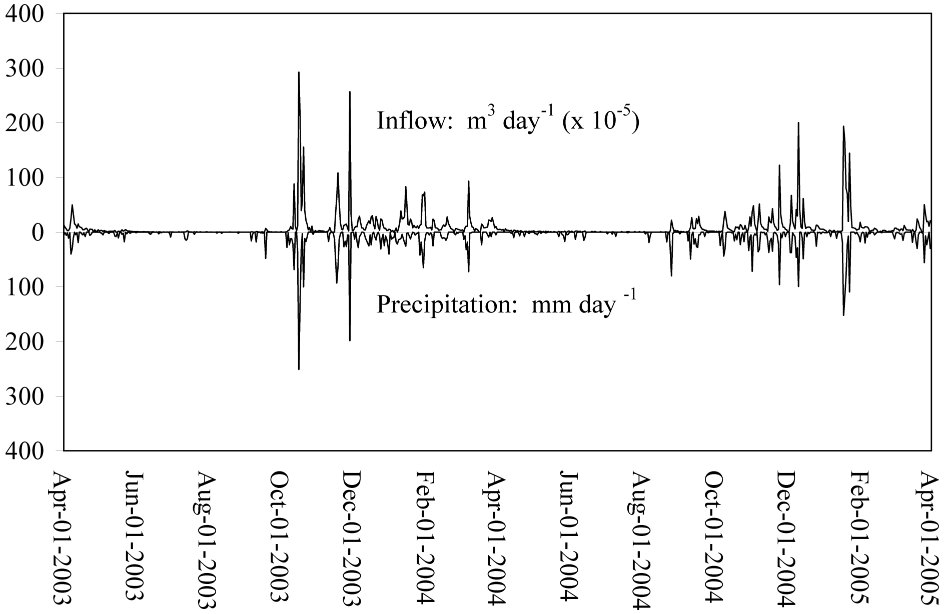

Figure 4.2: Daily precipitation vs. daily inflow for Jordan River on southwest Vancouver Island. The daily inflow (top graph, flow rate in m^3 day^{-1}) is determined from reservoir elevation and outflow measurements for the Jordan River reservoir system. The daily precipitation (bottom graph, in mm day^{-1}) was observed at a precipitation gauge in the middle of the watershed. The period of observations is from April 2003 through March 2005. The Pearson correlation coefficient for these two graphs is 0.88, indicating a high correlation between daily precipitation and daily reservoir inflow. For reservoirs throughout the study region the average Pearson correlation coefficient between daily precipitation and daily inflow is 0.70.

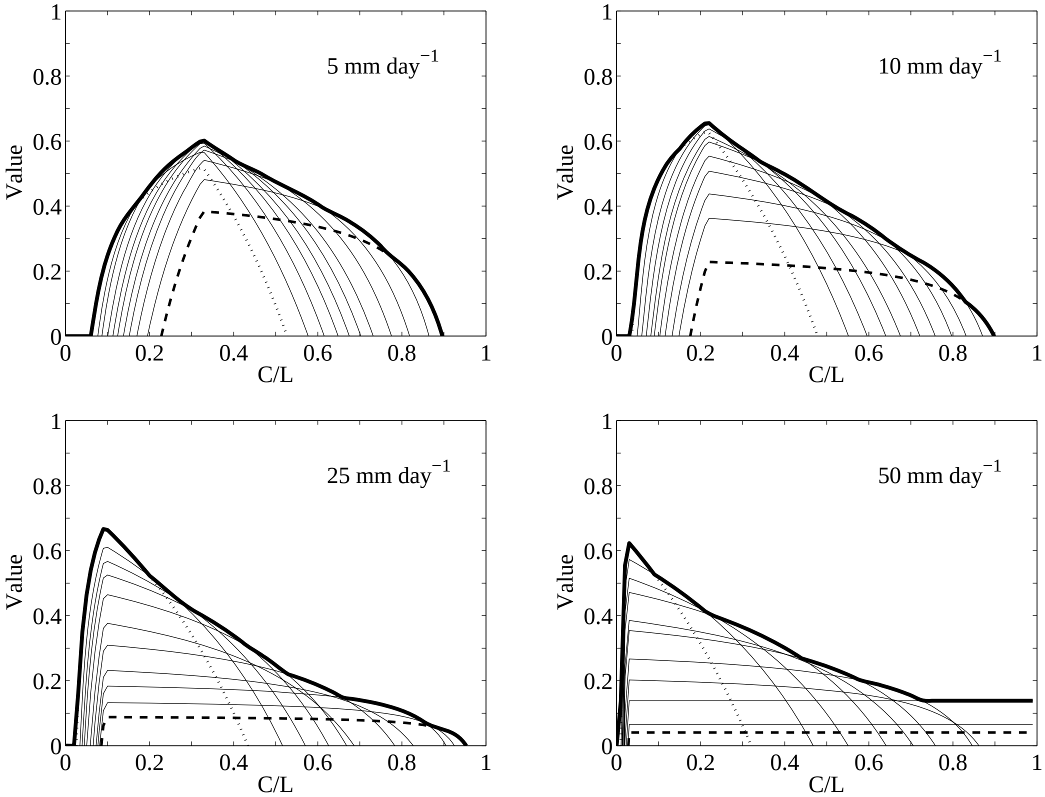

Figure 4.3: Forecast value as a function of cost/loss ratio (C/L) for day-one forecasts. Precipitation thresholds are as indicated on the figures. The thin curves represent forecast value calculated for different probability thresholds p_t. The thin curves begin with the lowest probability threshold $p_t = 1/11$ (dotted line) and continue through increasing values of p_t ending with the highest threshold $p_t = 11/11$ (dashed line). The heavy solid line shows the envelope curve of optimal value (where each user chooses the probability threshold that maximizes the value for their specific cost/loss ratio).

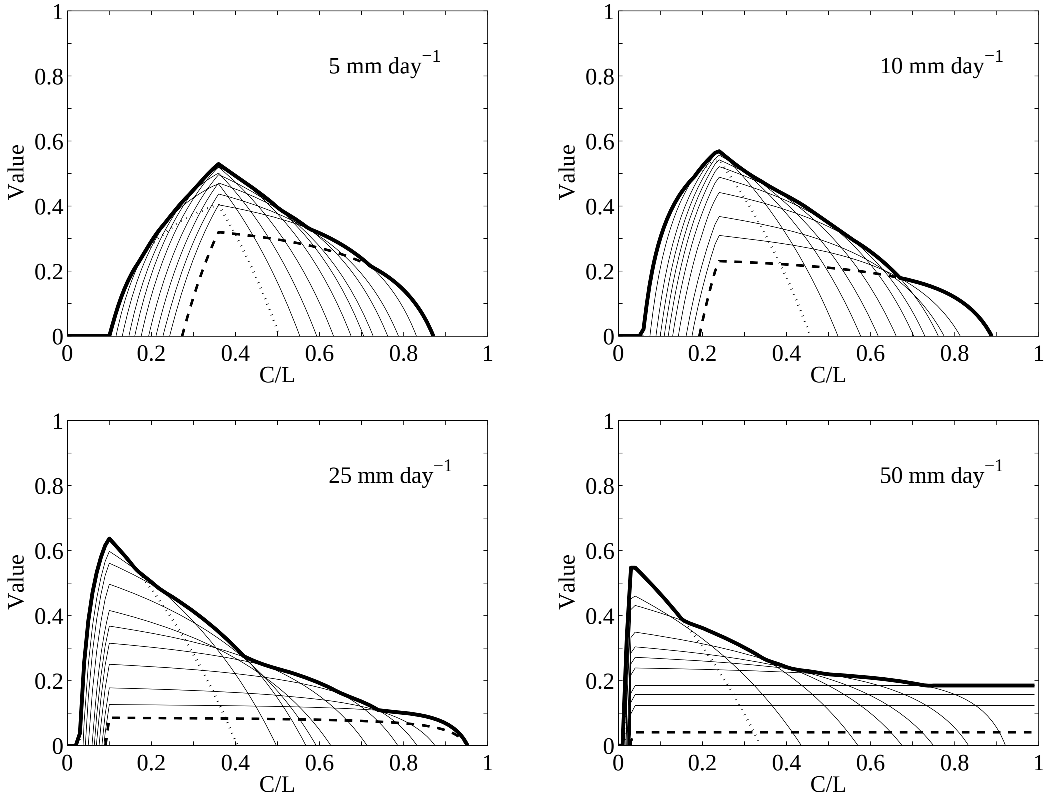

Figure 4.4: Forecast value as a function of cost/loss ratio (C/L) for day-two forecasts. The information presented is in the same form as for Fig. 4.3.

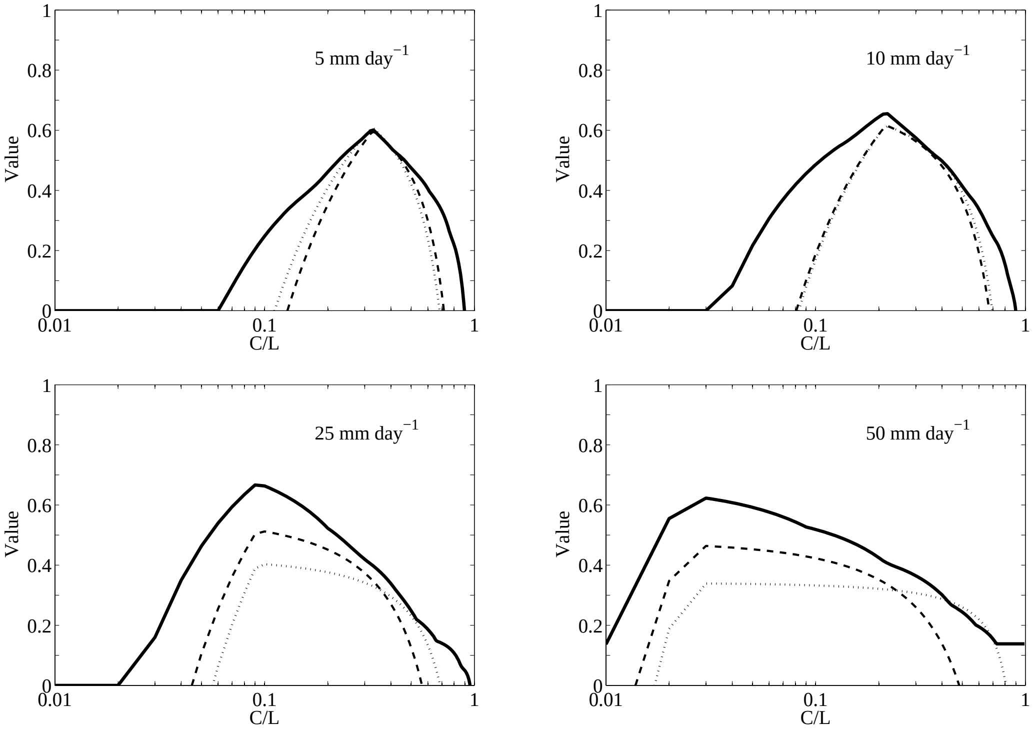

Figure 4.5: Forecast value envelope as a function of C/L for day-one forecasts. Precipitation thresholds are as indicated on the figures. The dashed line is the deterministic high-resolution forecast and the dotted line is the ensemble-average forecast (the abscissa is plotted with a logarithmic axis for C/L to enhance clarity). Since the maximum value occurs at the climate frequency of the threshold event, the curves for the more extreme events are concentrated around lower C/L values.

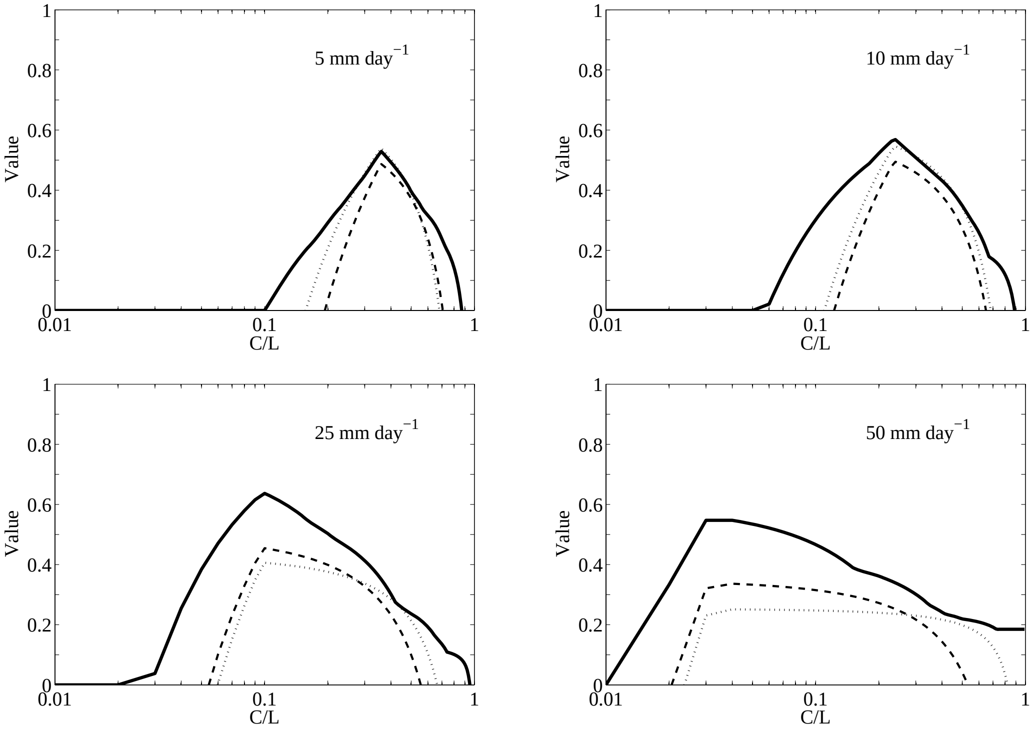

Figure 4.6: Forecast value envelope as a function of C/L for day-two forecasts. The information presented is in the same form as in Fig. 4.5.

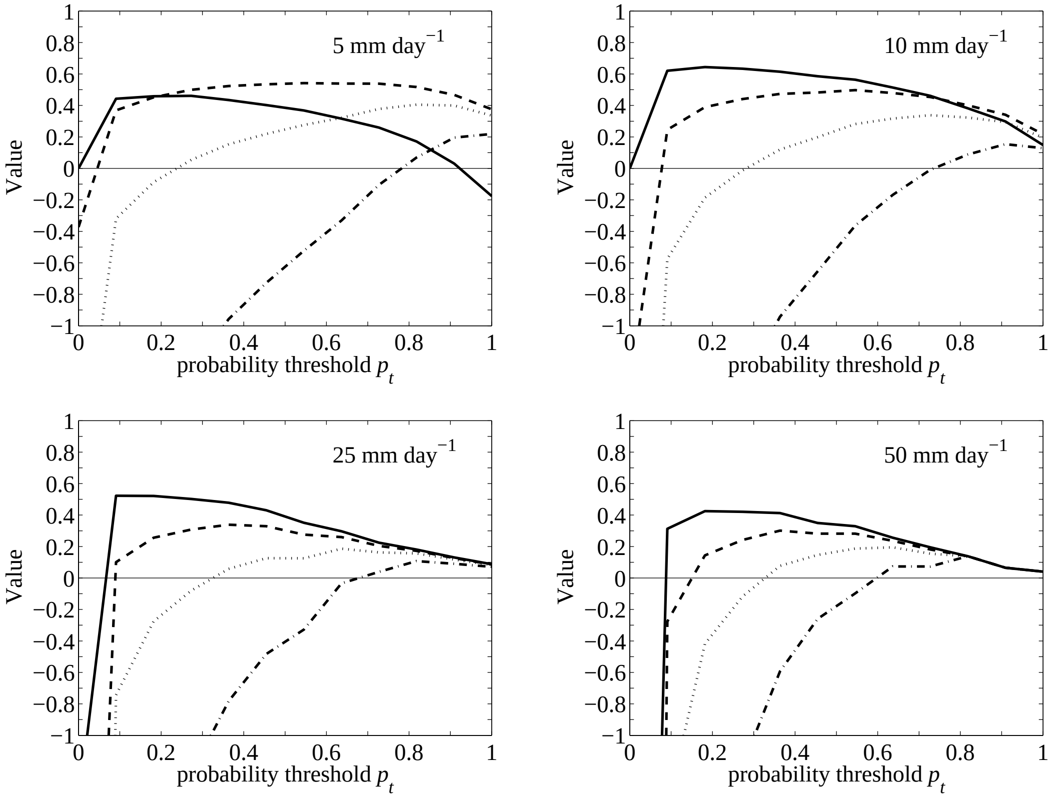

Figure 4.7: Forecast value as a function of probability threshold for day-one forecasts. Precipitation thresholds are as indicated on the figures. Different cost/loss ratios are shown as a solid line ($C/L = 0.2$), dashed line ($C/L = 0.4$), dotted line ($C/L = 0.6$), and dash-dot line ($C/L = 0.8$). A value of 1.0 indicates perfect forecasts for users while values below zero indicate that climatological forecasts are more valuable. The graphs show that users with higher C/L values must wait for a higher probability forecast of the given event from the suite of ensembles (i.e., the likelihood of the event occurring is more certain) before taking action to receive value from the forecasts.

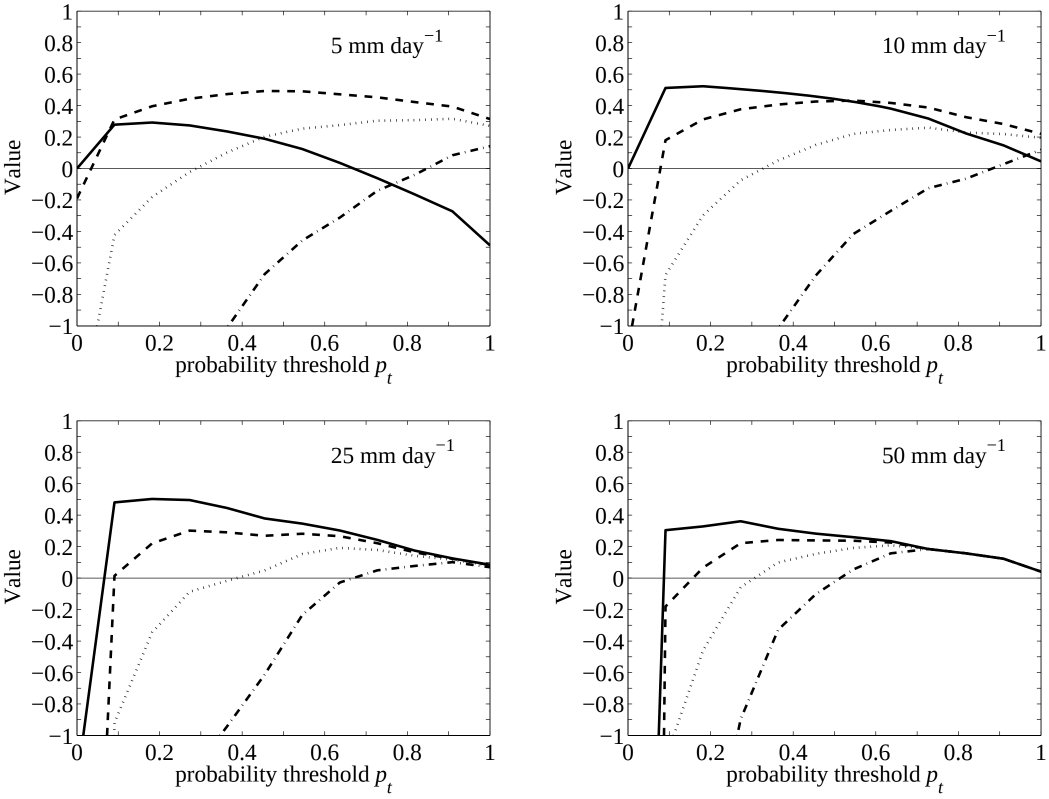

Figure 4.8: Forecast value as a function of probability threshold for day-two forecasts. The information presented is in the same form as for Fig. 4.7.

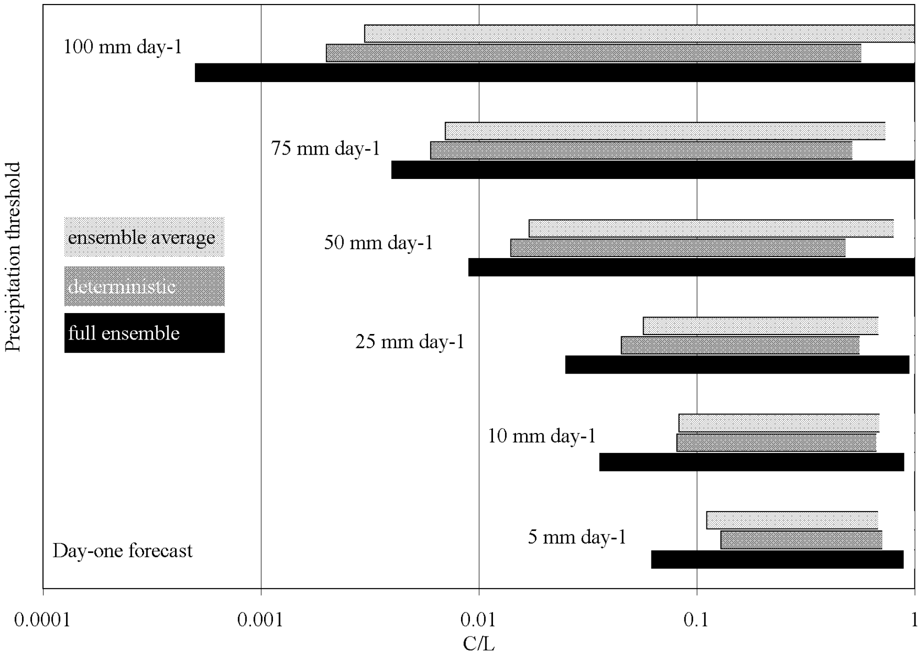

Figure 4.9: Each bar in the figure spans the portion of the C/L range having positive value for different SREF configurations and for different precipitation thresholds for the day-one forecast. The bars are labelled with the SREF configurations for the full ensemble, the very-high resolution deterministic forecast, and the ensemble-average forecast. Longer bars provide value to a wider range of users.

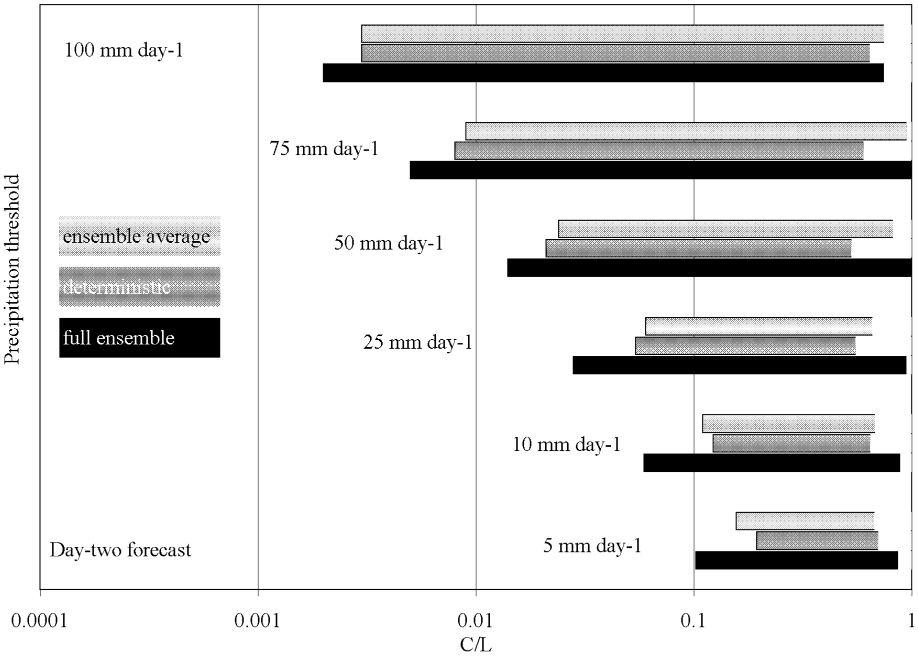

Figure 4.10: Same as Fig. 4.9 except for the day-two forecast.

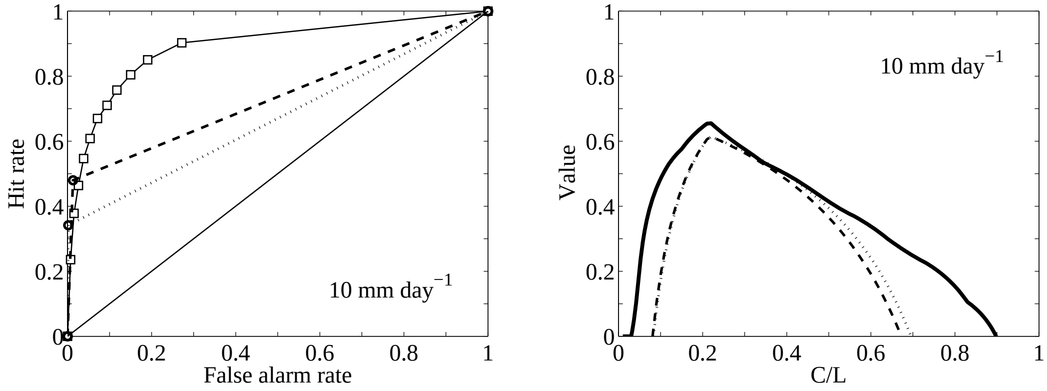

Figure 4.11: The relationship between ROC and forecast value. Curves are for day-one forecasts with precipitation threshold of 10 mm day^{-1}. The ROC curve in the left figure with the SREF probability threshold p_t square markers corresponds to the thick solid envelope value curve in the right figure. The dashed line in both figures is the very-high resolution deterministic forecast and the dotted line in both figures is the ensemble-average forecast. The open circles on the dashed line and the dotted line in the figure on the left represent the very-high resolution deterministic forecast point and the ensemble-average forecast point, respectively. The relative merit of ROC and forecast value as determined by these curves is discussed in the text.

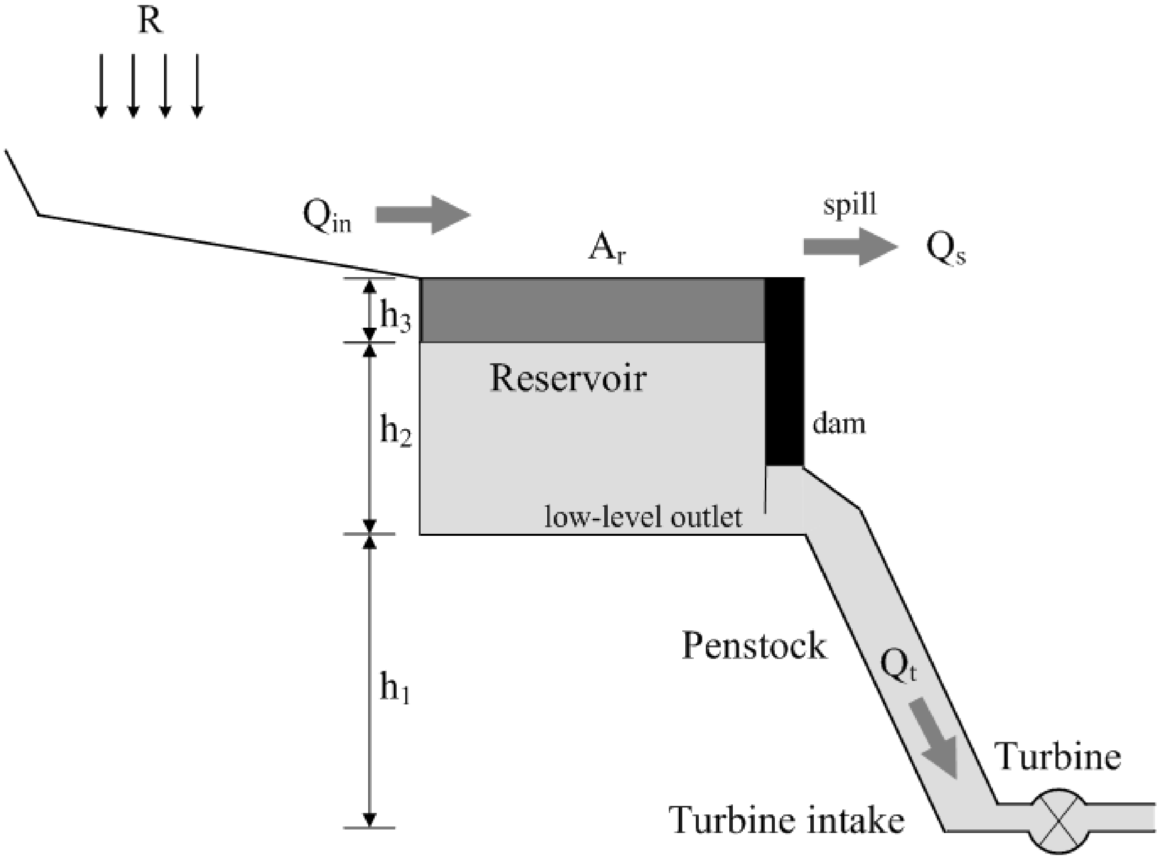

Figure 4.12: The reservoir schematic diagram for the cost/loss economic model discussed in the text.

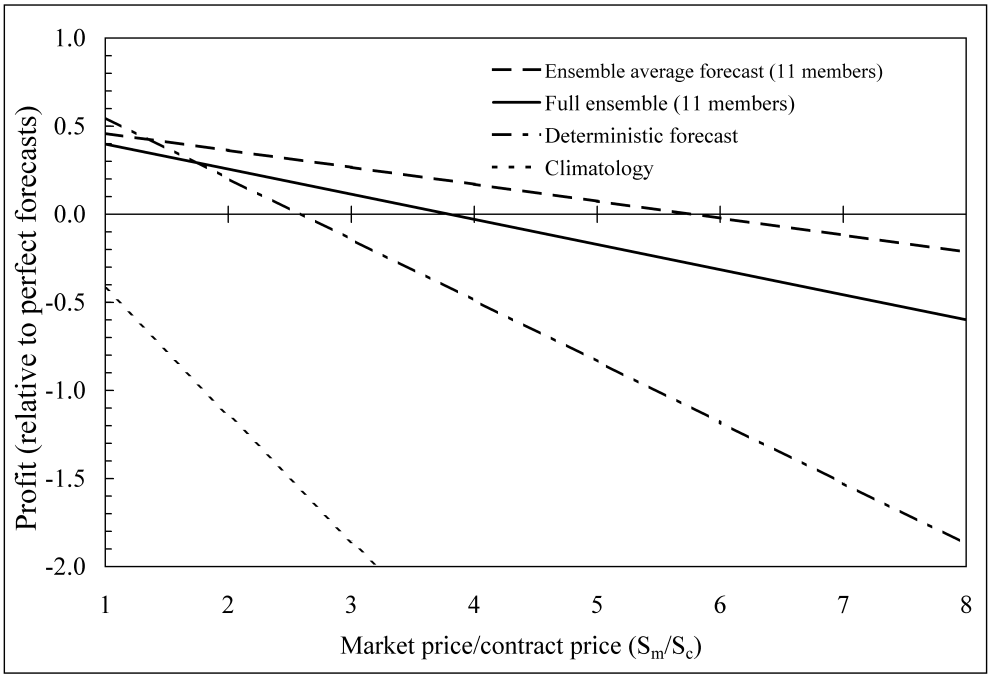

Figure 4.13: Profit as a function of the market/contract price ratio for different forecast configurations. The profit for each different configuration is calculated relative to a baseline maximum possible profit (if the weather forecasts were perfect).

Table 4.1: Cost/loss contingency table of precipitation forecasts and observations. The number of hits is given by a, b refers to false alarms, c is number of misses, and d is the number of correct rejections. Actions resulting from a forecast incur a cost C, while losses L result from each observed event that was not forecast. Correct rejections incur no cost and result in no loss. An event is when the precipitation exceeds a threshold.

	Observed	Not observed
Forecast	a (C)	b (C)
Not forecast	c (L)	d (0)

Table 4.2: Reservoir physical parameters for cost/loss calculations.

		Daily base	Nominal	Reservoir	Reservoir
		inflow	head	storage	area
		Q_b	h_1	G	A_r
Index	Reservoir	(x10^6 m^3)	(m)	(x10^6 m^3)	(km^2)
1	Alouette	1.8	38	155	16
2	Buntzen	2.0	116	202	1.8
3	Carpenter	4.4	340	928	48
4	Clowhom	3.1	44	45	8.0
5	Comox	2.9	104	106	30
6	Coquitlam	2.0	116	202	12
7	Daisy	4.2	291	46	43
8	Downton	3.5	54	722	24
9	Elsie	1.8	224	77	0.8
10	Hayward Lake	0.86	38	24	3.0
11	Jordan River	1.0	265	28	1.7
12	Seton	1.6	45	9.0	25
13	Stave	9.6	38	36	62
14	Upper Campbell	6.7	43	82	67
15	Wahleach	0.60	573	66	4.9

Table 4.3: Cost/loss ratios for the reservoirs in this study.

Index	Reservoir	basic model	including $\overline{o}$ (50 mm day^{-1} threshold)	including market price value $S_m/S_c = 2.5$ (50 mm day^{-1} threshold)
1	Alouette	0.0024	0.074	0.18
2	Buntzen	0.0048	0.15	0.37
3	Carpenter	0.00026	0.0080	0.020
4	Clowhom	0.0078	0.24	0.61
5	Comox	0.00091	0.029	0.071
6	Coquitlam	0.0012	0.038	0.094
7	Daisy	0.00034	0.011	0.026
8	Downton	0.0017	0.054	0.13
9	Elsie	0.0074	0.23	0.58
10	Hayward Lake	0.0063	0.20	0.50
11	Jordan River	0.0022	0.068	0.17
12	Seton	0.0014	0.044	0.11
13	Stave	0.0035	0.11	0.27
14	Upper Campbell	0.0018	0.057	0.14
15	Wahleach	0.00021	0.0066	0.016

Table 4.4: Cost/loss contingency table summary of actual precipitation forecasts and observations for Jordan River for the 50 mm day^{-1} precipitation threshold. The summary includes the full ensemble, the deterministic forecasts from the MM5-4 km model, and the ensemble average. The actual sample expense, E_{actual}, is determined as a function of L through the cost/loss ratio for Jordan River ($C/L = 0.068$). The cost is estimated by evaluating L as explained in the text.

	hit	false	miss	E_{actual}	cost
		alarm			estimate
	(a)	(b)	(c)		
Full ensemble	4	0	3	3.27L	$955,000
Deterministic	4	4	3	3.54L	$1,035,000
Ensemble average	2	0	5	5.14L	$1,500,000

Bibliography

Anderson-Berry, L., T. Keenan, J. Bally, R. Pielke Jr., R. Leigh, and D. King, 2004: The societal, social, and economic impacts of the World Weather Research Programme Sydney 2000 Forecast Demonstation Project (WWRP S2000 FDP). *Wea. Forecasting*, **19**, 168–178.

Ang, A. H.-S. and W. H. Tang, 1984: *Probability concepts in engineering planning and design*, John Wiley and Sons, New York, volume I, chapter 2.

Baldi, P., 2001: *The Shattered Self: The End of Natural Evolution*. The MIT Press, Cambridge MA.

Carpenter, T. M. and K. P. Georgakakos, 2001: Assessment of Folsom lake response to historical and potential future climate scenarios: 1. Forecasting. *J. Hydrol.*, **249**, 148–175.

Gulliver, J. S., 1991: Hydraulic conveyance design. *Hydropower Engineering Handbook*, J. S. Gulliver and R. E. A. Arndt, eds., McGraw-Hill, Inc., 5.1–5.81.

Hamill, T. M., S. L. Mullen, C. Snyder, Z. Toth, and D. P. Baumhefner, 2000: Ensemble forecasting in the short to medium range: Report from a workshop. *Bull. Amer. Meteor. Soc.*, **81**, 2653–2664.

Katz, R. W. and A. H. Murphy, 1997: *Economic Value of Weather and Climate Forecasts*. Cambridge University Press, England.

Laffont, J.-J., 1989: *The Economics of Uncertainty and Information*. MIT Press, Cambridge MA.

Legg, T. P. and K. R. Mylne, 2004: Early warnings of severe weather from ensemble forecast information. *Wea. Forecasting*, **19**, 891–906.

— 2007: Corrigendum. *Wea. Forecasting*, **22**, 216–219.

McCollor, D. and R. Stull, 2008: Hydrometeorological short-range ensemble forecasts in complex terrain. Part I: Meteorological evaluation. *Wea. Forecasting*, (in press).

Mullen, S. L. and R. Buizza, 2002: The impact of horizontal resolution and ensemble size on probabilistic forecasts of precipitation by the ECMWF ensemble prediction system. *Wea. Forecasting*, **17**, 173–191.

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

Pagowski, M. and G. A. Grell, 2006: Ensemble-based ozone forecasts: skill and economic value. *J. Geophys. Res.*, **111**, D23S30, doi:10.1029/2006JD007124.

ReVelle, C., E. Whitlatch, and J. Wright, 2004: *Civil and environmental systems engineering*, Prentice Hall, Englewood Cliffs, chapter 9. 2nd edition.

Richardson, D. S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **126**, 649–667.

— 2003: Economic value and skill. *Forecast verification: A practitioner's guide in atmospheric science*, I. Jolliffe and D. B. Stephenson, eds., Wiley, 164–187.

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

Roulston, M. S., G. E. Bolton, A. N. Kleit, and A. L. Sears-Collins, 2006: A laboratory study of the benefits of including uncertainty information in weather forecasts. *Wea. Forecasting*, **21**, 116–122.

Smith, L. A., M. S. Roulston, and J. von Hardenberg, 2001: End to end forecasting: Towards evaluating the economic value of the ensemble prediction system. Technical Report 336, ECMWF Technical Memorandum.

Stensrud, D. J. and N. Yussouf, 2003: Short-range ensemble predictions of 2-m temperature and dewpoint temperature over New England. *Mon. Wea. Rev.*, **131**, 2510–2524.

Thornes, J. E. and D. B. Stephenson, 2001: How to judge the quality and value of weather forecast products. *Meteorol. Appl.*, **8**, 307–314.

Yao, H. and A. Georgakakos, 2001: Assessment of Folsom Lake response to historical and potential future climate scenarios 2. Reservoir management. *J. Hydrol.*, **249**, 176–196.

Yuan, H., S. Mullen, X. Gao, S. Sorooshian, J. Du, and H.-M. H. Juang, 2005: Verification of probabilistic quantitative precipitation forecasts over the southwest United States during winter 2002/03 by the RSM ensemble system. *Mon. Wea. Rev.*, **133**, 279–294.

Zhu, Y., Z. Toth, R. Wobus, D. Richardson, and K. Mylne, 2002: The economic value of ensemble-based weather forecasts. *Bull. Amer. Meteor. Soc.*, **83**, 73–83.

Chapter 5

Evaluation of Probabilistic Medium-Range Temperature Forecasts from the North American Ensemble Forecast System[8]

5.1 Introduction

Accurate weather forecasts help mitigate risk to those economic sectors exposed to weather-related business losses. Astute business operators can also incorporate accurate weather forecasts for financial gain for their organization. Probabilistic forecasts are especially useful in making economic decisions because the expected gain or loss arising from a particular weather-related event can be measured directly against the probability of that event actually occurring. For the example of hydro-electric systems, temperature affects the demand for electricity (from electric heaters and air conditioners) and the resistance (therefore maximum electric load capacity) of transmission lines. In high-latitude countries river-ice formation affects hydro-electric generation planning during winter. Temperature effects on hydro-electric systems also include the height above the ground of transmission lines (lines sag closer to the ground as temperatures rise) and precipitation type (rain vs. snow), an important factor determining inflow response in watersheds.

Of course, weather-forecast accuracy diminishes as the forecast lead time increases; this is true for probabilistic forecasts as well as deterministic forecasts. This paper aims to analyze a set of mid-range probabilistic temperature forecasts and measure the quality and

[8] A version of this chapter has been submitted for publication. McCollor, D. and R. Stull: Evaluation of Probabilistic Medium-Range Temperature Forecasts from the North American Ensemble Forecast System.

value of the forecasts as a function of lead time. Decision makers using this forecast system will then have a valid estimate of how far into the future these particular forecasts provide valuable information.

This chapter builds on the probabilistic ensemble prediction system (EPS) analysis techniques incorporated in chapter 3 by extending the evaluation of surface-based numerical temperature forecasts through the mid-range and out to the limit of existing operational numerical weather guidance. This chapter includes a further refinement of the Brier Skill Score (the continuous ranked probability skill score) to analyze the forecasts. This chapter builds on the post-processing techniques introduced in chapter 2 to bias correct these mid- and long-range ensemble forecasts. This chapter also includes an economic evaluation component using the cost/loss model that was integrated into the analysis of chapter 4. A future research project is planned to investigate the meteorological skill and economic value of probabilistic precipitation forecasts from EPS mid-range forecasts.

Probability estimates of future weather are most often derived from an ensemble of individual forecasts. Many short-range ensemble forecast (SREF) systems, designed to provide forecasts in the day-one and day-two timeframe, provide high-quality probabilistic forecasts by varying models and/or model physics to derive their ensemble members (Stensrud et al., 2000; Eckel and Mass, 2005; Jones et al., 2007; McCollor and Stull, 2008b). Medium-range ensemble forecast (MREF) systems, necessary for probabilistic forecasts beyond day two, can be designed to represent the errors inherent in synoptic flow patterns in the mid-range period (Buizza et al., 2005; Roulin and Vannitsem, 2005; Tennant et al., 2007).

The idea of combining ensemble forecasts from different national meteorological centers, thus taking advantage of different modes of ensemble member generation and greatly increasing overall ensemble size, resulted in the formation of the North American Ensemble Forecast System (NAEFS) (Toth et al., 2006). We investigate the quality and value of NAEFS daily maximum and minimum temperature forecasts for ten cities in western North America.

Section 5.2 describes the forecasts and observations analyzed in the MREF study. Section 5.3 describes the verification metrics used to evaluate the probabilistic forecasts and discusses the results, and section 5.4 summarizes the findings and concludes the paper.

5.2 Methodology

Temperature forecasts were obtained and stored in real time via public access to the North American Ensemble Forecast System. The NAEFS is a joint project involving the Meteorological Service of Canada (MSC), the United States National Weather Service (NWS) and the National Meteorological Service of Mexico (NMSM). NAEFS was officially launched in November 2004 and combines state-of-the-art ensembles developed at MSC and NWS.

The grand- or super-ensemble provides weather-forecast guidance for the 1-16 day period, allowing for inclusion of ensemble members from different generating schemes. Note that at the time this study was done, NAEFS products were considered experimental in nature.

The original configuration of the NWS-produced ensembles consisted of one control run plus 14 perturbed ensemble members produced by a bred-vector perturbation method (Buizza et al., 2005; Toth and Kalnay, 1993). This initial-condition perturbation method assumed that fast-growing errors develop naturally in a data assimilation cycle and that errors will continue to grow through the medium-range forecast cycle. The original MSC EPS configuration consisted of one control run plus 16 perturbed ensemble members. The perturbation approach developed at MSC (Houtekamer et al., 1996; Buizza et al., 2005) generated initial conditions by assimilating randomly perturbed observations. This multi-model approach ran different model versions (involving variable physical parameterizations) incorporating a number of independent data assimilation cycles. At the time of this study, no forecasts were available from the NMSM.

In addition, it must be noted that ensemble system design, including strategies to sample observation error, initial condition error, and model error, as well as addressing the issue of member resolution vs. ensemble size, is continually evolving. Recent information describing current national ensemble prediction systems can be found in a report from a November 2007 workshop on ensemble prediction (available online at http://www.ecmwf.int/newsevents/meetings/workshops/2007/ensemble_prediction/wg1.pdf, accessed April 12, 2008). Current details on improvements to the MSC EPS are also available online (at http://www.ecmwf.int/newsevents/meetings/workshops/2007/MOS_11/presentations_files/Gagnon.pdf, accessed April 14 2008).

In May 2006 the NWS EPS changed its initial perturbation generation technique to the Ensemble Transform with rescaling (ETR) technique, where the initial perturbations are restrained by the best available analysis variance from the operational data assimilation system (Wei et al., 2008). In March 2007 the NWS ensemble increased to 20 members integrated from an 80-member ETR-based ensemble. In July 2007 MSC increased its ensemble size from 16 to 20 members (see http://www.smc-msc.ec.gc.ca/cmc/op_systems/doc_opchanges/genot_20070707_e_f.pdf, accessed April 17, 2008) to conform to the NWS ensemble size. MSC also increased the horizontal resolution of the members and extended the physical parameterization package. The complete current status of the MSC EPS can be found at http://www.ecmwf.int/newsevents/meetings/workshops/2007/ensemble_prediction/presentations/houtekamer.pdf (accessed April 17, 2008).

Forecasts for day one through day fifteen were retrieved from the on-line NAEFS over the period 15 February 2007 through 1 October 2007. An ensemble size of 32 members was maintained throughout this period, to maintain consistency in the process, despite operational changes at the national centers as described above. Ten city locations (five in Canada and five in the U.S.) were chosen to represent different climate regimes in far western North America (see Table 5.1 for the list of North American cities in the study). For each day and each city, the forecast valid at 00 UTC [4 pm local (Pacific Standard) time] was verified against the observed maximum temperature; the forecast valid at 12 UTC (4 am local time) was verified against the observed minimum temperature.

Forecast-system performance can be verified against actual point observations or against model-based gridded analyses. Gridded analyses directly from the models generally offer more data for investigation. However gridded analyses are subject to model-based data-estimation errors and smoothing errors. Even though point observations are subject to instrumentation error and site-representativeness error, it is preferable to verify forecasts against observations as opposed to model analyses (Hou et al., 2001).

The mean error of the maximum and minimum temperature forecasts for each ensemble member for each forecast day was evaluated from the sample. Random errors caused by initial-condition and model/parameterization errors can be reduced by using an ensemble of forecasts. Bias in the forecasts is largely attributable to local terrain and topographic effects poorly represented in the forecast models. Bias is evaluated by calculating the mean error in the forecasts, and can be reduced via post-processing of the direct model output (DMO).

Regarding post-processing, the authors previously found (McCollor and Stull, 2008a) that a 14-day moving-average filter was effective in reducing bias in a medium-range maximum and minimum temperature forecast sample. Therefore a 14-day moving-average filter is applied here to each of the ensemble member forecasts (for each station location and each forecast lead time) to minimize the bias in the DMO forecasts. Care was taken in this study to ensure the method is operationally achievable. That is, only observations that would be available in a real-time setting for each forecast period were used in the moving-average filter. For example, all day-five forecasts included a moving-average filter incorporating forecast-observation pairs from 19 days prior to the forecast valid time through 6 days prior to the forecast time to achieve a 14-day moving average bias correction.

This method of error reduction for ensemble prediction is known as the bias-corrected ensemble (BCE) approach (Yussouf and Stensrud, 2007). The scheme responds quickly to any new station locations, new models or model upgrades. The results of the filter application are presented in Fig. 5.1. Mean error in the DMO ensemble average daily maximum temperature forecasts, ranging from -3.6°C (day-one forecasts) to -2.6°C (day 15 forecasts), was reduced to 0.0°C (day-one forecasts) to +0.6°C (day 15 forecasts) in the post-processed forecasts. Similarly, mean error in the DMO ensemble average daily minimum temperature forecasts, near -1.5°C throughout the 15-day forecast cycle, was reduced to the 0.0°C to +0.3°C range in the post-processed forecasts. These bias-corrected forecasts are used in the subsequent ensemble evaluations in this paper.

The final phase in designing an ensemble prediction system lies in choosing a weighting scheme for the individual ensemble members in the computation of the ensemble average. In addition to applying equal weighting to all ensemble members by simply averaging all members, Yussouf and Stensrud (2007) explored performance-based weighted BCE schemes wherein unequal weights, based on individual member's past forecast accuracy, are assigned to each ensemble member prior to calculating the ensemble mean forecast. However, results from the Yussouf and Stensrud (2007) study indicated there is no significant improvement over original 2 m temperature BCE forecasts when performance-based weighting schemes are applied to the EPS individual members. Therefore a BCE approach computing the ensemble average from equally-weighted ensemble members is chosen for our study of medium-range temperature forecasts.

5.3 Results and Discussion

A single verification score is generally inadequate for evaluating all of the desired information about the performance of an ensemble prediction system (Murphy and Winkler, 1987; Murphy, 1991). Different measures, emphasizing different aspects and attributes of forecast performance, should be employed to assess the statistical reliability, resolution, and discrimination of an EPS. A standardized set of evaluation methods, scores, and diagrams to interpret ensemble forecasts is provided in Hamill et al. (2000) and is incorporated here to assess a MREF system for daily maximum and minimum temperature forecasts. The reader is referred to Wilks (2006) or Toth et al. (2003) for comprehensive descriptions of verification measures.

5.3.1 Skill Scores

Initial assessment of the forecasts utilized a root mean squared error skill score (RMSE_{SS}) and a Pearson correlation skill score (CORR_{SS}) evaluated for the ensemble average. A general skill score definition is given by:

$$\text{Skill score} = \frac{\text{score} - \text{score}_{ref}}{\text{score}_{perfect} - \text{score}_{ref}} \tag{5.1}$$

where score_{ref} is the score for the reference forecasts. In this case the reference forecasts are the climatological mean values of maximum and minimum temperatures for each individual city location, for each day, averaged over the 30 year normal period 1971-2000. The $\text{score}_{perfect}$ is the score the forecasts would have received if they were perfect. Skill scores range from zero (for no skill, or skill equivalent to the reference forecasts) to 1 (for perfect forecasts). Negative skill score values indicate the forecasts are worse than the reference forecasts.

RMSE_{SS} and CORR_{SS} were evaluated for the average of the 32 ensemble members for each of the forecast days one through fifteen. The results are shown in Fig. 5.2. Both the RMSE_{SS} and CORR_{SS} for the ensemble average forecast indicate similar results in terms of forecast skill. Forecast skill, as expected, is highest for the day-one forecast, and gradually diminishes. By day nine, forecast skill is negligible compared to the climatological reference forecasts. By day 10 the forecasts show no skill compared to the climatological reference forecasts, and for days 11 through 15 the NAEFS forecasts have worse skill than climatological forecasts.

Fig. 5.2 also compares the ensemble average forecast with the deterministic five-day forecast from the MSC operational Global Environmental Multiscale (GEM) model (Côté et al., 1998a,b). Forecast skill is comparable between the ensemble average and deterministic forecast early in the forecast period, but the ensemble average outperforms the deterministic forecast in terms of RMSE_{SS} and CORR_{SS} as the forecast period progresses.

A summary of the equations used for the meteorological statistical analysis of the MREF is provided in Appendix C.

5.3.2 Continuous Ranked Probability Score

The continuous ranked probability score (CRPS) (Hersbach, 2000) is a complete generalization of the Brier score. The Brier score (Brier, 1950; Atger, 2003; McCollor and Stull, 2008b) is a verification measure designed to quantify the performance of probabilistic forecasts of dichotomous event classes. The discrete ranked probability score (RPS) (Murphy, 1969; Toth et al., 2003) extends the range of the Brier score to more event classes. The CRPS extends the limit of the RPS to an infinite number of classes. The CRPS has several distinct advantages over the Brier score and the RPS. First, the CRPS is sensitive to the entire permissable range of the parameter. In addition, the CRPS does not require a subjective threshold parameter value (as does the Brier Score) or the introduction of a number of predefined classes (as does the RPS). Different choices of threshold value (Brier score) or number and position of classes (RPS) result in different outcomes, a weakness in these scores not shared with CRPS.

The CRPS can be interpreted as an integral over all possible Brier scores. A major advantage of the Brier score is that it can be decomposed into a reliability component, a resolution component, and an uncertainty component (Murphy, 1973; Toth et al., 2003). In a similar fashion, Hersbach (2000) showed how the CRPS can be decomposed into the same three-part components:

$$\mathrm{CRPS} = \mathrm{Reli} - \mathrm{Resol} + \mathrm{Unc} \tag{5.2}$$

The CRPS components can be converted to a positively-oriented skill score (CRPSS) in the same manner the Brier score components are converted to a skill score (BSS) (McCollor and Stull, 2008b):

$$\begin{aligned}\mathrm{CRPSS} &= \frac{\mathrm{Resol}}{\mathrm{Unc}} - \frac{\mathrm{Reli}}{\mathrm{Unc}} = \text{relative resolution} - \text{relative reliability}\\ &= \mathrm{CRPS}_{RelResol} - \mathrm{CRPS}_{RelReli}\end{aligned}$$

CRPSS results are shown in Fig. 5.3. The CRPSS was calculated on forecast and observation *anomalies*, defined as the difference between the forecast or the observation and the climatological value for that city location on a particular date. The forecasts are reliable throughout the forecast period, as indicated by the $\mathrm{CRPS}_{RelReli}$ term being near zero for all forecasts. The $\mathrm{CRPS}_{RelResol}$ term however, hence the CRPSS, is highest for the day-one forecast and exhibits diminishing skill as the forecast period extends through the day 12 forecast. For forecast days 13 through 15, both maximum and minimum daily temperature forecasts exhibit negligible or no skill compared to the reference term.

5.3.3 ROC Area

The relative operating characteristic (ROC) is a verification measure based on signal detection theory (Mason, 1982; Mason and Graham, 1999; Gallus et al., 2007) that provides a method of discrimination analysis for an EPS. The ROC curve is a graph of hit-rate versus false-alarm-rate for a particular variable threshold (see Appendix C for definitions of these terms and for the contingency table basis for this analysis).

Perfect discrimination is represented by a ROC curve that rises from (0,0) along the Y-axis to (0,1) then horizontally to the upper right corner (1,1). The diagonal line from (0,0) to (1,1) represents zero skill, meaning the forecasts show no discrimination among events. The area under the ROC curve can then be converted to a skill measure as in equation (5.1), where score$_{ref}$ is the zero-skill diagonal line (ROC$_{Area}$ = 0.5) and score$_{perfect}$ is the area under the full ROC curve, equal to 1. The ROC area skill score is then ROC$_{SS}$ = 2 ROC$_{Area}$ - 1.

The hit-rate and false-alarm-rate contingency table was evaluated on forecast and observation anomaly thresholds. The ROC daily maximum temperature anomaly threshold was defined as 5°C over the climatological value for the day (+5°C anomaly threshold), and the daily minimum temperature anomaly threshold was 5°C below the climatological value for the day (-5°C anomaly threshold).

A 5°C anomaly threshold was chosen because this value produces a significant deviation from normal electrical load in cities and is therefore an important factor for planning generation requirements and power import/export agreements. A 10°C anomaly threshold would produce an even greater impact on generation planning and power import/export, but there was not enough data available in our study to produce reliable results. In a future study we would like to acquire enough data to examine forecast skill for these temperature extremes.

Results for the ROC$_{SS}$ for the 5°C temperature anomaly threshold for the full range of the 15-day forecast period are shown in Fig. 5.4. Actual ROC area curves for the daily maximum forecasts for a subset of the forecast period is shown in Fig. 5.5.

The ROC$_{SS}$ indicates that both maximum- and minimum-daily forecasts exhibit consistent and very good discrimination through forecast day six. Both maximum- and minimum-daily forecasts exhibit diminishing, but measurable discrimination skill from forecast day seven through forecast day 12. The daily minimum temperature forecasts show slightly higher skill than the daily maximum temperature forecasts through this period. For forecast days 13, 14, and 15, the ROC$_{SS}$ is close to zero and somewhat noisy indicating no skill in the forecasts.

5.3.4 Economic Value

Potential economic value for the cost/loss decision model (Murphy, 1977; Katz and Murphy, 1997; Richardson, 2000; Zhu et al., 2002; Richardson, 2003; McCollor and Stull, 2008c) is uniquely determined from ROC curves. The cost/loss ratio (C/L) model shows how probabilistic forecasts can be used in the decision-making process, and provides a measure of the benefit of probabilistic forecasts over single ensemble-average forecasts.

Fig. 5.6 shows forecast value (Richardson, 2003) as a function of C/L value for daily maximum temperature forecasts with a +5°C anomaly threshold. Maximum value occurs at the climate frequency for the event, and users with a C/L value equal to the climate frequency derive most value from the forecasts.

Fig. 5.6 also shows the value calculated for the ensemble average forecasts. The maximum value for the ensemble average forecasts is less than the maximum value for the forecasts incorporating the full suite of ensemble members. The discrepancy between the maximum value for the ensemble forecasts and the maximum value for the ensemble average forecasts widens with projected forecast period. Additionally, the cast of users (described by different cost/loss ratios) that gain benefit (show positive value) from the forecasts is wider for the full ensemble forecasts than for the ensemble average forecasts.

A graph depicting the decay of forecast maximum value with projected forecast period is shown in Fig. 5.7. The maximum value is slightly higher for minimum temperature forecasts than for maximum temperature forecasts, and the forecasts lose very little of their initial value through forecast day six. Maximum value lessens but remains positive through forecast day eleven. Forecasts for days 12 through 15 indicate the value of these forecasts is little better than climatology.

Plotting forecast value as a function of probability threshold p_t for different C/L values (Fig. 5.8) is a way of showing how different forecast users can benefit individually from the same set of ensemble forecasts. The probability threshold p_t is the point that a decision maker should take action on the basis of the forecast to derive most value from the forecasts. Low cost/loss ratio forecast users ($C/L = 0.2$ in Fig. 5.8) can take advantage of the ensemble forecasts by taking action at very low forecast probabilities through forecast day eight. High C/L forecast users ($C/L = 0.8$) must wait until the forecast probability is much higher, and can gain positive value from the forecasts only through forecast day four.

5.3.5 Equal Likelihood

In a perfectly realistic and reliable EPS, the range of the ensemble forecast members represents the full spread of the probability distribution of observations, and each member of the ensemble exhibits an equal likelihood of occurrence. Actual EPSs are rarely that perfect. The rank histogram, also known as a Talagrand diagram (Anderson, 1996; Talagrand et al., 1998), is a useful measure of equal likelihood and of reliability (Hou et al., 2001; Candille and Talagrand, 2005) of an EPS.

A perfect EPS in which each ensemble member is equally likely to verify for any particular forecast exhibits a flat rank histogram. Ensemble forecasts with insufficient spread to adequately represent the probability distribution of observations exhibit a U-shaped histogram. Rank histograms derived from a particular EPS can also exhibit an inverted-U-shape, indicative of too much spread in the ensemble members compared to the observations. Rank histograms can also be L-shaped (warm bias), or J-shaped (cool bias). Operational ensemble weather forecasting systems in the medium lead-time range (3-10 days ahead) exhibit U-shaped analysis rank histograms, implying the ensemble forecasts underestimate the true uncertainty in the forecasts (Toth et al., 2003).

A subset of rank histogram plots for the EPS daily maximum temperature forecasts is provided in Fig. 5.9, for forecast days two, four, six, eight, ten, and twelve. These plots show that early in the forecast period, especially through forecast day four, the rank histograms do exhibit a 'U' shape, with excess values in the outermost bins. This means the EPS is under-dispersive through forecast day four; the spread in the EPS is not indicative of the spread in the actual observations. However the rank histogram is symmetric, confirming that the bias-corrected temperature forecasts are indeed unbiased.

Forecast days 6 through 12 display a progressively flatter histogram, indicating the EPS spread increases with forecast lead time and the EPS spread essentially matches the spread of the observations beyond forecast day six. Similar results (Fig. 5.10) were obtained from a rank histogram analysis of the EPS daily minimum temperature forecasts. However, this better spread is associated with worse skill, as shown previously in sections 5.3.1 and 5.3.2.

A measure of the deviation of the rank histogram from flatness is given in Candille and Talagrand (2005) for an EPS with M members and N available forecast/observation pairs. The number of values in the i_{th} interval of the histogram is given by s_i. For a flat histogram generated from a reliable EPS, $s_i = N/(M+1)$ for each interval i. The quantity

$$\Delta = \sum_{i=1}^{M+1} (s_i - \frac{N}{M+1})^2 \tag{5.3}$$

measures the deviation of the histogram from a horizontal line. For a reliable system, the base value is $\Delta_0 = NM/(M+1)$. The ratio

$$\delta = \Delta/\Delta_0 \tag{5.4}$$

is evaluated as the overall measure of the flatness of an ensemble prediction system rank histogram. A value of δ that is significantly larger than 1 indicates the system does not reflect equal likelihood of ensemble members. A value of δ that is closer to 1 indicates the ensemble forecasts are tending toward an equal likelihood of occurrence, and therefore are more reliable.

Fig. 5.11 provides an analysis of the rank histograms for the MREFs studied here. Early in the forecast period, for forecast days one, two, and three, δ is much greater than ten. Therefore the MREF system for forecast days one through three exhibits insufficient, but successively increasing spread to adequately represent the variability in the observations. For forecast days four and five the MREF exhibits δ values successively smaller, but still greater than five, so that evidence of slight under-dispersion remains. For forecast days beyond day five, the δ values derived from this MREF system (values between one and five), as indicated in Figs. 5.9-5.11, indicate ensemble spread is very slightly under-dispersive, though nearly sufficient spread is realized in the ensembles to represent the observations in the analysis. Spread does improve as the forecast period increases, however this improvement in spread is negated by the deterioration in skill at the longer forecast periods. This skill vs. spread relationship is demonstrated in a trade-off diagram, described in the following section.

Skill-spread Trade-off Diagram

We define a new spread score as:

$$\text{Spread score} = 1/\delta \tag{5.5}$$

A plot of skill score on the ordinate and this spread score on the abscissa (as the forecast projection progresses from day 1 through day 15) shows the trade-off between higher skill but lower spread early in the forecast period and degraded skill but better spread later in the forecast period. A skill-spread trade-off diagram with the continuous ranked probability skill score (CRPSS) vs. the spread score is shown in Fig. 5.12, and a similar trade-off diagram displaying the ROC skill score (ROC$_{SS}$, for the 5°C temperature anomaly threshold) is shown in Fig. 5.13.

Both Figs. 5.12 and 5.13 display the change of skill and spread with forecast projection, for both daily minimum and daily maximum temperature forecasts. Beyond forecast day 11 or 12 (indicated by a dotted line in these figures), the skill becomes negligible and the spread stops improving, or even lessens in some cases. The authors suggest that this relationship between skill and spread could be explored further with the aid of this skill-spread trade-off diagram, with different parameters, more data sets and other skill measures. One interesting application of this skill-spread diagram would analyze the variation of skill with spread as the number of individual members in an ensemble prediction system is varied. The diagram could be used, for example, to find the fewest number of members needed to reach a specific level of skill, for a specific forecast projection, in order to optimize the use of computer resources in ensemble system design.

5.4 Summary and Conclusions

The skill and value of medium-range maximum and minimum NAEFS temperature forecasts are evaluated here for the seven-month period of 1 March through 1 October 2007. However, caution must be used when objectively assessing the quality of ensemble forecasts. Ensemble forecast systems must be verified over many cases. As a result, the scoring metrics described here are susceptible to several sources of noise (Hamill et al., 2000). For example, improper estimates of probabilities will arise from small-sized ensembles. Also, insufficient variety and number of cases will lead to statistical misrepresentation. Finally, imperfect observations make true forecast evaluation impossible.

Two standard verification skill score measures designed for continuous deterministic forecast variables, a RMSE skill score and a Pearson correlation skill score, were applied to the mean of the bias-corrected ensemble forecasts. These measures both showed that the ensemble means of these city temperature forecasts from the NAEFS possessed skill (compared to climatology) through forecast day nine, though for forecast days eight and nine the forecast skill was marginal. These skill scores also indicated that daily maximum temperature forecasts were slightly more skillful than daily minimum temperature forecasts for the period analyzed.

This study then incorporated verification measures specifically designed for probabilistic forecasts generated from the bias-corrected set of all 32 ensemble members. The continuous ranked probability score (CRPS), particularly useful for evaluating a continuous ensemble variable, indicated skill in the EPS through day 12. The CRPS also supported the previous finding for the ensemble mean that daily maximum temperatures are slightly more skillful than the daily minimum forecasts.

The ROC area skill score, designed to indicate the ability of the EPS to discriminate among different events, showed that the EPS forecasts were indeed skillful (for 5°C temperature anomaly thresholds) in this regard through day 12. ROC evaluation is characterized by stratification by observation, and discrimination is based on the probabilities of the forecasts conditioned on different observed values, or likelihood-base rate factorization in the parlance of the Murphy and Winkler general framework for forecast verification (Murphy and Winkler, 1987; Potts, 2003).

ROC analysis leads to economic evaluation through a cost/loss measure of value. Cost/loss analysis indicated that overall, the EPS provides some economic value through day 12. However the maximum value, assigned to the user whose C/L value is equal to the climatological base rate for the event in question, diminishes markedly through this forecast range. Additionally, the range of users, defined by different cost/loss ratios, diminishes considerably beyond forecast day six. The C/L analysis also provides proof that the full suite of ensemble forecasts provides value to a greater range of users than the single ensemble average forecasts.

Rank histograms provide a method to test if the dispersion, or spread, of the ensemble forecasts represents the dispersion in the actual observations, a vital component in a reliable EPS. In the NAEFS forecast sample examined here, the forecasts were decreasingly underdispersive early in the forecast cycle, for forecast days one through five. The forecasts exhibited marginal underdispersion for forecast days six and beyond. A new trade-off diagram is proposed to show the interplay between skill and spread.

One aspect of interest uncovered in this study is the comparison of forecast skill between daily maximum and daily minimum temperature forecasts. The skill scores evaluated for the ensemble mean forecast (RMSE$_{SS}$ and CORR$_{SS}$) indicated that daily maximum temperature forecasts were somewhat more skillful than daily minimum temperature forecasts through the first nine days of the forecast period (Fig. 5.2). These skill scores indicate no skill in the forecasts, for either daily maximum or daily minimum forecasts, for forecast day 10 (compared to the climatological reference forecasts). Beyond day 10, the MREF forecasts fare worse than climatological forecasts, with daily maximum temperature forecasts now proving worse than daily minimum temperature forecasts.

Daily maximum temperature forecasts also prove slightly better than daily minimum temperature forecasts when compared using the CRPSS (Fig. 5.3), throughout the 15 day range of MREF forecasts. The ROC$_{SS}$, on the other hand, indicates daily minimum temperature forecasts show better discrimination than daily maximum temperature forecasts (Fig. 5.4). In terms of forecast spread, both daily maximum and daily minimum temperature forecasts possess similar spread for all forecasts beyond day one (Fig. 5.11). We do not have enough information to determine if the difference in skill between daily maximum and daily minimum temperature forecasts is an artifact of the scoring metric, the models in the MREF, or the fact that the 00 and 12 UTC forecasts don't necessarily correspond to the times of maximum and minimum daily temperatures. Investigation of this difference in skill between maximum and minimum temperature forecasts is an avenue for future research.

In summary, the bias-corrected NAEFS provides skill in the deterministic ensemble average, for the forecasts analyzed in this study, through forecast day nine. Employing the full range of 32 ensemble members to provide probabilistic forecasts of daily maximum and minimum temperature extended the skill of the NAEFS through an additional three forecast days, through day 12.

Summarizing the findings for the ensemble average forecasts, and in comparing MREF forecasts in a continuous mode, daily maximum temperature forecasts were slightly more skillful than daily minimum temperature forecasts. Shifting to discriminating ability (using +/-5°C anomaly thresholds), daily minimum temperature forecasts provided slightly better discriminating ability than daily maximum temperature forecasts.

Many sectors of the economy, such as hydro-electric energy generation, transmission, and distribution, are susceptible to financial risk driven by changing temperature regimes. This study has shown that business managers savvy enough to incorporate ensemble temperature forecasts provided by the NAEFS can mitigate that risk quite far into the future (potentially 12 days depending on user requirements).

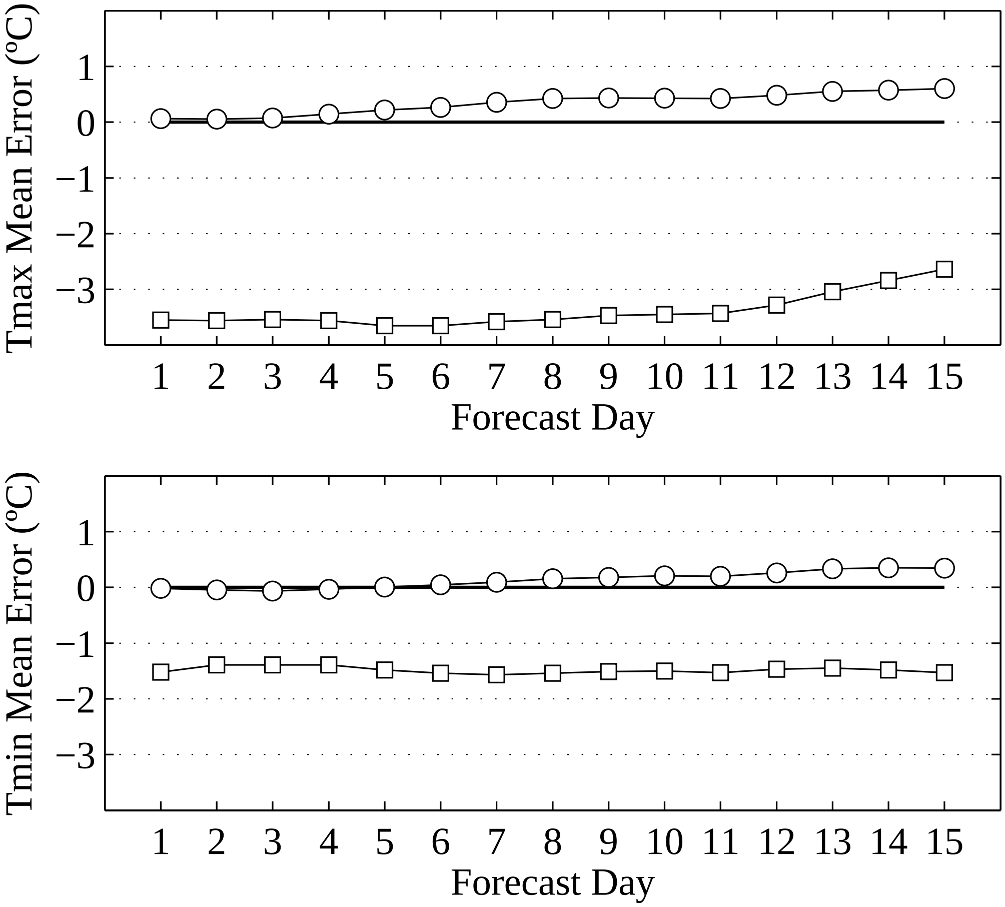

Figure 5.1: Mean error for ensemble average DMO forecasts (squares) and ensemble average post-processed forecasts (circles) for daily maximum temperatures (upper panel) and daily minimum temperatures (lower panel) for forecast days 1 through 15. Values near zero are best.

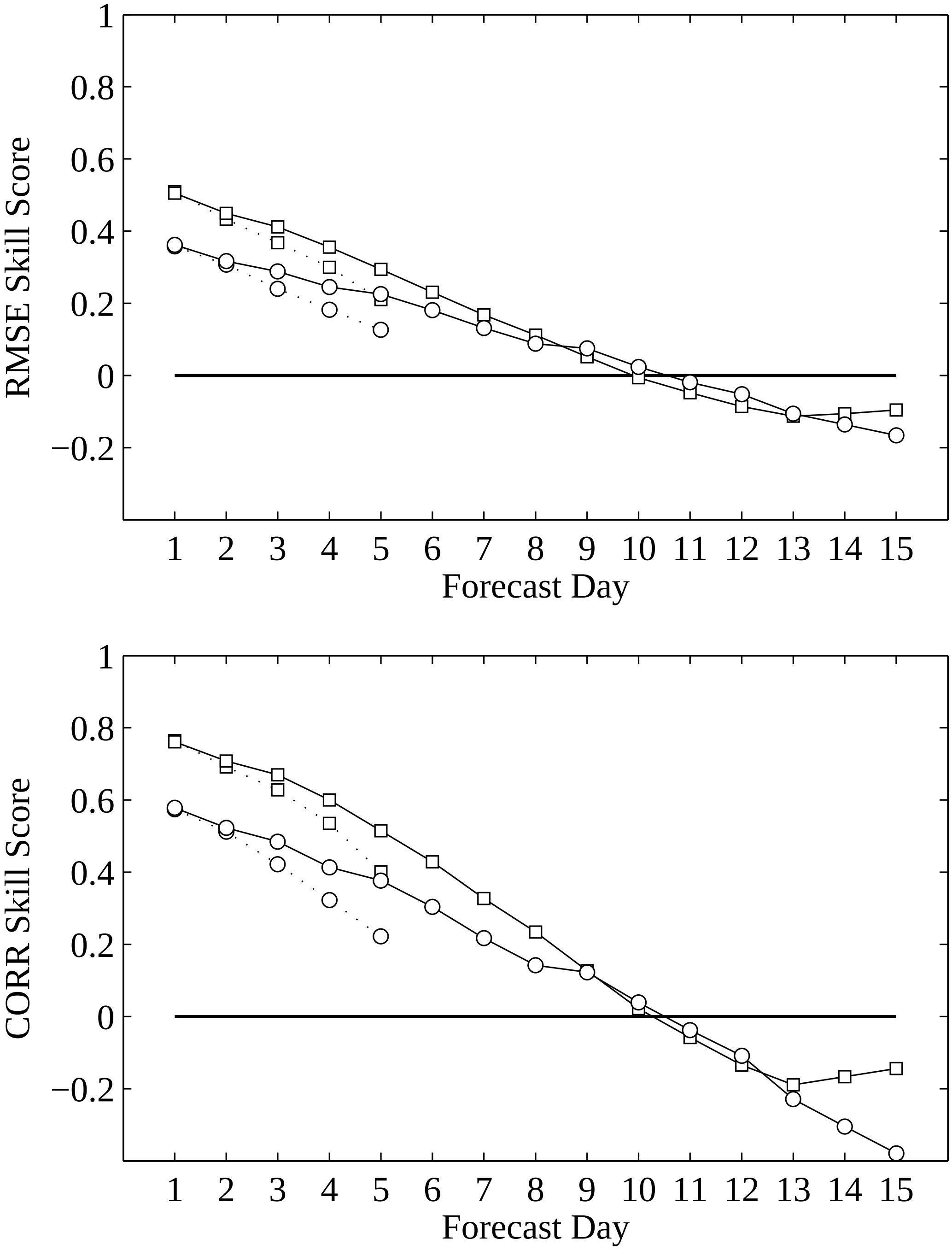

Figure 5.2: RMSE (upper panel) and correlation (lower panel) skill scores for daily maximum (squares) and minimum (circles) temperature forecasts for forecast days 1 through 15. The ensemble average forecast (solid line) is compared with a deterministic operational forecast (dashed line). A skill score of one is best.

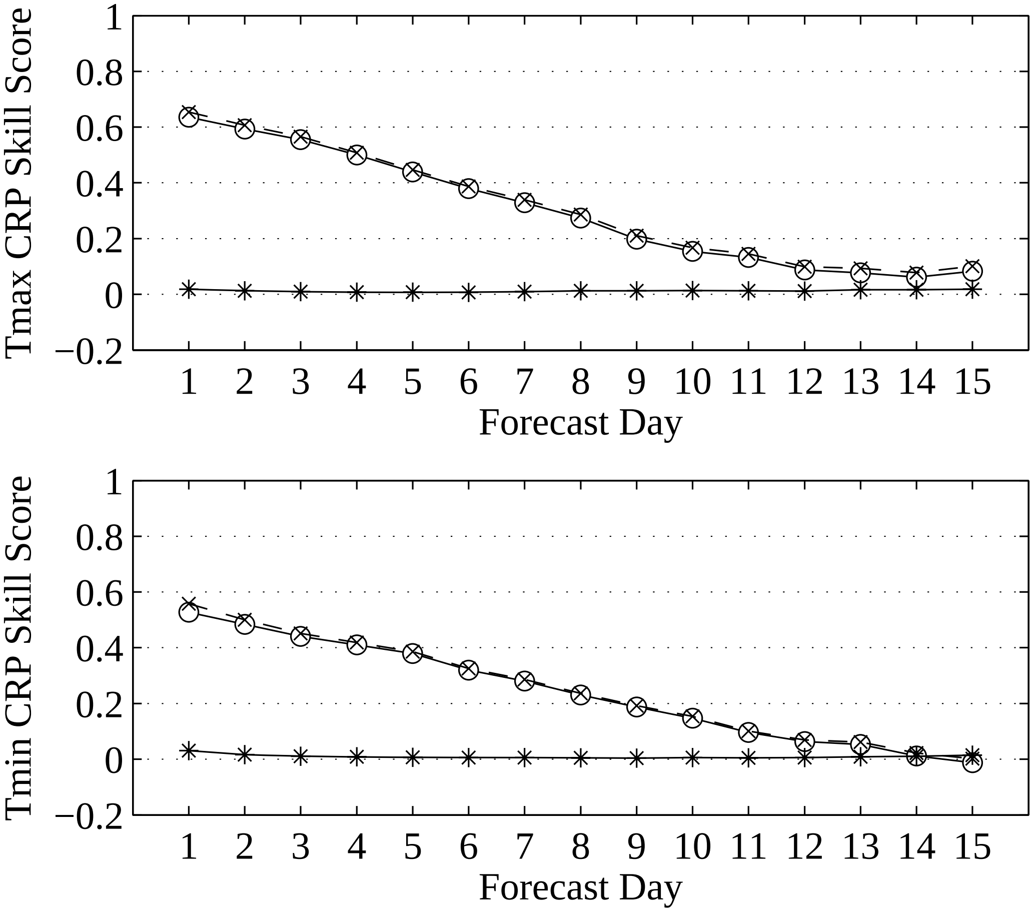

Figure 5.3: Continuous ranked probability skill score for ensemble daily maximum (upper panel) and minimum (lower panel) temperature anomaly forecasts. The CRP skill score (circles) equals the relative resolution (dashed line with x markers) less the relative reliability (* markers). A skill score of one is best.

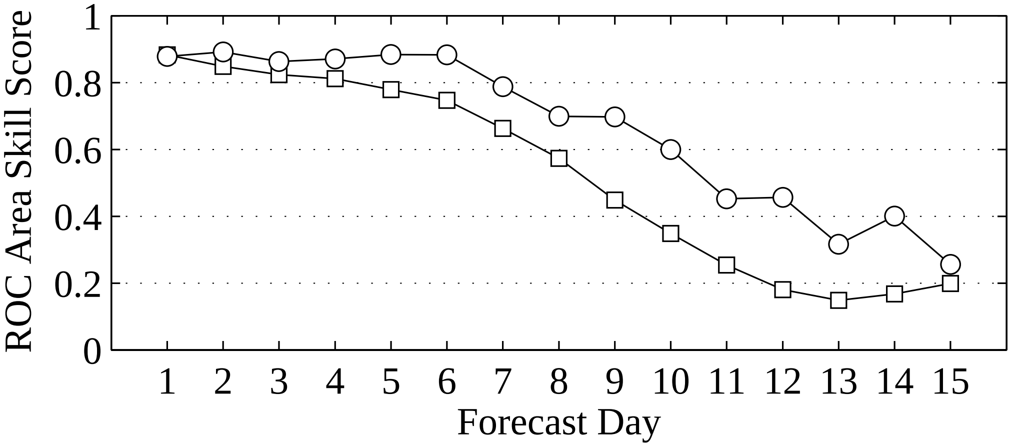

Figure 5.4: ROC area skill score for ensemble daily maximum temperature forecasts with a +5°C anomaly threshold (squares) and daily minimum temperature forecasts with a -5°C anomaly threshold (circles). A skill score of one is best.

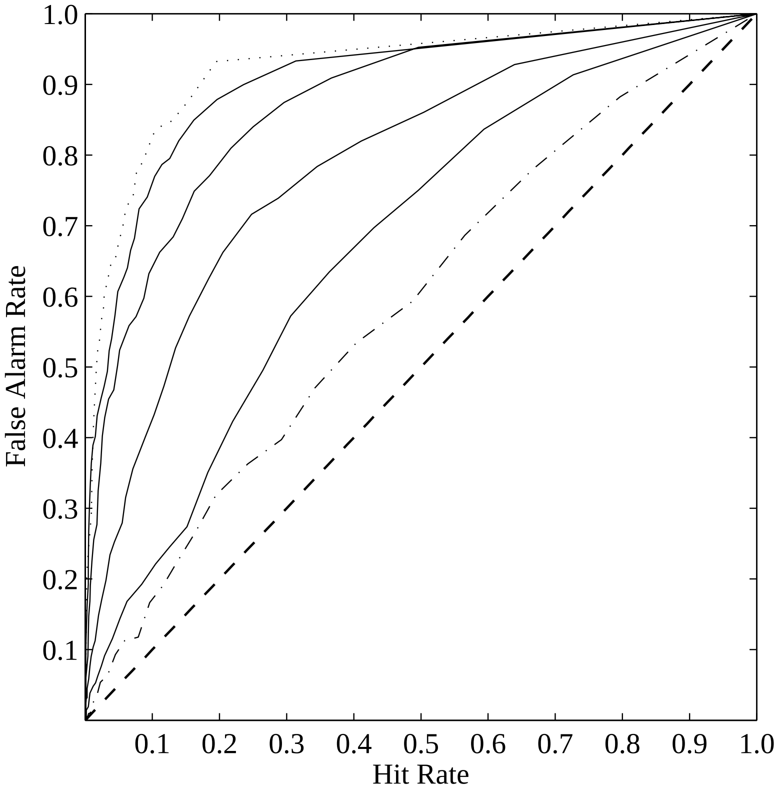

Figure 5.5: ROC Area plots for a subset of the forecast period for daily maximum temperature anomaly +5°C threshold. The plots range from ROC curve for day 2 (dot), day 4, day 6, day 8, day 10, and day 12 (dot-dash). The dashed line is the zero-skill line.

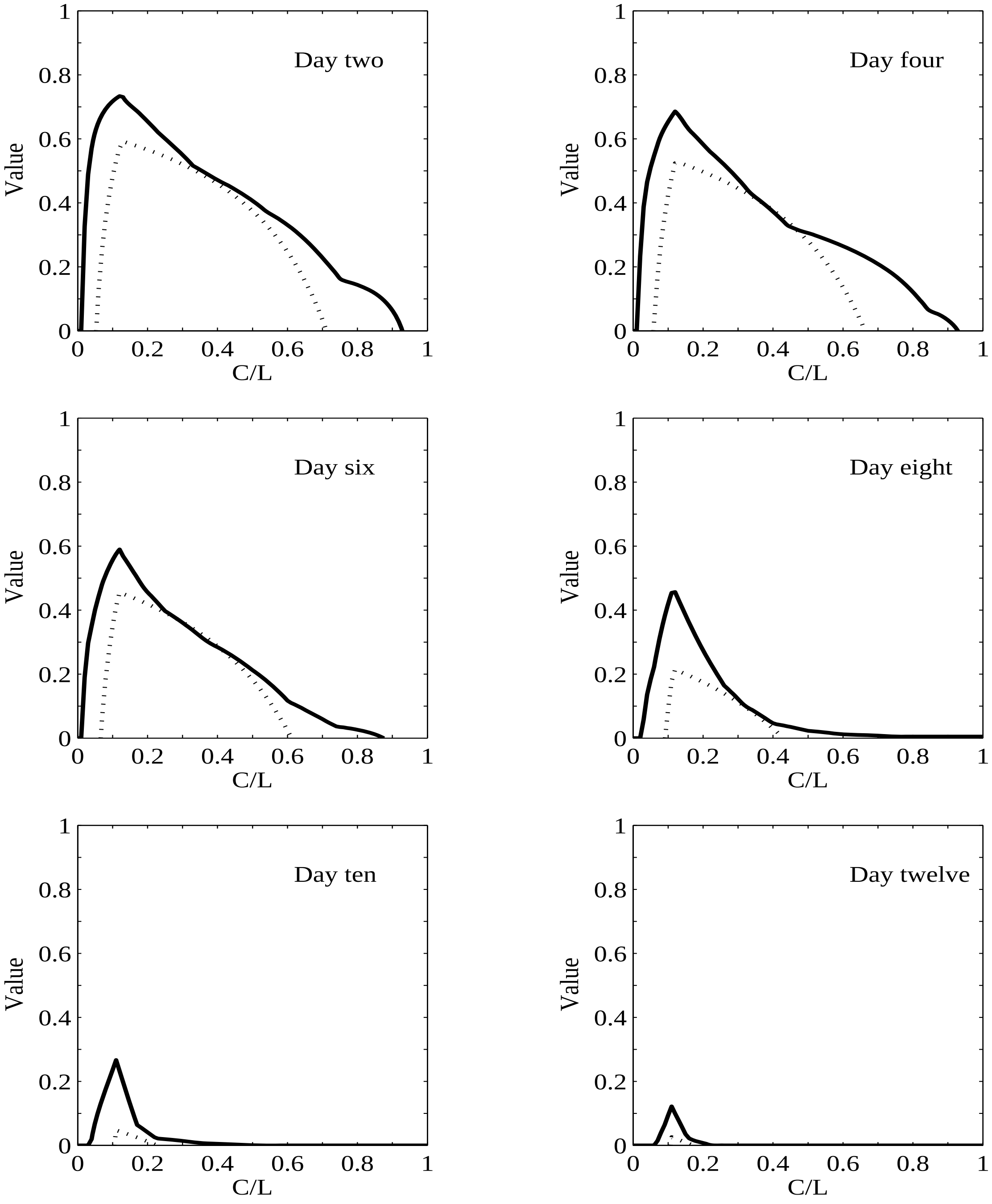

Figure 5.6: Forecast value as a function of cost/loss ratio (C/L) for ensemble daily maximum forecasts (+5°C anomaly threshold). The solid line represents the value of the full ensemble and the dotted line the value of the ensemble mean. A value of one is best.

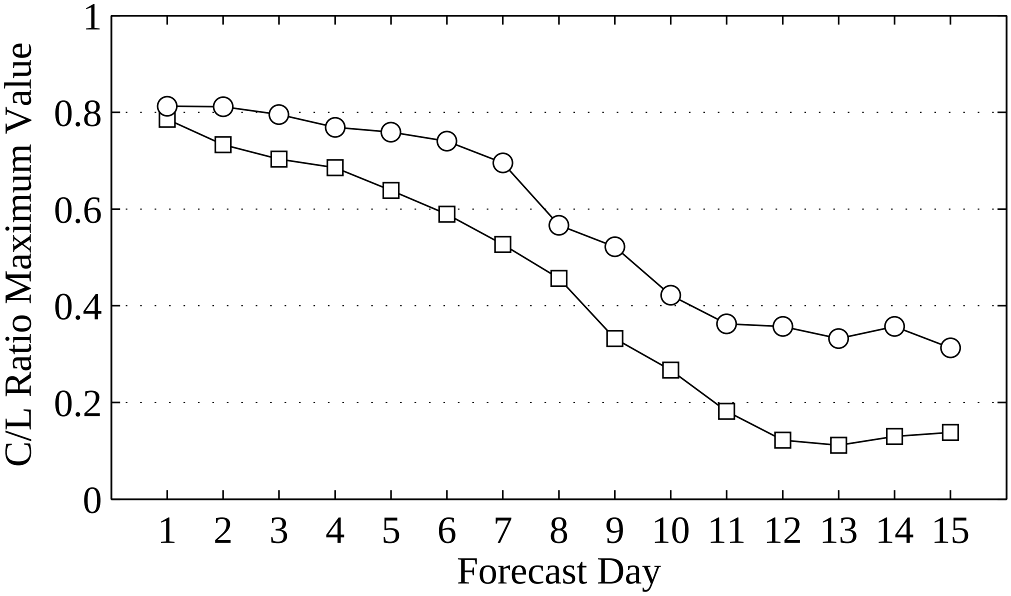

Figure 5.7: Cost/loss ratio (C/L) maximum value for ensemble daily maximum temperatures with a +5°C anomaly threshold (squares) and minimum temperatures with a -5°C anomaly threshold (circles). A value of one is best.

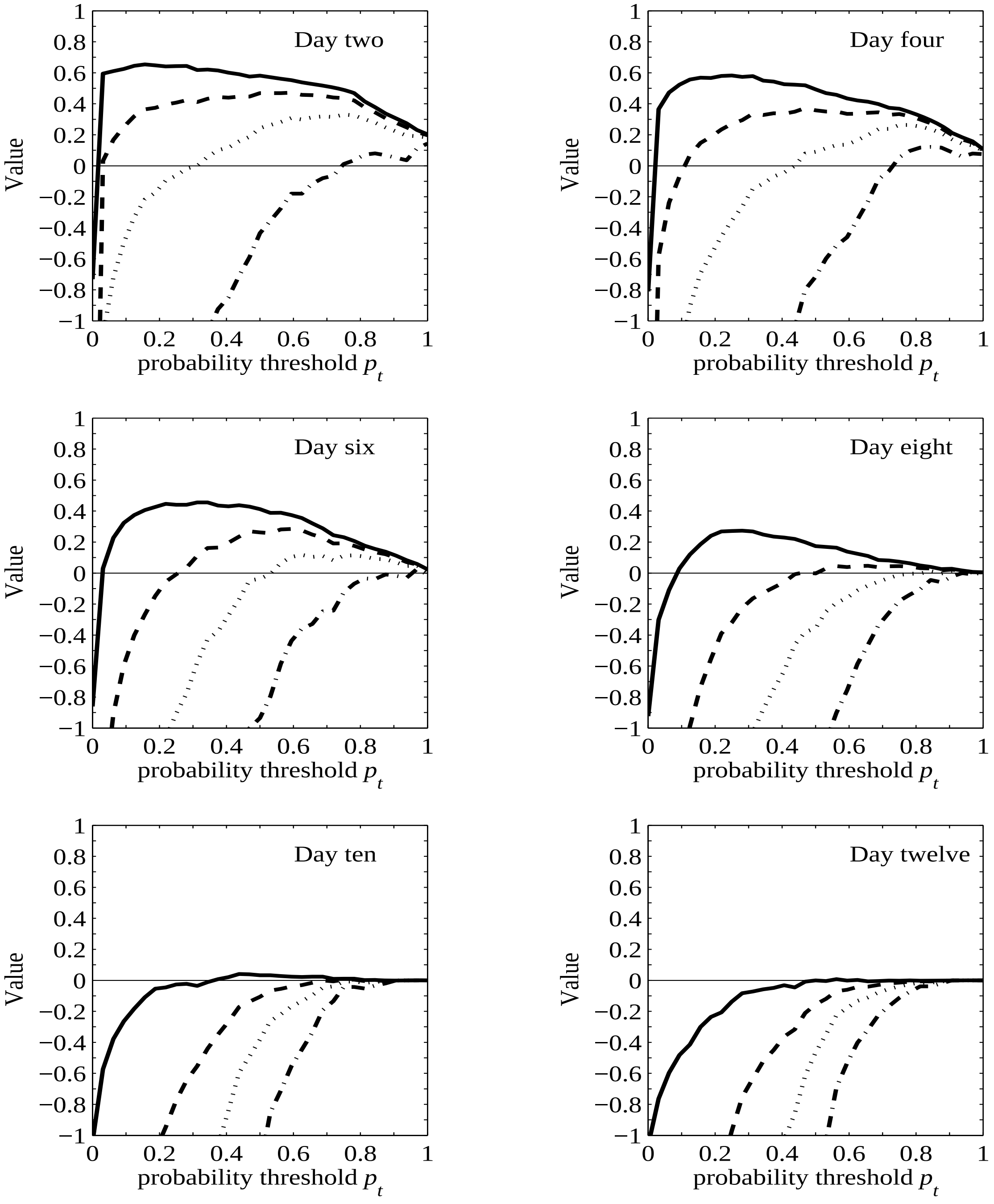

Figure 5.8: Value as a function of probability threshold p_t for cost/loss ratio C/L=0.2 (solid), C/L=0.4 (dash), C/L=0.6 (dot) and C/L=0.8 (dot-dash). The forecasts are daily maximum temperature forecasts with a +5°C anomaly threshold. A value of one is best.

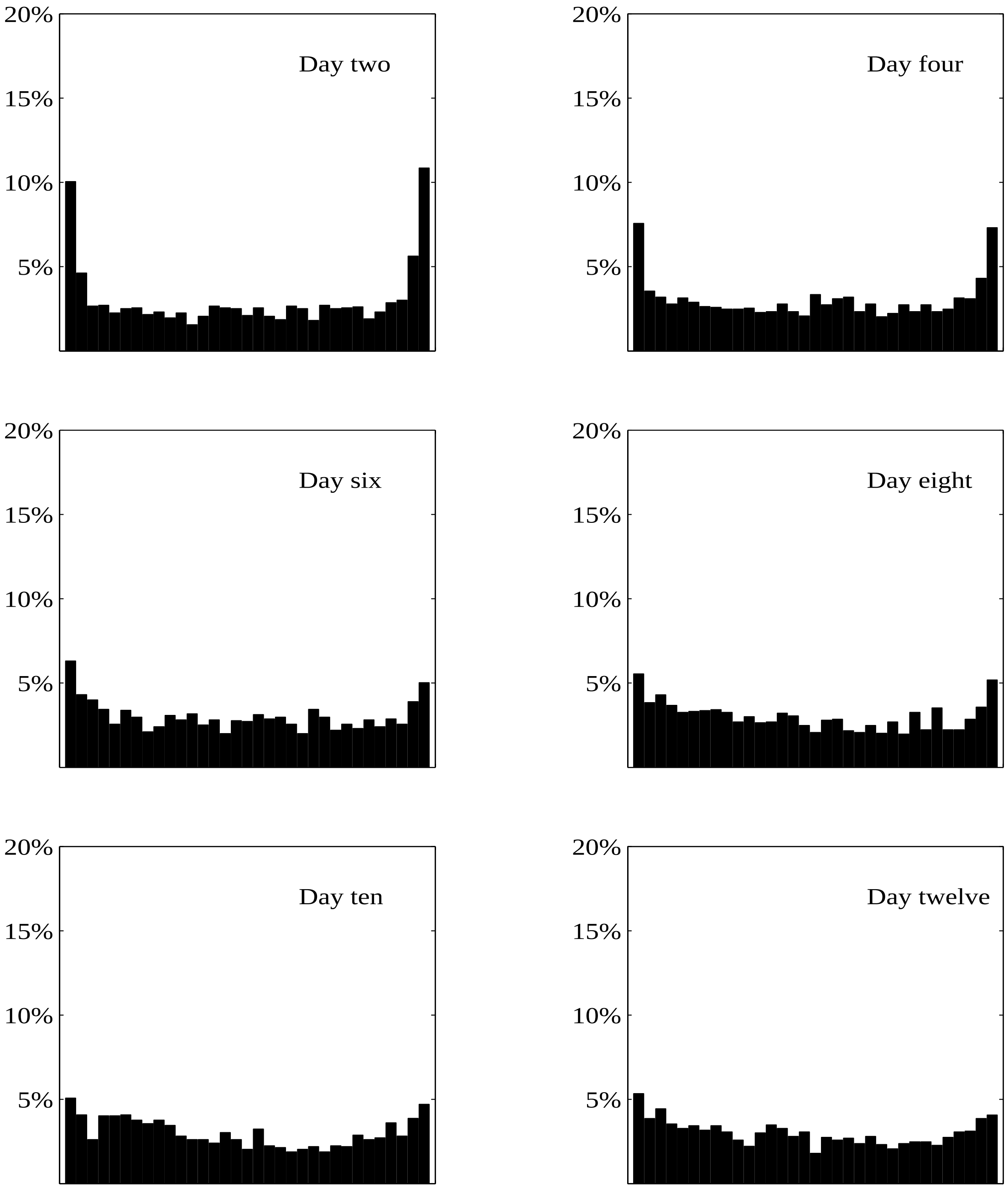

Figure 5.9: Rank histograms for the ensemble of daily maximum temperatures. A flat histogram indicates non-dispersive reliable forecasts. A U-shaped histogram indicates the ensemble forecasts are under-dispersive.

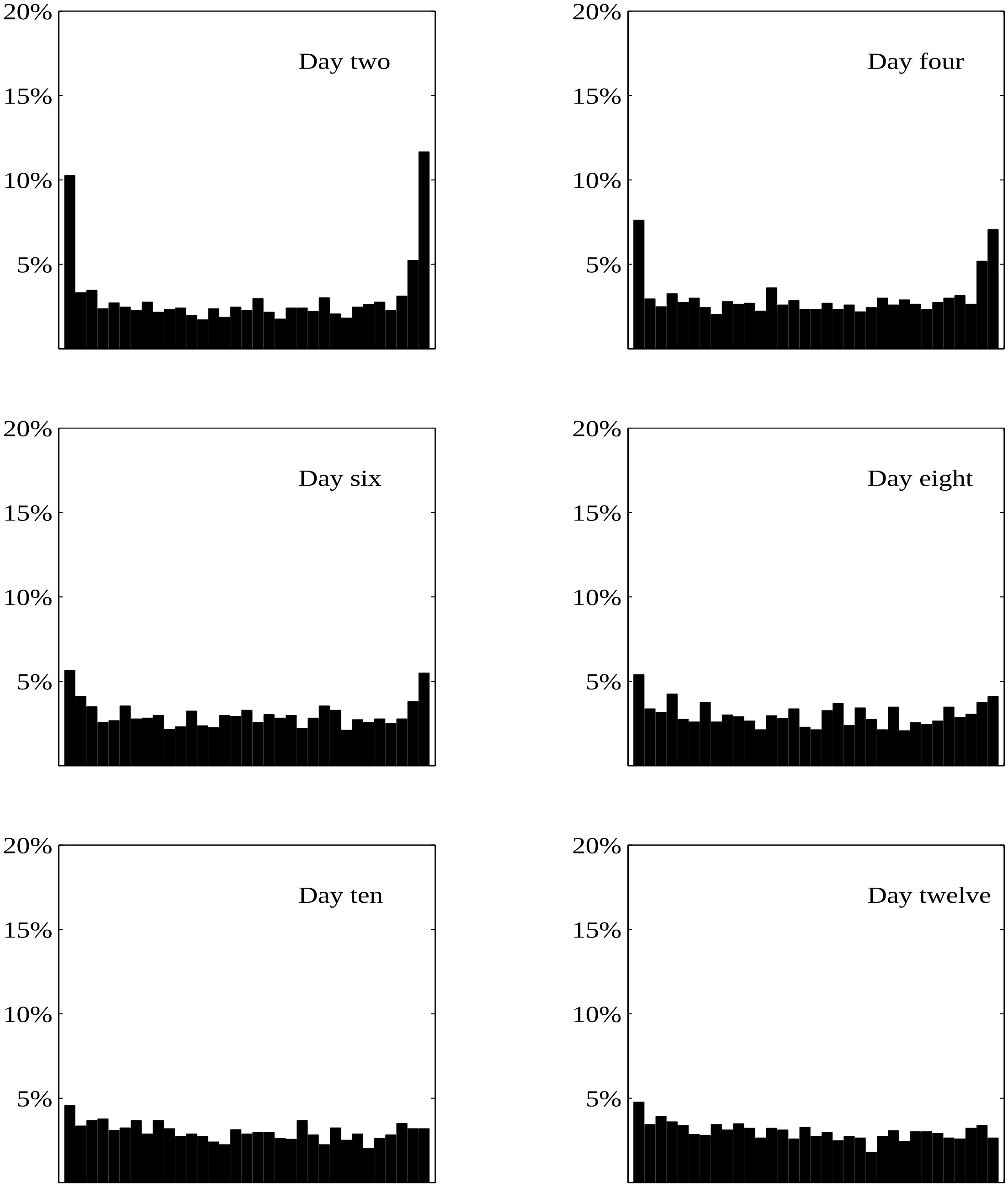

Figure 5.10: Rank histograms for the ensemble of daily minimum temperatures. A flat histogram indicates non-dispersive reliable forecasts. A U-shaped histogram indicates the ensemble forecasts are under-dispersive.

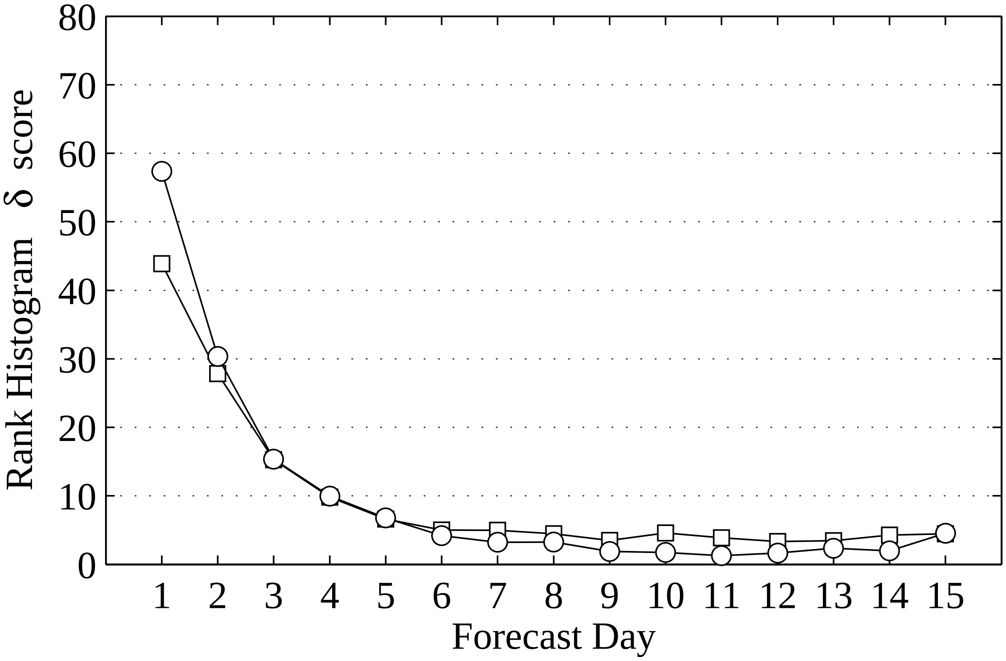

Figure 5.11: Rank histogram δ score for ensemble daily maximum (squares) and daily minimum (circles) temperature forecasts. A value of one is best.

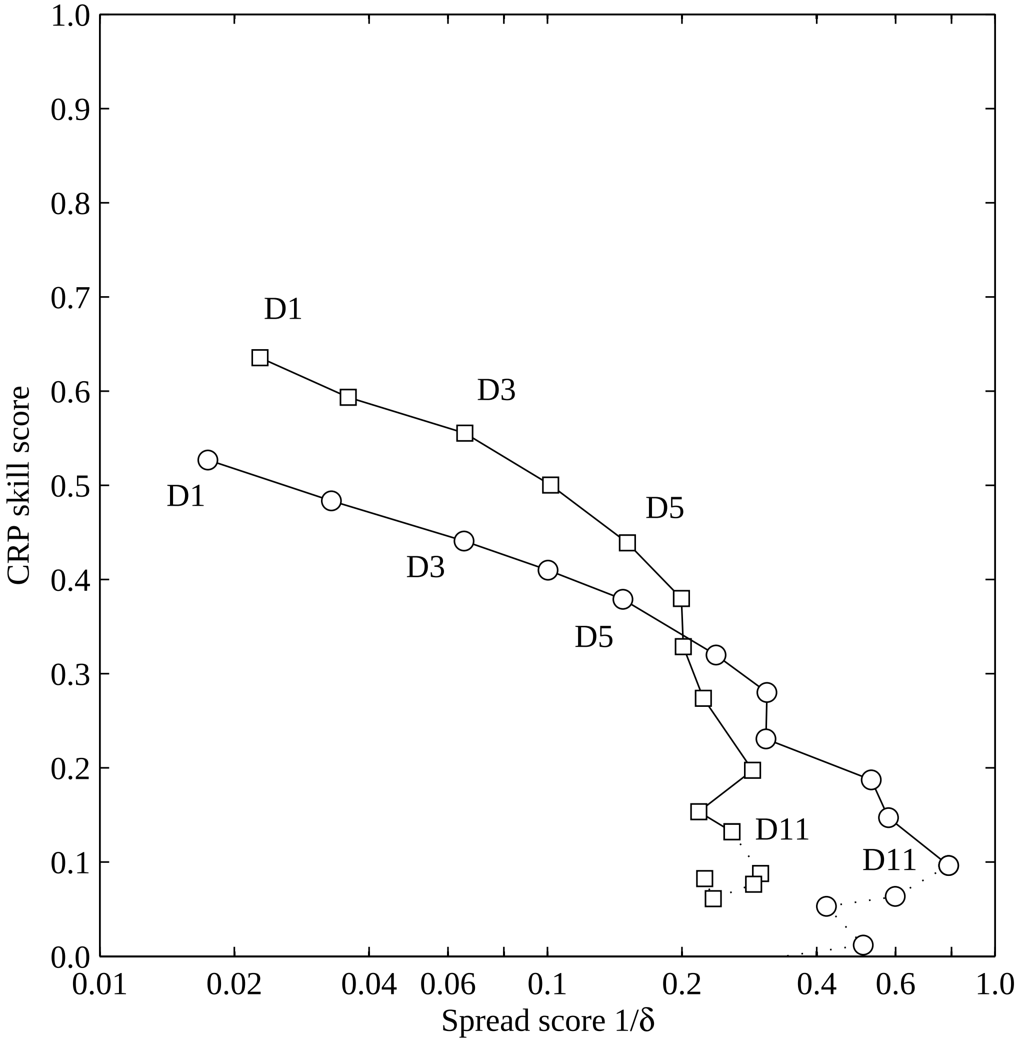

Figure 5.12: A trade-off diagram of CRPSS vs. spread score ($1/\delta$) for daily maximum temperature (squares) and daily minimum temperature (circles) forecast anomalies. The scores progress from the day 1 forecast (upper left) through the day 15 forecast (lower right). Values of one are best for both skill score and spread score, showing the trade-off between high skill but low spread early in the forecast period, and degraded skill but better spread later in the forecast period. After day 11 (dotted line) the skill is negligible and the spread stays the same or lessens.

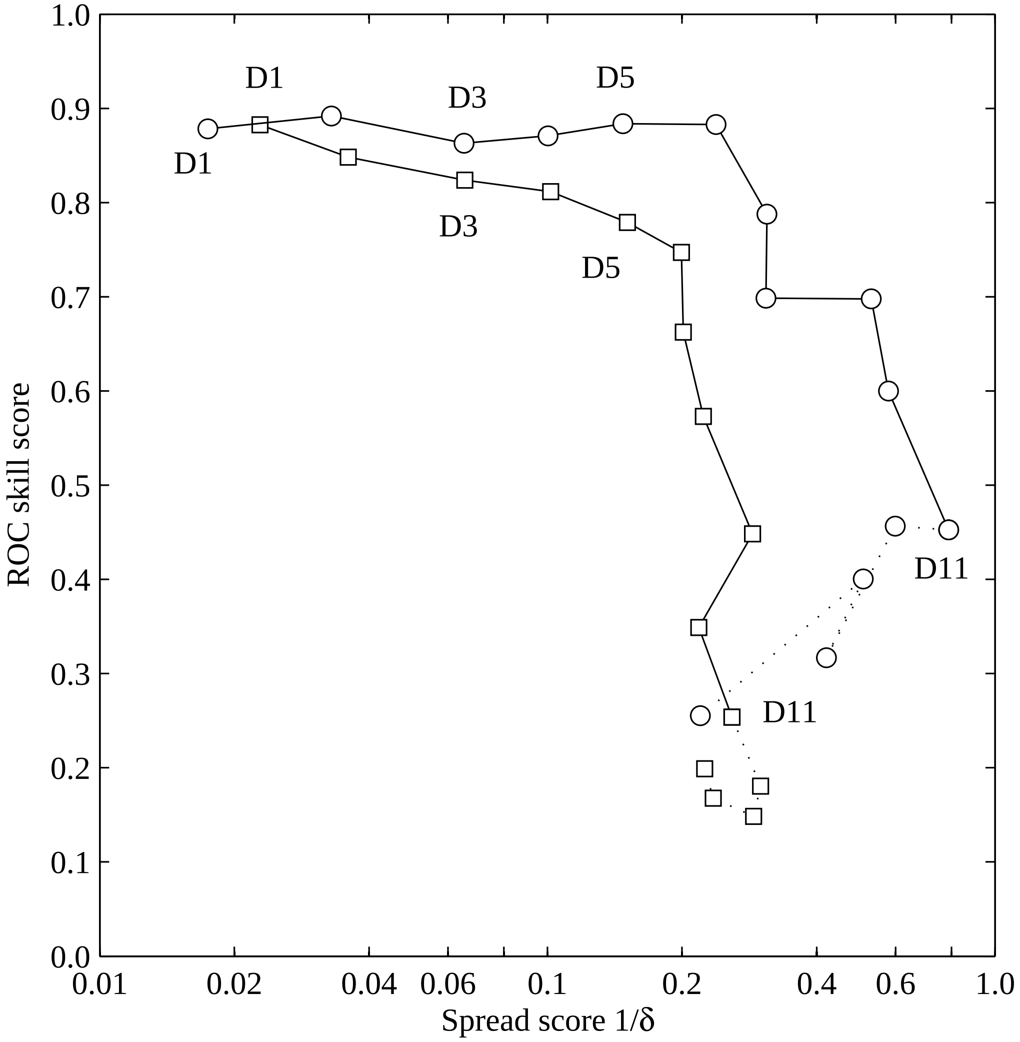

Figure 5.13: Similar to Fig. 5.12 but for ROC_{SS} vs. spread score ($1/\delta$).

Table 5.1: Ten cities for which forecast and associated verifying observations were evaluated. ICAO is the International Civil Aviation Organization.

City	ICAO Identifier	Latitude	Longitude
High Level, Alberta	CYOJ	58.6° N	117.2° W
Peace River, Alberta	CYPE	56.2° N	117.4° W
Fort St. John, British Columbia	CYXJ	56.2° N	120.7° W
Victoria, British Columbia	CYYJ	48.6° N	123.4° W
Vancouver, British Columbia	CYVR	49.2° N	123.2° W
Seattle, Washington	KSEA	47.4° N	122.3° W
Spokane, Washington	KGEG	47.6° N	117.5° W
Portland, Oregon	KPDX	45.6° N	122.6° W
Sacramento, California	KSAC	38.5° N	121.5° W
Bakersfield, California	KBFL	35.4° N	119.1° W

Bibliography

Anderson, J. L., 1996: A method for producing and evaluating probabilistic precipitation forecasts from ensemble model integrations. *J. Climate*, **9**, 1518–1529.

Atger, F., 2003: Spatial and interannual variability of the reliability of ensemble-based probabilistic forecasts: Consequences for calibration. *Mon. Wea. Rev.*, **131**, 1509–1523.

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

Buizza, R., P. L. Houtekamer, Z. Toth, G. Pellerin, M. Wei, and Y. Zhu, 2005: A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Mon. Wea. Rev.*, **133**, 1076–1097.

Candille, G. and O. Talagrand, 2005: Evaluation of probabilistic prediction systems for a scalar variable. *Quart. J. Roy. Meteor. Soc.*, **131**, 2131–2150.

Côté, J., J.-G. Desmarais, S. Gravel, A. Méthot, A. Patoine, M. Roch, and A. Staniforth, 1998a: The operational CMC-MRB global environmental multiscale (GEM) model: Part I – design considerations and formulation. *Mon. Wea. Rev.*, **126**, 1373–1395.

— 1998b: The operational CMC-MRB global environmental multiscale (GEM) model: Part II – results. *Mon. Wea. Rev.*, **126**, 1397–1418.

Eckel, F. A. and C. F. Mass, 2005: Aspects of effective mesoscale, short-range ensemble forecasting. *Wea. Forecasting*, **20**, 328–350.

Gallus, W. A. J., M. E. Baldwin, and K. L. Elmore, 2007: Evaluation of probabilistic precipitation forecasts determined from Eta and AVN forecasted amounts. *Wea. Forecasting*, **22**, 207–215.

Hamill, T. M., S. L. Mullen, C. Snyder, Z. Toth, and D. P. Baumhefner, 2000: Ensemble forecasting in the short to medium range: Report from a workshop. *Bull. Amer. Meteor. Soc.*, **81**, 2653–2664.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting*, **15**, 559–570.

Hou, D., E. Kalnay, and K. K. Droegemeier, 2001: Objective verification of the SAMEX ’98 ensemble forecasts. *Mon. Wea. Rev.*, **129**, 73–91.

Houtekamer, P. L., L. Lefaivre, J. Derome, H. Ritchie, and H. L. Mitchell, 1996: A system simulation approach to ensemble prediction. *Mon. Wea. Rev.*, **124**, 1225–1242.

Jones, M. S., B. A. Colle, and J. S. Tongue, 2007: Evaluation of a mesoscale short-range ensemble forecast system over the northeast United States. *Wea. Forecasting*, **22**, 36–55.

Katz, R. W. and A. H. Murphy, 1997: *Economic Value of Weather and Climate Forecasts*. Cambridge University Press, England.

Mason, I., 1982: A model for assessment of weather forecasts. *Aust. Meteor. Mag.*, **30**, 291–303.

Mason, I. and N. E. Graham, 1999: Conditional probabilities, relative operating characteristics, and relative operating levels. *Wea. Forecasting*, **14**, 713–725.

McCollor, D. and R. Stull, 2008a: Hydrometeorological accuracy enhancement via postprocessing of numerical weather forecasts in complex terrain. *Wea. Forecasting*, **23**, 131–144.

— 2008b: Hydrometeorological short-range ensemble forecasts in complex terrain. Part I: Meteorological evaluation. *Wea. Forecasting*, (in press).

— 2008c: Hydrometeorological short-range ensemble forecasts in complex terrain. Part II: Economic evaluation. *Wea. Forecasting*, (in press).

Murphy, A. H., 1969: On the ”Ranked probability score“. *J. Appl. Meteor.*, **8**, 988–989.

— 1973: A new vector partition of the probability score. *J. Appl. Meteor.*, **12**, 595–600.

— 1977: The value of climatological, categorical, and probabilistic forecasts in the cost-loss ratio situation. *Mon. Wea. Rev.*, **105**, 803–816.

— 1991: Forecast verification: Its complexity and dimensionality. *Mon. Wea. Rev.*, **119**, 1590–1601.

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

Potts, J. M., 2003: Basic concepts. *Forecast verification: A practitioner's guide in atmospheric science*, I. Jolliffe and D. B. Stephenson, eds., Wiley, 13–36.

Richardson, D. S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **126**, 649–667.

— 2003: Economic value and skill. *Forecast verification: A practitioner's guide in atmospheric science*, I. Jolliffe and D. B. Stephenson, eds., Wiley, 164–187.

Roulin, E. and S. Vannitsem, 2005: Skill of medium-range hydrological ensemble predictions. *J. Hydrometeorology.*, **6**, 729–744.

Stensrud, D. J., J. Bao, and T. T. Warner, 2000: Using initial conditions and model physics perturbations in short-range ensemble simulations of mesoscale convective systems. *Mon. Wea. Rev.*, **128**, 2077–2107.

Talagrand, O., R. Vautard, and B. Strauss, 1998: Evaluation of probabilistic prediction systems. *Workshop on predictability*, ECMWF, 1–25.

Tennant, W. J., Z. Toth, and K. J. Rae, 2007: Application of the NCEP ensemble prediction system to medium-range forecasting in South Africa: New products, benefits, and challenges. *Wea. Forecasting*, **22**, 18–35.

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

Toth, Z., L. Lefaivre, G. Brunet, P. L. Houtekamer, Y. Zhu, R. Wobus, Y. Pelletier, R. Verret, L. Wilson, B. Cui, G. Pellerin, B. A. Gordon, D. Michaud, E. Olenic, D. Unger, and S. Beauregard, 2006: The North American ensemble forecast system (NAEFS). *18th Conference on Probability and Statistics in the Atmospheric Sciences*, AMS.

Toth, Z., O. Talagrand, G. Candille, and Y. Zhu, 2003: Probability and ensemble forecasts. *Forecast verification: A practitioner's guide in atmospheric science*, I. Jolliffe and D. B. Stephenson, eds., Wiley, 137–163.

Wei, M., Z. Toth, R. Wobus, and Y. Zhu, 2008: Initial perturbations based on the ensemble transform (ET) technique in the NCEP global operational forecast system. *Tellus*, **60A**, 62–79.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*. Academic Press, San Diego, 2nd edition.

Yussouf, N. and D. J. Stensrud, 2007: Bias-corrected short-range ensemble forecasts of near surface variables during the 2005/2006 cool season. *Wea. Forecasting*, **22**, 1274–1286.

Zhu, Y., Z. Toth, R. Wobus, D. Richardson, and K. Mylne, 2002: The economic value of ensemble-based weather forecasts. *Bull. Amer. Meteor. Soc.*, **83**, 73–83.

Chapter 6

Conclusions

Improving NWP precipitation and temperature forecasts to benefit electricity generation summarizes the goal of this research project. The research presented in this dissertation has achieved this goal.

6.1 Summary of Methods and Procedures

Post-processing numerical temperature and precipitation forecasts and assessing the economic value of probabilistic hydrometeorological forecasts provides a measurable benefit to electricity generation systems. The first component of the method derived here established a viable and accurate technique of post-processing temperature and precipitation forecasts. Second, a full assessment of the meteorological quality and economic value of short-range ensemble precipitation forecasts provided the analysis necessary before synthesizing the forecasts into hydro-electric reservoir inflow models. Third, a complete evaluation of the quality and value of long-range bias-corrected ensemble temperature forecasts (required hydro-electric information for anticipating precipitation phase, electric load, energy market prices, and river-ice-front movement) concluded the research phase of this dissertation.

Both systematic error and random error degrade the quality and value of hydrometeorological numerical weather forecasts. Post-processing numerical model weather forecasts successfully minimizes systematic error attributable to poor representation of topography in the model, a technique especially suitable in regions of complex terrain. Post-processing relies on estimating past bias between forecasts and observations to predict future error in the forecasts. Bias-correcting results in reliable forecasts with negligible mean error.

Ideally the forecasts should be corrected for systematic error on a conditional basis, since different weather regimes and evolving models result in different, or conditional, bias factors. Conditional bias assessment remains difficult to achieve in a real-time operational forecast environment, so this study measured and assessed unconditional bias in the forecasts. A short window-length ensures that bias-correction adapts quickly to weather regime or model changes (emulating conditional bias-correction), while a long window-length promotes statistical stability. The study quantified optimal window-length facilitating these two competing factors.

The study compared seven competitive methods for daily maximum and minimum temperature forecasts, and three different methods for daily quantitative precipitation forecasts (in Chapter 2). The study assessed a set of forecasts produced for 19 observing sites within the complex terrain of southwest British Columbia, Canada. Achieving the goal of this first segment of the research required the development of an optimal post-processing scheme to calibrate systematic error in temperature and precipitation forecasts. The scheme must be easy to implement operationally, and must balance statistical stability with quick adaptation to changing weather regimes and model changes.

Post-processing cannot reduce random error in the forecasts. Only improving the actual forecast model reduces random error. Running a suite of different forecasts, termed a forecast ensemble, and assessing the resulting forecast probability distribution quantifies random error in the forecasts. Determining the user-dependent economic value of a forecast system requires such an assessment of forecast probability.

To generate a suite of different forecasts, a forecast ensemble designer will vary the computational components of a single model, will include output from different models, or will change the initial conditions in numerous runs of a single model. Combining all these methods into one probabilistic forecast results in a super-ensemble.

A judicious assessment of short-range ensemble forecasts of daily quantitative precipitation in this study utilized an arsenal of probabilistic forecast evaluation measures (in Chapter 3). A suite of three models running in different resolution configurations produced the forecasts. The forecasts were assessed over two consecutive 6-month wet seasons for 27 observing sites representing 15 hydro-electric watersheds in complex terrain. The evaluation measures extract important information for hydrometeorological users about the quality of probabilistic precipitation forecasts. Forecast quality comprises a complete analysis of reliability, resolution, skill, sharpness, discrimination, and spread.

Incorporating user needs into forecast evaluation upgrades an assessment of forecast quality to one of forecast value (see Chapter 4). Differing users, evaluating a set of forecasts assigned a particular level of quality, prescribe different measures of value to the same set of forecasts. Utilizing two different economic models provided a specific measurement of forecast value for the hydro-electric sector. A cost/loss model adapted here specifically for the hydro-electric energy sector initiated an overall assessment of forecast value dependent on particular watershed, reservoir, and hydraulic-head configuration. A decision-theory model adapted here to represent a single hydro-electric reservoir operation resulted in a direct monetary comparison between an ensemble forecast system and single deterministic forecasts.

Combining the successful post-processing approach devised in Chapter 2 (to minimize systematic error) with the meteorological and economic analyses presented in Chapters 3 and 4 (to assess probabilistic ensemble forecasts), a bias-corrected super-ensemble prediction system (presented in Chapter 5) refined a set of raw temperature forecasts into a valuable decision-making tool for hydro-electric generating system operators.

6.2 Summary of Findings

The results and findings of this research are summarized here:

- The best post-processing methods for minimizing mean error in daily maximum and minimum temperature forecasts included moving-weighted average methods and a Kalman-filter method. The optimal window-length for the moving-weighted average methods proved to be 14 days (McCollor and Stull, 2008b).

- The best post-processing methods for achieving mass balance in daily quantitative precipitation forecasts were a moving-average method and the best easy systematic estimator method. The optimal window-length for moving-average quantitative precipitation forecasts was 40 days (McCollor and Stull, 2008b).

- Given an operational precipitation forecast ensemble built from different models, each running at different resolutions, the best ensemble configuration included all models at high, medium, and low resolution to maximize reliability, resolution, skill, sharpness, discrimination, and spread performance (McCollor and Stull, 2008c).

- In rainfall-dominated regions of complex topography, reservoir operators managing high head, large storage, large areal reservoirs with low base inflows should drive their ensemble-based forecast decisions with reservoir-specific low cost/loss ratios (McCollor and Stull, 2008d).

- A reservoir-operation model based on decision theory and variable energy market pricing showed that applying an ensemble-average or full-ensemble precipitation forecast provided a much greater profit than using only a single deterministic high-resolution forecast (McCollor and Stull, 2008d).

- A bias-corrected super-ensemble prediction system consisting of 32 members producing daily maximum and minimum temperature forecasts for ten cities in western North America exhibited skill nine days into the future when using the ensemble average, and 12 days into the future when employing the full ensemble forecast (McCollor and Stull, 2008a).

6.3 Discussion and Recommendations

This dissertation proved that bias-correcting temperature and precipitation forecasts and providing users with an ensemble-generated probability distribution of forecasts provides concrete monetary benefits to the hydro-electric energy sector. This dissertation provides explicit details of methods to post-process forecasts and generate quality ensembles to build a valuable decision-making tool for this extremely important sector of any regional or national economy. The techniques documented in this dissertation work exceedingly well in the complex topography and extreme climatic regimes of western North America. These methods embody no intrinsic limitations preventing their similar application in any other geographic region with a viable energy generation sector, or in fact any temperature or precipitation dependent sector of the economy.

Future work to further improve the post-processing techniques presented in Chapter 2 would include a multivariate approach to weighting error estimates. Including variables such as integrated water vapour and flow direction in a multivariate approach would likely prove beneficial in a post-processing technique to improve precipitation forecasts. However much more training data is needed for multivariate approaches that calibrate forecasts conditioned on specific events.

A multi-model approach is also a cost-effective way to reduce error in forecasts. A combination of a model-averaging approach (Ebert, 2001) and the bias-correction techniques presented here may prove beneficial (Delle Monache et al., 2006), especially for precipitation forecasting.

Another approach that may provide positive results is autoregressive-moving average (ARMA) modelling of direct-model-output (DMO) error. ARMA models of meteorological time series gained popularity in the late 1970's and early 1980's, after Box and Jenkins (1976) published readily accessible statistical methodology for applying ARMA models to time series analysis. However meteorological ARMA models of the time were limited to seasonal analyses applications such as drought index modeling, monthly precipitation, and annual streamflow (Katz and Skaggs, 1981). ARMA techniques do provide a method to estimate future error between DMO and observations using optimized weighted past error measurements. Currently, however, major efforts in using ARMA and Box-Jenkins approaches are concentrated in the financial world. Complex ARMA schemes depend critically on developmental data (as does MOS) and hence perform poorly after changes to the underlying model.

The next stage of research, building on the work presented in this dissertation, should quantify the quality and value of probabilistic precipitation forecasts in a medium-range ensemble forecast system (day three and beyond). The results of a medium-range ensemble forecast system for precipitation would complement the results documented for temperature presented in Chapter 5.

Another important avenue for hydrometeorological research would adapt the methods prescribed in this dissertation to short-range high-inflow ensemble forecasting. The daily time-step for precipitation and temperature forecasts selected in this dissertation would change to a 6-hourly or 3-hourly timestep projected 24 or 48 hours into the future. Such a high-inflow forecasting system design should operate on a 6 or 12-hour update cycle for the duration of a high-rainfall event.

The techniques of numerical weather forecast post-processing and methods of forecast quality and value assessment adapted in this dissertation for hydro-electric generation could transfer to the wind-power generation sector. Wind-power generation capacity accounts for approximately 0.5% of the world's power generation (Gil, 2007). Over 60% of the world's installed wind power capacity resides in Europe, and 16% resides in the USA. India, Japan, and Canada produce only a small portion of their domestic electric power with windfarms. Wind power represents one of the fastest growing renewable-energy technologies, renowned for its many advantages: it is non-carbon based, it reduces a nation's dependence on conventional non-renewable carbon-based sources, and it diversifies a nation's energy portfolio. Wind power is touted as one of the most cost-effective methods of electricity generation available; costs have fallen around 90% during the last 20 years. Accurate wind-speed forecasts would improve efficiency and add value to an integrated wind-power electrical generation system in much the same way hydrometeorological forecasts aid hydropower systems.

Water resource managers operating a reservoir face many complex decisions. An automated, skillful, reliable ensemble streamflow forecast product is conceptually appealing (Schaake et al., 2007). Schaake (ibid.) described an "end-to-end" forecast system that explicitly accounts for the major sources of uncertainty in the forecasting process, including hydrologic model and initial basin-state error as well as atmospheric model and initial-condition error. The meteorological component of the end-to-end forecasting system provides atmospheric forcing for the hydrologic model that has been downscaled and converted to calibrated, bias-corrected temperature and quantitative precipitation forecasts. In this dissertation I have shown explicitly how to perform these crucial steps to improve probabilistic forecast quality.

Schaake (ibid.) stated that, as part of the Hydrologic Ensemble Prediction Experiment (HEPEX) project, end-to-end forecasting systems are just beginning to be assembled, and there are many basic and applied science questions that remain to be answered in order to build and promote the rapid development of useful systems. In this dissertation I have also shown that economic models, (e.g., the cost/loss and decision theory models adapted in this dissertation), integrated onto the stream of information flowing through atmospheric and hydrologic models, would truly constitute an end-to-end decision process proving extremely valuable to water resource managers.

Bibliography

Box, G. E. P. and G. M. Jenkins, 1976: *Time Series Analysis: Forecasting and Control (rev)*. Holden-Day, Oakland.

Delle Monache, L., X. Deng, Y. Zhou, and R. Stull, 2006: Ozone ensemble forecasts: 1. A new ensemble design. *J. Geophys. Res.*, **111**, D05307, doi:10.1029/2005JD006310.

Ebert, E. E., 2001: Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Wea. Rev.*, **129**, 2461–2480.

Gil, S. R., 2007: The socio-economic benefits of climatological services to the renewable energy sector. *Bulletin of the World Meteorological Organization*, **56**, 40–45.

Katz, R. W. and R. H. Skaggs, 1981: On the use of autoregressive-moving average processes to model meteorological time series. *Mon. Wea. Rev.*, **109**, 479–484.

McCollor, D. and R. Stull, 2008a: Evaluation of probabilistic medium-range temperature forecasts from the North American ensemble forecast system. *submitted to Wea. Forecasting*.

— 2008b: Hydrometeorological accuracy enhancement via postprocessing of numerical weather forecasts in complex terrain. *Wea. Forecasting*, **23**, 131–144.

— 2008c: Hydrometeorological short-range ensemble forecasts in complex terrain. Part I: Meteorological evaluation. *Wea. Forecasting*, (in press).

— 2008d: Hydrometeorological short-range ensemble forecasts in complex terrain. Part II: Economic evaluation. *Wea. Forecasting*, (in press).

Schaake, J. C., T. M. Hamill, R. Buizza, and M. Clark, 2007: HEPEX: The hydrologic ensemble prediction experiment. *Bull. Amer. Meteor. Soc.*, **88**, 1541–1547.

Appendix A

Kalman Filter Bias Correction

The Kalman filter (KF) is a recursive, adaptive technique to estimate a signal from noisy measurements (Bozic, 1994; Zarchan and Musoff, 2005). Kalman filter theory provides recursive equations for continuously updating error estimates via observations of a system involving inherent unknown processes. The KF has been employed as a predictor bias-correction method during post-processing of short-term NWP forecasts. Thorough descriptions of KF theory, equations, and applications to weather forecast models can be found in the meteorological literature (Homleid, 1995; Roeger et al., 2003; Delle Monache et al., 2006). The KF approach is adapted here for daily maximum and minimum temperature forecasts.

Recent bias values are used as input to the KF. The filter estimates the bias in the current forecast. As in the other post-processing methods examined in this study, the expected bias in the current forecast as estimated by the KF is removed from the current DMO forecast. The new corrected forecast should have improved error characteristics over the original DMO forecast.

Let x_k be the bias between the forecast and the verifying observation valid for time step *k*. The bias x_k is the signal we would like to predict for the next forecast period (at *k+1*). The future bias is determined by a persistence of the current bias plus a Gaussian-distributed random term w_k of variance σ_w^2: $x_{k+1} = x_k + w_k$.

Similarly our measurements y_k of the bias are assumed to be noisy, with random error term v_k (of variance σ_v^2): $y_k = x_k + v_k$. The objective is to get the best estimate of x_k, which is termed $\hat{x}_k$, by minimizing the expected mean-square error $p = E[(x - \hat{x})^2]$.

The recursive nature of the technique is characterized by a continuous update-predict cycle. The measurement update, or "corrector" portion of the cycle, is determined by the following equations:

Compute the Kalman gain β:

$$\beta_k = \frac{p_{k-1} + \sigma_w^2}{p_{k-1} + \sigma_w^2 + \sigma_v^2} \tag{A.1}$$

Update the estimate $\hat{x}_k$:

$$\hat{x}_k = \hat{x}_{k-1} + \beta_k[y_k - \hat{x}_{k-1}] \tag{A.2}$$

Finally, update the error covariance term p_k:

$$p_k = (p_{k-1} + \sigma_w^2)(1 - \beta_k) \tag{A.3}$$

The ratio r is defined as σ_w^2/σ_v^2. A value of $r = 0.01$ is used in this study, as suggested from previous studies (Roeger et al., 2003; Delle Monache et al., 2006).

The time update, or "predictor" portion of the cycle, is determined by:

$$\hat{x}_{k+1} = \hat{x}_k \tag{A.4}$$

$$p_{k+1} = p_k + \sigma_w^2 \tag{A.5}$$

Initial starting values of $\hat{x}(0)$ and $p(0)$ are chosen to start the process. The equations converge quickly so that the results are not sensitive to the particular initial values which begin the process.

Bibliography

Bozic, S. M., 1994: *Digital and Kalman Filtering, Second Ed.*. John Wiley and Sons, New York.

Delle Monache, L., T. Nipen, X. Deng, Y. Zhou, and R. Stull, 2006: Ozone ensemble forecasts: 2. A Kalman filter predictor bias correction. *J. Geophys. Res.*, **111**, D05308, doi:10.1029/2005JD006311.

Homleid, M., 1995: Diurnal corrections of short-term surface temperature forecasts using the Kalman Filter. *Wea. Forecasting*, **10**, 689–707.

Roeger, C., R. Stull, D. McClung, J. Hacker, X. Deng, and H. Modzelewski, 2003: Verification of mesoscale numerical weather forecasts in mountainous terrain for application to avalanche prediction. *Wea. Forecasting*, **18**, 1140–1160.

Zarchan, P. and H. Musoff, 2005: *Fundamentals of Kalman Filtering: A Practical Approach*. American Institute of Aeronautics and Astronautics, Inc., 2nd edition, United States.

Appendix B

Statistical Significance Hypothesis Test

Statistical significance tests are formulated to determine the likelihood that an observed outcome could have arisen by chance instead of design. For the 24-hour QPFs analyzed in this study, a non-parametric hypothesis test methodology designed for precipitation forecasts [see Hamill (1999) for details of the method] was employed.

The technique is based on resampling. The null hypothesis for the resampling test is that the difference in DMB between each of the SNL, MA, and BES forecasts and the baseline DMO forecast is zero:

$$H_0 : \qquad DMB_{PP} - DMB_{DMO} = 0.0 \tag{B.1}$$

and the alternative hypothesis:

$$H_A : \qquad DMB_{PP} - DMB_{DMO} \neq 0.0 \tag{B.2}$$

where DMB_{PP} is the DMB of the post-processing method SNL, MA, or BES, and DMB_{DMO} is the DMB of the direct model output. Assume a non-symmetric two-sided test with significant level $\alpha = 0.05$. Create a virtual time series by resampling the existing set of forecasts to form a test statistic consistent with the null hypothesis above. The test statistic,

$$(\widehat{DMB}_{PP} - \widehat{DMB}_{DMO}), \tag{B.3}$$

is calculated using equation (2.3) in Chapter 2 section 2.2.2. The virtual $\binom{*}{1}$ resampled test statistic time series consistent with the null hypothesis is generated by randomly selecting either the post-processed forecast or the DMO baseline forecast for each day and calculating the DMB for this time series. A second virtual $\binom{*}{2}$ randomly-selected time series is formed using the alternate selection of post-processed and DMO forecasts. The resampled test statistic,

$$(\widehat{DMB}_1^* - \widehat{DMB}_2^*), \tag{B.4}$$

is repeated 1000 times to build a null distribution.

The final step is to determine whether $(\widehat{DMB}_{PP} - \widehat{DMB}_{DMO})$ can be considered to fall within the distribution of $(\widehat{DMB}_1^* - \widehat{DMB}_2^*)$ values (the null hypothesis), or whether the null hypothesis can be rejected. The virtual time series resampled distributions are utilized to compute the locations $\widehat{t_L}$ and $\widehat{t_U}$ such that

$$\begin{aligned} Pr^*[(\widehat{DMB}_1^* - \widehat{DMB}_2^*) - \widehat{t_L}] &= \frac{\alpha}{2}\text{ , and} \\ Pr^*[(\widehat{DMB}_1^* - \widehat{DMB}_2^*) - \widehat{t_U}] &= 1 - \frac{\alpha}{2} \end{aligned} \tag{B.5}$$

where Pr^* represents the probabilities numerically calculated from this distribution. Then, H_0 is rejected if

$$(\widehat{DMB}_{PP} - \widehat{DMB}_{DMO}) < \widehat{t_L} \quad \text{or} \quad (\widehat{DMB}_{PP} - \widehat{DMB}_{DMO}) > \widehat{t_U}$$

In graphical displays of the post-processed forecast results, the values of t_U and t_L are represented by statistical significance error bars. Forecast differences outside the interval expressed by the error bars may be considered statistically significant for the particular α.

Bibliography

Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, **14**, 155–167.

Appendix C

Equations for Meteorological Statistical Analysis

Given f_k as forecast precipitation value of the k^{th} forecast, y_k as the corresponding observed value, $\overline{f}$ as mean forecast value, $\overline{y}$ as mean observed value, and N as the number of forecast-observation pairs, then the following definitions apply:

Degree of mass balance (DMB) (Grubišić et al., 2005):

$$DMB = \frac{\sum_{k=1}^{N} f_k}{\sum_{k=1}^{N} y_k} \tag{C.1}$$

DMB is the ratio of the predicted to the observed net water mass over the study period. Values of DMB < 1 indicate that 24-hour precipitation is under-forecast by the models. A value of DMB ≈ 1 indicates the 24-hour precipitation forecasts are in balance with the observations.

Mean error (ME):

$$ME = \overline{f} - \overline{y} \tag{C.2}$$

Mean absolute error (MAE):

$$MAE = \frac{1}{N}\sum_{k=1}^{N} |f_k - y_k| \tag{C.3}$$

Mean Square Error:

$$MSE = \frac{1}{N}\sum_{k=1}^{N} (f_k - y_k)^2 \tag{C.4}$$

Root Mean Square Error (RMSE):

$$RMSE = \sqrt{MSE} \tag{C.5}$$

Pearson correlation r:

$$r = \frac{\sum_{k=1}^{N}(f_k - \overline{f})(y_k - \overline{y})}{\sqrt{\sum_{k=1}^{N}(f_k - \overline{f})^2 \sum_{k=1}^{N}(y_k - \overline{y})^2}} \tag{C.6}$$

Linear error in probability space:

Linear error in probability space (LEPS) is defined as the mean absolute difference between the cumulative frequency of the forecasts and the cumulative frequency of the observations (Déqué, 2003). LEPS ensures that error in the center of the distribution is treated with more importance than error found in the extreme tail of the distribution. A LEPS skill score can be defined with the climatological median as a reference (Wilks, 2006).

Brier score (reliability and resolution):

The most widely used EPS evaluation score is the Brier score (Brier, 1950; Atger, 2003; Gallus et al., 2007), designed to quantify the performance of a probabilistic forecast of a dichotomous event. The Brier score (BS) is defined as the mean square error of the probability forecast:

$$BS = \frac{1}{N}\sum_{k=1}^{N}(p_k - o_k)^2 \tag{C.7}$$

where N is the number of forecasts, p_k is the probability that forecast precipitation will exceed a given threshold (estimated by the number of ensemble members that exceed that threshold), and o_k is the verifying observation (equal to 1 if the observed precipitation exceeds the threshold, 0 if it does not). The Brier score is negatively oriented, in that a score of 0 is a perfect forecast and increasing Brier score values, to a maximum of 1, indicate deteriorating performance.

Murphy (1973) showed how equation (C.7) could be partitioned into three parts, measuring the degree of reliability, resolution, and uncertainty in the forecasts and associated observations. In Murphy's decomposition, the verification dataset contains J different probability forecast values, where n_j is the number of cases in the jth forecast category. For each forecast category j, the average of the observations in that category is determined:

$$\overline{o}_j = \frac{1}{n_j}\sum_{k \in n_j} o_k \tag{C.8}$$

Additionally, the overall sample climatology of the defined event is measured as the average of all observations:

$$\overline{o} = \frac{1}{N}\sum_{k=1}^{N} o_k \tag{C.9}$$

Given definitions (C.8) and (C.9), equation (C.7) can be rewritten:

$$BS = \frac{1}{N}\sum_{j=1}^{J} n_j(p_j - \overline{o}_j)^2 - \frac{1}{N}\sum_{j=1}^{J} n_j(\overline{o}_j - \overline{o})^2 + \overline{o}(1-\overline{o}) \tag{C.10}$$

where the first term is the reliability component. Reliability is the correspondence between a given probability and the observed frequency of an event in those cases the event is forecast with the given probability. The reliability term quantifies the information provided in a reliability diagram.

The second term is the resolution component. The resolution term indicates the extent that the different forecast categories do in fact reflect different frequencies of occurrence of the observed event.

The last term in the decomposition is the uncertainty. The uncertainty term denotes the intrinsic difficulty in forecasting the event but depends on the observations only, not on the forecasting system.

The Brier score can be converted to a positively-oriented skill score:

$$BSS = 1 - \frac{BS}{BS_{ref}} \tag{C.11}$$

The reference system is often taken to be the low-skill *climatological forecast* in which the probability of the event is equal to $\overline{o}$ for all forecasts. The Brier score for such climatological forecasts is $BS_{ref} = BS_c = \overline{o}(1-\overline{o})$.

The Brier skill score is then expressed:

$$BSS = \frac{\text{resolution}}{\text{uncertainty}} - \frac{\text{reliability}}{\text{uncertainty}} = \text{relative resolution} - \text{relative reliability}$$

A perfect forecast would have the relative resolution equal to one and the relative reliability equal to zero.

An adjustment factor must be included in the calculations for Brier skill score to account for SREF systems with different ensemble sizes. Richardson (2001) provided a method to compensate for the effect of ensemble size on the Brier skill score. The relationship between BSS_M (the Brier skill score for an EPS represented by M members) and BSS_∞ (the Brier skill score for the full underlying probability distribution of forecasts as $M \to \infty$) is:

$$BSS_\infty = \frac{M \cdot B_M + 1}{M + 1} \tag{C.12}$$

The Brier skill score BSS_∞ can then be expressed as a sum of the relative resolution component and relative reliability component of the Brier score decomposition:

$$BSS_\infty = \frac{M \cdot BSS_{M_{rel\ res}} + 1}{M + 1} - \frac{M \cdot BSS_{M_{rel\ reliab}}}{M + 1} \tag{C.13}$$

Reliability:

Reliability is an essential attribute of the quality of probabilistic forecasts (Atger, 2003). Reliability is often represented with the aid of a reliability diagram of observed relative frequency o_k vs. forecast probability p_k (Wilks, 2006; Toth et al., 2003; Clark et al., 2004; Gallus et al., 2007). Ensemble forecasts are converted to probabilistic forecasts by determining what percentage of the ensemble members meet the specific event criterion. In the current study the event criterion is that 24-hour precipitation exceeds a threshold value of 5 mm, 10 mm, 25 mm, or 50 mm.

Ideally the probabilistic forecast/observation points lie on the diagonal of the reliability diagram, indicating the event is always forecast at the same frequency it is observed. The reliability component of the Brier score in graphical representation is the weighted, averaged, squared distance between the reliability curve and the 45° diagonal line. If the points lie above (below) the diagonal it means the event is underforecast (overforecast). Reliability curves with a zig-zag shape centered on the diagonal indicate good reliability represented by a small sample size. Poor reliability can be improved substantially by appropriate *a posteriori* calibration and/or *post*-processing of forecasts delivered from an established system, though it is a difficult task to achieve in a real-time operational EPS (Atger, 2003).

Associated with reliability is sharpness, which characterizes the relative frequency of occurrence of the forecast probabilities. Sharpness is often depicted in a histogram indicating the relative occurrence of each forecast probability category. If forecast probabilities are frequently near zero or near 1 then the forecasts are sharp, indicating the forecasts deviate significantly from the climatological mean, a positive attribute of an ensemble forecast system. Sharpness measures the variability of the forecasts alone, without regard to their corresponding observations; hence it is not a verification measure in itself. In a perfectly reliable forecast system, sharpness is identical to resolution.

Resolution:

A useful probabilistic forecast system must be able to *a priori* differentiate future weather outcomes, so that differing forecasts are, in fact, associated with distinct verifying observations. This is the most important attribute of a forecast system (Toth et al., 2003) and is called *resolution*.

Resolution cannot be improved through simple adjustment of probability values or statistical post-processing. Resolution can be gained only by improving the actual forecast-model engine that produces the forecasts.

ROC curve:

Discrimination, which is the converse of resolution, reflects an EPS's ability to distinguish between the occurrence and non-occurrence of forecast events; in other words the sensitivity of the probability that an event was forecast conditioned on whether or not the event was observed. In the case that observed frequencies of forecast events monotonically increase with increasing forecast probabilities, resolution and discrimination convey the same information about an EPS (Buizza et al., 2005). Resolution and discrimination are based on two different factorizations of the forecast/observed pair of events, as described in the distributions-oriented approach to Murphy and Winkler's (Murphy and Winkler, 1987) general framework for forecast verification.

The relative operating characteristic (ROC) is a verification measure based on signal detection theory that offers a way to examine the discrimination of an EPS (Mason, 1982; Mason and Graham, 1999; Gallus et al., 2007). The ROC curve is a graph of hit-rate versus false-alarm rate (see below for definitions of these terms and Table C.1 for the contingency table basis for this analysis). The ROC measure is based on stratification by observations, and therefore is independent of reliability and instead provides a direct measure of resolution. The ROC measure is particularly valuable in assessing the general issue of ensemble size/configuration versus model resolution. Additionally, potential economic value for the cost/loss decision model (Murphy, 1977; Katz and Murphy, 1997; Richardson, 2000; Zhu et al., 2002; Richardson, 2003) is uniquely determined from ROC curves (see Chapter 4 for an economic analysis of the probabilistic forecasts).

Contingency-table analysis equations:

Given event counts a, b, c, and d from Table C.1, the hit rate H and false alarm rate F are defined as:

Hit Rate H:

$$H = \frac{a}{a + c} \tag{C.14}$$

False Alarm Rate F:

$$F = \frac{b}{b + d} \tag{C.15}$$

Perfect discrimination is represented by a ROC curve that rises from (0,0) along the Y-axis to (0,1) then straight to (1,1). The diagonal line represents zero skill meaning the forecasts show no discrimination among events. The area under the ROC curve (ROC_{Area}) is a measure of skill that can be used to compare different probabilistic forecast systems. Perfect discrimination results in a ROC area value of 1.0 while the no-skill diagonal corresponds to a ROC area value of 0.5. From these values a ROC skill score ($ROCSS$) can be defined to range from 0 (no-skill forecasts) to 1 (perfectly discriminating forecasts):

$$ROCSS = 2 \cdot ROC_{Area} - 1 \tag{C.16}$$

ROC area is determined from A_z, the area under the *modelled* ROC on probability axes (Mason, 1982; Swets and Pickett, 1982; Swets, 1988; Mason, 2003). In practice, the ROC curve is transformed (see Figs. C.1 and C.2) to a plot on normal deviate axes that results in a straight line [under the assumption of a normal (Gaussian) probability distribution (Swets and Pickett, 1982)]. The straight line can be estimated from the data using a least-squares method of linear interpolation. The slope and y-axis-intercept of this line can be used to evaluate $z(A)$:

$$z(A) = \frac{s\Delta m}{(1 + s^2)^{1/2}} \tag{C.17}$$

where s is the slope of the line and Δm is the y-axis-intercept (with the sign reversed). Transforming $z(A)$ back to probability space through the use of the cumulative standardized normal distribution gives A_z as the area under the ROC curve on probability axes.

Equal likelihood:

Rank histograms, also known as Talagrand diagrams (Anderson, 1996; Talagrand et al., 1998), provide a necessary but not sufficient test for evaluating whether the forecast and verification are sampled from the same probability distribution. The rank histogram is a useful measure of the realism of an ensemble forecast system (Hou et al., 2001), providing a measure of reliability (Candille and Talagrand, 2005). For an ensemble forecast with M members, there will be $M + 1$ intervals defined by the members, including the two openended intervals. A rank histogram is built by accumulating the number of cases that an observation falls in each of the $M+1$ intervals.

In an ideal ensemble, the range of the ensemble forecast members represents the full spread of the probability distribution of observations, and each member exhibits an equal likelihood of occurrence; hence the resulting rank histogram is flat. Ensemble forecasts with insufficient spread to adequately represent the probability distribution of observations are indicated by a U-shaped histogram. Rank histograms computed from operational ensembles are commonly U-shaped, which is traditionally interpreted as the consequence of the lack of ensemble spread in the EPS (Anderson, 1996; Atger, 2003). Alternately, U-shaped rank histograms may be interpreted as a consequence of conditional model biases (Hamill, 2001), rather than point to a weakness of the method used for generating the ensemble.

A measure of the deviation of the rank histogram from flatness is described by Candille and Talagrand (2005) for an EPS with M members and N available forecast/observation pairs. The number of values in the i_{th} interval of the histogram is given by s_i. For a reliable system in which the histogram is flat, $s_i = N/(M+1)$ for each interval i. The quantity

$$\Delta = \sum_{i=1}^{M+1} (s_i - \frac{N}{M+1})^2 \tag{C.18}$$

measures the deviation of the histogram from flatness. For a reliable system, the base value is $\Delta_0 = NM/(M+1)$. The ratio

$$\delta = \Delta/\Delta_0 \tag{C.19}$$

is evaluated as the overall measure of the flatness of an ensemble prediction system rank histogram. A value of δ that is significantly larger than 1 indicates the system does not reflect equal likelihood of ensemble members.

An alternative summary index based on the rank histogram, called a “reliability index” (RI), has recently been introduced by Delle Monache (2005); see also Delle Monache et al. (2006). The RI measures the variation of the rank histogram from its ideal “flat” shape and is normalized so that ensembles of different sizes can be compared with each other.

Table C.1: Contingency table for hit rate and false alarm rate calculations. The counts a, b, c, and d are the number of events in each category, out of N total events.

	Observed	Not observed
Forecast	a	b
Not forecast	c	d

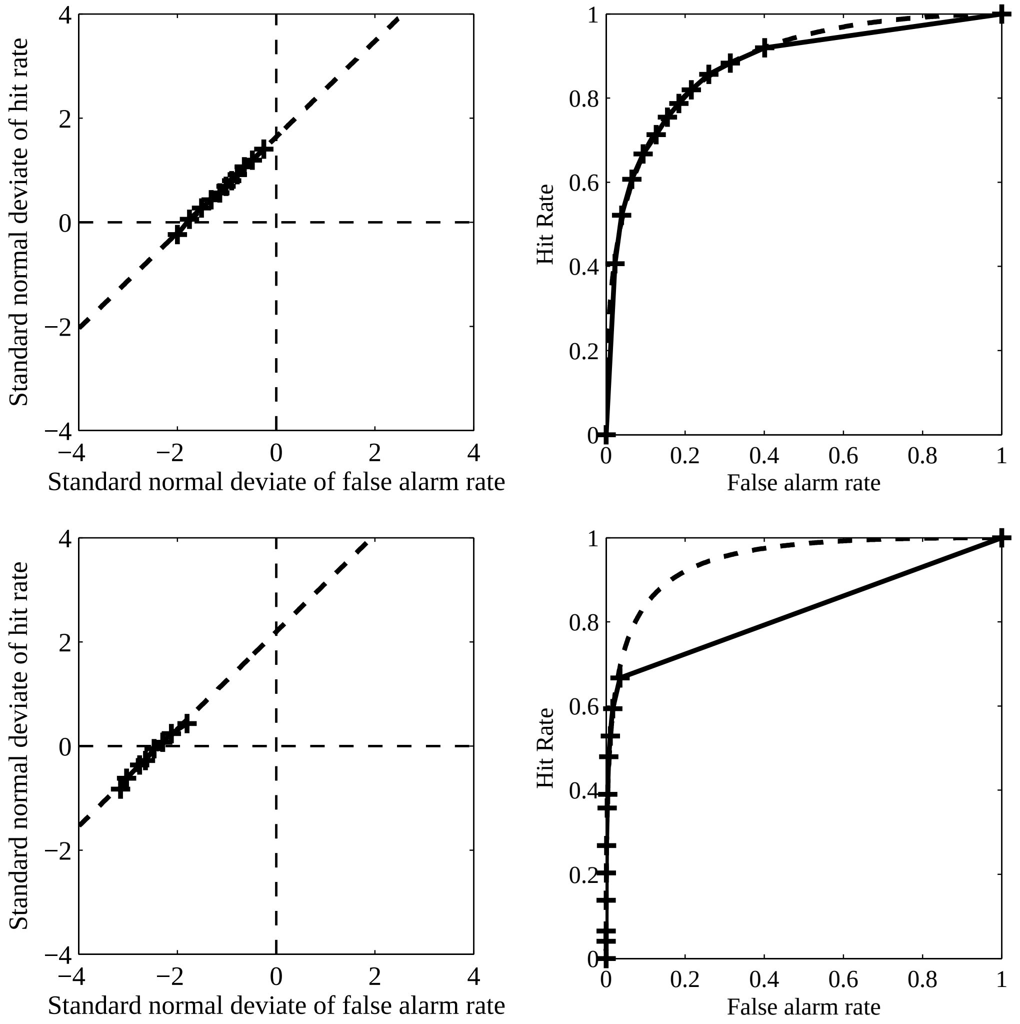

Figure C.1: Left: ROC values transformed to standard normal deviates of hit rate H and false alarm rate F. Crosses are the actual transformations, while the dashed line is the least squares straight-line fit. Right: ROC curves transformed back to probability axes. Original ROC points are shown by crosses connected together by a solid line. Transforming the straight-line from the plots on the left back to probability axes gives the dashed line in the plots on the right. The area under the transformed, or modelled ROC curve is A_z. Upper plots: day-one full 11-member ensemble forecasts with the 5 mm day^{-1} precipitation threshold. Lower plots: the same ensemble but for the rarer 50 mm day^{-1} precipitation threshold.

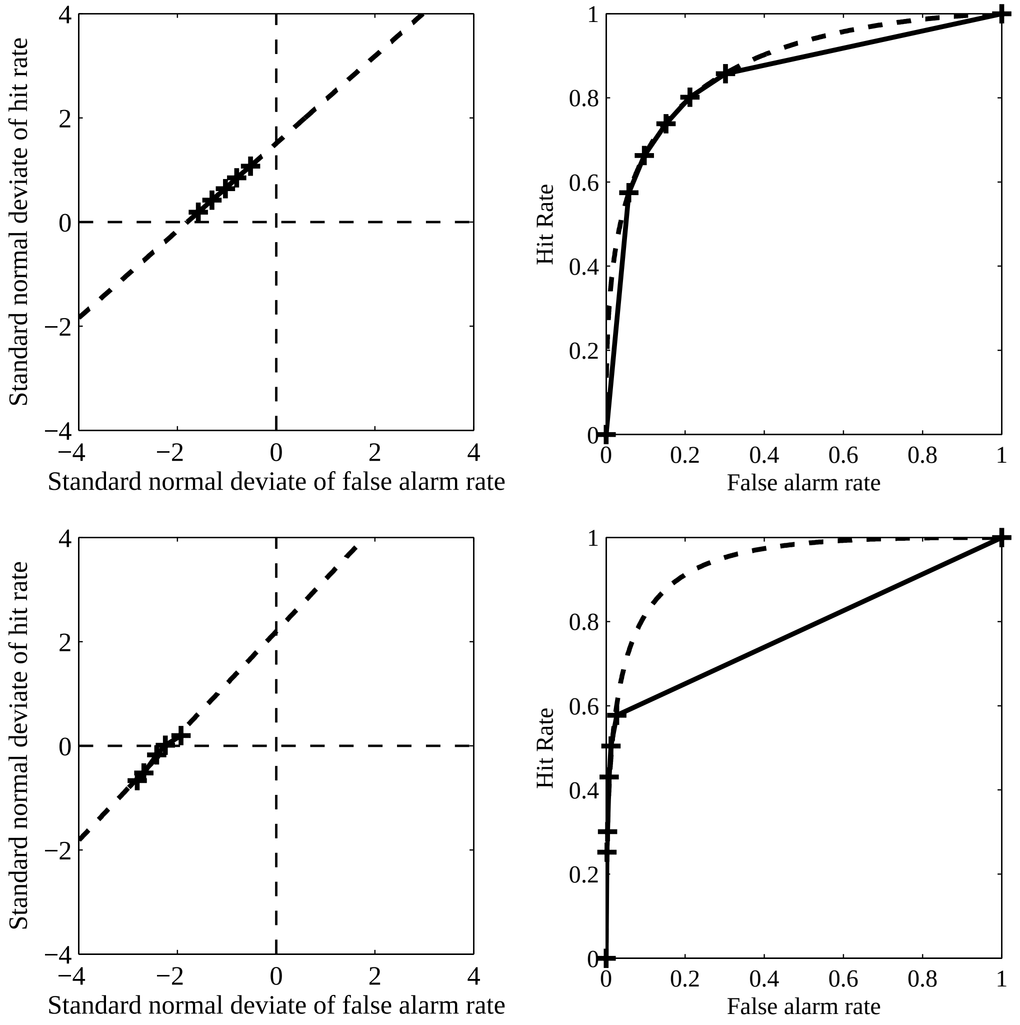

Figure C.2: Same as fig. C.1 except for the highest resolution ensemble vhires5.

Bibliography

Anderson, J. L., 1996: A method for producing and evaluating probabilistic precipitation forecasts from ensemble model integrations. *J. Climate*, **9**, 1518–1529.

Atger, F., 2003: Spatial and interannual variability of the reliability of ensemble-based probabilistic forecasts: Consequences for calibration. *Mon. Wea. Rev.*, **131**, 1509–1523.

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

Buizza, R., P. L. Houtekamer, Z. Toth, G. Pellerin, M. Wei, and Y. Zhu, 2005: A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Mon. Wea. Rev.*, **133**, 1076–1097.

Candille, G. and O. Talagrand, 2005: Evaluation of probabilistic prediction systems for a scalar variable. *Quart. J. Roy. Meteor. Soc.*, **131**, 2131–2150.

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. Hydrometeorology*, **5**, 243–262.

Delle Monache, L., 2005: *Ensemble-Averaged, Probabilistic, and Kalman-Filtered Regional Ozone Forecasts*. Ph.D. thesis, University of British Columbia.

Delle Monache, L., J. P. Hacker, Y. Zhou, X. Deng, and R. B. Stull, 2006: Probabilistic aspects of meteorological and ozone regional ensemble forecasts. *J. Geophys. Res.*, **111**, D24307, doi:10.1029/2005JD006917.

Déqué, M., 2003: Continuous variables. *Forecast verification: A practitioner's guide in atmospheric science*, I. Jolliffe and D. B. Stephenson, eds., Wiley, 97–119.

Gallus, W. A. J., M. E. Baldwin, and K. L. Elmore, 2007: Evaluation of probabilistic precipitation forecasts determined from Eta and AVN forecasted amounts. *Wea. Forecasting*, **22**, 207–215.

Grubišić, V., R. K. Vellore, and A. W. Huggins, 2005: Quantitative precipitation forecasting of wintertime storms in the Sierra Nevada: Sensitivity to the microphysical parameterization and horizontal resolution. *Mon. Wea. Rev.*, **133**, 2834–2859.

Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.*, **129**, 550–560.

Hou, D., E. Kalnay, and K. K. Droegemeier, 2001: Objective verification of the SAMEX '98 ensemble forecasts. *Mon. Wea. Rev.*, **129**, 73–91.

Katz, R. W. and A. H. Murphy, 1997: *Economic Value of Weather and Climate Forecasts*. Cambridge University Press, England.

Mason, I., 1982: A model for assessment of weather forecasts. *Aust. Meteor. Mag.*, **30**, 291–303.

Mason, I. and N. E. Graham, 1999: Conditional probabilities, relative operating characteristics, and relative operating levels. *Wea. Forecasting*, **14**, 713–725.

Mason, I. B., 2003: Binary events. *Forecast verification: A practitioner's guide in atmospheric science*, I. Jolliffe and D. B. Stephenson, eds., Wiley, 137–163.

Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.*, **12**, 595–600.

— 1977: The value of climatological, categorical, and probabilistic forecasts in the cost-loss ratio situation. *Mon. Wea. Rev.*, **105**, 803–816.

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

Richardson, D. S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **126**, 649–667.

— 2001: Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Quart. J. Roy. Meteor. Soc.*, **127**, 2473–2489.

— 2003: Economic value and skill. *Forecast verification: A practitioner's guide in atmospheric science*, I. Jolliffe and D. B. Stephenson, eds., Wiley, 164–187.

Swets, J. A., 1988: Measuring the accuracy of diagnostic systems. *Science*, **240**, 1285–1293.

Swets, J. A. and R. M. Pickett, 1982: *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*. Academic Press, New York.

Talagrand, O., R. Vautard, and B. Strauss, 1998: Evaluation of probabilistic prediction systems. *Workshop on predictability*, ECMWF, 1–25.

Toth, Z., O. Talagrand, G. Candille, and Y. Zhu, 2003: Probability and ensemble forecasts. *Forecast verification: A practitioner's guide in atmospheric science*, I. Jolliffe and D. B. Stephenson, eds., Wiley, 137–163.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*. Academic Press, San Diego, 2nd edition.

Zhu, Y., Z. Toth, R. Wobus, D. Richardson, and K. Mylne, 2002: The economic value of ensemble-based weather forecasts. *Bull. Amer. Meteor. Soc.*, **83**, 73–83.