AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE
10.1002/2014WR015462

Reliable probabilistic forecasts from an ensemble reservoir inflow forecasting system

Dominique R. Bourdin, Thomas N. Nipen, and Roland B. Stull

Department of Earth, Ocean and Atmospheric Sciences, University of British Columbia, Vancouver, British Columbia, Canada

Key Points:
- Shape of forecast PDF should match that of forecast errors to ensure reliability
- Forecast errors and therefore the ideal probability model may vary seasonally
- Probability calibration can introduce errors to already reliable forecasts

Correspondence to:
D. R. Bourdin,
dbourdin@eos.ubc.ca

Citation:
Bourdin, D. R., T. N. Nipen, and R. B. Stull (2014), Reliable probabilistic forecasts from an ensemble reservoir inflow forecasting system, *Water Resour. Res.*, *50*, 3108–3130, doi:10.1002/2014WR015462.

Received 16 FEB 2014
Accepted 19 MAR 2014
Accepted article online 25 MAR 2014
Published online 10 APR 2014

Abstract This paper describes a probabilistic reservoir inflow forecasting system that explicitly attempts to sample from major sources of uncertainty in the modeling chain. Uncertainty in hydrologic forecasts arises due to errors in the hydrologic models themselves, their parameterizations, and in the initial and boundary conditions (e.g., meteorological observations or forecasts) used to drive the forecasts. The Member-to-Member (M2M) ensemble presented herein uses individual members of a numerical weather model ensemble to drive two different distributed hydrologic models, each of which is calibrated using three different objective functions. An ensemble of deterministic hydrologic states is generated by spinning up the daily simulated state using each model and parameterization. To produce probabilistic forecasts, uncertainty models are used to fit probability distribution functions (PDF) to the bias-corrected ensemble. The parameters of the distribution are estimated based on statistical properties of the ensemble and past verifying observations. The uncertainty model is able to produce reliable probability forecasts by matching the shape of the PDF to the shape of the empirical distribution of forecast errors. This shape is found to vary seasonally in the case-study watershed. We present an "intelligent" adaptation to a Probability Integral Transform (PIT)-based probability calibration scheme that relabels raw cumulative probabilities into calibrated cumulative probabilities based on recent past forecast performance. As expected, the intelligent scheme, which applies calibration corrections only when probability forecasts are deemed sufficiently unreliable, improves reliability without the inflation of ignorance exhibited in certain cases by the original PIT-based scheme.

1. Introduction

Forecasts of hydrologic variables are subject to uncertainty due to errors introduced into the modeling chain via imperfect initial and boundary conditions, poor model resolution, and the necessary simplification of physical process representation in the model [e.g., *Palmer et al.*, 2005; *Bourdin et al.*, 2012]. Deterministic forecasts of streamflow 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 uncertainty by indicating the likelihood of occurrence of a range of forecast values. Additionally, probabilistic inflow forecasts enable water resource 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. However, in both weather and hydrologic forecasting applications, ensembles are often found to be underdispersive and therefore unreliable [e.g., *Eckel and Walters*, 1998; *Buizza*, 1997; *Wilson et al.*, 2007; *Olsson and Lindström*, 2008; *Wood and Schaake*, 2008; *Bourdin and Stull*, 2013]. 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 both the ensemble and past verifying observations. These theoretical distributions can potentially reduce the amount of data required to characterize the distribution (e.g., from 72 ensemble members to two parameters describing the mean and spread of a Gaussian distribution), and allow estimation of probabilities for future rare events that lie outside the range of observed behavior [*Wilks*, 2006].

A variety of different uncertainty models are available for generating probabilistic forecasts from ensembles. The simplest method is the binned probability ensemble (BPE), which makes the assumption that each ensemble member and the verifying observation are drawn from the same (unknown) probability

distribution [*Anderson*, 1996]. The verifying observation therefore has an equally likely probability of $(K+1)^{-1}$ (given an ensemble of K members) of falling between any two consecutive ranked ensemble members, or outside of their predicted range.

Alternatively, it can be assumed that verifying observations are drawn from a normal distribution centered on the ensemble mean (or, equivalently, that the ensemble mean forecast errors are normally distributed). In this Gaussian uncertainty model, distributional spread can be given by the variance of the ensemble members, implicitly assuming the existence of a spread-skill relationship. That is, that 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]. For variables with nonnormally distributed forecast errors, such as precipitation and streamflow, the Gamma distribution has also been applied in uncertainty modeling frameworks [e.g., *Hamill and Colucci*, 1998; *Sloughter et al*., 2007; *Vrugt et al*., 2008].

The Bayesian model averaging (BMA) uncertainty model assigns a probability distribution to each ensemble member, assuming the verifying observation to be drawn from one of these [*Raftery et al*., 2005]. The forecast distribution is taken to be a weighted average of these distributions, where weights are based on past performance of individual ensemble members.

In contrast with uncertainty models that fit distributions to ensembles, there exist sophisticated models that can be used to produce probabilistic forecasts from an individual hydrologic model. Such methods are commonly based on a sampling of the model's parameter uncertainty space. The generalized likelihood uncertainty estimation (GLUE) method is conceptually simple, easy to implement, and can handle a range of different error structures [*Kuczera and Parent*, 1998; *Blasone et al*., 2008]. Unlike GLUE, Bayesian recursive estimation (BaRE) makes strong, explicit assumptions about error characteristics [*Thiemann et al*., 2001]. The formal generalized likelihood function of *Schoups and Vrugt* [2010] builds on previous approaches, extending their applicability to situations where errors are correlated, heteroscedastic, and non-Gaussian, and resulting in improved forecast reliability.

If the assumptions made by the uncertainty model regarding error characteristics 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 (derived from an uncertainty model) directly. The BMA uncertainty model, as presented by *Raftery et al*. [2005] is an example of ensemble calibration, as it was developed specifically to produce sharp, calibrated probability forecasts by refining the spread parameters of the individual member distributions such that the continuous ranked probability score (CRPS) is minimized over a training period. Generalizations of BMA have also been developed for this purpose [e.g., *Johnson and Swinbank*, 2009].

The weighted ranks method [*Hamill and Colucci*, 1997] and its generalization, the Probability Integral Transform (PIT)-based calibration scheme 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 improve probabilistic forecasts generated using BMA when those forecasts were unreliable. Bayesian ensemble calibration methods have been applied successfully in hydrologic forecasting applications over a range of time scales [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 modeling 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].

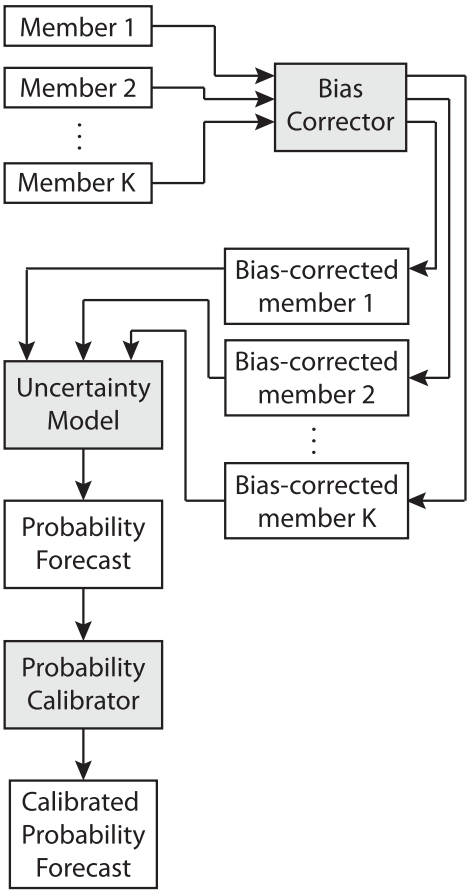

Figure 1. Flowchart illustrating how deterministic forecasts are transformed into probability forecasts and subsequently into calibrated probability forecasts. Postprocessing schemes are indicated by shaded boxes, while the inputs/outputs of these schemes (i.e., forecasts) are in white.

In this paper, we present a generalized methodology for producing probabilistic forecasts of reservoir inflow from an ensemble of deterministic forecasts. Prior to combination, each ensemble member is individually bias corrected using a simple degree-of-mass-balance scheme. An intelligent probability calibration scheme is employed to improve the reliability of the probability forecasts when necessary. The methods are applied to a 72 member ensemble and tested over a period of two water years.

2. From Ensembles to Calibrated Probability Forecasts

This section describes how an ensemble of deterministic forecasts becomes transformed into a probability forecast, and subsequently, a calibrated probability forecast. This process is also illustrated in Figure 1.

2.1. Bias Correction

Hydrologic forecasts contain both systematic and random errors. Systematic error, also known as (unconditional) bias, can arise due to differences between modeled and actual topography, deficiencies in model representation of physical processes, and errors in model parameterization. The aim of bias correction is to reduce the systematic error of future forecasts using statistical relationships between past forecasts and their verifying observations.

In a multimodel 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 multimodel 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 biases do not balance, then their combination can result in a degradation of forecast accuracy [*Wilson et al.*, 2007].

An appropriate measure of bias for volumetric quantities such as reservoir inflow is the degree of mass balance (DMB) [*McCollor and Stull*, 2008]. The DMB is a measure of the ratio of simulated or forecasted inflow to the observed inflow over a given period of time, with a DMB of one indicating a forecast or simulation that is free of volumetric bias. In this study, we apply a DMB correction scheme that allows more recent forecast errors to have greater impact on the bias correction [*Bourdin and Stull*, 2013]. In order to maintain computational efficiency, the DMB is calculated adaptively rather than over a moving window [*Nipen*, 2012]. Thus, only the last estimate of the DMB correction factor must be retrieved each day, along with the new forecast-observation pair to update the correction 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 ensemble 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:

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

where τ (>0) is a unitless time scale that describes how quickly the impact of new information $(\xi_{t,k}/x_t)$ diminishes over time. 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 or decay time in days. This formulation of the DMB is suitable for the case-study data (section 3), where inflows are always observed to be greater than zero.

A bias-corrected inflow forecast $\left(\hat{\xi}_{t,k}\right)$ is then calculated by:

$$\hat{\xi}_{t,k} = \frac{\xi_{t,k}}{DMB_{t,k}}. \tag{2}$$

The use of a multiplicative bias correction factor ensures that corrected inflow forecasts do not become negative.

Note that the bias correction is applied only to the flow forecast; NWP forecast fields are not bias-corrected prior to their use in driving the hydrologic models. The importance of postprocessing the inputs to hydrologic models has been discussed in the literature [e.g., *McCollor and Stull*, 2008; *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. Thus, the DMB corrector used in this study attempts to correct for a variety of different sources of bias in the modeling chain and their different characteristics [*Bourdin and Stull*, 2013].

2.2. Uncertainty Modeling

A commonly cited problem with both weather and hydrologic ensembles is unreliability due to underdispersion [e.g., *Eckel and Walters*, 1998; *Buizza*, 1997; *Wilson et al*., 2007; *Olsson and Lindström*, 2008; *Wood and Schaake*, 2008]. 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 how it compares to verifying observations. In this way, it is possible to implicitly account for any uncertainty that is neglected or simply 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 ensemble mean forecast errors (assuming the distribution is to be centered on the bias-corrected ensemble mean). Hydrologic variables and their errors are often described as being nonnormally distributed, and are therefore transformed into a space in which the errors become normally distributed, and the transformed variable can be modeled 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 to only positive forecast values.

In this study, we employ 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. The suitability of such a model with and without prior log-transformation of the case-study flow data is addressed in section 3.3. Incorporating the DMB bias-corrected ensemble member forecasts $\left(\hat{\xi}_{t,k}\right)$, the EMOS forecast PDF valid at time t is given by:

$$f_t \sim \mathcal{N}\left(\frac{1}{K}\sum_{k=1}^{K}\hat{\xi}_{t,k}, a_T s_t^2 + b_T\right), \tag{3}$$

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). As with the DMB bias correction scheme, computational simplicity is maintained in the uncertainty model scheme by updating these regression parameters adaptively [*Nipen*, 2012].

2.3. 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. 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 PIT values [*Gneiting et al*., 2007] in a PIT histogram, which, for perfectly reliable forecasts, should be approximately flat (Appendix A). PIT values are given by:

$$P_t = F_t(x_t), \tag{4}$$

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 defined as:

$$F_t(x) = \int_{-\infty}^{x} f_t(x)dx. \tag{5}$$

While the PIT histogram is a useful diagnostic tool, the calibration deviation metric D of *Nipen and Stull* [2011] (equation (A1)) provides a more objective measure of reliability. In addition to measuring reliability, 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). Like the ignorance score, the CRPS addresses both reliability and sharpness [*Gneiting et al*., 2005, 2007], but is also more robust. A description of verification metrics used in this study and their interpretation is given in Appendix A.

2.4. Probability Calibration Method

Two calibration schemes are compared in this study. One is the PIT-based calibration scheme (PITCal) described by *Nipen and Stull* [2011] with necessary modifications for computationally efficient adaptive parameter calculation [*Nipen*, 2012]. We also present a new "intelligent calibration" scheme (*inteliCal*) that improves on some of the shortcomings of the original PITCal method.

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 versus 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:

$$\Phi_T(p) = \frac{1}{||T||}\sum_{t\in T} H(p - F_t(x_t)), \tag{6}$$

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 of size $||T||$, p is a probability value between 0 and 1, and H is the Heaviside function given in equation (A6).

This calibration curve $\Phi_T(p)$ is generated by dividing the p interval [0, 1] into any number of bins. In this study, we use 10 bins and the individual PIT values along the calibration curve are updated with a time scale of $\tau = 90$. 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. Using fewer bins reduces sampling error but generates a very coarse calibration curve. Excluding the (constant) end points (0, 0) and (1, 1) of the calibration curve, these 10 bins are defined by nine interior "smoothing points" (p, Φ_p). Modifying equation (6) for adaptive updating of these points yields:

$$\Phi_{p,t+1}=\frac{\tau-1}{\tau}\Phi_{p,t}+\frac{1}{\tau}H(p-F_t(x_t)). \tag{7}$$

[*Nipen*, 2012].
The calibrated CDF is then calculated by:

$$\hat{F}_t(x)=\Phi_T(F_t(x)), \tag{8}$$

which amounts to a relabeling of CDF values $F_t(x)$ to form a new (calibrated) distribution $\hat{F}_t(x)$. The corrected forecast PDF $\left(\hat{f}_t\right)$ can be calculated by combining equations (5) and (8) and invoking the chain rule, yielding:

$$\hat{f}_t(x)=\Psi_T(F_t(x))f_t(x), \tag{9}$$

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 PIT-based calibration scheme implemented in this study 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 forecast 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 reliable forecasts actually worsens the forecast ignorance. This is caused by the introduction of sampling errors to 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. In this paper, we present the inteliCal calibration scheme, which is an evolution of the original PITCal method of *Nipen and Stull* [2011] that reduces the impact of sampling error and prevents inflation of ignorance scores.

InteliCal uses the calibration deviation metric, D (equation (A1)), which is a measure of how much the PIT histogram deviates from flatness or reliability, to decide when to apply the calibrator. This decision is based on a comparison of D (of a raw PIT histogram) to the calibration deviation expected for perfectly reliable forecasts given by $E[D_p]$ (equation (A2)) (which, due to sampling error, is nonzero) [*Nipen and Stull*, 2011]. When D is sufficiently greater than $E[D_p]$, we anticipate that the forecast will benefit from calibration. InteliCal will therefore apply the PIT-based calibration only when:

$$D > ICF\times E[D_p]. \tag{10}$$

The inteliCal adjustment factor (*ICF*) is used to adjust the sensitivity of inteliCal if necessary. In this study, we attempt to determine a suitable value for the *ICF*. InteliCal computes D from the same PIT bin counts that comprise the calibration curve ($\Phi_T(p)$; equation (7)), and is therefore updated adaptively over the same time scale of $\tau = 90$. Both the PITCal and inteliCal probability calibration schemes will be applied to the case-study probability forecasts. The performance of the two schemes will be compared using PIT histograms and calibration deviations in addition to the ignorance score.

3. Case Study

3.1. Study Dates and Data

In this study, two different uncertainty models and two probability calibration schemes 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 (Figure 2). 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

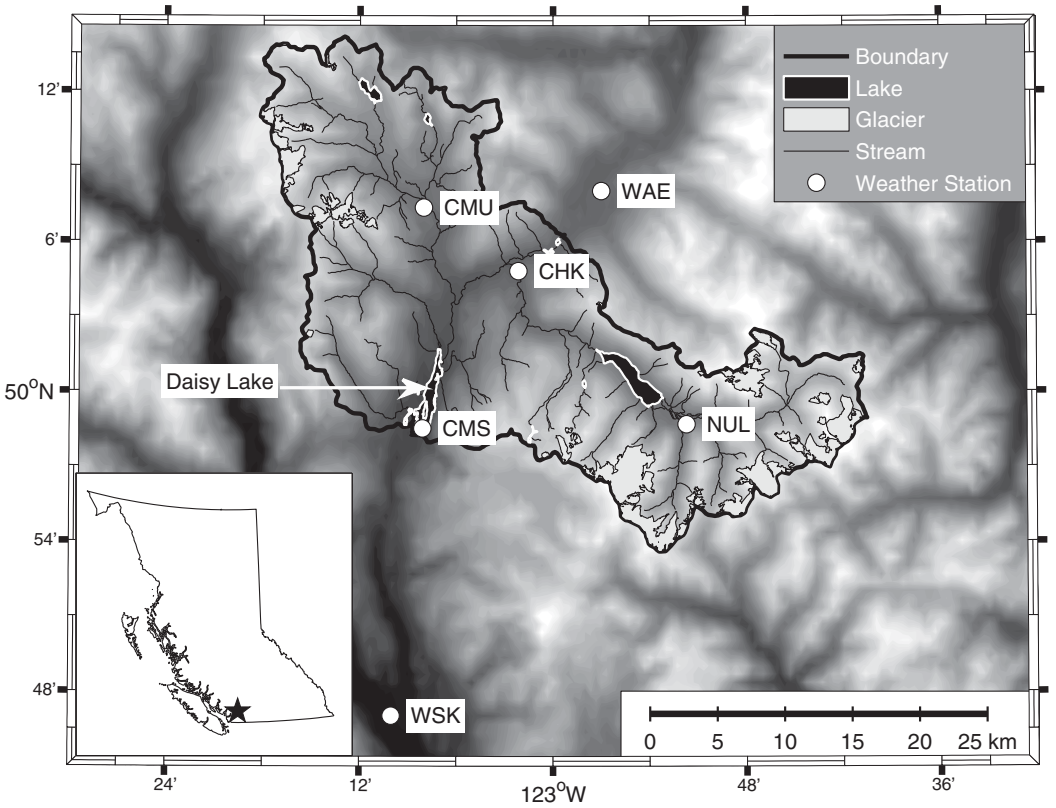

Figure 2. Map of the Cheakamus basin, which drains into the Daisy Lake hydroelectric reservoir. ASTER GDEM background map is a product of METI and NASA with higher elevations represented by lighter shades of gray. Reprinted from Journal of Hydrology, 502, D. R. Bourdin and R. B. Stull, Bias-corrected short-range Member-to-Member ensemble forecasts of reservoir inflow, 77–88, Copyright (2013), with permission from Elsevier.

defined as the period from 1 October to 30 September of the following year. 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. For the purposes of this study, these calculated inflow 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 EMOS uncertainty model parameters (section 2.2) and is excluded from evaluation.

3.2. 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 attempts to sample 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. A description of the various error-sampling components of the M2M ensemble is provided below. Interested readers are referred to *Bourdin and Stull* [2013] for further details on the M2M ensemble forecasting system.

3.2.1. Distributed Hydrologic Models

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.

Both WaSiM and WATFLOOD have been calibrated using the Dynamically Dimensioned Search (DDS) algorithm [*Tolson and Shoemaker*, 2007; *Graeff et al*., 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. This methodology is consistent with that of *Duan et al*. [2007], who likewise used objective functions favoring different parts of the hydrograph to optimize multiple hydrologic models. These different parameterizations attempt to sample the uncertainty in the hydrologic models' parameter values. Simulations during the 10 year calibration period (1997–2007) were driven by downscaled observed meteorological conditions at weather stations within the case-study watershed and surrounding area (Figure 2).

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], among others. 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.

The multi-state or multi-initial-condition component of the M2M ensemble forecasting system arises as a direct consequence of implementing a multiparameter 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 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 uncertainty in defining the hydrologic state of the case-study watershed.

While ensemble data assimilation methods are available for hydrologic modeling 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 multiparameter ensemble. This is because of the paucity of observed data available within the watershed for assimilation; flow data are available at the CMS and CHK stations (Figure 2), and continuous snow-water equivalent (SWE) observations are only available at a nearby location outside of the watershed. *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, 2005b; *DeChant and Moradkhani*, 2011b; *Leisenring and Moradkhani*, 2011].

3.2.2. Numerical Weather Models

The NWP models in the M2M ensemble 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 in Vancouver. The ensemble consists of three independent nested limited-area high-resolution mesoscale models with forecast domains centered over southwestern BC: (1) the Mesoscale Compressible Community (MC2) model [*Benoit et al*., 1997]; (2) the fifth-generation Pennsylvania State University-National Center for Atmospheric Research Mesoscale Model (MM5) [*Grell et al*., 1994]; and (3) Version 3 of the Weather Research and Forecasting (WRF) model

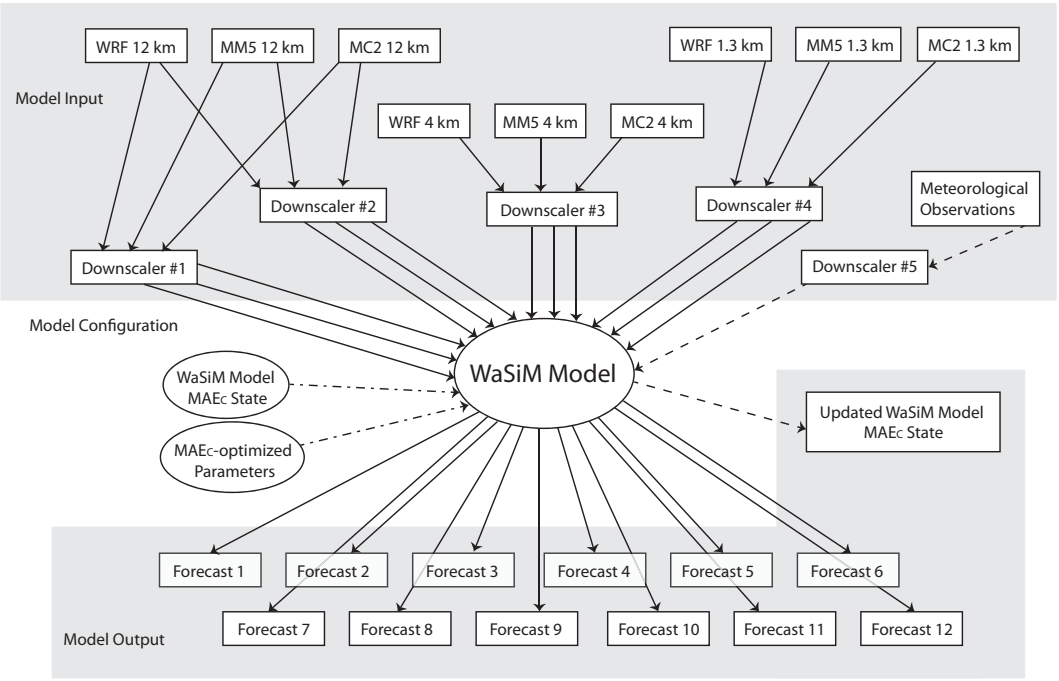

Figure 3. The flow of information into and out of the WaSiM model for generating forecasts with the MAE-optimized parameter set. This process is repeated for two watershed models, each with three parameterizations/states, yielding 72 unique inflow forecasts each day. Reprinted from Journal of Hydrology, 502, D. R. Bourdin and R. B. Stull, Bias-corrected short-range Member-to-Member ensemble forecasts of reservoir inflow, 77–88, Copyright (2013), with permission from Elsevier.

[*Skamarock et al.*, 2008]. Hourly model output fields with grid spacings of 12, 4, and 1.3 km are used for this study and produce inflow forecasts with lead times of up to 3 days.

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 and 1.3 km NWP fields are downscaled using a bilinear interpolation scheme. WaSiM uses gridded NWP output of temperature, precipitation, wind speed, humidity, and global radiation. WATFLOOD downscaling is done using IDW that incorporates elevation dependence using optional constant elevation adjustment rates for both temperature and precipitation (these being the only NWP fields required by WATFLOOD). The 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. Thus, each hydrologic model is driven by 12 different NWP inputs.

Figure 3 illustrates the process of updating hydrologic states and issuing forecasts from a particular parameterization of the WaSiM model (that was optimized using DDS with MAE as the objective function). 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 (where this configuration is specified by model components linked 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 (12 NWP inputs $\times$ 2 DH models $\times$ 3 parameterizations with associated states).

3.3. Application of Methods Using a COMponent-Based Postprocessing System

The methods described in section 2 are applied to the M2M ensemble using the COMponent-based Postprocessing 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/.

COMPS breaks down the process of generating calibrated probabilistic forecasts into a series of steps referred to as components (e.g., the gray-shaded boxes in Figure 1). The system contains components for bias correction, uncertainty modeling, 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 scheme to implement for each desired component, creating a specific configuration. COMPS can also be used to generate postprocessed deterministic forecasts by bypassing the uncertainty and calibration components.

We have implemented the DMB scheme as a bias-correction component in the COMPS framework and tested performance for a range of τ. An adaptive DMB bias corrector with $\tau = 3.0$ was found to be effective at removing M2M forecast bias for forecast horizons of 1–3 days. In a previous study [*Bourdin and Stull*, 2013], M2M inflow forecast bias was found to be strongly controlled by bias in the simulations used to generate the hydrologic state 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. *Bourdin and Stull* [2013] found a linearly weighted DMB correction factor calculated over a moving window of 3 days to be effective in removing bias and improving other aspects of the M2M ensemble's performance. The adaptive DMB scheme was selected for this case study for computational simplicity, and was found to be superior to the linearly weighted scheme over the case-study period.

The first step in generating a probabilistic forecast of inflows to the Daisy Lake reservoir from the bias-corrected M2M ensemble is choosing a suitable uncertainty model. The M2M ensemble is underdispersive; observations often fall outside of the range of inflows predicted by the ensemble members [*Bourdin and Stull*, 2013]. This implies that in spite of the M2M ensemble system explicitly attempting to sample major sources of error in the modeling chain (using simple methods), the amount of uncertainty captured by it is often inadequate. 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, 2005b; *DeChant and Moradkhani*, 2011b; *Leisenring and Moradkhani*, 2011], though such methods are computationally expensive.

Observed daily inflows at Daisy Lake exhibit a bimodal distribution, with storm-season flows forming a skewed distribution at low flows, and warm-season flows forming a second peak at higher flows. For this reason, forecast errors are analyzed by season: the storm season is defined as the period of October–April; the warm season runs from May to September. The distribution of storm-season errors at all lead times is characterized by high peaks and long, narrow tails with slight positive skew (Figure 4). Log-transformed storm-season errors, which are calculated by taking the natural logarithm of the forecasts and the observations prior to calculating the error, are well modeled by a normal distribution. Warm-season errors on the other hand do not require log-transformation prior to fitting a normal distribution.

The spread of the distributions fitted to the M2M ensemble should be related to forecast skill such that when the forecast is less skillful, the uncertainty (as represented by the spread of the PDF) is greater. Since the M2M ensemble is underdispersive, we expect that the distributional spread will be best represented by a combination of ensemble spread and information regarding recent errors as is done in the EMOS method described in section 2.2 [*Gneiting et al*., 2005; *Nipen and Stull*, 2011].

The Gaussian EMOS scheme implemented in COMPS includes an option to log-transform the forecast and observation data prior to fitting a normal distribution to the ensemble. This uncertainty model, which assumes forecast errors to be log-normally distributed, is referred to as log-EMOS. Both the EMOS and log-EMOS uncertainty models are tested on the M2M ensemble. We expect EMOS to produce calibrated forecasts during the warm season where forecast errors exhibit a normal distribution, and log-EMOS to likewise perform well during storm season, when errors are log-normally distributed (Figure 4). It is possible for the nontransformed EMOS 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 the M2M case study, the regression parameters in equation (3) are updated adaptively using a dimensionless time scale of $\tau = 30$ for both the EMOS and log-EMOS schemes. *Nipen* [2012] found this to be a suitable training period for various meteorological variables. While short training windows allow the uncertainty model to adapt quickly to changes in forecast regime or ensemble configuration, longer training 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 reasonable compromise between these competing criteria.

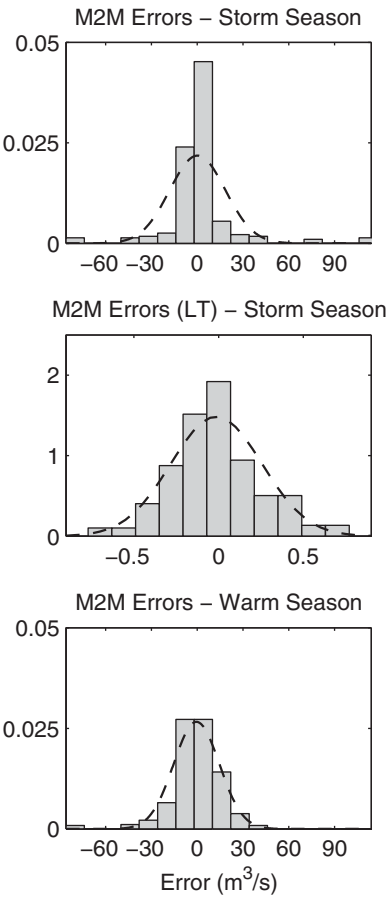

Figure 4. Empirical distributions of day 2 M2M ensemble mean forecast errors (m^3/s) during the 2009–2010 water year. Fitted Gaussian distributions are indicated by dashed lines. Storm-season errors are well modeled by the Gaussian distribution following log-transformation (LT), whereas warm-season forecast errors exhibit a Gaussian shape without transformation.

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. That is, we have ensured 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 reliable probability forecasts. However, at other times during the water year, or as evaluated over shorter time periods, these assumptions may be false, resulting in unreliable or poorly calibrated forecasts. It is during these times that probability calibration can offer improvements to the probabilistic forecasting system. Calibration in COMPS is done using both the original PITCal scheme and the new inteliCal method, as described in section 2.4. A comparison of the two methods is presented below.

4. Results and Discussion

4.1. Performance of the Uncertainty Models

Figure 5 shows the raw (not probability-calibrated) 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 reliable 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 can yield highly unreliable results. The PIT histograms for the storm season show that the EMOS uncertainty model does not concentrate enough probability density at the center of the distribution (i.e., more observations fall into these bins than are expected by the forecasted distributions). Indeed, this is readily anticipated given the distribution of storm-season forecast errors (Figure 4). During the warm season, which exhibits a more normal distribution of errors, the EMOS uncertainty model is able to produce reliable probabilistic forecasts. The importance of specifying a time period over which reliability is measured is evident in the PIT histograms for the full water year, which mask the excellent reliability during the warm season.

Figure 6 shows raw PIT histograms for forecast days 1–3 broken up by season for log-EMOS forecasts. This uncertainty model is, as expected, superior to the EMOS model during the storm season when errors are log-normally distributed, but produces slightly less reliable forecasts during the warm season.

The superior performance of the log-EMOS forecasts during the storm season is also reflected in this model's ignorance and continuous ranked probability scores (Figure 7). Ignorance scores for the EMOS model

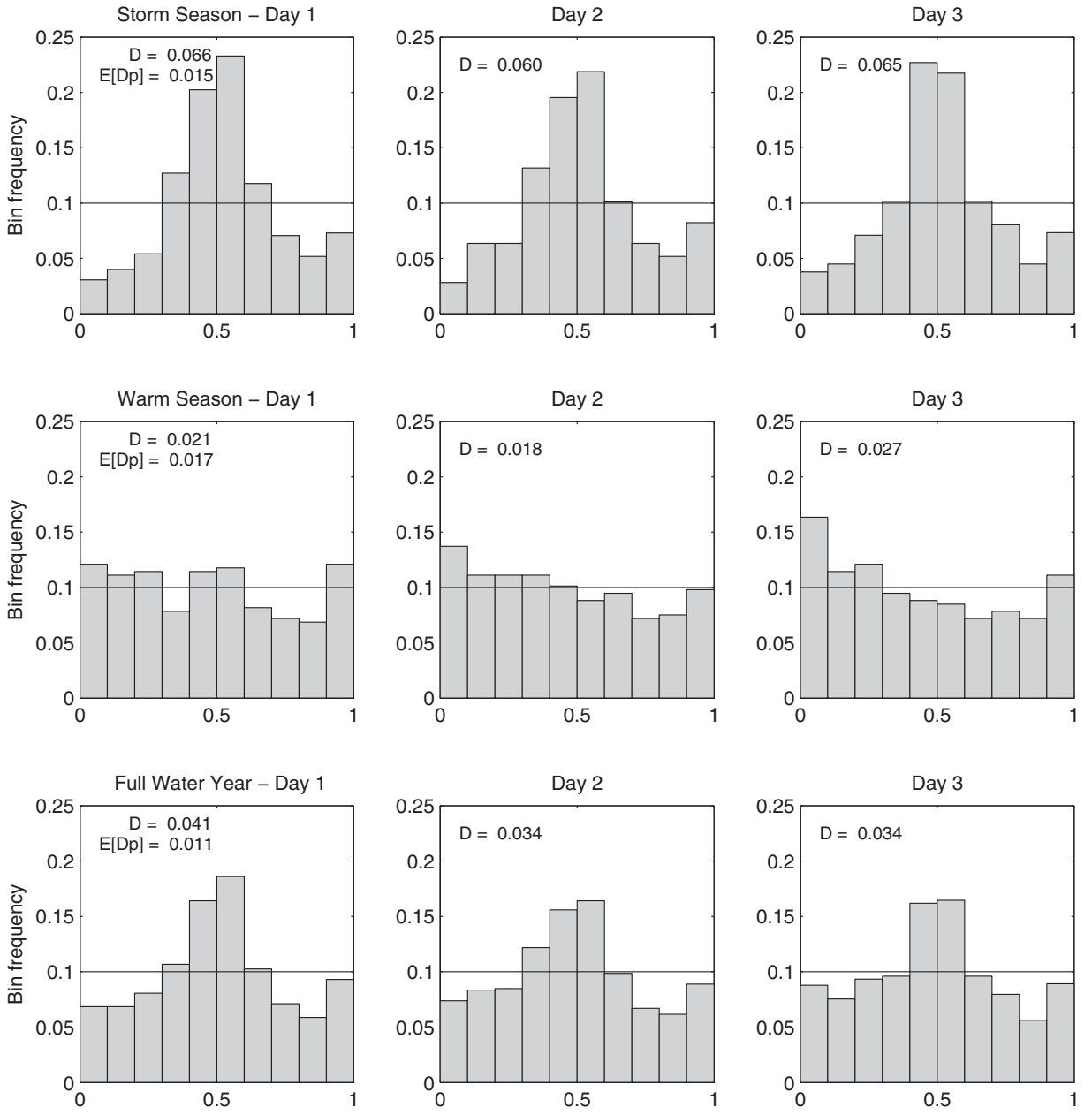

Figure 5. PIT histograms for the (top) storm seasons, (middle) warm seasons, and (bottom) full water years, pooled over the 2010–2011 and 2011–2012 water years. Results are for the raw (not probability-calibrated) EMOS uncertainty model. Calibration deviations *D* are shown for each histogram, with $E[D_p]$ for comparison. Flatter histograms and therefore lower *D* are preferred.

during the storm season are significantly worse due to the unsuitability of this model. Conversely, the EMOS model has slightly better ignorance scores than log-EMOS during the warm season for days 1 and 2. The fact that these ignorance scores are higher than EMOS storm-season forecasts is because the warm-season forecast PDFs from both uncertainty models have larger spread as shown in Figure 8. This is a result of the M2M ensemble members having larger forecast errors during the warm season. The CRPS values in Figure 7 are in agreement with ignorance results, indicating that the log-EMOS model performs best during the storm season and that both uncertainty models perform similarly during the warm season.

4.2. Effect of Probability Calibration

Figure 9 shows changes in ignorance scores (ΔIGN) and calibration deviations (*D*) following calibration with PITCal (black bars) and inteliCal (gray bars) with *ICF* adjustment factor ranging from 1.0 to 1.8. In these plots,

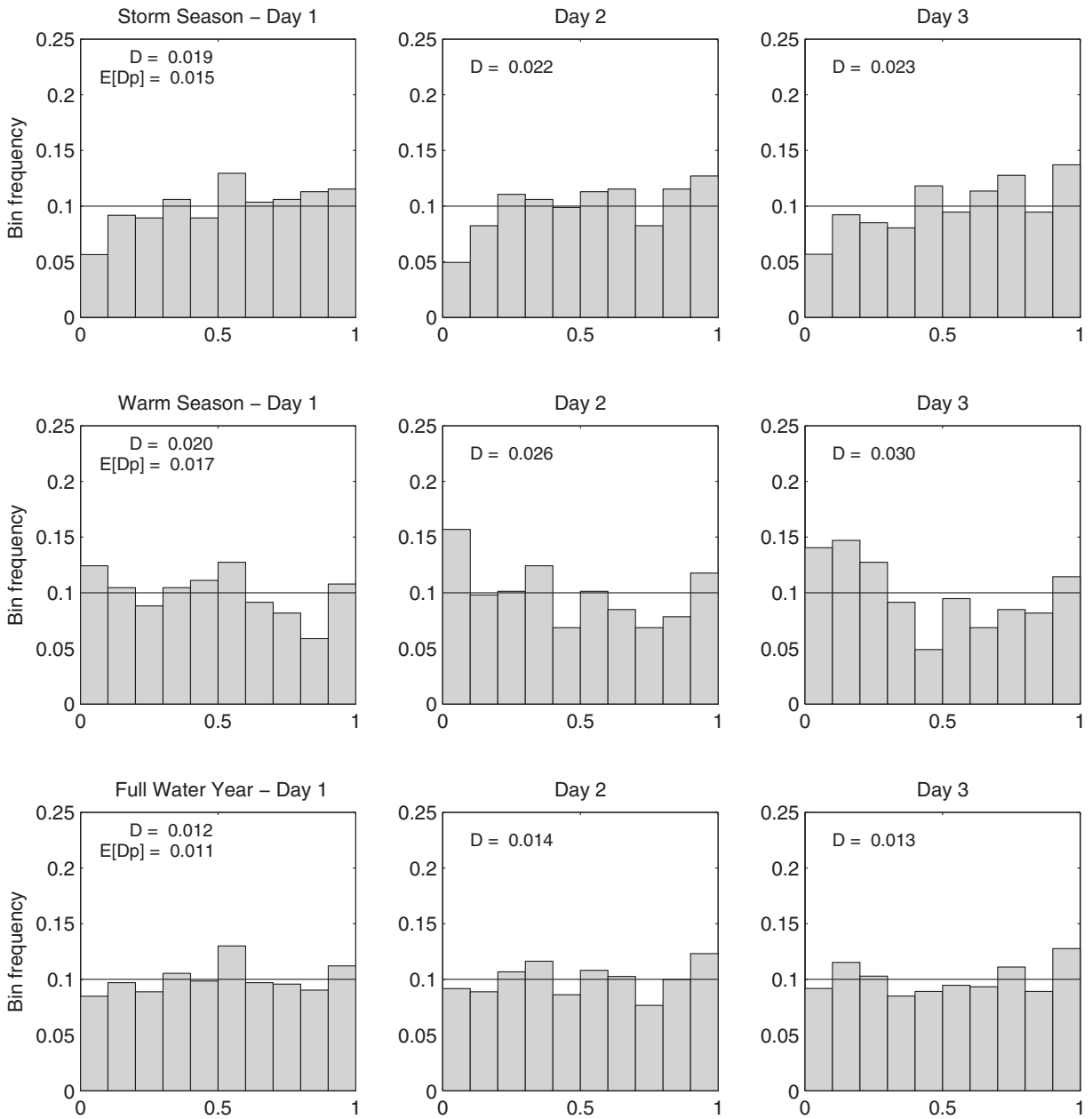

Figure 6. Same as Figure 5. Results are for the raw log-EMOS uncertainty model.

ΔIGN is calculated by subtracting the calibrated ignorance score from the raw score. Negative ΔIGN therefore indicates that the calibration method increases (worsens) forecast ignorance. In agreement with the results of *Nipen and Stull* [2011], Figure 9 shows that the PITCal probability calibration scheme increases ignorance when the original forecasts were reliable or nearly reliable. This is caused by the introduction of sampling error in the calibration curve. Note that sampling error is likely less significant in verification than it is in calibration, as the verification sample sizes are larger.

Due to the relatively long memory ($\tau = 90$) of the adaptive updating scheme used to generate the calibration curve, warm-season EMOS and log-EMOS forecasts become more unreliable with application of PITCal. This issue is illustrated and discussed further below. Storm-season log-EMOS forecast reliability is improved by PITCal, and all calibration strategies improve reliability and reduce ignorance for storm-season EMOS

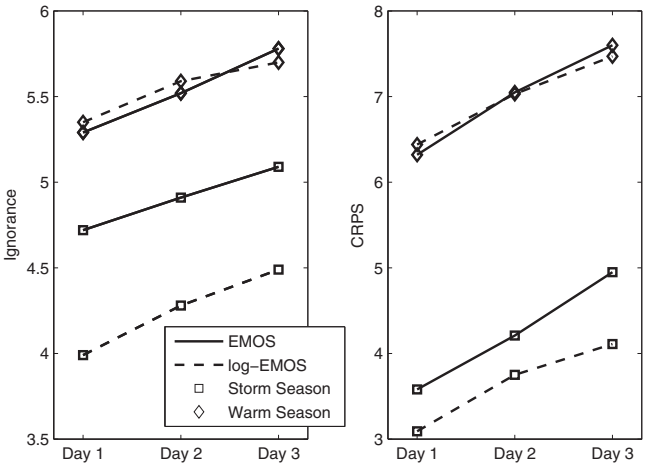

Figure 7. Ignorance and continuous ranked probability scores for the uncertainty models tested. Forecasts are divided by season for scoring, as each uncertainty model has different calibration characteristics during these times of year. Smaller ignorance scores and CRPS are preferred.

forecasts. *Nipen* [2012] derived a decomposition of the ignorance score for a set of 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}. In our comparison of raw and probability-calibrated EMOS and log-EMOS 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}.

As the inteliCal adjustment factor (*ICF*) is increased, the new inteliCal method is able to reduce and eventually eliminate any inflation of ignorance caused by the calibrator. This is achieved by applying the calibration correction less often, resulting in calibration deviations that approach their raw values. An *ICF* of 1.4 appears to balance the competing objectives of calibrating and improving the ignorance of unreliable forecasts (i.e., EMOS storm-season forecasts), without inflating the ignorance of forecasts that are already reliable or nearly so. An important result is that except in the case of very unreliable forecasts, the calibration schemes are unable to decrease ignorance scores beyond those of the raw forecasts. Thus, while calibration can lead to improvements in reliability and ignorance, the first line of defence is the selection of an appropriate uncertainty model. Changes in CRPS scores are not shown for PITCal and inteliCal calibrated forecasts because, as in Figure 7, their relative values were nearly identical to ignorance score results, and therefore provided no additional diagnostic information.

Figure 10a illustrates the power of the PITCal probability calibration scheme. 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 time scale of $\tau = 90$ applied here results in the calibration adjustments necessary during the storm season being propagated into the already reliable warm season. Namely, the calibration has successfully adjusted the storm-season predictive distributions to have higher peaks and thicker tails (causing flatter PIT histograms), but has carried this adjustment into the warm season, as indicated by the day 2 and day 3 distributions now being more underdispersive with too much probability density at the center of the distribution (causing PIT histograms to dip in the center). The same is true for the forecasts calibrated using

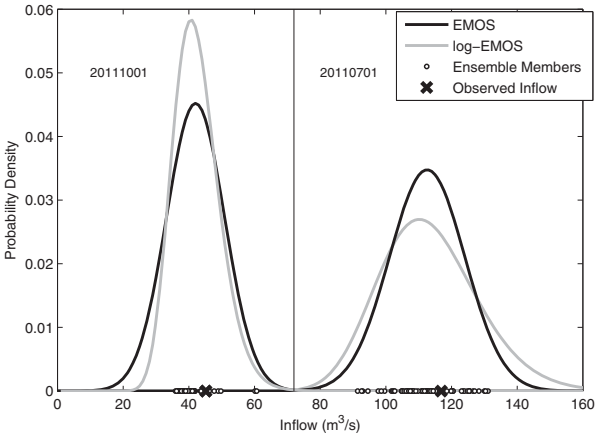

Figure 8. Day 1 forecast PDFs issued by the EMOS and log-EMOS uncertainty models on (left) 1 October (storm season) and (right) 1 July (warm season) 2011. Warm-season ignorance scores are higher than storm-season scores because the forecast PDFs have higher spread.

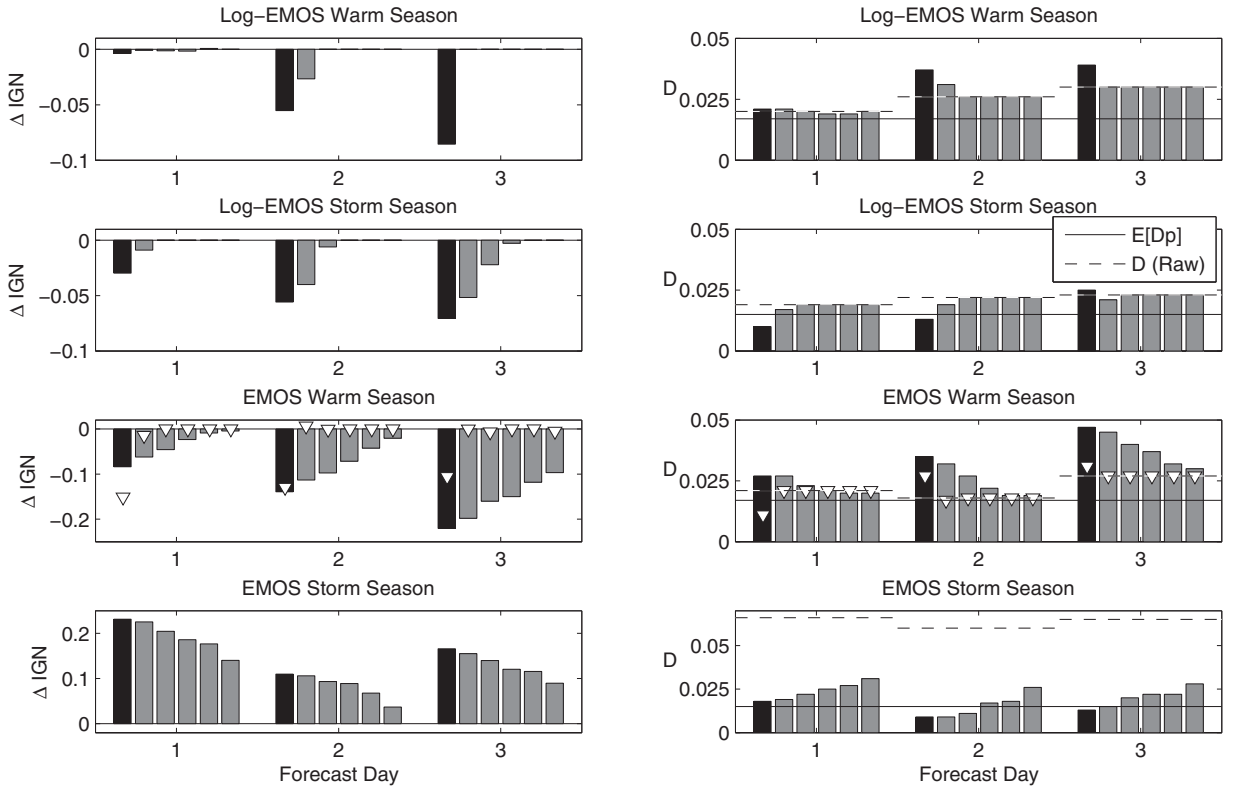

Figure 9. Changes to ignorance scores and calibration deviations after applying calibration using the PITCal (black bars) and inteliCal methods with adjustment factor *ICF* = 1.0, 1.2, . . .,1.8 (gray bars, left to right). Δ*IGN* values near to or greater than zero and *D* values near their expected values ($E[D_p]$; solid horizontal lines) are better. Scores for forecasts calibrated using the carry-forward strategy described in section 4.2 are shown by the triangle markers.

inteliCal with *ICF* = 1.4 (Figure 10b). Storm-season forecast reliability has been greatly improved as in the PITCal case, and the correction carries into the warm season and degrades calibration. In both seasons, since inteliCal applies the correction less often, the effects are less pronounced. It is this lag-time in updating the calibration curve between seasons that causes the increases in calibration deviation and ignorance scores following calibration shown in Figure 9.

The lag-time problem exhibited by calibrated warm-season EMOS forecasts is not as pronounced during the transition from the warm season to the storm season for either PITCal or inteliCal. Figure 11 illustrates that at the start of the storm season in 2011, the calibration curve still contains some information from the previous storm season (recall that in the adaptive updating scheme, old information is never forgotten, but becomes less important with time). Thus, PITCal applies the appropriate correction to the raw forecast, adjusting the normal distribution to a more log-normal shape. InteliCal on the other hand, with an *ICF* of 1.4, does not find the raw forecast to be sufficiently unreliable, and performs no correction. This period of no correction causes the storm-season PIT histograms in Figure 10b to exhibit signs of inadequate probability density in the center of the calibrated forecast PDFs (though the effect is minor relative to the raw results in Figure 5).

The PIT histograms for log-EMOS forecasts calibrated with the PITCal and inteliCal (*ICF* = 1.4) schemes are shown in Figures 12a and 12b, respectively. In the case of PITCal, the slight deterioration in warm-season calibration deviation (shown here and in Figure 9) 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. Since neither of these biases are particularly strong, inteliCal believes the warm-season forecasts to be already reliable, and applies these corrections less often, resulting in very little change in the PIT histogram relative to the raw log-EMOS forecasts.

The drastic change in shape of the EMOS uncertainty model's raw PIT histograms between seasons suggests an alternative calibration strategy for avoiding the decrease in reliability and associated increase in

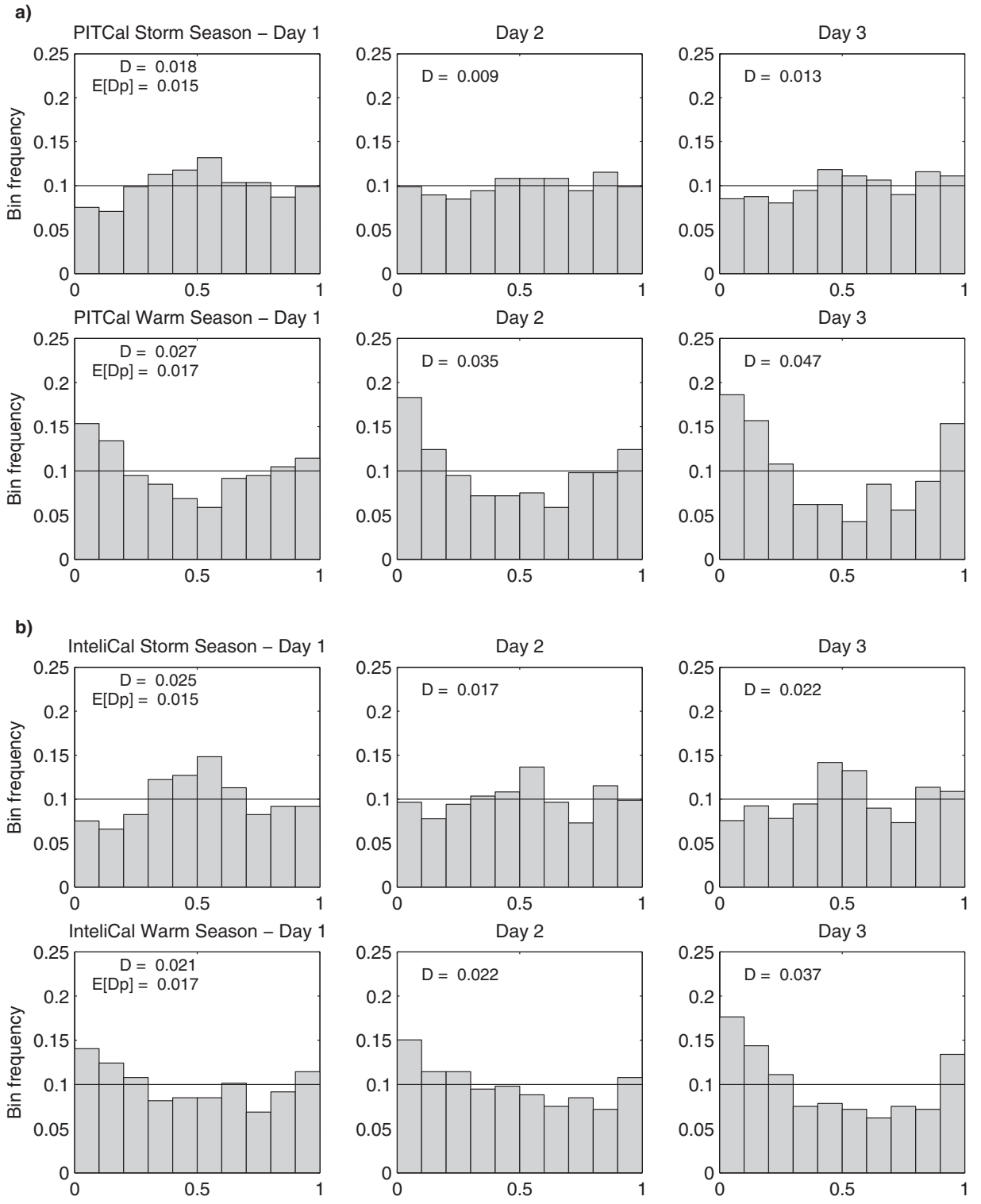

Figure 10. PIT histograms as in Figure 5 for EMOS forecasts following probability calibration using the (a) PITCal and (b) inteliCal (*ICF* = 1.4) schemes.

ignorance of warm-season EMOS forecasts. Indeed, even with *ICF* values of 1.6 and 1.8, warm-season EMOS ignorance is greatly increased (particularly at lead times of 2 and 3 days), even though the increase in calibration deviation is minor (Figure 9). To avoid the adaptive calibration scheme's long lag-time in generating

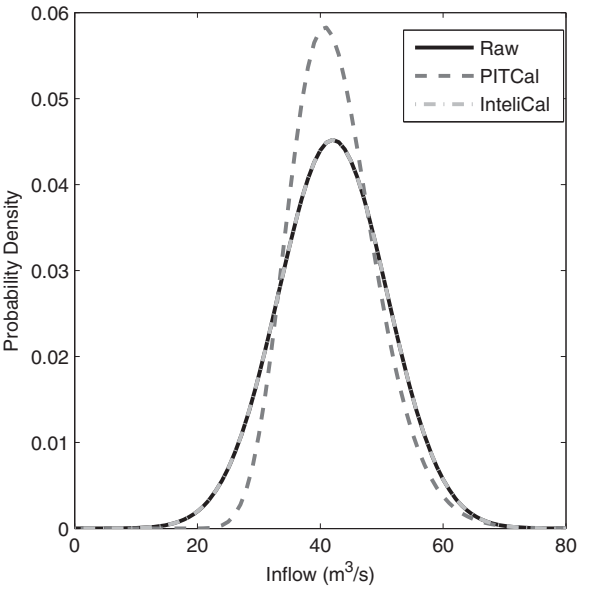

Figure 11. Sample raw and calibrated day 1 forecast PDFs from the EMOS uncertainty model for 1 October 2011. Following the nearly calibrated warm season, the calibration curve still contains information from the previous storm season (i.e., it attempts to transform the normal distribution into a log-normal distribution). The correction is small enough that inteliCal does not apply it and the inteliCal PDF is identical to the raw PDF.

representative calibration curves, we replaced the calibration curve parameters at the start of the warm season (taken to be 1 May) with those valid at some time during the previous year's warm season. 29 July was selected as the replacement date based on calibration statistics from the 2009–2010 water year. By this date, the adaptively updated PIT histogram is able to reflect the (reliable) characteristics of the raw EMOS warm-season probability forecasts. Note that the choice of 1 May 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, was carried out using PITCal and inteliCal with *ICF* values of 1.0–1.8. Changes in ignorance and calibration deviations from these methods are indicated in Figure 9 by the triangle markers. CF-PITCal increases the ignorance of these already reliable forecasts, whereas the CF-inteliCal method applies the correction very rarely, maintaining low *D* without increasing ignorance. PIT histograms are not shown for these CF-calibrated warm-season forecasts because they are indiscernible from those in Figure 5.

5. Concluding Remarks

In this paper, we have transformed a 72 member ensemble forecasting system that explicitly attempts to sample from major sources of error in the inflow modeling chain into a calibrated probabilistic forecasting system. This work was done using the COMponent-based Postprocessing System (COMPS) described and developed by *Nipen* [2012]. COMPS allows its users to implement and apply schemes for bias correction, uncertainty modeling, probability calibration, forecast updating using recent observations, and verification. Any of these components can alternatively be bypassed, making COMPS a flexible postprocessing 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 reliable results during the warm season when errors are normally distributed.

In agreement with previous results from *Nipen and Stull* [2011], the PITCal calibration scheme was generally found to improve reliability at the expense of increased ignorance of already reliable forecasts. This is caused by an overfitting of the calibration curve to sampling errors when the calibration correction is not very strong. As expected, the inteliCal calibration scheme, which builds upon the original PITCal scheme, was able to improve the reliability of poorly calibrated forecasts without inflating ignorance scores of forecasts that were already well calibrated. Testing of this new calibration scheme reveals that by applying the calibration correction when the calibration deviation is greater than 1.4 times its expected value, a balance can be achieved between these competing objectives.

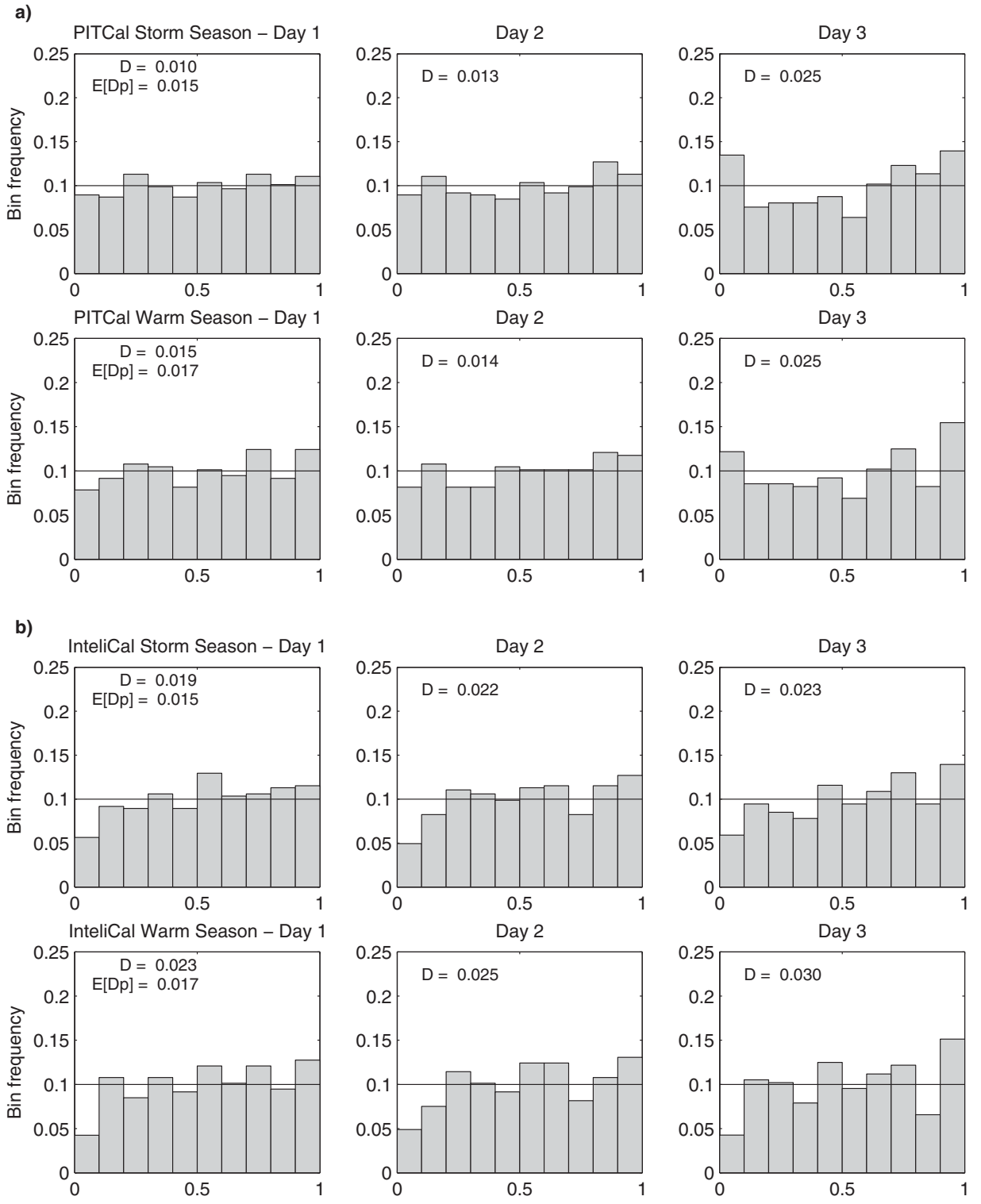

Figure 12. PIT histograms as in Figure 6 for log-EMOS forecasts following probability calibration using the (a) PITCal and (b) inteliCal (*ICF* = 1.4) schemes.

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 for both the PITCal and inteliCal schemes. 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 maintain small calibration deviation without inflating ignorance or CRPS. An alternative to the carry-forward strategy could include the use of different time scales (τ) for computing the calibration deviations and calibration curves. While the calibration curve is sensitive to sampling limitations and therefore requires a lengthy training period, the calibration deviation may be less sensitive. A shorter time scale could then be used to increase inteliCal's responsiveness to changes in forecast error characteristics. This is an area for future work in COMPS development.

This study clearly indicates that the ideal M2M-based probability forecasting system for Daisy Lake inflows during the storm season is the COMPS configuration that utilizes the log-EMOS uncertainty model. Even following the large reduction in ignorance achieved by calibrating the EMOS storm-season forecasts, log-EMOS ignorance remains far superior. Thus, while probability calibration can improve both reliability and ignorance, it is no match for choosing a suitable uncertainty model in the first place.

Selection of the ideal COMPS configuration for warm-season inflow forecasts is user-dependent; users who are more sensitive to forecast accuracy at forecast horizons of one to two days will benefit most from EMOS forecasts, whereas users sensitive to day 3 forecast quality will benefit from using log-EMOS. In all cases, inteliCal should be applied to the forecasts to ensure reliability without risking increased ignorance. In fact, day 1 log-EMOS forecast ignorance decreases by 0.038 following inteliCal calibration, indicating that there is some potential for improved ignorance with this new scheme. While this may seem a scant improvement, the gambling interpretation of ignorance (see Appendix A, equation (A4)) reveals that the expected number of bets required to double wealth using inteliCal log-EMOS forecasts against their raw (not calibrated) counterparts is only 27.

Testing of the newly implemented inteliCal calibration scheme in COMPS indicates that a suitable value for the inteliCal adjustment factor, *ICF*, may be approximately 1.4. This value appears to strike a compromise between improving reliability of unreliable forecasts, and reducing the impact of sampling error in nearly reliable forecasts. 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 paper 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 (and in generating low-ignorance forecasts). 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 (e.g., a rainy season and a snowmelt-driven season) may benefit from the use of different uncertainty models at different times of year.

In this work, determination of candidate PDF shapes was based on analysis of empirical ensemble mean forecast error distributions over 1 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 modeling, 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 may 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].

Another limitation of this study 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 predefined dates based on climatological flow characteristics) likely had very little impact on the verification. However, the change in forecasting system, if not correctly timed, could result in nonoptimal forecasts with significant impacts on reservoir operation. A smarter alternative would be to change the uncertainty model when flows are observed to have undergone the transition between seasons (i.e., when snowmelt contributions begin and end). Reservoir operators and planners could also have the option of seeing forecasts from both uncertainty models during the transition period to determine the ideal model.

Appendix A: Verification Metrics 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.

A1. Probability Integral Transform (PIT) Histogram

The PIT histogram is analogous to the rank histogram or Talagrand diagram [*Anderson*, 1996; *Talagrand et al*., 1997], and is used to assess reliability or calibration when a probability forecast is expressed as a fitted PDF. PIT values are given by equation (4). For perfectly calibrated forecasts, the PIT histogram will be approximately 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, a U-shaped histogram is an indication of underdispersion, or inadequate spread in the forecast PDF. Note that flatness is not always a guarantee of reliability; a verification sample that includes a combination of negatively and positively biased forecast distributions may yield a flat histogram despite being unreliable [*Hamill*, 2001].

A2. 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:

$$D=\sqrt{\frac{1}{B}\sum_{i=1}^{B}\left(\frac{b_i}{||T||}-\frac{1}{B}\right)^2}, \tag{A1}$$

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:

$$E[D_p]=\sqrt{\frac{1-B^{-1}}{||T||B}}. \tag{A2}$$

When referring to reliability or calibration, we will specify a time period over which the calibration metric is computed, and we will not require the forecast to exhibit reliability 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 reliability 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.

A3. Ignorance Score (IGN)

While reliability or 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 reliable, but far too vague for decision making. Therefore, we will also require our forecast PDFs (f_t) to concentrate probability in the correct area on each day. This property can be measured by the dimensionless ignorance score [*Roulston and Smith*, 2002], which is defined as:

$$IGN = -\frac{1}{||T||}\sum_{t\in T} log_2(f_t(x_t)), \tag{A3}$$

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.

The ignorance score can also be interpreted in terms of gambling returns, which offers a more intuitive feel for the relative quality of different forecasting systems [*Nipen and Stull*, 2011]. When placing bets on forecast outcomes, the optimal strategy is to distribute one's current wealth to each possible outcome according to the probability of that outcome being realized. Given two probabilistic forecasting systems, and assuming that forecasting system A has lower ignorance than system B, gamblers using system A to distribute their bets can expect to double their wealth against a user of system B in a number of bets given by:

$$N_{bets} = \frac{1}{IGN_B - IGN_A}. \tag{A4}$$

A4. 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 nonexceedance probabilities near zero or one. The continuous ranked probability score (CRPS) addresses both reliability and sharpness [*Gneiting et al.*, 2005, 2007] and is given by:

$$CRPS = \frac{1}{||T||}\sum_{t\in T}\int_{-\infty}^{\infty}[F_t(x) - H(x - x_t)]^2 dx, \tag{A5}$$

where H is the Heaviside function defined as:

$$H(s) = \begin{matrix} 1 & s \geq 0 \\ 0 & s < 0. \end{matrix} \tag{A6}$$

The CRPS is 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.

For a deterministic forecast, $F_t(x)$ is either zero or one, and the dimensionless 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 or sharpness.

Acknowledgments

Funding for this research was provided by the Canadian Natural Sciences and Engineering Research Council (NSERC) in the form of a postgraduate scholarship awarded to D.R.B. and a Discovery Grant to R.B.S. We wish to thank Scott Weston of BC Hydro for providing hydrometric data, and George Hicks and Henryk Modzelewski for aid in retrieving archived NWP forecasts. Nicholas Kouwen and Jörg Schulla provided indispensable guidance during the setup and calibration of the WATFLOOD and WaSiM models. Computer code for linking DDS with WaSiM was modified from that provided by Thomas Graeff. Numerical weather model data were provided by the Geophysical Disaster Computational Fluid Dynamics Center (GDCFDC) in the Department of Earth, Ocean and Atmospheric Sciences at the University of British Columbia. The authors are grateful to Greg West and three anonymous reviewers, who provided feedback that resulted in great improvements to the manuscript. Data used in this study are available upon request.

References

Anderson, J. L. (1996), A method for producing and evaluating probabilistic forecasts from ensemble model integrations, *J. Clim.*, *9*(7), 1518–1530.

Andreadis, K. M., and D. P. Lettenmaier (2006), Assimilating remotely sensed snow observations into a macroscale hydrology model, *Adv. Water Resour.*, *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, *Mon. Weather Rev.*, *125*(10), 2382–2415.

Blasone, R.-S., J. A. Vrugt, H. Madsen, D. Rosbjerg, B. A. Robinson, and G. A. Zyvoloski (2008), Generalized likelihood uncertainty estimation (GLUE) using adaptive Markov Chain Monte Carlo sampling, *Adv. Water Resour.*, *31*(4), 630–648, doi:10.1016/j.advwatres.2007.12.003.

Bourdin, D. R., and R. B. Stull (2013), Bias-corrected short-range Member-to-Member ensemble forecasts of reservoir inflow, *J. Hydrol.*, *502*, 77–88, doi:10.1016/j.jhydrol.2013.08.028.

Bourdin, D. R., S. W. Fleming, and R. B. Stull (2012), Streamflow modelling: A primer on applications, approaches and challenges, *Atmos. Ocean*, *50*(4), 507–536, doi:10.1080/07055900.2012.734276.

Box, G. E. P., and D. R. Cox (1964), An analysis of transformations, *J. R. Stat. Soc.*, *26*(2), 211–252.
Brocker, J., and L. A. Smith (2007), Increasing the reliability of reliability diagrams, *Wea. Forecasting*, *22*, 651–661, doi:10.1175/WAF993.1.
Buizza, R. (1997), Potential forecast skill of ensemble prediction and spread and skill distributions of the ECMWF ensemble prediction system, *Mon. Weather Rev.*, *125*(1), 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, *Mon. Weather Rev.*, *138*(11), 4268–4281, doi:10.1175/2010MWR3349.1.
Chow, V. T. (1954), The log-probability law and its engineering applications, *Proc. Am. Soc. Civ. Eng.*, *80*(5), 536-1-536-25.
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, *Adv. Water Resour.*, *31*, 1309–1324.
DeChant, C. M., and H. Moradkhani (2011a), Improving the characterization of initial condition for ensemble streamflow prediction using data assimilation, *Hydrol. Earth Syst. Sci.*, *15*, 3399–3410.
DeChant, C. M., and H. Moradkhani (2011b), Radiance data assimilation for operational snow and streamflow forecasting, *Adv. Water Resour.*, *34*, 351–364.
Duan, Q., N. K. Ajami, X. Gao, and S. Sorooshian (2007), Multi-model ensemble hydrologic prediction using Bayesian model averaging, *Adv. Water Resour.*, *30*(5), 1371–1386, doi:10.1016/j.advwatres.2006.11.014.
Eckel, F. A., and M. K. Walters (1998), Calibrated probabilistic quantitative precipitation forecasts based on the MRF ensemble, *Wea. Forecasting*, *13*(4), 1132–1147.
Gneiting, T., A. E. Raftery, A. H. Westveld, and T. Goldman (2005), Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation, *Mon. Weather Rev.*, *133*(5), 1098–1118, doi:10.1175/MWR2904.1.
Gneiting, T., F. Balabdaoui, and A. E. Raftery (2007), Probabilistic forecasts, calibration and sharpness, *J. R. Stat. Soc. Ser. B*, *69*, 243–268, doi:10.1111/j.1467–9868.2007.00587.x.
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, *Hydrol. Processes*, *26*(24), 3749–3769, doi:10.1002/hyp.8463.
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*, Natl. Cent. for Atmos. Res., Boulder, Colo.
Grimit, E. P., and C. F. Mass (2002), Initial results of a mesoscale short-range ensemble forecasting system over the Pacific Northwest, *Wea. Forecasting*, *17*, 192–205.
Hamill, T. M. (2001), Interpretation of rank histograms for verifying ensemble forecasts, *Mon. Weather Rev.*, *129*(3), 550–560.
Hamill, T. M., and S. J. Colucci (1997), Verification of Eta-RSM short-range ensemble forecasts, *Mon. Weather Rev.*, *125*(6), 1312–1327.
Hamill, T. M., and S. J. Colucci (1998), Evaluation of Eta-RSM probabilistic precipitation forecasts, *Mon. Weather Rev.*, *126*(3), 711–724.
Hashino, T., A. A. Bradley, and S. S. Schwartz (2007), Evaluation of bias-correction methods for ensemble streamflow volume forecasts, *Hydrol. Earth Syst. Sci.*, *11*, 939–950.
Hersbach, H. (2000), Decomposition of the continuous ranked probability score for ensemble prediction systems, *Wea. Forecasting*, *15*, 559–570.
Johnson, C., and R. Swinbank (2009), Medium-range multimodel ensemble combination and calibration, *Q. J. R. Meteorol. Soc.*, *135*, 777–794, doi:10.1002/qj.383.
Kouwen, N. (2010), *WATFLOOD/WATROUTE Hydrological Model Routing and Flow Forecasting System*, Univ. of Waterloo, Waterloo, Ont.
Krzysztofowicz, R. (2001), The case for probabilistic forecasting in hydrology, *J. Hydrol.*, *249*(1–4), 2–9, doi:10.1016/S0022–1694(01)00420-6.
Kuczera, G., and E. Parent (1998), Monte Carlo assessment of parameter uncertainty in conceptual catchment models: The Metropolis algorithm, *J. Hydrol.*, *211*, 69–84.
Leisenring, M., and H. Moradkhani (2011), Snow water equivalent prediction using Bayesian data assimilation methods, *Stochastic 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, *J. Hydrol.*, *240*(1–2), 106–117, doi:10.1016/S0022–1694(00)00337-1.
Madadgar, S., H. Moradkhani, and D. Garen (2012), Towards improved post-processing of hydrologic forecast ensembles, *Hydrol. Processes*, *28*, 104-122, doi:10.1002/hyp.9562.
Mascaro, G., E. R. Vivoni, and R. Deidda (2011), Impact of basin scale and initial condition on ensemble streamflow forecast uncertainty, paper presented at 25th Conference on Hydrology, Am. Meteorol. Soc., Seattle, Wash.
McCollor, D., and R. Stull (2008), Hydrometeorological accuracy enhancement via postprocessing of numerical weather forecasts in complex terrain, *Wea. Forecasting*, *23*, 131–144, doi:10.1175/2007WAF2006107.1.
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 Resour. Res.*, *41*, W05012, doi:10.1029/2004WR003604.
Moradkhani, H., S. Sorooshian, H. V. Gupta, and P. R. Houser (2005b), Dual state-parameter estimation of hydrological models using ensemble Kalman filter, *Adv. Water Resour.*, *28*, 135–147.
Murphy, A. H. (1973), A new vector partition of the probability score, *J. Appl. Meteorol.*, *12*, 595–600.
Nash, J. E., and I. V. Sutcliffe (1970), River flow forecasting through conceptual models. Part I: A discussion of principles, *J. Hydrol.*, *10*(3), 282–290, doi:10.1016/0022-1694(70)90255-6.
Nipen, T. (2012), A component-based probabilistic weather forecasting system for operational usage, PhD thesis, Dep. of Earth, Ocean and Atmos. Sci., Univ. of B. C., Vancouver.
Nipen, T., and R. Stull (2011), Calibrating probabilistic forecasts from an NWP ensemble, *Tellus, Ser. A*, *63*(5), 858–875, doi:10.1111/j.1600-0870.2011.00535.x.
Olsson, J., and G. Lindström (2008), Evaluation and calibration of operational hydrological ensemble forecasts in Sweden, *J. Hydrol.*, *350*(1–2), 14–24, doi:10.1016/j.jhydrol.2007.11.010.
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, *Annu. Rev. Earth Planet. Sci.*, *33*, 163–193, doi:10.1146/annurev.earth.33.092203.122552.
Parrish, M. A., H. Moradkhani, and C. M. DeChant (2012), Toward reduction of model uncertainty: Integration of Bayesian model averaging and data assimilation, *Water Resour. Res.*, *48*, W03519, doi:10.1029/2011WR011116.
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, *J. Hydrometeorol.*, *7*, 458–477.
Pinson, P., P. McSharry, and H. Madsen (2010), Reliability diagrams for non-parametric density forecasts of continuous variables: Accounting for serial correlation, *Q. J. R. Meteorol. Soc.*, *136*(646), 77–90, doi:10.1002/qj.559.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski (2005), Using Bayesian model averaging to calibrate forecast ensembles, *Mon. Weather Rev.*, *133*(5), 1155–1174, doi:10.1175/MWR2906.1.

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 Resour. Res.*, *45*, W02428, doi:10.1029/2007WR006758.

Roulin, E. (2007), Skill and relative economic value of medium-range hydrological ensemble predictions, *Hydrol. Earth Syst. Sci.*, *11*, 725–737, doi:10.5194/hess-11-725-2007.

Roulston, M. S., and L. A. Smith (2002), Evaluating probabilistic forecasts using information theory, *Mon. Weather Rev.*, *130*(6), 1653–1660.

Schoups, G., and J. A. Vrugt (2010), A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors, *Water Resour. Res.*, *46*, W10531, doi:10.1029/2009WR008933.

Schulla, J. (2012), Model description WaSiM (Water balance Simulation Model), Hydrol. Software Consult. J. Schulla, Zürich, Switzerland.

Skamarock, W. C., J. B. Klemp, J. Dudhia, D. O. Gill, D. M. Barker, M. G. Duda, X. Y. Huang, W. Wang, and J. G. Powers (2008), A description of the Advanced Research WRF version 3, *Tech. Rep. TN-475+STR*, Natl. Cent. for Atmos. Res., Boulder, Colo.

Sloughter, J. M., A. E. Raftery, T. Gneiting, and C. Fraley (2007), Probabilistic quantitative precipitation forecasting using Bayesian model averaging, *Mon. Weather Rev.*, *135*, 3209–3220.

Stedinger, J. R. (1980), Fitting log normal distributions to hydrologic data, *Water Resour. Res.*, *16*(3), 481–490, doi:10.1029/WR016i003p00481.

Steinschneider, S., and C. Brown (2011), Influences of North Atlantic climate variability on low-flows in the Connecticut River basin, *J. Hydrol.*, *409*(1–2), 212-224, doi:10.1016/j.jhydrol.2011.08.038.

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

Talagrand, O., R. Vautard, and B. Strauss (1997), Evaluation of probabilistic prediction systems, in *Proceedings of the Workshop on Predictability*, pp. 1–25, European Centre for Medium-Range Weather Forecasts, Reading, U. K.

Thiemann, M., M. Trosset, H. Gupta, and S. Sorooshian (2001), Bayesian recursive parameter estimation for hydrologic models, *Water Resour. Res.*, *37*(10), 2521–2535.

Tolson, B. A., and C. A. Shoemaker (2007), Dynamically dimensioned search algorithm for computationally efficient watershed model calibration, *Water Resour. Res.*, *43*, W01413, doi:10.1029/2005WR004723.

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 Resour. Res.*, *39*(8), 1214, doi:10.1029/2002WR001746.

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 Resour. Res.*, *39*(8), 1201, doi:10.1029/2002WR001642.

Vrugt, J. A., C. G. H. Diks, W. Bouten, and M. P. Clark (2008), Ensemble Bayesian model averaging using Markov Chain Monte Carlo sampling, *Environ. Fluid Mech.*, *8*, 579–595, doi:10.1007/s10652-008-9106-3.

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 Resour. Res.*, *45*, W05407, doi:10.1029/2008WR007355.

Wilks, D. S. (2006), *Statistical Methods in the Atmospheric Sciences*, 2nd ed., 627 pp., Academic Press, Boston, MA.

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, *Mon. Weather Rev.*, *135*(4), 1364–1385, doi:10.1175/MWR3347.1.

Wood, A. W., and J. C. Schaake (2008), Correcting errors in streamflow forecast ensemble mean and spread, *J. Hydrometeorol.*, *9*(1), 132–148, doi:10.1175/2007JHM862.1.

Yuan, H., J. A. McGinley, P. J. Schultz, C. J. Anderson, and C. Lu (2008), Short-range precipitation forecasts from the time-lagged multimodel ensembles during the HMT-West-2006 Campaign, *J. Hydrometeorol.*, *9*, 477–491.