Adv. Geosci., 29, 85–94, 2011
www.adv-geosci.net/29/85/2011/
doi:10.5194/adgeo-29-85-2011

Advances in Geosciences

A comparison between ensemble and deterministic hydrological forecasts in an operational context

M.-A. Boucher[1], **F. Anctil**[1], **L. Perreault**[2], **and D. Tremblay**[3]

[1]Chaire de recherche EDS en prévisions et actions hydrologiques, Département de génie civil et de génie des eaux, Université Laval, Pavillon Pouliot, Québec, G1K 7P4, Canada
[2]Hydro-Québec Research Institute, Varennes, J3X 1S1, Canada
[3]Hydro-Québec, Head Office, Montréal, H2Z 1A4, Canada

Received: 13 July 2010 – Revised: 23 January 2011 – Accepted: 27 January 2011 – Published: 3 March 2011

Abstract. Ensemble forecasts can greatly benefit water resources management as they provide useful information regarding the uncertainty of the situation at hand. However, weather forecasting systems are evolving and the cost for reanalysis and reforecasts is prohibitive. Consequently, series of ensemble weather forecasts from a particular version of the forecasting system are often short. In this case study, we consider a hydrological event that took place in 2003 on the Gatineau watershed in Canada and caused management difficulties in a hydropower production context. The weather ensemble forecasting system in place at that time is now obsolete, but we show that with minimal post-processing of the forecasts, it is still beneficial to exploit ensemble rather than deterministic forecasts, even if the latter emerge from a more advanced meteorological model and possess superior spatial resolution.

1 Introduction

Ensemble forecasts allow decision makers to analyze the uncertainty of the situation at hand, which potentially leads to improved management compared to using point forecasts (e.g. Krzysztofowicz, 2001). However, weather forecasts series from each version of the operational system are often short or outdated as the system improves over the years. In addition, it is virtually impossible to describe perfectly all sources of uncertainty associated with hydrological forecasting in order to directly obtain a perfect exhaustive estimate of the total uncertainty of the forecast. For instance, in large northern territories such as Canada, the precipitation gauges networks are usually too coarse relatively to the territory under study. This poses challenges regarding precipitation data, and in an even greater extent snow data, which in turn have great incidence on timing and magnitude of the spring melt.

This causes Canadian meteorological ensemble forecasts to be under-dispersed and has led many important industrial users to turn their back on the ensemble product, in favor of ESP-type forecasts (Day, 1985) forecasts. Dividing the forecasting horizon into stages, they use the deterministic forecasts in a rainfall-runoff model to obtain short-term streamflow forecasts, and then generate ensembles based on previous years climate for medium- to long-term streamflow forecasting, assuming that the climates of previous years are all equiprobable between themselves and compare to the actual climate. In the context of a warming climate, such palliative strategy may become increasingly misleading.

Considering this, it becomes necessary to post-process the ensemble forecasts before involving them into a decision-making process so that the predictive distributions are reliable and properly reflect real world uncertainty. Many post-processing strategies involve sophisticated statistical manipulations, which can deter their operational use. In this case study, we show that a kernel based post-processing method can, at least partially, compensate for the uncertainties that are not well captured by the hydrological ensemble forecasting process (rainfall-runoff parameterization, uncertainties in the observations, initial soil moisture or snowpack height, for instance) and for the under-dispersion of the meteorological ensembles used. Furthermore, we also show that even raw ensembles can be beneficial for decision-making compared with the use of a deterministic product.

Here we consider a flood event in fall 2003 in the Gatineau watershed in Canada, which caused management complications for Hydro-Québec, the major hydropower producer in the country. We use the weather ensemble forecasting system that was operational in 2003 at Environment Canada, which has a spatial resolution of 200 km and is formed by combining the outputs from two atmospheric models, the models SEF (Ritchie, 1991; Ritchie and Beaudoin, 1994) and GEM

Correspondence to: M.-A. Boucher
(marie-amelie_boucher@uqac.ca)

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

(Côté et al., 1998). Temperature and precipitation forecasts are then used to feed a physics-based distributed hydrological model, which in turn produces streamflow ensemble forecasts for the outlets of the six sub-catchments in the basin, for lead times between 48 h and 240 h. Because the raw streamflow forecasts suffer from under-dispersion and bias, the performance of the final product is then evaluated using the Continuous Ranked Probability Score (CRPS) and its decomposition into reliability and potential components, in addition to the logarithmic score and rank histograms. In order to assess the benefits of using ensemble streamflow forecasts, deterministic forecasts are also used in conjunction with the hydrological model. The 2003 ensemble forecasts have a 200 km spatial resolution, while the 2003 deterministic forecasts have a resolution of 45 km, the highest available. Since the mean CRPS reduces to the mean absolute error for deterministic forecasts (Gneiting and Raftery, 2007), we are able to compare deterministic and probabilistic forecasting systems.

The paper is organized as follows. First, the context of application is provided in Sect. 2, describing the watershed, the particular flood event considered, the meteorological ensembles as well as the hydrological model. The experimental protocol is explained in Sect. 3 and results are presented in Sect. 4, followed by a short conclusion.

2 Context of application

2.1 The Gatineau watershed

The Gatineau River watershed (1) covers 26 785 km^2 and it is 490 m a.s.l. at its highest. The climate is typically continental, with warm and humid summer seasons and cold, humid and cloudy winters. There are, however, important climatic variations between the downstream part of the catchment and its upstream part. The upper part of the catchment can be considered sub-polar, while the central part is described as mild sub-polar and the lower part presents a moderate climate. Although the climate exhibits some variations, the precipitation regime is the same throughout the catchment: mean total rain of 80–100 cm and mean total snow of 200–250 cm.

The Gatineau River crosses several urban areas, the largest of which is the town of Maniwaki. The system comprises two large upstream reservoirs (Cabonga and Baskatong). As illustrated by Fig. 1, the Gatineau watershed can be subdivided into six sub-catchments. Table 1 provides additional information regarding the mean observed daily streamflow for each sub-catchment.

2.2 The flood event of fall 2003

Because the Gatineau River basin comprises inhabited areas, certain operating constraints prevail over hydropower production. For instance, from 1 June to 15 September, the Baskatong reservoir must be filled almost to its capacity to allow boating and recreational activities for nearby residents. The river level must be kept above a specific level to ensure adequate drinking water supply for nearby towns. Finally, the river must also be kept below another level for flooding prevention.

Table 1. Mean daily streamflow for the six sub-catchments of the Gatineau River basin, from 1 January 1950 to 4 November 2004.

Basin	Mean daily streamflow ($m^3 s^{-1}$)
Chelsea	25.14
Paugan	92.14
Maniwaki	54.53
Baskatong	232.17
Cabonga	41.74
Ceizur	127.46

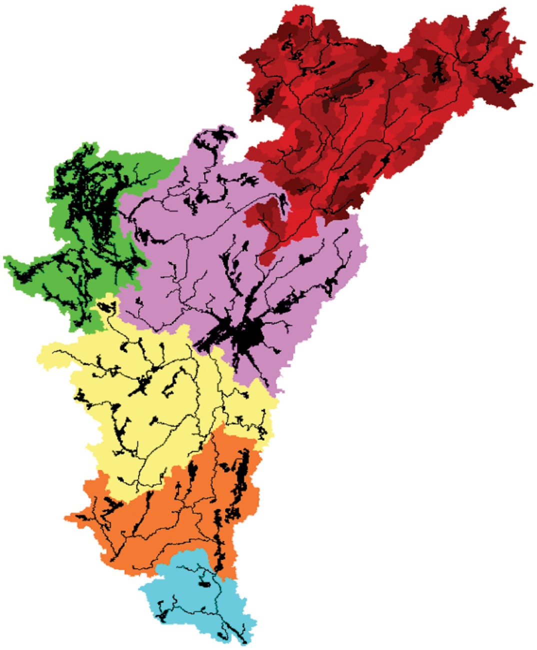

Fig. 1. The Gatineau watershed divided in six sub-catchments: Ceizur (red, 6840 km^2), Cabonga (green, 2662 km^2), Baskatong (purple, 6200 km^2), Maniwaki (yellow, 4145 km^2), Paugan (orange, 2790 km^2) and Chelsea (turquoise, 1148 km^2).

The average volume of a spring flood in the Baskatong reservoir is 3600 hm^3, but the capacity of the Baskatong reservoir is only 3049 hm^3. The routine strategy is to lower the level of the Baskatong reservoir at the end of the winter as much as possible and then let the level rise during spring. Then, the reservoir level is kept all summer within 2.5 to 1.3 m of its maximum level until mid-September. During the

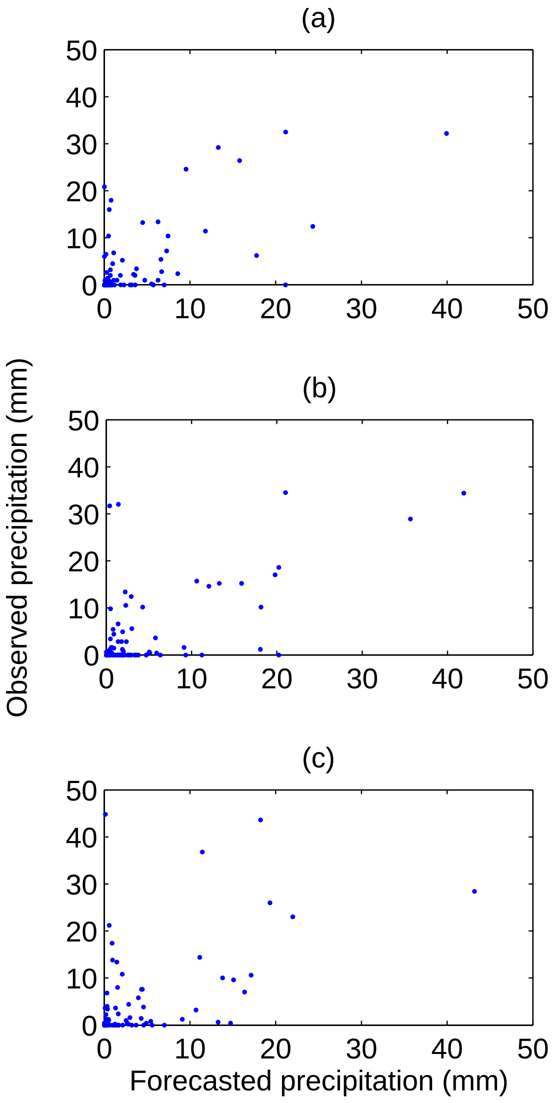

Fig. 2. Scatter plots of 48 h ahead deterministic precipitation forecasts and corresponding observations for **(a)** Station 7031360 (near outlet of the basin) **(b)** Station 7038885 (upper Baskatong, middle of the basin) and **(c)** Station 7038975 (Paugan).

fall, the reservoir is managed so that a sufficient water reserve is cumulated to anticipate electricity demand during winter.

The operating margin for the operation of the Baskatong reservoir is quite small considering the above mentioned constraints and the inflows it receives during certain periods. Consequently, spillage is sometimes inevitable at the hydropower stations in spring and fall. The most significant flood occurred in the spring of 1974. On this occasion, 3000 residents were required to evacuate the area, over one-third of Maniwaki was flooded and 2.9-million Canadian dollars had to be provided in disaster relief. More recently, heavy precipitation in fall 2003 caused important flooding in the municipality of Gracefield south of Maniwaki.

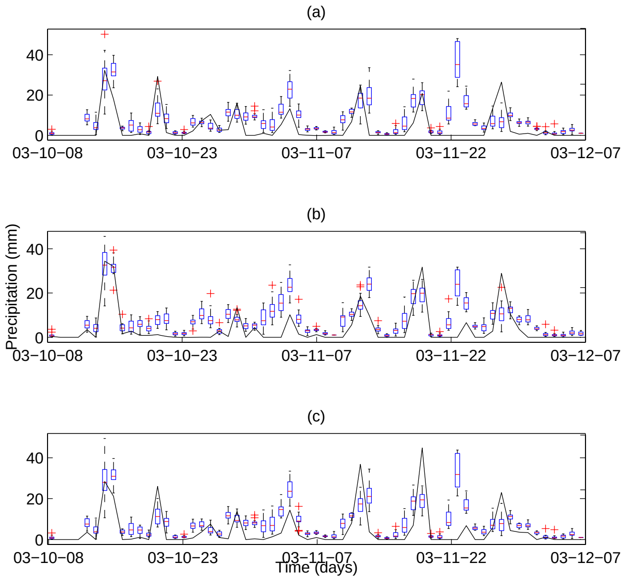

Fig. 3. Boxplots of 48 h ahead ensemble precipitation forecasts compared to 48 h ahead high resolution deterministic precipitation forecasts for **(a)** Station 7031360 (near outlet of the basin) **(b)** Station 7038885 (upper Baskatong, middle of the basin) and **(c)** Station 7038975 (Paugan).

2.3 Environment Canada’s meteorological forecasts

Two types of forecasts are compared for the events of fall 2003: the high resolution deterministic forecast (45 km) and the ensemble forecasts (200 km). The latter are known to be biased and underdispersed.

Forecasts were issued by Environment Canada, with the forecasting system that was in operation from January 1996 to July 2007. The ensemble forecasts are obtained by the outputs from two atmospheric models, SEF (regional) and GEM (global). Each model issues eight members, in addition to the control forecast.

Environment Canada’s forecasting system has since been improved, especially regarding the spatial resolution that is now 100 km.

Figure 2 shows scatter plots of the observed and forecasted precipitation for three measurement stations: 7031360, located in Chelsea sub-catchment, at the outlet of the Gatineau Watershed, station 7038885, in the middle of the basin, and station 7038975, in Paugan sub-catchment. Forecasted precipitation is often underestimated, especially at station 7038975.

As for ensemble forecasts, boxplots of the daily ensemble precipitation forecasts (17 members) are plotted in Fig. 3, against corresponding observations, for an excerpt of the fall 2003 period and for the same geographical locations (measurement stations) as Fig. 2. Forecasts for all stations announce a major precipitation event around 22 November, which appears to be a false alarm since it is not recorded

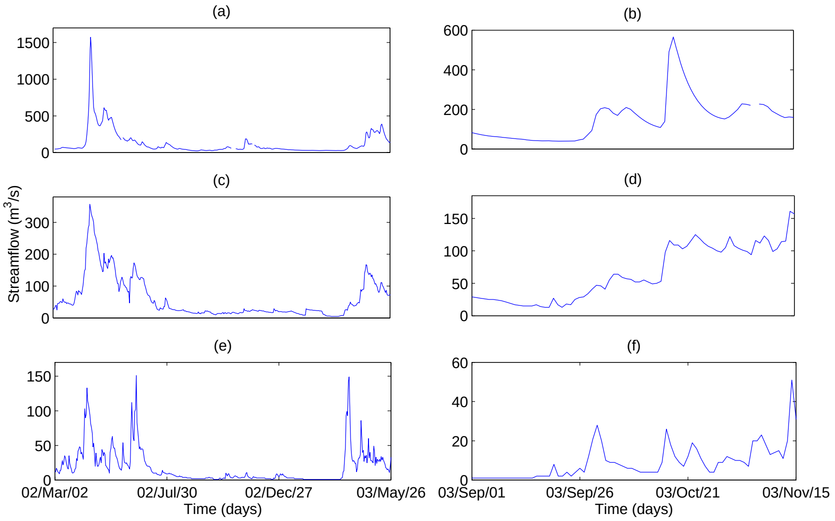

Fig. 4. Hydrograph for **(a)** Ceizur, calibration, **(b)** Ceizur, validation, **(c)** Maniwaki, calibration, **(d)** Maniwaki, validation, **(e)** Chelsea, calibration and **(f)** Chelsea, validation. Calibration data is from 2 March 2002 to 15 July 2003 and validation data is from 1 September 2003 to 17 December 2003.

at station 7031360 (Fig. 3a) and only a small amount of rainfall is recorded at the two other stations. Also, forecasts for station 7038885 over predict an event at the end of October. Generally speaking, ensemble precipitation forecasts seem to over predict low precipitation events. While the lower values of the ensembles are close to the observed precipitation, the ensemble mean is often too high (see for example Fig. 3b, between 15 and 25 October).

2.4 Hydrological model

HYDROTEL (Fortin et al., 1995, 2001) is a physics-based distributed hydrological model. It is used operationally as a short term forecasting tool by Hydro-Québec as well as by the Québec provincial government. The major strength of HYDROTEL is its capacity to directly use GIS inputs. For a physics-based model, it is also relatively unextensive regarding the amount of data needed, since it can be run using only precipitation and temperature observations. Like all hydrological models, it was designed from a deterministic point of view, so the generation of ensemble forecasts is not direct and requires a lot of additional manipulations.

Figure 4 shows the hydrographs for the calibration (2 March 2002 to 15 July 2003) and validation (1 September 2003 to 7 December 2003) periods. In both cases, these are daily streamflow observations and the upper plots are for Ceizur, the most upstream sub-catchment, the second row plots (Fig. 4c and d) correspond to Maniwaki, where the events of fall 2003 took place and the third row plots show data for Chelsea, the most downstream sub-catchment. For Chelsea and Ceizur, it can be noted that the maximal observed streamflow during the calibration period is higher that the one recorded during the validation period.

This study focuses on two to ten day ahead streamflow forecasts, which are used operationally at Hydro-Québec for short-term production management. The precipitation and temperature data needed by the model are available respectively every 12 and 6 h.

A basic updating method for the state of the model was used to ensure that the forecast process always starts from adequate hydrological state. The updating method consists in running HYDROTEL in simulation mode first, and correcting the inputs (precipitation and temperature) until a good fit between simulated and observed streamflow is achieved. Then, the states (soil moisture, snowpack height, surface runoff and streamflows) are saved and used later in the forecasting process. Considering the uncertainty related to rainfall observations and the low density of the gauging network, it is realistic to allow small corrections on those inputs. Regarding temperature corrections, they are mostly restricted to 1–3 degrees (subtracted or added) and used during spring to adjust snowmelt.

3 Experimental protocol

3.1 Weighting of the scenarios

Before comparing deterministic and ensemble forecasts, the question whether or not the forecasts issued by SEF and by GEM should have the same weight has to be raised. The analysis is based on the assumption that the observed value $x_{\mathrm{obs},t}$, given the ensemble forecasts $\boldsymbol{y}_{\mathrm{S},t} = y_{\mathrm{S},1,t}, y_{\mathrm{S},2,t}, \ldots, y_{\mathrm{S},8,t}$ and $\boldsymbol{y}_{\mathrm{G},t} = y_{\mathrm{G},1,t}, y_{\mathrm{G},2,t}, \ldots, y_{\mathrm{G},8,t}$ is drawn from a weighted mixture of two gamma distributions. $\boldsymbol{y}_{\mathrm{S},t}$ and $\boldsymbol{y}_{\mathrm{G},t}$ are respectively the ensemble members issued by SEF (S) and by

GEM (G) at time step t. First, a gamma distribution is fitted separately to $\boldsymbol{y}_{\mathrm{S},t}$ and to $\boldsymbol{y}_{\mathrm{G},t}$ using the method of moments. Second, the weighted mixture given by Eq. (1) is used to estimate w, the weight of model SEF

$$f(x_{\mathrm{obs},t}|\boldsymbol{y}_{\mathrm{S},t},\boldsymbol{y}_{\mathrm{G},t}) = w\,\mathrm{GAM}(x_{\mathrm{obs},t}|\hat{\alpha}_{\mathrm{S}},\hat{\beta}_{\mathrm{S}}) + (1-w)\mathrm{GAM}(x_{\mathrm{obs},t}|\hat{\alpha}_{\mathrm{G}},\hat{\beta}_{\mathrm{G}}) \qquad (1)$$

In this equation, $\hat{\alpha}_{\mathrm{S}}$, $\hat{\beta}_{\mathrm{S}}$, $\hat{\alpha}_{\mathrm{G}}$ and $\hat{\beta}_{\mathrm{G}}$ are the parameters of the gamma distribution (GAM) estimated with $\boldsymbol{y}_{\mathrm{S},t}$ and $\boldsymbol{y}_{\mathrm{G},t}$. The value of w is estimated over the entire forecasts-observation calibration archive between 3 March 2002 and 31 August 2003, using the maximum likelihood method. Consequently, there is only one estimated value of w, for each lead time, while the gamma distribution parameters change with time.

3.2 Evaluation of performance

The performances of the different types of forecasts were compared using three numerical criteria, in addition to the rank histogram for ensemble forecasts (Talagrand et al., 1997). In order to compare deterministic and ensemble forecasts, the Continuous Ranked Probability Score (CRPS) and the absolute error (AE) were used. As formally demonstrated by Gneiting and Raftery (2007), the mean CRPS is the probabilistic counterpart of the mean absolute error (MAE) for deterministic forecasts and the two scores are therefore directly comparable. Like for the MAE, the lower the CRPS, the better and both scores have a lower bound of zero.

The CRPS is a proper score, which implies that it can be separated into reliability and resolution components (Brocker, 2008). Its reliability and potential components can be evaluated following Hersbach (2000). The reliability component evaluates the extent to which probabilistic forecasts are reliable, meaning for example that the observed coverage probabilities of the confidence intervals correspond to the nominal confidence levels. The potential CRPS is the lowest possible CRPS that could be attained if the forecasts were made perfectly reliable (through post-processing, for example). The CRPS is a global score, meaning that its calculation involves the whole probability distribution. To complete our comparison of the performance of the ensemble forecasts, we also use the logarithmic score (Good, 1952), which is local. The calculation of a local score is based on the probability density function (pdf) evaluated at the observation $x_{\mathrm{obs},t}$ and hence does not involve the whole pdf. For the specific case of the logarithmic score, it is evaluated by computing the negative logarithm of the pdf evaluated at the observation for each forecast-observation group and then taking the average over all forecasts. Some authors suggested that "locality" is a desirable characteristic for a scoring rule (e.g. Bickel, 2007; Benedetti, 2010). There is no perfect value to use as a reference for the logarithmic score and no deterministic counterpart.

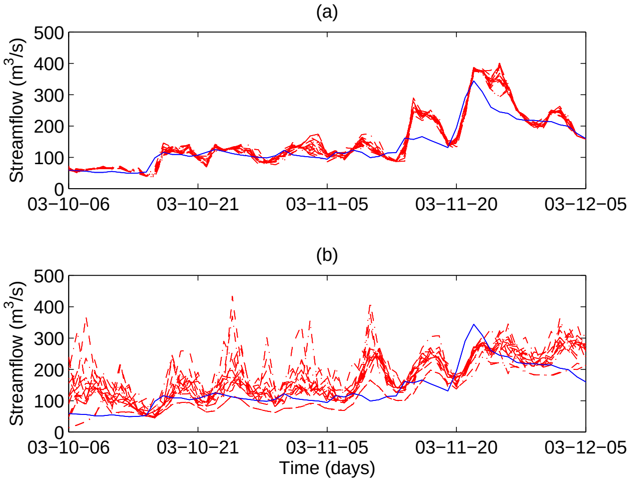

Fig. 5. Raw streamflow ensemble forecasts for Maniwaki sub-catchment, **(a)** 48 h ahead and **(b)** 240 h ahead.

To provide the reader with a more tangible view of the forecasts, Fig. 5 illustrates the raw ensemble forecasts (in red) and the observed streamflow (in black) for Maniwaki sub-catchment, both for 48 h (a) and 240 h (b) forecasts. As expected, the 48 h forecasts are less dispersed than the 240 h forecasts, for which some ensemble members seem to go astray sometimes and issue random streamflow peaks. This is clearly shown in Fig. 5b, where the lower members of the ensemble remain relatively close to the observed streamflow, while some members are much higher. This is observed in all other sub-catchments (not shown).

3.3 Post-processing of the ensemble forecasts

Ensemble forecasts post-processing follows the non-parametric kernel based method proposed by Roulston and Smith (2003). Generally speaking, a kernel based post-processing method consists in dressing each raw ensemble member with a probability function (the kernel) defined by a spread parameter (the bandwidth) and summing all the kernels to form a density mixture. This has the effect of increasing the spread of the ensemble, so such post-processing methods only suit under-dispersed ensembles. The extent to which the spread of the post-processed ensemble is greater than the spread of the raw ensemble depends on the bandwidth, so this parameter has to be calibrated. While kernel dressing can serve post-processing, it is a non parametric distribution fitting tool. Many textbooks can provide additional information about this technique, among which Wand and Jones (1995).

In the specific case of the best member method, the bandwidth is estimated through the errors between ensemble members and corresponding observations. First, for each

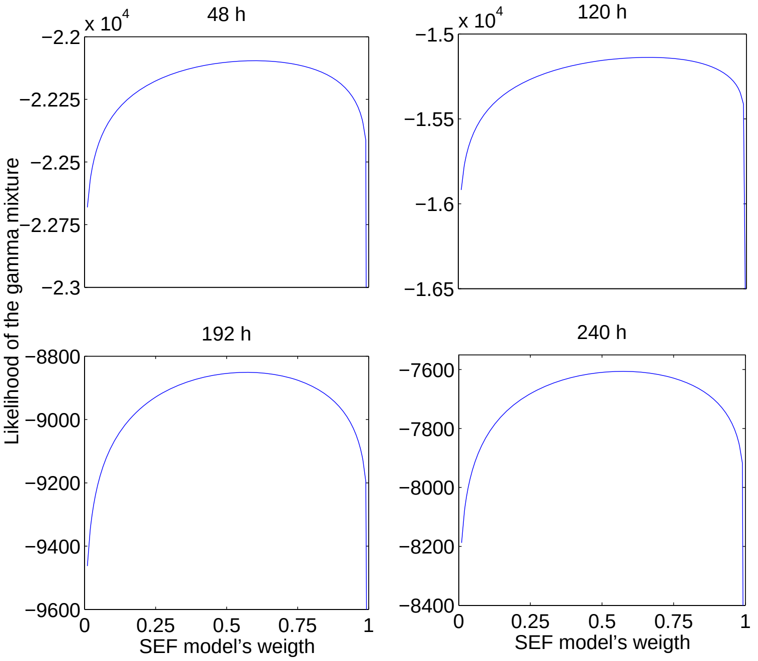

Fig. 6. Likelihood functions of the gamma mixture for the ensemble forecasts issued by models SEF and GEM, as a function of model SEF's weight, for Maniwaki sub-catchment and 48 h, 120 h, 192 h and 240 h lead time.

time step, the absolute difference between each ensemble member and the observation is computed. Note that this is done on a portion of the data (calibration data) which does not comprise the fall 2003 data that are used for validation and comparison. Here, the calibration data spans from 3 March to 31 August 2003. The calibration period is rather short, but as shown above by the hydrographs of Fig. 4, this may not be a problem since the highest observed streamflow values are included in this period.

Once all the absolute differences are obtained, the daily minimums are put in a vector, which constitute the "best member's errors". The errors made by the other ensemble members are greater and keeping them would increase the risk of obtaining a post-processed ensemble that is overdispersed.

Since there is a proportionality relation between the magnitude of the errors and the magnitude of the observations, it is not realistic to post-process forecasts of all magnitudes using a single kernel bandwidth. Ensemble forecasts for small streamflow values usually do not require as much correction than ensemble forecasts for extreme events. Consequently, the forecasts have to be categorized according to their magnitude, and corrected with an appropriate bandwidth.

In order to define the different categories, the streamflow observation database for each sub-catchment is used to fit a probability density function (gamma), from which the 25%, 50% and 75% percentiles are obtained. These percentiles become the limits of four categories, and a different bandwidth is estimated for each of those categories. To do so, the best member's errors are divided into those four categories depending on the magnitude of the corresponding observed flow. For instance, if the observed streamflow value is greater than the 50% percentile but less than the 75% percentile, the best member error of the corresponding ensemble forecast will be archived in a vector corresponding to the third category.

Subsequently, for each of these categories, the variance of the errors is calculated and serves as the bandwidth of the kernels in the smoothing method, which is applied to the remaining portion of the data. In this study, they are the fall 2003 data, from 1 September 2003 to 17 December 2003. The ensemble mean is used to divide the forecast into the same categories that were used to calibrate the bandwidth parameter, so the corresponding bandwidth is applied to obtain the post-processed ensemble. Table 2 presents the variance of the best member errors (the bandwidths) for all four streamflow categories for Maniwaki sub-catchment.

4 Results

The maximum likelihood estimates for the weight w of the two-component mixture of gamma distributions were obtained for each forecasting lead time. These are given in Table 3 for each sub-catchment. We notice that SEF's weight is higher than 0.5, but the corresponding likelihood functions, presented in Fig. 6 for Maniwaki, are quite flat, which indicates that the corresponding estimates display high

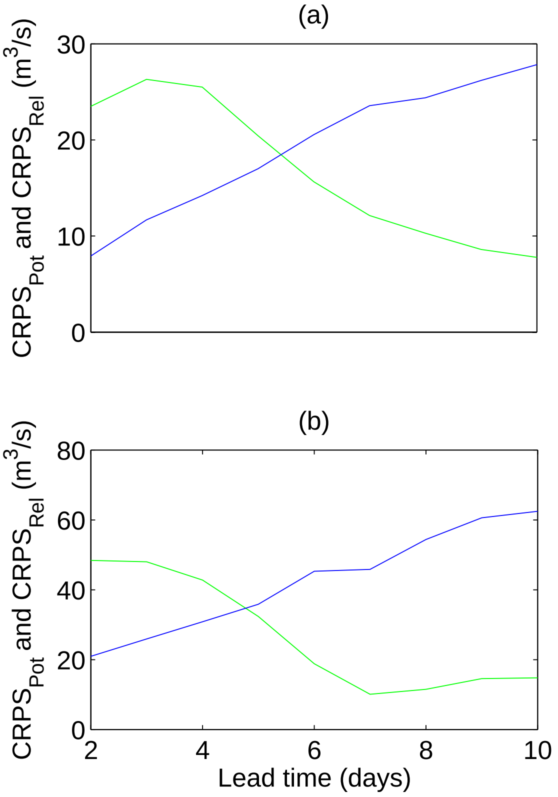

Fig. 7. Reliability (green) and Potential (blue) components of the mean CRPS before post-processing of the ensembles for **(a)** Paugan and **(b)** Baskatong.

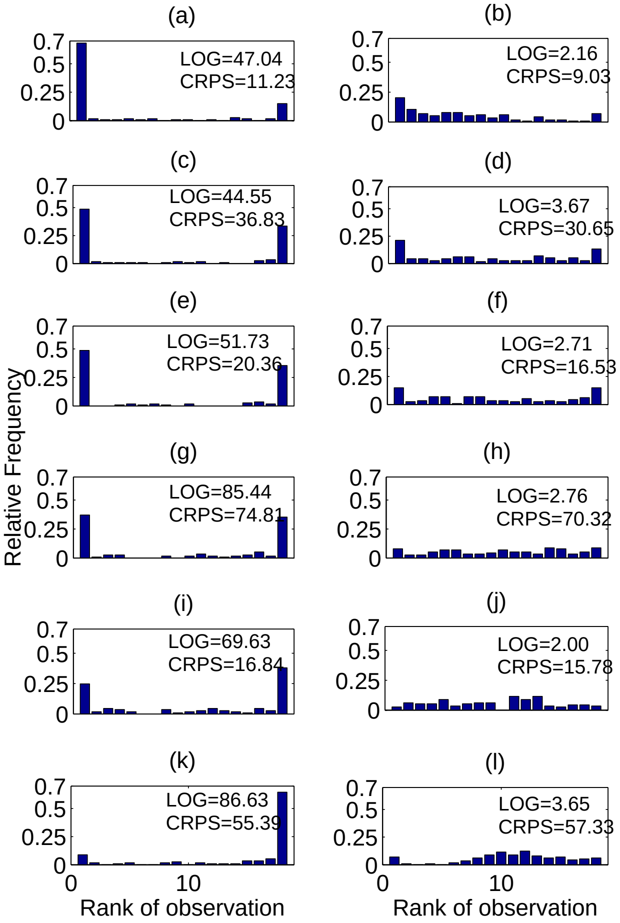

Fig. 8. Rank histograms for 2-days ahead ensemble hydrologic forecasts before and after post-processing, **(a)** Chelsea before post-processing **(b)** Chelsea after post-processing **(c)** Paugan before post-processing **(d)** Paugan after post-processing **(e)** Maniwaki before post-processing **(f)** Maniwaki after post-processing **(g)** Baskatong before post-processing **(h)** Baskatong after post-processing **(i)** Cabonga before post-processing **(j)** Cabonga after post-processing **(k)** Ceizur before post-processing **(l)** Ceizur after post-processing.

Table 2. Variance of the best member errors for Maniwaki subcatchment, divided according to the percentiles of the distribution of observed streamflow, for calibration period (2 March 2002 to 15 July 2003).

Horizon	Variance of best member errors			
	0 to 25%	25% to 50%	50% to 75%	75% to 100%
48-h	6.48	6.05	9.68	15.96
72-h	5.55	5.06	9.37	10.92
96-h	4.66	4.92	10.81	10.41
120-h	4.02	4.71	11.44	12.19
144-h	3.63	5.95	11.33	12.15
168-h	2.72	7.03	10.10	13.30
192-h	3.28	7.21	14.53	14.83
216-h	3.81	6.62	17.79	11.50
240-h	4.46	4.80	18.51	9.09

uncertainty and that it is not clear that one model outperforms the other. Consequently, SEF and GEM ensemble members are considered equiprobable hereafter.

Figure 7 shows the two components of the total CRPS as a function of the forecasting horizon, for unprocessed forecasts. The potential component of the CRPS (in blue) increases with the forecasting horizon. This means that the best attainable score gets higher (worst) as the lead time progress. For example, it is thus possible, through post-processing, to achieve a better score for two-day ahead forecasts than for ten-day ahead forecasts. As for the reliability component, it reveals that the forecasts become *more* reliable for longer lead times.

This behavior also illustrates the effect of a delay caused by the response time of the catchment (Velazquez et al., 2009). For forecasting horizons shorter than this response time, the hydrological state of the watershed prevails over the forecasts and therefore a gain could be achieved through

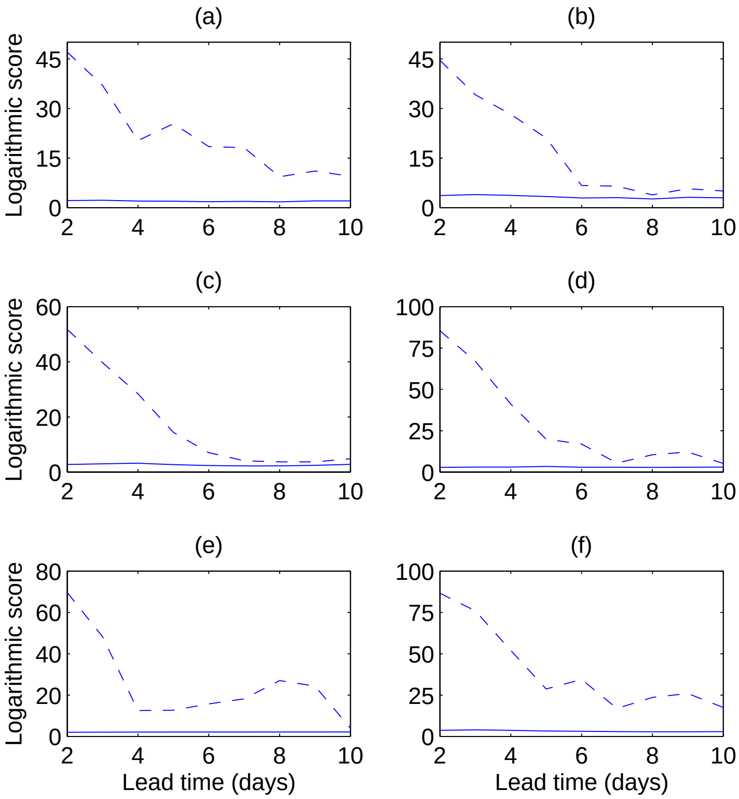

Fig. 9. Comparison between post-processed (blue solid line) and raw (dashed blue line) ensemble forecasts according to the logarithmic score **(a)** Chelsea **(b)** Paugan **(c)** Maniwaki **(d)** Baskatong **(e)** Cabonga **(f)** Ceizur.

Table 3. Maximum likelihood estimates for SEF model's weight in predictive distributions formed by a mixture of two gamma distributions.

Basin	Horizon (h)			
	48	120	192	240
Chelsea	0.70	0.68	0.62	0.61
Paugan	0.62	0.62	0.62	0.55
Maniwaki	0.63	0.67	0.61	0.59
Baskatong	0.63	0.58	0.59	0.60
Cabonga	0.66	0.64	0.68	0.65
Ceizur	0.61	0.65	0.70	0.67

post-processing, as shown by low potential CRPS. As the horizon lengthens, the forecasts become dominant over the observations. The potential CRPS becomes higher, meaning that even with appropriate post-processing methods, the lowest possible CRPS that could be attained is higher than for shorter lead times.

In order to assess the effect of the post-processing treatment, rank histograms were plotted for each lead-time. Figure 8 compares the raw ensemble (left-hand column) with the post-processed ensemble (right-hand column) for each sub-catchment and for two-day ahead forecasts. The logarithmic score and CRPS are also included in each plot.

All rank histograms are more uniformly distributed for post-processed forecasts and it is reflected in the scores, which are lower for the post-processed ensembles for all catchments except for Ceizur. However, for longer forecasting horizons and for some sub-catchments, the post-processing method does not significantly improve the forecasts (not shown). Figure 9 also shows that the logarithmic score systematically improves after post-processing. This is because the logarithmic score harshly penalizes forecasts that do not include the observed value as a possible outcome or attribute a very low probability to its occurrence. Consequently, when the dispersion increases and covers the observed value, the mean logarithmic score improves greatly.

Figure 10 compares the performance of the ensemble and deterministic forecasts. This is done using the CRPS and the MAE, for the post-processed ensembles (solid blue line) as well as for the unprocessed ensembles (dashed blue line) and high resolution deterministic forecasts (red line).The low resolution ensemble forecasts outperform the high resolution deterministic forecasts, except for the first forecasting horizons for some sub-catchments ((a) Chelsea, (b) Paugan and (c) Maniwaki). Moreover, this performance gap often increases with the lead time. As for the benefit of the

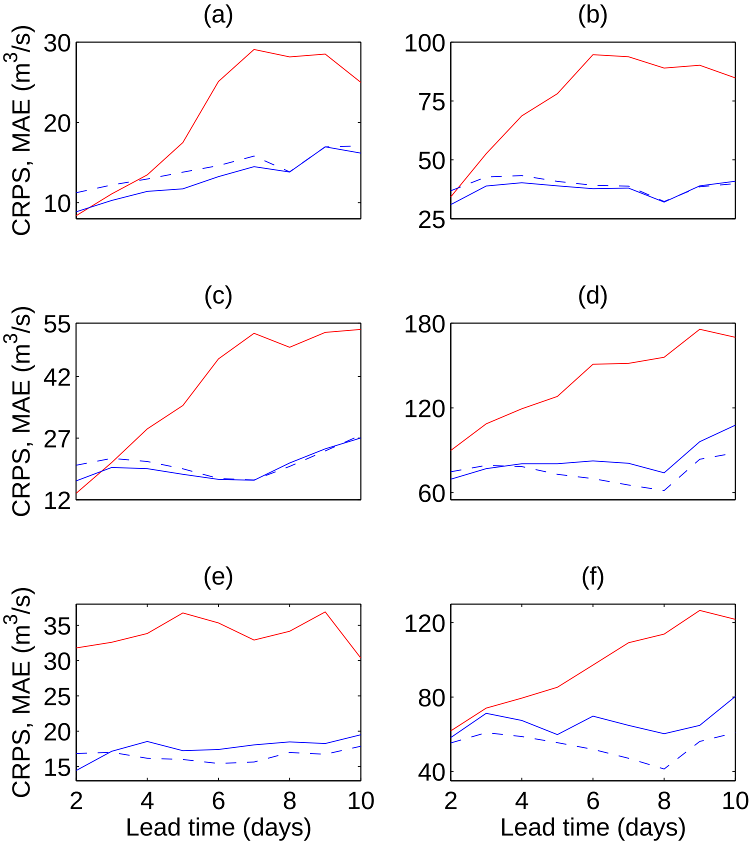

Fig. 10. Comparison between high resolution deterministic forecasts (red line) and low resolution post-processed (blue solid line) and raw (dashed blue line) ensemble forecasts in terms of CRPS and MAE **(a)** Chelsea **(b)** Paugan **(c)** Maniwaki **(d)** Baskatong **(e)** Cabonga **(f)** Ceizur.

post-processing method, it varies from one sub-catchment to another and also depends on the lead time. In Fig. 10a, it can be seen that the post-processed ensembles always have a lower (better) CRPS than the raw ensembles. Conversely, for Baskatong, Cabonga and Ceizur, which are the uppermost sub-catchments, the raw ensembles offer a greater performance than the processed ones. However, as illustrated in Fig. 9, according to the logarithmic score, the post-processed ensembles outperform the raw ensembles for all sub-catchment and forecasting horizons. This improvement is inversely proportional to the horizon and is greater for forecasting horizons shorter than four to six days. However, note that the CRPS and logarithmic score are not evaluated on the same scale and that the logarithmic scale emphasizes the difference of performance between raw and processed forecasts in Fig. 9 compared to Fig. 10.

5 Conclusions

In this case study, we show the benefit of choosing ensemble forecasts over deterministic forecasts, even when the spatial resolution of the ensemble forecasts is much lower than the deterministic forecasts for the Gatineau watershed in Canada.

A fairly simple post-processing method (Roulston and Smith, 2003) allows correcting the resolution and bias in the ensemble forecasts so that for at least some cases they outperform the high resolution determinist forecasts in terms of comparison with the observations for this basin. According to the CRPS, the use of a basic post-processing method for ensemble forecasts improves the results only for the first forecasting horizons. This is observed in most cases, except for two sub-catchments which are known to include large reservoirs. The influence of those reservoirs still has to be further investigated. According to the logarithmic score, the post-processed ensembles are systematically better than the raw ensembles, especially for the first four to six days. This difference of behavior between the two scores may be related to the fact that one is global while the other one is local, but this also requires further investigation.

In addition, future work on this basin should include the comparison of the old ensemble forecasting system with the new one. If no re-forecasts are made available, this could be done by comparing the fall 2003 flood event with a similar event that took place after July 2007. Also, it could be interesting to test and compare more sophisticated post-processing methods in order to compare their strengths and investigate the extent to which a particular post-processing method is suitable for different watersheds or for certain types of events. However, in this case study we have shown that, at least for the particular watershed at hand and the event considered, ensemble forecasts, even of poor quality and spatial resolution, can compete with more modern higher resolution deterministic products by means of minimal post-processing.

Edited by: R. M.-Helena
Reviewed by: two anonymous referees

References

Benedetti, R.: Scoring rules for forecast verification, Mon. Weather Rev., 138, 2033–211, 2010.

Bickel, J. E.: Some Comparisons among Quadratic, Spherical, and Logarithmic Scoring Rules., Decision Analysis, 4(7), 49–65, 2007.

Brocker, J.: Decomposition of Proper Scores., Technical Report, Max-Planck Institute für Physik komplexer Systeme, Näthnitzer Strasse 34, 01187 Dresden, Germany, 29 pp., 2008.

Côté, J., Gravel, S., Méthot, A., Patoine, A., Roch, M., and Staniforth, A.: The operational CMC-MRB Global Environmental Multiscale (GEM) model, Part I: Design considerations and formulation, Mon. Weather Rev., 126, 1373–1395, 1998.

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

Gneiting, T. and Raftery, A.: Strictly Proper Scoring Rules, Prediction, and Estimation, J. Am. Stat. Assoc., 102(477), 359–378, 2007.

Good, I. J.: Rational Decisions, J. Roy. Stat. Soc. B, 14, 107–114, 1952.

Gouvernement du Québec, Ministère des Ressources Naturelles, de la Faune et des Parcs: Portrait territorial de la région de l'Outaouais, 80 pp., 2005.

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

Fortin, J. P.,Turcotte, R., Massicotte, S., Moussa, R., and Fitzback, J.: A Distributed Watershed Model Compatible with Remote Sensing and GIS Data, Part I: Description of the model, J. Hydrol. Eng., 6(2), 91–99, 2001.

Hersbach, H.: Decomposition of the continuous ranked probability score for ensemble prediction systems, Weather Forecast., 15(5), 550–570, 2000.

Krzysztofowicz, R.: The case for probabilistic forecasting in hydrology, J. Hydrol., 249, 2–9, 2001.

Ritchie, H.: Application in the semi-Lagrangian method to a multilevel spectral primitive-equations model, Q. J. Roy. Meteor. Soc., 117, 91–106, 1991.

Ritchie, H. and Beaudoin, C.: Approximation and sensitivity experiments with a baroclinic semi-Lagrangian spectral model, Mon. Weather. Rev., 122, 2391–2399, 1994.

Roulston, M. S. and Smith, L. A.: Combining dynamical and statistical ensembles, Tellus, 55A, 16–30, 2003.

Talagrand, O., Vautard, R., and Strauss, B.: Evaluation of probabilistic prediction systems, ECMWF Workshop on Predictability, Shinfield Park, Reading, Berkshire, 1–25, 1997.

Velázquez, J. A., Petit, T., Lavoie, A., Boucher, M.-A., Turcotte, R., Fortin, V., and Anctil, F.: An evaluation of the Canadian global meteorological ensemble prediction system for short-term hydrological forecasting, Hydrol. Earth Syst. Sci., 13, 2221–2231, doi:10.5194/hess-13-2221-2009, 2009.

Wand, M.-P. and Jones, M.-C.: Kernel Smoothing, Chapman and Hall, London, 224 pp., 1995.