8.5 MEASURING THE PERFORMANCE OF HYDROLOGICAL FORECASTS FOR HYDROPOWER PRODUCTION AT BC HYDRO AND HYDRO-QUÉBEC

Frank Weber*
BC Hydro, Burnaby, British Columbia
Luc Perreault
Hydro-Quebec, Montreal, Québec
Vincent Fortin
Meteorological Service of Canada, Dorval, Québec

1. INTRODUCTION

The main source of electricity in Canada is hydroelectric energy. In particular in Quebec and British Columbia, the vast majority of the installed generation capacity is hydropower. Canada's largest utility is Hydro-Quebec with 96% of its total generation being from hydro generation. As the second largest hydro electric energy producer in Canada, BC Hydro provides approximately 94% of British Columbia's population with electrical energy, 90% of which is hydroelectric energy.

The purpose of forecasting is to support informed decision-making. Inflow forecasts at BC Hydro and Hydro-Quebec are issued to help operate reservoirs, assess resource capabilities, and determine pricing for the out-of-province sale of energy. The basic assumption of hydrologic forecasting is that having an uncertain forecast available is preferable to complete ignorance about future hydrologic events. Hydrologic forecasts are typically provided using hydrologic models driven by some estimates of future weather, flows observed during the previous time period, a.k.a. persistence forecasts, or average flows, a.k.a. climatology forecasts. BC Hydro and Hydro-Quebec make considerable efforts to provide accurate hydrologic forecasts for various lead times, using physically based conceptual and statistical hydrologic models, weather forecasts, and both deterministic and probabilistic techniques. The forecast lead times analyzed in this study are short-term 5-day forecasts and long-term seasonal water supply forecasts.

The aim of forecast evaluation is twofold. Firstly, forecast verification ensures that inflow forecasts are accurate and skilful from a technical viewpoint. Forecast verification is used to understand forecasts, establish a skill and accuracy reference against which subsequent changes in forecast procedures or the introduction of new technology can be measured, address strengths and weaknesses of the forecasting system, and justify funding for more research, training, and equipment. Secondly, forecast verification ensures that the forecast products meet user requirements. Since different users have different interests, verification schemes may even have to be tailored to different classes of users. For example, we can analyze the bias and accuracy of each individual day of the 5-day forecasts to generally address strengths and weaknesses of the forecasting system. Alternatively, we can analyze forecasts for high inflow events only to get a sense of the forecasting skill in critical conditions with potential of downstream flooding or overtopping the dam. For seasonal forecasts, some users only make use of the deterministic, most probable forecast. Hence, we may only analyze the accuracy of those. Other clients use each individual forecast trace and want confirmation that the possible outcomes represent the full uncertainty.

Advanced forecast verification measures are critical to the assessment of operational impacts of flow events. Insufficient knowledge of the forecast skill eventually translates into uncertainty on the level of risk adopted into operations. To communicate the forecast skill to decision-makers, but also to establish benchmarks of the forecasting systems, BC Hydro and Hydro-Quebec have developed fair, understandable, and relevant performance measures for their hydrologic forecasting systems. This paper explores the use of statistical verification measures and skill scores to quantify the quality of BC Hydro's deterministic and probabilistic hydrologic forecasts. Examples of BC Hydro's hydrologic forecasting skill are given.

BC Hydro's forecast team uses the UBC watershed model (Quick 1995) to deterministically forecast short-term inflows and probabilistically forecast long-term seasonal inflows. The UBC watershed model is a conceptual, continuous hydrologic simulation model, developed in the mid-60's to calculate streamflow from mountainous

* *Corresponding author's address:* Frank Weber, BC Hydro, 6911 Southpoint Dr., Burnaby, V3N 4X8, Canada; email: frank.weber@bchydro.com

watersheds. It is a semi-distributed model, which calculates runoff separately for lumped elevation bands and then linearly combines elevation band runoff to obtain total runoff. For a given watershed, the model simulates the various components of runoff using daily precipitation and daily minimum and maximum temperatures as input. Runoff components include surface runoff from rainfall, snowmelt, and glacier melt, interflow, and upper and lower groundwater flow. Together, these components represent the total inflow hydrograph for the historical and forecast time periods. The model also simulates basin state conditions, such as snow water equivalent and snow covered area. The UBC watershed model is calibrated for each basin using historical input and output data. The software, which is used to run short- and long-term inflow forecasts with the UBC watershed model, was developed at BC Hydro and is called the River Forecast System (RFS; Weiss 2001).

For the short-term forecast period, input data are forecast precipitation and temperature from the Canadian Meteorological Center's (CMC) Global Environmental Multiscale model (GEM). Deterministic input data are used from the high-resolution GEM regional model (0.1375º ~15 km resolution) for the 0 to 48 hr period and from the low-resolution GEM global model (0.9º ~100 km resolution) for the 48 to 120 hr period. Gridded forecasts from the CMC models are then downscaled to forecast locations using a spline function and subsequently re-calibrated (bias corrected). 5-day inflow forecasts are issued manually in the morning of each working day. A forecaster runs the RFS using quantitative precipitation and temperature forecasts, which may have been further adjusted by the BC Hydro meteorologist. The forecaster may also apply post-model adjustments to forecasted inflows to account for temporary model inaccuracies.

In the Pacific Northwest, meteorological forecasts are more accurate than climatological averages for forecast lead times of only a few days. For long-term forecasts, the Ensemble Streamflow Prediction (ESP) procedure (Day 1985), uses instead sequences of historic climatologic data as future UBC watershed model input. The ESP procedure assumes that meteorological events that occurred in the past are representative of events that may occur in the future. In a first step, the UBC watershed model is run up to the forecast date. The purpose of this tracking run is to create the correct basin state conditions in the model at the start of the forecast period. The current simulated basin-state conditions, such as snowpack, soil moisture, and

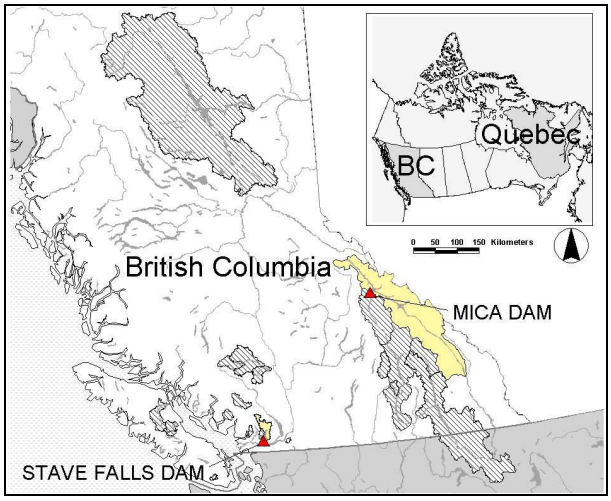

Figure 1 BC Hydro forecast basins and forecast points used in this study

groundwater conditions are the conditions from which the ESP run will start. Additionally, a data assimilation tool allows the forecaster to compare measured with simulated snow water equivalent and if necessary to adjust the simulated snow water equivalent accordingly. The ability to forecast the seasonal runoff lies in the fact that runoff from melting of the mountain snowpack is a major component of the seasonal water supply for many BC Hydro reservoirs. Historic weather sequences are then used as model input to simulate the runoff that would have occurred in these years given the current basin-state conditions. For the operations planning of its hydroelectric projects, BC Hydro requires seasonal inflow forecasts that extend from February to September. The model produces a number of possible future inflow hydrographs that are used directly for follow-up planning studies or are statistically analyzed to produce a mean volumetric forecast and associated error bounds. The accuracy of water supply forecasts increases significantly towards the end of the snow accumulation season, when the maximum basin snow storage is known best. In the Coast, Columbia, and Rocky Mountains of British Columbia this is typically in April or May. Figure 1 shows the location of BC Hydro's basins for which hydrologic forecasts are operationally provided. Due to the large number of basins and resulting performance measures, this study presents the results of 5-day inflow forecast verifications for one coastal and one interior basin. The coastal and interior basins are the Stave and Mica basins, respectively. Table 1 summarizes the basin characteristics of both basins. Inflows into the coastal Stave basin are of a pluvio-nival regime,

Table 1 Summary of basin characteristics

		Stave	Mica
Basin size	km^2	956	21287
Basin elevation	m	52 - 2307	381 - 3468
Normal inflows (1971-2000) in m^3/s			
Jan		111	114
Feb		106	106
Mar		99	116
Apr		105	239
May		145	828
Jun		151	1548
Jul		114	1535
Aug		63	1106
Sep		62	579
Oct		109	303
Nov		158	198
Dec		123	133
Annual		112	570

which is characterized by a slightly higher inflow maximum in November due to rainfall and a second maximum in June due to snowmelt. In BC Hydro's south-coastal basins approximately 30-40% of annual runoff are comprised of snowmelt. Inflows into the interior Mica basin are characterized by a nivo-glacial regime, which shows maximum inflows in June due to snowmelt and a prolonged period of relatively high summer flows in August and September due to glacial runoff. Approximately 70% of annual runoff are comprised of snowmelt. Since a long record of probabilistic water supply forecasts is only available for the Mica project, the analysis of ESP-style water supply forecasts focuses on this project.

2. METHODOLOGY

The forecast quality is traditionally determined by comparing disseminated forecasts with actual observations, and summary measures, which describe that relationship. Since there is no one best score that suits every data set in every situation and provides a complete picture of the overall performance of a forecasting system, the challenge is to find a few meaningful scores that address the specific purposes of the verification system despite the dimensionality of the verification task. The literature (e.g. Wilks 1995) describes several theoretical criteria of forecast quality and, therefore, which scores may be of relevance.

2.1 *Statistical scores and skill scores for deterministic forecasts of continuous variable*s

Reliability describes the statistical consistency between the probability distributions of the forecast and the observed values and can be quantified by the unconditional bias. It measures the correspondence between the average of the forecasts f_k and the average of the observations o_k of the predictand, k = 1, …, n, where n is the number of forecasts. The relative bias (RBias), which is calculated by dividing the unconditional bias by the mean observations, is used in this study to allow the comparison of scores between different forecast points:

$$RBias = \frac{1}{n}\sum_{k=1}^{n}(f_k - o_k) \Big/ \frac{1}{n}\sum_{k=1}^{n} o_k \qquad (1)$$

The perfect score is zero. Probabilistic forecasts are well calibrated if the forecast probability is equal to the long-term relative percentage. However, the bias gives no information about the typical magnitude of the forecasting error.

Accuracy refers to the average correspondence between individual forecasts and the events they predict. While some studies choose squared accuracy measures, such as the mean squared error (MSE; e.g. Mullusky et al. 2004, Pagnano 2005) or the root mean squared error (RMSE; e.g. Coulibaly et al. 2005, Wood et al. 2005), others determine forecast accuracy with the mean absolute error (MAE, Franz 2003). The MAE calculates the average correspondence between individual forecast/observation pairs:

$$MAE = \frac{1}{n}\sum_{k=1}^{n}\left|f_k - o_k\right| \qquad (2)$$

It is believed to be a reliable indicator of typical error magnitudes (Lettenmaier and Wood 1992) and is a robust measure. As a robust measure it is not unduly influenced by a few large residuals, which is a desirable quality if the full range of flows are analyzed rather than flood flows only. For the same reason, the MAE is less sensitive to timing errors in the predictions than squared error statistics (Beven 2001).

Although scores provide measures of forecasting reliability, most of them don't consider the forecasting difficulty. For example, basin characteristics or the predominant weather pattern affect the forecasting difficulty (Druce 1984).

Consequently, absolute error statistics can be better during recession and baseflow periods, and worse during extreme events. One possibility of how geographical and seasonal forecast difficulties can be taken into consideration and, thus, allow comparison amongst forecast projects, is to calculate verification scores separately for different periods of the year. For the purpose of this study, scores were calculated for the total period of interest, but also for 3-month winter, spring, summer, and fall seasons and for events above an arbitrarily chosen 90-percentile non-exceedance threshold. Thereby, winter is defined as the January to March period, spring as the April to June period, summer as the July to September period, and fall as the October to December period.

Another method to normalize forecasts is to calculate the relative performance of the forecast, the forecast skill, by comparing the score of the disseminated forecast with that of a reference forecast (Wilks 1995). In this study the forecasts were assessed using a MAE-based skill score:

$$SS_{MAE} = \left(1 - \frac{MAE_{forecast}}{MAE_{reference\,forecast}}\right) \times 100\% \qquad (3)$$

where SS_{MAE} is the MAE-based skill score relative to the reference forecast, $MAE_{forecast}$ is the score of the forecast to be evaluated and $MAE_{reference\ forecast}$ is the score of the reference forecast. The baseline forecasts used in this study were persistence forecasts and climatological average forecasts. The long-term daily mean was used as naïve climatological short-term reference forecast rather than the long-term annual mean. The MAE-based skill scores relative to climatology and persistence were labeled $SS_{MAE\text{-}CLIM}$ and $SS_{MAE\text{-}PERS}$, respectively. The skill score can be interpreted as a percentage improvement over the reference forecast. Skill scores can be negative if the benchmark forecasting system is better. They are 100% for a perfect forecast, while the reference forecast is imperfect and 0% if forecast and reference forecast are of the same skill.

Association is the overall strength of the linear relationship between the forecast and the observation and is typically measured by the linear correlation coefficient (R). It is calculated as the covariance of forecasts and observations divided by the product of the standard deviations of forecasts and observations. R is normalized to lie between –1 and +1 and as such is a non-dimensional measure. The coefficient of determination, R^2, describes which percentage of the variance is explained by a linear relationship between forecasts and observations (e.g. Zar 1996):

$$R^2 = \frac{\left(\sum_{k=1}^{n}\left(f_k - \frac{1}{n}\sum_{k=1}^{n} f_k\right)\cdot\left(o_k - \frac{1}{n}\sum_{k=1}^{n} o_k\right)\right)^2}{\sum_{k=1}^{n}\left(f_k - \frac{1}{n}\sum_{k=1}^{n} f_k\right)^2 \sum_{k=1}^{n}\left(o_k - \frac{1}{n}\sum_{k=1}^{n} o_k\right)^2} \qquad (4)$$

The RBias, SS_{MAE} relative to climatology and persistence, and R^2 were calculated for deterministic short-term forecasts. For deterministic mean ESP forecasts, the RBias, $SS_{MAE\text{-}CLIM}$ and R^2 were calculated.

2.2 *Statistical scores for deterministic forecasts of discrete variables*

To scrutinize short-term high flow forecasts, Corby and Lawrence (2002) provide a detailed description of a categorical flood forecast verification system. High non-exceedance thresholds are chosen, with which continuous variables are converted into categorical variables. Then the success of predicting high flows of different magnitudes is determined. In this study four flow thresholds were selected, which were based on the 90, 95, 98, and 99-percentile non-exceedance of the 1971-2000 daily flow records and resulted in four categories. These percentiles were chosen because on average about 36 daily observed flows fall into these categories per year. With 5-day forecasts, the number of data points analyzed then ranged from about 200 to 300. Very different results would have been obtained if different thresholds had been chosen. However, the thresholds chosen seemed reasonable as a benchmark for future comparison.

Table 2 shows the resulting 5 x 5 contingency table. If forecast and observed flows were below the 90-percentile, they were not analyzed. Forecast or observed flows above the 90-percentile were classified either as hits, misses, or false alarms. A hit is defined as a forecast, which is in the same category as the corresponding observation. A miss is a forecast, which is in one high flow category, while the corresponding observation is in another one. A false alarm is given when the forecast is in one of the four high flow categories, while the observation remained below the 90-percentile threshold.

Table 2 5 x 5 Contingency table for a categorical high-flow forecast verification

		Observed				
%-ile		<90	90-95	95-98	98-99	>99
Forecast	<90	Correct rejection	Miss	Miss	Miss	Miss
	90-95	False alarm	Hit	Miss	Miss	Miss
	95-98	False alarm	Miss	Hit	Miss	Miss
	98-99	False alarm	Miss	Miss	Hit	Miss
	>99	False alarm	Miss	Miss	Miss	Hit

Hits and misses are further labeled by the category of the corresponding observation, for example as a 'minor miss'. False alarms are further categorized into the category of the corresponding forecast. From the hits and misses, the probability of detection (POD) and the false alarm ratio (FAR) were calculated. POD describes the percent of observed high flow events that were correctly forecast and ranges from 0% to 100% with 100% being the best:

$$POD = \frac{Hits}{Hits + Misses} \times 100\% \quad (5)$$

FAR characterizes the number of false alarms per total number of event forecasts and ranges from 0% to 100% with 0% being the best:

$$FAR = \frac{False\ Alarms}{Alarms} \times 100\% \quad (6)$$

2.3 *Statistical scores for probabilistic forecasts of continuous variables*

Planning engineers use different levels of information content of probabilistic seasonal water supply forecasts. By only using the most probable outcome, typically the mean residual runoff volume, some decision-makers reduce the probabilistic forecast to a deterministic forecast. Others use all individual ensemble forecast sequences for reservoir routing studies. Hydro-Quebec and BC Hydro, therefore, recognize the need for a performance measure that reflects the overall performance of the probabilistic forecasts and, thereby, the uncertainty of the predictive inflow distribution (ESP), rather than only the quality of the ensemble mean to forecast future inflows.

Hydro-Quebec has investigated the use of the Continuous Ranked Probability Score (CRPS; Candille and Talagrand 2005) and other measures such as the logarithmic, quadratic and spherical scores (Gneiting and Raftery, 2004). In this paper, we concentrate only on the CRPS. The CRPS quantifies the overall performance of a probabilistic forecast. It is equivalent to the Brier score integrated over all possible thresholds and measures the area of squared differences between the predictive cumulative distribution function (CDF) F(x) of the forecast x and the CDF of a perfect deterministic forecast, as described by the Heaviside function $H(.)$, which takes the value 0 when $x < x_0$ and 1 otherwise:

$$CRPS = -\int_{-\infty}^{\infty} [F(x) - H(x - x_0)]^2 dx \quad (7)$$

where x_0 is the observation. As defined herein, the CRPS ranges from minus infinity for unreliable ensembles to zero for a perfect ensemble forecast. The CRPS is reported in the units of the observations, i.e. in this study in m^3/s. Due to the squared nature of the score it penalizes forecasts more severely when their probabilities are further from the observation. Consequently, credit is given for assigning high probabilities to values near the observation (near misses). As a derivation of the Brier score the CRPS is strictly proper. A proper scoring rule is one in which a forecaster maximizes the score (or minimizes the score if it is negatively oriented) by forecasting exactly his or her true beliefs about the upcoming situation: it encourages the forecaster to make careful assessments and to be honest (Gneiting and Raftery, 2004). The CRPS is sensitive to the whole range of values of forecast flows and it rewards small spread (sharpness) if the forecast is accurate (Ebert 2005). Finally, in negative orientation, the CRPS reduces to the mean absolute error if F(x) is a deterministic forecast, thereby allowing a direct comparison between probabilistic and deterministic forecasts.

To analyze ESP forecasts a theoretical CDF needs to be fitted to each ensemble. Seasonal hydrologic forecasts exhibit both unimodal and multimodal distributions. Therefore, several

theoretical distributions were tested, including mixtures of unimodal distributions.

CRPS scores were calculated for each forecast month of the February and April forecasts, which allows for an assessment of seasonal effects. Additionally, scores were calculated for the residual forecast periods of the February and April forecasts.

3. RESULTS AND DISCUSSION

3.1 *Performance measures for short-term forecasts*

The data used for the short-term forecast verification are the 2003 and 2004 calendar years. These years are the only complete calendar years available since the implementation of the RFS. Observed reservoir inflows are not measured, but calculated from recorded reservoir level changes and calculated project releases using the hydrologic continuity equation. Mica and Stave inflow data for the 2003 to 2004 period have not been quality controlled and are at times noisy. Figure 2, Figure 3, Figure 4, and Figure 5 summarize scores and skill scores as a function of lead time. The figures also distinguish between forecasts made in different climatological seasons and for threshold flows.

Figure 2 (a) shows that Stave 1 to 5-day lead time forecasts are typically over-forecast (RBias =15%). Figure 2 and Figure 6 (a) and (b) reveal that over-forecasting particularly occurs in fall when the jet stream moves over the area and Pacific frontal system trigger high inflows. However, very high inflows, i.e. inflows above the 210 m^3/s threshold are under-forecast for 2 to 5-day lead times. This is because, on average, large inflow events are not being picked up by the forecast until they occur on day 1. For example, the January 26, 2003, March 14, 2003, October 12, 2003 and the December 10, 2004 events were under-forecast with the 2 to 5-day lead time forecasts. Figure 6 (b) illustrates this for day 5 forecasts.

Figure 2 (b) shows that, on an annual basis, Mica forecasts are unbiased. Fall forecasts stand out as being negatively biased by a larger magnitude. This is due to a general under-simulation at the start of the winter low flow period as Figure 7 (a) and (b) illustrate.

The improvements of the RFS forecast accuracy over naïve forecasts, as measured by the MAE, are summarized in Figure 3 and Figure 4. Figure 3 (a) and (b) show that the forecast skill relative to climatology generally deteriorates from day 1 to day 5 as the accuracy of the RFS system declines. On an annual basis, RFS forecasts remain skillful for all five days of the forecast (Stave: 1 to 5 day lead time annual $SS_{MAE-CLIM}$=29%; Mica: average annual $SS_{MAE-CLIM}$ =35%). However, for some periods of the year the RFS forecast system is less skillful. For example, Figure 3 (a) suggests that Stave RFS forecasts don't provide any improvements over winter 4 to 5-day lead time climatology forecasts ($SS_{MAE-CLIM}$=0%). Also, climatology forecasts provide higher forecast skill than spring 2 to 5-day lead time RFS forecasts ($SS_{MAE-CLIM}$=-9%). The low forecasting skill in winter and spring is probably caused by poor snowmelt forecasts. Although the UBC watershed model uses a sophisticated energy balance method to compute snowmelt, the radiation data are not measured. Instead, they are estimated from daily minimum and maximum temperatures and precipitation. Short-term snowmelt forecasts suffer from this simplification.

Despite the large MAE for Stave threshold flows (not shown), the improvements over naïve climatology forecasts are considerable (1 to 5-day lead-time $SS_{MAE-CLIM}$=52%). This is because the variability of the climatological reference forecasts around the observations, which is used to normalize the forecast error, is even larger. The skill score, therefore, inherently takes the elevated forecasting difficulty for threshold flows under consideration.

For Mica, the improvements over naïve climatology forecasts are largest in spring during the annual freshet (1 to 5-day lead-time $SS_{MAE-CLIM}$=51%). In winter, which is the low flow period in the British Columbia interior, the RFS forecasts are little skillful (1 to 5-day lead-time $SS_{MAE-CLIM}$=4%). During this period the MAE of RFS, climatology and persistence forecasts are all low and differ only by small absolute amounts (not shown). Poor prediction accuracy is achieved for threshold 4 and 5-day lead time RFS forecasts ($SS_{MAE-CLIM}$=-3%).

On an annual basis, the RFS forecast performance relative to persistence forecast is high (Stave: 1 to 5-day lead-time $SS_{MAE-PERS}$=38%; Mica: 1 to 5-day lead-time $SS_{MAE-PERS}$=31%). Generally, the improvements over persistence forecasts increase from day 1 to day 5. For Stave, spring day 1 RFS forecasts are only marginally more accurate than persistence forecasts ($SS_{MAE-PERS}$=8%). For Mica, the RFS forecast skill relative to persistence is high for all periods analyzed (1 to 5-day lead-time $SS_{MAE-PERS}$=32%).

Figure 5 (a) presents the R^2 for the Stave project. On an annual basis, the correlation is

good on day 1 (R^2=80%), but drops to fair on day 5 (R^2=35%). Fall 1 to 4-day lead time forecasts correlate well with the observations (R^2=73%). Fair correspondence between forecasts and observations is obtained for spring forecasts (R^2=37%), 4 and 5-day lead time winter forecasts (R^2=24%) and 2 to 5-day lead time threshold forecasts (R^2=33%).

Mica 1 to 5-day lead time RFS forecasts correlate extremely well (R^2=93%), as do spring forecasts (R^2=86%). Correlation for winter RFS forecasts is low, which is not due to poor forecasts, but due to noise in the observed reservoir inflows (R^2=5%).

Table 3 summarizes the results of the categorical forecast verification for Stave and Mica high flows events. 1 to 5-day lead time forecasts were analyzed together. For the 2003 to 2004 period, the POD for all events above the 90-percentile non-exceedance was 29% for the coastal Stave project. The POD was better for events in the higher >99-percentile category (POD=44%) than in the lower 90-95-percentile category (POD=24%). In comparison, the total POD for the interior Mica project was 45%. No flows above the 98-percentile were recorded for the Mica project during the 2003 to 2004 period.

The FAR for all events above the 90-percentile non-exceedance was 69% and 33% for Stave and Mica, respectively. The FAR is higher in the lower categories, which are closer to the 90-percentile threshold, than in the extreme categories. The 90-95-percentile flow category showed the highest FAR for the Stave project (FAR=82%).

Also, we looked at the 1 to 2-day lead time forecasts to provide POD and FAR for the days for which numerical weather forecasts from the 48-hour higher resolution CMC-GEM model are used (not shown). The scores are slightly better.

To summarize, short-term Stave RFS forecasts are on average over-forecast, but generally provide skill relative to naïve reference forecasts. Exceptions are 2 to 5-day lead time spring forecasts and 4 to 5-day lead time winter forecasts, for which climatology forecasts are more accurate. The large under-forecast and poor correlation of 2 to 5-day lead time threshold forecasts indicates that large events are not being picked up in their magnitude until the event moves into the 1-day lead time horizon. For the same reason, the POD for above 90-percentile events is low (POD=29%). On the other hand, the FAR of is high (FAR=69%).

On an annual basis, short-term Mica RFS forecasts are unbiased, although fall flows are under-forecast. RFS forecasts provide skill throughout most periods of the year. Winter forecasts and threshold 4 and 5-day lead time forecasts are as accurate as naïve climatology forecasts. The Mica POD (45%) and FAR (33%) are significantly better than those for Stave. The discrepancy shows that it is more difficult to predict high flow events in coastal rainfall-dominated systems than in interior snowmelt-dominated systems.

Rainfall or rain-on-snow events predominantly trigger Stave basin's high flows. However, precipitation amounts are difficult to forecast for BC Hydro's coastal basins. Weather forecasting in the Pacific Northwest is difficult, because only a few reporting weather stations are located in the Pacific Ocean where the weather systems originate. Due to this Pacific data void, computer models fill the missing data points with an average between two known points and small errors in these estimates become magnified over time.

Additionally, numerical weather models' resolution is too coarse for accurate quantitative precipitation forecasts in mountainous regions, where the microclimate plays an important role. The complex terrain creates challenges for forecasting due to vertical precipitation and temperature gradients and rain shadow and upslope effects.

Furthermore, forecast precipitation amounts and freezing levels depend on the storm track and, for example, a shift of a few 10's of kilometers to the north or south can make a significant difference. In particular, incorrectly forecast storm tracks cause the relatively large amount of false alarms in the lower two categories for the Stave project.

In comparison, Mica's high flow events are predominantly caused by snowmelt and rain-on-snow events. Snowmelt events affect large regions and the streamflow response is less flashy than in small, rainfall-dominated, coastal basins. In part, the Columbia Wetlands at the Columbia River headwaters are responsible for the attenuation the Columbia River flows and, hence, the relative sluggish response of Mica inflows. In absolute terms, Stave threshold forecasts were less accurate than Mica forecasts, yet, considering the forecasting difficulty, they provided more skill.

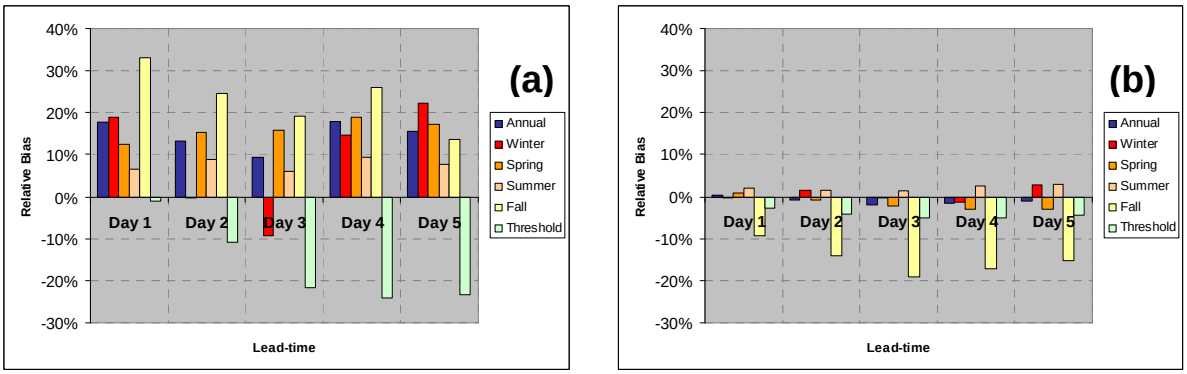

Figure 2 Relative bias (%) for short-term Stave (a) and Mica (b) forecast versus lead time

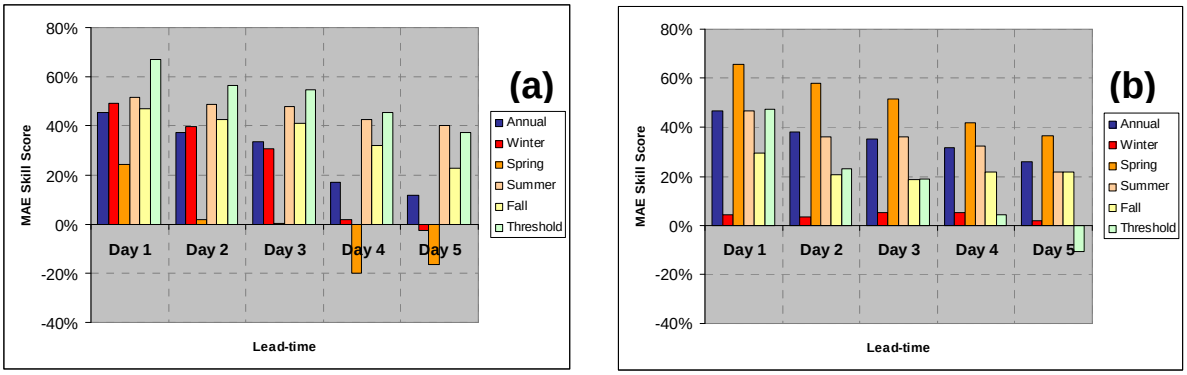

Figure 3 $SS_{MAE\text{-}CLIM}$ skill score for short-term Stave (a) and Mica (b) forecasts relative to climatology versus lead time

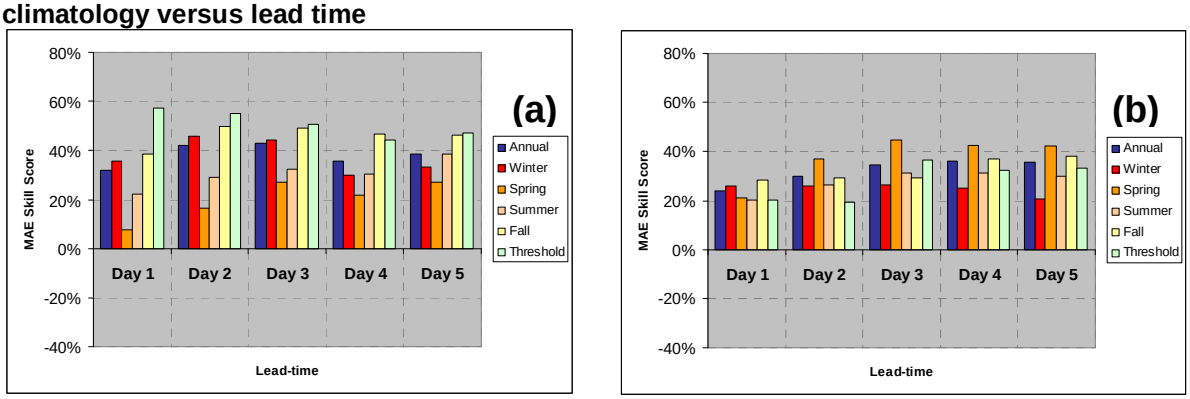

Figure 4 $SS_{MAE\text{-}PERS}$ skill score for short-term Stave (a) and Mica (b) forecasts relative to persistence versus lead time

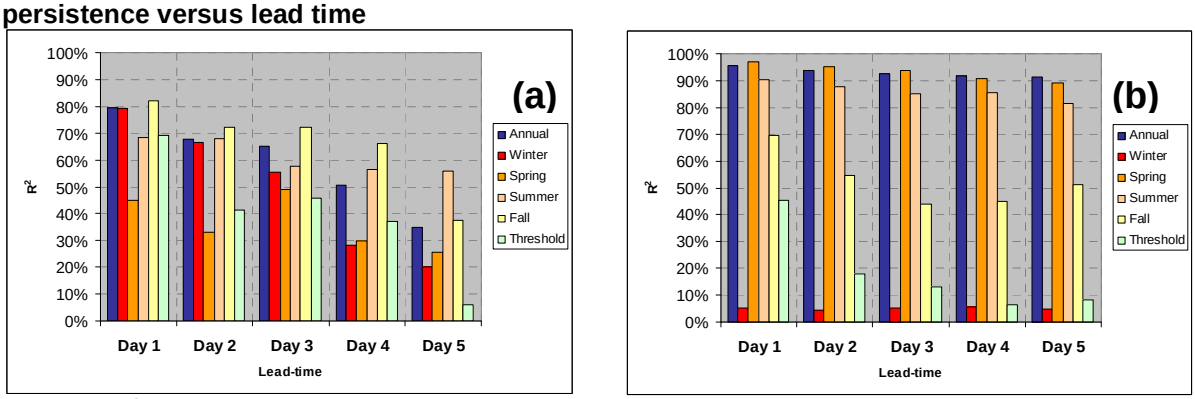

Figure 5 R^2 (%) for short-term Stave (a) and Mica (b) forecast versus lead time

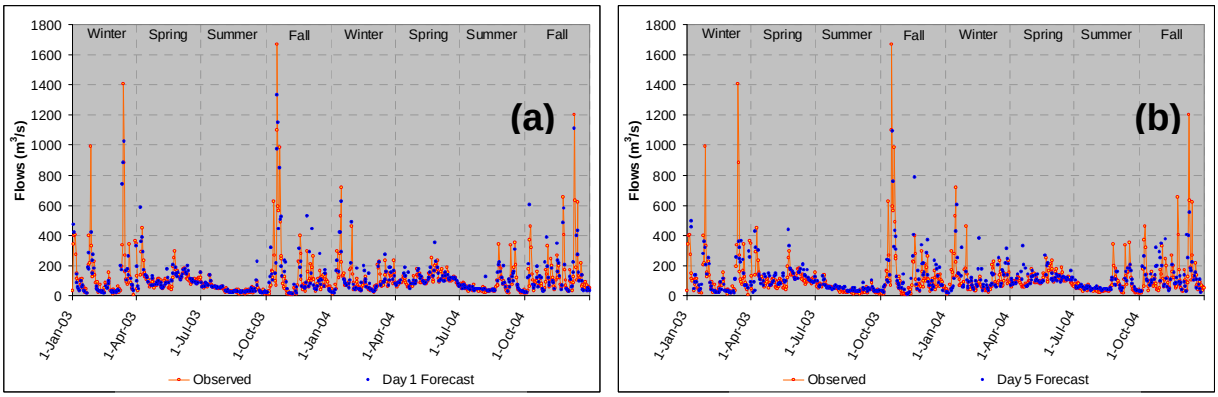

Figure 6 2003-2004 Stave hydrographs and-day 1 (a) and day 5 (b) forecasts

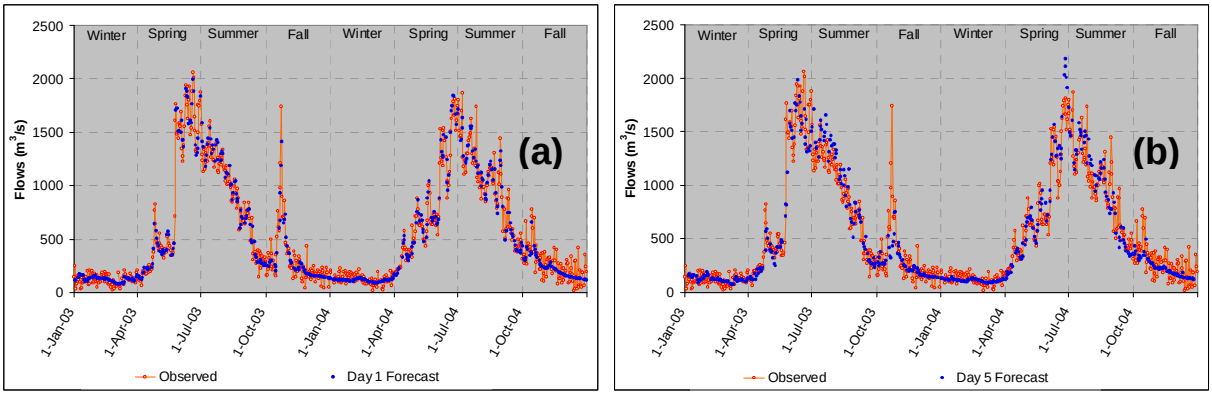

Figure 7 2003-2004 Mica hydrographs and day 1 (a) and day 5 (b) forecasts

Table 3 Comparative categorical high flow forecast verification statistics for the Stave and Mica 1 to 5-day lead time forecasts

	Probability of Detection (POD) Percentile category					False Alarm Ratio (FAR) Percentile category				
	Total	90-95	95-98	98-99	>99	Total	90-95	95-98	98-99	>99
Min Q (m^3/s)	225	225	323	485	650	1500	1500	1780	2086	2280
Max Q (m^3/s)	max	323	485	650	max	max	1780	2086	2280	max
Stave	29%	24%	25%	36%	44%	69%	82%	70%	44%	0%
n	216	88	64	28	36	203	116	53	18	16
Mica	45%	46%	42%	n/a	n/a	33%	38%	13%	n/a	n/a
n	171	121	50	0	0	115	91	24	0	0

3.2 *Performance measures for probabilistic seasonal forecasts*

The seasonal forecasts used in this study are ESP forecasts for the Mica project for the 1980 to 2005 period excluding 1983. The forecast ensembles included between 21 and 39 ensemble members depending on the forecast year. The observed Mica inflows for the 1980 to 2002 period have been quality controlled, while the 2003 to 2005 data are raw. The relative bias, SS_{MAE} relative to climatology, and R^2 of the mean ensemble forecasts and the reported standard errors about the ensemble mean were analyzed for all eight issue dates between January and August. The forecast period is the residual February to September period. The skill of the probabilistic forecasts was analyzed with the CRPS for the February and April forecast issue dates. The reference forecast ensemble is the naïve climatological ensemble, which comprised the monthly or residual distribution of observed 1980 to 2005 reservoir inflows.

Table 4 Comparative forecast performance measures for ensemble mean and ensemble residual Mica forecasts (1980-2005)

		Forecast Issue Month							
		January	February	March	April	May	June	July	August
Score / skill score of the ensemble mean									
RBias	%	2	2	0	0	0	1	2	5
$SS_{MAE-CLIM}$	%	14	27	43	50	56	54	51	73
R^2	%	22	50	62	72	69	74	63	62
Skill score of all ensemble members									
$SS_{CRPS-CLIM}$*	%	-1 (-3)	-4 (14)	14 (44)	19 (50)	18 (56)	29 (56)	16 (43)	3 (37)

* Average scores over n = 21 to 39 years depending on the forecast year; the median is given in parenthesis

Table 4 shows the results of the deterministic verification of the ensemble mean. The expected residual Mica forecasts are basically unbiased for all eight issue dates (average RBias=1.5%). The MAE-based skill score relative to climatology proves that mean ESP forecasts perform better than naive climatology forecasts (average $SS_{MAE-CLIM}$=46%). The improvements relative to climatology forecasts are small for January and February issue dates (January $SS_{MAE-CLIM}$=14%; February $SS_{MAE-CLIM}$=27%) and large from March onwards (March to August $SS_{MAE-CLIM}$=43% to 73%). The relationship between forecast and observations is fair in January (R^2=22%), but good throughout the remainder of the season (R^2=50% to 74%). The results suggest that ensemble mean residual forecasts are unbiased and skillful throughout the entire forecast season.

To calculate the CRPS, a theoretical distribution first needs to be fit to the discrete probability mass function of the probabilistic forecasts. For the type of data analyzed, a very flexible probability density function (PDF) is required since the distribution of the ESP forecasts varies with the issue date and lead time. This justifies the use of a mixture model of distributions (Titterington et al., 1985). Mixture distributions are typically used to model data in which each member (here: an ESP ensemble member) is assumed to have arisen from one of a number of different statistical populations. They also provide a convenient and flexible class of models for PDF estimation. The PDF of a mixture of p probability distributions $f(x|\theta_i)$ can be expressed as :

$$f(x) = \sum_{i=1}^{p} \pi_i f(x|\theta_i) \qquad (10)$$

where π_i are the mixture proportions and θ_i are the component specific parameters. In this study a mixture model of two normal distributions was used to describe the predictive distribution of each ESP forecast. Figure 8 and Figure 9 illustrate the fit of the normal mixture model for monthly forecasts issued in February 2003 and 2004. To test the sensitivity of the scores to different distributions, each CRPS calculation was made for the normal mixture model and for a single normal model. The calculated CRPS are very similar for both distributions (not shown), which shows that it is robust to the probabilistic model fitted to the ESP for the data used.

For the February and April issue dates, the monthly and residual CRPS skill scores were then calculated and their averages are shown in Figure 10 and Figure 11, respectively. The reference forecasts used to evaluate the skill scores are the climatological average forecasts. Figure 10(a) indicates that for the February issue date, the CRPS skill scores are only positive in July and in August. This means that only in these months ESP forecasts outperform the climatological ensemble. For the remainder of the forecast months, CRPS skill scores are negative. The results suggest that the forecast skill for early season RFS forecast ensembles is low for all months other than the snowmelt-dominated month of July. However, since this month bears heavy weight on the seasonal runoff volume it effectively offsets the inaccuracies of the other months of the residual forecast, as shown in Table 4 (February: $SS_{CRPS-CLIM}$ = -4%) .

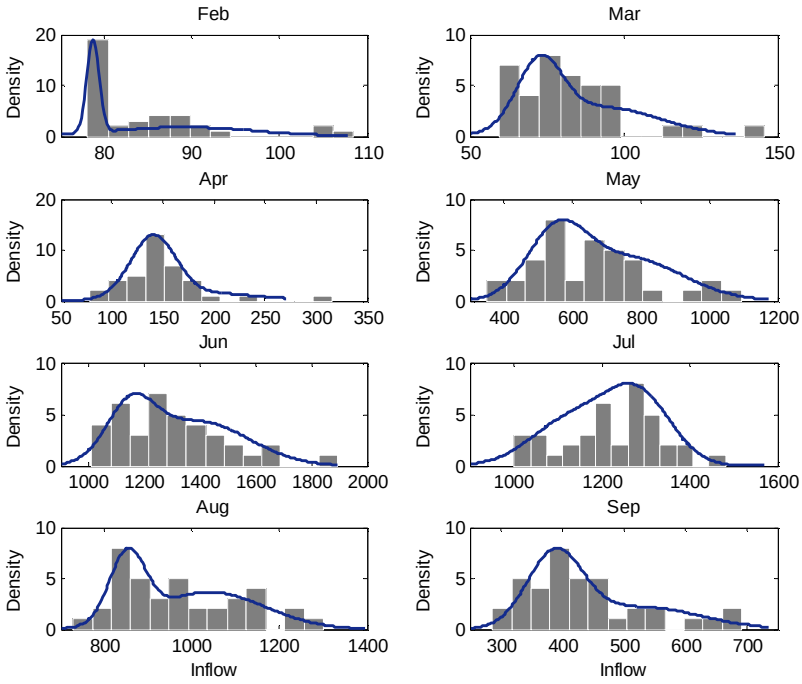

Figure 8 Mixture of two normal distributions fit to the discrete probability mass function of the monthly ESP forecasts for the February 2003 Mica forecast

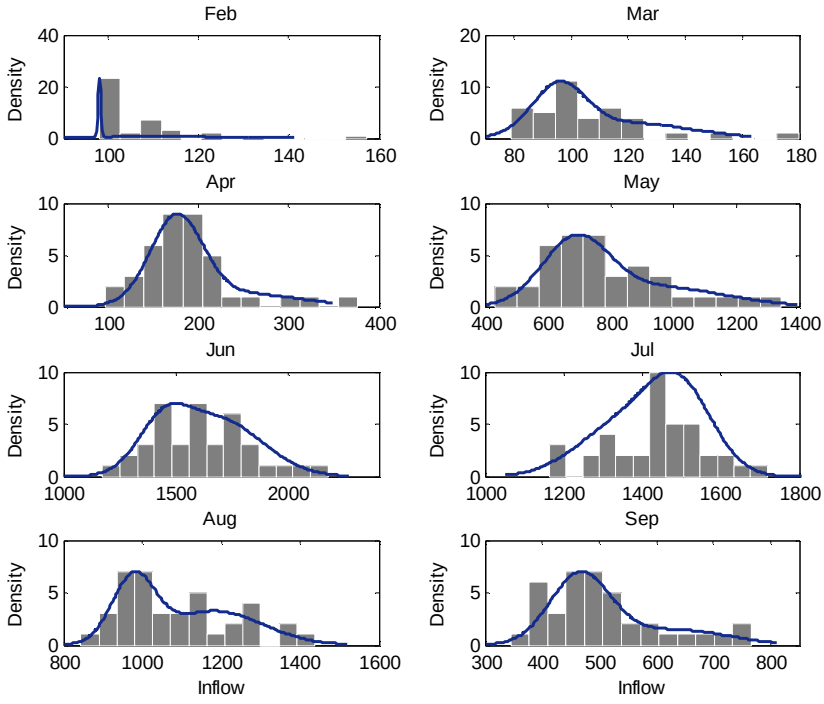

Figure 9 Mixture of two normal distributions fit to the discrete probability mass function of the monthly ESP forecasts for the February 2004 Mica forecast

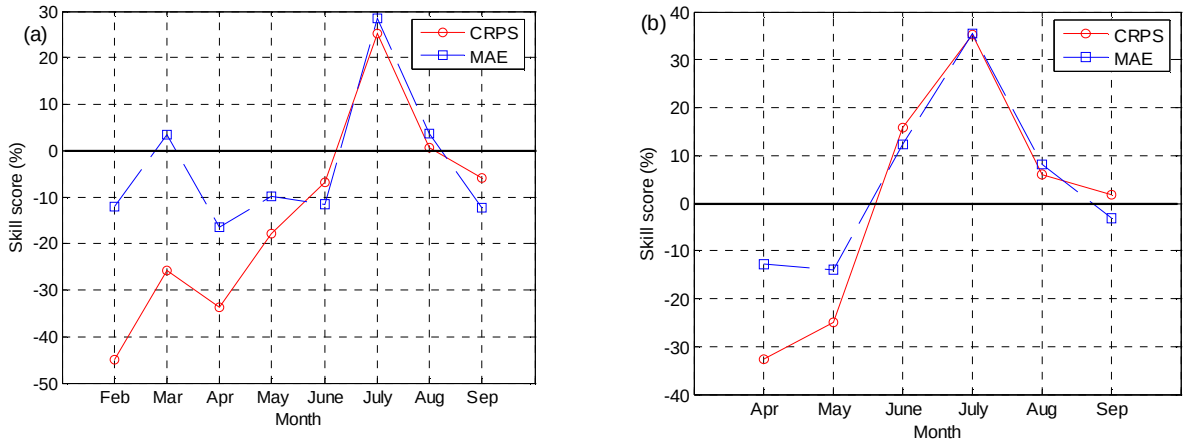

Figure 10 Monthly CRPS and MAE skill score for February (a) and April (b) forecasts (n=25)

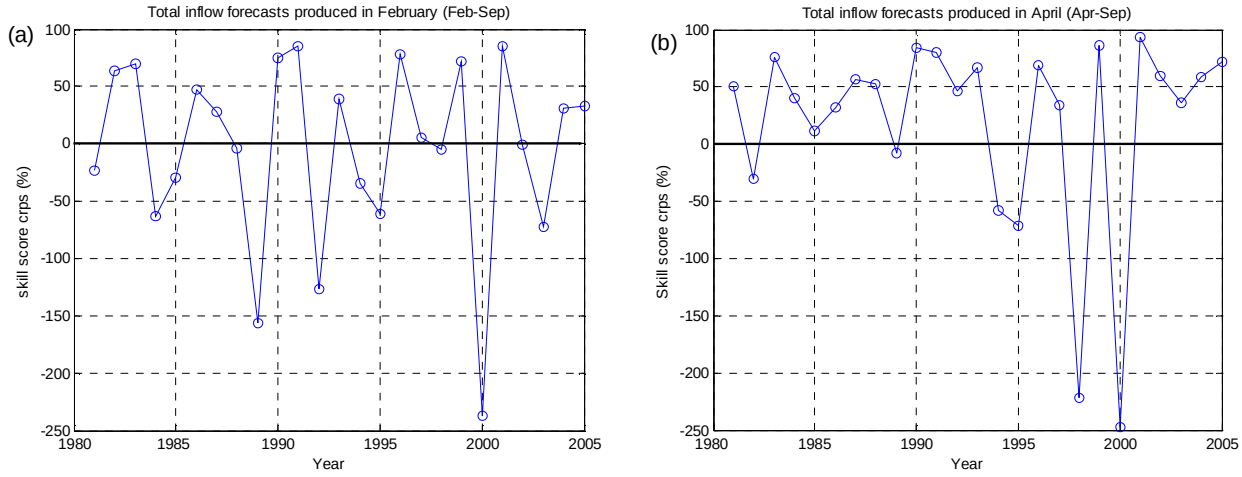

Figure 11 Residual skill score CPRS for the residual February (a) and April (b) forecasts (n=25)

Figure 10(b) illustrates that the April ensembles are more accurate than climatological ensembles during the snowmelt period in June, July, and August. Figure 10 also shows the corresponding MAE skill score obtained for the ensemble mean (dotted blue line). Since the MAE skill score for the deterministic ensemble mean forecast is, in the majority of the months, greater than the CRPS skill score for the ensemble forecasts, it can be concluded that the deterministic forecast outperforms the ensemble forecast. The reason for the relatively low score of the ensemble forecast is that the forecast uncertainty in the ensemble is underestimated.

Table 4 shows that the accuracy of the residual January and February ensemble forecasts is similar to that of a naïve climatological ensemble (January: $SS_{CRPS-CLIM}$=-1%; February: $SS_{CRPS-CLIM}$=-4%). The improved knowledge of the basin snow storage from March onwards results in an increase of the CRPS skill score from the February residual forecast ($SS_{CRPS-CLIM}$=-4%) to the March residual forecast ($SS_{CRPS-CLIM}$=14%). The CRPS skill score of the March to July residual forecasts remains relatively high, while the skill score for the August residual forecast is low.

In Figure 11 (a) and (b) the years 1998 and 2000 stand out due to relatively low CRPS skill scores. They were caused by unusual climatological conditions in the basin. 1998 was characterized by warm and dry El Niño winter conditions across the British Columbia. In

February, however, the forecast skill is typically still low and forecasts tend to be close to the climatological average. On April 1, 1998, the snowpack was well below normal resulting in a low water supply forecast. Contrary to expectations of a low runoff year, the glacier runoff contribution was much above normal due to a warm and dry summer, which offset the forecast water supply deficit. Therefore, the close to normal February ensemble performed better than the low April ensemble. In 2000, February and April climatological ensembles outperformed the RFS ensembles, because reservoir inflows were significantly over-forecast. This was due to a cool spring and summer, which prevented high elevation snow to be melted and glaciers to contribute to runoff. In comparison, the CRPS skill score for 2001 proves a large improvement over climatology. 2001 Mica seasonal inflows turned out to be the lowest since 1961. Forecasts picked up on the extremely dry conditions early on in the season.

The underestimation of the forecast uncertainty is thought to be caused by an under-representation of the weather and modeling uncertainty. Firstly, a limited sample of 21 to 39 historical weather sequences are used as model input to compute scenarios of future inflows. In a stationary climate, the sample variances of about 30 years of data are thought to be representative of the population variance. However, the assumption of climate stationarity is not valid (BC Ministry of Water Land and Air Protection 2002) and recent years' climate has been different from previous years' climate.

Secondly, continuous simulation studies show that, despite of using observed model input, simulations are not perfect. Druce (2001) demonstrated that the weather forecasting error is the dominant source of uncertainty for Mica January forecasts, while for the remainder of the season the weather forecasting and modeling errors are equally contributing factors.

The results suggest that the ensemble mean forecast provides more skill than the probability distribution of the individual forecast ensembles. Early-season residual forecasts demonstrate skill in the mean ensemble (February $SS_{MAE\text{-}CLIM}$=27%), but not for the distribution of individual ensembles (February $SS_{CRPS\text{-}CLIM}$=-4%). Mid-season residual forecasts are more accurate than naïve climatology forecasts for both the ensemble mean (April $SS_{MAE\text{-}CLIM}$=50%) and the ensemble distribution (April $SS_{CRPS\text{-}CLIM}$=17%). The skill in the April forecast ensemble stems from forecasting the snowmelt dominated months of June to August. The lower skill of the ensemble distribution compared to the mean ensemble is due to an under-representation of the forecast uncertainty.

4. SUMMARY AND CONCLUSION

The main goal of this study was to develop a framework for hydrologic forecast verification. Several performance measures were chosen to describe the quality of deterministic and probabilistic hydrologic forecasts. The performance of deterministic forecasts, which include short-term and ensemble mean seasonal forecasts, was measured with the relative bias, a mean absolute error-based skill score relative to naïve climatologic and persistence forecasts and the coefficient of determination. High flow forecasts were additionally scrutinized using the probability of detection and the false alarm ratio as categorical verification measures. The performance of probabilistic seasonal forecasts was assessed using the continuous ranked probability score. A secondary goal was to answer the question whether we are doing all right with our forecasts. The analysis allows the following conclusions:

1. Quantifying the relative bias, mean absolute error-based skill score, and coefficient of determination for deterministic forecasts is an effective way to highlight strengths and weaknesses of forecasting systems. Since these measures are unitless, they allow comparison between forecasts for different forecast points.

2. In this study, the probability of detection and false alarm ratio were used to describe the performance of high flow forecasts. With these two performance measures it is possible to reduce the large amount of verification information to two easily understandable measures. The probability of detection and false alarm ratio are not normalized by the difficulty of the forecast and are, therefore, measures that summarize both the performance of forecast system as well as the forecast difficulty.

3. It was found, that the deterministic short-term forecasts are generally over-forecast, but provide skill for most situations for the coastal Stave basin. Exceptions are 2 to 5-day lead time spring forecasts and 4 to 5-day lead time winter forecasts, where climatology forecasts are more accurate.

4. Due to the difficulty of correctly forecasting large rainstorms on the west coast of British Columbia, large flow events are not being forecast well until they move into the 1-day lead time

horizon. This is also reflected in a low probability of detection and a high false alarm ratio. However, since it is so difficult to accurately forecast these events, RFS high flow forecasts still add substantial forecast skill compared to naïve forecasts.

5. Deterministic short-term forecasts for the interior Mica basin are generally unbiased, except in fall, when they are under-forecast. Winter forecasts and 4 and 5-day lead time threshold forecasts are as accurate as naïve climatology forecasts. The probability of detection and false alarm ratio score better for the interior basin than for the coastal basin.

6. Due to the short period of record, scores for deterministic short-term forecasts are subject to sampling variations and affected by individual events. Therefore, the scores calculated in this study are somewhat representative of the 2003 to 2004 forecasting period. In order to draw general conclusions about the forecasting skill, more years would have to be analyzed.

7. Deterministic seasonal ensemble mean forecasts for the Mica basin are unbiased and, from February onwards, explain more than half of the variability around the observations. On average, the ensemble mean forecasts are 50% more skillful than naïve climatology forecasts.

8. The continuous ranked probability score summarizes the performance of the entire forecast ensemble. It is complementary to the scores for the deterministic ensemble mean forecast in that it helps to shed light on forecast skill of the predictive distribution.

9. The continuous ranked probability score shows that the skill of the ensemble distribution lies in forecasting the snowmelt-dominated months June, July, and August. For the remainder of the forecast months, climatological reference ensembles are as good if not better than RFS ensembles. The residual forecast ensemble does not provide any skill for early-season forecasts, but does for mid-season forecasts.

10. The CRPS seems to be robust to the parametric model chosen to represent the ESP distribution, which means from an operational point of view that a normality assumption is fine if one uses the CRPS to score forecasts.

11. The results suggest that, due to an under-representation of the forecast uncertainty, the empirical probability distribution of the ensemble members has less skill than the ensemble mean. This indicates that while the ESP procedure can give some information about the confidence one should have in the forecast, it clearly does not provide as such a reliable probabilistic forecast. Improvement to the ESP procedure and/or statistical adaptation of the ESP outputs is therefore necessary.

12. While the forecast quality is an effective performance measure for the hydrologist, the forecast value determines whether the forecaster is helping the operations-planning engineer to make better decisions. For this purpose, the optimal reservoir operation based on full hydrological foreknowledge should eventually be compared to historical operation and the value of hydrologic forecasts determined.

Acknowledgments. Special thanks are due to E. Weiss, who designed an early version of the short-term forecast verification tool, as well as to W. D. McCollor provided valuable comments on forecast verification.

5. REFERENCES

Beven K. J., 2001: Rainfall-runoff modeling - The primer, Chichester, United Kingdom, John Wiley & Sons Ltd.

Candille, G. and O. Talagrand, 2005: Evaluation of probabilistic prediction systems for a scalar variable. Q.J.R. Meteorol. Soc., 131, 2131-2150.

Corby, R.J. and W. E. Lawrence, 2002: A Categorical Forecvast Verification System for the Southern Region RFC River Forecasts. NWSRFC, Southern Region Headquaters, Tech. Memo, www.srh.noaa.gov/ssd/techmemo/sr220.pdf (accessed Oct 2005).

Coulibaly, P., M. Haché, V. Fortin, and B. Bobée, 2005: Improving daily reservoir forecasts with model combination, J. Hydrol. Eng., 10, no.2, 91-99.

Day, G. N., 1985: Extended streamflow forecasting using NWSRFS, J. Water Resour. Plan. Manage., 111, 157-170.

Druce D., 1984: Seasonal inflow forecasts by a conceptual hydrologic model for Mica Dam, British Columbia, Amer. Water Resour. Ass., June, 85-91.

Druce D., 2001: Insights from a history of seasonal inflow forecasting with a conceptual hydrologic model, J. Hydrology, 249, 102-112.

Ebert, B., 2005: Verification of ensembles, presentation at TIGGE workshop, 1-3 March 2005,

ECMWF, http://www.ecmwf.int/newsevents/meetings/workshops/2005/TIGGE/Ebert.pdf.

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. J. Hydrometeor., 4, 1105-1118.

Gneiting, T. and A. E. Raftery, 2004: Strictly proper scoring rules, prediction, and estimation. Tech. Report no. 463, Dept. of Statistics, Univ. of Wash.

Lettenmaier, D. P. and E. F. Wood, 1992: Hydrologic forecasting. In: D. R. Maidment (Ed.), Handbook of Hydrology. Chap. 26, McGraw-Hill Inc.

Ministry of Water, Land and Air Protection, 2002: Indicators of Climate Change for British Columbia

Mullusky, M., J. Demargne, E. Welles, L. Wu, and J. Schaake, 2004: Hydrologic applications of short and medium range ensemble forecasts in the NWS advanced hydrologic prediction services (AHPS)., 20th Conference on Weather Analysis and Forecasting/16th Conference on Numerical Weather Prediction Symposium on Forecasting the Weather and Climate of the Atmosphere and Ocean, 84th AMS Ann. Meeting, Seattle, WA.

Pagano,T., D. Garen, and S. Sorooshian, 2004: Evaluation of official western U.S. seasonal water supply outlooks, 1922-2002. J. Hydrometeor., 5, 896-909.

Quick, M. C., 1995: The UBC watershed model. In: V. P. Singh (Ed.), Computer models of watershed hydrology, Water Resources Publications, 233-280.

Titterington, D.M., Smith, A.F.M. and Makov, U.E., 1985: Statistical Analysis of Finite Mixture Distributions. Wiley, New York.

Weiss, E., 2001: Integrating the UBC watershed model into a river forecast system, Presented at the BC Branch CWRA May 9, 2001.

Wilks, D. S., 1995: Forecast verification – statistical methods in atmospheric sciences, Academic Press.

Wood A., A. Kumar, and D. Lettenmaier, 2005: A retrospective assessment of National Centers for environmental prediction climate model-based ensemble hydrologic forecasting in the western United States. J. Geophys. Res., 110, D04105, doi:10.1029/2004JD004508.

Zar, H. J., 1996: Biostatistical analysis. Upper Saddle River, New Jersey: Prentice Hall Inc.