A Probabilistic Inflow Forecasting System for Operation of Hydroelectric Reservoirs in Complex Terrain

by

Dominique R. Bourdin

B. Sc. Atmospheric Science, The University Of British Columbia, 2009

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

Doctor of Philosophy

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Atmospheric Science)

The University Of British Columbia
(Vancouver)

September 2013

Abstract

This dissertation presents a reliable probabilistic forecasting system designed to predict inflows to hydroelectric reservoirs. Forecasts are derived from a Member-to-Member (M2M) ensemble in which an ensemble of distributed hydrologic models is driven by the gridded output of an ensemble of numerical weather prediction (NWP) models. Multiple parameter sets for each hydrologic model are optimized using objective functions that favour different aspects of forecast performance. On each forecast day, initial conditions for each differently-optimized hydrologic model are updated using meteorological observations. Thus, the M2M ensemble explicitly samples inflow forecast uncertainty caused by errors in the hydrologic models, their parameterizations, and in the initial and boundary conditions (i.e., meteorological data) used to drive the model forecasts.

Bias is removed from the individual ensemble members using a simple degree-of-mass-balance bias correction scheme. The M2M ensemble is then transformed into a probabilistic inflow forecast by applying appropriate uncertainty models during different seasons of the water year. The uncertainty models apply ensemble model output statistics to correct for deficiencies in M2M spread. Further improvement is found after applying a probability calibration scheme that amounts to a re-labelling of forecast probabilities based on past performance.

Each component of the M2M ensemble has an associated cost in terms of time and/or money. The relative value of each ensemble component is assessed by removing it from the ensemble and comparing the economic gains associated with the reduced ensembles to those achieved using the full M2M system. Relative value is computed using a simple (static) cost-loss decision model in which the reservoir operator takes action (lowers the reservoir level) when significant inflows are predicted with probability exceeding some threshold.

The probabilistic reservoir inflow forecasting system developed in this dissertation is applied to the Daisy Lake hydroelectric reservoir located in the complex terrain of southwestern British Columbia, Canada. The hydroclimatic regime of the case study watershed is such that flashy fall and winter inflows are driven by Pacific frontal systems, while spring and summer inflows are dominated by snow and glacier melt. Various aspects of ensemble and probabilistic forecast performance are evaluated over a period of three water years.

Preface

The main body of this dissertation is comprised of work from two published journal papers (Chapter 1 and a combination of Chapters 2 and 3), one submitted paper (Chapter 4), and one in preparation (Chapter 5). Submitted and published papers have been reformatted to meet dissertation formatting requirements. Minor editing changes may have been made, but the content is otherwise unaltered.

Funding for this research was provided by the Canadian Natural Sciences and Engineering Research Council (NSERC) in the form of an Alexander Graham Bell Canada Graduate Scholarship. Additional funding was provided by a NSERC Discovery Grant to Professor Stull.

Chapter 1

Sections 1.1 and 1.2.1 of the introductory chapter are adapted from the following published paper:

Bourdin, D. R., S. W. Fleming and R. B. Stull, 2012: Streamflow modelling: A primer on applications, approaches and challenges. *Atmosphere-Ocean*, **50**, 507-536.

I wrote the original manuscript, with significant guidance and contributions from Dr. Fleming and editing by Professor Stull. I wrote the remainder of the chapter.

Chapter 2

Contents of Chapter 2 (Bias-Corrected Short-Range Member-to-Member Ensemble Forecasts of Reservoir Inflow) are adapted from the following published paper:

Bourdin, D. R. and R. B. Stull, 2013: Bias-corrected short-range Member-to-Member ensemble forecasts of reservoir inflow. *Journal of Hydrology*, **502**, 77-88.

I conducted the experiments and analysis and wrote the original chapter with editing by Professor Stull.

Chapter 3

Contents of Chapter 3 (Improving Ensemble Forecasts of Reservoir Inflow by Sampling Uncertainty in the Modelling Chain) are adapted from the following published paper:

Bourdin, D. R. and R. B. Stull, 2013: Bias-corrected short-range Member-to-Member ensemble forecasts of reservoir inflow. *Journal of Hydrology*, **502**, 77-88.

I conducted the experiments and analysis and wrote the original chapter with editing by Professor Stull.

Chapter 4

Bourdin, D. R., T. N. Nipen and R. B. Stull, 2013: Reliable Probabilistic Forecasts from an Ensemble Reservoir Inflow Forecasting System. Submitted for professional peer review on 6 June 2013. Undergoing revisions.

Additions and modifications to COMPS schemes were implemented by Dr. Thomas Nipen and myself. I carried out the experiments and analysis and wrote the original chapter. Editing was done by Dr. Nipen and Professor Stull.

Chapter 5

Bourdin, D. R. and R. B. Stull, 2013: On the Importance of Sampling Hydrologic Uncertainty: An Economic Analysis. In preparation.

I conducted the experiments and analysis and wrote the original chapter, with editing by Professor Stull.

Appendix A

I wrote the descriptions of verification measures used throughout the dissertation. Editing was done by Professor Stull.

Appendix B

As for Chapter 4.

Appendix C

As for Chapter 4.

Table of contents

List of tables

List of figures

Acknowledgments

Firstly, my endless thanks go to my supervisor, Dr. Roland Stull for his support and guidance. His enthusiasm and encouragement have been inspiring throughout the years. I also wish to thank my supervisory committee members, Dr. Sean Fleming and Dr. Doug McCollor, for their valuable input and expertise.

I would like to thank all my colleagues, past and present, at the University of British Columbia (UBC), Canada, firstly, for their input and expert assistance, and secondly, for making it such an enjoyable place to be. Thanks go to Dr. Rosie Howard, May Wong, Dr. Greg West, Roland Schigas, Bruce Thomson, Dr. Atoossa Bakhshaii, Jesse Mason and Katelyn Wells. This research would not have been completed in such a timely manner if not for the tireless efforts of George Hicks II and Dr. Henryk Modzelewski in retrieving archived weather forecasts. Thank you to Dr. Thomas Nipen for creating the COMPS framework and making it available for my use, and for answering many questions about probability modelling and calibration.

I am grateful to BC Hydro, and in particular, Scott Weston, for providing hydrometric data for the Cheakamus watershed. Dr. Nicholas Kouwen and Dr. Jörg Schulla provided indispensable guidance during the setup and calibration of the WATFLOOD and WaSiM hydrologic models, respectively. Computer code for linking DDS with WaSiM was modified from that graciously provided by Dr. Thomas Graeff. Computational resources and weather forecasts were provided by the Geophysical Disaster Computational Fluid Dynamics Centre at UBC.

I am especially grateful to my parents, my best friend and sister, Jillian, and family and friends. Your unwavering support has made this possible.

Finally, the primary funding for this research was provided by the Canadian Natural Sciences and Engineering Research Council (NSERC) in the form of an Alexander Graham Bell Canada Graduate Scholarship (CGS-D). Additional funding was provided by the Department of Earth, Ocean and Atmospheric Sciences at the University of British Columbia and a NSERC Discovery Grant awarded to Dr. Roland Stull.

Dedicated to my loving parents, Ron and Karen Bourdin

Chapter 1

Introduction

1.1 Uncertainty in Hydrologic Model Predictions

Hydrologic models are simplified representations of a complex physical system — the terrestrial component of the hydrologic cycle. The applications of hydrologic models are many and varied, ranging from the assessment of the impacts of long-term climate or land use change to operational forecasting of streamflows for flood forecasting or hydroelectric reservoir operation. Predictions derived from hydrologic models carry with them some amount of uncertainty. This uncertainty comes not only from the simplification of hydrologic process representation, but also from errors in input data, incomplete knowledge of antecedent conditions, and uncertainty in model parameters.

Model process uncertainty arises from the simplified or incorrect representation of hydrologic processes or their omission entirely. Imperfect process representation may be caused by the necessary use of simplified functional relations between hydrologic elements or, alternatively, by our insufficient knowledge of the physics that govern these processes (Kitanidis and Bras, 1980; Niehoff et al., 2002). As data availability and computational speed have increased, the model representation of hydrologic processes has become more accurate (Kouwen et al., 2005). In some cases, however, process understanding is still so limited that we are forced into black-box approaches (Sivapalan et al., 2003).

The equations that govern hydrologic processes contain parameters having values derived from observation, professional experience, or model calibration. Parameter calibration is carried out through trial-and-error by adjusting model parameters until the model output is sufficiently close to observations. Unfortunately, a common impediment to the successful calibration of model parameters is the availability of observations with which to compare the various model outputs (e.g., Brun and Band, 2000; Eckhardt and Ulbrich, 2003). The closeness of fit is measured by an objective function, the choice of which can have an impact on the resulting optimum parameter set (Özelkan and Duckstein, 2001; Wagener, 2003). Different objective functions may be sensitive to different parts of the hydrograph; the choice of a single function will necessarily lead to a biased calibration (Duan et al., 2007; Götzinger and Bárdossy, 2008). The non-unique dependence of model

error upon parameter values is commonly known in the hydrological literature as equifinality. The existence of multiple equally plausible parameter sets may suggest that none accurately represent watershed characteristics — the right runoff can be obtained for the wrong reasons.

Hydrologic state describes the conditions within the modelled watershed at any given time (e.g., soil moisture, groundwater storage, snow-water equivalent, lake and stream levels, etc.). This state forms the initial conditions from which hydrologic forecasts are started, and is an important source of uncertainty in the modelling chain. This uncertainty is related to hydrologic model and parameter uncertainty, because model states are updated by the model itself using observed meteorological or hydrologic data. Errors in the measurement of this data can lead to errors in hydrologic state.

In a short-term operational forecast setting, hydrologic models are generally driven by weather model output. It has been reported that the uncertainty in numerical weather prediction (NWP) model output is the largest source of uncertainty in NWP-driven flow forecasts with a time horizon beyond several days, whereas for shorter lead-times, uncertainties in the hydrologic model dominate prediction errors (Coulibaly, 2003; Cloke and Pappenberger, 2009). However, the comparative importance of the two forms of error over the two time scales depends on context; for an anticipated heavy rainstorm in a small and rapidly responding catchment, uncertainty around the amount of rainfall expected over the next day may have more impact on forecast quality than hydrologic model error.

Note that distributed hydrologic models are often run with much higher spatial resolution than atmospheric models, requiring the downscaling of meteorological fields from NWP to hydrologic model scale. This introduces an additional source of uncertainty to the modelling chain. Particularly in complex terrain, the process of distributing meteorological data across the watershed should account for temperature lapse rates and the rate of increase of precipitation with elevation, both of which may vary seasonally or on shorter time scales (Alila and Beckers, 2001). Elevation dependence can be incorporated into downscaled NWP fields using regression techniques (e.g., Daly, 2006; Kurtzman and Kadmon, 1999) or inverse-distance weighting with constant linear lapse rate adjustments (e.g., Leemans and Cramer, 1991; Willmott and Matsura, 1995; Westrick and Mass, 2001).

1.2 Previous Related Work

1.2.1 Ensemble Hydrologic Modelling

Research into probabilistic weather forecasts began in the 1960s, building on Lorenz's (1963) work in chaos theory. By the 1980s, ensembles of forecasts based on multiple initial conditions (multi-

analysis ensembles) were being made in research mode. The first operational Ensemble Prediction System (EPS) was generated by the US National Centers for Environmental Prediction (NCEP) in 1992 (Sivillo et al., 1997). Ensemble forecasts from the Meteorological Service of Canada combine the multi-analysis and varied-model ensemble approaches (in which the same model is run with alternative physics schemes or parameterizations) to sample a wide range of predictive uncertainty (Environment Canada, 2013). Super-ensembles or grand-ensembles, derived from the combination of ensembles from each of several forecast centres, comprise a truly probabilistic approach, accounting for uncertainties in initial conditions, parameterizations, and model structure (Ross and Krishnamurti, 2005). A similar approach is needed for hydrologic applications to increase ensemble spread and capture the full range of predictive uncertainty (Krzysztofowicz, 2001).

The success of ensemble weather forecasting has led to its adoption in hydrology, primarily through the use of ensemble NWP output to drive a deterministic hydrologic model (Cloke and Pappenberger, 2009). That is, the same hydrologic model is re-run using different weather predictions to generate an ensemble of hydrographs. The first efforts in ensemble streamflow prediction (ESP) used an ensemble of meteorological observations from the climate record for long-term prediction (Day, 1985). ESP methods of this type are still routinely used for seasonal to annual water supply forecasting purposes (i.e., for forecast time scales at which NWP models do not provide skill). Operational weather forecast ensembles such as those distributed by the European Centre for Medium-Range Weather Forecasts (ECMWF) have been applied to flood forecasting in research mode (e.g., Gouweleeuw et al., 2005; Roulin and Vannitsem, 2005).

In NWP, ensembles are commonly generated by running a weather model with multiple sets of varying initial conditions. This follows naturally from the existence of deterministic chaos — a highly non-linear sensitivity to initial conditions — in the weather, and our inability to know the exact state of the atmosphere at any given time (Lorenz, 1963). The possible existence of chaos in hydrologic processes has been investigated with inconclusive results (Sivakumar, 2000; Sivakumar et al., 2001; Khan et al., 2005). Nevertheless, errors in initial conditions are recognized as an important source of uncertainty in hydrologic modelling (Liu and Gupta, 2007). Estimates of state uncertainty are commonly made using the Ensemble Kalman Filter (EnKF) (e.g., Evensen, 1994; Andreadis and Lettenmaier, 2006; Clark et al., 2008), which generates an ensemble of hydrologic states. This method has shown promise in assimilation of remotely sensed snow coverage and snow-water equivalent data in complex terrain (Andreadis and Lettenmaier, 2006). An EnKF variant, the bias-aware Retrospective Ensemble Kalman Filter (REnKF) has been shown to successfully update state variables using observations of discharge by accounting for associated time lags (Pauwels and De Lannoy, 2006). The particle filter (PF) is an alternative data assimilation method that is not subject to the limitations of EnKF such as the use of Gaussian distributions to model non-normally

distributed hydrological errors (Moradkhani et al., 2005a; Moradkhani and Sorooshian, 2009). The PF has been used for assimilation of remotely-sensed and in-situ snow water equivalent data and observed streamflow, and has been shown to produce state estimates and subsequent streamflow forecasts of higher quality than the EnKF method (e.g., Moradkhani et al., 2005a; DeChant and Moradkhani, 2011b; Leisenring and Moradkhani, 2011).

The existence of equally likely sets of parameter values has long been recognized (Binley et al., 1991) and has led to the development of probabilistic and stochastic methods for estimating parameter uncertainty. For example, the Shuffled Complex Evolution Metropolis algorithm (SCEM-UA; Vrugt et al., 2003b) and the Multi-Objective Shuffled Complex Evolution Metropolis algorithm (MOSCEM-UA; Vrugt et al., 2003a) converge to an ensemble of parameter sets that can be used to infer probabilistic uncertainty. The Simultaneous Optimization and Data Assimilation (SODA) method combines SCEM-UA with an EnKF to improve the treatment of input, output, parameter uncertainty, and structural uncertainty, resulting in "meaningful prediction uncertainty bounds" (Vrugt et al., 2005, pg. 2). However, these methods require knowledge of a prior distribution of parameter values, which may be difficult to define. In practice, the prior distribution is usually taken to be a noninformative (uniform) distribution, though in some algorithms this can lead to slow convergence to the posterior target distribution (e.g., Kuczera and Parent, 1998). Parameter estimation is affected by uncertainty in measured model input and output (e.g., rainfall and streamflow), and ignoring this uncertainty can lead to biased and misleading model results. The Bayesian total error analysis (BATEA) methodology developed by Kavetski et al. (2006a) requires hydrologic modellers to incorporate all application-specific sources of data uncertainty into the modelling process. The method is effective at identifying and correcting input errors when the user-supplied error models are valid (Kavetski et al., 2006b). However, models of input uncertainty are poorly understood, and the method is computationally demanding. A more simple method for quantifying parameter uncertainty consists of using multiple objective functions for creating multiple differently-optimized parameter sets (Duan et al., 2007).

As outlined above, data assimilation applications have primarily focussed on updating hydrologic model states. Recent research has turned to simultaneous estimation of model states and model parameters (Liu et al., 2011, and sources cited therein). Real-time updating of states and parameters allows the hydrologic model to more closely reproduce observed system response (Moradkhani and Sorooshian, 2009). Unlike "batch" parameter calibration techniques, which seek to minimize long-term prediction error over some historical period of calibration data, dual state-parameter estimation improves flexibility and allows for the investigation of temporal variability in model parameters (Moradkhani et al., 2005a). Such methods can also be applied where long historical datasets are unavailable for batch calibration. Both the EnKF and PF data assimilation methods have been ap-

plied to the dual state-parameter estimation problem (e.g., Moradkhani et al., 2005a,b; DeChant and Moradkhani, 2011b; Leisenring and Moradkhani, 2011). Liu et al. (2011) note that data assimilation methods in general have not been adequately implemented in operational settings due to a number of challenges including the availability of observed data, the specification of uncertainty in the data, and computational burden. Recent research has attempted to make the particle filter method more viable for operational prediction, but computational expense is still an issue (Moradkhani et al., 2012).

Equifinality refers not only to the existence of different parameter sets within a model structure that produce acceptable simulation results but also to the existence of many possible suitable model structures. Beven and Binley (1992) define model structure as including model processes and other considerations such as spatial discretization. Structural uncertainty is typically handled through the use of multiple hydrologic models. Shamseldin et al. (1997) appear to have been the first to apply the multi-model ensemble approach in rainfall-runoff modelling, using four empirical models and a simple lumped, conceptual model. Their results showed that, in general, better discharge predictions could be obtained through model combination. Others have also shown that multi-model ensemble mean hydrologic forecasts are able to outperform even the best single-model forecast within the ensemble (e.g., Coulibaly et al., 2005; Ajami et al., 2006).

In order to generate a truly probabilistic forecasting system, it is necessary to sample all sources of uncertainty in the modelling chain (Krzysztofowicz, 2001). To date, operational and research efforts into probabilistic streamflow forecasts through the use of ensembles have neglected some sources of uncertainty. For example, the US National Weather Service River Forecast System generates operational probabilistic water supply forecasts using the original ESP method of Day (1985) (Franz et al., 2003). Thus, the forecasts account only for uncertainty in meteorological inputs, and ignore non-stationarity. Georgakakos et al. (2004) used multiple calibrated and uncalibrated hydrologic models, some with many parameter sets to assess streamflow prediction uncertainty. Duan et al. (2007) used three hydrologic models, each calibrated using three different objective functions to derive a nine-member ensemble that assessed uncertainty arising from model structure and parameter uncertainty. Carpenter and Georgakakos (2006) used a Monte Carlo sampling framework to account for both parametric and radar-rainfall uncertainty. BC Hydro's Absynthe modelling procedure for daily inflow likewise incorporates ensemble weather forecasts and multiple parameter sets for a single hydrologic model (Fleming et al., 2010). Other examples of hydrologic ensembles that incompletely sample uncertainty include but are not limited to: Vrugt et al. (2005); Moradkhani et al. (2005b); Randrianasolo et al. (2010); Thirel et al. (2010); Van den Bergh and Roulin (2010), and De Roo et al. (2011).

1.2.2 Uncertainty Modelling

Ensemble forecasting techniques are designed to sample the range of uncertainty in forecasts, but are often found to be underdispersive in both weather and hydrologic forecasting applications (e.g., Eckel and Walters, 1998; Buizza, 1997; Wilson et al., 2007; Olsson and Lindström, 2008; Wood and Schaake, 2008). In order to correct these deficiencies, uncertainty models can be used to fit a probability distribution function (PDF) to the ensemble, whereby the parameters of the distribution are estimated based on statistical properties of the ensemble and past verifying observations. These theoretical distributions reduce the amount of data required to characterize the distribution (for example, from n ensemble members to two parameters describing the mean and spread of a Gaussian distribution), and allow estimation of probabilities for events outside of the range of observed or modelled behaviour (Wilks, 2006).

Uncertainty models make different assumptions about how the ensemble members and observations are generated. A common method for producing probability forecasts is the binned probability ensemble (BPE; Anderson, 1996). The assumption in this case is that the N ensemble members and the unknown verifying observation are drawn from the same unknown probability distribution. The observation then has an equally likely probability of $(N + 1)^{-1}$ of falling between any two consecutive ranked ensemble members, or outside of this range. Alternatively, centering a Gaussian probability distribution on the ensemble mean with spread proportional to the ensemble variance makes the assumption that the ensemble mean forecast errors are normally distributed (or, equivalently, that the verifying observations are drawn from a normal distribution centred at the ensemble mean). This model also assumes the existence of a spread-skill relationship. That is, the spread of the ensemble members should be related to the accuracy (or skill) of the ensemble mean; when the forecast is more certain, as indicated by low ensemble spread, errors are expected to be small. However, this relationship is often tenuous (e.g., Hamill and Colucci, 1998; Stensrud et al., 1999; Grimit and Mass, 2002). Bayesian Model Averaging (BMA) is an alternative uncertainty model that assigns probability distributions to the individual ensemble members and takes the forecast PDF to be the weighted sum of these distributions (Raftery et al., 2005). The weights indicate the likelihood of each distribution being the correct one, and are based on past performance of the individual ensemble members.

In cases where a forecast PDF is fitted to the ensemble, the shape of the PDF should correspond to the shape of the empirical distribution of the forecast errors. For a simple Gaussian distribution centred on the ensemble mean, the errors of the ensemble mean forecast are used. In the case of BMA, the individual distributions should match the shape of the corresponding ensemble member's forecast errors. Hydrologic variables and their errors are often described as being non-normally

distributed, and are therefore transformed into a space in which the errors become normally distributed, and the transformed variable can be modelled using the simple Gaussian PDF (e.g., Duan et al., 2007; Reggiani et al., 2009; Wang et al., 2009). The log-normal distribution, which amounts to fitting a simple Gaussian distribution to log-transformed data, has a long history of use in hydrology, and is still commonly applied today (e.g., Chow, 1954; Stedinger, 1980; Lewis et al., 2000; Steinschneider and Brown, 2011). This distribution is particularly well-suited to streamflow and inflow forecasting, as it assigns probabilities only to positive forecast values.

If the uncertainty model assumptions are valid, the resulting probability forecasts should be statistically reliable or *calibrated*, meaning that an event forecasted to occur with probability p will, over the course of many such forecasts, be observed a fraction p of the time (Murphy, 1973). Otherwise, the probabilistic forecasts cannot be used for risk-based decision making, since the probabilities cannot be taken at face value. Reliability is easily corrected using probability calibration methods, discussed next. Note that the probabilistic definition of calibration differs from that used in hydrologic modelling. In the latter field, calibration is the process of obtaining hydrologic model parameters tuned for a particular watershed (described in Section 1.1). Both probability calibration and hydrologic model calibration are addressed in this dissertation. Thus, to avoid ambiguity, hydrologic model calibration will be referred to as parameter optimization or simply optimization, and the term *calibration* will be used in the probabilistic sense.

1.2.3 Statistical Postprocessing

Meteorological and hydrologic forecasts contain both systematic and random errors. Systematic error, also known as (unconditional) bias, can arise due to differences between modelled and actual topography, and due to deficiencies in model representation of physical processes. The objective of bias correction is to reduce the systematic error of future forecasts by using statistical relationships between past forecasts and their verifying observations. Random error can be reduced through ensemble averaging, though the full ensemble contains valuable information regarding probabilities of possible future outcomes (Anderson, 1996).

In a multi-model ensemble context, ensemble members derived from different dynamical (NWP and/or hydrologic) models should be corrected by computing bias correction factors for the individual members. In an ensemble where multiple realizations of a single dynamical model are used, application of a single correction factor (e.g., the bias of the ensemble mean) to all members is appropriate. If bias correction is not done prior to multi-model combination, spread and other measures of ensemble performance can be artificially inflated due to the interaction of opposing model biases (Johnson and Swinbank, 2009; Candille et al., 2010). If component EPS biases do not balance, then their combination can result in a degradation of forecast accuracy (Wilson et al., 2007).

For this reason, many frameworks for generating reliable probabilistic forecasts through model combination begin with a bias correction step. A linear regression-type bias corrector is built into the BMA framework described by Raftery et al. (2005) to correct the individual meteorological ensemble members prior to combining them into a probabilistic weather forecast. Likewise, the more generalized approach of Johnson and Swinbank (2009) uses the bias of the ensemble mean forecast to correct the individual members from a single dynamical model. Vrugt and Robinson (2007) have suggested that the global regression-based correction of the original BMA framework is too simple to be useful in hydrology where model errors are non-Gaussian and heteroscedastic (i.e., proportional to flow level), and that local non-linear bias correction should be applied based on modelling errors in the immediate past.

Various methods of statistical calibration have been devised to correct for conditional or distributional biases in probabilistic forecasts. These can generally be split into two groups: ensemble calibration, which adjusts individual ensemble members in order to produce reliable forecasts; and probability calibration, which adjusts the probabilities directly. Examples of ensemble calibration include BMA (Raftery et al., 2005) and generalizations thereof (e.g., Johnson and Swinbank, 2009).

When the BPE is used to generate a probabilistic forecast, information contained in the rank histogram (Anderson, 1996; Talagrand et al., 1997) can be used for probability calibration (Hamill and Colucci, 1997). The probability mass between consecutive ensemble members is adjusted based on how often historical observations fell into that bin. This amounts to shifting the cumulative distribution function (CDF) at each ensemble member to the frequency of historical observations falling below that particular rank. This weighted ranks (WR) method has been found to produce more reliable and generally higher quality probabilistic quantitative precipitation forecasts than the raw ensembles or even model output statistics (MOS) (e.g., Eckel and Walters, 1998; Hamill and Colucci, 1998). The WR method can be generalized to calibrate probabilistic forecasts generated by other probability models, where reliability is assessed using a probability integral transform (PIT) histogram (Gneiting et al., 2005). In this case, the forecast CDF values are relabelled based on the distributions of past PIT values. Nipen and Stull (2011) have shown that this method can improve the reliability and other scores of probabilistic forecasts generated using BPE and even BMA.

In hydrologic EPSs, bias correction has focused on the removal of unconditional bias from long-term (e.g., monthly or seasonal) forecasts (Hashino et al., 2007; Wood and Schaake, 2008). Post-processing of short-term hydrologic EPSs has focused on the correction of conditional or distributional bias of probabilistic forecasts to generate reliable probabilities (Seo et al., 2006; Zhao et al., 2011). Bayesian methods have been applied successfully in hydrologic forecasting applications over a range of timescales (e.g., Duan et al., 2007; Reggiani et al., 2009; Wang et al., 2009; Parrish et al., 2012). Probability calibration on the other hand, has not yet been widely adopted

by the hydrologic modelling community. Olsson and Lindström (2008) provide an example of a very simple probability calibration used to improve ensemble spread. Roulin (2007) applied the weighted ranks method to medium-range forecasts of streamflow and found very little improvement to the already reliable forecasting system.

When weather model output is used to drive a hydrologic model, bias correction is often applied to the precipitation forecasts (Kouwen et al., 2005; Yoshitani et al., 2009; Westrick et al., 2002). The importance of post-processing the inputs to hydrologic models has been discussed in the literature (e.g., McCollor and Stull, 2008a; Yuan et al., 2008). Mascaro et al. (2010) have shown that well calibrated precipitation forecasts will indeed yield reliable probabilistic streamflow predictions despite the nonlinearities in the hydrologic model. They also found that underdispersive precipitation forecasts do not necessarily lead to underdispersive streamflows. Thirel et al. (2008) have shown that underdispersive precipitation forecasts can lead to even more highly underdispersive streamflow forecasts in both short- and medium-range applications. Mascaro et al. (2011) have demonstrated that dispersion in a streamflow forecast is highly dependent on the antecedent rainfall; when the watershed has more initial wetness, the streamflow forecast is increasingly controlled by the deterministic nature of a previous rainfall event. This indicates that the uncertainties in the hydrologic model and its initial conditions are extremely important, and thus further calibration of the downstream model output may be necessary. Olsson and Lindström (2008) have suggested that separate treatment of meteorological and hydrologic errors may be desirable from a scientific standpoint, but that operationally, only adjustment of the final hydrologic output is necessary.

1.3 Dissertation Case Study and Contributions

The ensemble and probabilistic forecasting methods developed in this dissertation are used to predict inflows to the Daisy Lake reservoir, a hydroelectric facility on the upper Cheakamus River in southwestern British Columbia (BC), Canada. This reservoir is operated by the British Columbia Hydro and Power Authority (BC Hydro). The total area of the Cheakamus watershed upstream of the reservoir is 721 km^2, approximately 8% of which is glaciated. Elevation within the study basin ranges from 341 m to 2677 m above sea level with a median elevation of 1401 m. Inflows to the Daisy Lake reservoir are primarily driven by snowmelt during spring and summer with a small glacial melt component. A secondary inflow peak occurs during the fall and winter storm season when Pacific frontal systems can bring significant inflows, particularly in the case of rain-on-snow events that can be difficult to predict. The watershed responds rapidly to such events, generating inflow time series with steep rising and falling limbs; watersheds with this type of response are commonly referred to as *flashy*.

This watershed was selected because it presents various modelling challenges. NWP forecasts are complicated by the region's complex terrain, which can lead to strong orographic gradients in precipitation fields (which can in turn be highly dependent on storm track). Other challenges include cold air damming episodes and difficulties in forecasting temperature profiles and therefore precipitation phasing. High-resolution NWP models may be able to capture these processes, as they are able to represent complex topography more accurately than models using coarse grid scales. In order to take direct advantage of high-resolution NWP fields to drive the inflow forecasts, distributed hydrologic models are required (as opposed to lumped, conceptual or empirical modelling approaches) (Bourdin et al., 2012). For application to the case study watershed, these hydrologic models must be capable of modelling snow and glacier melt processes and lakes in complex terrain given relatively limited input data.

The main goal of this dissertation is to generate reliable probabilistic forecasts of inflow for a hydroelectric reservoir in complex terrain. This will be achieved by: sampling all sources of error in the inflow modelling chain, thereby creating an ensemble of inflow forecasts; and by applying statistical post-processing techniques including simple bias correction, uncertainty models, and probability calibration. The dissertation components are described presently.

1.3.1 Sampling Uncertainty in Inflow Forecasts

The primary contribution of this dissertation is the creation of a Member-to-Member (M2M) ensemble forecasting system that explicitly attempts to sample all sources of error in the hydrologic modelling chain.

In Chapter 2, a M2M forecasting system is generated by using individual members of a multi-model, multi-grid scale NWP ensemble to drive an ensemble of distributed hydrologic models. The NWP fields are downscaled using multiple interpolation schemes, thereby generating multiple meteorological forcings from a single NWP output. This ensemble therefore explicitly samples uncertainty in the NWP forecasts and the processes used to downscale them to the hydrologic model scale, and the uncertainty in the hydrologic model structures.

This M2M ensemble is expanded in Chapter 3 with the addition of multiple hydrologic model parameterizations and multiple hydrologic states or initial conditions, which are used to begin each daily forecast during the case study period. The multi-parameter M2M component is created by optimizing the parameters of the two hydrologic models using three different objective functions, thereby taking advantage of equifinality. The multi-state component is generated by updating the hydrologic state in the watershed each day using meteorological observations to drive each hydrologic model with each parameterization.

The full M2M ensemble developed in Chapter 3 consists of 72 ensemble members, each of

which represents a different possible scenario of inflow to the Daisy Lake reservoir. It is believed that this ensemble is the first example of a short-term hydrologic forecasting system that attempts to explicitly sample all sources of error in the inflow modelling chain, thereby comprising a truly probabilistic forecasting system as defined by Krzysztofowicz (2001).

1.3.2 Bias Correction

Prior to combining the M2M ensemble members into probabilistic or ensemble mean forecasts, a bias correction factor is applied to each individual ensemble member. In a multi-model ensemble context, ensemble members derived from different dynamical (numerical weather prediction and/or hydrologic) models should be corrected by computing bias correction factors for the individual members.

In Chapter 2, a simple bias correction scheme is developed that makes use of the degree of mass balance (DMB). The DMB is simply the ratio of past forecast inflows to past observed inflows, calculated over a moving window. A bias-corrected forecast is generated by dividing the raw forecast by the DMB factor. The use of a multiplicative bias corrector ensures that bias-corrected inflow forecasts never become negative. Unlike regression-based bias correction schemes, which can require lengthy training periods, the method employed in this study is able to handle the heteroscedastic nature of hydrologic forecast errors (Vrugt and Robinson, 2007).

1.3.3 Calibrated Probability Forecasts

In Chapter 4, the 72-member M2M ensemble forecasting system is transformed into a reliable probabilistic forecast by using suitable uncertainty models and applying probability calibration when necessary.

The uncertainty models are based on the Ensemble Model Output Statistics (EMOS) method of Gneiting et al. (2005). EMOS fits a probability distribution function to the ensemble, whereby the parameters describing the spread of the distribution are estimated based on statistical properties of the ensemble and the verifying observations. In this way, it is possible to implicitly account for any uncertainty that is neglected or simply underestimated by the M2M ensemble.

An intelligent calibration strategy is employed to correct for calibration deficiencies during periods when the uncertainty model produces unreliable forecasts. Calibration is done using the PIT-based method of Nipen and Stull (2011), which relabels forecast probabilities based on the distribution of past PIT values accumulated over a training window. This is the first application of the method to hydrologic forecasting.

1.3.4 Economic Analysis

Since the sum total of the costs associated with the full M2M ensemble (e.g., time spent setting up a hydrologic model or price paid for high-resolution NWP fields) may be prohibitive for operational forecasting applications, it is prudent to evaluate the economic value of each M2M component. If the price paid for each component is known, such an analysis can be used to determine whether or not they are cost-effective. Murphy (1993) identified three types of forecast "goodness": consistency (i.e., between a forecaster's best judgement and the actual forecast), quality, and value. Value, which is concerned with economic worth to the forecast end user, is the focus of Chapter 5.

In order to determine the economic value of each component of the M2M ensemble (multiple NWP models and grid scales, multiple distributed hydrologic models, multiple model parameterizations and multiple hydrologic states), a simple cost-loss decision model is developed for the Daisy Lake reservoir based on the work of Richardson (2000) and McCollor and Stull (2008b). By comparing the economic value of the full M2M ensemble probability forecasts to that achieved by M2M configurations with individual ensemble components removed, it is possible to estimate the value added by the individual components.

Chapter 2

Bias-Corrected Short-Range Member-to-Member Ensemble Forecasts of Reservoir Inflow

2.1 Introduction

Since the first efforts in ensemble streamflow prediction over two decades ago (Day, 1985), ensemble hydrologic forecasting has grown immensely in scope. Hydrologic ensemble prediction systems (EPSs) have been developed that include multiple weather inputs (Gouweleeuw et al., 2005; Roulin and Vannitsem, 2005), multiple assimilated hydrologic states or initial conditions (Pauwels and De Lannoy, 2006; Clark et al., 2008), multiple parameter sets (Vrugt et al., 2003a,b) and multiple hydrologic models (Shamseldin et al., 1997; Ajami et al., 2006). Such ensembles attempt to sample the range of uncertainty in hydrologic prediction that is caused by errors in these components of the modelling chain (Bourdin et al., 2012).

Meteorological and hydrologic forecasts contain both systematic and random errors. Systematic error, also known as (unconditional) bias, can arise due to differences between modelled and actual topography, deficiencies in model representation of physical processes, and errors in model parameterization. The objective of bias correction is to reduce the systematic error of future forecasts by using statistical relationships between past forecasts and their verifying observations. Random error can be reduced through ensemble averaging, though the full ensemble contains valuable information regarding probabilities of possible future outcomes (Anderson, 1996).

In a multi-model ensemble, members derived from different dynamical (numerical weather prediction and/or hydrologic) models should be individually bias-corrected prior to their combination. In an ensemble where multiple realizations of a single dynamical model are used, a single correction factor (e.g., the bias of the ensemble mean) should be applied to all members. If bias correction is not done prior to multi-model combination, spread and other measures of ensemble performance

can be artificially inflated due to the interaction of opposing model biases (Johnson and Swinbank, 2009; Candille et al., 2010). If component EPS biases do not balance, then their combination can result in a degradation of forecast accuracy (Wilson et al., 2007).

For this reason, many frameworks for generating reliable probabilistic forecasts through model combination begin with a bias correction step. For example, a linear regression-type bias corrector is built into the Bayesian Model Averaging (BMA) framework described by Raftery et al. (2005) to correct the individual meteorological ensemble members prior to combining them into a probabilistic weather forecast. Likewise, the more generalized approach of Johnson and Swinbank (2009) uses the bias of the ensemble mean forecast to correct the individual members from a single dynamical model. Vrugt and Robinson (2007) suggested that the global regression-based correction of the original BMA framework is too simple to be useful in hydrology where model errors are non-Gaussian and heteroscedastic (i.e., proportional to flow level), and that local non-linear bias correction should be applied based on errors in the immediate past. In some applications of BMA, additive bias correction schemes have been used in place of global correction (e.g., Schmeits and Kok, 2010). Parrish et al. (2012) have suggested the combined use of data assimilation methods with BMA as a way of dealing with non-Gaussian error and other shortcomings of the method.

In hydrologic EPSs, bias correction has focused on the removal of unconditional bias from long-term (e.g., monthly or seasonal) forecasts (Hashino et al., 2007; Wood and Schaake, 2008). Post-processing of short-term hydrologic EPSs has focused on the correction of conditional or distributional bias of probabilistic forecasts to generate reliable probabilities (Seo et al., 2006; Zhao et al., 2011). Madadgar et al. (2012) recently developed an ensemble post-processing method suitable for seasonal hydrologic forecasting that is able to remove both unconditional and conditional bias while additionally improving other aspects of ensemble quality. When weather model output is used to drive a hydrologic model, bias correction is often applied to the precipitation forecasts (Kouwen et al., 2005; Yoshitani et al., 2009; Westrick et al., 2002). The importance of post-processing the inputs to hydrologic models has been discussed in the literature (e.g., McCollor and Stull, 2008a; Yuan et al., 2008). However, Mascaro et al. (2011) have demonstrated that dispersion in streamflow ensemble forecasts is highly dependent on hydrologic state, suggesting that further correction of the end forecast is likely to be required. Indeed, Olsson and Lindström (2008) have suggested that while separate treatment of meteorological and hydrologic errors may be desirable from a scientific standpoint, from an operational point of view, only adjustment of the final hydrologic forecast is strictly necessary.

In this study, we develop a Member-to-Member (M2M) ensemble inflow forecasting system that incorporates multiple weather models driving multiple hydrology models for a total of 24 ensemble members. That is, individual NWP forecast ensemble members are used to drive the individual

members of the DH ensemble — hence the term 'Member-to-Member'. We then assess the quality of both probabilistic and deterministic ensemble mean forecasts before and after applying different bias correction schemes to individual ensemble traces. Evaluation is based on daily inflow forecasts for a hydroelectric reservoir in the complex terrain of southwestern British Columbia, Canada.

2.2 Case Study Area and Data

The M2M ensemble is used to forecast inflows to the Daisy Lake reservoir, a hydroelectric facility on the upper Cheakamus River in southwestern British Columbia (BC), Canada, operated by the British Columbia Hydro and Power Authority (BC Hydro) (Figure 2.1). The total basin area upstream of the reservoir is 721 km^2, approximately 8% of which is glaciated. Elevation within the study basin ranges from 341 m to 2677 m above sea level with a median elevation of 1401 m. Hindcasts from the M2M system are tested over the 2009–2010 water year, which was characterized by El Niño conditions during the winter months that weakened throughout the spring and shifted into a La Niña state by late summer. In southwestern BC, it is well documented that El Niño episodes generally bring warmer, drier weather, while La Niña episodes are characterized by cooler and wetter than normal conditions (e.g., Mantua et al., 1997; Dettinger et al., 1998; Fleming et al., 2007; Fleming and Whitfield, 2010). The El Niño episode that occurred during the case-study water year was relatively wet for the region of interest and winter snow accumulation at low elevations was below normal due to above-average temperatures. Note that for this particular hydroclimatic regime, a water year is defined as the period from October 1 of the starting year through September 30 of the following year.

Inflows to the Daisy Lake reservoir are primarily driven by snowmelt during spring and summer with a small glacial melt component. A secondary inflow peak occurs during the fall and winter storm season when Pacific frontal systems can bring significant inflows, particularly in the case of rain-on-snow events that can be difficult to predict. Daily average inflow rates are calculated by BC Hydro using a water balance based on observed reservoir levels and outflows. The calculated inflows used in this study have undergone quality control and are considered to be of high quality. For the purposes of this study, these values will be referred to as observed inflows. The Water Survey of Canada (WSC) collects streamflow data for the Cheakamus River above Millar Creek (CHK) location (Figure 2.1). This data source was used in various stages of watershed model parameter optimization, but is not used to verify forecasts made by the M2M inflow forecasting system.

The Cheakamus basin upstream of Daisy Lake has limited coverage with respect to meteorological observations, especially at high elevation. BC Hydro operates three data collection platforms (DCP) within the watershed. These are located at the Daisy Lake Dam (CMS), on the Cheakamus

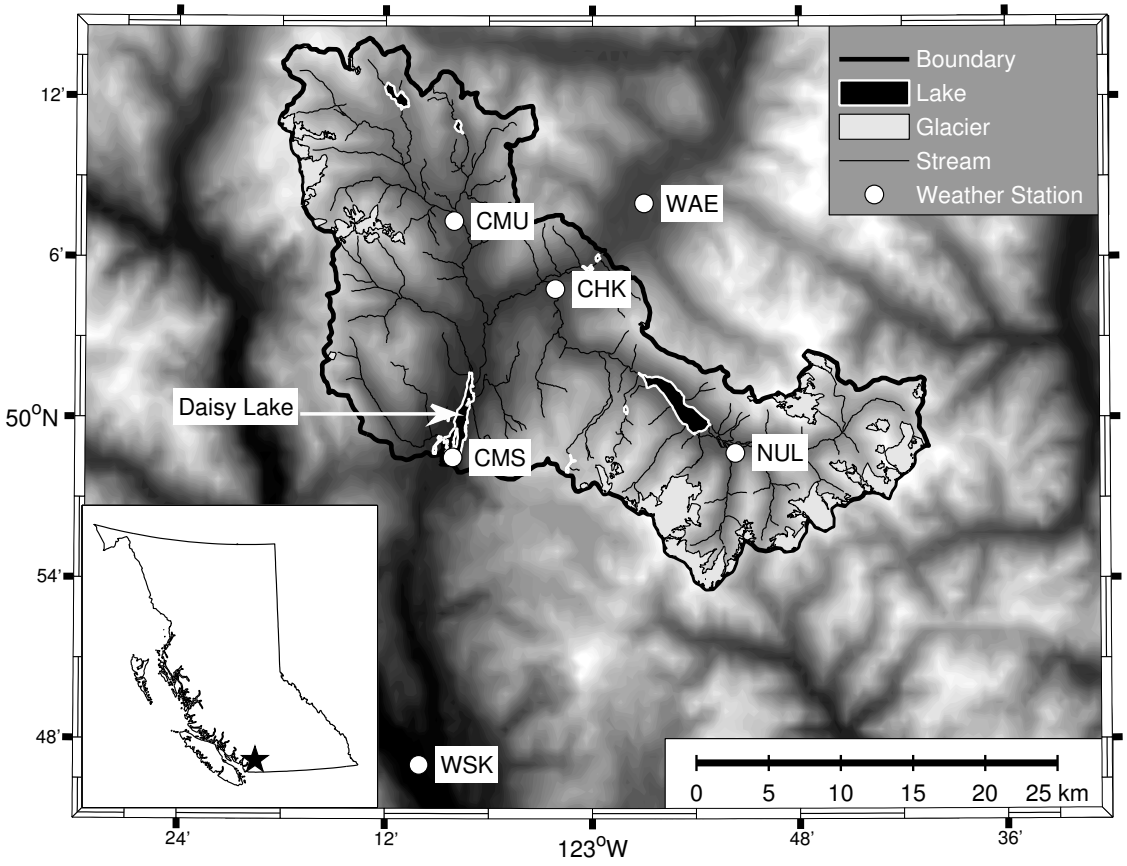

Figure 2.1: Map of the Cheakamus basin above the Daisy Lake reservoir, located in southwestern BC. ASTER global digital elevation model background map is a product of the Japanese Ministry of Economy, Trade and Industry (METI) and the National Aeronautics and Space Administration (NASA), with higher elevations represented by lighter shades of grey.

River above Millar Creek (CHK) and the Upper Cheakamus site (CMU). Additional coverage is provided by observing stations at the Whistler (WAE) and Squamish Airport (WSK) stations operated by Environment Canada (EC). All weather station locations and identifiers are shown in Figure 2.1. These observing platforms range in elevation from 52 m above sea level at WSK to 880 m at CMU.

Daily maximum and minimum temperatures and 24-hour accumulated precipitation observations are available for all stations. WAE and WSK additionally provide observations of wind speed and humidity. During the model optimization and validation periods (1986 – 1997), missing humidity and wind speed observations were filled in using linear regression when one of WSK or WAE were available, and using monthly climatological values when both were missing. BC Hy-

dro DCP data undergoes stringent quality control and has a complete record of observations during this period. An imaginary weather station (designated as "NUL" in Figure 2.1) has been placed in the eastern portion of the watershed to improve meteorological data coverage. Data from the CMU site were copied to the NUL site, which was carefully selected based on elevation, aspect, surrounding terrain, and comparison of PRISM (Parameter-elevation Regressions on Independent Slopes Model) 1961–1990 climate normal data at the two sites, which was downscaled to 400 m resolution by ClimateBC (PRISM Climate Group, 2012; Wang et al., 2006). We expect that without NUL, the inverse-distance downscaling of meteorological observations to hydrologic model grid scale (discussed in Section 2.3.3) would be less accurate in the eastern half of the watershed. Since the nearest real weather stations are located in regions of lower elevation with different aspect, precipitation measurements at these sites will likely do a poor job of characterizing rainfall events in the mountainous terrain of eastern Cheakamus, even following adjustments for elevation.

Continuous snow pillow observations are available from the WSC for all of the parameter optimization period and six out of ten years during the validation period for the Squamish Upper site, located outside of the western boundary of the Cheakamus watershed at an elevation of 1340 m. A proxy site was selected for verification of simulated snow water equivalent (SWE) at a location just inside the western watershed boundary. Site selection was again based on elevation, aspect, terrain, and a comparison PRISM-ClimateBC data at the real and proxy locations.

2.3 A Member-to-Member (M2M) Ensemble Forecasting System

The M2M ensemble inflow forecasting system developed and applied in this study incorporates multiple Numerical Weather Prediction (NWP) models, which are downscaled using multiple interpolation schemes, and finally used to drive multiple Distributed Hydrologic (DH) models. That is, individual members of the NWP ensemble drive individual members of the hydrologic ensemble. A description of each of these components follows.

2.3.1 Numerical Weather Prediction Models

The NWP models are taken from the operational ensemble suite run by the Geophysical Disaster Computational Fluid Dynamics Centre (GDCFDC), in the Department of Earth, Ocean and Atmospheric Sciences at the University of British Columbia. The ensemble consists of three independent nested limited-area high-resolution mesoscale models with forecast domains centered over southwestern BC.

The Mesoscale Compressible Community (MC2) model is a fully compressible, semi-implicit, semi-Lagrangian, non-hydrostatic mesoscale model (Benoit et al., 1997). The fifth-generation Penn-

sylvania State University-National Center for Atmospheric Research Mesoscale Model (MM5) is a fully compressible, non-hydrostatic model designed for mesoscale and regional-scale atmospheric simulation (Grell et al., 1994). Version 3 of the Weather Research and Forecasting (WRF) mesoscale model is also fully compressible and non-hydrostatic and has been developed as a community model (Skamarock et al., 2008).

The coarse resolution (108-km horizontal grid spacing) outer nests of these three NWP models are initialized using the National Centers for Environmental Prediction (NCEP) North American Mesoscale (NAM) model, which also provides time-varying boundary conditions. All three NWP models produce forecast output at horizontal grid spacings of 36, 12, 4 and 1.3 km. The finer grids, which have smaller model domains due to computational time constraints, are nested inside of the coarse grids from which they receive their time-varying boundary conditions. Due to the relatively small size of the case-study watershed, only NWP output from the three finest grids are used to drive the DH models. The NWP models are initialized at 00UTC and run out to 60 hours (the 1.3-km MC2 model runs for only 39 hours due to operational time constraints). NWP model forecast hours beginning from 00PST (08UTC) are used to drive the DH models. A multi-model, multi-grid-scale ensemble of weather forecasts consisting of at least six of the total nine members was issued every day throughout the study period except for a five-day interval (April 9–13, 2010) in which only WRF members were available.

2.3.2 Distributed Hydrologic Models

In order to take advantage of the high-resolution distributed NWP output available, two physically-oriented, distributed hydrologic models have been selected for use based on their suitability to the case-study watershed. Specifically, the models must be able to simulate snowmelt and glacier melt processes and lakes in complex terrain given relatively limited input data. The DH models selected for this study are the Water balance Simulation Model (WaSiM; Schulla, 2012) and WATFLOOD (Kouwen, 2010).

WaSiM is fully distributed and uses physically based algorithms for most process descriptions. Algorithms of varying complexity may be selected by the model developer based on data constraints and knowledge of processes operating in the study watershed. For the current application, potential evapotranspiration (PET) is based on the Penman-Monteith equation (Monteith, 1965), the infiltration model is based on the Green and Ampt approach (Green and Ampt, 1911; Peschke, 1987), and soil water modelling and runoff are based on the TOPMODEL approach of Beven and Kirkby (1979). For sub-daily time steps, snowmelt can be modelled using a simple temperature index algorithm (Anderson, 1973) or a temperature-wind index approach in which melt rate is proportional to wind speed (Schulla, 2012). Because of poor coverage (both spatial and temporal) for wind speed

observations in the Cheakamus basin, and because of the importance of snowmelt contributions to Daisy Lake inflows, we use the temperature index algorithm. The model was run using a 1 km grid spacing and an hourly time step. WaSiM uses gridded NWP output of hourly precipitation, temperature, wind speed, humidity and global radiation (the latter three variables being used by the PET module). For the duration of this study, global radiation (total direct and diffuse solar radiation at the ground surface) fields are only available from the MM5 models. Therefore WaSiM model forecasts from all NWP models incorporate MM5 global radiation fields of corresponding NWP model grid scale.

The WATFLOOD model similarly incorporates mainly physically based process descriptions, but operates in a semi-distributed nature using the Grouped Response Unit (GRU; Kite and Kouwen, 1992.) The GRU approach lumps together Hydrologic Response Units (HRU), which are areas of similar land cover residing within one model grid square. Hydrologic processes are modelled identically for each group of HRU, and the responses of each group are weighted and summed to generate a total GRU outflow (Kouwen et al., 1993). This allows WATFLOOD to preserve sub-grid scale hydrologic variability (for example, that described by a high-resolution digital elevation model), while computing flows at a grid scale selected based on availability of meteorological inputs or the desired level of output detail. WATFLOOD uses the Hargreaves equation to estimate PET (Hargreaves and Samani, 1982). Infiltration is modelled by the Philip formula (Philip, 1954), which is identical to the Green and Ampt approach except that it includes the effects of surface ponding. Snowmelt is modelled using the temperature index algorithm (Anderson, 1973). The NWP model output fields utilized by WATFLOOD are hourly precipitation and temperature.

Parameters of both DH models were optimized on observations of inflows at CMS and streamflows at CHK, using inputs of observed meteorological quantities from the DCP and EC stations (Figure 2.1) to drive model simulations for a period of ten water years (October 1997 – September 2007). To run the models at an hourly time step, daily minimum (TMIN) and maximum (TMAX) temperatures were transformed into hourly temperatures using a sine curve connecting TMIN at 0400 PST to TMAX at 1600 PST. Daily total precipitation was disaggregated for WaSiM by dividing the daily total into 24 equal hourly amounts. WATFLOOD incorporates a built-in disaggregation method whereby 1 mm of precipitation accumulates each hour until the daily total is met and equal hourly amounts are used if the daily total is greater than 24 mm. Global radiation inputs for the WaSiM model were calculated for each DCP and EC station location using equations from Stull (2000) with adjustment for atmospheric conditions based on Spokas and Forcella (2006).

Parameter optimization consisted of a multi-stage process beginning with manual tuning of a parameter set previously used for a similar application. Then a series of automated optimizations were run using the Dynamically Dimensioned Search (DDS) algorithm (Tolson and Shoemaker,

2007; Graeff et al., 2012; Francke, 2012). First, each DH model was auto-optimized with 500 DDS runs to tune parameters expected to impact high flows in the basin (e.g., rain/snow partitioning and snowmelt parameters) using the Nash-Sutcliffe Efficiency (NSE; Nash and Sutcliffe, 1970) of simulated flow as an objective function. Continuing from the resulting parameter set, an additional 500 DDS runs were done using the NSE of log-transformed flows to tune parameters affecting low flow periods (e.g., soil parameters). Finally, three separate trials of 1000 DDS runs each were executed to optimize all previously-tuned parameters on the MAE of simulated CMS inflows only. While parameter optimization should in theory seek to optimize all characteristics of the hydrologic regime, this study is concerned only with generating high-quality inflow forecasts. Simulated CHK streamflow performance statistics still improved for both hydrologic models during the final optimization stage (with the exception of WATFLOOD mean absolute error, which increased by 1.6%).

The best of the three final-stage parameter optimization trials was selected based on performance during an independent validation period of ten water years spanning October 1986 through September 1996. Optimization and validation results for the best trial are shown in Table 2.1. Observed and simulated SWE at the proxy site for the Squamish Upper snow pillow were in good agreement during the optimization and validation periods, with coefficient of determination (R^2) values of 0.91 (optimization) and 0.77 (validation) for WaSiM. Values for WATFLOOD were 0.87 and 0.83 respectively. The timing of annual peak and zero-SWE conditions was also acceptable for both models. Snow pillow observations were not used during optimization due to extremely limited coverage, and because the use of a proxy watershed location introduces uncertainty. Also, because there is only one SWE measurement site, optimization of WATFLOOD SWE would be limited to the particular land class in which the proxy site was located, which is somewhat arbitrary. SWE was excluded from WaSiM optimization in order to create a level playing field for the two models.

Table 2.1: Performance of simulated inflows from the WaSiM and WATFLOOD hydrologic models during optimization (1997–2007) and validation (1986–1996) periods. Measures of model performance are described in Section 2.5 and Appendix A.

Performance Measure	WaSiM		WATFLOOD	
	Optimization	Validation	Optimization	Validation
NSE	0.79	0.72	0.75	0.79
LNSE	0.73	0.75	0.79	0.81
R^2	0.79	0.73	0.77	0.79
DMB	1.01	0.92	1.07	0.97
MAE (m^3/s)	12.9	15.3	12.9	13.8
RMSE (m^3/s)	19.8	24.8	21.4	21.6

These optimized models were then used to make ex-post-facto NWP-driven forecasts (or hindcasts) for the 2009–2010 water year to enable an independent verification in a setting similar to real-time operational forecasts. To begin, WaSiM and WATFLOOD were spun up from uniform, snow-free initial conditions for the period of September 1–30, 2009 using observed meteorological data to drive the models. The simulated hydrologic state for each model was saved at the end of this period to be used as an initial condition for the first NWP-driven M2M forecast run on October 1, 2009. Each day of the study period, observed meteorological data are used to drive the hydrologic models to update the model states, producing initial conditions for the day's forecasts. This updating is done to ensure that large hydrologic state errors do not accumulate due to poor NWP forecasts (Westrick et al., 2002). Observation-driven simulated inflows are created as a by-product of the state-updating process for WaSiM and WATFLOOD. Figure 2.2 illustrates this process of generating updated hydrologic states, simulated inflows (driven by observed meteorological data), and forecasted inflows (driven by NWP forecasts) for an individual DH model. Grey arrows represent input used to drive the model from a particular hydrologic state. Black arrows represent model runs initialized from this state. Solid lines show the flow of meteorological observations into the model and the resulting model runs that produce simulated inflows and updated hydrologic states for the following day. Dashed lines show the flow of NWP forecast fields into the model and the resulting 2-day inflow forecasts. The flow of time is indicated by the dash-dotted line along the top of the figure.

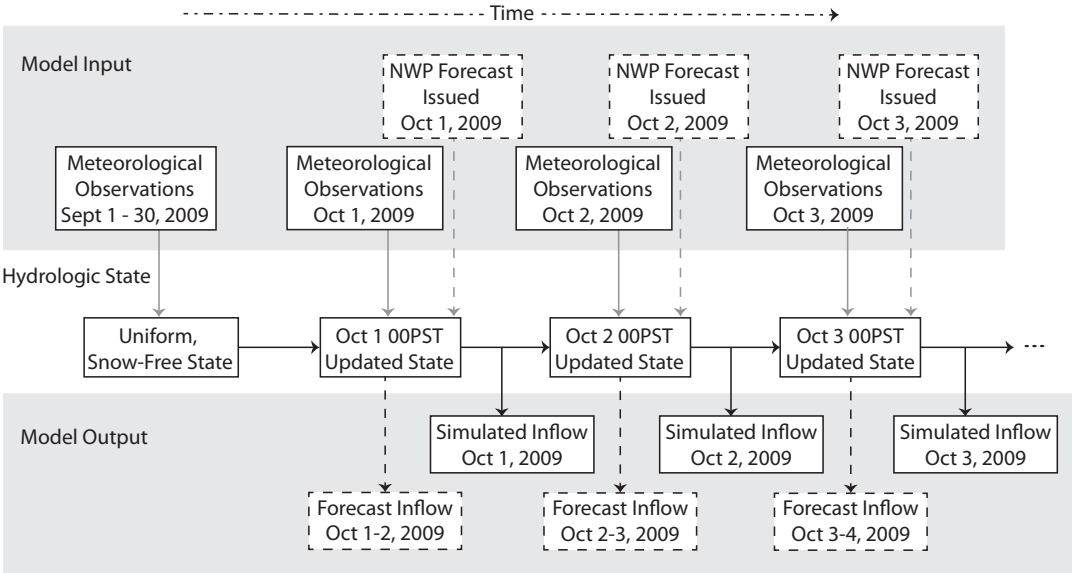

Figure 2.2: Flowchart illustrating the process of generating updated hydrologic states, simulated inflows, and forecasted inflows for a particular hydrologic model.

2.3.3 Downscaling of Meteorological Input

Each DH model incorporates built-in methods for downscaling weather station data or gridded NWP forecast fields to the DH model grid scale. The downscaling or interpolation process introduces uncertainty into the forecasting chain, particularly for low-resolution meteorological inputs. Therefore, an ensemble of downscaling methods has been incorporated into the M2M ensemble in an attempt to account for such errors.

WaSiM has several downscaling methods available for use, a few of which have been applied here: bilinear interpolation, Inverse-Distance Weighting (IDW) where the weight parameter is set to two (i.e., distance-squared), altitude-dependent regression (REGR), and weighted combinations of IDW and REGR. DCP and EC station observations of temperature, wind speed, humidity and global radiation were downscaled using REGR, while precipitation was downscaled using a combination of 25% IDW (with a search radius of 30 km) and 75% REGR. These methods were selected based on the characteristics of each meteorological variable (Klok et al., 2001).

NWP output fields at the 12 km grid scale are downscaled to the WaSiM grid using two different methods (where the same method is used for all meteorological variables): IDW with a search radius of 12 km; and REGR. Outputs from the 4 and 1.3 km grids are downscaled using the bilinear interpolation. Other methods were initially included in the M2M ensemble but were found to yield results too similar to those listed above.

WATFLOOD similarly offers an inverse-distance weighting interpolation with a default weighting of distance-squared. This built-in scheme offers an option to incorporate elevation-dependence using a constant elevation adjustment rate (EAR); additional smoothing parameters can be specified. For downscaling of station observations, EARs for temperature and precipitation were respectively set to 3 °C/km (decreasing with height) and -0.3 mm/km (increasing with height). The IDW search radius was set to 20 km, and fields were smoothed over a distance of 5km. These EAR values were selected based on examination of PRISM-ClimateBC fields surrounding DCP stations within the watershed. The temperature EAR is slightly less than slope-air lapse rates measured during clear nights and saturated frontal conditions at Whistler Mountain by Erven (2012).

NWP grids are downscaled using a search radius equal to their grid spacing. Smoothing is applied to the 12-km fields, and two different sets of EARs are used: those used in downscaling meteorological observations for model optimization and state updating, and a measured clear-day temperature slope-air lapse rate of 8 °C/km (Erven, 2012). The 4-km and 1.3-km NWP grids are downscaled without elevation adjustment or smoothing.

In summary, the full M2M ensemble consists of 24 different combinations of NWP models, downscaling procedures and hydrologic models. WaSiM is driven by 12 different sets of NWP

inputs: 12-km NWP model output from MC2, MM5 and WRF downscaled using IDW and REGR, 4-km outputs downscaled using bilinear interpolation, and 1.3-km outputs also downscaled using the bilinear algorithm. There are 12 additional members from the WATFLOOD model: 12-km NWP output from the three NWP models downscaled using IDW with two different temperature lapse rates, 4-km outputs downscaled using no elevation adjustment, and 1.3-km outputs likewise downscaled without elevation dependence. To the best of our knowledge, this is the first example of a short-range ensemble inflow or streamflow forecasting system to incorporate multiple NWP models, multiple downscaling schemes, and multiple hydrologic models.

2.4 A Simple Bias Correction Method

An appropriate measure of bias for volumetric quantities such as precipitation and reservoir inflow is the degree of mass balance (DMB; McCollor and Stull, 2008a). The DMB is a measure of the ratio of simulated or forecasted inflow to the observed inflow over a given period of time and is given by:

$$DMB_N = \frac{\sum_{k=1}^{N} f_k}{\sum_{k=1}^{N} o_k}, \tag{2.1}$$

where DMB_N is the degree of mass balance over an interval of N days and f_k and o_k are the forecasted and observed inflows, respectively, for the kth day prior to the current day. A DMB of one indicates a forecast or simulation that is free of volumetric bias.

A bias-corrected inflow forecast is calculated by:

$$F_{BC} = \frac{F_{Raw}}{DMB_N}, \tag{2.2}$$

where F_{BC} is today's bias-corrected daily inflow forecast, F_{Raw} is today's raw (uncorrected) daily inflow forecast, and DMB_N is the correction factor applied to the raw forecast. Forecast days 1 and 2 are treated separately (i.e., the day 1 forecasts are corrected using a DMB of the day 1 forecasts valid over the past N days, while the day 2 forecasts are corrected using the DMB of the day 2 forecasts valid over the past N days). The use of a multiplicative bias correction factor ensures that corrected inflow forecasts do not become negative.

In order to allow more recent forecast errors to have a bigger impact on the bias correction, an additional bias correction scheme is applied in which the DMB correction factor is a linearly-weighted average of the previous errors. This linearly-weighted DMB, calculated over an interval

of N days (denoted $LDMB_N$) is given by:

$$LDMB_N = \sum_{k=1}^{N} w_k \frac{f_k}{o_k}, \tag{2.3}$$

where k, f_k and o_k are as previously defined and w_k is the weight applied to the error for day k. The weight, given by:

$$w_k = \frac{N - k + 1}{\sum_{i=1}^{N} i} \tag{2.4}$$

is normalized such that the sum of applied weights is equal to one. The $LDMB_N$ correction factor is applied to the forecast using Eq. (2.2) and replacing DMB_N with $LDMB_N$.

This bias corrector handles only the unconditional forecast bias, or the difference between the central locations of the forecasts and observations. Conditional bias, also known as distributional bias or reliability (Appendix A) can be corrected using probability calibration methods (e.g., Hamill and Colucci, 1997; Seo et al., 2006; Zhao et al., 2011; Nipen and Stull, 2011; Madadgar et al., 2012). It will be shown that correction of unconditional bias improves forecast resolution, or the ability of the forecasting system to *a priori* differentiate future weather outcomes such that different forecasts are associated with distinct verifying observations. This is an important aspect of ensemble forecast quality that cannot be improved by conditional bias correction via probability calibration (Toth et al., 2003).

The DMB and LDMB bias correction schemes described above are applied to each inflow ensemble member separately. Recall that the purpose of bias correction is to correct for systematic errors in the dynamic NWP and DH models. Since each ensemble member is derived from a different NWP model driving a different DH model, individual member bias correction is appropriate in this context. Moving windows of lengths N equal to 3, 7, 15, 30, 45 and 60 days are applied to the M2M ensemble members and compared in Section 2.6. If there are missing forecasts or observations during the past N days, the N most recent days with available forecast-observation pairs are used instead. Thus, even very short training windows are not overly sensitive to missing data.

2.5 Verification Approach

Forecasted hourly inflow rates from each M2M ensemble member were averaged over each forecast day for verification against daily observed inflow rates. Measures-oriented verification statistics are calculated for the M2M ensemble mean at each forecast horizon. Such measures of forecast quality include the DMB as a measure of forecast bias (a DMB of one indicating no bias), and the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) as measures of accuracy

(with perfect forecasts having MAE and RMSE of zero). Forecast bias and accuracy can also be determined by a visual assessment of inflow hydrographs. M2M skill is measured relative to a zero-skill persistence forecast using the RMSE Skill Score (RMSESS). Statistical association is measured by the Nash-Sutcliffe Efficiency (NSE) and the NSE of Log-transformed flows (LNSE), which emphasizes forecast quality during periods of low flow.

While measures-oriented verification scores are useful for evaluating the quality of the ensemble mean forecast or individual ensemble members, the true value of an ensemble is best described using distributions-oriented measures (Murphy and Winkler, 1987). Here, we employ the rank histogram, Brier Skill Score (BSS), and the Relative Operating Characteristic (ROC). A description of all measures- and distributions-oriented verification measures is provided in Appendix A.

Both the Brier scores and ROC are calculated for forecast and observation anomaly thresholds relative to climatological inflow values. In order to ensure that the ensemble is not unduly rewarded for making high inflow forecasts during the snowmelt period where little skill is required to do so, we subtract climatology from the forecasts and observations. This daily climatology is derived from the median of observations on each calendar day over the period 1986–2008. A 15-day running mean is then used to generate a smoothed climatology. BSS, its decomposition, and ROC curves will be calculated for anomaly thresholds having inflow rates of -5.0, 2.7, and 19.5 m^3/s. These correspond to the quartiles of 2009–2010 observed inflow anomalies.

When comparing bias correction windows of different lengths, the verification periods include only days where all methods had enough prior forecast-observation pairs for calculation of DMB_N or $LDMB_N$. This ensures that shorter moving window corrections that are available earlier in the water year are not penalized (rewarded) for difficult (easy) forecast cases during this period.

2.6 Results and Discussion

The raw ensemble traces for each ensemble member forecast are shown for the entire study period in Figure 2.3. The consistency in forecast bias among WATFLOOD ensemble members and among WaSiM ensemble members indicates bias in the simulations used to generate their initial conditions. Periods of strong positive (negative) M2M forecast bias are consistent with periods during which the daily simulated inflows exhibit positive (negative) bias relative to observed inflows.

This failure to accurately simulate the watershed state may be due to incorrect distribution of meteorological observations during the winter El Niño and summer La Niña episodes. Errors in the wintertime simulations for both models are largely consistent with errors in winter simulations during the optimization and validation periods where El Niño conditions prevailed; the snowmelt-related errors are likewise consistent with those seen during years with La Niña summers.

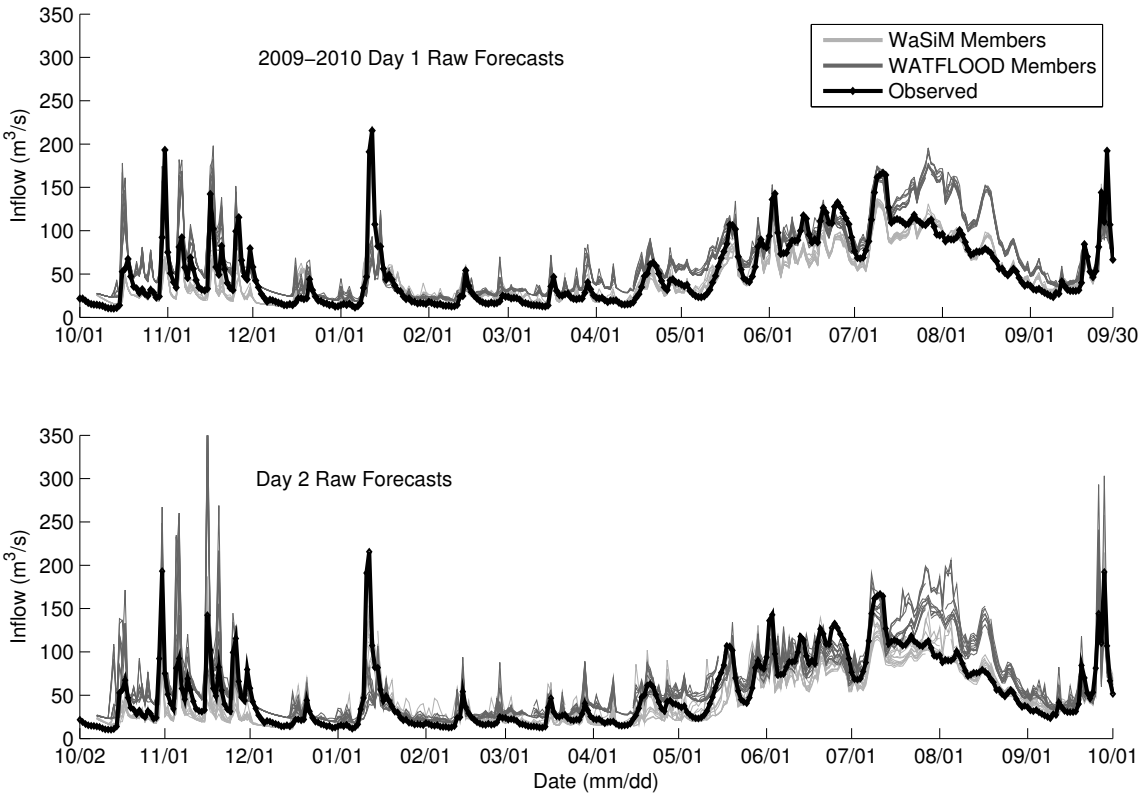

Figure 2.3: Raw ensemble traces for day 1 (top) and day 2 (bottom) forecasts during the 2009–2010 water year. Traces from the individual hydrologic models exhibit consistent bias, indicating a failure to accurately simulate the hydrologic state within the watershed.

For example, the WATFLOOD model (as set up for this particular study) tends to simulate erroneously high inflows during El Niño winters, and to be late in simulating snowmelt during La Niña summers. The winter errors could be due to distributed temperatures being too warm at high elevations where observations are not available. This would result in wintertime precipitation too-often falling as rain, and in an underestimation of the high-elevation snow pack. This hypothesis is supported by snowmelt-driven flows being undersimulated during spring/summer following several of these El Niño winters. Conversely, high-elevation distributed temperatures may be too cold during the La Niña spring and summer. The use of alternative downscaling techniques or temperature EARs for different climate indices may improve these simulations; the investigation of such alternatives is beyond the scope of this work. Data assimilation methods that update hydrologic state using observed SWE have shown promise for seasonal forecasting (DeChant and Moradkhani, 2011a), but may perform poorly for the Cheakamus basin due to the paucity of representative SWE data.

The DMB and LDMB bias correction methods result in dramatic improvements in M2M ensemble mean forecast quality, with best results for a 3-day moving window (Figure 2.4). For both forecast horizons and all window lengths, the LDMB correction offers improvement over the equally-weighted DMB correction. Moving windows of 45 and 60 days were found to produce bias-corrected ensemble mean forecasts that were worse than the raw output for some performance metrics, and are therefore not shown (raw forecast scores are indicated by the horizontal lines in Figure 2.4). The relatively good DMB of the raw ensemble mean forecasts is likely a result of performing model combination prior to bias correction of the individual ensemble members. That is, a balance is achieved by combining the WATFLOOD ensemble members, which are generally too wet (with DMB values ≈ 1.2 for days 1 and 2), with the WaSiM members, which have a dry bias (DMB ≈ 0.9 for both days).

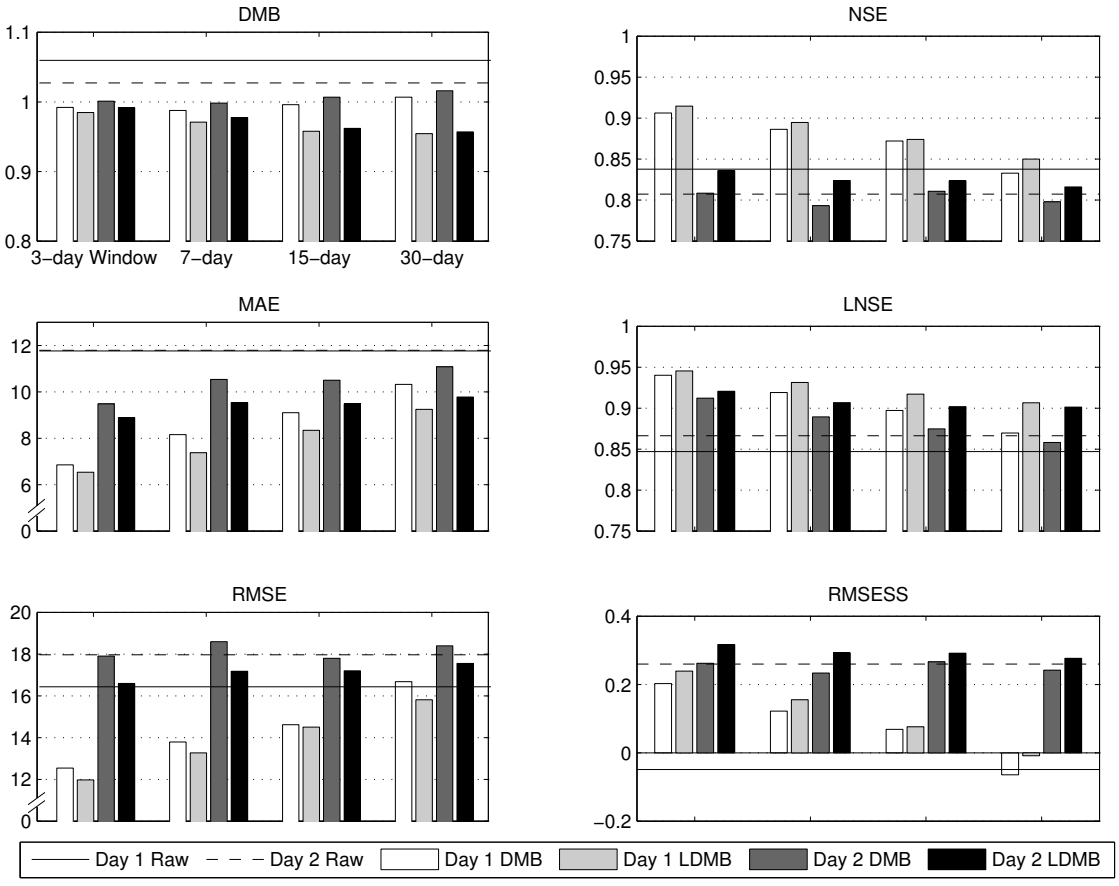

Figure 2.4: Results of applying bias correction schemes with varying window lengths to day 1 and day 2 forecasts as measured by ensemble mean verification metrics. Perfect forecasts have DMB, NSE, LNSE and RMSESS of one, and MAE and RMSE of zero.

The bias in the hydrologic state used to start each NWP-driven forecast was found to be the primary contributor to forecast bias. By correcting the individual M2M traces using the bias of the corresponding simulated inflows calculated over the same window lengths, similar improvements to those shown in Figure 2.4 were found for most performance metrics. The importance of this bias source is likely the reason that short correction windows perform so well; since the Cheakamus watershed is mountainous and flashy in nature, only recent forecast errors are likely to play an important role in bias correction for short-term (1–2 day) forecasts.

The $LDMB_3$ bias-corrected ensemble traces (Figure 2.5) no longer exhibit strong bias. They do, however, show an erroneous forecast spike on January 14–15, 2010 that is not as pronounced in Figure 2.3. An examination of NWP ensemble mean forecasts and observations at the CMU weather station reveals that this is caused by a combination of NWP failure and subsequent bias correction. On January 11 and 12, raw inflow forecasts from all models were too low likely because NWP forecasts were colder and drier than observations, leading to snow accumulation rather than a rain-on-snow inflow event. The raw inflow forecast on January 15 is slightly larger than observed because the NWP forecasts were too warm and wet. This forecast failure, coupled with the large DMB correction resulting from the January 11–12 forecast failure, leads to the false alarm issued by the $LDMB_3$-corrected forecast on January 14–15.

The absence of strong bias in the $LDMB_3$ forecasts is also evident in the bias-corrected rank histograms in Figure 2.6. The raw forecast rank histograms exhibit an overall L shape, indicating an over-forecasting bias. The peak in the middle of the raw histograms is due to the fact that the WATFLOOD and WaSiM ensemble members are tightly clustered and have opposing biases. Thus, observations are most likely to fall outside of the range of these ensembles, or somewhere between the clusters. Following bias correction, the L shape is far less pronounced. Both the raw and bias-corrected ensembles are underdispersive; bias correction causes a slight reduction in dispersion.

ROC diagrams (Figure 2.7) for the day 1 raw and $LDMB_3$ bias-corrected ensembles indicate that the bias-corrected ensemble is better able to discriminate between the occurrence and non-occurrence of inflow events of various magnitudes. The DMB_3 bias-corrected ensemble performs similarly to the $LDMB_3$ corrected ensemble, with slightly less area under each curve.

Figure 2.8 shows the BSS, relative reliability and relative resolution of raw and bias-corrected forecasts for the 19.5 m^3/s (75th percentile) inflow anomaly threshold. $LDMB_3$ is better than DMB_3 for day 1 in terms of all three metrics. For day 2 forecasts, however, DMB_3 performs better than $LDMB_3$ as measured by the BSS. The decomposition shows that this is because of a deterioration in $LDMB_3$ reliability, which can easily be corrected by further post-processing. Day 2 resolution of $LDMB_3$ forecasts remains superior to the DMB_3, and it is in this attribute that we find the intrinsic value of the forecasting system. These results point to the importance of separately handling uncon-

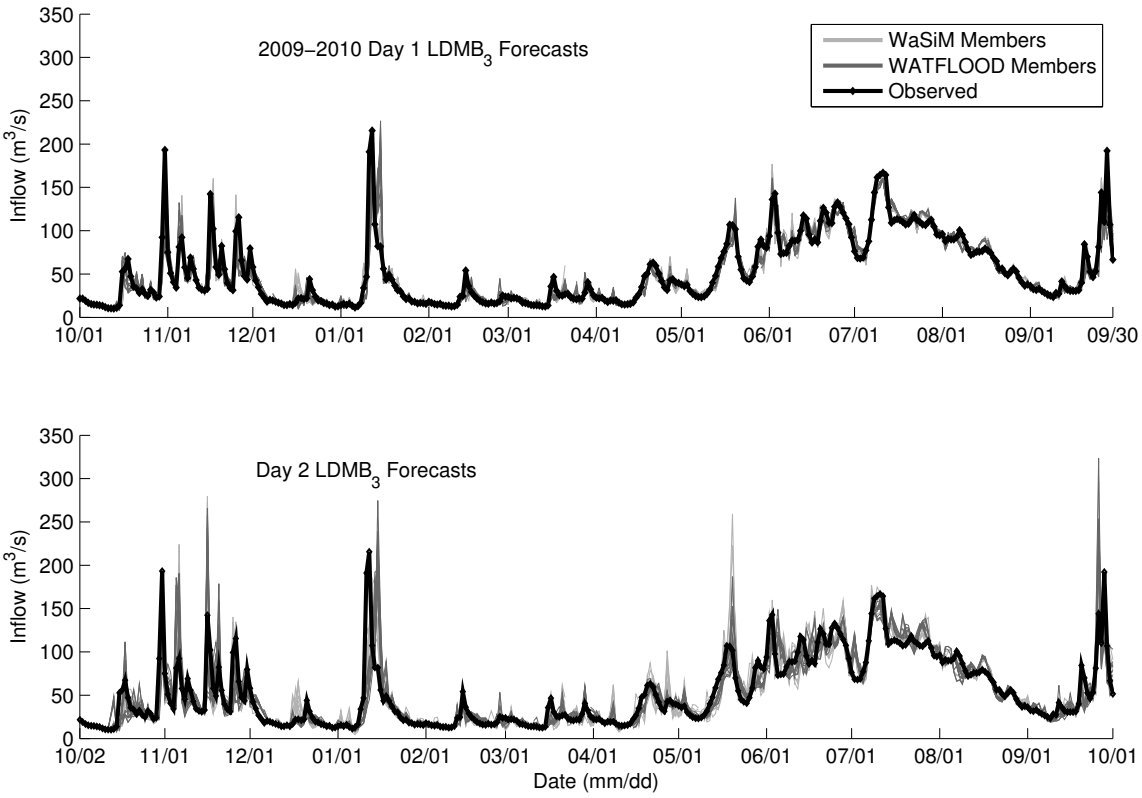

Figure 2.5: Ensemble traces for day 1 (top) and day 2 (bottom) forecasts during the 2009–2010 water year following LDMB_3 bias correction.

ditional and conditional (distributional) bias or reliability. Reliability can be improved by removing conditional bias, whereas resolution can only be corrected by improving the forecasting "engine" used to generate the ensemble, for example, through unconditional bias correction.

2.7 Conclusions

Two different bias correction schemes, each trained using windows of varying lengths, have been applied to a 24-member ensemble inflow forecasting system developed for the Daisy Lake reservoir in southwestern British Columbia, Canada. Both bias correction schemes use the degree of mass balance between past inflow forecasts and observations to correct future forecasts. Based on examination of a suite of measures- and distributions-oriented verification metrics, we determined that a linearly-weighted combination of past DMB errors (with more recent errors being weighted more heavily) performs slightly better than an equally-weighted combination, with both methods

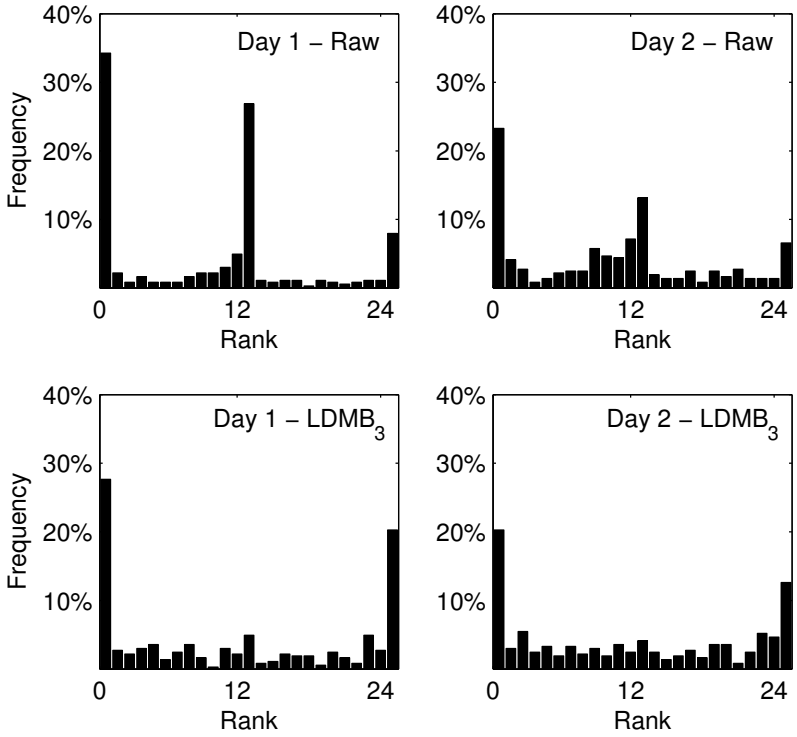

Figure 2.6: Rank histograms for day 1 and day 2 raw and LDMB_3 bias-corrected ensemble forecasts.

producing forecasts that are superior to the raw forecasts. The best improvement was obtained for a 3-day bias correction training window, likely due to the importance of hydrologic state bias and the flashy, mountainous nature of the case study watershed.

The bias correction schemes used in this study are simple and easily implemented in an operational setting. They can be applied to any watershed, but the ideal training window is likely to be basin-dependent. For example, larger basins with slow response times may have better results with longer training periods. There is very little overhead involved in calculating the bias correction factors, which require only $(N + forecast\ length)$ days of past forecast-observation pairs. This presents an advantage over regression-based correction methods, which can require years of data for training. Additionally, unlike global regression, the DMB and LDMB methods described herein can handle the heteroscedasticity of errors in hydrologic forecasts (Vrugt and Robinson, 2007).

The NWP models used to drive the WATFLOOD and WaSiM hydrologic models in this study had forecast horizons of two days. At longer forecast horizons, longer training windows may be necessary in order to balance the goals of minimizing both short-memory hydrologic state bias and

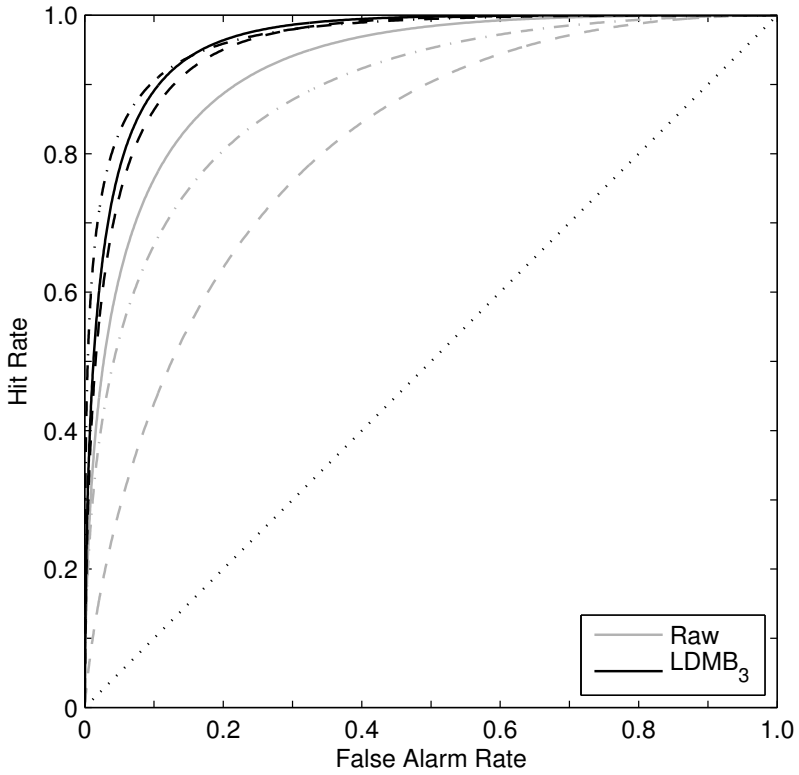

Figure 2.7: ROC curves for day 1 ensemble forecasts for forecasted inflow anomalies greater than -5.0 m^3/s (dot-dashed line), 2.7 m^3/s (dashed line) and 19.5 m^3/s (solid line). The dotted line is the zero-skill line.

NWP errors that may require longer learning periods (McCollor and Stull, 2008a). Using knowledge of observation-driven simulated inflow bias, it is possible to separate the bias in the M2M ensemble forecasts into that caused by bias in initial conditions, and that caused by the interaction of the NWP and DH models (i.e., total forecast bias is the product of the DMB in the hydrologic state and of the interacting models). This is an area for potential future study.

To date, operational and research applications of ensemble forecasting have dealt with only one or two error sources at a time (e.g., Vrugt et al., 2005; Moradkhani et al., 2005b; Randrianasolo et al., 2010; Thirel et al., 2010; Van den Bergh and Roulin, 2010; De Roo et al., 2011) and are therefore underdispersive, failing to sample the full range of possible hydrologic outcomes. The M2M system presented in this chapter is likewise underdispersive. This will be handled by adding to the M2M system a multi-state component and multi-parameter component. This upgraded M2M ensemble will thereby attempt to explicitly account for all sources of uncertainty in the modelling

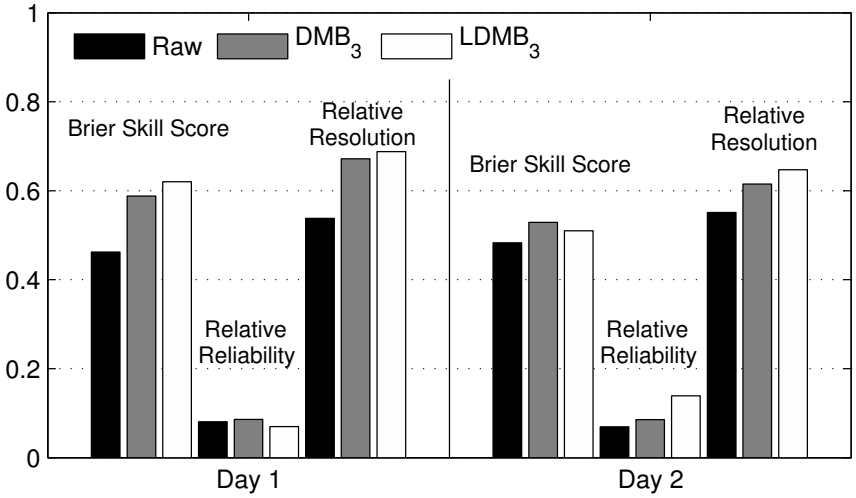

Figure 2.8: Brier skill score (BSS = 1 is perfect), relative reliability (zero is perfect) and relative resolution (one is perfect) for raw and bias-corrected forecasts for days one and two. The inflow anomaly threshold evaluated here is 19.5 m^3/s.

chain (Bourdin et al., 2012) and constitute a truly probabilistic forecast as defined by Krzysztofowicz (2001). It is anticipated that further probabilistic calibration (e.g., Nipen and Stull, 2011) will be needed in order to improve the statistical reliability of the ensemble.

Chapter 3

Improving Ensemble Forecasts of Reservoir Inflow by Sampling Uncertainty in the Modelling Chain

3.1 Introduction

As simplified representations of complex processes, hydrologic models and their predictions are subject to uncertainty. Making an ensemble of multiple forecasts is a way to sample the range of this uncertainty. Since Lorenz's (1963) work in chaos theory, the quantification of uncertainties in both initial conditions and model processes has become common practice in ensemble weather forecasting. The use of multi-model, multi-analysis super-ensembles represents a truly probabilistic approach, accounting for uncertainties in initial conditions, parameterizations, and model structure (Ross and Krishnamurti, 2005). A similar approach is needed for hydrologic applications to increase ensemble spread and capture the full range of predictive uncertainty.

In order to generate a truly probabilistic forecasting system, it is necessary to sample all sources of uncertainty in the modelling chain (Krzysztofowicz, 2001). In the case of reservoir inflow forecasting, errors are incorporated into the forecasting system by way of the hydrologic models themselves, their parameterizations, and the initial and boundary conditions (i.e., meteorological data) used to drive the models. To date, operational and research efforts into probabilistic streamflow forecasts through the use of ensembles have neglected some sources of uncertainty. For example, the US National Weather Service River Forecast System generates operational probabilistic water supply forecasts using the original Ensemble Streamflow Prediction (ESP) method of Day (1985), whereby streamflow simulation is driven by historical sets of temperature and precipitation data (Franz et al., 2003). Thus, the ESP forecasts account only for uncertainty in meteorological inputs, ignoring non-stationarity. Georgakakos et al. (2004) used multiple hydrologic models, some with many parameter sets to assess streamflow prediction uncertainty. Duan et al. (2007) used three hy-

drologic models, each optimized using three different objective functions to derive a nine-member ensemble that assessed uncertainty arising from model structure and parameter error. Carpenter and Georgakakos (2006) used a Monte Carlo sampling framework to account for both parametric and radar-rainfall uncertainty. The Absynthe modelling procedure used by the BC Hydro and Power Authority (BC Hydro) for daily inflow forecasts likewise incorporates ensemble weather forecasts and multiple parameter sets for a single hydrologic model (Fleming et al., 2010). Other examples of hydrologic ensembles that incompletely sample uncertainty include but are not limited to: Vrugt et al. (2005); Moradkhani et al. (2005b); Randrianasolo et al. (2010); Thirel et al. (2010); Van den Bergh and Roulin (2010), and De Roo et al. (2011).

In this chapter, we present an ensemble reservoir inflow forecasting system that samples forecast uncertainty arising from all sources of error in the modelling chain. The ensemble consists of multiple Numerical Weather Prediction (NWP) model output grids downscaled using multiple interpolation schemes and subsequently used to drive multiple Distributed Hydrologic (DH) models. Each of these DH models makes use of multiple differently-optimized parameter sets and begins each day's forecast from a set of different initial conditions or hydrologic states. This ensemble thereby comprises a truly probabilistic forecasting system as defined by Krzysztofowicz (2001).

To the best of our knowledge, this Member-to-Member (M2M) ensemble is currently the only example of a short-term hydrologic forecasting system that explicitly attempts to sample all sources of model error. In a previous study (Chapter 2), an ensemble inflow forecasting system consisting of a multi-NWP, multi-DH ensemble with multiple downscaling schemes was evaluated. The focus of this chapter is on evaluating the impact of adding the multi-parameter and multi-state components to this ensemble.

3.2 Case Study Area and Data

The hydrologic ensemble forecasting system developed in this study is used to forecast inflows to the Daisy Lake reservoir, a hydroelectric facility on the upper Cheakamus River in southwestern British Columbia (BC), Canada. The reservoir is operated by BC Hydro. Evaluation of the ensemble is carried out over the 2009–2010 water year, defined as the period from 1 October, 2009 to 30 September, 2010. Fall and winter storm season inflows are primarily driven by precipitation from Pacific frontal systems. Rain-on-snow events can cause significant inflows during this period. During spring and summer, inflows are snowmelt-driven, with some late-season glacier melt contributions.

Daily average inflow rates are calculated by BC Hydro using a water balance based on observed reservoir levels and outflows. The calculated inflows employed in this study have undergone quality

control and are considered to be of high quality. For the purposes of this study, these values will be referred to as observed inflows. Hourly forecast inflows to the Daisy Lake reservoir are transformed into daily average inflow rates for verification against these observations.

Evaluation of the ensemble forecasting system is based on a suite of measures- and distributions-oriented verification metrics. Measures-oriented scores are used to evaluate the performance of the ensemble mean, and include the Degree of Mass Balance (DMB) as a measure of forecast bias, the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) as measures of accuracy, and the RMSE Skill Score (RMSESS), which measures forecast skill relative to a zero-skill reference forecast, taken here to be persistence. Statistical association of ensemble mean forecasts is evaluated using the Nash-Sutcliffe Efficiency (NSE; Nash and Sutcliffe, 1970), which emphasizes high-flow periods, and the NSE of Log-transformed flows (LNSE) to evaluate low-flow performance. The full ensemble is evaluated using distributions-oriented measures including the Brier Skill Score (BSS) and its decomposition into relative reliability and relative resolution, the ensemble rank histogram, and the Relative Operating Characteristics (ROC) diagram. Detailed descriptions of these verification scores are given in Appendix A.

3.3 A Member-to-Member (M2M) Ensemble Forecasting System

The Member-to-Member (M2M) ensemble forecasting system used for forecasting inflows to the Daisy Lake reservoir explicitly samples uncertainty arising from errors in the Numerical Weather Prediction (NWP) fields used to drive the Distributed Hydrologic (DH) models, the hydrologic models themselves and their parameterizations, and the hydrologic states or initial conditions used to begin each daily forecast run. The result is an ensemble of 72 unique daily inflow forecasts. A description of each of the M2M ensemble components follows.

3.3.1 A Multi-NWP Ensemble

The NWP models are from the operational ensemble run by the Geophysical Disaster Computational Fluid Dynamics Centre (GDCFDC) in the Department of Earth, Ocean and Atmospheric Sciences at the University of British Columbia. The ensemble includes three independent nested limited-area high-resolution mesoscale models with forecast domains centred over southwestern BC.

The Mesoscale Compressible Community (MC2) model is a fully compressible, semi-implicit, semi-Lagrangian, non-hydrostatic mesoscale model (Benoit et al., 1997). The fifth-generation Pennsylvania State University-National Center for Atmospheric Research Mesoscale Model (MM5) is a fully compressible, non-hydrostatic model designed for mesoscale and regional-scale atmospheric simulation (Grell et al., 1994). Version 3 of the Weather Research and Forecasting (WRF) mesoscale

model is also fully compressible and non-hydrostatic and has been developed as a community model (Skamarock et al., 2008).

The coarse resolution (108 km horizontal grid spacing) outer nests of these three NWP models are initialized using the National Centers for Environmental Prediction (NCEP) North American Mesoscale (NAM) model, which also provides time-varying boundary conditions. All three NWP models produce forecast output at horizontal grid spacings of 36, 12, 4 and 1.3 km. The finer grids (which have smaller model domains due to computational constraints) are nested inside of the coarse grids from which they receive their time-varying boundary conditions. Due to the small size of the case-study watershed, NWP output from only the three finest grids are used to drive the DH models. The NWP models are initialized at 00UTC and run out to 60 hours (the 1.3-km MC2 model runs for only 39 hours due to operational time constraints). NWP model forecasts beginning from 00PST (08UTC) are used to drive the DH models from a particular set of initial conditions, or a hydrologic state. The creation of initial conditions for the daily inflow forecasts is described in Section 3.3.4.

3.3.2 A Multi-Hydrologic Model Ensemble

The DH models applied to the case-study watershed are the Water balance Simulation Model (WaSiM; Schulla, 2012) and WATFLOOD (Kouwen, 2010). These models were selected because they are distributed, and therefore able to take direct advantage of high-resolution NWP input. They are also able to simulate snowmelt and glacier melt processes and lakes in complex terrain given relatively limited input data. These features are critical for modelling in the case-study watershed. The optimization of model parameters for each DH model is described in Section 3.3.3.

Both DH models are run at 1 km grid spacing at an hourly time step. The NWP fields are downscaled to the DH model grid using interpolation schemes built into each DH model. For the WaSiM model, 12-km NWP fields are downscaled using two methods: inverse-distance weighting (IDW); and elevation-dependent regression (Schulla, 2012). The 4-km and 1.3-km NWP fields are downscaled using a bilinear interpolation scheme. WATFLOOD downscaling is done using IDW that incorporates elevation dependence using an optional elevation adjustment rate for both temperature and precipitation. 12-km fields are downscaled using IDW with two different elevation adjustments, while the 4- and 1.3-km fields do not use the elevation adjustment.

3.3.3 A Multi-Parameter Hydrologic Ensemble

Parameters of both DH models were optimized on observed inflows using meteorological data from weather stations located within the case-study watershed and surrounding area to drive model simulations for a period of ten water years (1997–2007). Parameter optimization consisted of a multi-

stage process beginning with manual tuning of a parameter set previously used for a similar application. Then, to generate different parameter sets, a series of automated optimizations were run using the Dynamically Dimensioned Search (DDS) algorithm (Tolson and Shoemaker, 2007; Graeff et al., 2012; Francke, 2012) with different objective functions: the mean absolute error (MAE) of simulated inflow, to minimize overall errors; Nash-Sutcliffe Efficiency (NSE; Nash and Sutcliffe, 1970) of inflow, to emphasize performance during high-flow events; and the NSE of log-transformed flows (LNSE), to optimize during low-flow periods. This methodology is consistent with that of Duan et al. (2007), who likewise used objective functions favouring different parts of the hydrograph to optimize multiple hydrologic models. The different model parameterizations attempt to explicitly sample uncertainty in the parameter values. The parameter sets optimized using NSE and LNSE of inflows are referred to as the NSE_o and $LNSE_o$ parameter sets.

The third parameter set, designated MAE_o, was generated using a four-step procedure to produce a 'best' model parameterization for each hydrologic model. Following the manual tuning step, DDS was applied to optimize parameters expected to impact high flows in the basin (e.g., rain/snow partitioning and snowmelt parameters) using the NSE of simulated flow as an objective function. Continuing from the resulting parameter set, an additional DDS optimization was carried out using the NSE of log-transformed flows to optimize parameters affecting low flow periods (e.g., soil parameters). These optimizations were based on performance of simulated inflows to Daisy Lake and streamflows at an upstream location (Cheakamus Upper), which are collected by the Water Survey of Canada (WSC). Finally, three separate DDS trials were executed to optimize all tuning parameters on the MAE of simulated inflows only. The best of the three trials was selected based on performance during an independent validation period of ten water years (1986–1996). In this chapter, the full ensemble consisting of forecasts from all three model parameterizations is compared to an ensemble comprised of the single best (MAE_o) parameterization for each DH model. This MAE_o-only configuration is identical to the multi-NWP, multi-DH ensemble evaluated in Chapter 2.

The set of model parameters selected for optimization in each DH model is the same in all three optimized model parameterizations. These are parameters related to watershed soil properties and the accumulation and melt of snow and glaciers. WATFLOOD allows different land-use types to be assigned different values of parameters (Kite and Kouwen, 1992; Kouwen et al., 1993). WATFLOOD soil parameters were optimized for four different land classes comprising the majority of the watershed area: barren/alpine, old forest, young forest, and logged (Figure 3.1). Model parameters impacting snow and glacier processes were also optimized for the glacier land class. Parameters corresponding to the other land classes in WATFLOOD (wetlands, water and impervious surfaces) were excluded from optimization because they have well-defined values or because they contribute relatively little to the total watershed area.

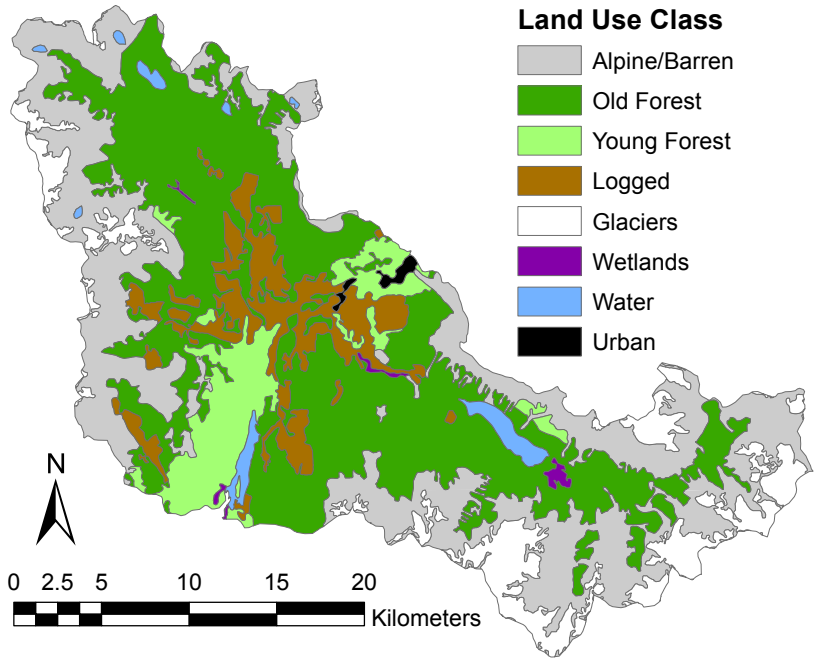

Figure 3.1: Map of the Cheakamus watershed showing land-use/land cover classes utilized in the WATFLOOD model. Map derived from data provided by BC Hydro.

A selection of readily-interpreted, optimized WaSiM model parameters is shown in Table 3.1. By examining these parameters, we can anticipate how the NSE_o and $LNSE_o$ WaSiM model runs will differ from those made using the MAE_o parameter set. Since the NSE_o rain-snow threshold ($T_{R/S}$) is lower, we should expect less snow accumulation and flashier precipitation-driven inflow events during the fall and winter. A slightly lower threshold temperature for snowmelt (T_{melt}) means that melt may begin earlier in the year, and a lower degree-day melt factor (MF) should result in a slower melt rate. Similarly, we anticipate $LNSE_o$ simulations to have less snow accumulation than the MAE_o and the NSE_o, with more rain events during the fall and winter. The timing of spring snowmelt should be approximately the same as the NSE_o simulations, but snowmelt will be slower.

Table 3.1: Selected model parameters for the WaSiM hydrologic model, as optimized by the DDS algorithm using different objective functions.

Model Parameter	MAE_o	NSE_o	$LNSE_o$
$T_{R/S}$ (°C)	-1.44	-1.75	-1.88
T_{melt} (°C)	1.68	1.28	0.87
MF (mm/day/°C)	1.59	1.48	1.08

Figure 3.2 displays snow water equivalent (SWE) simulations made by WaSiM for the 2009–2010 water year. Observed SWE is from the Squamish Upper snow pillow site operated by the Water Survey of Canada, located just outside the western boundary of the watershed. Simulated SWE is at a proxy location in the watershed selected based on elevation, aspect, terrain, and a comparison of PRISM (Parameter-elevation Regressions on Independent Slopes Model) 1961–1990 climate normal data at the real and proxy sites (PRISM Climate Group, 2012). PRISM data was downscaled to 400 m resolution by ClimateBC (Wang et al., 2006). Results in Figure 3.2 confirm that the NSE_o- and $LNSE_o$-based simulations accumulate less snow than MAE_o. The melt rate for these simulations is lower than MAE_o simulations and observations. The timing of the onset of spring snowmelt is similar for all model runs, and is in good agreement with observations.

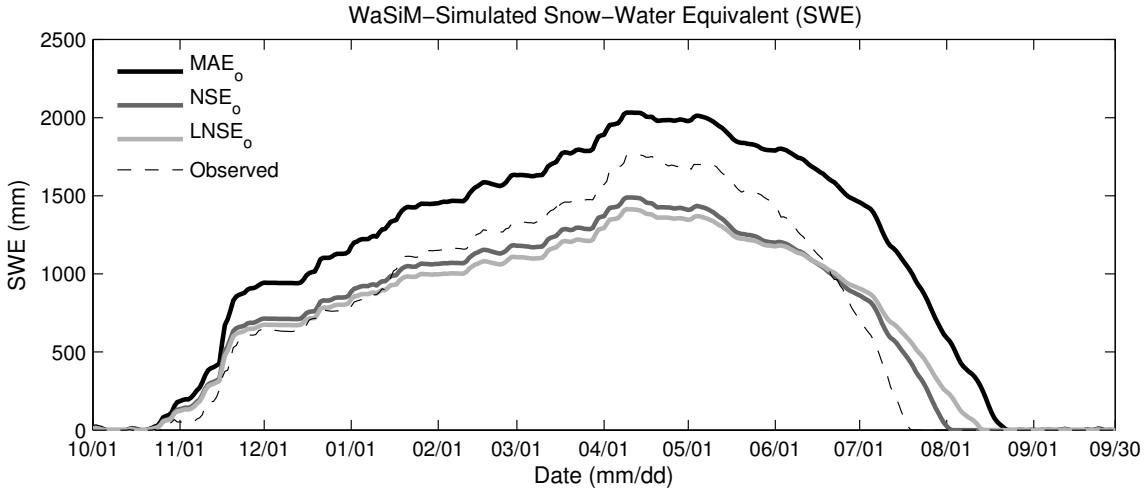

Figure 3.2: Snow-water equivalent at the Squamish Upper proxy site as simulated by the WaSiM hydrologic model using the MAE_o, NSE_o and $LNSE_o$ parameter sets.

The impact of different parameter values can also be seen in WaSiM simulated inflows. These simulations are driven by observed meteorological data and are a by-product of updating the daily hydrologic state or initial condition. Figure 3.3 illustrates the process of generating updated hydrologic states, simulated inflows, and forecasted inflows (which are driven by NWP fields) for an individual DH model. As anticipated, the fall/winter period displayed in Figure 3.4 shows NSE_o and $LNSE_o$ simulated inflows to be flashier than the MAE_o simulated inflows due to different rain/snow partitioning of precipitation events. These results have not been bias-corrected. The impact of varying soil parameters is more difficult to discern from the inflow record in such a flashy watershed. WaSiM model output allows for a detailed analysis of water storage in the unsaturated and saturated soil zones among many other diagnostic products; such output was not analyzed in this study.

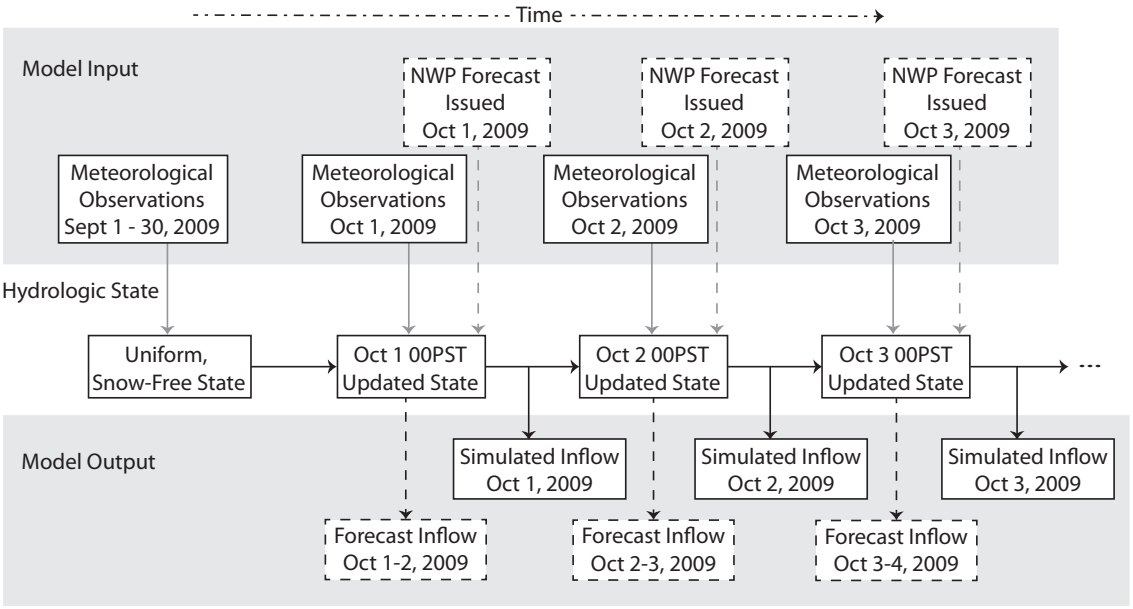

Figure 3.3: Flowchart illustrating the process of generating updated hydrologic states, simulated inflows, and forecasted inflows for a particular hydrologic model. Solid lines show the flow of meteorological observations into the model and the production of simulated inflows and updated hydrologic states for the following day. Dashed lines show the flow of NWP forecasts into the model and the resulting 2-day inflow forecasts.

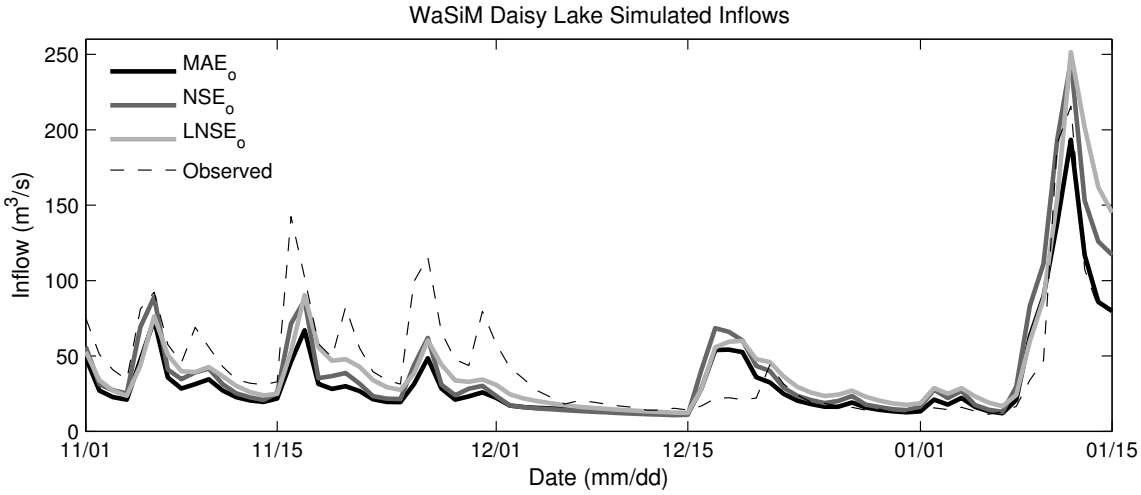

Figure 3.4: Daisy Lake inflows during fall and early winter of the 2009–2010 water year simulated by the WaSiM model using the MAE_o, NSE_o and $LNSE_o$ parameter sets.

WATFLOOD model parameters for alpine and old forest land-use classes are shown in Table 3.2. These classes are examined because they are the largest classes in the watershed (Figure 3.1), and because the proxy site for comparing modelled and observed SWE is in the old forest class. Note that in WATFLOOD, $T_{R/S}$ is taken to be T_{melt}, whereas WaSiM allows a rain/snow mix to occur over a range of temperatures and rain/snow and snowmelt temperature thresholds can differ. WATFLOOD model parameters for the old forest land class indicate that relative to MAE_o, both NSE_o and $LNSE_o$ models may begin to accumulate snow earlier in the fall, and will do so at a greater rate. Flashy rain-on-snow or winter rainfall events are less likely due to higher threshold temperatures for rain/snow partitioning. NSE_o and $LNSE_o$ snowmelt should begin later in the spring, with $LNSE_o$ having a greater rate of melting. In alpine areas, T_{melt} for NSE_o and $LNSE_o$ is only slightly lower than for the MAE_o parameter set, so the differences in rain/snow partitioning are likely insignificant. The NSE_o melt factor is less than that of the other models, so this land class will contribute less to snowmelt-driven inflows, but its contribution may last longer into the summer.

Table 3.2: Same as Table 3.1, but for the two primary land classes in the WATFLOOD model.

Model Parameter	Land Class	MAE_o	NSE_o	$LNSE_o$
T_{melt}, $T_{R/S}$ (°C)	Alpine	-1.90	-2.12	-2.12
	Old Forest	1.96	2.71	2.84
MF (mm/h/°C)	Alpine	0.24	0.17	0.24
	Old Forest	0.11	0.10	0.12

Figure 3.5 shows that SWE simulated at the proxy location for the 2009–2010 water year varies between the different model simulations as expected. That is, the rate of snow accumulation at the site is slightly higher for NSE_o and $LNSE_o$ than for MAE_o, and melt begins later in the season. Once melt begins, the $LNSE_o$ SWE drops off faster than the NSE_o and MAE_o SWE due to the higher degree-day melt factor (note that this is a simplified explanation, as the melt rate is also related to the difference between T_{melt} and actual air temperature). In a previous study (Chapter 2), WATFLOOD was found to be late in simulating the onset of snowmelt during La Niña summers; this phenomenon is visible in Figure 3.5. Simulated inflows are not shown because the interaction of inflow contributions from the various land classes makes it difficult to assess the impact of a small number of model parameters on this variable. However, it appears that the MAE_o simulated inflows are slightly flashier than the NSE_o and $LNSE_o$ inflows during the fall and winter, and this could be due to the lower MAE_o $T_{R/S}$ temperature in the old forest land class. Additional diagnostic model output available for examining the impacts of soil parameters has not been examined.

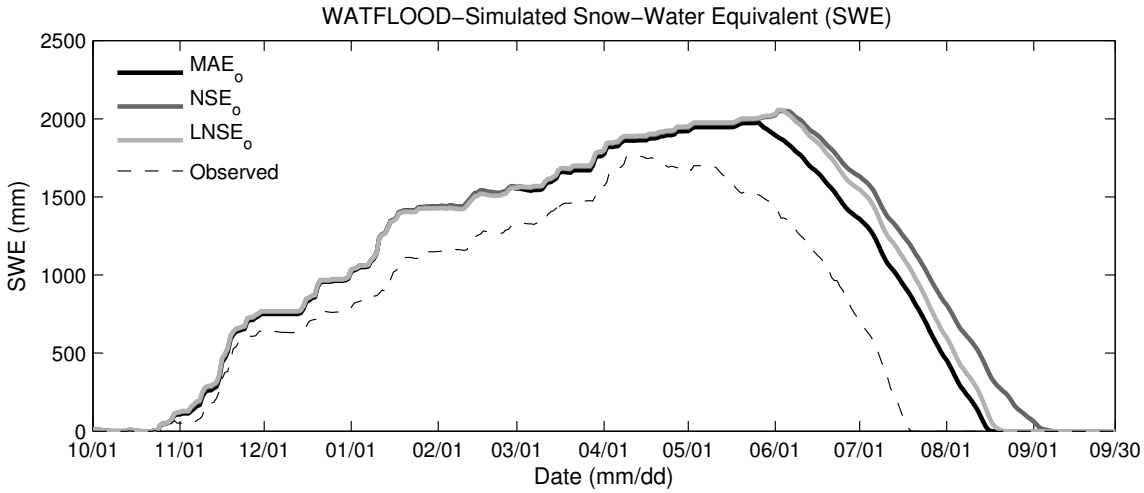

Figure 3.5: As in Figure 3.2, but simulations done by the WATFLOOD hydrologic model.

3.3.4 A Multi-State Hydrologic Ensemble

The multi-state or multi-initial-condition component of the M2M ensemble forecasting system arises as a direct consequence of implementing a multi-parameter component. On each forecast day, the model state is updated by simulating watershed processes using yesterday's meteorological observations to drive the DH models (Figure 3.3). In order to maintain equilibrium in the hydrologic models, the model parameterization used to drive the forecast must match that used in generating these initial conditions. Consider, for example, the WaSiM model parameter m, which describes water recession in the saturated zone. This parameter has a direct impact on the soil saturation deficit, so if the model state is updated with a particular m value, it may generate a hydrologic state with a deficit. If the forecast beginning from this state uses a different value for m, this deficit may suddenly become a surplus, resulting in an immediate release of water from the saturated zone.

To avoid such discontinuities, the initial model spin-up from uniform, snow-free conditions at the beginning of the case-study water year is done for each hydrologic model using each of the three different parameter sets with meteorological observations used to drive the models. WATFLOOD and WaSiM each have three different sets of initial conditions — a MAE$_o$ state, a NSE$_o$ state, and a LNSE$_o$ state. Each forecast day, these states are updated using the corresponding parameter sets and newly observed meteorological data, and then forecasts are made from these updated states, again using the corresponding model parameterization. This process is illustrated in Figure 3.3 for a single model/parameterization/state. Forecasts made using the MAE$_o$ parameter set and beginning from the MAE$_o$ state will be referred to as MAE$_o$ forecasts, and so forth. These different hydrologic

states comprise a limited sampling of the uncertainty space in the models' initial conditions.

While ensemble data assimilation methods are available for hydrologic modelling applications (e.g., Andreadis and Lettenmaier, 2006; Clark et al., 2008), we have opted to limit the component of the M2M ensemble that samples hydrologic state uncertainty to just those states necessitated by the multi-parameter ensemble. This is because of the paucity of observed data available within the watershed for assimilation. DeChant and Moradkhani (2011a) have had some success using assimilation of observed SWE to update hydrologic state in seasonal forecasting. However, the method was found to be sensitive to the availability of representative observations and would therefore possibly fail to produce an accurate state for the Cheakamus watershed. The Retrospective Ensemble Kalman Filter (REnKF; Pauwels and De Lannoy, 2006) may be worth exploring, though the computational expense of ensemble data assimilation can be prohibitive in an operational forecasting framework. Dual state-parameter estimation frameworks that incorporate data assimilation could also be used for a more complete handling of parameter and initial condition uncertainty (e.g., Moradkhani et al., 2005a; DeChant and Moradkhani, 2011b; Leisenring and Moradkhani, 2011).

The full ensemble including the six different hydrologic models (two distinct DH models, each with three parameterizations/states) and driven by the multi-model, multi-grid scale NWP ensemble with different downscaling schemes has a total of 72 ensemble members. A sample workflow for generating the MAE_o WaSiM forecasts is illustrated in Figure 3.6. Each day, each model configuration (consisting of a hydrologic model and a parameter set/hydrologic state) is driven by 12 different downscaled NWP forecast fields, generating 12 different inflow forecasts. This forecast workflow is indicated by the solid arrows. Dashed arrows illustrate how meteorological observations are used to update the model configuration's hydrologic state for the following day's forecasts. The model configuration is indicated by dash-dotted arrows. This process is repeated for each watershed model (WaSiM and WATFLOOD) and each parameterization/state (MAE_o, NSE_o and $LNSE_o$), yielding 72 unique inflow forecasts each day.

3.3.5 Bias Correction of Inflow Forecasts

Prior to combination and evaluation, each of the 72 inflow forecast ensemble members is post-processed to remove unconditional bias. The purpose of bias correction is to correct for systematic errors in the dynamic NWP and DH models. Since each ensemble member is derived from a different NWP model driving a different DH model, individual member bias correction is necessary in this context.

An appropriate measure of bias for volumetric quantities such as precipitation or reservoir inflow is the degree of mass balance (DMB; McCollor and Stull, 2008a). The DMB is a measure of the ratio of simulated or forecasted inflow volume to the observed inflow volume over a given period of

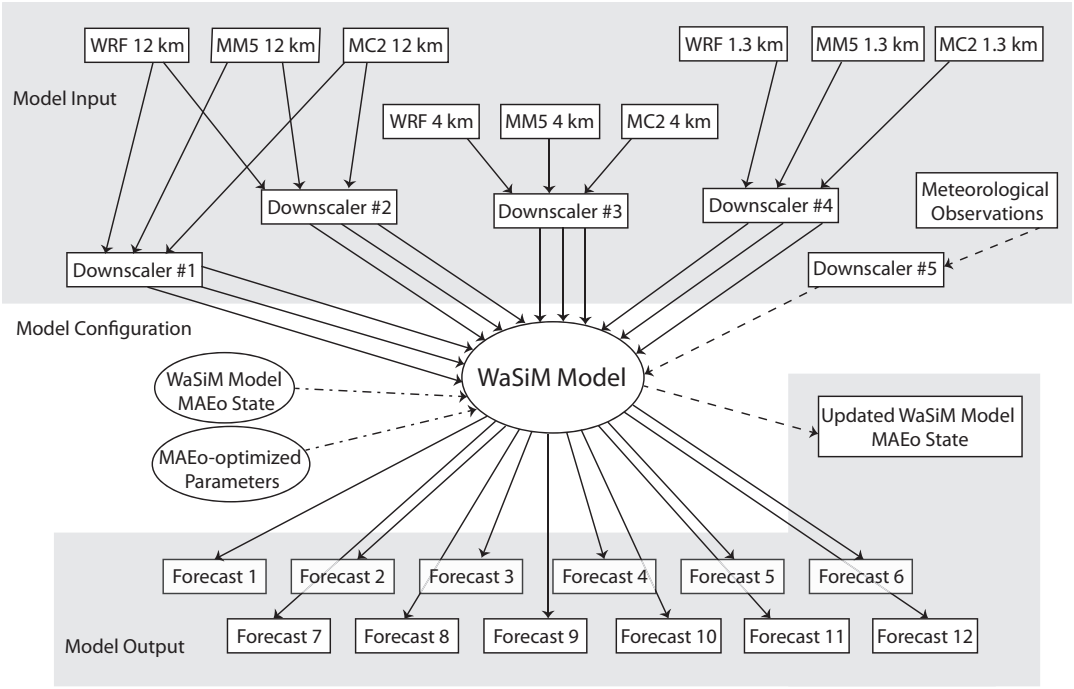

Figure 3.6: The flow of information into and out of the WaSiM model for generating MAE_o forecasts. Each model (WaSiM and WATFLOOD) and each parameterization (MAE_o, NSE_o and $LNSE_o$) generates 12 different daily forecasts in this way for a combined total of 72 unique daily forecasts.

time (see Appendix A). The use of a multiplicative bias corrector prevents corrected inflows from becoming negative.

A linearly-weighted DMB bias correction factor calculated over a moving window of three days was found to be ideal in removing bias from the multi-NWP, multi-DH M2M ensemble forecasts evaluated in Chapter 2. This study found bias in the hydrologic state used to start each inflow forecast to be the main contributor to forecast bias. The importance of this bias source is likely the reason that a short correction window performs so well; due to the flashy nature of the case-study watershed, only recent forecast errors are likely to play an important role in bias correction for short-term forecasts. The linearly-weighted three-day DMB bias corrector ($LDMB_3$) developed in Chapter 2 is applied in this study. Uncorrected forecasts are referred to herein as 'raw' forecasts.

3.4 Results and Discussion

In Chapter 2, a M2M ensemble consisting of the multi-NWP and multi-DH ensemble components with a single 'best' (MAE_o) model parameterization was evaluated. In this chapter, the impact

of adding the multi-parameter and multi-state ensemble components is evaluated by comparison against the smaller ensemble.

An initial analysis of individual forecast ensemble members indicated that the NSE_o and $LNSE_o$ members performed poorly relative to the MAE_o members for many measures-oriented verification scores. Figure 3.7 shows the performance of day 1 inflow forecasts driven by the 4-km WRF model output, evaluated over the 2009–2010 water year. Note that MAE, NSE and LNSE without the subscript '$_o$' refer to verification measures (see Appendix A), while those with the subscript refer to model forecasts from a particular model parameterization/hydrologic state. The differences in performance among differently-optimized forecasts driven by this particular NWP model are consistent with those driven by output from other NWP models and grid scales. Thus, this can be considered a representative sample of the relative performance of the various DH model parameterizations.

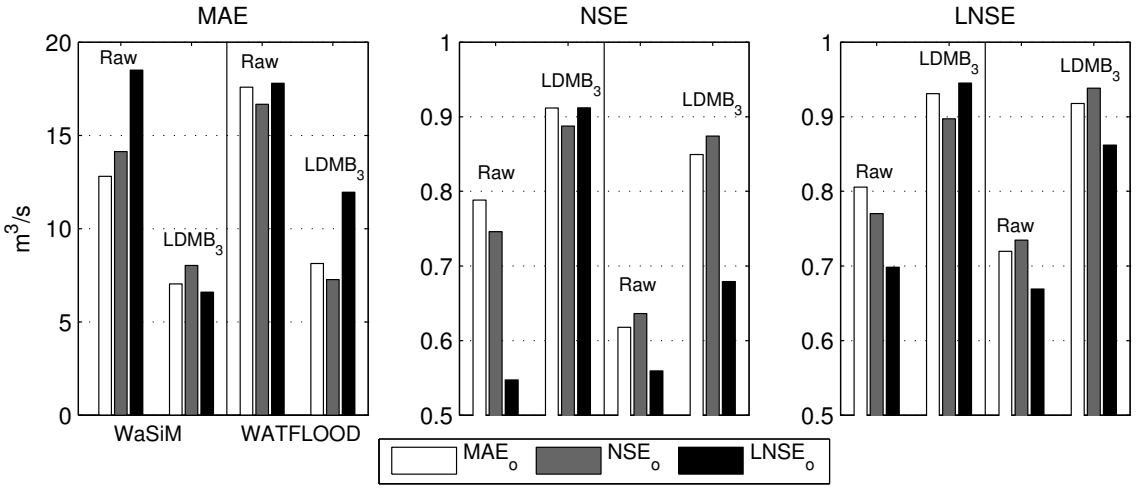

Figure 3.7: Performance of day 1 MAE_o, NSE_o and $LNSE_o$ forecasts from the WaSiM and WATFLOOD models driven by the 4-km WRF NWP output fields. Perfect inflow forecasts have NSE and LNSE equal to one (unitless), and MAE of zero m^3/s.

The raw $LNSE_o$ WaSiM forecasts perform poorly relative to the other optimizations, especially in terms of MAE and NSE. The differences among the raw WATFLOOD members are not as large. Following bias correction using the $LDMB_3$ correction factor, the performance of the WaSiM members is roughly equal. In the case of WATFLOOD, the bias corrector is unable to improve the $LNSE_o$ ensemble member performance to the same degree as the MAE_o and NSE_o members (in terms of MAE and NSE). A comparison of the forecast hydrographs for these ensemble members (not shown) reveals that the characteristics of snowmelt in the $LNSE_o$ forecast cause contributions to inflow during the rising limb of the freshet to fluctuate more rapidly than those from the MAE_o

forecast. This erratic behaviour could hamper the ability of the bias correction scheme to reduce errors during this period.

Based on this initial analysis, it is not immediately obvious whether addition of the NSE_o and, in particular, the $LNSE_o$ ensemble members to the M2M inflow forecasting system will improve any characteristics of ensemble performance. Figure 3.8 shows how the bias-corrected ensemble mean forecast performance changes as the NSE_o members and $LNSE_o$ members are added to the MAE_o-only ensemble evaluated in Chapter 2. Adding the NSE_o ensemble members improves all measures of ensemble mean performance at all forecast lead times, while the additional inclusion of the $LNSE_o$ members diminishes these improvements. The ensemble configuration that includes all three model parameterizations is referred to as the full ensemble.

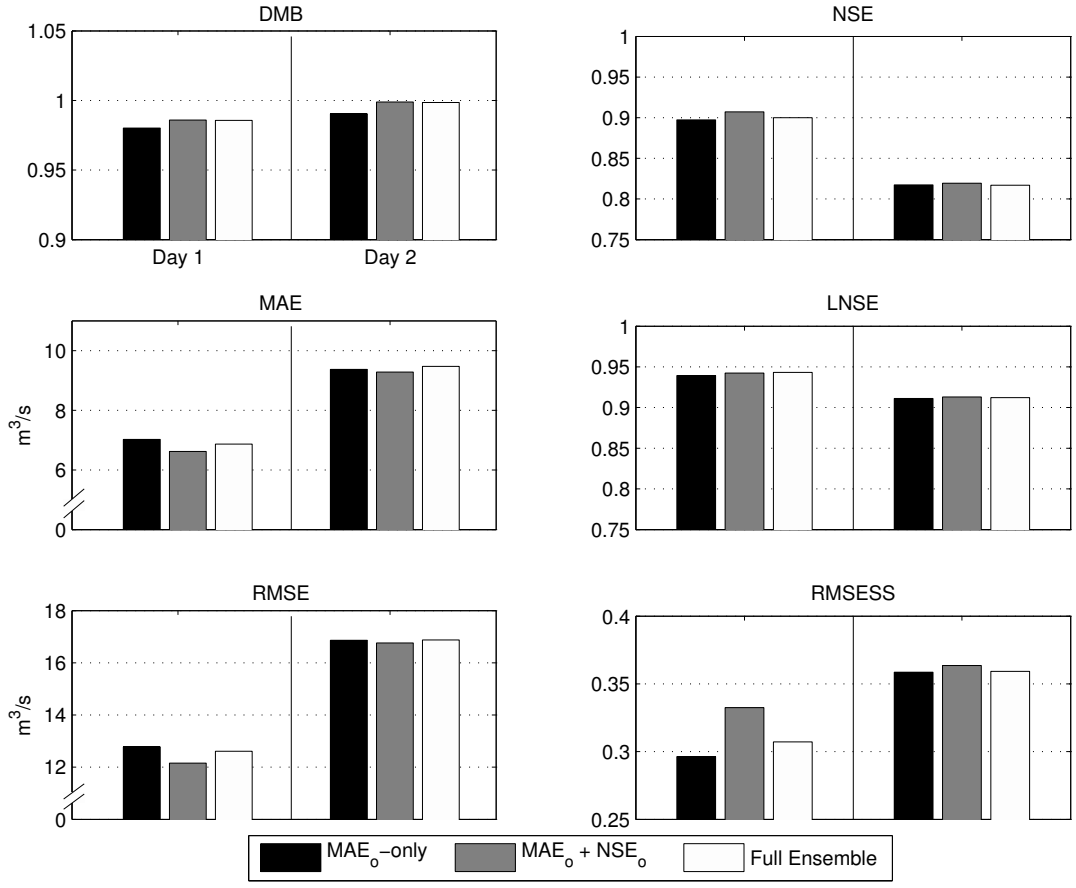

Figure 3.8: Performance of the bias-corrected M2M ensemble mean with MAE_o ensemble members only, and with the addition of the NSE_o and $LNSE_o$ ensemble members. Perfect inflow forecasts have DMB, NSE, LNSE and RMSESS equal to one (unitless), and MAE and RMSE of zero m^3/s.

The Brier Skill Scores (BSS) shown in Figure 3.9 indicate that the addition of the NSE_o members to the MAE_o-only ensemble improves the BSS for the 19.5 m^3/s inflow anomaly threshold on forecast days 1 and 2. Scores are shown for bias-corrected forecasts (indicated by bar heights) and also for raw forecasts (indicated by triangles). Anomalies are calculated by subtracting the daily climatological median inflow from the forecast, and are used so that the forecasting system is not rewarded for making high inflow forecasts during the snowmelt season when little skill is required to do so. Adding the $LNSE_o$ members offers little to no improvement to the BSS. Decomposition of the BSS into relative reliability and relative resolution components shows that this is because the full ensemble has poor (higher) reliability relative to the MAE_o-plus-NSE_o ensemble, but greater resolution. In fact, the addition of the NSE_o members to the MAE_o members results in almost no improvement to the day 2 forecast resolution, which is the most important attribute of an ensemble forecasting system (Toth et al., 2003). Reliability can be corrected by further post-processing to remove conditional bias, whereas resolution can only be corrected by improving the forecast "engine" used to generate the ensemble, for example, through unconditional bias correction. This is clearly indicated by the difference between raw and bias-corrected forecast scores in Figure 3.9. This result demonstrates that while inclusion of more diverse ensemble members to the M2M ensemble is important for forecast quality, the forecast is only able to reach its full potential following bias correction.

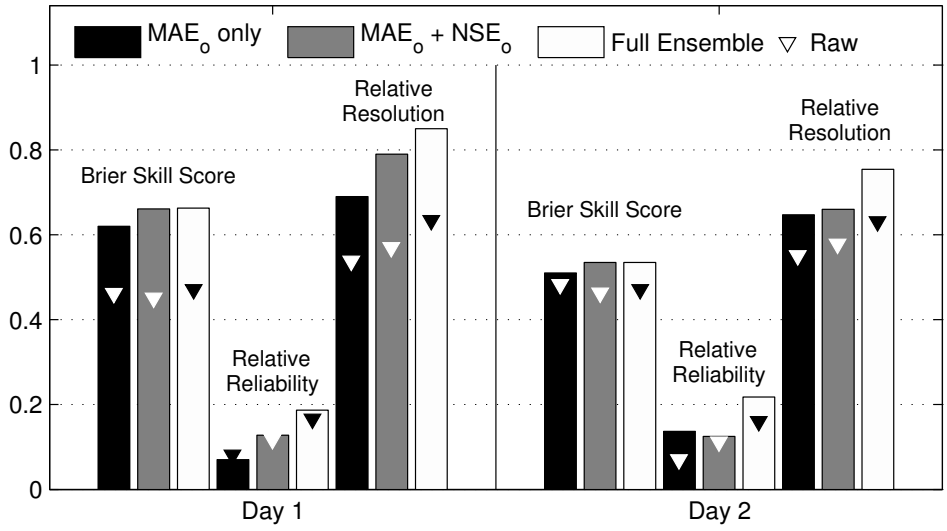

Figure 3.9: Brier skill score (BSS = 1 is perfect), relative reliability (zero is perfect) and relative resolution (one is perfect) for different ensemble forecasts for days 1 and 2. Scores for bias-corrected forecasts are indicated by bar heights, while those for raw forecasts are indicated by triangles. The inflow anomaly threshold evaluated here is 19.5 m^3/s.

Relative to the MAE_o-only ensemble, inclusion of the NSE_o and $LNSE_o$ ensemble members also results in improved forecast discrimination for a range of anomaly thresholds as indicated by the ROC curves in Figure 3.10. Results are shown for both raw and bias-corrected ($LDMB_3$) ensemble forecasts and clearly indicate the importance of bias correction of individual ensemble members prior to combination.

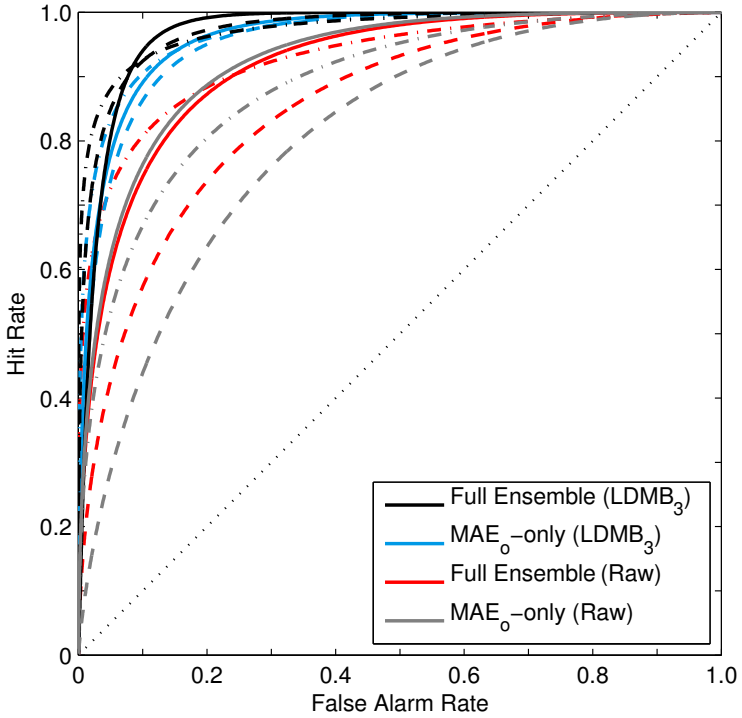

Figure 3.10: ROC diagrams for raw and $LDMB_3$ bias-corrected day 1 forecasts for inflow anomalies greater than -5.0 m^3/s (dot-dashed line), 2.7 m^3/s (dashed line) and 19.5 m^3/s (solid line). The dotted line is the zero-skill line.

Finally, the ensemble rank histograms shown in Figure 3.11 indicate that the full ensemble is, as intended, more dispersive than the MAE_o-only ensemble on forecast days 1 and 2 as a result of its improved error sampling. The upper histograms show 48% and 33% of observations falling outside of the range of day 1 and day 2 MAE_o-only forecasts, respectively. These values drop to 26% and 16% respectively with the addition of the NSE_o and $LNSE_o$ ensemble members. Thus, while the 72-member ensemble attempts to explicitly sample all sources of error in the hydrologic modelling chain, it still fails to capture the full range of forecast uncertainty. We expect that the limited sampling of hydrologic state uncertainty is the main cause of this underdispersion. Dual state-parameter estimation methods could be employed for more complete handling of parameter

and initial condition uncertainty if additional observed data were available within the case study watershed (Moradkhani et al., 2005a,b; DeChant and Moradkhani, 2011b; Leisenring and Moradkhani, 2011), though such methods are computationally expensive.

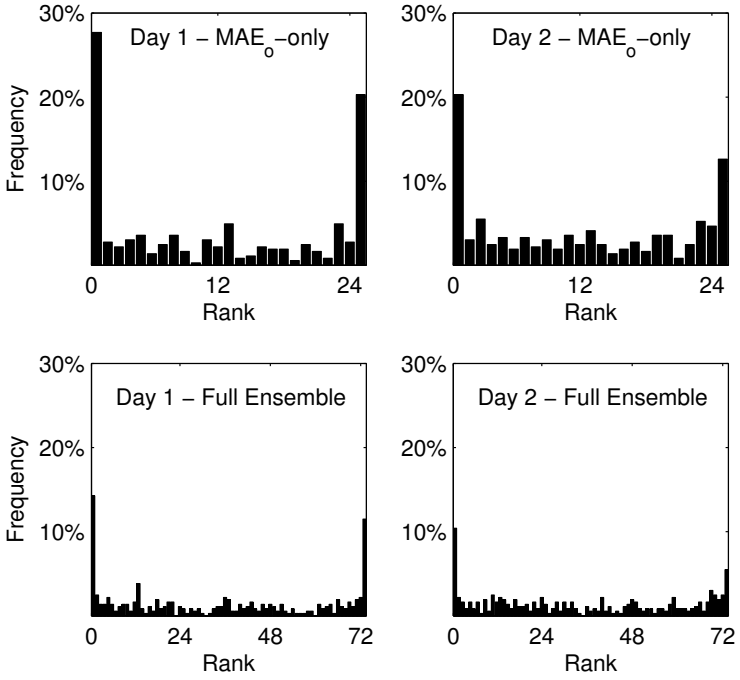

Figure 3.11: Rank histograms for the bias-corrected MAE$_o$-only and full ensembles. The full ensemble has greater dispersion as indicated by a smaller percentage of observations falling into the extreme bins of the histogram.

Increased spread is also evident in the raw ensemble hydrograph traces shown in Figure 3.12 as compared with those in Figure 2.3. Ensemble members derived from the same hydrologic model parameterizations have a tendency to cluster together, supporting the finding in Chapter 2 that bias in the model simulation used to generate the daily hydrologic state is the primary contributor to overall forecast bias. This clustering is most visible near the peak of summertime snowmelt-driven inflows in July and August. The spread of the full ensemble is greatly reduced following LDMB$_3$ bias correction (Figure 3.13), but remains larger than that of the MAE$_o$-only ensemble (Figure 2.5).

3.5 Concluding Remarks

In this chapter, we have evaluated the impact of incorporating a multi-parameter, multi-state component into a M2M ensemble forecasting system consisting of a multi-NWP, multi-hydrologic

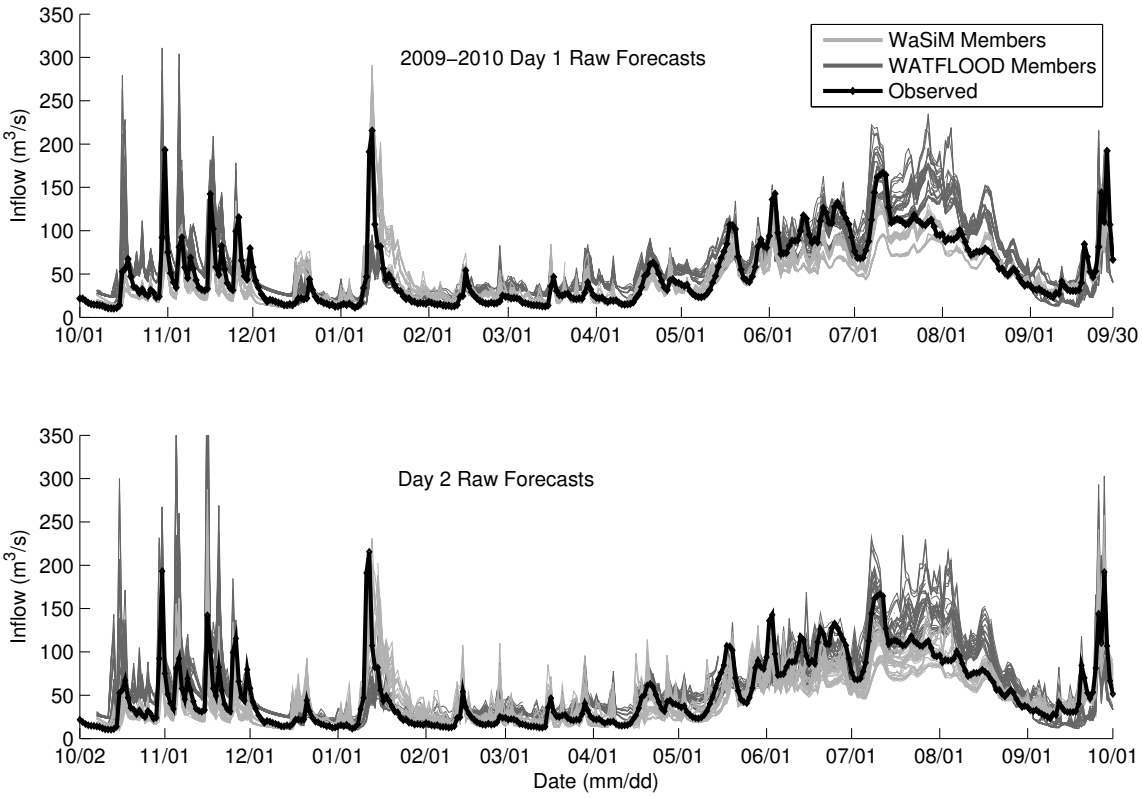

Figure 3.12: Raw ensemble traces for day 1 (top) and day 2 (bottom) forecasts during the 2009–2010 water year for all hydrologic model parameterizations.

model component. The multi-parameter component was achieved by optimizing the WaSiM and WATFLOOD hydrologic models with different objective functions (MAE, NSE and NSE of log-transformed flows). The multi-state component is necessitated by the use of multiple parameterizations in order to avoid discontinuities in the inflow forecasts that could occur due to suddenly changing model parameters. To the best of our knowledge, this is the first example of a short-term hydrologic forecasting ensemble that explicitly attempts to sample all sources of hydrologic uncertainty.

Initial analysis of inflow forecast performance indicated that the addition of the LNSE_o ensemble members had a negative impact on the performance of the ensemble mean relative to an ensemble mean comprised of MAE_o and NSE_o members alone. However, examination of distributions-oriented measures of forecast performance revealed that while the inclusion of the LNSE_o ensemble members did result in a deterioration of ensemble reliability, it significantly improved the ensemble resolution. Recall that reliability can easily be corrected using calibration methods (e.g., Hamill

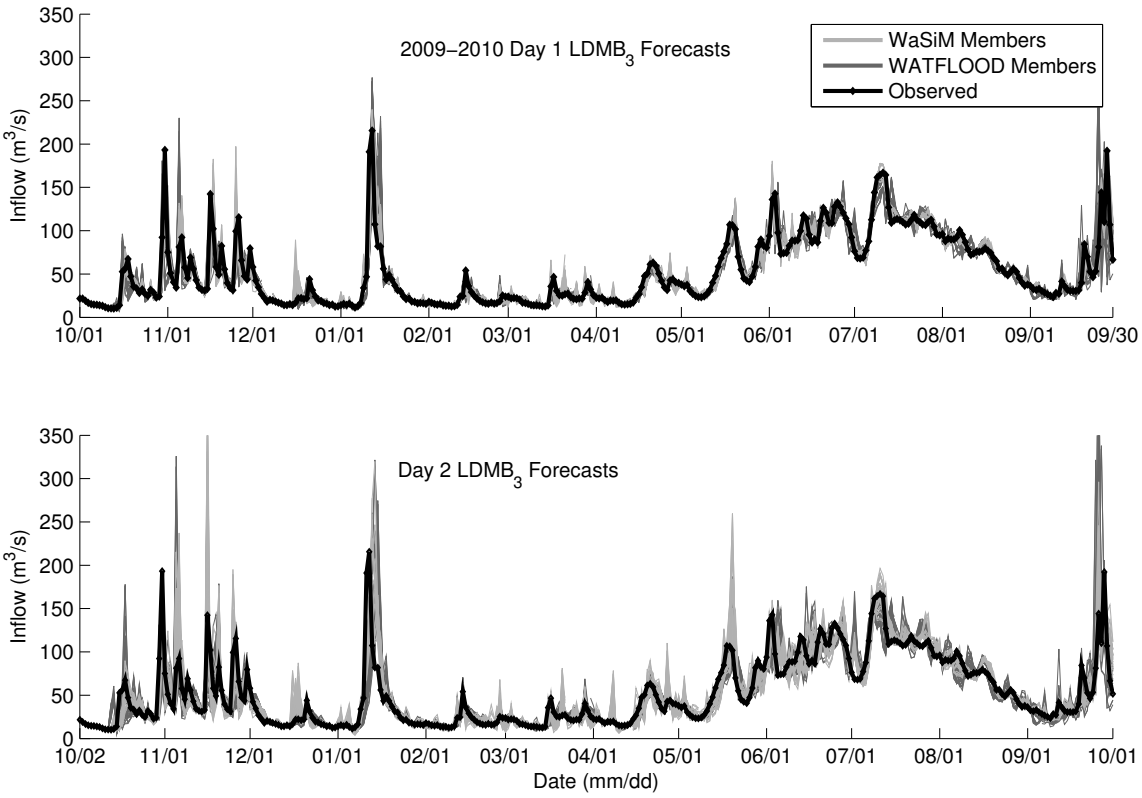

Figure 3.13: As in Figure 3.12, but following $LDMB_3$ bias correction.

and Colucci, 1997; Nipen and Stull, 2011), whereas resolution can only be corrected by improving the forecasting "engine" used to generate the ensemble. Adding diversity to the M2M ensemble by explicitly attempting to sample the full range of forecast uncertainty is important in improving this most important aspect of ensemble quality. However, it was found that the removal of unconditional forecast bias offered significant additional gains, enabling the ensemble to reach its true potential.

Based on these results, future work on the M2M ensemble forecasting system will include bias-corrected ensemble members from all three model parameterizations. While the spread of the full ensemble is greater than the MAE_o-only ensemble evaluated in a previous study (Chapter 2), it remains underdispersive. This will be improved upon by applying an appropriate uncertainty model to transform the ensemble into a reliable probabilistic forecast, and by performing probability calibration if and when necessary to further improve reliability.

Chapter 4

Reliable Probabilistic Forecasts from an Ensemble Reservoir Inflow Forecasting System

4.1 Introduction

Forecasts of weather and hydrologic variables are subject to uncertainty due to errors introduced into the modelling chain via imperfect initial and boundary conditions, poor model resolution of terrain and small-scale processes, and the necessary simplification of physical process representation in the models themselves (e.g., Palmer et al., 2005; Bourdin et al., 2012). Deterministic forecasts ignore these errors and may provide forecast users with a false impression of certainty. Probabilistic forecasts expressed as probability distributions are a way of quantifying this uncertainty by indicating the likelihood of occurrence of a range of forecast values. Additionally, reliable probabilistic inflow forecasts enable hydroelectric reservoir managers to set risk-based criteria for decision making and offer potential economic benefits (Krzysztofowicz, 2001).

Ensemble forecasting techniques are designed to sample the range of uncertainty in forecasts, but are often found to be unreliable, with underdispersiveness being a frequently cited deficiency in both weather and hydrologic forecasting applications (e.g., Eckel and Walters, 1998; Buizza, 1997; Wilson et al., 2007; Olsson and Lindström, 2008; Wood and Schaake, 2008). In order to correct these deficiencies, uncertainty models can be used to fit a probability distribution function (PDF) to the ensemble, whereby the parameters of the distribution are estimated based on statistical properties of the ensemble and the verifying observations. These theoretical fitted distributions reduce the amount of data required to characterize the distribution (for example, from 72 ensemble members to two parameters describing the mean and spread of a Gaussian distribution), and allow estimation of probabilities for events that lie outside of the range of observed or modelled behaviour (Wilks, 2006).

Uncertainty models make different assumptions about how the ensemble members and observations are generated. For example, centering a Gaussian probability distribution on the ensemble mean with spread proportional to the ensemble variance makes the assumption that the ensemble mean forecast errors are normally distributed (or, equivalently, that the verifying observations are drawn from a normal distribution centred at the ensemble mean). This model also assumes the existence of a spread-skill relationship. That is, the spread of the ensemble members should be related to the accuracy (or skill) of the ensemble mean; when the forecast is more certain, as indicated by low ensemble spread, errors are expected to be small. However, this relationship is often tenuous (e.g., Hamill and Colucci, 1998; Stensrud et al., 1999; Grimit and Mass, 2002). If the uncertainty model assumptions are valid, the resulting probability forecasts should be statistically reliable or *calibrated*, meaning that an event forecasted to occur with probability p will, over the course of many such forecasts, be observed a fraction p of the time (Murphy, 1973). Otherwise, the probabilistic forecasts cannot be used for risk-based decision making, since the probabilities cannot be taken at face value.

Various methods of statistical calibration have been devised to correct for deficiencies in probabilistic forecasts. These can generally be split into two groups: ensemble calibration, which adjusts individual ensemble members in order to produce reliable forecasts; and probability calibration, which adjusts the probabilities directly. Examples of ensemble calibration include Bayesian Model Averaging (BMA; Raftery et al., 2005) and generalizations thereof (e.g., Johnson and Swinbank, 2009). The weighted ranks method (Hamill and Colucci, 1997) and its generalization, the Probability Integral Transform (PIT)-based calibration of Nipen and Stull (2011) are examples of probability calibration that have been shown to improve the reliability and value of forecasts of precipitation, temperature, wind speed, and other meteorological variables. Nipen and Stull (2011) also demonstrated that their method was able to further improve forecasts generated using BMA. Bayesian methods have been applied successfully in hydrologic forecasting applications over a range of timescales (e.g., Duan et al., 2007; Reggiani et al., 2009; Wang et al., 2009; Parrish et al., 2012). Probability calibration on the other hand, has not yet been widely adopted by the hydrologic modelling community. Olsson and Lindström (2008) provide an example of a very simple probability calibration used to improve ensemble spread. Roulin (2007) applied the weighted ranks method to medium-range forecasts of streamflow and found very little improvement to the already reliable forecasting system. Quantile mapping (QM) is a similar probability calibration technique, but is suited to seasonal hydrologic forecasting, as it maps forecast probabilities to their corresponding climatological values (Hashino et al., 2007; Madadgar et al., 2012).

In this chapter, we apply two simple uncertainty models to a 72-member ensemble of bias-corrected reservoir inflow forecasts in order to generate probabilistic forecasts. We then test the

application of a probability calibration scheme to improve reliability where necessary. All post-processing applied to the ensemble is done via the COmmunity Modular Post-processing System (COMPS). This system was originally described and implemented by Nipen (2012) and is now available as open-source at http://wfrt.github.io/Comps/. We have contributed a number of new and existing schemes within the COMPS framework for bias correction, uncertainty modelling and "intelligent" calibration; a description of COMPS and the applied schemes follows.

4.2 Case Study

4.2.1 Study Dates and Data

In this study, various uncertainty models and probability calibration strategies are tested on a 72-member ensemble reservoir inflow forecasting system developed for the Daisy Lake reservoir, a hydroelectric facility on the upper Cheakamus River in southwestern British Columbia (BC), Canada. The reservoir is operated by the BC Hydro and Power Authority (BC Hydro). Evaluation of the ensemble is carried out over the 2010–2011 and 2011–2012 water years. For this particular hydroclimatic regime, a water year is defined as the period from October 1 to September 30. Fall and winter storm season inflows are primarily driven by precipitation from Pacific frontal systems. Rain-on-snow events can result in significant inflows during this period. During the spring and summer, inflows are snowmelt-driven, with some late-season glacier melt contributions.

Daily average inflow rates are calculated by BC Hydro using a water balance based on observed reservoir levels and outflows. The calculated inflows employed in this study are considered to be of high quality. For the purposes of this study, these values will be referred to as observed inflows. Hourly forecasts of inflows to the Daisy Lake reservoir are transformed into daily average inflow rates for verification against these observations.

Ensemble and probability forecasts were generated continuously from the beginning of the 2009–2010 water year through the end of the study period. The first water year (2009–2010) was used to spin up the COMPS model parameters (described in Section 4.2.3 and Section 4.3.1) and is excluded from evaluation.

4.2.2 The Member-to-Member (M2M) Ensemble Forecasting System

The Member-to-Member (M2M) ensemble forecasting system used for forecasting inflows to the Daisy Lake reservoir explicitly samples uncertainty arising from errors in the Numerical Weather Prediction (NWP) input fields used to drive the Distributed Hydrologic (DH) models, the hydrologic models themselves and their parameterizations, and the hydrologic states or initial conditions used

to begin each daily forecast run. The result is an ensemble of 72 unique daily inflow forecasts.

The NWP models are taken from the operational ensemble suite run by the Geophysical Disaster Computational Fluid Dynamics Centre (GDCFDC), in the Department of Earth, Ocean and Atmospheric Sciences at the University of British Columbia. The ensemble consists of three independent nested limited-area high-resolution mesoscale models with forecast domains centred over south-western BC: the Mesoscale Compressible Community model (MC2; Benoit et al., 1997); the fifth-generation Pennsylvania State University-National Center for Atmospheric Research Mesoscale Model (MM5; Grell et al., 1994); and Version 3 of the Weather Research and Forecasting (WRF) model (Skamarock et al., 2008). Hourly model output fields with grid spacing of 12, 4 and 1.3 km are used for this study.

From the start of the modelling period (October 2009) through March 2012, all NWP models were initialized at 00UTC using the National Centers for Environmental Prediction (NCEP) North American Mesoscale (NAM) model, which also provides time-varying boundary conditions. In March 2012, the initial/boundary condition for the MM5 and WRF was switched to the NCEP Global Forecast System (GFS) model, while MC2 continued to make use of the NAM.

NWP (and therefore inflow) forecast horizon varies during the case-study period. From the start of the modelling period through April 2010, all NWP models were run out to 60 hours except for the 1.3-km MC2 model runs, which are limited to 39 hours due to operational time constraints. All WRF grids began producing 84-hour forecasts in late April, 2010. In March 2011, the MM5 12-km and 4-km forecasts were extended to 84 hours, enabling them to drive a 3-day inflow forecast. In March 2012, 1.3-km MM5 model output was also made available out to 84 hours.

The Distributed Hydrologic (DH) models applied to the case-study watershed are the Water balance Simulation Model (WaSiM; Schulla, 2012) and WATFLOOD (Kouwen, 2010). These models were selected because they are distributed, and therefore able to take advantage of high-resolution NWP input. They are also able to simulate snow and glacier melt processes and lakes in complex terrain given relatively limited input data. Both DH models are run at 1 km grid spacing at an hourly time step. The required NWP fields are downscaled to the DH model grids using interpolation schemes built into each DH model. For the WaSiM model, 12-km NWP fields (temperature, precipitation, wind speed, humidity, and global radiation) are downscaled using two methods: inverse-distance weighting (IDW) and elevation-dependent regression (Schulla, 2012). The 4-km and 1.3-km NWP fields are downscaled using a bilinear interpolation scheme. WATFLOOD downscaling is done using IDW that incorporates elevation dependence using an optional constant elevation adjustment rate for both temperature and precipitation (these being the only required NWP fields). 12-km fields are downscaled using IDW with two different elevation adjustments, while the 4- and 1.3-km fields are downscaled without elevation adjustment.

Both WaSiM and WATFLOOD model parameters have been optimized using the Dynamically Dimensioned Search (DDS) algorithm (Tolson and Shoemaker, 2007; Graeff et al., 2012; Francke, 2012). Optimization of each model was done using three different objective functions: the mean absolute error (MAE) of simulated inflow, to minimize overall errors; Nash-Sutcliffe Efficiency (NSE; Nash and Sutcliffe, 1970) of inflow, to emphasize performance during high-flow events; and the NSE of log-transformed flows, to optimize performance during low-flow periods. These different parameterizations attempt to sample the uncertainty in the hydrologic models' parameter values. Simulations during the ten-year optimization period (1997–2007) were driven by observed meteorological conditions at several weather stations within the case-study watershed and surrounding area (Figure 2.1).

The multi-state or multi-initial-condition component of the M2M ensemble forecasting system arises as a direct consequence of implementing a multi-parameter component. In forecast mode, the hydrologic state for each model and each model parameterization is updated at the start of the forecast day by driving the model with observed meteorological data. This resulting simulated state is used as the initial condition for the day's forecast run. In order to avoid discontinuities early in the daily forecast cycle, the parameter set used in updating the hydrologic state must match that used in the forecast. Thus, each parameter set has its own hydrologic state for each model, resulting in the creation of six different hydrologic states each day. While each hydrologic model/parameterization is initialized from a deterministic hydrologic state, these initial conditions still provide a small sampling of the hydrologic state uncertainty space. Figure 4.1 illustrates the update/forecast process for a particular parameterization of the WaSiM model. The forecast workflow is indicated by the solid arrows. Dashed arrows illustrate how meteorological observations are used to update the model configuration's hydrologic state for the following day's forecasts. This process is repeated for each watershed model (WaSiM and WATFLOOD) and each parameterization/state, yielding 72 unique inflow forecasts each day.

During the 731-day evaluation period, ensemble forecasts were issued every day for forecast days 1 and 2, while day 3 forecasts were issued on 729 days. Due to NWP model failures, the size of the ensemble forecast issued each day is variable: the day 1 forecasts consisted of a full 72-member ensemble on 456 days, while the day 2 forecasts were complete on 446 days. In the majority of cases, the number of missing day 1 and 2 ensemble members was small (3-6 missing members). The smallest ensemble size for forecast days 1 and 2 during the case-study period is 39 members and occurred on 2 forecast days. Day 3 NWP failures are more common, as model instabilities can result in shortened forecast lead time. A full day 3 ensemble forecast was issued on 684 of the 731 case-study days. There were 12 forecast days when the day 3 ensemble was less than half of its intended size. Probabilistic forecasts are issued regardless of ensemble size.

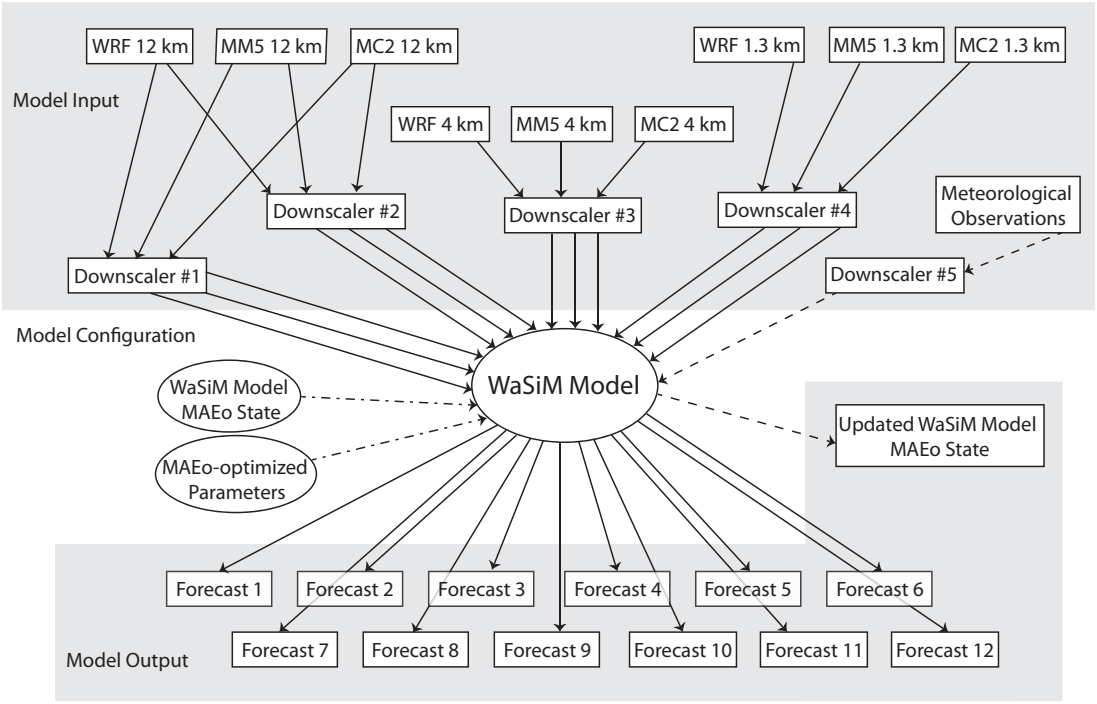

Figure 4.1: The flow of information into and out of the WaSiM model for generating forecasts with the MAE-optimized parameter set. The forecast workflow is indicated by the solid arrows. Dashed arrows illustrate how meteorological observations are used to update the model configuration's hydrologic state for the following day's forecasts. The model configuration is specified by the dash-dotted arrows.

4.2.3 A COmmunity Modular Post-processing System (COMPS)

The COmmunity Modular Post-processing System (COMPS) breaks down the process of generating calibrated probabilistic forecasts into a series of steps referred to as components. As implemented by Nipen (2012), COMPS contains components for bias correction, uncertainty modelling, probability calibration, forecast updating (not applied in this study), and verification. The input to the system is a set of predictors: ensemble forecasts of, for example, weather or hydrologic variables at a specific geographical location. The COMPS user selects the schemes to implement for each desired component, creating a specific configuration. COMPS can also be used to generate post-processed deterministic forecasts by bypassing the uncertainty and calibration components (bypass schemes exist for each component).

Each component scheme relies on model parameters that evolve over time. Consider for example a simple degree-of-mass-balance [DMB; Eq. (A.1)] bias correction scheme. DMB values less than one indicate that inflows are underforecast, while DMB greater than one indicates an

overforecasting bias. Using a moving window approach, each day, the previous N days of forecast-observation pairs are retrieved in order to calculate the bias correction factor DMB_N. Today's bias-corrected forecast is generated by dividing the raw forecast by this value. COMPS maintains computational efficiency by requiring the parameters of its various component models to be computed adaptively rather than over a moving window. Thus, only the last estimate of the parameter value must be retrieved each day, along with the new forecast-observation pair to update the parameter for the next forecast cycle.

Let an ensemble of K raw inflow forecasts be denoted as $\xi_{t,k}$, where t is a particular time and k is an index between 1 and K. The verifying observation at time t is x_t. An adaptive calculation of the DMB correction factor for ensemble member k is then given by:

$$DMB_{t+1,k} = \frac{\tau - 1}{\tau} DMB_{t,k} + \frac{1}{\tau}\left(\frac{\xi_{t,k}}{x_t}\right) \tag{4.1}$$

where τ is a unitless time scale that describes how quickly the impact of new information ($\xi_{t,k}/x_t$) diminishes over time. Recent information is weighted more heavily; older information ($DMB_{t,k}$) is never forgotten by the adaptive scheme but becomes less important with time. While τ is necessarily unitless, for a daily adaptive update it can be interpreted as an e-folding time in days.

We have implemented this scheme in the COMPS framework; results of testing a range of dimensionless time scales (τ) against the moving-window DMB calculation described in Chapter 2 are given in Appendix B. An adaptive DMB bias corrector with $\tau = 3.0$ was found to be effective at removing bias for forecast horizons of 1-3 days. Inflow forecast bias is strongly controlled by bias in the hydrologic states from which each day's forecast is begun. This, coupled with the flashy, mountainous nature of the study watershed, suggests that only very recent errors are likely to aid in bias correction, and explains why such a short e-folding time is so effective. Other components of the COMPS system used in this study are described in Sections 4.3.1 and 4.3.3.

4.3 From Ensembles to Calibrated Probability Forecasts

The first step in generating a probabilistic forecast of inflows to the Daisy Lake reservoir from the M2M ensemble is choosing a suitable uncertainty model. As indicated by the rank histograms in Figure 4.2, the M2M ensemble is underdispersive. That is, observations often fall outside of the predicted range of inflows (interpretation of rank histograms is described in Appendix A). This result implies that in spite of the M2M ensemble system explicitly attempting to sample all sources of error in the modelling chain, the amount of uncertainty captured by it is often inadequate. This is a common problem in both weather and hydrologic ensembles (e.g., Eckel and Walters, 1998;

Buizza, 1997; Wilson et al., 2007; Olsson and Lindström, 2008; Wood and Schaake, 2008). We suspect that a more complete handling of parameter and, in particular, initial condition uncertainty would improve this characteristic of the M2M ensemble. Dual state-parameter estimation frameworks that incorporate data assimilation could be used to accomplish this goal, though the paucity of observed hydrologic state data within the case study watershed confounds their use, and such methods are computationally demanding (Moradkhani et al., 2005a; DeChant and Moradkhani, 2011b; Leisenring and Moradkhani, 2011).

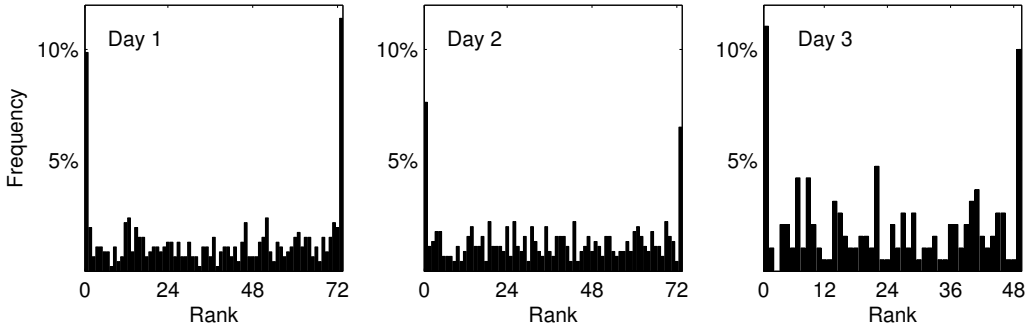

Figure 4.2: Rank histograms for the M2M ensemble forecasts at lead times of 1–3 days. The ensemble forecasting system is underdispersive for all forecast horizons as indicated by the large percentage of observations that fall outside the range of the ensemble.

In order to correct this deficiency, uncertainty models can be used to fit a probability distribution function (PDF) to the ensemble, whereby the parameters describing the spread of the distribution are estimated based on statistical properties of the ensemble and the verifying observations. In this way, it is possible to implicitly account for any uncertainty that is neglected or underestimated by the ensemble. The shape of the PDF fitted to the ensemble should correspond to the shape of the empirical distribution of the bias-corrected M2M ensemble mean forecast errors (because we plan to centre the distribution on the bias-corrected M2M mean). Hydrologic variables and their errors are often described as being non-normally distributed, and are therefore transformed into a space in which the errors become normally distributed, and the transformed variable can be modelled using a simple Gaussian PDF (e.g., Duan et al., 2007; Reggiani et al., 2009; Wang et al., 2009). The log-normal distribution, which amounts to fitting a Gaussian distribution to log-transformed data, has a long history of use in hydrology, and is still popular today (e.g., Chow, 1954; Stedinger, 1980; Lewis et al., 2000; Steinschneider and Brown, 2011). This distribution is particularly well-suited to streamflow and inflow forecasting, as it only assigns probabilities to positive forecast values.

Observed daily inflows at Daisy Lake exhibit a bimodal distribution, with storm season flows forming a skewed distribution at low flow values, and warm season flows forming a second peak at

higher flows. For this reason, forecast errors are analyzed by season. Figure 4.3 shows the distribution of M2M ensemble mean forecast errors during the 2009–2010 storm season (October through April), warm season (May through September) and full water year before and after log transformation. The full 72-member bias-corrected ensemble described in Section 4.2.2 has been used to calculate the ensemble mean. Log-transformed errors are calculated by taking the natural logarithm of the forecasts and the observations prior to calculating the error ($observed - forecast$). A Gaussian distribution is plotted on each empirical distribution, centred over the mean forecast error with the standard deviation given by that of the errors. Despite the small sample size (358 forecast-observation pairs for the full water year and 206 and 152 for the storm season and warm season, respectively), we can draw some useful conclusions about the distribution of M2M ensemble forecast errors. The non-transformed forecast errors comprise a slightly positively skewed distribution with the mean of the errors consistently greater than the median. The error distributions during the storm season are characterized by high peaks and long, narrow tails, and are therefore not well modelled by the normal distribution. Day 2 warm season errors do not exhibit skewness and appear to be well modelled by the superposed normal distribution. The log-transformed errors are likewise more normally distributed with much smoother peaks than their raw counterparts.

Based on these results and on the above-cited literature, we will test the performance of two different uncertainty models for producing reliable inflow forecasts for Daisy Lake: a log-normal uncertainty model is expected to perform well during the storm season; the Gaussian shape of warm season forecast errors suggests that a non-transformed normal PDF may produce calibrated probability forecasts during this time. The spread of these distributions should be related to forecast skill; when the forecast is less skillful, the uncertainty (as represented by the spread of the PDF) is greater. Since the M2M ensemble is underdispersive (Figure 4.2), we expect that the distributional spread will be best represented by a combination of ensemble spread and information regarding recent errors (Nipen and Stull, 2011; Gneiting et al., 2005).

4.3.1 Uncertainty Modelling in the COMPS Framework

COMPS includes an uncertainty model scheme in which a Gaussian distribution $\mathcal{N}$ is fitted to the ensemble using the Ensemble Model Output Statistics (EMOS) method of Gneiting et al. (2005). This uncertainty model makes the assumption that the forecast errors are normally distributed. Incorporating the DMB bias correction scheme, which is applied individually to each of K ensemble member forecasts ($\xi_{t,k}$) at time t, the EMOS forecast PDF (f_t) is given by:

$$f_t \sim \mathcal{N}\left(\frac{1}{K}\sum_{k=1}^{K}\frac{\xi_{t,k}}{DMB_{t,k}}, a_T s_t^2 + b_T\right), \tag{4.2}$$

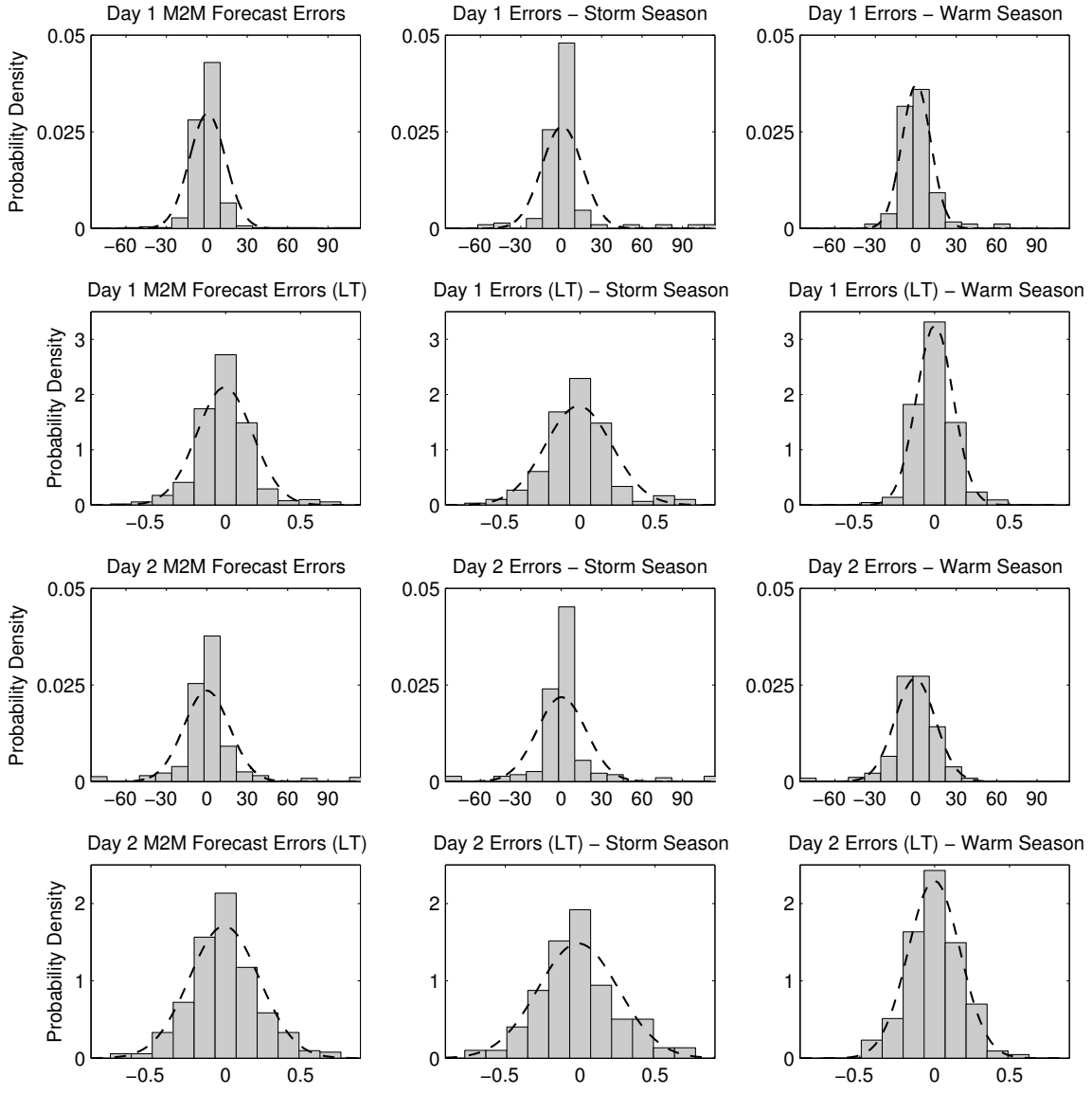

Figure 4.3: Empirical distributions of M2M ensemble mean forecast errors (m^3/s) for forecast days 1 and 2 during the 2009–2010 water year. Errors computed after a log transformation (LT) of forecasts and observations are generally more Gaussian, though the raw day 2 warm season forecast errors exhibit a Gaussian shape.

where s_t^2 is the ensemble variance. The first parameter of the Gaussian distribution is the bias-corrected ensemble mean, while the second represents the spread of the distribution and is determined by a least squares linear regression fit to the variance of the ensemble. The regression parameters a_T and b_T are determined based on past values of the square error of the bias-corrected ensemble mean during a training period (i.e., they describe the ensemble spread-skill relationship). Users of COMPS can also choose to have this scheme find a linear relationship between the forecast error and the mean of the ensemble, which has been shown to be a good predictor of error for precipitation (Hamill and Colucci, 1998).

We have modified the Gaussian EMOS scheme in COMPS to be able to fit a normal distribution to log-transformed data. This uncertainty model, which assumes forecast errors to be log-normally distributed, will be used to transform the M2M ensemble into a probabilistic forecast, and will be referred to as log-EMOS. We will also test this scheme without log transformation; we expect this uncertainty model (which we refer to simply as EMOS) to produce calibrated forecasts during the warm season when forecast errors exhibit a normal distribution. It is possible for the non-transformed uncertainty model to assign positive probabilities to negative inflow rates, which makes this model unsuitable for prediction during low-flow periods; this is not a concern during the warm season when snowmelt-driven inflows are relatively high.

In both the EMOS and log-EMOS schemes, the regression parameters in Eq. (4.2) are updated adaptively using a dimensionless timescale of $\tau = 30$. Nipen (2012) found this to be a suitable training period for various meteorological variables. While short training periods allow the uncertainty model to adapt quickly to changes in forecast regime or ensemble configuration, longer periods allow for a more robust estimation of the parameters. Gneiting et al. (2005) similarly found a moving window of 40 days to be a suitable compromise between these competing criteria.

An adaptive updating scheme was also implemented in COMPS for computing the weights in Bayesian Model Averaging (BMA), which can be used to produce calibrated probabilistic forecasts (Raftery et al., 2005). For various reasons, the method was found to be unsuitable for application to the M2M ensemble; results of testing the method are given in Appendix C.

4.3.2 Metrics of Probabilistic Forecast Quality

So long as the assumptions made by the uncertainty model hold true, it will produce calibrated probability forecasts. Probabilistic calibration, or reliability (Murphy, 1973) is a measure of consistency between forecast probabilities and the frequency of occurrence of observed values. That is, events forecasted with probability p should, over the course of many such forecasts, be observed to occur a fraction p of the time. This property is evaluated by visualizing the distribution of Probability Integral Transform (PIT) values (Gneiting et al., 2007) in a PIT histogram, which, for perfectly

calibrated forecasts, should be approximately flat. PIT values are given by:

$$P_t = F_t(x_t) \tag{4.3}$$

where x_t is the verifying observation at time t, and F_t is the corresponding forecast cumulative distribution function (CDF). The forecast CDF of variable x at time t is given by:

$$F_t(x) = \int_{-\infty}^{x} f_t(x)dx. \tag{4.4}$$

The calibration deviation metric D of Nipen and Stull (2011) provides a more objective measure of calibration (Appendix A). While calibration is a desirable characteristic of probabilistic forecasts, it is not an adequate measure of the usefulness of a forecast. Consider, for example, an uncertainty model that always issues a climatological forecast (i.e., the forecast PDF is always taken as the distribution of the climatological record). Assuming stationarity, such a forecasting system would be perfectly calibrated, but far too vague for decision making. Therefore, we will also require our forecast PDFs to concentrate probability in the correct area (i.e., near the verifying observation) on each day. This property can be measured by the ignorance score (Roulston and Smith, 2002). We also employ the Continuous Ranked Probability Score (CRPS), which addresses both calibration and sharpness (Gneiting et al., 2005, 2007). A description of verification metrics used in this chapter and their interpretation is given in Appendix A.

4.3.3 Probability Calibration Method

We have carefully selected candidate uncertainty models for the M2M ensemble forecasts of inflows to the Daisy Lake reservoir based on characteristics of the forecast errors. Namely, ensuring that the uncertainty models' assumptions (regarding how the ensemble and verifying observations are realized) are true at certain times of the year. During these times, the uncertainty model should be able to produce calibrated forecasts. However, at other times during the water year, or as evaluated over shorter time periods, these assumptions may be false, resulting in poorly calibrated forecasts. It is during these times that probability calibration can offer improvements to the probabilistic forecasting system.

The PIT-based probability calibration scheme implemented within COMPS is that described by Nipen and Stull (2011) with necessary modifications for adaptive parameter calculation (Nipen, 2012). Recall that a necessary condition for reliability is a flat PIT histogram. This is equivalent to requiring the cumulative distribution of PIT values to lie along the 1:1 line of PIT values vs. observed relative frequencies. By constructing an empirical cumulative distribution of PIT values

accumulated over a moving window of time points T, we can derive the PIT-based calibration function as:

$$\Phi_T(p) = \frac{1}{\|T\|} \sum_{t \in T} H(p - F_t(x_t)) \tag{4.5}$$

where the PIT value $F_t(x_t)$ is the forecast CDF value at the verifying observation x_t at a time t in the training set T, p is a probability value between 0 and 1, and H is the Heaviside function given in Eq. (A.17).

The probability calibrated CDF is then calculated by:

$$\hat{F}_t(x) = \Phi_T(F_t(x)), \tag{4.6}$$

which amounts to a relabelling of CDF values $F_t(x)$ to form a new distribution $\hat{F}_t(x)$. The corrected forecast PDF ($\hat{f}_t$) can be calculated by combining Eq. (4.4) and Eq. (4.6) and invoking the chain rule, yielding:

$$\hat{f}_t(x) = \Psi_T(F_t(x)) f_t(x), \tag{4.7}$$

where $\Psi_T(p)$ is defined as the derivative of the calibration function $\Phi_T(p)$ with respect to p, and serves as an amplification function to the raw PDF $f_t(x)$.

The calibration curve $\Phi_T(p)$ is generated by dividing the p interval [0,1] into any number of bins. Unless otherwise stated, we will use 10 bins and the individual PIT values along the calibration curve will be updated with a time scale of τ = 90. Note that using more bins requires a longer training period T in order to reduce the curve's sensitivity to sampling errors. Nipen and Stull (2011) found that the calibrator required on the order of 100 data points for optimal results (i.e., balancing the competing objectives of reducing sampling error and using recent data for training the calibrator). Note that using fewer bins would reduce sampling error, but that the calibration curve would be very coarse. Excluding the (constant) end points (0,0) and (1,1) of the calibration curve, these ten bins are defined by nine interior "smoothing points" (p,Φ_p). Modifying Eq. (4.5) for adaptive updating of these points yields:

$$\Phi_{p,t+1} = \frac{\tau - 1}{\tau} \Phi_{p,t} + \frac{1}{\tau} H(p - F_t(x_t)), \tag{4.8}$$

(Nipen, 2012).

The PIT-based calibration scheme as implemented in COMPS uses a monotonically increasing cubic spline to create a smooth $\Phi_T(p)$ curve with a continuous derivative to connect the smoothing points. This allows it to generate a smoothly varying adjusted PDF for calculating the ignorance score.

An important finding of Nipen and Stull (2011) is that applying the calibration during periods when the uncertainty model already produces calibrated forecasts actually degrades the reliability. This is due to sampling errors in the PIT histogram used to generate the calibration curve (Brocker and Smith, 2007; Pinson et al., 2010). In such cases, the probability forecast is best left unadjusted. An option for "intelligent calibration" (*inteliCal*) has therefore been added to the COMPS PIT-based calibration scheme. InteliCal determines whether or not the calibration should be applied by comparing the calibration deviation metric, D [Eq. (A.8)] to the value of calibration deviation expected for perfectly calibrated forecasts (caused by sampling error), given by $E[D_p]$ [Eq. (A.9)] (Nipen and Stull, 2011). When D is sufficiently greater than $E[D_p]$, we anticipate that the forecast will benefit from calibration.

The calibration deviation metric D in Eq. (A.8) is calculated using bin counts computed over a moving window of length $\|T\|$. In the adaptive updating framework of COMPS, the only change to the formulation of D is that the bin frequencies $b_i\|T\|^{-1}$ are updated adaptively. We also replace $\|T\|$ in Eq. (A.9) with the dimensionless time scale τ. To correct for scaling differences that occur with the replacement of the moving window update with the adaptive updating scheme, a factor of 1.5 is required when comparing D and $E[D_p]$. This conversion factor was determined through trial-and-error by comparing D computed using a moving window to that computed adaptively over the range $60 \leq \tau \leq 250$. InteliCal then applies the PIT-based calibration only when:

$$1.5D > ICF \times E[D_p]. \tag{4.9}$$

The inteliCal adjustment Factor (ICF) allows COMPS users to adjust the sensitivity of inteliCal if necessary. The default ICF is 1.0, but as the performance of the inteliCal scheme has not yet been thoroughly tested, the ideal ICF is not known. In this study, we attempt to determine a suitable value for the ICF, though results may be case-specific.

4.4 Results and Discussion

4.4.1 Performance of the Uncertainty Models

Figure 4.4 shows the uncalibrated PIT histograms for the 3-day EMOS uncertainty model forecasts made during the case-study storm seasons (first row), the warm seasons (second row), and for the full water years (bottom row). Calibration deviation (D) is shown on each histogram, with the expected deviation for a perfectly calibrated forecast (E[D_p]) also shown on the day 1 plots (E[D_p] does not change with lead time as it is only a function of the number of bins in the PIT histogram and the sample size). This figure clearly illustrates how selecting an inappropriate uncertainty model

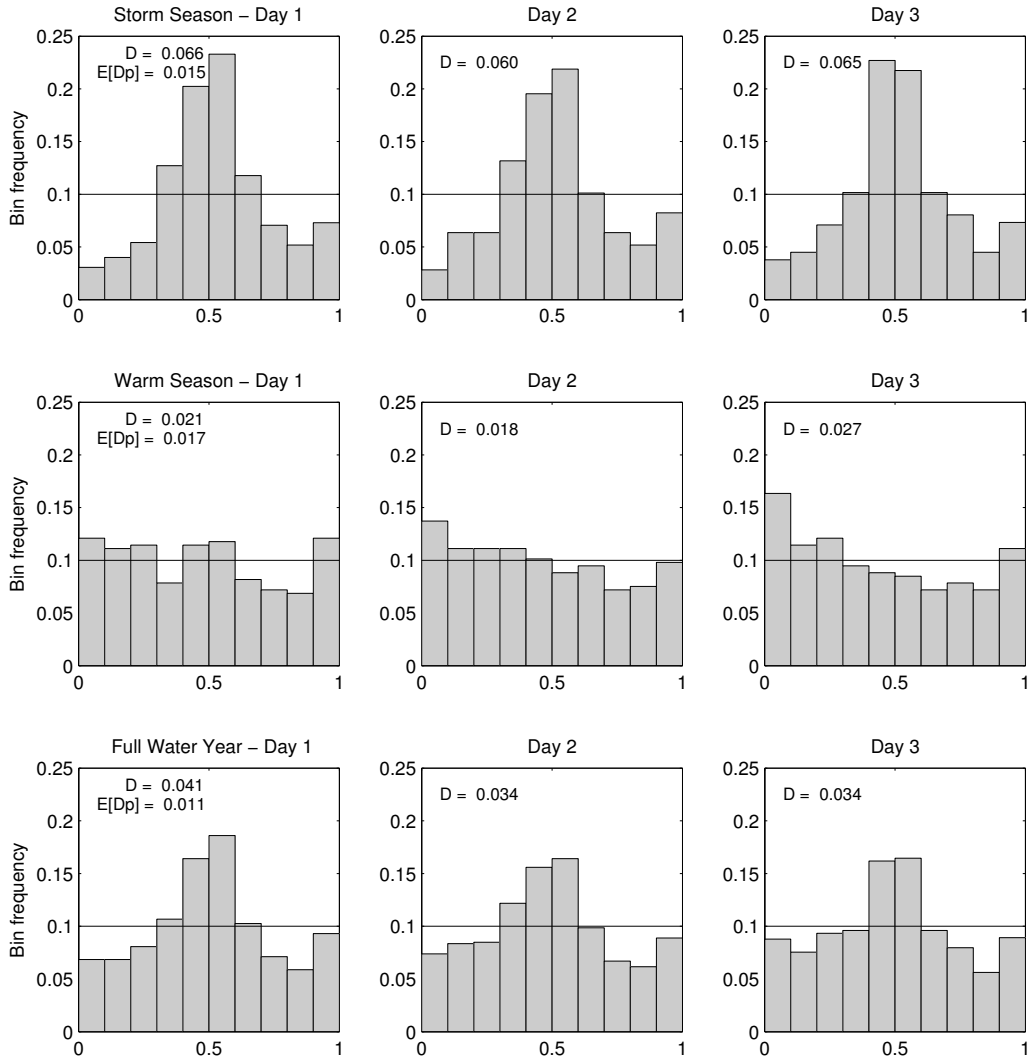

Figure 4.4: PIT histograms for the storm seasons (top row), warm seasons (middle row), and full water years (bottom row), pooled over the 2010–2011 and 2011–2012 water years. Results are for the uncalibrated EMOS uncertainty model. Calibration deviations D are shown for each histogram, with $E[D_p]$ for comparison. Flatter histograms and therefore lower D are preferred.

can yield highly uncalibrated results. The PIT histograms for the storm season show that the EMOS uncertainty model does not concentrate enough probability density at the centre of the distribution. This is readily anticipated given the empirical storm season error distribution in Figure 4.3. During the warm season, which exhibits a more normal distribution of errors, the EMOS uncertainty model

is able to produce nearly calibrated probabilistic forecasts. The importance of specifying a time period over which calibration is measured is evident in the PIT histograms for the full water year, which mask the excellent calibration during the warm season.

We have applied the log-EMOS uncertainty model with two different configurations: using ensemble variance as a predictor of the Gaussian spread (log-EMOS$_v$); and using ensemble mean as a predictor of this spread (log-EMOS$_m$). Precipitation uncertainty has been found to be better explained by ensemble mean than by measures of ensemble spread (Hamill and Colucci, 1998). Since reservoir inflows are so strongly influenced by precipitation, it was anticipated that storm season inflow uncertainty would likewise be better represented by the ensemble mean.

Figure 4.5 shows uncalibrated PIT histograms for forecast days 1 through 3 broken up by season for log-EMOS$_v$ forecasts. This uncertainty model is, as expected, superior to the EMOS model during the storm season when errors are log-normally distributed (Figure 4.3), but produces slightly less calibrated forecasts during the warm season. Using the log-EMOS$_m$ uncertainty model during the storm season results in more observations falling in the tails of the forecast distribution (Figure 4.6). This may be caused by the behaviour of inflows to the Daisy Lake reservoir during the fall and winter. During the storm season, observed and forecasted ensemble mean inflows can be quite low for several days or even weeks (due to dry weather patterns, or precipitation falling as snow), and can then increase very suddenly when a rain or rain-on-snow event occurs. The ensemble mean-as-spread uncertainty model will have difficulty training for these sudden changes. Also, ensemble mean forecast misses and false alarms will result in the distribution having spread completely unrelated to forecast skill.

The superior performance (relative to EMOS forecasts) of the log-EMOS$_v$ forecasts during the storm season is also reflected in this model's ignorance and continuous ranked probability scores (Figure 4.7). Ignorance scores for the EMOS model during the storm season are significantly worse due to the unsuitability of this model during periods when forecast errors are not normally distributed. Conversely, the EMOS model has slightly better ignorance scores during the warm season for days 1 and 2. The fact that these ignorance scores are still higher than those for the EMOS model during the storm season appears to be caused by the uncertainty model generating warm season forecast PDFs with large spread. The bias-corrected ensemble members themselves exhibit large spread, suggesting that this is a failure of both the M2M ensemble forecasting system and of the regression-based EMOS model as the forecast error characteristics change between seasons. The CRPS values in Figure 4.7 clearly show that the log-EMOS$_v$ model performs best during the storm season. The EMOS forecasts have the best day 1 CRPS, but log-EMOS$_v$ is better at longer lead times due to these forecasts having better sharpness.

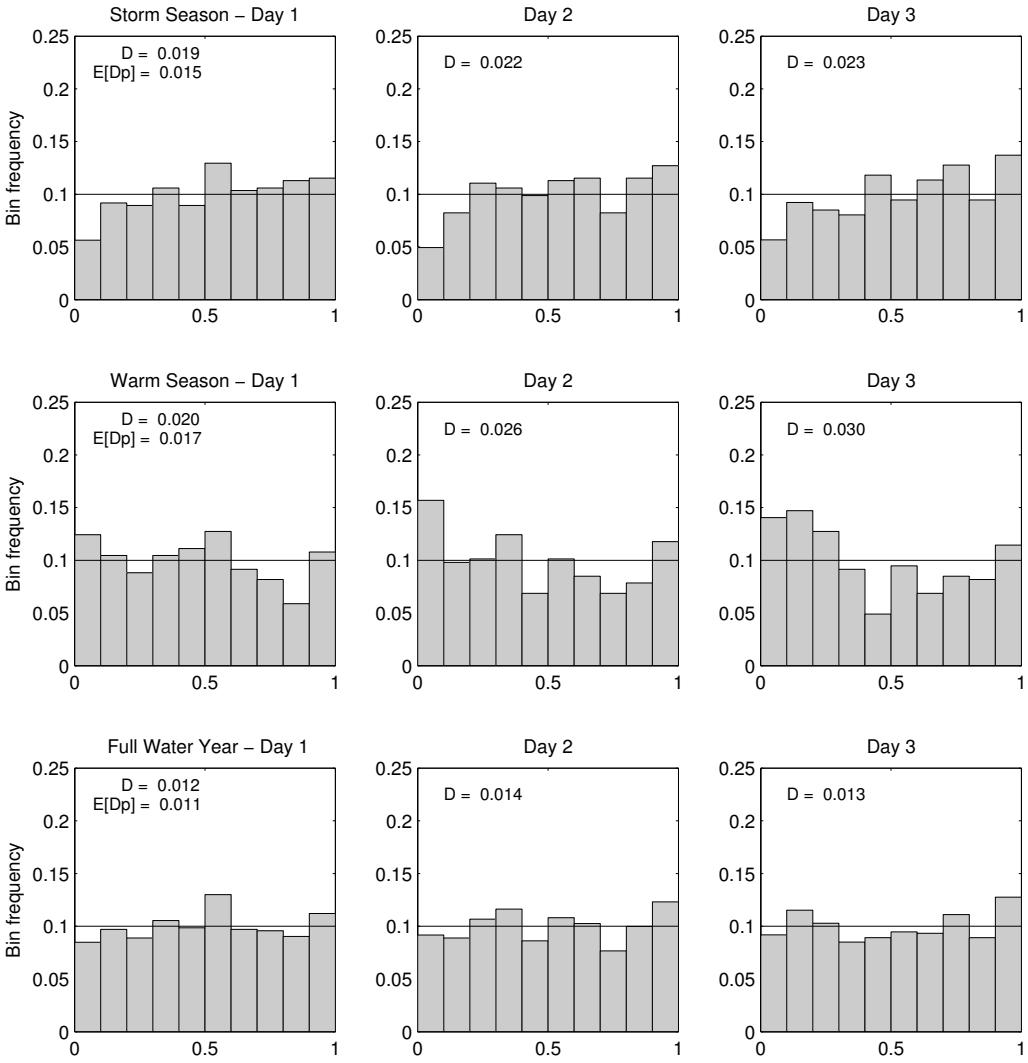

Figure 4.5: PIT histograms for the storm seasons (top row), warm seasons (middle row), and full water years (bottom row), pooled over the 2010–2011 and 2011–2012 water years. Results are for the uncalibrated log-EMOS$_v$ uncertainty model. Calibration deviations D are shown for each histogram, with $E[D_p]$ for comparison.

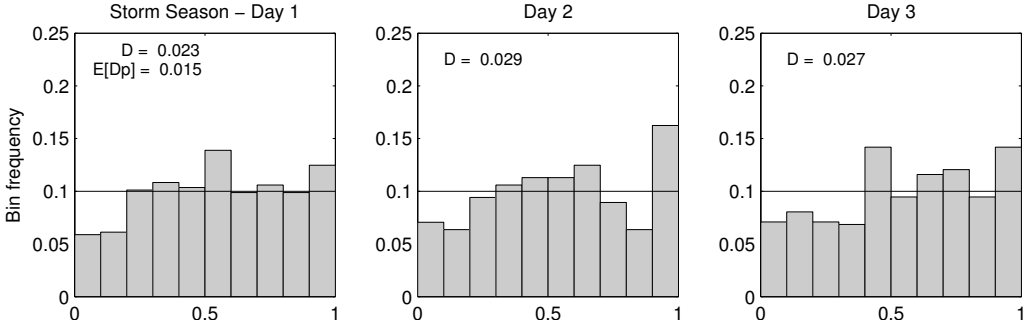

Figure 4.6: PIT histograms for all forecast horizons during the 2010–2011 and 2011–2012 storm seasons using the uncalibrated log-EMOS$_m$ uncertainty model.

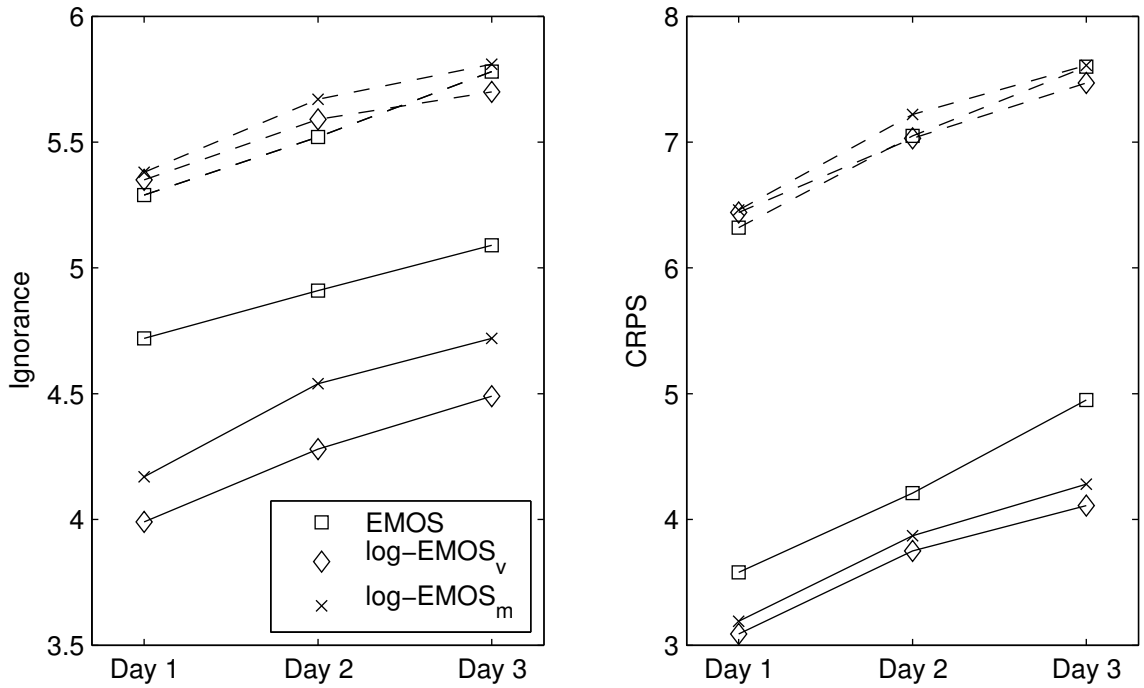

Figure 4.7: Ignorance and continuous ranked probability scores (CRPS) for the various uncertainty models tested. Forecasts are divided into storm season (solid lines) and warm season (dashed lines) for scoring, as each uncertainty model has different calibration characteristics during these times of year. Smaller ignorance scores and CRPS are preferred.

4.4.2 Effect of Probability Calibration

Figure 4.8 illustrates the power of the PIT-based probability calibration scheme (with the default ICF of 1.0). The method is able to correct for the EMOS uncertainty model's failed assumption of Gaussian forecast errors during the storm season. However, the dimensionless timescale of $\tau = 90$ applied here results in the calibration adjustments necessary during the storm season being propagated into the already well-calibrated warm season. Namely, the calibration has successfully adjusted the storm season predictive distributions to have higher peaks and thicker tails, but has carried this adjustment into the warm season, as indicated by these distributions now being underdispersive with too much probability density at the centre of the distribution. This problem does not occur during the transition from the warm season to the storm season. This is because by the end of the already well-calibrated warm season, the calibration deviation as measured by the adaptive scheme is very small, and inteliCal does not apply any calibration to the forecasts.

The PIT histograms for probability calibrated log-EMOS$_v$ forecasts (with an ICF of 1.0) are shown in Figure 4.9. The deterioration in warm season calibration deviation may again be caused by the lengthy learning period of the scheme and the fact that the raw storm season and warm season PIT histograms exhibit different distributional biases; there is a slight underforecasting bias during the former period, and a tendency to overforecast in the latter. A more likely explanation is that the inteliCal scheme is applying the calibration correction too often, resulting in the introduction of additional calibration deviation caused by sampling error. The raw log-EMOS$_v$ calibration deviations D are not significantly different from $E[D_p]$; these forecasts (particularly during the storm season) may therefore be considered calibrated to within sampling error.

While storm season calibration is improved by the PIT-based calibration scheme in both uncertainty models, the ignorance scores indicate that calibration is somehow shifting the highest concentration of probability in the forecast PDF away from the verifying observation (Figure 4.10). Examination of forecast CDFs on a handful of forecast days (not shown) reveals that while calibration of the EMOS uncertainty model can yield excellent results with respect to calibration deviation during the storm season, it does so by shifting the forecast PDF such that the verifying observation falls nearer to the tails of the distribution. Warm season EMOS ignorance scores are only worse during the (lengthy) period when the calibrator is adjusting to the new regime. The opposite is true for the log-EMOS$_v$ model, where the ignorance scores plotted in Figure 4.10 are similar or better for the probability calibrated forecasts during the storm season, but are consistently higher during the warm season where the calibration is having to do the most adjustments. Calibration of the EMOS forecasts results in improved CRPS during the storm season because these forecasts have more probability mass near the centre of the distribution and are therefore sharper. EMOS warm

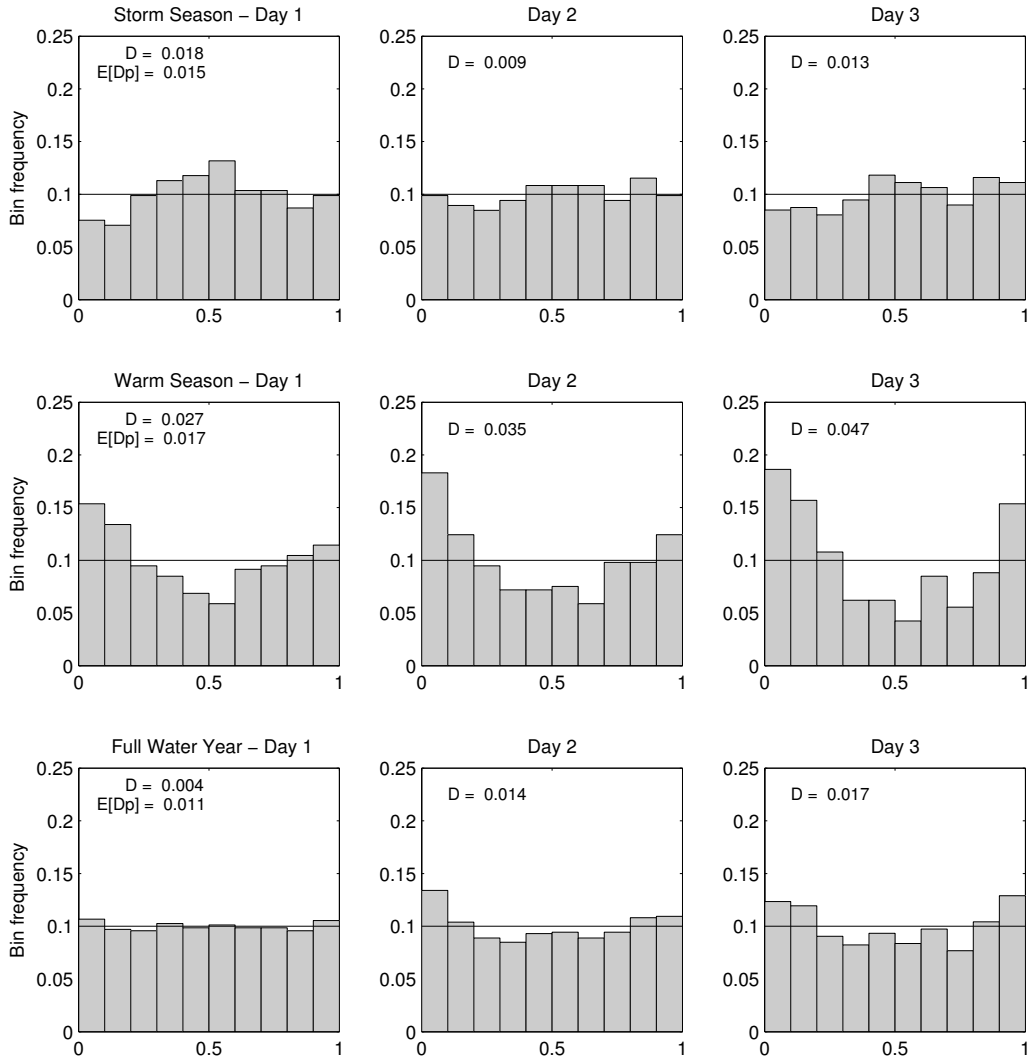

Figure 4.8: PIT histograms for EMOS uncertainty model forecasts as in Figure 4.4, but following PIT-based probability calibration with nine smoothing points and $\tau = 90$.

season scores deteriorate as a result of the introduction of significant calibration deviation. The increased calibration deviation of log-EMOS$_v$ forecasts during the warm season contributes to the increased CRPS during this season. During the storm season, CRPS for this model does not change significantly after application of the PIT-based calibrator.

Nipen (2012) derived a decomposition of the ignorance score for a set of raw forecasts into two parts: (1) the potential ignorance score of a perfectly calibrated forecast (IGN$_{pot}$), and (2)

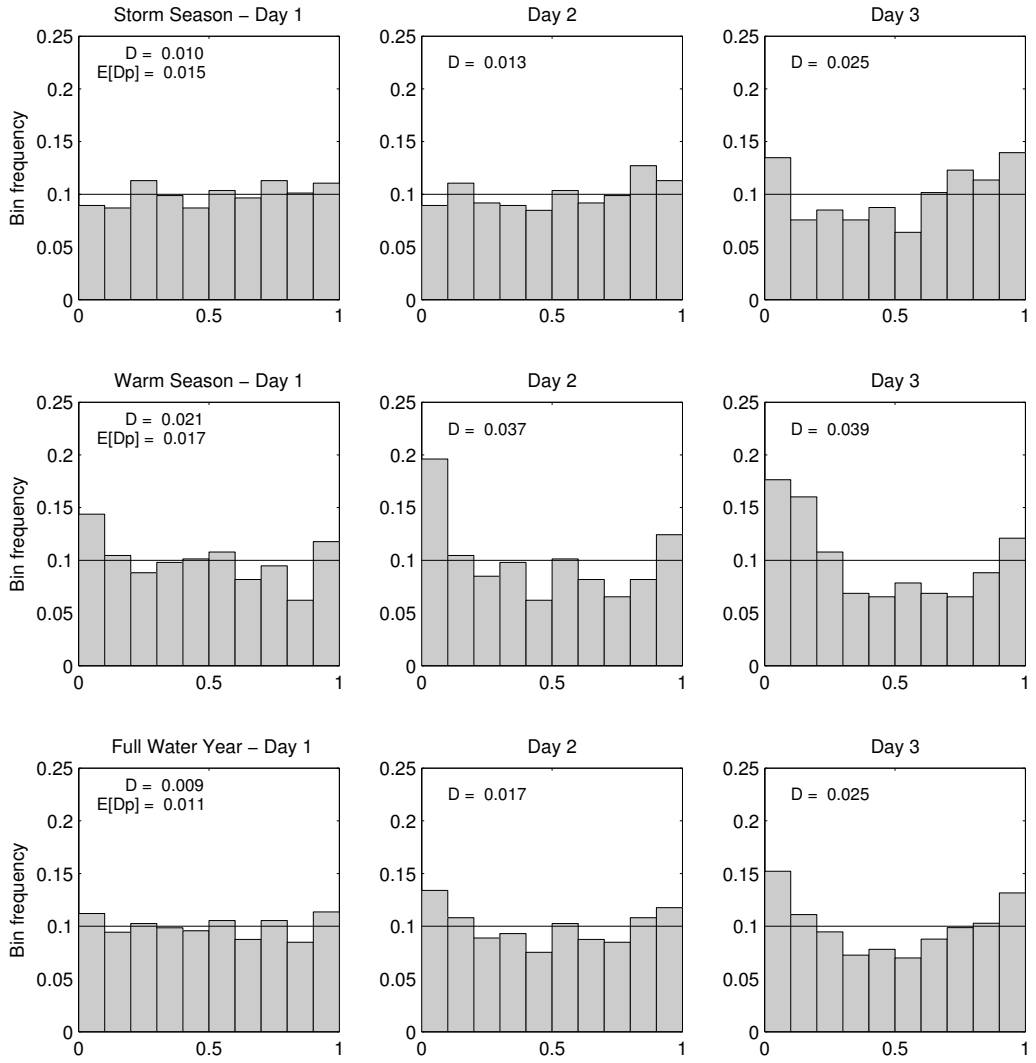

Figure 4.9: PIT histograms for log-EMOS$_v$ uncertainty model forecasts as in Figure 4.5, but following PIT-based probability calibration with nine smoothing points and $\tau = 90$.

extra ignorance caused by a lack of calibration (IGN$_{uncal}$). Ignorance can therefore be reduced by improving the ensemble forecasting system, applying bias correction, or using a more suitable uncertainty model to reduce IGN$_{pot}$, or by calibrating the forecast to reduce IGN$_{uncal}$. In our comparison of raw and probability calibrated EMOS and log-EMOS$_v$ forecasts, the bias correction and uncertainty model schemes have not undergone any changes. Therefore changes in ignorance scores can be attributed to changes in IGN$_{uncal}$. This suggests that the increased ignorance exhibited

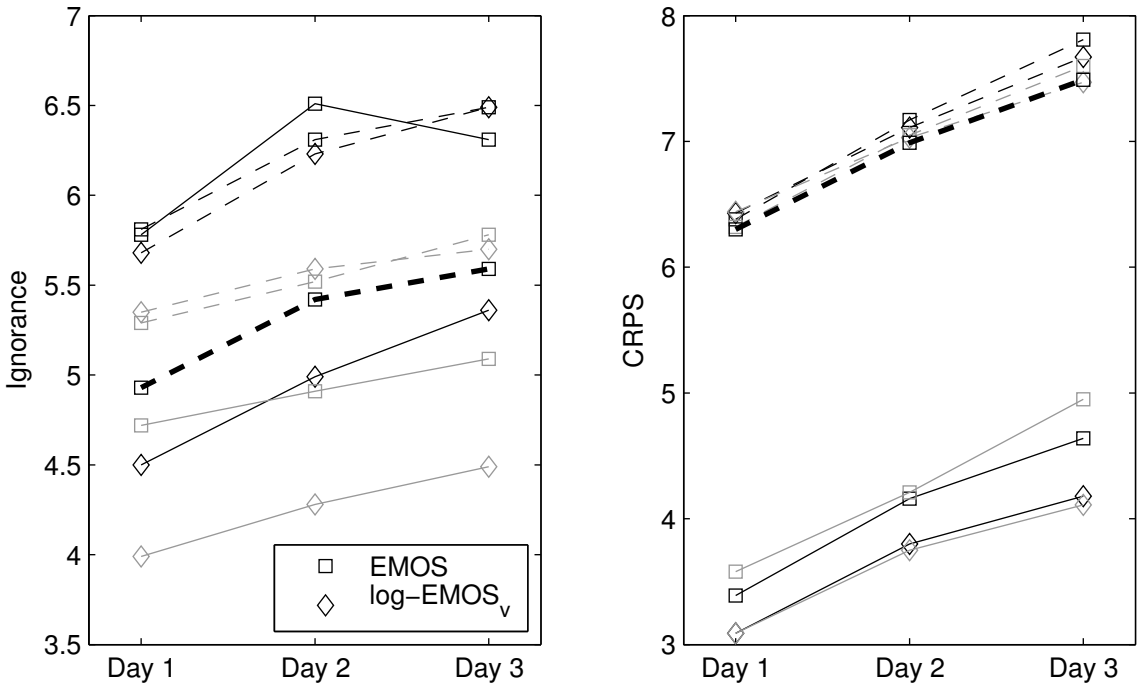

Figure 4.10: Same as Figure 4.7, but scores are computed after applying the PIT-based calibration scheme (black). Uncalibrated results (grey) are plotted for comparison. Results for warm season EMOS forecasts probability calibrated using the 'carry-forward' method are indicated by the heavy dashed line.

by the calibrated log-EMOS$_v$ forecasting system can be attributed to overfitting of the calibration curve to sampling errors. While overfitting may play a role in the deterioration of EMOS uncertainty model ignorance after calibration, the main problem in this case is the long lag-time in updating the PIT histogram when the forecasting system's error characteristics transition between seasons. Note that sampling error is likely less significant in verification than it is in calibration, as the verification sample sizes are larger.

The adjustment factor (ICF) in Eq. (4.9) allows users of COMPS to adjust the inteliCal sensitivity. Since the probability calibrated log-EMOS$_v$ forecasts made with the default ICF of 1.0 exhibit signs of overfitting, we may expect better results with a larger ICF. We tested inteliCal calibration of the storm season log-EMOS$_v$ forecasts with ICF ranging from 1.1 to 2.0. This experiment revealed that the most improvement to calibration deviation, which occurs for ICF in the range of 1.0 to 1.33, is accompanied by an increase in ignorance. Higher values of ICF prevent the ignorance score from being inflated due to the introduction of sampling error in the calibration, but this is achieved by applying the calibration correction sparingly. The fact that the log-EMOS$_v$

storm season forecast ignorance scores are lowest without any probability cailbration supports the earlier suggestion that these forecasts can be considered to be calibrated (within sampling error).

While we can likewise consider the EMOS warm season forecasts to be calibrated, the drastic change in shape of the EMOS uncertainty model's raw (uncalibrated) PIT histograms between seasons suggests an alternative calibration strategy for improving the sharpness of EMOS forecasts. Indeed, even with an ICF of 2.0, the calibrated warm season EMOS PIT histograms exhibit the 'U' shape evident in Figure 4.8. This confirms that the problem is not caused by overfitting to sampling error in the calibration curve, but rather by the long lag-time in updating its shape. To avoid the adaptive calibration scheme's long lag-time in generating representative calibration curves, we replaced the calibration parameters at the start of the warm season (taken to be May 1) with those valid at some time during the previous year's warm season. July 29 was selected as the replacement date based on calibration statistics from the 2009–2010 water year. By this date, the PIT histogram is able to reflect the (well-calibrated) characteristics of the EMOS warm season probability forecasts. Note that the choice of May 1 for the start of the warm season is based solely on examination of climatological inflows (i.e., it is the approximate start of the rising limb of the climatological freshet). Whether this date coincides with the start of the snowmelt season in any given year is not known ahead of time.

This calibration strategy, which we refer to as *carry-forward* (*CF*) calibration, resulted in significant improvements to warm season forecasts derived from the EMOS uncertainty model (using the default ICF of 1.0). The calibration deviation for day 1 forecasts dropped to 0.011, while deviations for days 2 and 3 dropped to 0.015 and 0.021 respectively. Additionally, ignorance scores for these CF-calibrated forecasts were greatly improved as shown in Figure 4.10 (heavy dashed line). CRPS also improved as a result of forecasts becoming sharper after calibration. A comparison of raw and CF-calibrated distributional spread (not shown) reveals that the calibration scheme reduces spread, particularly early in the warm season. As the warm season progresses, the change in spread is reduced, and toward the end of the warm season, the calibration scheme slightly increases the distributional spread for forecast horizons of 2–3 days. This supports the suggestion in Section 4.4.1 that the regression-based distributional spread fitting has difficulty during the transition period between the storm season and warm season.

Examination of forecast PDFs for a handful of dates throughout the warm season reveals that IGN and CRPS improvements are due to the amplification function [Eq. (4.7)] increasing the height of the forecast PDF near the centre, and reducing the height in the tails (i.e., reducing the spread). Results from the CF-calibration were found to be insensitive to increases in ICF up to 1.43. For ICF values greater than this, the calibration is applied too sparingly and ignorance scores show very little improvement. In terms of both calibration deviation and ignorance, an ICF of 1.0 gives

the best results in the carry-forward calibration framework.

Based on these results, the ideal M2M-based probability forecasting system for Daisy Lake inflows is a combination of two different COMPS configurations: (1) the raw (uncalibrated) log-EMOS$_v$ forecasts during the storm season; and (2) the carry-forward-calibrated EMOS forecasts during the warm season. Figure 4.11 shows the PIT histogram resulting from pooling the forecasts from the these two model configurations over the 2010–2011 and 2011–2012 water years. The corresponding ignorance scores for forecast days 1, 2 and 3 are 4.38, 4.76 and 4.95, respectively, while CRPS values are 4.44, 5.11 and 5.53.

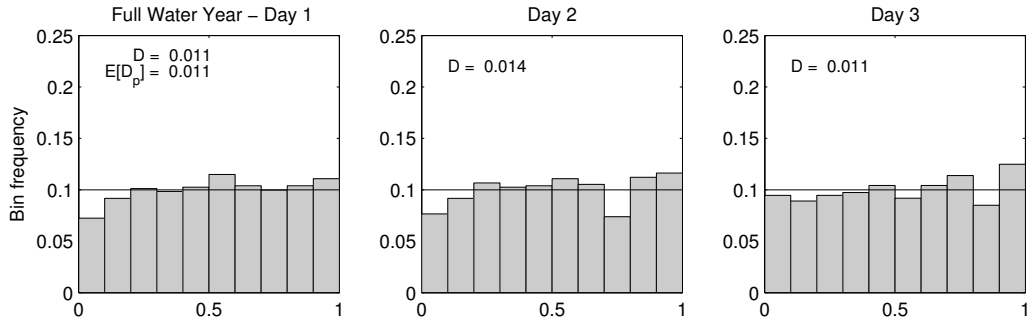

Figure 4.11: PIT histogram for full water years after combining raw (no PIT-based calibration applied) storm season forecasts from the log-EMOS$_v$ uncertainty model with carry-forward-calibrated EMOS forecasts during the warm season for ideal forecast reliability and sharpness.

4.5 Concluding Remarks

In this chapter, we have transformed a 72-member ensemble forecasting system that explicitly samples all sources of error in the inflow modelling chain into a calibrated probabilistic forecasting system. This work was done exclusively using the COmmunity Modular Post-processing System (COMPS) described and developed by Nipen (2012). COMPS allows its users to implement and apply schemes for bias correction, uncertainty modelling, probability calibration, forecast updating using recent observations, and verification. Any of these components can alternatively be bypassed, making COMPS a flexible post-processing tool for point forecasts of almost any observed phenomenon.

An analysis of inflow forecast error characteristics at the Daisy Lake Reservoir enabled us to implement and apply COMPS uncertainty models appropriate at different times of year. During the storm season, a log-normal uncertainty model fit to the M2M ensemble using EMOS yields

reliable or calibrated forecasts; a simple normal EMOS distribution yields calibrated results during the warm season when errors are normally distributed.

The PIT-based calibration scheme of Nipen and Stull (2011) was generally found to improve calibration at the expense of forecast ignorance. In the case of the well-calibrated log-EMOS$_v$ uncertainty model, this is caused by an overfitting of the calibration curve to sampling errors. Seasonal changes in PIT histogram shape for the EMOS uncertainty model caused continuous updating of calibration curve parameters to produce poorly calibrated forecasts during the warm season. This is because of the long lag-time in adaptively updating the calibration curve. By replacing these calibration parameters at the start of the warm season with those valid late in the previous year's warm season (a process referred to as carry-forward calibration), we were able to produce sharper forecasts with slightly less calibration deviation, and greatly reduced ignorance and CRPS.

The ideal approach to probabilistic forecasting of inflows to the case-study watershed is therefore a combination of two different configurations: raw (not calibrated) log-normal EMOS uncertainty model forecasts during the storm season (October through April), and Gaussian EMOS uncertainty model forecasts with carry-forward calibration (with an ICF of 1.0) during the warm season (May through September). This combined configuration is easily achieved in an operational forecast setting, whereby both uncertainty models are run continuously throughout the year, and the forecast output from the COMPS system is switched at pre-determined dates, or, alternatively, when the observed flow characteristics begin to transition. Testing of the newly-implemented inteliCal calibration scheme in COMPS indicates that a suitable value for the inteliCal adjustment factor, ICF, may be in the range of 1.43 to 1.67. Whether these results are specific to the case study data is unknown; future work should include further testing of the inteliCal scheme.

While the methods applied and results shown in this chapter are specific to the case-study watershed, there are some general lessons that can be applied in other studies. First and foremost, an analysis of forecast error characteristics goes a long way in determining the ideal uncertainty model. We have shown that when the uncertainty model makes correct assumptions about how forecast errors are distributed around the ensemble mean, probabilistic forecasts derived from the model are reliable or very nearly so. We have also shown that error characteristics can be strongly regime-dependent. Thus, applications of probabilistic forecasting methods in watersheds with distinct seasonality (for example, a rainy season and a snowmelt-driven season) may benefit from the use of different uncertainty models at different times of year.

Chapter 5

On the Importance of Sampling Hydrologic Uncertainty: An Economic Analysis

5.1 Introduction

Deterministic forecasts can give forecast users a false impression of certainty. In risk-based decision making, deterministic forecast failures can lead to significant economic and societal losses (e.g., Glassheim, 1997). Forecasts expressed in terms of reliable probabilities of a range of possible events can enable rational decision making and provide economic benefits to the decision maker and to society as a whole (Krzysztofowicz, 2001). Roulston et al. (2006) have shown that when provided with weather forecasts and quantitative estimates of forecast uncertainty, even nonspecialists are able to make decisions that increase economic reward and reduce exposure to risk. Research has repeatedly illustrated that, over a range of time scales, even imperfect probabilistic weather and hydrologic forecasts are able to provide positive economic value to a wider range of users than deterministic forecasts and that for most users reliable probability forecasts provide increased economic value (e.g., Richardson, 2000; Zhu et al., 2002; Palmer, 2002; Stensrud and Yussouf, 2003; Roulin, 2007; McCollor and Stull, 2008b).

The Member-to-Member (M2M) ensemble evaluated in this chapter consists of various components, each sampling a different source of uncertainty in the hydrologic modelling chain. The M2M forecasting system includes multiple Numerical Weather Prediction (NWP) models and multiple (nested) NWP grids that are downscaled using multiple interpolation schemes to drive multiple Distributed Hydrologic (DH) models. Each DH model has multiple model parameterizations and uses multiple hydrologic states to begin each daily forecast. Each of these components comes at a price, whether measured in terms of money, hours worked, or computational costs.

Many gridded NWP model output fields are freely available from national forecast centres

such as the U.S. National Centers for Environmental Prediction, and the Meteorological Service of Canada, though there are computational costs associated with handling these large data sets. Other, generally higher resolution or custom products may be obtained through a contract with a private or academic weather modelling group, or by purchasing expensive high-performance computers and hiring IT staff to make in-house NWP forecasts. Many hydrologic models are freely available, but support may not be, and setup of these models for a specific watershed can be a time-consuming, and therefore costly undertaking. Automated model parameter optimization schemes require some manual setup time, but can generally be left to run for the days or weeks required for tuning, so long as the computing resources are available. A multi-parameter ensemble may necessitate a multi-state component to avoid forecast discontinuities when model parameters change suddenly between model runs. The multi-state component is then essentially free, aside from costs associated with additional model run-time, which could be significant.

Since the sum total of the price paid for the full M2M ensemble may be prohibitive for some operational forecasting applications, it is prudent to evaluate the economic value of each M2M component. If the price paid for each component is known, such an analysis can be used to determine whether or not they are cost-effective. Murphy (1993) identified three types of forecast "goodness": consistency (i.e., between a forecaster's best judgement and the actual forecast), quality, and value. Value, which is concerned with economic worth to the forecast end user, is the focus of this chapter.

In this study, the full 72-member M2M (hereafter identified as 'Full') ensemble is reduced by eliminating various ensemble components. Each reduced ensemble forecast is transformed into a probabilistic forecast using uncertainty models that fit a probability distribution to the ensemble. The relative economic values of the probabilistic forecasts from the reduced M2M configurations are then compared to those from the Full M2M system to ascertain the value added by each component. Any sources of uncertainty that are not explicitly or adequately sampled by the ensemble may be implicitly accounted for by the uncertainty model or by subsequent probability calibration. Using this strategy, it may be possible to reduce the ensemble setup cost and computational complexity (and therefore operationally critical forecast run-time) while continuing to generate reliable probabilistic forecasts.

Economic value of the probabilistic forecasts used in this study is estimated based on costs and losses associated with operating the reservoir under the guidance of each ensemble configuration. This value provided to the forecast end user does not include any reductions in value due to computational, time, or monetary costs associated with the various ensemble components. It is therefore up to the individual forecast end user to weigh the value of each ensemble component against the price paid for that component.

5.2 Economic Value of Forecasts

The economic value of reservoir inflow forecasts is controlled by complex interactions between forecast quality, reservoir operation constraints, transmission constraints, demand for electricity, and the highly variable electricity market, among other factors. Dynamic economic models that seek to capture these processes have been developed for specific regions or markets, and are typically used for determining economically optimal strategies for water management.

For example, energy production in the Columbia River system, which is controlled by a number of major storage and run-of-river dams in southwestern Canada and the United States, is modelled by ColSim (Hamlet et al., 2002; Hamlet and Lettenmaier, 1999). ColSim handles a variety of competing system objectives such as hydropower, flood control, flow targets and recreational constraints. The CALVIN (California Value Integrated Network) model (Jenkins et al., 2001; Draper et al., 2003; Pulido-Velazquez et al., 2004) similarly balances different objectives for optimal operation of California's major water supply system.

The Short-Term Optimization Model (STOM) developed by Shawwash (2000) for the British Columbia Hydro and Power Authority (BC Hydro) focuses on operations planning that optimizes hydroelectric resource utilization and trade opportunities at time scales of one day to a week for the entire BC Hydro generating system. STOM determines the optimal tradeoff between the long-term value of water and the returns from spot trading transactions in a competitive electricity market. Operating decisions are driven by the need to meet system electricity demand while meeting other requirements and constraints that are often in competition with the main objective. Other decision support tools have been developed for BC Hydro operations planning at longer timescales for specific power complexes (Druce, 1990) and for system-wide management (Fane, 2003). Although models such as ColSim, CALVIN and STOM allow a thorough, realistic examination of energy production and revenues for different operating strategies (whether driven by weather forecasts or changes operating on longer time scales), their use also entails "enormous data requirements" (Draper et al., 2003, p. 160).

In the absence of suitable complex, dynamic models like those described above, or the data required to drive them, the economic value of forecasts can still be estimated with respect to hydroelectricity production using more simplified decision-making models. For example, McCollor and Stull (2008b) developed such a model for daily reservoir operation using the static cost-loss model of Richardson (2000). This cost-loss model has also been employed in the evaluation of forecasts of temperature for the energy sector (Stensrud and Yussouf, 2003), road-weather forecasts (Thornes and Stephenson, 2001), and severe weather forecasts (Legg and Mylne, 2004) among many other applications.

Krzysztofowicz and Duckstein (1979) employed a similar model of decision making for optimizing the conflicting objectives of flood control and hydroelectric production. However, in this model, decision criteria vary with the decision maker's preferences with respect to these objectives, and can therefore differ from purely economic criteria. Roulin (2007) presents a more dynamic model of decision making in which decisions and actions can change as the event draws nearer and new forecast information becomes available. Georgakakos and Yao (2001) have shown that reservoir management models that make use of forecast ensembles have potential to improve system performance only if the management model or process uses the forecast information effectively. This can be done by employing adaptive decision systems to determine dynamic operational policies given uncertain forecasts.

In this chapter, economic value of the M2M ensemble and various reduced configurations thereof will be evaluated using the simple (static) cost-loss model for reservoir operation developed by McCollor and Stull (2008b). A description of the general cost-loss decision-making model follows in Section 5.2.1. Refinement of the decision-making problem for economical operation of the case-study reservoir is illustrated in Section 5.3.4.

5.2.1 A Simple Cost-Loss Decision Model

In the simplified model of reservoir operation developed herein, the operator, when faced with a forecast of a significant inflow event, must decide whether the forecast probability is great enough to warrant taking mitigative action. This action amounts to drafting (i.e., lowering the water level of) the reservoir by routing the water through the turbines and generating electricity, thereby making room for subsequent inflow. If the event does not occur, this action results in an economic cost due to the lowered hydraulic head, which reduces the energy that can be derived from a given volume of water. Conversely, if the reservoir operator does not draft the reservoir and the inflow event does occur, an economic loss is incurred by spilling water rather than running it through the generators.

The reservoir operator's choice depends on: the capacity of the reservoir to take in additional water, the inflow forecast, and operational constraints such as maintaining constant reservoir levels for maximum hydroelectric production or recreational usage, or meeting minimum flow requirements for aquatic habitat. By taking the appropriate action for each forecast, the operator can expect to minimize costs and losses over the long run. The decision-making process can be simplified by making certain assumptions that are outlined in Section 5.3.4.

There are four possible combinations of mitigative action and occurrence of an event, each with its own net cost. These are summarized in Table 5.1. If the forecast probability of a particular event (where an event is the exceedance of some significant inflow threshold) exceeds some threshold value (p_t), the reservoir operator takes action, incurring a cost C. A loss L occurs if the event was

not forecast, but was observed to occur. Operational expenses that occur as a result of using the forecasting system are related to the number of forecast hits a (correct forecasts), the number of false alarms b (no occurrence when forecast), and the number of misses c (event occurs but was not forecast). Correct rejections d (event was neither forecast nor observed) do not result in any expenses.

Table 5.1: Cost-loss contingency table of inflow forecasts and observations. The number of forecast hits is given by a, b is the number of false alarms, c the number of misses, and d the number of correct rejections. Action is taken when a particular inflow exceedance event is forecast to occur, incurring a cost C, while events that were not forecast result in losses L. Correct rejections result in no costs or losses.

	Observed	Not observed
Forecast/Action	$a(\$C)$	$b(\$C)$
Not forecast/No action	$c(\$L)$	$d(\$0)$

The mean expense associated with using a particular forecasting system is then given by:

$$E_f = \frac{a}{n}C + \frac{b}{n}C + \frac{c}{n}L, \tag{5.1}$$

where n is the total number of forecast-observation pairs in the evaluation period ($n = a+b+c+d$) (Richardson, 2000).

The economic value of the forecasts is assessed by comparing E_f to the mean expenses associated with perfect forecasts (E_p), and those associated with operating without any forecast information (E_c):

$$V = \frac{E_c - E_f}{E_c - E_p}. \tag{5.2}$$

Note that this relative value definition is equivalent to a skill score, with maximum V of one indicating perfect forecasts, and V less than zero indicating forecasts that are less skillful/valuable than climatology (Wilks, 2001; Richardson, 2003).

In the absence of any forecast information, the decision maker will either take protective action every day, incurring a mean expense of C, or they will choose to never protect, in which case losses occur at a rate equal to the climatological base rate of an event (s), resulting in mean expense of sL. Since this choice depends on which course of action results in the minimum economic risk, E_c is simply $\min(C, sL)$. Given perfect forecasts the decision maker would choose to protect only when the event occurred, yielding a mean expense of $E_p = sC$.

Upon defining the user-specific cost-loss ratio $\alpha = C/L$, Eq. (5.2) can be expressed as:

$$V = \frac{min(\alpha, s) - \frac{\alpha}{n}(a+b) - \frac{c}{n}}{min(\alpha, s) - s\alpha}. \tag{5.3}$$

Incorporating the definitions of hit rate (H) and false alarm rate (F) given in Appendix A, and setting the climatological base rate to $s = (a+c)/n$, Eq. (5.3) can be rewritten as:

$$V = \frac{min(\alpha, s) - F\alpha(1-s) + Hs(1-\alpha) - s}{min(\alpha, s) - s\alpha}. \tag{5.4}$$

5.3 Case Study

5.3.1 Study Dates and Data

The M2M ensemble (described in Section 5.3.2) is used to forecast inflows to the Daisy Lake reservoir, a hydroelectric facility on the upper Cheakamus River in the mountainous terrain of southwestern BC, Canada. The reservoir is operated by BC Hydro. Evaluation is carried out over the 2010–2011 and 2011–2012 water years. Forecasts were also generated for the 2009–2010 water year, but these forecasts are excluded from evaluation because their quality may be impacted by the spin-up of uncertainty model and probability calibration parameters.

For this particular hydroclimatic regime, a water year is defined as the period from October 1 to September 30 of the following year. Fall and winter storm season (October – April) inflows are primarily driven by precipitation from Pacific frontal systems. Rain-on-snow events can result in significant inflows during this period. During the spring and summer warm season (May – September), inflows are snowmelt-driven, with some late-season glacier melt contributions. Daily average inflow rates are calculated by BC Hydro using a water balance based on observed reservoir levels and outflows. These calculated inflows are considered to be of high quality for this basin, and will be referred to as observed inflows for the purposes of this study. Hourly forecast inflows to the Daisy Lake reservoir are transformed into daily average inflow rates for verification against these observations.

In the simple cost-loss model developed herein, it is assumed that the hypothetical Daisy Lake reservoir operator is sensitive to daily average anomaly inflow rates of 70 m^3/s and 100 m^3/s . Anomalies are calculated by subtracting the daily climatological median inflow from the forecast, and are used so that the forecasting system is not unduly rewarded for making high inflow forecasts during the snowmelt season when relatively little skill is required to do so. Instead, the forecasting

system is rewarded for correctly forecasting events that are significantly different from climatology, or a readily anticipated inflow value. The choice of inflow rate threshold is limited by the small sample size of two water years in which no extreme events occurred (Figure 5.1). An anomaly threshold of 100 m^3/s corresponds to an absolute inflow threshold of approximately 130–150 m^3/s depending on time of year; based on the full climatological record, this threshold corresponds to a one-in-one-month inflow event. For comparison, inflow events requiring pre-generation or drafting of the Daisy Lake reservoir occur on average once per year (Doug McCollor, personal communication, April 17, 2013).

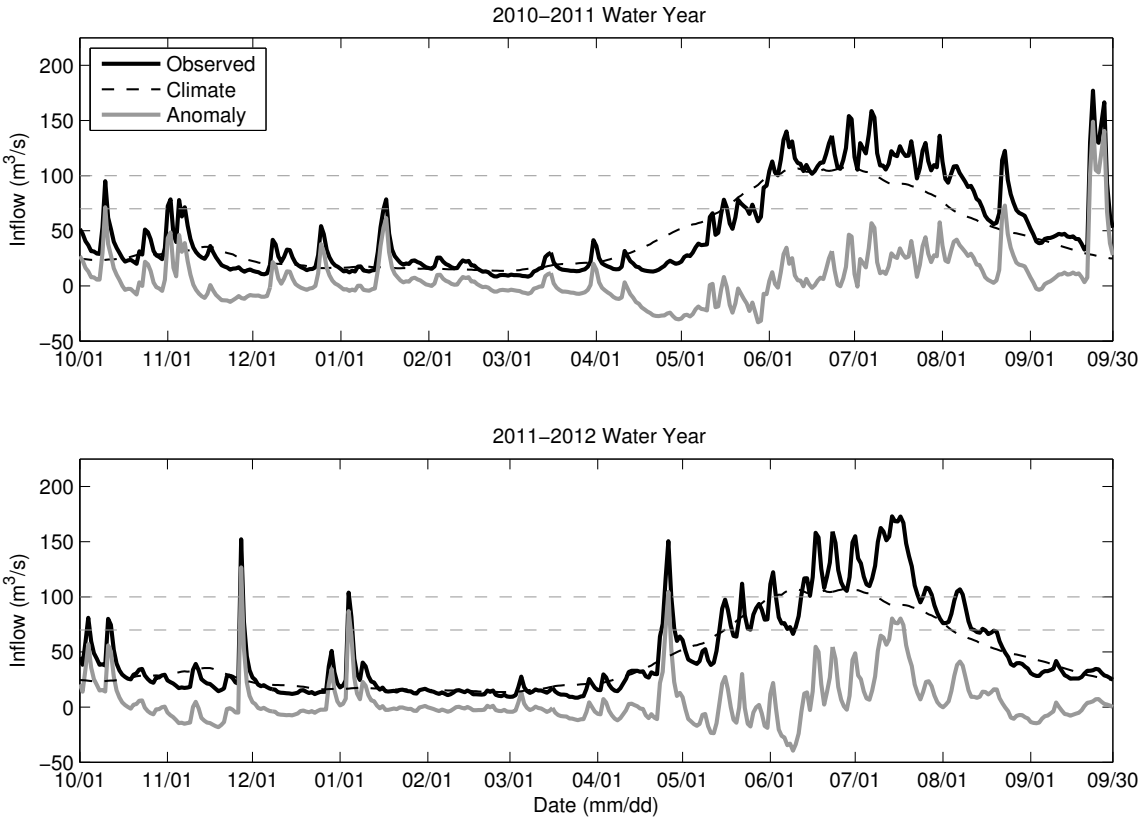

Figure 5.1: Observed inflows (solid black line) for the 2010–2011 and 2011–2012 years. Anomaly inflow values (solid grey line) are calculated by subtracting the climatological inflows (dashed black line) from the observations. The anomaly thresholds of 70 m^3/s and 100 m^3/s are indicated by the horizontal dashed grey lines.

During the evaluation period, the 100 m^3/s threshold is exceeded on eight days. Six of these inflow events occur during the spring/summer warm season, and the remaining two events occur during the fall/winter storm season. The climatological base rate (or exceedance probability), s, for

inflow event anomalies exceeding 100 m^3/s is approximately 0.011 during the evaluation period. The corresponding base rate for the 70 m^3/s threshold is approximately 0.023. During the evaluation period, there are five such storm season events and twelve during the warm season.

In addition to assessing the economic value of the M2M configurations over the case-study period, probabilistic forecast quality will be evaluated using the ignorance score (IGN) and continuous ranked probability score (CRPS). Deterministic ensemble median forecast quality and skill will be measured using Mean Absolute Error (MAE) and the Root Mean Squared Error Skill Score (RMSESS) for each configuration. For a description of these verification measures and their interpretation, see Appendix A.

5.3.2 The Member-to-Member (M2M) Ensemble Forecasting System

The M2M ensemble forecasting system is designed such that all sources of uncertainty in the inflow modelling chain are sampled. Uncertainty in the forecasts comes from the NWP models used to drive the hydrologic models, the hydrologic models themselves and their parameterizations, and the initial conditions or hydrologic states from which the forecasts are started.

The NWP models are taken from the operational ensemble suite run by the Geophysical Disaster Computational Fluid Dynamics Centre (GDCFDC), in the Department of Earth, Ocean and Atmospheric Sciences at the University of British Columbia. The ensemble consists of three independent nested limited-area high-resolution mesoscale models with forecast domains centred over southwestern BC: the Mesoscale Compressible Community model (MC2; Benoit et al., 1997); the fifth-generation Pennsylvania State University-National Center for Atmospheric Research Mesoscale Model (MM5; Grell et al., 1994); and Version 3 of the Weather Research and Forecasting (WRF) model (Skamarock et al., 2008). Hourly model output fields with grid spacing of 12, 4 and 1.3 km are used for this study.

The NWP models are initialized at 00UTC. Forecast run time varies during the case-study period. From the start of the evaluation period (October 2010) all NWP models were run out to at least 60 hours except for the 1.3-km MC2 model runs, which are 39 hours due to operational time constraints. The WRF model produced 84-hour forecasts for all grids throughout the study period. In March 2011, the MM5 12-km and 4-km forecasts were extended to 84 hours, enabling them to generate a 3-day inflow forecast. In March 2012, 1.3-km MM5 model output was made available out to 84 hours, resulting in a day-3 inflow forecast ensemble consisting of up to 48 members; forecast days 1 and 2 had at most 72 ensemble members available throughout the three-year forecast period. Due to occasional NWP model failures, the size of the ensemble forecast issued each day is variable.

From the beginning of the case-study period through March 2012, the coarse resolution (108 km

horizontal grid spacing) outer nests of the three NWP models were initialized using the National Centers for Environmental Prediction (NCEP) North American Mesoscale (NAM) model, which also provides time-varying boundary conditions. In March 2012, the initial/boundary condition for the MM5 and WRF was switched to NCEP's Global Forecast System (GFS) model, while MC2 continued to make use of the NAM.

The Distributed Hydrologic (DH) models applied to the case-study watershed are the Water balance Simulation Model (WaSiM; Schulla, 2012) and WATFLOOD (Kouwen, 2010). These models were selected because they are distributed, and therefore able to take direct advantage of high-resolution NWP input, and because they are able to simulate snow and glacier melt processes and lakes in complex terrain given relatively limited input data. Both DH models are run at 1 km grid spacing with an hourly time step. The NWP fields are downscaled to the DH model grid using interpolation schemes built into each DH model. For the WaSiM model, 12-km NWP fields are downscaled using two methods: inverse-distance weighting (IDW); and elevation-dependent regression (Schulla, 2012). The 4-km and 1.3-km NWP fields are downscaled using a bilinear interpolation scheme. WATFLOOD downscaling is done using IDW that incorporates elevation dependence using optional elevation adjustment rates for both temperature and precipitation. The 12-km fields are downscaled using IDW with two different sets of elevation adjustments, while the 4- and 1.3-km fields do not use the elevation adjustment.

Both WaSiM and WATFLOOD model parameters have been optimized using the Dynamically Dimensioned Search (DDS) algorithm (Tolson and Shoemaker, 2007; Graeff et al., 2012; Francke, 2012). Three parameter sets were generated for each model by using three different objective functions for DDS optimization: the mean absolute error (MAE) of simulated inflow, to minimize overall errors; Nash-Sutcliffe Efficiency (NSE; Nash and Sutcliffe, 1970) of inflow, to emphasize performance during high-flow events; and the NSE of log-transformed flows, to optimize during low-flow periods. Simulations during the ten-year optimization period (1997–2007) were driven by observed meteorological data at several weather stations within the case-study watershed and surrounding area.

The multi-state or multi-initial-condition component of the M2M ensemble forecasting system arises as a direct consequence of implementing a multi-parameter component. In forecast mode, the hydrologic state for each model and each model parameterization is updated at the start of the forecast day by driving the model with observed meteorological data. This resulting simulated state is used as the initial condition for the day's forecast run. In order to avoid discontinuities early in the daily forecast cycle, the parameter set used for the updating of hydrologic state must match that used in the forecast. Thus, each parameter set has its own daily hydrologic state for each model. Figure 5.2 illustrates the update/forecast process for a particular parameterization of the

WaSiM model. The forecast workflow is indicated by the solid arrows. Dashed arrows illustrate how meteorological observations are used to update the model configuration's hydrologic state for the following day's forecasts. The model configuration is specified by dash-dotted arrows. This process is repeated for each watershed model (WaSiM and WATFLOOD) and each parameterization/state, yielding 72 unique inflow forecasts each day.

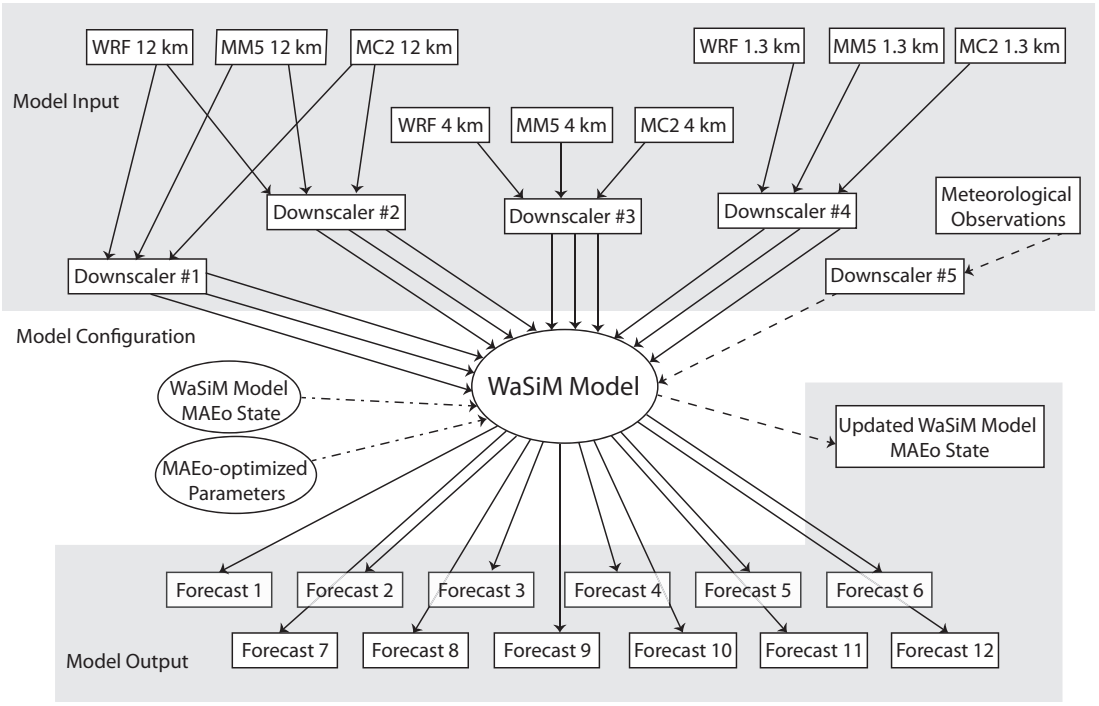

Figure 5.2: The flow of information into and out of the WaSiM model for generating forecasts with the MAE-optimized parameter set. This process is repeated for each watershed model (WaSiM and WATFLOOD) and each parameterization/state, yielding 72 unique inflow forecasts each day.

Ensemble forecasts from the M2M forecasting system are transformed into probabilistic forecasts using the Ensemble Model Output Statistics (EMOS) method of Gneiting et al. (2005) (see Chapter 4). EMOS fits a normal distribution to the bias-corrected ensemble mean whereby distributional spread is based on a combination of ensemble variance and past ensemble mean errors. During the storm season, ensemble mean inflow errors to the Daisy Lake dam were found to exhibit a distribution with a high, narrow peak and slight positive skewness. Thus, prior to fitting the normal distribution to the ensemble, the data are log-transformed. During the warm season, when inflows are driven by snowmelt and glacier melt, forecast errors were found to be normally distributed, and no transformation is required prior to fitting the normal probability distribution. In Chapter 4,

these seasonal uncertainty models were shown to produce reliable forecasts (i.e., events forecasted with probability p are, over the course of many such forecasts, observed to occur a fraction p of the time). The probability calibration method of Nipen and Stull (2011) was able to significantly increase warm-season forecast sharpness, thereby decreasing ignorance.

5.3.3 Ensemble Reduction Test Cases

In order to compare the cost-effectiveness of each of the M2M components described above, the performance of the ensemble with all components, and with individual components removed is evaluated. Note that the multi-state, multi-parameter (MSP) hydrologic modelling elements described in Section 5.3.2 are inextricably linked and therefore comprise a single M2M component. Since interpolation schemes are built into the hydrologic models, these are considered to be free ensemble components. Recall that the 12-km NWP model inputs to both DH models are downscaled using two different methods, while the higher resolution NWP grids are each downscaled using one method per DH model. In the M2M configurations tested in this economic analysis, 12-km NWP grids are downscaled using both schemes unless otherwise specified.

The NWP ensemble itself is comprised of two ensembles: a multi-model (MM) ensemble and a multi-grid scale (MGS) ensemble. It is likely that ensemble configurations including a MM NWP ensemble would be most commonly exploited in operational settings, as low resolution models are available free-of-charge more commonly than their higher-resolution counterparts — hence the term "the poor-man's ensemble" (Ebert, 2001). In configurations consisting of a single NWP input, the WRF model output has been selected because it is the most current NWP model in the M2M ensemble, and, as a community model, is subject to ongoing development and support. In order to remove any impact of ensemble size on comparisons of high-resolution and low-resolution single-NWP configurations, the 12-km NWP fields are downscaled using one interpolation scheme.

The following reduced M2M configurations are evaluated in this study (where the name of the configuration specifies the component that has been removed):

- –MGS: multi-DH, multi-MSP using 12 km multi-model NWP fields (36 members)
- –MM: multi-DH, multi-MSP using multi-grid scale WRF NWP fields (24 members)
- –MSP: multi-NWP, multi-DH with MAE-optimized parameterizations (24 members)
- –DH (WFLD): multi-NWP, multi-MSP, WATFLOOD DH model (36 members)
- –DH (WaSiM): multi-NWP, multi-MSP, WaSiM DH model (36 members)
- –NWP (HR): multi-DH, multi-MSP with WRF 1.3-km NWP fields (6 members)

- –NWP (LR): multi-DH, multi-MSP with WRF 12-km NWP fields downscaled one way (6 members)

Each of these ensemble configurations is transformed into a probabilistic forecast in the form of a probability density function (PDF) using the EMOS method described in Section 5.3.2 (including warm-season probability calibration). Relative value of each of the reduced ensembles is compared to that of the Full M2M probabilistic forecasts.

In order to be useful for risk-based decision making, probabilistic forecasts must be reliable. Otherwise, their probabilities cannot be taken at face value. Thus, calibration deviations Eq. (A.8) for the probability forecasts derived from each reduced M2M configuration were examined during the storm season and warm season. Forecasts with calibration deviations less than 50% greater than their expected values (Nipen and Stull, 2011) were taken to be sufficiently reliable. Using this criterion, all of the configurations listed above produced calibrated forecasts except for –MSP and –DH (WFLD) storm season forecasts. The calibration method of Nipen and Stull (2011) was applied during storm season periods when the forecasts were deemed to be sufficiently unreliable. The "intelligent" calibration scheme described in Chapter 4 (with $ICF = 1.67$ [Eq. (4.9)]) improves reliability and prevents ignorance scores from being inflated by sampling error (Nipen, 2012).

5.3.4 Cost-Loss Model Development for Daisy Lake

In this section, the cost-loss model described in Section 5.2.1 is refined, providing a specific model for the Daisy Lake reservoir. While this represents a great simplification of reservoir management and operational constraints, it can be used to examine the relative value of probabilistic inflow forecasts to the reservoir. This model is taken from McCollor and Stull (2008b).

A schematic of a hydroelectric reservoir is presented in Figure 5.3. The physical characteristics of the reservoir are given by:

h_1 = nominal head (m), the height difference between the reservoir outlet and the turbine,

h_2 = the difference between the lowered (to prevent spillage) reservoir elevation and the outlet (m),

h_3 = the difference between the full reservoir and lowered reservoir elevations (m),

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

Inflows (Q_{in}) to the reservoir are either spilled (Q_s), or channeled to the turbine via the penstock at a rate Q_t. The power P (W) produced by the turbine is given by:

$$P = \eta\gamma Q_t H, \tag{5.5}$$

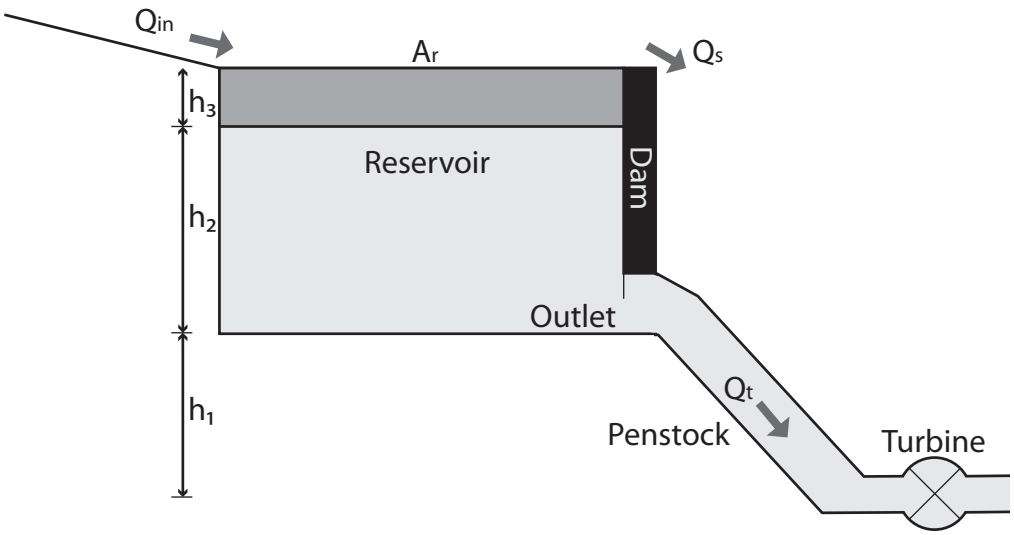

Figure 5.3: Reservoir schematic diagram for the cost-loss economic model developed in Section 5.3.4 for Daisy Lake. Water that does not spill can be channeled through the penstock to the turbines to produce power and therefore revenue. Figure based on McCollor and Stull (2008b).

where η is the turbine efficiency (expressed as a fraction), γ is the specific weight of water (9807 N/m^3 at 5°C), Q_t is the flow through the turbine (m^3/s), and H is the head, or the difference in elevation (m) between the reservoir level and the turbine. Assuming constant flow conditions, the energy K (J) produced by running the generator for time T (in seconds) is:

$$K = PT. \tag{5.6}$$

Given a selling value of S (\$J^{-1}), the market value ν (\$) of this energy is:

$$\nu = KS = PTS. \tag{5.7}$$

In this decision model, losses occur in conjunction with large inflow events that are not forecast. If it is assumed that the reservoir is full when this event occurs (i.e., $H = h_1 + h_2 + h_3$), then this loss is equivalent to the value of the water that spills past the generator without producing power or revenue. The amount of this loss, $L(\$)$, which is related to lost power P_L, is given by:

$$L = P_L TS = \eta\gamma Q_s(h_1 + h_2 + h_3)TS. \tag{5.8}$$

Assuming that lowering the reservoir level by h_3 would have prevented this spill, the spilled volume

Q_s can be defined as:

$$Q_s = A_r h_3. \tag{5.9}$$

In order to prevent losses, the reservoir should be lowered preceding large inflow events. The costs of this preventive action result from operating the reservoir at a lower head $H = h_1 + h_2$ relative to operation at $H = h_1 + h_2 + h_3$:

$$C = \eta\gamma Q_t(h_1 + h_2 + h_3)TS - \eta\gamma Q_t(h_1 + h_2)TS = \eta\gamma Q_t h_3 TS. \tag{5.10}$$

If we further assume that flow through the turbine, Q_t, is at least equal to the base inflow rate, Q_b, and let G represent the reservoir storage volume [$Ar(h_2 + h_3)$], then the cost/loss ratio ($\alpha = C/L$) is:

$$\alpha = \frac{Q_b}{A_r h_1 + G}. \tag{5.11}$$

Eq. (5.11) can be evaluated for Daisy Lake using the physical parameters listed in Table 5.2, yielding $\alpha = 0.00033$. This number indicates that the Daisy Lake reservoir operator is highly sensitive to losses associated with spilling, and that they will benefit from taking mitigative action even when events are forecasted with very small probabilities.

Table 5.2: Physical parameter values for the Daisy Lake reservoir for cost-loss calculations. Values are taken from McCollor and Stull (2008b).

Physical Parameter	Value
Nominal head h_1 (m)	291
Reservoir storage G (m^3)	46×10^6
Reservoir area A_r (m^2)	43×10^6
Daily base inflow Q_b (m^3/day)	4.2×10^6

As noted in McCollor and Stull (2008b), it is possible to further refine the basic cost-loss model by assuming that the costs associated with operating the reservoir at a lowered head may be realized until the next inflow event occurs. The length of time until the next event is given by the inverse of the climatological base rate of the event (s). This refined cost-loss ratio is:

$$\alpha = \frac{Q_b}{s(A_r h_1 + G)}. \tag{5.12}$$

Further refinement can be made by allowing for a dynamic energy market. That is, we assume that power from spilled water could have been sold at a contract price S_c, while additional power

required to meet demand during periods of lowered head will need to be purchased at a market price of S_m. This yields a cost-loss ratio of:

$$\alpha = \frac{S_m Q_b}{S_c \; s(A_r h_1 + G)}. \tag{5.13}$$

Specific cost-loss ratios for the Daisy Lake reservoir calculated using Equations (5.12) and (5.13) are given in Table 5.3 for inflow anomaly thresholds of 70 m^3/s and 100 m^3/s. Following from McCollor and Stull (2008b), the ratio of market to contract price (S_m/S_c) is taken to be 2.5 for the case study. In reality, this ratio is highly variable, and market prices can spike such that S_m/S_c approaches 20, though it is more commonly in the range of 1 to 10 (Doug McCollor, personal communication, April 17, 2013). Based on these values, the cost-loss ratio for Daisy Lake can be seen to vary over a range of 0.00033 (using the basic model) to approximately 0.3 (incorporating s for the 100 m^3/s threshold and S_m/S_c of 10).

Table 5.3: Cost-loss ratios for the Daisy Lake reservoir calculated using the basic model [Eq. (5.11)], including climatological frequency s [Eq. (5.12)], and including a variable electricity market [Eq. (5.13)] with $S_m/S_c = 2.5$.

C/L Model	Anomaly Threshold	
	70 m^3/s	100 m^3/s
Basic Model	0.00033	0.00033
Including s	0.014	0.030
Including S_m/S_c	0.036	0.075

Note that the dynamic cost-loss model developed by Roulin (2007) is based on the idea that users may be able to reduce the cost of taking action if they have more time to prepare. This reduction in cost results in a reduction in α toward the region where maximum value is achieved (Richardson, 2000). Such a cost reduction would provide little benefit in the case-study region, where α is already very low. Additionally, if the size of the event to which the reservoir operator is sensitive is assumed constant, it follows from Eq. (5.10) that the cost associated with taking action earlier is actually higher, because the reservoir will operate at lower head for a longer period of time T prior to the next inflow event. A more complex dynamic model that allows the inflow threshold to change with time and that accounts for the duration of an inflow event would be more suitable, but is beyond the scope of this research.

5.4 Results and Discussion

5.4.1 Quality and Skill of Reduced Ensemble Forecasts

As shown in Figure 5.4, ensemble median forecasts from the –MGS M2M configuration have the lowest MAE for forecast day 1. As forecast horizon increases, the importance of including high-resolution NWP guidance becomes more apparent, as illustrated by the superior performance of both –MM and Full M2M forecasts. The same conclusions can be drawn from an examination of RMSESS. At lead times of 1-2 days, the inclusion of a multi-model NWP ensemble is more important than a multi-grid scale NWP ensemble (–MM is worse than –MGS). Sampling any aspect of NWP uncertainty has the potential for significant gains in forecast quality as compared to ensemble configurations where NWP uncertainty is neglected altogether (–NWP).

It has been reported that the uncertainty in NWP model output is the largest source of error in NWP-driven flow forecasts with a time horizon beyond several days, whereas for shorter lead-times, uncertainties in the hydrologic model dominate prediction errors (Coulibaly, 2003; Cloke and Pappenberger, 2009). Note that the comparative importance of NWP and DH model error over different time scales depends on context; for an anticipated heavy rainstorm in a small and flashy catchment, uncertainty around the amount of rainfall expected over the next day may have considerably more impact on tomorrow's streamflow forecast than uncertainty introduced by the hydrologic models. Indeed, the deterministic scores in Figure 5.4 indicate no clear winner with respect to –NWP and –DH or –MSP ensembles. On average, the –NWP, –DH and –MSP ensembles are very similar for days 1 and 2. Ignoring the large MAE of –DH (WaSiM), it appears that NWP error is most important at a lead time of 3 days. RMSESS also reveals that removal of the NWP ensemble component introduces more large forecast errors at this lead time. Excluding the MSP component does not introduce many large forecast errors. Overall, including an ensemble of different hydrologic modelling approaches is more advantageous than including a multi-state and multi-parameter component for a single hydrologic model for this reservoir.

The reliability and sharpness of the various probabilistic forecasting systems are assessed using the ignorance and continuous ranked probability scores shown in Figure 5.5. These scores indicate that the –MGS and Full ensemble probabilistic forecasts are of equally high quality for a lead time of one day, and that beyond this forecast horizon, the inclusion of high-resolution NWP models is important. Also, these results support the finding that including a multi-model NWP ensemble component is more important than a multi-grid scale NWP component, and that either is far superior to ignoring NWP uncertainty altogether.

Given that all of the configurations yield probabilistic forecasts with comparable (small) cali-

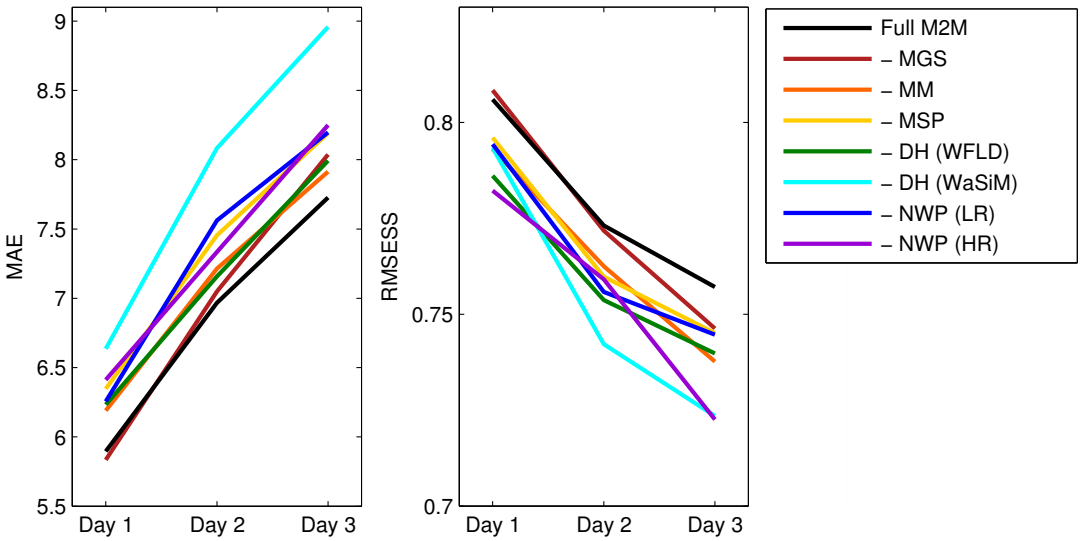

Figure 5.4: MAE and RMSESS for ensemble median forecasts derived from the various M2M configurations. Perfect deterministic forecasts have MAE of zero and RMSESS of one.

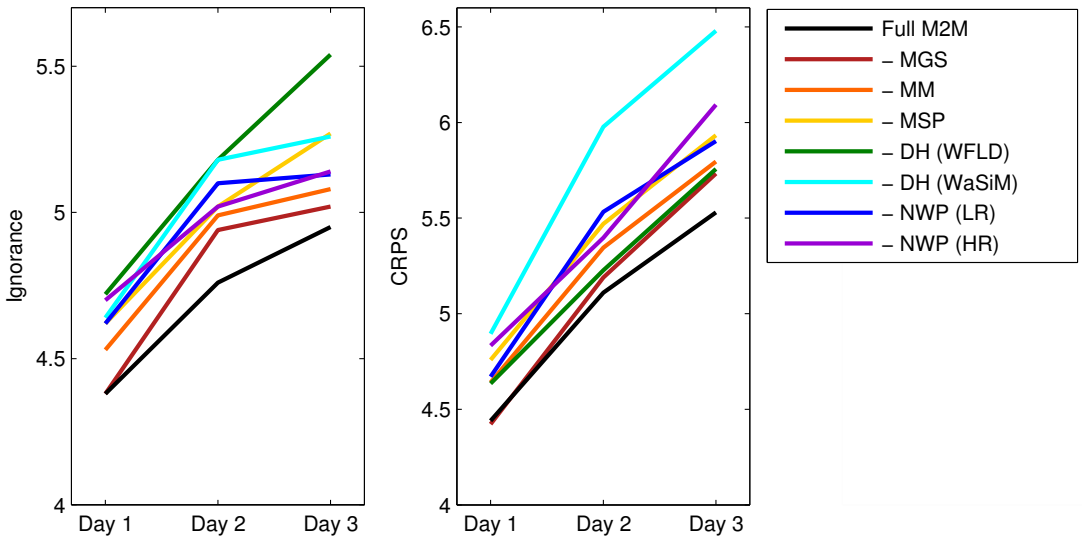

Figure 5.5: Ignorance and continuous ranked probability scores for probabilistic forecasts derived from the various M2M configurations. Lower values are preferred for these scores

bration deviations (D), the lower ignorance scores of the Full M2M forecasts indicate that they are better able to concentrate probability density in the right area (i.e., near the verifying observation) each day. Since the forecast PDFs are centred at the ensemble mean, lower ignorance scores are, as expected, generally associated with lower MAE and higher RMSESS. The slightly inflated ignorance scores of the –DH (WFLD) forecasts relative to the –DH (WaSiM) forecasts are likely caused by the introduction of sampling error by the probability calibration scheme applied to storm-season forecasts. Ensembles that ignore error associated with the hydrologic models (–DH and –MSP) have generally higher ignorance scores (i.e., are worse) than –NWP ensembles.

The lower CRPS of the Full M2M ensemble forecasts indicates that its forecast PDFs are sharper than those derived from any of the reduced M2M configurations. Again, the exception occurs for day 1 forecasts, where the –MGS ensemble CRPS is slightly lower. The –DH (WaSiM) forecast PDFs have consistently higher spread than the –DH (WFLD) configuration, and this is reflected in their higher CRPS. In fact, forecasts from this configuration have generally higher and more variable spread than any other configuration; analysis of EMOS spread parameters reveals that distributional spread is often taken to be up to $8\times$ the –DH (WaSiM) ensemble variance and is further inflated by the large ensemble mean errors (Figure 5.4). Removal of any multi-model NWP component (–MM or –NWP) also results in high spread, and this is again reflected by higher CRPS values. These results support the findings of Stensrud and Yussouf (2003) who illustrated that for temperature forecasting, multi-model weather ensembles improve the spread-skill relationship (the spread of the ensemble members should be related to the accuracy or skill of the ensemble mean; when the forecast is more certain, as indicated by low ensemble spread, errors are expected to be small).

5.4.2 Economic Value of Ensemble Components

Value curves for an inflow anomaly threshold of 70 m^3/s are shown in Figure 5.6 for the Full 1-day M2M probability forecasts. Curves are plotted for probability thresholds p_t of 0.02 (dashed line), 0.1, 0.2,...0.9 (solid lines), and 0.98 (dash-dotted line). The heavy solid line is the envelope curve of optimal value achieved when each user chooses the probability threshold that maximizes the value for their α.

For calibrated forecast probabilities, users will maximize the value they get from the forecasting system by choosing $p_t = \alpha$ (Richardson, 2000). Deviations are caused by sampling variability; value curves for operationally insignificant inflows with greater sample sizes (e.g., 40 m^3/s, not shown) are able to match p_t values to their corresponding α ranges more closely. Note that probability thresholds above 0.7 are rarely exceeded even for the relatively low inflow anomaly threshold of 70 m^3/s; the false alarm rate is zero, and it follows from Eq. (5.4) that the value converges to the hit rate H for $\alpha > s$, indicated by the horizontal line segments in Figure 5.6. Forecasts issued with

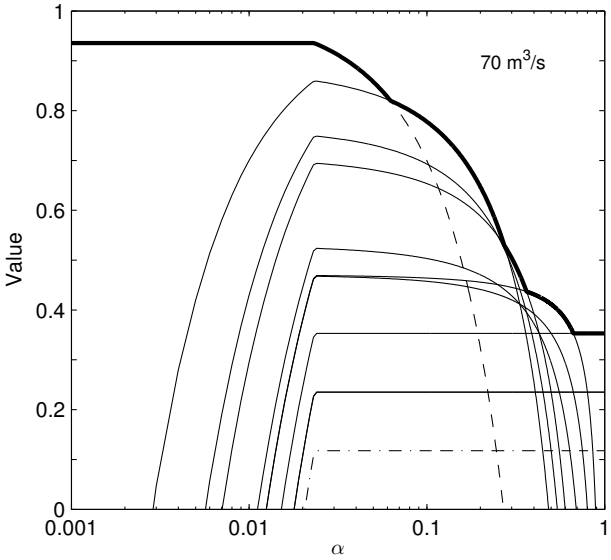

Figure 5.6: Forecast value as a function of user-specific cost-loss ratio α for the Full 1-day M2M probability forecast. Relative value of zero indicates that the forecasting system offers no benefits over climatology, while perfect forecasts have relative value of one.

probability thresholds at or below 0.08 never result in forecast misses, thus the hit rate is one, and it follows from Eq. (5.4) that the value is equal to $1 - F$ for $\alpha < s$. The value curves for p_t of 0.8 and 0.9 are identical. These characteristics of the value curves are due to sampling limitations.

The impacts on relative value of removing various components of the M2M ensemble are shown in Figures 5.7 and 5.8. Envelope value curves are plotted for the Full ensemble probabilistic forecasts (solid black line) as well as for each reduced M2M configuration (coloured lines). Note that the envelope curve in Figure 5.7 does not match that in Figure 5.6 because the maximum value has not been constrained by the few p_t shown in Figure 5.6. The range of α valid for operation of the Daisy Lake reservoir at each inflow threshold for S_m/S_c from 1 to 10 is indicated by the grey shaded area. Since the uncertainty models used in this study generate inflow forecasts over a continuous range of probability thresholds, differences in relative value between ensembles of varying size are due only to associated changes in forecast quality and not in the available resolution of p_t (Richardson, 2000, 2001).

For events exceeding the 70 m^3/s threshold, it can be seen in Figure 5.7 that day 1 forecast value is insensitive to most changes in M2M ensemble composition over the range of α valid for Daisy Lake. Only the –DH (WFLD) and –NWP (HR) ensembles result in forecast misses at low p_t, resulting in significantly reduced value. The impact of the large spread of the –DH (WaSiM) ensemble can be seen in its slightly lower forecast value at low p_t where there are no forecast

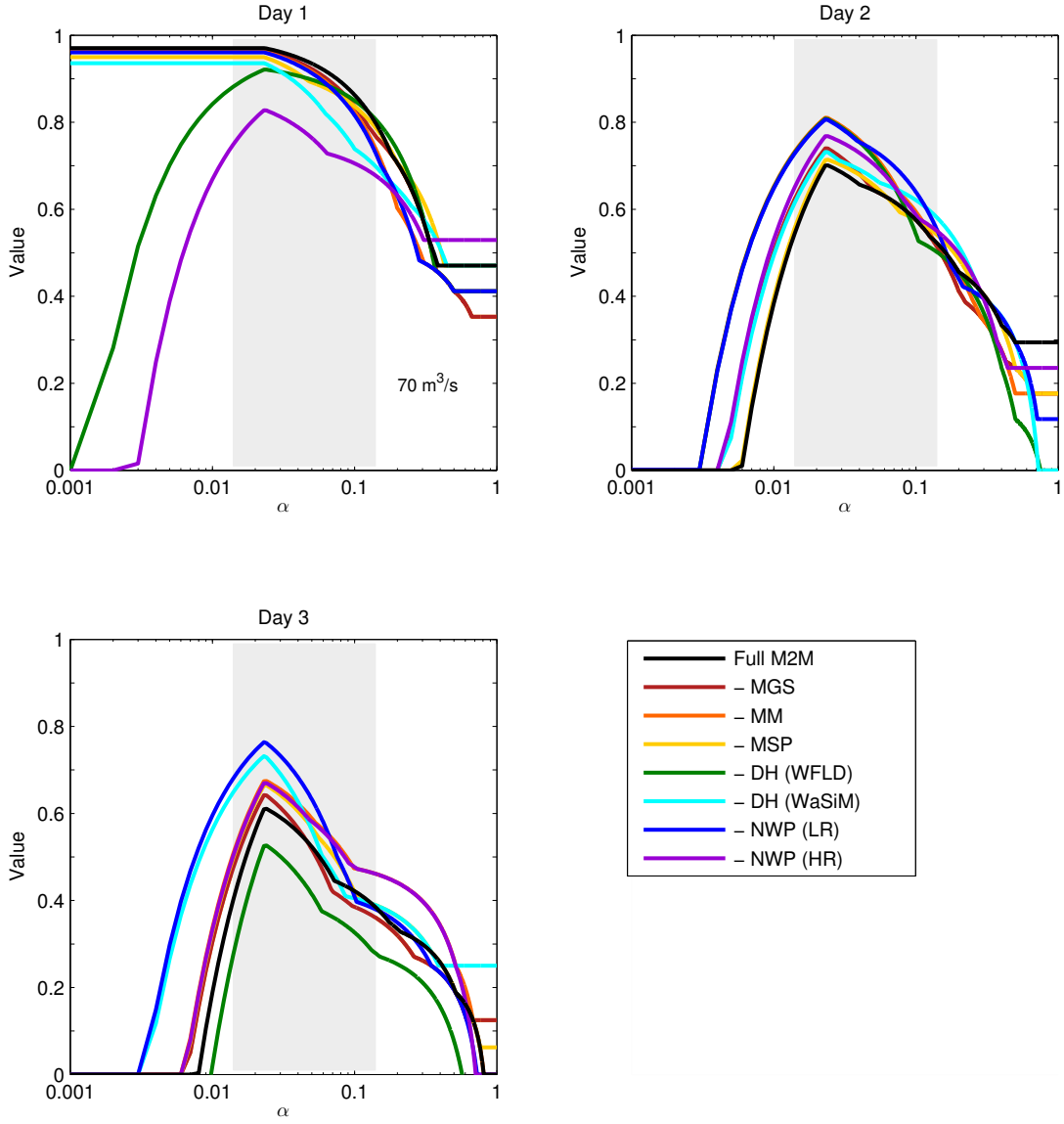

Figure 5.7: Forecast value as a function of user-specific cost-loss ratio α for the Full M2M probability forecast (black line), and the various reduced ensemble configurations (coloured lines). The range of α valid for Daisy Lake reservoir operation for S_m/S_c from 1 to 10 are indicated by grey-shading.

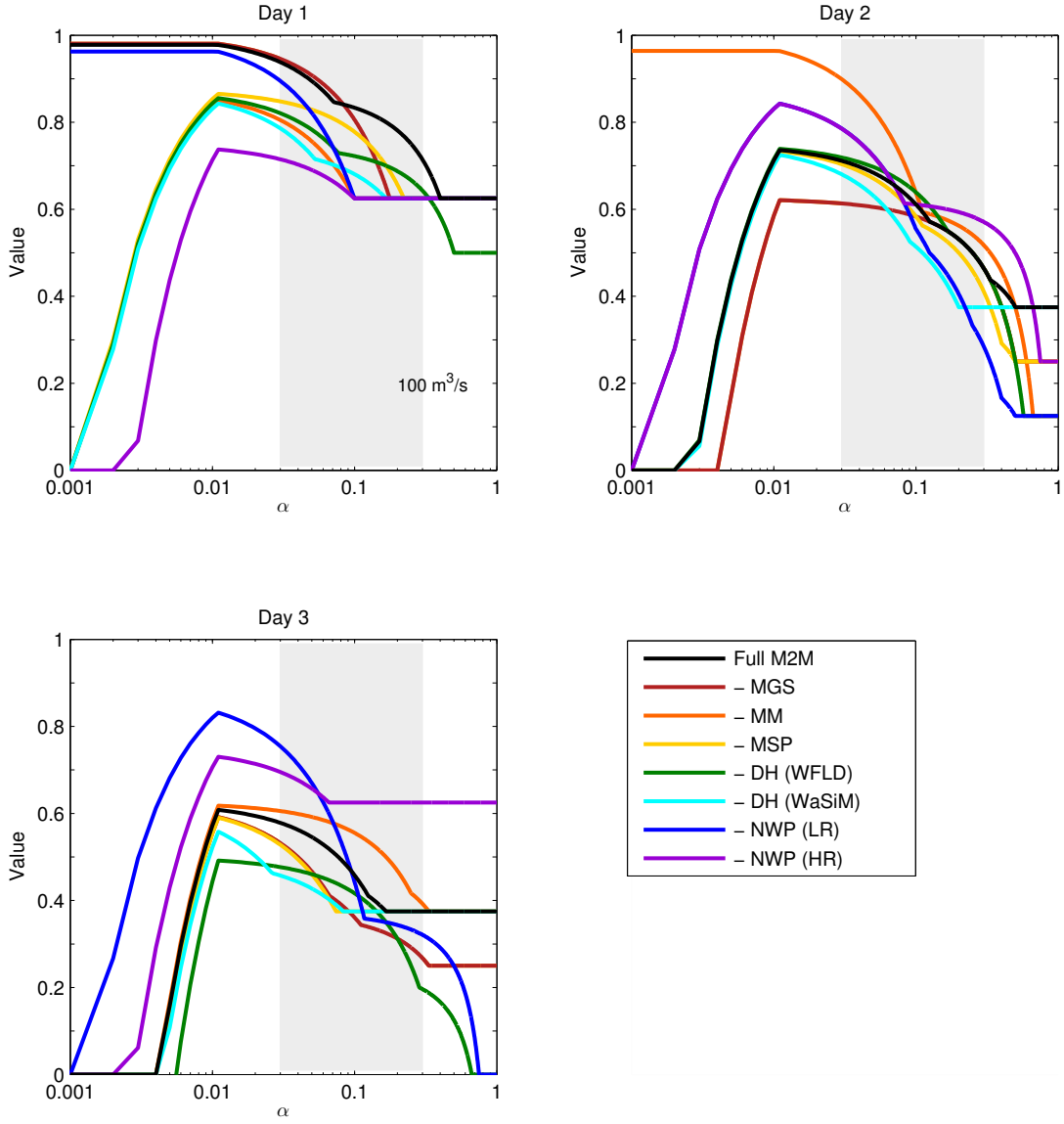

Figure 5.8: Same as Figure 5.7, but for an inflow anomaly threshold of 100 m^3/s .

misses, but a large number of false alarms, indicating poor discrimination. The importance of including ensemble NWP input is apparent for users with α greater than approximately 0.08, where value decreases due to lower hit rates for inflow events of this size.

At forecast lead times of two and three days, –NWP (LR) shows high value for a range of users and significantly greater value for low-α users. This is a result of the high spread of the model's fitted PDF resulting in fewer forecast misses at low probabilities. The same is true of the other

–NWP ensembles and the –DH (WaSiM) configuration, relative to the Full ensemble. The increase in spread associated with these configurations also results in the forecasting systems issuing more false alarm forecasts. For low α, it follows from Eq. (5.4) that false alarms are not penalized very heavily, and value is impacted very little. For this reason, most of the M2M configurations show greater value than the Full ensemble over the range of α valid for the Daisy Lake reservoir operator.

Figure 5.8 shows that day 1 forecast value is more sensitive to ensemble configuration at an anomaly inflow threshold of 100 m^3/s , particularly at low probability thresholds where most ensembles miss event forecasts more frequently. Over the range of α valid for the case study, the Full ensemble has more value than many other configurations. The –MGS has less value than –MM ensemble at lead times of 2 and 3 days, supporting the conclusion that high-resolution NWP input becomes increasingly important with forecast lead time, particularly for high-impact events. At lead times of two and three days, ensembles that neglect sources of NWP uncertainty have higher value than the Full ensemble, as high forecast spread leads to few forecast misses and many false alarms. Additionally, calibration deficiencies (or sampling limitations) for forecasts of high inflows cause maximum value curves to follow the individual value curves for very low p_t for a wide range of α.

The actual expense associated with the various ensemble configurations over the two-year evaluation period can be estimated by evaluating the contingency table counts (Table 5.1) for a particular threshold. From Eq. (5.1), actual expense is:

$$E_{actual} = L(a\alpha + b\alpha + c). \tag{5.14}$$

In this example, the cost-loss ratio that incorporates climatological frequency for the 70 m^3/s and 100 m^3/s inflow anomaly thresholds and variable market pricing with $S_m/S_c = 2.5$ is used. Table 5.4 shows the contingency table elements and estimated costs for various ensemble configurations for the 70 m^3/s threshold. Contingency table counts are based on forecasts with p_t of 0.04 (recall that maximum value occurs for users who take action at p_t equal to their α). Results for the 100 m^3/s threshold are shown in Table 5.5 for $p_t = 0.08$. In calculating the one-day loss L as in Eq. (5.8), turbine efficiency (η) is taken to be 0.9 (Gulliver, 1991), and the cost of energy (S) is assumed constant at \$50/MWh[1]. Since the goal of this study is in estimating the relative value of the various M2M ensemble components, the use of these generic values (which may not be appropriate for the case study watershed and dates) is not a critical issue. Accurate estimates of E_{actual} are only necessary for comparison against the price paid for ensemble components, which is beyond the scope of this work.

[1]Source: U.S. Energy Information Administration (2012). 2011 Brief: Wholesale electricity prices mostly lower in 2011. URL: http://www.eia.gov/todayinenergy/detail.cfm?id=4530. Retrieved May 21, 2013. Price quoted is the approximate average wholesale spot electricity price for the United States.

Table 5.4: Actual expenses incurred over the two-year evaluation period by using various M2M configurations for decision making at the 70 m^3/s inflow anomaly threshold. Expenses are calculated for $\alpha = 0.036$ using the probability threshold, p_t, of 0.04. The loss L incurred for each missed forecast at this threshold is \$216,543.

	Ensemble configuration	Hit (a)	False Alarm (b)	Miss (c)	E_{actual}	Cost estimate (\$)
Day 1	Full Ensemble	17	35	0	$1.87L$	405,400
	–MGS	17	33	0	$1.80L$	389,800
	–MM	17	33	0	$1.80L$	389,800
	–MSP	17	36	0	$1.91L$	413,200
	–DH (WFLD)	16	37	1	$2.91L$	629,700
	–DH (WaSiM)	16	39	1	$2.98L$	645,300
	–NWP (LR)	17	35	1	$2.87L$	621,900
	–NWP (HR)	15	38	2	$3.91L$	846,300
Day 2	Full Ensemble	13	41	4	$5.94L$	1,287,100
	–MGS	13	46	4	$6.12L$	1,326,100
	–MM	15	50	2	$4.34L$	939,800
	–MSP	13	46	4	$6.12L$	1,326,100
	–DH (WFLD)	12	43	5	$6.98L$	1,511,500
	–DH (WaSiM)	14	64	3	$5.81L$	1,257,700
	–NWP (LR)	15	53	2	$4.45L$	963,200
	–NWP (HR)	14	52	3	$5.38L$	1,164,100
Day 3	Full Ensemble	11	45	5	$7.02L$	1,519,300
	–MGS	10	56	6	$8.38L$	1,813,800
	–MM	12	52	4	$6.30L$	1,365,100
	–MSP	12	59	4	$6.56L$	1,419,700
	–DH (WFLD)	8	51	8	$10.12L$	2,192,300
	–DH (WaSiM)	13	85	3	$6.53L$	1,413,600
	–NWP (LR)	12	60	4	$6.59L$	1,427,500
	–NWP (HR)	12	55	4	$6.41L$	1,388,500

At the 70 m^3/s threshold, Table 5.4 shows reduced costs associated with the –MGS ensemble at the 1-day forecast horizon. This is in agreement with results from Figure 5.4 and Figure 5.5, which show the –MGS ensemble to be at least as good as the Full ensemble at this lead time. For forecast horizons of 2 and 3 days, ensembles with lower costs than the Full ensemble are those with high spread (e.g., –MM, –NWP and –DH (WaSiM)). These forecasting systems have fewer forecast misses but also a significantly greater number of false alarms, which are not heavily penalized because of the low cost-loss ratio.

Table 5.5: Actual expenses incurred over the two-year evaluation period by using various M2M forecasts for decision making at the 100 m^3/s inflow anomaly threshold. Expenses are calculated for $\alpha = 0.075$ using a probability threshold, p_t, of 0.08. L for this threshold is $309,348.

	Ensemble configuration	Hit (a)	False Alarm (b)	Miss (c)	E_{actual}	Cost estimate ($)
Day 1	Full Ensemble	7	7	1	$2.05L$	634,200
	–MGS	5	8	3	$3.98L$	1,229,700
	–MM	5	11	3	$4.20L$	1,299,300
	–MSP	7	10	1	$2.28L$	703,800
	–DH (WFLD)	6	12	2	$3.35L$	1,036,300
	–DH (WaSiM)	6	9	2	$3.13L$	966,700
	–NWP (LR)	5	10	3	$4.13L$	1,276,100
	–NWP (HR)	6	9	2	$3.13L$	966,700
Day 2	Full Ensemble	6	12	2	$3.35L$	1,036,300
	–MGS	5	15	3	$4.50L$	1,392,100
	–MM	5	13	3	$4.35L$	1,345,700
	–MSP	6	15	2	$3.58L$	1,105,900
	–DH (WFLD)	6	15	2	$3.58L$	1,105,900
	–DH (WaSiM)	5	17	3	$4.65L$	1,438,500
	–NWP (LR)	6	14	2	$3.50L$	1,082,700
	–NWP (HR)	5	10	3	$4.13L$	1,276,100
Day 3	Full Ensemble	4	12	4	$5.20L$	1,608,600
	–MGS	4	13	4	$5.28L$	1,631,800
	–MM	5	12	3	$4.28L$	1,322,500
	–MSP	4	12	4	$5.20L$	1,608,600
	–DH (WFLD)	4	7	4	$4.83L$	1,492,600
	–DH (WaSiM)	4	19	4	$5.73L$	1,771,000
	–NWP (LR)	4	17	4	$5.58L$	1,724,600
	–NWP (HR)	5	8	3	$3.98L$	1,229,700

At the 100 m^3/s threshold, the Full ensemble is markedly superior to the reduced configurations for day 1 forecasts, and slightly less so for day 2 forecasts. At a lead time of 3 days, some of the M2M configurations that neglect NWP uncertainty lead to cost reductions of up to 24%. The fact that forecast misses are reduced without inflation of false alarms suggests that these typically high-spread ensembles have a tendency to underforecast extreme events. Observed anomaly inflows above the 100 m^3/s threshold during the case-study period are all driven by precipitation events with strong orographic gradients and, in two cases, rain-on-snow contributions. This again points

to the importance of including NWP uncertainty in the ensemble, particularly as forecast lead time increases. Unfortunately, given the small sample size evaluated for this inflow anomaly threshold, it is difficult to draw any strong conclusions.

Note that the relative costs in Tables 5.4 and 5.5 do not necessarily correspond to the maximum value curves shown in Figures 5.7 and 5.8. Despite all of the ensembles being calibrated as measured over all inflows, calibration deviation can still exist for particular thresholds. This may lead to non-optimal decision making when action is taken at $p_t = \alpha$. For example, Figure 5.7 suggests that users taking action using day 2 forecasts with $\alpha = 0.04$ will achieve the least forecast value by using the Full ensemble, and that value would be maximized by using one of –MM, –DH (WFLD), or –NWP (LR). However, Table 5.4 reveals that due to calibration deficiencies in all of the ensembles at this high threshold (which may be artifacts of sampling limitations), the Full ensemble actually has less associated cost than several configurations, and –DH (WFLD) has the highest cost.

For low α, it follows from Eq. (5.3) that false alarms result in very little decrease in value, whereas forecast misses result in significant costs. As discussed by Murphy (1994), it is impossible to know the economic value of forecasts to a particular user unless you know that a forecast resulted in the user taking action, have detailed knowledge of the decision-making processes of the user, and know the skill of the forecasts. Thus, it is reasonable, given the assumptions made to simplify the Daisy Lake cost-loss model, to assume that the decision maker won't necessarily take action at the low probabilities used in Tables 5.4 and 5.5. In spite of the economic value analysis shown here, the high number of false alarms issued by some of the forecasting systems would likely lead the reservoir operator to begin to ignore these warnings, which would result in significant losses when the event finally did occur.

5.5 Conclusions

It has been argued that in order to produce truly probabilistic forecasts of hydrologic phenomena, it is necessary to sample all sources of error (Krzysztofowicz, 2001). Unfortunately, accounting for each these sources of error comes at a price, either in terms of money or time spent. Therefore, this chapter has been devoted to an economic analysis of each error-sampling component of the M2M ensemble.

In order to evaluate the economic value of each component of the M2M ensemble, a simple, static cost-loss model has been applied to simulate a hypothetical reservoir operator at the Daisy Lake reservoir who is assumed to be sensitive to certain inflow thresholds. By modelling the decision-making process based on the use of the Full ensemble and on ensemble configurations with individual components removed, it is possible to draw some general conclusions about the

relative value of each component. This value provided to the forecast end user does not include any reductions in value due to computational, time, or monetary costs associated with the various ensemble components. It is therefore up to the individual forecast end user to weigh the value of each ensemble component against the price paid for that component.

The impacts of some M2M components cannot be isolated. For example, the cost difference between the Full ensemble and the –DH (WFLD) configuration cannot be attributed solely to the addition or removal of the WaSiM hydrologic model because each model uses different schemes to downscale the driving NWP fields, and the differently-optimized parameter sets and initial conditions for the models are not equivalent. Thus, it is only possible to make general observations about the relative importance of different types of ensemble configurations.

Based on a comparison of actual operating costs associated with taking action using the variety of M2M ensemble configurations, it can be seen that for inflow anomalies above the 70 m^3/s threshold, the inclusion of multiple distributed hydrologic models is worth approximately $120,000 annually for forecast lead times of 1 day (excluding the price paid for this ensemble component). Other configurations give inconclusive results. For the 100 m^3/s threshold, sampling aspects of NWP uncertainty is, on average, worth $280,000/year at a lead time of 1 day, and $119,000/year at a lead time of 2 days. DH uncertainty at this threshold is worth slightly less, at $180,000 and $118,000 annually for forecast days 1 and 2, respectively. The MSP component is worth $35,000/year for both lead times at the 100 m^3/s threshold. This analysis could be used to determine whether the case-specific price associated with including multiple DH models and/or multi-model or multi-grid scale NWP ensembles are offset by the benefits provided by their inclusion.

This economic analysis, though based on a relatively small evaluation sample size, can be used in conjunction with more robust metrics of deterministic and probabilistic forecast quality and skill, to draw some useful conclusions for probability forecasting in the case-study watershed. Across all lead times, the Full M2M ensemble is generally superior to any of the reduced ensemble configurations. Exclusion of the multi-grid scale NWP ensemble component is not detrimental at lead times of one day, though the importance of high-resolution NWP model output through the use of a multi-grid scale ensemble is apparent as lead time increases. Including a multi-hydrologic model component is important at all lead times, and is more important than including multiple parameter sets and hydrologic states (which cannot be separated in this particular case). Forecast false alarms can be avoided by forecast PDFs that do not overestimate spread. In this case study, the Full M2M ensemble is superior to the other configurations in its skill in predicting appropriate forecast spread using the EMOS uncertainty model. Note that while larger ensemble sizes tend to perform better in this respect, ensemble size does not appear to be the primary driver; the –NWP configurations have the smallest ensemble sizes, whereas the distributional spread predicted by the –DH (WaSiM)

configuration is often the greatest. Ensemble mean errors, which are used by EMOS to increase PDF spread when ensemble variance is inadequate, appear to be more important in this respect.

Uncertainty in hydrologic model predictions has led to numerous recommendations for the quantification of this uncertainty and the use of probabilistic forecasting frameworks (e.g., Kitanidis and Bras, 1980; Krzysztofowicz, 2001; Beven, 2006; Liu and Gupta, 2007; De Roo et al., 2011). The use of weather ensembles has been shown to provide added skill for flood warning as compared to deterministic forecasts (e.g., Gouweleeuw et al., 2005; Roulin and Vannitsem, 2005; Thirel et al., 2010), and Pappenberger et al. (2008) have shown that super-ensembles that combine the ensembles of various forecasting centres have significant added reliability and value. Ensemble mean forecasts derived from multiple hydrologic models have been demonstrated to have overall superior performance to the best individual member, even when models with non-optimized parameters are included (e.g., Shamseldin et al., 1997; Coulibaly et al., 2005; Ajami et al., 2006). Forecast quality has also been shown to improve with the inclusion of different hydrologic states (e.g., McMillan et al., 2013), differently-optimized hydrologic model parameterizations (e.g., Duan et al., 2007; Vrugt et al., 2003a,b), and combinations of the two (Moradkhani et al., 2005b; Vrugt et al., 2005).

The results presented in this chapter, while specific to the case-study watershed and the models applied, are generally in agreement with the above-cited literature. That is, the importance of sampling all sources of error is clear in terms of forecast quality, skill, and to a lesser extent, value. This study has gone a step beyond its predecessors and attempted to assign an actual monetary value to the importance of sampling the various sources of uncertainty. Whether or not each ensemble component is worthwhile is likely to be case-dependent, but it is anticipated that for rainfall-driven flows, the inclusion of multiple NWP models will be worthwhile. The necessity for potentially costly high-resolution NWP model output may be restricted to predictions in complex terrain where orographic effects are particularly important or in regions subject to convective rainfall. Including multiple hydrologic models is likely to be advantageous in many applications. In applications where ensemble data assimilation methods are feasible and provide hydrologic state estimates of high quality (e.g., Andreadis and Lettenmaier, 2006; Clark et al., 2008; Pauwels and De Lannoy, 2006), they may contribute added value. This is an area for potential future exploration.

Chapter 6

Conclusions

The goal of this dissertation was to generate reliable probabilistic forecasts of inflow for a hydroelectric reservoir in complex terrain. This has been achieved through a combination of explicit and implicit sampling of the various sources of uncertainty in the hydrologic modelling chain.

6.1 Summary of Methods and Procedures

Predictions derived from hydrologic models are uncertain due to errors in: (1) the hydrologic model's structure; (2) the meteorological data (observed or modelled) used to drive the hydrologic model; (3) the initial conditions or hydrologic state used to start the forecast run; and (4) the parameterization of the hydrologic models.

In order to generate a probabilistic forecast of reservoir inflows to the Daisy Lake reservoir located in southwestern British Columbia, Canada, a forecasting framework has been created in which all of these sources of error are explicitly sampled. The probabilistic forecasting framework was built incrementally throughout the dissertation:

- Reservoir inflow forecast uncertainty stemming from the hydrologic models themselves and the meteorological data used to drive these models was addressed in Chapter 2. This was done by using the individual members of a multi-model, multi-grid scale numerical weather prediction (NWP) ensemble to drive two different distributed hydrologic models. Uncertainty introduced into the modelling chain by the procedures used to downscale the driving data from NWP to hydrologic model grid scale was sampled by using multiple interpolation schemes for the lowest resolution NWP fields, where this uncertainty is greatest.

- In Chapter 3, this Member-to-Member (M2M) forecasting system was expanded to account for errors introduced via hydrologic model parameterization and the hydrologic state or initial conditions used to begin each daily inflow forecast. Parameter uncertainty was sampled by optimizing inflows simulated by the two hydrologic models using three different objective functions to improve different aspects of inflow forecast quality. The multi-state component

of this expanded M2M ensemble arises as a direct result of the multi-parameter component because of the way in which the daily hydrologic state is updated.

- Prior to combining the individual ensemble members into ensemble mean or probabilistic forecasts, a simple bias correction scheme was applied to each ensemble member to remove systematic errors introduced by the dynamical models. The bias correction factor was determined by comparing the total forecasted inflow volume over a training period of specified length to the corresponding total observed inflow volume. A variety of training period lengths were tested in Chapter 2.

- The resulting 72-member inflow forecast ensemble was transformed into a probabilistic forecasting system in Chapter 4 by applying suitable uncertainty models. The uncertainty models fit a probability distribution function (PDF) to the ensemble whereby the spread of the distribution was related to the variance of the ensemble members and recent ensemble mean errors.

- An intelligent probability calibration scheme was applied to the probability forecasts to improve reliability during periods when the uncertainty model produced forecasts deemed to be sufficiently uncalibrated. The probability calibrator relabels forecast probabilities based on the distribution of probability integral transform (PIT) values over some past training period.

- The price paid for generating an ensemble hydrologic forecasting system that explicitly samples all sources of uncertainty in the modelling chain may be excessive for operational forecasting applications. Therefore, the approximate economic value added by each of the M2M components developed in Chapters 2 and 3 was estimated in Chapter 5 using a simple cost-loss decision model adapted specifically for the hydroelectric energy sector.

6.2 Summary of Findings

A number of findings were made based on evaluation of the probabilistic inflow forecasting system and its various components:

- The addition of the multi-state, multi-parameter M2M components to the multi-NWP, multi-distributed-hydrologic-model ensemble increases forecast resolution by improving the forecasting "engine" that generates the ensemble. This is the most important aspect of ensemble forecast quality because unlike reliability, it cannot be corrected using probability calibration methods. The full 72-member M2M ensemble was found to be underdispersive in spite of attempting to account for all sources of error (Chapter 3).

- Bias in the hydrologic state used to begin each daily inflow forecast was found to be the primary source of bias in the forecast. Because of this, and the flashy mountainous nature of the case study watershed, a short bias-correction training window of three days was found to be ideal for correcting forecast bias and other measures of deterministic (ensemble mean) forecast quality. The bias corrector also significantly improved ensemble forecast resolution and discrimination. A degree-of-mass-balance bias correction scheme that weights more recent information more heavily performed better than a scheme with equal weighting (Chapter 2).
- In the case study watershed, forecast error characteristics were found to change with the seasons. During the fall-winter storm season, errors are approximately log-normally distributed; a log-normal PDF fitted to the ensemble during this period produced reliable forecasts. Spring-summer inflows are driven by snowmelt, and since forecast errors are normally distributed during this season, reliable forecasts were generated using a simple Gaussian uncertainty model. These ensemble model output statistics (EMOS) uncertainty models were able to correct for the spread-deficiency of the M2M ensemble (Chapter 4).
- Since the probabilistic forecasts were already well calibrated, the PIT-based calibrator had a tendency to increase forecast ignorance scores by introducing sampling error. Examination of forecast error characteristics led to the development of an alternative "carry-forward" calibration strategy that was able to improve forecast sharpness and therefore decrease ignorance during the warm season (Section 4.4.2).
- A sensitivity comparison of the full 72-member M2M ensemble forecasting system to alternative M2M configurations with individual ensemble components removed revealed that explicit sampling of all sources of error improves many facets of forecast quality. At short lead times, a multi-NWP ensemble component was not critical, but was found to become important with increasing lead time (Chapter 5).
- Using a simple cost-loss decision model, NWP uncertainty sampling was found to have the greatest economic value for management of the case study watershed, followed by uncertainty introduced by hydrologic model structure. Ensembles with poor spread-skill relationship were found to have high value in spite of often issuing forecast false alarms, which were not heavily penalized by the cost-loss model (Chapter 5).

6.3 Potential Applications

The methods outlined in this dissertation are simple and can be easily adopted for any number of hydrologic modelling applications. Predictions in ungauged basins (PUB) are highly uncertain, as

they are generally based on the premise that data from a gauged basin can be applied in other locations (Sivapalan et al., 2003). The M2M error sampling approach would be a viable method for estimating uncertainty in PUBs because of its relatively small data requirements. Another hydrologic application for the ensemble methods discussed herein is in predicting the impacts of climate change on hydrologic processes using global climate model (GCM) output. Uncertainty in GCM prediction is caused by errors in estimations of future greenhouse gas emissions, climate sensitivity, regional responses, and changes in the intensity and frequency of weather extremes (Eckhardt and Ulbrich, 2003; Wilby et al., 2006). The choices made in estimating these model parameters as well as the choice of global climate model result in a broad range of possible future scenarios (Christensen et al., 2004). In cases where climate change may result in land cover changes (e.g., glacier retreat or changes in vegetation), uncertainty in land cover data used by the hydrologic models should also be incorporated. Note that in climate change applications, statistical post-processing methods would need to assume stationarity, as adaptive updating of correction parameters would be impossible. Statistical corrections based on past data may be invalid in future climates (Hay et al., 2002; Fowler et al., 2007).

Applications of the M2M probabilistic forecasting framework are certainly not limited to hydrology. All uncertain forecasts should be expressed probabilistically in order to convey this uncertainty; any forecasting system involving multiple uncertain components could make use of the simple error-sampling strategy employed in this dissertation, particularly if these uncertainties interact non-linearly. As an example, consider modelling the transport of forest fire smoke. These forecasts are subject to uncertainty in the plume model structure (for example, whether it tracks large-scale puffs or individual particles), in the meteorological forecasts used to drive the plume model, in the assimilation of forest fire data and errors in fire fuel loading estimates and the resulting emissions forecasts. A M2M ensemble strategy would be a suitable and relatively simple method of estimating uncertainty in these and other air-quality forecasts.

6.4 Limitations and Recommendations for Future Work

The primary limitation of the analysis in this dissertation is the short forecast lead-time afforded by the high-resolution NWP models used in the M2M ensemble. Building on these results, the next stage of research should examine the relative importance of various sources of inflow forecast error at longer lead times. This would be done by adding medium-range NWP forecasts to the M2M ensemble.

At longer lead times, it is anticipated that a different bias correction strategy from that developed in Chapter 2 would be necessary. As forecast horizon increases, end forecast bias arising from NWP

models will likely begin to outweigh that caused by bias in hydrologic state. Longer bias correction training windows may be necessary to correct the NWP errors (e.g., McCollor and Stull, 2008a). Different bias correction schemes should also be tested, such as seasonal degree-of-mass-balance (DMB) calculation or the robust best easy systematic estimator of Woodcock and Engel (2005), both of which are suitable for correcting daily hydrometeorological forecasts (McCollor and Stull, 2008a).

Another expansion of the M2M ensemble forecasting framework that warrants examination is the addition of conceptual and soft computing hydrologic models. Because of the choices made by the developers of physically oriented models such as WaSiM and WATFLOOD, each is good at simulating different parts of the hydrologic cycle. In addition, the subjective choices made in developing models based on soft computing approaches (e.g., auto-regressive methods, artificial neural networks, and fuzzy expert systems) can result in models that perform well at different times (Han et al., 2007). By combining predictions from different models and from different modelling approaches, it is possible to take advantage of the expertise of each of them, theoretically resulting in better overall predictive capabilities.

The methods used in this dissertation to sample uncertainty arising from hydrologic model parameterization and hydrologic state were extremely simple and likely contributed to the lack of dispersiveness exhibited by the full M2M ensemble. Future work should evaluate the merits of more advanced methods of ensemble data assimilation and parameter optimization in a M2M forecasting framework.

The use of data assimilation methods such as ensemble Kalman filtering and particle filters (e.g., Moradkhani et al., 2005b,a; Moradkhani and Sorooshian, 2009; DeChant and Moradkhani, 2011b; Leisenring and Moradkhani, 2011) is confounded by a lack of observed hydrologic state data within the case study watershed. It is anticipated that the use of such methods would greatly improve the sampling of initial condition uncertainty in the M2M framework, and could potentially correct the underdispersiveness of the ensemble developed in this study. The deployment of additional observing stations within the watershed would make such methods feasible (though their computational complexity is of concern), and could potentially result in forecast (and therefore economic) improvements that outweigh the cost of installation and maintenance of these stations.

Other ensemble parameter optimization methods are also available, including the Shuffled Complex Evolution Metropolis algorithm (SCEM-UA; Vrugt et al., 2003b) and its extension, the Multi-Objective Shuffled Complex Evolution Metropolis algorithm (MOSCEM-UA; Vrugt et al., 2003a). Such methods are based on the idea that a search of the feasible parameter space near the optimum parameter set will reveal many sets that are equally capable of producing simulations and forecasts of high quality. Whether it is preferable to perturb parameter values around their optimum

values or to use different objective functions to optimize an ensemble of parameterizations is an area needing further research. Additionally, the dynamically dimensioned search algorithm used in this dissertation was developed specifically for high-dimensional optimization problems associated with distributed hydrologic models, and may be hard to beat in practice (Tolson and Shoemaker, 2007). Dual state-parameter estimation methods that allow parameters to evolve in time could be employed for more complete handling of parameter and initial condition uncertainty if additional observed hydrologic state data were available within the case study watershed (Moradkhani et al., 2005a,b; DeChant and Moradkhani, 2011b; Leisenring and Moradkhani, 2011). Parameter optimization that incorporates knowledge of climate signals such as the El Niño-Southern Oscillation state may also be a worthwhile area of future research, as recommended in Chapter 2.

Another likely contributor to the underdispersiveness of the M2M ensemble is its (implicit) assumption of perfect observations. Ideally, model input uncertainty should be incorporated into the model parameter optimization procedure to avoid biased or misleading model output (Kavetski et al., 2006a). Data uncertainty can be incorporated into hydrologic modelling using Bayesian methods, but computational complexity is a concern (Kavetski et al., 2006b). Neglect of observational error has an additional impact on forecast verification — an impact that could actually have an opposite effect on apparent M2M dispersiveness. That is, it is entirely possible that after accounting for errors in the calculated Daisy Lake inflow 'observations', they would fall within the bounds of the M2M ensemble more often.

Determination of candidate PDF shapes in Chapter 4 was based on analysis of empirical ensemble mean forecast error distributions over one year. The storm-season forecast error distribution was found to have a very high peak and a slight positive skew. Based on this and on a review of the literature on probabilistic hydrologic modelling, the log-normal distribution was selected to model the forecast PDF during the storm season, and the method did indeed produce reliable forecasts. An area of potential future study should include testing the performance of other PDF shapes such as the Gamma or Weibull distributions (Wilks, 2006). Alternatively, the Gaussian PDF could be used following data reexpression using a power transformation such as the Box-Cox transformation (Box and Cox, 1964).

Since Chapter 4 is the first application of the inteliCal probability calibrator, it is not yet known how large the calibration deviation should be relative to the expected deviation before the PIT-based calibration is applied. This sensitivity is controlled by the inteliCal adjustment factor (ICF) in Eq. (4.9), whereby higher values of ICF result in less frequent calibration. For nearly-calibrated storm season forecasts in this case study, an ICF of approximately 1.67 seems to balance the apparently competing objectives of improving calibration without increasing ignorance. During the warm season, when calibration is good but ignorance is high, an ICF of 1.0 to 1.43 provides

great improvements to forecast ignorance. While an ICF in the range of 1.43 to 1.67 therefore appears to be a suitable compromise for maximizing both storm season and warm season forecast quality, further testing of the inteliCal scheme is required to determine whether these results are case-specific.

Another limitation of Chapter 4 is the way in which the warm season and storm season were defined, and therefore how the uncertainty model was changed between seasons. The strategy employed (whereby the models were switched on pre-defined dates based on climatological flow characteristics) likely had very little impact on the verification metrics in Chapters 4 and 5. However, the change in forecasting system, if not correctly timed, could result in non-optimal forecasts with significant impacts on reservoir operation. An alternative would be to change the uncertainty model when flows are observed to have undergone the transition between seasons.

The simple (static) cost-loss decision model used in Chapter 5 to estimate the economic value provided by the various M2M components exhibited a tendency to give high value to forecasts with large uncertainty bounds. These forecasting systems give fewer forecast misses than their sharper counterparts, but also issue many forecast false alarms. Making use of a wide variety of verification scores can help to weed out such poor forecasting systems. However, estimates of economic value should ideally incorporate more knowledge about the reservoir operator's decision-making process and how they react to false alarms. It is likely that after many instances of the forecasting system "crying wolf", the decision maker will begin to ignore such forecasts. This could result in large economic losses when the event finally did occur.

The analysis in Chapter 5 would also benefit from a more robust verification sample size for operationally significant inflows. This would require a longer record of forecasts and observations. This analysis also ignored costs associated with setting up the full M2M ensemble forecasting system (e.g., time spent on model setup, price paid for high-resolution NWP forecasts), because they were a non-issue in this research setting. Future economic analyses of this type for operational forecasting will require an estimate of the price paid for each component in order to determine the feasibility of the M2M approach.

Verification in this dissertation was based on a comparison of daily average inflow rates. This was necessary given the "observed" inflow data available for verification. These observations are actually calculated using a water balance based on observed reservoir levels and outflows; the hourly data are extremely noisy. Daily averages are of much higher quality and therefore suitable for verification. The use of daily averages likely results in an inflation of apparent forecast quality relative to what might be achieved by verifying over shorter averaging periods. Indeed, for quantitative precipitation forecasts (QPF) derived from high-resolution NWP models, lengthening the accumulation period used for verification significantly increases QPF skill because timing differences become less

important (Stensrud and Yussouf, 2007; Mass et al., 2002). Hydrologic forecasting applications that make use of high-resolution NWP forecasts and that require forecasts with sub-daily time steps should consider these caveats when carrying out forecast evaluation.

Bibliography

Ajami, N. K., Q. Duan, X. Gao, and S. Sorooshian, 2006: Multimodel combination techniques for analysis of hydrological simulations: Application to distributed model intercomparison project results. *Journal of Hydrometeorology*, **7**, 755–768.

Alila, Y. and J. Beckers, 2001: Using numerical modeling to address hydrologic forest management issues in British Columbia. *Hydrological Processes*, **15**, 3371–3387.

Anderson, E. A., 1973: *National weather service river forecast system - snow accumulation and ablation model*. Tech. Rep. NOAA Technical Memorandum NWS Hydro-17, U.S. Department of Commerce, 217 pp.

Anderson, J. L., 1996: A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *Journal of Climate*, **9**, 1518–1530.

Andreadis, K. M. and D. P. Lettenmaier, 2006: Assimilating remotely sensed snow observations into a macroscale hydrology model. *Advances in Water Resources*, **29**, 872–886.

Benoit, R., M. Desgagné, P. Pellerin, S. Pellerin, Y. Chartier, and S. Desjardins, 1997: The Canadian MC2: A semi-Lagrangian, semi-implicit wideband atmospheric model suited for finescale process studies and simulation. *Monthly Weather Review*, **125**, 2382–2415.

Beven, K., 2006: On undermining the science? *Hydrological Processes*, **20**, 3141–3146.

Beven, K. and A. Binley, 1992: The future of distributed models: Model calibration and uncertainty prediction. *Hydrological Processes*, **6**, 279–289.

Beven, K. J. and M. J. Kirkby, 1979: A physically based, variable contributing area model of basin hydrology. *Hydrol. Sci. Bull.*, **24**, 43–69.

Binley, A. M., K. J. Beven, A. Calver, and L. G. Watts, 1991: Changing responses in hydrology: Assessing the uncertainty in physically based model predictions. *Water Resources Research*, **27**, 1253–1261.

Bourdin, D. R., S. W. Fleming, and R. B. Stull, 2012: Streamflow modelling: A primer on applications, approaches and challenges. *Atmosphere-Ocean*, **50**, 507–536.

Box, G. E. P. and D. R. Cox, 1964: An analysis of transformations. *Journal of the Royal Statistical Society*, **26**, 211–252.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, **78**, 1–3.

Brocker, J. and L. A. Smith, 2007: Increasing the reliability of reliability diagrams. *Weather and Forecasting*, **22**, 651–661.

Brun, S. E. and L. E. Band, 2000: Simulating runoff behavior in an urbanizing watershed. *Computers, Environment and Urban Systems*, **24**, 5–22.

Buizza, R., 1997: Potential forecast skill of ensemble prediction and spread and skill distributions of the ECMWF ensemble prediction system. *Monthly Weather Review*, **125**, 99–119.

Candille, G., S. Beauregard, and N. Gagnon, 2010: Bias correction and multiensemble in the NAEFS context or how to get a "free calibration" through a multiensemble approach. *Monthly Weather Review*, **138**, 4268–4281.

Carpenter, T. M. and K. P. Georgakakos, 2006: Intercomparison of lumped versus distributed hydrologic model ensemble simulations on operational forecast scales. *Journal of Hydrology*, **329**, 174–185.

Chow, V. T., 1954: The log-probability law and its engineering applications. *Proceedings of the American Society of Civil Engineers*, **80**, 536–1–536–25.

Christensen, N. S., A. W. Wood, N. Voisin, D. P. Lettenmaier, and R. N. Palmer, 2004: The effects of climate change on the hydrology and water resources of the Colorado River basin. *Climatic Change*, **62**, 337–363.

Clark, M. P., D. E. Rupp, R. A. Woods, X. Zheng, R. P. Ibbitt, A. G. Slater, J. Schmidt, and M. J. Uddstrom, 2008: Hydrological data assimilation with the ensemble Kalman filter: Use of streamflow observations to update states in a distributed hydrological model. *Advances in Water Resources*, **31**, 1309–1324.

Cloke, H. L. and F. Pappenberger, 2009: Ensemble flood forecasting: A review. *Journal of Hydrology*, **375**, 613–626.

Coulibaly, P., 2003: Impact of meteorological predictions on real-time spring flow forecasting. *Hydrological Processes*, **17**, 3791–3801.

Coulibaly, P., M. Haché, V. Fortin, and B. Bobée, 2005: Improving daily reservoir inflow forecasts with model combination. *Journal of Hydrologic Engineering*, **10**, 91–99.

Daly, C., 2006: Guidelines for assessing the suitability of spatial climate data sets. *International Journal of Climatology*, **26**, 707–721.

Day, G. N., 1985: Extended streamflow forecating using NWSRFS. *Journal of Water Resources Planning and Management*, **111**, 157–170.

De Roo, A., et al., 2011: Quality control, validation and user feedback of the European Flood Alert System (EFAS). *International Journal of Digital Earth*, **4**, 77–90.

DeChant, C. M. and H. Moradkhani, 2011a: Improving the characterization of initial condition for ensemble streamflow prediction using data assimilation. *Hydrology and Earth System Sciences*, **15**, 3399–3410.

DeChant, C. M. and H. Moradkhani, 2011b: Radiance data assimilation for operational snow and streamflow forecasting. *Advances in Water Resources*, **34**, 351–364.

Dettinger, M. D., D. R. Cayan, H. F. Diaz, and D. M. Meko, 1998: North-south precipitation patterns in western North America on interannual-to-decadal timescales. *Journal of Climate*, **11**, 3095–3111.

Draper, A. J., M. W. Jenkins, K. W. Kirby, J. R. Lund, and R. E. Howitt, 2003: Economic-engineering optimization for California water management. *Journal of Water Resources Planning and Management*, **129**, 155–164.

Druce, D. J., 1990: Incorporating daily flood control objectives into a monthly stochastic dynamic programming model for a hydroelectric complex. *Water Resources Research*, **26**, 5–11.

Duan, Q., N. K. Ajami, X. Gao, and S. Sorooshian, 2007: Multi-model ensemble hydrologic prediction using Bayesian model averaging. *Advances in Water Resources*, **30**, 1371–1386.

Ebert, E. E., 2001: Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Monthly Weather Review*, **129**, 2461–2480.

Eckel, F. A. and M. K. Walters, 1998: Calibrated probabilistic quantitative precipitation forecasts based on the MRF ensemble. *Weather and Forecasting*, **13**, 1132–1147.

Eckhardt, K. and U. Ulbrich, 2003: Potential impacts of climate change on groundwater recharge and streamflow in a central European low mountain range. *Journal of Hydrology*, **284**, 244–252.

Environment Canada, 2013: Ensemble forecasts: Definition of the control model and perturbed models. URL http://weather.gc.ca/ensemble/verifs/model_e.html, retrieved May 22, 2013.

Erven, L. N., 2012: *An observational study of slope air and free air wintertime temperatures in Whistler Valley, British Columbia, Canada*. M.S. thesis, Department of Earth and Ocean Sciences, University of British Columbia, 112 pp.

Evensen, G., 1994: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research*, **99**, 10143–10162.

Fane, L. A., 2003: *Generalized optimization in the British Columbia hydroelectric system*. M.S. thesis, Department of Civil Engineering, University of British Columbia, 109 pp.

Fleming, S. W., F. A. Weber, and S. Weston, 2010: Multiobjective, manifoldly constrained Monte Carlo optimization and uncertainty estimation for an operational hydrologic forecast model. *American Meteorological Society Annual Meeting*, Atlanta, Georgia.

Fleming, S. W. and P. H. Whitfield, 2010: Spatiotemporal mapping of ENSO and PDO surface meteorological signals in British Columbia, Yukon, and Southeast Alaska. *Atmosphere-Ocean*, **48**, 122–131.

Fleming, S. W., P. H. Whitfield, R. D. Moore, and E. J. Quilty, 2007: Regime-dependent streamflow sensitivities to Pacific climate modes across the Georgia-Puget transboundary ecoregion. *Hydrological Processes*, **21**, 3264–3287.

Fowler, H. J., S. Blenkinsop, and C. Tebaldi, 2007: Linking climate change modelling to impacts studies: recent advances in downscaling techniques for hydrological modelling. *International Journal of Climatology*, **27**, 1547–1578.

Francke, T., 2012: Particle swarm optimization and dynamically dimensioned search, optionally using parallel computing based on Rmpi. URL www.rforge.net/ppso/, retrieved May 13, 2013.

Franz, K. J., H. C. Hartmann, S. Sorooshian, and R. Bales, 2003: Verification of National Weather Service ensemble streamflow predictions for water supply forecasting in the Colorado River Basin. *Journal of Hydrometeorology*, **4**, 1105–1118.

Georgakakos, K. P., D.-J. Seo, H. Gupta, J. Schaake, and M. B. Butts, 2004: Towards the characterization of streamflow simulation uncertainty through multimodel ensembles. *Journal of Hydrology*, **298**, 222–241.

Georgakakos, K. P. and H. Yao, 2001: Assessment of Folsom Lake response to historical and potential future climate scenarios 2. Reservoir management. *Journal of Hydrology*, **249**, 176–196.

Glassheim, E., 1997: Fear and loathing in North Dakota. *Natural Hazards Observer*, **21**, 1–4.

Gneiting, T., F. Balabdaoui, and A. E. Raftery, 2007: Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society*, **69**, 243–268.

Gneiting, T., A. E. Raftery, A. H. Westveld, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, **133**, 1098–1118.

Götzinger, J. and A. Bárdossy, 2008: Generic error model for calibration and uncertainty estimation of hydrological models. *Water Resources Research*, **44**, W00B07.

Gouweleeuw, B. T., J. Thielen, G. Franchello, A. P. J. DeRoo, and R. Buizza, 2005: Flood forecasting using medium-range probabilistic weather prediction. *Hydrology and Earth System Sciences*, **9**, 365–380.

Graeff, T., E. Zehe, T. Blume, T. Francke, and B. Schröder, 2012: Predicting event response in a nested catchment with generalized linear models and a distributed watershed model. *Hydrological Processes*, **26**, 3749–3769.

Green, W. H. and G. A. Ampt, 1911: Studies of soil physics. Part 1: The flow of air and water through soils. *J. Agric. Soc.*, **4**, 1–24.

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

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

Gulliver, J. S., 1991: *Hydraulic conveyance design*, 5.1–5.81. Hydropower Engineering Handbook, McGraw-Hill, Washington, D.C.

Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, **129**, 550–560.

Hamill, T. M. and S. J. Colucci, 1997: Verification of Eta-RSM short-range ensemble forecasts. *Monthly Weather Review*, **125**, 1312–1327.

Hamill, T. M. and S. J. Colucci, 1998: Evaluation of Eta-RSM probabilistic precipitation forecasts. *Monthly Weather Review*, **126**, 711–724.

Hamlet, A. F., D. Huppert, and D. P. Lettenmaier, 2002: Economic value of long-lead streamflow forecasts for Columbia River hydropower. *Journal of Water Resources Planning and Management*, **128**, 91–101.

Hamlet, A. F. and D. P. Lettenmaier, 1999: Effects of climate change on hydrology and water resources in the Columbia River basin. *Journal of the American Water Resources Association*, **35**, 1597–1623.

Han, D., L. Chan, and N. Zhu, 2007: Flood forecasting using support vector machines. *Journal of Hydroinformatics*, **9**, 267–276.

Hargreaves, G. H. and Z. A. Samani, 1982: Estimating potential evapotranspiration. *ASCE J. Irrigation Drainage Div.*, **108**, 225–230.

Hashino, T., A. A. Bradley, and S. S. Schwartz, 2007: Evaluation of bias-correction methods for ensemble streamflow volume forecasts. *Hydrology and Earth System Sciences*, **11**, 939–950.

Hay, L. E., et al., 2002: Use of Regional Climate Model Output for Hydrologic Simulations. *Journal of Hydrometeorology*, **3**, 571–590.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, **15**, 559–570.

Jenkins, M. W., et al., 2001: *Improving California water management: Optimizing value and flexibility*. Tech. rep., Center for Environmental and Resources Engineering, University of California.

Johnson, C. and R. Swinbank, 2009: Medium-range multimodel ensemble combination and calibration. *Quarterly Journal of the Royal Meteorological Society*, **135**, 777–794.

Kavetski, D., G. Kuczera, and S. W. Franks, 2006a: Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory. *Water Resources Research*, **42**, W03407.

Kavetski, D., G. Kuczera, and S. W. Franks, 2006b: Bayesian analysis of input uncertainty in hydrological modeling: 2. Application. *Water Resources Research*, **42**, W03408.

Khan, S., A. R. Ganguly, and S. Saigal, 2005: Detection and predictive modeling of chaos in finite hydrological time series. *Nonlinear Processes in Geophysics*, **12**, 41–53.

Kitanidis, P. K. and R. L. Bras, 1980: Real-time forecasting with a conceptual hydrologic model: 1. Analysis of uncertainty. *Water Resources Research*, **16**, 1025–1033.

Kite, G. W. and N. Kouwen, 1992: Watershed modeling using land classifications. *Water Resources Research*, **28**, 3193–3200.

Klok, E. J., K. Jasper, K. P. Roelofsma, J. Gurtz, and A. Badoux, 2001: Distributed hydrological modelling of a heavily glaciated alpine river basin. *Hydrological Sciences Journal*, **46**, 553–570.

Kouwen, N., 2010: WATFLOOD/WATROUTE hydrological model routing and flow forecasting system. Tech. rep., University of Waterloo. URL www.civil.uwaterloo.ca/watflood/downloads/manual10.pdf, retrieved May 13, 2013.

Kouwen, N., M. Danard, A. Bingeman, W. Luo, F. R. Seglenieks, and E. D. Soulis, 2005: Case study: Watershed modeling with distributed weather model data. *Journal of Hydrologic Engineering*, **10**, 23–38.

Kouwen, N., E. D. Soulis, A. Pietroniro, J. Donald, and R. A. Harrington, 1993: Grouped response units for distributed hydrologic modeling. *Journal of Water Resources Planning and Management*, **119**, 289–305.

Krzysztofowicz, R., 2001: The case for probabilistic forecasting in hydrology. *Journal of Hydrology*, **249**, 2–9.

Krzysztofowicz, R. and L. Duckstein, 1979: Preference criterion for flood control under uncertainty. *Water Resources Research*, **14**, 513–520.

Kuczera, G. and E. Parent, 1998: Monte Carlo assessment of parameter uncertainty in conceptual catchment models: the Metropolis algorithm. *Journal of Hydrology*, **211**, 69–84.

Kurtzman, D. and R. Kadmon, 1999: Mapping of temperature variables in Israel: A comparison of different interpolation methods. *Climate Research*, **13**, 33–43.

Leemans, R. and W. P. Cramer, 1991: *The IIASA database for mean monthly values of temperature, precipitation, and cloudiness on a global terrestrial grid*. Tech. Rep. RR-91-18, International Institute for Applied Systems Analysis.

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

Leisenring, M. and H. Moradkhani, 2011: Snow water equivalent prediction using Bayesian data assimilation methods. *Stoch. Environ. Res. Risk Assess.*, **25**, 253–270.

Lewis, D., M. J. Singer, R. A. Dahlgren, and K. W. Tate, 2000: Hydrology in a California oak woodland watershed: a 17-year study. *Journal of Hydrology*, **240**, 106–117.

Liu, Y. and H. V. Gupta, 2007: Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework. *Water Resources Research*, **43**, W07401.

Liu, Y., et al., 2011: Advancing data assimilation in operational hydrologic forecasting: progresses, challenges, and emerging opportunities. *Hydrology and Earth System Sciences*, **16**, 3863–3887.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, **20**, 130–141.

Madadgar, S., H. Moradkhani, and D. Garen, 2012: Towards improved post-processing of hydrologic forecast ensembles. *Hydrological Processes*, doi:10.1002/hyp.9562.

Mantua, N. J., S. R. Hare, Y. Zhang, J. M. Wallace, and R. C. Francis, 1997: A Pacific interdecadal climate oscillation with impacts on salmon production. *Bulletin of the American Meteorological Society*, **78**, 1069–1079.

Mascaro, G., E. R. Vivoni, and R. Deidda, 2010: Implications of ensemble quantitative precipitation forecast errors on distributed streamflow forecasting. *Journal of Hydrometeorology*, **11**, 69–86.

Mascaro, G., E. R. Vivoni, and R. Deidda, 2011: Impact of basin scale and initial condition on ensemble streamflow forecast uncertainty. *25th Conference on Hydrology*, Seattle, WA.

Mass, C. F., D. Ovens, K. Westrick, and B. A. Colle, 2002: Does increasing horizontal resolution produce more skillful forecasts? *Bulletin of the American Meteorological Society*, **83**, 407–430.

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

McCollor, D. and R. Stull, 2008b: Hydrometeorological short-range ensemble forecasts in complex terrain. Part II: Economic evaluation. *Weather and Forecasting*, **23**, 557–574.

McMillan, H., E. Hreinsson, M. P. Clark, S. K. Singh, C. Zammit, and M. J. Uddstrom, 2013: Operational hydrological data assimilation with the recursive ensemble Kalman filter. *Hydrology and Earth System Sciences*, **17**, 21–38.

Monteith, J. L., 1965: Evaporation and environment. *Symposium of the Society for Experimental Biology*, **19**, 205–234.

Moradkhani, H., K.-L. Hsu, H. V. Gupta, and S. Sorooshian, 2005a: Uncertainty assessment of hydrologic model states and parameters: Sequential data assimilation using the particle filter. *Water Resources Research*, **41**, W05012.

Moradkhani, H., S. Sorooshian, H. V. Gupta, and P. R. Houser, 2005b: Dual state-parameter estimation of hydrological models using ensemble Kalman filter. *Advances in Water Resources*, **28**, 135–147.

Moradkhani, H. and S. Sorooshian, 2009: *General review of rainfall-runoff modeling: Model calibration, data assimilation, and uncertainty analysis*, 1–24. Hydrological Modelling and the Water Cycle, Coupling the Atmospheric and Hydrological Models, Springer, Berlin, Germany.

Moradkhani, H., C. M. DeChant, and S. Sorooshian, 2012: Evolution of ensemble data assimilation for uncertainty quantification using the particle filter-Markov chain Monte-Carlo method. *Water Resources Research*, **48**, W12520.

Murphy, A. H., 1973: A new vector partition of the probability score. *Journal of Applied Meteorology*, **12**, 595–600.

Murphy, A. H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather and Forecasting*, **8**, 293–293.

Murphy, A. H., 1994: Assessing the economic value of weather forecasts: An overview of methods, results and issues. *Meteorological Applications*, **1**, 69–73.

Murphy, A. H. and R. L. Winkler, 1987: A general framework for forecast verification. *Monthly Weather Review*, **115**, 1330–1338.

Nash, J. E. and I. V. Sutcliffe, 1970: River flow forecasting through conceptual models: Part I - A discussion of principles. *Journal of Hydrology*, **10**, 282–290.

Niehoff, D., U. Fritsch, and A. Bronstert, 2002: Land-use impacts on storm-runoff generation: scenarios of land-use change and simulation of hydrological response in a meso-scale catchment in SW-Germany. *Journal of Hydrology*, **267**, 80–93.

Nipen, T., 2012: *A component-based probabilistic weather forecasting system for operational usage*. Ph.D. thesis, Department of Earth, Ocean and Atmospheric Sciences, University of British Columbia, 101 pp.

Nipen, T. and R. Stull, 2011: Calibrating probabilistic forecasts from an NWP ensemble. *Tellus A*, **63**, 858–875.

Olsson, J. and G. Lindström, 2008: Evaluation and calibration of operational hydrological ensemble forecasts in Sweden. *Journal of Hydrology*, **350**, 14–24.

Özelkan, E. C. and L. Duckstein, 2001: Fuzzy conceptual rainfall-runoff models. *Journal of Hydrology*, **253**, 41–68.

Palmer, T. N., 2002: The economic value of ensemble forecasts as a tool for risk assessment: From days to decades. *Quarterly Journal of the Royal Meteorological Society*, **128**, 747–774.

Palmer, T. N., G. J. Shutts, R. Hagedorn, F. J. Doblas-Reyes, T. Jung, and M. Leutbecher, 2005: Representing model uncertainty in weather and climate prediction. *Annual Review of Earth and Planetary Sciences*, **33**, 163–193.

Pappenberger, F., J. Bartholmes, J. Thielen, H. L. Cloke, R. Buizza, and A. De Roo, 2008: New dimensions in early flood warning across the globe using grand-ensemble weather predictions. *Geophysical Research Letters*, **35**, L10404.

Parrish, M. A., H. Moradkhani, and C. M. DeChant, 2012: Toward reduction of model uncertainty: Integration of Bayesian model averaging and data assimilation. *Water Resources Research*, **48**, W03519.

Pauwels, V. R. N. and G. J. M. De Lannoy, 2006: Improvement of modeled soil wetness conditions and turbulent fluxes through the assimilation of observed discharge. *Journal of Hydrometeorology*, **7**, 458–477.

Peschke, G., 1987: Soil moisture and runoff components from a physically founded approach. *Acta Hydrophysica*, **31**, 191–205.

Philip, J. R., 1954: An infiltration equation with physical significance. *Soil Science*, **77**, 153–158.

Pinson, P., P. McSharry, and H. Madsen, 2010: Reliability diagrams for non-parametric density forecasts of continuous variables: Accounting for serial correlation. *Quarterly Journal of the Royal Meteorological Society*, **136**, 77–90.

PRISM Climate Group, 2012: Latest PRISM data. URL http://www.prism.oregonstate.edu/, retrieved May 13, 2013.

Pulido-Velazquez, M., M. W. Jenkins, and J. R. Lund, 2004: Economic values for conjunctive use and water banking in southern California. *Water Resources Research*, **40**, W03401.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, **133**, 1155–1174.

Randrianasolo, A., M. H. Ramos, G. Thirel, V. Andréassian, and E. Martin, 2010: Comparing the scores of hydrological ensemble forecasts issued by two different hydrological models. *Atmospheric Science Letters*, **11**, 100–107.

Reggiani, P., M. Renner, A. H. Weerts, and P. A. H. J. M. van Gelder, 2009: Uncertainty assessment via Bayesian revision of ensemble streamflow predictions in the operational River Rhine forecasting system. *Water Resources Research*, **45**, W02428.

Richardson, D. S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, **126**, 649–667.

Richardson, D. S., 2001: Measures of sklll and value of ensemble predictions systems, their interrelationship and the effect of ensemble size. *Quarterly Journal of the Royal Meteorological Society*, **127**, 2473–2489.

Richardson, D. S., 2003: *Economic Value and Skill*, 165–187. Forecast Verification: A Practitioner's Guide in Atmospheric Science, Wiley, Chichester, England.

Ross, R. S. and T. N. Krishnamurti, 2005: Reduction of forecast error for global numerical weather prediction by the Florida State University (FSU) superensemble. *Meteorology and Atmospheric Physics*, **88**, 215–235.

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

Roulin, E. and S. Vannitsem, 2005: Skill of medium-range hydrological ensemble predictions. *Journal of 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. *Weather and Forecasting*, **21**, 116–122.

Roulston, M. S. and L. A. Smith, 2002: Evaluating probabilistic forecasts using information theory. *Monthly Weather Review*, **130**, 1653–1660.

Schmeits, M. J. and K. J. Kok, 2010: A comparison between raw ensemble output, (modified) Bayesian model averaging, and extended logistic regression using ECMWF ensemble precipitation forecasts. *Monthly Weather Review*, **138**, 4199–4211.

Schulla, J., 2012: *Model description WaSiM (Water balance Simulation Model)*. Tech. rep., Hydrology Software Consulting J. Schulla, 305 pp.

Seo, D.-J., H. D. Herr, and J. C. Schaake, 2006: A statistical post-processor for accounting of hydrologic uncertainty in short-range ensemble streamflow prediction. *Hydrology and Earth System Sciences Discussions*, **3**, 1988–2035.

Shamseldin, A. Y., K. M. O'Connor, and G. C. Liang, 1997: Methods for combining the outputs of different rainfall-runoff models. *Journal of Hydrology*, **197**, 203–229.

Shawwash, Z. K. E., 2000: *A decision support system for real-time hydropower scheduling in a competetive power market environment*. Ph.D. thesis, Department of Civil Engineering, University of British Columbia, 315 pp.

Sivakumar, B., 2000: Chaos theory in hydrology: important issues and interpretations. *Journal of Hydrology*, **227**, 1–20.

Sivakumar, B., R. Berndtsson, and J. Olsson, 2001: Evidence of chaos in the rainfall-runoff process. *Hydrological Sciences Journal*, **46**, 131–145.

Sivapalan, M., et al., 2003: IAHS decade on Predictions in Ungauged Basins (PUB), 2003-2012: Shaping an exciting future for the hydrologic sciences. *Hydrological Sciences Journal*, **48**, 857–880.

Sivillo, J. K., J. E. Ahlquist, and Z. Toth, 1997: An ensemble forecasting primer. *Weather and Forecasting*, **12**, 809–818.

Skamarock, W. C., et al., 2008: *A description of the Advanced Research WRF version 3*. Tech. Rep. TN-475+STR, NCAR, 113 pp.

Spokas, K. and F. Forcella, 2006: Estimating hourly incoming solar radiation from limited meteorological data. *Weed Science*, **54**, 182–189.

Stedinger, J. R., 1980: Fitting log normal distributions to hydrologic data. *Water Resources Research*, **16**, 481–490.

Steinschneider, S. and C. Brown, 2011: Influences of North Atlantic climate variability on low-flows in the Connecticut River basin. *Journal of Hydrology*, **409**, 212–224.

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

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

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

Stull, R. B., 2000: *Meteorology for Scientists and Engineers*. 2d ed., Brooks/Cole Thomson Learning, Pacific Grove, CA, 502 pp.

Talagrand, O., R. Vautard, and B. Strauss, 1997: Evaluation of probabilistic prediction systems. *Proc. Workshop on Predictability*, Reading, United Kingdom, ECMWF, Reading, United Kingdom, 1–25.

Thirel, G., F. Regimbeau, E. Martin, J. Noilhan, and F. Habets, 2010: Short- and medium-range hydrological ensemble forecasts over France. *Atmospheric Science Letters*, **11**, 72–77.

Thirel, G., F. Rousset-Regimbeau, E. Martin, and F. Habets, 2008: On the impact of short-range meteorological forecasts for ensemble streamflow predictions. *Journal of Hydrometeorology*, **9**, 1301–1317.

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

Tolson, B. A. and C. A. Shoemaker, 2007: Dynamically dimensioned search algorithm for computationally efficient watershed model calibration. *Water Resources Research*, **43**, W01413.

Toth, Z., O. Talagrand, G. Candille, and Y. Zhu, 2003: *Probability and Ensemble Forecasts*, 137–163. Forecast Verification: A Practitioner's Guide in Atmospheric Science, Wiley, Chichester, England.

Van den Bergh, J. and E. Roulin, 2010: Hydrological ensemble prediction and verification for the Meuse and Scheldt basins. *Atmospheric Science Letters*, **11**, 64–71.

Vrugt, J. A., C. G. H. Diks, H. V. Gupta, W. Bouten, and J. M. Verstraten, 2005: Improved treatment of uncertainty in hydrologic modeling: Combining the strengths of global optimization and data assimilation. *Water Resources Research*, **41**, W01017.

Vrugt, J. A., H. V. Gupta, L. A. Bastidas, W. Bouten, and S. Sorooshian, 2003a: Effective and efficient algorithm for multiobjective optimization of hydrologic models. *Water Resources Research*, **39**, 1214.

Vrugt, J. A., H. V. Gupta, W. Bouten, and S. Sorooshian, 2003b: A shuffled complex evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters. *Water Resources Research*, **39**, 1201.

Vrugt, J. A. and B. A. Robinson, 2007: Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging. *Water Resources Research*, **43**, W01411.

Wagener, T., 2003: Evaluation of catchment models. *Hydrological Processes*, **17**, 3375–3378.

Wang, Q. J., D. E. Robertson, and F. H. S. Chiew, 2009: A Bayesian joint probability modeling approach for seasonal forecasting of streamflows at multiple sites. *Water Resources Research*, **45**, W05407.

Wang, T., A. Hamann, D. Spittlehouse, and S. N. Aitken, 2006: Development of scale-free climate data for western Canada for use in resource management. *International Journal of Climatology*, **26**, 383–397.

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. *Journal of Hydrometeorology*, **2**, 161–180.

Westrick, K. J., P. Storck, and C. F. Mass, 2002: Description and evaluation of a hydrometeorological forecast system for mountainous watersheds. *Weather and Forecasting*, **17**, 250–262.

Wilby, R. L., P. G. Whitehead, A. J. Wade, D. Butterfield, R. J. Davis, and G. Watts, 2006: Integrated modelling of climate change impacts on water resources and quality in a lowland catchment: River Kennet, UK. *Journal of Hydrology*, **330**, 204–220.

Wilks, D. S., 2001: A skill score based on economic value for probability forecasts. *Meteorological Applications*, **8**, 209–219.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*. 2d ed., Academic Press, 627 pp.

Willmott, C. J. and K. Matsura, 1995: Smart interpolation of annually averaged air temperature in the United States. *Journal of Applied Meteorology*, **34**, 2577–2586.

Wilson, L. J., S. Beauregard, A. E. Raftery, and R. Verret, 2007: Calibrated surface temperature forecasts from the Canadian ensemble prediction system using Bayesian model averaging. *Monthly Weather Review*, **135**, 1364–1385.

Wood, A. W. and J. C. Schaake, 2008: Correcting errors in streamflow forecast ensemble mean and spread. *Journal of Hydrometeorology*, **9**, 132–148.

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

Yoshitani, J., Z. Q. Chen, M. L. Kavvas, and K. Fukami, 2009: Atmospheric model-based streamflow forecasting at small, mountainous watersheds by a distributed hydrologic model: Application to a watershed in Japan. *Journal of Hydrologic Engineering*, **14**, 1107–1118.

Yuan, H., J. A. McGinley, P. J. Schultz, C. J. Anderson, and C. Lu, 2008: Short-range precipitation forecasts from time-lagged multimodel ensembles during the HMT-West-2006 campaign. *Journal of Hydrometeorology*, **9**, 477–491.

Zhao, L., Q. Duan, J. Schaake, A. Ye, and J. Xia, 2011: A hydrologic post-processor for ensemble streamflow predictions. *Advances in Geosciences*, **29**, 51–59.

Zhu, Y., Z. Toth, R. Wobus, D. Richardson, and K. Mylne, 2002: The economic value of ensemble-based weather forecasts. *Bulletin of the American Meteorological Society*, **83**, 73–83.

Appendix A

Forecast Verification Metrics

A.1 Measures-Oriented Verification for Deterministic Forecasts

Let m_t be the deterministic (i.e., ensemble mean or median) forecasted inflow at time t. Scores are calculated over all t in the set of time points T. The size of this set is given by $\|T\|$, which, for the case of daily inflow forecasts, can be interpreted as the number of days over which the forecast is evaluated. The verifying observation is given by x_t, and the mean observed value over all t in T is given by $\bar{x}$.

Degree of mass balance (DMB)

An appropriate measure of bias for volumetric quantities such as precipitation and reservoir inflow is the DMB (McCollor and Stull, 2008a). The DMB is a measure of the ratio of simulated or forecasted inflow to the observed inflow over a given period of time and is given by:

$$DMB = \frac{\sum_{t \in T} m_t}{\sum_{t \in T} x_t}. \tag{A.1}$$

DMB values less than one indicate that inflows are underforecast, while DMB values greater than one indicate a wet forecast bias. A DMB of one is achieved for forecasts that are free of bias as measured over T.

Mean Absolute Error (MAE)

$$MAE = \frac{1}{\|T\|} \sum_{t \in T} |m_t - x_t| \tag{A.2}$$

For perfect forecasts MAE is zero.

Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\frac{1}{\|T\|}\sum_{t\in T}(m_t - x_t)^2} \tag{A.3}$$

Perfect forecasts have an RMSE of zero. This score places more emphasis on large inflow forecast errors than does MAE.

Root Mean Square Error Skill Score (RMSESS)

Forecast skill is determined by comparing the RMSE of the forecast to that of a zero-skill reference forecast ($RMSE_{ref}$). We take the reference forecast to be persistence (i.e., the forecast issued today for all lead times is taken to be yesterday's observed inflow).

$$RMSESS = 1 - \frac{RMSE}{RMSE_{ref}} \tag{A.4}$$

A skill score of zero indicates that the forecast and reference forecast scores are the same and that therefore the forecasting system offers no advantage over persistence. A positive skill score indicates that the forecast has more skill than persistence, while a negative skill score indicates the opposite. A perfect forecasting system has a skill score of one.

Nash-Sutcliffe Efficiency (NSE)

The Nash-Sutcliffe Efficiency (NSE) is an indicator of statistical association (Nash and Sutcliffe, 1970) that gives a score of one for a perfect forecast. An NSE of zero indicates that the model performs no better than a climatological constant forecast given by $\bar{x}$. Emphasis is placed on forecast performance during periods of high inflows.

$$NSE = 1 - \frac{\sum_{t\in T}(x_t - m_t)^2}{\sum_{t\in T}(x_t - \bar{x})^2} \tag{A.5}$$

The NSE of log-transformed flows (LNSE) provides a measure of low-flow forecast quality.

A.2 Distributions-Oriented Verification for Ensemble and Probabilistic Forecasts

In the following, a forecast probability density function (PDF) of variable x, valid at time t is given by $f_t(x)$. The verifying observation is designated as x_t. Scores are calculated for all t in the set of time points T. The size of this set is given by $\|T\|$, which, for the case of daily inflow forecasts, can

be interpreted as the number of days over which the forecast is evaluated.

Reliability

Reliability or calibration (Murphy, 1973) is a measure of consistency between forecast probabilities and the frequency of occurrence of observed values. That is, events forecasted with probability p should, over the course of many such forecasts, be observed to occur a fraction p of the time. A reliable forecasting system exhibits a flat rank histogram or a flat probability integral transform (PIT) histogram (see below), however, a flat histogram is not always a guarantee of a reliable ensemble (Hamill, 2001). Reliability is easily corrected using probability calibration methods (e.g., Hamill and Colucci, 1997; Nipen and Stull, 2011).

Resolution

Resolution is a measure of how well a forecasting system can *a priori* differentiate future weather outcomes such that different forecasts are associated with distinct verifying observations. This is the most important attribute of a forecast system (Toth et al., 2003), as it cannot be improved through adjustment of probability values. Resolution can only be corrected by improving the forecasting "engine" used to generate the ensemble or probability forecasts. This measure of forecast performance is conditioned on the forecasts in that it examines whether or not an event occurred given that it was predicted with a particular probability. The converse of resolution is discrimination, which is conditioned on observations (see discussion of the Relative Operating Characteristic).

Rank Histogram

The rank histogram or Talagrand diagram is used to assess calibration when the Binned Probability Ensemble (BPE) uncertainty model is used to generate probabilistic forecasts (Anderson, 1996; Talagrand et al., 1997). In the BPE uncertainty model, it is assumed that the ensemble members and verifying observation are pulled from the same probability distribution. Therefore, each ensemble member is equally likely, and when the observation and K ensemble members are pooled together and ranked, the rank of the observation is a random integer between 1 and $K + 1$. The rank histogram indicates the frequency with which the observation falls into each bin, defined as the regions between two consecutive ensemble members and above and below the highest and lowest ensemble members. If the BPE assumption is true, and the ensemble members represent the full spread of the probability distribution of the observations, then the probabilities derived from this uncertainty model should be calibrated, and the rank histogram should be flat. A U-shaped diagram indicates a lack of ensemble spread, while L- and J-shaped diagrams indicate over- and under-forecasting biases, respectively.

Probability Integral Transform (PIT) Histogram

The PIT histogram (Gneiting et al., 2005) is analogous to the rank histogram, and is used to assess calibration when a probability forecast is expressed as a fitted PDF. PIT values are given by:

$$P_t = F_t(x_t), \tag{A.6}$$

where F_t is the corresponding forecast cumulative distribution function (CDF). The forecast CDF of variable x at time t is given by:

$$F_t(x) = \int_{-\infty}^{x} f_t(x)dx. \tag{A.7}$$

For perfectly calibrated forecasts, the PIT histogram will be flat, with equal numbers of observations falling into each equally sized bin. The number of bins is arbitrary and not constrained by ensemble size; for our PIT histograms, we divide the interval [0,1] into 10 equally sized bins. If the PIT histogram is not flat, its shape can be used to diagnose problems with the uncertainty model. For example, as with the rank histogram, a U-shaped histogram is an indication of underdispersion, or inadequate spread in the forecast PDF. Note that for both diagrams, flatness is not always a guarantee of calibration, as opposing biases at different times during the evaluation period can result in flat histograms (Hamill, 2001).

Calibration Deviation (D)

A more objective measure of calibration is the calibration deviation metric D of Nipen and Stull (2011), which measures the degree of deviation from a flat PIT histogram:

$$D = \sqrt{\frac{1}{B}\sum_{i=1}^{B}\left(\frac{b_i}{\|T\|} - \frac{1}{B}\right)^2} \tag{A.8}$$

where i is an integer between 1 and the number of bins B and b_i is the bin count or number of observations in bin i. Bin frequencies are given by $b_i\|T\|^{-1}$. Low values of D are preferred, and indicate a small degree of deviation from a flat PIT histogram.

Perfectly reliable forecasts can be expected to exhibit some calibration deviation as a result of sampling limitations (Brocker and Smith, 2007; Pinson et al., 2010). The expected calibration deviation for a perfectly calibrated forecast is given by:

$$E[D_p] = \sqrt{\frac{1 - B^{-1}}{\|T\|B}}. \tag{A.9}$$

When referring to calibration, we will specify a time period over which the calibration metric is computed, and we will not require the forecast to exhibit calibration over shorter time scales. This is important because, as Hamill (2001) points out, a forecast can have different distributional biases during different times of year. Thus, when calibration is computed over a set of time points T, an overforecasting bias during the first half of T combined with an underforecasting bias during the second half can balance to produce a flat histogram.

Ignorance Score (IGN)

While reliability/calibration is a desirable characteristic of probabilistic forecasts, it is not an adequate measure of the usefulness of a forecast. Consider, for example, an uncertainty model that always issues a climatological forecast (i.e., the forecast PDF is always taken as the distribution of the climatological record). Assuming stationarity, such a forecasting system would be perfectly calibrated, but far too vague for decision making. Therefore, we will also require our forecast PDFs (f_t) to concentrate probability in the correct area (i.e., near the verifying observation) on each day. This property can be measured by the ignorance score (Roulston and Smith, 2002), which is defined as:

$$IGN = -\frac{1}{\|T\|}\sum_{t\in T} log_2(f_t(x_t)), \tag{A.10}$$

with lower ignorance scores being preferred. Forecasts are rewarded with low ignorance scores for placing high probability in the vicinity of the verifying observation. Due to the use of the logarithm in the definition of IGN, arithmetic differences between two ignorance scores are more relevant than their ratios.

Nipen (2012) derived a decomposition of the ignorance score for a set of raw forecasts into two parts: (1) the potential ignorance score of a perfectly calibrated forecast (IGN_{pot}), and (2) extra ignorance caused by a lack of calibration (IGN_{uncal}). Ignorance can therefore be reduced by improving the ensemble forecasting system, applying bias correction, or using a more suitable uncertainty model to reduce IGN_{pot}, or by calibrating the forecast to reduce IGN_{uncal}.

Brier Score (BS) and Brier Skill Score (BSS)

The Brier Score (Brier, 1950) is one of the most frequently used evaluation scores for ensemble prediction systems; it is defined as the mean square error of the probability forecast:

$$BS = \frac{1}{\|T\|}\sum_{t\in T}(p_t - x_t)^2, \tag{A.11}$$

where p_t is the probability that the forecasted inflow will exceed a given inflow threshold, and x_t is equal to one if the observed inflow exceeds the threshold, or zero otherwise. The exceedance forecast probability is given by the number of ensemble members that exceed the threshold or the probability of exceedance given by the forecast CDF. A BS of zero indicates a perfect deterministic forecast. The Brier score can also be converted into a skill score:

$$BSS = 1 - \frac{BS}{BS_{ref}} \tag{A.12}$$

where BS_{ref} is the Brier score of a low-skill climatological forecast in which the probability of the event for each forecast is equal to $\bar{x}$. Thus, $BS_{ref} = \bar{x}(1 - \bar{x})$.

The BS can also be decomposed into uncertainty, reliability and resolution components as illustrated by Murphy (1973). Uncertainty is a measure of the difficulty in forecasting the event and depends only on observations. Following decomposition, the BSS can be reformulated as:

$$\begin{aligned} BSS &= \frac{resolution}{uncertainty} - \frac{reliability}{uncertainty} \\ &= Relative\,Resolution - Relative\,Reliability. \end{aligned} \tag{A.13}$$

A perfect forecast has a BSS and relative resolution equal to one and a relative reliability of zero.

Brier scores are calculated for forecast and observation anomaly thresholds relative to climatological inflow values. In order to ensure that the ensemble is not unduly rewarded for making high inflow forecasts during the snowmelt period where little skill is required to do so, we subtract climatology from the forecasts and observations. This daily climatology is derived from the median of observations on each calendar day over the period 1986–2008. A 15-day running mean is then used to generate a smoothed climatology.

Relative Operating Characteristic (ROC)

Given event counts a, b, c and d from Table A.1, the hit rate (H) and false alarm rate (F) are defined as:

$$H = \frac{a}{a + c} \tag{A.14}$$

$$F = \frac{b}{b + d} \tag{A.15}$$

The ROC diagram (Toth et al., 2003) compares H to F at different forecast probability levels and is an indicator of an ensemble's ability to discriminate between the occurrence or non-occurrence of a forecast event (e.g., exceedance of some threshold inflow). Discrimination is the converse of

Table A.1: Contingency table for calculating hit rates and false alarm rates. The number of forecast hits is given by a, b is the number of false alarms, c the number of misses, and d the number of correct rejections.

	Observed	Not observed
Forecast	a	b
Not forecast	c	d

resolution, and examines the probabilities that were predicted conditioned on whether or not an event was observed to occur.

At a probability threshold of $p_t = 0$, an exceedance forecast is issued as long as long as there is at least a 0% chance of exceeding the event threshold (i.e., always). Thus, H and F are both equal to one. As p_t increases, more points are created along the ROC curve. At $p_t = 1$, H and F are both zero. A ROC curve lying along the 1:1 line indicates no skill. Ideally, the curve should travel along the left and upper axes (i.e., it should have a low F and high H at all probability levels).

As with the Brier Score [Eq. (A.11)], H and F are calculated for inflow anomaly thresholds relative to climatological inflows. Exceedance probabilities are again given by the number of ensemble members that exceed the threshold or the probability of exceedance given by the forecast CDF.

Continuous Ranked Probability Score (CRPS)

According to Gneiting et al. (2005), probabilistic forecasts should aim to maximize sharpness subject to calibration. Sharpness refers to the spread of the forecast PDFs; forecasts are sharp if their PDFs are narrow relative to low-skill forecasts derived from climatology, for example. A sharp probabilistic forecasting system is more likely to generate binary event exceedance or non-exceedance probabilities near zero or one. The Continuous Ranked Probability Score (CRPS) addresses both calibration and sharpness (Gneiting et al., 2005, 2007) and is given by:

$$CRPS = \frac{1}{\|T\|} \sum_{t \in T} \int_{-\infty}^{\infty} [F_t(x) - H(x - x_t)]^2 dx, \tag{A.16}$$

where H is the Heaviside function defined as:

$$H(s) = \begin{cases} 1 & s \geq 0 \\ 0 & s < 0. \end{cases} \tag{A.17}$$

This score can be interpreted as an integral of Brier Scores [Eq. (A.11)] over the range of all

possible forecast thresholds x. As with the Brier Score Eq. (A.11), these thresholds are taken to be anomaly thresholds relative to climatology. For a deterministic forecast, $F_t(x)$ is either zero or one, and the CRPS reduces to the mean absolute error. Hersbach (2000) has shown that the CRPS can be decomposed into a reliability component and a 'potential' CRPS component that measures sharpness. Thus, lower CRPS values are preferred, and can be achieved by improving probabilistic forecast reliability and sharpness.

Appendix B

Testing an Adaptive Bias Corrector for Daisy Lake Inflow Forecasts

Recall from Chapter 2 that a linearly-weighted DMB bias corrector with a moving window of 3 days ($LDMB_3$) was found to be ideal for removing bias and improving other measures of forecast error in the M2M ensemble mean inflow forecasts for the Daisy Lake reservoir. This linear weighting of past errors is similar to the weighting in the adaptive parameter updating scheme employed by COMPS [Eq. (4.1)]. Indeed, a dimensionless time scale of τ = 2.0 gives the same weighting to yesterday's forecast error as the $LDMB_3$ scheme [Eq. (2.4)].

A range of τ from 2.0 to 5.0 has been tested on the M2M inflow forecasts from the 2010–2011 water year. Using this year enabled an evaluation of the impact of bias correction on day 3 forecasts; the previous water year has only a short record of day 3 forecasts. Comparison of the various time scales was based on mean absolute error (MAE) and root mean squared error (RMSE) of ensemble mean forecasts (see Appendix A for a description of these verification measures). This comparison is shown in Table B.1. Other metrics of ensemble mean performance were found to be relatively insensitive to the choice of τ over the specified range.

For all forecast horizons, the adaptive schemes outperform $LDMB_3$ for $\tau < 5.0$. Day 1 forecasts show the best improvement for the shortest time scales. Day 2 does best for τ between 3.0 and 3.5, while the ideal τ for day 3 appears to be close to 3.0. Comparison for the 2009–2010 water year yielded similar results for days 1 and 2. Based on the superior performance of the adaptive bias corrector to the $LDMB_3$ corrector, and on the comparison of the various time scales, M2M COMPS configurations for the Daisy Lake inflow forecasting system will use the adaptive DMB bias corrector with $\tau = 3.0$.

Table B.1: A comparison of ensemble mean inflow forecast performance after applying a DMB bias correction computed adaptively for a range of time scales (τ) and computed over a 3-day moving window using the linearly-weighted corrector described in Chapter 2. Smaller values of MAE and RMSE are preferred.

Forecast Day	Metric	τ = 2.0	τ = 2.5	τ = 3.0	τ = 3.5	τ = 4.0	τ = 5.0	$LDMB_3$
Day 1	MAE	5.6	5.6	5.8	5.9	6.0	6.2	6.2
	RMSE	8.8	8.9	9.0	9.1	9.2	9.5	9.7
Day 2	MAE	7.3	7.2	7.2	7.2	7.2	7.3	8.1
	RMSE	12.0	11.9	11.8	11.8	11.9	12.1	12.9
Day 3	MAE	8.6	8.4	8.3	8.4	8.4	8.5	9.0
	RMSE	13.2	12.9	12.9	12.9	13.0	13.1	14.1

Appendix C

Bayesian Model Averaging and the M2M Ensemble

An adaptive updating scheme was implemented in COMPS (Chapter 4) for computing the weights in Bayesian Model Averaging (BMA), which can be used to produce calibrated probabilistic forecasts (Raftery et al., 2005). In the BMA uncertainty model, an observation is assumed to be drawn from one of several candidate distributions centred at each ensemble member. A forecast probability density function (PDF) is then taken to be the weighted sum of these distributions, where the weights are a measure of past forecast performance based on the value of the candidate model's forecast PDF at past verifying observations. BMA has been applied successfully in hydrologic forecasting applications over a range of timescales, where predictors are transformed prior to fitting normal distributions (e.g., Duan et al., 2007; Reggiani et al., 2009; Wang et al., 2009; Parrish et al., 2012). In order to reduce the number of parameters that must be fit given the training data, most applications of BMA take the spread parameters of the individual distributions to be identical.

Early tests of the method indicated that it was not a suitable uncertainty model for the M2M ensemble. This is likely due to the multi-parameter component of the ensemble's formulation. Evidence of this can be seen in Figure C.1, which shows how the model weights evolve throughout the 2009–2010 water year for a small sample of M2M WaSiM ensemble members (weights have been calculated for the full 72-member ensemble, but only a small sample is plotted for readability; day 1 forecasts weights are shown). In the upper frame of Figure C.1, the weights have been calculated adaptively using a dimensionless timescale of $\tau = 30$. The centre frame shows how the weights evolve when calculated using a moving window of 150 days. A long window is required because the algorithm is attempting to fit 73 parameters (72 weights and a common spread parameter). The lower frame shows observed inflows and observation-driven model runs made with the three different parameterizations (MAE_o, NSE_o, and $LNSE_o$; see Chapter 3). These observation-driven runs are used to generate the initial conditions for the inflow forecasts each day, and have a direct impact on forecast quality.

The $LNSE_o$ models, which perform slightly better than the others during the early part of the

water year, are given slightly larger weights (indicated by thicker shaded areas) early on by the adaptive algorithm. At the start of the warm season, the MAE_o and NSE_o begin to outperform the $LNSE_o$ models. The moving window calculation of BMA weights is able to increase the weights applied to these models, but it does so with a significant time lag. The adaptive weight calculation on the other hand is unable to resurrect the weights of the MAE_o and NSE_o members, even with a relatively small τ value.

A possible solution would be to allow each member to have its own spread parameter so that models with small weights would be permitted to have large spread. Since the weights are calculated based on the value of the model's forecast PDFs at the observed values during training, this greater spread would allow the weights to recover more quickly during periods of transition between ideal parameterizations. However, for the 72-member M2M ensemble, this would require fitting 144 parameters, and would require a lengthy training period to ensure a good fit, resulting in a significant time lag in weight changes. Another issue with implementing BMA for the M2M ensemble is that the number of ensemble members for each forecast day is not necessarily constant (e.g., due to numerical weather prediction model instabilities). When calculating weights using a moving window, this issue can be solved by calculating weights based only on the performance of the models that are actually available for the next forecast. In the adaptive framework, this would require tracking thousands of possible ensemble member combinations and updating the weights for each combination daily. We have therefore opted to exclude BMA from further consideration.

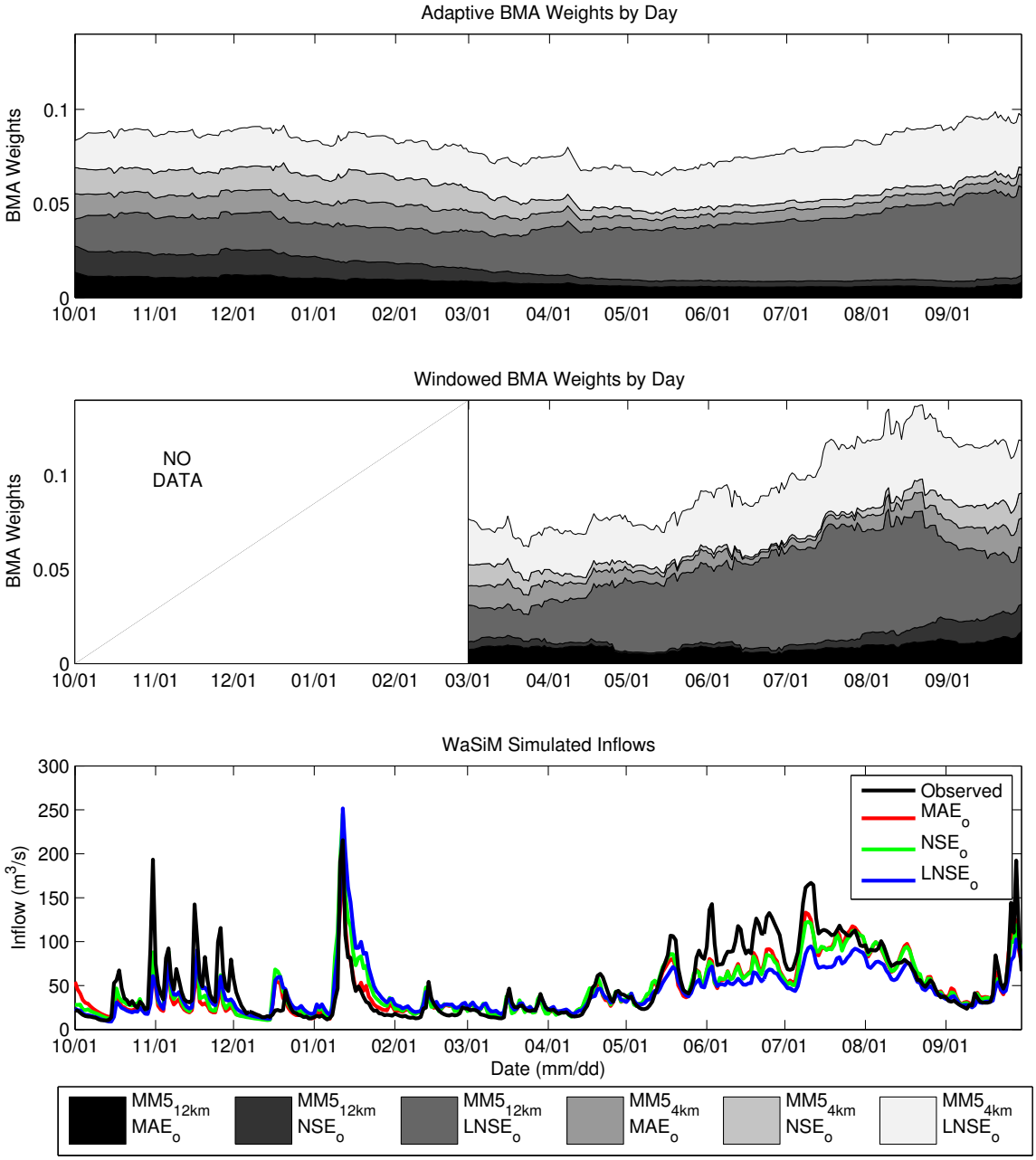

Figure C.1: A subset of BMA weights calculated using an adaptive updating scheme (upper panel) and a moving window (middle panel). The weights are stacked such that thicker areas represent larger weights. Results from observation-driven model runs (simulations) made with the different model parameterizations are shown in the lower panel with observed inflows for comparison. Weights calculated using the moving window change with model performance, but with a significant time lag.