Journal of Hydrology 416–417 (2012) 133–144

Contents lists available at SciVerse ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Hydro-economic assessment of hydrological forecasting systems

M.-A. Boucher [a,*], D. Tremblay [b], L. Delorme [c], L. Perreault [c], F. Anctil [d]

[a] Applied Sciences Department, Université du Québec à Chicoutimi, Saguenay, Canada
[b] Hydro-Québec, Head Office, Montréal, Canada
[c] Hydro-Québec Research Institute, Varennes, Canada
[d] Chaire de recherche EDS en prévisions et actions hydrologiques, Université Laval, Pavillon Pouliot, Québec, Canada

ARTICLE INFO

Article history:
Received 18 May 2011
Received in revised form 7 October 2011
Accepted 20 November 2011
Available online 26 November 2011
This manuscript was handled by Geoff Syme, Editor-in-Chief, with the assistance of Ram Ranjan, Associate Editor

Keywords:
Ensemble forecasts
Economic value
Hydropower production
Optimization

SUMMARY

An increasing number of publications show that ensemble hydrological forecasts exhibit good performance when compared to observed streamflow. Many studies also conclude that ensemble forecasts lead to a better performance than deterministic ones. This investigation takes one step further by not only comparing ensemble and deterministic forecasts to observed values, but by employing the forecasts in a stochastic decision-making assistance tool for hydroelectricity production, during a flood event on the Gatineau River in Canada. This allows the comparison between different types of forecasts according to their value in terms of energy, spillage and storage in a reservoir. The motivation for this is to adopt the point of view of an end-user, here a hydroelectricity production society. We show that ensemble forecasts exhibit excellent performances when compared to observations and are also satisfying when involved in operation management for electricity production. Further improvement in terms of productivity can be reached through the use of a simple post-processing method.

1. Introduction

Since the beginning of the 1990s, most major meteorological agencies issue both deterministic and ensemble (probabilistic) forecasts (e.g. Buizza et al., 2005). In hydrology, Extended Streamflow Predictions (ESP, Day, 1985) have been operational for almost 30 years. ESP are created from an analysis of historical streamflow observations. With the more modern approach of hydrological ensemble prediction systems (see Cloke and Pappenberger, 2009), ensemble weather forecasts are forced into the hydrological model to produce a more accurate ensemble, often with reduced spread compared to ESP. The reason for the paradigm shift from deterministic to probabilistic forecasts the need for a better, more detailed evaluation and description of the uncertainty linked to these forecasts. However, the use of hydrological ensemble forecasts is still hampered by many obstacles (forecast communication, bias and under-dispersion issues, etc.). Nonetheless, some promising experiments have shown that their wider use could result in more efficient water management under various contexts (e.g. Richardson, 2000; Mylne, 2002; Roulin, 2007; Kim et al., 2007; McCollor and Stull, 2008).

Traditionally, performance assessment of forecasting systems consists in a comparison between the forecast and the observation over a certain period of time (e.g. Nash and Suthcliffe, 1970). Equivalent comparison tools were developed to address the challenge of comparing probabilistic values with scalar observations (e.g. Good, 1952; Hamill and Colucci, 1997; Talagrand et al., 1997; Roulston and Smith, 2002; Gneiting and Raftery, 2007; Casati et al., 2008). However, from an end-user point of view, it could be more interesting to evaluate performance according to operational variables such as electricity production, inundation reduction, or even irrigation capacity. Such variables reflect reservoir operations pending hydrological forecasts and may thus be much more relevant for the selection of an operational meteorological product. It has been commented that although ensemble forecasts are becoming more frequent among meteorological and hydrological agencies, many users still stick with deterministic forecasts for their operational needs (e.g. Shaake et al., 2007). This is in part due to the difficulty of communicating the uncertainty and interpreting it in terms of operations (e.g. Ramos et al., 2010).

Following this chain of thoughts, the experiment presented in this paper concerns an important flood event that occurred in fall 2003 on the Gatineau River, in south-eastern Canada and follows the work already begun on this in Boucher et al. (2011), where deterministic and ensemble forecasts were confronted using more conventional performance assessment tools, involving the comparison between forecasts and observations. The Gatineau basin

* Corresponding author. Address: Applied Sciences Department, Université du Québec à Chicoutimi, Saguenay, Canada G7H 2B1. Tel.: +1 418 545 5011x5057; fax: +1 418 545 5012.
E-mail address: marie-amelie_boucher@uqac.ca (M.-A. Boucher).

0022-1694/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhydrol.2011.11.042

comprises two upstream reservoirs and a series of three in-stream plants that belong to Hydro-Québec, the major hydropower producer in eastern Canada. HYDROTEL, a physics-based distributed hydrological model (Fortin et al., 1995) is fed with meteorological forecasts issued by Environment Canada to issue both deterministic and ensemble hydrological forecasts over a 10-day-ahead horizon and a 24-h time increment. Hydrological forecasts are then passed into a stochastic decision-making assistance tool (Krau, 2005; Bibeau et al., 2006) designed for optimizing electricity production while avoiding spillage and inundation. Performance of the hydrological forecasts can then be assessed from two complementary angles. Forecasts and observations are first confronted; performance is next quantified in terms of electricity production and spillage reduction. Deterministic hydrological forecasts are compared to raw ensemble forecasts as well as post-processed and weighted ones. Real operations recorded in fall 2003 are available, and compared to simulated operations that arose from different types of forecast.

The first section of the paper describes the context of application, including a portrait of the basin and a description of the selected flood event, as well as a brief introduction to SOHO, a stochastic decision-making assistance tool. Section 3 presents the hydrological model HYDROTEL and Section 4 details the different types of forecasts that are compared in the experiment. Results are presented separately in terms of statistical performance and in terms of operational performance in Section 5. Conclusions are gathered in Section 6.

2. Context of application: hydropower production on the Gatineau watershed

2.1. Portrait of the basin

The Gatineau River is 443 km long, beginning from the Pain de Sucre Lake and ending in the Des Outaouais River. The basin is located in the south of the province of Quebec, Canada and covers 23,785 km^2. It is a highly reactive catchment regarding meteorological variations, with a response time of 1–3 days depending on the sub-watersheds. Moreover, the observed streamflows cover a wide range of possibilities, ranging from as low as 5 m^3/s to a maximum of 357 m^3/s for Maniwaki. The catchment comprises hydropower production structures located in inhabited areas, highlighting the importance of optimal management. In consideration of those characteristics, this basin is considered as a test bed by Hydro-Québec, the major hydropower producer in Canada.

The basin can be divided into six sub-catchments, as illustrated by Fig. 1. Meteorological and hydrological stations are identified respectively by black dots and grey stars on this figure. The small turbine-like symbols on the correspond to the location of the hydropower plants.

The head basin is Ceizur and covers 6840 km^2. Cabonga, with its 2662 km^2, is the the most challenging for simulation and forecasting, since a large portion of its area is occupied by lakes and reservoirs. This situation implies that the gauging curve used to convert water level to streamflow can be affected by wind-induced level variations on the surface and generates measurement errors (roughly estimated to an average of ±225 m^3/s). Sub-catchments Baskatong, Maniwaki, Paugan and Chelsea account respectively for 6200 km^2, 4145 km^2, 2790 km^2 and 1148 km^2.

Fig. 2 shows the observed hydrographs from 2002 March 1st to 2003 December 31st. Meteorological and hydrological data up to August 30th 2003 were used for calibration while 2003 September 1st to December 31st is the validation period. As can be seen on the hydrographs, the largest streamflow observations for Ceizur, Cabonga and Maniwaki belong to the calibration portion of the dataset. For other sub-catchments, the calibration and validation period contain similar streamflow values. Consequently, even if the calibration period is short, it is representative of a variety of events and includes, for most sub-watersheds, a large fall flood with streamflow almost as high as the one we want to model.

Fall 2003 was characterized by a series of consecutive rainfall events, some of which more intense than what had been forecasted. This situation turned out difficult to manage because Baskatong reservoir had been maintained too high and could not contain incoming precipitation. This caused flooding in the municipality of Gracefield south of Maniwaki. From this point on, this flooding event has been used for the development and testing of new forecasting techniques and production planning strategies.

Table 1 enumerates the main hydropower production installations on the Gatineau basin that existed in 2003. A new plant (Mercier) was added in 2009.

2.2. SOHO: a stochastic decision-making assistance tool

The flood of fall 2003 encouraged Hydro-Québec developing an experimental stochastic decision-making tool named SOHO (Krau, 2005). SOHO s principal advantage is that it optimizes the use of water considering the level of all reservoirs for the entire period, in opposition to a day-to-day type water management that does not account for the long-term benefits.

SOHO is a linear stochastic programming model (Birge and Louveaux, 1997), which is a direct extension of deterministic linear programming. The main objective of SOHO is to maximize the hydraulic generation of a valley which consists of run-of-river power plants, over a given horizon of several days. The decision variables are daily reservoir volumes, river flows and turbined outputs from plants. SOHO takes into account water spillages (which are minimized), water delays, water head variations at plants (which are approximated via successive linear programming method), hydraulic bounds and reservoir target levels at the end of the planning horizon. Most constraints are modelled as penalty cost functions which are included in the objective function.

The main difference of SOHO with standard hydraulic generation planning models lies in the representation of hydrological forecasts in the form of a probabilistic scenario tree. Each branch corresponds to an inflow scenario, the first one being related to the "known horizon". Scenario trees typically consists of two stages, as illustrated by Fig. 3, but the use of additional stages is allowed. Each branch of the tree in Fig. 3 can be weighted using the conditional probability associated with the related inflow scenario. At each branch is associated a set of decision variables and all related constraints. The resulting mathematical formulation is equivalent to a large and unique deterministic model, where weights appeared only in the objective function; the function can then be viewed as the "expected cost function" of the whole generation plan. When using ensemble forecasts (Fig. 3a), all weights can be considered equal if the members of the ensemble are equiprobable or the weights can be different in order to reflect a probability of occurrence that varies from one scenario to another. SOHO can also be used with deterministic hydrological forecasts. In such case, the scenario trees only have one branch, which has a weight of one (Fig. 3b).

The different forecast options as well as the four weighting scenarios are passed through the available stochastic management system (SOHO). It is then possible to compare deterministic and ensemble forecasts in term of energy production. SOHO is used as the core of a simulation tool, which reproduces the generation planning of the Gatineau valley over several months

Fig. 4 provides a reference of the filling of Baskatong reservoir, the spilling and the turbinate flow associated to real operations

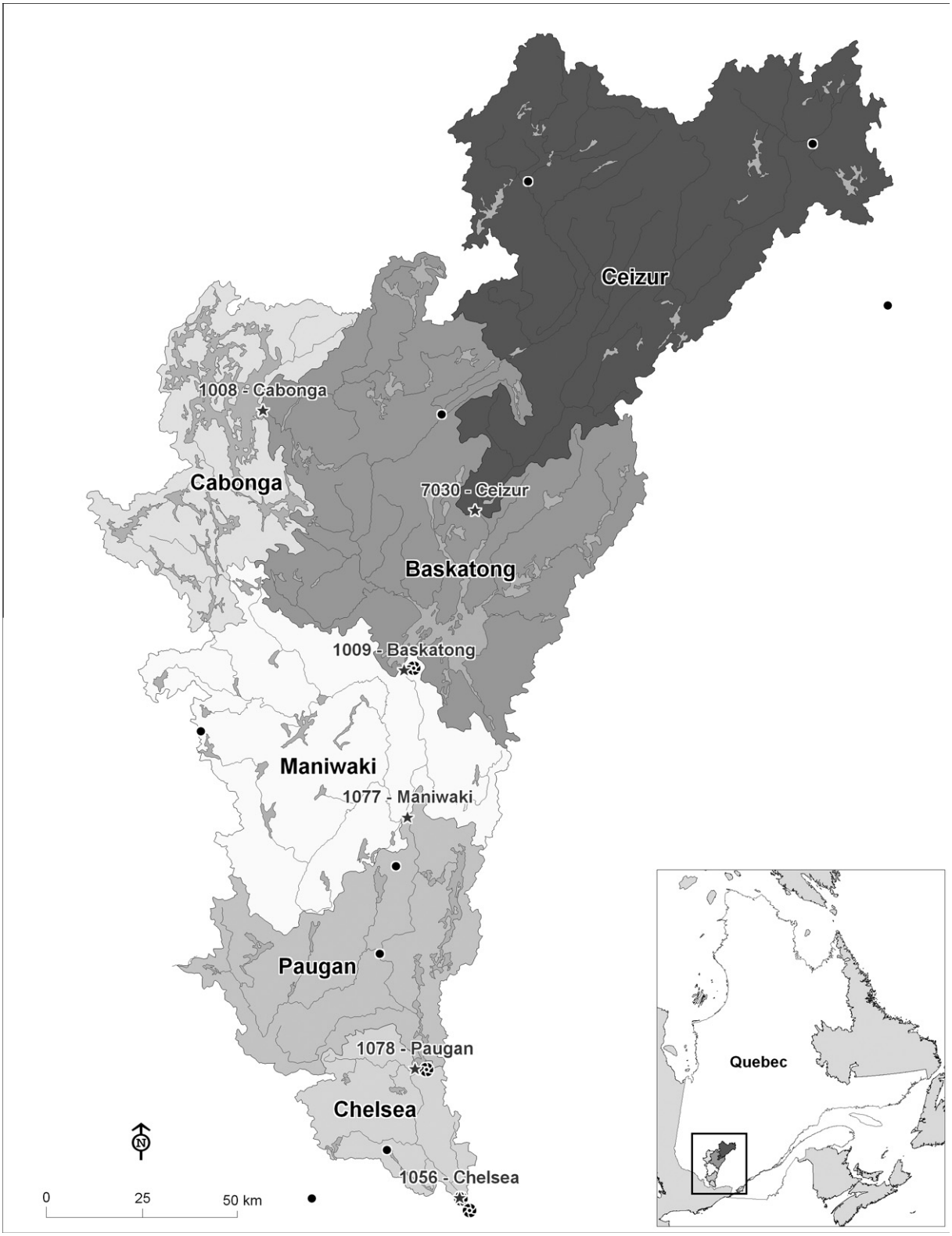

Fig. 1. Subdivision of the Gatineau catchment into six sub-catchments.

made during fall 2003. The reader will note that the maximal turbinate flow value for the entire system is 1948.31 m^3/s, which corresponds to the total capacity of the three plants, Paugan, Chelsea and Rapides-Farmers. This value was almost reached around 2003 October 10th when an important spilling occurred.

3. 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

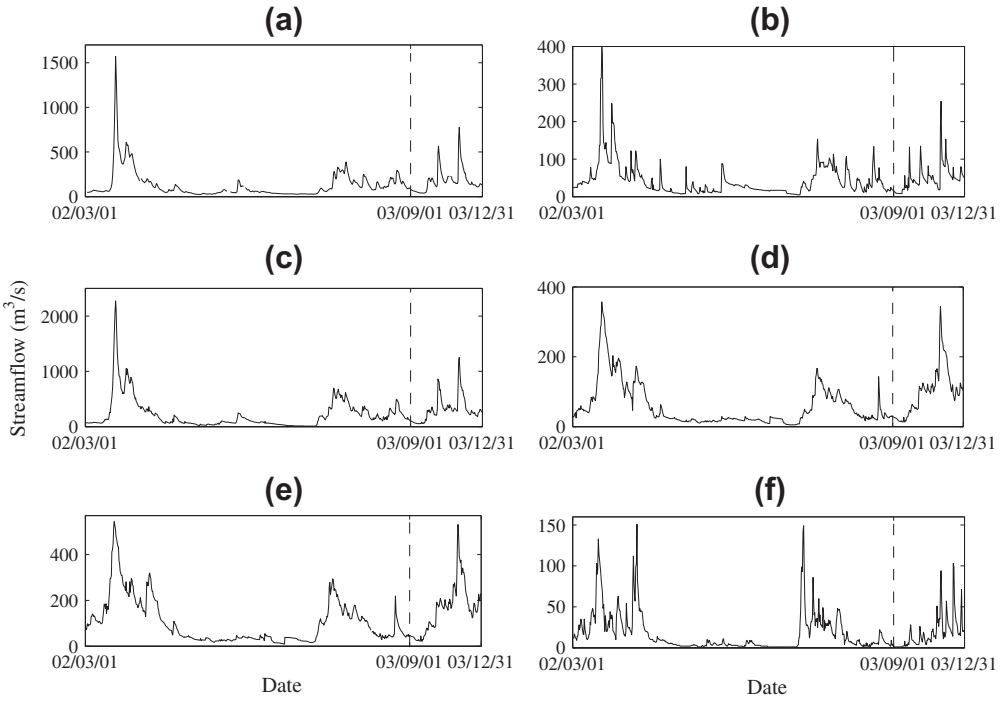

Fig. 2. Daily observed streamflows at stations (a) 7030-Ceizur, (b) 1008-Cabonga, (c) 1009-Baskatong, (d) 1077-Maniwaki, (e) 1078-Paugan, (f) 1056-Chelsea from 2002 March 1st to 2003 December 31st. The vertical dashed line on the hydrographs divides the data into training (before August 31st 2003) and validation datasets (from August 31st 2003).

Table 1
Characteristics of the main operating hydropower installations on the Gatineau watershed in fall 2003.

Name	Type	Area (km^2)	No. of groups	Installed power (MW)	Drop height (m)
Cabonga	Controled reservoir	434			
Baskatong	Uncontroled reservoir	413			
Paugan	Instream plant	30	8	202	40.5
Chelsea	Instream plant	24	5	153	28.4
Rapides-Farmers	Instream plant	1	5	98	20.1

provincial government. This model was originally proposed to allow the direct use of remote sensing data. Gridded inputs (precipitation, snow cover, etc.) can be directly incorporated into the model.

The physical features (elevation map, hydrographic network, vegetation, soil types, land use, etc.) of the catchment must be described separately in another model called PHYSITEL. It relies on this information to divide the catchment into small units called RHHU (Relatively Homogenous Hydrological Units). For each one of those territorial units, topography, land use and soil type are assumed homogenous. The physical description of the watershed, divided into those RHHU is then passed on to HYDROTEL, which used it to compute the vertical and horizontal flows.

One interesting feature of the model is that it offers a selection of sub-models of different levels of complexity and data requirement for each hydrological process. Consequently, the total number of parameters to be calibrated depend on the choice of sub-models. For instance, to model potential evapotranspiration, the user either exploits direct observation when available, or chose between five sub-models, among which an empirical simplification (developed at Hydro-Qubec and used in this study) if only temperature data are available.

For some processes, only one sub-model is available. This is the case with vertical flow, which is modelled through a routine called BV3C. It divides the soil layer into three layers to represent surface runoff and fast and slow sub-surface runoff as well as infiltration and recharge of the aquifer. BV3C comprises four parameters to be calibrated and are associated to snow cover, soil humidity, water transiting through the three soil layers and water flowing to the drainage network.

In this study, HYDROTEL is driven by various types of meteorological forecasts and the resulting streamflows in turn become inputs for SOHO. This allows the comparison of those different forecasts in terms of corresponding potential production gains. It can be seen as a form of performance assessment that could be particularly appealing to dam managers and decision-makers.

4. Testing different types of forecasts

4.1. High resolution deterministic forecasts

The deterministic forecasts, obtained through the Canadian Meteorological Centre Global Environmental Multi-scale (GEM) model (Côté et al., 1998), have a spatial resolution of about 45 km. Precipitation forecasts are available twice a day, at midnight and at noon, while temperature forecasts are available at midnight, 6 AM, noon and 6 PM. All forecasts have a 10-day horizon. As HYDROTEL requires maximum and minimum temperature forecasts for each time step, we opted for a daily (24-h) time step for the hydrological modelling and used the minimum and maximum of the three daily forecasts of temperature as inputs, as well as the non-cumulative midnight (local time) precipitation forecast.

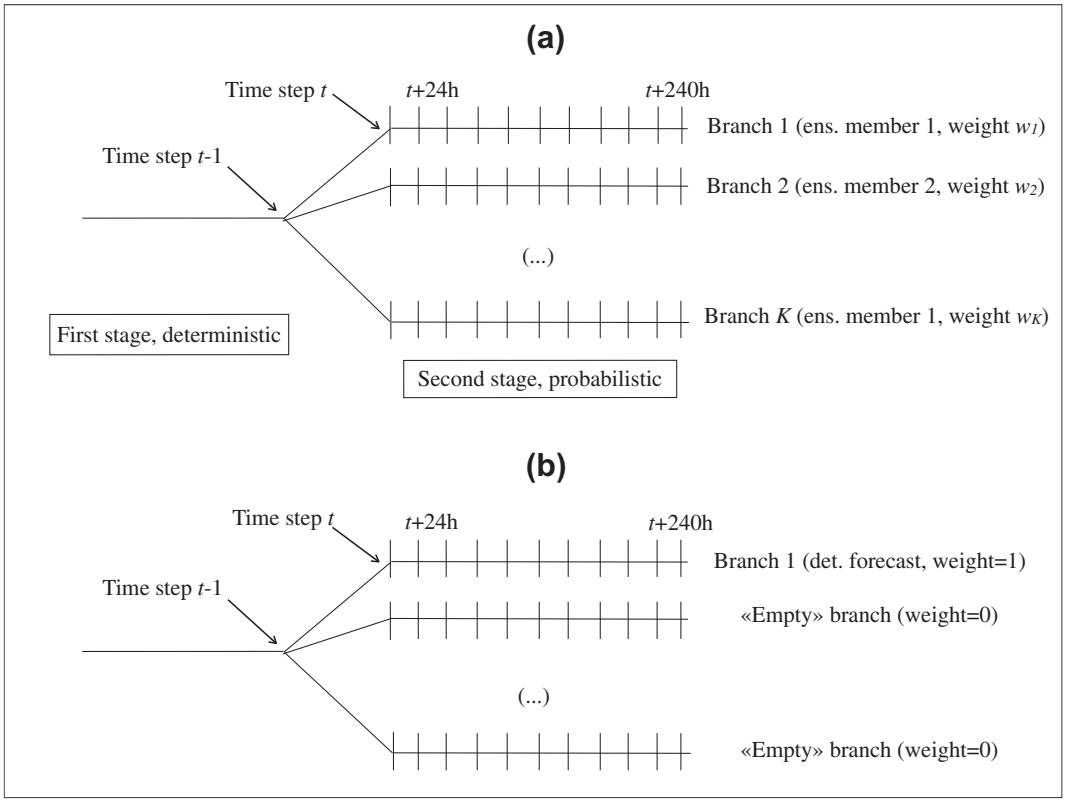

Fig. 3. Two stages scenario trees for (a) ensemble forecasts and (b) deterministic forecasts.

Streamflow forecasts are issued with a 24-h time step, always at midnight (local time).

HYDROTEL, like most hydrological models, is designed for deterministic applications: producing a deterministic hydrological forecast from deterministic temperature and precipitation forecasts. The scenario tree fed to SOHO, as described in Fig. 3a, is then modified so that the second stage of the tree comprises only one branch with a weight of one (Fig. 3b).

4.2. Low resolution ensemble forecasts

Forecasts were issued by Environment Canada from the system in operation from January 1996 to July 2007. Ensembles originated from two atmospheric models, SEF (regional) and GEM (global), both with spatial resolution of 1.2 °. Each model issued eight members, in addition to the control forecast, over a 10-day horizon and a time step of 6 h for precipitation and 12 h for temperature. However, this system is known to lead to biased and under-dispersed ensembles (Evora, 2005). For instance, the temperature bias is evaluated at +5 °C.

Environment Canada's forecasting system has since been improved. It is now solely based on the GEM model, which physics has been improved, especially regarding the representation of condensation and precipitation as well as the land-surface data assimilation scheme (Bélair et al., 2008). Its spatial resolution is now 0.6° and it comprises 21 members. No re-forecasts based on this new system were available to this study.

Even if HYDROTEL was not originally designed to use and issue ensemble forecasts, it exploits a file system that can be modified through an external script to successively change repository names and corresponding ensemble members, then to regroup all results and reconstitute hydrological ensemble forecasts in the appropriate format for SOHO. HYDROTEL allows the use of both gridded data and point measures, an interesting feature here since the precipitation and temperature observations are available at some points on the watershed (the meteorological stations) whereas the forecasts are distributed over a grid covering the basin.

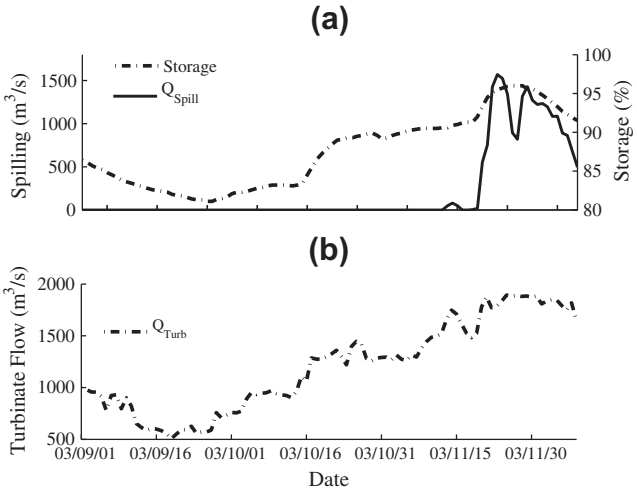

Fig. 4. Measured management variables for the Gatineau watershed hydropower system in fall 2003: (a) Storage in Baskatong reservoir and spilled flow. (b) Turbinate flow.

4.3. Post-processed ensemble forecasts

Hydrological ensemble forecasts based on meteorological ensemble ones do not account for all possible sources of uncertainty. For example, when only one hydrological model is used, the uncertainty linked to the choice of a particular hydrological

representation is not depicted. This generally leads to under-dispersed predictive distributions, a frequently encountered situation (e.g. Hamill and Colucci, 1997; Toth et al., 2001; Buizza et al., 2005). However, this situation can be overcome through post-processing.

Two kernel-smoothing based post-processing methods have been used here and compared: the best member method by Roulston and Smith (2003) and a variant proposed by Fortin et al. (2006). Kernel dressing is mainly a non parametric distribution fitting tool. However, it can serve post-processing for under-dispersed ensemble forecasts because it increases the spread of the ensemble through 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.

The best member method consists in exploiting the errors between ensemble members and corresponding observations to estimate the bandwidth. First, for each time step t, the absolute difference between each of the $k = 1,\ldots,K$ ensemble member $y_{t,k}$ and the observation x_t is computed. Note that this is done on a portion of the data that spans from 2002 March 3rd to 2003 August 31st. As noted before, this calibration period is rather short, but this may not be a problem since it comprises a wide variety of streamflow values, ensuring that the post-processing methods parameters can be estimated both for very low and very high streamflow, as well as for more standard flows.

Once all the absolute differences are obtained, the forecast with the minimum error y_t^* is computed for each time step as stated in Eq. (1), and represents the best member for this time step, while ξ_t is the corresponding error.

$$\xi_t = y_t^* - x_t = min|y_{t,k} - x_t| \tag{1}$$

The errors ξ_t are stored in a vector, which constitute the "best member's errors" and $\bar{\xi}$ is the average of all the ξ_t. Since there is a proportionality relation between the magnitude of the errors and the magnitude of the observations, the forecasts have to be categorized according to their magnitude, and corrected with an appropriate bandwidth. Ensemble forecasts for small streamflow values usually do not require as much correction as ensemble forecasts for large events.

To define 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. They are used to define limits of four categories, and then 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. Finally, a bandwidth is estimated for each of those categories by computing the variance of the best member errors. This is given by

$$\sigma^2 = \frac{1}{N-1}\sum_{t=1}^{N}(\xi_t - \bar{\xi})^2 \tag{2}$$

where N is the number of forecast–observation groups and σ^2 is the bandwidth. This bandwidth computed on the calibration portion of the data can then be applied to the remaining portion of the data (validation), which corresponds to 2003 September 1st to December 17th. 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. The general expression for the post-processed ensemble obtained through kernel dressing is given by

$$\hat{f}(z_t) = \frac{1}{K}\sum_{k=1}^{K}\mathcal{K}(y_t - y_{t,k}, \sigma) \tag{3}$$

where f is the density function from which the post-processed ensemble members, z_t, are drawn and $\mathcal{K}$ is the kernel, here chosen to be a normal distribution. To verify whether or not the forecasts in the testing portion of the dataset are correctly categorized, classification of the forecasts in the testing dataset according to the ensemble mean was compared to the classification of the corresponding observations. It revealed that the classification is accurate (the classification of the forecasts correspond to the classification of the corresponding observations) most of the time, especially for high streamflow (fourth category, for streamflows superior to the 75% percentile of the distribution of observations. Fig. 5 summarizes those results.

Fortin et al. (2006) suggested that it could be beneficial to use kernels of different bandwidths depending on the order statistic for each ensemble member. The initial procedure is identical to Roulston and Smith (2003). Instead of defining categories based on percentiles of the distribution of observations, best member errors ξ_t are categorized according to the rank of the corresponding member in the ensemble. For instance, if there are 17 ensemble members, there will be 17 categories. If, at a certain time step t the best member occupies the 15th rank of ordered ensemble members, the corresponding error ξ_t will belong to the 15th category. This is described by Eq. (4),

$$\xi_{t,(k)} = \left\{ |y_t^* - x_t| y_t^* = y_{t,(k)}, \quad t = 1,2,\ldots,N \right\} \tag{4}$$

where (k) represents the order statistic of the K ensemble members. Again, the variance of the errors $\xi_{t,(k)}$ is computed for each category (order statistic) and applied as the bandwidth for kernel dressing on the validation data. This time, the members of each forecast for the validation data must first be ordered and the corresponding bandwidth is used to fit a different kernel around each ensemble member. The post-processed ensemble forecast $\hat{f}(z_t)$ is given by Eq. (5)

$$\hat{f}(z_t) = \frac{1}{K}\sum_{k=1}^{K}\mathcal{P}_k\mathcal{K}(y_t - y_{t,(k)}, \sigma_{(k)}) \tag{5}$$

where $\mathcal{P}_k$ is the probability that the kth member be the best member and $\sigma_{(k)}$ is the standard deviation of the best members errors, when the best member occupies rank k. This is given by

$$\sigma_{\mathcal{K}_{FF}} = \sqrt{\frac{1}{N'-1}\sum_{t=1}^{N'}(\xi_{y^*_{t,(k)}} - \bar{\xi}_{y^*_{(k)}})^2} \tag{6}$$

where $\xi_{y^*_{t,(k)}}$ correspond to the best member's errors when it occupies rank k, $\bar{\xi}_{y^*_{(k)}}$ is the mean of the errors and N' is the number of forecasts (among the N forecasts-observation groups in the calibration database) for which the best members occupies rank k.

Finally, $\mathcal{P}_k = \mathcal{P}\left[y_t^* = y_{t,(k)}\right]$ is the probability that the best member be $y_{t,(k)}$. In the original method proposed by Fortin et al. (2006), this probability is computed parametrically by fitting a beta density function to the observed frequencies of the ranks occupied by the best member. To keep things as simple and operationally appealing as possible, here the probabilities $\mathcal{P}_k$ are estimated empirically using observed relative frequencies directly.

The main drawback of Fortin et al. (2006)'s method is that it requires long data records, a rare luxury in hydrology. In addition, hydrological forecasting mostly relies on meteorological forecasts issued by atmospheric models, which benefit from regular improvements. Each time the atmospheric model is updated, it creates a break point in the data archive, in the sense that the forecasts issued by the previous version of the model cannot be merged with the forecasts issued by the new version. The consequence of this is to shorten the available dataset so methods which

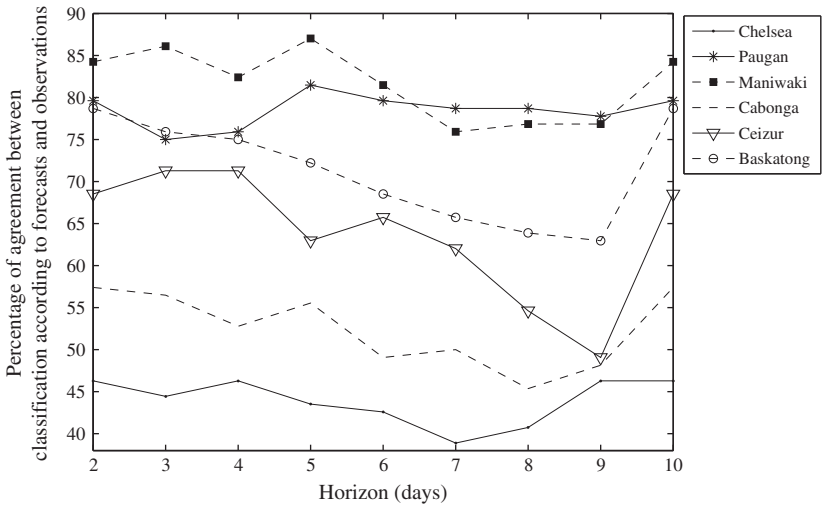

Fig. 5. Comparison between classification according to the ensemble forecast mean and the observed streamflow for the testing portion of the data, as a function of the forecasting horizon and for each sub-catchment. Chelsea: solid line with black dots, Paugan: solid line with stars, Maniwaki: dashed line with black squares, Cabonga: dashed line without markers, Ceizur: solid line with white triangles and Baskatong: dashed line with white circles.

rely on long data chronicles may be less convenient in practice for that reason.

4.4. Weighting of the scenarios

Before constructing the various scenario trees that will be provided to SOHO, the weight associated to each scenario must be estimated and this requires an hypothesis regarding the distribution of the forecast members issued by SEF and GEM models. If both models are equiprobable representations of the reality, then all ensemble members can be attributed the same weight. However, there is a possibility that either SEF or GEM is a more accurate representation of reality for this basin, or at least in some particular circumstances. For instance, it is possible that one of the two models performs better than the other for a particular season. If it is the case, then all members of the ensemble forecasts cannot be considered equiprobable and have to be weighted differently according to the model that issued them. Considering this, it must be verified, before feeding the scenarios to SOHO, whether or not those scenarios have to be weighted differently.

The analysis is based on the assumption that the observed streamflow value x_t, given the ensemble forecasts $\vec{y}_{S,t} = y_{S,1,t}, y_{S,2,t}, \ldots, y_{S,8,t}$ and $\vec{y}_{G,t} = y_{G,1,t}, y_{G,2,t}, \ldots, y_{G,8,t}$ is drawn from a mixture of two gamma distributions given by

$$f(x_t|\vec{y}_{S,t}, \vec{y}_{G,t}) = wGAM(x_t|\hat{\alpha}_S, \hat{\beta}_S) + (1-w)GAM(x_t|\hat{\alpha}_G, \hat{\beta}_G) \quad (7)$$

$\vec{y}_{S,t}$ and $\vec{y}_{G,t}$ are respectively the ensemble members issued by SEF (S) and by GEM (G) at time step t. The assumption of gamma distributions comes naturally since it is a very flexible density function, restricted to positive values and that allows a positive asymmetry. This assumption was verified by visual inspection of the histograms of the members from each model, for hydrological forecasts of various magnitudes, separately for each sub-catchment. First, a gamma distribution is fitted separately to $\vec{y}_{S,t}$ and to $\vec{y}_{G,t}$ using the method of moments to obtain estimates for the parameters $\hat{\alpha}_S$, $\hat{\beta}_S$, $\hat{\alpha}_G$ and $\hat{\beta}_G$. Second, the weights of the 2-component gamma mixture are estimated using the maximum likelihood.

Post-processed hydrological ensemble forecasts are given to SOHO in the same fashion as raw forecasts and then their respective performance is assessed in terms of management variables such as turbinate flow, spilling, flooding, and amount of produced energy.

4.5. Forecast assessment

This study focuses on the performance assessment of hydrological ensemble forecasts from an operational point of view, through the benefit they can bring to hydroelectricity production. Nonetheless, the usual comparison between forecasts and observations is also carried.

The Continuous Ranked Probability Score (CRPS) is the probabilistic equivalent of the mean absolute error (MAE). Further information regarding the computation of the CRPS can be found in numerous publications, among which Gneiting and Raftery (2007). The mathematical proof of the equivalence between CRPS and MAE can be found in Baringhaus and Franz (2004) and Székely and Rizzo (2005). Since one aims at minimizing the difference between forecasts and observations, a perfect forecast would score a MAE or CRPS of zero.

One very interesting property of the CRPS is that it can be decomposed into reliability and potential components (Hersbach, 2000). The reliability component measures the ability of the forecasting system to be coherent with the frequency of the observations. For instance, one can use ensemble forecasts to compute daily confidence intervals for the streamflow, say 90% confidence interval. If the forecasting system is reliable, then the 90% confidence interval should comprise the observed streamflow 9 times out of 10 on average. This is the most important quality for a forecasting system because unreliable forecasts give an estimate of the uncertainty that is not representative of the truth, which can in turn lead to non optimal management decisions. As for the potential component of the CRPS, it is the lowest (best) score that could be achieved if the system was made perfectly reliable. It includes two sub-components, the uncertainty and the resolution. The uncertainty component is linked to the available database and the intrinsic characteristics of the river and climate system. It cannot be improved by post-processing the forecasts. The resolution refers to the ability of the forecasting system to discriminate between two streamflow events that are similar. One could see it as a measure of the level of precision of the forecasts. Since the uncertainty component remains fixed for a given database, the potential CRPS mostly reflects the resolution.

To further investigate the various trade-offs made between production, storage, and spilling, the following production ratio is proposed:

$$R_{Prod} = \frac{Production}{Production + Spilling + Storage} \quad (8)$$

The ratio indicates the portion of the water available each day that is dedicated to energy production. It equals 1 when both storage and spilling are nil (what gets in, gets out), it is larger than 1 when storage is negative (no spilling occurred here whenever the reservoir was emptied), and it is less than 1 when some water is stored or spilled.

5. Results

5.1. Weight of SEF and GEM models

It has first been verified whether or not there is a dependency between SEF model's weight and the forecasting horizon. According to Fig. 6, there is no such clear relationship between the horizon and weights for all sub-catchments, as SEF's weight fluctuates from 0.55 to almost 0.7. Those values have been obtained from a maximum likelihood estimation and the 90% level confidence intervals are also included on the figure (dots above and below the curve). They were obtained using the bootstrap technique applied separately for SEF and GEM forecast members, and also a second bootstrap on the forecast–observation archive in order to include a temporal component on the uncertainty.

SEF's weights for Maniwaki sub-catchment are shown in Table 2 per horizon and season. Similar results were obtained for the other sub-catchments, which is favourable considering that SOHO does not allow weighting specific to geographical location. Consequently, one has to resort to the mean weight of the six sub-catchments. As can be seen from Table 2, SEF model's weight is always superior but close to 0.5, especially considering the large 90% confidence interval shown in Fig. 6. For the winter season, SEF's weight is always greater than 0.6, indicating the superiority of SEF for that time of the year.

Table 2
Weight of model SEF as a function of the horizon and season, for hydrological ensemble forecasts at Maniwaki.

Horizon	All	Winter	Spring	Summer–fall
2-day	0.63	0.61	0.72	0.54
3-day	0.57	0.71	0.59	0.54
4-day	0.61	0.69	0.58	0.62
5-day	0.67	0.73	0.61	0.68
6-day	0.66	0.77	0.61	0.67
7-day	0.64	0.68	0.55	0.63
8-day	0.61	0.63	0.52	0.62
9-day	0.60	0.64	0.62	0.61
10-day	0.59	0.70	0.55	0.59

As can be seen from the confidence intervals in Fig. 6, the uncertainty on SEF model's weight is quite large and it is not clear if both models should be attributed equal weights or not. To further verify the effect of weighting the members differently depending on which model they come from, two supplemental simulations are analyzed: attributing excessive weight to SEF members (0.8, divided into the ensemble members plus the control forecast) and similarly for GEM. Ultimately, there are four weighting scenarios: all members equiprobable, SEF model attributed a weight of 0.6 (according to Fig. 6 and Table 2), SEF model attributed a weight of 0.8 and finally SEF model attributed a weight of 0.2. Those four weighting scenarios are next passed into SOHO in order to quantify the influence of weighting on operational performance.

5.2. Forecasts performance

Before providing the streamflow forecasts to SOHO, their performance relative to the observations is evaluated using the CRPS with its decomposition as well as the MAE. First, the mean absolute error (MAE) between deterministic forecasts and corresponding observations as a function of the forecasting horizon is drawn on

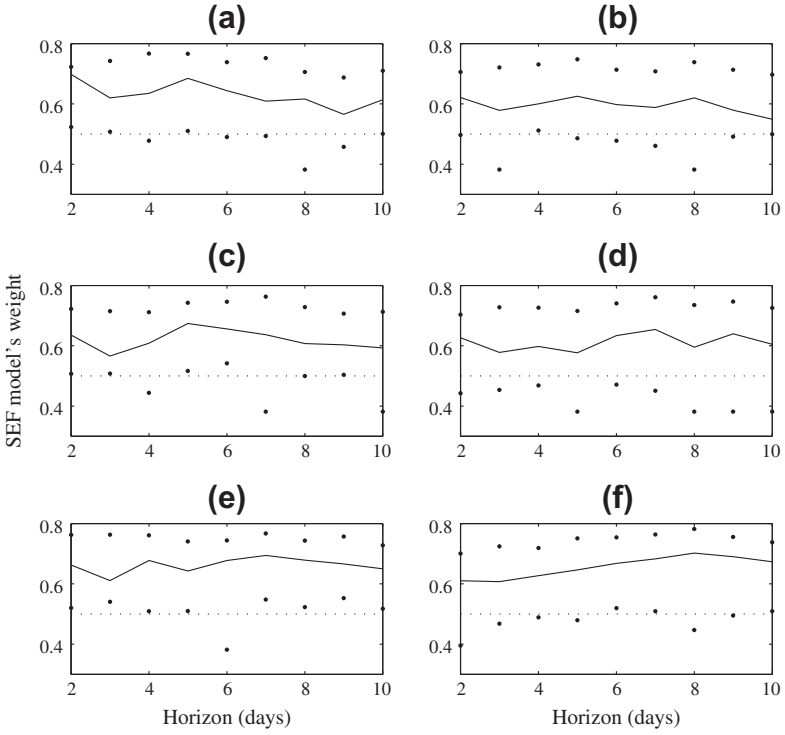

Fig. 6. Weight of SEF model as a function of forecasting horizon in hydrological ensemble forecasts for sub-catchments (a) Chelsea, (b) Paugan, (c) Maniwaki, (d) Baskatong, (e) Cabonga, and (f) Ceizur.

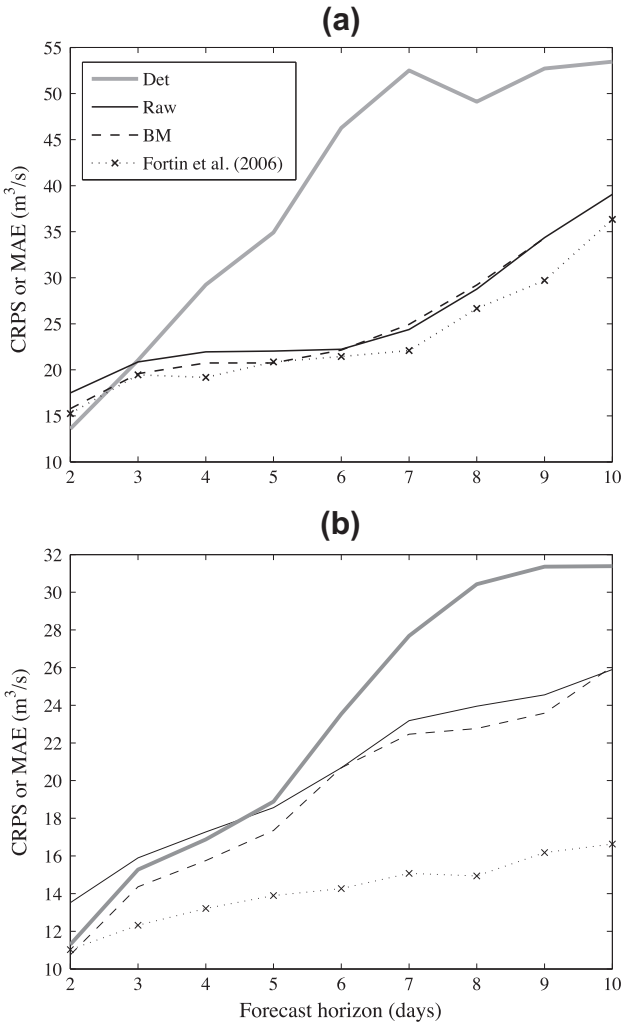

Fig. 7. Comparison of the MAE of the deterministic forecasts (thick grey line) and CRPS of the ensemble forecasts (raw ensembles: black line, post-processed by best member: dashed line and by Fortin et al. (2006) method: dotted line with *x*'s markers) as a function of the forecasting horizon, for (a) Maniwaki sub-catchment and (b) Chelsea sub-catchment.

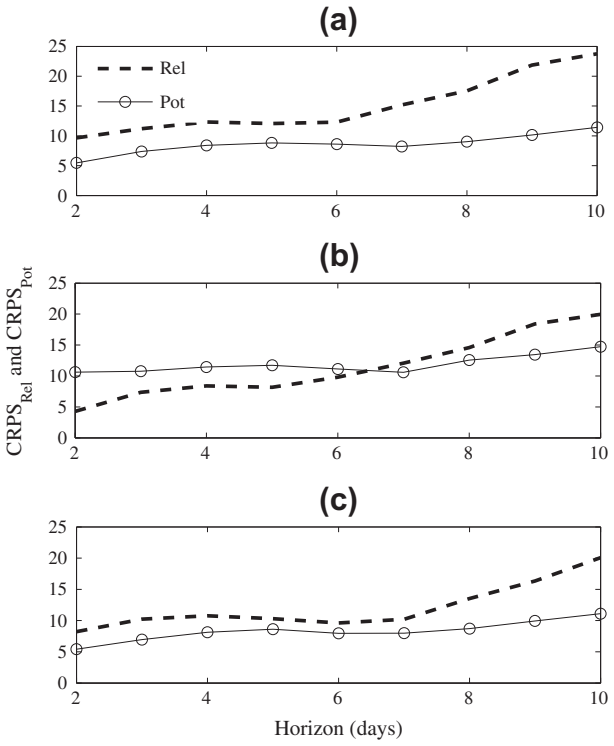

Fig. 8. Decomposition of the CRPS for Maniwaki sub-watershed for (a) raw ensemble forecasts, (b) forecasts post-processed by the best member method and (c) by Fortin et al. (2006) method. The dashed line corresponds to the reliability component and the solid line with circle markers corresponds to the potential component.

Fig. 7 for the Maniwaki (a) and Chelsea (b) sub-catchments, along with the Continuous Ranked Probability Score (CRPS).

In Fig. 7, the thick grey line corresponds to the MAE from the high resolution forecasts, the solid black line represents the CRPS from the original ensemble forecasts, whereas the dashed line and the dotted line with *x*'s markers represent respectively the ensembles post-processed using the best member and Fortin et al. (2006)'s methods. The deterministic forecasts score a lower MAE than any of the ensembles CRPS, for short-term forecasts. But for longer horizons (from 3 days for Maniwaki and from 5 days for Chelsea), the MAE increases more rapidly and surpasses all the ensemble variants, implying the general superiority of the probabilistic approach. The fact remains that both the CRPS and the MAE increase with the forecasting horizon, which means that all forecasts accuracy decreases as the horizon lengthens. As for the comparison between raw and post-processed ensembles, Fortin et al. (2006) method better improves the CRPS for all horizons, while the best member method improves the CRPS only in the short-term for Maniwaki, up to about five days. It does, however improve the forecasts for most horizons in Chelsea's case.

Fig. 8 illustrates the evolution of the reliability and potential components of the CRPS as a function of the horizon, for raw forecasts (a), after best member post-processing (b) and after Fortin et al. (2006)'s post-processing (c). For all three cases, the reliability component increases with the forecasting horizon, which means that the forecasts become less reliable as the horizon lengthens. The potential component also increases with the horizon, but less markedly. One remarkable feature of Fig. 8b is that, for forecast horizons shorter than 6-day, the reliability component of the CRPS is lower than the potential component, which means that the forecasts are reliable but have a very poor resolution. This could happen, for instance, when the predictive distribution has a very wide spread so it most often comprises the observed value, but the spread could be reduced and the observed value would still fall inside the distribution. Also, this figure shows that best member post-processed ensembles possess a potential CRPS component that is higher than for the raw forecasts. This means that the post-processing method, although improving reliability by almost 50%, also significantly reduces resolution, which is not suitable. This does not happen with Fortin et al. (2006)'s method (Fig. 8c), for which the potential component of the CRPS remains unchanged (compared to raw forecasts) while reliability improves (it becomes lower). Although the improvement of reliability is not large, it could still translate into improved decision making. Finally, note that for the longest horizons, both post-processing methods have similar reliability components but Fig. 8b shows that the best member methods leads to a higher (worst) potential component.

5.3. Economic performance

Once all types of forecasts have been passed through SOHO their performance can be compared in term of production, spilling and storage in Baskatong reservoir. The time period for verification spans from 2003 September 1st to 2003 December 17th.

Fig. 9 shows what would have been observed for fall 2003 if reservoir operations had been solely based upon SOHO and deterministic forecasts. The spilling is then greatly reduced because SOHO computes a final reward parameter associated with water left in the reservoir, which favours optimal long-term management and compromises between immediate usage and future ones. In opposition, as in Fig. 4, the management of the system was performed in a day-to-day fashion without any decision support tool. The main concern is then maximizing the drop height of the water, which in turn encourages maintaining Baskatong reservoir at a high level. This practice may be inefficient for long-term considerations in the eventuality of important unexpected rainfall events, because it does not leave much margin for operations. Note that since Baskatong reservoir is located in the upper part of the catchment, in-stream flow is directly dependent on the quantities released from this reservoir and so is the amount of power that can be produced by the three downstream plants.

Fig. 10 presents the evolution of the same management variables as in Figs. 9 and 4, but for the case of weighted raw ensemble forecasts. The eight SEF members have a weight of 0.075 each and GEM members have a weight of 0.044 each. In comparison to Fig. 9, the spilling is reduced further, but the behaviour of the other management variables does not change drastically. The same comment applies to all the other simulations: raw ensemble forecasts with various weighting strategies or post-processed with the best member method or Fortin et al. (2006) method. They are thus not shown here.

SOHO was designed to prioritize the respect of the exploitation constraints, before optimizing water use. Consequently, as long as all ensemble members (streamflow scenarios) remain inside the limits of exploitation (respect of exploitation constraints), all management solutions are similar. This causes SOHO to be insensitive to over or under dispersion of streamflow ensemble forecasts, to some extent. SOHO performs especially well for situations that are very near exploitation limits. We believe that this aspect of SOHO merits improvements. The decision support system should of course lead to the respect of exploitation constraints, but its discrimination capacity between streamflow scenarios that all lie in the acceptable range of exploitation limits for optimal water use could be improved.

Tables 3 and 4 contain numerical values taken from the management experiment. Mean values of the turbinate flow over the period

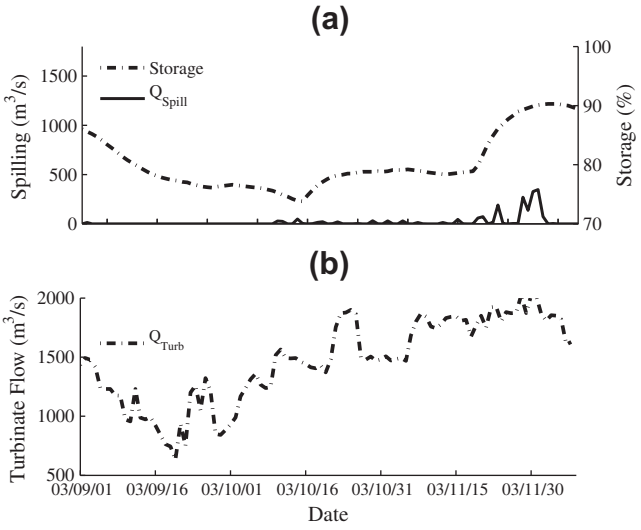

Fig. 9. Simulated management variables for the Gatineau watershed hydropower system in fall 2003 with high resolution deterministic forecasts: (a) Storage in Baskatong reservoir and spilled flow. (b) Turbinate flow.

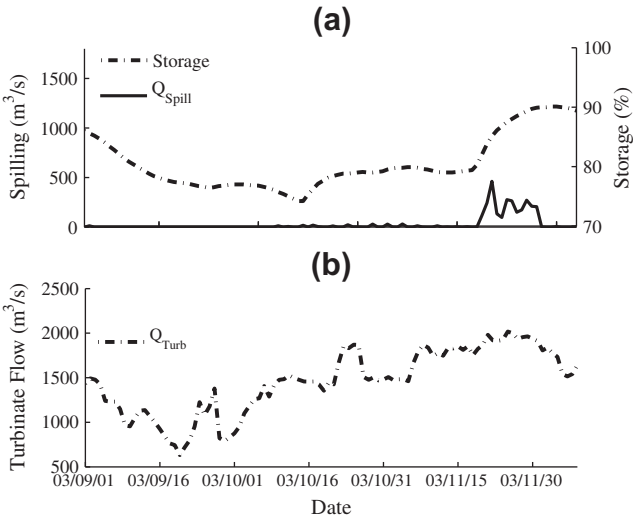

Fig. 10. Simulated management variables for the Gatineau watershed hydropower system in fall 2003 with low resolution weighted ensemble forecasts: (a) Storage in Baskatong reservoir and spilled flow. (b) Turbinate flow.

from 2003 September 1st to 2003 December 17th are presented, as well as the total production, spilling and storage (in Baskatong reservoir), all converted in energy equivalent units (GW h). The excess volume (inundations) at Maniwaki and number of inundation days are also given. Note that the turbinate flow is the sum of the water passing in succession through Paugan, Chelsea and Rapides-Farmers plants, which all have similar production capacity.

Results given in Table 3 demonstrate the great potential improvement achievable through the inclusion of SOHO in Hydro-Québec management practice. In fact, for all cases, excess water volumes at Maniwaki are prevented, increasing public safety. The spilling is also considerably reduced, probably because of an optimization scheme that takes longer term information into account. Raw ensemble forecasts offer slightly lower energy production and increased spilling than deterministic forecasts. Using the best member post-processing improves turbinate flow and production to levels higher than for the deterministic forecasts. Fortin et al. (2006)'s method, despite better performance in terms of CRPS, does not lead to comparable improvements, but increases the storage.

Results obtained through different weighting possibilities for forecast members from SEF and GEM are shown in Table 4. The term "weighted ensemble" (W. Ens.) refers to the situation where a total weight of 0.6 is distributed among SEF members, as suggested by Fig. 6 and Table 2. This approach improves turbinate flow and production, but not as much as can be reached using the best member method. Interestingly, giving important weights to SEF members (third column) leads to a rise in turbinate flow, but also increases the spilling whereas the opposite is true for attributing important weights to GEM members (second column). Therefore, it is possible that GEM model leads to more "conservative" decision making, prioritizing relatively high storage in Baskatong reservoir whereas SEF may prioritize immediate power production. To verify those assumptions, the average of SEF and GEM daily forecasts were computed separately for each sub-watershed and each forecasting horizon, as well as the corresponding standard deviation. It revealed that SEF generally issues higher streamflow forecasts, with a larger standard deviation compared to GEM members. This difference increases with the forecasting horizon. It is therefore consequent that SEF forecasts lead to increased turbinate flow and spilling, as there is more water in the system. Further investigation regarding this issue would require a deeper knowledge of the detailed mechanics of SOHO, which is beyond the objective of this work.

Table 3
Comparison of different forecast types for the hydropower production management on the Gatineau watershed.

Criteria	Real	Det.	Ens.	Best member	Fortin et al.
Mean turbinate flow (m^3/s)	1202	1467	1453	1470	1451
Production (GW h)	722	839	831	840	831
Spillage (GW h)	129	10	16	14	14
Storage (GW h)	18	8	8	6	9
Tot. excess. vol. Maniwaki (h m^3)	295	0	0	0	0
Number of inundation days:	17	0	0	0	0

Table 4
Comparison of different weighting scenarios for the hydropower production management on the Gatineau watershed. Column 2 (W. Ens.): weight of 0.6 to SEF members and 0.4 to GEM members, Column 3 (GEM): weight of 0.8 to GEM members and Column 4 (SEF): weight of 0.8 to SEF members.

Criteria	W. Ens.	GEM	SEF
Mean turbinate flow (m^3/s)	1458	1455	1462
Production (GW h)	833	830	835
Spillage (GWh)	16	14	15
Storage (GW h)	7.45	8.46	6.76
Tot. excess. vol. Maniwaki (h m^3)	0	0	0
Number of inundation days:	0	0	0

Fig. 11 compares the production ratio for real operations in 2003 (thick solid line) with those for deterministic forecasts (thick dashed grey line), raw ensemble forecasts (fine line with circle markers) and best member post-processed ensemble forecasts (fine line with square markers). In general, for real operations, the decision was to empty Baskatong reservoir gradually until mid-October, followed by a week of intense storage. The water level was next kept more or less constant up to mid-November when large storage events occurred, especially around the 20th. The overall lower production achieved during real operations (Table 3) explains why this production ratio is generally lower than the others. It is also more peaky than the others, possibly reflecting day-to-day short-term decisions where large quantities of water were withdrawn from Baskatong reservoir on particular occasions. Such withdrawals may also correspond to real life constraints that were not accounted for by SOHO. The most striking divergence between real operations and the simulated ones occurred at the end of the testing period when large spilling events push the real production ratio downward. Otherwise, the three curves corresponding to management simulations with deterministic, raw and best member ensembles are globally closer to unity than the black curve representing real operations, which could be associated to wiser long-term management of the system. There is not much difference between those last three curves, except that production ratios for ensembles post-processed with the best member method (square markers) and deterministic forecasts (dashed thick grey line) are generally closer to each other than raw ensembles (circle markers) with deterministic forecasts.

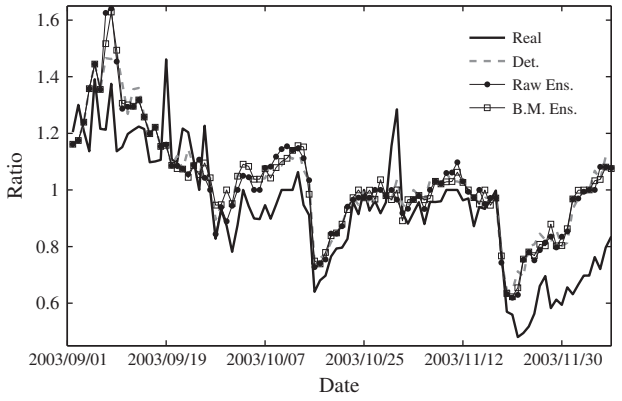

Fig. 11. Production ratios for real operations in 2003 (black solid line), deterministic forecasts (black dashed line), raw ensemble forecasts (fine line with circle markers) and best member ensembles (fine line with square markers).

6. Conclusions

This experiment on the management of the Gatineau hydropower production complex allows the comparison of different hydrological forecasting systems in terms of production variables directly linked to economic value for the electricity producer. To our knowledge, such a comparison is not frequent in the literature. It should be of great interest to forecasts users because gains in terms of electricity production, reduction of spillage, and prevention of inundations are much more convincing to them than improvements in more abstract numerical scores such as the CRPS or various curvatures of rank histograms. Amongst the most promising results, it is shown that ensemble forecasts outperform deterministic ones for all horizons greater than 48-h and that post-processing improves further that performance. Fortin et al. (2006)'s method leads to the best Continuous Ranked Probability Score (CRPS) results, for horizons greater than 5 days. Post-processing of the ensembles by the best member method (Roulston and Smith, 2003) improves the hydroelectric power production over the one derived from the deterministic forecasts, which leaves us believing in their operational potential. As for the hydroelectricity production, a noticeable improvement is achieved through the use of the stochastic decision-making assistance tool.

Comparison between fall 2003 real operations and simulated ones, based on operational hydrological forecasts and the SOHO decision system, promotes implementation of a stochastic management tool for decision-making assistance over large complex hydro-systems. Human seem to have a tendency to favour short-time gains and have difficulties evaluating and comparing long-term benefits of different scenarios. In the case of the Gatineau watershed, in fall 2003, the level of the Baskatong reservoir was maintained too high, so the operators were unsuccessful to contain excess water when large precipitation occurred.

On some occasions, improvements identified by performance scores based on the comparison between ensemble forecasts and observations, such as the CRPS, did not directly translate into improved energy production. The case of the ensemble forecasts post-processed with Fortin et al. (2006) is eloquent in that sense. Operations linked to electricity production on a large system are complex while mathematical subtleties of the management operation optimization tool (SOHO) belong more to the field of operational research than hydrology. One has to keep in mind that SOHO was designed for deterministic forecasts, and does not take into account the entire predictive distribution nor does it allow the weighting of the streamflow scenarios to vary with geographical location. Furthermore, a two-stage decision tree was used in this study and it would be interesting to test a multi-stages tree. It is therefore probable that SOHO is unable, in its actual form, of exploiting all the features of hydrological ensemble forecasts. As a

consequence, improvements in reliability or resolution of the ensemble forecasts, significant in terms of the CRPS, do not necessarily translate in terms of reservoir operations. In a multi-stage tree, the probability for each streamflow scenario would evolve with the forecasting horizon. This would be more in line with the forecasts, as their dispersion depends on the forecasting horizon. A multi stage tree would consequently allow a more refined description of scenarios probabilities.

Two main conclusions are drawn from this experiment. There is room for improvement of management operation optimization tools so that they can use more of the information carried by hydrological ensemble forecasts. Also, there is a demand for testing and comparing the performance of hydrological ensemble forecasts in terms of operational attributes. Performance assessment metrics based strictly on a comparison between forecasts and observations are necessary but not sufficient to establish the value of a forecasting system, especially from an operational point of view, because such metrics may not be exhaustively representative of the whole complexity of the problem.

Finally, as previously noted, high resolution deterministic forecasts outperformed most ensembles in terms of turbinate flow and production. Further investigation is needed to evaluate the extent to which this effect can only be attributed to higher spatial resolution or if the meteorological model also influence this issue. Knowing that the spatial resolution of deterministic forecasts is better that the resolution of ensemble forecasts, it would be interesting to measure the extent to which spatial resolution is a factor in improving the management performance of a forecasting system over another. Further investigation will be needed to assess this. It will require the identification of interesting precipitation events that occurred after 2007, when the most recent ensemble forecasting system was put into operation. Those events would have to resemble the one observed in fall 2003, with flows of comparable magnitude and a similar duration. This issue emphasizes the importance of precipitation and temperature reforecasts. Reforecasts would help gaining a better understanding of how the improvement of meteorological models, improved meteorological forecast spatial resolution as well as increased number of members translate into better streamflow forecasts. Also, further collaboration with the different stakeholders, especially from the private and corporate sectors, would be of great help. For instance, access to additional economic data related to water management and decision making based on forecasted streamflow would be relevant to improve streamflow ensemble forecasting and ensure that it is helpful for operational applications.

Future studies should involve a management operation optimization tool with the possibility of modifying its parameters and structure to allow deeper understanding of the interactions between the management optimization tool and the different types of forecasts.

References

Baringhaus, L., Franz, C., 2004. On a new multivariate two-sample test. J. Multivariate Anal. 88, 190–206.

Bélair, S., Roch, M., Leduc, A.M., Vaillancourt, P., Laroche, S., Mailhot, J., 2008. Medium-range quantitative precipitation forecasts from Canada's new 33-km deterministic global operational system. Weather Forecast 24, 690–705.

Bibeau, L., Krau, S., Latraverse, M., Tremblay, D., 2006. Construction d'un arbre de scénarios pour le bruitage des grilles de prévisions météorologiques. Technical Report IREQ-2006-0199. Hydro-Québec, Institut de Recherche.

Birge, J.R., Louveaux, F., 1997. Introduction to Stochastic Programming. Springrer Series in Operations Research, New-York.

Boucher, M.A., Tremblay, D., Perreault, L., Anctil, F., 2011. A comparison between ensemble and deterministic hydrological forecasts in an operational context. Adv. Geosci. 29, 85–94.

Buizza, R., Houtekamer, P.L., Toth, Z., Pellerin, G., Wei, M.Z., Zhu, Y.J., 2005. A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. Mon. Weather Rev. 133, 1076–1097.

Casati, B., Wilson, L.J., Stephenson, D.B., Nurmi, P., Ghelli, A., Pocernich, M., Damrath, U., Ebert, E.E., Brown, B.G., Mason, S., 2008. Forecast verification: current status and future directions. Meteorol. Appl. 15, 3–18.

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

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

Day, G.N., 1985. Extended streamflow forecasting using NWSRFS. J. Water Plan. Manage. ASCE III, 157–170.

Evora, N. 2005. Valorisation des prévisions météorologiques d'ensemble. Technical Report IREQ-2005-065. Hydro-Québec, Institut de Recherche.

Fortin, J.P., Moussa, R., Bocquillon, C., Villeneuve, J.P., 1995. 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, 97–124.

Fortin, J.P., Turcotte, R., Massicotte, S., Moussa, R., Fitzback, J., 2001. A distributed watershed model compatible with remote sensing and GIS data. Part I: Description of the model. J. Hydrol. Eng. Am. Soc. Civil Eng. 6, 91–99.

Fortin, V., Favre, A.C., Saïd, M., 2006. Probabilistic forecasting from ensemble prediction systems: improving upon the best-member method by using a different weight and dressing kernel for each member. Quart. J. Roy. Meteorol. Soc. 132, 1349–1369.

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

Good, I.J., 1952. Rational decisions. J. Roy. Statist. Soc. B 14, 107–114.

Hamill, T.M., Colucci, S.J., 1997. Verification of ETA–RSM short-range ensemble forecasts. Mon. Weather Rev. 125, 1312–1327.

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

Kim, Y.O., Eum, H.I., Lee, E.G., Ko, I.H., 2007. Optimizing operational policies of a korean multireservoir system using sampling stochastic dynamic programming with ensemble streamflow prediction. J. Water Resour. Plan. Manage. 10, 14.

Krau, S., 2005. Présentation générale de SODAD, Super Outil D'Aide la Décision. Technical Report. IREQ-2005-077.

McCollor, D., Stull, R., 2008. Hydrometeorological short-range ensemble forecasts in complex terrain. Part II: Economic evaluation. Weather Forecast. 23, 557–574.

Mylne, K.R., 2002. Decision-making from probability forecasts based on forecast value. Meteorol. Appl. 9, 307–315.

Nash, J.E., Suthcliffe, J.V., 1970. River flow forecasting through conceptual models : Part I: A discussion of principles. J. Hydrol. 10, 282–290.

Ramos, M.H., Mathevet, T., Thielen, J., Pappenberger, F., 2010. Communicating uncertainty in hydro-meteorological forecasts: mission impossible? Meteorol. Appl., 223–235.

Richardson, D., 2000. Skill and relative economic value of the ECMWF ensemble prediction system. Quart. J. Roy. Meteorol. Soc. 126, 649–667.

Roulin, E., 2007. Skill and relative economic value of medium-range hydrological ensemble predictions. Hydrol. Earth Syst. Sci. 11, 725–737.

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

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

Shaake, J., Hamill, T., Buizza, R., Clark, M., 2007. HEPEX – the Hydrological Ensemble Prediction Experiment. Bull. Am. Meteorol. Soc., 1541–1547.

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

Talagrand, O., Vautard, R., Strauss, B., 1997. Evaluation of Probabilistic Prediction Systems, ECMWF Workshop on Predictability, Shinfield Park, Reading, Berkshire. pp. 1–25.

Toth, Z., Zhu, Y., Marchok, T., 2001. The use of ensembles to identify forecasts with small and large uncertainty. Weather Forecast. 16, 463–477.