Hydrol. Earth Syst. Sci., 13, 2221–2231, 2009
www.hydrol-earth-syst-sci.net/13/2221/2009/

Hydrology and Earth System Sciences

An evaluation of the Canadian global meteorological ensemble prediction system for short-term hydrological forecasting

J. A. Velázquez[1], **T. Petit**[1], **A. Lavoie**[1], **M.-A. Boucher**[1], **R. Turcotte**[2], **V. Fortin**[3], **and F. Anctil**[1]

[1]Chaire de recherche EDS en prévisions et actions hydrologiques, Université Laval, Québec, Canada
[2]Centre d'expertise hydrique du Québec, Québec, Canada
[3]Recherche en prévision numérique environnementale, Environnement Canada, Montréal, Canada

Received: 19 June 2009 – Published in Hydrol. Earth Syst. Sci. Discuss.: 7 July 2009
Revised: 23 October 2009 – Accepted: 30 October 2009 – Published: 25 November 2009

Abstract. Hydrological forecasting consists in the assessment of future streamflow. Current deterministic forecasts do not give any information concerning the uncertainty, which might be limiting in a decision-making process. Ensemble forecasts are expected to fill this gap.

In July 2007, the Meteorological Service of Canada has improved its ensemble prediction system, which has been operational since 1998. It uses the GEM model to generate a 20-member ensemble on a 100 km grid, at mid-latitudes. This improved system is used for the first time for hydrological ensemble predictions. Five watersheds in Quebec (Canada) are studied: Chaudière, Châteauguay, Du Nord, Kénogami and Du Lièvre. An interesting 17-day rainfall event has been selected in October 2007. Forecasts are produced in a 3 h time step for a 3-day forecast horizon. The deterministic forecast is also available and it is compared with the ensemble ones. In order to correct the bias of the ensemble, an updating procedure has been applied to the output data. Results showed that ensemble forecasts are more skilful than the deterministic ones, as measured by the Continuous Ranked Probability Score (CRPS), especially for 72 h forecasts. However, the hydrological ensemble forecasts are under dispersed: a situation that improves with the increasing length of the prediction horizons. We conjecture that this is due in part to the fact that uncertainty in the initial conditions of the hydrological model is not taken into account.

Correspondence to: J. A. Velázquez
(juan-alberto.velazquez.l@ulaval.ca)

1 Introduction

Most short-term hydrological forecasting systems are deterministic, providing a single value per time step, and hence no information on the uncertainty associated with this forecast. A hydrological ensemble prediction system (H-EPS) seeks to assess and communicate the uncertainty of the forecast by proposing an ensemble of possible forecasts, from which one can estimate the probability distribution of the predictand at each time step (the probabilistic forecast) instead of a single estimate of the flow (the deterministic forecast). A H-EPS offers many benefits: it informs the user about the uncertainty, and it allows decision-makers to determine criteria for alarms based on the probability of exceeding certain thresholds and to test emergency measures on the scenarios proposed by the H-EPS. In short, it allows users to manage the risk associated with decisions based on a forecast.

An ideal probabilistic forecasting system describes all sources of uncertainty. In hydrology, uncertainty arises from (Beck, 1987):

– uncertainty in the values of the parameters that appear in the identified structure of the dynamic model for the system behaviour (model parameter error);

– uncertainty in the model structure i.e. uncertainty about the relationships among the variables characterizing the dynamic behaviour of systems and uncertainty associated with the predictions of the future behaviour of the system (model structure error);

– numerical errors, truncation errors, rounding errors and typographical mistakes in the numerical implementations;

Published by Copernicus Publications on behalf of the European Geosciences Union.

– boundary conditions uncertainties;

– sampling errors, when the data does not represent the required spatial and temporal averages;

– measurement errors;

– human reliability and mistakes.

Efforts towards an H-EPS began in the early 70s. For example, the California-Nevada River Forecast Center developed a procedure that involved running deterministic hydrologic model simulations over the time period for which the discharge forecast was desired, using the historical climate record as input. This procedure provides an ensemble of possible streamflow, given the historical conditions (Day, 1985; Pica, 1997).

Clark and Hay (2004) used 40 years data from the National Center of Environmental Prediction (NCEP) as an hydrological model input to four basins with different hydrological conditions, with areas ranging from 530 to 3630 km^2. The prediction of streamflow was then based on the climatic H-EPS procedure of Day (1985). They found improvements in streamflow forecasts for the snowmelt-dominated basins as compared to the climatic-based ensemble streamflow predictions.

Other studies assessed hydrological forecast improvements using a meteorological ensemble prediction system (M-EPS). For instance, Roulin et al. (2005) evaluated an H-EPS relying on the 50 members ensemble precipitation forecast from the European Center for Medium Range Weather Forecast (ECMWF) for two catchments located in Belgium (1616 and 1775 km^2) over a six-year period. The forecast quality of the hydrological ensemble system was then compared to the probabilistic ensemble based on the climatology. Using the Brier Score and the root-mean-square-error, this study has concluded that the skill of this H-EPS is much better than the one based on historical precipitation inputs.

The coupling of a numerical weather prediction system and a hydrological model was also explored by Bartholmes et al. (2005), in a case of the River Po, an Italian watershed spanning 37 000 km^2. One extraordinary flood event was studied by using deterministic and probabilistic input data in order to compare their performance to predict the magnitude as well as the time to peak discharge. In this case study, the probabilistic forecast was negatively biased in time as well as in discharge magnitude for this particular flood event.

Other hydrological applications of M-EPS have been recently reported. For instance, Jaun et al. (2008) have studied an extreme flood event in August 2005, for a Rhine subcatchment of 34 500 km^2, for which the precipitation had a return period over 10 to 100 years. They coupled the meteorological operational system COSMO-LEPS, which downscales the ECMWF-EPS to a resolution of 10 km, with a semi-distributed hydrological model, and concluded that their H-EPS is effective and provides additional guidance for extreme event forecasting in comparison to a deterministic forecasting system. These conclusions are confirmed by a second analysis reported by Jaun et al. (2009), based on a longer duration (two years). Another case study, over a 9-month period, is presented by Renner et al. (2009). It evaluates the performance of a H-EPS for various Rhine stations: catchments areas ranging from 4000 to 160 000 km^2. Two meteorological ensembles are then confronted: the low resolution ECMWF-EPS and the high resolution COSMO-LEPS ensemble. Results showed that the increased resolution meteorological model provides higher scores, particularly in the short term precipitation forecasts. The authors concluded that there is a need for the downscaling of the ensemble forecast in order to obtain a more representative scale for the sub-basins in the hydrological model.

There are some difficulties by using the H-EPS for flood forecasting. Cloke et al. (2009) discuss some of these problems as, for example, the difficulties in assessing flood forecasts because of their rarity and the difficulties to compare consecutive floods because of the spatial and temporal non-stationarity of the catchments. The authors suggest that there is no other option than to analyse the performance of an EPS driven flood forecast on a case by case basis, and gradually, over the decades, to build up a database of several hundred of flood events on which to base a more thorough flood analysis. This paper also presents an extensive list of recent studies applying ensemble approaches for runoff forecasts with a variety of catchment areas, periods, hydrological models and meteorological EPS.

In July 2007, the Meteorological Service of Canada has improved its M-EPS based on the Global Environmental Multiscale Global Model (GEM), which has been operational since 1998. The purpose of this study is to evaluate the use of this improved Canadian M-EPS as a tool to produce short-range (1–3 days) hydrological predictions and to analyze the uncertainty using probabilistic forecasting. In the next section, the test catchments are described, as well as the hydrometric and the meteorologic data used and the applied methodology. Section 3 presents the results and conclusions are given in Sect. 4.

2 Methodology

2.1 Meteorological forecasting

In the present set-up, the hydrological ensemble prediction system (H-EPS) relies on the output of Environment Canada (EC) meteorological ensemble prediction system (M-EPS), which provides a flow dependent assessment of uncertainty that continuously varies with the state of the atmosphere. Currently, EC's operational M-EPS has a horizontal resolution of 100 km at mid-latitudes and contains 20 ensemble members which are obtained by perturbing the initial conditions and physical parameterizations of the GEM

atmospheric model using an ensemble Kalman filter technique (Houtekamer et al., 2005). In comparison, EC's operational deterministic forecasting systems use a single integration of the GEM atmospheric model on a 33 km grid (at mid-latitudes). For convenience, each 100-km ensemble member was linearly interpolated to the same 33-km resolution grid as the deterministic prediction system that will serve here as benchmark to the M-EPS.

2.2 Hydrological forecasting

The study resorts to the operational flow forecasting system put together by the Centre d'expertise hydrique du Québec (CEHQ) for public dam management (Turcotte et al., 2004), which relies on the hydrological model Hydrotel (Fortin et al., 1995). This system uses 3-h time steps to perform short-term forecasts on small watersheds located upstream of dams with quick hydrological responses. In practice, CEHQ operators combine various objective and subjective procedures to update the system prior of issuing forecasts (Turcotte et al., 2004) – see O'Connell and Clarke (1981) and Refsgaard (1997) for reviews concerning updating methodologies. Here, the operational flow forecasting system is used along with a simple objective output updating based on the last known forecast error. This approach exploits the usually strong autocorrelation of hydrological model errors. For example, Lauzon et al. (1997) have reported that such a simple objective procedure lead to better performance than the application of a Kalman filter.

Hydrotel is a spatially distributed hydrological model with physical bases that performs independent simulations on relatively homogenous hydrological unit (RHHU), taking into account the spatial variability of topography, land use, soil type and meteorological variables within a basin (Fortin et al., 1995). A three-layer vertical water budget takes into account most of the macro-processes in action for infiltration and redistribution of soil-water within RHHU columns. Surface and sub-surface runoff occurs on each RHHU until water reaches the river network. In practice, this runoff consists of water flowing on the soil surface through vegetation and other obstacles, natural and artificial channels too small to be considered as part of the river network, and laterally flowing soil-water. River routing is either based on a kinematic wave or on a diffusing wave.

2.3 Watershed description

The study resorts to twelve watersheds, which areas range from 355 to 5820 km^2 (Table 1) that are parts of five river systems located in the Province of Québec (Canada): Chaudière, Châteauguay, Du Nord, Kénogami and Du Lièvre (Fig. 1). Two criteria dictated this selection: the area of the watersheds had to be suitable for 3-day-ahead forecasts and the watersheds had to be geographically dispersed for the single selected storm producing different rainfall patterns.

The Chaudière River drains 6682 km^2 to the St. Lawrence River south of Québec City: 63% forest and 33% crop. Sites at study in this watershed are the Chaudière at Saint Lambert with two of its sub-catchments, Famine and Chaudière at Sartigan. The Châteauguay River has its source in New York State (USA) and flows towards Lake Saint-Louis, south-west of Montréal. It drains 2543 km^2, of which 30% is forest and 68% is crop. We also consider in this study its Des Anglais sub-catchment. The Du Nord River flows into the Ottawa River, draining 2213 km^2 mostly covered by forest (70%) and crop (10%). The Du Lièvre River also empties into the Ottawa River, draining 9542 km^2 of forested land (75%), with two sites at study, Du Lièvre at Lac Saint Paul and its sub-catchment Lac Mitchinamecus. Finally, the Kénogami Lake drains 1950 km^2, 150 km north of the Saint Lawrence River with three sites at study, Cyriac, Pikauba and Aux Écorces.

Daily streamflow observations are available at all sites from 11 to 31 October 2007 – complete availability is detailed in Table 1. Note that the data for Mitchinamecus consists in reservoir inflow computed using a water budget approach (Haché et al., 1994; Poirier et al., 2005) Streamflows for the selected period are drawn in Fig. 2, after standardization by the historical average value over that same period, in order to assess the severity of the hydrological conditions at hand. Flows in the Chaudière and Châteauguay Rivers exceeded their historical average by up to seven folds. Flows in Du Nord River and flows to Kénogami Lake exceeded their historical average by about two folds. Flows in Du Lièvre River were below average. Site Mitchinamecus on the Du Lièvre River is not shown, because only four years of on-site historical observations are available.

2.4 Experimental set-up

The study resorts to the calibrated Hydrotel model parameters used by CEHQ for operational daily forecasting. A continuous simulation is first performed up to 11 October 2007, based on climate observations and CEHQ state variables. For the test period, the model is again driven by the climate observations, but the state variables are left unconstrained. Flow forecasts are performed for the next 17 days, using EC meteorological predictions. Probabilistic flow forecasts are produced from the M-EPS runs (20 members) and deterministic flow forecasts, from the deterministic run. The H-EPS has a prediction horizon of 72 h consisting of 24 successive 3-h forecasts. The prediction database available for evaluation thus consists of 24 forecasts per site, spanning over 17 days.

Nowadays, performance evaluation of deterministic forecasts is a routine matter. The well known absolute error (AE) score is selected here because it is equivalent the Continuous Ranked Probability Score (CRPS) – described next – for probabilistic forecasts (Gneiting and Raftery, 2007). It thus provides a way to compare the performance of ensemble

Table 1. Streamflow data analyzed.

Code Name	Station Name	Water Survey of Canada Code	Latitude	Longitude	Watershed Area (km^2)	Mean daily discharge (annual) ($m^3 s^{-1}$)	Mean daily discharge (October) ($m^3 s^{-1}$)	Years of available discharge data	Downstream watershed
Chaudière T6	Chaudière at St-Lambert	02PJ005	46°35′16″ N	71°12′59″ W	5820	115.0	91.3	1936–2007	–
Chaudière T7	Beaurivage	02PJ007	46°39′33″ N	71°17′19″ W	709	14.2	11.3	1925–2007	–
Chaudière T63	Famine	02PJ030	46°09′51″ N	70°38′23″ W	691	15.3	13.7	1964–2007	Chaudière T6
Chaudière T106	Chaudière at Sartigan	02PJ014	46°05′52″ N	70°39′22″ W	3070	59.6	60.7	1979–2007	Chaudière T6
Châteauguay T7	Châteauguay	02OA054	45°19′55″ N	73°45′43″ W	2940	37.7	24.5	1970–2007	–
Châteauguay T56	Des Anglais	02OA057	45°10′30″ N	73°50′42″ W	643	8.5	5.68	1974–2007	Châteauguay T7
Du Nord T33	Du Nord	02LC008	45°47′35″ N	74°00′46″ W	1170	23.5	16.5	1930–2007	–
Kénogami T15	Cyriac	02RH066	48°14′07″ N	71°17′23″ W	355	8.7	7.8	1997–2007	–
Kénogami T173	Pikauba	02RH027	47°56′28″ N	71°22′55″ W	495	11.1	14.5	1970–2007	–
Kénogami T323	Aux Écorces	02RH035	48°10′56″ N	71°38′43″ W	1110	28.3	31.5	1971–2007	–
Du Lièvre T34	Lièvre at Lac Saint-Paul	02LE024	46°47′03″ N	75°18′50″ W	4530	85.2	72.1	1979–2007	–
Du Lièvre T50	Mitchinamecus	02LE014	47°12′41″ N	75°10′39″ W	932	14.4	9.6	1963–1966	Du Lièvre T34

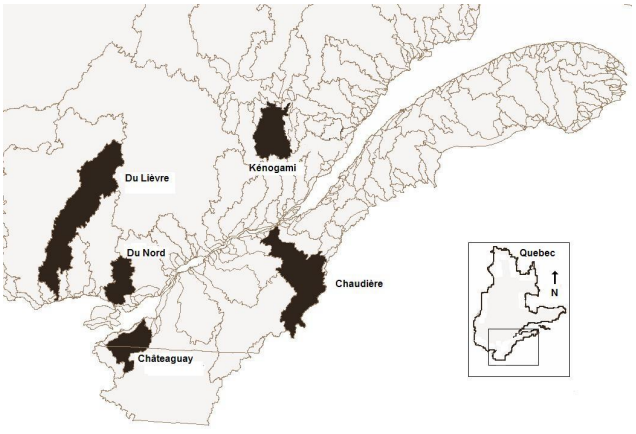

Fig. 1. Localization of the five selected river systems (Province of Québec, Canada).

forecasts against the performance of deterministic forecasts for the same watershed.

Performance evaluation of probabilistic forecasts implies the verification of probability distributions functions against scalar observations. Then, the forecast error can be estimated from a comparison between the forecast value and the verifying value. Performance measure chosen must depend on the reliability, which is the correspondence between the predicted probability and the actual frequency of occurrence (Atger, 1999). Various methods have been proposed to assess the quality of ensemble and probabilistic methods from the meteorological science, and one may chose a probabilistic score that best suits his needs. For the present study they will be described next. However, one should concentrate on scores that are proper (Wilks, 1995; Gneiting and Raftery, 2007). A score that is not proper can favour certain types of forecasts and therefore encourage forecasters to make forecasts that do not represent their true judgement but for which they know that they will obtain a high mark, a practice called "hedging".

The score selected for this study is the Continuous Ranked Probability Score (CRPS) (Matheson and Winkler, 1976), which is a proper score widely used in atmospheric and hydrologic sciences (e.g. Gneiting et al., 2005; Candille and Talagrand, 2005; Weber et al., 2006; Boucher et al., 2009). The CRPS is defined, in its negative orientation, as:

$$\mathrm{CRPS}(F_t, x_t) = \int_{-\infty}^{\infty} (F_t(x) - H\{x \geq x_t\})^2 dx \quad (1)$$

where F_t is the cumulative predictive distribution function for the time t, x is the predicted variable (here the streamflow) and x_t is the corresponding observed value. The function $H\{x \geq x_t\}$ is the Heaviside function which equals 1 for predicted values larger than the observed value and 0 for predicted values lower than the observation. The value that would be taken by the CRPS for a perfect forecasting system is zero. Therefore, one must aim to minimize this score, which is not bounded on the upper side. A known analytical solution of Eq. (1) exists only for normal predictive distributions (Gneiting and Raftery, 2007). Probability plots of the predictive distributions were then drawn to assess normality and, because this hypothesis is not always true, the following Montecarlo approximation to Eq. (1) has been used instead (Székely et al., 2005; Gneiting et al., 2007):

$$\mathrm{CRPS} = E|X - x_t| - 0.5E\left|X - X'\right| \quad (2)$$

where X and X' are independent vectors consisting of 1000 random values from a gamma distribution adjusted to the predictive function.

One interesting properties of the CRPS is that it reduces to the AE in the case of a deterministic forecast. However, because the score obtained by a particular ensemble forecast for one time step has no meaning, we rather consider the average of all individual scores as a measure of the quality of the forecasting system, thus comparing the mean AE (MAE) and mean CRPS ($\overline{\mathrm{CRPS}}$), which values are directly proportional to the magnitude of the observations.

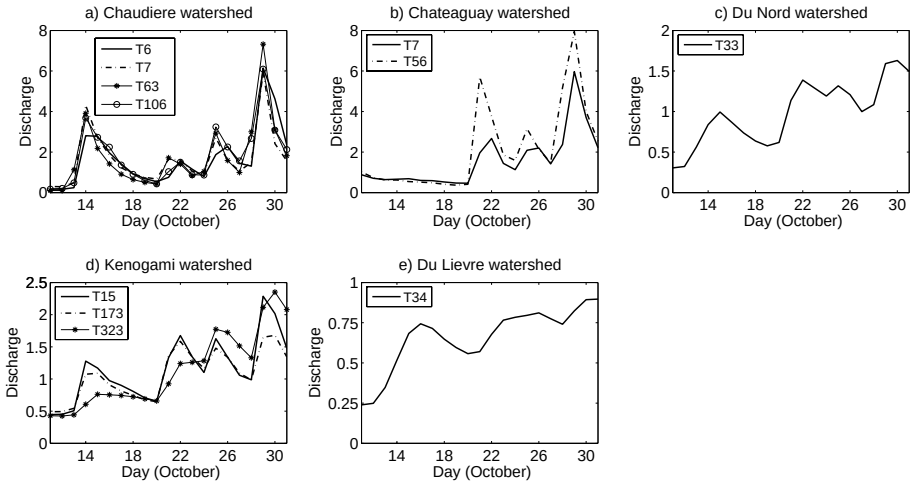

Fig. 2. Standardized streamflow observations from 11 to 31 October 2007.

Hersbach (2000) has shown that the CRPS combines two measures: a reliability component and a "potential CRPS" component. Reliability refers to the statistical consistency between the forecasts and the observations. For instance, a reliable 90% confidence interval calculated using the predictive distribution f_t should on average contain the observed value in 9 cases out of 10. On the other hand, the potential CRPS corresponds to the best possible CRPS value that could be obtained with the database and the particular forecasting system that is used, if the latter was made to be perfectly reliable. Because of the complex nature of the CRPS, other means of assessing the reliability is often used in parallel, such as the rank histogram and the reliability diagram. Unreliable forecasts can be misleading and should be used with caution, if at all. Statistical procedures exist to calibrate unreliable probabilistic forecasts (e.g. Raftery et al., 2005; Fortin et al., 2006; Stensrud and Yussouf, 2007).

The rank histogram or Talagrand diagram (Talagrand et al., 1999; Hamill, 2001), allows one to visually assess the reliability of the predictive distribution. To construct it, the observed value x_t is added to the ensemble forecast. That is, if the forecast has n members, the new set consists of $n+1$ values. Then, the rank associated with the observed value is determined. This operation is repeated for all forecasts and corresponding observations in the archive. The rank histogram is obtained by constructing the histogram of the resulting N ranks. The interpretation of the rank histogram is based on the assumptions that all the members of the ensemble forecast along with the observations are independent and identically distributed; under these hypotheses, if the predictive distribution is well calibrated, then the rank histogram should be close to flat. An asymmetrical histogram is usually an indication of a bias in the mean of the forecasts. If the rank histogram is symmetric and "U" shaped, it may indicate that the predictive distribution is under dispersed. If it has an arch form, the predictive distribution may be over dispersed.

A numerical indicator, linked to the rank, has been proposed by Candille and Talagrand (2005): the ratio δ reflects the squared deviation from flatness of the rank histogram. It is given by

$$\delta = \frac{\Delta}{\Delta_0} \tag{3}$$

where

$$\Delta = \sum_{k=1}^{n+1} \left(s_k - \frac{N}{n+1} \right)^2 \tag{4}$$

and s_k is the number of elements in the kth interval of the rank histogram. For a reliable system, s_k has expectation $N/(n+1)$. Then, Δ_0 is the ratio that would be obtained by a perfectly reliable system, which is

$$\Delta_0 = \frac{Nn}{n+1} \tag{5}$$

Rank histogram requires long time series in order to divide up the observations among $n+1$ ranks. When, as in the present study, one is dealing with a single storm, there is just not enough information to compute rank histograms in the usual way. We are thus suggesting to modify slightly the procedure to allow the computation from say a 20-member 17-d time series. The idea is first to reduce the number of target ranks to $n*+1$ (say 10 or 12) and then to use a bootstrap technique to randomly select $n*$ members from the n-member quasi equiprobable probability distribution. The bootstrapping procedure is then repeated a number of times at each time steps – a number that has to be determined experimentally – and the rank histogram is computed from the combination of all random realizations.

Finally, the reliability diagram is used to graphically represent the performance of probability forecasts of dichotomous events. A reliability diagram consists of the plot of observed relative frequency as a function of forecast probability and the 1:1 diagonal perfect reliability line (Wilks, 1995). In the present study, nine confidence intervals have been calculated with nominal confidence level of 10% to 90%, with an increment of 10% for each emitted forecast. Then, for each forecast and for each confidence interval, it was established whether or not each confidence intervals covered the observation. This is repeated for all forecast-observation pair and the mean effective coverages are then plotted against the nominal confidence levels (Boucher et al., 2009).

3 Results

The first question this study is attempting to answer is: Is there any added value in the Canadian global M-EPS compared with its deterministic counterpart? This question may be answered after comparing the MAE and the $\overline{\text{CRPS}}$ at each 3-h time steps, i.e. independently for prediction horizon spanning from 3 h to 72 h. Each MAE and $\overline{\text{CRPS}}$ estimations are thus computed out of 17 AE and CRPS values, one for each day of the duration of the selected storm. The AE scores describe the hydrological performance based on the deterministic meteorological forecasts, while the CRPS, the hydrological performance based on the 20-member probabilistic meteorological forecasts.

A graphical comparison of the evolution of the MAE and of the $\overline{\text{CRPS}}$, as a function of prediction horizon, is drawn in Fig. 3. All four watersheds of the Chaudière River system present deterministic and probabilistic scores that are close to one another for prediction horizon up to about 48 h (Fig. 3a–d). Then, the supplemental information carried by the M-EPS kicks in: the MAE starts rising substantially while the $\overline{\text{CRPS}}$ remains about the same. In fact, the difference between the MAE and the $\overline{\text{CRPS}}$ at the 72-h horizon is quite remarkable, clearly indicating the superiority of the probabilistic meteorological forecasts over the deterministic one for longer prediction horizon. Similar results are obtained for both watersheds of the Châteauguay River system (Fig. 3e–f), with the exception that the demarcation between the MAE and the $\overline{\text{CRPS}}$ starts earlier, indicating that after about 24 h the M-EPS already provides more information to the hydrological model than the deterministic meteorological forecast.

As already discussed (Fig. 2), the Chaudière and Châteauguay River systems were hit by the selected storm, leading to observed flows up to 4 to 8 times larger than the historical average, at times when dam managers typically have difficulties meeting management objectives. It is thus noteworthy that the M-EPS proves to be particularly useful in such situations.

In all other cases, the storm had a lesser hydrological impact, leading to observed flows twice the historical averages or less (Fig. 2). Results are then mixed. For instance, the superiority of the M-EPS is still quite striking for the watershed on the Du Nord River system (Fig. 3g) and both ones on the Du Lièvre River system (Fig. 3k–l). For the Du Nord River, the $\overline{\text{CRPS}}$ is even substantially lower than the MAE for all prediction horizon. However, performance gains are in general less important for all three watersheds on the Kenogami Lake system (Fig. 3h–j). Nonetheless, even for more standard hydrological events, the $\overline{\text{CRPS}}$ is smaller than the MAE for all watersheds and prediction horizon (except sometimes at 3 h), confirming the superiority of the M-EPS especially for a longer prediction horizon.

The second question this study is attempting to answer is: does the Canadian global M-EPS, used in conjunction with the CEHQ operational flow forecasting system, lead to reliable hydrological forecasts at all time steps? Lack of reliability may orient managers making non optimal decisions. For example, an under dispersed probability distribution will prevent managers appreciating the full uncertainty range of a forecast. This issue is analysed based on ranks histograms and reliability plots.

Figure 4 presents examples of rank histograms computed after 100, 200 and 400 bootstrap repetitions. All those rank histograms are quite similar, which confirms that $n^* = 10$ rank histograms may be successfully drawn for the 17-d storm at hand. Consequently, all rank histograms presented next will be computed from 200 bootstrap repetitions.

The analysis of all rank histograms reveals that, even if they all show signs of under dispersion, this issue improved considerably as the length of the prediction horizon expands. The ratio δ associated with deviation from flatness in a rank histogram is a good way to illustrate this reality. In Fig. 5, it may be seen that for all sites, notwithstanding the intensity of the watershed hydrological response, the ratio δ diminishes from values around 100 to values around 25, for the longest prediction horizon. At this latter stage, the rank histograms, drawn in Fig. 6, indicate that the under dispersion is then small enough.

These findings are confirmed by the reliability diagrams, which incidentally do not resort to a bootstrapping procedure. For example, the 24-h, 48-h, and 72-h reliability diagrams for a selected watershed substantiate that the under dispersion is improving for longer prediction horizons (Fig. 7). At 72-h, however, the under dispersion issue is not completely resolved (Fig. 8).

We also evaluated the lack of reliability due to the M-EPS and to the hydrological model. The M-EPS ratio δ drawn in Fig. 9 shows the flatness of the M-EPS rank histograms for all prediction horizons. The M-EPS may be responsible for part of the reported under dispersion. This is assessed in Fig. 10 that shows the 72 h reliability diagrams for four sites and two approaches: the first one evaluates the reliability of the updated flow forecasting against the observed discharge, as in Fig. 8; the second one evaluates a no-updated forecast against a base line simulation, (a simulation produced with

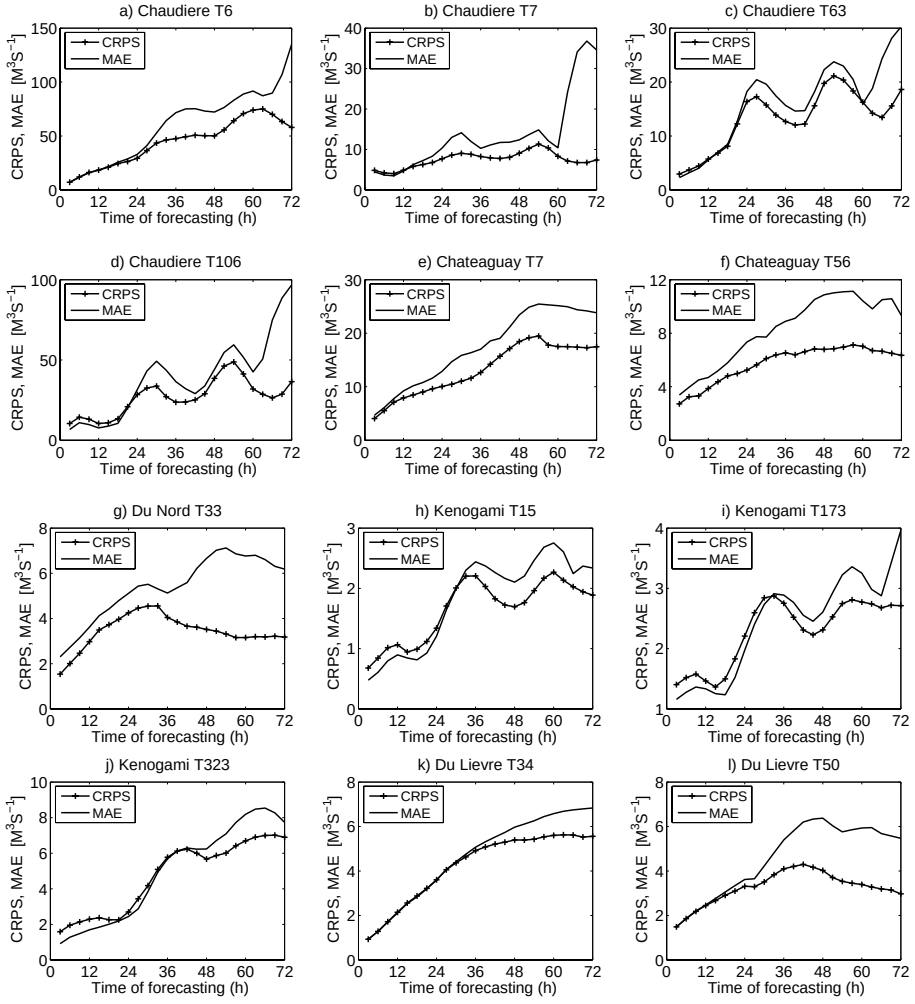

Fig. 3. H-EPS mean probabilistic and deterministic score comparison as a function of the prediction horizon.

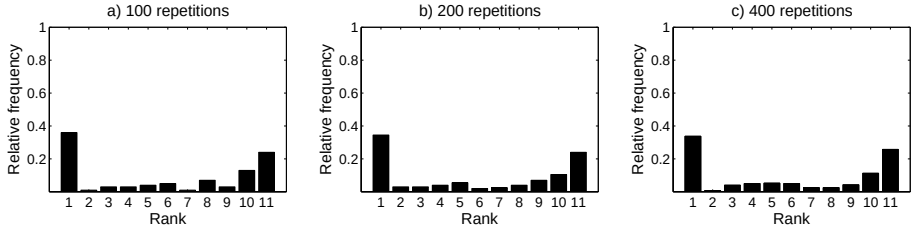

Fig. 4. H-EPS Rank histograms after 100, 200 and 400 bootstrap repetitions: Chaudière T6 watershed, 72-h horizon.

observed precipitation). The first approach is used to evaluate the reliability of the ensemble forecast, including meteorological, hydrological and observational uncertainties, while the second one is used to evaluate reliability due to meteorological model only (e.g. Renner et al., 2009). For all sites, there is an improvement of the reliability, especially for the site Du Lièvre T50, for which it could be inferred that the bias originates from the hydrological model more than in the meteorological forecast. In other cases, like Kénogami T15, both lines are very close to each other and present a more marked under dispersion. This is an indication that the meteorology is biased. In the case of the hydrological model, we could conjecture that part of the lack of reliability of the H-EPS is due to the fact that uncertainty in the initial conditions is not taken into account.

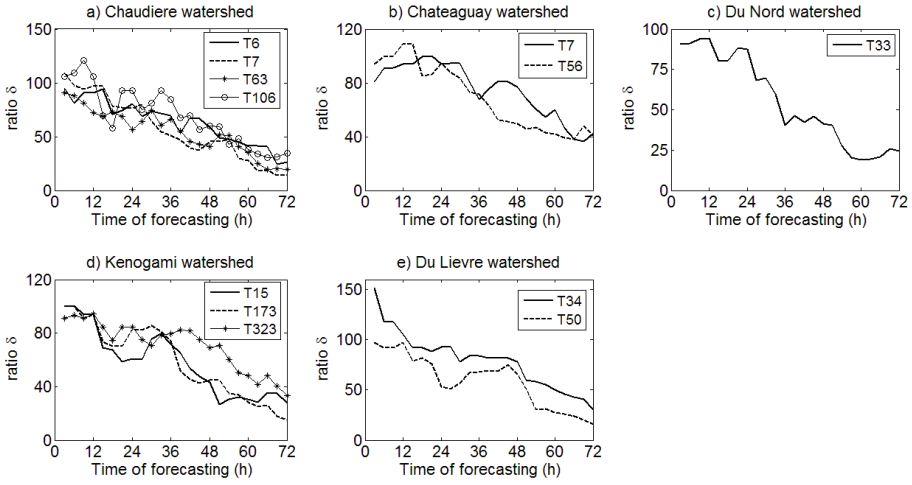

Fig. 5. H-EPS Ratio δ as a function of the prediction horizon.

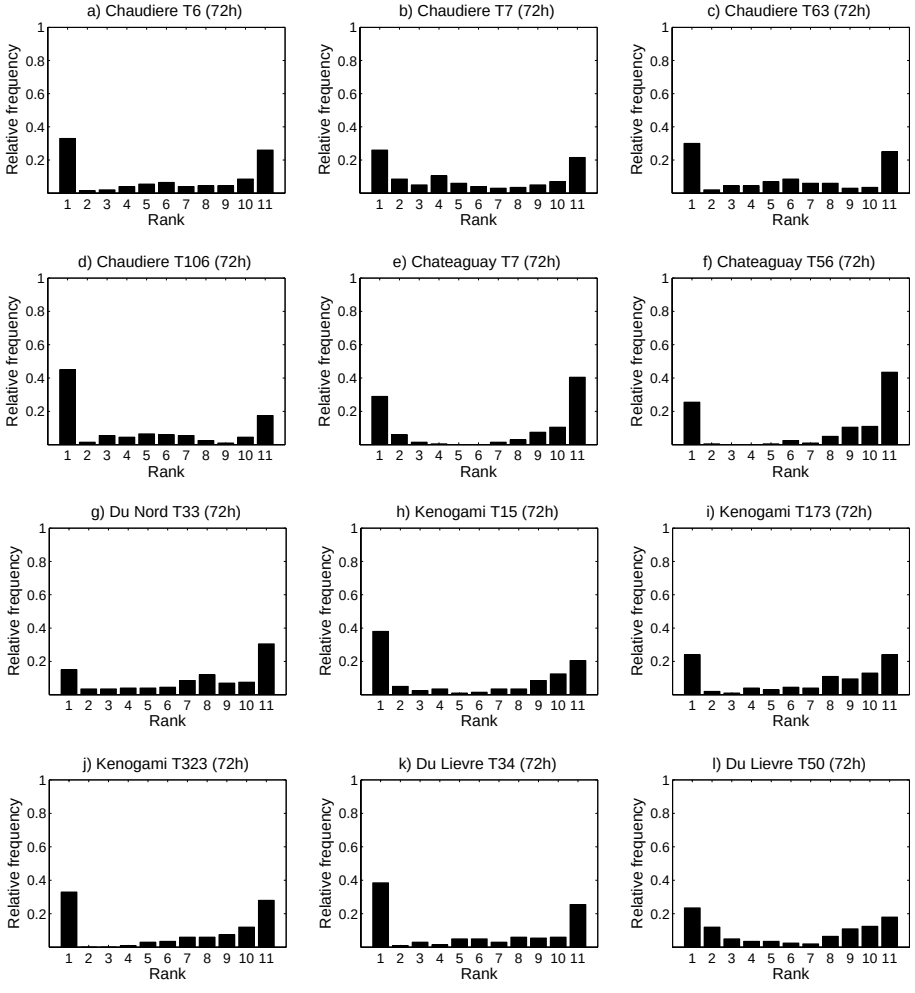

Fig. 6. H-EPS 72-h rank histograms.

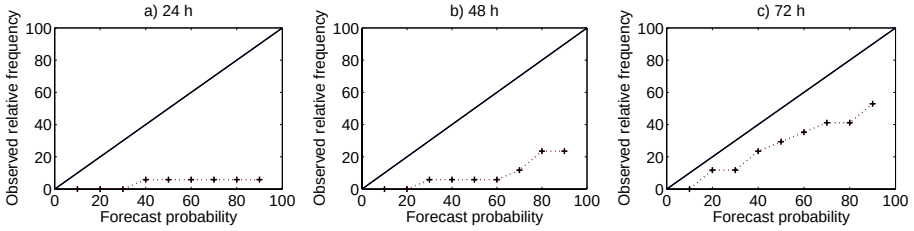

Fig. 7. H-EPS 24-h, 48-h, and 72-h reliability diagrams: Chaudière T6 watershed.

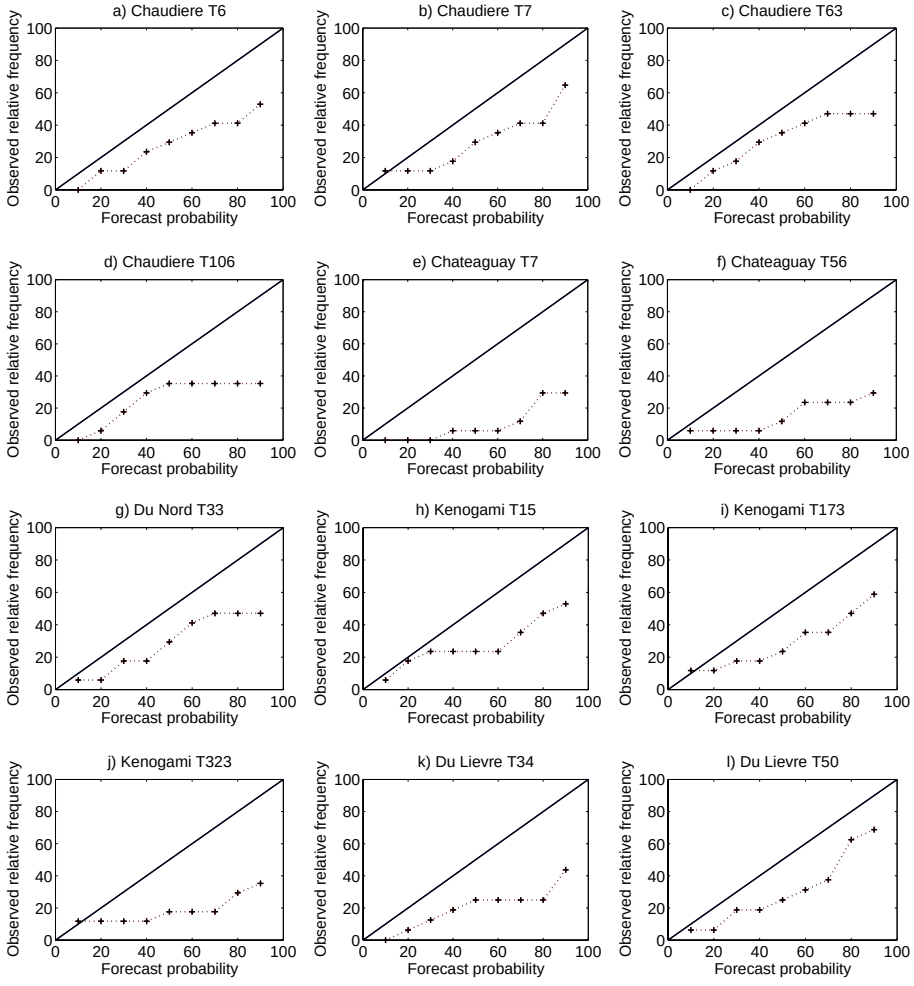

Fig. 8. H-EPS 72-h reliability diagrams.

4 Conclusions

The main scope of this work is assessing the performance and the reliability of a H-EPS based of the latest implementation of the Canadian global M-EPS. The M-EPS encapsulated 20 ensemble members which were obtained by perturbing the initial conditions and physical parameterizations of the GEM atmospheric model using an ensemble Kalman filter technique. The H-EPS resorted to the operational flow forecasting system put together by the CEHQ for public dam management, implemented for twelve watersheds, which are parts of five river systems. A deterministic forecast was also computed for comparison from the EC's operational deterministic forecasting system.

Results, based on a single rain storm, confirmed that the Canadian global M-EPS did contain valuable additional information for hydrological forecasting than its deterministic counterpart. Indeed, the $\overline{\text{CRPS}}$ was lesser than the MAE for all twelve watersheds and for all prediction horizons, clearly indicating the usefulness of the probabilistic meteorological

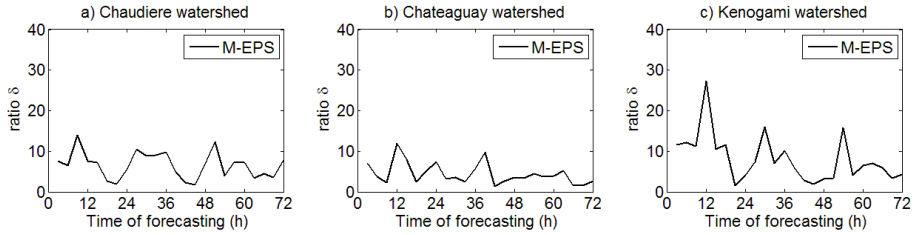

Fig. 9. M-EPS Ratio δ as a function of the prediction horizon.

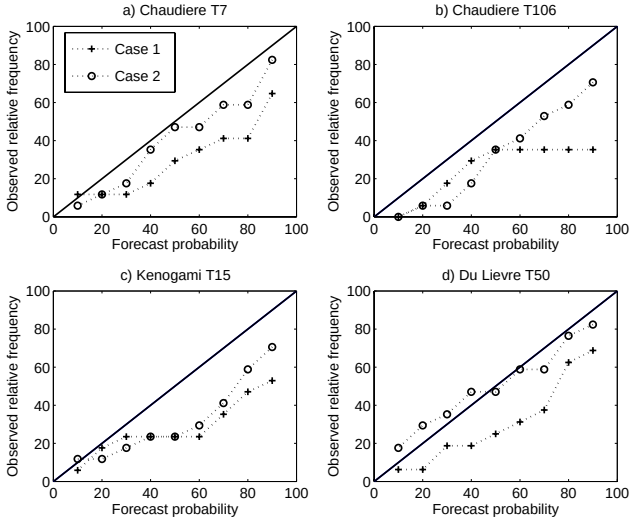

Fig. 10. H-EPS 72-h reliability diagrams. Case 1 evaluates the reliability of an updated flow forecasting against the observed discharge. Case 2 evaluates a no-updated forecast against the base line simulation.

forecasts. The performance gain of the H-EPS was especially important for a longer prediction horizon and for larger hydrological events.

Results also revealed that the Canadian global M-EPS, used in conjunction with CEHQ's operational flow forecasting system, lead to unreliable hydrological forecasts. Indeed, all hydrological forecasts turned out to be under dispersed; however, less so as the prediction horizon increased up to 72 h for which the reliability became reasonable. We have also distinguished the lack of reliability due to the M-EPS and to the hydrological model. Results showed that the meteorology is biased and may be responsible for part of the reported under dispersion. In the case of the hydrological model, we conjecture that the lack of reliability of the H-EPS is in part due to the fact that uncertainty in the initial conditions of the hydrological model is not taken into account.

Further work may generalize the results through an extended simulation period, a selection of hydrological models and a smaller grid resolution of the M-EPS when available.

Acknowledgements. The authors acknowledge the fruitful revisions of J. Schaake and anonymous reviewer. Financial support for the undertaking of this work has been provided the Natural Sciences and Engineering Research Council of Canada, MITACS, and CONACYT (Consejo Nacional de Ciencia y Tecnología, México). François Anctil holds the Chaire de recherche EDS en prévisions et actions hydrologiques.

Edited by: L. Pfister

References

Atger, F.: The skill of ensemble prediction systems, Mon. Weather Rev., 127, 1941–1953, 1999.

Bartholmes, J. and Todini, E.: Coupling meteorological and hydrological models for flood forecasting, Hydrol. Earth Syst. Sci., 9, 333–346, 2005, http://www.hydrol-earth-syst-sci.net/9/333/2005/.

Beck, M. B.: Water quality modeling – a review of the analysis of uncertainty, Water Resour. Res., 23(8), 1393–1442, 1987.

Boucher, M. A., Perreault, L., and Anctil, F.: Tools for the assessment of hydrological ensemble forecasts obtained by neural networks, J. Hydroinform., 11, 297–307, 2009.

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

Clark, M. and Hay, L. E.: Use of Medium-Range Numerical Weather Prediction Model Output to Produce Forecasts of Streamflow, J. Hydrometeorol., 5, 15–32, 2004.

Cloke, H. L. and Pappenberger, F.: Ensemble flood forecasting: a review, J. Hydrol., 375, 613–626, 2009.

Day, G. N.: Extended Streamflow Forecasting using NWSRFS, J. Water Res. Pl.-ASCE, 111(2), 157–170, 1985.

Fortin, J. P., Moussa, R., Bocquillon, C., and Villeneuve, J. P.: HYDROTEL, un model hydrologique distribué pouvant bénéficier des données fournies par la détection et les systèmes d'information géographique, Revue des sciences de l'eau, 8(1), 97–124, 1995.

Fortin, V., Favre, A. C., and Said, M.: Probabilistic forecasting from ensemble prediction systems: Improving upon the best-member method by using a different weight and dressing kernel for each member, Q. J. Roy. Meteor. Soc., B132, 1349–1369, 2006.

Gneiting, T., Raftery, A. E., Westveld III, A. H., and Goldman, T.: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation, Mon. Weather Rev., 133(5), 1098–1118, 2005.

Gneiting, T. and Raftery, A. E.: Strictly proper scoring rules, prediction, and estimation, J. Am. Stat. Assoc., 102, 359–378, 2007.

Hamill, T. M.: Interpretation of Rank Histograms for Verifying Ensemble Forecasts, Mon. Weather Rev., 129, 550–560, 2001.

Haché, M., Larouche, B., Perreault, L., Mathier, L., and Bobée, B.: Validation des apports non contrôlés historiques, INRS-Eau, Sainte-Foy, Québec, Rapport de recherche R-423, 65 pp., 1994.

Hersbach, H.: Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems, Weather Forecast., 15, 559–570, 2000.

Houtekamer, P. L., Mitchell, H. L., Pellerin, G., Buehner, M., Charron, M., Spacek, L., and Hansen, M.: Atmospheric data assimilation with an ensemble Kalman filter: Results with real observations, Mon. Weather Rev., 133(3), 604–620, 2005.

Jaun, S., Ahrens, B., Walser, A., Ewen, T., and Schär, C.: A probabilistic view on the August 2005 floods in the upper Rhine catchment, Nat. Hazards Earth Syst. Sci., 8, 281–291, 2008, http://www.nat-hazards-earth-syst-sci.net/8/281/2008/.

Jaun, S. and Ahrens, B.: Evaluation of a probabilistic hydrometeorological forecast system, Hydrol. Earth Syst. Sci., 13, 1031–1043, 2009, http://www.hydrol-earth-syst-sci.net/13/1031/2009/.

Lauzon, N., Birikundavyi, S., Gignac, C., and Rouselle, J.: Comparaison de deux procédures d'amélioration des prévisions à court terme des apports naturels d'un modèle déterministe, Can. J. Civil Eng., 24, 723–735, 1997.

Matheson, J. E. and Winkler, R. L.: Scoring rules for continuous probability distributions, Manage Sci., 22, 1087–1096, 1976.

O'Connell, P. E. and Clarke, R. T.: Adaptative hydrological forecasting: A review, Hydrolog. Sci. Bulletin, 26(2), 179–205, 1981.

Pica, J.: Review of Extended Streamflow Prediction of the National Weather Service NWSRFS ESP, in: CE505 Conference Course, Civil Engineering, Portland State University, 1 July 1997.

Poirier, C., Turcotte, R., and Lacombe, P.: Procédure de reconstitution d'apports historiques, in: Congrès de l'Association canadienne des barrages, Calgary, 3–5 October 2005.

Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M.: Using bayesian model averaging to calibrate forecast ensembles, Mon. Weather Rev., 133, 1155–1174, 2005.

Refsgaard, J. C.: Validation and intercomparison of different updating procedures for real-time forecasting, Nord. Hydrol., 28, 65–84, 1997.

Renner, M., Werner, M. G. F., Rademacher, S., and Sprokkereef, E.: Verification of ensemble flow forecast for the River Rhine, J. Hydrol., 376, 463–475, 2009.

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

Stensrud D. J. and Yussouf, N.: Reliable probabilistic quantitative precipitation forecasts from a short-range ensemble forecasting system, Weather Forecast., 22(1), 3–17, 2007.

Székely, G. J. and Rizzo, M. L.: A new test for multivariate normality, J. Multivariate Anal., 93(1), 58–80, 2005.

Talagrand, O., Vautard, R., and Strauss, B.: Evaluation of the probabilistic prediction systems, in: Proceedings, ECMWF Workshop on Predictability, Shinfield Park, Reading, Berkshire, ECMWF, 1–25, 1999.

Turcotte, R., Lacombe, P., Dimnik, C., and Villeneuve, J. P.: Prévision hydrologique distribuée pour la gestion de barrages publics du Québec, Can. J. Civil Eng., 31, 308–320, 2004.

Weber, F., Perreault, L., Fortin, V., and Gaudet, J.: Performance measures for probabilistic hydrologic forecasts used at BC-Hydro and Hydro-Québec, in: EGU Conference, April 2006.

Wilks, D. S.: Statistical Methods in the Atmospheric Sciences, Academic Press, San Diego, CA, 465 pp., 1995.