A Comparison of the Canadian Global and Regional Meteorological Ensemble Prediction Systems for Short-Term Hydrological Forecasting

MABROUK ABAZA AND FRANÇOIS ANCTIL

Chaire de Recherche EDS en Prévisions et Actions Hydrologiques, Université Laval, Quebec, Canada

VINCENT FORTIN

Recherche en Prévision Numérique Environnementale, Environnement Canada, Montreal, Canada

RICHARD TURCOTTE

Centre d’Expertise Hydrique du Québec, Quebec, Canada

(Manuscript received 26 July 2012, in final form 1 March 2013)

ABSTRACT

Meteorological ensemble prediction systems (M-EPS) are generally set up at lower resolution than for their deterministic counterparts. Operational hydrologists are thus more prone to selecting deterministic meteorological forecasts for driving their hydrological models. Limited-area implementation of meteorological models may become a convenient way of providing the sought after higher-resolution meteorological ensemble forecasts. This study aims to compare the Canadian operational global EPS (M-GEPS) and the experimental regional EPS (M-REPS) for short-term operational hydrological ensemble forecasting over eight watersheds, for which performance and reliability was assessed. Higher-resolution deterministic forecasts were also available for the study. Results showed that both M-EPS provided better performance than their deterministic counterparts when comparing their mean continuous ranked probability score (MCRPS) and mean absolute error (MAE), especially beyond a 24-h horizon. The global and regional M-EPS led to very similar performance in terms of RMSE, but the latter produced a larger spread and improved reliability. The M-REPS was deemed superior to its operational global counterpart, especially for its ability to better depict forecast uncertainty.

1. Introduction

Hydrological ensemble prediction systems (H-EPS), which rely heavily on meteorological ensemble prediction systems (M-EPS) that are devised to assess the uncertainty of the forecast (Leith 1974; Ehrendorfer 1997), are gradually replacing current operational deterministic prediction systems that provide no information on uncertainty. By generating a series of forecasts rather than a single deterministic forecast, ensemble prediction systems (EPS) provide a means to quantify and study the propagation of uncertainty in forecasting, in the form of probabilistic forecasts and probability density functions—an ability that has already been well received (Palmer 2002). Accordingly, an H-EPS allows situations having a high level of predictability and those with a low level of predictability (Verret 2010).

It is of particular interest to hydrologists to study both the potential and the involvement of rainfall and temperature inputs to hydrological modeling systems with respect to the propagation of uncertainty. The ensemble weather forecast is defined as a forecasting technique in which we spin a numerical model prediction several times for a given situation in order to make predictions from slightly different initial conditions (here called initial perturbations) that are consistent with existing knowledge of the uncertainties of the initial state of the atmosphere (Palmer et al. 2002). These prediction systems appeared toward the end of the 1980s. Since 1992, agencies such as the European Centre for Medium-Range Weather Forecasts (ECMWF; Palmer et al. 1993; Buizza and Palmer 1995; Buizza et al. 2007), the National Centers for Environmental Prediction (NCEP; Toth and
Kalnay 1993; 1997), and the Canadian Meteorological Center (Houtekamer et al. 1996) have been involved in issuing operational ensemble forecasts (Boucher et al. 2009). However, methodology differs from one center to another, either by the number of members or by the method of perturbing the initial conditions (Buizza et al. 2007). H-EPS development has been promoted by European [the European Flood Awareness System (EFAS)] and international [the Hydrological Ensemble Prediction Experiment (HEPEX)] initiatives. EFAS was launched in 1993 by the European Commission while HEPEX was launched in 2004 (Schaake et al. 2007; Thielen and Schaake 2008), bringing together the meteorological and hydrological communities around the world to conduct research focused on the promotion of probabilistic hydrologic forecast techniques. This favors the testing of ensemble forecasts in different climatic conditions while increasing the scientific community involved in this subject.

In July 2007, the Meteorological Service of Canada improved its M-EPS, which has been in operation since 1998. It now relies on the Global Environmental Multiscale (GEM) model for generating a set of 20 members on a 100-km grid (at midlatitudes) over a 15-day forecast horizon (Velázquez et al. 2009). GEM is an operational model that can be used for a wide range of spatial scales and for a variety of meteorological applications (Toth et al. 2010). In parallel, the Meteorological Service of Canada has been experimenting with a limited-area version of the GEM to produce regional meteorological ensembles on a 33-km grid over a 3-day horizon (Charron et al. 2011). The purpose of this study is to compare the global (operational) and regional (experimental) Canadian M-EPS under an H-EPS frame, when producing short-range (up to 3 days) hydrological predictions with uncertainty. The regional EPS (M-REPS) became operational in September 2011, but only experimental products were available at the time this study was performed.

In this study, the watersheds, database, and methodology for the comparison are described in section 2, results are shown in section 3, and a discussion is provided in section 4.

2. Material and methodology

a. Watershed description and observation availability

Eight watersheds were selected for the present study from four river systems located in the province of Québec (Canada). Two criteria dictated the selection: the areas of the watersheds had to be suitable for 3-day-ahead forecasts and the watersheds had to be geographically dispersed for the selected storms producing different rainfall patterns.

The four selected river systems—Chaudière, Châteauguay, Kénogami, and Saumon—are illustrated in Fig. 1. The Chaudière River, located on the south shore of the St. Lawrence River near Québec City, drains an area of 6682 km². The river takes its source from Lake Mégantic at the southern end of the basin and flows 185 km north to empty into the St. Lawrence. Land use consists of 69.5% forest, 22.9% urbanized areas, 3.6% water, and 3.8% wetlands (Thibault 2008). The river provides three watersheds to the study, as detailed in Table 1. The Châteauguay River drains an area of 2543 km², which straddles Canada and the United

FIG. 1. Localization of the four selected river systems (Source: CEHQ).
States in the respective proportions of 57% and 43%. The river rises in the state of New York and empties into Lake St-Louis on the south shore of the St. Lawrence River. Land use in this watershed is 59.7% agriculture, 33.1% forest, 3.4% water and wetlands, and 3.8% urban areas (Simoneau 2007). This river provides one watershed to the study. Kénogami Lake drains a forest-dominated area of 1950 km$^2$ located 150 km to the north of the St Lawrence River. It also provides three watersheds to the study. The Saumon River, about 80 km in length, drains 738 km$^2$ of mostly forested land.

Daily rainfall hyetographs for each watershed are shown in Fig. 2. Local accumulation covers a period from 23 September to 25 October 2010 for the Saumon watershed and from 23 September to 27 October 2010 for the others. Date selection was largely imposed by the availability of the experimental regional ensembles described below. Totals vary from 116 mm for Châteauguay to 237 mm for Kénogami at Cyriac, with daily maxima ranging from 30 mm for Chaudière at Beaurivage to 58 mm for Kénogami at Cyriac.

The accompanying standardized hydrographs are shown in Fig. 3, where values of the observed streamflow are divided by the mean flow over the respective periods in order to assess the severity of the flood events. For instance, observations surpassed the normal about 21-fold on the Saumon River, from eightfold to 23-fold on the Chaudière River, sevenfold on Kénogami Lake, and 18-fold on the Châteauguay River.

All observations were provided by the Centre d’Expertise Hydrique du Québec including 3-h streamflow time series at the outlet of the watersheds and 3-h

<table>
<thead>
<tr>
<th>Code name</th>
<th>Name of the gauging station</th>
<th>Water Survey of Canada code</th>
<th>Qc code</th>
<th>Lat</th>
<th>Lon</th>
<th>Watershed area (km$^2$)</th>
<th>Years of data availability</th>
<th>Downstream watershed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chaudière T6</td>
<td>Chaudière at Saint Lambert</td>
<td>02PJ005</td>
<td>023402</td>
<td>46°35’16”N</td>
<td>71°12’59”W</td>
<td>5820</td>
<td>1915–2010</td>
<td>—</td>
</tr>
<tr>
<td>Chaudière T7</td>
<td>Beaurivage</td>
<td>02PJ007</td>
<td>023401</td>
<td>46°39’33”N</td>
<td>71°17’19”W</td>
<td>709</td>
<td>1925–2010</td>
<td>—</td>
</tr>
<tr>
<td>Chaudière T106</td>
<td>Chaudière at Sartigan</td>
<td>02PJ014</td>
<td>023429</td>
<td>46°05’52”N</td>
<td>70°39’22”W</td>
<td>3070</td>
<td>1969–2010</td>
<td>Chaudière T6</td>
</tr>
<tr>
<td>Châteauguay T7</td>
<td>Châteauguay</td>
<td>02OA054</td>
<td>030905</td>
<td>45°19’55”N</td>
<td>73°45’43”W</td>
<td>2940</td>
<td>1970–2010</td>
<td>—</td>
</tr>
<tr>
<td>Kénogami T15</td>
<td>Cyriac</td>
<td>02RH066</td>
<td>061024</td>
<td>48°14’07”N</td>
<td>71°17’23”W</td>
<td>355</td>
<td>1997–2010</td>
<td>—</td>
</tr>
<tr>
<td>Kénogami T173</td>
<td>Pikauba</td>
<td>02RH027</td>
<td>061022</td>
<td>47°56’28”N</td>
<td>71°22’55”W</td>
<td>495</td>
<td>1970–2010</td>
<td>—</td>
</tr>
<tr>
<td>Kénogami T323</td>
<td>Écorces</td>
<td>02RH035</td>
<td>061020</td>
<td>48°10’56”N</td>
<td>71°38’43”W</td>
<td>1110</td>
<td>1971–2010</td>
<td>—</td>
</tr>
<tr>
<td>Saumon T219</td>
<td>Saumon</td>
<td>02OE032</td>
<td>030282</td>
<td>45°34’47”N</td>
<td>71°23’06”W</td>
<td>738</td>
<td>1975–2010</td>
<td>—</td>
</tr>
</tbody>
</table>

**Fig. 2.** Rainfall (bars) and accumulated rainfall (lines) observations for the study period.
rainfall and temperature time series kriged to a 0.1° resolution grid (about 10 km × 15 km) encompassing each river systems.

b. Meteorological forecasting

As mentioned previously, this study compares the hydrological interest of the global (operational) and regional (experimental) M-EPS of the Canadian Meteorological Center. Both systems rely on the GEM model for issuing 20-member predictions at a 3-h time step: the global at a 100-km resolution (at midlatitudes) and a 16-day forecast horizon and the regional at a 33-km resolution and a 3-day horizon. A global ensemble Kalman filter (Houtekamer et al. 2009) allows the perturbation of initial conditions and physical parameters of the atmospheric model and provides initial conditions to the regional system (Charron et al. 2011). The regional ensembles were issued once a day at midnight (UTC).

Deterministic meteorological forecasts were also available. They consist of regional 15-km forecasts for the first 3 days and global 35-km forecasts for the rest of the time series.

For spatial consistency with the observed climatic time series used for the initialization of the hydrological model (see below), all meteorological forecasts were bilinearly interpolated to the same 0.1° resolution grid.

In contrast with many implementations of H-EPS, no preprocessing of the meteorological forcing is performed. This is for two reasons. First, there is no long reforecast database available for Canadian ensemble prediction systems, making it very challenging to assess and correct biases. The regional EPS was still experimental at the time the study was performed, so there existed no archive whatsoever of past operational forecasts. Second, the main objective of the paper is to establish whether or not the increase in horizontal resolution of the EPS translates into more skillful and/or more reliable hydrological forecasts. Preprocessing the input would have likely blurred this diagnostic.

c. Hydrological forecasting

For coherence with their operational nature, meteorological ensembles need to be evaluated through an operational hydrological setup (Velázquez et al. 2009). In this study we resort 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 distributed hydrological model Hydrotel (Fortin et al. 2001). This semiphysical model performs short-term forecasts at a 3-h time step 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 to issuing forecasts (Turcotte et al. 2004); see O’Connell and Clarke (1981) and Refsgaard (1997) for reviews concerning updating methodologies. We have proceeded in a similar way, adjusting the observed rainfall intensities in order for

FIG. 3. Standardized streamflow observations for the study period.
the Hydrotel output to follow as closely as possible the observed streamflow time series prior to issuing the forecasts (Boucher et al. 2012); streamflow assimilation was performed over the one-month period that precedes the forecasts. The overall procedure is detailed in Fig. 4. Hydrological forecasts were issued from 0000 UTC 23 September to 0000 UTC 17 October 2010 for Châteauguay, Chaudière, and Kénogami, but only up to 0000 UTC 15 October 2010 for Saumon (due to lack of streamflow data for 15 and 16 October).

d. Implementation and verification

Ensemble forecasts are probabilistic in nature. Their verification and interpretation can be performed using appropriate scores (Martin et al. 2009), which involve the verification of the probability distribution function (the forecast) against a scalar (the observation) at each time step. Many scoring techniques have been proposed (Stanski et al. 1989; Wilks 1995). Selected scores for the present study consist of the continuous ranked probability score (CRPS), the spread-skill plot, the rank histogram, and the reliability diagram. The value of the system also assessed using a cost–loss decision model. The CRPS is a common scoring technique for evaluating probabilistic forecasts in atmospheric and hydrologic sciences (Matheson and Winkler 1976; Stanski et al. 1989; Hersbach 2000). It is presented as follows:

\[
\text{CRPS}(F_t, y) = \int_{-\infty}^{+\infty} [F_t(x) - H(x \geq y)]^2 dx, \tag{1}
\]

where \(F_t\) is the cumulative predictive distribution function for the time \(t\), and \(x\) and \(y\) are the predicted variable (in our case streamflow) and the corresponding observed value, respectively. Also, \(H(x \geq y)\) is the Heaviside function, which equals 1 for predicted values larger than the observed value and 0 for predicted values lower than the observed value. The value of the CRPS lies between 0 and \(+\infty\) and is considered perfect if it is equal to 0. Supposing that the gamma distribution is well suited for ensemble forecasts (which is a plausible hypothesis for streamflows), the CRPS may then be estimated as the following Monte Carlo approximation (Gneiting and Raftery 2007):

\[
\text{CRPS} = E|X - y| - 0.5E|X - X'|, \tag{2}
\]

where \(X\) and \(X'\) are two vectors consisting of 1000 random values from a gamma distribution adjusted to the predictive function \(F\). The symbol \(E\) represents the expected value.

The CRPS is the probabilistic equivalent to the absolute error (AE), as shown by Gneiting and Raftery (2007). It has become a common procedure to use the mean value of these two metrics over an archive: the mean CRPS (MCRPS) and the mean AE (MAE), which is simply given by the following expression:

\[
\text{AE} = |x_i - y_i|. \tag{3}
\]

The spread-skill plot evaluates the ability of the ensemble spread (variance) to depict the forecast error of the archive expressed as the root-mean-square error (RMSE) of the ensemble means (Holt et al. 2009). The RMSE and spread \(\sigma\) are given by

\[
\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - y_i)^2} \tag{4}
\]

The spread-skill plot evaluates the ability of the ensemble spread (variance) to depict the forecast error of the archive expressed as the root-mean-square error (RMSE) of the ensemble means (Holt et al. 2009). The RMSE and spread \(\sigma\) are given by
and
\[ \sigma = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n-1} \sum_{k=1}^{n} (x_{k,i} - \bar{x}_i)^2, \tag{5} \]
where \( N \) is the number of time steps of the archive in order to compare the performance of an EPS to a standard nonprobabilistic system, \( \bar{x}_i \) is the member average at time step \( i \), \( x_{k,i} \) is the value of member \( k \) at time step \( i \), and \( y_i \) is the observed streamflow at time step \( i \). One can conclude that the ensemble successfully predicts the forecast error when the RMSE and spread are equal, but it is underdispersive whenever the RMSE is superior to the spread (Palmer et al. 2005).

The rank histogram (Hamill 2001) is produced from the time series of the rank of the observations (verification) when considered as an additional member sorted in ascending order at each time step (Boucher et al. 2009). The diagram, therefore, has one more category than the number of members that forms the prediction archive. The rank histogram is a simple method for detecting systematic defects of an EPS, such as problems associated with the dispersion of the members (Randrianasolo 2009) or bias. If the range of the members is too narrow, there is a high probability that the observed value will fall outside the predicted values. The first and the last rank of the rank histogram will then be overloaded, resulting in a U-shaped histogram. On the other hand, if the same range is too broad, the observed value will almost always be around the center of the predicted values. An asymmetrical histogram is often an indication of a bias in the mean of the forecasts.

"The flatness of the corresponding rank histogram is therefore a measure of the reliability of the prediction system" (Candille and Talagrand 2005). The ratio \( \delta \) measures the reliability of an ensemble prediction system for a scalar variable and is given by the following expression:
\[ \delta = \frac{\Delta}{\Delta_0}, \tag{6} \]
where
\[ \Delta = \sum_{k=1}^{n+1} \left( s_k - \frac{N}{n+1} \right)^2 \tag{7} \]
and \( s_k \) is the number of elements in the \( k \)th interval of the histogram \( (k \in [1, n+1]) \). For a reliable system, \( s_k \) has expectation \( N/(n+1) \), while \( \Delta_0 \) represents a ratio that would be obtained by a perfectly reliable system:
\[ \Delta_0 = \frac{N \times n}{n+1}. \tag{8} \]

Rank histograms require long enough archives of probabilistic forecasts and corresponding observations in order to distribute prediction ranks along their \( n+1 \) categories. Whenever the archive is short, as in this study, where we explore 20 members that extend over 25 days, the ranking method has to be modified following Velázquez et al. (2009). The number of categories is then reduced from 21 to 11 and a bootstrap technique is applied to randomly select 10 members from the available 20 quasi-equiprobable members. This procedure is repeated many times for each prediction [200 repetitions were deemed sufficient by Velázquez et al. (2009)] and the rank histogram is computed from the combination of all random realizations.

Reliability diagrams propose a graphical evaluation of an EPS, depicting the ability of its predictive distribution to encompass the observation at each time step (Wilks 1995). For example, a reliable 50\% confidence interval calculated using the predictive distribution function should, on average, contain the observed value 5 cases out of 10. In our study, confidence intervals from 10\% to 90\% are explored with an increment of 10\% (Boucher et al. 2009) and drawn against their nominal value. A reliable EPS should, in theory, lead to the representation of a line of slope 1 and intercept 0. A lower curve indicates underdispersion and a higher one, overdispersion. However, it has been shown that the quality of a reliability diagram may also be affected by the length of the available prediction archive (Bradley et al. 2003; Clark and Slater 2006). To assess this situation, we assigned confidence limits to the reliability diagram using a bootstrap technique.

Finally, we must not only look at performance but also at the usefulness of the hydrological forecasting system. We proposed to evaluate the value from the cost–loss decision model as defined by Richardson (2000). Consider a decision maker who must choose either to take action or to do nothing, with the choice depending exclusively on the belief that a given weather event \( X \) will occur or not. If the event occurs and no action has been taken, then a cost \( L \) is incurred. Taking action has a cost \( C \) whatever the outcome, but if the event occurs, a part of the loss \( L_1 \) is prevented. The decision maker is interested in following a strategy that minimizes the cost. When only climatological information is available, there are two choices: always protect or never protect. The first instance results in an average expense of \( C + \mu(L - L_1) \), where \( \mu \) is the frequency of occurrence of the event in the climatological record. If protective action is never taken, the average expense is \( \mu L \). Therefore, the optimal strategy consists in taking protective action if \( C + \mu(L - L_1) < \mu L \) (that is if \( C < \mu L_1 \)) and never taking protective action otherwise. The average expense is then.
\[ E_{\text{climate}} = \min\{C + \mu(L - L_1), \mu L\}. \]  

In the case of a perfect forecast, it is possible to take action only when the event is going to occur. The average expense would then be

\[ E_{\text{perfect}} = \mu(C + L - L_1). \]  

We can define the relative value \( V \) of a forecast system as the reduction in expense as a proportion of what would be obtained by a perfect forecast:

\[ V = \frac{E_{\text{climate}} - E_{\text{forecast}}}{E_{\text{climate}} - E_{\text{perfect}}}. \]  

So the decision maker will benefit from the forecast when \( V > 0 \).

Consider the case of a deterministic forecast: the hit rate \( H \) is defined as the frequency of occurrence of the event after a correct forecast and the false-alarm rate \( F \) as the frequency of occurrence when the event was forecast but did not occur. According to Richardson (2000), \( V \) is a function of \( H, F, \mu, \) and \( \alpha \):

\[ V = \frac{\min(\alpha, \mu) - F\alpha(1 - \mu) + H\mu(1 - \alpha) - \mu}{\min(\alpha, \mu) - \mu\alpha}, \]  

where \( \alpha = C/L_1 \) is the cost of taking an action expressed as a fraction of that part of the potential loss which is protected by that action, which is known as the cost/loss ratio. So the relative value depends only on the parameter \( \alpha \) that describes the decision making situation \( \mu \) that characterizes the hydrologic (climatic) context, and \( H \) and \( F \), which quantify the performance of the forecast system.

When working with a probabilistic forecast, as in the case of this study, a decision maker has to choose the probability threshold \( p_t \) at which he will take an action. From this threshold, the probabilistic forecast is changed into a deterministic one. For each value of \( p_t \), the relative value of the probabilistic forecast can be calculated with Eq. (12). The user can then choose that value of \( p_t \) that results in the largest value of \( V \).

### 3. Results

#### a. Performance

As mentioned earlier, the MAE is the deterministic equivalent of the MCRPS on probabilistic forecasts. MAE and MCRPS scores are thus computed independently for each prediction horizon, spanning from 3 to 240 h at a 3-h time step. To calculate the MAE and MCRPS, a database of streamflows of 25 days for three watersheds (Chaudière, Châteauguay, and Kénogami) and 23 days for the Saumon river watershed are used.

Figure 5a illustrates the evolution of the MAE and MCRPS against the prediction horizon. The MAE originates from the deterministic forecasts while the MCRPS originates from both probabilistic forecasts: the 100-km global forecast and the 33-km regional forecast (only available up to 72 h). To better see the differences between the global and regional products, Fig. 5b presents the same information, but only for the first 72 h. In most basins, deterministic and probabilistic forecasts lead to similar performance for about the first 24 h because a period is required before the meteorological forecast may influence the hydrograph, which remains dominated for some time by the initial conditions imposed by the simulation/assimilation process. During this transition period from deterministic to probabilistic, ensembles are poorly diversified and tend to be reduced to a single deterministic forecast (members are almost all equal). The advantage of the ensemble forecasts is then quite limited. For horizons longer than 24 h, MAE deteriorates faster than MCRPS. The performance of probabilistic forecasts is better than that of deterministic forecasts, confirming the superiority of the H-EPS over its deterministic counterpart. The latter will not be discussed further.

Comparing both types of Canadian M-EPS products, one may notice small MCRPS gains favoring the regional H-EPS. In fact, the difference between the global and regional MCRPS is most remarkable in basins such as Châteauguay and Kénogami at Cyriac, where the regional H-EPS clearly surpasses the global one. We evaluated the evolution of the bias in the mean values of the ensemble forecasts and in the deterministic forecast as a function of forecast horizon and we did not observe any systematic bias. The probabilistic forecast was negatively biased over the whole time period. Figure 5b shows that the MAE of the ensemble forecast is, in most cases, close to the MAE of the deterministic forecast, which is logical because there is no noticeable difference in bias between these products.

#### b. Reliability of the H-EPS

Many authors adopted the point of view of Gneiting and Raftery (2007), in which a good ensemble forecast maximizes sharpness (performance), subject to calibration (reliability). While sharpness refers to the precision of the ensemble members, calibration refers to the statistical consistency between the forecasts and the observations. An ensemble forecasting system is accurate if the ensemble mean is close to the observed value. An ensemble forecasting system is also well calibrated.
(reliable) if the dispersion of the ensemble reflects the true uncertainty of the situation.

Spread skill plots are a simple means of simultaneously assessing performance and reliability, where in a well-calibrated system the spread about the ensemble mean should equal the error of the ensemble mean (Palmer et al. 2005). Figure 6 presents spread-skill plots comparing the global and regional H-EPS. For all sites, the spread increases with horizon, confirming that H-EPS progresses gradually from a deterministic setup to a probabilistic one. At the same time, performance diminishes with horizon, as already depicted by the MCRPS. Most sites provide reliable forecasts around the 72-h horizon (similar RMSE and spread values), with the notable exception of Chaudière at Sartigan and Châteauguay, which may need a longer horizon. Comparing the global and regional H-EPS in Fig. 6, the spread of the regional H-EPS is, in general, closer to the RMSE,
which characterizes a better reliability; but its performance is not necessarily superior to its global counterpart. In fact, superiority, in terms of the RMSE, varies from one site to the other and is quite evenly distributed between sites. Both H-EPS implementations would still require a postcalibration scheme (Boucher et al. 2011) before an operational implementation, especially for short horizons.

Rank histogram is also a common method for visualizing the reliability of probabilistic forecasts. Figure 7 compares the evolution of the ratio $\delta$ for the global and regional H-EPS as a function of the prediction horizon.
The general decrease of ratio $\delta$, with increasing horizon, confirms once again the gradual improvement in H-EPS reliability, with a general advantage to the regional H-EPS in most situations.

Rank histograms for three typical sites and the 72-h horizon are drawn in Fig. 8. For Écorces, the regional H-EPS offers a noticeable advancement in reliability over the global H-EPS, with the $\delta$ ratio improving from 14.85 to 8.09. This is a clear situation under which improved spatial resolution of the limited-area model leads to a better description of the uncertainty of the forecasts. For Saumon, the global H-EPS already offers a reasonable account of the uncertainty while the regional H-EPS fails to improve on it; the small bias is even larger. Finally, both H-EPS implementations at Châteauguay lead to biased forecasts. The limited-area regional model does not correct such a situation. The reliability plots drawn in Fig. 9 concur with the rank histograms of Fig. 8, giving an edge to the regional H-EPS over the global one.

Figures 10–13 provide hyetographs and hydrographs of outstanding forecasts issued by the global and regional M-EPS and H-EPS for the Saumon and Saint Lambert watersheds on 30 September 2010 and 7 October 2010. The cumulative rainfall ensemble predictions are presented in the top panels as a series of box plots, for which the box depicts the central 50% of the predictive distribution, the line inside the box shows the median forecasts, and the "x" illustrates the observation. The streamflow ensemble predictions are drawn in the lower panels as thin lines, while the observation is the thicker line. For 30 September, the global and regional M-EPS differ in the intensity and uncertainty of the predicted rainfall. The M-GEPS issued larger rainfall and uncertainty, which turned out to be closer to reality (Figs. 10 and 11). When integrated in the H-EPS, the global forecasts lead to overall better streamflow predictions than the regional one, except for at the beginning of the flood, which is identified by two regional members, but no global ones. The opposite situation occurred on 7 October (Fig. 12), when the M-REPS provided a better rainfall pattern than the M-GEPS, which reflects on the streamflow forecasts. The global H-EPS considerably underestimated considerably the 8 October peak. This situation is even more evident for the Chaudière at Saint-Lambert watershed, for which the global forecasts largely underestimated the storm, while the regional forecasts were much closer to what actually did happen (Fig. 13).
EPS, and in particular high-resolution EPS, can produce nonrealistic weather events because of the nonlinear response of the model to random perturbations. More work needs to be done to address these particular shortfalls of hydrologic ensemble prediction systems.

c. Economic value of the streamflow predictions

The relative economic value of the hydrologic ensemble prediction system is evaluated for a streamflow exceeding a threshold corresponding to the 2-yr return period at each site and for the studied period

FIG. 9. H-EPS 72-h reliability diagrams provided by (left) global and (right) regional M-EPS. The gray area indicates the 90% confidence interval.

FIG. 10. Cumulative rainfall and streamflow predictions for the Saumon watershed issued on 30 Sep 2010. (a),(b) Cumulative rainfall ensemble predictions are shown as a series of box plots, for which the box depicts the central 50% of the predictive distribution, the line inside the box is the median forecast, and the “x” illustrates the observation. (c),(d) The streamflow ensemble predictions are shown as thin lines, while the observation is the thicker line.
(September–October). Figure 14 illustrates the overall relative value \( V \) of both the hydrological ensemble forecast (obtained by choosing the optimal value of \( p_t \) for each value of the cost/loss ratio \( \alpha \), which represents the envelope of the different curves of different probability thresholds) and the deterministic forecast with a 72-h forecast horizon. It can be seen that the usefulness to a decision maker depends greatly on the particular cost/loss ratio. It is also apparent that the hydrological ensemble prediction system has a positive value for most cost/loss situations and most watersheds. This value varies from one basin to another. For example, for the Saumon watershed, the global H-EPS is useful only when \( \alpha < 0.66 \) and the regional HEPS is useful only for \( \alpha < 0.8 \).

Figure 14 demonstrates the advantage of the ensemble forecast over the deterministic forecast in all watersheds (the economic value \( V \) is always higher when using an ensemble forecast compared to deterministic forecast). The relative value depends largely on the appropriate choice of \( p_t \).

During the course of this study, we have noticed that the M-REPS sometimes locally predicts unrealistically high precipitation amounts, in comparison to climatological data. In actuality those findings were outside of our study areas. It is speculated that this issue is related to the stochastic perturbation scheme of the physical tendencies, which can have nonlinear effects on precipitation. Nonetheless, it is difficult to separate one anomalous local precipitation spike from another that would allow for a reliable description of the rainfall uncertainty. In our case, considering the physical distance between our test watersheds (see Fig. 1), we feel that it would be dubious that a precipitation spike would be so large as to simultaneously encompass all our watersheds. Such a situation occurred once on 30 September, when member 14 turned out to be much higher (and temporally lagged) in comparison to all other members on all sites (see Figs. 10 and 11). It is thus warranted that additional testing (development) of the M-REPS be performed before its operational
implementation, at least for short-term hydrological forecasting.

4. Discussion and conclusions

The comparison focused on the 72-h prediction horizon of the MCRPS (for both H-EPS implementations) and of the MAE (for deterministic meteorological forecasts) as a function of the prediction horizon. It was validated that there was added value for the probabilistic streamflow forecasts over its deterministic counterpart. These basic issues being settled, the rest of the work focused on the comparison of the performance and reliability of the Canadian global and regional (experimental) M-EPS for short-term hydrological forecasting.

As expected, the performance of the global and regional H-EPS diminished for longer prediction horizons, as depicted by the MCRPS and the RMSE applied to the ensemble mean. The analysis based on the RMSE revealed no clear winner between the two H-EPS, with superiority at one site being overshadowed by inferiority at another site. In this respect, the contest turned into a draw.

The situation changed when comparing the MCRPS, which combines performance and reliability, and the ensemble spread. Indeed, the MCRPS of the regional
October 2013 A B A Z A E T A L . 3475

H-EPS turned out to be lower than or equal to its global counterpart for all eight sites. Therefore, the regional H-EPS is considered to be better than or similar to the global H-EPS: a gain that we attribute to the larger spread of the regional ensembles, which overall better depicted the uncertainty of the predictions. So, while both H-EPS implementations progressed gradually from a deterministic setup to a probabilistic one, as revealed by a steadily increasing spread, only the regional H-EPS consistently provided reliable forecasts around the 72-h horizon, showing similar RMSE and spread values for four sites out of eight. This situation was seen only twice in the case of the global H-EPS. The advantage of the regional H-EPS, in terms of reliability, was also confirmed by rank histograms and reliability plots. For example, the regional H-EPS generally lead to lower δ ratios than the global one. Evaluation of the economic value of probabilistic and deterministic forecasts revealed that the forecasts are useful for most cost/loss ratios and most basins. Decision makers must therefore have a good perception or knowledge of the situation and select an appropriate probability threshold.

All in all, the issue discussed here can be synthesized in one simple question: to what level does the M-EPS provide pertinent information on rainfall intensity, timing, and uncertainty? We have drawn attention to some specific forecasts, achieved in the context of this study, for which it is evident that an underestimation or an overestimation of the rainfall intensity immediately impacted the ability of the H-EPS to reproduce a flood event. The same was true for the timing of the storm and for the range of its associated uncertainty. It is not that the operational flow forecasting system used here is exempt of any problems that may have diminished the performance and reliability of our predictions (it surely has); rather, we are just stressing the importance of high-quality M-EPS for streamflow prediction. In this regard, the experimental M-REPS tested in the context of this study was deemed superior to its operational global counterpart, especially in its ability to better depict uncertainty. However, the M-REPS was found to produce anomalous local precipitation spikes, which is an issue that has to be addressed in more detail.

As for the hydrological modeling, some issues remain open such as problems resulting from not accounting for uncertainties associated with the initial conditions of the prediction and uncertainties related to the structure of the hydrological model.

Acknowledgments. The authors acknowledge financial support from Mitacs/Mprime, Hydro-Québec, Manitoba Hydro, and NSERC. The authors also acknowledge constructive comments for anonymous reviewers.

REFERENCES


CORRIGENDUM

MABROUK ABAZA AND FRANÇOIS ANCTIL

Chaire de Recherche EDS en Prévisions et Actions Hydrologiques, Université Laval, Québec, Canada

VINCENT FORTIN

Recherche en Prédiction Numérique Environnementale, Environnement Canada, Montréal, Canada

RICHARD TURCOTTE

Centre d’Expertise Hydrique du Québec, Québec, Canada

(Manuscript received and in final form 20 December 2013)

Spread-skill plots in Abaza et al. (2013) are in error because the RMSE is compared to the average standard deviation instead of the square root of the average variance. Indeed, two different and inconsistent methodologies have been used over the last few years in the meteorological and hydrological literature to compute the average ensemble spread: in some cases, the square root of the average ensemble variance is used; in other cases the average of the ensemble standard deviation is computed instead. The second option, used in Abaza et al. (2013), is incorrect and may lead to false diagnostics of underdispersion. The correct Eq. (5) and Fig. 6 are given below:

\[
\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \frac{1}{n-1} \sum_{k=1}^{n} (x_{k,i} - \bar{x})^2}.
\]

We regret having contributed to the existing confusion about spread-skill plots. The correct Fig. 6 shows lesser underdispersion than before: results that are more consistent with the other assessment tools used in Abaza et al. (2013), namely, rank histograms and reliability diagrams. Interpretations in Abaza et al. (2013) remain valid despite the error.

Please also notice that in Figs. 10–13 of Abaza et al. (2013), the global forecasts stand on the left side and the regional ones stand on the right side.

REFERENCE


**Corresponding author address:** Mabrouk Abaza, Université Laval, Civil and Water Engineering, 1065 Avenue de la Médecine, Pavillon Adrien-Pouliot, Québec, QC G1V 0A6, Canada.

E-mail: mabrouk.abaza.1@ulaval.ca

DOI: 10.1175/MWR-D-14-00018.1

© 2014 American Meteorological Society
Fig. 6. Evolution of RMSE and spread of H-EPS as a function of time of forecasting.