Tools for the assessment of hydrological ensemble forecasts obtained by neural networks

Marie-Amélie Boucher, Luc Perreault and François Anctil

ABSTRACT

The increasing demand for uncertainty assessment in streamflow forecasts has drawn the hydrological community's interest toward ensemble forecasting techniques. The widespread deterministic hydrological forecasting point of view focuses to a great extent on the search for a hydrological model that would come as close as possible to "perfection" (i.e. the aim is to implement a model that produces a point forecast that is as close as possible as the observed outcome). On the other hand, ensemble forecasting departs from the deterministic point of view by avoiding the assumption that the "perfect" model exists and instead focuses on issuing a type of forecast that accounts explicitly for the uncertainty inherent to the forecasting process as a whole. In this paper, one-day-ahead hydrological ensemble forecasts obtained by stacked neural networks are presented and analysed. To do so, three simple performance assessment criteria are presented. Those criteria were originally developed in the meteorological and statistical communities to accommodate the need for a quality assessment methodology that is coherent with the probabilistic nature of ensemble weather forecasts. It will be shown that, even though the ensemble forecasts suffer from underdispersion, they outperform point forecasts.

Key words | continuous ranked probability score, hydrological ensemble forecasts, neural networks, stacking

Marie-Amélie Boucher (corresponding author)
François Anctil
Department of Civil Engineering,
Université Laval,
Pavillon Adrien Pouliot,
1065 avenue de la Médecine,
Québec G1V 0A6,
Canada
Tel.: +1 418 656 2131X8727
E-mail: *marie-a.boucher.1@ulaval.ca*

Luc Perreault
Hydro-Quebec, IREQ,
Varennes J3X 1S1,
Canada

INTRODUCTION

The last few years have seen an increasing demand for ensemble and probabilistic streamflow forecasts emerging from the user community. Rather than just a point streamflow forecast, there is a need for an estimation of the forecast's uncertainty (e.g. Krzysztofowicz 2001). Ensemble forecasts provide this type of information and allow for the calculation of the probability of being over a certain threshold as well as the evaluation of confidence intervals to be associated with the forecast. Since 2004, the issues concerning ensemble forecasts have been the object of an international project called the Hydrological Ensemble Prediction Experiment (HEPEX) (e.g. Schaake *et al.* 2007). HEPEX promotes collaboration between meteorologists, hydrologists and users of forecasts. Although this project does not exploit physics-based models, as is the case with most HEPEX applications, it concurs with the ensemble philosophy and promotes the idea that ensemble forecasting could be beneficial not only for meteorology and hydrology but also among the neural network community.

In ensemble forecasting, the focus is taken off finding the one best estimate of the streamflow (i.e. finding a "perfect" model), and drawn to finding the best possible estimate of the forecast's uncertainty. Therefore, instead of forecasting a single streamflow value for each lead time, an ensemble forecasting system produces n members' forecasts that can be used to fit a probability density function (pdf) which in turn can be used to assess confidence intervals as well as other measures of the forecasts uncertainty.

The uncertainty originates mainly from the hydrometeorological data, from the model, and from the lack of

doi: 10.2166/hydro.2009.037

knowledge concerning the initial conditions. In a certain manner, with ensemble forecasting, the uncertainty pertaining to the forecasting situation is embedded in the forecast itself.

However, because of its probabilistic nature, the performance of an ensemble forecasting system cannot be evaluated using criteria such as the mean absolute error (MAE) or the root mean squared error (RMSE).

Various methods have been proposed to assess the quality of ensemble and probabilistic forecasts in the meteorological community (e.g. Stanski *et al.* 1989; Wilks 1995) and in statistical decision theory (e.g. Good 1952). The three tools presented in this paper are the Continuous Ranked Probability Score (CRPS), the rank histogram and the confidence interval reliability diagram. The latter two are graphical methods which can provide a diagnostic concerning the bias and the dispersion of the predictive distributions. The CRPS (Matheson & Winkler 1976) is a numerical criterion that assesses the quality of the ensemble forecasts. It takes into account the calibration of the distribution as well as its sharpness and reduces to the mean absolute error (MAE) for a point forecast (Gneiting & Raftery 2007). This quality allows the comparison between point forecasts and ensemble forecasts. In the present study, the CRPS values obtained lead to the conclusion that neural-network-based ensemble forecasts, even when uncorrected for bias and dispersion, perform better than their deterministic counterparts. The latter were taken as the mean of each daily ensemble, as suggested by Breiman (2000).

The methodology employed here was inspired by the work of Weber *et al.* (2006*a*,*b*) in the context of hydroelectricity production. In the present case, 50-member one-day-ahead streamflow ensemble forecasts were issued using neural networks for three watersheds of diverse characteristics. The networks used were multilayer perceptrons and the intrinsic instability of the backpropagation optimisation process is exploited to obtain the ensembles. Although ensemble methods are already used within the neural network community, they are mostly used as generalisation-enhancing techniques, such as stacking (Wolpert 1992) or different model combination strategies (e.g. Hansen & Salamon 1990). Therefore, the ensemble of solutions is not used as a whole but instead summarised in a point measure. Here, all the ensemble members are kept and the forecasts consist of the fitted probability density functions, which allow the calculation of various confidence intervals for the forecasts. Eventually, the predictive distributions could also be used to evaluate probabilities of the streamflow exceeding different thresholds.

In the following section, the databases are described as well as the methodology used to obtain neural-network-based ensemble forecasts. Then, three verification tools for ensemble forecasts are presented, namely the CRPS, the rank histogram and the confidence interval reliability diagram. The results and discussion are presented in the third section and the conclusion follows.

DATABASES AND ENSEMBLE FORECASTS

Description of the databases

The databases for three watersheds of diverse hydrologic and climatologic characteristics were exploited to produce one-day-ahead ensemble forecasts.

The Leaf River, located in the United States (Mississippi), has a catchment's size of 1,949 km^2 under a temperate climate. Forty years of daily precipitation and streamflow values are available for this basin. The Serein River is located in France (east of Paris) and has a catchment's size of 1,120 km^2. There are 43 years of data available for this basin. Finally, the Volpajola River, which is also located in France (Corsica), has an area of 930 km^2 and 18 years of data are available. Figure 1 shows the annual means of the daily streamflow and precipitation for the three basins. The horizontal line represents the overall annual mean of the daily streamflow.

Those three watersheds present strong interannual variability in the annual means of streamflow and precipitation, this being especially true for the Serein and Leaf Rivers. The data presented in Figure 1 also shows that a variation in precipitation is not always guaranteed to lead to a proportional variation in streamflow. Figure 1(a), for instance, has an annual mean streamflow peak of almost 3 mm for year 1961 and it decreases to reach approximately 0.5 mm in 1963, while for the same years the annual mean precipitation decreases from approximately 5.8 mm to 2.5 mm. The ratio is thus much greater for the annual mean daily streamflow than for the annual mean

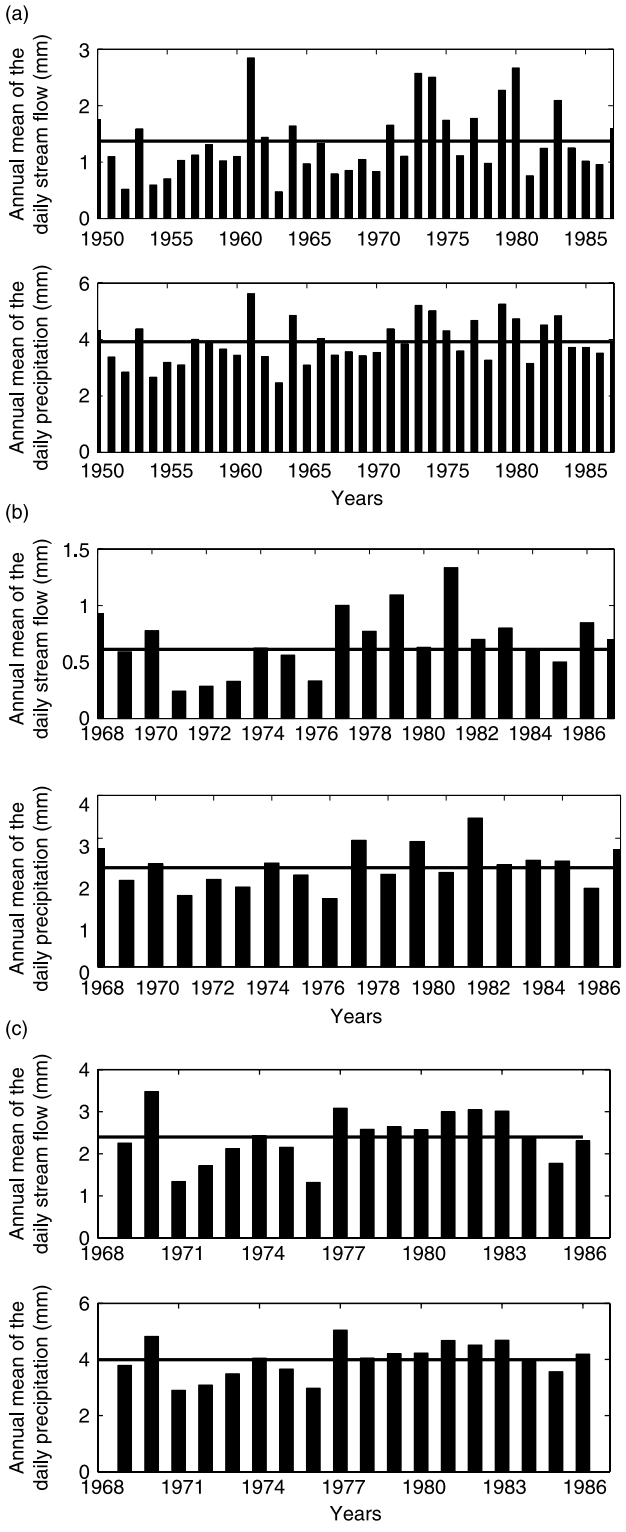

Figure 1 | Annual means of the daily streamflow and precipitation for (a) the Leaf River, (b) the Serein River and (c) the Volpajola River.

precipitation, so it can be concluded that a change in precipitation is reflected nonlinearly in the corresponding streamflow.

Ensemble forecasts

An ensemble forecast is a forecast for which n values of the predicted variable are emitted for the same lead time. On the other hand, what we define here as a deterministic forecast is a forecast consisting of a unique predicted value, a practice which implicitly assumes the forecasts to be error-free.

In the context of hydrological forecasting, the n ensemble members can be obtained either by giving a model n sets of equally likely initial conditions or parameter sets or by using n different models running in parallel. Another possibility is to combine meteorological ensemble forecasts with a hydrological model. Consequently, an ensemble forecast consists of an ensemble of n possible values of the predicted variable for the same lead time. A probability density function (pdf) can then be fitted to the ensemble members at each time step.

For each of the basins mentioned earlier, one-day-ahead ensemble forecasts were produced using neural networks. Each of the three databases was first split in two parts using a Kohonen network (Kohonen 1990). This ensures that the training dataset and the validation dataset have the same statistical properties (Klemeš 1986). Next, a multilayer perceptron (MLP) was used as a rainfall–runoff model. This kind of network is popular in the hydrological community mainly because of its great suitability for vectorial information treatment (Rosenblatt 1958). The networks used here consist of one input layer, a five-neurons hidden layer and an output layer consisting of one neuron. The four inputs are the precipitation for time t, $t-1$ and $t-2$ and the streamflow for time t, while the output is the streamflow at time $t+1$. The transfer function for the five hidden neurons is the sigmoid tangent and the output transfer function is linear. The cost function is the mean squared error, used with Bayesian regulation (Foresee & Hagan 1997). When Bayesian regulation is used, the database only needs to be split in two instead of three, like in the case with other generalisation-enhancing methods.

This architecture was proven to be efficient with those databases in a previous study (Anctil & Lauzon 2004).

Since the backpropagation optimisation method used in the training of the networks is unstable, a different parameter set will be obtained each time the optimisation process is launched. This variability can be exploited to obtain an ensemble, as illustrated in Figure 2. In atmospheric sciences, ensembles are usually generated by a Monte Carlo procedure that exploits the chaotic characteristics of the atmosphere. The model's initial conditions are perturbed and a set of equally likely ensemble members are obtained. Unlike meteorological models, neural networks are not physics-based. A neural network, rather, constructs a nonlinear relationship between given inputs and outputs. This is done by adjusting its parameters during an optimisation phase that exploits a portion of the available data (the remaining data are used to test the capability of the network). Here, we choose not to perturbate the initial state of a model, but rather to initiate the optimisation process from a different initial location for each neural network (each ensemble member). A different local minimum of the optimisation function is thus found at each repetition of this process. The consequence of this is that each of the 50 neural networks has different weights and biases, which are the model's parameters. Therefore, our method for constructing the ensembles can be viewed as a perturbation of the model's parameters rather than the initial conditions. Another possibility to generate ensembles would be to perturb the precipitation data.

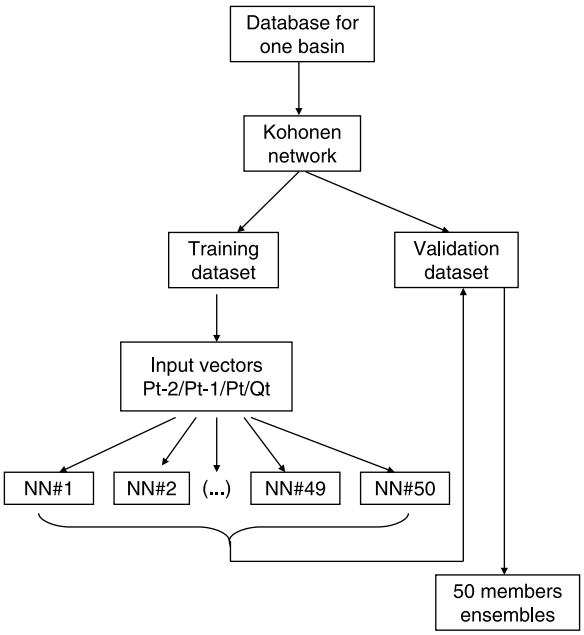

Figure 2 | Neural-network-based stacked ensemble forecasts procedure.

Ensembling methods are now gaining popularity among the neural network community. However, neural network ensemble procedures are, most of the time, seen as a means of stabilising the final output. The most common practice in this community is to average the ensemble members to obtain a point forecast (e.g. Hansen & Salamon 1990; Zhou *et al.* 2002; Zhang 2007). This averaged output of many neural networks trained on the same data has less variability than the outcome of a single network. We argue that information concerning the forecast's uncertainty is lost in the averaging process and that all the ensemble members should be considered.

Here, the training process was repeated 50 times for neural networks with identical architecture, leading to 50 different sets of weights and biases.

The choice of neural networks to do the modelling was motivated by their easy and fast implementation. It is important to emphasise that the tools presented in this paper could be used to assess the quality of ensemble forecasts produced by any type of continuous hydrological model.

The neural-network-based hydrological forecasts presented here also differ from those obtained by Weber *et al.* (2006*a*,*b*). The predictive distributions obtained here are close to normality, which facilitates certain types of operations, for example *a posteriori* calibration of the forecasts (e.g. the inflation factor method presented in Gneiting & Raftery (2007)).

VERIFICATION

Once the daily streamflow observation of day t, x_t, is available, we want to assess the ability of the probabilistic forecast system. This involves comparison of the observation x_t with the corresponding predictive distribution, say F_t. The main difficulty is that F_t is a function whereas the corresponding observation is a scalar. Therefore, one cannot use standard measures of performance such as the MAE.

Performance assessment methods developed in atmospheric sciences and in statistical decision theory can be considered to overcome this difficulty. Three of them are

described in the following subsections. But beforehand, one must specify what is meant by a "good" ensemble forecast. We adopted the perspective of Gneiting & Raftery (2007) who contended that a good probabilistic forecast must maximise the sharpness of the predictive distribution F_t, subject to calibration. A probabilistic forecast is well calibrated if the forecasts and the observations are statistically compatible. For example, the 80% confidence interval calculated using the predictive distribution F_t should, on average, contain the observed value in 8 cases out of 10. Maximising the sharpness of the distribution means that its spread should be reduced to a minimum. This can be done by improving the model as well as all hydrometeorological data used in the forecasting process, so that each member of the ensemble forecast is more accurate.

In what follows, we briefly present the tools we considered to assess the one-day-ahead ensemble forecasts produced by our neural networks model.

Scores

A *score* is a numerical criterion which evaluates the forecast's quality, whether this forecast is deterministic or probabilistic. The *scoring rule* is the name given to the equation used to calculate the score. The absolute error (AE) is an example of a score that is suitable for deterministic forecasts. It is given by $\mathrm{AE} = |\hat{x}_t - x_t|$, where $\hat{x}_t$ is the deterministic forecast for day t and x_t is the corresponding observation. In the case of ensemble or probabilistic forecasts, there are scores that are appropriate for forecasts of discrete variables and others that are appropriate for continuous variables. Those scoring rules usually require that a pdf be fitted to the ensemble forecasts. As mentioned above, this pdf, denoted F_t for the forecast of day t, is called the predictive distribution.

Each score has specific characteristics, and one may chose a score that is suited to his needs. However, one should concentrate on scores that are proper. Improper scores lead to conclusions inconsistent with common sense. A proper score is internally consistent in the sense that the forecast distribution is given an optimal expected score when the observation is, in fact, drawn from that probability distribution (Gneiting & Raftery 2007). This justifies the focus only on proper scoring rules.

The score used in this study is the Continuous Ranked Probability Score (CRPS) (Matheson & Winkler 1976), which is proper. One of its interesting properties is that the CRPS reduces to the AE in the case of a deterministic forecast (Gneiting & Raftery 2007). Therefore, it is possible to compare the performance of probabilistic forecasts with those of deterministic forecasts. In order to make such a comparison, we considered as a deterministic forecast the mean value of each daily ensemble forecast.

Along with the ignorance score (Roulston & Smith 2002), the CRPS is widely used in atmospheric sciences (e.g. Hersbach 2000; Candille & Talagrand 2005; Gneiting *et al.* 2005).

It is important to stress that the standalone score obtained by the forecast for a particular day has no meaning. The only way one can possibly assess the performance of a probabilistic forecasting system is by using statistical accumulation, meaning that the statistical correspondence between the forecasted pdf F_t and the point observation x_t has to be evaluated over a large number of forecast–observation pairs. Therefore, the scores have to be averaged over all forecast–observation pairs in the archive.

The continuous ranked probability score

The CRPS is suitable for probabilistic forecasts of continuous variables. It is defined as

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

where, as above, F_t is the predictive cumulative distribution function (cdf) for day t, x is the predicted variable (here the streamflow) and x_t is the corresponding observed value. The function $H\{x \geq x_t\}$ is the Heaviside function which equals 1 for predicted values over the observed value and 0 for predicted values lower than the observation. The value of the perfect score is 0. Therefore, one must aim to minimise the score. The CRPS is not bounded on the upper side.

The CRPS is the integral of the Brier score (Brier 1950) for an infinity of thresholds. The Brier score is

$$\mathrm{BS} = \frac{1}{N} \sum_{t=1}^{N} (p_t(x) - o_t)^2$$

where $p_t(x)$ is the forecasted probability of occurrence of a particular event x and o_t is equal to 1 if the event is observed and 0 if it is not. N represents the total number of forecast–observation pairs in the archive.

The CRPS evaluates simultaneously the calibration and the sharpness of the predictive distributions. It represents the area under a curve delimited by the squared difference between the predictive cdf and the Heaviside function, as defined previously. Therefore, if the observed value corresponds to a high forecasted probability while the probability of other events is minimal (sharp forecast), the area under this curve is minimised and so is the score. Figure 3 illustrates this explanation. Figure 3(a) shows a forecast (the pdf) with the corresponding observation (Heaviside function). In Figure 3(b), we consider the cdf for the same forecast with the observation. Then, in Figure 3(c), the CRPS is obtained by computing the squared area under the curve.

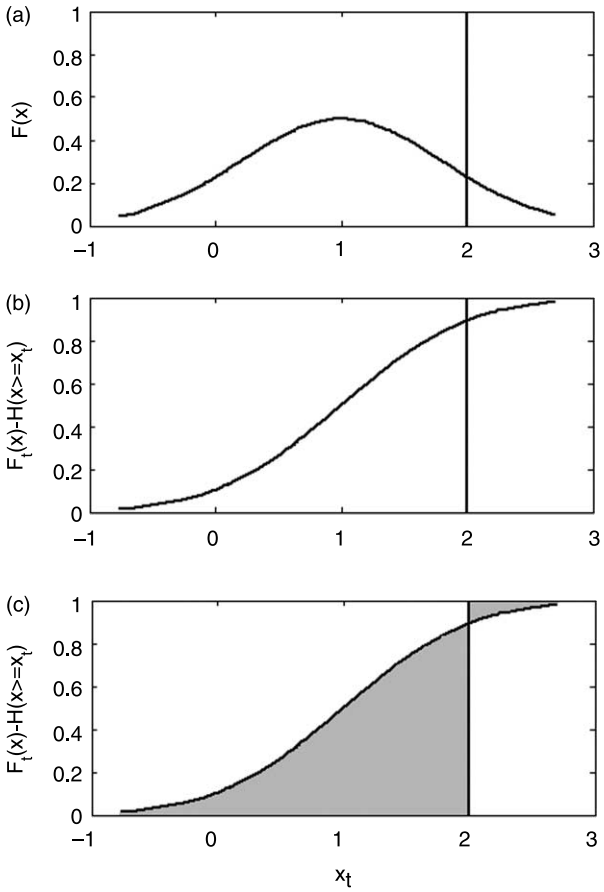

Figure 3 | Graphical step-by-step description of the CRPS for a normal predictive distribution with $\mu_t = 1$, $\sigma_t = 0.8$ and an observed streamflow of 2 mm.

Probability plots of the neural network ensemble members obtained in this study revealed that a normal distribution could be fitted to most of the ensemble forecasts. The calculation of the CRPS was then performed using the explicit equation of the CRPS for F_t being a normal distribution with mean μ_t and variance σ_t^2:

$$\mathrm{CRPS}(F_t, x_t) = \sigma_t\left[\frac{1}{\sqrt{\pi}} - 2\varphi\left(\frac{x_t - \mu_t}{\sigma_t}\right) - \frac{x_t - \mu_t}{\sigma_t}\left(2\Phi\left(\frac{x_t - \mu_t}{\sigma_t}\right) - 1\right)\right]$$

where ϕ and Φ stand, respectively, for the reduced standard pdf and cdf.

As mentioned above, one should evaluate the score for many cases and then take the average, which will be denoted by $\overline{\mathrm{CRPS}}$. As for all statistics estimated using a sample of limited size, there is an uncertainty associated with the $\overline{\mathrm{CRPS}}$. Standard deviation and confidence intervals can be estimated using resampling techniques such as the jackknife and the bootstrap (Effron & Tibshirani 1998). We used the jackknife technique. To do so, one forecast is withdraw from the archive and the mean CRPS is calculated with the remaining ones. This procedure is repeated N times, N being the total number of forecast–observation pairs in the archive. Each time, a different forecast is withdrawn from the archive, which leads to a different $\overline{\mathrm{CRPS}}$ value. Then, the standard deviation associated with $\overline{\mathrm{CRPS}}$ can be estimated using the following equation:

$$\overline{\mathrm{CRPS}}_{\mathrm{jack}} = \left[\frac{N-1}{N}\sum_{k=1}^{N}\left(\overline{\mathrm{CRPS}}_k - \overline{\overline{\mathrm{CRPS}}}\right)^2\right]^{1/2}$$

where $\overline{\mathrm{CRPS}}_k$ is the average CRPS calculated from the kth jackknife sample (the one consisting of the original series with the kth forecast–observation pair removed), N being the total number of forecasts in the archive and

$$\overline{\overline{\mathrm{CRPS}}} = \frac{1}{N}\sum_{k=1}^{N}\overline{\mathrm{CRPS}}_k.$$

The rank histogram

A useful graphical tool called the rank histogram, or Talagrand diagram (Talagrand *et al.* 1999) allows one to visually assess the calibration of the predictive distribution.

To construct it, the observed x_t value is added to the ensemble forecast. That is, if the forecast has n members, the new set consists of $n + 1$ values. Then, the rank associated with the observed value is determined. This operation is repeated for all N forecasts and corresponding observations in the archive. The rank histogram is obtained by constructing the histogram of the resulting N ranks.

The interpretation of the rank histogram is based on the assumptions that all the members of the ensemble forecasts, along with the observation, are independent and identically distributed (iid). Under these hypotheses, if the predictive distribution is well calibrated, then the rank histogram should be close to flat. An asymmetrical histogram is usually an indication of a bias in the mean of the forecasts. If the rank histogram is symmetric and U-shaped, it may indicate that the predictive distribution is under-dispersed. If it has an arch form, the predictive distribution may be over-dispersed.

Some authors have pointed out a few flaws of the rank histogram. A U-shaped histogram, usually taken as a sign of underdispersion, can sometimes be caused by conditional bias (Hamill 2001). Figure 4, which is an adaptation from Figure 4 of Hamill (2001), shows that selecting the ensemble members with equal probabilities in two biased (one positively and one negatively) predictive distributions has the same impact on the rank histogram as selecting the ensemble members from an under-dispersed predictive distribution. Observation errors can have the same effect (Saetra *et al.* 2004). Generally, in hydrology, observed streamflow values are obtained from a gauging curve and measures of the water level. Clearly, they are not error-free, especially for large streamflow values, since they often correspond to an extrapolation of the gauging curve. Therefore, the rank histogram must be interpreted carefully and used together with other verification tools.

Recently, some numerical indicators closely linked to the rank histogram have been proposed. First, Candille & Talagrand (2005) proposed to use the ratio

$$\delta = \frac{\Delta}{\Delta_O}$$

as a measure of the "flatness" of the rank histogram. In this ratio, Δ is a measure of the squared deviation from flatness, given by

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

where s_k is the number of elements in the kth bin of the rank histogram. Δ_O is the ratio that would be obtained by a

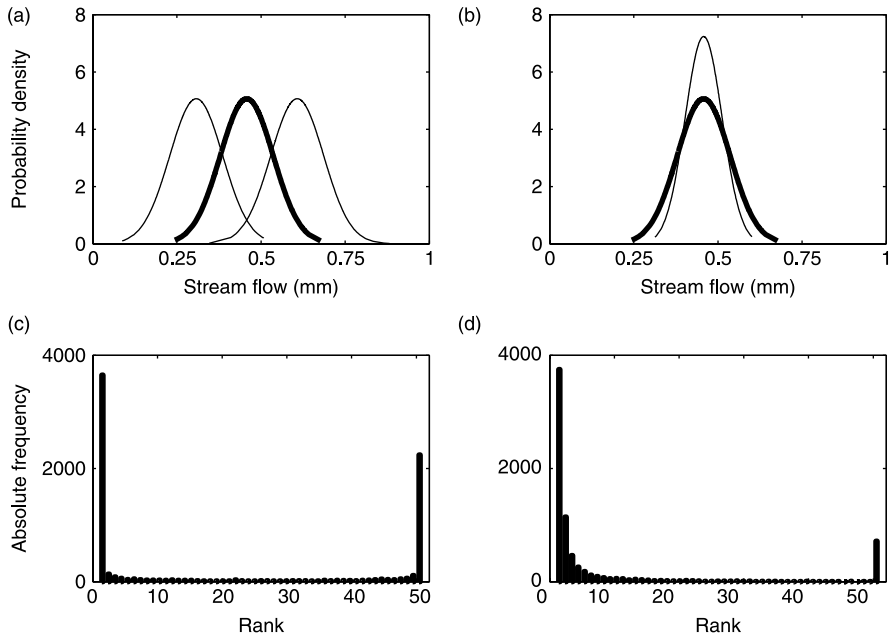

Figure 4 | The effect of conditional bias on the rank histogram. (a) The 50 ensemble members are taken from the two biased distributions with equally likely probability. (b) The members are taken from an under-dispersed distribution. (c) Rank histogram corresponding to (a) and (d) rank histogram corresponding to (b).

perfectly reliable system and is given by $\Delta_O = \frac{n}{n+1}$.

Candille & Talagrand (2005) mentioned that "A value of Δ significantly larger than one is a proof of unreliability".

Candille *et al.* (2007) proposed another indicator derived from the reduced centered random variable (RCRV), itself proposed by Talagrand *et al.* (1999). This indicator is defined as

$$y_t = \frac{x_t - \bar{x}_t}{\sqrt{\sigma_t^2 + \sigma_N^2}} \quad t = 1, \ldots, N$$

where x_t is the observation at time t, $\bar{x}_t$ and σ_t^2 the corresponding mean and variance of the ensemble forecasts and σ_N^2 stands for the variance of the observations.

The indicator y_t can be used to study the bias and the dispersion of the forecasts. To do so, one must calculate the sample mean $\bar{y}$ and standard deviation s_y of y_t, $t = 1, \ldots, N$. If the value of $\bar{y}$ is different from zero, the forecasts may be biased. Additionally, if s_y is smaller (greater) than one, the forecasts may be underdispersive (overdispersive).

The confidence interval reliability diagram

The reliability diagram (e.g. Stanski *et al.* 1989; Wilks 1995) is a plot of the observed relative frequency of events predicted with a certain occurrence probability against this occurrence probability. The confidence interval reliability diagram is a variant of this concept which aims at assessing the quality of the confidence interval that can be calculated using the predictive distribution.

In the present study, 10 confidence intervals have been calculated with nominal confidence level of 50–95%, with an increment of 5% for each emitted forecast. Then, for each forecast and for each confidence interval, it was established whether or not each confidence interval covered the observation. This was repeated for all the forecast–observation pairs in the archive and the computed coverage (in percentage) of each confidence interval was plotted against the corresponding nominal confidence level. If the predictive distribution is well calibrated, these two values should be close. For example, for the 95% confidence interval, 95% of the N intervals should cover the observed value. Therefore, a graph of the observed percentages plotted against the values of the nominal confidence level should take the form of a linear relationship with a slope of 1 and an ordinate at the origin of 0.

Table 1 | Numerical measures of the forecasts' quality for all basins, along with their standard deviations estimated with the jackknife technique (50-member ensembles)

Basin	$\overline{\text{CRPS}}$ (mm)	MAE (mm)
Leaf	0.234 (0.009)	0.281 (0.010)
Serein	0.047 (0.002)	0.057 (0.002)
Volpajola	0.268 (0.009)	0.311 (0.009)

RESULTS AND DISCUSSION

CRPS and AE

The mean CRPS and AE values obtained for the three watersheds under investigation are shown in Table 1. The mean AE is denoted MAE and can be compared to the mean CRPS. Here, the MAE was calculated using the daily ensemble mean as the point forecast. As mentioned before, a perfect ensemble forecast would achieve a $\overline{\text{CRPS}}$ of zero. Therefore, it must be minimised, like the MAE.

For all three watersheds, the mean values of the $\overline{\text{CRPS}}$ are lower than the MAE, even if the standard deviation is considered. It indicates that the ensemble forecasts perform better overall than the point forecasts. However, both the $\overline{\text{CRPS}}$ and the MAE being proportional to the magnitude of the streamflow, it is not possible to compare one basin with another.

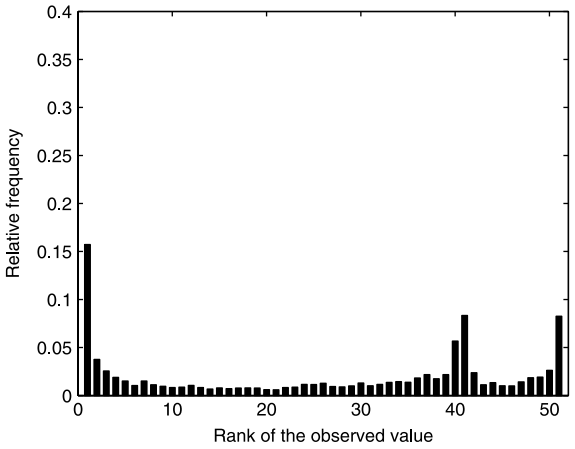

Figure 5 | Rank histogram for the Leaf River streamflow forecasts, 50 members.

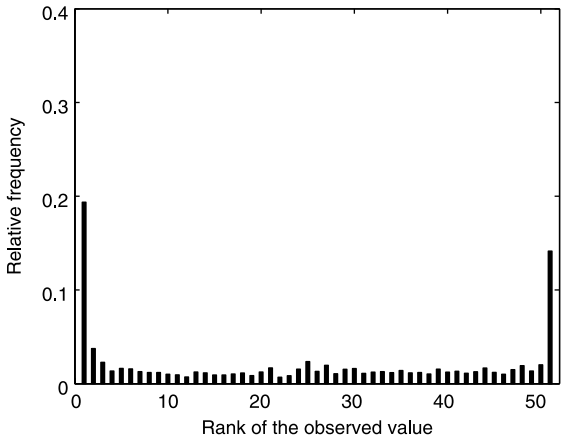

Figure 6 | Rank histogram for the Serein River streamflow forecasts, 50 members.

Rank histogram

Figure 5 presents the rank histogram obtained for the Leaf River while Figures 6 and 7 present the rank histograms obtained for the Serein and Volpajola Rivers, respectively. All ensembles have 50 members.

All rank histograms are U-shaped, which points out a possible under-dispersion of the predictive distributions. In the three cases, the first rank is the one with the highest relative frequency, meaning that the observation is usually lower than all the ensemble members (underestimation). The Volpajola River is particularly under-dispersed since the relative frequencies of the first and last ranks are, respectively, 0.4 and 0.3 compared with 0.16 and 0.12 for the Leaf River or 0.19 and 0.14 for the Serein River. In addition, there appears to be a small bias in the forecasts for the Leaf basin since the histogram is asymmetric. The ranks 40 and 41 have relative frequencies superior to the adjacent ranks.

Table 2 | Indicators associated to the rank histogram

Basin	δ (flatness)	$\bar{y}$ (bias)	s_y (dispersion)
Leaf	141	−0.002	0.23
Serein	257	−0.007	0.16
Volpajola	517	−0.016	0.23

Table 2 presents the values obtained for the various indicators mentioned in the methodology section. It can be seen that the flatness ratio (δ) is much greater than 1, which is coherent with the non-uniform shape of the rank histograms. The negative bias $\bar{y}$ for all basins are all very small and the values of s_y are smaller than 1 for all watersheds, which also indicate under-dispersion. According to this indicator, the forecasts for Leaf would be the most under-dispersed while Serein would be the less under-dispersed. This does not seem to agree with the conclusions that can be drawn from the rank histograms, which indicate that the forecasts for Volpajola seem more under-dispersive than the forecasts for Leaf.

Confidence interval reliability diagrams

Figures 8–10 show the confidence interval reliability diagrams obtained with the ensemble forecasts for each basin. They confirm the under-dispersion of the distributions that was pointed out by the rank histograms as well

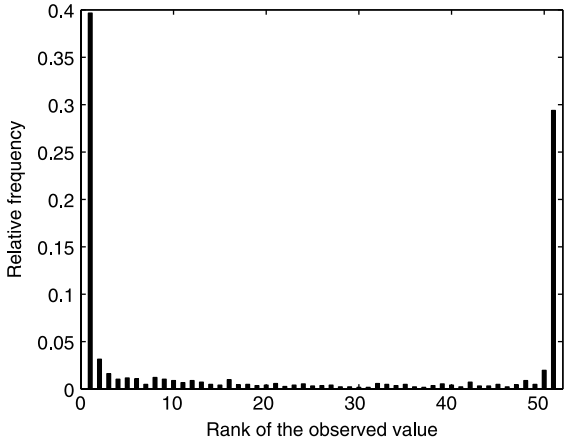

Figure 7 | Rank histogram for the Volpajola River streamflow forecasts, 50 members.

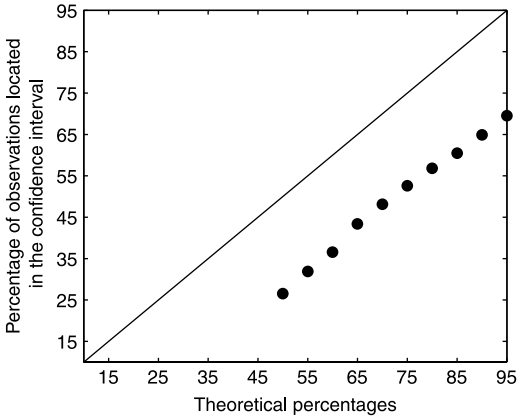

Figure 8 | Confidence interval reliability diagram for Leaf River streamflow forecasts, 50 members.

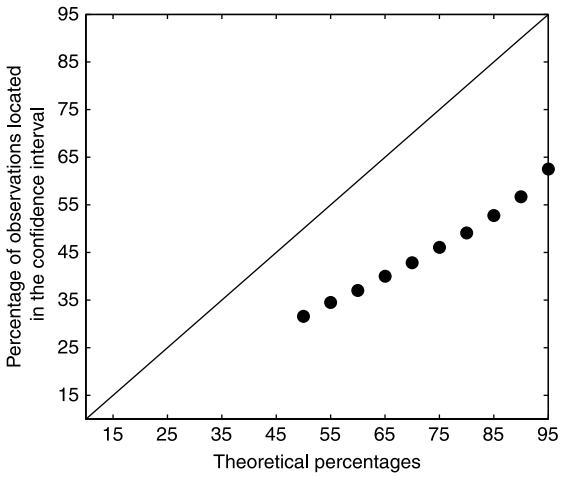

Figure 9 | Confidence interval reliability diagram for Serein River streamflow forecasts, 50 members.

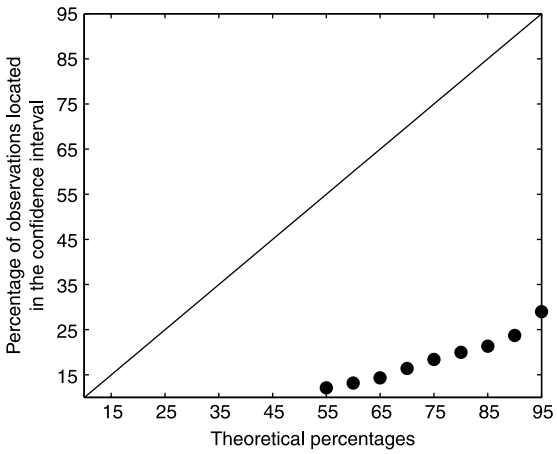

Figure 10 | Confidence interval reliability diagram for Volpajola River streamflow forecasts, 50 members.

as a possible bias. The percentages of observations located inside the confidence interval are consistently lower than their nominal value, for all watersheds and for all confidence levels. In accordance with the rank histograms, distributions for the Volpajola River are the most under-dispersed, the percentage of observations located in the 95% confidence interval being only 44%.

CONCLUSION

The main objective of this paper was to assess the quality of stacked neural-network-based hydrological ensemble forecasts using tools that have primarily been developed in the meteorological and statistical communities. Daily ensemble streamflow forecasts for three watersheds located in France and the United States have been evaluated. Two graphical tools, the rank histogram and the confidence interval reliability diagram, were used, in addition to a numerical criterion, the CRPS. This criterion allows us to compare ensemble forecasts to point forecasts since the CRPS reduces to the MAE for a point forecast.

Both the rank histograms and the confidence interval reliability diagrams have shown that the ensemble forecasts produced for all three catchments are under-dispersed and may be biased, a condition that has also been encountered with other ensemble-generating systems (e.g. Buizza 1997; Hamill 2001; Toth *et al.* 2001; Gneiting *et al.* 2005). It indicates that our ensemble generation method does not cover all the sources of uncertainty and may be improved. The random initialisation of the neural networks' weights and biases mainly accounts for the uncertainty linked to the optimisation of the model's parameters. The Volpajola River is the most affected by under-dispersion. Under-dispersion, or equivalently underestimation of forecast uncertainty, may lead to wrong decisions for water resources management. The results obtained with the catchments under study showed that ensemble forecasts produced by neural networks underestimate uncertainty. Clearly, there is a need for a methodology that would enhance the spread of the ensembles. This could be achieved by improving our neural networks model and/or the way the ensemble was constructed. If this cannot be done, a statistical post-processing of the forecasts can be applied (e.g. Fischhoff 1982; Stewart & Reagan-Cirincione 1991; Goodwin & Lawton 2003; Gneiting *et al.* 2005; Raftery *et al.* 2005; Stensrud & Yussouf 2007). One could also work on improving the selection of the ensemble members (e.g. Zhou *et al.* 2002; Granitto *et al.* 2005). In further research, attention will be focused on ensemble dispersion. Nevertheless, for all the basins, the mean CRPS is lower than the mean AE, which leads to the conclusion that ensemble streamflow forecasts, even flawed, outperform point forecasts for the three basins under study. Moreover, they provide the users and decision-makers with useful information regarding the forecast's uncertainty.

REFERENCES

Anctil, F. & Lauzon, N. 2004 Generalisation for neural networks through data sampling and training procedures, with applications to streamflow predictions. *Hydrol. Earth Syst. Sci.* **8** (5), 940–958.

Breiman, L. 2000 Randomizing outputs to increase prediction accuracy. *Mach. Learn.* **40**, 229–242.

Brier, G. W. 1950 Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.* **78**, 1–3.

Buizza, R. 1997 Potential forecast skill of ensemble prediction and spread and skill distributions of the ECMWF ensemble prediction system. *Mon. Weather Rev.* **133**, 1076–1097.

Candille, G., Côté, C., Houtekamer, P. L. & Pellerin, G. 2007 Verification of an ensemble prediction system against observations. *Mon. Weather Rev.* **135**, 2688–2698.

Candille, G. & Talagrand, O. 2005 Evaluation of probabilistic prediction systems for a scalar variable. *Q. J. R. Meteorol. Soc.* **131** (609), 2131–2150. Part A.

Efron, B. & Tibshirani, R. J. 1998 *An Introduction to the Bootstrap. Monographs on Statistics and Applied Probability*. Vol. 57. Chapman and Hall/CRC. Boca Raton, FL.

Fischhoff, B. 1982 *Debiasing, Judgment under Uncertainty: Heuristics and Biases* (ed. D. Kahneman, P. Slovic & A. Tversky). Cambridge University Press, Cambridge.

Foresee, F. D. & Hagan, M. T. 1997 Gauss-Newton approximation to Bayesian learning. In *Proceedings, 1997 IEEE International Conference on Neural Networks, Houston, TX*. Vol. 3, pp. 1930–1935. IEEE Press, Hillsdale, New Jersey.

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

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

Good, I. J. 1952 Rational decisions. *J. R. Stat. Soc. B* **14**, 107–114.

Goodwin, P. & Lawton, R. 2003 Debiasing forecasts:how useful is the unbiasedness test? *Int. J. Forecast.* **19**, 467–475.

Granitto, P. M., Verdes, P. F. & Ceccatto, H. A. 2005 Neural network ensembles: evaluation of aggregation algorithms. *Artif. Intell.* **163**, 139–162.

Hamill, T. M. 2001 Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Weather Rev.* **129**, 550–560.

Hansen, K. H. & Salamon, P. 1990 Neural network ensembles. *IEEE Trans. Pattern Anal. Mach. Intell.* **12** (10), 993–1001.

Hersbach, H. 2000 Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather Forecast.* **15** (5), 559–570.

Klemeš, V. 1986 Operational testing of hydrological simulation models. *Hydrol. Sci. J.* **31** (1), 13–24.

Kohonen, T. 1990 The self-organizing map. *Proc. IEEE* **79**, 1464–1480.

Krzysztofowicz, R. 2001 The case for probabilistic forecasting in hydrology. *J. Hydrol.* **249** (1), 2–9.

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

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

Rosenblatt, F. 1958 The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol. Rev.* **65**, 386–408.

Roulston, M. S. & Smith, L. 2002 Evaluating probabilistic forecasts using information theory. *Mon. Weather Rev.* **130** (6), 1653–1660.

Saetra, O., Hersbach, H., Bidlot, J.-R. & Richardson, D. S. 2004 Effects of observations errors on the statistics for ensembles spread and reliability. *Mon. Weather Rev.* **132**, 1487–1501.

Schaake, J. C., Hamill, T. M., Buizza, R. & Clark, M. 2007 The hydrological ensemble prediction experiment. *Bull. Am. Meteorol. Soc.* **88** (10), 1541–1547.

Stanski, H. R., Wilson, L. J. & Burrows, W. R. 1989 Survey of common verification methods in meteorology. *WMO World Weather Watch Tech. Report* 8, WMO TD 358.

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

Stewart, T. R. & Reagan-Cirincione, P. 1991 Coefficients for debiasing forecasts. *Mon. Weather Rev.* **119**, 2047–2051.

Talagrand, O., Vautard, R. & Strauss, B. 1999 Evaluation of probabilistic prediction systems. In *Proceedings, ECMWF Workshop on Predictability, Shinfield Park, Reading, Berkshire ECMWF*, pp. 1–25. Shinfield Park, Reading.

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

Weber, F., Perreault, L. & Fortin, V. 2006*a* Measuring the performance of hydrological forecasts for hydropower production at BC Hydro and Hydro-Québec. In: *Proceeding of the 18th Conference on Climate Variability and Change, AMS, Atlanta, 30 January–2 February* AMS, Boston, 8.5.

Weber, F., Perreault, L., Fortin, V. & Gaudet, J. 2006*b* Performance measures for probabilistic hydrologic forecasts used at BC-Hydro and Hydro-Québec. In: *EGU Conference, April 2nd–7th, EGU, Katienburg-Lindau, Germany, SRef-ID: 1607-7962/gra/EGU06-A-09273*.

Wilks, D. S. 1995 *Statistical Methods in the Atmospheric Sciences*. Academic Press, San Diego.

Wolpert, D. H. 1992 Stacked generalization. *Neural Netw.* **5**, 241–259.

Zhang, G. P. 2007 A neural network ensemble method with jittered training data for time series forecasting. *Inf. Sci.* **177**, 5329–5346.

Zhou, Z. H., Wu, J. & Tang, W. 2002 Ensembling neural networks: many could be better than all. *Artif. Intell.* **137**, 239–263.

First received 30 April 2008; accepted in revised form 11 September 2008